Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 5, April-2019, ISSN 2230-7540

An Attempt to Understand the Queueing-Inventory System and the Consideration of M/M/1/1 Queueing-Inventory Model with Retrial of Unsatisfied Customers

Devendra Kumar Pandey*

Professor & Director, Unique Institute of Management & Technology, Ghaziabad, India

Abstract – Research on queueing systems with inventory control has captured much attention of researchers in the course of the last decades. The proposed issue is comprehended through lining theory for a solitary item. Characterization of queueing model is also given in the exhibited article. Further, extraordinary queueing models are elaborated clearly for the correct understanding of the models. In this case, transitional probabilities are calculated in steady state. This paper thinks about a M/M/1/1 queueing-inventory framework with retrial of unsatisfied customers. Arrivals taking place when server is busy, continue to an orbit of infinite capacity. From the orbit, the head of the queue alone resigns to access the server. An optimization issue is investigated numerically.

- x -

I. INTRODUCTION

In this framework, customers arrive at the service facility one by one and require service. So as to finish the client service, an item from the inventory is required. A served client departs immediately from the framework and the on-hand inventory decreases by one right now of service finish. The inventory is provided by an outside provider. This framework is called a queueing-inventory framework. The queueing-inventory framework is not quite the same as the traditional queueing framework because the attached inventory impacts the service. In the event that there is no inventory on hand, the service will be interfered. Also, it is unique in relation to the traditional inventory management because the inventory is expended at the serving rate rather than the customers' arrival rate when there are customers queued up for service.

Queueing theory is a well-created branch of applied probability theory. Historically, the subject of queueing theory has been grown largely in the. setting of phone traffic building. In the course of recent decades, steady advancement has been made towards illuminating increasingly troublesome and realistic queueing models.

Queueing theory is a branch of mathematics that reviews and models the act of waiting in lines. This paper will take a concise investigate the formulation of lining theory along with examples of the models and applications of their utilization. The goal of the paper is to furnish the reader with enough background so as to appropriately model a basic lining framework into one of the categories we will take a gander at, when conceivable. Also, the reader should start to understand the basic ideas of how to decide helpful information, for example, average waiting times from a particular lining framework.

A queueing model is usually characterized as far as three characteristics- - the info procedure, the service mechanism and the queue discipline. The info procedure depicts the arrangement of solicitations for service. Often the info procedure is indicated as far as the distribution of the time allotments between consecutive client arrival instants. The service mechanism is the category that incorporates such characteristics as the number 01' servers and the time spans that customers hold the servers. The queue discipline deals with the rule by which ' customers are taken for service.

For the single server queue a busy period is the time interval amid which the server is consistently busy i.e. it is the time allotment from the instant the (previously idle) server is seized until it next becomes idle. The time between the starting points of two consecutive busy periods is called a busy cycle. The actual waiting time in the queue of a client is characterized as the time between the snapshot of his arrival and the minute at which his service starts. The virtual waiting time at time t is the actual waiting time of a client in the event that he had arrived at time t.

The primary paper on lining theory, "The Theory of Probabilities and Telephone Conversations" was published in 1909 by A.K. Erlang, presently thought about the father of the field. His work with the Copenhagen Telephone Company is what provoked his initial foray into the field. He considered the issue of deciding what number of phone circuits were necessary to give telephone service that would keep customers from waiting too long for an available circuit. In building up an answer for this issue, he began to realize that the issue of limiting waiting time was applicable to many fields, and began building up the theory further.

Erlang's switchboard issue laid the path for current lining theory. The chapters on lining theory and its applications in the book "Operations Research: Applications and Algorithms" by Wayne L. Winston illustrates many expansions

Queueing models

The basic queueing model is appeared in figure 1. It can be utilized to model, e.g., machines or operators processing requests or communication equipment processing information

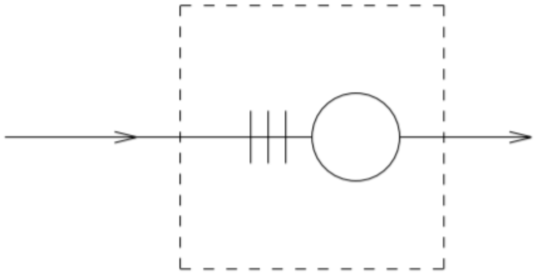


Figure 1: Basic queueing model

Among others, a queueing model is characterized by:

The arrival process of customers

Usually we assume that the interarrival times are independent and have a typical distribution. In many practical situations customers arrive according to a Poisson stream (for example exponential interarrival times). Customers may arrive one by one, or in batches. An example of batch arrivals is the traditions office at the outskirt where travel records of transport passengers have to be checked.

The behaviour of customers

Customers may be patient and willing to wait (for quite a while). Or on the other hand customers may be impatient and leave after some time. For example, in call focuses, customers will hang up when they have to wait excessively some time before an operator is available, and they potentially attempt again after some time.

