Queueing Systems
https://doi.org/10.1007/s11134-020-09665-9

Strategic behaviour in a tandem queue with alternating server

Nimrod Dvir[1] · **Refael Hassin**[1] · **Uri Yechiali**[1]

Received: 30 October 2019 / Revised: 27 April 2020

Abstract

This paper considers an unobservable two-site tandem queueing system attended by an alternating server. We study the strategic customer behaviour under two threshold-based operating policies, applied by a profit-maximizing server, while customers' strategic behaviour and server's switching costs are taken into account. Under the *Exact-N* policy, in each cycle the server first completes service of N customers in the first stage (Q_1), then switches to the second stage (Q_2) and then serves those N customers before switching back to Q_1 to start a new cycle. This policy leads to a mixture of *Follow-the-Crowd* and *Avoid-the-Crowd* customer behaviour. In contrast, under the *N-Limited* policy, the server switches from Q_1 to Q_2 also when the first queue is emptied, making this regime work-conserving and leading only to *Avoid-the-Crowd* behaviour. Performance measures are obtained using matrix geometric methods for both policies and any threshold N, while for sequential service ($N = 1$) explicit expressions are derived. It is shown that the system's stability condition is independent of N, and of the switching policy. Optimal performances in equilibrium, under each of these switching policies, are analysed and compared through a numerical study.

Keywords Strategic behaviour · Tandem queues · Alternating server · Threshold policy · Switching cost · Equilibrium · Follow-the-Crowd · Avoid-the-Crowd

Mathematics Subject Classification 60K25 · 60K30 · 90B22 · 91A10

✉ Nimrod Dvir
dvirnimrod@gmail.com

Refael Hassin
hassin@tauex.tau.ac.il

Uri Yechiali
uriy@tauex.tau.ac.il

[1] Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

Published online: 09 September 2020

1 Introduction

This study analyses a queueing system where customers are served in two phases by the same server. There is a separate queue for each phase, and the server alternates between the two queues. The server incurs a switching cost for every change in the queue being served.[1] Arriving customers are strategic and act to maximize their utility. Their decision to join or not is based on the system's known parameters, while the system's state is unobservable. Customers are homogeneous; they incur a waiting cost which is linear in their sojourn time in the system and gain a fixed value upon service completion in the second phase. Both queues are *first-come first-served* (FCFS), whereas the server determines the operating policy (that is, when to serve in each queue).

We consider two common threshold-based operating policies (regimes): (i) *Exact-N* and (ii) *N-Limited*. The first is a strict policy in which the server switches the queue operated only after the number of customers served reaches a fixed threshold. The second is a more adaptable policy, where the server switches when reaching the threshold or when the first queue is emptied, whatever occurs first.

A simple example for such a model is a food stand where a single operator serves each arriving customer at two tandem stations. First is receiving an order from the customer, and second is processing the order. Obviously, accumulating several orders, and then serving these customers, can be more efficient. This attribute is expressed as a switching cost in our model. Another example is a safety-concerned double system of gates, as operated in a safari or in high-security establishments, where at most one gate can be open concurrently.

Observing these examples or similar ones, intuition may consider the *N-Limited* policy as a superior regime, due to its work-conserving quality, where the server never idles when the system is not empty, in contrast to the *Exact-N* policy. However, the latter may be justified when a significant switching cost is incurred. For example, in an industry where a product is processed at two tandem stations operated by the same machinery but under different setups.

Our model belongs to the strategic queueing literature which has been studied extensively since Naor's pioneering work [22], where server and customer strategies were first considered in an observable classical M/M/1 system. Yechiali [27] studied the observable GI/M/1 queue and showed that, among all randomized customers' joining policies, the non-randomized threshold policy of Naor is indeed optimal. Edelson and Hilderbrand [11] examine the unobservable case of Naor's model, and further on numerous extensions of this idea have been published. Hassin and Haviv [15] and Hassin [14] provide surveys of this field. Queueing systems with an alternating server (also known as 'polling systems') have been studied extensively in the literature (for example, Boxma et al. [9], Takagi [25] and Yechiali [28]). For a survey of this subject, see Boon et al. [6].

Arachenkov et al. [5] and Perel and Yechiali [24] present a two-queue system when an alternating server uses a threshold-based switching policy. Jolles et al. [17] extend the model to include switchover times. Similarly, we define our model as a

[1] Often a switch is accompanied by a switching time. We simplify the model by assuming that it can practically be substituted by an appropriate switching cost.

three-dimensional Markovian process, examine both non-work-conserving and work-conserving policies and use matrix geometric techniques and probability generating functions as the main mathematical methods to derive the multi-dimensional probability distribution function of the system's states in stationarity, from which the system's performance measures are obtained.

The topic of tandem queues is well studied. Usually, there is a server at each stage. Nair [21] and Taube-Netto [26] introduced the idea of two queues in tandem where a single alternating server operates both queues. There are several subsequent works (for example, Katayama [18] and Iravani et al. [16]), where mostly the optimization problem considered is minimizing the server's expenses.

Additional closely related works are Bountali and Economou [7] and Bountali and Economou [8], where strategic behaviour in a two-stage service system with batch processing is studied. Their methods and results have some resemblance to ours.

As perceivable from this literature review, many works have considered the subjects of alternating servers, tandem queues or strategic behaviour in queueing systems. Furthermore, there are examples for equilibrium strategies in tandem queues (for example, D'Auria and Kanta [10] and Allon and Bassamboo [2]) and in polling systems (for example, Altman and Shimkin [3], Atar and Saha [4] and Adan et al. [1]). However, in spite of the extensive study in these fields we are not aware of a paper considering all three subjects combined. This is where our work is positioned.

In this work, we study the equilibrium behaviour under steady-state conditions of the system in the strategic game among the agents (the customers and the server). The server determines the operating policy, threshold and price, in order to maximize profit (net income). Our goal is to compute, for given price and policy, the equilibrium effective arrival rate, and then, using this information, to compute the maximal profit and the corresponding price and threshold level, and compare the outcomes for the two policies (*Exact-N* and *N-Limited*).

The structure of the paper is as follows: In Sect. 2, the model is described and defined as a three-dimensional continuous-time Markov chain (CTMC) and formulated as a two-dimensional quasi birth-and-death (QBD) process, and the strategic aspect is explained. In Sect. 3, a matrix geometric approach is employed to derive the system's steady-state probabilities by which the expected sojourn time for each of the above-mentioned policies is obtained. In Sect. 4, customers' strategic behaviour is analysed and possible equilibria are detailed. In Sect. 5, the special case of sequential service (when the threshold is $N = 1$) is explored and analytical results are derived. Section 6 presents extensive numerical study and its inferences are discussed. Finally, main conclusions along with suggestions for further research are provided in Sect. 7.

2 The model

2.1 Model description

We study a two-site tandem queueing system, where a single server alternates between the two queues. Each site Q_i is a *FCFS* queue with an exponential service time with rate μ_i $(i = 1, 2)$. Customers are served one by one in each site. A customer first arrives

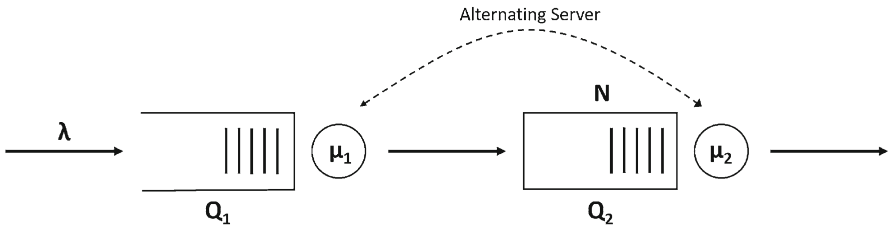


Fig. 1 A flow diagram of the system

at queue Q_1, which has an unlimited buffer, and is requested to pay p, a service fee (price) determined and collected by the server. Upon service completion at Q_1, the customer immediately proceeds to Q_2. There, the customers await for the server to switch over from Q_1 and are then served in the order of arrival. A customer who completes service at Q_2 obtains a fixed reward V (the service value) and leaves the system. When Q_2 is emptied, the server switches back to Q_1 and so forth. Customers incur a cost of C_W per unit time they spend in the system (waiting or being served). The server incurs a switching cost of C_S for every double switch (from Q_1 to Q_2 and back) between the queues.

Customers are homogeneous and the potential Poisson arrival rate Λ is greater than the server can handle. Both queues are unobservable, and arriving customers decide to either join the system or balk, based on the server's policy and price. We denote the joining rate or the effective arrival rate, by λ. See Fig. 1 for an illustration of the system.

We study two operating policies (regimes):

1. *Exact-N*: The server attends Q_1 until exactly N customers are served, then switches to Q_2 and serves continuously the N customers accumulated there. Upon service completion at Q_2, the server switches back to Q_1, resides there until exactly N customers are served, switches again to Q_2, and so forth. Note that this is a **non-work-conserving** regime.
2. *N-Limited*: The server switches to Q_2 after serving continuously up to a maximum number of N customers at Q_1 or when Q_1 is first depleted (in any case, at least one customer is served). After switching to Q_2, the server serves the customers accumulated there and then switches back. This is a **work-conserving** regime.

2.2 Setting as a QBD process

Denote by $L_i(t)$ the number of customers in Q_i at time t, and let $I(t) = i$ if at time t the server attends Q_i $(i = 1, 2)$. The triple $(L_1(t), L_2(t), I(t))$ defines a irreducible continuous-time quasi birth-and-death (QBD) process. Let $L_i = \lim_{t\to\infty} L_i(t)$ $(i = 1, 2)$ and $I = \lim_{t\to\infty} I(t)$. A transition-rate diagram for the *Exact-N* policy is depicted in Fig. 2 and for the *N-Limited* policy in Fig. 3. The numbers within each node indicate (L_1, L_2), while the queue the server is operating at, i.e. I, is marked by the shape and colour of the node, a blue circle for $I = 1$ and red rectangle for $I = 2$.

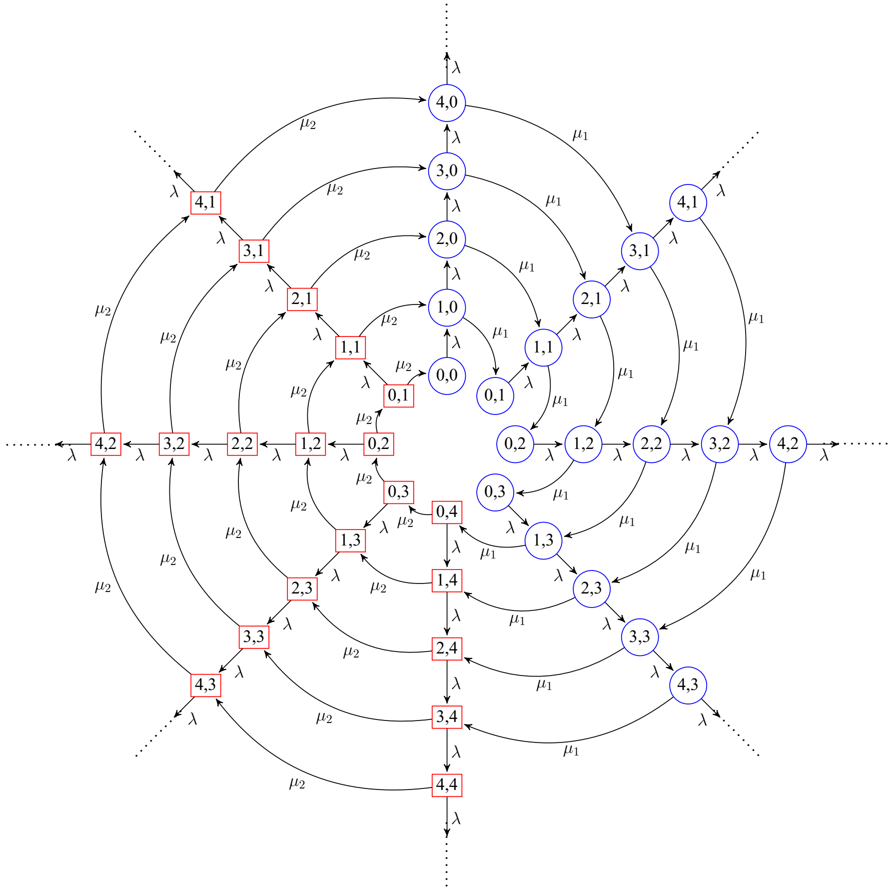


Fig. 2 Transition rate diagram for the *Exact-N* policy, with $N = 4$

Further on we use the notation $P_{nj}^{(i)}$ for the steady-state probability that the system is in the state $(L_1 = n, L_2 = j, I = i)$, $n = 0, 1, 2, \ldots$; $j = 0, 1, \ldots, N$; $i = 1, 2$.

2.3 Strategic view

In the game among the customers and the server, players maximize their own benefit. We denote a customer's expected utility by $U(N, p)$ and the server's expected profit by $r(N, p)$, when the decision variables determined by the server are the operating policy, which is a tuple of the threshold and the queue principle, $(N \in \mathbb{N}) \times \{Exact - N, N - Limited\}$ and the price $p \in \mathbb{R}_+$. Customers make one decision—whether or not to join the system. Customers join if their expected utility is positive, balk if it is negative and are indifferent if it is zero. The joining rate is denoted by $\lambda(N, p)$. A customer's expected utility is a function of the price and mean sojourn time in the system, $W = W(N, \lambda(N, p))$. To simplify notation, we omit the decision variables (N, p) when no ambiguity arises.

