Strategic overtaking in a monopolistic M/M/1 queue

Jenny Erlichman and Refael Hassin

***Abstract*—This paper analyzes strategic overtaking equilibria in a single server queue, where customers observe the queue length and have the option of overtaking some of the customers already present in the queue by paying a fixed amount per overtaken customer. Customers incur linear waiting costs, and act to minimize their expected total cost. Characterizing the symmetric Nash equilibrium strategies is much harder than in other priority queueing systems analyzed in the literature. The paper generates two sets of results: (i) customer equilibrium characterization for a fixed overtaking fee C_o. The set of equilibrium symmetric strategies is rich and includes surprisingly odd strategies. Some may look counter-intuitive. (ii) selling a position in the queue could be more profitable to the server than selling a priority.**

***Index Terms*—Queues: Priority, optimization, Nash equilibrium**

I. Introduction

The subject of this paper is a new mechanism for pricing and ordering service in the M/M/1 observable queue, such as the main attraction in a theme park. In such systems, customers make strategic decisions based on the state of the queue at the time of their arrival. We focus on a *monopolistic model* where customers have no choice but to obtain the service from this server. (This terminology is borrowed from [1], and it simply means that customers must buy the service,) This contrasts *non-monopolistic models* where customers attribute some value to obtaining the service, and can balk (possibly to obtain the service elsewhere) if they find that their expected waiting costs exceed that value. We start by surveying the literature on non-monopolistic and monopolistic queueing models. This survey helps in positioning our contribution, but it also contains novel observations regarding profit maximization in observable queues.

A. Non-monopolistic systems

Naor [29] was the first to consider pricing in such systems. Naor's model assumes an M/M/1 system with service rate μ, homogeneous customers, a first-come first-served (FCFS) regime, a fixed value of service, and linear cost of C per time unit of waiting. Naor showed that if the overall (social) welfare (which is defined as the total expected net benefit of the members of the society, including both customers and servers) is to be maximized then for some threshold value n^*, customers should join the queue iff its length is at most n^*-1. However, self-interested customers use, in general, a higher threshold. A single admission price is sufficient to induce optimal joining behavior of self-optimizing customers. This price is different from the profit maximizing price, because the server cannot fully extract the customers' welfare by a single price, and therefore its objective differs from the social objective. In particular, customers who arrive while the queue is short enjoy higher net utility than those who arrive to a longer queue (utility functions of individual customers are identical and additive, from the public (social) point of view).

The server's profit is bounded by the maximum social welfare generated by the system. A server can collect this amount iff two conditions are satisfied: (i) The socially optimal behavior is maintained, in particular customers join according to the threshold n^*, and (ii) the server can fully extract the customers' welfare. This is not achievable in Naor's system with a single price. However, we make the straightforward observation that if the server could set dynamic prices, then by charging $p(n) = R - C \cdot \frac{n+1}{\mu}$ from a customer who observes $n < n^*$ customers upon arrival, and a higher price otherwise, the two conditions will be achieved (R is a customers benefit from completed service). Thus, dynamic pricing can achieve the upper bound on server's profits, and also social optimality (admission fees are considered transfer payments and do not affect social welfare). This scheme is not restricted to the FCFS regime. We refer to its implementation in the last-come first-served (LCFS) queue below.

Following a natural objection to price discrimination, one may claim that dynamic pricing is unfair in addition to being hard to implement, and hence Naor's profit maximizing price can be considered as second-best optimization.

Priority sale in queueing systems is commonly 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 overtake others who arrived earlier. Of course, a low priority customer may be overtaken by later arrivals who purchase higher priority, and this serves as a further incentive to purchase priority. Customers take all this into consideration when choosing their purchase strategy. Since a customer's strategy responds to other customers' strategy, the result is a (Nash) equilibrium strategic behavior.

Adiri and Yechiali [1] considered another second-best optimization problem. There is a given number, m, of priority classes. The server sets prices for these classes. Arriving customers observe the number of customers in each class, and decide what priority to buy. The equilibrium has the following structure: For some vector of m thresholds, one for each priority class, a new customer buys the lowest priority such that the number of existing customers in this class is smaller than its threshold.

Hassin [14] proved that social optimality can be achieved *without the use of any prices*, simply by implementing a LCFS regime with service preemption. The last customer in the queue (the one who arrived first among current customers)

School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel, {jennybro,hassin}@post.tau.ac.il

Research supported by Israel Science Foundation Grants no. 526/08 and 1015/11.

decides whether or not to balk, and this decision has no impact on other customers. Hence the decision affects the customer exactly in the same way that it affects social welfare, and it is carried out in the socially optimal way. A consequence is that if the server installs a LCFS regime and charges a *single non-refundable* entry fee equal to the expected utility of a joining customer, then the joining process will be optimal and the servers profit equal to its upper bound (the maximum possible social welfare). Hassin discussed the drawbacks of such a model: it might be considered unfair, it is difficult to maintain, and customers may leave the system and try to return as new arrivals.

Alperstein [4] considered profit maximization in the same model. While the derivation there is technical, the qualitative results can be easily explained. Suppose that the thresholds are 1 for priority classes $i = 1, \ldots, n^*$, so that an arriving customer buys the lowest priority that has no current customer, and balks if all n^* priorities are taken. To achieve this strategy set the price for priority i to be the expected utility of a customer who buys this priority assuming that all others behave according to the unit threshold strategy. This behavior is an equilibrium under the stated strategy: Buying the lowest available priority (or balking when all priorities have at least one present customer) gives zero net expected utility, while any other act gives non-positive net expected utility. The result is a LCFS regime, customers behavior is socially optimal, and the server's profit attains its upper bound. An advantage of this model is that although the outcome is again LCFS among the customers who obtain service, customers may not feel it is unfair because they choose the type of priority to purchase. Also, those who pay eventually obtain service and those who balk do not incur any costs, whereas under the LCFS regime with a single price, the waiting costs of reneging customers are not refunded.

Alperstein's price for class i is $\theta_i = R - \frac{C}{\mu}\sum_{j=0}^{i-1}\rho^j$, $i = 1, \ldots, n^*$. Note that $\frac{1}{\mu}\sum_{j=0}^{i-1}\rho^j$ is the expected length of a busy period in an M/M/1/i system, which is exactly the expected waiting time of a customer who buys i-th priority. Therefore, the net expected utility of the customer is 0. It can be viewed as a *two-part tariff* consisting of a fixed cost $\theta_0 = R - \frac{C}{\mu}\sum_{j=0}^{n^*-1}\rho^j$, and *nonlinear* overtaking charge. The cost of the first overtaking is $\frac{C}{\mu}\rho^{n^*-1}$, the k-th overtaking costs $\frac{C}{\mu}\rho^{n^*-k}$. Thus, the cost of overtaking a customer increases with the number of overtaken customers.

Alperstein did not prove uniqueness of the claimed equilibrium. In fact, the equilibrium is not unique, since buying any higher than the lowest available priority is also a best response. But it can be made unique by an appropriate small perturbation of the prices that slightly discourages overtaking. After such a perturbation the claimed equilibrium is unique.

This solution is that it requires implementing a whole menu of n^* priorities and prices, and this is a well recognized drawback. For example, the paper of Cachon and Zhang [7] is devoted to replacing an optimal menu of contracts in a supply chain by a suboptimal but simple alternative mechanism. Hall, Kopalle, and Pyke [19] (p. 421) write that "the constant pricing policy is much easier to compute and communicate to relevant managers ... (and) is less likely to cause dissonance among customers who experience widely varying prices ..." Cachon and Feldman [6] (p. 245) write that "a two-part tariff may not be desirable ... a customer may dislike being charged twice for the same service ... firms might prefer to forgo the additional revenues from using a two-part tariff to save the transaction costs and the administrative burden required for its implementation. Disney, for example, initially charged consumers both to get into the park and for specific rides within the park, but later it abandoned per-ride charges." Similarly, Tong and Rajagopalan [34] is focused on obtaining second-best solutions that use a single price (fixed or per time unit) rather than the non-practical complicated optimal solution.

B. Monopolistic systems

The prices set by monopolies are usually regulated. A common rule is that the lowest quality/priority service is free, or almost free, of charge. This is the case, for example, in the observable model of [1], and in the unobservable models of [17] §4.2, Gilland and Warsing [12], and Guo, Lindsey and Zhang [13], but it is not satisfied in Alperstein's two-part tariff. Assuming that customers are homogeneous, and since balking is not possible, any time conserving priority rule is associated with the same expected social welfare. The main question in such models is therefore profit maximization.

Adiri and Yechiali gave special attention to the monopolistic version of their model with $m = 2$, and the lower priority is costless. Therefore, as in Naor's model, there is just a single price to be set. They numerically solved an example depicting the server's profit as a function of the fee and identified its optimum. Hassin and Haviv [16] showed that the equilibrium strategy is not unique, and in fact the number of such strategies is unbounded. They left the question of profit maximization under multiplicity of equilibria open, and we treat it below.

The subject of this paper is a different mechanism that allows overtaking, and we focus our research on the monopolistic version. The overtaking model which we offer in this paper can also be viewed as having many classes with one customer per each class, but this is achieved with just a single price. 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 sojourn time in the system, and their goal is to minimize their own expected total cost. As more customers overtake others, the more inclined an individual should be to do so himself. In other words, the nature of the situation is of *follow the crowd* (FTC) behavior [16] which typically leads to a multiplicity of equilibrium solutions.

