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

Roei Engel *
Department of Statistics and Operation Research
Tel Aviv University

A thesis submitted in partial satisfaction of the requirements
for the degree master in operational research (M.Sc.)

Supervised by Prof. Rafi Hassin

February 1, 2016

*RoeiEngel@Gmail.com

Abstract

Consider a non-preemptive $M/M/1$ queuing system with two first-come first-served queues, virtual (VQ) and system (SQ), differing by their waiting cost per unit time. An arriving customer who finds the server idle enters service immediately. If the server is busy the customer chooses which queue to enter. Customers in the SQ have non-preemptive priority over those in the VQ, but waiting in the SQ is more costly than waiting in the VQ.

We study two information models of the system. In the unobservable model customers are notified only whether the server is busy or not, and on the observable model customers are also informed about the number of customers currently waiting in the SQ. For each model we investigate the stationary distributions, the Nash equilibrium of the joining strategies and the socially optimal joining strategy.

Contents

1 Introduction and literature review **4**

2 Modeling the single-server virtual queuing system **7**

3 The unobservable model **8**

4 The observable model **15**
4.1 Pure strategy customers 15
4.2 Mixed strategy customers 21
4.3 Numerical investigation of the mixed strategy 29

5 Concluding remarks **39**

6 Notations and definitions **41**

1 Introduction and literature review

Companies worldwide use call centers as an important channel of interaction with their customers. As a company grows the high unpredictability of the number of customers contacting the call center at any given moment rises, thus increasing workload and applying additional strain on its limited resources. Customers generally do not like to wait in queues and feel it is a waste of their precious time. Naturally as waiting grows so does their anger towards the company and its agents which will influence their evaluation of the service (eg., Taylor [25]) and increase the chances of their leaving the company and switching to a competitor.

One way call centers can handle the inbound workloads and customer demands is employing the "virtual queuing" concept. Virtual queuing systems offer an arriving customer the option of a callback. When busy, the system allows for choosing whether to stay on hold and wait in the "system queue" or to secure a place in a "virtual queue" and hang up. A customer choosing this option will receive a callback when his time for service arrives, therefore reducing waiting costs by freeing him to perhaps do some other business while waiting, and hopefully reduce his level of anger.

In this thesis 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 enter, based on the waiting costs and the available information. Our goal is to determine the Nash equilibrium and social optimal strategy of our single server virtual queuing system in two information models, an unobservable model where the customer is only informed whether the server is busy, and an observable model where the customer is also informed of the number of customers in the SQ. Our systems differs from the basic two priority queues system by having different waiting cost for each queue and by having partial information model where only the SQ is observable. For the unobservable model we find explicit equations for the expected number of customers and the expected waiting time in both queues as well as the equilibrium joining and social optimal joining strategies to the SQ. For the observable model we find explicit equations of the stationary probabilities, the mean busy period in the SQ, the expected number of customers and expected waiting time in the VQ and the social optimal strategy. We also present numerical investigation of the best response and equilibrium customers (mixed) strategy for different cost values, utilization rates (ρ) and the number of customers in the SQ.

Customer decision-making and Nash Equilibrium in queues were initially defined and investigated by Naor [23], Littlechild [24] and Edelson and Hildebrand [14], fundamentals and basic concepts can be found in Hassin,and Haviv [18] and in Adan and Resing [2].

Virtual queues, and their use in real life applications are extensively discussed in the literature. A few examples relate to such applications are hospitality sector in Dickson, Ford, and Laval [12], airports in Burgain, Feron, and Clarke

[7], de Lange, Samoilovich, and van der Rhee [11], amusement parks in Robert, Rachelle and Harold [26], Aravkin, Lovejoy, and Schneider-Mizell [21], and call centers by Camulli [8] and Armony and Maglaras [5], [6].

Several papers introduce customer decision-making when facing a choice between an SQ and a VQ. Guijarro, Pla, and Tuffin [15] in their second model investigate an unobservable multi-server system with an SQ and a VQ with different entrance costs where the queues are managed by different servers (vendors) that compete with each other and each profits from its own queue separately. The authors investigate two-stage sequential game where in the first stage the servers choose their entrance prices and in the second stage the packet (customer) chooses which queue to enter based on the entrance price only. Hassin [16] models a system with two servers (gas stations) each with a different SQ, there is no VQ and the costs are the same in both SQ per time unit, as in our observable information model, an arriving customer (driver) can only see the length of one queue and needs to choose which queue to enter based on the expected length of the queue. The author investigates the equilibrium thresholds in pure and mixed strategies. this model is later described in a different context by Altman, Jiménez, núñez-Queija, and Yechiali [4] as individual optimal routing model and the authors provides a verifiable necessary condition to check whether an equilibrium exists within the threshold policies. Mandelbaum and Yechiali [22], investigate the optimal strategy of an arriving "smart" customer in a single server system, the arriving customer can choose to enter the queue, leave the system or to wait outside of the queue for reduced cost - which can be viewed as entering a private virtual queue. Aguir, Karaesmen, and Chauvet [3], investigate a multi-server call center system with an SQ and a VQ (orbit, not FCFS) where the queues do not have waiting costs and the customers are impatient and can balk or abandon the SQ to the VQ. The paper proposes a fluid approximation to the stationary and non-stationary settings of the system and estimate the arrival rates based on real demand data. Chakravarthy, Krishnamoorthy and Joshua [9] model a multi-server system with only an orbit VQ where there is no waiting cost and customers compete by sending out signals at random times until a server is free, and when the server is free it searches for waiting customers in the VQ. Iravani and Balcioǵlu [19] consider a multi-server system with an SQ and a VQ where impatient customers choose which queue to enter with exogenous probability and with no waiting costs. The authors investigate the steady-state performance measures of the SQ and the factorial moments of the VQ length (third model). Wüchner, Sztrik and de Meer [27] provide numerical analysis on a system with an SQ and a VQ with where customers are allowed to balk and move between the queues, but with no waiting costs. Kostami, and Ward [20] model a single server with inline (system) and offline (virtual) queues where arriving customers choose which queue to enter according to the waiting time estimations by the server and the offline customers can leave the system without the server knowledge. Armony and Maglaras investigate two close models in [5] and [6] of a multi-server call center system with an SQ and a VQ and customers can choose which queue to enter, or to balk, but customers who enter the VQ have a grantee of their maximum delay, there is

no waiting costs difference between the queues and after each service the server decides whether to take a caller from the SQ or from the VQ.

We can also consider our system as a priority-based system where the customers who enter the SQ have high priority (class 1) and those who enter the VQ are the low-priority (class 2) customers. The fundamental model involving customer decisions in queues with priorities is defined by Adiri and Yechiali [1]. Hassin and Haviv [17] investigate the observable case of a system with different payment method, where both queues are observable and the waiting cost in both is identical but the admission price is different. Hassin and Haviv [18], §4.2 investigate the unobservable case of that system.

Economou and Kanta [13] investigate the closest model to our own, they model a single-server system with no waiting space (no SQ), where customers who find the server busy automatically enter the VQ. After finishing service the server seeks a customer in the VQ with an exponentially distributed search time (not immediate as in our model). If a new customer arrives to the system during the search time the server interrupts the search and servers 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.

This thesis is structured as follows. Section 2 presents the basic model and assumptions. In section 3 we model and investigate the unobservable model. Section 4 is divided to three parts, sub-sections 4.1 and 4.2 investigate the observable model when the customers use pure strategies and mixed strategies respectively, and sub-section 4.3 contains our numerical investigation of the mixed strategies case. Section 5 discusses future work that can be done by expanding this model or by investigating interesting variations of it. All notations and definitions can be found at Section 6

2 Modeling the single-server virtual queuing system

Consider a system modeled as a non-preemptive single-server M/M/1 virtual queuing system. The system has two queues, a System Queue (SQ) and a virtual queue (VQ) which is often referred to as orbit queue. An arriving customer who finds the server idle enters the service immediately, and when the server is busy the customer decides whether to enter 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 of the queues is FCFS, customers arrive according to a Poisson process at rate λ, their service times are exponentially distributed random variables with rate μ. 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.
An arriving customer who finds the server idle will enter the service immediately, will not be effected by any other customers and his expected sojourn time is $\frac{1}{\mu}$ regardless to any information provided. Therefore we conduct our study on the behavior of the customers who find the server busy and need to decide which queue to enter. We consider two information cases, an unobservable queue in which an arriving customer is only informed of the state of the server - busy or ideal, and an observable queue where the arriving customer is also informed of the number of customers currently waiting in the SQ.

A special case of our system is when the system has only a VQ (or when $C_s \gg C_v$), both the observable and unobservable cases have been described and investigated by Economou and Kanta [13]. Another special case is when the system has only a SQ (when $C_s = C_v$).

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 a mixed strategy (r_s, r_v) were r_s (r_v) is the probability for entering the SQ (VQ). Since the customers cannot balk we have $r_v = 1 - r_s$ and the mixed strategy is defined only by one of its elements r_s. $S(r_s)$ is defined as the expected net benefit for a customer following the strategy r_s.

In order to calculate the equilibrium strategy we first need to find the expected waiting time of an arriving customer in each of the queues. Define $E(W_s^q)$ and $E(W_v^q)$ as the expected waiting time in the queue of a joining customer to SQ and VQ accordingly and $E(L_s^q)$ and $E(L_v^q)$ as the expected number of customers in each of the queues. We use the indicator function I for the server state (1-busy, 0-idle).

The arrival rates to the SQ and VQ when the server is busy are λr_s and λr_v, and the fraction of times that the server is working (the occupation rates) for the entire system and for each of the queues separately are ρ, ρ_s, ρ_v where

$$\rho = \frac{\lambda}{\mu} \ , \ \rho_s = \frac{\lambda r_s}{\mu} \ , \ \rho_v = \frac{\lambda r_v}{\mu} \ . \tag{3.1}$$

We represent the state of the system at time t in which the server is busy by a random vector $(L_s(t), L_v(t))$ where $L_s(t)$ and $L_v(t)$ are the number of customers in SQ and the VQ. This stochastic process is a continuous-time Markov chain with state space S, where $S = 0, 1, 2 \ldots \times 0, 1, 2 \ldots$. We represent the state of the system when the server is idle with the special tag $0'$. The transitions rates for the entire system are

$$q_{(0'),(0,0)} = \lambda \tag{3.2}$$

$$q_{(0,0),(0')} = \mu \tag{3.3}$$

$$q_{(0,i),(0,i-1)} = \mu \quad i = 0, 1, 2, \ldots \tag{3.4}$$

$$q_{(j,i),(j-1,i)} = \mu \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 1, 2, 3, \ldots \end{matrix} \tag{3.5}$$

$$q_{(j,i),(j+1,i)} = r_s\lambda \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 0, 1, 2, \ldots \end{matrix} \tag{3.6}$$

$$q_{(j,i),(j,i+1)} = r_v\lambda \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 0, 1, 2, \ldots \end{matrix} \tag{3.7}$$

The transition rate diagram of the system is shown in Figure 1.

We denote the stationary distribution as P_{ji} where $(j, i) \in S$ and the special state where the server is idle as $P_{0'}$. Also define "$\bullet$" as the sum of an entire column or a row, i.e.,

$P_{0\bullet} = \sum_{i=0}^{\infty} P_{0i}$, $P_{\bullet 0} = \sum_{j=0}^{\infty} P_{j0}$, $P_{\bullet\bullet} = \sum_{j,i=0}^{\infty} P_{ji} (= 1 - P_{0'})$ and so on.

We start by finding the probabilities that the server is busy and idle via the transition rates.

