Customer Equilibrium in a Single-Server System with Virtual and System Queues

Roei Engel and Refael Hassin*
Department of Statistics and Operations Research
Tel Aviv University

June 25, 2017

Abstract

Consider a non-preemptive M/M/1 system with two first-come first-served queues, virtual (VQ) and system (SQ). An arriving customer who finds the server busy decides which queue to join. Customers in the SQ have non-preemptive priority over those in the VQ, but waiting in the SQ is more costly. We study two information models of the system. In the unobservable model customers are notified only whether the server is busy, and in the observable model they are also informed about the number of customers currently waiting in the SQ. We characterize the Nash equilibrium of joining strategies in the two models and demonstrate a surprising similarity of the solutions.

Keywords: Virtual queues, equilibrium behavior in a queueing system, observable queues

1 Introduction

Virtual queueing is an innovative method which is now commonly used to improve customer service satisfaction. Virtual queues, and their use in real life applications are extensively discussed in the literature. Examples include hospitality organizations, restaurants, amusement parks, airports, and call centers; see Armony and Maglaras [4, 5], Burgain, Feron, and Clarke [6], Camulli [7], Cope, Cope and Davis [10], Dickson, Ford, and Laval [11], Lovejoy, Aravkin, and Schneider-Mizell [25], de Lange, Samoilovich, and van der Rhee [24]. These systems offer an arriving customer the option of receiving a call when his time

*This research was supported by the Israel Science Foundation (grants No. 1015/11 and 355/15)

for service arrives, thereby reducing waiting costs by freeing him to perhaps perform useful activities while waiting.

Understanding the impact a virtual queue has on customer behavior and customer satisfaction is a central managerial issue for service systems of this type. However, modeling and analyzing such systems is not simple because it involves customers' strategic decisions. Moreover, a customer's decision and welfare strongly depend on the strategies adopted by the other customers of the system, and for this reason we look for *equilibrium behavior*.

Some virtual queueing systems maintain the first-come first-served order (calling back the customer when his turn arrives), whereas other virtual queueing systems give priority to customers who choose to wait in the system queue and call back virtual queue customers when the server become idle. Our model considers a system of the latter type.

We model a service system with two first-come first-served queues, a system queue (SQ) and a virtual queue (VQ) with a different waiting cost per unit of time. An arriving customer who finds the server idle enters service immediately, but if the server is busy the customer chooses which queue to join, based on the waiting costs and the available information. Our goal is to determine the (Nash) equilibrium of the system in two information models: an *unobservable model* where the customer is only informed whether the server is busy, and an *observable model* where customers are also informed on the length of the SQ and they follow a threshold strategy, joining the SQ if its length is below a critical value, and possibly randomizing at that value.

Let C_s and C_v denote the waiting cost per time unit in the SQ and VQ, respectively. The system is defined by two normalized parameters, the cost ratio $\varphi = C_v/C_s$, and the system utilization ρ. Our main results are the following:

- **The unobservable model.** We characterize the equilibrium solutions. Joining the SQ is a dominant strategy if $\varphi > 1 - \rho$, joining the VQ is dominant if $\varphi < 1 - \rho$, and any pure or mixed joining strategy is an equilibrium when $\varphi = 1 - \rho$.
- **The observable model.** We derive formulas for the stationary probabilities, the mean busy period in the SQ, the expected number of customers and expected waiting time in the VQ. We derive a linear-time algorithm for the truncated steady-state distribution and conduct a numerical investigation of the best response and equilibrium customers strategy. We conclude that, similar to the unobservable case, if φ is significantly greater than $1 - \rho$, joining the SQ is a dominant strategy; if it is significantly smaller, joining the VQ is dominant. In other cases, where φ and $1-\rho$ are approximately of the same size, we obtain the "follow-the-crowd" (FTC) behavior which is typical in priority systems, and leads to

multiple equilibria [17, 18].

The stationary distribution in the observable case with a pure threshold strategy and preemption is investigated by Haviv [19] §12.2.4 as a special case of a model analyzed by Kopzon, Nazarathy and Weiss [22]. In our generalized analysis of these results customers apply a mixed threshold strategy and we compute the resulting customer equilibrium.

Customer decision-making and Nash equilibrium in queues were initially defined and investigated by Naor [27], Littlechild [28] and Edelson and Hildebrand [13]. The literature on strategic behavior in queueing systems is surveyed in [18, 16].

Our paper is the first to consider customer strategic behavior and the resulting equilibrium in a virtual queueing system where waiting costs in the VQ are lower but the SQ obtains priority. We now describe the most relevant literature and emphasize where these papers differ from ours.

Guijarro, Pla, and Tuffin [14] (in their second model) investigate an unobservable multi-server system with an SQ and a VQ with different admission costs. The queues are managed by *competing servers* and each profits from its own queue. The authors investigate a two-stage sequential game where the servers choose admission prices and the customer chooses the queue to join. Hassin [15] and Altman, Jiménez, Núñez-Queija, and Yechiali [3] consider a system *with two servers* each with a different queue and *identical waiting costs*. An arriving customer can only see the length of one queue and decides which queue to join based on the conditional expected length of the other queue. Mandelbaum and Yechiali [26], investigate the optimal strategy of a *single arriving "smart" customer* in a single server system. The customer can join the queue, leave the system, or delay his decision and wait outside of the queue at a reduced cost, which can be viewed as joining a private virtual queue. Economou and Kanta [12] consider a single-server system with no waiting space (no SQ), where *customers who find the server busy automatically join the VQ*. After completing service the server seeks a customer in the VQ, with an exponentially distributed search time. If a new customer arrives during the search time the server interrupts the search and serves the new customer. The authors solve the social-optimization and profit-maximization problems of both the observable and the unobservable cases. This model with immediate search time can be considered as a special case of our model where only the VQ exists.

Some other papers analyze virtual queueing systems but customer behavior is not a result of strategic self-optimization and therefore *no equilibrium behavior is considered*. Aguir, Karaesmen, Akşin, and Chauvet [2] investigate a multi-server call center system with an SQ

and a VQ where customers are impatient and can balk or jockey from the SQ to the VQ. Chakravarthy, Krishnamoorthy and Joshua [8] model a multi-server system where new arrivals who find all servers busy join a VQ and retry after exponentially distributed time intervals. Moreover, upon service completion, with a given probability p the server serves a customer from the VQ if one exists. Our model, in contrast, assumes a single server, no retrials, $p = 1$, and customers choosing between the VQ and an SQ. Iravani and Balcioǵlu [21] consider a multi-server system with an SQ and a VQ where impatient customers choose a queue with an exogenous probability. Wüchner, Sztrik and de Meer [29] numerically analyze a system with an SQ and a VQ where customers are allowed to balk and move between the queues. Kostami, and Ward [23] model a single server with inline (system) and offline (virtual) queues where arriving customers choose which queue to join according to waiting time estimates by the server, and offline customers may leave the system without informing the server. Armony and Maglaras [4, 5] investigate two close models of a multi-server call center system with an SQ and a VQ, customers choose a queue to join or balk, and those joining the VQ are guaranteed an upper bound on their delay. There is no waiting-cost difference between the queues, and after each service the server decides whether to take a caller from the SQ or from the VQ.