As in any priority discipline, the benefit of using our proposed scheme may be enhanced when customers are heterogeneous, but we prefer to emphasis the potential benefits of this novel regime by assuming homogeneous customers.

C. Main goals and results

1) We are interested in a simple single price mechanism that increases the server's profits.
2) We assume a monopolistic firm, and therefore customers must have the option of obtaining the service at a fixed cost which we assume without loss of generality to be 0.
3) The proposed model results in a rich collection of equilibrium solutions, and we would like to characterize the symmetric equilibrium strategies. However, it turns out that this mission is much harder in our system than in related priority queueing systems analyzed in the literature. We consider several types of strategies and find out that the set of equilibrium 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. 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.
4) One of the counter-intuitive findings in our work is that sometimes it is worthwhile to arrive to a longer queue since then the customers' expected cost is lower.
5) We solve the maximum profit among all possible prices and equilibria in the Adiri and Yechiali model.
6) We show that charging a single price for overtaking gives higher profits than charging a single price for priority.
7) Our model is fundamentally different from that of Alperstein, though some of the equilibrium policies we investigate resemble the behavior induced by the optimal pricing in Alperstein's paper. In particular, when the number of priority classes is restricted to k the customers' induced strategy has unit thresholds at all but the highest priority, exactly as the strategy Σ_k which we consider in Section IV. However, this behavior results in Alperstein's paper from a completely different game, in which customers buy priorities, and a higher priority supplies a stronger protection against overtaking by future arrivals. For example, a customer who purchases the highest priority is assured of never being overtaken. In our model, this is not possible. Also, in contrast to the (nonlinear) costs assumed by Alperstein, our analysis concentrates on the case where the cost of overtaking a customer is at least c/μ. Thus the question is how many of the existing customers it is worth overtaking in equilibrium, whereas with Alperstein's prices, all of the existing customers are always overtaken. For example, not overtaking any customer is always an equilibrium in our model, but never so under optimal priority pricing.
8) It is important to emphasize that overtaking in our model is a result of strategic choice of customers and not a consequence of a given structure, such as, for example, in the model of Whitt [36] and the work which followed it.

D. Additional literature

We now describe additional literature on equilibrium in priority queues. Dolan [9] considers an observable queue with customers that are homogeneous except for that they have different waiting costs which are their private information. Balking is not allowed, and social welfare is maximized when customers with higher waiting cost obtain priority over those with a lower cost. This can be achieved by the use of *Clarke prices*: each customer declares his waiting cost, obtains priority accordingly, and pays for the externalities he imposes assuming that all customers truthfully declare their costs. It turns out that under these prices, truthful declaration is an equilibrium strategy. Mendelson and Whang [27] analyze the equilibrium behavior in an unobservable queue where heterogeneous customers choose their priority out of a menu set by the server. They show that the $c\mu$ priority rule can be used with an appropriate price menu to maximize social welfare even when the customer's type is her private information. Afèche [2] considers the same model and shows that in certain cases revenue maximization is obtained by the same priority rule but while also artificially inflating the service time of a low priority class. Gilland and Warsing [12] consider revenue maximization in a multi-priority unobservable system with waiting cost rates that are uniformly distributed over [0,1], and show that the solution also minimizes the expected total delay costs. Lui [25] Glazer and Hassin [11], Hassin [15], and Afèche, and Mendelson [3] consider an auctioning scheme, where 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. Myrdal [28] claimed that corrupt officials may deliberately cause administrative delays in service so as to attract more bribe payments. Hassin [15] 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, the server's profit is higher. Rosenblum [31] 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 the socially optimal order. This model is a kind of overtaking model where a customer overtakes other customers only if both agree to this overtaking. Larson [24] describes an example of tugboats that may increase their speed to the maximum (while increasing their fuel consumption) to avoid being overtaken, resulting in a socially suboptimal equilibrium. Similarly, Hassin and Haviv [17] §4.2 analyze the unobservable version of Adiri and Yechiali's model, and show that it might be that all customers purchase priority in equilibrium to protect themselves from being overtaken. The result is FCFS discipline, but all customers are now worse-off. In our model, costly overtaking of others in order to avoid being overtaken is a key phenomenon. In Section IV we describe an infinite set of equilibrium solutions that may be viewed as generalizations of the one observed by Larson. Shenker [32] considers a finite number of customers whose utility is a convex function of

their demand rate and their expected queueing demand. A *fair share service discipline* is used to regulate the equilibrium rates of demand: Every customer obtains the highest priority for a portion of its demand of the size of the smallest demand rate of any customer. Then, recursively, the customers with higher demand are given the lower priority levels. Hassin and Haviv citeHH06 and Hayel and Tuffin [21] show that using relative priorities can improve both social welfare and profit maximization when the choice of prices is restricted. Hassin, Puerto and Fernandez [20] consider a multiclass model with relative priorities, 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 both the server's and customers' costs. Sun, Guo, Tian, and Li SGT09 extend this model allowing class-dependent service rates. Equilibrium behavior in priority queues is the subject of Chapter 4 of Hassin and Haviv [17]. Özekici, Li and Chou [30] consider a model of *impolite customers*, where a customer who arrives when there are m customers in the queue joins the k-th position ($1 \leq k \leq m+1$) with probability $P(m,k)$. They show that the more impolite are the customers the bigger is the variance of the waiting times.

E. Organization of the paper

In Section II we formally present our model. In Section III we numerically compute equilibrium strategies in our model. In Section IV we consider strategies of overtaking k customers if there are at least k customers in the system, and overtaking all of them otherwise. In Subsection IV-A we ask whether it is worthwhile to encounter a longer queue when all customers follow this strategy, and the answer is sometimes positive. In Subsections IV-B and IV-C we consider mixed strategies equilibria where at most one customer is overtaken. In Section V we compare the server's maximum expected profit per customer under equilibrium conditions in two models. The first is our model, and the second is the model analyzed in [1] and [17], in which there are two priority classes. Finally, Section VI contains suggestions for future research.

II. Model description

In our observable M/M/1 model, customers purchase priority, and this priority enables overtaking present customers. 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. We assume that there is no balking, and a customer cannot renege or 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 (waiting or in service). All customers have the same value C_w. We denote the rate of arrival by λ, and the service rate by μ.

Remark 2.1: There are four parameters in our model, namely, λ, μ, C_o and C_w. By normalizing time and cost we remain with just two, $\rho = \frac{\lambda}{\mu}$ and $\nu = \frac{C_o \mu}{C_w}$, as in Naor's model. In some cases, especially when the recursive equations are involved, we prefer the four parameter representation, but the main results and computational experiments are better expressed with the normalized parameters.

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 $\frac{C_w}{\mu} < C_o$, or equivalently

$$\nu > 1. \tag{1}$$

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 a queue is sequentially formed but the number of customers to be served is fixed and they are all present at the time the service begins, and 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 by (1), 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 there is no overtaking. In the sequel we show that while never overtaking is always an equilibrium strategy, in the dynamic model there are numerous other equilibrium strategies.

III. Pure equilibrium strategies

In our model, customers observe the queue and then decide how many to overtake. In our terminology, the number of customers that an arriving customer *observes* includes the customer in service, but not the new customer himself. In this section we analyze strategies defined by a vector $(k_1, k_2, \ldots)$, where k_i is the number of customers that an arriving customer who observes i customers, overtakes. Clearly, $k_i \leq i$.

Let $f_{i,j}$ denote 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, so that the total number of customers in the system is $i+1+j$. In addition define $f_{-1,j} = 0$. Then,

$$\begin{aligned} f_{i,j} &= \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{i-1,j} + \frac{\lambda}{\lambda+\mu} f_{i+1,j}, \quad k_{i+j+1} > j, \\ f_{i,j} &= \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{i-1,j} + \frac{\lambda}{\lambda+\mu} f_{i,j+1}, \quad k_{i+j+1} \leq j. \end{aligned}$$

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

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 strategy $(k_1, k_2, k_3, \ldots)$ defines an equilibrium if overtaking k_i customers is a best response of a new customer who observes i customers for $i = 1, 2, \ldots$. Therefore, the conditions for equilibrium are: $C_w f_{i-k_i,k_i} + k_i C_o \leq C_w f_{i-k,k} + kC_o$, for $i = 1, 2, \ldots$ and $k = 0, 1, \ldots, i$.

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 $\frac{C_o}{C_w}$.

Table I contains a list of all strategies $(k_1, k_2, k_3, k_4, k_5, k_6)$ with $k_i = 0, \forall i \geq 7$, i.e $7! = 5040$ options, such that at least for some values of $1 \leq \nu \leq 2$ and $0.1 \leq \rho \leq 0.9$ they define an equilibrium. Figure 1 shows the values of $\left(\rho, \frac{1}{\nu}\right)$ 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. Actually, table I and figure 1 present you the potential to many equilibrium strategies.

IV. Overtaking k customers

In this section we consider the strategy Σ_k 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. We observe that if the strategy of all customers is Σ_0, i.e., not overtaking others, then from (1) it follows that the best response of a new customer is also not overtaking. In addition, we have already mentioned that Σ_∞, or equivalently, $k_i = i$ for all i, is the only equilibrium strategy when (1) is violated, i.e., $C_o < \frac{C_w}{\mu}$. Theorem 4.1 states that Σ_k, $k = 1, 2, \ldots$ are equilibrium strategies when $\frac{C_w}{\mu} \leq C_o \leq \frac{C_w}{\mu - \lambda}$. These results are summarized in Figure 2.