The service times

Usually we assume that the service times are independent and identically distributed, and that they are independent of the interarrival times. For example, the service times can be deterministic or exponentially distributed. It can also happen that service times are dependent of the queue length. For example, the processing rates of the machines in a generation framework can be increased once the quantity of occupations waiting to be processed becomes excessively large.

The service discipline

Customers can be served one by one or in batches. We have many possibilities for the request in which they enter service. We notice

- first started things out served, for example arranged by arrival;
- random request;
- last started things out served (for example in a PC stack or a shunt support in a generation line);
- priorities (for example surge arranges first, most limited processing time first);
- processor sharing (in PCs that equally separate their processing control over all occupations in the framework).

The service capacity

There may be a solitary server or a gathering of servers helping the customers.

The waiting room

There can be limitations concerning the quantity of customers in the framework. For example, in a data communication network, just finitely many cells can be cradled in a switch. The determination of good cushion sizes is an important issue in the plan of these networks.

Queuing Disciplines

It is easy for one to think about all queues operating like a basic supply checkout line. That is to say, when an arrival happens, it is added as far as possible of the queue and service is not performed on it until all of the arrivals that came before it are served in the request they arrived. Although this a typical strategy for queues to be handled, it is far from the main way. The strategy wherein arrivals in a queue get processed is known as the queuing discipline. This particular example plots a first-start things out serve discipline, or a FCFS discipline. Other potential disciplines incorporate last-start things out served or LCFS, and service in random request, or SIRO. While the particular discipline picked will probably greatly affect waiting times for particular customers (no one wants to arrive early at a LCFS discipline), the discipline generally doesn't affect important results of the queue itself, since arrivals are constantly getting service regardless.

Kendall-Lee Notation

Since portraying all of the characteristics of a queue inevitably becomes exceptionally longwinded, an a lot less complex notation (known as Kendall-Lee notation) can be utilized to depict a framework. Kendall-Lee notation gives us six abbreviations for characteristics listed all together separated by slashes. The first and second characteristics depict the arrival and service processes based on their separate probability distributions. For the first and second characteristics, M speaks to an exponential distribution, E speaks to an Erlang distribution, and G speaks to a general distribution. The third characteristic gives the quantity of servers cooperating at the same time, also known as the quantity of parallel servers. The fourth depicts the queue discipline by it's given acronym. The fifth gives the maximum number of number of customers allowed in the framework. The 6th gives the extent of the pool of customers that the framework can draw from. For example, M/M/5/F CF S/20/inf could speak to a bank with 5 tellers, exponential arrival times, exponential service times, a FCFS queue discipline, a total capacity of 20 customers, and an infinite population pool to draw from.

II. QUEUEING SYSTEMS WITH INVENTORY:

(i) Single Server Queueing Systems with Inventory

Maike Schwarz proposed the model M/M/1 Queueing systems with inventory. They inferred stationary distributions of joint queue length and inventory processes in explicit item structure for various M/M/1-systems with inventory under persistent audit and distinctive inventory management strategies, and with lost sales. Here demand pursues Poisson distribution, service times and lead times exponential distribution.

(ii) Queueing Inventory System with Stochastic Environment

A mathematician displayed an inventory model joined with queueing theory and considered demand and lead time as stochastic parameters. They proposed Poisson distribution for demands and exponential for generation times in a solitary item make-to stock creation system and M/M/1/S queuing system for modelling. He addressed inventory control of a multi-provider strategy in a two-level supply chain. They considered random arrivals for customers and random conveyance time for providers and spoke to the system as a queuing network.

An another author created two distinct models for contemplating inventory systems with constant generation and perishable items. The perishable items have a deterministic usable life after which they should be outdated. For each of the models, analytical articulations got from queueing theory, are found for the steady-state distribution of system inventory. Information of this steady-state behaviour may be utilized for evaluation of system performance, and for consideration of alternatives for improving system performance. The analysis for the two models exploits the similarity of the inventory system with a solitary server queueing system

(iii) Queueing Inventory System with Substitution Flexibility

S.M. presented an application of queueing theory in inventory systems with substitution flexibility which can improve profits in many multi-item inventory systems. They prepared a comprehensive substitution inventory model, where an inventory system with two substitute products with ignorable lead time has been considered, and impacts of simultaneous ordering have been examined and demands of customers for both of the products have been regarded as stochastic parameters, and queuing theory has been utilized to build a mathematical model.

A researcher presented a consistent level generation rate with a base-stock level inventory strategy subject to fluctuating demand. The inventory level at time t is signified by I(t) and is the time required to deliver the items presently in stock given a generation rate of unity. Negative inventory levels I(t) < 0, reflect orders that cannot be satisfied by current inventory and are backlogged assumed that the inventory is delivered at a rate which relies upon the present inventory level and may vary from unity. Consequently, the generation rate is controlled by the maker, which presents a criticism mechanism for inventory control. They characterize the base-stock level M as the level of inventory at which creation stops. Creation possibly restarts when I(t) < M. Requests or demands arrive randomly with a between arrival time that is exponentially distributed with a parameter that relies upon the present inventory level. The extent of the requests is exponentially distributed. The reliance of the arrival rate on the present inventory level can have several interpretations. The model is appropriate if creation is very automated and easily controlled.