Our system is priority-based, and customers in the SQ have priority over those who join the VQ. However, the main difference between our system and the common two-priority system is in our assumption that there is a different waiting cost for each queue and only the SQ is observable. The fundamental model involving customer decisions in queues with priorities is analyzed by Adiri and Yechiali [1] and by Hassin and Haviv [17]. Both queues are observable, and the waiting cost in both is identical but the admission price is different. Hassin and Haviv [18], §4.2 solve the unobservable case of that system.

This paper is structured as follows. Section 2 presents the basic model and assumptions. In section 3 we solve the unobservable model. Section 4 investigates the observable model when the customers use mixed strategies. Section 5 suggests directions for future work. An appendix with a table summarizing our notation and detailed derivations of the stationary probabilities in the observable case concludes the paper.

2 The system

We consider a non-preemptive single-server M/M/1 system with two queues, a System Queue (SQ) and a Virtual Queue (VQ) which is often called in the literature an orbit queue or a standby queue. An arriving

customer who finds the server idle enters service immediately. If the server is busy, the customer decides whether to join the VQ or the SQ (no balking allowed). Each queue has a different waiting cost per unit time, C_s for the SQ and C_v for the VQ ($C_v < C_s$). The discipline in each queue is FCFS, customers arrive according to a Poisson process at rate λ, and service times are exponentially distributed with rate μ. The system parameters λ, μ, C_s and C_v are known to the customers. (We will see that it is sufficient that they know the ratios λ/μ and C_v/C_s.) If the SQ is not empty, the first customer in it will be the next to be served when the current service terminates. Only if the SQ is empty the server calls the first customer from the VQ. We consider two information cases, an unobservable queue in which an arriving customer is only informed of the state of the server - busy or idle, and an observable queue where the arriving customer is also informed of the number of customers currently waiting in the SQ.

3 The unobservable model

In the unobservable model arriving customers are only informed if the server is busy or not. An arriving customer who finds the server busy will use, in general, a mixed strategy and join the SQ with some probability r_s. The arrival rates to the SQ and VQ when the server is busy are λr_s and $\lambda(1-r_s)$, respectively. We define the following normalized parameters:

$$\varphi = \frac{C_v}{C_s} \ , \ \rho = \frac{\lambda}{\mu} \ , \ \rho_s = \frac{\lambda r_s}{\mu} \ , \ \rho_v = \frac{\lambda(1-r_s)}{\mu} \ .$$

For stability we assume $\rho < 1$.

We denote the state of the system as (l_s, l_v) where l_s and l_v are the number of customers in SQ and the VQ, respectively. We represent the state when the server is idle with the special tag $0'$. The transition rate diagram of the system is shown in Figure 1.

Theorem 3.1.

$$r_s^e = \begin{cases} 1 & \varphi > 1-\rho \ , \\ \\ 0 & \varphi < 1-\rho \ . \end{cases}$$

If $\varphi = 1-\rho$, any strategy $0 \le r_s^e \le 1$ defines an equilibrium.

Proof. Let B denote the event that the server is busy. Let W_s and W_v denote the queueing time in the SQ and VQ, respectively. Given that the server is busy, the SQ is an M/M/1 system with arrival rate λr_s, and therefore the expected time in the system of a joining customer is

$$E(W_s|B) = \frac{1}{(1-\rho_s)\mu}. \tag{1}$$

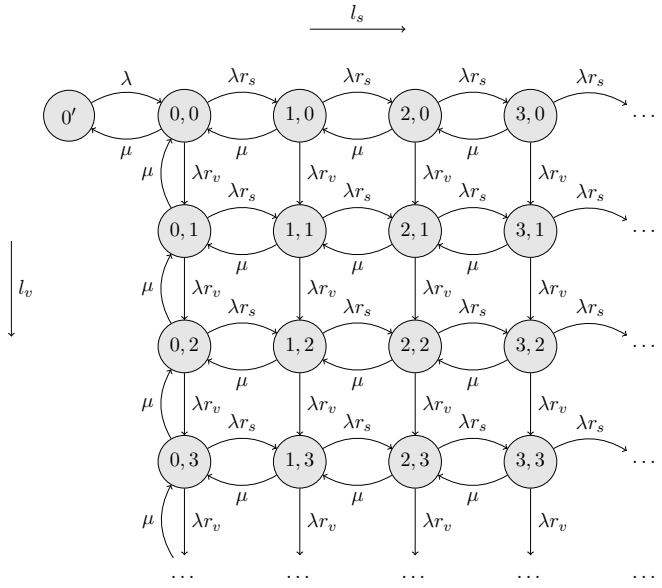

Figure 1: Transition rate diagram for the unobservable model

The total number, L_q, of customers in the two queues does not depend on the order in which the customers are served. Hence, as in the M/M/1 queue,

$$E(L_q|B) = \frac{\rho}{1-\rho},$$

and

$$E(L_v|B) = E(L_q|B) - E(L_s|B) = \frac{\rho_v}{(1-\rho)(1-\rho_s)} \ ,$$

where L_s and L_v denote the number of customers in the SQ and VQ, respectively. By Little's law

$$E(W_v|B) = \frac{1}{(1-\rho)(1-\rho_s)\mu} \ . \tag{2}$$

Define the cost-difference function

$$f(r_s) = E(W_v|B)C_v - E(W_s|B)C_s \ .$$

An arriving customer will be indifferent in his choice when $f(r_s) = 0$, will choose the SQ when $f(r_s) > 0$ and the VQ when $f(r_s) < 0$. Substituting (1) and (2) gives

$$\frac{f(r_s)}{C_s} = \frac{\varphi + \rho - 1}{(1-\rho)(1-\rho_s)\mu} \ ,$$

from which the claim follows. □

4 The observable model

In the observable case, an arriving customer is informed about the state of the server and the number of customers, l_s, in the SQ. We assume that customers follow a threshold strategy defined as follows (see [15]):

A threshold strategy $s(l_s)$ to join the SQ with threshold $T = n + r$ $(n \in \mathbb{N}, r \in [0,1))$ is defined by

$$s(l_s) = \begin{cases} 1 & l_s < n \\ r & l_s = n \\ 0 & l_s > n \ . \end{cases}$$

By following this strategy a customer always joins the SQ when l_s is at most $n-1$, joins the VQ if it is greater than n, and joins the SQ with probability r when $l_s = n$. When $r > 0$, the number of customers in the SQ is at most $n+1$. When $r = 0$, $s(l_s)$ is a pure strategy and a customer who observes a queue of length n joins the VQ.

4.1 Steady-state solution

The transition rate diagram for a threshold strategy $T = n+r$ is shown in Figure 2.

We denote the stationary distribution as P_{ji} where j corresponds to l_s and i corresponds to l_v. The probability of the special state where the server is idle is denoted $P_{0'}$.

