

Equilibrium solutions in the observable $M/M/1$ queue with overtaking

Jenny Erlichman and Refael Hassin

School of Mathematical Sciences

Tel-Aviv University

jennybro@post.tau.ac.il, hassin@post.tau.ac.il

ABSTRACT

The subject of this paper is a mechanism that allows customers to overtake others. In our system, customers observe the queue length upon arrival, and have the option of overtaking some or all of the customers already present in the queue. Overtaking is associated with a fixed price per overtaken customer. If a customer chooses to overtake some but not all of the present customers, overtaking applies to the last customers in the queue. Customers incur a fixed cost per every unit of time in the system, and their goal is to minimize their own expected total cost. We would like to characterize the symmetric equilibrium strategies of our model. However, it turns out that this mission is much harder in our system than in the other priority queueing systems analyzed in the literature. We consider several types of symmetric strategies and find out that the set of equilibrium symmetric strategies is quite reach and includes surprisingly odd strategies. In addition, we compare overtaking with ordinary (absolute) priority systems. We assume that the server can induce the customers to choose among the equilibrium strategies, the one which maximizes its profits. Under this assumption, we compare the server's profits in the two models and find, somewhat surprisingly, that the system of overtaking gives the server higher profits.

1. INTRODUCTION

Priority sale in queueing systems is a common mechanism used to improve service and increase profits. In such regimes, a customer has the option of purchasing priority, out of a menu of options, and obtain service before some others who arrived earlier. Of course, later arrivals who purchase higher priority may overtake the customer, and this may serve as a further incentive to purchase priority. It is expected that customers take all this into consideration when choosing their purchase strategy. Since a customer's strategy responds to other customers strategy, the result is an equilibrium strategic behavior.

The subject of this paper is a different mechanism that al-

lows customers to overtake others. In our system, customers observe the queue length upon arrival, and have the option of overtaking some or all of the customers already present in the queue. Overtaking is associated with a fixed price per overtaken customer. If a customer chooses to overtake some but not all of the present customers, overtaking applies to the last customers in the queue. Customers incur a fixed cost per every unit of time in the system, and their goal is to minimize their own expected total cost. Our model assumes a single server Markovian queue where customers are identical in all parameters except for their arrival times.

We exclude balking, and hence there is no question of social optimality here. However, different strategies affect the flow of profit to the server.

We set up two main goals:

- We would like to characterize the symmetric equilibrium strategies of our model. However, it turns out that this mission is much harder in our system than in the other priority queueing systems analyzed in the literature. We consider several types of symmetric strategies and find out that the set of equilibrium symmetric strategies is quite rich and includes surprisingly odd strategies. We characterize some particular families of equilibrium strategies, but it is clear that these are not the only equilibrium strategies. We found some surprising results of unexpected strategies which for some parameters are equilibrium strategies. For example, a strategy like: overtaking a single customer when observing one customer in the system upon arrival, overtaking none when observing two or three, overtaking four customers when observing four, overtaking three customers when observing five, and not overtaking any customer otherwise, can be an equilibrium.
- We compare overtaking with ordinary (absolute) priority systems. We assume that the server can induce the customers to choose among the equilibrium strategies, the one which maximizes its profits. Under this assumption, we compare the server's profits in the two models and find, somewhat surprisingly, that the system of overtaking gives the server higher profits.

One of the interesting findings in our work is that sometimes it is worthwhile to observe a longer queue since then the customers's expected cost is lower.

The paper is organized as follows: In Section 2 we formally present our model. In Section 3 we survey relevant literature on strategic behavior in queueing systems with priorities. In Section 4 we try to characterize the equilib-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Gamecomm October 23, 2009 - Pisa, Italy.

Copyright 2009 ICST 978-963-9799-70-7

rium strategies of our model. We find that it is difficult to characterize all strategies that can be equilibrium and therefore consider some attractive special cases. In Section 5 we analyze equilibria where at most one customer is overtaken. We consider pure and mixed threshold strategies. We give necessary and sufficient conditions for these strategies to define an equilibrium. In Section 6 we consider strategies of overtaking k customers if there are at least k customers, and overtaking all of them otherwise. In Subsection 6.1 we ask whether it is worthwhile to observe a longer queue when all customers follow this strategy and the answer is sometimes yes. In Subsections 6.2 and 6.3 we consider mixed strategies. In Subsection 6.2 we consider mixtures between overtaking k or $k - 1$ customers, and in Subsection 6.3 we consider mixtures with overtaking k , $k - 1$ or $k - 2$ customers. In both cases we give necessary and sufficient conditions for these strategies to define an equilibrium. In Section 7 we compare the server's maximum expected profit per customer under equilibrium conditions in two models. The first is our model, and the second model which is analyzed by Adiri and Yechiali [1], and Hassin and Haviv [9] has two priority classes. In this model two FCFS queues are formed in front of a single server, one for priority customer and the other for ordinary customers. An arriving customer buys priority if and only if the number of customers in the queue is at least a threshold n . There may be numerous equilibria in the both models. We assume that the server can choose the equilibrium which maximizes its expected profit. Under this assumption we prove that the server's expected profit per arrival in our model is greater than the server's expected profit per arrival in the second model. Our proof is partially analytic and partially based on a numerical computations. Finally, Section 8 contains concluding remarks and open problems.

2. MODEL DESCRIPTION

In our observable $M/M/1$ model, customers purchase priority. This priority enables overtaking present customers. Upon arrival a new customer observes the queue length and announces the number of customers that he overtakes. There is a fixed cost C_o per overtaken customer, and customers have homogeneous waiting costs. There is no balking or reneging, and a customer cannot overtake after joining the queue. The service discipline is preemptive resume. Let C_w denote the cost per unit of time to a customer for staying in the system (either waiting or being served). Let C_o denote the cost per overtaken customer. All customers have the same waiting time value C_w . We denote the rate of arrival by λ , and the service rate by μ . Note that the case $C_o < \frac{C_w}{\mu}$ has a trivial unique equilibrium since overtaking all present customers is clearly a dominant strategy. Therefore, we assume $\frac{C_w}{\mu} < C_o$.

In a Nash equilibrium no customer has anything to gain by changing his or her own strategy unilaterally. In a symmetric equilibrium all customers use the same strategy.

Consider a static version of the model, where the number of customers to be served is fixed and they are all present at the time the service begins. Moreover, no future arrivals are expected. In this case there is a unique equilibrium in which no customer overtakes any other customer. To see why this is true note that from the assumption of our model that $\frac{C_w}{\mu} < C_o$, a dominant best response of the *last customer* is not overtaking, therefore a best response of the

customer whose position before the last one is not overtaking either, and if we continue this way the result is that no one is overtaking. In the sequel we show that while never overtaking is always an equilibrium strategy, in the dynamic model there are numerous of the equilibrium strategies as well.