Theorem 4.1: Σ_k, $k = 1, 2, \ldots,$ defines an equilibrium iff

$$1 \leq \nu \leq \frac{1}{1-\rho}. \tag{2}$$

The proof is given in the thesis [10], Theorem 5.1.

A. Overtaking k customers -Is it preferable to encounter a longer queue?

One of the interesting questions about Σ_k strategy is: Is it preferable to encounter a longer queue? Strategy Σ_k prescribes a customer who observes at least k customers upon arrival to overtake k of them, and by that to ensure that future customers will not overtake him. In contrast, a customer who observes less than k customers upon arrival cannot ensure that. He can only 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.

Theorem 4.2: Suppose that $1 \leq \nu \leq \frac{1}{1-\rho}$, and that all customers follow the Σ_k strategy. Denote the number of observed customers by j. Then:

1) The expected cost as a function of j is built of two linear functions, one for $j < k$, and the second for $j \geq k$.
2) If $\nu \geq \frac{\rho}{1-\rho}$, then the function is monotone increasing for any k (Figure 3***a***).
3) If $\nu < \frac{\rho}{k(1-\rho)}$, then the global minimum is at k (Figure 3***b***). Otherwise, if $\nu > \frac{\rho}{k(1-\rho)}$, then the global minimum is at 0 (Figure 3***c***).
4) If $\nu < \frac{\rho-(1-\rho)(j-k)}{(1-\rho)(k-j')}$ for some $j \geq k$ and $j' < k$, then a new customer prefers to observe j to j' customers, in other words prefers to observe a longer queue. (For example, Figure 3***c***, with $j' = 3$ and $j = 13$).

The proof is given in the thesis [10], Theorem 5.2.

For example, in Figure 3***a*** (where $k = 4$), the function is monotone increasing which means that an arriving customer always prefers to observe a shorter queue. In Figure 3***b*** (again with $k = 4$) an arriving customer prefers to observe 4 rather

k_1	k_2	k_3	k_4	k_5	k_6
0	0	0	0	0	0,6
0	0	0	0	1	1
0	0	0	0	2	2
0	0	0	0	3	3
0	0	0	0	4	3,4
0	0	0	0	5	0,4,5,6
0	0	0	1	1	0,1
0	0	0	2	2	0,1,2
0	0	0	3	3	0,2,3
0	0	0	4	0	0,6
0	0	0	4	3	0
0	0	0	4	4	0,3,4
0	0	0	4	5	0,5
0	0	1	1	0	0
0	0	1	1	1	0,1
0	0	2	2	0,1	0
0	0	2	2	2	0,1,2
0	0	2	2	3	3
0	0	3	3	3	1,2,3
0	0	3	3	4	4
0	0	3	4,0	0	0
0	0	3	0	5	5
0	0	3	3	0,2,3	0
0	0	3	4	4	0,3,4
0	0	3	4	5	5
0	1	1	0,1	0	0
0	1	1	1	1	0,1
0	1	1	2	2	2
0	2	0	0	0	0
0	2	0	0	5	5,6
0	2	0	4	4	0,4
0	2	0	4	5	5
0	2	2	0,1	0	0
0	2	2	2	0,1	0
0	2	2	2	2	0,1,2
0	2	2	3	3	0,3
0	2	2	4	4	4
0	2	3	3	0	0
0	2	3	3	3	0,1,2,3
0	2	3	4	4	0,4
0	2	3	4	5	5
1	0	0	0	0	0
1	0	0	0	4	3,4
1	0	0	0	5	4,5,6
1	0	0	4	3	0
1	0	0	4	4	0,3,4
1	0	0	4	5	0,5
1	0	3	3	0	0
1	0	3	3	3	0,2,3
1	0	3	4	4	4
1	0	3	4	5	5
1	1	0	0	0	0
1	1	0	0	2	2
1	1	1	0,1	0	0
1	1	1	1	1	0,1
1	1	2	2	2	0,2
1	1,2	3	3	3	0,2,3
1	1	3	4	4	4
1	2	0,2	0	0	0
1	2	2	2	0,1	0
1	2	2	2	2	0,1,2
1	2	2	3	3	3
1	2	2	4	4	4
1	2	3	3	0	0
1	2	3	4	4	0,4

TABLE I
EACH OF THE ABOVE IS AN EQUILIBRIUM FOR SOME ρ AND ν.

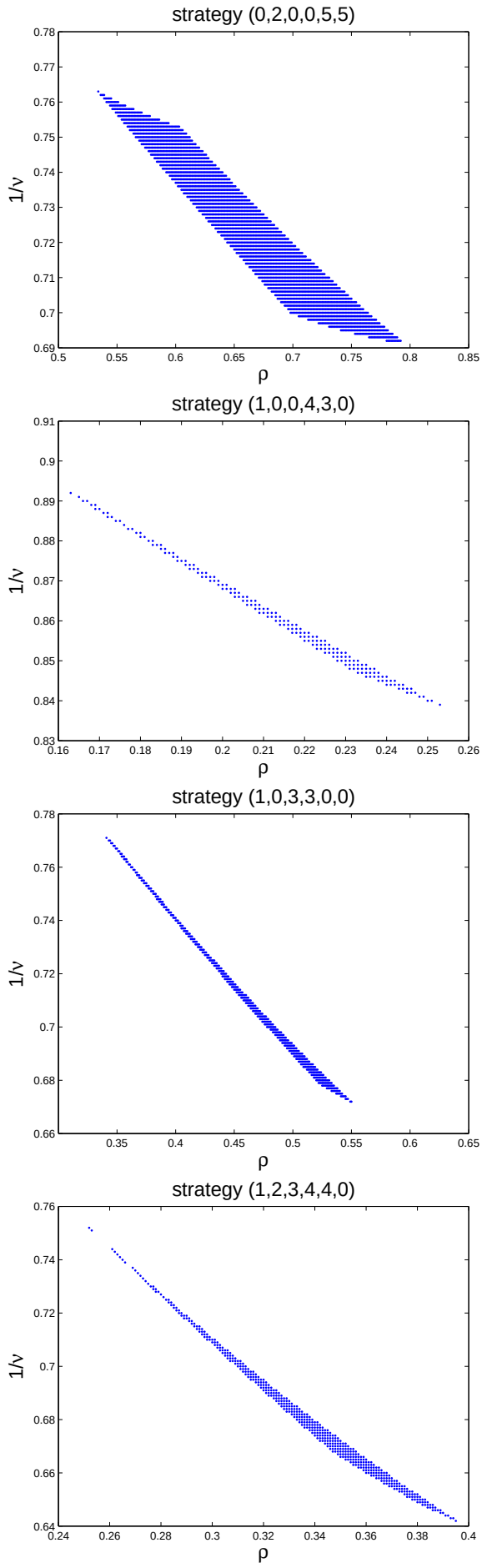

Fig. 1. The region in which the prescribed strategy defines an equilibrium.

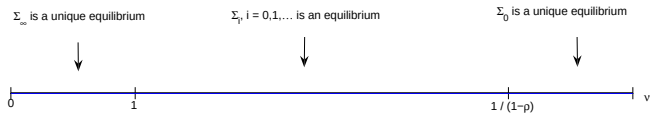

Fig. 2. The region in which Σ_i, $i = 0, 1, \ldots$ is an equilibrium.

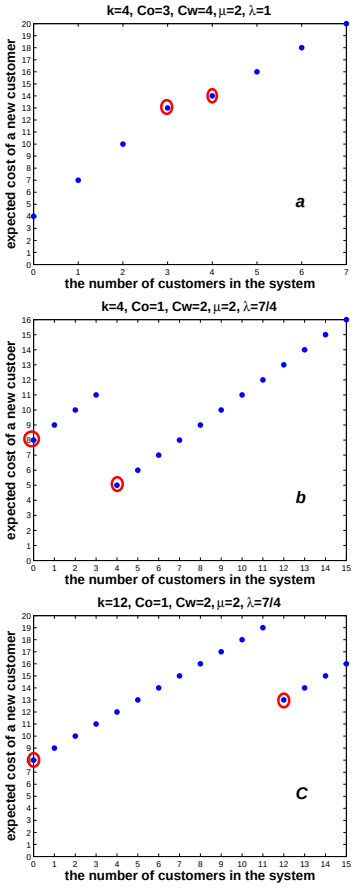

Fig. 3. Expected cost as a function of the number of observed customers (the red circles represent the comparison in example that is detailed after theorem 4.2).

than 0 customers in the system. If he observes 4 customers, then he overtakes all of them and all future arrival customers do not overtake him, and his expected cost is $\frac{C_w}{\mu} + kC_o = 5$. Otherwise, if he observes 0 customers, then all future arrival customers overtake him, and his expected cost is $\frac{C_w}{\mu-\lambda} = 8$. In Figure 3***c*** (where $k = 12$) an arriving customer prefers to observe 0 rater than 12 customers in the system. If he observes 12 customers, then he overtakes all of them and all future arrival customers do not overtake him, and his expected cost is $\frac{C_w}{\mu} + kC_o = 13$. Otherwise, If he observes 0 customers, then all of them and all future arrival customers overtake him, and his expected cost is $\frac{C_w}{\mu-\lambda} = 8$.

B. Overtaking k customers - two actions mixed strategy

The mixed strategy $\Sigma_{k,\mathbf{p}}$ is defined as follows: For a given integer $k \geq 1$ and a vector $\mathbf{p} = (p_k, p_{k+1}, \ldots)$ such that $p_i \in [0,1]$ for every $i = k, \ldots$, 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.