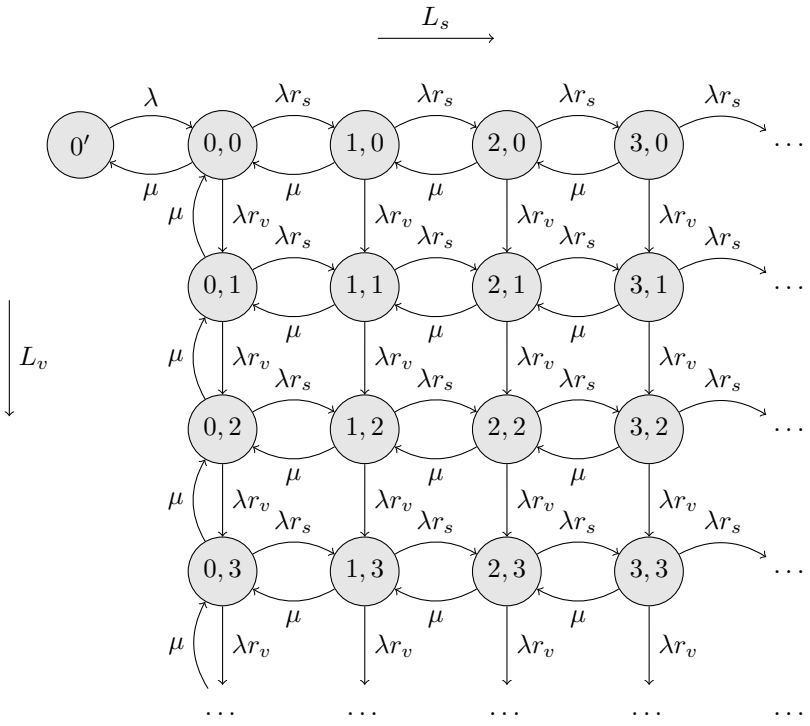

Figure 1: Transition rate diagram for the unobservable model

Proposition 3.1.

$$P_{0'} = \frac{\mu - \lambda}{\mu} = 1 - \rho \tag{3.8}$$

$$P_{\bullet\bullet} = \frac{\lambda}{\mu} = \rho \tag{3.9}$$

Remark This result is compatible with a regular M/M/1 priority queue system with two queues where the probability that the server is idle is $\pi_0 = 1 - \rho_1 - \rho_2$ and the probability that the server is busy is $\pi_1 = \rho_1 + \rho_2$

Proof. From the transition rate diagram,
the vertical cuts are

$$\mu P_{j\bullet} = \lambda r_s P_{j-1,\bullet} \ , \ j = 1, 2, 3, \dots \ , \tag{3.10}$$

the horizontal cuts are

$$\lambda r_v P_{\bullet i} = \mu P_{0,i+1} \ , \ i = 0, 1, 2, \dots \ , \tag{3.11}$$

the general equation is

$$P_{0'} + P_{\bullet\bullet} = 1 \ , \tag{3.12}$$

and the cut around $P_{0'}$ is

$$\lambda P_{0'} = \mu P_{00} \ . \tag{3.13}$$

From the vertical cuts (3.10)

$$P_{j\bullet} = \left(\frac{\lambda r_s}{\mu}\right)^j P_{0\bullet} \ , \ j = 1, 2, 3, \dots \tag{3.14}$$

and since $\frac{\lambda r_s}{\mu} \leq 1$

$$P_{\bullet\bullet} = \sum_{j=0}^{\infty} P_{j\bullet} = P_{0\bullet} \sum_{j=0}^{\infty} \left(\frac{\lambda r_s}{\mu}\right)^j = \frac{1}{1 - \frac{\lambda r_s}{\mu}} P_{0\bullet} = \frac{\mu}{\mu - \lambda r_s} P_{0\bullet} \ . \tag{3.15}$$

The horizontal cuts (3.11) gives

$$P_{\bullet\bullet} = \sum_{i=0}^{\infty} P_{\bullet i} = \frac{\mu}{\lambda r_v} \sum_{i=0}^{\infty} P_{0,i+1}$$

and since

$$\sum_{i=0}^{\infty} P_{0,i+1} = P_{0\bullet} - P_{00}$$

with equation (3.13) we have

$$P_{\bullet\bullet} = \frac{\mu P_{0\bullet} - \lambda P_{0'}}{\lambda r_v} \ . \tag{3.16}$$

From equations (3.15) and (3.16)

$$P_{0\bullet} = \frac{\lambda(\mu - \lambda r_s)}{\mu(\mu - \lambda(r_s + r_v))} P_{0'} \tag{3.17}$$

therefore

$$P_{\bullet\bullet} = \frac{\lambda}{\mu - \lambda} P_{0'}$$

and from the general equation (3.12) we have equations (3.8) and (3.9).

□

We now find the expected waiting times in the queues.

Proposition 3.2.
Given that the server is busy, the expected waiting times in the system of a joining customer to each of the queues are

$$E(W_s^q | I = 1) = \frac{1}{(1 - \rho_s)\mu} \tag{3.18}$$

$$E(W_v^q | I = 1) = \frac{1}{(1 - \rho)(1 - \rho_s)\mu} \ . \tag{3.19}$$

Remark This result is compatible with the regular M/M/1 priority queue system with two queues and preemptive regime, where expected waiting time is $E(W_1) = \frac{1}{(1-\rho_1)} \frac{1}{\mu}$ and $E(W_2) = \frac{1}{(1-\rho_1)(1-\rho_1-\rho_2)} \frac{1}{\mu}$.

Proof. We start by finding the expected number of customers in each of the queues ($E(L_s^q)$, $E(L_v^q)$) when the server is busy. From equations (3.8),(3.14) and (3.17) we have

$$P_{j\bullet} = \frac{\lambda}{\mu^2(\mu - \lambda r_s)} \left(\frac{\lambda r_s}{\mu} \right)^j \ , \ j = 1, 2, 3, \dots \ . \tag{3.20}$$

Given that the server is busy the probability that there are j customers in the SQ queue is

$$P(X = j | I = 1) = \frac{P(X = j \wedge I = 1)}{P(I = 1)} = \frac{P_{j\bullet}}{P_{\bullet\bullet}} = \frac{\mu - \lambda r_s}{\mu} \left(\frac{\lambda r_s}{\mu} \right)^j$$

and the expected number of customers in the SQ is

$$E(L_s^q | I = 1) = \sum_{j=1}^{\infty} j P(X = j | I = 1) = \frac{\mu - \lambda r_s}{\mu} \sum_{j=1}^{\infty} j \left(\frac{\lambda r_s}{\mu} \right)^j = \frac{\lambda r_s}{\mu - \lambda r_s} \ .$$

In terms of ρ we have

$$E(L_s^q | I = 1) = \frac{\rho_s}{1 - \rho_s} \ . \tag{3.21}$$

In order to find $E(L_v^q)$ we start by looking at the expected number of waiting customers in the entire system, $E(L^q)$. Since the service times are exponentially distributed with the same mean, and the total number of waiting customers in the system does not depend on the order in which the customers are served, $E(L^q)$ is the same as in M/M/1 FIFO system with no queues. Hence

$$E(L^q) = \frac{\rho^2}{1-\rho} \ .$$

Since $E(L^q) = P_{0'}E(L^q|I=0) + P_{\bullet\bullet}E(L^q|I=1)$ and $E(L^q|I=0) = 0$ we have

$$E(L^q|I=1) = \frac{E(L^q)}{P_{\bullet\bullet}} = \frac{\rho}{1-\rho}$$

and for $E(L_v^q|I=1)$,

$$E(L_v^q|I=1) = E(L^q|I=1) - E(L_s^q|I=1)$$

$$= \frac{\rho}{1-\rho} - \frac{\rho_s}{1-\rho_s} = \frac{\rho-\rho_s}{(1-\rho)(1-\rho_s)} = \frac{\rho_v}{(1-\rho)(1-\rho_s)} \ . \tag{3.22}$$

Now by applying Little's law we can find the expected conditional waiting times in the queues,

$$E(W_s^q|I=1) = \frac{E(L_s^q|I=1)}{\lambda r_s} = \frac{1}{(1-\rho_s)\mu}$$

and

$$E(W_v^q|I=1) = \frac{E(L_v^q|I=1)}{\lambda r_v} = \frac{1}{(1-\rho)(1-\rho_s)\mu} \ .$$

□

Now we can find the symmetric equilibrium strategy for an arriving customer. We assume that when the customer is indifferent in his choice between the queues he will enter the SQ to receive the advantage of the high priority.

Theorem 3.3.
When the server is busy, there exists a unique symmetric equilibrium strategy of joining the SQ, r_s^e, that is given by

$$r_s^e = \begin{cases} 1 & \frac{C_v}{C_s} + \rho \geq 1^* \ , \\ \\ 0 & \frac{C_v}{C_s} + \rho < 1 \ . \end{cases} \tag{3.23}$$

Proof. If all customers follow the same joining strategy r_s then the system expected waiting times are as described in proposition 3.2.
Define the cost function $\frac{f(x)}{C_s}$ where

$$f(x) = E(W_v^q|I=1)C_v - E(W_s^q|I=1)C_s \ ,$$

an arriving customer will be indifferent in his choice to enter the queues when $f(r_s) = 0$ will choose to enter the SQ when $f(r_s) > 0$ and to the VQ when $f(r_s) < 0$. With proposition 3.2 we have

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

Checking when $\frac{f(r_s)}{C_s} = 0$ we get

$$\frac{C_v}{C_s} + \rho = 1 \ .$$

Therefore,

$$\begin{aligned} &\text{when } \frac{C_v}{C_s} + \rho > 1 \quad \text{then } \frac{f(r_s)}{C_s} > 0 \ , \\ &\text{when } \frac{C_v}{C_s} + \rho < 1 \quad \text{then } \frac{f(r_s)}{C_s} < 0 \end{aligned}$$

and we get

$$r_s^e = \begin{cases} 1 & \frac{C_v}{C_s} + \rho \geq 1 \\ \\ 0 & \frac{C_v}{C_s} + \rho < 1 \end{cases}$$

meaning a pure strategy. When $\frac{C_v}{C_s} + \rho \geq 1$ arriving customer will choose to enter the SQ and when $\frac{C_v}{C_s} + \rho < 1$ he will enter the VQ. □

We can now find r_s^{soc}, the socially optimal probability of joining the SQ.

Theorem 3.4.
When the server is busy, the unique mixed strategy of joining the SQ, r_s^{soc}, that maximizes the social net benefit per unit time is

$$r_s^{soc} = 0 \ . \tag{3.24}$$

Proof. For a given joining strategy r_s we define the social cost per unit time function as $\frac{f(r_s)}{C_s}$, where

$$f(r_s) = [C_s \lambda r_s E(W_s^q | I = 1) + C_v \lambda r_v E(W_v^q | I = 1)]$$

The socially-optimal strategy minimizes this function, meaning

$$\begin{aligned} &\min_{r_s} \frac{\lambda r_s}{(1-\rho_s)\mu} + \frac{\frac{C_v}{C_s}\lambda r_v}{(1-\rho)(1-\rho_s)\mu} \\ &= \min_{r_s} \frac{(1-\rho)r_s + (1-r_s)\frac{C_v}{C_s}}{(1-\rho)(1-\rho r_s)}\rho \ . \end{aligned}$$

The derivative is

$$\frac{(1-\frac{C_v}{C_s})\rho}{(1-\rho r_s)^2} > 0 \ ,$$

and so the minimum of the function will be at the ends of the domain,

$$\begin{aligned} \text{when } r_s = 0 \quad &\text{we get } \frac{\frac{C_v}{C_s}\rho}{1-\rho} > 0 \\ \text{and when } r_s = 1 \quad &\text{we get } \frac{\rho}{1-\rho} > 0 \ . \end{aligned}$$

Since $\frac{C_v}{C_s} < 1$ we get that the minimum, and the socially optimal strategy is

$$r_s^{\text{soc}} = r_s = 0 \ .$$

□

Therefore the social perspective would be to encourage all the customers to enter the VQ. This can be done by decreasing the waiting costs of the VQ or increasing them at the SQ.