It is straightforward to see that $P_{0'} = 1 - \rho$ and $P_{00} = (1-\rho)\rho$. The proofs of Propositions 4.1 and 4.2 can be found in the appendix:

Proposition 4.1.

$$P_{j0} = \frac{(1+\rho)\rho^j + (1+\rho-r)\left(\rho^j \sum_{k=1}^{n} \rho^k - \rho^n \sum_{k=1}^{j} \rho^k\right)}{(1+\rho)\sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k} P_{00} \ , \ j = 1, \ldots, n \ , \tag{3}$$

$$P_{n+1,0} = \frac{\rho^{n+1} r}{(1+\rho)\sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k} P_{00} \ . \tag{4}$$

We now present a recursive expression for the stationary probability of a generic state ji, P_{ji}.

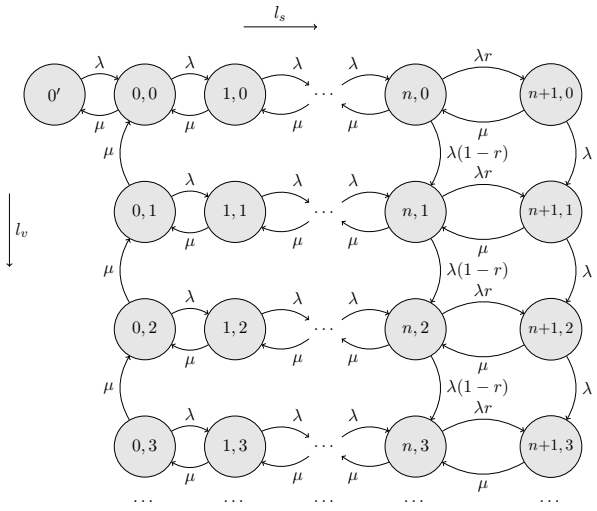

Figure 2: Transition rate diagram in the observable model when customers follow the mixed threshold strategy $T = n + r$.

Proposition 4.2.
For $i = 1, 2, \ldots$ *and* $0 \le r < 1$

$$P_{ni} = XZ^{i-1}P_{n+1,0} + YP_{n,i-1} + \frac{\rho r}{1+\rho}XY\sum_{k=0}^{i-2} Z^k P_{n,i-2-k} \ , \tag{5}$$

$$P_{n+1,i} = \frac{\rho r}{1+\rho}Y^i P_{n0} + ZP_{n+1,i-1} + \frac{\rho r}{1+\rho}XY\sum_{k=0}^{i-2} Y^k P_{n+1,i-2-k} \ , \tag{6}$$

$$P_{ji} = \sum_{k=0}^{j} \rho^k P_{0i} - (P_{n+1,i} + (1-r)P_{ni})\sum_{k=1}^{j} \rho^k \ , \ j = 1, 2, \ldots, n-1 \ , \tag{7}$$

$$P_{0i} = (1-r)\rho P_{n,i-1} + \rho P_{n+1,i-1} \ , \tag{8}$$

where

$$X = \frac{(1-\rho^{n+2})\rho}{(1+\rho)(1-\rho^{n+1}) - (1-\rho^n)\rho r} \ , \tag{9}$$

$$Y = \frac{(1-r)(1+\rho)(1-\rho^{n+1})\rho}{(1+\rho)(1-\rho^{n+1}) - (1-\rho^n)\rho r} \ , \tag{10}$$

$$Z = (1+Xr)\frac{\rho}{1+\rho} \ . \tag{11}$$

Remark 4.3. *For $i = 1$ the sums $\sum_{k=0}^{i-2} Y^k P_{n+1,i-2-k}$ and $\sum_{k=0}^{i-2} Z^k P_{n,i-2-k}$ are over empty sets, therefore we have*

$$P_{n1} = YP_{n0} + XP_{n+1,0} \ , \tag{12}$$

$$P_{n+1,1} = ZP_{n+1,0} + \frac{\rho r}{1+\rho} YP_{n0} \ . \tag{13}$$

Remark 4.4. *The stationary probabilities of row i can be computed in $O(n+i)$ time. By defining the functions*

$$F_i = \sum_{k=0}^{i-2} Z^k P_{n,i-2-k} \ , \ F'_i = \sum_{k=0}^{i-2} Y^k P_{n+1,i-2-k}$$

we have

$$F_i = P_{n,i-2} + ZF_{i-1} \ , \ F'_i = P_{n+1,i-1} + YF'_{i-1} \ ,$$

with this we express P_{ni} and $P_{n+1,i}$ as

$$P_{ni} = XZ^{i-1}P_{n+1,0} + YP_{n,i-1} + \frac{\rho r}{1+\rho} XYF_i \ , \tag{14}$$

$$P_{n+1,i} = \rho r Y^i P_{n0} + ZP_{n+1,i-1} + \frac{\rho r}{1+\rho} XYF'_i \ . \tag{15}$$

We first pre-calculate Y, Z, $\frac{\rho r}{1+\rho}XY$, $\frac{\rho r}{1+\rho}P_{n0}$ and $XP_{n+1,0}$. Then we calculate $P_{n,i}$ and $P_{n+1,i}$ and from these we calculate the probabilities P_{ji} $j = 0, \ldots, n$. The pre-calculation is done in $O(n)$ time by using Equations (25), (4), (9), (10) *and* (11). *We calculate $P_{n,i}$ and $P_{n+1,i}$ by using Equations* (15), (14) *and the pre-calculated values. By saving the variables F_{k-1}, F'_{k-1}, Z^{k-2} and Y^{k-1} when we calculate P_{nk} and $P_{n+1,k}$ $(0 < k < i)$ we can calculate $P_{n,k+1}$ and $P_{n+1,k+1}$ in $O(1)$ time and therefore find $P_{n,i}$ and $P_{n+1,i}$ in $O(i)$ time. We calculate the stationary probabilities of the row using Equation* (7). *By saving the value of the sums when calculating $P_{k-1,i}$ we can find $P_{k,i}$ in $O(1)$ time and calculate all the stationary probabilities in $O(n)$. Therefore the calculation of a row i takes $O(n+i)$ time.*

For the SQ when the server is busy, the probability that the queue length is l_s is

$$P(l_s) = \frac{\sum_{i=0}^{\infty} P_{l_s i}}{1 - P_{0'}} \ .$$

and the expected queueing time of a customer joining it, given l_s, is

$$E(W_s) = \frac{l_s + 1}{\mu} \ . \tag{16}$$

The expected length of the VQ when the server is busy and the length of the SQ is l_s can be calculated from propositions 4.1, 4.2 and $P(l_s)$

$$E(L_v|l_s) = \sum_{i=1}^{\infty} iP_{l_s i} \Big/ \sum_{i=0}^{\infty} P_{l_s i} \ . \tag{17}$$

We define a busy-period type variable $b(f)$, $f = 0, \ldots, n$ denoting the expected time it takes the SQ to decrease its length from $n+1-f$ to $n-f$. Similarly, $b(n+1)$ denotes the expected time it takes a busy server to be ready to serve a customer from the VQ given that $l_s = 0$. Then:

$$b(0) = \frac{1}{\mu} \ ,$$

$$b(1) = \frac{1}{\lambda+\mu} + \frac{\lambda r}{\lambda+\mu}(b(1)+b(0)) + \frac{\lambda(1-r)}{\lambda+\mu}b(1) \ ,$$

$$b(f) = \frac{1}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda}(b(f)+b(f-1)) \quad f = 2, 3, \ldots, n+1 \ .$$

These equations are explained as follows: When $f = 0$ new arrivals do not join the SQ and therefore $b(0) = \frac{1}{\mu}$. When $f > 0$, the expected time until the next event is $\lambda + \mu$. For $f = 1$, if the next event is an arrival then with probability r the arriving customer joins the SQ and needs to wait $b(0)$ and an additional $b(1)$, and with probability $1-r$ the arriving customer joins the VQ and we stay at the same state. When $f = 2, 3, \ldots$, if the next event is an arrival then we enter state $f-1$ and need expected time $b(f-1)$ to return to state f and then another $b(f)$ until the end of the busy period. If the next event is the end of service then the busy period terminates.