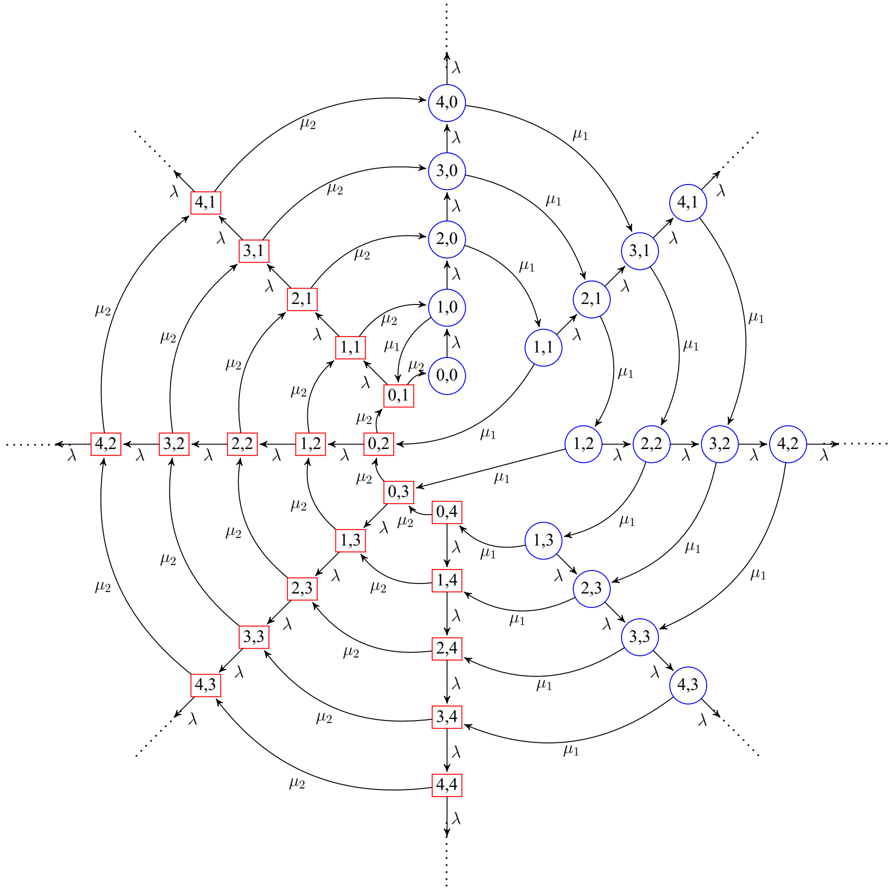


Fig. 3 Transition rate diagram for the *N-Limited* policy, with $N = 4$

A customer's expected utility from joining the system is

$$U = V - p - C_{\mathrm{W}} W \ . \tag{1}$$

Notice that

$$V - p - C_{\mathrm{W}} \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} \right) > 0 \tag{2}$$

has to hold for customers' utility to be positive (otherwise, under any positive joining rate, the customers' revenue from the service minus the price is not worth the cost of their own service time).

Since customers are homogeneous, there exists a symmetric equilibrium where all customers expect equal utility. Because of our assumption that the potential arrival rate is higher than the server can handle, necessarily some customers balk. Thus, the common expected utility in equilibrium is $U = 0$.

In many simple queueing systems, a growth in customer expected utility U motivates an increase in joining rate, resulting in an increase in expected sojourn time and therefore in a decrease in expected utility. This is an outcome of negative externalities caused by a customer when joining. Similarly, a reduction in customer expected utility causes a decrease in joining rate, followed by a decrease in expected sojourn time and therefore an increase in expected utility. Such a progression around equilibrium makes it stable.

However, as will be elaborated in Sect. 4, this behaviour does not happen in every point where $U = 0$. In the *Exact-N* Scenario, it is possible that a more congested system will lead to a **decrease** in waiting time and then to an increase in the expected utility, making this equilibrium unstable. This is a consequence of the coexistence of positive and negative externalities inflected by customer behaviour under this policy. We show that in the case of multiple equilibria with positive arrival rate only one of them is stable and denote the joining rate in this equilibrium by λ_e. If there is no such equilibrium, $\lambda_e = 0$. We analyse the system in a stable equilibrium with positive arrival rate.

The profit of the server is the revenue minus the expenses, which are calculated differently for each of the scenarios. In the *Exact-N* scenario, the server incurs a switching cost, C_S, for every N arrivals. Therefore, the server's expected profit per unit time is

$$r = \lambda p - C_S \frac{\lambda}{N} = \lambda \left(p - \frac{C_S}{N} \right) . \tag{3}$$

In the *N-Limited* scenario the switching expenses (per customer) are not exclusively dependent on N. It is possible to calculate the average number of switches executed by the server per unit time by looking at either one of the two directions of switching. The first option is by multiplying the proportion of time the server spends in the states leading from $I = 1$ to $I = 2$ by the transition rate μ_1, the second is by multiplying the proportion of time the server spends in the states leading from $I = 2$ to $I = 1$ by the transition rate μ_2. Then (see Fig. 6)

$$r = \lambda p - C_S \mu_1 \left(\sum_{j<N} P_{1j}^{(1)} + \sum_{n>1} P_{n,N-1}^{(1)} \right) = \lambda p - C_S \mu_2 \sum_{n} P_{n1}^{(2)} . \tag{4}$$

We consider a monopolistic profit-maximizing server. As in Edelson and Hildebrand [11], the profit-maximization policy leads to social optimization, where the server gains all the welfare.

Given the parameters μ_1, μ_2, C_S, C_W, and V, our goal is to find the maximal profit r^* and the corresponding optimal values for the decision variables N^*, p^* and the operating policy. This requires a few steps for each of the policies:

1. Calculation of $W(N, \lambda(N, p))$. This step is achieved by using the *matrix geometric* method to obtain the system's steady-state probabilities.
2. Finding the equilibrium effective arrival rate $\lambda_e(N, p)$ for any pair (N, p) of policy and price.

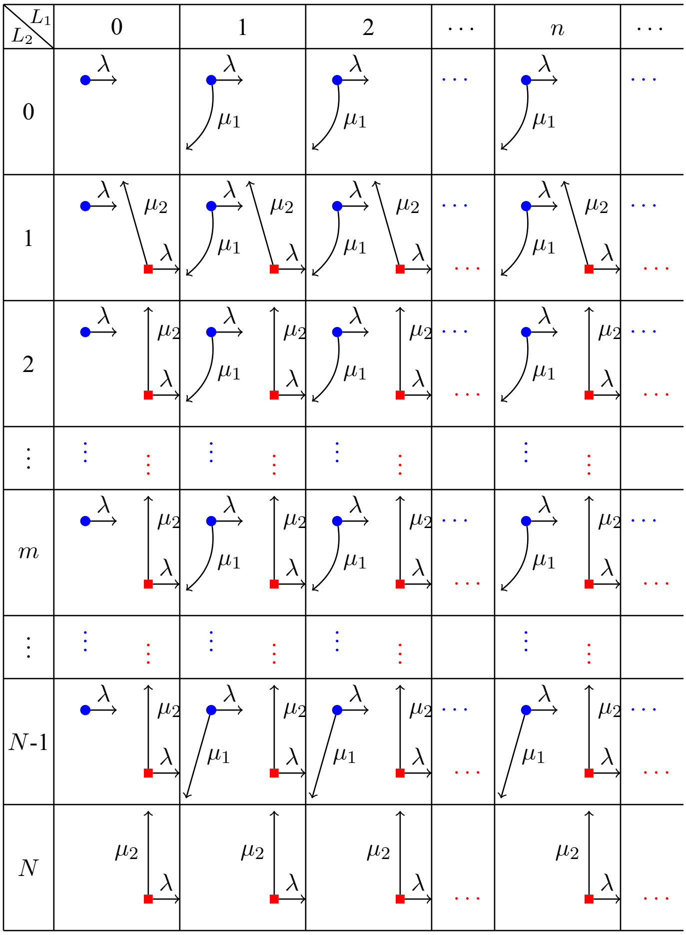


Fig. 4 Transition rate diagram for the *Exact-N* policy

3. Finding the maximal profit r^* and the matching optimal pair (N^*, p^*).

For the sequential service case (when $N = 1$), we obtain a closed-form solution of the expected sojourn time while using the *probability generating functions* (PGF) method, by which we manage to reach a close-form solution for the optimal price and corresponding effective arrival rate and profit.

3 Performance measures

3.1 Exact-N scenario

The triple (L_1, L_2, I) defines a QBD process at stationarity, where L_1 denotes the *'level'* and the pair (L_2, I) indicates the *'phase'* of the process. In Fig. 4, we provide an alternative (traditional) representation of the transition rate diagram of the process.

The infinite-state space S is ordered as follows: We start with column $L_1 = 0$ and go down the boxes from $L_2 = 0$ to $L_2 = N$, where, in each box, we specify first the state (if any) associated with $I = 1$ (marked by a blue round dot at the upper-left corner), and then the state (if any) associated with $I = 2$ (marked by a red square dot at the lower-right corner). We proceed similarly with columns $L_1 = 1, 2, \ldots, n, \ldots$. Thus, the state space is

$$\begin{aligned} S = \{&(0,0,1), (0,1,1), (0,1,2), (0,2,1), (0,2,2), \ldots, (0,N-1,1), (0,N-1,2), (0,N,2);\\ &(1,0,1), (1,1,1), (1,1,2), (1,2,1), (1,2,2), \ldots, (1,N-1,1), (1,N-1,2), (1,N,2); \ldots\\ &(n,0,1), (n,1,1), (n,1,2), (n,2,1), (n,2,2), \ldots, (n,N-1,1), (n,N-1,2), (n,N,2); \ldots\}. \end{aligned}$$

The generator matrix Q is given by

$$Q = \begin{pmatrix} B_0 & A_0 & \mathbf{0} & \cdots & \cdots & \cdots \\ A_2 & A_1 & A_0 & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where $\mathbf{0}$ is a matrix of zeros and B_0, A_0, A_1, A_2 are the following matrices, all of size $(2N) \times (2N)$, with $\alpha_1 = \lambda + \mu_1$ and $\alpha_2 = \lambda + \mu_2$:

$$B_0 = \begin{pmatrix} -\lambda & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & -\lambda & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ \mu_2 & 0 & -\alpha_2 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mu_2 & 0 & -\alpha_2 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & -\lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \mu_2 & 0 & -\alpha_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & \mu_2 & -\alpha_2 \end{pmatrix},$$

$A_0 = \lambda I$,

$$A_1 = \begin{pmatrix} -\alpha_1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & -\alpha_1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ \mu_2 & 0 & -\alpha_2 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -\alpha_1 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mu_2 & 0 & -\alpha_2 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & -\alpha_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \mu_2 & 0 & -\alpha_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & \mu_2 & -\alpha_2 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0 & \mu_1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_1 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_1 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & \mu_1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \end{pmatrix}.$$

Let $A = A_0 + A_1 + A_2$, Then

$$A = \begin{pmatrix} -\mu_1 & \mu_1 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 \\ 0 & -\mu_1 & 0 & \mu_1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 \\ \mu_2 & 0 & -\mu_2 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mu_1 & 0 & \mu_1 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_2 & 0 & -\mu_2 & 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & -\mu_1 & 0 & \mu_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \mu_2 & 0 & -\mu_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & \mu_2 & -\mu_2 \end{pmatrix},$$

where the states are the phases of the process (L_2, I). The underlying process defined by A is a cyclic-state process, as depicted in Fig. 5.

Let $\vec{\pi} = \left(\pi_0^{(1)}, \pi_1^{(1)}, \pi_1^{(2)}, \ldots, \pi_{N-1}^{(1)}, \pi_{N-1}^{(2)}, \pi_N^{(2)}\right) \in [0,1]^{2N}$ be the stationary probability vector of the matrix A, i.e. it satisfies

$$\begin{cases} \vec{\pi} A = \vec{0} \\ \vec{\pi} \cdot \vec{e} = 1 \end{cases}.$$

Then, $\vec{\pi} = (\theta_2, \theta_2, \theta_1, \theta_2, \theta_1, \ldots, \theta_2, \theta_1, \theta_1)$, i.e., the first element is θ_2, the last element is θ_1, and in between there are $N-1$ pairs of (θ_2, θ_1), where $\theta_1 = \frac{\mu_1}{N(\mu_1+\mu_2)}$ and $\theta_2 = \frac{\mu_2}{N(\mu_1+\mu_2)}$. Following Neuts [23], the stability condition $\vec{\pi} A_0 \vec{e} < \vec{\pi} A_2 \vec{e}$ becomes

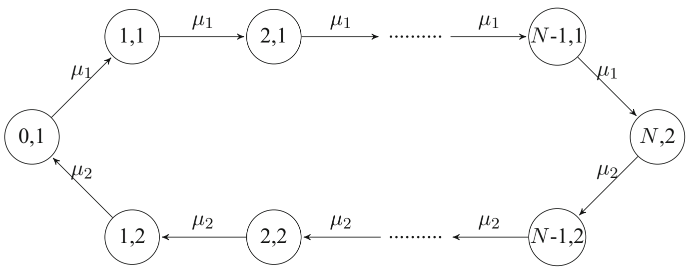


Fig. 5 The underlying process defined by A

$$\lambda < \frac{\mu_1 \mu_2}{\mu_1 + \mu_2} ,$$

which is equivalent to

$$\frac{1}{\lambda} > \frac{1}{\mu_1} + \frac{1}{\mu_2} . \tag{5}$$

Notice that this condition is **independent of** N and requires that the mean inter-arrival time should be greater than the mean total service time given to each individual customer.

Denote the proportion of time the server is busy by $\rho = \lambda(\frac{1}{\mu_1} + \frac{1}{\mu_2})$. The number of states where the server is idle is N and the sum of their stationary probabilities is $P^{(1)}_{0\bullet} = \sum_{j=0}^{N-1} P^{(1)}_{0j} = 1 - \rho$. When N increases, the probability that at least one new customer arrives while the server serves the N customers at Q_2 increases accordingly. Thus, when $N \to \infty$ the probability that no new customers will join tends to 0, so $P^{(1)}_{00} \to 0$, and the probability that the server is idle while there are customers in the system is $P^{(1)}_{0\bullet} - P^{(1)}_{00} \to 1 - \rho$.

