International Journal of Innovative Mathematics, Statistics & Energy Policies 4(4):1-7, Oct.-Dec. 2016

© SEAHI PUBLICATIONS, 2016 www.seahipaj.org ISSN: 2467-852X

Equilibrium Behavior in Queueing Systems

NAJEEB AL-MATAR
Management Information System Department,
AlBaha University, Saudi Arabia
Email: nalmatar2006@my.fit.edu

ABSTRACT

The research paper analyzes the decisions made by customers at the service center when deciding to join or cut the M/M/1 queue model whereby they can choose to cut the queue and later be overtaken after joining. In this case, the equilibrium is demonstrated among the patient customers in queueing games. The agents are asked to make decisions independently as to when to or not to join the queuing line. In this stage, repetition of this behavior game results to the introduction of a mixed-strategy approach to deal with the problem. This is used to obtain a solution to the strategies regarding joining arrivals. However, there are methods employed to prevent the delay times such as First Come First Served (FCFS), Egalitarian Processor Sharing (EPS) and any other equitable disciplines to facilitate the queuing solution.

Keywords: Equilibrium, Queueing model; Equilibrium solution, Discipline

INTRODUCTION

In service center systems, customers act independently such as deciding to cut the queue or join it with the aim of maximizing their individual welfare. In such case, the remedy taken by customers and the service facility authorities will significantly affect the rest of the remainder of the client's optimal behavior in the queue. Consequently, the resultant outcome is an aggregate equilibrium expression pattern. At the face of the society, this equilibrium behavior pattern not seem to be an optimal solution or the best payoff to the affiliated. For a long time, after the publication of a pamphlet by Naor (1969), similar observations have been made by economists but being only concerned with queuing theory. Before Naor (1969) publication, the scope of queuing theory is well reflected in another book by Leeman (1964); in his "Letter to the Editor" that ends with the following except as below:

> *"It is quite surprising that in the process of minimizing queues, the capitalist economy has limited itself to the recommendations of administrative measures through the adoption of the queuing theory limits. Such acts are expected to be seen in a planned economy and an economy where markets and prices play a significant role."*

According to Leeman (1964), there are three primary objectives that service centers can be attained by pricing the queuing mechanism. First, the existing allocation service facilities should be improved through allocating the real central priorities and the shift of demand from the spatial-temporal bottlenecks which can replace the notion of the First- Come-First Served rule. Further, Leeman proposes that management decisions should be centralized, and lastly through the guidance of the long-term investment decisions. However, Leeman (1964) had missed another important aim that is later addressed by Naor groundbreaking paper. Naor proposes that the demand process should be regulated and without such an action, the service facility could be extensively used. Besides, there have been further extensive surveys done on the same equilibrium behavior in queuing systems that followed Naor's work. However, most of the works have regarded the equilibrium behavior in queuing system implicitly and thus the concept is still interesting to carry a study.

For instance, the economic principle requires that for the optimal allocation condition to be achieved concerning the usage of the economic principles, an economic cost should be put on those using the scarce economic resources. Again, Knudsen (1972) realized that the meaning of queuing model behavior is broader for stochastic queuing models compared to the usual deterministic and static economic theory models. With this view, an economic resource will not be considered as scarce if the capacity of the system can adequately accommodate the expected demand for the service in a deterministic model. In this case, the social welfare of the customers can still be increased given that such condition prevails. This follows from the probability that at one given time, there exists a non-negative probability that the service facility will be completely exploited. Moreover, this condition can still prevail even when the service rate exceeds the arrival rate so that the server ca handle all entries. Besides, queues form due to unevenness in inter-arrival and service times. Therefore, from the above analysis, a queue can be regarded as a price that the service facility has to pay for it to guarantee some degree of server exploitation. Hence, the deterministic and stochastic models are significantly different from the criteria of economic optimality. This paper seeks to come up with a solution to the problem of deciding the appropriate joining strategy among the homogeneous customers by placing it at the intersection of two areas of literature. This is done by facing a system of M/M/1 queue model discipline such as First Come First Served but, overtaking is possible among the potential customers. However, this will take a Last Come First Served discipline if overtaking is allowed. Once it is granted, customers delay times will depend on the service rate apart from the length of the queue.