4 The observable model

In the observable case, an arriving customer is informed about the state of the server **and** of the exact number of customers in the SQ. This new information is useful when the server is busy since now the customer can evaluate his expected waiting time the queues and use this in order improve his decision. It is reasonable for the customers to consider using threshold strategies in this case, meaning an arriving customer will enter the SQ when its length is shorter than a given threshold, will use a mixed strategy when the length equals the threshold, and enter the VQ when it is longer. We define the threshold strategy as defined by Hassin (1996) [16].

A threshold strategy $s(l_s)$ to join the SQ with threshold $T = N + r$ ($r \in [0,1), N \in \mathbb{N}$), where l_s is the current number of customers in the SQ, 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 will always join the SQ when the number of customers is at most $N-1$, will join the VQ if it is longer than N, and will randomize with probability r when there are exactly N customers in the SQ. In the case where $r = 0$ then $s(l_s)$ is a pure strategy and when a customer observes the length N it will join the VQ. Therefore the number of customers in the SQ is at most $N+1$ and when the server is busy the SQ has a similar behavior to an M/M/1/$N+1$ system.

4.1 Pure strategy customers

In this section we derive the stationary probabilities when all arriving customers follow a pure strategy ($r = n$). i.e.,

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

Although the pure strategies solution is not required in order to solve the mixed strategies case we have decided to add this section to our thesis since it has a clean closed solution that can be used when only pure strategies are used by the customers.

Remark The special cases when $n \to \infty$ and when $n = 0$ are the cases where all the customers always enter only one of the queues, and therefore both are similar to an M/M/1 queue.

For a threshold strategy with threshold $T = n$ the balance equations of the system are

$$q_{(0'),(0,0)} = \lambda \tag{4.1}$$

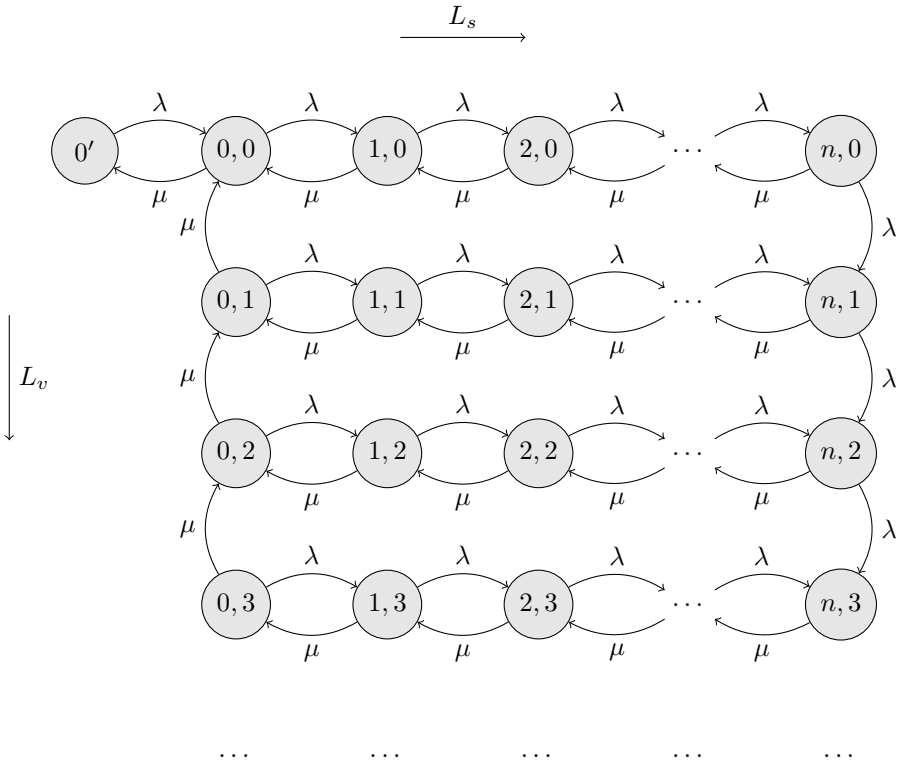

Figure 2: Transition rate diagram in the observable model when customers follow the pure threshold strategy $T = n$.

$$q_{(0,0),(0')} = \mu \tag{4.2}$$

$$q_{(n,i),(n,i+1)} = \lambda \quad i = 0, 1, 2, \ldots \tag{4.3}$$

$$q_{(0,i),(0,i-1)} = \mu \quad i = 0, 1, 2, \ldots \tag{4.4}$$

$$q_{(j,i),(j-1,i)} = \mu \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 1, 2, 3, \ldots, n \end{matrix} \tag{4.5}$$

$$q_{(j,i),(j+1,i)} = \lambda \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 0, 1, 2, \ldots n-1 \ . \end{matrix} \tag{4.6}$$

The transition rate diagram is shown in Figure 2.

From the same considerations we used in the proof of Proposition 3.1 in the unobservable model we have

$$P_{0'} = 1 - \rho \ , \tag{4.7}$$

$$P_{\bullet\bullet} = \rho \tag{4.8}$$

and from the transition rate (4.1) we immediately have

$$P_{00} = (1-\rho)\rho \ . \tag{4.9}$$

We start by finding the stationary probabilities $P_{01}, P_{10}, \ldots, P_{n0}$.

Proposition 4.1.

$$P_{n0} = \frac{\rho^n}{\sum\limits_{k=0}^{n} \rho^k} P_{00} \ , \tag{4.10}$$

$$P_{01} = \frac{\rho^{n+1}}{\sum\limits_{k=0}^{n} \rho^k} P_{00} \ , \tag{4.11}$$

and

$$P_{j0} = \frac{\rho^j \sum\limits_{k=0}^{n} \rho^k - \rho^n \sum\limits_{k=1}^{j} \rho^k}{\sum\limits_{k=0}^{n} \rho^k} P_{00} \ , \ j = 1, \ldots, n \ . \tag{4.12}$$

Proof. From the transition rate diagram, the horizontal cuts are

$$\lambda P_{ni} = \mu P_{0,i+1} \ , \ i = 0, 1, 2, \ldots \ . \tag{4.13}$$

A cut that contains the nodes $0', 00, 10, \ldots, j0$ gives the equation

$$\lambda P_{j0} = \mu P_{01} + \mu P_{j+1,0} \ , \ j = 0, 1, \ldots, n-1 \ . \tag{4.14}$$

Using equation (4.13) for $i = 0$ we have

$$\lambda P_{j0} = \lambda P_{n0} + \mu P_{j+1,0} \ , \ j = 0, 1, \ldots, n-1 \ ,$$

and by re-indexing we get

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

By recursively positioning this equation for all the partial cuts of the cut $0', 00, 10, \ldots, j0$ we can explicitly write the probability P_{j0} in terms of P_{00}

$$P_{j0} = \rho^j P_{00} - \sum_{k=1}^{j} \rho^k P_{n0} \ . \tag{4.15}$$

Specifically for P_{n0}

$$P_{n0} = \rho^n P_{00} - \sum_{k=1}^{n} \rho^k P_{n0}$$

giving equation (4.10). By using (4.13) we get equation (4.11). From equations (4.10) and (4.15)

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

which gives (4.12). □

We now find the stationary probability for a generic node ji, P_{ji}, in terms of P_{00}.

Proposition 4.2.

$$P_{ji} = \frac{\rho^{n+i}}{\sum_{k=0}^{n} \rho^k} P_{00} \ , \quad \begin{matrix} i = 1, 2, \dots \\ j = 0, 1, \dots n \end{matrix} \tag{4.16}$$

Proof. We start with P_{0i}, a cut that contain the nodes $0i, 1i, \dots, ji$, the cut has the equation is

$$\lambda P_{ji} + \mu P_{0i} = \mu P_{j+1,i} + \mu P_{0,i+1} \quad \begin{matrix} i = 1, 2, \dots \\ j = 0, 1, 2, \dots n-1 \ . \end{matrix} \tag{4.17}$$

By using equation (4.13)

$$\lambda P_{ji} + \mu P_{0i} = \mu P_{j+1,i} + \lambda P_{ni} \quad \begin{matrix} i = 1, 2, \dots \\ j = 0, 1, 2, \dots n-1 \end{matrix}$$

and by re-indexing we get

$$P_{ji} = P_{0i} - \frac{\lambda}{\mu} P_{ni} + \frac{\lambda}{\mu} P_{j-1,i} = P_{0i} - \rho P_{ni} + \rho P_{j-1,i} \quad \begin{matrix} i = 1, 2, \dots \\ j = 1, 2, \dots n \end{matrix} \ .$$

When we use these equations of all the partial cuts of the cut $0i, 1i, \dots, ji$ we can explicitly write the probability in terms of P_{0i}

$$P_{ji} = (P_{0i} - \rho P_{ni}) \sum_{k=0}^{j-1} \rho^k + \rho^j P_{0i} \ . \tag{4.18}$$

Specifically for P_{n0} we have

$$P_{ni} = (P_{0i} - \rho P_{ni}) \sum_{k=0}^{n-1} \rho^k + \rho^n P_{0i} \ ,$$

which gives

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

and therefore

$$P_{ni} = P_{0i} \ , \ i = 1, 2, \dots \ . \tag{4.19}$$

Substituting P_{ni} from (4.13) gives

$$P_{0i} = \rho P_{0,i-1} \ , \ i = 2, 3, \dots$$

and summing these equation for all node $01, 02, \dots, 0i$ gives

$$P_{0i} = \rho^{i-1} P_{01} \ , \ i = 2, 3, \dots \ .$$

This result and equation (4.11) give

$$P_{0i} = \frac{\rho^{n+i}}{\sum\limits_{k=0}^{n} \rho^k} P_{00} \ \ i = 1, 2, 3, \dots \ . \tag{4.20}$$

Now for P_{ji}, substitute (4.19) in (4.18) to get

$$P_{ji} = (P_{0i} - \rho P_{0i}) \sum_{k=0}^{j-1} \rho^k + \rho^j P_{0i} = \left[(1-\rho) \sum_{k=0}^{j-1} \rho^k + \rho^j \right] P_{0i}$$

and therefore

$$P_{ji} = P_{0i} \ . \tag{4.21}$$

Finally, substituting equation (4.20) in (4.21) gives (4.16). □

From proposition 4.2, the probability of an entire column j is

$$P_{j\bullet} = \sum_{i=0}^{\infty} P_{ji} = P_{j0} + \sum_{i=1}^{\infty} P_{ji} = P_{j0} + \sum_{i=1}^{\infty} \frac{\rho^{n+i}}{\sum\limits_{k=0}^{n} \rho^k} P_{00} =$$

$$P_{j0} + \frac{\rho^n}{\sum\limits_{k=0}^{n} \rho^k} \frac{\rho}{1-\rho} P_{00} = P_{j0} + \frac{\rho^{n+2}}{\sum\limits_{k=0}^{n} \rho^k} \ . \tag{4.22}$$

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

$$P(l_s) = P(L_s^q = l_s | I = 1) = \frac{P_{l_s\bullet}}{P_{\bullet\bullet}} = \frac{P_{l_s 0} + \frac{\rho^{n+2}}{\sum\limits_{k=0}^{n} \rho^k}}{\rho} \ .$$

The loss probability, $P(B)$, is

$$P(B) = P(x = n | I = 1) = \frac{P_{n\bullet}}{P_{\bullet\bullet}} = \frac{P_{n0} + \frac{\rho^{n+2}}{\sum\limits_{k=0}^{n} \rho^k}}{\rho}$$

and given that the length of the SQ is l_s, the expected waiting time in this queue is

$$E(W_s^q) = \frac{l_s}{\mu} \ . \tag{4.23}$$

For the VQ when the server is busy, the arrival rate is $\lambda P(B)$ and the expected queue length is

$$\begin{aligned} E(L_v^q|l_s) &= \sum_{i=1}^{\infty} i \frac{P(x = i \wedge l_s | I = 1)}{P(l_s)} = \frac{1}{P_{l_s \bullet}} \sum_{i=1}^{\infty} i P_{l_s i} \\ &= \frac{\sum_{k=0}^{n} \rho^k}{P_{l_s 0} \sum_{k=0}^{n} \rho^k + \rho^{n+2}} \sum_{i=1}^{\infty} i \frac{\rho^{n+i}}{\sum_{k=0}^{n} \rho^k} P_{00} \\ &= \frac{\rho^{n+1}}{\left(P_{l_s 0} \sum_{k=0}^{n} \rho^k + \rho^{n+2} \right) (1-\rho)^2} P_{00} \ . \end{aligned} \tag{4.24}$$