The solution to these equations is

$$b(1) = \frac{\rho r + 1}{\mu} \ ,$$

$$b(f) = \left(\sum_{k=0}^{f-1} \rho^k + \rho^f r \right) \frac{1}{\mu}, \quad f = 2, 3 \ldots, n+1 \ . \tag{18}$$

Finally, the expected queueing time of a customer who joins the VQ when the length of the SQ is l_s is

$$E[W|l_s] = \sum_{k=0}^{l_s} b(n+1-k) + E(L_v|l_s)b(n+1) \ , l_s \leq n+1 \ . \tag{19}$$

The first term is the expected time until the server is ready to serve the first customer from the VQ, consisting of l_s+1 consecutive busy periods with increasing values of f, i.e., for the first busy period $f = n+1-l_s$,

for the second $f = n+1-(l_s-1)$, and so on, concluding with $b(n+1)$ which is the time it takes until the first customer from the VQ is called for service given that the SQ is empty. The second term is the time it takes until the server is ready to serve the new VQ customer given that it just started serving a VQ customer.

Remark 4.5. *The case where the customers follow a pure strategy has a closed-form solution:*

$$P_{n0} = \frac{\rho^n}{\sum_{k=0}^{n} \rho^k} P_{00} \ , \ P_{01} = \frac{\rho^{n+1}}{\sum_{k=0}^{n} \rho^k} P_{00} \ ,$$

$$P_{j0} = \frac{\rho^j \sum_{k=0}^{n} \rho^k - \rho^n \sum_{k=1}^{j} \rho^k}{\sum_{k=0}^{n} \rho^k} P_{00} \ , \ j = 1, \ldots, n \ ,$$

$$P_{ji} = \frac{\rho^{n+i}}{\sum_{k=0}^{n} \rho^k} P_{00} \ , \quad \begin{array}{l} i = 1, 2, \ldots \\ j = 0, 1, \ldots n \ , \end{array}$$

$$E(L_v|l_s) = \frac{\rho^{n+1}}{\left(P_{l_s 0} \sum_{k=0}^{n} \rho^k + \rho^{n+2} \right) (1-\rho)^2} P_{00} \ ,$$

and

$$E[W|l_s] = \sum_{k=0}^{l_s} b(n-k) + E(L_v|l_s) b(n) \ , \ l_s = 0, \ldots, n \ .$$

4.2 Numerical investigation of the equilibrium strategies

A complete analysis of the equilibrium strategies is not possible and therefore we present in this section the results of a numerical study. We conclude from this study that the conditional expected waiting time in the VQ is approximately linear and provide an explanation for this phenomenon. This conclusion enables us to obtain important insights about the equilibrium behavior. Specifically we provide conditions for dominant strategies, and show that when these conditions are violated there are, in general, multiple equilibrium solutions.

We define the normalized expected waiting time in the VQ

$$\hat{E}[W|l_s] = \frac{E[W|l_s]}{\frac{1}{\mu}} = \sum_{k=0}^{l_s} b'(n+1-k) + E(L_v|l_s) b'(n+1) \ , l_s \leq n+1 \ . \tag{20}$$

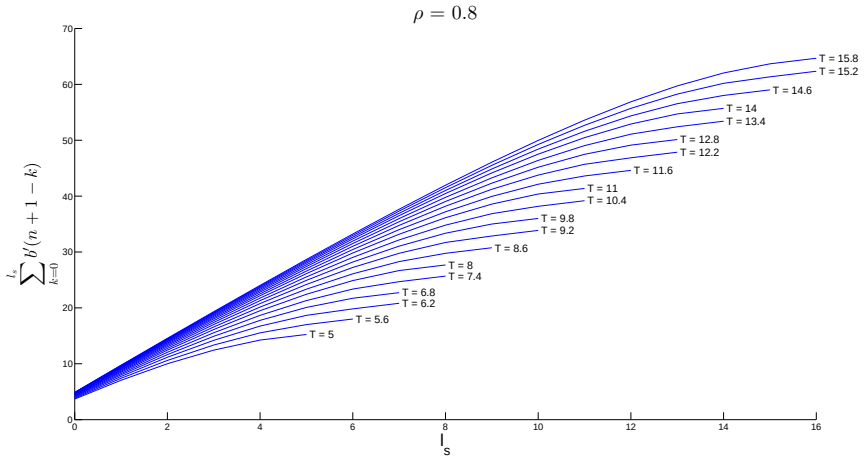

Figure 3: $\sum_{k=0}^{l_s} b'(n+1-k)$ when $\rho = 0.8$.

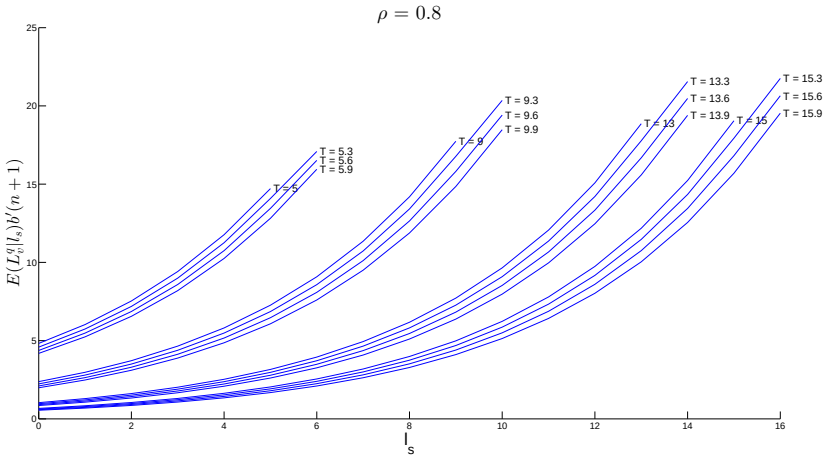

Figure 4: $E(L_v|l_s)b'(n+1)$ when $\rho = 0.8$.

