Abstract
Cloud computing is a well-known dynamic platform for providing on-demand services of resources such as software, infrastructure, platform over the internet. It delivers flexible and time-to-value performance all over the world. It is the most emerging field from business perspective and its significance in other fields. The important aspects of cloud computing are its accurate implementation, its performance and availability of resources to maintain QoS. The one way to develop cloud computing system is with the concept of queueing theory. Much work on queueing theory-based cloud computing has been reported in the literature. In literature, various queueing models are implemented and performance based on various parameters considered, are evaluated. This paper does the review of all such model implementations by various researchers and scientists. Critical analysis of the survey of literature is done and the previous work done can be compared in order to improve the efficiency of the cloud system by evaluating the best proposal for its development.

Keywords: Cloud computing; queuing models; waiting time; impatient users; throughput rate

1. INTRODUCTION
In the early time with the advancement of the technology man started using utility services in day to day life for the fulfilling of the daily routine. They are so frequently used that they must be available at all the times. In the 21st century computing services will be readily available on demand, like other utility services. Cloud computing is the practice of using a network of remote servers hosted on the internet to store, manage, and process data, rather than a local server or a personal computer. Among the larger companies in this space, Amazon was the first with its Amazon Web Services (AWS) division. The total investment in cloud computing is estimated to be $110 billion and in cloud computing is estimated to be $110 billion and increasing at the rate of 28% per year. Cloud computing has a service oriented architecture in which services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), where equipment such as hardware, storage, servers, and networking components are made accessible over the Internet; Platform-as-a-Service (PaaS), which includes computing platforms- hardware with operating systems, virtualized servers, and the like; and Software-as--a-Service (SaaS), which includes software applications and other hosted services [1].

1.1 Queuing Theory
As given in [2] when either units requiring services i.e. customers wait for service or the service facilities stand idle and wait for users, a queue is formed. Some users wait when the number of service facilities are less and number of customers or users requesting for service exceeds that number. Some service facilities remain idle when number of users requesting service is less than the total number of service facilities.

There are few key points about queueing theory as given under: -
- Queuing theory is applied to various operational scenarios in which imperfect matching between service facilities and users is occurred due to system’s inability to predict the arrival and service time of customers accurately.
• Queuing theory consists of mathematical models used for modelling various real-life situations.
• It is the study of waiting lines or queues, mathematically dealing with the prediction of average delay faced by the customers in the queue, estimation of average number of customers in the queue.
• Its main objective is to provide best level of service at minimum possible cost.

The paper outlines following sections - section II covers the related work, section III discusses the queuing models briefly, section IV provides a detailed comparison with the evaluation of these models based on certain parameters. Finally, section V presents the overall conclusion and future scope based on the previous sections.

