Hosseini, B.* and **Tan, B.**, "Modelling and Analysis of a Cooperative Service Network," *Computer and Industrial Engineering*, Vol.161, 107620, 2021.

https://doi.org/10.1016/j.cie.2021.107620

# Modeling and analysis of a cooperative service network

Behnaz Hosseini

*Department of Industrial Engineering, Koç University, Istanbul, Turkey,34450*
*bhosseini13@ku.edu.tr*

Barış Tan

*College of Administrative Sciences and Economics, Koç University, Istanbul, Turkey, 34450,*
*btan@ku.edu.tr*

**Abstract**

With the advances in technology and changes in customers' attitude towards different service delivery formats, it is important for the service providers to deliver online services in addition to the traditional face-to-face services. In the cooperative service network presented in this study, service providers cooperate to serve online service requests received by the network in addition to their own customers. Designing and managing the cooperative network effectively increase the utilization of the involved servers, provide an adequate service for the external customers, and increase the profit for both the network and service providers. From the operational perspective, the number and utilization of the members to be included in the network and the price that will be paid to each member for a directed request are the main design questions. In order to answer these questions, we present a stochastic model that captures the dynamics of customer arrivals, assignment, and admission control. To establish this model, we first derive the solution of the dynamic admission control problem for the servers who decide how to admit their own customers and the external online customers using a Markov decision process. We then analyze the operation of the whole network with the servers who use the optimal admission control policy and obtain the system performance measures depending on the members' operational parameters. These results are used to determine the optimal number of servers in the network and the service price to be paid to the participating servers in order to maximize the obtained profit. We show that a cooperative service network is an effective way of utilizing the idle capacity of the servers while providing an adequate service level for the external online customers and increasing the profit for both the network and service providers.

*Keywords:* Cooperative service network, Stochastic models, Dynamic admission control, Markov decision process, Queueing analysis

## 1. Introduction

Technology has improved the ways of delivering services to the customers (Snyder et al., 2016). As one of the significant advances, online services lead to a more convenient access for the customers and a utilization improvement for the providers (Zhang and Prybutok, 2005; Ostrom et al., 2015). Online services enable the service providers to serve not only the customers who come to the service centers but also the external customers over a distance. Although some customers still prefer to have a direct in person interaction with the service providers, online services help to target the customers who want to be served without any cost of traveling (Fernández-Sabiote and Román, 2012). Hence, there is a significant potential to offer online services in addition to traditional face-to-face services by the service providers (Berry et al., 2002; Ostrom et al., 2015).

Based on this motivation, we focus on a cooperative service network which is formed by a group of independent cooperating service providers. Each server serves the stream of his own customers independently of the network structure at the service facility. The cooperative network enables the servers to access online external customers who request for the remote service by using the network online platform. A telemedicine platform is a good example of cooperative service networks. It is a group of cooperating healthcare providers who deliver healthcare services over a distance through an easy to use on-line network using the means of information technology (Roine et al., 2001; Norris, 2002; Craig and Petterson, 2005; Whitten et al., 2010). It is crucial to design these platforms in a way to provide profitable and efficient service for the customers (Körpeoğlu et al., 2014). Based on the operational parameters of the service providers, designers should decide on the characteristics of the servers to be included in the cooperative network as well as the optimal number of participants. These decisions are dependent on the servers' admission policy to serve their own customers and external customers. Moreover, the network needs to decide about the service price to pay the members for serving external customers. This price should be high enough to convince the servers to participate in the network while it should result in a profitable network operation with respect to the market price set externally for online services. In order to answer these questions about designing and managing a cooperative service network, we develop a stochastic model that captures the operation of the network with the dynamics of customer arrivals, assignment, and admission control. We use this model to determine the optimal number of servers and the service price to be offered to the service providers in order to maximize the expected profit while achieving an adequate service level for the external customers. We then make general observations about the effects of the servers' utilization rate, the external customers' arrival rate, and the service price on the optimal decisions. We show that in an optimally designed cooperative network, the members benefit from pooling their excess capacity to serve external customers and hence

increasing their utilization and obtained profit.

There are a large number of studies focused on the cooperation-based business models for service networks by pooling the individual streams of customers. Our paper differs from the other ones in the literature since we consider a type of cooperative operation that allows the members to receive and serve their own customers independent of the cooperation structure. This assumption is consistent with the industrial practice for the service providers who wish to add another channel to offer a service in addition to their previous mode of serving customers. The service providers in the proposed network decide about their availability for the external customers. In response to the servers' admission policy, the network defines the optimal number of members and the service price for the external customers. Different from the other works that focus on the profit of the whole service network, we discuss profit of each member at the first step. Accordingly, we establish a relatively simple sufficient condition for the service price to identify whether participation in the network is economically feasible for the servers. We then turn into the whole network's profit to define the optimal design.

The main contribution of this work is two-fold. First, we provide a detailed operational model of a cooperative network among a number of independent servers with embedded optimal dynamic admission decisions. This operational model is then used for determining the optimal design of the network. To the best of our knowledge, this is the first study that addresses the optimal design of a cooperative network based on the analytical analysis of the structure with the optimal operational decisions . Second, we extend the literature on the analysis of queueing systems with non-preemptive service order and heterogeneous customer streams. We prove that the optimal dynamic admission policy is a non-preemptive threshold-type policy in these kind of systems.

The organization of the remaining parts of this paper is as follows: In Section 2, we review the related literature. We present the model, its assumptions and also explain the two main stages of our analysis in Section 3. In Section 4, we give the dynamic programming formulation for each server's operation and derive the optimal admission policy for the customers. Then, we perform the stationary analysis of this operation using a queueing model and obtain the stationary optimal admission policy for the service providers. Based on our results from Section 4, we analyze the operation of the network in Section 5 and determine the optimal number of servers to achieve the desired performance of the network. Our numerical results and discussion regarding the efficient design of the network are given in Section 6. Finally, Section 7 is devoted to concluding remarks.

## 2. Literature review

Using a cooperative network structure to offer services is gaining acceptance as a promising approach to utilize the resources more effectively and achieve cost efficiency. Hence, dynamic resource allocation and admission control policies in network structures are studied in different settings such as communication, video services, radio access and manufacturing systems (Niu et al., 2016; Bagci and Tekalp, 2018; Buyakar et al., 2020; Mourtzis et al., 2020; Feng et al., 2020).

From an operational point of view, our study is related to two streams of the literature in queueing systems: cooperative networks, and dynamic resource allocation and admission decisions in service systems.

### *2.1. Cooperative networks*

The papers related to the cooperation in the service operations are mainly focused on the queueing systems with server pooling. One of the earliest works in this area is presented by Stidham (1970), who considered a design problem to define the optimal number of parallel servers and the service rate of each server to minimize the service and waiting cost in the system. Later on Benjaafar (1995), Buzacott (1996) and Mandelbaum and Reiman (1998) studied pooled systems and discussed the effectiveness of several pooling scenarios. There is also another stream of researches that focus on pooling, capacity sharing and cost allocation decisions among a number of companies (González and Herrero, 2004; García-Sanz et al., 2008; Anily and Haviv, 2010) by applying cooperative game theory. Yu et al. (2015) extended these studies by considering the level of information exchange about the servers' parameters. Similar studies are given by Karsten et al. (2015), Anily and Haviv (2017), Zeng et al. (2018), Bendel and Haviv (2018) and Liu et al. (2021).

In the studies reviewed above, the servers agree to operate as a common server for all arriving customers. However, in our setting, the cooperative network provides another channel for the servers to receive additional customers. Hence, the members need to manage their service policy not only for the network customers but also for their own customers independently from the others. Different from these studies, we focus on the operational decisions for each independent server first and then use the obtained results to discuss about the network's optimal design. Our results enable us to address the optimal design of the cooperative network based on the number of service providers, their utilization level, the customers' arrival rates and the service price.

Table 1 shows the summary of the reviewed literature related to cooperative networks and the contribution of this study.

Table 1: Summary table for the reviewed studies related to cooperative networks

| Study | Server pooling | capacity sharing (Independent firms) | cooperate to serve all customers | cooperate to serve additional customers | Objective function | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Maximize total profit | Minimize total cost | Performance evaluation | Optimal cost allocation (cost sharing game) |
| Stidham (1970) | ✓ | | ✓ | | | ✓ | | |
| Benjaafar (1995) | ✓ | | ✓ | | | | ✓ | |
| Buzacott (1996) | ✓ | | ✓ | | | | ✓ | |
| Mandelbaum and Reiman (1998) | ✓ | | ✓ | | | | ✓ | |
| González and Herrero (2004) | | ✓ | ✓ | | | | | ✓ |
| García-Sanz et al. (2008) | | ✓ | ✓ | | | | | ✓ |
| Anily and Haviv (2010) | | ✓ | ✓ | | | | | ✓ |
| Yu et al. (2015) | | ✓ | ✓ | | | ✓ | | |
| Karsten et al. (2015) | | ✓ | ✓ | | | | | ✓ |
| Anily and Haviv (2017) | | ✓ | ✓ | | | | | ✓ |
| Zeng et al. (2018) | | ✓ | ✓ | | | | | ✓ |
| Liu et al. (2021) | | ✓ | ✓ | | | ✓ | | |
| This paper | | ✓ | | ✓ | ✓ | | | |

### *2.2. Dynamic resource allocation and admission decisions in service systems*

One of the important operational issues in service operations management is the resource allocation between heterogeneous customer streams and the admission decisions that should be made upon each arrival and service completions. In our model, determining the individual server's admission policy is crucial for designing the cooperative network. In one of the earliest studies, Harrison (1975) showed that the $r\mu$ rule is optimal for scheduling a single server queue with two classes of jobs characterized by different service rates $\mu_i$, and rewards, $r_i$. Later on Örmeci et al. (2001) consider the dynamic admission control problem in a Markovian loss queueing system with two classes of customers and show the optimality of a threshold type admission policy in this structure. This model is also analyzed under different kinds of assumptions regarding the arrival process. For example, Örmeci and Burnetas (2004) show the optimality of sequential threshold policy in the system receiving random batches and Örmeci and van der Wal (2006) establish the existence of the optimal acceptance thresholds for all job classes in the system with general inter-arrival times.