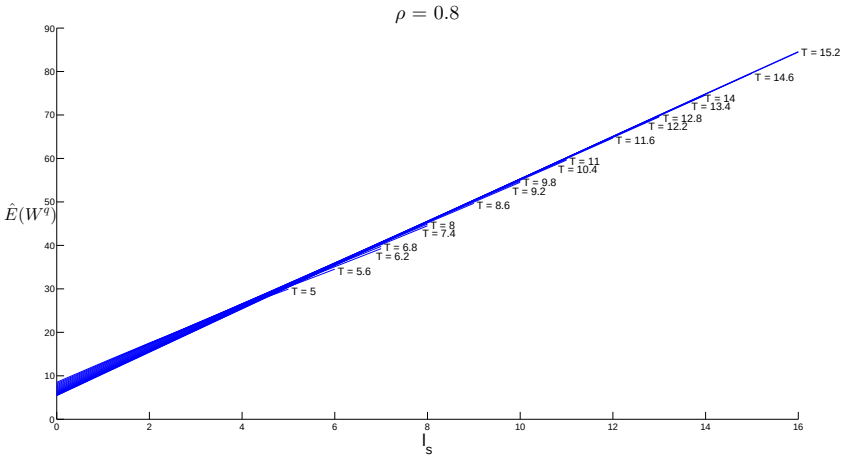

Figure 5: Normalized expected waiting time in the VQ ($\hat{E}(W|l_s)$) when $\rho = 0.8$.

Figure 3 shows the first term of Equation (20), $(\sum_{k=0}^{l_s} b'(n+1-k))$, which is the expected time it will takes until the server is ready to serve a VQ customer. **Figure 4** shows the second term $(E(L_v|l_s)b'(n+1))$ which is the expected extra time it will take a customer joining the VQ to leave the system. **Figure 5** shows the total expected waiting time. The graphs in Figure 3 are monotone increasing and concave, being a sum of such functions. In Figure 4 the graphs are monotone increasing and convex. In Figure 5 the graphs are monotone increasing very close to each other and almost linear, as the sum of the convex and concave functions cancels most of the slope.

The asymptotic slope, when $T \to \infty$, corresponds to the case where the VQ is always empty, and therefore a customer who joins it would wait in the VQ $l_s + 1$ M/M/1 busy periods with expected duration $(l_s + 1)/\mu(1 - \rho)$. Therefore the slope of the normalized curves in the figure is approximately $1/(1 - \rho)$ and $\hat{E}[W|l_s] \approx (l_s + 1)/(1 - \rho)$.

For a given threshold strategy $T = n + r$ followed by all customers we define the cost-difference function

$$f(l_s) = \frac{l_s + 1}{\mu} C_s - E(W|l_s)C_v. \tag{21}$$

A customer arriving when there are l_s customers in the SQ is indifferent between the two queues when $f(l_s) = 0$ and chooses to join the SQ when $f(l_s) < 0$. Therefore,

$$\frac{\mu f(l_s)}{C_s} = \hat{E}(W|l_s)\varphi - (l_s + 1) \ .$$

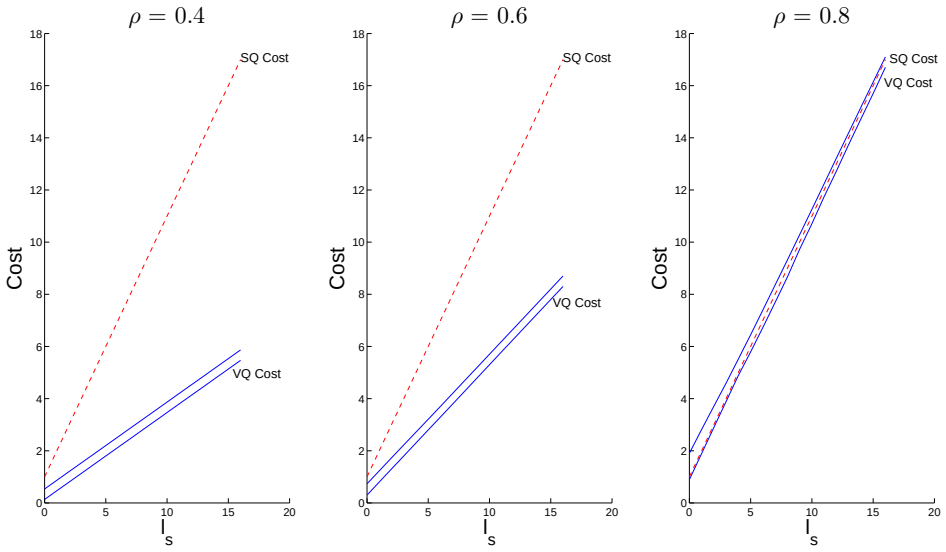

Figure 6: Cost of joining the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.2$

Substituting $\hat{E}(W|l_s) \approx (l_s + 1)/(1 - \rho)$ gives

$$\frac{\mu f(l_s)}{C_s} \approx (l_s + 1)\left(\frac{\varphi}{1-\rho} - 1\right). \tag{22}$$

Figures 6-8 show the two cost components of Equation (21) in each of the queues for different values of φ and ρ. The VQ costs are almost linear in l_s and very close to each other for close values of T, in accordance to the VQ expected waiting times shown in Figure 5. Therefore we have marked the area in the graph were all the VQ costs reside by two solid lines. The broken line is the expected cost associated with joining the SQ. When the VQ cost area is above the SQ cost line, the cost of joining the SQ is always smaller than the cost of joining the VQ and therefore all arriving customers will join the SQ regardless of its size. As one expects from Equation (22), this case occurs when φ is significantly greater than $1 - \rho$. Customers will always join the VQ in the opposite case, when φ is significantly smaller than $1 - \rho$. When the SQ line passes the VQ costs area or resides inside it there are values of T for which $f(l_s) = 0$ and therefore are candidates to be equilibrium strategies. This case is obtained when $\varphi \approx 1 - \rho$, because, as implied by Equation (22), in this case the expected waiting times at the SQ and the VQ are approximately equal. For example, in Figure 7, when $\rho = 0.4$ joining the VQ is a dominating strategy, when $\rho = 0.8$ joining