Theorem 4.3: $\Sigma_{k,\mathbf{p}}$ defines an equilibrium iff $1 \leq \nu \leq \frac{1}{1-\rho}$, and for some x such that, $\max\left\{0, \left(\sqrt{\rho}+\frac{1}{\sqrt{\rho}}\right)^2\left(\frac{1}{1+\rho}-\frac{1}{\nu}\right)\right\} \leq x \leq \min\left\{1, \left(\sqrt{\rho}+\frac{1}{\sqrt{\rho}}\right)^2\left(1-\frac{1}{\rho}\right)\right\}$:

$p_k = x$,

$p_{k+1} = \left(1+\frac{1}{\rho}\right)\left(1-\frac{1}{\nu}\right) - \frac{1}{1+\rho}x$,

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

In particular, the condition on x is:

1) $0 \leq x \leq \left(\sqrt{\rho}+\frac{1}{\sqrt{\rho}}\right)^2\left(1-\frac{1}{\nu}\right)$ if $1 \leq \nu \leq \frac{(1+\rho)^2}{1+\rho+\rho^2}$.
2) $0 \leq x \leq 1$ if $\frac{(1+\rho)^2}{1+\rho+\rho^2} \leq \nu \leq 1+\rho$.
3) $\left(\sqrt{\rho}+\frac{1}{\sqrt{\rho}}\right)^2\left(\frac{1}{1+\rho}-\frac{1}{\nu}\right) \leq x \leq 1$ if $1+\rho \leq \nu \leq \min\left\{\frac{1}{1-\rho}, (1+\rho)^2\right\}$.

The proof is given in the thesis [10], Theorem 5.3.

Observation 4.4: Theorem 4.3 shows that for any $k \geq 1$ and $j \geq 2$, $p_{k+j} = 1-\frac{1}{\nu}$. This can be explained as follows: A customer who observes $j+k-1$ customers, and overtakes only $k-1$ of them will be overtaken till a new arrival chooses to overtake $k-1$ customers. The time until this happens has a geometric distribution, with probability $1-p_{j+k}$ for success, and probability p_{j+k} for failure. Hence, the expected number of customers who overtake k is $\frac{1}{1-p_{j+k}} - 1$. 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 these customers, plus j service times of customers that were before him, plus one service time of himself. Hence, when $j \to \infty$ $f_{j,k-1} = \frac{j+\left(\frac{1}{1-p_{j+k}}-1\right)+1}{\mu} = \frac{j+\frac{1}{1-p_{j+k}}}{\mu}$. Substituting $f_{j,k-1}$ from (8) gives when $j \to \infty$ $p_j = 1-\frac{C_w}{C_o\mu} = 1-\frac{1}{\nu}$ which is proved in Theorem 4.3 .

A (symmetric) equilibrium strategy is, by definition, a best response against itself. However, it need not be the unique best response. Specifically, let y be an equilibrium strategy. There may be a best response strategy $z \neq y$ such that z is strictly a better response against itself than y is. In this case, y is unstable in the sense that when starting with y, it may be that the players adopt the best response z, and then a new equilibrium, at z, will be reached. If no such z exists then y is said to be an *evolutionarily stable strategy* or ESS. Note that if y is an equilibrium strategy and it is the unique best response against itself, then it is necessarily ESS.

Observation 4.5: The equilibrium mixed strategy $\Sigma_{k,\mathbf{p}}$ ($p \neq 0,1$) is not ESS.

C. 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 $\mathbf{P} = \begin{pmatrix} p_{k-1}^{k-1} & 0 \\ p_{k-1}^{k-1} & p_k^k \\ p_k^{k-1} & p_k^k \\ p_{k+1}^{k-1} & p_{k+1}^k \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix}$ such that $p_i^{k-1}, p_j^k \in [0,1]$ for every $i = k-1, \ldots$ and $j = k, \ldots$. 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 4.6: : $\Sigma_{k,\mathbf{P}}$ defines a unique equilibrium where customers are indifferent between overtaking k, $k-1$ or $k-2$ customers iff $1 \leq \frac{1}{1-\rho}$,

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

$p_{k+1}^k = \left(1+\frac{1}{\rho}\right)\left(1-\frac{1}{\nu}\right) - \frac{1}{1+\rho}x$,

$p_{k+1}^{k-1} = \frac{(1+\rho)^2\left(1-\frac{1}{\nu}\right)-\rho[\rho x+(1+\rho)y]}{\rho(1+\rho)(1+\rho+\rho^2)} - \frac{1+\rho}{1+\rho+\rho^2}z$,

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

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

$0 \leq x \leq 1,\ 0 \leq y \leq 1,\ 0 \leq z \leq 1,\ 0 \leq p_{k+1}^k \leq 1,\ 0 \leq p_{k+1}^{k-1} \leq 1$,

$x+z \leq 1$,

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

The proof is given in the thesis [10], Theorem 5.6.

We note that the equilibrium conditions are not an empty range.

D. Overtaking a single customer

We consider pure and mixed threshold strategies with at most one overtaken customer, and give necessary and sufficient conditions for these strategies to define an equilibrium.

Under the pure threshold strategy σ_n, a new customer overtakes one customer if there are n or more customers in the system, and does not overtake any customer otherwise.

Theorem 4.7: σ_n defines an equilibrium iff $\frac{1}{1-\rho} \leq \nu \leq \frac{1+\rho}{1-\rho}$.

The proof is given in the thesis [10], Theorem 7.1.

The mixed threshold strategy $\sigma_{n,p}$ where $0 < p < 1$, is defined as follows: a new customer overtakes one customer if there are at least $n+1$ customers in the system, does not overtake any customer if there are at most $n-1$ customers in the system, and if there are exactly n customers in the system he overtakes one customer with probability p, and does not overtake any customer otherwise.

Theorem 4.8: $\sigma_{n,p}$ defines an equilibrium iff $\frac{1}{1-\rho} \leq \nu \leq \frac{1+\rho}{1-\rho}$ and $p = \frac{1+\rho}{\rho(1-\rho)}\left(1-\rho-\frac{1}{\nu}\right)$.

The proof is given in the thesis [10], Theorem 7.2.

V. PROFIT MAXIMIZATION

In this section we compare two models of profit maximization using a single price. 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.

A. Maximum profit in the current priority (CP) model

The first model is the model of Section II. 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 a fixed cost per each overtaken customer. 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 have already shown that there may be numerous equilibria in this models. For example, always overtaking k customers, i.e., Σ_k, $k = 0, 1, 2 \ldots$, are equilibrium strategies. In particular strategy Σ_∞, in which an arriving customer overtakes all customers who are currently in the system, 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.

The profit from a customer is the cost which this customer pays. In the **CP** model it is C_o per each overtaken customer. The server's expected profit per customer under the Σ_∞ strategy is $C_o L$, where $L = \frac{\rho}{1-\rho}$ is the expected number of customers in the system. By Theorem 4.1, the maximum price for overtaking any customer such that Σ_∞ still defines an equilibrium is $\frac{C_w}{\mu-\lambda}$. This gives the following theorem:

Theorem 5.1: The expected maximum profit in the **CP** model among all Σ_k, $k = 0, 1, 2, \ldots$, strategies is

$$\Pi^{\mathrm{CP}} = \frac{\lambda}{(\mu-\lambda)^2} C_w. \tag{3}$$

It is received from Σ_∞ with $C_o = \frac{C_w}{\mu-\lambda}$ (or, $\nu = \frac{1}{1-\rho}$).
The proof is given in the thesis [10], Theorem 6.1.

We assume that the server can choose the equilibrium which maximizing its expected profit. This assumption is standard, see for example [26] p. 867 and p. 910. Hence, among Σ_k, $k = 0, 1, 2, \ldots$, it will choose Σ_∞ strategy in the **CP** model.

Remark 5.2: It may come as a surprise that it is possible that pure strategies, not one of the Σ_k type, give higher profit than Π^{CP} for some parameter values. We found pure strategies from Table I and parameters ν, ρ that satisfy $C_o L > \Pi^{\mathrm{CP}} = \frac{\lambda}{(\mu-\lambda)^2} C_w$, or $\nu > \frac{1}{1-\rho}$. For example, strategy (0,0,3,3,3,1,0,0,…) with $\rho = 2$ and $\nu = 4.77$ satisfies the condition.

B. Maximum profit in the absolute priority (AP) model with threshold n=0

The second model is that of [1], [16] In this model two FCFS queues are formed in front of a single server, one for priority customers and the other for ordinary customers. For a given threshold value $n \geq 0$, an arriving customer buys priority iff the number of customers in the ordinary queue is at least n. This is an absolute priority discipline, and therefore we 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, and 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 in [16]:

Theorem 5.3: The integer threshold strategy n, $n \geq 1$, specifies an equilibrium iff $\theta + \frac{C_w}{\mu} - \frac{C_w}{\mu-\lambda} \leq C_w W(n) \leq \theta + \frac{C_w}{\mu}$. The threshold $n = 0$ specifies an equilibrium iff $\theta + \frac{C_w}{\mu} \leq \frac{C_w}{\mu-\lambda}$.

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 $\Pi^{\mathrm{AP}}(n)$ 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. Suppose that all customers use the pure threshold strategy $n = 0$, i.e., the strategy is always buying priority. From Theorem 5.3, $\theta_{max} = \frac{\lambda}{\mu(\mu-\lambda)} C_w$, so that $\Pi^{\mathrm{AP}}(0) = \theta_{max}$.