All these studies formulate the dynamic programming for the admission problem based on the assumption of having a preemptive service discipline. There are also a number of studies that use preemption as a control tool (Brouns and Van Der Wal, 2006; Ulukus et al., 2011; Turhan et al., 2012). However, the main difference of our work with the other studies in the literature is that the operation of each server is defined as a non-preemptive process. Moreover, we have different settings for different class of customers to join the system: the servers' own customers join a queue and wait for their admission time whereas the external customers are assigned to one of the available servers without waiting in the queue. There are a few studies dealing with non-preemptive models. Iravani et al. (2007) consider the optimal single-server scheduling in a non-preemptive finite-population queueing system with heterogeneous customers. They show that the optimal service policy is a simple static priority policy for those classes that are served. Bispo (2013) study a single server scheduling problem with the non-preemptive service

discipline, different classes of customers, and convex costs depending on the arrival rates. He establishes a policy to reach near optimal performances and presents the policy as a function of individual loads in the system. Zhao and Wang (2009) investigate the performance of a non-preemptive $M/M/1$ queueing system with two priority classes, and derive the probability that the server is busy or idle in the system.

In this study, we contribute to the admission control literature by extending the theoretical results of the dynamic admission control problem with non-preemptive service and two groups of customers joining in two different ways. We change the standard formulation of the resulting Markov decision process to accommodate our model where the server's state needs to be captured in the overall state description. We establish results on the structure of the optimal admission policy for this problem and give an approximate analysis of the cooperative network accordingly.

Table 2 shows the summary of the reviewed literature related to dynamic resource allocation and admission control in service systems and the contribution of this study.

Table 2: Summary table for the reviewed studies related to dynamic resource allocation and admission control in service systems

| Study | Single server | Server pooling | cooperate to serve all customers | Model assumptions | | Objective function | |
|---|---|---|---|---|---|---|---|
| | | | | preemptive | non-preemptive | admission control | termination control |
| Örmeci et al. (2001) | | ✓ | ✓ | ✓ | | ✓ | |
| Örmeci and Burnetas (2004) | | ✓ | ✓ | ✓ | | ✓ | |
| Örmeci and van der Wal (2006) | | ✓ | ✓ | ✓ | | ✓ | |
| Brouns and Van Der Wal (2006) | ✓ | | | ✓ | | ✓ | ✓ |
| Ulukus et al. (2011) | | ✓ | ✓ | ✓ | | ✓ | ✓ |
| Iravani et al. (2007) | ✓ | | | ✓ | | ✓ | |
| Bispo (2013) | ✓ | | | ✓ | ✓ | ✓ | |
| This paper | ✓ | | | | ✓ | ✓ | |

## 3. Model description

In this study, we consider a cooperative service network which connects on-line customers to a network of servers who serve their own customers and also the customers sent through the online network. Figure 1 depicts the network along with the servers participating in the network and receiving both groups of customers.

**Network**. The network operates with $N$ homogeneous servers. The online customers of the network, referred as the *external customers* place their requests on the network according to a Poisson process with the rate of $\lambda_t$ and pay $p_t$ for receiving an on-line consultation to the network. The price $p_t$ is determined externally based on the market conditions. That is, the network cannot dictate $p_t$ to its customers due to the competitive environment it operates in. The network pays $p_p$ to the server who serves the arrived external customer. Hence, the network earns $p_t - p_p$ from each customer who receives an on-line service. We assume that $p_t$ is

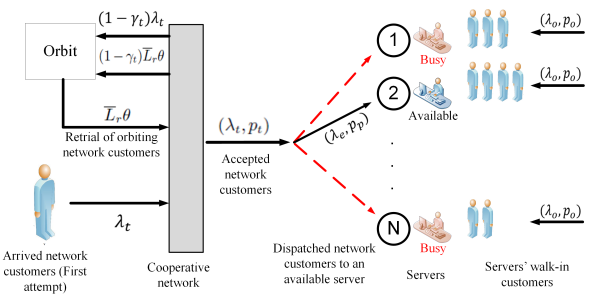


Figure 1: The cooperative service network

a given benchmark price for the online service in the market and the $p_p$ is set by the network. The price that will be paid to the servers must be less than $p_t$, $p_t > p_p$ to form an economically feasible cooperative network.

We assume that the network incurs a cost of $c$ for each server working in the cooperative network to cover operational and technological costs. When a request from an external customer is received, the customer is dispatched to one of the available servers who is available to accept the external customer at the time of arrival and the service starts immediately. Due to the non-preemptive feature of the service process, the servers can be available for the network only when they are idle. If more than one server is available at the time of arrival of an external customer, the customer is dispatched to one of the available servers randomly with equal probability. If none of the servers are available to admit an external customer when the request is received, then the request is unsatisfied for this time and the customer goes in to the pool of unsatisfied customers referred as the orbit (Keilson et al., 1968). An orbiting external customer repeats her request repeatedly until she is accepted by the network and assigned to an available server. The network incurs a cost of $c_r$ for each unsuccessful attempt of an external customer. The total amount of this cost is defined by $c_t$ and referred as *rejection cost* in this study. The inter-retrial times of the orbiting external customers are assumed to be exponentially distributed with the rate of $\theta$ and the average number of orbiting external customers in the system is defined by $\overline{L}_r$.

The long run proportion of the external customers whose requests are accepted by the network at each single attempt is defined by $\gamma_t$. Therefore, unsatisfied external customers enter to the orbit with the rate of $(1-\gamma_t)\lambda_t$. We define $\gamma_t$ as the service level offered by the network to its customers. Although, the network does not lose any customer, $\gamma_t$ is interpreted as the network service level to show the performance of the network setting. Offering a high service level to the external customers decreases the number of unsatisfied customers and hence provides a desired service experience for this group of customers.

The effective arrival rate of the external customers to each server in the network is defined by $\lambda_e$. The servers' admission policy and the number of servers in the network are the main factors to determine the effective arrival rate of the external customer $\lambda_e$.

**Service providers**. The servers in the network serve their own customers in addition to the external customers and they do not have information about the other servers in the network, the total number of servers and the network operating strategy.

The server's own customers are served in person and referred as the *own customers* in this study. Each server receives own customers independently of the network and according to a Poisson process with the arrival rate of $\lambda_o$. When a server's own customer arrives, the customer joins a queue and wait to be served by the server. The servers' own customers pay $p_o$ for the service and each server incurs a holding cost of $h$ for each of her own customers waiting in the queue per unit time.

Each idle server uses an admission policy to decide whether to start serving a waiting own customer from the queue or stay available to receive an arriving external customer. The optimal admission policy for a server to admit her waiting own customer or an external customer or stay idle for the future customer arrivals depending on the state of the system is denoted with $u$. We give the full description about $u$ in Section 4.

The servers' own customers are served by the order of their arrival to the queue and the average number of the customers waiting in the queue is denoted by $\overline{L}_Q$. The average waiting time of the servers' own customers in the queue is also defined by $\overline{W}_Q$ in this study.

The service times for the servers' own customers and also for the external customers are independently and exponentially distributed with the mean of $1/\mu$. The utilization of the servers due to their own customers is defined by $\rho = \lambda_o/\mu$. Their utilization due to both groups of customers is defined by $\xi$. Once a server starts to serve a customer, the service will not be preempted, and upon completion of a service, the customer being served leaves the system.

The service level provided to the external customers by a server in the cooperative network is the long run proportion of the time that a server is available to admit the external customers and denoted by $\gamma_p$. Therefore, the service level offered by the network to the external customers ($\gamma_t$) is dependent on the value of $\gamma_p$.

### *3.1. The network's and the servers' decision problems*

The two main decision problems in this study are deciding on the number of servers $N$ and determining the price to be paid to the servers $p_p$, depending on the system parameters. In order to examine this problem, we need to analyze the operation of the network model. Hence, the first step is to derive the servers' optimal admission policy to serve their own customers and the external customers.

**The network's optimization problem**. The cooperative network makes a profit of $p_t - p_p$ from each arriving customer who is served by an available server in the network. The orbiting proportion of the external customers retry their requests until they are accepted and hence the network does not lose any external customer and make profit from serving all of them. On

the other hand, the network incurs a total rejection cost due to unsatisfied customers requests. Rejection happens with the probability of $1 - \gamma_t$ at each trial of an external customer and the value of $\gamma_t$ is dependent on two decision factors. The first one is the servers' optimal admission policy ($u$), which determines the availability of the servers for the network, and the second one is the number of servers in the cooperative network ($N$). Therefore, the total rejection cost is also a function of $u$ and $N$, and defined by $c_t(u, N)$. Moreover, the network incurs a total cost of $cN$ for operating $N$ servers in the setting. Thus, the network's decision problem is deciding on $N$ and $p_p$ to maximize its average profit $\Pi_t$:

$$\Pi_t = \lambda_t(p_t - p_p) - c_t(u, N) - cN, \tag{1}$$

and

$$\Pi_t^* = \max_{N, p_p} \quad \Pi_t \tag{2}$$

The optimal decision factors, $N$ and $p_p$ are determined by solving the optimization problem given in Equation (2). While the optimal value of $p_p$ should be less than the market benchmark price $p_t$ to form a profitable network, it should be also high enough to convince the members to participate in the network. If the price that needs to be paid to the servers to convince them to join the network ($p_p$) is determined to be greater than the market price $p_t$ ($p_p > p_t$), then the network cannot operate in a feasible way.