We analyze pure and mixed threshold strategies, and also other strategies in the dynamic model. We look for equilibrium conditions in each of these strategies.

We compare overtaking with ordinary (absolute) priority systems. We assume that the server can induce the customers to choose among the equilibrium strategies, the one which maximizes its profits. Hence, server's goal is to induce all customers to overtake all customers who are currently in the system.

3. RELATED LITERATURE

There is an extensive literature on priority queues. In this section we review the literature on strategic behavior in queueing systems with priorities and server's profit maximizing under different priority regimes.

There are three basic priority disciplines [?, 14]: preemptive resume, preemptive repeat and non preemptive. In a preemptive resume discipline the service of a customer is interrupted when a customer belonging to a higher priority class arrives, and will be resumed from the point of interruption. In a preemptive repeat discipline the service of a customer is interrupted when on arrival of customer belonging to a higher priority class arrives, and will start from the beginning. In a non preemptive discipline the service of the customer is completed even if a customer of higher priority may arrive. In our model the priority discipline is preemptive resume.

- Adiri and Yechiali [1] analyzed an $M/M/1$ model with two priority classes. In their model the priority discipline is preemptive resume and two FCFS queues are formed in front of a single server, one line for priority customers and one for ordinary customers. Upon arrival and after observing the length of the two queues, a customer decides whether to purchase priority. Customers cannot purchase priority while waiting. Adiri and Yechiali assumed a given price for priority and computed an equilibrium solution of the following type: buy priority if and only if the number of customers in the queue is at least a threshold n . We compare the server's expected profit in this model to the server's expected profit in our model.

Hassin and Haviv [9] continued this line of research. They extended the set of possible strategies to include mixed strategies of the following kind: for some non-negative real number $x = n + p$, where n is an integer and $0 \leq p < 1$, a customer who observes a total of k customers in the system joins the ordinary queue if $k \leq n - 1$, does so with probability p if $k = n$, and otherwise buys priority. They show that multiplicity of equilibria is possible. Moreover, in general there is an interval of integer thresholds that define stable equilibria and between each pair of such equilibria there is an unstable mixed equilibrium.

- Lui [11], Glazer and Hassin [5] and Hassin [8] consider a scheme of auctioning or bribery in an unobservable

queue, i.e., at time a customer's need for service arises, he irrevocably either joins the queue or balks. It is not possible for him to observe the queue length before making this decision. In this model, each customer chooses the amount he wishes to pay for priority and then he is placed in the queue ahead of those who paid smaller amounts. It turns out that this decentralized scheme can be used to induce a socially optimal joining rate.

- Myrdal [12] claimed that corrupt officials may deliberately cause administrative delays in service so as to attract more bribes. Lui [11] referred to this claim as Myrdal's hypothesis, and argued that the hypothesis is not always true. For example, if increasing the rate of service is costly to the server, then without a bribe the server has no incentive to supply service, and bribes induce faster service. However, Hassin [8] compared the service rate chosen by a profit maximizer to the socially optimal rate, showing that from this point of view Myrdal's hypothesis is correct. In this paper we show that when the service is slower, i.e., μ is lower, then the server's profit is higher.
- Rosenblum [13] explores a market model where customers trade queue positions. The result is that the customers will be served in decreasing order of value of time, which is known the socially optimal order. But there is a strong assumption in this model that customers do not consider profits that might be gained from transactions in the future, but consider only the reward they receive for the service and their cost of waiting. This model is a kind of overtaking model, but as opposed to our model a customer overtakes other customers only if both he and the overtaken customers agree to this overtaking.
- Hassin, Puerto and Fernandez [10] consider a relative priority approach, where the priority given to a class also depends on state variables associated with other classes. They show that relative priority in an n -class queueing system can reduce server and customer costs. This property is demonstrated in a single server Markovian model where the goal is to minimize a non-linear cost function of class expected waiting times. The priority regimes which they consider are: FCFS, absolute preemptive priorities and DPS (discriminatory processor sharing). Special attention is given to minimizing server's costs when the expected waiting time of each class is restricted.

4. PURE EQUILIBRIUM STRATEGIES

In our model, customers observe the number of people who are already in the system, and then decide how many customers to overtake. We refer to the number of customers that an arriving customer observes as including the customer in service but not including the new customer himself. In this section we analyze pure strategies defined by a vector (k_1, k_2, \dots) , where k_i is the number of customers that a customer who observes i customers in the system upon arrival overtakes. Clearly, $k_i \leq i$.

Define $f_{i,j}$ to be the expected waiting time of a customer given that there are i customers in front of him (including

a customer in service), and j customers behind him. In addition define $f_{-1,j} = 0$.

The expected time till either a service completion or a new arrival occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\mu}{\lambda+\mu}$ the service completion occurs before a new customer arrives, then his expected waiting time is $f_{i-1,j}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs, then the new arrival overtakes the present customer if $k_{i+j+1} > j$, and doesn't overtake if $k_{i+j+1} \leq j$. Hence, we get

$$f_{i,j} = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{i-1,j} + \frac{\lambda}{\lambda+\mu} f_{i+1,j}, \quad k_{i+j+1} > j, \quad (1a)$$

$$f_{i,j} = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{i-1,j} + \frac{\lambda}{\lambda+\mu} f_{i,j+1}, \quad k_{i+j+1} \leq j. \quad (1b)$$

If $k_i \leq K$ for all i for some K , this provides boundary conditions, namely $f_{i,j} = \frac{i+1}{\mu}, \forall j \geq K$.

If a new customer observes i customers, and decides to overtake k customers, his expected waiting cost is $C_w f_{i-k,k} + kC_o$.

The pure strategy (k_1, k_2, k_3, \dots) defines an equilibrium if overtaking k_i customers is a best response of a new customer who observes i customers for $i = 1, 2, \dots$. Therefore, the conditions for equilibrium are:

$$C_w f_{i-k_i,k_i} + k_i C_o \leq C_w f_{i-k,k} + k C_o, \text{ if } i = 1, 2, \dots \text{ and } k = 0, 1, \dots, i.$$

We could not give analytic characterization to the equilibrium strategies. However, we applied numerical analysis to see which strategies are equilibrium for some values of λ , μ and $\frac{C_o}{C_w}$.

We compute strategies $(k_1, k_2, k_3, k_4, k_5, k_6)$ with $k_i = 0, \forall i \geq 7$, i.e. 7! options. In particular, our study show that even strategies like - (1, 0, 0, 4, 3, 0) or (0, 2, 0, 0, 5, 5) can be equilibrium strategies.

For example, Figure 1 shows the values of $(\lambda, \frac{C_w}{C_o})$ for which the strategies (0, 2, 0, 0, 5, 5), (1, 0, 0, 4, 3, 0), (1, 0, 3, 3, 0, 0), and (1, 2, 3, 4, 4, 0) are equilibrium.