LITERATURE REVIEW

Naor (1969) introduced an appropriate strategy for FCFS to assist in maximizing the social welfare of the queueing customers. The recommendation was to start charging a 'fee' to the joining customers who could minimize the Nash threshold to optimal social one. This was generalized to waiting and arrival time distributions and a reasonable number of servers (Knudsen, 1972). With Naor's paper, more articles explaining and examining customers' queueing reactions followed such as M/M/1 FCFS strategies. Since then, more articles have been published regarding the subject such as Burnetas and Economou (2007) and Sun et al. (2009). Research beyond FCFS was rarer as it was more mathematically involved. According to Guillemin and Boyer (2001), customers can overtake others in the queue and hence covering the FCFS M/M/1 model. Further, from Guillemin and Boyer's view, customers' overtaking do not happen because of interactions with others but pay some fee to the servers for them to be allowed. Also, the delay time is not examined as it can be done in FCFS model through the application of EPS approach. This method evaluates the expected delay time and also the construction of a left-multiplication in determining the joining strategy at entrance (Larson, 1987).

The difference in Mailath and Samuelson (2006), and the present work is that the likelihood of overtaking comes from repeated mingling among customers rather than using the designed rules by the station managers. This brings some classification of queueing systems where Parsons (1995) regards it to be a social system since it brings about interactions among individuals involved. These kinds of interactions can be modeled to be investigated in a well -defined standard theoretic game tool framework of repeated games (Okuno $ Postlewaite, 1995).

Borrowing from the sociological and psychological perspective, according to Burnetas and Economou (2007), the two social thinking play a significant role in governing the queueing norms and strategies. The sociological evaluation of waiting for a service and customers' interpretations in the fairness of queueing rules has been ever the guiding principles. However, overtaking can meet with failure because most of them may not be successful. The avoidance of FCFS discipline is termed to be unfair or unjust and at the same time interpreting the aspect of overtaking to be overwhelmingly negative (Guillemin and Boyer, 2001). Moreover, this is reinforced by study in reactions to avoidance of FCFS strategy in an overnight queue for a concert in *U*2 whereby negative results were found as a result of little outcomes (Helweg and LoMonaco, 2008). However, payment was made for letting a customer cut the queueing line which was accepted and hence which would conclude that self-interest was the dominant factor in analyzing the behavior in the queue. This leads to the introduction of a model of queue governing customers with different interest and priorities and concluded that patient customers will give a chance to the one with high priority to cut the line (Milgram, 1996).

For instance, as put by Allon and Hanany (2012) in a repeated game dramatic analysis where overtaking is possible in the queue as set by the present report. This deals with customers of the same interest in a queue and assists in the simplifying computation of delay time and also the process of identifying possible cutting attempts possible. Further, this analysis allows an equilibrium behavior presentation where patient customers allow others to overtake them providing dependent queue strategies which do not affect the anticipated state of equilibrium. The overtaking attempts are repeated in this case and hence trigger procedure in a single equilibrium presentation.

According to Yu, Tang and Wu (2014), the method of relieving customer delay time in cutting points is done once equilibrium state is established by employing the EPS discipline and from this case the joining strategy can be analyzed and investigated. According to Naor's findings, M/M/1 queue model follows when overtaking systems are presented after which joining strategies and delay times are obtained which contrasts with those for EPS and FCFS disciplines. However, the results prevent presentation of general conclusions since the simulations indicate the line cutting as an action which reduces the social welfare and should not be allowed (Guillemin and Boyer, 2001).