Since the participating servers in the network are homogeneous and they do not have any information about the network, there is no competition between the servers to serve the external customers and they use the same optimal admission policy. The network has full information regarding the participants and solve its decision problem by assuming that the servers in the network use the optimal admission policy to serve both groups of customers.

**The server's decision problem**. Each participating server in the network receives a payment from her own customers and also from the external customers who are admitted and served. The servers also incur a waiting cost due to keeping their own customers waiting in the queue. Therefore, they should manage their own customers' admission and their availability to receive and serve the external customers in a way to maximize the average profit rate, $\Pi_D$:

$$\Pi_D = \lambda_o p_o + \lambda_e(u, N)\gamma_p p_p - h\overline{L}_Q(u). \tag{3}$$

In this equation, the first term is the revenue from an arrival of the server's own customers. The second term is the revenue from an admitted external customer. The external customers' revenue is generated only when the server is available, and hence the value of $\gamma_p$ is dependent on the admission policy, $u$. The state-dependent arrival rate of the external customers to each

server ($\lambda_e$) is a function of both $u$ and $N$ since it depends on the availabilities of the other servers in the network. A server will receive an arriving external customer surely if she is the only available server at the time of arrival. However, if there are many servers available at the requested time, her chance of receiving the external customer will be lower. Since the servers do not have any information regarding the network and other servers, they will consider $\lambda_e$ as an average arriving rate of the external customers that they can receive at their available time, and solve their optimal admission problem accordingly. In Section 4.1, we will capture the dependency of $\lambda_e$ on $u$ and $N$ from the network's perspective who has full information regarding the system. Finally, the last term of Equation (3) depicts the waiting cost of the server's own customers in the queue. $\overline{L}_Q$ is also dependent on the server's decision about the way of serving their own customers.

Therefore, the servers should decide about the optimal admission policy $u$ to maximize the average profit $\Pi_D$,

$$\Pi_D^* = \max_u \quad \Pi_D. \tag{4}$$

The servers' optimal admission policy of serving their own and the external customers must be determined in order to address the network's decision problem. As a result, the decisions of the network designers and the servers must be solved jointly.

In order to simplify the notation in the remaining part, we use $\gamma_p$, $c_t$ and $\overline{L}_Q$ instead of $\gamma_p(u)$, $c_t(u, N)$ and $\overline{L}_Q(u)$ respectively.

## 4. The servers' optimal admission policy

We first consider the servers' optimal admission policy to serve two streams of customers arriving with exponentially distributed inter-arrival times. Each server's arriving own customers join to a queue and wait for the server's admission to get an in-person service. However, an arriving external customer is assigned to the server only if the server is in the available state for the network. The service times for both groups of customers are also exponentially distributed.

Since this system evolves as a continuous-time Markov chain, each server in the network should decide on the admission policy depending on the current state and not on the prior history. Hence, we employ a Markov decision process (MDP) framework to find the optimal admission control policy. The objective is to obtain a dynamic scheduling policy for each server to maximize the discounted profit with discount parameter $\beta$ in an infinite time horizon as well as the long-run average profit.

We define the state of the system as $S(t) = (x(t), j(t))$, where $x(t)$ is the number of the server's customers, including the server's own customers and the external customers present in the system at time $t$, and $j(t) \in \{0, 1\}$ is the server's state at time $t$. If $j(t) = 0$, the server is

idle, and $j(t) = 1$ means that the server is busy by serving a customer. This state description captures the server state, which is needed for modeling of a non-preemptive system. The only decision points for the server are the arrival times of the server's own customers when the server is idle ($j(t) = 0$), and the departure times of the served customers when the server is busy ($j(t) = 1$).

We use the standard tools of uniformization to convert the continuous time Markov process into a discrete-time equivalent. The uniformization rate is defined by $\delta = \mu + \lambda_o + \lambda_e + \beta$. Without loss of generality, we normalized $\delta$ and set $\delta = 1$.

Let $v(x, j)$ be the maximum expected $\beta$-discounted reward of the system over the infinite time horizon and $v_n(x, j)$ be the maximum expected $\beta$-discounted reward of the system over a finite time horizon starting in state $(x, j)$ when $n$ observation points remain in the horizon. There exists an optimal stationary policy for the infinite horizon problem and $v(x, j) = \lim_{n\to\infty} v_n(x, j)$ whenever $\beta > 0$ (Puterman (2014), chapter 6). Therefore, the structural properties of the system over a finite time horizon holds for the infinite time horizon as well.

The value iteration expressions for the dynamic scheduling of a server in the cooperative network over a finite number of transitions, $n$, can be written as follows:

$$\begin{aligned} v_{n+1}(x,0) =& \lambda_o \max\Big\{v_n(x+1,0) + p_o, v_n(x+1,1) + p_o\Big\} \\ &+ \lambda_e\Big[v_n(x+1,1) + p_p\Big] + \mu v_n(x,0) - hx \quad \forall x, \end{aligned} \tag{5a}$$

$$\begin{aligned} v_{n+1}(x,1) =& \lambda_o\Big[v_n(x+1,1) + p_o\Big] + \lambda_e v_n(x,1) \\ &+ \mu \max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} - h(x-1) \quad \forall x \geq 1. \end{aligned} \tag{5b}$$

The maximization term in Equation (5a) corresponds to an available server's decision when the server's own customer arrives at the queue. At this point, the server should decide whether to start serving the first customer in the queue or stay idle to receive and serve an external customer who may arrive later. According to the second term in the right-hand side of Equation (5a), the available server serves the arriving external customer immediately upon the arrival time. The third term shows that the service completion rate does not have any effect on the state of system since the server is idle. Finally, the last term in Equation (5a) shows the holding cost for the server's own customers waiting in the queue to receive a service.

According to Equation (5b), the server is not available to serve any customer, when she is busy. Thus, the only decision time is the completion of the customer's service, when the server should decide whether to serve her own customer from the queue or stay idle and wait for an external customer's arrival at a later time.

It can be argued that it should be less profitable for the servers to stay idle and accept

the external customers when they have a number of their own customers already in the queue waiting for the service and incurring a waiting cost. The optimality of the threshold policy for the admission control problem has been proven for different systems where the system state is based on the number of jobs in the system but not on the state of the server (Altman et al., 1998; Koole, 1998; Örmeci et al., 2001). We modify the related results for our model and show the optimality of a threshold policy for the proposed system in Theorem 4.1.

**Theorem 4.1.** *The optimal admission policy for the servers participating in the cooperative network is a threshold-type policy: there exists a threshold level $R$, such that it is optimal to stay idle, and serve the arriving external customers if the total number of customers already in the system is under the threshold level $R$. Otherwise, it is optimal for the server to serve her own customers, and reject the network requests.*

The proof is given in the Appendix

*4.1. Analysis of the system in the steady state*

All the results proven for a finite number of transition $n$ in previous section are also true as $n$ goes to infinity. Therefore, the corresponding conclusions are valid when the total expected discounted reward over an infinite horizon is maximized. Moreover, since the state space and the action space in each state are finite, the results also hold for $\beta = 0$; so, we have the same conclusions for the long-run average reward. In this section, we evaluate the average profit rate of the servers under the threshold policy.

The explicit expressions for the average profit can be obtained by analyzing the related queueing model. To obtain the queueing model for each server in the network, we define the state of the system as $(x_1, x_2)$, where $x_1$ indicates the number of server's own customers in the system and $x_2 \in \{0, 1\}$ shows whether there is an external customer in service ($x_2 = 1$) or not ($x_2 = 0$).

Figure 2 shows the state space diagram of the resulting queueing model for the server's operation. As depicted in Figure 2, each server in the network serves only the received network requests if the number of the waiting customers is less than the threshold level, i.e. $x_1 \leq R-1$. Consequently, the server starts to serve her own waiting customers from the queue whenever the number of customers reaches the threshold level.

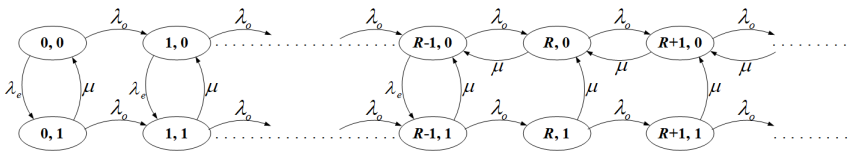


Figure 2: The state space diagram of queueing model for server's operation

In this case, an incoming external customer is served whenever the number of waiting customers in the queue is less than the threshold level $R$ and there is not any external customer already in the system. Analyzing the queueing model depicted in Figure 2 analytically, the service level offered by each server to the external customers is determined from the steady state probabilities as

$$\gamma_p = Pr(x_1 \leq R-1, x_2 = 0) = \frac{1-\rho}{1+\lambda_e/\mu} = \frac{\mu-\lambda_o}{\mu+\lambda_e}. \tag{6}$$

Equation (6) indicates that the steady state probability of accepting the external customers does not depend on the value of the threshold level $R$. The reason for this result is that all the states under the threshold value are transient states and the value of $\gamma_p$ is equal to $Pr(x_1 = R-1, x_2 = 0)$ and independent of all the states that the number of waiting customers is smaller than $R-1$ ($x_1 < R-1$).