5. A THRESHOLD STRATEGY FOR OVERTAKING A SINGLE CUSTOMER

5.1 Overtaking one customer - threshold strategy

The pure threshold strategy σ_n is defined as follows: a new customer overtakes k customers if there are n or more customers in the system, and does not overtake any customer otherwise.

THEOREM 5.1. *The pure threshold strategy σ_n defines an equilibrium if and only if $\frac{1}{\mu-\lambda} \leq \frac{C_o}{C_w} \leq \frac{\mu+\lambda}{\mu(\mu-\lambda)}$.*

The mixed threshold strategy $\sigma_{n,p}$ where $0 < p < 1$, is defined as follows: a new customer overtakes k customers if there are at least $n+1$ customers in the system, overtakes $k-1$ customers if there are at most $n-1$ customers in the system, and if there are exactly n customers in the system he overtakes k customers with probability p , and $k-1$ customers otherwise.

THEOREM 5.2. *The mixed threshold strategy $\sigma_{n,p}$ defines an equilibrium if and only if $\frac{1}{\mu-\lambda} \leq \frac{C_o}{C_w} \leq \frac{\mu+\lambda}{\mu(\mu-\lambda)}$ and $p = p_e$, where $p_e = \frac{(\mu+\lambda)(C_o(\mu-\lambda)-C_w)}{C_o\lambda(\mu-\lambda)}$.*

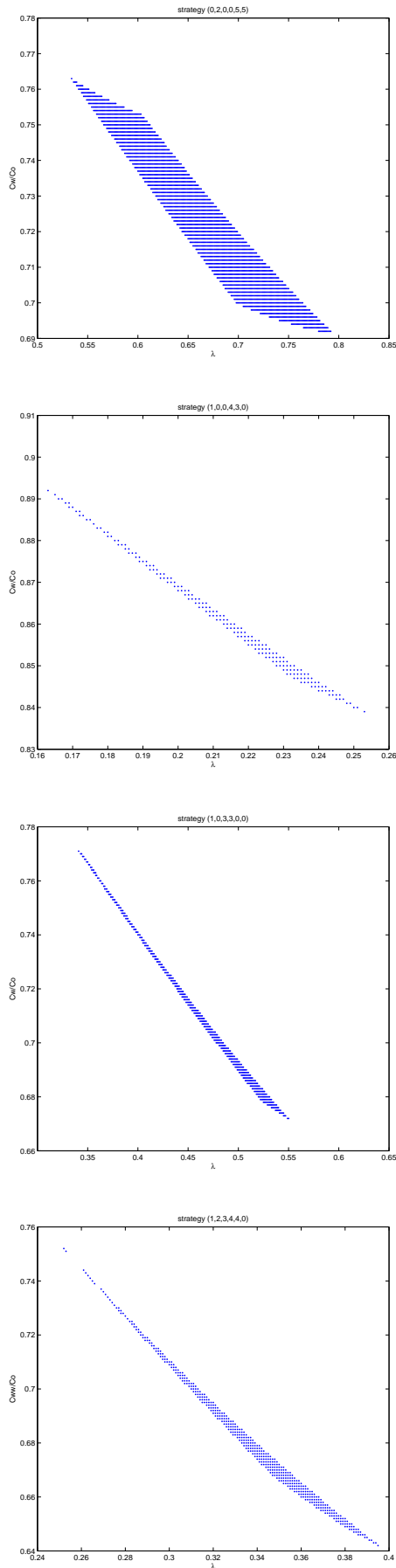


Figure 1: The graphs represent the region of $(\lambda, \frac{C_w}{C_o})$ in which the prescribed strategy defines an equilibrium. $\mu = 1, C_o = 1$.

The proofs are given in Appendix-A.

6. OVERTAKING K CUSTOMERS

In this section we consider strategies of the form $k_i = \min\{k, i\}$, i.e., overtaking k customers if there are at least k customers, and overtaking all of them otherwise. Denote this strategy by Σ_k . We observe that if the strategy of all customers is Σ_0 , i.e., not overtaking others, then from the assumption of our model that $\frac{C_w}{\mu} < C_o$, the best response of a new customer is also not overtaking. Therefore Σ_0 is always an equilibrium. In addition, we show that Σ_∞ is a unique equilibrium when $\frac{C_w}{\mu} < C_o$, $\Sigma_i, i = 0, 1, \dots$ are equilibrium when $\frac{C_w}{\mu} \leq C_o \leq \frac{C_w}{\mu - \lambda}$, otherwise Σ_0 is a unique equilibrium.

THEOREM 6.1. *The strategy $\Sigma_k, k = 1, 2, \dots$ defines an equilibrium if and only if*

$$\frac{C_o}{C_w} \leq \frac{1}{\mu - \lambda}. \quad (2)$$

The proof is given in Appendix-B.

6.1 Overtaking k customers -Is it worthwhile to observe a longer queue?

One of the interesting questions about Σ_k strategy is: Is it worthwhile to observe a longer queue?

The Σ_k strategy enables a customer who observes at least k customers upon arrival to overtake k of them, and by that to ensure that all future customers will not overtake him. In contrast, a customer who observes less than k customers upon arrival cannot ensure that. He just can overtake all present customers, but all future customers will overtake him till his service completion. We find that there are input parameters for which a customer prefers to observe a longer queue.

Denote the number of observed customers by j .

THEOREM 6.2. *Suppose that $\frac{1}{\mu} \leq \frac{C_o}{C_w} \leq \frac{1}{\mu - \lambda}$ and that all customers follow the Σ_k strategy. Then:*

1. *The expected cost as a function of j is built from two linear functions, one for $j < k$, and the second for $j \geq k$.*
2. *If $\frac{\lambda}{\mu(\mu - \lambda)} \leq \frac{C_o}{C_w}$, then the the function is monotone increasing for any j , Figure 2-a.*
3. *If $\frac{C_o}{C_w} < \frac{\lambda}{\mu(\mu - \lambda)}$, then the global minimum is at k , Figure 2-b.*
4. *If $\frac{\lambda}{k\mu(\mu - \lambda)} < \frac{C_o}{C_w}$, then the global minimum is at 0. In addition if $\frac{C_o}{C_w} \leq \frac{\lambda - (\mu - \lambda)(j - k)}{\mu(\mu - \lambda)(k - j')}$, when $j \geq k$ and $j' < k$ then a new customer prefers to observe a longer queue, i.e., j grater than smaller, Figure 2-c.*

6.2 Overtaking k customers - two actions mixed strategy

In this section we consider the mixed strategy $\Sigma_{k,p}$. The mixed strategy $\Sigma_{k,p}$ is defined as follows: For a given integer $k \geq 1$ and a vector $\mathbf{p} = (p_k, p_{k+1}, \dots)$ such that $p_i \in [0, 1]$ for every $i = k, \dots$, a customer who observes upon arrival $i \geq k$ customers in the system (including the one in service) overtakes k customers with probability p_i and $k - 1$ customers otherwise. If there are at most $k - 1$ customers in the system, the customer overtakes them all.