2. RELATED WORK

Many studies and surveys have been conducted highlighting various aspects of IoT protocols regarding their suitability and workability under constrained environments and how they led the communication flow seamlessly without any encumbrance. In [3] a single web server was model using M/G/1/K*PS queue. Authors have assumed the arrival of customers according to Poisson process. At most k servers can be processed; after kth customer, if any customer arrives, it is blocked. The service time of the server is based on general distribution and processes are shared in its operation. This shows that M/M/1/K is simpler than general M/G/1/K queue. The authors have proposed that their approach and model used by them is better than other models because of its simplicity. They have estimated model parameters like average service time $\bar{x}$ and number of service places $K$. They generated a random workload by setting maximum TCP-connections the server could process and checked for many data sets, the average response time for virtual customers. The average response time is estimated several times and this estimation is put into normal probability function. The values of function are multiplied and with the varying values of $\bar{x}$and $K$, they have tried to maximize the results. To obtain the varying values of these, they have used brute force optimization algorithm. This model [3] is advantageous in the sense of its simplicity but it is not valid for overloaded work region. In [4] authors have presented an improved model of [3] which is valid for overload region. They have used a two-state Markov Modulated Poisson Process (MMPP) for simulating burst traffic with random peaks. MMPP comprise of two states in simple Markov chain. The Markov chain outputs intensity $\lambda_1$ for incoming customers in state one and intensity $\lambda_2$ ($\lambda_2 >> \lambda_1$) in state second and the Markov chain changes in these states with intensities $r_1$ and $r_2$. The model used is same as in [3] so the system is MMPP/G/1/K*PS. The same method is used for obtaining the mean service time and number of service places. This paper presents a more realistic model than in [3] but there is no comparison of real server traffic with the traffic generated by MMPP. In [5] the authors have given a closed queuing network for modelling a 3-tier web-server. A web-server consist of three layers- a database server layer which stores the important data and records, an application server layer for implementing business logic and a web server for handling customer requests. They modelled each of these layers as a service with separate queue attached in series so that the customers being served by server are advanced to application servers. They suggested a simplified scenario like assuming the exponential service time which is workload independent. Queuing network parameters such as mean delay, throughput, etc are computed with the help of mean-value analysis. They performed tests, measured the parameters and compared them to the mathematical deduced ones and found that the error between the empirical & theoretical values is minute which suggests that the model could represent real setting precisely but this also lacks the enquiry about real traffic and service. To generate workload, they used a TPC-W but there are no justifications about that. In [6] M/G/s queueing model is used for modelling the cloud centre with arrivals of single tasks and with a task buffer of infinite capacity. The authors have analysed the performance based on the combination of approximate Markov chain model and transform-based analytical model which enabled them to obtain complete probability distribution of request, response time and number of tasks in system. In [7] M/M/s model is used for two servers which proved to increase the performance over using single server by minimizing the waiting time and queue length guarantying the QoS demands of the Cloud Computing User’s (CCU) jobs and making utmost profits.
for Cloud Computing Service Provider (CCSP). [8] has shown how a single server and multiple server system varies on the basis of the factors such as time delay and throughput in cloud computing environment. In [9] [(M/G/1 \mid \infty /GD MODEL)] queuing model is used for modelling cloud centre with arrivals of single tasks and with a task buffer of infinite capacity. An analytical model is used for measuring the performance of queuing system and with its help important performance factors are evaluated like mean number of tasks in the system. In [10] (GE/G/m/k) queuing model is used to model cloud centre consisting of task arrivals as GE distribution, huge number of single servers, a finite capacity task buffer and a general service time for requests. Various performance factors are calculated such as probability of immediate service, average of response time, blocking probability and average number of tasks in the system. In [11] M/M/m queuing system consisting of task request buffer of infinite capacity and multiple task arrivals, is used for modelling cloud centre. This model is composed of homogeneous servers and it evaluated the aspects related to QoS.

3. QUEUEING THEORY IN CLOUD COMPUTING

The queuing system in cloud service system consist of input source i.e. source of requests, queuing process which has waiting requests in the queue to be served, service process which comprises of servers to process the various requests in the queue. There can be finite capacity queuing systems or infinite capacity queuing systems and the various characteristics of queuing system are shown below [12].

4. QUEUING MODELS

This section compares and analyses the different queuing models and their study in order to enhance the clarity of usage of these models along with their specifications. Queuing models helps in estimating the performance of service systems when there is unpredictability in service times and arrival times [14].

4.1 M/M/1 MODEL

This model is derived on the basis of certain assumptions about queuing system. It comprises of exponential distribution of inter-arrival time or Poisson’s distribution of arrivals with mean rate ‘\( \lambda \)’. The inter-arrival times are independently, identically and exponentially distributed in parameter \( \lambda \).
This type of queuing system consists of only one service unit and the service times are independently, identically and exponentially distributed with parameter ‘\( \mu \)’. The capacity of the system is infinite and queue of requests are served in FIFO (First In First Out) fashion [13][14]. Table I compares various studies on M/M/1, M/M/2 models extensively with respective parameters deployed for the cloud service system based on the nature of the application. For achieving efficiency in systems for providing better QoS, M/M/1 is considered a good choice. But request/response takes more time due to single service unit serving and long waiting line [15][16].