Consequently, the first two terms in the average profit rate given in Equation (3) are independent of the threshold level $R$ while the last term, the average waiting cost of the customers waiting in the queue increases with $R$. As a result, setting $R$ as low as possible will maximize the average profit rate for the servers. This observation yields the following proposition on the non-preemptive priority policy with the proof given in the Appendix.

**Proposition 4.2.** *The optimal threshold level is equal to one, $R = 1$ and the servers are available to receive the network requests whenever there is not any customer in the queue. The stationary optimal admission policy for the servers in the network, $u$ is a non-preemptive-priority policy: the servers serve their own customers in a non-idling manner whenever there is at least one customer in the queue.*

Since it is equally likely for all the available servers to receive the arriving customer, we can also find the effective state dependent arrival rate of $\lambda_e$ which is dependent on the servers admission policy and the number of participants in the network:

$$\lambda_e = \frac{\lambda_t}{N\gamma_p}. \tag{7}$$

When the servers decide to work with the cooperative network and use the optimal admission policy to serve their customers, the average waiting time of their own customers in the queue is determined from the steady-state probabilities of the system in Figure 2 with $R = 1$:

$$\overline{W}_Q = \frac{\lambda_o + \lambda_e}{(\mu+\lambda_e)(\mu-\lambda_o)}. \tag{8}$$

Based on Equation (8), When $\lambda_e$ goes to infinity, the average waiting time of the own

customers becomes equal to $1/(\mu - \lambda_o)$. Hence, accepting the external customers will increase the server's own customers' average waiting time by $1/\mu$ at most.

We can also derive the servers' utilization due to the both network and their own customers as follows:

$$\xi = \frac{\lambda_o + \gamma_p \lambda_e}{\mu} = \frac{\lambda_o + \lambda_e}{\mu + \lambda_e}. \tag{9}$$

Therefore, the percentage of the server's utilization increase due to the external customers will be equal to $\frac{\lambda_e(1-\rho)}{\rho(\mu+\lambda_e)}$. When $\lambda_e$ goes to infinity, the percentage of the utilization increase becomes equal to $\frac{1-\rho}{\rho}$, which means that the server's utilization approaches 1 in this case.

## 5. Approximate analysis of the cooperative network under the optimal admission policy

Based on the result of the optimal admission policy for the servers in the network, we will analyze the performance of the network approximately in this section and discuss the optimal number of participants as well as the pricing strategy for this setting.

The network will reject an arriving request only if there is no available server in the system. In order to derive a closed-form expression for the network service level, we assume that the servers operate independently as an approximation. Therefore, the approximate value of the network service level, $\gamma_t$, is given by:

$$\gamma_t \cong 1 - (1 - \gamma_p)^N. \tag{10}$$

Accordingly, we derive the total rejection cost, $c_t$ as:

$$c_t = \sum_{k=1}^{\infty} (1 - \gamma_t)^k \lambda_t c_r = \frac{1 - \gamma_t}{\gamma_t} \lambda_t c_r, \tag{11}$$

where $k$ is the number of unsuccessful trials for the external customers before admission.

The average number of orbiting customers who are repeating their requests to enter the system is given by

$$\overline{L}_r = \frac{\lambda_t(1 - \gamma_t)}{\gamma_t \theta}. \tag{12}$$

As a result, the network's decision problem in Equation(2) is written as follows:

$$\Pi_t^* = \max_{N, p_p} \quad \lambda_t(p_t - p_p) - \frac{1 - \gamma_t}{\gamma_t} \lambda_t c_r - cN. \tag{13}$$

Our numerical experiments given in Section 6.1 show that this approximation is accurate to determine the optimal number of servers.

Next, we discuss the optimal pricing strategy along with the optimal number of servers to achieve the optimal profit and enhance a desired performance of the cooperative network.

### *5.1. Pricing*

From the network designers' perspective, the network should pay the minimum level that will give the servers the incentive to accept the network customers. Paying a higher price does not have any effect on the service level offered by the servers for the external customers while it decreases the network's profit. The servers will participate in the cooperative network only if the network brings them financial benefits compared to the case that they serve only their own customers independently. This means that they should compare the financial benefit of getting external customers with the increase in the waiting cost due to higher utilization associated with the external customers.

**Proposition 5.1.** *Participating in the cooperative network is economically feasible for the servers if and only if:*

$$p_p > h\frac{\rho}{\mu(1-\rho)}. \tag{14}$$

The proof is given in the Appendix.

Proposition 5.1 states that each external customer must bring a revenue higher than $h\frac{\rho}{\mu(1-\rho)}$ for each member to work in the cooperative network and accept network's on-line customers. Note that this lower bound is independent of the network and it is fully defined by the server's operational parameters. The servers expect a higher value of $p_p$ when their utilization level $\rho$ is higher. Moreover, the network should offer a higher price of $p_p$ if the customers' service time is higher.

The inequality in (14) defines the condition to form an economically feasible network. Considering the value of $p_t$, for a group of servers if $h\frac{\rho}{\mu(1-\rho)} > p_t$, it is infeasible to form a cooperative service network with this group of servers.

### *5.2. Deciding on the number of participating servers*

In this section, we discuss the network design problem in Equation (13) to define the optimal number of servers in the cooperative network leading to the network and its members profitability.

According to Equations (6) and (7), increasing the number of participants in the network will result in a lower effective arrival rate of the external customers for each server and then the servers become more available to receive the external customers. Figure 3 shows the effect of the number of servers in the cooperative network on $\gamma_p$ and $\lambda_e$.

As a result of increasing $N$, the cooperative network will be able to provide a higher service level for its customers, $\gamma_t$, and decrease the rejection probability of the customers on each

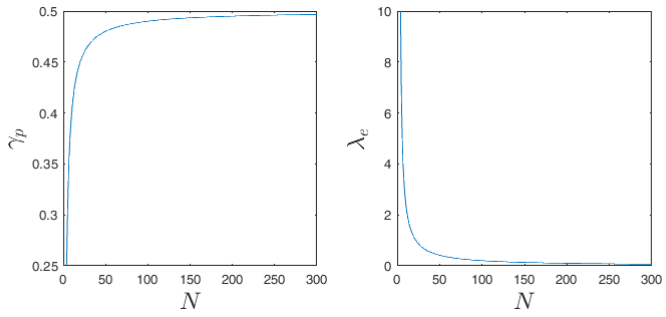


Figure 3: The offered service level for the external customers by each server in the network ($\gamma_p$) and the effective arrival rate of the external customers ($\lambda_e$). ($\rho = 0.5, \lambda_t = 10, \mu = 10$)

trial. All these changes lead to a lower rejection cost for the network based on Equation (11). However, there is a trade-off between the rejection cost, $c_t$, and the cost of operating servers in the cooperative network defined by $cN$ in the profit function. Achieving a balance between these two aspects of the profit function in Equation (13) leads to a unique optimal solution, $N^*$, in a way to give the maximum possible profit for the network.

Furthermore, the cost of $c_r$ determines the network's sensitivity regarding the unsatisfied customers since the higher level of $c_r$ causes a higher optimal solution, $N^*$, and hence a higher service level for the external customers. It means that by setting an adequate value of $c_r$, the network assures a desired service level for the external customers.

**Proposition 5.2.** *The network's maximization problem has a unique optimal solution for the number of participating servers in the network. In other words, there is a unique optimal number of participating servers in the cooperative network which assures the maximum total profit for the network and a desired service level for the external customers.*

The proof of this proposition includes showing the dependency depicted in Figure 3 formally and it is given in the Appendix.

## 6. Numerical experiments

The important parameters in determination of the optimal number of participants in the cooperative network are the utilization of the servers ($\rho$) and their service rate ($\mu$), the network's technological cost ($c$), the external customers' rejection cost ($c_r$) and their arrival rate to the network ($\lambda_t$). The network designers should consider all these parameters to achieve an efficient cooperative network with the maximum profit and the desired performance.

In this section, we are going to show the accuracy of our proposed model compared to the simulation and then discuss the effects of changing parameters on the optimal number of servers, the lower bound of the price $p_p$ and the feasibility of the network depending on the market price $p_t$.

### 6.1. Accuracy of the approximate model

Assuming that the servers operate independently from each other and receive the external customers with a Poisson arrival rate, may lead to errors in calculating $\gamma_p$, $\gamma_t$ and hence the total rejection cost $c_t$. In order to evaluate the accuracy of the approximate model, we compared our obtained numerical results with the simulation of the system. The simulation length is set to get results with the same number of significant digits as the approximate model.

We run the simulation for different cases with different utilization levels $\rho \in \{0.4, 0.5, 0.6, 0.7, 0.8\}$, and also different levels of network cost $c_r/c \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The optimal number of participants are determined by using the approximate model, $N^*$, and also by using simulation, $N_s^*$. Our obtained results with the approximate model show a high accuracy compared to the simulation results. Figure 4 shows that in 68% of the total 25 instances we considered, the model gives the same optimal number of participants as the simulation. In 28% of the instances the difference between the result of simulation and the approximate model is 1, and only in one of the instances where $\rho = 0.8$ and $c_r/c = 0.5$ the difference between the results is 2.