Next, we calculate the stationary probability of each state. Define the steady-state probability vector $\vec{\mathcal{P}} = \left(\vec{P}_0, \vec{P}_1, \ldots, \vec{P}_n, \ldots\right)$, satisfying

$$\vec{\mathcal{P}} Q = \vec{0}, \tag{6}$$

$$\vec{\mathcal{P}} \cdot \vec{e} = 1, \tag{7}$$

where $\vec{0}$ is a vector of zeros, $\vec{e}$ is a vector of ones and the $2N$-dimensional probability vectors are

$$\vec{P}_n = \left(P^{(1)}_{n0}, P^{(1)}_{n1}, P^{(2)}_{n1}, P^{(1)}_{n2}, P^{(2)}_{n2}, \ldots, P^{(1)}_{n,N-1}, P^{(2)}_{n,N-1}, P^{(2)}_{nN}\right), n \geq 0 . \tag{8}$$

Now we rewrite the balance equations (6) as a set of matrix equations:

$$\vec{P}_0 B_0 + \vec{P}_1 A_2 = \vec{0}, \tag{9}$$

$$\vec{P}_{n-1} A_0 + \vec{P}_n A_1 + \vec{P}_{n+1} A_2 = \vec{0} , \quad n \geq 1 . \tag{10}$$

As in Neuts [23] we recursively express $\vec{P}_n$ in terms of $\vec{P}_0$ and a matrix R:

$$\vec{P}_n = \vec{P}_0 R^n , \quad \forall n \geq 0 , \tag{11}$$

where R is the minimal non-negative solution of the matrix quadratic equation

$$A_0 + R A_1 + R^2 A_2 = \mathbf{0} ,$$

or,

$$R = -(R^2 A_2 + A_0) A_1^{-1} . \tag{12}$$

The matrix R is calculated via a successive-substitutions algorithm (see, for example, Harchol-Balter [13], Section 21.4.3, page 370). We note that there are occasions where R can be determined explicitly (Latouche and Ramaswami [20] for special cases, and Hanukov and Yechiali [12] for more general cases). The next step is finding the vectors $\left(\vec{P}_0, \vec{P}_1, \ldots, \vec{P}_n, \ldots\right)$. The cornerstone is reaching $\vec{P}_0$ and onward, by using (11), every $\vec{P}_n$ can be calculated. There are two equations involving $\vec{P}_0$: The first is

$$\vec{P}_0(B_0 + RA_2) = \vec{0}\,, \tag{13}$$

which is obtained by substituting (11) into the first matrix balance equation (9), and the second is the normalizing equation (7), that can be rewritten as

$$\sum_{n=0}^{\infty} \vec{P}_n \vec{e} = 1\,. \tag{14}$$

Substituting (11) in (14) results in

$$\vec{P}_0 \left(\sum_{n=0}^{\infty} R^n \right) \vec{e} = \vec{P}_0 (I - R)^{-1} \vec{e} = 1\,. \tag{15}$$

We find $\vec{P}_0$ by solving the set of equations (13) with (15), and from there, by using (11) we can calculate every $\vec{P}_n$. For further details turn to Appendix 1. The mean queue sizes are given by

$$E[L_1] = \sum_{n=0}^{\infty} n \vec{P}_n \vec{e} = \sum_{n=1}^{\infty} n \vec{P}_0 R^n \vec{e} = \vec{P}_0 R (I - R)^{-2} \vec{e}\,, \tag{16}$$

$$E[L_2] = \sum_{n=0}^{\infty} \vec{P}_n \vec{Z}\,, \quad \vec{Z} = (0, 1, 1, 2, 2, \ldots, N-1, N-1, N). \tag{17}$$

Placing the sum of those two equations in *Little's Law*, we can obtain $W(N, \lambda)$, and by using Eq. (1) calculate U.

3.2 N-Limited scenario

For the *N-Limited* scenario, we use the same triple (L_1, L_2, I) to define the QBD process, but the state space changes. The sole difference is the removal of the boundary states $(0, m, 1)$, $m > 0$, as is seen in Fig. 6. The resulting state space is

$$\begin{aligned} S = \{&(0,0,1), (0,1,2), (0,2,2), \ldots, (0,N-1,2), (0,N,2); \\ &(1,0,1), (1,1,1), (1,1,2), (1,2,1), (1,2,2), \ldots, (1,N-1,1), (1,N-1,2), (1,N,2); \ldots \\ &(n,0,1), (n,1,1), (n,1,2), (n,2,1), (n,2,2), \ldots, (n,N-1,1), (n,N-1,2), (n,N,2); \ldots\}\,. \end{aligned}$$

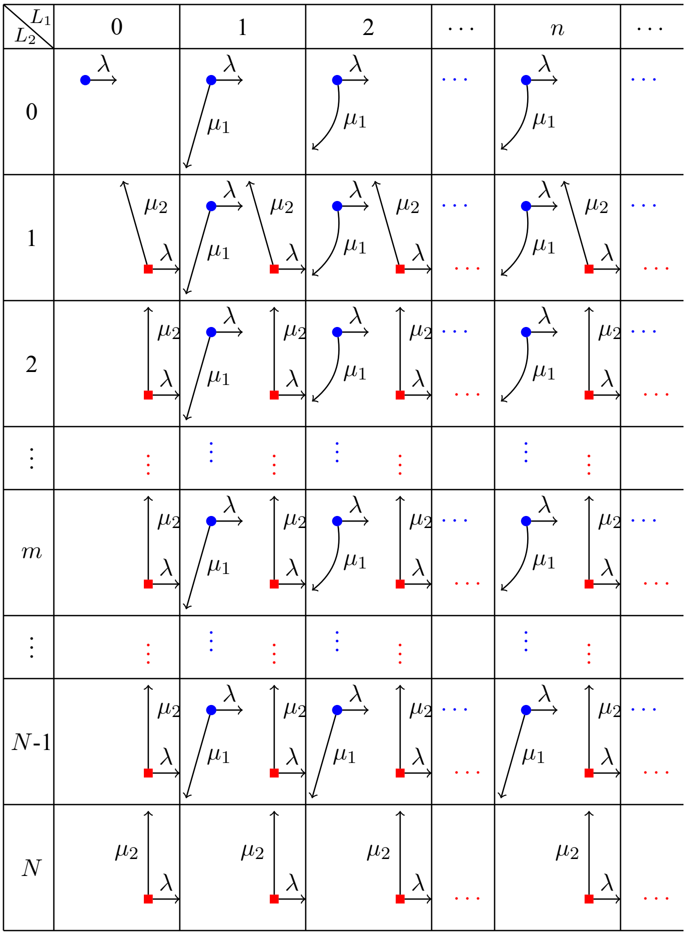


Fig. 6 Transition rate diagram for the *N-Limited* policy

The corresponding generator matrix Q is

$$Q = \begin{pmatrix} B_0 & C_1 & \mathbf{0} & \cdots & \cdots & \cdots & \cdots \\ B_1 & A_1 & A_0 & \mathbf{0} & \cdots & \cdots & \cdots \\ \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & A_2 & A_1 & A_0 & \mathbf{0} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where the matrices $\mathbf{0}$, A_0, A_1, A_2 are as in the *Exact-N* scenario. B_0 is different and now of size $(N+1) \times (N+1)$. There are two additional matrices: B_1 of size $(2N) \times (N+1)$ and C_1 of size $(N+1) \times (2N)$:

$$B_0 = \begin{pmatrix} -\lambda & 0 & 0 & \cdots\cdots\cdots & 0 & 0 \\ \mu_2 & -\alpha_2 & 0 & \cdots\cdots\cdots & 0 & 0 \\ 0 & \mu_2 & -\alpha_2 & \cdots\cdots\cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots\ \ddots\ \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots\cdots\ \mu_2 & -\alpha_2 & 0 \\ 0 & 0 & 0 & \cdots\cdots\cdots & \mu_2 & -\alpha_2 \end{pmatrix}, B_1 = \begin{pmatrix} 0 & \mu_1 & 0 & \cdots\cdots\cdots & 0 & 0 \\ 0 & 0 & \mu_1 & \cdots\cdots\cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 \\ 0 & 0 & 0 & \mu_1\ \cdots\cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots\ \ddots\ \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots\cdots\cdots & 0 & \mu_1 \\ 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} \lambda & 0 & 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & \cdots\cdots\cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & \cdots\cdots\cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots\ \ddots\ \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots\cdots\cdots & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots\cdots\cdots & 0 & 0 & \lambda \end{pmatrix}.$$

Since A_0,A_1,A_2 are identical to the *Exact-N* case, it follows that the underlying process defined by $A = A_0 + A_1 + A_2$ is the same (Fig. 5) and so is $\vec{\pi}$, the stationary probability vector of A. Therefore, we conclude that the stability condition in the *N-Limited* Scenario is the same as in the *Exact-N* [Eq. (5)]. This is in spite of the fact that *N-Limited* is a work-conserving regime and *Exact-N* is not. Note that, in contrast to the *Exact-N* scenario, in the *N-Limited* case the server can be idle only when there are no customers in the system. Hence, $P_{00}^{(1)} = 1 - \rho$ for every N.

The steady-state probability vectors $\vec{P}_n$ for $n \geq 1$ are the same as in Eq. (8), but $\vec{P}_0$ changes to

$$\vec{P}_0 = (P_{00}^{(1)}, P_{01}^{(2)}, P_{02}^{(2)}, \ldots, P_{0,N-1}^{(2)}, P_{0N}^{(2)}) . \tag{18}$$

The matrix equation (9) is replaced by two new equations (19) and (20):

$$\vec{P}_0 B_0 + \vec{P}_1 B_1 = \vec{0}, \tag{19}$$

$$\vec{P}_0 C_1 + \vec{P}_1 A_1 + \vec{P}_2 A_2 = \vec{0} , \tag{20}$$

while the matrix equation (10) holds for $n \geq 2$, namely,

$$\vec{P}_{n-1} A_0 + \vec{P}_n A_1 + \vec{P}_{n+1} A_2 = \vec{0} , \quad n \geq 2 . \tag{21}$$

In this scenario $\vec{P}_n = \vec{P}_{n-1} R$ holds for $n \geq 2$ and, instead of (11), we get

$$\vec{P}_n = \vec{P}_1 R^{n-1} , \quad \forall n \geq 1 , \tag{22}$$

while R is calculated by successive substitutions as in Sect. 3.1 [Eq. (12)]. The vectors $\left(\vec{P}_0, \vec{P}_1, \ldots, \vec{P}_n, \ldots\right)$ are calculated by the same method used in the *Exact-N* scenario.

Placing (22), first in the second balance matrix equation (20), we get

$$\vec{P}_0 C_1 + \vec{P}_1 (A_1 + RA_2) = \vec{0}\,, \tag{23}$$

and second, in the normalization equation (14), we get

$$\vec{P}_0 \vec{e} + \sum_{n=1}^{\infty} \vec{P}_1 R^{n-1} \vec{e} = \vec{P}_0 \vec{e} + \vec{P}_1 (I - R)^{-1} \vec{e} = 1\,. \tag{24}$$

(Notice that the first $\vec{e}$ is of size $N + 1$ and the second is of size $2N$.) The procedure to calculate $\vec{P}_0$ is specified in Appendix 1. By using (22) one can calculate any $\vec{P}_n$.

The calculation of the expected queue sizes is similar to the *Exact-N* case:

$$E[L_1] = \sum_{n=0}^{\infty} n \vec{P}_n \vec{e} = \sum_{n=1}^{\infty} n \vec{P}_1 R^{n-1} \vec{e} = \vec{P}_1 (I - R)^{-2} \vec{e}\,, \tag{25}$$

$$E[L_2] = \vec{P}_0 \vec{Z}_0 + \sum_{n=1}^{\infty} n \vec{P}_n \vec{Z}\,, \quad \begin{array}{l} \vec{Z}_0 = (0, 1, 2, \ldots, N), \\ \vec{Z} = (0, 1, 1, 2, 2, \ldots, N-1, N-1, N). \end{array} \tag{26}$$

Finally, $W(N, \lambda)$ and U are obtained in the same way as in the *Exact-N* scenario.

4 Utility analysis

In order to determine the effective arrival rate in equilibrium, we analyse the utility U as a function of the effective arrival rate λ. From Eq. (1), $U(N, \lambda)$ is a linear function of the expected sojourn time $W(N, \lambda)$, and therefore, analysing the latter would apply immediate conclusions on the former. Due to intricate, direct and indirect, dependence of W on N, there is no closed formula of W for every N. However, as specified in Sect. 3, one can numerically calculate W for any given N and λ. Thus, by numeric methods, we are able to provide evidence of the convexity of $W(\lambda)$ for every N. In Sect. 5, we present an analytical proof of the convexity of $W(\lambda)$ for the sequential service ($N = 1$).

The mean waiting times in Q_1 and Q_2 as a function of λ are depicted in Fig. 7 for the *Exact-N* scenario, and in Fig. 8 for the *N-Limited* case.

Figure 7 demonstrates that in the *Exact-N* Scenario the expected sojourn time in the first queue, W_1, is a convex increasing function of the effective arrival rate, and in the second queue the expected sojourn time, W_2, is a convex decreasing function of the effective arrival rate (where, for the $N = 1$ case, W_2 is constant). Figure 8 demonstrates that in the *N-Limited* Scenario the expected sojourn time in both queues is a convex increasing function of the effective arrival rate (and again, when $N = 1$, W_2 is constant). An extensive numerical study verifies that these properties are kept for larger values of N. We conclude that in both scenarios the expected total sojourn time in the system is a convex function of the effective arrival rate. Hence, from (1), the expected utility is a concave function of the effective arrival rate. In the *Exact-N*

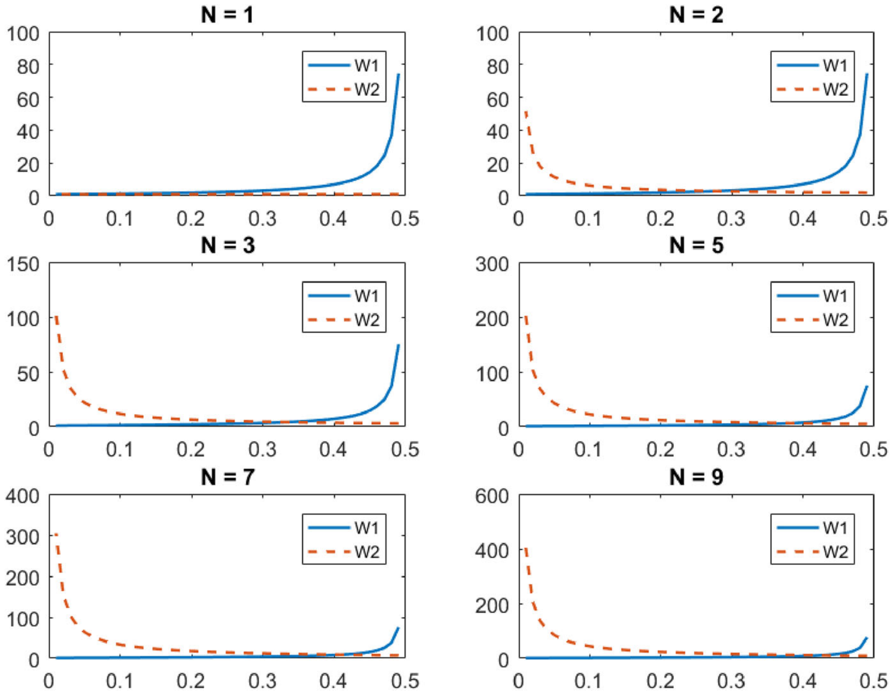


Fig. 7 *Exact-N* Mean sojourn time W_i in each queue as a function of the effective arrival rate λ for different values of N, where $\mu_1 = 1$ and $\mu_2 = 1$ ($\rho = 2\lambda$)

scenario for $N \geq 2$, $U(\lambda)$ is unimodal with a maximum, whereas in the *N-Limited* scenario, or under the sequential service (when $N = 1$), it is monotone decreasing and concave. As explained in Sect. 2.3, customers' expected utility in equilibrium is zero. We identify a few options for each of the two possible patterns (unimodal or monotone decreasing) of $U(\lambda)$:

1. The *Exact-N* (Unimodal) Case:
    (a) *One Equilibrium* When the maximum expected utility is **negative**, the only equilibrium is when no customers join the system, i.e. $\lambda_e = 0$ (interpreted as a scenario where the server decides not to operate the system). This equilibrium is **stable**, in the sense that if a positive fraction of the customers change their strategy and join, the individual's expected utility is still negative and therefore such a change will not affect the strategy of the rest of the customers. Notice that this equilibrium always exits.
    (b) *Two Equilibria* When the maximum expected utility is **exactly zero**, we get an additional equilibrium. This equilibrium is stable in the positive direction (an increase in the joining rate leads to a utility diminution which leads back to a decrease in the joining rate) and unstable in the negative direction (a decrease in the joining rate leads to a utility diminution which leads to a growing decrease in the joining rate). Hence, the only **stable** equilibrium is $\lambda_e = 0$.