(iv) Queueing System with Production-Inventory

He examined the inventory replenishment strategy of a M/M/1 make to-arrange inventory-creation system with zero lead times. They investigated the structure of the optimal replenishment approach which limits the average total cost per item. For the M/PH/1 make-to-arrange inventory-generation system with Erlang distributed lead times, Heet. quantified the value of information utilized in inventory control. A logically related model has been examined.

"The M/M/1 queue with a production inventory system and lost sales". The authors considered an augmentation of the queueing system with inventory in which the stocks are conveyed both by an outside provider and an internal generation and called the proposed queueing system as a M/M/1 queue with an attached creation inventory system. Customers arrive in the system according to a Poisson process, and a solitary server serves the customers. The customers leave the system with exactly one item from the inventory at his service fulfilment age. In the event that there is no inventory item, all arriving customers are lost. The stocks are replenished by (1) an external request under (r, Q)- approach, or (2) an internal generation. The internal generation process is assumed to be a Poisson process. They inferred the stationary joint distribution of the queue length and the on-hand inventory in item structure. Utilizing the joint distribution, they presented long-run performance measures and a cost model. Then, they established numerical examples, which limit the long-run cost per unit time.

They considered random arrivals for customers and random conveyance time for providers and spoke to the system as a queuing network. The goal of inventory management is to balance clashing goals, for example, stock costs and shortage costs.

(v) Queueing System with Service Inventory

In inventory management perspective, the assembly like queue can be applied to a service-inventory system in which the customers can be served just when the level of the attached inventory is positive. The least complex example is a retail market where customers invest energy to pay (the service time) for the item (the inventory) that they want to purchase. A supply network model was created for a service facility system with perishable inventory by considering a two-dimensional stochastic process of the structure (L, X) = $\{(L(t), X(t)), t \geq 0\}$, where L (t) is the level of the on hand inventory and X (t) is the quantity of customers at time t. The between arrival time to the service station is assumed to be exponentially distributed with mean 1/λ and the service time for each client is exponentially distributed with mean 1/μ. The maximum inventory level is S and the maximum capacity of the waiting space is N. The replenishment process is assumed to be (S1, S) with a replenishment of just a single unit at any level of the inventory. Lead time is exponentially distributed with parameter β. The items are replenished at a rate of β whose mean replenishment time is $1/\beta.$ Item in inventory is perishable when it's utility drops to zero or the inventory item turned out to be useless while in storage. Perishable of any item happens at a rate of γ. Once entered a queue, the client may leave the queue at a rate of α on the off chance that they have not been served after a certain time. They determined the steady state probability distributions for the system states.

(vi) Continuous Review Inventory Systems with Server Vacation

A few authors presented the idea of server vacation in the inventory system with two servers. They had examined a (s,S) inventory system in which the server took a rest when the level of the inventory became zero. They assumed that the demands that happened amid stock-out period or the server rest periods were lost. The between event times between progressive demands, the lead times, and the rest times were assumed to pursue general distributions. Utilizing renewal and convolution systems, they obtained articulations including the steady state transition probabilities. An inventory system with random positive service time was considered. Customers arrived at the service station according to a Markovian arrival process and the service time for each client had phase-type distribution. They assumed correlated lead time for the requests and an infinite waiting hall for the customers. The customers who wait for service may renege after a random time. The server got away whenever there was no client waiting in the system or the inventory level was zero. Under the above assumptions, they analysed the level dependent quasi birth-death process.

(vii) Queueing Inventory System with Postponed Demands/Customers.

A finite source (s,S) inventory system with postponed demands and server vacation was considers. Their altered M vacation arrangement is characterized as: Whenever the inventory level reaches zero, the server goes to inactive period which comprises the inactive-idle and vacation period. In the event that replenishment happens amid the inactive-idle time frame, the server becomes active immediately, or otherwise he goes for a vacation period. The server can take at most M inactive periods repeatedly until replenishment takes place. This inactive-idle time, the vacation time and lead time pursue independent PH distributions. After the Mth inactive period, the server remains dormant in the system independent of the replenishment of request. Demands that happen amid stock out or inactive periods, enter the pool and these demands are chosen if the inventory level is above. The between choice time pursues exponential distribution. The joint distribution of the method of the server, server status, the inventory level and the quantity of demands in the pool is obtained in the steady state. They have determined several system performance measures and total expected cost function.