**TABLE I**

<table>
<thead>
<tr>
<th>#</th>
<th>AUTHOR</th>
<th>TITLE</th>
<th>RESULTS</th>
<th>PARAMETERS</th>
<th>DRAWBACKS</th>
</tr>
</thead>
</table>
| 1 | K.RuthEvangelin and V.Vidhya | Performance Measures of Queuing Models Using Cloud Computing | • This paper compared M/M/1 and M/M/c models for number of servers 1,2 & 3 and observed the factors C_s and E_s.  
  • It was observed that for M/M/1, there was more C_s for more E_s as compared to that for M/M/c. | C_s (waiting time) and E_s (number of customers at the system). | No improvement scope is given. |
| 2 | Bashir Yusuf Bichi, TuncayErcan | An Efficient Queuing Model for Resource Sharing in Cloud Computing | • M/M/1 and M/M/c queuing models are compared and various factors for both the models are analysed.  
  • It was observed that MQL and waiting time are lesser for M/M/c than that in case of M/M/1 and throughput is obtained to be more for M/M/c. | Mean queue length, waiting time (time delay), throughput. | Behaviour of customers is not considered. |
| 3 | SuneetaMohanty, Prasant Kumar Pattnaik and Ganga BishnuMund | A Comparative Approach to Reduce the Waiting Time Using Queuing Theory in Cloud Computing Environment | • Comparative study of M/M/c and Erlang-C is done in order to reduce waiting time by increasing number of servers.  
  • Erlang gives better results to reduce waiting time when number of servers is increased than M/M/c model.  
  • Average waiting time is calculated for both the models taking n (no. of servers) =1,2,3.  
  • It introduces probability of delay as the function of n (no. of servers) and x (traffic intensity) which is the function of Erlang B function. | Average waiting time, number of servers, probability of delay, traffic intensity. | The pitfalls of Erlang methods are not contemplated. |
4. N.N. Bharkad, Dr. M.H. Durge

The application of queue theory in cloud computing to reduce the waiting time

- Dynamic behaviour of the system with infinite servers analysed by finding effective measures. Various measures are evaluated for M/M/1, M/M/2 and M/M/3.
- It is concluded that to reduce the waiting time of customer, increasing the number of service channels is needed.

| Average number of units in the system, Average number of units in the queue (queue length), average waiting time in the system, av. waiting time in the queue, variance of queue length, utilization factor. | This work doesn’t include both waiting time as well as service cost. |

5. Yuxiang Shi et al.

An Energy-Efficient Scheme for Cloud Resource Provisioning Based on Cloud Sim

- Energy-efficient scheme for cloud resource provisioning is proposed using M/M/1 queuing theory predicting model.
- Linear predicting method and flat period reservation-reduced method are used to get information of resource utilization.
- Better response time and less energy consumption is there with the help of proposed model.

| Response time and energy consumption, average queue length, resource allocation. | Only one server is being worked on whereas several servers if examined could give more insight to the results. |

4.2 M/M/C MODEL

This model is based on similar assumptions as that of M/M/1 model except that it comprises of multiple servers in parallel equivalent to C. It is assumed that customers arrive according to a Poisson process at an average rate of ‘λ’ customers per unit time. The requests are served in FCFS (First Come First Serve) basis at any service unit. Servers are identical and customers are served with an average rate of ‘μ’ customers per unit time. [16][17] For large number of service units this model may be deployed for measuring the performance of the service system whereas for one or two service units M/M/1 or M/M/2 may be deployed. Table II discusses the effect of modelling an application of cloud service system as M/M/C model on its performance based on the various performance measures. These can be critically examined in order to analyse the model and its efficiency properly [18][19]. Table II compares various studies on M/M/C model in respect to cloud service system. Implementing M/M/C model obviously increased the throughput and lessen the response time but practically infinite buffer capacity might not be possible in applications involving cloud servers. Performance measures have shown better results for M/M/C model as compared to that of M/M/1 or M/M/2 [20].

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>TITLE</th>
<th>RESULTS</th>
<th>PARAMETERS</th>
<th>DRAWBACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 N.N. Bharkad, Dr. M.H. Durge</td>
<td>The application of queue theory in cloud computing to reduce the waiting time</td>
<td>- Dynamic behaviour of the system with infinite servers analysed by finding effective measures. Various measures are evaluated for M/M/1, M/M/2 and M/M/3. - It is concluded that to reduce the waiting time of customer, increasing the number of service channels is needed.</td>
<td>Average number of units in the system, Average number of units in the queue (queue length), average waiting time in the system, av. waiting time in the queue, variance of queue length, utilization factor.</td>
<td>This work doesn’t include both waiting time as well as service cost.</td>
</tr>
<tr>
<td>5 Yuxiang Shi et al.</td>
<td>An Energy-Efficient Scheme for Cloud Resource Provisioning Based on Cloud Sim</td>
<td>- Energy-efficient scheme for cloud resource provisioning is proposed using M/M/1 queuing theory predicting model. - Linear predicting method and flat period reservation-reduced method are used to get information of resource utilization. - Better response time and less energy consumption is there with the help of proposed model.</td>
<td>Response time and energy consumption, average queue length, resource allocation.</td>
<td>Only one server is being worked on whereas several servers if examined could give more insight to the results.</td>
</tr>
</tbody>
</table>

