Managing Queues with Reentrant Customers in Support of Hybrid Healthcare

Noa Zychlinski
Faculty of Data and Decision Sciences,
Technion – Israel Institute of Technology, Haifa 3200003, Israel
noazy@technion.ac.il

The Covid-19 pandemic has profoundly boosted the use of hybrid healthcare settings, which orchestrate face-to-face services together with virtual ones. The advantages of virtual healthcare services are clear: they are less costly, less disruptive for patients who can receive the service in the comfort of their home, and reduce patients' exposure to illnesses prevalent in healthcare facilities. Nevertheless, there is evidence that patients are likely to require a supplementary in-person service upon completion of their virtual service. Motivated by such settings, we study a multi-service queueing system with face-to-face, virtual and supplementary service channels. The service operator needs to allocate service capacity among the three classes and decide how to prioritize the patients when a service provider becomes available. The strong dependency between virtual and supplementary visits makes the problem challenging. Based on a fluid relaxation, we develop an index-based policy, the $\mathcal{R}-c\mu/\theta$ rule (or the $\mathcal{R}$ rule in short), which in addition to the holding cost, service time, abandonment rate and service reward, also carefully balances the return probability and associated penalty. The theoretical results along with numerical experiments demonstrate the effectiveness of the proposed policy and the importance of capacity coordination when managing hybrid service settings. Our work provides insights on the trade-off between convenience and the value of care when offering virtual healthcare services.

Key words: Hybrid/virtual healthcare, capacity allocation, scheduling, queues with rework, fluid models

1. Introduction

The Covid-19 pandemic has dramatically affected healthcare worldwide, accelerating broad acceptance of telemedicine and transforming the provision of medical care (Bokolo 2020, Kadir 2020). Telemedicine refers to technologies that enable the provision of remote clinical services via real-time communication between patients and healthcare providers, using video conferencing and patient monitoring (Monaghesh and Hajizadeh 2020). Telemedicine and virtual care can be integrated into the healthcare system to maximize the efficiency of healthcare delivery (Kadir 2020, Hur and Chang 2020). Indeed, telehealth adoption has significantly increased across the 50 countries most affected by Covid-19 (Wong et al. 2021).

Telemedicine advantages include the reduction of contagion risk and emergency room/clinic visits (Chauhan et al. 2020, Doshi et al. 2020). Virtual visits promote social distancing and help circumvent prolonged waiting times. Moreover, by minimizing in-person visits, telehealth can help reduce spread of today's virus and future ones, and protect medical practitioners from infection (Hollander and Carr 2020). Clearly, healthcare provision through telehealth will have a permanent role in traditional healthcare delivery long after Covid-19 becomes endemic (Ahmed et al. 2020).

Despite the obvious advantages of virtual visits, when operating and designing such systems, one must be aware of their potential for providing low-value instead of quality care (O'Reilly-Jacob et al. 2021). In this paper we focus on one important aspect of ensuring that patients receive quality care, which has significant operational implications – the requirement for a supplementary face-to-face visit. Specifically, there is evidence that virtual visits are likely to lead to a follow-up in-person visit (Ashwood et al. 2017, Shi et al. 2018). This usually has to do with the fact that some medical examinations/procedures cannot be executed remotely (McConnochie et al. 2015, Uscher-Pines et al. 2016). Indeed, an empirical study from a large healthcare system in the United States revealed that e-visits remove the gatekeepers between patients and specialist physicians and trigger a 6% increase in in-person visits (Bavafa et al. 2018).

To ground our model, we consider the real-world case of an urgent care center or a community clinic that patients visit only when they feel sick. An example of such a center, mentioned in Çakıcı and Mills (2021), is Northwell Health, `https://www.gohealthuc.com/northwell`, the largest healthcare provider in New York state. When patients seek urgent care, they are offered the opportunity to book a telehealth visit instead of heading to the clinic for an in-person visit. Then, they can meet with any available physician on call. If they choose a telehealth visit and require a supplementary in-person visit, they arrive at the clinic, and see one of the available physicians there.

To better understand the effect of these supplementary visits on system performance and decision making, we study a *hybrid* healthcare system, which provides virtual and in-person visits. The first visit to the system can be either in person or virtual. In a face-to-face visit, patients wait in a waiting room and are thus exposed to other illnesses. In a virtual visit, patients wait in the comfort of their home; nevertheless, they might require

a supplementary in-person visit. We model the patients requiring a supplementary visit as a different class of patients (i.e., the patient's class switches when physically entering the clinic). This is because the supplementary visit might have different characteristics in terms of service times, holding costs and abandonment rates than the first (face-to-face or virtual) visit. In particular, since some information has already been collected, the service time of the supplementary visit might be shorter than a full in-person visit. The holding cost and abandonment rate might be different as well, since the patient is now required to wait a second time and, therefore, may be more agitated and less patient.