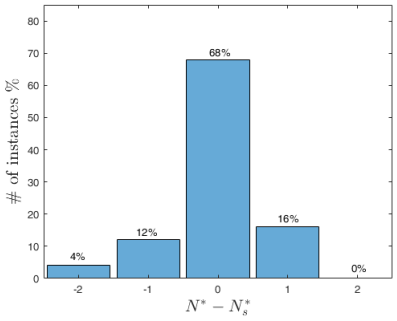


Figure 4: The difference between the optimal number of participants obtained by the approximate model ($N^*$) and simulation ($N_s^*$) in solving 25 different instances. ($\rho \in \{0.4, 0.5, 0.6, 0.7, 0.8\}, c_r/c \in \{0.1, 0.2, 0.3, 0.4, 0.5\}, \lambda_t = 20, p_t = 80, p_p = 20$)

### 6.2. Effect of the servers' utilization level on the optimal number of servers and the lower bound of $p_p$ in the cooperative network

The servers in a selected segment or in a selected geographic area can have similar utilization. The network chooses these servers depending on their utilization level and define the optimal number of participants accordingly. Figure 5 shows the optimal number of servers for different utilization levels. This figure is given for different values of $c_r$ to highlight the effect of rejection cost on the optimal number of participants and hence the network's offered service level, $\gamma_t$.

According to Figure 5, for the higher values of the rejection cost, the network may need to increase the number of participants to reduce the rejection probability and then reach the maximum possible profit. In other words, because the total rejection cost is larger for higher $c_r$, the network tries to decrease the rejection probability, thereby increases the offered service

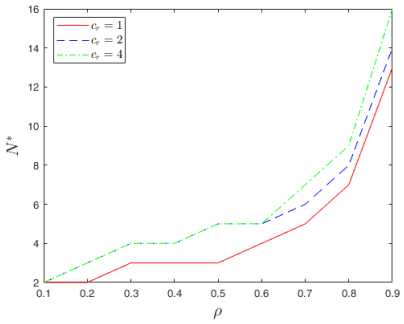


Figure 5: The optimal number of participating servers in the cooperative network ($N^*$) for different values of servers' utilization ($\rho$) and the external customers rejection cost ($c_r$). ($\mu = 10, \lambda_t = 10, p_t = 80, p_p = 20, c = 10$)

level. In summary, setting an adequate value of $c_r$ ensures the desired service level for the external customers. For instance, when $\rho = 0.7$, the network offered service level is equal to $\gamma_t = 0.4$, $\gamma_t = 0.57$ and $\gamma_t = 0.69$ respectively for $c_r = 1$, $c_r = 2$ and $c_r = 4$. On the other hand, the optimal number of participants is the same when $\rho$ is less than 0.6 for $c_r = 2$ and $c_r = 4$. The reason for this equality is that the technological cost of adding one more server (associated with $c$) is more than the increase of total rejection cost in these networks. Therefore, choosing the servers with higher utilization values leads to a higher optimal number of participants. This could be even more if the network wants to achieve a higher service level.

Figure 6 shows the lower bound of the price $p_p$ for different values of servers' utilization level. According to this figure, the network needs to pay a higher price to the servers with higher utilization level in order to convince them to cooperate within the network. Considering an arbitrarily set value of $p_t = 30$, all the servers with a higher utilization than 0.9 are infeasible to participate in the cooperative network since their required lower bound of $p_p$ is greater than the $p_t$ price. That is why, we defined the infeasible region by a dashed line where $p_p > p_t$.

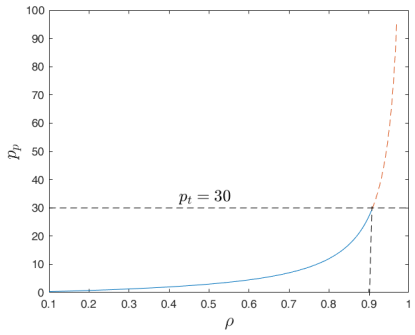


Figure 6: The lower bound of the price offered by the network to the servers for serving external customers ($p_p$) for different values of servers' utilization level($\rho$). ($\mu = 5, h = 15, p_t = 30$)

*6.3. Effect of the servers' service rate on the optimal number of servers and the lower bound of $p_p$ in the cooperative network*

The servers' service rate is another parameter in defining the optimal number of participants in the network and the lower bound for the price $p_p$. Figure 7 shows that the optimal number of servers in the network is decreasing in response to the higher values of servers' service rate and the rate of decreasing is higher for the smaller value of rejection cost. The reason for this effect, is that the rejection cost grows with a lower rate whenever the value of $c_r$ is smaller.

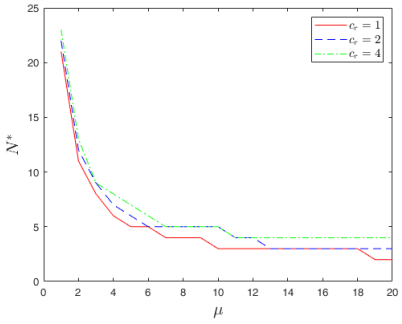


Figure 7: The optimal number of participating servers in the cooperative network ($N^*$) for different values of servers' service rate ($\mu$) and the external customers rejection cost ($c_r$). ($\rho = 0.5, \lambda_t = 10, p_t = 80, p_p = 20, c = 10$)

Similar to our discussion in 6.2, for higher $c_r$ the network tries to reduce the rejection probability and then the rejection cost by keeping $N$ at a higher level. Again, this happens only if the rejection cost is higher than the technological cost of having more servers. For example, when $10 \leq \mu \leq 18$ the optimal number of participants is the same and equal to 3 for $c_r = 1$. However, for $18 < \mu \leq 20$, the optimal number of participants decreases to 2 since the technological cost of having one more server is greater than the rejection cost associated by having one less server in the network. Note that, for $18 < \mu \leq 20$, when $c_r = 2$ and $c_r = 4$, the optimal number of participants remains the same as before due to higher rejection cost.

Figure 8 presents the lower bound of price $p_p$ for different values of the server's service rate. As it is mentioned in Section 5.1, The network should offer a higher $p_p$ to the servers when they have a lower service rate and need more time to serve the customers. Similar to Figure 6, we defined the infeasible part by a dashed line where the required value of $p_p$ is greater than $p_t$. It means that the network operation is infeasible with the members whose service rate is less than 1.17.

*6.4. Effects of the technological and rejection costs on the optimal number of servers in the network*

The technological and rejection costs in the cooperative network have important roles in defining the optimal number of servers to create a profitable network with a desired performance

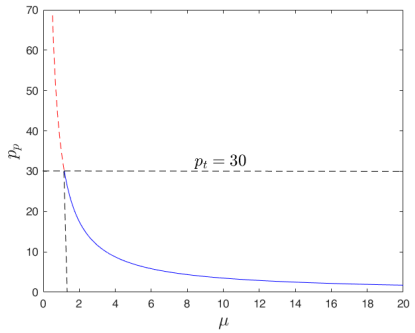


Figure 8: The lower bound of the price offered by the network to the servers for serving external customers ($p_p$) for different values of servers' service rate($\mu$). ($\rho = 0.7, h = 15, p_t = 30$)

level.

Figure 9 presents the effect of increasing the operational cost on the optimal number of participants for different values of rejection costs. According to this figure, in response to the higher values of the technological cost, the network will decrease the number of servers down to the points which gives the maximum possible profit as the result of the existing trade of between the technological cost and the rejection cost. In terms of having a comparison between the network performances in Figure 9, the network's offered service levels are equal to $\gamma_t = 0.68$, $\gamma_t = 0.83$ and $\gamma_t = 0.91$ for $c_r = 6$, $c_r = 12$ and $c_r = 18$ respectively, when $c/p_t = 0.16$.

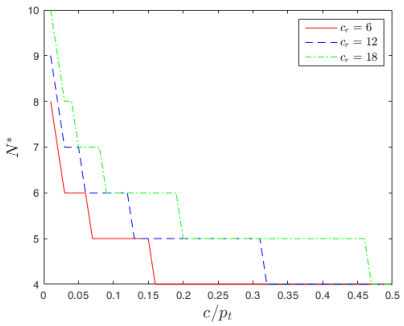


Figure 9: The optimal number of participating servers in the cooperative network ($N^*$) for different values of $c/p_t$ and the external customers rejection cost ($c_r$). ($\rho = 0.5, \mu = 10, \lambda_t = 10, p_t = 100, p_p = 30$)

Definitely, the value of $c_r$ is the main designing parameter for achieving the desired performance of the different kinds of networks with different characteristics regarding their members. Figure 10 shows the effect of the rejection cost on the networks with different utilization levels of the participants. Again, we can see that the network will try to choose higher number of servers when the rejection cost is higher and the increasing rate of $N$ is lower for the networks with low utilization participants

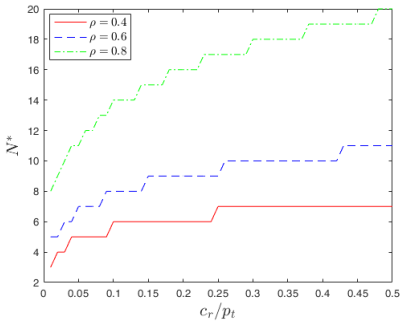


Figure 10: The optimal number of participating servers in the cooperative network ($N^*$) for different values of $c_r/p_t$ and the servers' utilization level ($\rho$). ($\mu = 10, \lambda_t = 10, p_t = 100, p_p = 30, c = 5$)

### *6.5. Effect of the external customers' traffic on the optimal number of servers in the network*

The main uncertainty in the network design is the customers' on-line request rate which may be subject to change due to many different reasons. Hence, it is essential for the designers to consider how the number of members in the network should be changed in response to an increase in $\lambda_t$ in order to have an efficient and profitable network. Figure 11 shows this effect for different networks containing the members with three different utilization levels.