TABLE II
M/M/C MODEL

<table>
<thead>
<tr>
<th>#</th>
<th>AUTHOR</th>
<th>TITLE</th>
<th>RESULTS</th>
<th>PARAMETERS</th>
<th>DRAWBACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>N.N. Bharkad, Dr. M.H. Durge</td>
<td>The application of queue theory in cloud computing to reduce the waiting time</td>
<td>- Dynamic behaviour of the system with infinite servers analysed by finding effective measures. Various measures are evaluated for M/M/1, M/M/2 and M/M/3. - It is concluded that to reduce the waiting time of customer, increasing the number of service channels is needed.</td>
<td>Average number of units in the system, Average number of units in the queue (queue length), average waiting time in the system, av. waiting time in the queue, variance of queue length, utilization factor.</td>
<td>This work doesn’t include both waiting time as well as service cost.</td>
</tr>
<tr>
<td>5</td>
<td>Yuxiang Shi et al.</td>
<td>An Energy-Efficient Scheme for Cloud Resource Provisioning Based on Cloud Sim</td>
<td>- Energy-efficient scheme for cloud resource provisioning is proposed using M/M/1 queuing theory predicting model. - Linear predicting method and flat period reservation-reduced method are used to get information of resource utilization. - Better response time and less energy consumption is there with the help of proposed model.</td>
<td>Response time and energy consumption, average queue length, resource allocation.</td>
<td>Only one server is being worked on whereas several servers if examined could give more insight to the results.</td>
</tr>
</tbody>
</table>
| 1 | Yonal Kirs et al. | Analytical Modelling and Performance Analysis for Cloud Computing Using Queuing System | • M/M/c queuing model as performance model is used with MRM (Markov Reward Model). MRM is used as availability model for c no. of servers & infinite buffer capacity.  
• The performance measures are passed as reward rates to the availability model.  
• System performance degradation becomes more evident as repair rate decreases. MQL increases in case of quick break down and long repair time. | MQL (Mean Queue Length), Vs \( \lambda \) (requests/sec) is shown. | This work didn’t introspect the heterogeneous nature of servers. |
| 2 | Dan Liao et al. | Energy and Performance Management in Large Data Centers: A Queuing Theory Perspective | • M/M/n+m_1+m_2 model is used for analytical modelling where n+m_1+m_2 is the group of servers.  
• Power consumption can be minimized by switching off/on certain group of servers.  
• It determines activation thresholds for server. Power performance trade-off and problem of dynamic scheduling scheme. | Mean service rate, energy efficiency, waiting time, number of servers. | Not dynamic in nature. |
| 3 | Lizheng Guo et al. | Dynamic Performance Optimization for Cloud Computing Using M/M/m Queueing System | • M/M/m model is used with infinite task buffer capacity to design an optimization function and a synthetical optimization method for cloud computing. Proposed optimization method improves the performance of data centers.  
• QoS can be gained for the given service rate, customer’s arrival rate and number of servers.  
• Performance of optimization strategy is compared with that of FIFO and SSF priority queues. | Average waiting time, Utilization, average queue length, amount of service customer. | Cost analysis on the basis of optimization methods are not evaluated. |
| 4 | G. Vijaya Lakshmi et al. | A Queuing Model to Improve Quality Service by Reducing Waiting Time in Cloud computing | • (M/M/c): (\( \infty \)/FIFO) model is proposed for multiple servers to reduce mean waiting time, decreasing queue length and improving QoS in cloud computing environment.  
• The results show that when there are more number of servers, waiting time decreases & therefore, QoS can be achieved otherwise CCU has to wait till its service or may enter balking or reneging state. | Mean queue length, waiting time. | There’s no concept of impatient customers. |
Jordi Vila plana et al.  