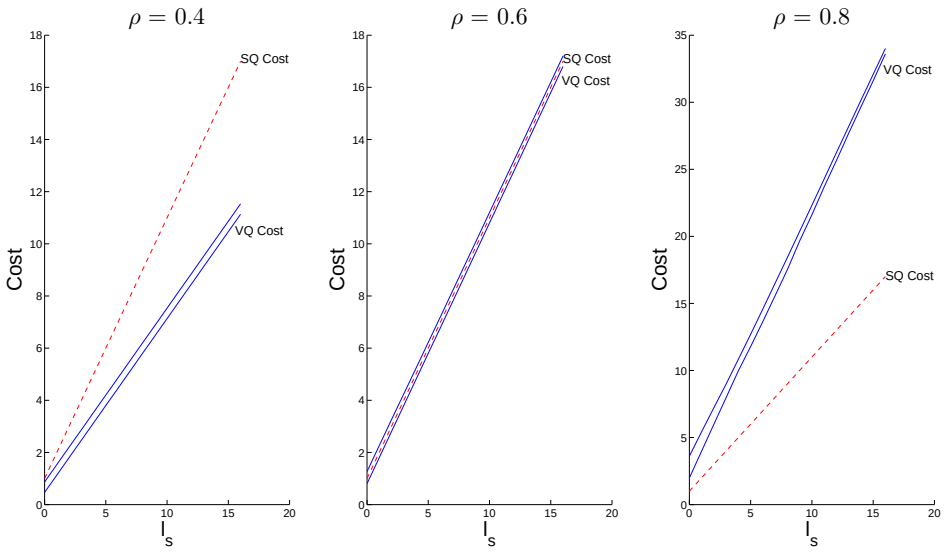

Figure 7: Cost in the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.4$

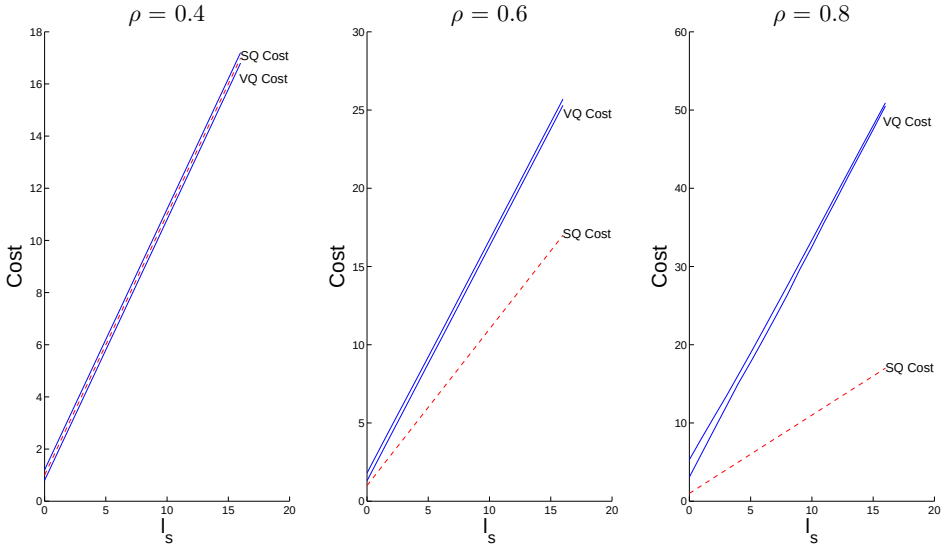

Figure 8: Cost in the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.6$

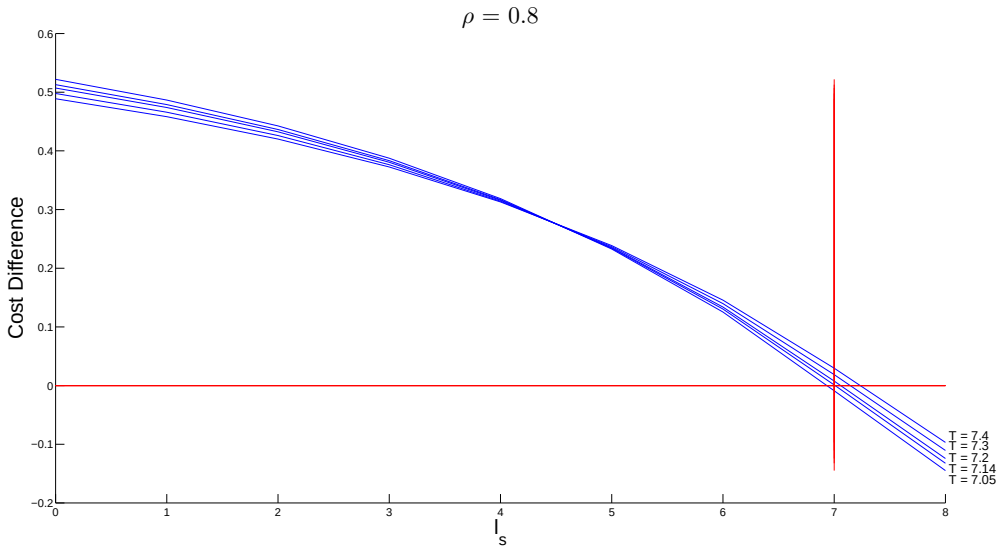

Figure 9: Expected cost difference SQ−VQ for thresholds $T = 7 + r$, $r = (0.05, 0.14, 0.2, 0.3, 0.4)$, $\varphi = 0.2$ and $\rho = 0.8$.

the SQ dominates, and when $\rho = 0.6$ we have $\varphi = 1 - \rho$ and (multiple) non-dominating equilibria exist.

We search for best response and equilibrium best response strategies for a given n by looking at the expected cost-difference graph. **Figure 9** plots the expected cost difference (SQ-VQ), a horizontal line at $y = 0$ and a vertical line $x = n$ for $n = 7$. In a pure equilibrium the cost-difference function is negative at $n - 1$ and positive at n. In a mixed equilibrium it is zero at n, as obtained for $T \approx 7.14$ in the figure.

Figures 10-12 show the best response for three pairs $(\rho, \varphi = 1 - \rho)$. As explained above, we expect non-dominating equilibria for such pairs. The intersection of the best response function and the 45° line are either equilibrium points or points at the end of the action space. For example, for $\varphi = 0.2$, $\rho = 0.8$ (Figures 10), if all others use a threshold $T = 1$ then the queue length is at most 2 and the best response of a customer is to join even when arriving to a system queue of this length. Thus $T = 1$ is not an equilibrium. The pure equilibria are at $T = 3, \ldots, 12$ and between every two pure equilibria n and $n+1$ there is a mixed equilibrium $n + r$ for some $0 < r < 1$. (The first ones are difficult to observe in the figure as r is close to 0.) These mixed solutions occur at the intersection of the 45° line with a vertical "jump" of the best response function, indicating indifference between

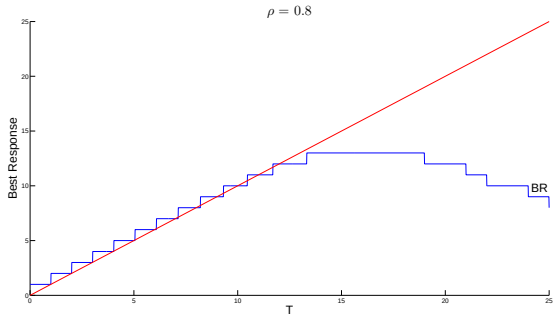

Figure 10: Best response for $\varphi = 0.2, \rho = 0.8$

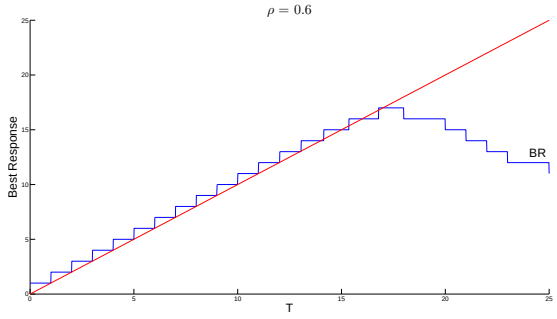

Figure 11: Best response for $\varphi = 0.4, \rho = 0.6$

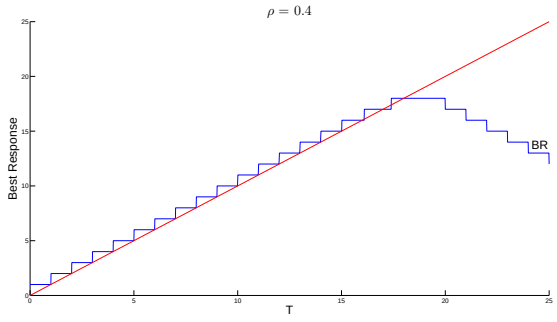

Figure 12: Best response for $\varphi = 0.6, \rho = 0.4$

two integer thresholds. In Figure 11 the pure equilibria are $12, \ldots, 17$, and in Figure 12 they are $15, \ldots, 18$.