(viii) Queuing Systems with New Inventory Models

Narayanan contemplated a Markovian inventory system with positive service time and Krishnamoorthy considered a creation inventory with service time. In both model, server got away because of lack of inventory or the lack of client or both. Queuing systems with server vacation have been generally examined in the literature. They have contemplated various vacation strategies, for example, single and different vacation approach for single server and also synchronous and asynchronous with single and numerous vacation for multi-server queueing system. An another have considered M[X]/G/1 queue with J vacation approach in which after the exhaustive service, the server takes at most J vacation of constant length repeatedly until at least one client present in the system. A few authors considered a nonstop survey perishable inventory system with service facility consisting of finite waiting hall and a solitary server. The primary customers arrive according to a Markovian arrival process. An arriving client, who finds the waiting hall is full, is viewed as lost. The individual client's unit demand is satisfied after a random time of service which is distributed as exponential. The existence time of each item is assumed to be exponentially distributed. The items are replenished based on variable requesting strategy. The lead time is assumed to have phase type distribution. After the service finish, the primary client may choose either to join the secondary (feedback) queue, which is of infinite size, or leave the system according to a Bernoulli trial and the server chooses to serve either for primary or feedback client according to a Bernoulli trail. The primary and secondary services are at various counters. The service time for feedback customers is assumed to be independent exponential distribution. After the service finish for feedback client, the server starts immediately for primary client's service, whenever the inventory level and the primary client level is positive, otherwise the server becomes idle for an exponential duration. On the off chance that the primary client level and inventory level becomes positive amid the server idle period then he starts service for primary client immediately. After finishing his idle period, the server goes to secondary counter to serve for feedback client, assuming any. The joint probability distribution of the system is obtained in the steady state. Important system performance measures are inferred and the long-run total expected cost rate is also calculated.

III. QUEUING MODELS

With our foundation laid for the investigation of important characteristics of queuing systems, we can start to analyse particular systems themselves. We will start by taking a gander at one of the least complex systems, the M/M/1/GD/∞/∞ framework.

The M/M/1/GD/∞/∞ Queuing System

A M/M/1/GD/∞/∞ framework has exponential interarrival times, exponential service times, and one server. This framework can be modeled as a birth-death process where

$$\lambda_j = \lambda \, for \, (j = 0, 1, 2 \ldots)$$

$$\mu_0 = 0$$

$$\mu_j = \mu \, for \, (j = 1, 2, 3 \ldots)$$

Substituting this in to the equation for the steady-state probability, we get

$$\pi_j = \frac{\lambda^j \pi_0}{\mu^j}$$

We will define p = λ/μ as the traffic intensity of the system, which is a ratio of the arrival and service rates. Knowing that the sum of all of the steady state probabilities is equal to one, we get

$$\pi_0(1 + p + p^2 + \ldots + p^j) = 1$$

If we assume $0 \leq p \leq 1$ and let the sum $S = (1 + p + p 2 + \ldots + p j)$, then $\mathrm{S} = 1 \; 1{-}\mathrm{p}$ and $\pi 0 = 1 - \mathrm{p}$. This yields

$$\pi j = p \, j \, (1 - p)$$

as the steady-state probability of state j. Note that if p ≥ 1, S approaches infinity, and in this manner no steady state can exist. Intuitively, in the event that p ≥ 1, then it must be that λ ≥ μ, and in the event that the arrival rate is greater than the service rate, then the state of the framework will develop without end.

With the steady-state probability for this framework calculated, we can now settle for L. In the event that L is the average number of customers present in this framework, we can speak to it by the formula $L = \sum_{j=0}^{\infty} j\pi_j = (1-p)\sum_{j=0}^{\infty} jp^j$

Let $S = P\infty\, j = 0 = p + 2p\,2 + 3p\,3 + \ldots.$ Then $pS = p\,2 + 2p\,3 + 3p\,4 + \ldots.$ If we subtract, we get

$$S - pS = p + p^{2} + p^{3} \ldots = \frac{p}{1 - p}$$

And $\mathrm{S = p\,(1{-}p)\,2}$. Substituting this into the equation for L will get us

$$L = (1 - p)\,\frac{p}{(1 - p)^{2}} = \frac{p}{1 - p} = \frac{\lambda}{\mu - \lambda}$$

To unravel for Ls, we have to decide what number of customers are in service at any given minute. In this particular framework, there will always be one client in service with the exception of when there are no customers in the framework. Accordingly, this can be calculated as

$$L_q = 0\pi_0 + 1(\pi_1 + \pi_2 + \pi_3 + \ldots) = 1 - \pi_0 = 1 - (1 - p) = p$$

From here, Lq is an easy calculation.

$$L_q = L - L_s = \frac{p}{1-p} - p = \frac{p^2}{1-p}$$

Utilizing Little's queuing formula, we can also illuminate for W, Ws, and Wq by separating each of the relating L values by λ.

The M/M/1/GD/c/∞ Queuing System

A M/M/1/GD/c/∞ queuing system has exponential interarrival and service times, with rates λ and μ separately. This system is fundamentally the same as the past system, then again, actually whenever c customers are available in the system, all additional arrivals are barred from entering, and are thereafter never again considered. For example, if a client were to walk up to a fast sustenance restaurant and see that the lines were unreasonably long for him to want to wait there, he would go to another restaurant instead.