A Queuing Theory Model for Cloud Computing

- Computer service QoS model for the cloud architecture is presented.
- The cloud architecture is modelled with an open Jackson network of M/M/m and M/M/1 interconnected servers.
- The utilization rate of queue decreases and the response time stabilizes to the same value as the service time.

Response time, utilization, arrival rate, file size, number of requests, bandwidth.
Open Jackson network not truly imitate the cloud service system.

4.3 M/M/C/N MODEL

In this model, there are multiple servers in parallel to provide service to customers’ requests. It is assumed that only one queue is formed and requests are served on a first-come, first serve basis by any of the servers. The inter-arrival times are distributed exponentially with parameter ‘\( \lambda \)’ and the service times are distributed exponentially with an average of ‘\( \mu \)’ customers per unit time [21] [22]. It has N number of requests capacity of the buffer. Also, if there are ‘N’ requests in the queueing system at any point of time, then two cases may arise i.e. if N<C, there will be no queue and if N>=C, then all the servers will be busy and a queue will be formed. Some example application areas of this model are:

- Counters in library to address the service of issuing/returning books,
- Counters in telephone exchange to service the bill requests,
- Counters at the frontier to check the passports,
- Counters at tax consulting offices to receive requests concerning income and sales tax.

Table III compares the studies of M/M/C/N model in cloud service system and performance measures are evaluated based on this model by various authors in their study. So far, this model provides energy efficiency, dynamicity and better performance efficiency as well. This model serves a good purpose in maintaining QoS [23].

<table>
<thead>
<tr>
<th>#</th>
<th>AUTHOR</th>
<th>TITLE</th>
<th>RESULTS</th>
<th>PARAMETERS</th>
<th>DRAWBACKS</th>
</tr>
</thead>
</table>
| 1 | Yi-Ju Chiang, Yeh-Chieh Ouyang, Ching-Hsien (Robert) Hsu | Performance and Cost-Effectiveness Analyses for cloud Services Based on Rejected and Impatient Users | • M/M/R/K cloud server with finite capacity buffer and R identical servers is used.  
• Multi-server queuing system with impatient users, a cost model is developed and the cost-effective policy (CEA) is presented to solve constrained optimization problems.  
• Dynamic system controls are used to alleviate system losses. | Queue length, waiting time, Loss probability, requests arrival rate, service rates, throughput, cost. | But retention, penalty as such not considered. |
| 2 | Yi-Ju Chiang and Yen-Chieh Ouyang | Profit Optimization in SLA-Aware Cloud Services with a Finite Capacity Queuing Model | • A cloud server farm provided with finite capacity is modelled as an M/M/R/K queuing system. Revenue losses are estimated according to the system controls and impatient customer behaviours.  
• A profit function is developed in which both the system blocking loss and the user abandonment loss are evaluated in total revenue.  
• A trade-off between system performances and reducing operating costs is conducted.  
• The effects of system capacity control and utilization on various performance measures are evaluated. | Operational costs, waiting time, loss probability, arrival rate. | It doesn’t assess the heterogeneity of servers and factors such as system blocking probability. |
| 3 | Wendy Ellens et al. | Performance of Cloud Computing Centers with Multiple Priority Classes | • Problem of resource provisioning in cloud computing is investigated.  
• M/M/c/c queuing system with different priority cases to support decision making for resource allocation is used to model the cloud center.  
• Performance is measured by analysing blocking probability for different customer classes. | Number of customers, rejection probability, number of servers, arrival rate. | An appropriate priority scheduler needs to be embedded within the suggested system. |
| 4 | V. Goswami, S. S. Patra, G. B. Mund | Performance Analysis of Cloud with Queue Dependent Virtual Machines | • Analytical queuing-based model for performance management is given. Web applications are modelled as queues and virtual machines are modelled as service centers.  
• A finite buffer queueing system with queue dependent multi-heterogeneous VMs server is considered.  
• The probability P(j) of server j being busy increases with arrival rate \( \lambda \) but decreases by increasing the number of servers and average no. of customers \( (L) \) in the system increases monotonically with the increase of arrival rate. | Average number of customers, Inter-arrival rate, number of servers, probability of server being busy. | An effective method may be imparted based on the analytical results. |
4.4 M/G/S MODEL