There are two basic questions that come to mind when considering such a hybrid setting. First, there is a design question – how do we allocate capacity among the three types of services? The second question is an operative one – how do we schedule/prioritize the three classes or decide whom to admit when a physician becomes available? This paper attempts to address these two questions by studying a multi-server queueing model with three customer classes: face-to-face, virtual and supplementary. (We use the terms patients and customers interchangeably.) In Section 5 we discuss two model extensions for more classes and different supplementary classes for teletriage systems.

The optimal scheduling of multi-class queues has been studied extensively in the literature (See Section 1.1). The main take-way from these studies is the need to carefully balance the holding cost and the service and abandonment rates. Our work captures an additional feature in multi-class queueing systems: customer return and class transition while returning. The analysis suggests that in addition to the holding cost, and the service and abandonment rates, we also have to take into account the return probability and associated penalty. How to balance these factors can be highly non-trivial. The optimality of the $c\mu/\theta$ rule, for example, is only achieved asymptotically (Atar et al. 2010). Moreover, solving the Markov Decision Process (MDP) exactly often leads to limited structural insights and suffers from the curse of dimensionality especially in large systems (Papadimitriou and Tsitsiklis 1999).

Using a fluid framework, we study the optimal scheduling and capacity allocation policies. Specifically, the strong dependency between the virtual and supplementary visits, requires an integrative approach. To this end, we develop an effective index-based policy, the $\mathcal{R} - c\mu/\theta$ rule (or the $\mathcal{R}$ rule in short), which captures this dependency. We demonstrate how important it is to consider the return probability and penalty when scheduling

and capacity allocating service systems with returns. In particular, using policies that neglect the dependency between first- and second-time visitors can lead to unsatisfactory performances.

Our main contributions can be summarized as follows:

- **Modeling**. We study a multi-server queuing model with reentrant customers that have different characteristics than first-time visitors to the system. The main motivation for the model is facilitating a hybrid healthcare setting that provides face-to-face, virtual and supplementary in-person services. Nevertheless, the model is relevant to other service systems such as technical support centers, in which some repairs must be handled through an in-person service center. We provide two model extensions: the first includes multiple classes for each channel, and the second refers to the virtual channel as a teletriage system that classifies patients according to the supplementary service they require. We use a deterministic fluid model to approximate the system dynamics and derive scheduling and capacity allocation policies that shed light on the convenience versus low-value trade-off of virtual healthcare services.

- **The $\mathcal{R}$ rule.** For maximizing the fluid long-run profit of a hybrid healthcare setting, we introduce an index-based policy that incorporates the return probability and associated penalty along with the holding cost, service time, abandonment rate and service completion reward from each service channel. The $\mathcal{R}$ rule, which performs well in different parameter regimes, utilizes the $\mathcal{R}$ index of each class together with an integrated index for virtual and returning patients. We demonstrate that the $\mathcal{R}$ rule, which is optimal for the fluid problem, performs well – very close to optimal – in the corresponding stochastic system. Moreover, we show that our policy performs much better than other known policies that neglect the dependency between classes, even under non-stationary arrival rates. The simplicity of the policy together with its strong performance and the lack of other good policies for this setting, make the policy appealing and implementable.

- **Service capacity coordination.** Our work underscores the need for an integrated view of patients' first and following (if any) visits. In terms of system design, we show the importance of joint capacity allocation, in particular, for virtual and returning patients. This is done by carefully balancing the service allocation for these two classes, while incorporating the return probability and associated penalty. We identify the cases where this

coordination has the largest impact. In these cases, the superiority of the $\mathcal{R}$ rule is most significant when compared to other benchmark policies.

The rest of the paper is organized as follows. This section is concluded with a brief relevant literature review. In Section 2 we introduce our model and assumptions. In Section 3 we develop the $\mathcal{R}$ index rule and discuss its properties, optimality and implications in terms of system design and scheduling. In Section 4 we provide numerical experiments for the $\mathcal{R}$ index rule including a comparison to the optimal solution from the MDP solution, the $c\mu/\theta$ rule and the max-weight policies. This section also considers the transient profit-maximization problem under non-stationary arrivals. In Section 5 we discuss two model extensions: multiple supplementary services that consider the virtual channel as a teletriage system, and a multiple class model. Section 6 offers concluding remarks and future research directions.

1.1. Literature Review

This paper is related to two main bodies of literature. The first is the OR/OM literature on virtual/e-visits in healthcare systems. The second includes scheduling and capacity planning of queues with different customer classes.