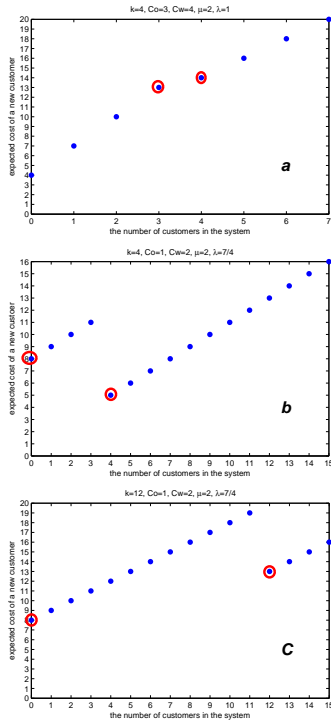


Figure 2: Expected cost as a function of the number of observed customers.

THEOREM 6.3. *The mixed strategy $\Sigma_{k,\mathbf{p}}$ defines an equilibrium if and only if $\frac{1}{\mu} \leq \frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda}$, and for some $x \in \left[\max \left\{ 0, \frac{(\lambda+\mu)^2}{\mu\lambda} \left[\frac{\mu}{\lambda+\mu} - \frac{C_w}{C_o\mu} \right] \right\}, \min \left\{ 1, \frac{(\lambda+\mu)^2}{\mu\lambda} \left[1 - \frac{C_w}{C_o\mu} \right] \right\} \right]$*

$$p_k = x,$$

$$p_{k+1} = \frac{\lambda+\mu}{\lambda} \left[1 - \frac{C_w}{C_o\mu} \right] - \frac{\mu}{\lambda+\mu} x,$$

$$p_{k+j} = 1 - \frac{C_w}{C_o\mu}, \quad \forall j \geq 2.$$

In particular, such x satisfies the following:

$$1. \quad 0 \leq x \leq \frac{(\lambda+\mu)^2}{\mu\lambda} \left[1 - \frac{C_w}{C_o\mu} \right], \text{ if } \frac{1}{\mu} \leq \frac{C_o}{C_w} \leq \frac{(\lambda+\mu)^2}{\mu(\mu^2+\mu\lambda+\lambda^2)}.$$

$$2. \quad 0 \leq x \leq 1, \text{ if } \frac{(\lambda+\mu)^2}{\mu(\mu^2+\mu\lambda+\lambda^2)} \leq \frac{C_o}{C_w} \leq \frac{\lambda+\mu}{\mu^2}.$$

$$3. \quad \frac{(\lambda+\mu)^2}{\mu\lambda} \left[\frac{\mu}{\lambda+\mu} - \frac{C_w}{C_o\mu} \right] \leq x \leq 1, \\ \text{ if } \frac{\lambda+\mu}{\mu^2} \leq \frac{C_o}{C_w} \leq \min \left\{ \frac{1}{\mu-\lambda}, \frac{(\lambda+\mu)^2}{\mu^3} \right\}.$$

For example, suppose that $\frac{C_o}{C_w} = \frac{\lambda+\mu}{\mu^2}$, then $p_k = 0$ and $p_{k+1} = 1$.

We omit the full proof of the Theorem 6.3 but here is the idea why the probabilities behave this way for large values of j . We say that a customer is in state (i, j) if there are exactly i customers in front of him (including the one in service) and exactly j customers behind him. We denote by $f_{i,j}$ the expected (residual) waiting time of a customer in state (i, j) given that all future customers adopt the strategy $\Sigma_{k,\mathbf{p}}$. Consider now customer who observes $k+j-1$ customers upon arrival, where $j \geq 1$. If he overtakes k customers, he guarantees his position in the queue and his expected cost is $C_w \frac{j}{\mu} + kC_o$. Otherwise, if he overtakes only $k-1$ customers, his expected cost is $C_w f_{j,k-1} + (k-1)C_o$. For $\Sigma_{k,\mathbf{p}}$ to define an equilibrium strategy, it must be that the

customer is indifferent between the two options, hence

$$f_{j,k-1} = \frac{C_o}{C_w} + \frac{j}{\mu}, \quad \forall j \geq 1. \quad (3)$$

OBSERVATION 6.4. *A customer who observes $j+k-1$ customers, and overtakes $k-1$ of them will be overtaken till a new arrival customer chooses to not overtake k customers with probability $1-p_{j+k}$. In other words, the time till the choice of an arriving customer is not overtaking has a geometric distribution, with probability $1-p_{j+k}$ for success, and probability p_{j+k} for failure. Hence, the number of customers who arrive till the choice of the arriving customer is not overtaking, is $\frac{1}{1-p_{j+k}} - 1$ (not including the customer who chooses not to overtake). Therefore, the residual expected waiting time of a customer who observes $j+k-1$ customers, and overtakes $k-1$ of them, consists of the service times of all customers who arrive till the first time he isn't overtaken, plus j service times of customers that were before him, plus one service time of himself. Hence, when*

$$j \rightarrow \infty \quad f_{j,k-1} = \frac{j + \left(\frac{1}{1-p_{j+k}} - 1 \right) + 1}{\mu} = \frac{j + \frac{1}{1-p_{j+k}}}{\mu}.$$

Substituting $f_{j,k-1}$ from (3) gives when $j \rightarrow \infty$ $p_j = 1 - \frac{C_w}{C_o\mu}$.

6.3 Overtaking k customers - three actions mixed strategy

In this section we check whether there is an equilibrium strategy, where customers are indifferent between overtaking k , $k-1$ or $k-2$ customers.

Now the mixed strategy $\Sigma_{k,\mathbf{p}}$ is defined as follows: For a

given integer $k \geq 1$ and a matrix $\mathbf{p} = \begin{pmatrix} p_{k-1}^{k-1} & 0 \\ p_{k-1}^{k-1} & p_k^k \\ p_{k-1}^{k-1} & p_{k+1}^k \\ \vdots & \vdots \end{pmatrix}$ such

that $p_i^{k-1}, p_j^k \in [0, 1]$ for every $i = k-1, \dots$ and $j = k, \dots$. A customer who observes upon arrival $i \geq k$ customers in the system (including the one in service) overtakes k customers with probability p_i^k , $k-1$ customers with probability p_i^{k-1} , and $k-2$ customers with probability $p_i^{k-2} = 1 - p_i^k - p_i^{k-1}$. A customer who observes upon arrival $k-1$ customers in the system (including the one in service) overtakes $k-1$ customers with probability p_{k-1}^{k-1} , and $k-2$ customers otherwise. If there are at most $k-2$ customers in the system, the customer overtakes them all.