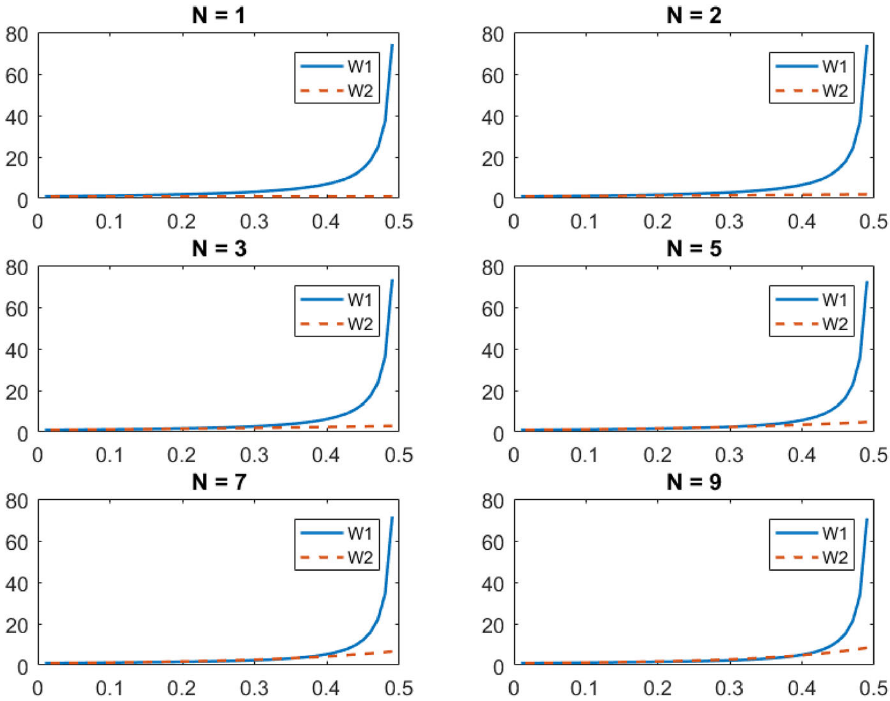
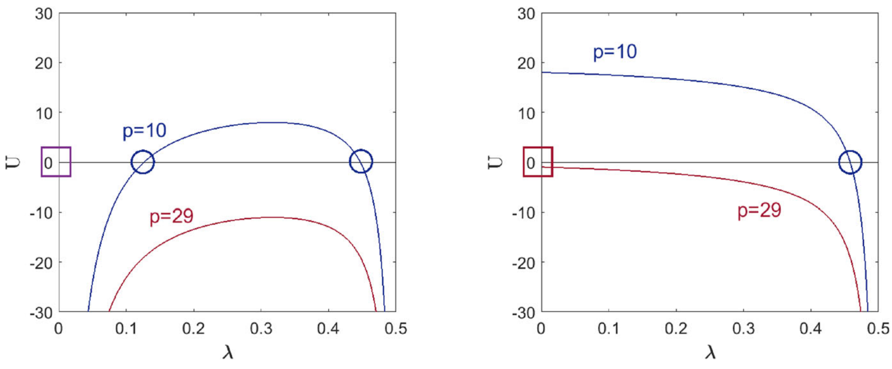


Fig. 8 *N-Limited* Mean sojourn time W_i in each queue as a function of the effective arrival rate λ for different values of N, where $\mu_1 = 1$ and $\mu_2 = 1$ ($\rho = 2\lambda$)

   (c) *Three Equilibria* When the maximum expected utility is **positive** there are two equilibria, $\lambda_2 > \lambda_1 > 0$, in addition to the equilibrium at $\lambda = 0$. In the neighbourhood of λ_1, the utility is increasing in the joining rate, and thus every drift is sharpened and this equilibrium is unstable. For λ_2, the utility is decreasing in the joining rate, and therefore the effect is restraining and this equilibrium is **stable**. Thus, in this case we denote $\lambda_e = \lambda_2$.

2. The *N-Limited/Sequential* (Monotone-Decreasing) Case:
   (a) *A Positive Equilibrium* When $V - p - C_W(\frac{1}{\mu_1} + \frac{1}{\mu_2}) > 0$ [see (2)] there exists $\lambda_e > 0$ where $U(\lambda_e) = 0$. Similar to λ_2 in the previous case, this is a **stable** equilibrium. $\lambda = 0$ is not an equilibrium, because an individual would gain a positive utility from deviating from it and therefore will do so.
   (b) *The Zero Equilibrium* When $V - p - C_W(\frac{1}{\mu_1} + \frac{1}{\mu_2}) \leq 0$ the individual gains no profit from joining the service even when there is no queueing time. That is, $\forall\lambda > 0 : U(\lambda) < 0$ and the only equilibrium is when $\lambda_e = 0$.

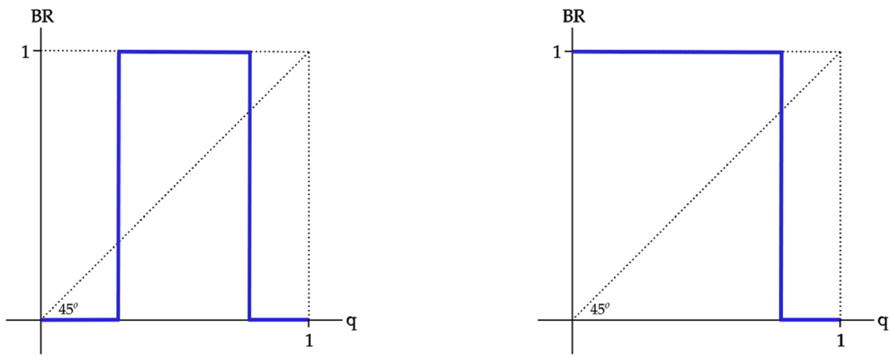
Figure 9a shows two examples of the *Exact-N* case. For $p = 10$ there are three equilibria (marked with a circle or a square), as in case (1c), and for $p = 29$ there is one equilibrium (marked by a square), as in case (1a). In Fig. 9b, there are two examples of the *N-Limited* case. For $p = 10$ the “positive” equilibrium (marked with



(a) *Exact-N* scenario.

(b) *N-Limited* scenario.

Fig. 9 $U(\lambda)$ in the different scenarios with $N = 5$, $\mu_1 = 1$, $\mu_2 = 1$, $C_W = 1$, $V = 30$ and $p = 10$ or $p = 29$



(a) Unimodal $U(\lambda)$.

(b) Monotone Decreasing $U(\lambda)$.

Fig. 10 The best response versus the joining probability

a circle), as in case (2a) and for p = 29 the "zero" equilibrium (marked with a square), as in case (2b).

Let $q \in [0, 1]$ be the probability of joining the service chosen as a strategy by all customers. The *best response* ($\text{BR}(q) \in [0, 1]$) is the best strategy for an individual, assuming all other customers execute joining strategy q. An individual who expects positive utility $U > 0$ joins the system, i.e. BR $= 1$, and one who expects negative utility $U < 0$ does not join, so BR $= 0$. If $U = 0$, the individual is indifferent between joining and baulking. In Fig. 10, we display two examples of a best-response graph as a function of the common joining probability: In Fig. 10a, an example that fits the unimodal case of $U(\lambda)$, and in Fig. 10b an example that fits the monotone decreasing case (in both $\exists\lambda : U(\lambda) > 0$). Customers are homogeneous and therefore equilibrium is reached when all execute the same strategy, thus, the equilibrium strategies are at the values where the graph meets the 45° line.

Notice that in the monotone-decreasing case there is one equilibrium, which is typical to an *Avoid-the-Crowd* (ATC) situation. Compared to the unimodal case, where there are multiple (three) equilibria, which is typical to a *Follow-the-Crowd* (FTC) situation. However, as implicit from Fig. 10a, for small values of q FTC is indeed the case, but for large values, it is an ATC situation (unlike typical FTC cases, the third equilibrium is not $q = 1$).

5 Sequential service

An elementary special case is the sequential service, i.e. when $N = 1$. In this case, the *Exact-N* and *N-Limited* scenarios coincide.

Denote by X_n the number of customers in the system (in fact, at Q_1) at the instant of the nth customer's end of service at Q_2 (there are no customers in Q_2 and the server switches back to Q_1). The law of motion is

$$X_{n+1} = \begin{cases} X_n + \xi - 1 + \eta \,, & X_n \geq 1 \,, \\ 1 + \xi - 1 + \eta = \xi + \eta \,, & X_n = 0 \,, \end{cases} \tag{27}$$

where ξ denotes the number of customers who joined the system (at Q_1) during the service of the customer in Q_1 and η is the number of customers who joined (at Q_1) during the service of the customer in Q_2.

Denote by B_i the service duration of a customer in Q_i $(i = 1, 2)$ with *Laplace-Stieltjes Transforms (LST)* $\tilde{B}_i(s) \equiv E[e^{-sB_i}]$ (note that in this section B_i is a random variable, not the matrix appearing in Sect. 3). Then the *probability generating functions (PGFs)* of ξ and η are given, respectively, by (as in Kleinrock [19] 5.46):

$$\hat{\xi}(z) = E[z^{\xi}] = \tilde{B}_1[\lambda(1 - z)] \,, \quad \hat{\eta}(z) = E[z^{\eta}] = \tilde{B}_2[\lambda(1 - z)] \,.$$

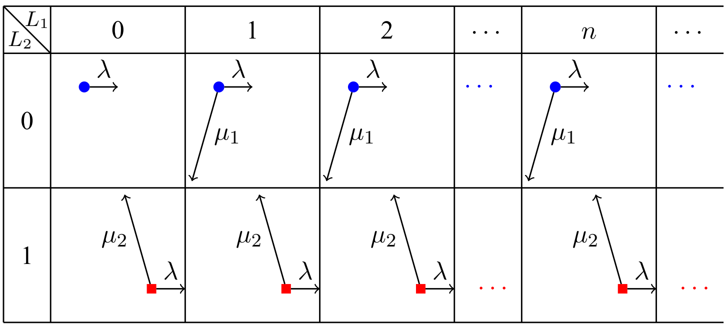


Fig. 11 The transition rate diagram for $N = 1$

Equation (27) resembles the law of motion of the classical $M/G/1$ queue, leading to the *PGF* of the steady-state number of customers in the system, X:

$$\hat{X}(z) = (1-\rho)\frac{1-z}{\hat{\xi}(z)\hat{\eta}(z)-z}\hat{\xi}(z)\hat{\eta}(z)\ , \tag{28}$$

where

$$\rho = \lambda(E[B_1]+E[B_2]) = \lambda\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right) < 1\ . \tag{29}$$

Clearly, condition (29) is identical to the stability condition (5) we have found in Sect. 3 for general N. When $\mu_1 = \mu_2 = \mu$, this condition becomes $\lambda < \frac{\mu}{2}$.

By manipulating the usual balance equations for the probabilities $P_{nj}^{(i)}$, $n = 0, 1, 2, \ldots;\ j = 0, 1;\ i = 1, 2$ corresponding to Fig. 11, we obtain recursive formulas for the stationary probability for every state of the system as a function of the boundary probability $P_{00}^{(1)}$:

$$\begin{aligned}
P_{10}^{(1)} &= \frac{\lambda+\mu_2}{\mu_1}\frac{\lambda}{\mu_2}P_{00}^{(1)}\ ,\\
P_{01}^{(2)} &= \frac{\lambda}{\mu_2}P_{00}^{(1)}\ ,\\
P_{n0}^{(1)} &= \frac{1}{\mu_1}\left[(\lambda+\mu_2)P_{n-1,1}^{(2)} - \lambda P_{n-2,1}^{(2)}\right]\ , \quad \forall n \geq 2\ ,\\
P_{n1}^{(2)} &= \frac{1}{\mu_2}\left[(\lambda+\mu_1)P_{n,0}^{(1)} - \lambda P_{n-1,0}^{(1)}\right]\ , \quad \forall n \geq 1\ .
\end{aligned} \tag{30}$$

Our next goal is to find an expression for the expected number of customers in the system, $E[X] \equiv E[L]$. This is possible by setting $z = 1$ in the derivative of $\hat{X}(z)$ [Eq. (28)], or by exploiting the *Partial Generating Functions* method to calculate separately the two queue sizes, while using the balance equations (as in Yechiali and Naor [29]). But, due to the resemblance to the M/G/1 case, this result can be obtained also by using the *Khinchine–Pollaczek* formula:

$$E[L] = \rho + \frac{\lambda^2 E[B^2]}{2(1-\rho)} = \frac{\rho}{1-\rho} - \frac{\lambda^2}{\mu_1\mu_2(1-\rho)}\ , \tag{31}$$

where $\rho = \lambda(\frac{1}{\mu_1}+\frac{1}{\mu_2})$, $B = B_1 + B_2$ and $E[B^2] = 2(\frac{1}{\mu_1^2}+\frac{1}{\mu_1\mu_2}+\frac{1}{\mu_2^2})$.