Before we derive the expected waiting time in the VQ we define the *busy period in the SQ* as the duration of time from the instant there are l customers in the SQ until the instant when the SQ has $l-1$ customers and $b(f)$ as the mean busy period starting when there are $l = n - f$ customers in the SQ (the current amount of available places in the SQ given l).
We have

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

$$b(f) = \frac{1}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda}(b(f) + b(f-1)) \ f = 1, 2, \dots \ .$$

When $f = 0$ there are no free places and we need to wait until the service ends therefore $b(0) = \frac{1}{\mu}$, when there are $f = 1, 2, \dots$ free places then the expected time until the next event is $\lambda + \mu$, if it is an arrival then we enter state where there are $f - 1$ free places and need time $b(f-1)$ to return to state f and then another $b(f)$ time until the end of the busy period. If the next event is the end of service then the busy period terminates and we have $0 \frac{\mu}{\lambda + \mu}$.
Separating $b(f)$ gives

$$b(f) = \frac{1}{\mu} + \rho b(f-1)$$

which gives

$$b(f) = \frac{1}{\mu} \sum_{k=0}^{f-1} \rho^k + b(0) \ f = 1, 2, \dots$$

therefore

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

$$b(f) = \frac{1}{\mu}\sum_{k=0}^{f}\rho^k \;\; f = 1, 2, \ldots \; .^1 \tag{4.25}$$

Finally, the expected waiting time in the VQ is

$$E[W^q|l_s] = \sum_{k=0}^{l_s} b(n-k) + E(L_v^q|l_s)b(n) \;\; , \;\; l_s = 0, \ldots, n \; . \tag{4.26}$$

The first term of the equation is the service time for the customers currently in the SQ, which is l_s consecutive busy periods with increasing number of empty places in the SQ, i.e., for the first busy period $f = n - l_s$, for the second $f = n - (l_s - 1)$, the third $f = n - (l_s - 2)$ and so on. The second term is the time it takes for all customers currently in the VQ to be served, plus the time it takes for all the customers that will arrive to the SQ during these service times - $E(L_v^q|l_s)$ busy periods starting when the SQ is empty.

For a given threshold $T = n$ followed by all customers we define a normalized cost function $\frac{\mu f(l_s)}{C_s}$ where

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

An arriving customer that sees l_s customers in the SQ will enter the SQ when $\frac{\mu f(l_s)}{C_s} \geq 0$ and to the VQ when $\frac{\mu f(l_s)}{C_s} < 0$.

4.2 Mixed strategy customers

We now investigate the stationary probabilities when the customers use mixed strategies.

For a threshold strategy $T = n + r$ the balance equations are

$$q_{(0'),(0,0)} = \lambda \tag{4.27}$$

$$q_{(0,0),(0')} = \mu \tag{4.28}$$

$$q_{(n,i),(n,i+1)} = \lambda(1-r) \quad i = 0, 1, 2, \ldots \tag{4.29}$$

$$q_{(n,i),(n+1,i)} = \lambda r \quad i = 0, 1, 2, \ldots \tag{4.30}$$

$$q_{(n+1,i),(n+1,i+1)} = \lambda \quad i = 0, 1, 2, \ldots \tag{4.31}$$

$$q_{(0,i),(0,i-1)} = \mu \quad i = 0, 1, 2, \ldots \tag{4.32}$$

$$q_{(j,i),(j-1,i)} = \mu \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 1, 2, 3, \ldots, n+1 \end{matrix} \tag{4.33}$$

$$q_{(j,i),(j+1,i)} = \lambda \quad \begin{matrix} i = 0, 1, 2, \ldots \\ j = 0, 1, 2, \ldots, n-1 \; . \end{matrix} \tag{4.34}$$

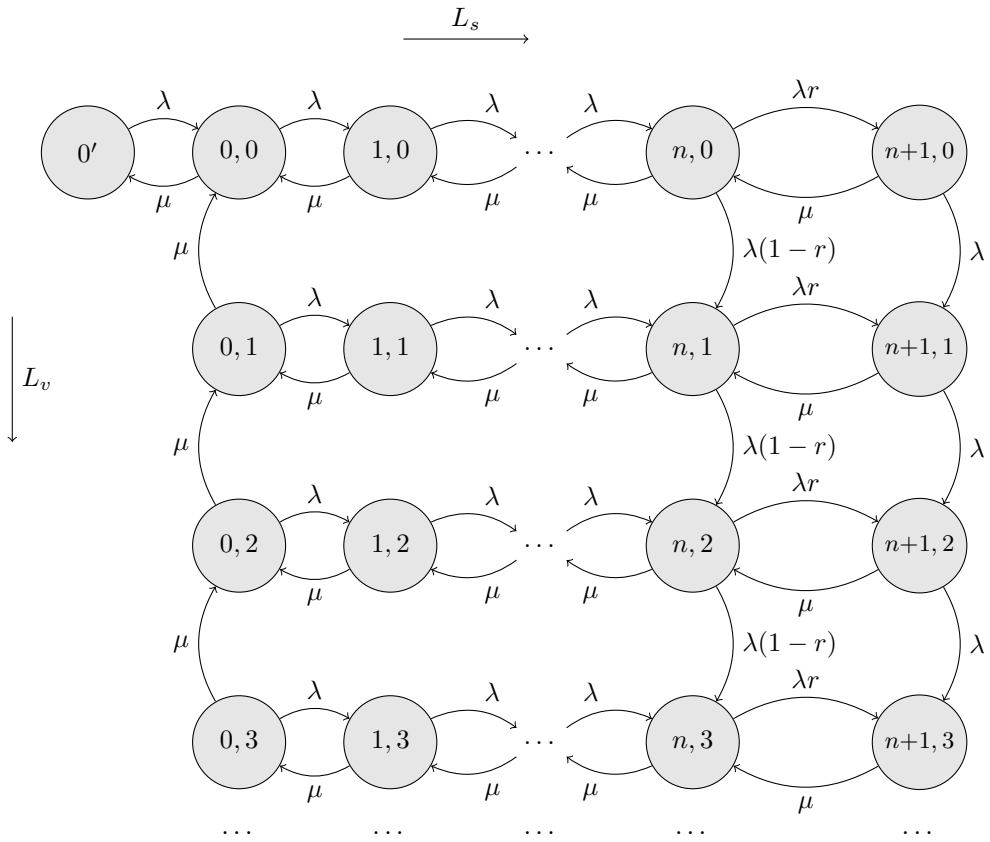

Figure 3: Transition rate diagram in the observable model when customers follow the mixed threshold strategy $T = n + r$.

The transition rate diagram is shown in Figure 3.

From the same considerations that we have in the proof of proposition 3.1 in the unobservable model

$$P_{0'} = 1 - \rho \ , \tag{4.35}$$

$$P_{\bullet\bullet} = \rho \tag{4.36}$$

and from equation (4.27)

$$P_{00} = (1-\rho)\rho \ . \tag{4.37}$$

We now find the stationary probabilities $P_{01}, P_{10}, \ldots, P_{n0}$.

Proposition 4.3.

$$P_{n0} = \frac{(1+\rho)\rho^n}{(1+\rho)\sum\limits_{k=0}^{n}\rho^k - r\sum\limits_{k=1}^{n}\rho^k} P_{00} \ , \tag{4.38}$$

$$P_{n+1,0} = \frac{\rho^{n+1} r}{(1+\rho)\sum\limits_{k=0}^{n}\rho^k - r\sum\limits_{k=1}^{n}\rho^k} P_{00} \ , \tag{4.39}$$

$$P_{01} = \frac{(1+\rho-r)\rho^{n+1}}{(1+\rho)\sum\limits_{k=0}^{n}\rho^k - r\sum\limits_{k=1}^{n}\rho^k} P_{00} \ , \tag{4.40}$$

and

$$P_{j0} = \frac{(1+\rho)\rho^j + (1+\rho-r)\left(\rho^j \sum\limits_{k=1}^{n}\rho^k - \rho^n \sum\limits_{k=1}^{j}\rho^k\right)}{(1+\rho)\sum\limits_{k=0}^{n}\rho^k - r\sum\limits_{k=1}^{n}\rho^k} P_{00} \ , \ j = 1, \ldots, n \ . \tag{4.41}$$

Proof. From the transitions rate diagram the cut around node $(n+1)0$ is

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

and therefore

$$P_{n+1,0} = \frac{\lambda r}{\lambda+\mu} P_{n0} = \frac{\rho r}{\rho+1} P_{n0} \ . \tag{4.42}$$

The horizontal cuts are

$$\lambda(1-r)P_{ni} + \lambda P_{n+1,i} = \mu P_{0,i+1} \ , \ i = 0, 1, 2, \ldots \ . \tag{4.43}$$

A cut that contains the nodes $0', 00, 10, \ldots, j0$ gives the equation

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

[1]This result corresponds with the busy period of an M/G/1/n queue. As shown in [10] page 239

and the cut that contains the nodes $0', 00, 10, \ldots, n0$ gives the equation

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

Therefore

$$\lambda P_{j0} = \mu P_{01} + \mu P_{j+1,0} \ , \ j = 0, 1, \ldots, n \ . \tag{4.44}$$

By using equation (4.43) for $i = 0$ and equation (4.42) we have

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

and after re-indexing we get

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

By recursively positioning this equation for all the partial cuts of the cut $0', 00, 10, \ldots, j0$ we can explicitly write the probability P_{j0} in terms of P_{00}

$$P_{j0} = \rho^j P_{00} - \frac{1+\rho-r}{1+\rho} P_{n0} \sum_{k=1}^{j} \rho^k \ . \tag{4.45}$$

Specifically for P_{n0} we have

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

and we obtain equation (4.38). From this and by using (4.42) we obtain equation (4.39) and with (4.42),(4.43) when $i = 0$ we obtain equation (4.40).
From equations (4.38) and (4.45) 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 (4.41). □

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

Proposition 4.4.
For $i = 1, 2, \ldots$ *and* $0 \le r < 1$

$$P_{0i} = (1-r)\rho P_{n,i-1} + \rho P_{n+1,i-1} \ , \tag{4.46}$$

$$P_{ni} = XZ^{i-1}P_{n+1,0} + YP_{n,i-1} + \frac{\rho r}{1+\rho} XY \sum_{k=0}^{i-2} Z^k P_{n,i-2-k} \ , \tag{4.47}$$

$$P_{n+1,i} = \frac{\rho r}{1+\rho} Y^i P_{n0} + Z P_{n+1,i-1} + \frac{\rho r}{1+\rho} XY \sum_{k=0}^{i-2} Y^k P_{n+1,i-2-k} \ , \qquad (4.48)$$

$$P_{ji} = \sum_{k=0}^{j} \rho^k P_{0i} - (P_{n+1,i} + (1-r)P_{ni}) \sum_{k=1}^{j} \rho^k \ , \ j = 1, 2, \ldots, n \ . \qquad (4.49)$$

Where

$$X = \frac{(1-\rho^{n+2})\rho}{(1+\rho)(1-\rho^{n+1}) - (1-\rho^n)\rho r} \ , \qquad (4.50)$$

$$Y = \frac{(1-r)(1+\rho)(1-\rho^{n+1})\rho}{(1+\rho)(1-\rho^{n+1}) - (1-\rho^n)\rho r} \ , \qquad (4.51)$$

$$Z = \left(1 + \frac{(1-\rho^{n+2})\rho r}{(1+\rho)(1-\rho^{n+1}) - (1-\rho^n)\rho r}\right) \frac{\rho}{1+\rho} \ . \qquad (4.52)$$

Remark 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

$$P_{n1} = Y P_{n0} + X P_{n+1,0} \ , \qquad (4.53)$$

$$P_{n+1,1} = Z P_{n+1,0} + \frac{\rho r}{1+\rho} Y P_{n0} \ . \qquad (4.54)$$

Proof. From the horizontal cut between the rows $i, i-1$ (equation (4.43)) we immediately have equation (4.46).
A cut that contains the nodes $0i, 1i, \ldots, ji$ gives the equation

$$\lambda P_{ji} + \mu P_{0i} = \mu P_{j+1,i} + \mu P_{0,i+1} \quad \begin{matrix} i = 1, 2, \ldots \\ j = 0, 1, 2, \ldots, n-1 \end{matrix} \ . \qquad (4.55)$$

By using equation (4.43) we have

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

or in terms of ρ

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

and by re-indexing we get

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

When we use all the equations of partial cuts of the cut $0i, 1i, \ldots, ji$ for $j < n$ we can write the probability in terms of P_{0i}, P_{ni} and $P_{n+1,i}$:

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

which is equation (4.49).