Since $\frac{\lambda}{\mu(\mu-\lambda)} C_w < \frac{\lambda}{(\mu-\lambda)^2} C_w$, it follows that, $\Pi^{\mathrm{AP}}(0) < \Pi^{\mathrm{CP}}$, 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 $n = 0$.

C. Maximum profit in the absolute priority (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. $P_n = P(L \geq n) = (\frac{\lambda}{\mu})^n$. We assume that all customers use the pure threshold strategy $n \geq 1$. From Theorem 5.3, 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 iff the number of customers in the queue is at least n, $\Pi^{\mathrm{AP}}(n) = \theta_{max} P_n = \theta_{max} \left(\frac{\lambda}{\mu} \right)^n$, or equivalently

$$\Pi^{\mathrm{AP}}(n) = C_w \left[W(n) + \frac{\lambda}{\mu(\mu-\lambda)} \right] \left(\frac{\lambda}{\mu} \right)^n. \tag{4}$$

Lemma 5.4:
$W(1) = \frac{1}{\mu-\lambda}$ and $\Pi^{\mathrm{AP}}(1) = C_w \frac{\lambda+\mu}{\mu(\mu-\lambda)} \frac{\lambda}{\mu}$;
$W(2) = \frac{2\mu+\lambda}{\mu^2-\lambda^2}$ and $\Pi^{\mathrm{AP}}(2) = C_w \frac{2\mu^2+2\lambda\mu+\lambda^2}{\mu(\mu^2-\lambda^2)} (\frac{\lambda}{\mu})^2$;
$W(3) = \frac{3\mu^3+7\lambda\mu^2+4\lambda^2\mu+\lambda^3}{(\lambda+\mu)^2(\mu^2-\lambda^2)}$ and $\Pi^{\mathrm{AP}}(3) = C_w \frac{3\mu^4+8\lambda\mu^3+7\lambda^2\mu^2+4\lambda^3\mu+\lambda^4}{\mu(\lambda+\mu)^2(\mu^2-\lambda^2)} (\frac{\lambda}{\mu})^3$.
In these cases $\Pi^{\mathrm{AP}}(n) < \Pi^{\mathrm{CP}}$.

The proof is given in the thesis [10], Lemma 6.3.

ρ	$\Pi^{\mathrm{AP}}(4)$	$\Pi^{\mathrm{AP}}(7)$	$\Pi^{\mathrm{AP}}(10)$	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 II
SERVER'S EXPECTED PROFIT PER CUSTOMER IN **CP** AND **AP** MODELS, $C_w = 1$

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

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

In addition, we see in Figure 5, 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.

VI. CONCLUDING REMARKS

In this paper, we formulated a novel mechanism for allocating priorities in a queue. As with other mechanisms that are now widely acceptable in theme parks, communication systems and other complex queueing systems, this new regime may seem odd at first. We prefer to express in this paper some of its advantages, for example the fact that it uses a single price and increases the system's profits, together with interesting theoretical results associated with it, which are quite different from regular priority regimes.

A natural variation of our model allows customers to overtake not just at their arrival time but also later. Such opportunities make the game more complicated and different solution concepts may be required.

Another natural continuation of our research includes overtaking when balking is possible and when customers are heterogeneous. Such models also involve interesting questions regarding social optimization that were not raised in our model.

Fairness among customers is a fundamental issue for queueing systems. 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 [5] propose a fairness measure enabling to quantitatively measure and compare the level of fairness associated with various queueing systems. Other approaches are described in the survey [35]. 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.

REFERENCES

[1] Adiri, I. and U.Yechiali (1974), "Optimal priority purchasing and pricing decisions in nonmonpoly and monopoly queues," *Operations Research* **22**, 1051-1066.

[2] Afèche, P. (2010), "Incentive-compatible revenue management in queueing systems: capacity and optimal strategic delay."

[3] Afèche, Philipp and Haim Mendelson (2004), "Pricing and priority auctions in queueing systems with a generalized delay cost structure," *Management Science* **50** 896-882.

[4] Alperstein H. (1988), "Optimal Pricing Policy for the Service Facility Offering a Set of Priority Prices," *Management Science*, **34** 666-671.

[5] Avi-Itzhak, B. and H. Levy (2004), "On measuring fairness in queues," *Advances in Applied Probability* **36**, 919-936.

[6] Cachon, G. P. and P. Feldman (2011), "Pricing services subject to congestion: charge per-use fee or sell subscription?," *Manufacturing & Service Operations Management* **13** 244-260.

[7] Cachon, G.P. and F. Zhang (2006), "Procuring fast delivery: Sole sourcing with information asymmetry," *Management Science* **52**, 881-896.

[8] Conway, R.W., W.L. Maxwell, and L. W. Miller, *Theory of scheduling*, Addison-Wesley Pub. Co., Reading Mass, 1967.

[9] Dolan, R. J. (1978), "Incentive mechanisms for priority queuing problems," *The Bell Journal of Economics* **9**, 421-436.

[10] Erlichman, Jenny, (2009), "Equilibrium solutions in the observable M=M=1 queue with overtaking," *M.Sc. thesis, Department of Statistics and Operations Research, Tel Aviv University*, www.math.tau.ac.il/∼hassin/jenny_thesis.pdf.

[11] Glazer, A. and R. Hassin (1986), "Stable priority purchasing in queues," *Operations Research Letters* **4**, 285-288.

[12] Gilland, W.G. and D.P. Warsing (2009), "The impact of revenue-maximizing priority pricing on customer delay costs," *Decision Sciences* **40**, 89-120.

[13] Guo, Pengfei, Robin Lindsey, and Zhe George Zhang, "Pricing and capacity decisions for a regulated service provider in a two-tier service system," 2012.

[14] Hassin, R. (1985), "On the optimality of first-come last-served queues," *Econometrica* **53**, 201-202.

[15] Hassin, R. (1995), "Decentralized regulation of a queue," *Management Science* **41**, 163-173.

[16] Hassin, R. and M. Haviv (1997), "Equilibrium threshold strategies: the case of queues with priorities," *Operations Research* **45**, 966-973.

[17] Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer.

[18] Hassin, R. and M. Haviv (2006) "Who should be given priority in a queue?" *Operations Research Letters* **34**, 191-198.

[19] Hall, Joseph M., Praveen K. Kopalle, and David F. Pyke, "Static and dynamic pricing of excess capacity in a make-to-order environment," *Production and Operations Management* **18**(4) (2009) 411-425.

[20] Hassin, R., J. Puerto, and F. R. Fernandez (2009), "The use of relative priorities in optimizing the performance of a queueing system," *European J. of Operations Research* **193**, 476-483.

[21] Hayel, Y. and B. Tuffin (2005), "Pricing for heterogeneous services at a discriminatory processor sharing queue," *NETWORKING 2005, Lecture Notes in Computer Science* **3462**, 816-827,

[22] Kleinrock, L., *Queueing Systems* (Vol. 1: Theory, Vol.2: Computer Applications), Wiley, 1975 and 1976.

[23] Kleinrock, L. (1967), "Optimal Bribing for Queuing Position," *Operations Research* **15**, 304-318.

[24] Larson, R. C. (1987), "Perspectives on queues: social justice and the psychology of queueing," *Operations Research* **35**, 895-905.

[25] Lui, F.T (1985), "An equilibrium queuing model of bribery," *J. of Political Economy* **93**, 760-781.

[26] Mas-Collel, A., M.D. Whinston, and J.R. Green (1995), *Microeconomic Theory*, Oxford University Press.

[27] Mendelson, H. and S. Whang (1990), "Optimal Incentive-compatible priority pricing for the M/M/1 queue," *Management Science* **38**, 870-883.

[28] Myrdal, G. (1968), *Asian Drama: An Inquiry into the Poverty of Nations*, Pantheon, New York.

[29] Naor, P., "The regulation of queue size by levying tolls," *Econometrica* **37** (1969) 15-24.

[30] Özekici S. ,J. Li, and F. S. Chou (1994), "Queues with impolite customers," *Queueing Systems* **15** 261-277.

[31] Rosenblum, D.M. (1992), "Allocation of waiting time by trading in position on a G/M/s queue," *Operations Research* **40**, 338-342.

[32] Shenker, S. J. (1995), "Making greed work in networks: a game-theoretic analysis of switch service disciplines," *IEEE/ACM Transactions on Networking* **3**, 819-831.

[33] Sun, W. P. Guo, N. Tian, and S. Li (2009), "Relative policies for minimizing the cost of queueing systems with service discrimination," *Applied Mathematical Modelling* **33** 4241-4258.

[34] Tong, C. and S. Rajagopalan (2012), "Pricing and operational performance in discretionary services,".

[35] Wierman, A. (2011), "Fairness and scheduling in single server queues," *Surveys in Operations Research and Management Science* **16** 39-48.

[36] Whitt, W. (1983) "The amount of overtaking in a network of queues," *Networks* **14**, 411-426.