From Eq. (31), it follows that the expected number of customers in the system is smaller than in a corresponding standard M/M/1 queue ($E[L_{\mathrm{M/M/1}}] = \frac{\rho}{1-\rho}$). This deserves further examination. In our case, the service time for a customer is composed of two independent lengths, each distributed exponentially, B_1 with mean $\frac{1}{\mu_1}$ and B_2 with mean $\frac{1}{\mu_2}$. Consider an M/M/1 queue with arrival rate λ and exponential service time B_{exp} with mean $E[B_{\text{exp}}] = \frac{1}{\mu_1} + \frac{1}{\mu_2}$. Clearly, these two systems have the same work rate $\rho = \lambda(\frac{1}{\mu_1} + \frac{1}{\mu_2})$ and the same mean service time

$$E[B_1 + B_2] = \frac{1}{\mu_1} + \frac{1}{\mu_2} = E[B_{\text{exp}}] \,.$$

However,

$$\text{Var}[B_1 + B_2] = \frac{1}{\mu_1^2} + \frac{1}{\mu_2^2} < \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)^2 = \text{Var}[B_{\text{exp}}] \,,$$

thus, $E[L]$ in (31) is smaller than $E[L_{\text{M/M/1}}]$.

Applying *Little's Law*, we obtain the expected sojourn time of a customer in the system:

$$W = \frac{\mu_1 + \mu_2 - \lambda}{\mu_1\mu_2(1-\rho)} \,. \tag{32}$$

Proposition 1 *In the N=1 case, while the stability condition holds, the expected sojourn time is an increasing and convex function of the effective arrival rate.*

Proof The claim follows since, under the stability condition, the function's first and second derivatives are positive:

$$W'(\lambda) = \frac{\mu_1^2 + \mu_1\mu_2 + \mu_2^2}{(\mu_1\mu_2(1-\rho))^2} > 0 \,, \tag{33}$$

$$W''(\lambda) = \frac{2(\mu_1^2 + \mu_1\mu_2 + \mu_2^2)(\mu_1 + \mu_2)}{(\mu_1\mu_2(1-\rho))^3} > 0 \,. \tag{34}$$

□

Now we are able to express the effective arrival rate in equilibrium, λ_e, as a function of the parameters and the decision variable p. As explained in Sect. 2.3, in equilibrium $U = 0$ for all customers; hence, by substituting (32) in (1) we get

$$V - p = C_W W(\lambda_e) = C_W \frac{\mu_1 + \mu_2 - \lambda_e}{\mu_1\mu_2\left[1 - \lambda_e\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)\right]} \,,$$

and isolating λ_e:

$$\lambda_e = \frac{C_W(\mu_1 + \mu_2) - \mu_1\mu_2(V-p)}{C_W - (\mu_1 + \mu_2)(V-p)} \,. \tag{35}$$

As derived from Proposition 1, and elaborated in Sect. 4, Case 2a, when condition (2) holds, there exits one positive value of λ_e, and indeed, the expression in (35) is positive.

Our ultimate goal is to find the maximal profit for the server, $r^*(p)$, and the corresponding optimal price p^*. For $N = 1$ Eq. (3) is

$$r = \lambda_e(p - C_S) ,$$

and its derivative with respect to p is

$$\begin{aligned} r'(p) &= \lambda_e'(p)p + \lambda_e(p) - C_S\lambda_e'(p) \\ &= \lambda_e(p) + \lambda_e'(p)(p - C_S) \\ &= \frac{C_W(\mu_1 + \mu_2) - \mu_1\mu_2(V - p)}{C_W - (\mu_1 + \mu_2)(V - p)} + \frac{C_W[\mu_1\mu_2 - (\mu_1 + \mu_2)^2]}{[C_W - (\mu_1 + \mu_2)(V - p)]^2} . \end{aligned}$$

From $r'(p^*) = 0$, we eventually get two solutions:

$$p^*_{1,2} = V - \frac{C_W}{\mu_1 + \mu_2} \pm \frac{\sqrt{[\frac{(\mu_1+\mu_2)^2}{\mu_1\mu_2} - 1][C_W(\mu_1 + \mu_2)(V - C_S) - C_W^2]}}{\mu_1 + \mu_2} . \quad (36)$$

Notice that

$$\frac{(\mu_1 + \mu_2)^2}{\mu_1\mu_2} - 1 > 0 , \quad (37)$$

which is the same as

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} > \frac{1}{\mu_1 + \mu_2} .$$

Using the above and the fact that in a positive equilibrium condition (2) holds, we get

$$p^* < V - C_W\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right) < V - \frac{C_W}{\mu_1 + \mu_2} ,$$

so we conclude that there is one possible optimal price:

$$p^* = V - \frac{C_W}{\mu_1 + \mu_2} - \frac{\sqrt{[\frac{(\mu_1+\mu_2)^2}{\mu_1\mu_2} - 1][C_W(\mu_1 + \mu_2)(V - C_S) - C_W^2]}}{\mu_1 + \mu_2} . \quad (38)$$

From (37), the discriminant for (38) is non-negative when

$$C_W(\mu_1 + \mu_2)(V - C_S) - C_W^2 \geq 0 ,$$

that is,

$$V \geq \frac{C_{\mathrm{W}}}{\mu_1+\mu_2} + C_{\mathrm{S}} \,. \tag{39}$$

The meaning of Eq. (39) is that for a set of parameters which would not satisfy this condition the system is not profitable for $N = 1$.

Finally, the optimal profit can be calculated by using Eq. (35):

$$r^* = \frac{C_{\mathrm{W}}(\mu_1+\mu_2) - \mu_1\mu_2(V-p^*)}{C_{\mathrm{W}} - (\mu_1+\mu_2)(V-p^*)}(p^* - C_{\mathrm{S}}) \,, \tag{40}$$

where p^* is given in (38). The next proposition specifies the sufficient (and for *N-Limited* also necessary) conditions on the parameters for the optimal threshold to be $N^* = 1$.

Remark 1 Proposition 3 in Iravani et al. [16] provides a more general sufficient condition than stated in Proposition 2 below. We provide it here for completeness.

Proposition 2 *For all values of* V *such that* $\lambda_{\mathrm{e}} > 0$ *,*

1. *For the* ***N-Limited*** *policy:* $N^* = 1$ *if and only if* $\frac{\mu_1 C_{\mathrm{S}}}{C_{\mathrm{W}}} \leq 1$.
2. *For the* ***Exact-N*** *policy:* $N^* = 1$ *if* $\frac{\mu_1 C_{\mathrm{S}}}{C_{\mathrm{W}}} \leq 1$.

Proof We have already argued that the monopolistic policy is socially optimal. Hence, when the cost of switching is smaller than the cost of waiting for one service ($C_{\mathrm{S}} \leq \frac{C_{\mathrm{W}}}{\mu_1}$), it will always be better to switch after one service and not wait for more customers. This explains why $\frac{\mu_1 C_{\mathrm{S}}}{C_{\mathrm{W}}} \leq 1$ is a sufficient condition for $N^* = 1$ for both policies.

Notice that the only states where the $N = 1$ policy and the *2-Limited* policy yield different actions by the server are the states of the type $(L_1 > 0, L_2 = 1, I = 1)$. In this situation, by applying the *2-Limited* policy rather than the $N = 1$ policy, the server saves one switching cost C_{S} while the customer in Q_2 will incur an additional cost of $\frac{C_{\mathrm{W}}}{\mu_1}$. All other costs are similar (all customers incur $\frac{C_{\mathrm{W}}}{\mu_1} + \frac{C_{\mathrm{W}}}{\mu_2}$ for each customer that has come before them and for themselves). So, if $C_{\mathrm{S}} = \frac{C_{\mathrm{W}}}{\mu_1}$ the server is indifferent between the two policies with respect to social welfare. For $C_{\mathrm{S}} = \frac{C_{\mathrm{W}}}{\mu_1} - \varepsilon$ the optimal policy is $N^* = 1$ and for $C_{\mathrm{S}} = \frac{C_{\mathrm{W}}}{\mu_1} + \varepsilon$ it is not, because *2-Limited* yields a higher social welfare. Therefore, if $C_{\mathrm{S}} > \frac{C_{\mathrm{W}}}{\mu_1}$ then $N^* \neq 1$, which proves that $C_{\mathrm{S}} \leq \frac{C_{\mathrm{W}}}{\mu_1}$ is a necessary condition for $N^* = 1$. □

Remark 2 Proposition 2 does not depend on the arrival distribution or the service time distribution as long as the stability condition (5) holds (where $\frac{1}{\lambda}$ is the mean inter-arrival time and $\frac{1}{\mu_i}$ $(i = 1, 2)$ is the mean service time at Q_i).

6 Numerical results for optimal values

In this section, we present numerical results and shed light on the system behaviour. For different sets of parameters, we calculate the optimal values of the objective function (the server's profit) and the corresponding decision variables (the price and threshold) under each of the operating policies. To reduce the number of free parameters, we assume the same service rate for both of the phases, i.e. $\mu_1 = \mu_2 = \mu$ ($\rho = 2\frac{\lambda}{\mu}$). As long as a positive profit is reachable, we assume a positive stable equilibrium has been reached (see Sect. 4). Where no positive profit is possible, there is no service, the system is not operating, and the optimal profit is $r^* = 0$ (an exception takes place in § 6.1, where a negative profit is displayed to illustrate the behaviour of the profit's function).

Following Naor [22], we normalize monetary values by setting the unit to be the expected cost of waiting for a single service-phase completion, i.e. $\frac{C_W}{\mu}$. We show the optimal values achieved as a function of the normalized service value $\hat{V} = \frac{\mu V}{C_W}$ and the normalized switching cost $\hat{C}_S = \frac{\mu C_S}{C_W}$. Consequently, the profit and the price presented in this section are normalized in the same manner. Computationally, this normalization is equivalent to assuming $\mu = 1$ and $C_W = 1$.

6.1 Equilibrium effective arrival rate λ_e and profit $\hat{r}$ as functions of price $\hat{p}$ when threshold N is fixed

We gain several insights based on graphs like those in Figs. 12 and 13 (note that a change in the value of the switching cost parameter does not affect the equilibrium arrival rate for a fixed N):

1. As expected, the equilibrium joining rate is a decreasing function of the price and of N. Furthermore, for the *Exact-N* policy, except for the case of $N = 1$, this function has a point of discontinuity where λ_e drops down to zero (Fig. 12a). For the *N-Limited* policy, the function is continuous for all values of N (Fig. 13a). The reason for this difference lies in the form of the $U(\lambda)$ function, as we now explain. Denote by $\bar{\lambda}(\hat{p})$ the maximizer of the function $U(\lambda)$ for a certain $\hat{p}$ (while N is fixed). As $\hat{p}$ increases, the function's maximal value $U(\bar{\lambda}(\hat{p}))$ decreases. In the unimodal case, as it is possible to see in Fig. 9a, there exists $\hat{p}_{\max}$ such that for all $\hat{p} < \hat{p}_{\max}$ the maximal value $U(\bar{\lambda})$ is positive; for $\hat{p} = \hat{p}_{\max}$ it is exactly zero; for all $\hat{p} > \hat{p}_{\max}$ it is negative. As explained in Sect. 4, as long as $\hat{p} < \hat{p}_{\max}$ there are three equilibria, where one, $\lambda_2 > \bar{\lambda}$, is a positive stable equilibrium, and therefore, the equilibrium joining rate is $\lambda_e = \lambda_2$. So, on the one hand, $\lambda_e \downarrow \bar{\lambda}$ as $\hat{p} \uparrow \hat{p}_{\max}$. On the other hand, when $\hat{p} \geq \hat{p}_{\max}$, the only stable equilibrium is $\lambda = 0$, so $\lambda_e = 0$ and hence the discontinuity of $\lambda_e(\hat{p})$ at $\hat{p}_{\max}$. In the monotone decreasing case, as illustrated in Fig. 9b, there is a positive equilibrium when $\hat{p} < \hat{p}_{\max}$ and a zero equilibrium otherwise. In this case, $\lambda_e \downarrow 0$ as $\hat{p} \uparrow \hat{p}_{\max}$, and $\lambda_e = 0$ when $\hat{p} \geq \hat{p}_{\max}$. Hence, there is no point of discontinuity.
2. The server's net income for $\hat{p} = 0$, presuming the system is operating, is of course negative, and for $\hat{p} > 0$ it is an increasing function of the price until

a local maximum point (for example, all the graphs in Figs. 12b and 13b, the graphs for $3 \leq N \leq 6$ in Fig. 12c and for $N \geq 3$ in Fig. 13c), or, if it comes first, until the equilibrium joining rate becomes zero, and therefore also the profit (as in the graphs for $N \leq 2, N \geq 7$ in Fig. 12c and for $N \leq 2$ in Fig. 13c). It is not assured that for every set of parameters there exists $\hat{p}$ such that the net income is positive. In general, the profit grows slowly with an increase in $\hat{p}$ until the maximum point and decreases fast. Considering this, it is better for the server to make an under-assessment of the optimal price than to guess too high. This conclusion is enhanced in the *Exact-N* case where the profit may decrease to zero by any small deviation from the optimal value (as brought out in Fig. 12b).

3. Intuitively, an increase in the value of N (while $\hat{p}$ is fixed) decreases the joining rate, but there is a salient difference between the two policies. For the *N-Limited* regime, increasing N has a minor effect on both functions $\lambda_e(\hat{p})$ and $\hat{r}(\hat{p})$ (Fig. 13). Furthermore, the larger the value of N, the smaller the effect. In contrast, due to the *Exact-N* regime's strictness, the impact of increasing N is much stronger under this policy. There are two main outcomes for this phenomenon (both are noticeable in Fig. 12). The first is that the maximal profit is achieved with a significantly lower price with any increase of N, and the second is that the maximal price allowing a profitable service decreases as well.

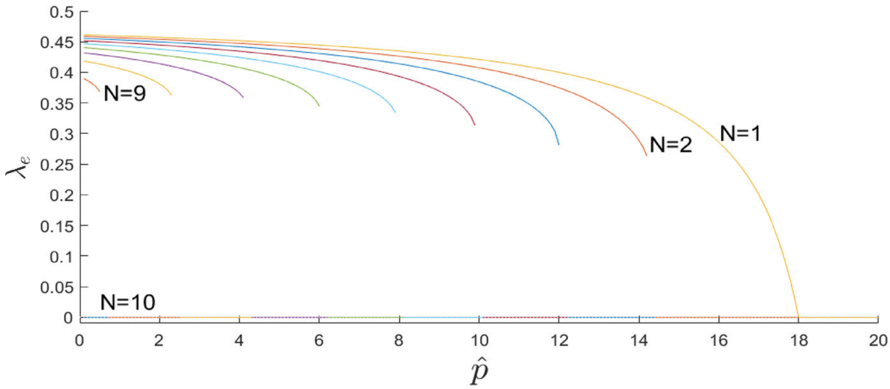
6.2 Maximal profit $\hat{r}^*$ as a function of threshold N

The empirical results indicate that for the *Exact-N* policy the optimal profit $\hat{r}^*$, while positive, is a unimodal concave function of N with one maximal point, followed by a decrease, until it is no longer profitable to operate the system.