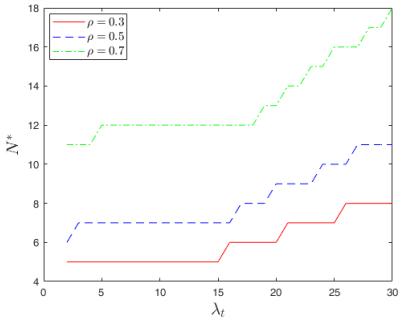


Figure 11: The optimal number of participating number of servers in the cooperative network ($N^*$) for different values of external customers request rate and for three different values of servers' utilization ($\mu = 10, p_t = 80, p_p = 20, c = 10, c_r = 5$)

According to Figure 11, whenever the members of the network have higher utilization, we should increase the number of participants with a higher rate to obtain the maximum profit. We can also consider the changes in the network offered service level ($\gamma_t$) as the result of change in the value of $\lambda_t$ in the cooperative networks with different numbers of participants.

Figure 12 shows that the small-size cooperative networks with lower number of members are more sensitive to the changes in the on-line request rate than the larger networks with a higher number of members. In summary, we observe that the networks with high utilization level of participating servers or lower number of members, show higher sensitivity to fluctuating external customers request rate, and their offered service level decreases with a higher rate in response to increasing on-line request rate.

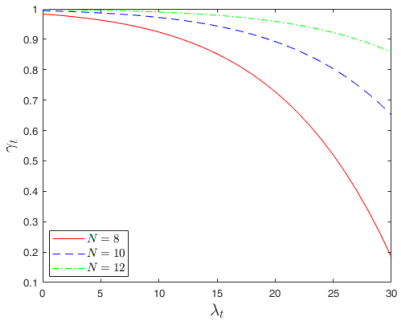


Figure 12: The network offered service level for different values of external customers request rate and for three different numbers of participating members, $N$ ($\mu = 10, \rho = 0.6, p_t = 80, p_p = 20, c = 10$)

## 7. Concluding remarks

In this paper, we present a detailed operational model for a cooperative service network among a number of independent servers and discussed the optimal design of the proposed setting.

The model developed in this study allows us to examine the relationship among important network designing factors which are the number of members, their utilization level, external customers arrival rate, service price and the operational and rejection costs. Based on these findings, we discuss the network design problem to find the optimal number of participants. We prove that there is a unique optimal number of participants in the network which assures the maximum total profit for the network and a desired service level for the external customers. Additionally, we derive a sufficient lower bound on the service price to show the economic feasibility of the network for each independent server. The obtained lower bound is increasing in the waiting cost, customers' arrival rate and the servers' utilization level while it is decreasing in the servers' service rate. Furthermore, we used this condition to define whether forming a cooperative network is feasible by considering the benchmark service price in the market, $p_t$.

According to the obtained results, the cooperative network increases the utilization of its members serving their own customers together with the external customers when it is designed and operated effectively. Our numerical results show the optimal number of servers in the cooperative network for different values of the rejection cost and the servers' utilization level. We also present the effect of changing external customers arrival rate on the offered service level and discuss this issue for the networks with different number of participants and members' utilization levels. Our results indicate that a cooperative network benefits the servers, as they enjoy both a higher utilization and a higher revenue. We also demonstrate that accepting the external customers will increase the servers' own customers waiting time, $1/\mu$ at the most, which does not affect their satisfaction.

This study shows that a cooperative service network is an effective way of utilizing the idle

capacity of the servers while providing an adequate service level for the external customers and increasing the profit for both the network and service providers.

The model investigated in this study can be extended to examine the network with heterogeneous servers possessing different operational parameters. Also, the service rates can be different for various groups of customers. This will affect the optimal admission policy and hence the optimal designing factors. The model can also be extended to consider a waiting queue for the external customers, and to derive the optimal dispatching policy for the network to improve the performance. Considering the utility functions, we can incorporate the preference of customers in to the proposed model as well. These are left for the future research.

## Appendix A. Structure of the optimal admission policy

In this section, we show that an optimal threshold policy indeed maximizes the expected $\beta$-discounted reward over a finite horizon using induction.

We first show that when a new customer enters the system, the expected discounted profit decreases:

$$v_n(x+1,1) \leq v_n(x,1) \quad \forall x \geq 1; \tag{A.1}$$

$$v_n(x+1,0) \leq v_n(x,0) \quad \forall x. \tag{A.2}$$

The Inequalities (A.1), (A.2) mean that an additional customer incurs a positive opportunity cost.

We then prove that the opportunity cost of having an additional customer when the server is idle is always greater than having an additional customer when the server is busy:

$$v_n(x,1) - v_n(x+1,1) \leq v_n(x,0) - v_n(x+1,0) \quad \forall x \geq 1. \tag{A.3}$$

Inequality (A.3) implies the optimality of threshold policy for each server in the cooperative network if $v_n(x,1)$ and $v_n(x,0)$ are concave and decreasing in $x$, i.e.

$$v_n(x+2,1) - v_n(x+1,1) \leq v_n(x+1,1) - v_n(x,1) \quad \forall x \geq 1, \tag{A.4}$$

$$v_n(x+2,0) - v_n(x+1,0) \leq v_n(x+1,0) - v_n(x,0) \quad \forall x. \tag{A.5}$$

*Appendix A.1. Proof of Theorem 4.1*

We use induction to prove the structural properties for all finite $n$. To start the induction, we set $v_0(x,0) = 0$ and $v_0(x,1) = 0$. For $n = 1$ all Inequalities (A.1), (A.2), (A.3), (A.4) and (A.5) are true. We assume that all these inequalities hold for $n$ and prove them for $n+1$. Hence,

1) For $n+1$, we can write Inequality A.1 as follows:

$$\begin{aligned}&\lambda_o\Big[v_n(x+2,1)+p_o\Big]+\lambda_e v_n(x+1,1)+\mu\max\Big\{v_n(x,0),v_n(x,1)\Big\}-h(x)\leq\\&\lambda_o\Big[v_n(x+1,1)+p_o\Big]+\lambda_e v_n(x,1)+\mu\max\Big\{v_n(x-1,0),v_n(x-1,1)\Big\}-h(x-1)\end{aligned}\tag{A.6}$$

The inequalities in the first two lines can easily be shown to hold and the last line is trivially true. Therefore, $v_n(x,1)$ is decreasing in $x$. The same argument also applies to show the $v_n(x,0)$ is decreasing in $x$ in Inequality (A.2).

2) For $n+1$, we can write Inequality (A.3) as follows:

$$v_{n+1}(x,1)-v_{n+1}(x+1,1)\leq v_{n+1}(x,0)-v_{n+1}(x+1,0).$$

According to Inequalities (5a) and (5b), we have:

$$\begin{aligned}&\lambda_o\Big[v_n(x+1,1)-v_n(x+2,1)\Big]+\lambda_e\Big[v_n(x,1)-v_n(x+1,1)\Big]\\&+\mu\Big[\max\Big\{v_n(x-1,0),v_n(x-1,1)\Big\}-\max\Big\{v_n(x,0),v_n(x,1)\Big\}\Big]\leq\\&\lambda_o\Big[\max\Big\{v_n(x+1,0),v_n(x+1,1)\Big\}-\max\Big\{v_n(x+2,0),v_n(x+2,1)\Big\}\Big]\\&+\lambda_e\Big[v_n(x+1,1)-v_n(x+2,1)\Big]+\mu\Big[v_n(x,0)-v_n(x+1,0)\Big].\end{aligned}\tag{A.7}$$

In order to show that Inequality (A.7) is correct, we must show that the inequality holds for all the components associated with the same multipliers on the both sides. For the multipliers of $\lambda_o$, we must show that:

$$v_n(x+1,1)-v_n(x+2,1)\leq\max\Big\{v_n(x+1,0),v_n(x+1,1)\Big\}-\max\Big\{v_n(x+2,0),v_n(x+2,1)\Big\}.\tag{A.8}$$

Case 1: If $x+1\geq R$, then $\max\Big\{v_n(x+1,0),v_n(x+1,1)\Big\}=v_n(x+1,1)$ and $\max\Big\{v_n(x+2,0),v_n(x+2,1)\Big\}=v_n(x+2,1)$, so Inequality (A.8) is true.

Case 2: If $x+2<R$, then $\max\Big\{v_n(x+1,0),v_n(x+1,1)\Big\}=v_n(x+1,0)$ and $\max\Big\{v_n(x+2,0),v_n(x+2,1)\Big\}=v_n(x+2,0)$, so Inequality (A.8) is true due to our assumption in Inequality (A.3).

Case 3: If $x+2=R$, then $\max\Big\{v_n(x+1,0),v_n(x+1,1)\Big\}=v_n(x+1,0)$ and $\max\Big\{v_n(x+2,0),v_n(x+2,1)\Big\}=v_n(x+2,1)$, so Inequality (A.8) is true due to our assumption in Inequality (A.3).

For the multipliers of $\lambda_e$, we must show that:

$$v_n(x,1) - v_n(x+1,1) \leq v_n(x+1,1) - v_n(x+2,1). \tag{A.9}$$

Inequality (A.9) is true due to our assumption in Inequality (A.4) which implies that $v_n(x,1)$ is concave.