PROOFS

Proof of Theorem 4.1: 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 (1) there is no reason to overtake more than k customers.
  If he overtakes k customers, his expected cost is $C_w\frac{j+1-k}{\mu} + kC_o$. 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 $C_w\frac{j+1-i}{\mu-\lambda} + iC_o$. The strategy defines an equilibrium iff overtaking k customers is a best response of a new customer. Hence, $C_w\frac{j+1-k}{\mu} + kC_o \leq C_w\frac{j+1-i}{\mu-\lambda} + iC_o$, or $\frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda} + \frac{\lambda(j+1-k)}{\mu(\mu-\lambda)(k-i)}$ for $i = 0, 1, \ldots, k-1$ and $j = k, k+1, \ldots$. The minimum of $\left\{\frac{1}{\mu-\lambda} + \frac{\lambda(j+1-k)}{\mu(\mu-\lambda)(k-i)}\right\}$ over $i = 0, 1, \ldots, k-1$ and $j = k, k+1, \ldots$ 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 $j = 1, 2, \ldots, k-1$ customers, and chooses to overtake all of them. His expected cost is $C_w\frac{1}{\mu-\lambda} + jC_o$. Otherwise, if he chooses to overtake i customers, $i = 0, 1, \ldots, j-1$, 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 $j = 1, 2, \ldots, k-1$ and $i = 0, 1, \ldots, j-1$, or $\frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda}$.

■

Proof of Theorem 4.2: If a new customer observes $j \geq k$ customers, then according to Σ_k he overtakes k of them. Future customers do not overtake him, and therefore, his expected cost is $C_w\frac{j+1-k}{\mu} + kC_o$.

If a new customer observes $j < k$ customers, then according to Σ_k he chooses overtaking all of them. Future customers overtake him, and therefore, his expected cost is $C_w\frac{1}{\mu-\lambda} + jC_o$.

Hence, the expected cost of a new customer is

$$C_w\frac{j+1-k}{\mu} + kC_o, \quad j \geq k,$$

$$C_w\frac{1}{\mu-\lambda} + jC_o, \quad j < k.$$

The functions $C_w\frac{j+1-k}{\mu} + kC_o$ and $C_w\frac{1}{\mu-\lambda} + jC_o$ are both monotone increasing in j. If $C_w\frac{1}{\mu-\lambda} + j'C_o \leq C_w\frac{j+1-k}{\mu} + kC_o$ when $j' = k-1$ and $j = k$, or equivalently $\frac{\lambda}{\mu(\mu-\lambda)} \leq \frac{C_o}{C_w}$ then the expected cost of a new customer as a function of the number of customers in the system is a monotone and increasing. Otherwise, this function is built from two monotone increasing functions with a break-point at k.

Since both functions are monotone increasing then k is a global minimum if a new customer prefers to observe k rather than an empty queue, i.e., $C_w\frac{1}{\mu} + kC_o < C_w\frac{1}{\mu-\lambda}$, or equivalently, $\frac{C_o}{C_w} < \frac{\lambda}{k\mu(\mu-\lambda)}$. If a new customer prefers to observe an empty queue rather than k then 0 is a global minimum and then $C_w\frac{1}{\mu-\lambda} < C_w\frac{1}{\mu} + kC_o$, or equivalently, $\frac{\lambda}{k\mu(\mu-\lambda)} < \frac{C_o}{C_w}$.

A new customer prefers to observe a longer queue, i.e., prefers to observe $j \geq k$ customers rather than $j' < k$, if $C_w\frac{j+1-k}{\mu} + kC_o < C_w\frac{1}{\mu-\lambda} + j'C_o$, or equivalently if $\frac{C_o}{C_w} < \frac{\lambda-(\mu-\lambda)(j-k)}{\mu(\mu-\lambda)(k-j')}$. ■

Proof of Theorem 4.3: 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 $\Sigma_{k,\mathbf{p}}$. In addition let $f_{-1,j} = 0$.

- Consider a customer in state (i, j) where $j \leq k-2$:

$$f_{i,j} = \frac{i+1}{\mu-\lambda}, \quad \forall j \leq k-2. \tag{5}$$

- Consider a customer in state $(j-1, k-1)$, $j \geq 1$: $f_{j-1,k-1} = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}f_{j-2,k-1} + \frac{\lambda p_{j+k-1}}{\lambda+\mu}f_{j,k-1} + \frac{\lambda(1-p_{j+k-1})}{\lambda+\mu}\frac{j}{\mu}$, $j \geq 1$, or equivalently,

$$f_{j,k-1} = \frac{1}{\lambda p_{j+k-1}}[(\lambda+\mu)f_{j-1,k-1} - \mu f_{j-2,k-1} - 1 - (1-p_{j+k-1})\frac{\lambda j}{\mu}], \quad j \geq 1. \tag{6}$$

  In particular,

$$f_{1,k-1} = \frac{1}{\lambda p_k}\left[(\mu+\lambda)f_{0,k-1} - 1 - (1-p_k)\frac{\lambda}{\mu}\right]. \tag{7}$$

- Consider now a 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

$$f_{j,k-1} = \frac{C_o}{C_w} + \frac{j}{\mu}, \quad \forall j \geq 1. \tag{8}$$

Substituting $f_{1,k-1}$ from (8) in (7) gives

$$p_k = \frac{1}{\lambda}\frac{C_w}{C_o}\left[(\lambda+\mu)f_{0,k-1} - 1 - \frac{\lambda}{\mu}\right]. \tag{9}$$

Substituting $f_{1,k-1}$ from (8) in (6) for $j=2$ gives

$$p_{k+1} = 1 + \frac{\mu}{\lambda} - \frac{1}{\lambda}\frac{C_w}{C_o}\left[\mu f_{0,k-1} + \frac{\lambda}{\mu}\right]. \tag{10}$$

For $j \geq 2$, substituting $f_{j,k-1}$, $f_{j-1,k-1}$, and $f_{j-2,k-1}$ from (8) in (6) gives,

$$p_{k+j} = 1 - \frac{C_w}{C_o\mu}, \quad \forall j \geq 2. \tag{11}$$

We denote p_k by x. Substituting p_k from (9), and p_{k+1} from (10), in $p_k + \frac{\lambda+\mu}{\mu}p_{k+1}$ gives $p_{k+1} = \frac{\lambda+\mu}{\lambda}\left[1 - \frac{C_w}{C_o\mu}\right] - \frac{\mu}{\lambda+\mu}x$. Since $0 \leq p_{k+1} \leq 1$, we get that $\frac{(\lambda+\mu)^2}{\mu\lambda}\left[\frac{\mu}{\lambda+\mu} - \frac{C_w}{C_o\mu}\right] \leq x \leq \frac{(\lambda+\mu)^2}{\mu\lambda}\left[1 - \frac{C_w}{C_o\mu}\right]$. $x = p_k$, hence we must get that $0 \leq x \leq 1$. Therefore,
$\max\left\{0, \frac{(\lambda+\mu)^2}{\mu\lambda}\left[\frac{\mu}{\lambda+\mu} - \frac{C_w}{C_o\mu}\right]\right\} \leq x \leq$
$\min\left\{1, \frac{(\lambda+\mu)^2}{\mu\lambda}\left[1 - \frac{C_w}{C_o\mu}\right]\right\}$.
We consider these cases:

1) $\frac{1}{\mu} \leq \frac{C_o}{C_w} \leq \frac{(\lambda+\mu)^2}{\mu(\mu^2+\mu\lambda+\lambda^2)}$. In this case $0 \leq x \leq \frac{(\lambda+\mu)^2}{\mu\lambda}\left[1 - \frac{C_w}{C_o\mu}\right]$.
2) $\frac{(\lambda+\mu)^2}{\mu(\mu^2+\mu\lambda+\lambda^2)} \leq \frac{C_o}{C_w} \leq \frac{\lambda+\mu}{\mu^2}$. In this case $0 \leq x \leq 1$.
3) $\frac{\lambda+\mu}{\mu^2} \leq \frac{C_o}{C_w} \leq \min\left\{\frac{1}{\mu-\lambda}, \frac{(\lambda+\mu)^2}{\mu^3}\right\}$. In this case $\frac{(\lambda+\mu)^2}{\mu\lambda}\left[\frac{\mu}{\lambda+\mu} - \frac{C_w}{C_o\mu}\right] \leq x \leq 1$.

We now analyze the other equilibrium conditions and show that they are satisfied iff $\frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda}$.

- The best response of a new customer who observes $j \leq k-2$ customers is overtaking all of them. Hence, $C_w f_{0,j} + C_o j \leq C_w f_{j-l,l} + C_o l$, $l = 0, 1, 2, \ldots, j-1$. Substituting $f_{0,j}$ and $f_{j-l,l}$ from (5), this gives

$$\frac{C_0}{C_w} \leq \frac{1}{\mu-\lambda}. \tag{12}$$

- Consider the best response for a new customer who observes $k-1$ customers. If he overtakes all of them, his expected cost is $C_w f_{0,k-1} + C_o(k-1)$. Otherwise, if he overtakes only $l \leq k-2$ customers, his expected waiting time is $f_{j-l,l}$, substituting $f_{j-l,l}$ from (5) we get that his expected cost is $C_w\frac{k-l}{\mu-\lambda} + C_o l$. In equilibrium the best response is overtaking all customers, hence $C_w f_{0,k-1} + C_o(k-1) \leq C_w\frac{k-l}{\mu-\lambda} + C_o l$, or $\frac{C_0}{C_w} \leq \frac{k-l}{(\mu-\lambda)(k-1-l)} - \frac{f_{0,k-1}}{k-1-l}$. $f_{0,k-1}$ is bounded from above by the expected length of a busy period, because it is the maximum time till a customer in service leaves the system, even if all new arrival customers overtake him. Therefore, $\frac{k-l}{(\mu-\lambda)(k-1-l)} - \frac{f_{0,k-1}}{k-1-l} \geq \frac{k-l}{(\mu-\lambda)(k-1-l)} - \frac{1}{\mu-\lambda}\frac{1}{k-1-l} = \frac{1}{\mu-\lambda}$, and we get (12).
- The last case that we should check is if a new customer observes $j \geq k$ customers and chooses overtaking only $m \leq k-2$ customers, then his expected cost is $C_w f_{j-m,m} + mC_o$. In equilibrium the best response is overtaking k customers and not less. Therefore $C_w f_{j-m,m} + mC_o \geq C_w\frac{j-k+1}{\mu} + kC_o$, or $f_{j-m,m} \geq \frac{C_o}{C_w}(k-m) + \frac{j-k+1}{\mu}$. Substituting $f_{j-m,m}$ from (5) we get, $\frac{j-m+1}{\mu-\lambda} \geq \frac{C_o}{C_w}(k-m) + \frac{j-k+1}{\mu}$, or $\frac{C_o}{C_w} \leq \frac{1}{\mu-\lambda} + \frac{\lambda(j-k+1)}{\mu(\mu-\lambda)}\frac{1}{k-m}$, and $\frac{1}{\mu-\lambda} + \frac{\lambda(j-k+1)}{\mu(\mu-\lambda)}\frac{1}{k-m} > \frac{1}{\mu-\lambda}$. Therefore we get the condition (12).