THEOREM 6.5. *The mixed strategy $\Sigma_{k,\mathbf{p}}$ defines a unique equilibrium where customers are indifferent between overtaking k , $k-1$ or $k-2$ customers if and only if $\frac{1}{\mu} \leq \frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda}$,*

$$p_k^k = x, \quad p_{k-1}^{k-1} = y, \quad p_k^{k-1} = z,$$

$$p_{k+1}^k = \frac{\lambda+\mu}{\lambda} \left[1 - \frac{C_w}{C_o\mu} \right] - \frac{\mu}{\lambda+\mu} x,$$

$$p_{k+1}^{k-1} = \frac{\mu^2(\lambda+\mu)^2 - \frac{C_w}{C_o}\mu(\lambda+\mu)^2 - \mu^2\lambda(\lambda x + (\lambda+\mu)y)}{\lambda(\lambda+\mu)(\lambda^2+\mu\lambda+\mu^2)} - \frac{\mu(\lambda+\mu)}{\lambda^2+\mu\lambda+\mu^2} z,$$

$$p_{k+j}^k = 1 - \frac{C_w}{C_o\mu}, \quad \forall j \geq 2,$$

$$p_{k+j}^{k-1} = 0 \quad \forall j \geq 2,$$

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1, \quad 0 \leq z \leq 1,$$

$$0 \leq p_{k+1}^k \leq 1, \quad 0 \leq p_{k+1}^{k-1} \leq 1,$$

$$x + z \leq 1,$$

$$p_{k+j}^{k-1} + p_{k+j}^k \leq 1 \quad \forall j \geq 1.$$

The equilibrium conditions are not an empty range.

7. PROFIT MAXIMIZATION

In this section we compare two models. In both of them the customers purchase priority, customers are identical except their arrival time, there is no balking or reneging, and decisions are made upon arrival and cannot be changed later. The service disciplines are preemptive resume. We compare the server's maximum expected profit *per customer* under equilibrium conditions.

7.1 Maximum profit in the CP model

The first model is the model of Section 2. In this model purchasing priority enables overtaking present customers. Upon arrival, a new customer decides on the number of current customers that he overtakes, and pays per each overtaken customer a fixed cost. In this model an arriving customer overtakes customers who are currently in the system, but future customers may overtake him. Therefore, we call this discipline *current priority discipline* and denote this model by CP.

We already showed that there may be numerous equilibria in this models. For example, always overtaking k customers, i.e., Σ_k , $k = 0, 1, 2, \dots$, are equilibrium strategies. In particular strategy Σ_∞ , in which an arriving customer overtakes all customers who are currently in the system. Note that Σ_∞ induces a last-come first-served order of service.

Notice that C_o is a parameter that can be changed by a server, as opposed to C_w which is a given parameter. Denote by C_o^* the value which the server chooses in order to maximize its expected profit.

THEOREM 7.1. *The maximum profit in the CP model among all equilibrium strategies is received from Σ_∞ with $C_o^* = \frac{C_w}{\mu - \lambda}$.*

We assume that the server can choose the equilibrium which maximizing its expected profit. Hence, it will choose Σ_∞ strategy in the CP model. The profit from a customer is the cost which this customer pays. In the CP model it is C_o per each overtaken customer. Denote by Π^{CP} the server's expected profit *per customer* in the CP model where an arriving customer always overtakes all present customers, i.e., Σ_∞ strategy, and the price per overtaken customer is the maximum price which satisfies the equilibrium conditions, i.e., C_o^* . The server's expected profit per customer is C_o^*L , where $L = \frac{\lambda}{\mu - \lambda}$ is the expected number of customers in the queue. Therefore,

$$\Pi^{CP} = \frac{\lambda}{(\mu - \lambda)^2} C_w. \tag{4}$$

7.2 Maximum profit in the AP model with threshold $n=0$

The second model which is analyzed by Adiri and Yechiali [1], and Hassin and Haviv [9] has two priority classes. In this model two FCFS queues are formed in front of a single server, one for priority customer and the other for ordinary customers. For a given threshold value $n \geq 0$, an arriving customer buys priority if and only if the number of customers in the ordinary queue is at least n . In other words, this is an absolute priority discipline, and therefore denote this model by AP. If a customer purchases priority then he overtakes all customers in the ordinary queue, and becomes the last customer in the priority queue. The price for becoming a lower priority ordinary customer is 0, this

assumption is without loss of generality since there is no balking or reneging.

Denote by θ the price of purchasing priority, and by $W(n)$ the expected time in the system of the last customer in the ordinary queue when there are no customers in the priority queue and n in the ordinary one, and all use the pure threshold strategy n . The following theorem is proved by Hassin and Haviv [9]:

THEOREM 7.2. *The integer threshold strategy n , $n \geq 1$, specifies an equilibrium if and only if $\theta + \frac{C_w}{\mu} - \frac{C_w}{\mu - \lambda} \leq C_w W(n) \leq \theta + \frac{C_w}{\mu}$. The threshold $n = 0$ specifies an equilibrium if and only if $\theta + \frac{C_w}{\mu} \leq \frac{C_w}{\mu - \lambda}$.*

The profit from a customer is the cost which this customer pays. In the AP model it is θ , if a customer buys priority, otherwise it is zero. Denote by $\Pi^{AP}(n)$ the server's expected profit *per customer* in the AP model as a function of a threshold n , and by θ_{max} the maximum price for buying priority which satisfies the equilibrium conditions. In this section we compute the maximum profit in the AP model with threshold $n = 0$. In this case all customers use the pure threshold strategy $n = 0$, therefore the strategy is always buying priority, and from Theorem 7.2, $\theta_{max} = \frac{\lambda}{\mu(\mu - \lambda)} C_w$, so that $\Pi^{AP}(0) = \theta_{max}$.

Since $\frac{\lambda}{\mu(\mu - \lambda)} C_w < \frac{\lambda}{(\mu - \lambda)^2} C_w$, it follows that, $\Pi^{AP}(0) < \Pi^{CP}$, i.e., the server's expected profit per arrival in the CP model is greater than the server's expected profit per arrival in the AP model with threshold $n = 0$.