In this model, the queuing system involves multiple servers ‘S’ in parallel. This model is an extension of M/M/C or M/G/1 queue where service times are exponentially distributed and a single server system respectively. The inter-arrival times in this model is exponentially distributed with parameter ‘λ’ and service times follow general probability distribution instead of an exponential one. It is assumed that length of inter arrival times and service periods are independent statistically [24]. This model can be deployed in systems comprising of self-service mechanisms such as restaurants with self-service filling and refilling activity, traffic light systems, etc. In table IV various researches carried out by different authors taking into consideration M/G/S and its extensions, deployed in a cloud service system is shown. It can be seen that M/G/S queue enables increase in service rates and allows more utilization in same amount of time. But in general cases requests and customers may not arrive by general probability instead there has been more cases of requests being random in nature. In comparison to M/M/C, M/G/S is suitable for handling more traffic [25]. M/G/S doesn’t provide better analytical insights with its mathematical formulation than M/M/C plus it isn’t suited for numerical computations in some cases [26].

<table>
<thead>
<tr>
<th>#</th>
<th>AUTHOR</th>
<th>TITLE</th>
<th>RESULTS</th>
<th>PARAMETERS</th>
<th>DRAWBACKS</th>
</tr>
</thead>
</table>
| 1 | Mohammad ali Safvati | Analytical Review on Queuing Theory in Clouds Environments | • M/G/m/m+r model is proposed with FCFS queue fashion. M/G/m/m+r and M/M/1 models are compared.  
• Results obtained show that queue size and queue delay are less for M/G/m/m+r model.  
• And efficiency of this model is more. | Queue size, queue delay (packet waiting time), Laplace queue length, Laplace waiting time, mean response time. | Model proposed may not be deployed in real cloud systems. |
| 2 | R. Murugesan et al. | Resource Allocation in Cloud Computing with M/G/s Queueing System | • Resource allocation techniques to reduce the resource cost and minimize service response time. It uses M/G/s model with infinite capacity buffer.  
• It takes into account general service time in cloud center and Cloud Computing Network (CCN) is Open Jackson Queueing Network. It is seen that waiting time decreases for increased number of servers and length of queue decreases. It can be extended to G/M/s. | Response time, task, Blocking probability, probability of immediate services, mean number of tasks for QoS, waiting time, queue length and average number of servers. | It could be extended with the concept of cost estimation, power consumption and impatient behaviour of requests based on timing constraints. |
| 3 | Hamzeh Khazaei et al. | Performance Analysis of Cloud Computing Centers | • The system has finite buffer, the performance of the model is evaluated using a combination of transform based analytical model. | Mean number of requests, Blocking probability of the | General distribution may not emerge to the dynamicity of a |

International Journal of Pure and Applied Mathematics Special Issue
Many of the authors have worked on the similar parameters such as response time, waiting time, queue length and in few of the papers blocking probability have also been considered but there is a lot of scope of improvement. Factors like balking, reneging need to considered in depth so as to enhance the service levels and customer satisfaction. The efficiency thus can be determined in terms of revenue, number of customers registered, etc. QoS can be maintained with the amelioration in customer satisfaction.

5. CONCLUSION

The cloud system applications and requests require efficient and reliable modelling over constrained environments. This study highlights the characteristics and applicability of existing queuing models in the real-world applications. The results showed M/M/C as the most empowering model used primarily for simulating cloud environment. Although M/G/S offers, but lack behind on realistic basis. Designing a successful service system requires not only some system control factors but need to study all the factors [27]. However previous studies witnessed various failures to offer a realistic service
system which is effective in nature. To overcome these failures, a model needs to be proposed inclusive of all the relevant considerations. In comparison to M/M/1, M/M/C and M/G/S, M/M/C/N is good for power efficient service systems. Consequently, M/M/C/N surpasses others through its effectiveness in combating power consumption and cost while maintaining QoS and SLA. Although M/G/m/m queues provide support for easy analyses, wide range of network topologies making them suitable for scalable IOT applications can be deployed to attain better results. These models if amalgamated together provides a platform to be implemented in queuing networks where multiple queues are deployed with multiply service stations [29]. However, as suggested, in future direction, a model needs to be recommended considering all the factors affecting the service system embedded with buffers which are dynamic in nature. This will privilege the service systems to be analysed more precisely and effectively further improvements can be done in any curtailed environment. The new model thus formulated may be practically deployed to check their efficacy in providing reliable services in constrained service systems.

6. REFERENCES