We now find an expression for P_{ni} and $P_{n+1,i}$.
Equation (4.49) for P_{ni} yields

$$\left[\sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k\right] P_{ni} = \sum_{k=0}^{n} \rho^k P_{0i} - \sum_{k=1}^{n} \rho^k P_{n+1,i} \ , \tag{4.56}$$

in terms of ρ, the cut around the node $(n+1)i$ is

$$(1+\rho)P_{n+1,i} = \rho P_{n+1,i-1} + \rho r P_{ni} \ . \tag{4.57}$$

By using equations (4.46), (4.56) and (4.57) we have

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

and therefore

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

giving

$$P_{ni} = \frac{(1+\rho)(1-r) \sum_{k=1}^{n+1} \rho^k}{\sum_{k=1}^{n+1} \rho^k + \sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k} P_{n,i-1} + \frac{\sum_{k=1}^{n+2} \rho^k}{\sum_{k=1}^{n+1} \rho^k + \sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k} P_{n+1,i-1} \ . \tag{4.58}$$

With this result and equation (4.57) we also get

$$P_{n+1,i} = \left[1 + \frac{\rho r \sum_{k=1}^{n+2} \rho^k}{\sum_{k=1}^{n+1} \rho^k + \sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k}\right] \frac{\rho}{\rho+1} P_{n+1,i-1} + \frac{(1+\rho)(1-r) r \sum_{k=2}^{n+2} \rho^k}{\sum_{k=1}^{n+1} \rho^k + \sum_{k=0}^{n} \rho^k - r \sum_{k=1}^{n} \rho^k} P_{n,i-1} \ . \tag{4.59}$$

In terms of X, Y, Z as defined in equations (4.50) - (4.52) we have

$$P_{ni} = Y P_{n,i-1} + X P_{n+1,i-1} \ , \ P_{n+1,i} = Z P_{n+1,i-1} + \frac{\rho r}{1+\rho} Y P_{n,i-1}$$

and 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 equations (4.47), (4.48), (4.53) and (4.54). □

Remark The stationary probabilities of a 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

$$F_i = P_{n,i-2} + Z F_{i-1} \ , \ F'_i = P_{n+1,i-1} + Y F'_{i-1} \ ,$$

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

$$P_{ni} = X Z^{i-1} P_{n+1,0} + Y P_{n,i-1} + \frac{\rho r}{1+\rho} XY F_i \ , \tag{4.60}$$

$$P_{n+1,i} = \rho r Y^i P_{n0} + Z P_{n+1,i-1} + \frac{\rho r}{1+\rho} XY F'_i \ . \tag{4.61}$$

We first pre-calculate Y, Z, $\frac{\rho r}{1+\rho} XY$, $\frac{\rho r}{1+\rho} P_{n0}$ and $X P_{n+1,0}$. Then we calculate $P_{n,i}$ and $P_{n+1,i}$ and from these we calculate the all the stationary probabilities in the row, $P_{j,i}$ when $j = 0, \ldots, n$. The pre-calculation is done in $O(n)$ time by using equations (4.38), (4.39), (4.50), (4.51) and (4.52). We calculate $P_{n,i}$ and $P_{n+1,i}$ by using equations (4.61), (4.60) 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 finding $P_{n,i}$ and $P_{n+1,i}$ in $O(i)$ time. We calculate the stationary probabilities of the row using equation (4.49). By saving the value of the sums when calculating $P_{k-1,i}$ we can find $P_{k,i}$ in $O(1)$ steps and calculating all the stationary probabilities in $O(n)$. Therefore the full calculation of a row i is done in $O(n+i+n) = O(n+i)$ time.

As in the pure strategies case, for the SQ when the server is busy, the probability that the length of queue is l_s is

$$P(l_s) = P(L_s^q = l_s | I = 1) = \frac{P_{l_s \bullet}}{P_{\bullet\bullet}}$$

and given that the length of the SQ is l_s, the expected waiting time

$$E(W_s^q) = \frac{l_s}{\mu} \ . \tag{4.62}$$

For the VQ, the expected length of the queue when the server is busy and the length of the SQ is l_s can be calculated from propositions 4.3, 4.4 and $P(l_s)$

$$E(L_v^q | l_s) = \sum_{i=1}^{\infty} i \frac{P(x = i \wedge l_s | I = 1)}{P(l_s)} = \frac{1}{P_{l_s \bullet}} \sum_{i=1}^{\infty} i P_{l_s i} \ . \tag{4.63}$$

We use the same definitions for the *busy period in the SQ* as in the pure strategies case. The *busy period in the SQ* is the duration of time from the instant there are l customers in the SQ until the instant when the SQ has $l-1$ customers. $b(f)$ is now defined as the busy period starting when there are $l = n+1-f$ customers in the SQ (the current number of available places in the SQ given l). We have

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

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

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

When $f = 0$ there are no free places and we need to wait until the service ends therefore $b(0) = \frac{1}{\mu}$. The expected time until the next event is $\lambda+\mu$. For $f = 1$, if the next event is an arrival then with potability r the arriving customer joins the SQ and we enter the state where there are no free places and need to wait $b(0)$ and additional $b(1)$, and with probabilities $1-r$ the arriving customer joins the VQ and we stay at the same state. When there are $f = 2,3,\dots$ free places, if the next event is an arrival then we enter state where there are $f-1$ free places and need time $b(f-1)$ to return to state f and then another $b(f)$ time until the end of the busy period. If the next event is the end of service then the busy period terminates.
Separating $b(1)$ gives

$$\mu b(1) = 1 + \lambda r b(0)$$

which gives

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

and separating $b(f)$ gives

$$b(f) = \frac{1}{\mu} + \rho b(f-1) = \frac{1}{\mu} \sum_{k=0}^{f-2} \rho^k + \rho^{f-1} b(1) \ f = 2, 3, \dots \ .$$

Therefore,

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

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

$$b(f) = \left(\sum_{k=0}^{f-1} \rho^k + \rho^f r \right) \frac{1}{\mu} \ f = 2, 3 \dots \ . \tag{4.64}$$

Finally, the expected waiting time in the VQ is

$$E[W^q | l_s] = \sum_{k=0}^{l_s} b(n+1-k) + E(L_v^q | l_s) b(n+1) \ , l_s \leq n+1 \ . \tag{4.65}$$

The first term of the equation is the service time for the customers currently in the SQ, which is l_s consecutive busy periods with increasing number of empty places in the SQ, i.e., for the first busy period $f = n + 1 - l_s$, for the second $f = n + 1 - (l_s - 1)$, the third $f = n + 1 - (l_s - 2)$ and so on. The second term is the time it takes for all customers currently in the VQ to be served, plus the time it takes for all the customers that will arrive to the SQ during these service times - $E(L_v^q | l_s)$ busy periods starting when the SQ is empty.

4.3 Numerical investigation of the mixed strategy

As stated in Proposition 3.2, the total number of customers in the system is equal to the total number of customers in an equivalent M/M/1 system with one queue. Therefore, the probability of having 40 customers is very small even with a high ρ (for $\rho = 0.9$ the probability of having 40 customers is 0.0015). We set the maximum number of customers in the system to be $M = 40$. We set the number of places in the SQ to be n according to the threshold T ($n = 1, 2, \dots, M$, $0 < r < 1$) and the number of customers in the VQ is set to be $M - n$ accordingly.

Figure 4 shows the difference $\Delta = l_s - E(L_v^q | l_s)$ between the length of the SQ and the expected length of the VQ for different thresholds when $\rho = 0.8$. We observe that the difference is monotone increasing in both l_s and T and also the graphs do not intersect. We obtain similar results when changing the value of ρ.

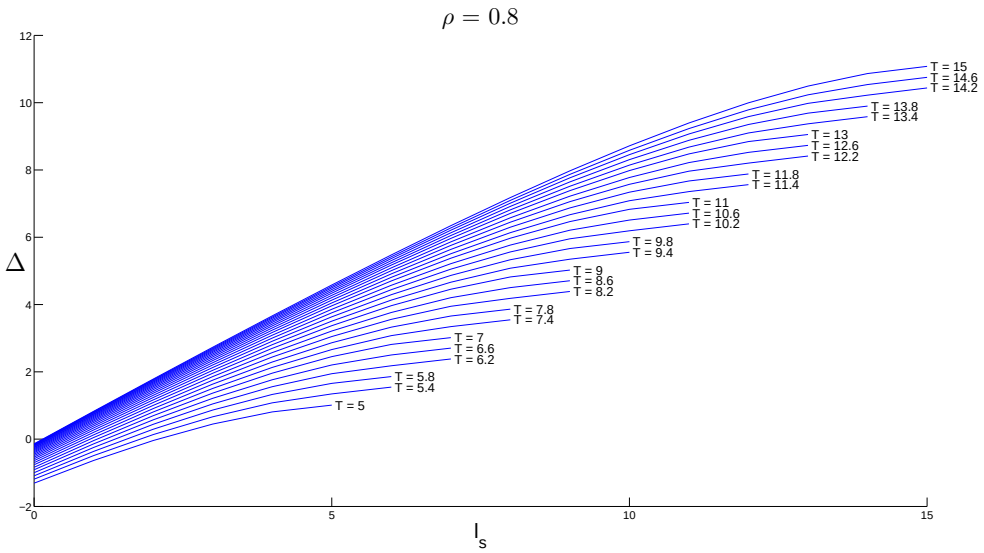

Figure 4: Queue-length difference SQ - VQ ($\Delta = l_s - E(L_v^q|l_s)$) for $\rho = 0.8$.

Let $b'(f)$ be the normalized mean busy period in the SQ in units of customer service time. Then, $b'(f) = \frac{b(f)}{\frac{1}{\mu}}$ and from equation (4.64)

$$b'(0) = 1$$

$$b'(1) = \rho r + 1 \ ,$$

$$b'(f) = \sum_{k=0}^{f-1} \rho^k + \rho^f r \ \ f = 2, 3, \ldots \tag{4.66}$$

and the normalized expected waiting time in the VQ $\hat{E}[W^q|l_s]$ is

$$\hat{E}[W^q|l_s] = \frac{E[W^q|l_s]}{\frac{1}{\mu}} = \sum_{k=0}^{l_s} b'(n+1-k) + E(L_v^q|l_s)b'(n+1) \ , l_s \leq n+1 \ . \tag{4.67}$$

Figure 5 shows the normalized mean busy period in the SQ for different values of r when $\rho = 0.8$. The graphs are valid for every threshold strategy T since $b'(f)$ is independent of n. We observe that the expected difference is increasing in l_s, the graphs non intersecting and also all the graphs have the same horizontal asymptote at the value $y = \frac{1}{1-\rho}$, which is the mean busy period for $M/M/1$ queue with unlimited buffer ($T \to \infty$). We obtain similar results when changing the value of ρ.