A system like this can be modelled as a birth-death process with these parameters:

$$\lambda j = \lambda\, for\, j = 0, 1, \ldots, c - 1$$

$$\lambda c = 0$$

$$\mu 0 = 0$$

$$\mu j = \mu\, for\, j = 1, 2, \ldots, c$$

The limitation λc = 0 is what separates this from the past system. It makes it with the goal that no state greater than c can ever be reached. Because of this limitation, a steady state will always exist. This is because regardless of whether λ ≥ μ, there will never be more than c customers in the system.

Seeing formulas got from the investigation of birth-death processes and by and by letting p = λ μ , we can determine the accompanying steady-state probabilities:

$$\pi_0 = \frac{1 - p}{1 - p^{c+1}}$$

$$\pi_j = p^j \pi_0 \text{ for } j = 1, 2, \ldots, c$$

$$\pi_j = 0 \text{ for } j = c + 1, c + 2, \ldots, \infty$$

A formula for L can be found in a similar fashion, yet is omitted because of the muddled calculations. The system is similar to the one utilized in the past segment.

Calculating W is another issue. This is because in Little's queuing formula, λ speaks to the arrival rate, yet in this system, not all of the customers who arrive will join the queue. In fact, λπc arrivals will arrive, however leave the system. Therefore, just λ − λπc = λ(1 − πc) arrivals will ever enter the system. Substituting this into Little's queuing formula gives us

$$W = \frac{L}{\lambda(1 - \pi_c)}$$

The M/M/s/GD/∞/∞ Queuing System.

A M/M/s/GD/∞/∞ queuing system, similar to the past system we took a gander at, has exponential interarrival and service times, with rates λ and μ. What separates this system is that there are s servers willing to serve from a solitary line of customers, as perhaps one would discover in a bank. On the off chance that $j \leq s$ customers are available in the system, then every client is being served. On the off chance that j > s customers are in the system, then s customers are being served and the remaining j − s customers are waiting in the line.

To model this as a birth-death system, we have to see that the death rate is dependent on what number of servers are actually being utilized. In the event that each server finishes service with a rate of μ, then the actual death rate is μ times the quantity of customers actually being served. Parameters for this system are as per the following:

$$\lambda_j = \lambda\, for\, j = 0, 1, \ldots, \infty$$

$$\mu_j = j\mu\, for\, j = 0, 1, \ldots, s$$

$$\mu_j = s\mu\, for\, j = s + 1, s + 2, \ldots, \infty$$

In unraveling the steady-state probabilities, we will characterize p = λ sμ . Notice that this definition also applies to the other systems we took a gander at, since in the other two systems, s = 1. The steady-state probabilities can be found in this system in the same manner as for other systems by utilizing the stream balance equations. I will also omit these particular steady-state equations because they are rather awkward.

The M/G/∞/GD/∞/∞ and GI/G/∞/GD/∞/∞ Queuing Systems.

These systems are separate in that they have an infinite number of servers, and in this way, a client never has to wait in a queue for their service to start. One way to think about this is as a self-service, such as shopping on the web, for example. In this system, it can be demonstrated that W = 1 µ and L = λ µ. It can also be demonstrated that the steady state probability at state j is

$$\pi_j = \frac{(\frac{\lambda}{\mu})^j e^{-(\frac{\lambda}{\mu})}}{j!}$$

The Machine Repair model

The machine repair model is a M/M/R/GD/K/K queue system, where R is the quantity of servers, and K is both the span of the client population and the maximum number of customers allowed in the system. This model can explain a situation where there are K machines that each break down at rate λ and R repair workers who can each fix a machine at rate µ. This means that both λ and µ are dependent on either what number of machines are remaining in the population or what number of repair workers are in service.

How about we model this as a birth-death process. Since λj relies upon the quantity of machines left in the population that are not in service, we can say that

t $\lambda_j = (K - j)\lambda$

We can calculate µj by taking a gander at the quantity of repair workers as of now in service. In the event that a machine breaks down when all servers are busy, it waits in a queue to be served. We can calculate µj as pursues:

$$\mu j \ = \ j\mu \, for \, j \ = \ 0, 1, \ldots, R$$

$$\mu j \ = \ R\mu \, for \, j \ = \ R \ + \ 1, R \ + \ 2, \ldots, K$$

The steady-state probability for a machine repair system, derived from page 1081 of the Winston text is

$$\pi_j = \binom{K}{j} p^j \pi_0 \text{ for } J = 0, 1, ..., R$$

$$\pi_j = \frac{\binom{K}{j} j! p^j \pi_0}{R! R^{j-R}} \text{ for } J = R+1, R+2, ..., K$$

The M/G/s/GD/s/∞ Queuing System

Another reasonable model of a queue is one where on the off chance that a client arrives and sees all of the servers busy, then the client exits the system totally without getting service. In this case, no actual queue is ever framed, and we say that the blocked customers have been cleared. Since no queue is ever shaped, $Lq \ = \ Wq \ = \ 0$. On the off chance that λ is the arrival rate and 1/µ is the mean service time, then $W \ = \ Ws \ = \ 1/\mu$

In this system, arrivals are dismissed whenever s customers are available, so πs is equal to the fraction of all arrivals who are dismissed by the system. This means that an average of λπs arrivals per unit of time will never enter the system, and therefore, λ(1 − πs) arrivals per unit of time will actually enter the system. This leads us to the end based on Little's queuing formula that $L \ = \ Ls \ = \ \lambda(1 - \pi s) \ \mu$.