METHODOLOGY

First, after reviewing and applying M/M/1/0 problem methodology and results, this analysis illustrates the case study of patients at the health service and further includes the assumptions of the simulation model, health service facility with one server. Further, the study includes the equilibrium behavior of the system when one server is used and then gives the obtained empirical results for the optimal social condition, and, lastly, and conclusions based on the analyses made.

A Simple Rule for M/M/1/∞.

According to Sun, Guo, and Tian (2009), the transient behavior of M/M/1/∞ is estimated at Q (*t*), using a technique that numerically solves the Kolmogorov differential equations of a system. In this case, the system describes the transition rate of a continuous- time, discrete- time Markov Process. The process has been solved by the use of the International Math-Science Library subroutine DVERK, where two main reasons led to the selection of the empirical approach to obtain the desired outcome (Sun and Li, 2014). First, there were difficulties in getting the closed-form and the general solutions to the simplest types of queuing systems. Further, the need to establish specific solutions from which they would draw conclusions. With the analysis of the above conditions, authors have confirmed that the equilibrium is bounded by the decaying exponential function given as;

$Q\ (t) = Q\ (\infty)\ *[1- e]\ -t\ /\ \tau$

Where,

$Q\ (\infty)$ = value regarding the function approaches as t increases.

τ = Relaxation time regarding the system parameters.

And,

τ is not contingent on the initial state but directly proportional to the service rate.

Thus,

$\tau = C\ (2, a) + C\ (2, s)/\ \{2,\ 8\mu s\ (1- \sqrt{\rho})^{2}\}$

Where, $C\ (2, a)$ + C (2, *s*) are inter-arrival coefficients of variation and service times respectively.

$1/\mu s$ = mean service time.

Further, from M/M/1/∞, if C (2, *a*) = C (2, *s*) =1

For practicality, if *Q (t)* is within 2% of *Q* (∞), it is then regarded as sufficiently closed to the steady state value (Erlichman and Hassin, 2009). For this is subtle, it can then be ignored.

Again, where servers are infinite,

$Q\ (t) = \rho\ *[1- e\ -\mu t]$,

Thus, from the function, the transient time can be estimated, and at the same time computing the upper bounds for the steady-state to be reached.

Which can be calculated as,

$Ub = 4\tau$

By taking the following theoretical tested different values by Erlichman, and Hassin (2009) at the health service facility where customers want to maximize their tradeoffs, then various parameters can be calculated to back up the function $Q\ (t) = \rho\ *[1- e\ -\mu t]$.

Table 1. Queuing behavior of customers

P	*Q*(∞)	*#C*	*R*	*#COR*
0, 0.4	0, 06	4	56, 60	6
0, 0.5	0, 01	4	36, 36	8
0, 0.10	0, 012	6	25, 00	10
0, 0.20	0, 26	9	18, 60	15
0, 0.30	0, 44	13	18, 92	20
0, 0.40	0, 66	21	20,00	24
0, 0.50	1, 00	31	20, 00	56
0, 0.60	1,26	69	26, 68	106
0, 0.70	2,00	145	38,50	256
0, 0.80	5,00	330	36, 68	1086

Where,
ρ = the utilization rate.
$Q(\infty) = \rho/(1-\rho)$ = expected number of customers in the queue when the state is attained.
#C = arrival rate.
$R = \#C/Q(\infty)$,
#COR- it is computed using *Ub* and the relaxation time.
Thus, the arrival number of customers served when the system is at steady state,
From the fourth column,
$\#C \approx 5\,Q(\infty)$ and the first value should be zero for the first customer.
From the equations given herein, the transient time T can always be computed using the formula
$T = \#C/\lambda$.

Assumptions made from the above;
1. The initial state of the queuing system strongly affects the equilibrium behavior of Q (t) during the start -up period.
2. *Q (t)* approaches *Q* (∞) after the start-up period through a decaying exponential system.
3. ρ is the only parameter that influences *Q* (∞).