Figures 6-8 show the normalized expected waiting time in the VQ, $\hat{E}[W^q|l_s]$, for $\rho = 0.8$ and different values of T. Figure 6 shows the first sum of the equation

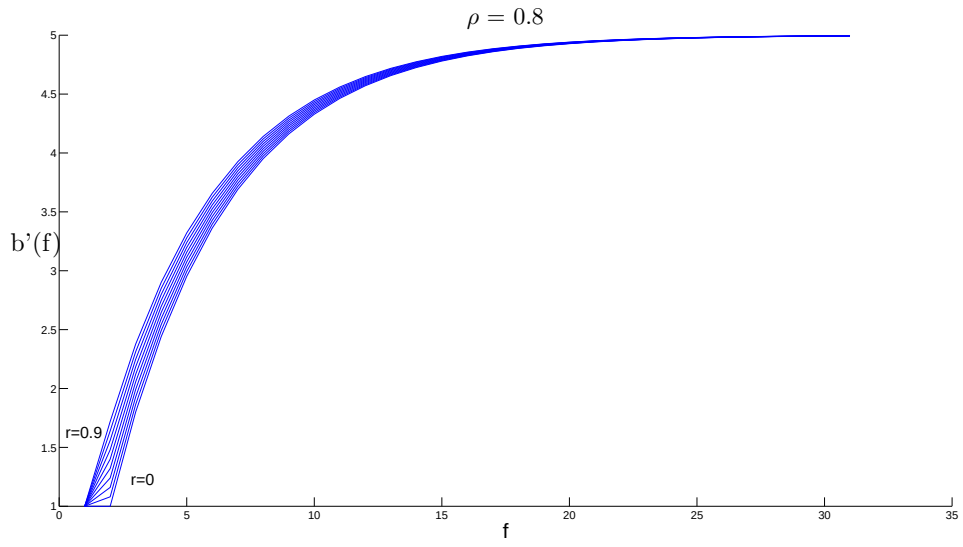

Figure 5: Normalized mean busy period in the SQ in time units per customer, $b'(f)$, for $r = 0, 0.1, \ldots, 0.9$ and $\rho = 0.8$.

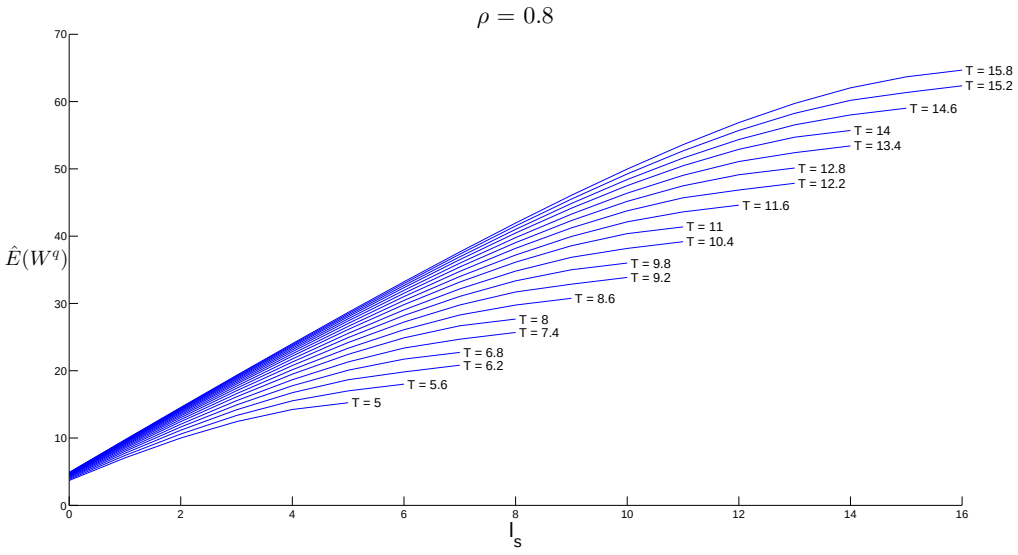

Figure 6: First sum of the normalized expected waiting time in the VQ, $\sum_{k=0}^{l_s} b'(n+1-k)$ when $\rho = 0.8$.

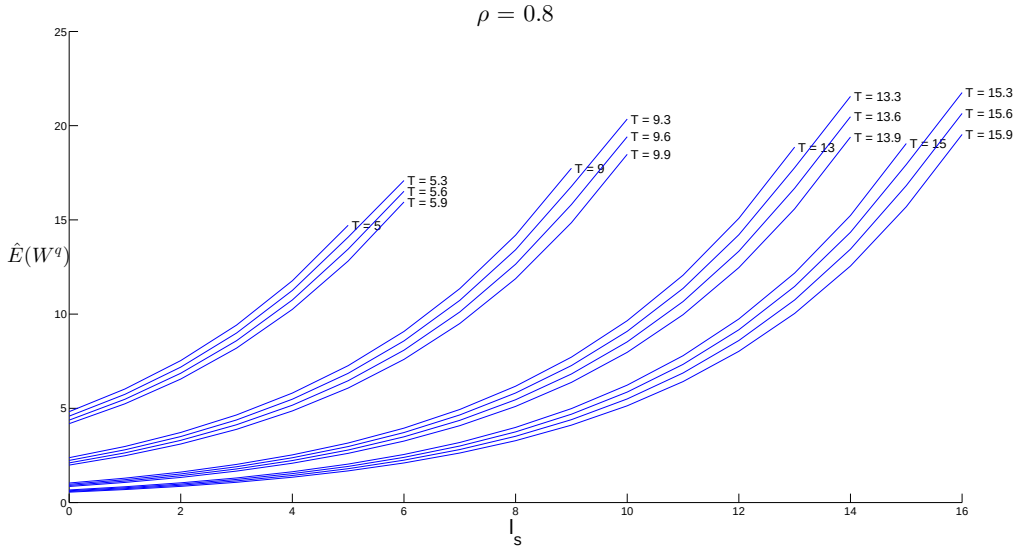

Figure 7: Second sum of the normalized expected waiting time in the VQ, $E(L_v^q|l_s)b'(n+1)$ when $\rho = 0.8$.

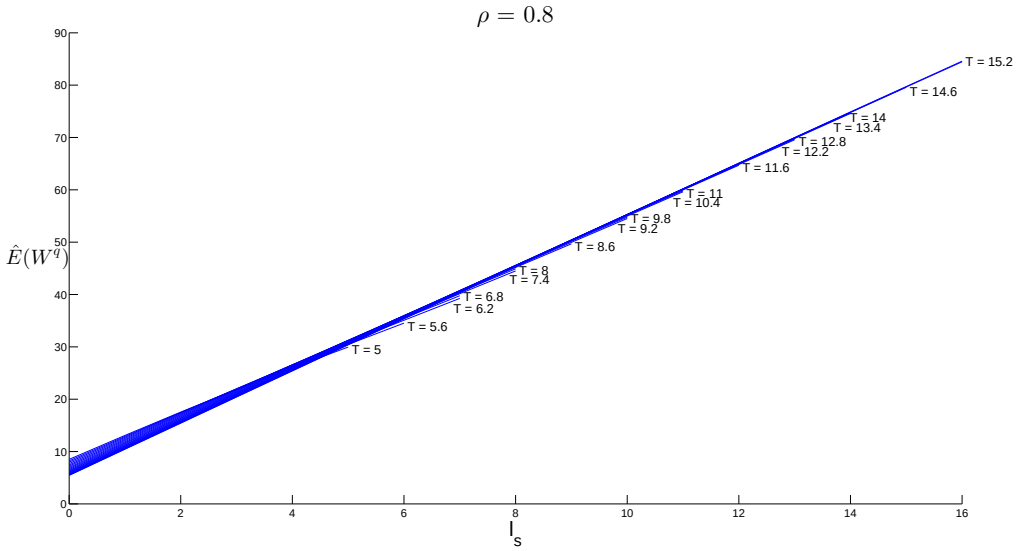

Figure 8: Normalized expected waiting time in the VQ ($\hat{E}(w^q|l_s)$) when $\rho = 0.8$.

$(\sum_{k=0}^{l_s} b'(n+1-k))$ which is the expected time it will take the l_s existing customers in the SQ to leave the system, Figure 7 shows the second sum $(E(L_v^q|l_s)b'(n+1))$ which is the expected time it will take the customers in the VQ to leave the system and Figure 8 shows the total expected waiting time. We note that in Figure 6 the graphs are continuous in T, monotone increasing in l_s and in T and concave, which is expected as a sum of concave functions. In Figure 7 the graphs are monotone increasing, convex and continuous - where the jumps in the graphs between two thresholds with the same n (for example $5, 5.5$) are since the domain is discontinuity changing. In Figure 8 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.

For a given threshold strategy of joining the SQ when the server is busy, $T = n + r$, followed by all customers we define a normalized cost-difference function $\frac{\mu f(l_s)}{C_s}$ where

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

An arriving customer that sees l_s customers in the SQ will be indifferent in his choice to enter the queues when $\frac{\mu f(l_s)}{C_s} = 0$ and will choose to enter the SQ when $\frac{\mu f(l_s)}{C_s} > 0$. Define $\varphi = \frac{C_v}{C_s} < 1$ and we have,

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

Figures 9-11 show the expected costs in each of the queues for different values of ϕ and ρ. For given values of ϕ and ρ the expected costs in the VQ are almost linear and very close to each other for close values of T, in resemblance to the VQ expected waiting times shown in Figure 8. Therefore we have marked the area in the graph were all the VQ costs reside by two blue (solid) lines and the red (broken) line is the expected cost in the SQ. When the VQ cost area is above the SQ cost line, the cost in the SQ is always smaller than the cost in the VQ and therefore all arriving customers will enter the SQ regardless of the its size. They will always enter the VQ in the opposite case. We note that, as expected, when φ, l_s or ρ increase so does the cost in the VQ, and therefore increasing the customers motivation to enter the SQ. When the SQ line passes the VQ costs area or resides inside it (as shown in the right graph in Figure 9) there are values of T for which $f(l_s) = 0$ and therefore are suspected as equilibrium strategies, since the customer will be indifferent in choosing between the queues.

Figures 12-14 show the expected costs in the VQ for different T values. The continuous part of each line shows the cost when l_s is at the threshold value $(l_s = n)$.

For selected φ and ρ we calculated the best response threshold of a customer. The best response is the first l_s in which a customer will choose to enter the VQ, meaning $\min\left[l_s|\frac{\mu f(l_s)}{C_s} < 0\right]$.