Since the accommodation of virtual/e-visits in healthcare systems is a relatively new practice, there are only a few OR/OM papers in this area. Rajan et al. (2019) studied the impact of telehealth on the quality–speed trade-off for chronic patients. By considering an M/M/1 queue with strategic behavior, the authors showed that telemedicine can contribute to the specialists' productivity and to overall social welfare. Nevertheless, some patients that continue to use in-person visits may be worse off. In a recent study, Bavafa et al. (2021) focused on the physician compensation scheme (pricing) of e-visits in a primary care setting. The authors demonstrated that patients requiring intermediate healthcare may improve or worsen when e-visits are introduced, and identified settings in which system outcomes worsen under e-visits. Çakıcı and Mills (2021) recently studied the use of teletriage, a telemedicine service that allows patients to consult about their health condition. By analyzing an MDP to model patients' choices under triage errors, the authors find that for patients with high uncertainty regarding their health condition, teletriage can be beneficial in terms of cost outcomes. Nevertheless, due to the general overtriage rate, adding teletriage may increase the ED arrival rate and produce a negative cost outcome.

In our paper, we focus on a different aspect of telemedicine, which is the supplementary in-person visit and its effect on scheduling and capacity allocation decisions.

The scheduling of multiple customer classes in stochastic processing networks is a broad literature area. Cox and Smith (1961) proved the optimality of a simple index-based policy, known as the $c\mu$ rule for a single server queue with linear holding costs. Many generalizations have been offered for the rule; their optimality, however, is mostly obtained asymptotically (e.g., Van Mieghem 1995, Mandelbaum and Stolyar 2004, Huang et al. 2015).

In a multi-server system, Harrison and Zeevi (2004) and Atar et al. (2004) studied the scheduling of multiple classes with customer abandonment under the critically loaded regime. Atar et al. (2010) derived the asymptotic optimality of the $c\mu/\theta$ rule for many-server queues with abandonment under the many-server heavy traffic regime. More recently, Long et al. (2020) suggested an extension of the rule to general queue length cost functions and customer patience time distributions. Puha and Ward (2019) provided a tutorial on scheduling policies of many-server queues with impatient customers under the overloaded regime. Other recent extensions include scheduling of customers with different resource requirements (Zychlinski et al. 2020, 2022) and scheduling of proactive services (Hu et al. 2022). The latter, as our current paper, is studied under the conventional heavy-traffic regime.

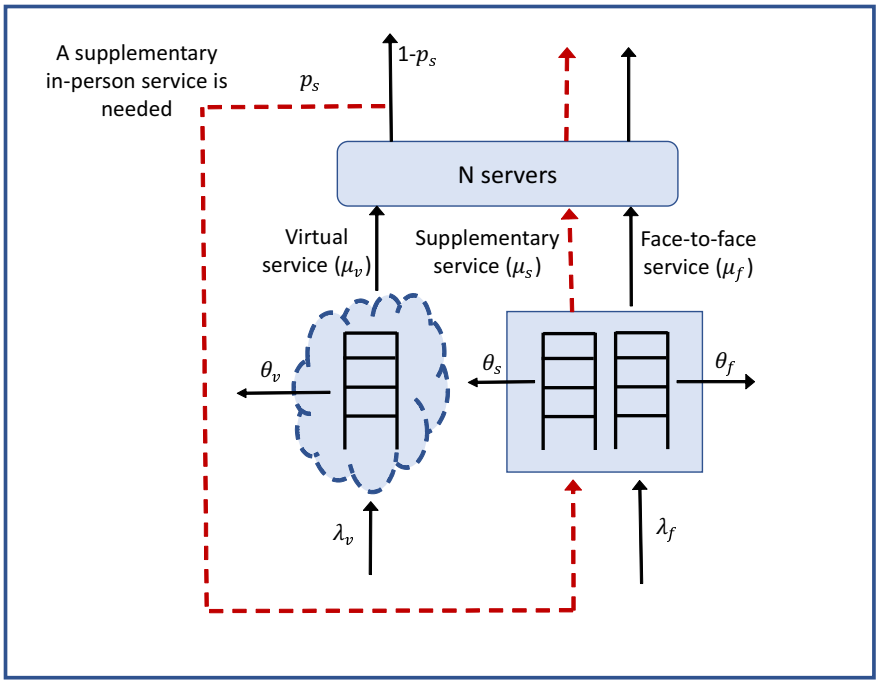
In this work, we complement this literature body by studying the scheduling of new and reentrant customers. This feature is relevant for hybrid healthcare settings as well as other services that may require a supplementary in-person visit. We allow returning patients to have different characteristics than the first-time visitor, and address the questions of how to schedule and allocate capacity among the different service channels.