Social optimization

According to Erlichman and Hassin (2009), deriving the equilibrium threshold strategy, this threshold coincides with n^*, under the LCFS-PR regime. Assuming that n is the maximum length of the queue including the one in the service already, that is, the customer reneges whenever there are n customers in front of him in the service facility, and letting F_n to be the expected welfare of the n^{th} customer in the LCFS-PR queue, then the customer reneges from the current position $n+1$. Under this condition, F_n is monotonic decreasing in *n*.

Also, n^* is the largest n such that $F_n \geq 0$.
Then, the model's parameters becomes;

$$F_n = R\,[(1-\rho)/(1-\rho^{n+1}) - C/\mu\,(1-\rho)\,\{n-(n+1)\,\rho\,[1-\rho^{n}]/1-\rho^{n+1}\}$$

The problem at stake is addressed by looking for an appropriate model adaptation the one presented by Allon and Hanany (2008). The type variety was the key driver for sustainability of the line cutting discipline according to Mailath and Samuelson (2006). Customers with the high cost of time were allowed to overtake the patient ones. In the present case, consumers are homogeneous, but still, line cutting strategy is repeatedly done for patient customers showing that heterogeneity is not needed for analysis of this result. Nevertheless, the statistical findings indicate that there is a negative impact on social welfare, as opposed by Allon and Hanany (2012) as it termed to be beneficial. The issues concerning the social welfare suggest the application of FCFS by the manager to enhance the social well-being of the customers.

A Poisson model is used to evaluate this problem in an individual service from an M/M/1 queueing line whereby the utility function of customers is represented as;

$U= v - c\,E\,[W]$,

Where v is service value, c, cost of time, and expected sojourn time as;E [W].

Poisson process can model this with rate λ. Customers are assumed to be identical and obtain a value, v at the end of service after incurring a waiting cost, c. This further borrows from the analysis of the M/M/1 queue with processor sharing via spectral theory by Guillemin and Boyer (2001). Then, the single equilibrium game is obtained as follows:

Customers are assigned in their arrival order naturally

The customer arrives the queue and observes the length (N^{th} customer is N-1). Then the customer decides to join or balk. If the balking is done the game end for them.

The customer will attempt to cut the line or choose to join the end of the queue (action P and step J).

A customer who receives the request of cutting can accept it (action A). However, an incumbent can also reject the cutting measures (action R) whereby the arrival would always join at the end of the line.

The full criteria of the i$^{\text{th}}$ customer can be given by (E_i, I_i). Whereby, $E_i \in \{J,P\}$, $I_i \in \{R,A\}$

Where;

E_i describes the choices available to customers who arrive, and I_i , is the available incumbents. In this case, one shot equilibrium game can be determined. Also in this shot game, there is no guarantee for allowing customers to cut a line by the available incumbents.

Now consider the case where participants require the services repeatedly. The assumption of no concurrent requests is made between the service delivery and waiting times. However, FCFS strategy which is the equilibrium at the single stage is also the same for the repeated game. Punishment policy is triggered to the one who refuses to agree with the requests hence it is anonymous (Mailath and Samuelson, 2006).

Further, from Mailath and Samuelson's approach to repeated game theory, Let W^R and W^A denote the expected delay time experienced by a patient customer when denying or accepting the cut ahead, respectively in a condition that all others follow the criteria of reducing requests (P, A). Moreover, according to Sun and Li's (2014) approach to the egalitarian patient customer, let V^A be expected payoff when all participants follow the (P, A) the long-term expectation of discounted payoff follows FCFS strategy in each time period, (J, R). With queue-length-dependent strategies, the actions of either cutting ahead or joining the queue have been considered. In this section, equilibrium is explored to analyze whether the procedures vary with the queueing length. This is achieved by focusing on dependent strategies discussed by Mailath and Samuelson (2006).