7.3 Maximum profit in the AP model with threshold $n \geq 1$

In this section we compute the maximum profit in the AP model with threshold $n \geq 1$. Denote by P_n the probability that the number of customers in the system (both ordinary and priority queues) is at least n , in the AP model under the threshold strategy n . We assume that all customers use the pure threshold strategy $n \geq 1$. From Theorem 7.2, in this case $\theta_{max} = C_w \left[W(n) + \frac{1}{\mu - \lambda} - \frac{1}{\mu} \right] = C_w \left[W(n) + \frac{\lambda}{\mu(\mu - \lambda)} \right]$. Since an arriving customer buys priority if and only if the number of customers in the queue is at least n , $\Pi^{AP}(n) = \theta_{max} P_n = \theta_{max} \left(\frac{\lambda}{\mu} \right)^n$, or equivalently

$$\Pi^{AP}(n) = C_w \left[W(n) + \frac{\lambda}{\mu(\mu - \lambda)} \right] \left(\frac{\lambda}{\mu} \right)^n. \tag{5}$$

LEMMA 7.3.

$$\begin{aligned} W(1) &= \frac{1}{\mu - \lambda} \text{ and } \Pi^{AP}(1) = C_w \frac{\lambda + \mu}{\mu(\mu - \lambda)} \frac{\lambda}{\mu}; \\ W(2) &= \frac{2\mu + \lambda}{\mu^2 - \lambda^2}, \text{ and } \Pi^{AP}(2) = C_w \frac{2\mu^2 + 2\lambda\mu + \lambda^2}{\mu(\mu^2 - \lambda^2)} \left(\frac{\lambda}{\mu} \right)^2; \\ W(3) &= \frac{3\mu^3 + 7\lambda\mu^2 + 4\lambda^2\mu + \lambda^3}{(\lambda + \mu)^2(\mu^2 - \lambda^2)}, \\ \text{and } \Pi^{AP}(3) &= C_w \frac{3\mu^4 + 8\lambda\mu^3 + 7\lambda^2\mu^2 + 4\lambda^3\mu + \lambda^4}{\mu(\lambda + \mu)^2(\mu^2 - \lambda^2)} \left(\frac{\lambda}{\mu} \right)^3. \end{aligned}$$

In these cases $\Pi^{AP}(n) < \Pi^{CP}$.

Since it is difficult to find general expressions to $W(n)$'s, we numerically compute these values. In all cases we found that $\Pi^{AP}(n) < \Pi^{CP}$. Some results are illustrated in the next Subsection 7.4.

7.4 Numerical Analysis of profit maximization

The graphs in Figure 4 present the server's expected profit per customer in the AP model as a function of the threshold

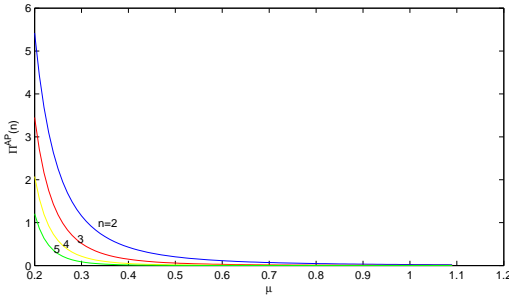


Figure 3: The server's expected profit per customer in AP model as a function of μ .

λ	$\Pi^{AP}(4)$	$\Pi^{AP}(7)$	$\Pi^{AP}(10)$	Π^{CP}
0.1	0.0004	0.0000	0.0000	0.0123
0.2	0.0073	0.0001	0.0000	0.0625
0.3	0.0406	0.0018	0.0001	0.1837
0.4	0.1452	0.0143	0.0012	0.4444
0.5	0.4149	0.0768	0.0126	1.0000
0.6	1.0556	0.3241	0.0895	2.2500
0.7	2.5746	1.2067	0.5149	5.4444
0.8	6.5393	4.4083	2.7315	16.0000
0.9	20.8894	19.3777	16.6511	81.0000

Table 1: Server's expected profit per customer in CP and AP models, $\mu = 1$.

n and arrival rate λ . For every λ the server's expected profit is higher when the threshold n is smaller. There are λ values for which the function is convex, for example $\lambda = 0.3$. There are λ values for which the function is concave, for example $\lambda = 0.99$, and there are λ values for which the function is neither convex nor concave, for example $\lambda = 0.9$.

As presented in Table 1 Π^{CP} is much greater than $\Pi^{AP}(n)$ for all presented parameters. Therefore, the server can obtain a higher profit in our model.

In addition, we see in Figure 3, as expected, that when the service is slower, i.e., μ is lower, the server's profit is higher. This result is expected since there is no balking.

8. CONCLUDING REMARKS

- We did not analyze social optimization since the customers are identical, i.e., we assume homogeneous time values, and in addition we do not allow balking. Hassin [7] observed that if there is an option for balking then priorities have a positive influence on social welfare even if all customers are identical. If the customers are not identical, for example, have heterogeneous time values then apparently in our model the customers with lower time values will be at the end of the queue and others will overtake them. These issues are out of the scope of this study and are left to future research.
- When assuming heterogeneous time values, the simplest and most intuitive model of an incentive compatible pricing scheme is that of Ghanem [6]. Ghanem

proved the intuitive result that for social optimization higher priority should be given to customers with a higher time value.

- Fairness among customers is a fundamental issue for queueing systems. In many situations we notice that customers wish for fair service and fair waiting time. The issue of fairness is raised frequently in the context of evaluating queueing policies and its resolution is not simple at all. Avi-Itzhak and Levy [2] propose a fairness measure enabling to quantitatively measure and compare the level of fairness associated with various queueing systems. They propose yardsticks that can be used as standards for evaluating the fairness of various queueing systems with one class of customers, and for comparing different disciplines to each other. They focused on the issue of customer seniority which is crucial in many queueing systems and used an axiomatic approach to develop fairness measure that is based on this notion. Raz, Avi-Itzhak, and Levy [3] develop a quantitative model for studying priority and classification systems, focusing on the relative fairness of these mechanism. Their analysis provides a measure of fairness for these systems. They limit the discussion to systems where job classification is based only on service characteristics. One of their results is that from fairness perspective, providing preferential service to shorter jobs may be justified in many cases. For comparison, they provide the fairness analysis for an equivalent system where jobs are served in the order of arrivals (FCFS). They conclude that in many cases prioritization of short jobs over the long jobs leads to higher fairness (than that of FCFS), nonetheless, in some cases FCFS is more fair. In queueing systems with priorities the issue is how priorities and preferential service affect fairness. In queueing systems with priorities which involve costs (waiting cost, priority cost) such as our model the issue of how priorities and preferential service affect fairness has not been explored and evaluated at all. This is an interesting subject for future research.

Appendix-A

The proof of Theorem 5.1:

PROOF. When all others apply the pure threshold strategy σ_n , a new customer's best response does not overtake more than one customer, since by overtaking one customer the new customer guarantees his place in the queue and from the assumption of our model that $\frac{C_w}{\mu} < C_o$ there is no benefit in overtaking more than one. In addition, if a customer observes $n - j$ customers, $j = 2, 3, \dots, n$ then not overtaking any customer is the best response since he will never be overtaken, and from the assumption of our model that $\frac{C_w}{\mu} < C_o$ there is no benefit in overtaking any customer.