IV. MATHEMATICAL FORMULATION

With arrival constituting a Poisson process of rate λ, service time independent identically distributed exponential random variables with parameter µ, lead time for replenishment having exponential distribution with parameter β and between retrial time of head of the queue in the orbit following exponential distribution with parameter θ, the process $\{(N\,(t), C(t), I(t))|t \ \geq \ 0\}$ forms a CTMC on the state space Ω given by

$$(Z + \ [\,\{0\}) \ \times \ \{0, 1\} \ \times \ \{1, 2, \ldots, S\} \, [\,(Z + \ [\,\{0\}) \ \times \ \{0\} \ \times \ \{0\}$$

It is to be noticed that we make a solid assumption on customers getting into the system: when inventory level is zero, no client joins the system. When the inventory level reaches a pre-indicated value s > 0, a replenishment request is placed for Q units with Q > s. We fix S = Q+s as the maximum number of items that could be held in the system at any given time. That is the replenishment arrangement pursued is (s, Q). Further, as considered in Krishnamoorthy et.al it is assumed that at the finish of service a client is given one unit of the item from inventory with probability γ. We expected, "the assumption that no client joins the system when inventory is zero" would enable us to arrive at, in the least, a shut structure arrangement of the system state distribution, if not decomposition of the system. Nevertheless, it ended up being otherwise. Subsequently we are compelled to adopt algorithmic approach for the analysis of the system depicted. The state space of the CTMC is partitioned in to levels L(i) characterized as,

$$L(i) = \{(0,0,j)|0 \le j \le S\} \cup \{(0,1,j)|1 \le j \le S\}$$
$$\cup \{(i,k,j)\,/i \ge 1;\ k = 0,1;\ 0 \le j \le S\}$$

The transitions in the Markov chain are listed below:

(a) Transitions due to arrival of customers:

$$(i,0,j) \rightarrow (i,1,j) : \textit{the rate is } \lambda, \textit{for } i \ge 0;\ 1 \le j \le S.$$
$$(i,1,j) \rightarrow (i+1,1,j) : \textit{the rate is } \lambda, \textit{for } i \ge 0;\ 1 \le j \le S.$$

(b) Transitions due to service completion of customers:

$$(i,1,j) \rightarrow (i,0,j-1) : \textit{the rate is } \gamma\mu, \textit{for } i \ge 0;\ 1 \le j \le S.$$
$$(i,1,j) \rightarrow (i,0,j) : \textit{the rate is } (1-\gamma)\,\mu, \textit{for } i \ge 0;\ 1 \le j \le S.$$

(c) Transitions due to replenishments:

$$(i,0,j) \rightarrow (i,0,Q+j) : \textit{the rate is } \beta, \textit{for } i \ge 0;\ 0 \le j \le s.$$
$$(i,1,j) \rightarrow (i,1,Q+j) : \textit{the rate is } \beta, \textit{for } i \ge 0;\ 0 \le j \le s.$$

(d) Transitions due to retrial of customers:

$$(i,0,j) \rightarrow (i-1,1,j) : \textit{the rate is } \theta, \textit{for } i \ge 1;\ 1 \le j \le S.$$

All other transition pairs have rate zero. The infinitesimal generator Q of this CTMC is given by

$$Q = \begin{bmatrix} B_0 & B_1 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \cdots \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

where B0, B1 and B2 contain transition rates within L(0), transition from L(0) to L(1) and transition from L(1) to L(0) respectively; A0 represents transitions from L(i) to L(i + 1), i ≥ 1; A1 represents transitions within L(i) for i ≥ 1, and A2 represents transitions from L(i) to L(i − 1), i ≥ 2. All these matrices are square matrices of order 2S + 1.

V. SYSTEM STABILITY AND COMPUTATION OF STEADY-STATE PROBABILITY VECTOR

The Markov chain under consideration is a LIQBD process. For this chain to be stable it is necessary and sufficient that

$$\xi A0e < \xi A2e.$$

where ξ is the unique non negative vector satisfying,

$$\xi A = 0, \xi e = 1$$

and A = A0 + A1 + A2, is the infinitesimal generator of the finite state CTMC. Let ξ = (ξ0(0), ξ0(1), . . ., ξ0(S), ξ1(1), ξ1(2), . . . ξ1(S)) be the steady-state vector of the generator matrix A. Then ξA = 0 gives the following equations

$$-\beta\xi_0(0) + \gamma\mu\xi_1(1) = 0$$

$$-(\lambda+\theta+\beta)\xi_0(j) + (1-\gamma)\mu\xi_1(j) + \gamma\mu\xi_1(j+1) = 0,\ 1 \le j \le s$$