The best response functions are increasing and then decreasing in l_s which reflect follow-the-crowd (FTC) and avoid the crowd (ATC) strategies respectively. FTC behavior is common in priority queues, but in our case when the threshold adopted by others is very high the individual prefers not to compete with them and joins the SQ instead. Another example of such behavior has been found by Haviv and Ravner [20] while investigating an accumulating priority queue system with different costs per unit time per customer class.

5 Concluding remarks

We investigated a non-preemptive single-server M/M/1 system with a System Queue (SQ) and a Virtual Queue (VQ). When the server is busy an arriving customer chooses between joining the VQ or the SQ. Waiting in the VQ is less costly. In the unobservable case customers are notified only whether the server is busy or not, and in the observable case they are also informed about l_s, the number of customers in the SQ.

For each case we compute the expected waiting time in each of the queues and the equilibrium joining strategy of an arriving customer. In the observable case, the conditional expected waiting time in the VQ turns out to be almost linear in l_s. We use this fact to characterize the equilibrium behavior, that shows similarities to that of the unobservable case. Multiple equilibrium strategies may exist, all in the monotone increasing part of the best response function, and this is compatible with a follow-the-crowd (FTC) behavior.

Clearly, since the customers' joining strategy has no effect on their expected waiting time, *it is socially optimal, in both models, that all customers join VQ*, and a social planner would encourage all the customers to join the VQ. This can be done in many ways involving penalties, subsidies, or a change in the service priority regime.

Our motivation for investigating service systems with virtual queues arises from call centers and similar systems where reaching VQ customers is practically costless and instantaneous. We believe however that analyzing similar systems in which contacting a customer from the VQ requires non-negligible time or expenditure is an interesting direction for future research. Such a system, but without an SQ, is solved by Economou and Kanta [12]. Another interesting variation of our model would allow customers to move from one queue to the other.

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] Aguir, M.S., F. Karaesmen, O.Z. Akşin, and F. Chauvet, (2004) *The impact of retrials on call center performance*, OR Spectrum 26, 353–376.

[3] Altman, E., T. Jiménez, R. Núñez-Queija, and U. Yechiali, (2004) *Optimal routing among ·/M/1 queues with partial information*, Stochastic Models 20, 149–171.

[4] Armony, M. and C. Maglaras, (2004) *On customer contact centers with a call-back option: customer decisions, routing rules, and system design*, Operations Research, 52, 271–292.

[5] Armony, M. and C. Maglaras, (2004) *Contact centers with a call-back option and real-time delay information*, Operations Research, 52, 527–545.

[6] Burgain, P., E. Feron, and J.-P Clarke, (2009) *Collaborative virtual queue: benefit analysis of a collaborative decision making concept applied to congested airport departure operations*, Air Traffic Control Quarterly 17, 195–222.

[7] Camulli, E., (2007) *Answer my call: technology helps utilities get customers off hold*, Electric Light and Power 2, 56.

[8] Chakravarthy, R.S., A. Krishnamoorthy, and V.C. Joshua, (2006) *Analysis of a multi-server retrial queue with search of customers from the orbit*, Performance Evaluation, 63, 776–798.

[9] Cooper, B.R., (1981) *Introduction to Queueing Theory - Second Edition*, North Holland.

[10] Cope III, R.F., R.F. Cope and H.E. Davis, (2008) *Disney's virtual queues: a strategic opportunity to co-brand services?*, Journal of Business and Economics Research 6, 13–20.

[11] Dickson, D., R.C. Ford, and B. Laval, (2005) *Managing real and virtual waits in hospitality and service organizations*, Cornell Hotel and Restaurant Administration Quarterly, 46, 52–68.

[12] Economou, A. and S. Kanta, (2011) *Equilibrium customer strategies and social-profit maximization in the single-server constant retail queue*, Naval Research Logistics 58, 107–122.

[13] Edelson, N.M. and K. Hildebrand, (1975) *Congestion tolls for Poisson queuing processes*, Econometrica 43, 81–92.

[14] Guijarro, L., V. Pla, and B. Tuffin, (2013) *Entry game under opportunistic access in cognitive radio networks: a priority queue model*, Wireless Days (WD), 1–6.

[15] Hassin, R., (1996) *On the advantage of being the first server*, Management Science, 42, 618–623.

[16] Hassin, R., (2016) Rational Queueing, CRC Press.

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

[18] Hassin, R. and M. Haviv, (2003) To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Kluwer Academic Publishers

[19] Haviv, M., (2013) Queues - A Course in Queueing Theory, Springer-Verlag, 191.

[20] Haviv, M. L. and Ravner, (2016) *Strategic bidding in an accumulating priority queue: equilibrium analysis*, Annals of Operations Research (to appear).

[21] Iravani, F. and B. Balcioǵlu, (2008) *On priority queues with impatient customers*, Queueing Systems, 58, 239–260.

[22] Kopzon, A., Y. Nazarathy, and G. Weiss, (2009) *A push pull queueing with infinite supply of work*, Queueing Systems: Theory and Applications, 66, 75–111.

[23] Kostami, V. and R.A. Ward, (2009) *Managing service systems with an offline waiting option and customer abandonment*, Manufacturing & Service Operations Management 11, 644–656.

[24] de Lange, R., I. Samoilovich, and B. van der Rhee, (2013) *Virtual queuing at airport security lanes*, European Journal of Operational Research 225, 153–165.

[25] Lovejoy, T.C., S. Aravkin, and C. Schneider-Mizell, (2004) *Kalman queue: an adaptive approach to virtual queuing*, The UMAP Journal 25, 337–352.

[26] Mandelbaum, A. and U. Yechiali, (1983) *Optimal entering rules for a customer with wait option at an M/G/1 queue*, Management Science, 29, 174–187.

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

[28] Littlechild, S.C., (1974) *Optimal arrival rate in a simple queueing system*, International Journal of Production Research 12, 391–397.

[29] Wüchner, P., J. Sztrik, and H. de Meer, (2009) *Finite-source M/M/S retrial queue with search for balking and impatient customers from the orbit*, Computer Networks 53, 1264–1273.