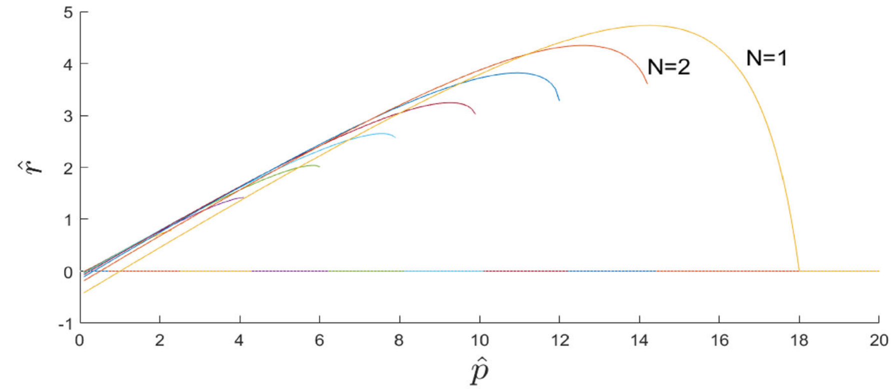
Observing the *N-Limited* policy, we find a similar behaviour for small values of N, but with a more moderate slope around the maximal point, such that adjacent values of N yield approximately the same earnings. Also for the latter policy, after the peak, while N grows $\hat{r}^*$ decreases, but the decrease is convex and converges to a value not much smaller than the maximal value. This matches our conclusion from the previous subsection, that under the *N-Limited* policy, a change in the chosen threshold has a negligible effect on the profit for large values of N. The intuitive explanation for this phenomenon (as elaborated in § 6.4), is that as N increases, the probability that this threshold will be reached decreases. Notice that for $N \to \infty$ this policy is in fact the well-studied *Exhaustive* regime (see, for example, Yechiali [28]).

Some intuitive conclusions, as illustrated in Fig. 14:

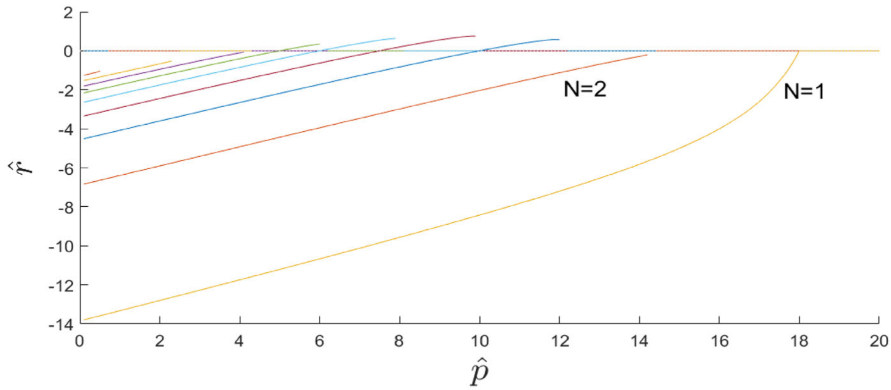
1. The choice of N is much more critical for the *Exact-N* policy.
2. An over-assessment of the optimal threshold under the *N-Limited* policy has minor consequences.
3. For the same values of the parameters, the optimal value N^* under the *N-Limited* policy is at least as large as the N^* under the *Exact-N* policy.



(a) $\lambda_e(\hat{p})$

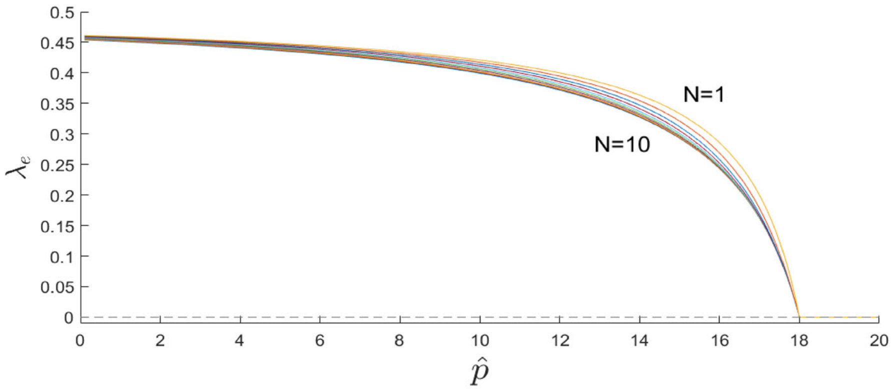


(b) $\hat{r}(\hat{p})$ for $\hat{C}_S = 1$

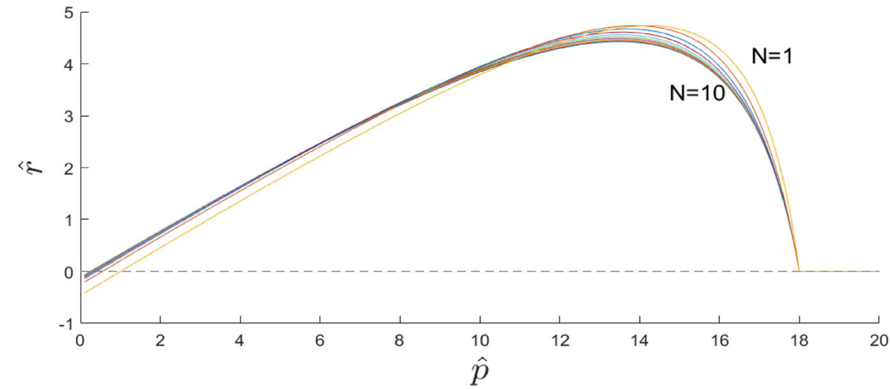


(c) $\hat{r}(\hat{p})$ for $\hat{C}_S = 30$

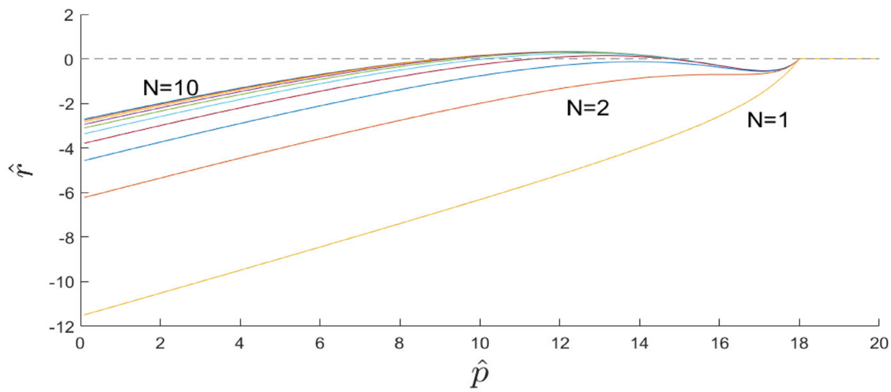
Fig. 12 $\lambda_e(\hat{p})$ and $\hat{r}(\hat{p})$ for $N = 1, 2, \ldots, 10$, $\hat{V} = 20$, under the *Exact-N* policy



(a) $\lambda_e(\hat{p})$

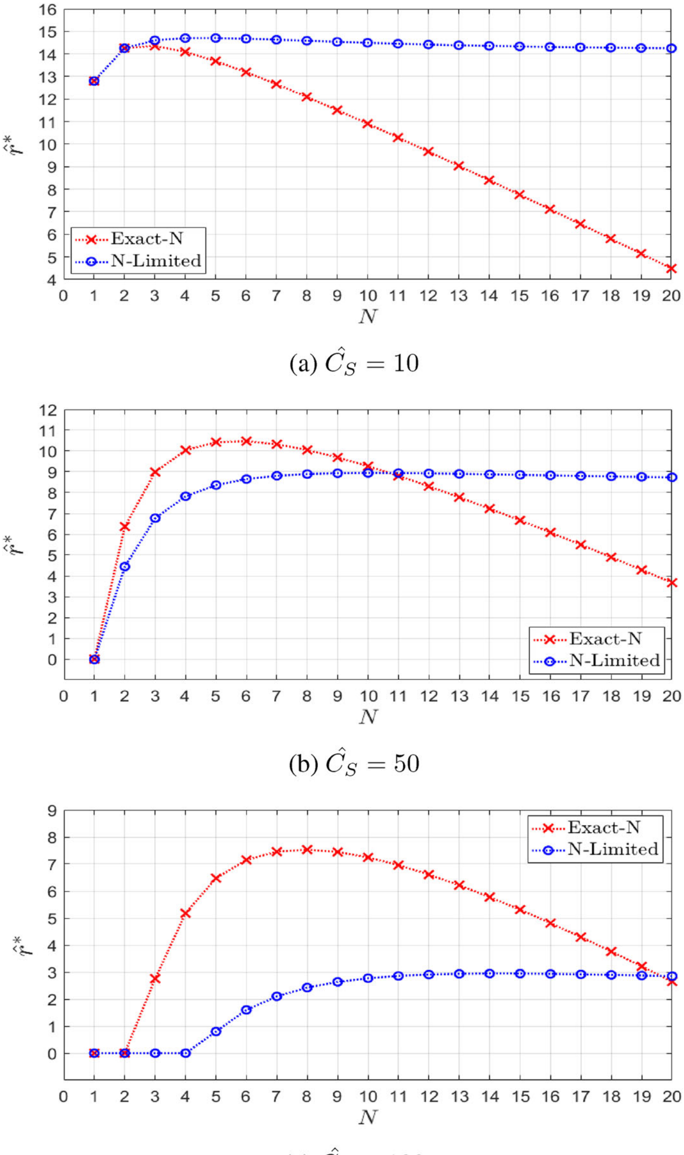


(b) $\hat{r}(\hat{p})$ for $\hat{C_S} = 1$



(c) $\hat{r}(\hat{p})$ for $\hat{C_S} = 25$

Fig. 13 $\lambda_e(\hat{p})$ and $\hat{r}(\hat{p})$ for $N = 1, 2, \ldots, 10$, $\hat{V} = 20$, under the *N-Limited* policy



(a) $\hat{C}_S = 10$

(b) $\hat{C}_S = 50$

(c) $\hat{C}_S = 100$

Fig. 14 $\hat{r}^*(N)$ under each operating policy for $\hat{V} = 50$ and various values of $\hat{C}_S$

4. As proved in Proposition 2, when $\hat{C}_S \leq 1$ the optimal threshold is $N^* = 1$ and the two policies are identical. Generally speaking, for a fixed value of $\hat{V}$, small values of $\hat{C}_S$ yield a higher maximal profit $\hat{r}^*$ under the *N-Limited* policy while large values of $\hat{C}_S$ yield a higher maximal profit $\hat{r}^*$ under the *Exact-N* policy. In the next subsection, we explore this property, elaborate on it and identify some exceptions.

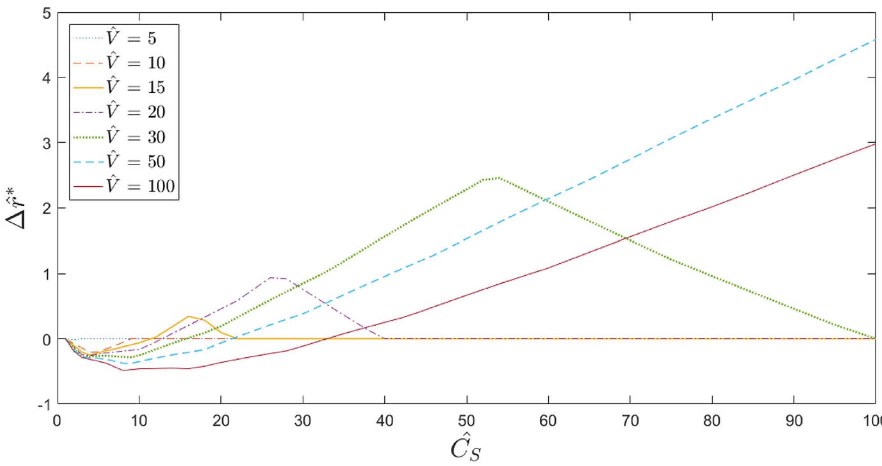


Fig. 15 $\Delta\hat{r}^*(\hat{C}_\text{S})$ for different values of $\hat{V}$

6.3 Maximal profit $\hat{r}^*$ as a function of service value $\hat{V}$ and switching cost $\hat{C}_\text{S}$

In this subsection, we analyse and compare the maximal profit under the two operating policies. Let $\Delta\hat{r}^*$ denote the difference between $\hat{r}^*$ under the *Exact-N* and under the *N-Limited* regimes. Figure 15 confirms Conclusion 3 from Sect. 6.2, that for smaller values of $\hat{C}_\text{S}$ the *N-Limited* regime is more profitable, and for larger values the *Exact-N* regime is more profitable. In fact, we can divide each graph into 3–5 parts, the first two and the last one exist for all values of $\hat{V}$ and the third and fourth are not found for small values of $\hat{V}$ (approximately, $\hat{V} < 10$) :

A. $\Delta\hat{r}^* = 0$: Here $0 < \hat{C}_\text{S} \leq 1$, $N^* = 1$ for both policies (see Proposition 2) and therefore they are identical.
B. $\Delta\hat{r}^* < 0$: The values of $\hat{C}_\text{S}$ where the *N-Limited* policy is more profitable than *Exact-N*.
C. $\Delta\hat{r}^* > 0$ and increasing: Both policies are profitable but *Exact-N* is more profitable.
D. $\Delta\hat{r}^* > 0$ and decreasing: Only *Exact-N* is profitable.
E. $\Delta\hat{r}^* = 0$: $\hat{C}_\text{S}$ is very large, preventing both of the policies profiting.