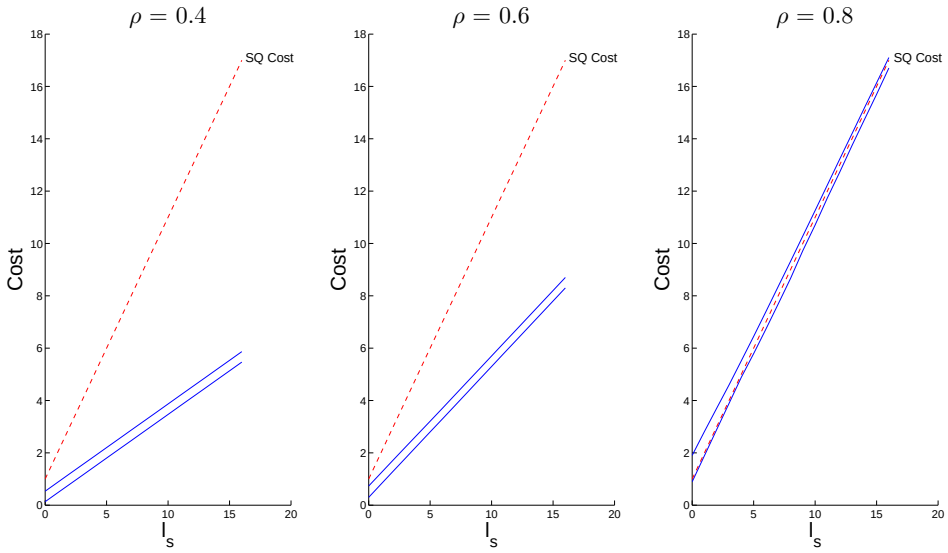

Figure 9: Cost in the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.2$

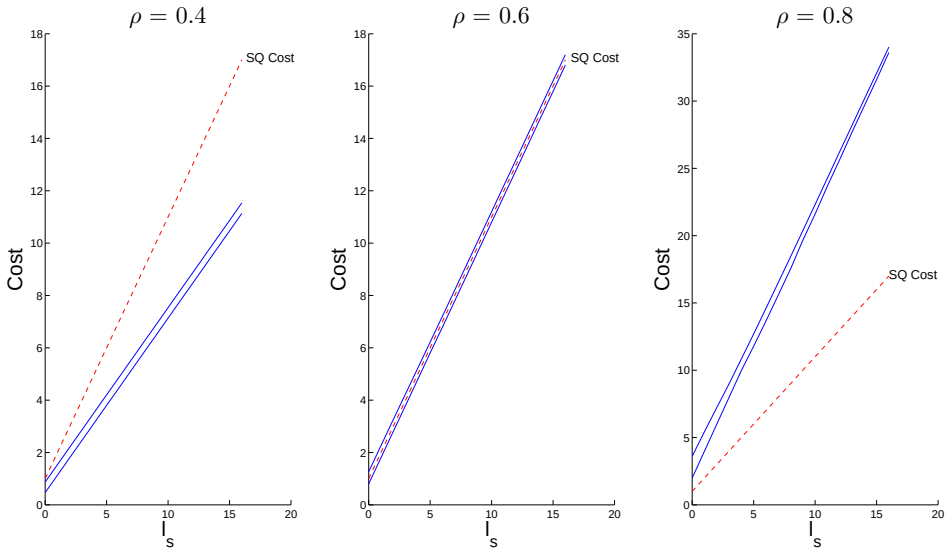

Figure 10: Cost in the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.4$

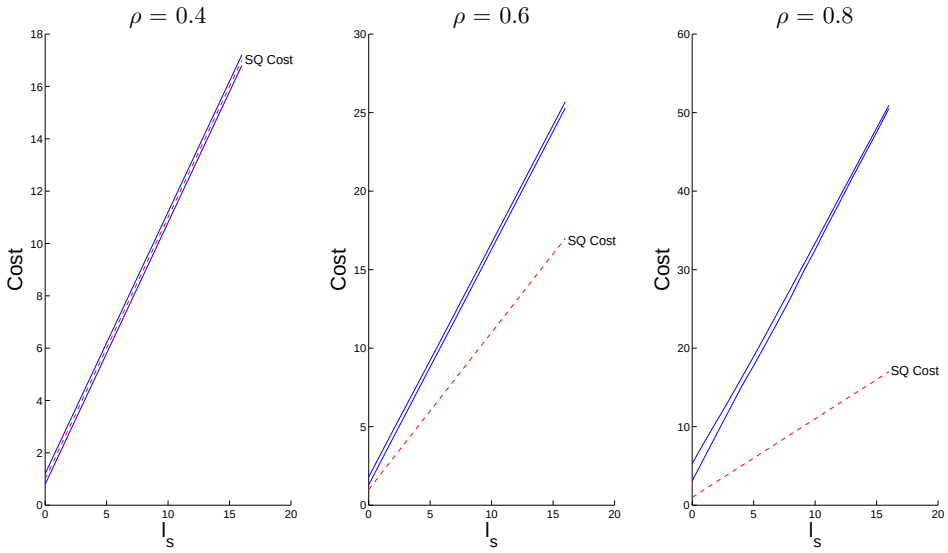

Figure 11: Cost in the SQ and cost in the VQ for $5 \leq T \leq 15$ and $\varphi = 0.6$

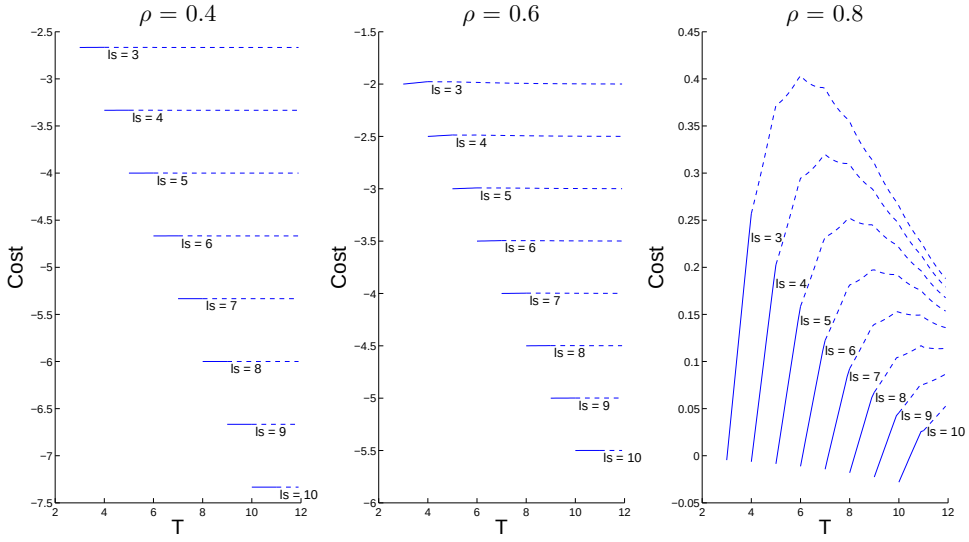

Figure 12: Cost in the VQ for different l_s when $\varphi = 0.2$

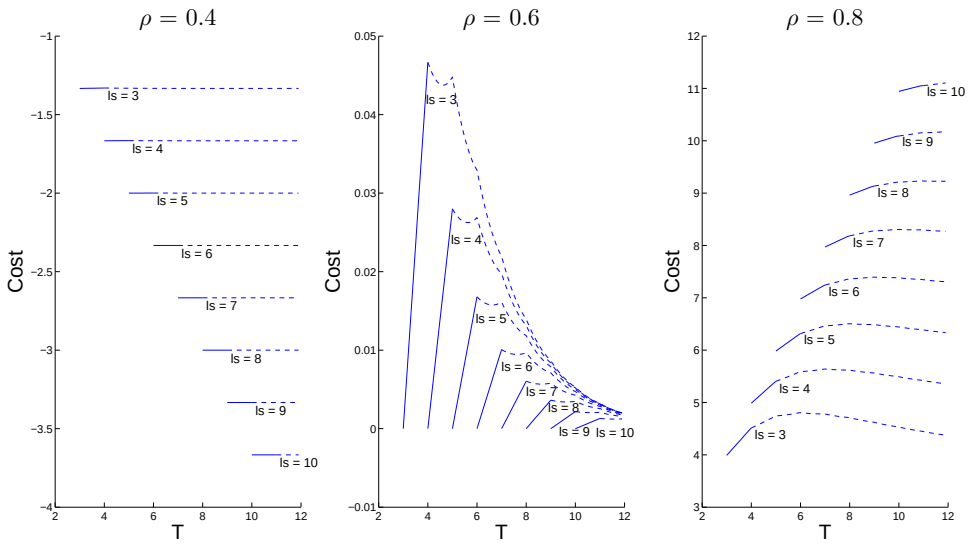

Figure 13: Cost in the VQ for different l_s when $\varphi = 0.4$

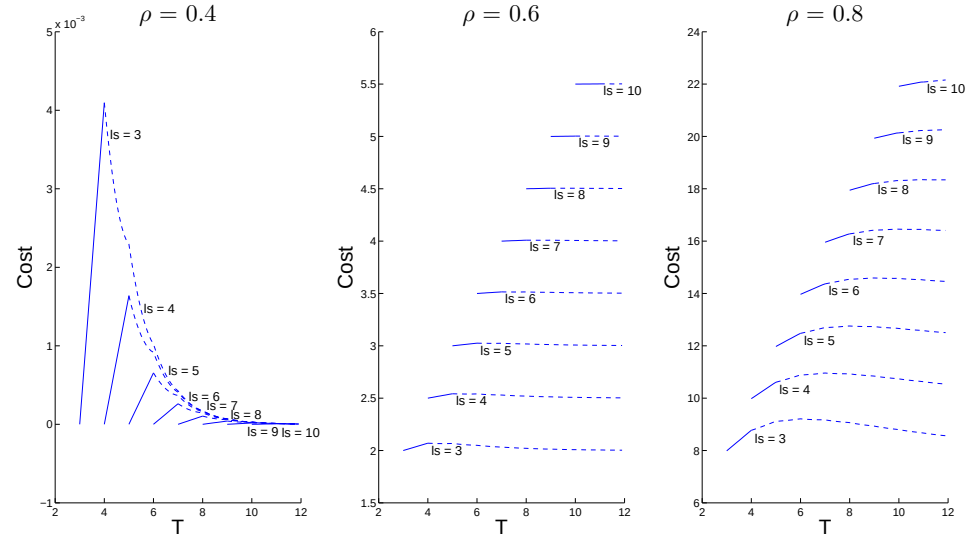

Figure 14: Cost in the VQ for different l_s when $\varphi = 0.6$

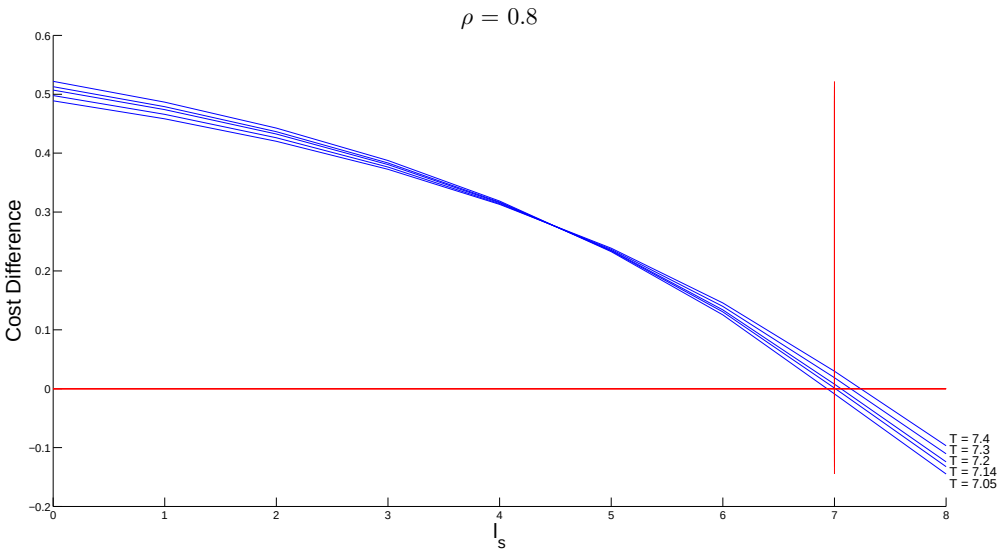

Figure 15: Expected cost difference SQ−VQ for different thresholds when $\varphi = 0.2$, $\rho = 0.8$ and $n = 7$.

Remark We can search for best response and equilibrium best response strategies at specific n by looking at the expected cost difference graph. We plot the expected cost difference (SQ-VQ) lines for all the T values that we would like to check, horizontal line at $y = 0$ and the vertical line $x = n$ as shown in **Figure 15**. The best response strategies have positive values at the second quadrant, negative values at the fourth quadrant and are passing from the second to the fourth quadrant though the first quadrant (meaning passing the vertical line with positive values), the equilibrium best response strategies are the best response strategies that passes through the origin. In Figure 15, $T = 7.05$ is not best response strategy, $T = 7.2, 7.3, 7.4$ are best response strategies and $T \approx 7.14$ is an equilibrium best response strategy.

Figures 16-18 show the best response for $(\varphi, \rho) = (2, 0.8), (4, 0.6), (6, 0.4)$ the intersection of the best response function and the linear line are the equilibrium points, when an equilibrium point has $r = 0$ then it is a pure strategy and mixed when $0 < r < 1$.