Appendix: Notations and proofs

Notations and Definitions

SQ	System queue.
VQ	Virtual queue.
$C_s,\ C_v$	Waiting costs in the SQ and VQ.
φ	The cost ratio, $\varphi = \frac{C_v}{C_s}$.
λ	Mean arrival rate of customers to the system.
μ	Mean service rate.
r_s	Probability of joining the SQ in the unobservable case.
$S(r_s)$	Expected net benefit for a customer following the strategy r_s.
$L_s(t),\ L_v(t)$	The number of customers in SQ and the VQ at time t.
$E(L),\ E(W_s),\ E(W_v)$	Expected waiting time in the system, SQ and VQ, respectively.
$\hat{E}[W\|l_s]$	The expected waiting time in the VQ in time units per customer.
$E(L_s),\ E(L_v)$	Expected number of customers in the SQ and VQ.
$\rho,\ \rho_s,\ \rho_v$	Occupation rates in the entire system, the SQ and the VQ.
P	Stationary probabilities.
l_s	Number of customers in the SQ.
$s(l_s)$	Threshold strategy of joining the SQ.
T	Threshold strategy, $T = n + r$ $(r \in [0,1], n \in \mathbb{N})$.
f	Number of unoccupied places in the SQ
$b(f)$	Mean busy period when there are $n + 1 - f$ customers in the SQ.
$b'(f)$	Normalized $b(f)$.

Proof of Proposition 4.1 From the transition rate diagram

$$(\lambda + \mu)P_{n+1,0} = \lambda r P_{n0}$$

and therefore

$$P_{n+1,0} = \frac{\rho r}{\rho + 1} P_{n0} \ . \tag{23}$$

A cut around nodes $0', 00, 10, \ldots, j0$ gives

$$\lambda P_{j0} = \mu P_{01} + \mu P_{j+1,0} \ , \ j = 0, 1, \ldots, n \ ,$$

or, after reindexing,

$$P_{j0} = \rho P_{j-1,0} - P_{01}.$$

We now substitute P_{01} from the upper horizontal cut equation

$$\lambda(1 - r)P_{n0} + \lambda P_{n+1,0} = \mu P_{01}$$

and $P_{n+1,0}$ from (23) and obtain

$$P_{j0} = \rho P_{j-1,0} - \frac{1 + \rho - r}{1 + \rho} \rho P_{n0} \ \ , \ j = 1, \ldots, n + 1 \ .$$

A recursive application of this equation gives

$$P_{j0} = \rho^j P_{00} - \frac{1+\rho-r}{1+\rho} P_{n0} \sum_{k=1}^{j} \rho^k \ , \ j = 1, \ldots, n+1 \ . \tag{24}$$

Specifically for $j = n$

$$P_{n0} = \rho^n P_{00} - \frac{1+\rho-r}{1+\rho} P_{n0} \sum_{k=1}^{n} \rho^k \ ,$$

giving

$$P_{n0} = \frac{(1+\rho)\rho^n}{(1+\rho)\sum\limits_{k=0}^{n} \rho^k - r \sum\limits_{k=1}^{n} \rho^k} P_{00} \ . \tag{25}$$

Substituting P_{n0} in (23) gives (4).

From Equations (25) and (24) we obtain

$$P_{j0} = \rho^j P_{00} - \frac{1+\rho-r}{1+\rho} \frac{(1+\rho)\rho^n}{(1+\rho)\sum\limits_{k=0}^{n} \rho^k - r \sum\limits_{k=1}^{n} \rho^k} \sum_{k=1}^{j} \rho^k P_{00}$$

which gives (3).

Proof of Proposition 4.2

Equation (8) follows from the horizontal cut between the rows $i, i-1$

$$\lambda(1-r)P_{ni} + \lambda P_{n+1,i} = \mu P_{0,i+1} \ , \ i = 0, 1, 2, \ldots \ . \tag{26}$$

A cut that contains the nodes $0i, 1i, \ldots, ji$ gives

$$\lambda P_{ji} + \mu P_{0i} = \mu P_{j+1,i} + \mu P_{0,i+1} \quad \begin{matrix} i = 1, 2, \ldots \\ j = 0, 1, 2, \ldots, n-1 \end{matrix} \ . \tag{27}$$

By Equation (26) and re-indexing we have

$$P_{ji} = P_{0i} - (1-r)\rho P_{ni} - \rho P_{n+1,i} + \rho P_{j-1,i} \quad \begin{matrix} i = 1, 2, \ldots \\ j = 1, 2, \ldots, n \end{matrix} \ .$$

A recursive application of this relation leads to Equation (7).

We now find an expression for P_{ni} and $P_{n+1,i}$. Equation (7) for P_{ni} yields

$$\left[\sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k \right] P_{ni} = \sum_{k=0}^{n} \rho^k P_{0i} - \sum_{k=1}^{n} \rho^k P_{n+1,i} \ , \tag{28}$$

and the cut around the node $(n+1)i$ is

$$(1+\rho)P_{n+1,i} = \rho P_{n+1,i-1} + \rho r P_{ni} \ . \tag{29}$$

By Equations (8), (28) and (29) we have

$$P_{ni} = \frac{(1+\rho)(1-r)\sum_{k=1}^{n+1}\rho^k}{\sum_{k=1}^{n+1}\rho^k + \sum_{k=0}^{n}\rho^k - r\sum_{k=1}^{n}\rho^k} P_{n,i-1} + \frac{\sum_{k=1}^{n+2}\rho^k}{\sum_{k=1}^{n+1}\rho^k + \sum_{k=0}^{n}\rho^k - r\sum_{k=1}^{n}\rho^k} P_{n+1,i-1} \ . \tag{30}$$

With this result and Equation (29) we also get

$$\begin{aligned} P_{n+1,i} = & \left[1 + \frac{\rho r \sum_{k=1}^{n+2}\rho^k}{\sum_{k=1}^{n+1}\rho^k + \sum_{k=0}^{n}\rho^k - r\sum_{k=1}^{n}\rho^k} \right] \frac{\rho}{\rho+1} P_{n+1,i-1} \\ & + \frac{(1+\rho)(1-r)r\sum_{k=2}^{n+2}\rho^k}{\sum_{k=1}^{n+1}\rho^k + \sum_{k=0}^{n}\rho^k - r\sum_{k=1}^{n}\rho^k} P_{n,i-1} \ . \end{aligned} \tag{31}$$

In terms of X, Y, Z as defined in Equations (9) - (11) we have

$$P_{ni} = YP_{n,i-1} + XP_{n+1,i-1} \ , \ P_{n+1,i} = ZP_{n+1,i-1} + \frac{\rho r}{1+\rho} YP_{n,i-1}$$

therefore

$$P_{ni} = Y^i P_{n0} + X\sum_{k=0}^{i-1} Y^k P_{n+1,i-1-k} \ ,$$

$$P_{n+1,i} = Z^i P_{n+1,0} + \frac{\rho r}{1+\rho} Y \sum_{k=0}^{i-1} Z^k P_{n,i-1-k}$$

and we get (5) and (6).