The larger $\hat{V}$, the larger $|\Delta\hat{r}^*|$ gets at its extrema points (while the maximal advantage of applying *Exact-N* is always larger than the maximal advantage of applying *N-Limited*), and these points are reached at a larger $\hat{C}_\text{S}$ (parts 2–4 are wider). The meaning of this is that large values of $\hat{V}$ lead to large differences in potential profit and intensify the importance of choosing the right policy. Furthermore, Fig. 15 confirms the intuitive insight that for every value of $\hat{V}$ there exits a sufficiently large value of $\hat{C}_\text{S}$ denying both of the regimes the potential to be profitable. We provide an analytical proof for this claim:

Proposition 3 *For every value of $\hat{V}$, there exists a corresponding value of $\hat{C}_S$ such that the system is not profitable under any operating policy, and when $\mu_1 = \mu_2$ this value is smaller than: $\hat{V}^2 - 3\hat{V} + 2$.*

Proof Denote by $\tilde{N}$ the average number of customers served at Q_1 before the server switches to Q_2 (under the *Exact-N* regime $\tilde{N} = N$). The server earns p and incurs an average cost of $\frac{C_S}{\tilde{N}}$ for each served customer, so for the service to be profitable it must be that [the second inequality is from Eq. (2)]

$$\frac{C_S}{\tilde{N}} < p < V - C_W\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right),$$

yielding a lower bound for $\tilde{N}$:

$$\tilde{N} > \frac{C_S}{V - C_W\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)}. \tag{41}$$

Suppose that for a certain operating policy the threshold for an arbitrary event of switching queues is M, so the expected sojourn time W_M of the nth customer in that cycle ($1 \le n \le M$) satisfies

$$W_M > \frac{M - n + 1}{\mu_1} + \frac{n}{\mu_2},$$

and, when $\mu_1 = \mu_2 = \mu$,

$$W_M > \frac{M+1}{\mu}.$$

Since, by definition, the average M is $\tilde{N}$, we get for the general case that the mean sojourn time satisfies

$$W > \frac{\tilde{N}+1}{\mu}.$$

Using Eq. (41) we get

$$W > \frac{\frac{C_S}{V - \frac{2}{\mu}C_W} + 1}{\mu} = \frac{\mu V - 2C_W + \mu C_S}{\mu^2 V - 2\mu C_W}. \tag{42}$$

For customers to join the service, the next condition must hold [using (42)]:

$$V > C_W W > \frac{\mu C_W V - 2C_W^2 + \mu C_W C_S}{\mu^2 V - 2\mu C_W}.$$

We multiply by the positive denominator ($V > C_{\mathrm{W}}(\frac{1}{\mu_1} + \frac{1}{\mu_2})$), divide by C_{W}^2 and rearrange:

$$\frac{\mu C_{\mathrm{S}}}{C_{\mathrm{W}}} < \left(\frac{\mu V}{C_{\mathrm{W}}}\right)^2 - 3\frac{\mu V}{C_{\mathrm{W}}} + 2\,,$$

which is

$$\hat{C}_{\mathrm{S}} < \hat{V}^2 - 3\hat{V} + 2\,. \tag{43}$$

This upper bound for $\hat{C}_{\mathrm{S}}$ implies that for every set of parameters which do not agree to this term (43) the system is not profitable. □

Remark 3 By Condition (2), $\hat{V} > 2$ when $\mu_1 = \mu_2$, and therefore the upper bound in (43) is always positive.

Remark 4 Proposition 3 does not depend on the arrival distribution or the service time distribution as long as the stability condition (5) holds (where $\frac{1}{\lambda}$ is the mean inter-arrival time and $\frac{1}{\mu_i}$ ($i = 1, 2$) is the mean service time at Q_i).

Figure 16 shows which policy is better for each set of (normalized) parameters and how much more profitable it is. Using the same partition as defined with regard to Fig. 15, we label segments *A*–*E* in the next figure. For better intelligibility, segment '*D*' is coloured red and segment '*E*' yellow, while the upper bound in Eq. (43) is marked by a dashed line inside the yellow area. This figure provides a graphical illustration of our findings, from which we make several observations:

1. For small switching costs $\hat{C}_{\mathrm{S}}$ and value of service $\hat{V}$, the *N-Limited* policy is more profitable compared to the *Exact-N* policy, as long as the system can be profitable. Interestingly, it shows that there is an additional segment where only *N-Limited* is profitable (the small area, around $6 \le \hat{V} \le 10$ and $4 \le \hat{C}_{\mathrm{S}} \le 9$, which is coloured blue and labelled '*F*' in Fig. 16).
2. As displayed in Fig. 15, it is also noticeable in Fig. 16 that for a fixed value of $\hat{V}$, increasing the value of $\hat{C}_{\mathrm{S}}$ will eventually lead to a better result applying the *Exact-N* policy, then to a situation where only *Exact-N* is profitable, a further increase would lead to a non-profitable system. In Fig. 16, it is also shown that increasing $\hat{V}$, while $\hat{C}_{\mathrm{S}}$ is fixed, will eventually lead to a better result applying the *N-Limited* policy.
3. For $\hat{V} > 10$ (approximately) there is a relation between $\hat{V}$ and $\hat{C}_{\mathrm{S}}$, which is the upper line marked with zeros, that separates the values of the parameters where the *N-Limited* policy is more profitable (section '*B*') from those where the *Exact-N* policy is more profitable (section '*C*'). It is recognizable that a linear function can be a good fit for this relation. We find that (approximately) below the line

$$\hat{C}_{\mathrm{S}} = 0.24\hat{V} + 8.49 \tag{44}$$

the *N-Limited* policy is more profitable, whereas above it the *Exact-N* policy is more profitable.

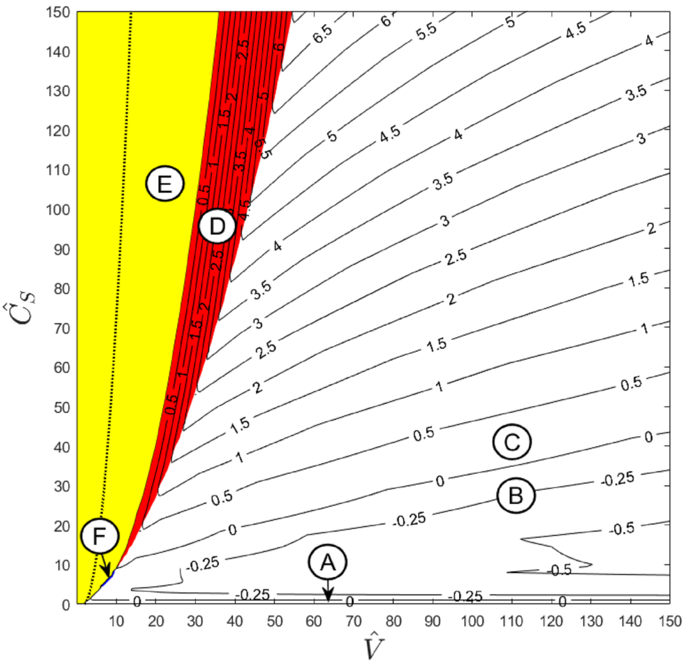


Fig. 16 $\Delta\hat{r}^*(\hat{V}, \hat{C}_S)$ and profitability segmentation

4. There is another relation between these parameters, which is the border of the yellow coloured area (section 'E'), that distinguishes between a profitable system and a futile one. In this case, a quadratic function is a decent fit for that relation, so in the same manner, below (i.e. to the right of) the curve

$$\hat{C}_S = 0.14\hat{V}^2 - 0.79\hat{V} + 3.24 \tag{45}$$

the system is profitable, and above (i.e. to the left of) it, the system is not profitable.

Remark 5 Due to the use of normalized parameters, Fig. 16 contains the entire set of possibilities for this model (with the restriction $\mu_1 = \mu_2$). Because of the interpretation of this normalization, the range considered, between the values 0–150 for both axes, is a satisfying scope for the majority of real world applications. We measure the value of service and the switching cost in units of customers' mean cost while waiting for one service. For example, the meaning of $\hat{V} = 150$ is that the value a customer benefits from the service is equal to the cost of waiting for 150 services, that is, a queue of 75 customers.

6.4 Optimal threshold N^* as a function of service value $\hat{V}$ and switching cost $\hat{C}_S$

In the next two subsections, we discuss the optimal values for the decision variables, the threshold N^* and the price $\hat{p}^*$. From our empirical study, we extract several inferences, as reflected in the example in Table 1:

1. The optimal value for N depends almost solely on the value of $\hat{C}_S$, while the magnitude of $\hat{V}$ mainly determines whether the system can be profitable for the given $\hat{C}_S$ (an empty cell in the table represents a non-profitable system).
2. Of course, for both policies, N^* increases in $\hat{C}_S$, but the changes under *N-Limited* are faster. Notice that an *N-Limited* server does not always wait for the number of customers in Q_2 to reach N and that the bigger this N the bigger the gap between it and $\tilde{N}$, the average number of customers in Q_2 when switching. As seen from comparing the left third of Table 1 with the middle third, N^* is larger under *N-Limited* compared to under *Exact-N*. However, as noticeable in the right third of the table, $\tilde{N}^*$ under the *N-Limited* policy is generally smaller than N^* under the *Exact-N* policy for the same parameters (with exceptions for small values of $\hat{C}_S$).
3. Even though N^* under the *N-Limited* policy is a non-increasing function of $\hat{V}$, $\tilde{N}^*$ does increase in $\hat{V}$. This is because $\tilde{N}$, the average number of customers served between every two consecutive switches under this policy, clearly increases in the equilibrium effective arrival rate, and the latter grows in the normalized value of service, as we show in Sect. 6.6.

6.5 Optimal price p^* as a function of service value $\hat{V}$ and switching cost $\hat{C}_S$

From the empirical results (for example, Figs. 17, 18) we see that the dependence of $\hat{p}^*$ on $\hat{V}$ is much stronger than the dependence of $\hat{p}^*$ on $\hat{C}_S$. $\hat{p}^*(\hat{V})$ is approximately linear (with a slope between 0.85 and 0.9 for both policies and all different values of $\hat{C}_S$) with discontinuities at the values of $\hat{V}$ where N^* changes. Consider, for example, Fig. 17 with $\hat{C}_S = 50$. Between $\hat{V} = 33$ and $\hat{V} = 33.5$ there is a point of discontinuity where the value of the optimal price $\hat{p}^*$ decreases, while at the same time the optimal threshold N^* increases from 5 to 6. Our interpretation of this phenomenon is that, from the server's point of view, the aggravation in customers' utility caused by the increase in N^* is compensated by a decrease in the service fee. Because N^* increases with $\hat{C}_S$, this can also explain the decrease of $\hat{p}^*$ while $\hat{C}_S$ increases. Another interesting observation is that the optimal prices are similar under the two policies, as long as the systems are profitable.

6.6 Optimal equilibrium effective arrival rate λ_e^* as a function of service value $\hat{V}$ and switching cost $\hat{C}_S$

The equilibrium effective arrival rate under the optimal price and threshold, denoted λ_e^*, is more strongly affected by $\hat{V}$ than by $\hat{C}_S$ (mainly for low values of $\hat{V}$), but a more prominent difference is the direction.

Table 1 Optimal thresholds N^* under *Exact-N*, N^* and $\tilde{N}^*$ under *N-Limited*

$\hat{C}_S$	N^* under *Exact-N*			N^* under *N-Limited*			$\tilde{N}^*$ under *N-Limited*		
	$\hat{V}=15$	$\hat{V}=30$	$\hat{V}=100$	$\hat{V}=15$	$\hat{V}=30$	$\hat{V}=100$	$\hat{V}=15$	$\hat{V}=30$	$\hat{V}=100$
3	1	2	2	3	3	3	1.664	1.925	2.296
10	2	3	3	5	5	5	1.783	2.239	2.954
20	3	4	4		7	6		2.438	3.190
30		4	5		8	8		2.594	3.513
40		5	5		9	9		2.768	3.677
50		5	6		10	10		2.946	3.835
60		6	6			11			3.988
70		6	7			12			4.135
80		7	7			13			4.274
90		7	8			14			4.412
100			8			14			4.510

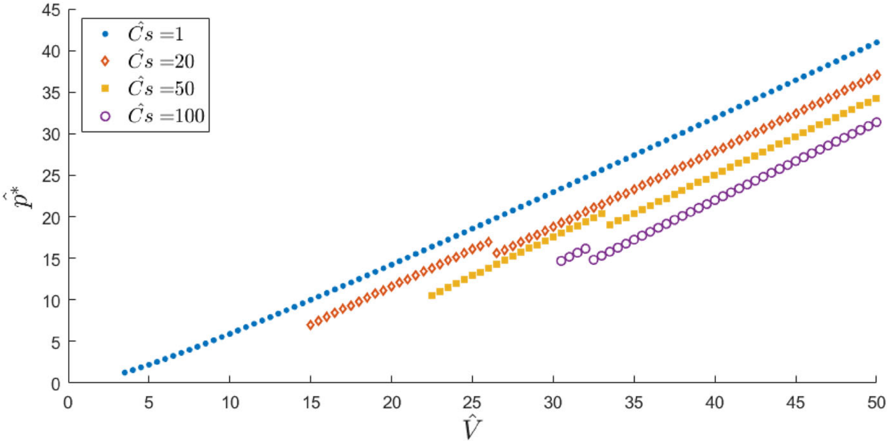


Fig. 17 $\hat{p}^*(\hat{V})$ for various values of $\hat{C}_{\mathrm{S}}$, under the *Exact-N* policy

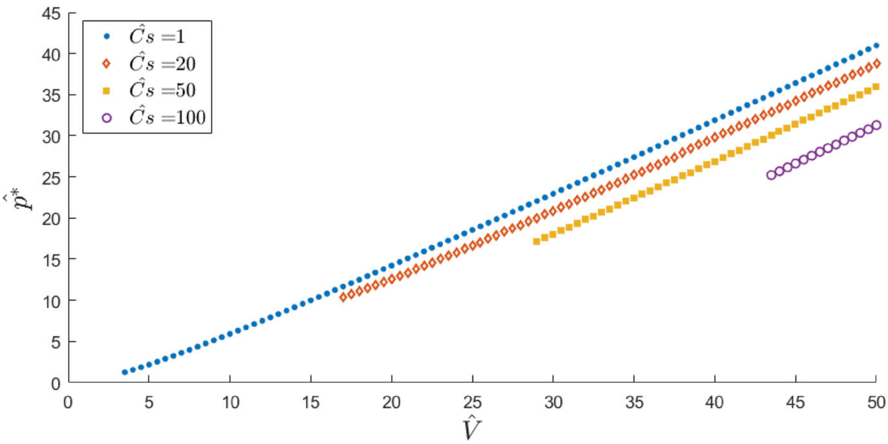


Fig. 18 $\hat{p}^*(\hat{V})$ for various values of $\hat{C}_{\mathrm{S}}$, under the *N-Limited* policy

For both policies, as $\hat{V}$ grows λ_{e}^* naturally increases (see, for example, Figs. 19, 20). When the reward to the customers grows infinitely, the equilibrium joining arrival rate grows to the edge of stability: $\hat{V} \to \infty \implies \lambda_{\mathrm{e}}^* = \frac{\rho_{\mathrm{e}}^*}{2} \to \frac{1}{2}$. However, for a fixed $\hat{V}$, when $\hat{C}_{\mathrm{S}}$ grows, the tendency is opposite under the two policies. Under *Exact-N* λ_{e}^* is generally decreasing (as in Fig. 19) and under *N-Limited* it is increasing (as in Fig. 20). That is, the bigger $\hat{C}_{\mathrm{S}}$ the bigger the difference in λ_{e}^* under the two regimes.