■

Proof of Theorem 4.7: When all others apply the pure threshold strategy σ_n, a new customer's best response overtakes at most one customer, since by overtaking one customer the new customer guarantees his place in the queue and by the assumption (1) there is no benefit in overtaking more than one. In addition, if a customer observes $n-j$ customers, $j = 2, 3, \ldots, n$, then not overtaking any customer is the best response since the customer will never be overtaken, and by the same assumption, in this case there is no benefit in overtaking.

Suppose that a new customer observes $n-1$ customers. In equilibrium the best response of a new customer is not overtaking. Hence, $\frac{C_w}{\mu}(\frac{1}{1-\rho} + n - 1) \leq C_w\frac{n-1}{\mu} + C_o$, 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, $j = 0, 1, 2, \ldots$, and doesn't overtake any customer. 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, his expected cost is $C_w\frac{n+j}{\mu} + C_o$. In equilibrium the best response 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 equilibrium.

Proof of Theorem 4.8: 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 $\sigma_{n,p}$. Then, $f_n(p) = \frac{1}{\mu+\lambda} + \frac{\mu}{\mu+\lambda}\frac{n-1}{\mu} + \frac{\lambda}{\mu+\lambda}\left[pf_{n+1}(p) + (1-p)\frac{n}{\mu}\right]$. Substituting, $f_{n+1}(p) = \frac{1}{\mu}\frac{1}{1-\rho} + f_n(p) = \frac{1}{\mu-\lambda} + f_n(p)$ gives

$$f_n(p) = \frac{1}{\mu-\lambda-\lambda p}\left[n + \frac{\lambda p}{\mu-\lambda} + (1-p)\frac{\lambda n}{\mu}\right]. \tag{13}$$

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 (1), there is no benefit in overtaking more than one.

- Suppose that a new customer observes n customers. In equilibrium he 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 (13) gives the price $p_e = \frac{(\mu+\lambda)(C_o(\mu-\lambda)-C_w)}{C_o\lambda(\mu-\lambda)}$.

- 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 $(\mu+\lambda)\left(C_o(\mu-\lambda) - C_w\right) < C_o\lambda(\mu-\lambda)$, 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: $C_w f_n(p) < C_w \frac{n-1}{\mu} + C_o$, or $\frac{C_w\lambda}{C_o(\mu-\lambda)} > 0$, and this is always true.

Proof of Lemma 5.4: A customer in the ordinary queue is in state (i,j) if there are exactly i customers in front of him in the ordinary queue (including the one in service), exactly j customers behind him in the ordinary queue, and no customers in the priority queue. We denote by $f_{i,j}(n)$ the expected (residual) waiting time of a customer in state (i,j) given that all future customers adopt the pure threshold strategy n. In addition let $f_{-1,j}(n) = 0$. Hence $W(n) = f_{n-1,0}(n)$.
We now express the equations for calculating $W(n)$.

- Suppose that the state is (i,j) such that $i+j < n-1$, $i = 1,\ldots,n-2$ and $j = 0,\ldots,n-2$. The expected time till the next arrival or service completion is $\frac{1}{\lambda+\mu}$. With probability $\frac{\mu}{\lambda+\mu}$ the service completion occurs before a new arrival, and then the customer's expected residual waiting time is $f_{i-1,j}(n)$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs and since $i+j < n-1$, i.e., there are less than n customers in the system, a new customer does not overtake any customer and the customer's expected residual waiting time is $f_{i,j+1}(n)$. Therefore, for i, j such that $i+j < n-1$, $i = 1,\ldots,n-2$ and $j = 0,\ldots,n-2$,

$$f_{i,j}(n) = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{i-1,j}(n) + \frac{\lambda}{\lambda+\mu} f_{i,j+1}(n). \tag{14}$$

- Suppose that the state is $(i, n-i-1)$, $i = 1,\ldots,n-1$, i.e., there are n customers in the system. Then, all future arrivals will overtake the customer, till the number of customers in the queue is reduced by one. This is a busy period. Hence, for $i = 0,\ldots,n-2$,

$$f_{i,n-i-1}(n) = \frac{1}{\mu-\lambda} + f_{i-1,n-i-1}(n). \tag{15}$$

  In particular,

$$W(n) = f_{n-1,0}(n) = \frac{1}{\mu-\lambda} + f_{n-2,0}(n). \tag{16}$$

- If the state is $(0,j)$, $j \in 0,1\ldots,n-2$, the expected time till the next arrival or service completion is $\frac{1}{\lambda+\mu}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs and since $j < n-1$, i.e., there are less than n customers in the system, a new arriving customer does not overtake any customer and the customer's expected residual waiting time is $f_{0,j+1}(n)$. Therefore,

$$f_{0,j}(n) = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} f_{0,j+1}(n), \quad j \in 0,1,\ldots,n-2. \tag{17}$$

Now we give the proof to the Lemma:
$W(1) = \frac{1}{\mu-\lambda}$ and $\Pi^{\mathrm{AP}}(1) = C_w \frac{\lambda+\mu}{\mu(\mu-\lambda)} \frac{\lambda}{\mu}$;
$W(2) = \frac{2\mu+\lambda}{\mu^2-\lambda^2}$ and $\Pi^{\mathrm{AP}}(2) = C_w \frac{2\mu^2+2\lambda\mu+\lambda^2}{\mu(\mu^2-\lambda^2)} (\frac{\lambda}{\mu})^2$;
$W(3) = \frac{3\mu^3+7\lambda\mu^2+4\lambda^2\mu+\lambda^3}{(\lambda+\mu)^2(\mu^2-\lambda^2)}$ and $\Pi^{\mathrm{AP}}(3) = C_w \frac{3\mu^4+8\lambda\mu^3+7\lambda^2\mu^2+4\lambda^3\mu+\lambda^4}{\mu(\lambda+\mu)^2(\mu^2-\lambda^2)} (\frac{\lambda}{\mu})^3$.
In these cases $\Pi^{\mathrm{AP}}(n) < \Pi^{\mathrm{CP}}$.

If $n = 1$, then all new arrivals buy priority and overtake the present ordinary customer. In this case when the ordinary customer's service ends the system becomes empty. Thus his waiting time amounts to a busy period. Therefore, $W(1) = \frac{1}{\mu-\lambda}$. Substituting in (4) gives $\Pi^{\mathrm{AP}}(1) = C_w \left[\frac{1}{\mu-\lambda} + \frac{\lambda}{\mu(\mu-\lambda)}\right] \frac{\lambda}{\mu}$, or equivalently,

$$\Pi^{\mathrm{AP}}(1) = C_w \frac{\lambda+\mu}{\mu(\mu-\lambda)} \frac{\lambda}{\mu}. \tag{18}$$

Comparing $\Pi^{\mathrm{AP}}(1)$ from (18) to Π^{CP} from (3) we get that $\Pi^{\mathrm{AP}}(1) < \Pi^{\mathrm{CP}}$.

Observation 6.1: $W(2) = f_{1,0}(2)$.
Now we compute $W(2)$.

- Suppose a customer is in state (0,0). The expected time till the next arrival or service completion occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs. Then, the new arrival observes one customer upon arrival, therefore he does not buy priority and does not overtake the present customer in the ordinary queue. Hence,

$$f_{0,0}(2) = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} f_{0,1}(2). \tag{19}$$

- Suppose a customer is in state (0,1). All future arrivals will observe two customers or more upon arrival, therefore, they will buy priority and overtake the present customers in the ordinary queue till the number of customers in the ordinary queue is reduced by one, and it is equal to a busy period. Hence,

$$f_{0,1}(2) = \frac{1}{\mu-\lambda}. \tag{20}$$

- Suppose a customer is in state (1,0). All future arrivals will observe two customers or more upon arrival, therefore, they will buy priority and will overtake the present customers in the ordinary queue till the number of customers in the ordinary queue is reduced by one, and it is equal to a busy period. Hence, $f_{1,0}(2) = \frac{1}{\mu-\lambda} + f_{0,0}(2)$. Substituting in (19) gives,

$$f_{0,0}(2) = \frac{1}{\lambda+\mu} + \frac{\lambda}{\mu^2-\lambda^2} = \frac{\mu}{\mu^2-\lambda^2}. \tag{21}$$

  Substituting (21) in (20) gives, $f_{1,0}(2) = \frac{\lambda+2\mu}{\mu^2-\lambda^2}$. Since $W(2) = f_{1,0}(2)$,

$$W(2) = \frac{\lambda+2\mu}{\mu^2-\lambda^2}. \tag{22}$$

Substituting (22) in (4) gives $\Pi^{\mathrm{AP}}(2) = C_w \left[\frac{\lambda+2\mu}{\mu^2-\lambda^2} + \frac{\lambda}{\mu(\mu-\lambda)}\right] \left(\frac{\lambda}{\mu}\right)^2$, or equivalently,

$$\Pi^{\mathrm{AP}}(2) = C_w \frac{2\mu^2 + 2\lambda\mu + \lambda^2}{\mu(\mu^2-\lambda^2)} \left(\frac{\lambda}{\mu}\right)^2. \quad (23)$$

Comparing $\Pi^{\mathrm{AP}}(2)$ from (23) to Π^{CP} from (3) we get that $\Pi^{\mathrm{AP}}(2) < \Pi^{\mathrm{CP}}$.