For the multipliers of $\mu$, we must show that:

$$\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} - \max\Big\{v_n(x,0), v_n(x,1)\Big\} \leq v_n(x,0) - v_n(x+1,0). \tag{A.10}$$

Case 1: If $x < R$, then $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,0)$ and $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,0)$, so Inequality (A.10) is true due to our assumption in Inequality (A.5) which implies that $v_n(x,0)$ is concave.

Case 2: If $x-1 > R$, then $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,1)$ and $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,1)$, so Inequality (A.10) is true due to our assumption in Inequalities (A.3) and (A.5).

Case 3: If $x = R$, then $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,0)$ and $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,1)$, so Inequality (A.10) is true due to our assumption in Inequality (A.5) and knowing that $v_n(x,1) \geq v_n(x,0)$.

3) For $n+1$, we can write Inequality (A.4) as follows:

$$v_n(x+2,1) - v_n(x+1,1) \leq v_n(x+1,1) - v_n(x,1)$$

and we have:

$$\begin{aligned}
&\lambda_o\Big[v_n(x+3,1) - v_n(x+2,1)\Big] + \lambda_e\Big[v_n(x+2,1) - v_n(x+1,1)\Big] \\
&+\mu\Big[\max\Big\{v_n(x+1,1), v_n(x+1,0)\Big\} - \max\Big\{v_n(x,0), v_n(x,1)\Big\}\Big] \leq \\
&\lambda_o\Big[v_n(x+2,1) - v_n(x+1,1)\Big] + \lambda_e\Big[v_n(x+1,1) - v_n(x,1)\Big] \\
&+\mu\Big[\max\Big\{v_n(x,1), v_n(x,0)\Big\} - \max\Big\{v_n(x-1,1), v_n(x-1,0)\Big\}\Big]
\end{aligned} \tag{A.11}$$

In order to show that Inequality (A.11) is correct, we must show that the inequality holds for all the components associated with the same multipliers on the both sides. For the multipliers of $\lambda_o$, we must show that:

$$v_n(x+3,1) - v_n(x+2,1) \leq v_n(x+2,1) - v_n(x+1,1). \tag{A.12}$$

Inequality (A.12) is true due to our assumption in Inequality (A.4) which implies that $v_n(x,1)$

is concave.

For the multipliers of $\lambda_e$, we must show that:

$$v_n(x+2,1) - v_n(x+1,1) \leq v_n(x+1,1) - v_n(x,1). \tag{A.13}$$

Inequality (A.13) is also true due to our assumption in Inequality (A.4).

For the multipliers of $\mu$, we must show that:

$$\begin{aligned} &\max\Big\{v_n(x+1,1), v_n(x+1,0)\Big\} - \max\Big\{v_n(x,0), v_n(x,1)\Big\} \\ &\leq \max\Big\{v_n(x,1), v_n(x,0)\Big\} - \max\Big\{v_n(x-1,1), v_n(x-1,0)\Big\}. \end{aligned} \tag{A.14}$$

Case 1: If $x+1 < R$, then $\max\Big\{v_n(x+1,1), v_n(x+1,0)\Big\} = v_n(x+1,0)$, $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,0)$ and $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,0)$ so Inequality (A.14) is true due to our assumption in Inequality (A.5) which implies that $v_n(x,0)$ is concave.

Case 2: If $x-1 > R$, then $\max\Big\{v_n(x+1,1), v_n(x+1,0)\Big\} = v_n(x+1,1)$, $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,1)$ and $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,1)$ so Inequality (A.14) is true due to our assumption in Inequality (A.4).

Case 3: If $x = R$, then $\max\Big\{v_n(x+1,1), v_n(x+1,0)\Big\} = v_n(x+1,1)$, $\max\Big\{v_n(x,0), v_n(x,1)\Big\} = v_n(x,1)$ and $\max\Big\{v_n(x-1,0), v_n(x-1,1)\Big\} = v_n(x-1,0)$. Therefore, we should show that the following inequality holds in this case:

$$v_n(x+1,1) - v_n(x,1) \leq v_n(x,1) - v_n(x-1,0). \tag{A.15}$$

According to the concavity conditions for $v_n(x,0)$ and $v_n(x,1)$, we can write:

$$v_n(x+1,1) + v_n(x+1,0) - v_n(x,1) - v_n(x,0) \leq v_n(x,1) + v_n(x,0) - v_n(x-1,1) - v_n(x-1,0). \tag{A.16}$$

On the other hand, we know that the following inequality is true because $v_n(x-1,0) \geq v_n(x-1,1)$.

$$v_n(x+1,0) - v_n(x,0) \leq v_n(x,0) - v_n(x-1,1). \tag{A.17}$$

From Inequalities (A.16) and (A.17) we can conclude that Inequality (A.15) is true because both $v_n(x,0)$ and $v_n(x,1)$ are decreasing functions.

The same argument also applies to show the concavity of $v_n(x,0)$ and the same proof procedure can be used to show that Inequality (A.5) is true.

Based on our discussion in Section 4, the structural properties obtained for $v_n(x,j)$ hold for $v(x,j)$, too. In order to address the long-run average profit, we use the result in Weber and

Stidham (1987) which shows that under certain conditions, that are valid for our model, the average reward could be obtained as the limit of the value function in the discounted problem when the discounted factor, $\beta$ goes to zero. Furthermore, in this setting, the average reward problem possesses the identical structural properties as the discounted cost problem (Weber and Stidham, 1987).

## Appendix B. Proof of Proposition 4.2

Using the steady-state queue length probabilities, we can write the profit rate function in Equation (3) as follows:

$$\begin{aligned} \Pi_D = &\quad \lambda_o p_o \quad + \lambda_e Pr(x_1 < R, x_2 = 0) p_p - h\overline{L}_Q. \\ = &\quad \lambda_o p_o \quad + \lambda_e \frac{1-\rho}{1+\lambda_e/\mu} p_p - h\overline{L}_Q. \end{aligned} \tag{B.1}$$

Based on the queueing model with the threshold level $R$ in Figure 2, we can obtain the steady state probabilities of the system as follows:

$$Pr(x_1 = l, x_2 = 0) = \Big(\frac{1-\rho}{1+\lambda_e/\mu}\Big)\Big(\frac{\rho^{l-R}(\lambda_o+\lambda_e)}{\mu} - \frac{\lambda_e \lambda_o^{l-R}}{(\mu+\lambda_o)^{l-R+1}}\Big) \qquad \forall l > R-1. \tag{B.2}$$

$$Pr(x_1 = l, x_2 = 1) = \Big(\frac{1-\rho}{1+\lambda_e/\mu}\Big)\Big(\frac{\lambda_e \lambda_o^{l-R+1}}{(\mu+\lambda_o)^{l-R+2}}\Big) \qquad \forall l \geq R-1. \tag{B.3}$$

Therefore, the value of $\overline{L}_Q$ is equal to:

$$\overline{L}_Q = \frac{(R-1)(\mu-\lambda_o) + (\lambda_o+\lambda_e)(\mu(R-1)+\lambda_o(2-R))}{(\mu+\lambda_e)(\mu-\lambda_o)}. \tag{B.4}$$

Due to the independency of Equation (6) from the value of $R$, we should minimize the average number of server's own customers waiting in the queue ($\overline{L}_Q$) to maximize the average long run profit obtained by each server in the network given in Equation (3). Since the value of $\overline{L}_Q$ is increasing in the value of $R$, the optimal threshold level to maximize the long run profit for each member in the network is equal to 1. Setting $R = 1$ maximizes the profit and leads to the following formulation for $\overline{L}_Q$:

$$\overline{L}_Q = \frac{\lambda_o(\rho + \lambda_e/\mu)}{(\mu+\lambda_e)(1-\rho)}, \tag{B.5}$$

and the value of $\gamma_p$ becomes equal to $Pr(x_1 = 0, x_2 = 0)$.

## Appendix C. Proof of Proposition 5.1

Regarding the independent operation of the servers with just their own customers, we face an $M/M/1$ queueing model where the customers join the queue with the rate of $\lambda_o$, pay the

price of $p_o$, and get the service with the rate of $\mu$. Hence, we can derive the minimum value of the price $p_p$ by ensuring that the long run average profit for each member of the network in Equation (B.1) is always greater than or at least equal to the resulting long run average profit of the server's independent operation:

$$\lambda_o p_o + \lambda_e \frac{1-\rho}{1+\lambda_e/\mu} p_p - h\frac{\lambda_o(\rho+\lambda_e/\mu)}{\mu-\lambda_o+\lambda_e(1-\rho)} \geq \lambda_o p_o - h\frac{\lambda_o}{\mu-\lambda_o} \tag{C.1}$$

As the result, the servers accept to work in the cooperative network if and only if:

$$p_p \geq h\frac{\rho}{\mu-\lambda_o}. \tag{C.2}$$

## Appendix D. Proof of Proposition 5.2

According to Equations (10) and (13) and for a defined value of $p_p$, we can write the network's optimization problem as:

$$\Pi_t^* = \max_N \quad \lambda_t(p_t - p_p) - \frac{(1-\gamma_p)^N}{1-(1-\gamma_p)^N}\lambda_t c_r - cN. \tag{D.1}$$

The revenue obtained from the external customers is independent of $N$ and thus the network profit maximization problem is the same as the costs minimization problem as follows:

$$\Pi_t^* = \min_N \quad \frac{(1-\gamma_p)^N}{1-(1-\gamma_p)^N}\lambda_t c_r + cN. \tag{D.2}$$

The objective function of the problem in (D.2) contains two parts. The first part is the rejection cost, $\frac{(1-\gamma_p)^N}{1-(1-\gamma_p)^N}\lambda_t c_r$, we define this part with $f$. The second part is the technological and maintenance cost for all the servers in the network, $(cN)$, we define this part with $g$ .