The reason for this phenomenon is not intuitive and we suggest the following speculation: We have seen that N^* increases with $\hat{C}_{\mathrm{S}}$ (see Sect. 6.4) and that $\hat{p}^*$ decreases with $\hat{C}_{\mathrm{S}}$ (see Sect. 6.5). Hence, there are two opposing effects on the optimal equilibrium joining rate λ_{e}^*. On the one hand, an increase in λ_{e}^* due to the decrease in $\hat{p}^*$ and on the other hand a decrease in λ_{e}^* because of the increase in waiting time that accompanies the increase in the selected threshold N^*. Our numerical study

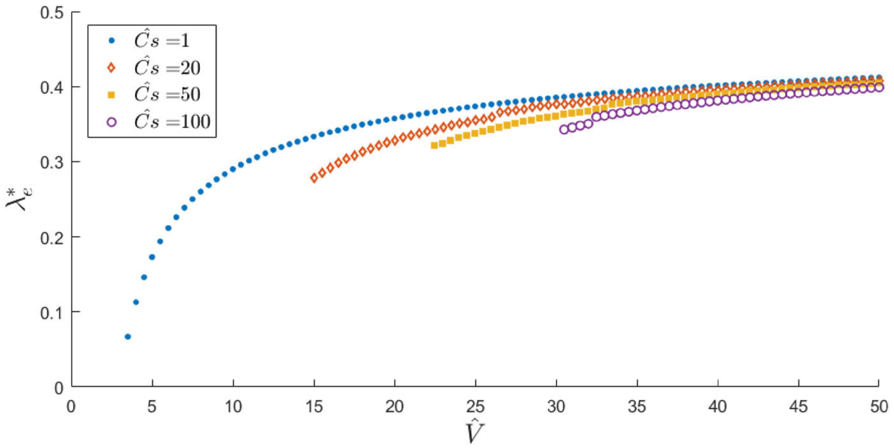


Fig. 19 $\lambda_e^*(\hat{V})$ for various values of $\hat{C}_S$, under the *Exact-N* policy

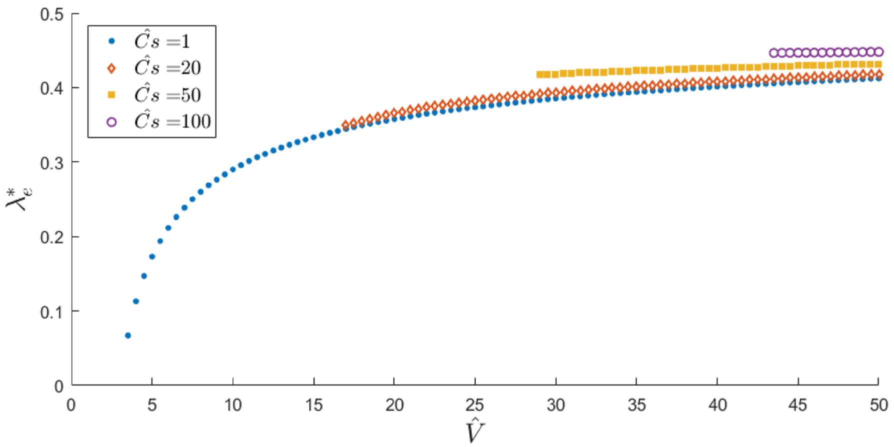


Fig. 20 $\lambda_e^*(\hat{V})$ for various values of $\hat{C}_S$, under the *N-Limited* policy

concludes that under the *Exact-N* policy the effect of changing the threshold is stronger compared to that of changing the price, while under the *N-Limited* policy the leverage of the threshold is much smaller, due to the policy's adaptability.

7 Summary

The present paper considers an unobservable, two-phase, tandem queueing system with an alternating server. We study the strategic customer behaviour under two threshold-based policies, applied by a profit-maximizing server, while waiting and switching costs are taken into account. Optimization performances in equilibrium, under each of these regimes, are analysed and compared.

By defining the system as a QBD process, we derive the system's steady-state probabilities and obtain the mean sojourn time for each policy. Interestingly, the stability condition of the system is independent of the switching policy and/or the chosen threshold N and requires that the mean inter-arrival time should be greater than the mean total service time given to each individual customer.

Next, we analyse the equilibrium behaviour. We learn that under the *N-Limited* policy, or a sequential service ($N = 1$), the system is in a typical *Avoid-the-Crowd* (ATC) situation with one equilibrium. In contrast, under the *Exact-N* policy (for $N \geq 2$), the system is in a *Follow-the-Crowd* (FTC) situation for low joining rates and ATC for high joining rates, with one, two or three equilibria. We see this kind of behaviour in Bountali and Economou [7,8] in tandem two-node assembly service systems with batch features.

Delving deeper into the sequential policy, we prove that the necessary condition for the optimal threshold to be $N^* = 1$ is $\frac{\mu_1 C_S}{C_W} \leq 1$. Under the *N-Limited* policy this is also a sufficient condition.

From an extensive numerical study, we learn about behaviours of the optimal profit, the optimal equilibrium joining rate and the optimal decision variables. Here are some of the more conspicuous ones:

1. The server exploits an increase in the (normalized) service value $\hat{V}$ to directly raise the optimal (normalized) price $\hat{p}^*$. An increase in the (normalized) switching cost $\hat{C}_S$ leads to an obvious increase in N^* (to minimize expenses), and a compensating decrease in $\hat{p}^*$.
2. Seemingly, under the *N-Limited* policy, the optimal threshold, determined by the server, increases faster with the switching cost. However, the expected number of customers served between every two adjacent switches (which we denote by $\tilde{N}$) is mostly smaller than the optimal threshold under the *Exact-N* policy for the same case (with exceptions for small values of $\hat{C}_S$).
3. The equilibrium joining rate increases with the service value, limited only by the stability of the system. An increase in switching cost yields contrasting behaviours under the different policies: a decrease in equilibrium joining rate under the *Exact-N* policy and an increase under *N-Limited*.

This empirical study also yields managerial implications for the strategic calibration of the decision variables. For instance, concerning the service fee determined by the server, an under-assessment of the optimal price is better than over-assessment, particularly under the *Exact-N* policy where a slightly excessive price leads to an immediate halt of the joining rate. Another prominent example is that the choice of the threshold is less crucial for the *N-Limited* policy, due to a minor deterioration associated with exceeding the optimal value. A very interesting question is which policy is more profitable. The answer depends on the parameters: A sufficiently large value of service will lead the server to prefer the *N-Limited* policy, whereas a sufficiently large switching cost will divert the server to the *Exact-N* policy. An excessive value of switching cost would preclude the system from being profitable. In fact, there is a certain relation of the switching cost to the value of service that determines the regions in which one policy is superior to the other. Another relation between these parameters distinguishes a profitable system from a non-profitable one. Approximated functions are fitted for this relations, linear for the first and quadratic for the second (Eqs. (44), (45), respectively).

This work intends to fill the gap in the literature on strategic behaviour in tandem queueing systems with an alternating server. Subsequent research is desired in many interesting courses. Here are some primary leads:

1. While a model that considers switching cost is a good foundation, subsequent research applying switching time is needed for a better fit to many real-life applications.
2. Our numerical study is focused on the elementary case where the service rates and waiting costs are the same for both service phases. Relaxing this constraint may lead to additional interesting conclusions.
3. Further work should consider other operating policies. An especially captivating policy to consider is an extension of *N-Limited* regime where a threshold for a minimal number of services before switching is added. This addition would presumably improve its performances dealing with high switching cost. In [16] Iravani et al. presented the *Triple-Threshold* (TT) policy, which is a similar, more general, idea. They show that this relatively simple switching policy yields near-optimal performance.
4. Of course, studying the model under different levels of information, where arriving customers are fully informed or partially informed about the state of the system, is also a required sequel. In D'Auria and Kanta [10] there are some good examples for different levels of information that can be considered.

Appendix A: Calculating $\vec{P}_o$

In this appendix, we elaborate the process of calculating $\vec{P}_0$ by replacing one of the non-repeating matrix balance equations with the normalizing equation and by that obtaining a system of equations with a unique solution.

A.1 Exact-N scenario

For notational simplicity let

$$\phi = B_0 + RA_2,$$
$$\psi = (I - R)^{-1}\vec{e}\,,$$

where ϕ is a matrix of size $(2N) \times (2N)$ and ψ is a column vector of size $(2N)$. Thus (13) and (15) become

$$\begin{cases} \vec{P}_0\phi = \vec{0} \\ \vec{P}_0\psi = 1 \end{cases}.$$

Denote ϕ_j as the jth column of the matrix ϕ and expand the first equation:

$$\vec{P}_0\begin{bmatrix}\phi_1 & \phi_2 & \dots & \phi_{2N}\end{bmatrix} = \langle 0, 0, \dots, 0\rangle \ .$$

Now replace the first column of the matrix ϕ by the second equation:

$$\vec{P}_0\begin{bmatrix}\psi & \phi_2 & \dots & \phi_{2N}\end{bmatrix} = \langle 1, 0, \dots, 0\rangle \ .$$

This system has a unique solution for $\vec{P}_0$.

A.2 N-Limited scenario

The notation in this case is as follows:

$$\begin{aligned}\phi &= \begin{pmatrix}B_0 & C_1\\ B_1 & A_1 + RA_2\end{pmatrix},\\ \psi &= \langle \vec{e}, (I - R)^{-1}\vec{e}\rangle \ ,\end{aligned}$$

where ϕ is a matrix of size $(3N + 1) \times (3N + 1)$ and ψ is a column vector of size $3N + 1$, in which the first $N + 1$ entries are ones. Similarly to the *Exact-N* scenario, (19), (23) and (24) become

$$\begin{cases}\langle \vec{P}_0, \vec{P}_1\rangle\phi = \vec{0},\\ \langle \vec{P}_0, \vec{P}_1\rangle\psi = 1,\end{cases}$$

and with a similar outcome:

$$\langle \vec{P}_0, \vec{P}_1\rangle\begin{bmatrix}\psi & \phi_2 & \dots & \phi_{2N}\end{bmatrix} = \langle 1, 0, \dots, 0\rangle \ .$$

This gives a solution for $\vec{P}_0$ and $\vec{P}_1$.

References

1. Adan, I.J.B.F., Kulkarni, V.G., Lee, N., Lefeber, E.: Optimal routeing in two-queue polling systems. J. Appl. Probab. **55**(3), 944–967 (2018)
2. Allon, G., Bassamboo, A.: The impact of delaying the delay announcements. Oper. Res. **59**(5), 1198–1210 (2011)
3. Altman, E., Shimkin, N.: Worst-case and Nash routing policies in parallel queues with uncertain service allocations (1993) (Unpublished manuscript)
4. Atar, R., Saha, S.: An ε-Nash equilibrium with high probability for strategic customers in heavy traffic. Math. Oper. Res. **42**(3), 626–647 (2016)
5. Avrachenkov, K., Perel, E., Yechiali, U.: Finite-buffer polling systems with threshold-based switching policy. Top **24**(3), 541–571 (2016)
6. Boon, M.A.A., van der Mei, R., Winands, E.M.M.: Applications of polling systems. Surv. Oper. Res. Manag. Sci. **16**(2), 67–82 (2011)

7. Bountali, O., Economou, A.: Equilibrium joining strategies in batch service queueing systems. Eur. J. Oper. Res. **260**(3), 1142–1151 (2017)
8. Bountali, O., Economou, A.: Strategic customer behavior in a two-stage batch processing system. Queueing Syst. **93**, 3–29 (2019)
9. Boxma, O.J., Levy, H., Weststrate, J.A.: Optimization of polling systems. Department of Operations Research and System Theory [BS] (R 8932) (1989)
10. D'Auria, B., Kanta, S.: Equilibrium strategies in a tandem queue under various levels of information (2011) (unpublished manuscript)
11. Edelson, N.M., Hilderbrand, D.K.: Congestion tolls for Poisson queuing processes. Econom. J. Econom. Soc. **43**(1), 81–92 (1975)
12. Hanukov, G., Yechiali, U.: Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method. Probab. Eng. Inf. Sci. (2020). https://doi.org/10.1017/S0269964819000470:1-16
13. Harchol-Balter, M.: Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge University Press, Cambridge (2013)
14. Hassin, R.: Rational Queueing. Chapman and Hall/CRC, Boca Raton (2016)
15. Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Springer, Berlin (2003)
16. Iravani, S.M.R., Posner, M.J.M., Buzacott, J.A.: A two-stage tandem queue attended by a moving server with holding and switching costs. Queueing Syst. **26**(3–4), 203–228 (1997)
17. Jolles, A., Perel, E., Yechiali, U.: Alternating server with non-zero switch-over times and opposite-queue threshold-based switching policy. Perform. Eval. **126**, 22–38 (2018)
18. Katayama, T.: Analysis of a tandem queueing system with gate attended by a moving server. Rev. Electr. Commun. Lab. NTT **29**(3), 254–267 (1981)
19. Kleinrock, L.: Queueing Systems, Volume 1: Theory. Wiley, Hoboken (1975)
20. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling, vol. 5. Siam, Philadelphia (1999)
21. Nair, S.S.: Two queues in series attended by a single server. Bull. Belg. Math. Soc. **25**, 160–176 (1973)
22. Naor, P.: The regulation of queue size by levying tolls. Econom. J. Econom. Soc. **37**, 15–24 (1969)
23. Neuts, M.F: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press (1981)
24. Perel, E., Yechiali, U.: Two-queue polling systems with switching policy based on the queue that is not being served. Stoch. Models **33**(3), 430–450 (2017)
25. Takagi, H.: Analysis of Polling Systems. MIT Press, Cambridge (1986)
26. Taube-Netto, M.: Two queues in tandem attended by a single server. Oper. Res. **25**(1), 140–147 (1977)
27. Yechiali, U.: On optimal balking rules and toll charges in the GI/M/1 queuing process. Oper. Res. **19**(2), 349–370 (1971)
28. Yechiali, Uri: Analysis and control of polling systems. In: Donatiello, L., Nelson, R. (eds.) Performance Evaluation of Computer and Communication Systems, pp. 630–650. Springer, Berlin (1993)
29. Yechiali, U., Naor, P.: Queuing problems with heterogeneous arrivals and service. Oper. Res. **19**(3), 722–734 (1971)