In similarity to the unobservable case, the optimal social net benefit is received when all customers enter the VQ in any case, meaning when $T = 0$ and then

$$\lambda \frac{C_v}{\mu - \lambda} = \frac{C_v}{1 - \rho} \rho \ . \tag{4.68}$$

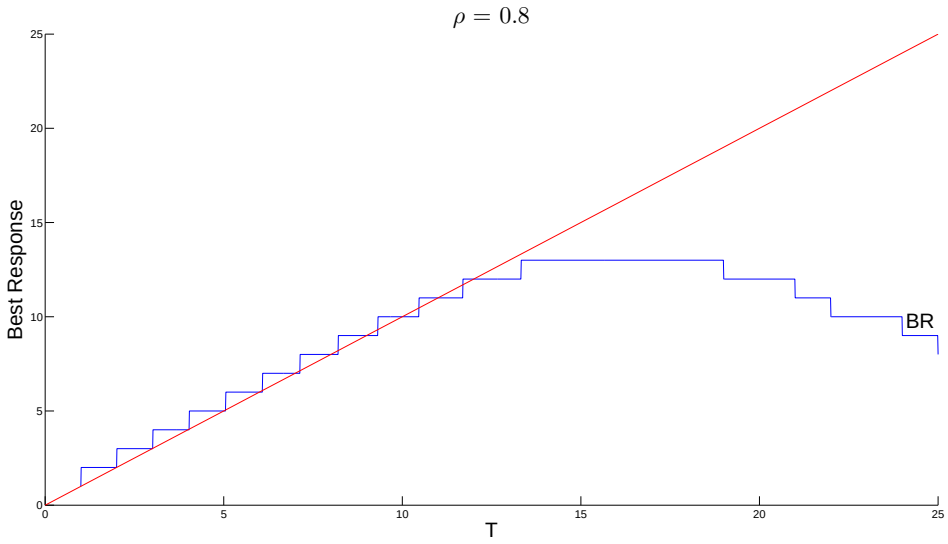

Figure 16: Best response for $\varphi = 0.2, \rho = 0.8$

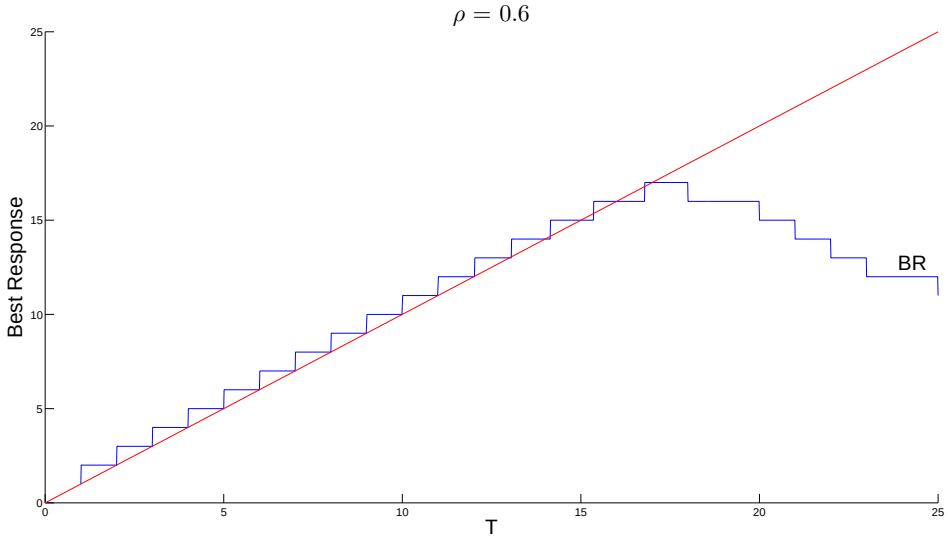

Figure 17: Best response for $\varphi = 0.4, \rho = 0.6$

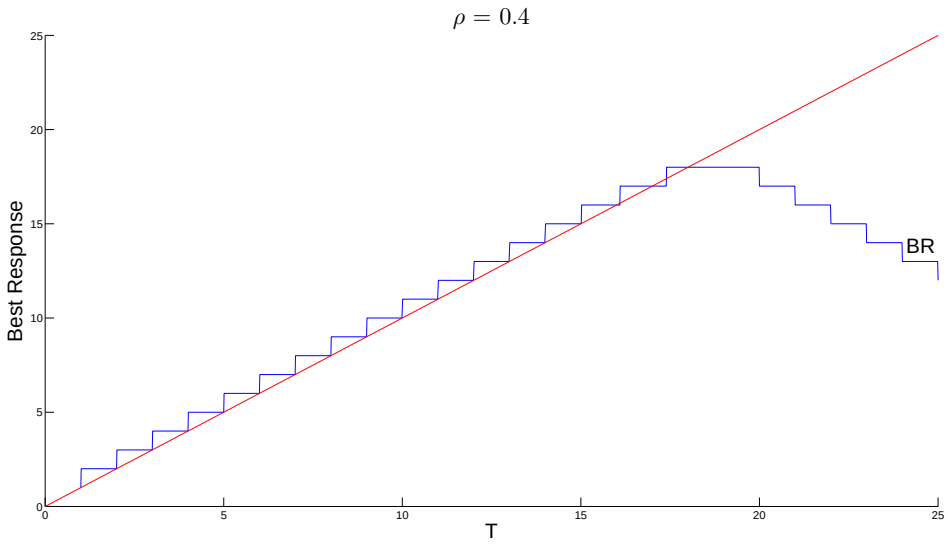

Figure 18: Best response for $\varphi = 0.6, \rho = 0.4$

5 Concluding remarks

In this thesis we investigate a service system modeled as a non-preemptive single-server M/M/1 virtual queuing system with two first-come first-served queues, a System Queue (SQ) and a virtual queue (VQ). An arriving customer who finds the server idle enters service immediately, and when the server is busy the customer chooses whether to enter the VQ or the SQ. Each queue has a different waiting cost for a single time unit, C_s for the SQ and C_v for the VQ ($C_v < C_s$). We study two information models of the system, an unobservable where customers are notified only where the server is busy or not, and an observable where customers are also informed about the number of customers (l_s) currently waiting in the SQ.

For the unobservable case we define the expected number of customers and expected waiting time in each of the queues. A result which is compatible with the regular M/M/1 priority queue system with two queues and preemptive regime. We have also found the equilibrium strategy of an arriving customer to join the SQ which is to join iff

$$\frac{C_v}{C_s} + \rho \geq 1 \ .$$

The social optimal strategy is to always join the VQ

For the observable case we define the stationary probabilities, the mean busy period in the SQ and the expected number of customers and expected waiting time in the VQ for the pure and mixed strategy customers. The expected

waiting time in the VQ is shown in the numerical investigation part to be almost linear in l_s, as a sum of convex and concave functions. We also conclude from the numerical investigation that multiple best response equilibrium strategies exist, all in the monotone increasing part of the best response function. This is compatible with follow-the-crowd (FTC) strategies.

The social optimal strategy for both information models is

$$\lambda \frac{C_v}{\mu - \lambda} = \frac{C_v}{1 - \rho} \rho \ ,$$

and from the perspective we would recommend to encourage all customers to enter the VQ.

In this model we do not cover systems that have more than one system queue, virtual queue or both. Systems that provide their customers the option to renege or to balk, and systems in which the time it takes the server to contact a customer from the virtual queue is not negligible, as by Economou and Kanta [13].

Interesting variation of our model would be to allow impatient (or patient) customers to move from one queue to the other, and to negotiate for their place with the other customers who are currently waiting. Another variation could be to provide the customers in the VQ a quality-of-service guarantee for systems with high traffic intensity (high ρ).

6 Notations and definitions

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.
$r_s,\ r_v$	Entering probability to the SQ and VQ.
$S(r_s)$	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.
$E(L^q),\ E(W_s^q),\ E(W_v^q)$	Expected waiting time in the system, the SQ and VQ of an arriving customer.
$\hat{E}[W^q\|l_s]$	The expected waiting time in the VQ in time units per customer.
$E(L_s^q),\ E(L_v^q)$	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.
I	Indicator function for the server state (1-busy, 0-idle).
S	State space of the continues time Markov chain of the system
P_{ji}	Stationary distribution, where $(j,i) \in S$.
$P_{0'}$	Stationary distribution when the server is idle.
r_s^{soc}	The social optimal strategy of joining the SQ.
l_s	Current number of customers in the SQ.
$s(l_s)$	Threshold strategy of joining the SQ.
T	Threshold of a threshold strategy, $T = n + r$ $(r \in [0,1], n \in \mathbb{N})$.
$P(B)$	The loss probability of the SQ.
f	Number of free places in the SQ
$b(f)$	The mean busy period starting when there are l customers in the SQ and $f = n - l$ (pure strategies) or $f = n + 1 - l$ (mixed strategies)
$b'(f)$	Mean busy period in the SQ in time units per customer.
M	Maximum number of places in the system in the numerical section.

References

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

[2] Adan, I. and Resing, J. (2015) *Queueing Theory*, Eindhoven University of Technology.

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

[4] Altman, E., Jiménez, T., núñez-Queija, R. and Yechiali, U. (2004) *Optimal routing among */M/1 queues with partial information*, Stochastic Models 20, 2, 149-171.

[5] Armony, M. and Maglaras, C. (2004) *On customer contact centers with a call-back option: customer decisions, routing rules, and system design*, Operations Research, 52(2), 271-292.

[6] Armony, M. and Maglaras, C. (2004) *Contact centers with a call-back option and real-time delay information*, Operations Research, 52(4), 527-545.

[7] Burgain, P., Feron, E., and Clarke, J.-P (2009) *Collaborative virtual queue: benefit analysis of a collaborative decision making concept applied to congested airport departure operations*, Air Traffic Control Quaterly 17(2), 195-222.

[8] Camulli, E. (2007) *Answer my call: technology helps utilities get customers off hold*, Electric Light and Power 2(6), 56.

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

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

[11] de Lange, R., Samoilovich, I. and van der Rhee, B. (2013) *Virtual queuing at airport security lanes*, European Journal of Operational Research 225(1), 153-165.

[12] Dickson, D., Ford, R. C. and Laval, B. (2005) *Managing real and virtual waits in hospitality and service organizations*, Cornell Hotel and Restaurant Administration Quarterly, 46, 52- 63.

[13] Economou, A. and Kanta, S. (2011) *Equilibrium customer strategies and social-profit maximization in the single-server constant retail queue*, Navel Research Logistics 58(2), 107-122.

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

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

[16] Hassin, R. (1996) *On the advantage of being the first server*, Management Science, 42 (4) 618-623.

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

[18] Hassin, R. and Haviv, M. (2006) *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems*, Kluwer Academic Publishers

[19] Iravani, F. and Balcioǵlu, B. (2008) *On priority queues with impatient customers*, Queueing Systems, 58(4):239-260.

[20] Kostami, V. and Ward, R.A. (2009) *Managing service systems with an ofine waiting option and customer abandonment*, Manufacturing & Service Operations Management 11(4) 644-656.

[21] Lovejoy, T.C., Aravkin, S. and Schneider-Mizell, C. (2004) *KalmanQueue: an adaptive approach to virtual queuing*, The UMAP Journal 25(3), 337-352.

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

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

[24] S. C. Littlechild (1974) *Optimal arrival rate in a simple queueing system*, International journal of production research 12, 391-397.

[25] Taylor. S, (1994) *Waiting for service: the relationship between delays and evaluations of service*, Journal of Marketing 58(2) 56-69.

[26] R.F. Cope III, 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(10), 13-20.

[27] Wüchner, P., Sztrik, J., and de Meer, H. (2009) *Finite-source M/M/S retrial queue with search for balking and impatient customers from the orbit*, Computer Networks 53(8), 1264-1273.