It is quite easy to see that $g$ is an increasing function of $N$ due to positive slope. We need to show that $f$ is a non-increasing function of $N$ and intersects $g$ in one point to prove the uniqueness of the optimal solution.

The only dependent part of function $f$ on the value of $N$ is $\frac{(1-\gamma_p)^N}{1-(1-\gamma_p)^N}$ and we assume that $a = 1-\gamma_p$. Let us consider the continuous form of this formulation and take the derivative of $\frac{a^N}{1-a^N}$ with respect to $N$:

$$\frac{\partial(\frac{a^N}{1-a^N})}{\partial N} = \frac{Na^{N-1}}{(1-a^N)^2}(\frac{\partial a}{\partial N}). \tag{D.3}$$

From Equations (6) and (7), we can write the value of $\gamma_p$ as follows:

$$\gamma_p = 1 - \rho - \frac{\lambda_t}{N\mu}, \tag{D.4}$$

and then:

$$\frac{\partial a}{\partial N} = \frac{\partial(\rho + \frac{\lambda_t}{N\mu})}{\partial N} = \frac{-\lambda_t}{N^2\mu}. \tag{D.5}$$

By substituting the value of $\frac{\partial a}{\partial N}$ in Equation (D.3), we can derive the value of $\frac{\partial f}{\partial N}$, which is always negative according to the following equation:

$$\frac{\partial f}{\partial N} = -c_r \frac{{\lambda_t}^2}{N\mu} \frac{(\rho + \frac{\lambda_t}{N\mu})^{N-1}}{(1 - (\rho + \frac{\lambda_t}{N\mu})^N)^2} \leq 0. \tag{D.6}$$

We show that $f$ is a non-increasing function for continuous values of $N$. Therefore, we can also say that $f$ is a non-increasing function for discrete values of $N$ and conclude that the network's maximization problem has a unique optimal solution, $N^*$.

## References


Altman, E., Jimenez, T., and Koole, G. (1998). On optimal call admission control. In *Decision and Control, 1998. Proceedings of the 37th IEEE Conference on Decision and Control*, volume 1, pages 569–574. IEEE.

Anily, S. and Haviv, M. (2010). Cooperation in service systems. *Operations Research*, 58(3):660–673.

Anily, S. and Haviv, M. (2017). Line balancing in parallel m/m/1 lines and loss systems as cooperative games. *Production and Operations Management*, 26(8):1568–1584.

Bagci, K. T. and Tekalp, A. M. (2018). Dynamic resource allocation by batch optimization for value-added video services over sdn. *IEEE Transactions on Multimedia*, 20(11):3084–3096.

Bendel, D. and Haviv, M. (2018). Cooperation and sharing costs in a tandem queueing network. *European Journal of Operational Research*, 271(3):926–933.

Benjaafar, S. (1995). Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research*, 87(2):375–388.

Berry, L. L., Seiders, K., and Grewal, D. (2002). Understanding service convenience. *Journal of marketing*, 66(3):1–17.

Bispo, C. F. (2013). The single-server scheduling problem with convex costs. *Queueing Systems*, 73(3):261–294.

Brouns, G. A. and Van Der Wal, J. (2006). Optimal threshold policies in a two-class preemptive priority queue with admission and termination control. *Queueing Systems*, 54(1):21–33.

Buyakar, T. V. K., Agarwal, H., Tamma, B. R., and Franklin, A. A. (2020). Resource allocation with admission control for gbr and delay qos in 5g network slices. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 213–220. IEEE.

Buzacott, J. A. (1996). Commonalities in reengineered business processes: models and issues. *Management Science*, 42(5):768–782.

Craig, J. and Petterson, V. (2005). Introduction to the practice of telemedicine. *Journal of Telemedicine and Telecare*, 11(1):3–9.

Feng, J., Pei, Q., Yu, F. R., Chu, X., Du, J., and Zhu, L. (2020). Dynamic network slicing and resource allocation in mobile edge computing systems. *IEEE Transactions on Vehicular Technology*, 69(7):7863–7878.

Fernández-Sabiote, E. and Román, S. (2012). Adding clicks to bricks: A study of the consequences on customer loyalty in a service context. *Electronic Commerce Research and Applications*, 11(1):36–48.

García-Sanz, M. D., Fernández, F. R., Fiestras-Janeiro, M. G., García-Jurado, I., and Puerto, J. (2008). Cooperation in markovian queueing models. *European Journal of Operational Research*, 188(2):485–495.

González, P. and Herrero, C. (2004). Optimal sharing of surgical costs in the presence of queues. *Mathematical Methods of Operations Research*, 59(3):435–446.

Harrison, J. M. (1975). Dynamic scheduling of a multiclass queue: Discount optimality. *Operations Research*, 23(2):270–282.

Iravani, S. M., Krishnamurthy, V., and Chao, G. H. (2007). Optimal server scheduling in nonpreemptive finite-population queueing systems. *Queueing Systems*, 55(2):95–105.

Karsten, F., Slikker, M., and Van Houtum, G.-J. (2015). Resource pooling and cost allocation among independent service providers. *Operations Research*, 63(2):476–488.

Keilson, J., Cozzolino, J., and Young, H. (1968). A service system with unfilled requests repeated. *Operations Research*, 16(6):1126–1137.

Koole, G. (1998). Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 30(3-4):323–339.

Körpeoğlu, E., Kurtz, Z., Kılınç-Karzan, F., Kekre, S., and Basu, P. A. (2014). Business analytics assists transitioning traditional medicine to telemedicine at virtual radiologic. *Interfaces*, 44(4):393–410.

Liu, H., Yu, Y., Benjaafar, S., and Wang, H. (2021). Price-directed cost sharing and demand allocation among service providers with multiple demand sources and multiple facilities. *Manufacturing & Service Operations Management*.

Mandelbaum, A. and Reiman, M. I. (1998). On pooling in queueing networks. *Management Science*, 44(7):971–981.

Mourtzis, D., Zervas, E., Boli, N., and Pittaro, P. (2020). A cloud-based resource planning tool for the production and installation of industrial product service systems (ipss). *The International Journal of Advanced Manufacturing Technology*, 106(11):4945–4963.

Niu, B., Zhou, Y., Shah-Mansouri, H., and Wong, V. W. (2016). A dynamic resource sharing mechanism for cloud radio access networks. *IEEE Transactions on Wireless Communications*, 15(12):8325–8338.

Norris, A. C. (2002). *Essentials of telemedicine and telecare*. Wiley Online Library.

Örmeci, E. L. and Burnetas, A. (2004). Admission control with batch arrivals. *Operations Research Letters*, 32(5):448–454.

Örmeci, E. L., Burnetas, A., and van der Wal, J. (2001). Admission policies for a two class loss system. *Stochastic Models*, 17(4):513–539.

Örmeci, E. L. and van der Wal, J. (2006). Admission policies for a two class loss system with general interarrival times. *Stochastic Models*, 22(1):37–53.

Ostrom, A. L., Parasuraman, A., Bowen, D. E., Patrício, L., and Voss, C. A. (2015). Service research priorities in a rapidly changing context. *Journal of Service Research*, 18(2):127–159.

Puterman, M. L. (2014). *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Roine, R., Ohinmaa, A., and Hailey, D. (2001). Assessing telemedicine: a systematic review of the literature. *Canadian Medical Association Journal*, 165(6):765–771.

Snyder, H., Witell, L., Gustafsson, A., Fombelle, P., and Kristensson, P. (2016). Identifying categories of service innovation: A review and synthesis of the literature. *Journal of Business Research*, 69(7):2401–2408.

Stidham, J. S. (1970). On the optimality of single-server queuing systems. *Operations Research*, 18(4):708–732.

Turhan, A., Alanyali, M., and Starobinski, D. (2012). Optimal admission control in two-class preemptive loss systems. *Operations Research Letters*, 40(6):510–515.

Ulukus, M. Y., Güllü, R., and Örmeci, E. L. (2011). Admission and termination control of a two class loss system. *Stochastic Models*, 27(1):2–25.

Weber, R. R. and Stidham, S. (1987). Optimal control of service rates in networks of queues. *Advances in applied probability*, 19(1):202–218.

Whitten, P., Holtz, B., and Nguyen, L. (2010). Keys to a successful and sustainable telemedicine program. *International Journal of Technology Assessment in Health Care*, 26(2):211–216.

Yu, Y., Benjaafar, S., and Gerchak, Y. (2015). Capacity sharing and cost allocation among independent firms with congestion. *Production and Operations Management*, 24(8):1285–1310.

Zeng, Y., Zhang, L., Cai, X., and Li, J. (2018). Cost sharing for capacity transfer in cooperating queueing systems. *Production and Operations Management*, 27(4):644–662.

Zhang, X. and Prybutok, V. R. (2005). A consumer perspective of e-service quality. *IEEE transactions on Engineering Management*, 52(4):461–477.

Zhao, G.-x. and Wang, S.-y. (2009). New research on a serving system under non-preemptive priority protocol. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on Information Engineering and Computer Science*, pages 1–4. IEEE.