Observation 6.2: $W(3) = f_{2,0}(3)$.

Now we compute $W(3)$.

- Suppose a customer is in state (0,0). The expected time till the next arrival or service completion occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs. Then, the new arrival observes one customer upon arrival, therefore he does not buy priority and does not overtake the present customer in the ordinary queue. Hence,

$$f_{0,0}(3) = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} f_{0,1}(3). \quad (24)$$

- Suppose a customer is in state (0,1). The expected time till the next arrival or service completion occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs. Then, the new arrival observes two customer upon arrival, therefore he does not buy priority and does not overtake the present customer in the ordinary queue. Hence,

$$f_{0,1}(3) = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} f_{0,2}(3). \quad (25)$$

- Suppose a customer is in state (0,2). All future arrivals will observe three customers or more upon arrival, therefore, they will buy priority and will overtake the present customers in the ordinary queue till the number of customers in the ordinary queue is reduced by one, and it is equal to a busy period. Hence,

$$f_{0,2}(3) = \frac{1}{\mu-\lambda}. \quad (26)$$

- Suppose a customer is in state (1,0). The expected time till the next arrival or service completion occurs is $\frac{1}{\lambda+\mu}$. With probability $\frac{\lambda}{\lambda+\mu}$ a new customer arrives before a service completion occurs. Then, the new arrival observes two customer upon arrival, therefore he does not buy priority and does not overtake the present customer in the ordinary queue. With probability $\frac{\mu}{\lambda+\mu}$ a service completion occurs before an arrival of a new customer, then the customer's expected time is $f_{0,0}(3)$. Hence,

$$f_{1,0}(3) = \frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu} f_{0,0}(3) + \frac{\lambda}{\lambda+\mu} f_{1,1}(3). \quad (27)$$

- Suppose a customer is in state (1,1). All future arrivals will observe three customers or more upon arrival, therefore, they will buy priority and overtake the present customers in the ordinary queue till, the number of customers in the ordinary queue is reduced by one, and it is equal to a busy period. Hence,

$$f_{1,1}(3) = \frac{1}{\mu-\lambda} + f_{0,1}(3). \quad (28)$$

- Suppose a customer is in state (2,0). All future arrivals will observe three customers or more upon arrival, therefore, they will buy priority and will overtake the present customers in the ordinary queue till the number of customers in the ordinary queue is reduced by one, and it is equal to a busy period. Hence,

$$f_{2,0}(3) = \frac{1}{\mu-\lambda} + f_{1,0}(3). \quad (29)$$

Substituting (26) in (25), gives

$$f_{0,1}(3) = \frac{\mu}{\mu^2-\lambda^2}. \quad (30)$$

Substituting (30) in (28), gives

$$f_{1,1}(3) = \frac{2\mu+\lambda}{\mu^2-\lambda^2}. \quad (31)$$

Substituting (30) in (24), gives

$$f_{0,0}(3) = \frac{\mu^2-\lambda^2+\lambda\mu}{(\lambda+\mu)(\mu^2-\lambda^2)}. \quad (32)$$

Substituting (32) and (31) in (27) and the result in (29) gives, $f_{2,0}(3) = \frac{3\mu^3+7\lambda\mu^2+4\lambda^2\mu+\lambda^3}{(\lambda+\mu)^2(\mu^2-\lambda^2)}$. Since $W(3) = f_{2,0}$,

$$W(3) = \frac{3\mu^3+7\lambda\mu^2+4\lambda^2\mu+\lambda^3}{(\lambda+\mu)^2(\mu^2-\lambda^2)}. \quad (33)$$

Substituting (33) in (4) gives $\Pi^{\mathrm{AP}}(3) = C_w \left[\frac{3\mu^3+7\lambda\mu^2+4\lambda^2\mu+\lambda^3}{(\lambda+\mu)^2(\mu^2-\lambda^2)} + \frac{\lambda}{\mu(\mu-\lambda)}\right] \left(\frac{\lambda}{\mu}\right)^3$, or equivalently,

$$\Pi^{\mathrm{AP}}(3) = C_w \frac{3\mu^4+8\lambda\mu^3+7\lambda^2\mu^2+4\lambda^3\mu+\lambda^4}{\mu(\lambda+\mu)^2(\mu^2-\lambda^2)} \left(\frac{\lambda}{\mu}\right)^3. \quad (34)$$

Comparing $\Pi^{\mathrm{AP}}(3)$ from (34) to Π^{CP} from (3) we get that $\Pi^{\mathrm{AP}}(3) < \Pi^{\mathrm{CP}}$.

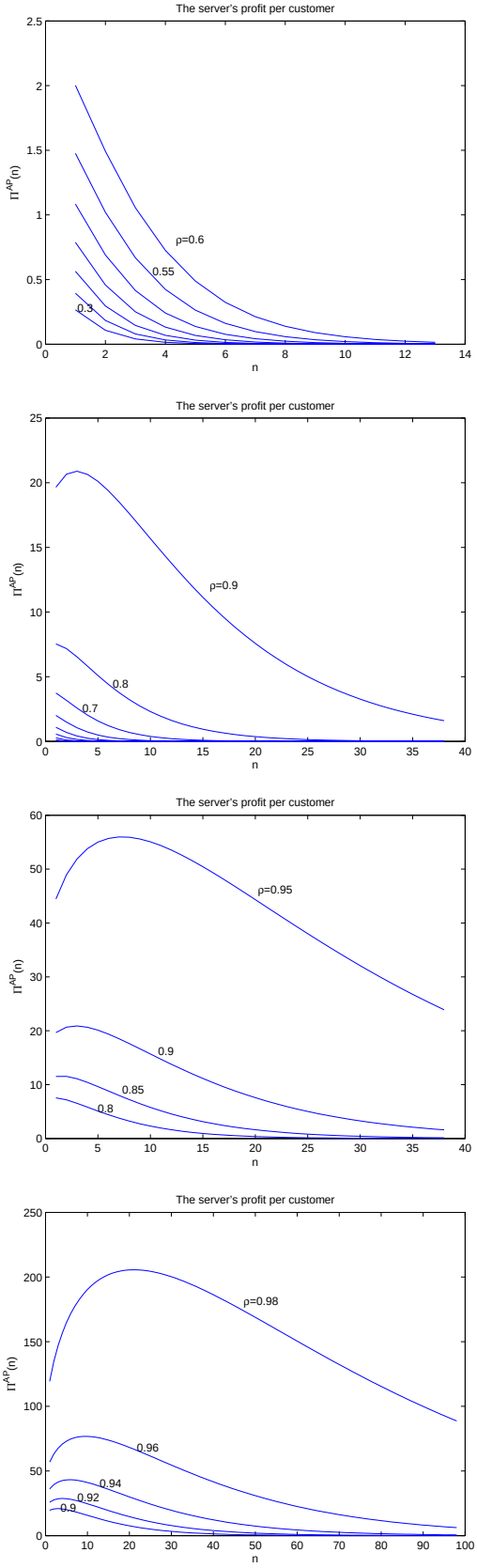

Fig. 4. The server's expected profit $\Pi^{\mathrm{AP}}(n)$ as a function of the threshold n and ρ, $C_w = 1$.

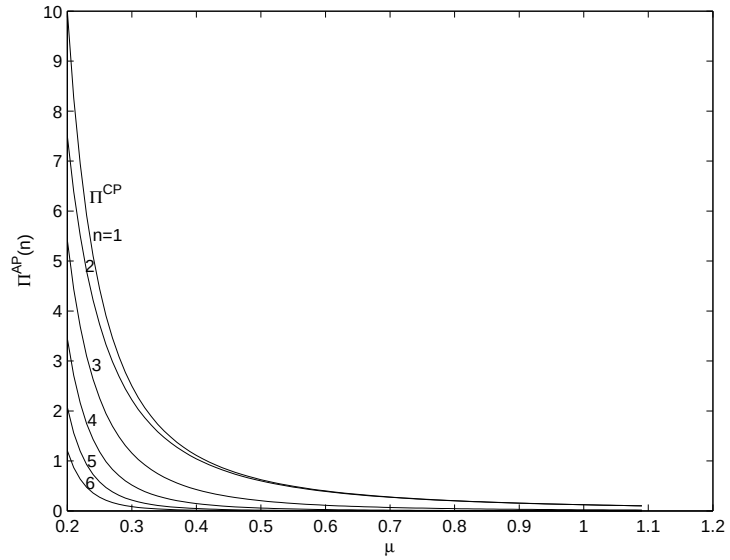

Fig. 5. The server's expected profit per customer in **AP** model as a function of μ, $C_w = 1$, $\lambda = 0.1$.