$$-(\lambda+\theta)\xi_0(j) + (1-\gamma)\mu\xi_1(j) + \gamma\mu\xi_1(j+1) = 0,\ s+1 \le j \le Q-1$$

$$\beta\xi_0(j) - (\lambda+\theta)\xi_0(Q+j) + (1-\gamma)\mu\xi_1(Q+j) + \gamma\mu\xi_1(Q+j+1) = 0,\ 0 \le j \le s-1$$

$$\beta\xi_0(s) - (\lambda+\theta)\xi_0(S) + (1-\gamma)\mu\xi_1(S) = 0$$

$$(\lambda + \theta)\xi_0(j) - (\beta + \mu)\xi_1(j) = 0, 1 \le j \le s$$
$$(\lambda + \theta)\xi_0(j) - \mu\xi_1(j) = 0, s + 1 \le j \le Q$$
$$\beta\xi_1(j) + (\lambda + \theta)\xi_0(Q + j) - \mu\xi_1(Q + j) = 0, 1 \le j \le s$$

The LIQBD process with infinitesimal generator Q is stable if and only if $\xi A0e < \xi A2e$. That is,

$$\theta\,(\xi_0(1) + \xi 1(2) + \cdots + \xi_0(S)) > \lambda\,(\xi_1(1) + \xi_1(2) + \cdots + \xi_1(S)),$$

which simplifies to λ μ < θ λ+θ . Thus we have the following lemma for the stability of the system:

VI. PERFORMANCE MEASURES

- Mean number of customers in the orbit,

$$L_O = \left(\sum_{i=1}^{\infty}\sum_{j=0}^{Q+s} i x_i(0,j) + \sum_{i=1}^{\infty}\sum_{j=1}^{Q+s} i x_i(1,j)\right)$$

- Mean inventory level,

$$E_{inv} = \sum_{i=0}^{\infty}\sum_{j=0}^{Q+s} j x_i(0,j) + \sum_{i=1}^{\infty}\sum_{j=1}^{Q+s} j x_i(1,j)$$

- Depletion rate of inventory,

$$D_{inv} = \gamma\mu_2\left(\sum_{i=1}^{\infty}\sum_{j=1}^{Q+s} x_i(1,j)\right)$$

- Mean number of replenishments per unit time,

$$R_r = \beta\left(\sum_{j=0}^{s}\left(\sum_{i=0}^{\infty} x_i(0,j) + \sum_{i=1}^{\infty} x_i(1,j)\right)\right)$$

- Expected loss rate of customers,

$$E_{loss} = \lambda\left(\sum_{i=1}^{\infty} x_i(0,0)\right)$$

- Probability that the server is busy,

$$P_{busy} = \sum_{i=0}^{\infty} \sum_{j=1}^{Q+s} x_i(1, j).$$

- Successful rate of retrials,

$$E_{retrial} = \theta \left(\sum_{i=1}^{\infty} \sum_{j=1}^{Q+s} x_i(0, j) \right)$$

- Mean number of departures per unit time,

$$D_m = \mu \left(\sum_{i=0}^{\infty} \sum_{j=1}^{Q+s} x_i(1, j) \right)$$

- Mean number of customers waiting in the orbit when inventory is available,

$$\widetilde{W_O} = \left(\sum_{i=1}^{\infty} \sum_{j=1}^{Q+s} i x_i(0, j) + \sum_{i=1}^{\infty} \sum_{j=1}^{Q+s} i x_i(1, j) \right)$$

- Mean number of customers waiting in the orbit during the stock out period,

$$\widetilde{\widetilde{W_O}} = \left(\sum_{i=1}^{\infty} i x_i(0, 0) \right)$$

VII. OPTIMIZATION PROBLEM

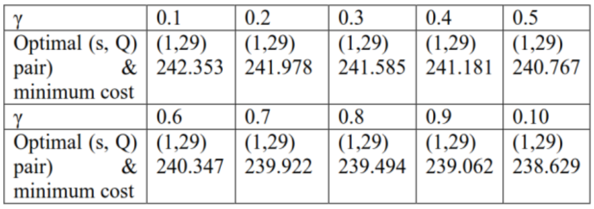
In this area we give the optimal values of the inventory level s and the fixed request quantity Q of the model. For checking the optimality of s and Q, the accompanying cost function is developed. Characterize F(s, Q) as the normal total cost per unit time over the long haul. Then

$$\mathcal{F}(s, Q) = h.E_{inv} + c_1.E_{loss} + c_2.(1 - P_{busy}) + (K + Q.c_3).R_r,$$

where K is the fixed cost for placing a request, c1 is the cost acquired because of misfortune per client, c2 is the waiting cost per unit time per client amid the stock out period, c3 is the variable obtainment cost per item and h is the unit holding cost of inventory for one unit of time. Table 1 gives the optimal pair (s, Q) and the comparing least cost (in Dollars) by utilizing MATLAB program. Here γ is varied from 0.1 to 1, at an interval of 0.1. The values for the information parameters are given as pursues λ = 2, μ = 5, θ = 4, β = 3, K = \$500, c1 = \$25, c2 = \$50, c3 = \$35, h = \$3.5. We provide a numerical comparison based on a few performance measures in Table 3.