- Suppose that a new customer observes $n - 1$ customers. If he does not overtake, all future arrivals will overtake him as long as his position is n or more. The time it takes for the new customer till he reaches position $n - 1$ is equal to a busy period. So the customer's expected cost is $\frac{C_w}{\mu} \left(\frac{1}{1-\rho} + n - 1 \right)$. Otherwise, if the new customer overtakes a single customer, he guarantees his place in the queue and his expected cost is

$C_w \frac{n-1}{\mu} + C_o$. In symmetric equilibrium the best response of a new customer is not overtaking. Hence, $\frac{C_w}{\mu} \left(\frac{1}{1-\rho} + n - 1 \right) \leq C_w \frac{n-1}{\mu} + C_o$, and this inequality gives the first condition for an equilibrium, $\frac{1}{\mu-\lambda} \leq \frac{C_o}{C_w}$.

- Suppose that a new customer observes $n+j$ customers, $j = 0, 1, 2, \dots$, and doesn't overtake any customer. Then all future arrivals will overtake him as long as his position is n or more. Hence his waiting time consists of $j+2$ busy periods (reducing the number of customers by $j+2$), plus $n-1$ service periods. His expected cost is then $\frac{C_w}{\mu} \left(\frac{j+2}{1-\rho} + n - 1 \right)$. Otherwise, if a new customer overtakes a single customer, he guarantees his place in the queue and his expected cost is $C_w \frac{n+j}{\mu} + C_o$. In symmetric equilibrium the best response of a new customer is overtaking. Hence, $C_w \frac{n+j}{\mu} + C_o \leq \frac{C_w}{\mu} \left(\frac{j+2}{1-\rho} + n - 1 \right)$, or equivalently $\frac{C_o}{C_w} \leq \frac{\mu+\lambda+j\lambda}{\mu(\mu-\lambda)}$, and $\frac{\mu+\lambda+j\lambda}{\mu(\mu-\lambda)}$ is minimum for $j = 0$. Therefore, $\frac{C_o}{C_w} \leq \frac{\mu+\lambda}{\mu(\mu-\lambda)}$ is the second condition for an equilibrium.

□

The proof of Theorem 5.2:

PROOF. Define $f_i(p)$ to be the expected waiting time in position i , when i is the last customer in the queue, given that all customers follow the strategy $\sigma_{n,p}$.

The expected waiting time in position n is computed as follows: the expected time till either a service completion or a new arrival occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\mu}{\lambda+\mu}$ the service completion occurs before a new customer arrives, in this position his place is guaranteed, and the expected waiting time $f_{n-1}(p)$, consists of $n-1$ service periods. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs, then the new arrival overtakes the n -th customer with probability p and his expected waiting time is $f_{n+1}(p)$. Otherwise, if the new customer doesn't overtake the n -th customer, the n -th customer's position is guaranteed, and his expected waiting time is n service periods. The expected waiting time in position $n+1$ is computed as follows: in this position all future arrivals will overtake the last customer, therefore his waiting time will be the busy period (which means to reduce the number of customers in the system to n), plus $f_n(p)$. Hence, we get

$$f_n(p) = \frac{1}{\mu+\lambda} + \frac{\mu}{\mu+\lambda} \frac{n-1}{\mu} + \frac{\lambda}{\mu+\lambda} \left[p f_{n+1}(p) + (1-p) \frac{n}{\mu} \right]. \quad (A1a)$$

$$f_{n+1}(p) = \frac{1}{\mu} \frac{1}{1-\rho} + f_n(p) = \frac{1}{\mu-\lambda} + f_n(p). \quad (A1b)$$

Substituting (A1b) in (A1a) gives

$$f_n(p) = \frac{1}{\mu-\lambda-\lambda p} \left[n + \frac{\lambda p}{\mu-\lambda} + (1-p) \frac{\lambda n}{\mu} \right]. \quad (A2)$$

Under the pure threshold strategy $\sigma_{n,p}$ the new customer does not overtake more than one customer, since by overtaking one customer the new customer guarantees his place in the queue and from the assumption of our model that $\frac{C_w}{\mu} < C_o$ there is no benefit in overtaking more than one.

- Suppose that a new customer observes n customers. If he overtakes a single customer, he guarantees his

place in the queue and his expected waiting time is n service periods, plus C_o . Otherwise, if he does not overtake, all future arrivals will overtake him, therefore his expected waiting time is a single busy period, plus $f_n(p)$. In equilibrium a new customer is indifferent between overtaking a single customer or not overtaking. Hence, $C_w f_{n+1}(p) = C_w \frac{n}{\mu} + C_o$, or $C_w \left[\frac{1}{\mu-\lambda} + f_n(p) \right] = C_w \frac{n}{\mu} + C_o$. Substituting $f_n(p)$ from (A2) gives $p_e = \frac{(\mu+\lambda)(C_o(\mu-\lambda)-C_w)}{C_o\lambda(\mu-\lambda)}$.

- Because p_e is a probability, we require that $0 < p_e < 1$. The denominator of p_e is always positive, so the numerator must be positive too. Therefore $C_o(\mu-\lambda) - C_w > 0$, or $\frac{C_o}{C_w} > \frac{1}{\mu-\lambda}$, and this is one of the conditions for an equilibrium in a pure threshold strategy. The condition for $p_e < 1$ is $(\mu+\lambda)(C_o(\mu-\lambda) - C_w) < C_o\lambda(\mu-\lambda)$, or $\frac{C_o}{C_w} < \frac{\mu+\lambda}{\mu(\mu-\lambda)}$, and this is the additional condition for an equilibrium in a pure threshold strategy.
- If p_e is an equilibrium strategy, then the best response of a new customer who observes $n-1$ customer is not to overtake. If he does not overtake, he will be n in the system, and his expected waiting time will be $f_n(p)$. Otherwise, if he overtakes, he guarantees his place in the queue and his expected waiting time is $n-1$ service periods, plus C_o . Therefore we should get $C_w f_n(p) < C_w \frac{n-1}{\mu} + C_o$, or $\frac{C_w\lambda}{C_o(\mu-\lambda)} > 0$, and this is always true.

□

Appendix-B

The proof of Theorem 6.1:

PROOF. We divide the proof into two parts.

- Suppose that a new customer observes $j \geq k$ customers. By overtaking k , he guarantees his place in the queue, because behind him there are k customers, and only they will be overtaken by new customers. Overtaking any additional customer costs C_o and saves $\frac{C_w}{\mu}$. By assumption $C_o > \frac{C_w}{\mu}$ there is no reason to overtake more than k customers.