Also, considering another equilibrium threshold strategies by Sun, Guo and Tian (2009), Let the system size be N at arrival to be N-1, and assume that all customers decide to join rather than cutting ahead, $ne(N)$ be expected place in the queue of the patient customers. Moreover, let p be the threshold defining the queue dependent strategy such that;

$n \leq p$. For $n > p$ they will play action R. Hence W^A, $p\ (n)$ denote the expected delay time for the incumbent and W^R, p for those who refuse to comply with the requests. The discounted expected utility for all periods will be V^A, p.

This enables formulation of a random process which will determine the incumbents whom the cutting ahead requests are made to, by the following probability distribution function,

ƶ$n\ (N) = \{1/n\ (n\text{-}1)$ if n ϵ $\{2...\ N - 1\}$ and $1/\ (n\ \text{-}1)$ if $n = N$.

As N is drawn from the distribution, the arriving customer does not get a chance to cut ahead. The choice of the probability distribution was chosen because the likelihood of being overtaken is of independent queue length. Moreover, the function retains a level of intuitive appeal. Henceforth, the customers' decision will choose to join the line as they arrive when the expected utility for the action is positive that is;
$U = v – c\ E\ [W^N]$, where W^N, (N = 1, 2…) is the delay time for the customer in the system containing N -1 customers as further discussed by Guillemin and Boyer (2001).
Hence customer N will join if and only if $U \geq 0$ which can be represented as
$E\ [W^N] \leq v/c$. However, the most challenge to be encountered here is the determination of sojourn time of customers as elaborated by Okuno-Fujiwara and Postlewaite (1995). Then, let Wn^N denote the expected short stay time of a client to be joining in the n^{th} position in a system of length N and C be maximum line size, which is given as a result of computing of total delay times. A and S can denote the expected arrival time for the next customer, the expected completion time of service for the client's n = 1. Moreover, let I_B = {1, if event B occurs, and 0 if event B does not happen}. To perform the decomposition process, the following workings are done.

$E[\ \mathrm{e}\text{-}^{Sw1N}]= E[\ \mathrm{e}\text{-}^{\mathrm{Sw1N}}\ I\ \{\mathrm{S} < A\}] + E[\ \mathrm{e}\text{-}^{\mathrm{Sw1N}}\ I\{A < S \wedge \text{no cut}\}]$, $n = 1, N \epsilon \{1,\dots, \acute{N} – 1\}$,

$\mathrm{E}[\ \mathrm{e}\text{-}^{SwnN}]\ \ = E[\ \mathrm{e}\text{-}^{Sw1N}\ \mathrm{I}\ \{S < A\}] + E[\ \mathrm{e}\text{-}^{Sw1N}\ I\{A < S\ {}^\wedge\text{cut}\}] + E[\ \mathrm{e}\text{-}^{Sw1N}\ I\ \{S < A\}] + E[\ \mathrm{e}\text{-}^{Sw1N}\ I\{A < S\ {}^\wedge\text{no cut}\}]$ for all $n \epsilon \{2,\dots,N\}$, $N \epsilon \{2,\dots, \acute{N} – 1\}$ which result to

$\mathrm{E}\ [\mathrm{e}\text{-}^{Sw1N}] = \mathrm{E}\ [\mathrm{e}\text{-}^{Sw1N}\ \mathrm{I}\ \{S < A\}]$, n = 1, $N = \acute{N}$ whereby the queue length is given by $\acute{N}$.

This brings about the following computation;

$E\ [Wn^N] = 1/\ (\lambda+\mu) +\mu/\ (\lambda+\mu)\ E\ [W^N\text{-1n-1}] + (1\text{-}1/\mathrm{n})\ \lambda/\ (\lambda+\mu)\ \mathrm{E}\ [W^N+1n+1] + 1/n\ \lambda/\ (\lambda+\mu)\ E\ [W^N+1\mathrm{n}]$.