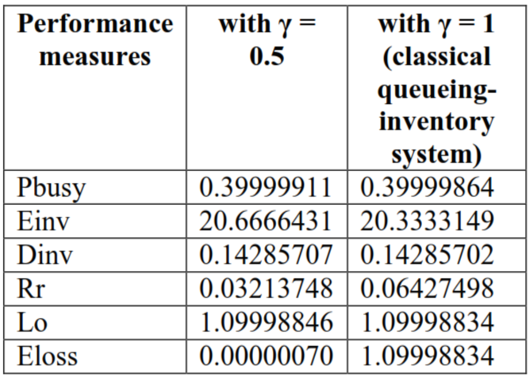
For numerical comparison we assign the same input values as for Table 1with s = 10 and S = 31. For example, we observe from Table 2 that the mean number of replenishments and loss rate of customer is larger for γ = 1 compared to that for γ (= 0.5). Further P_{busy} and E_{inv} are higher for γ = 0.5 compared to that for γ = 1. These are all on expected lines.

Table 1: Optimal (s, Q) pair and minimum cost



γ	0.1	0.2	0.3	0.4	0.5
Optimal (s, Q) pair) & minimum cost	(1,29) 242.353	(1,29) 241.978	(1,29) 241.585	(1,29) 241.181	(1,29) 240.767
γ	0.6	0.7	0.8	0.9	0.10
Optimal (s, Q) pair) & minimum cost	(1,29) 240.347	(1,29) 239.922	(1,29) 239.494	(1,29) 239.062	(1,29) 238.629

Table 2: Effect of γ on various performance measures



Performance measures	with γ = 0.5	with γ = 1 (classical queueing-inventory system)
Pbusy	0.39999911	0.39999864
Einv	20.6666431	20.3333149
Dinv	0.14285707	0.14285702
Rr	0.03213748	0.06427498
Lo	1.09998846	1.09998834
Eloss	0.00000070	1.09998834

VIII. CONCLUSION

In this paper we discussed a M/M/1/1 queueing-inventory system with retrial of unsatisfied customers. Arrivals taking place when server busy, continue to an orbit of infinite capacity. From the orbit, the head of the queue alone resigns to access the server. Failed attempts to access an idle server with positive inventory results in the retrial client coming back to orbit. The between retrial times are independent identically distributed exponential random variables, regardless the quantity of customers in the orbit. We figured the condition for stability and then utilized algorithmic approach to obtain the system steady-state probability. The normal waiting time of a client in the orbit, distribution of the time until the principal client goes to orbit and the probability of no client going to orbit in a given interval of time were processed. An optimization issue is also numerically investigated.

REFERENCES:

1. Z. Melikov, A. M. Rustamov & L. A. Ponomarenko (2017). “Approximate analysis of a queueing–inventory system with early and delayed server vacations” Springer November 2017, Volume 78, Issue 11, pp 1991–2003

2. Gabi Hanukova, Tal Avinadava, Tatyana Chernonoga, Uriel Spiegel A. B., Uri Yechiali (2017). “A queueing system with decomposed service and inventoried

preliminary services" Applied Mathematical Modelling 47, pp. 276–293

3. K. Karthikeyan, R. Sudhesh (2016). "Recent review article on queueing inventory systems" Research J. Pharm. and Tech. 9(11): November 2016, pp. 1451-1461

4. A. Krishnamoorthy, R. Manikandan, and Dhanya Shajin (2014). "Analysis of a Multiserver Queueing-Inventory System" Advances in Operations Research Volume, Article ID 747328, 16 pages

5. Reza Rashid, Seyed Farzad Hoseini, M. R. Gholamian & Mohammad Feizabad (2015). "Application of queuing theory in production-inventory optimization" J Ind Eng Int 11: pp. 485–494 DOI 10.1007/s40092-015-0115-9

6. Krishnamoorthy, A., Lakshmy, B., and Manikandan, R. (2011). A Survey on Inventory Models with Positive Service Time, OPSEARCH, 2011, Vol. 48, No. 2, pp. 153–169.

7. Seyedhoseini S., Rashid R., Kamalpour I., Zangeneh E. (2015). Application of queuing theory in inventory systems with substitution flexibility. J Ind Eng Int, pp. 1–8

8. Anoop N. Nair and M.J. Jacob, (S, S) inventory system with positive service time and retrial of demands: an approach through multiserver queues, ISRN Operations Research, Article ID 596031, http://dx.doi.org/10.1155/2014/596031.

9. Baek J.W., Moon S.K. (2014). The M/M/1 queue with a productioninventory system and lost sales. Appl Math Comput 233: pp. 534–544

Corresponding Author

Devendra Kumar Pandey*

Professor & Director, Unique Institute of Management & Technology, Ghaziabad, India

devkp60@rediffmail.com