Hence, his expected cost is

$$C_w \frac{j+1-k}{\mu} + k C_o. \quad (B1)$$

Otherwise, if he overtakes i customers, $i < k$, all future customers overtake him till he finishes his service and leaves the system. Therefore his expected waiting time is $j+1-i$ busy periods, and his expected cost is

$$C_w \frac{j+1-i}{\mu-\lambda} + i C_o. \quad (B2)$$

The strategy defines an equilibrium if and only if overtaking k customers is a best response of a new customer. Hence, $C_w \frac{j+1-k}{\mu} + k C_o \leq C_w \frac{j+1-i}{\mu-\lambda} + i C_o$, or $\frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda} + \frac{\lambda(j+1-k)}{\mu(\mu-\lambda)(k-i)}$ for $i = 0, 1, \dots, k-1$ and $j = k, k+1, \dots$. The minimum of $\left\{ \frac{1}{\mu-\lambda} + \frac{\lambda(j+1-k)}{\mu(\mu-\lambda)(k-i)} \right\}$ over $i = 0, 1, \dots, k-1$ and $j = k, k+1, \dots$ is obtained

at $i = 0$ and $j = k$. Therefore the condition is $\frac{C_o}{C_w} \leq \frac{1}{\mu - \lambda} + \frac{\lambda}{\mu(\mu - \lambda)k}$.

- Suppose that a new customer observes $j = 1, 2, \dots, k - 1$ customers, and chooses overtaking all of them. His expected cost is $C_w \frac{1}{\mu - \lambda} + jC_o$. Otherwise, if he chooses overtaking i customers, $i = 0, 1, \dots, j - 1$, his expected waiting time is $j + 1 - i$ busy periods, and his expected cost is $C_w \frac{j+1-i}{\mu - \lambda} + iC_o$. In equilibrium overtaking all customers in the queue should be a best response of a new customer. Therefore $C_w \frac{1}{\mu - \lambda} + jC_o \leq C_w \frac{j+1-i}{\mu - \lambda} + iC_o$ for $j = 1, 2, \dots, k - 1$ and $i = 0, 1, \dots, j - 1$, or $\frac{C_o}{C_w} \leq \frac{1}{\mu - \lambda}$.

□

9. REFERENCES

- [1] Adiri, I. and U. Yechiali (1974), "Optimal priority purchasing and pricing decisions in nonmonopoly and monopoly queues," *Operations Research* **22**, 1051-1066.
- [2] Avi-Itzhak, B. and H. Levy (2004), "On measuring fairness in queues," *Advances in Applied Probability* **36**, 919-936.
- [3] Raz, D., B. Avi-Itzhak and H. Levy (2004), "Classes, priorities and fairness in queueing systems," *Technical Report RRR-21-2004*, RUTCOR, Rutgers University.
- [4] Edelson, N.M. and K. Hildebrand (1975), "Congestion tolls for Poisson queueing process," *Econometrica* **43**, 81-92.
- [5] Glazer, A. and R. Hassin (1986), "Stable priority purchasing in queues," *Operations Research Letters* **4**, 285-288.
- [6] Ghanem, S.B. (1975), "Computing central optimization by a pricing priority policy," *IBM Systems Journal* **14**, 272-292.
- [7] Hassin, R. (1985), "On the optimality of first-come last-served queues," *Econometrica* **53**, 201-202.
- [8] Hassin, R. (1995), "Decentralized regulation of a queue," *Management Science* **41**, 163-173.
- [9] Hassin, R. and M. Haviv (1997), "Equilibrium threshold strategies: the case of queues with priorities," *Operations Research* **45**, 966-973.
- [10] Hassin, R., J. Puerto, and F. R. Fernandez (2009), "The use of relative priorities in optimizing the performance of a queueing system," *European Journal of Operations Research* **193**, 476-483.
- [11] Lui, F.T (1985), "An equilibrium queueing model of bribery," *Journal of Political Economy* **93**, 760-781.
- [12] Myrdal, G. (1968), *Asian Drama: An Inquiry into the Poverty of Nations*, Pantheon, New York.
- [13] Rosenblum, D.M. (1992), "Allocation of waiting time by trading in position on a G/M/s queue," *Operations Research* **40**, 338-342.
- [14] Conway, R.W., W.L. Maxwell, and L. W. Miller, *Theory of scheduling*, Addison-Wesley Pub. Co., Reading Mass, 1967.

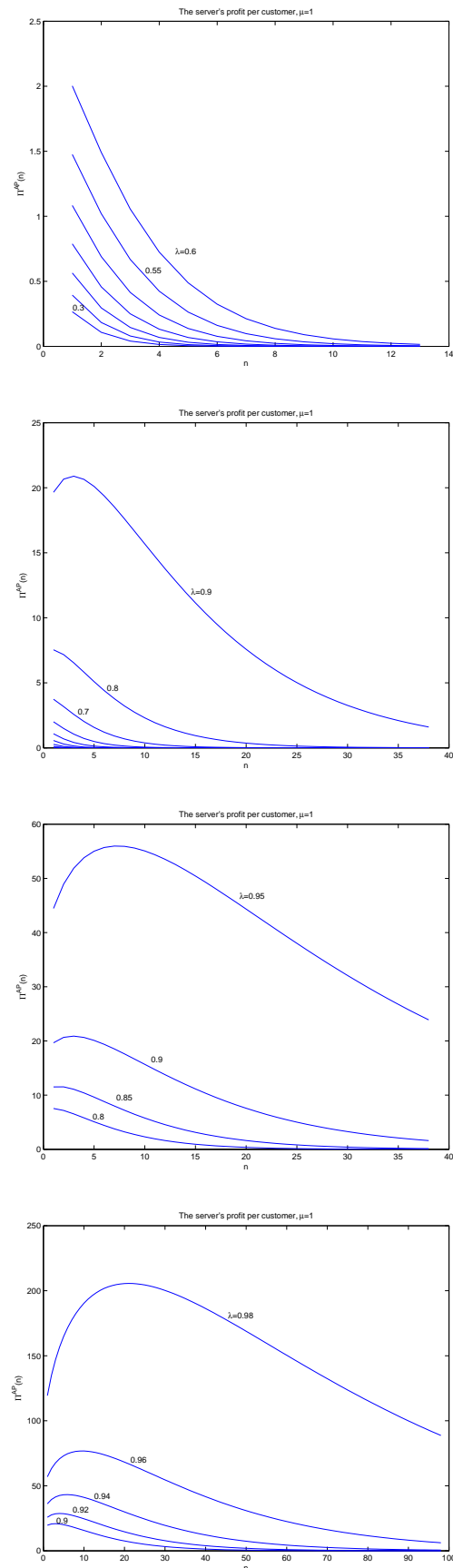


Figure 4: The server's expected profit $\Pi^{AP}(n)$ as a function of the threshold n and arrival rate λ .