In this situation, the probability of the service completion fall by one implying that $\lambda/\ (\lambda+\mu)$ is the likelihood of a new event that is the arrival of a new customer in the queue.
Social optimization considers a view of social perspective to maximize the utility accruing to the line of clients (cf, Naor, 1969). Individual optimization is the same no matter the service discipline used given the M/M/1 system. This can be evaluated as;

$\pi N = \ (\rho\ N(1\text{-}\rho))/(1\text{-}\rho\ N+ 1)$, I = 0, 1, 2… $\acute{N}$

Numerical investigations apply the numerical simulations to assist in comparing the variety of strategies about sojourn times. It includes the analysis of customer utility in the threshold line cutting arrangement, in queueing systems. With the combination of the discipline, the cutting points, and social optimization will eventually determine the investigations of the behavior of customers.

CONCLUSION

In brief, as presented by Naor and the rest of the previous works the social-welfare of clients is unimodal on queueing line. However, biasedness is due at the front of the queue, and hence it aggravates into the worst condition of FCFS queue to a certain degree imposing external negativity to all future arrivals of customers. This is significant in consideration of social welfare aspect of the clients in repeated game results. This research contributes to the findings of Naor (1969) in analyzing the queueing systems which apply the FCFS discipline where arriving customers can overtake the necessary one. The used methods and formulae are adapted to produce the expected delay or waiting time when analyzing the joining threshold as previously explained. The joining threshold is numerically compared with EPS and FCFS to obtain optimal threshold optimally. On the other hand, this problem would still be achieved when the queue system cutting is less biased. However, the motivation behind this research is to come up with a program to improve the social welfare of the customers through queue-re-ordering. Moreover, a lot of the investigation would widen the scope of solving the health results while investigating this problem.

REFERENCES

Allon, G., and Hanany, E. (2012). Cutting in Line: Social Norms in Queues. Management Science 58, 493–506.

Boudali, O. and Economou, A. (2012). Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes. European Journal of Operational Research 218, 708–715.

Burnetas, A. and Economou, A. (2007). Equilibrium customer strategies in a single server Markovian queue with setup times. Queueing Systems 56, 213–228.

Erlichman, J., and Hassin, R. (2009). Equilibrium Solutions in the Observable M/M/1 Queue with Overtaking. In Proceedings of the Fourth International ICST Conference on Performance

Guillemin, F., and Boyer, J. (2001). Analysis of the M/M/1 queue with processor sharing via spectral theory. Queueing Systems 39, 377–397.

Helweg-Larsen, M., and LoMonaco, B. L. (2008). Queuing among U2 fans: Reactions to social norm violations. Journal of Applied Social Psychology 38, 2378–2393.

Knudsen, N. C. (1972). Individual and Social Optimization in a Multiserver Queue with a single server. Queueing Systems39, 366–377

Larson, R. C. (1987). Perspectives on queues: Social justice and the psychology of queueing. Operations Research 35, 895–905.

Mailath, G. L. and Samuelson, L. (2006). Repeated Games and Reputations: Long-Run Relationships. Oxford University Press, New York.

Milgram, S., Liberty, H. J., Toledo, R. and Wackenhut, J. (1986). Response to intrusion into waiting lines. Journal of Personality and Social Psychology 51, 683–689.

Naor, P. (1969). The Regulation of Queue Size by Levying Tolls. Econometrica 37, 15–24.

Oberholzer-Gee, F. (2006). A market for time fairness and efficiency in waiting lines. Kyklos 59, 427–440.

Okuno-Fujiwara, M. and Postlewaite, A. (1995). Social norms and random matching games. Operations Research 35, 895–905.

Sun, W. and Li, S. (2014). Equilibrium and optimal behavior of customers in Markovian queues with multiple working vacations. TOP 22, 694–715.

Sun, W., Guo, P., and Tian, N. (2009). Equilibrium threshold strategies in observable queueing systems with setup/close down times. Central European Journal of Operations Research 18, 241 268. Systems. Kluwer Academic Publishers, Norwell, Massachusetts.

Yu, M., Tang, Y. and Wu, W. (2014). Individually and socially optimal joining rules for an egalitarian processor-sharing queue under different information scenarios. Computers and Industrial Engineering 78, 26–32.