2. The Hybrid Queuing Model

We consider a Markovian N-server queuing model and three classes of visits: face-to-face (f), virtual (v) and supplementary (s), as illustrated in Figure 1. The in-person and virtual classes arrive to the system according to a time-homogeneous Poisson process with rate λ_f and λ_v, respectively. Upon completion of a virtual visit, with probability p_s, the patient will require a follow-up in-person visit. To allow returning patients to have different characterizations than the face-to-face or virtual patients, and to support the decision of

how to schedule/prioritize the different classes, we consider returning patients to be a separate class of patients. Service and patience times of each class are exponential with rates μ_i and θ_i, $i = f, v, s$, respectively.

Figure 1 A hybrid system illustration with three patient classes: face-to-face (f), virtual (v) and patients returning for a supplementary in-person visit (s).



We assume that patients may return for an additional service at most once. If their service requirements have not been filled after that, they permanently leave the system. In the context of a hybrid emergency clinic, patients are usually referred to an ED if their health requirements have not been met during the in-person visit. Moreover, patients arriving for their virtual visit whose health condition is critical are referred to the ED. On the other hand, patients arriving for their virtual visit who are mildly ill do not require a supplementary in-person visit. Consequently, we assume that there is no significant difference in terms of health criticality between first-time and second-time (following a virtual visit) in-person patients. This allows us to focus on a profit/cost-effective metric.

Note that we refer to face-to-face patients as ones that do not reenter the system immediately upon service completion. After a while, if such patients require service, we consider them to be new arrivals.

Let $X_i(t)$ and $Q_i(t)$, $i = f, v, s$, denote the number of Class i customers in the system and in the queue, respectively, at time t, $t \geq 0$. Moreover, we use the notation $X(t) = (X_i(t), i = f, v, s)$ and $Q(t) = (Q_i(t), i = f, v, s)$. Let $Z_i(t)$ denote the number of servers assigned to Class i at time t; $Z(t) = (Z_i(t), i = f, v, s)$ are the decision variables. A scheduling policy π determines the allocation of servers to customers. We consider Markovian non-anticipating policies; that is, server allocations are made based on the current state $(X; Q)$ only. Under these scheduling policies, $\{(X(t); Q(t)) : t \geq 0\}$ is a Markov process. Finally, we also denote by $\Gamma_i(t)$, $i = f, v, s$, the cumulative number of Class i patients who abandoned the queue by time t.

Each completed service is associated with a profit of r_i, $i = f, v, s$. To capture a variety of reimbursement/pricing schemes, we do not impose any restrictions on these profits or their relationship. Each class incurs a holding cost of h_i, $i = f, v, s$, per patient per unit of time. It makes sense to assume that the holding cost of a face-to-face visit is higher than that of a virtual visit, due to the inconvenience and higher exposure to other illnesses prevalent in clinics and waiting rooms. Nevertheless, to keep the analysis as general as possible, we do not impose such a restriction.

Lastly, we incur an abandonment cost α_i, $i = f, v, s$ for each Class i patient that abandons the queue while waiting, and a return cost $\gamma \geq 0$ for each virtual patient that requires a supplementary in-person service. The aggregated profit up to time T is, therefore,

$$\mathbb{E}\left[\int_0^T \left[\sum_{i=f,v,s} [r_i \mu_i Z_i(t) - h_i Q_i(t)] - \gamma p_s \mu_v Z_v(t)\right] \mathrm{d}t - \sum_{i=f,v,s} \alpha_i \Gamma_i(T)\right], \tag{1}$$

where under the Markovian modeling assumption,

$$\mathbb{E}\left[\Gamma_i(T)\right] = \theta_i \mathbb{E}\left[\int_0^T Q_i(t)\mathrm{d}t\right], \quad i = f, v, s.$$

Equation (1) can then be rewritten as follows,

$$\mathbb{E}\left[\int_0^T \sum_{i=f,v,s} [r_i \mu_i Z_i(t) - (h_i + \alpha_i \theta_i) Q_i(t)] - \gamma p_s \mu_v Z_v(t)\mathrm{d}t\right].$$

For simplicity of notation and similar to Hu et al. (2022), we introduce the "generalized" holding costs $c_i = h_i + \alpha_i \theta_i$, $i = f, v, s$.

Our goal is, therefore, to find a scheduling policy π that maximizes the total expected long-run average profit; specifically:

$$\max_{\pi\in\Omega} \liminf_{T\to\infty} \frac{1}{T}\,\mathbb{E}\left[\int_0^T \sum_{i=f,v,s} [r_i\mu_i Z_i(t) - c_i Q_i^\pi(t)] - \gamma p_s \mu_v Z_v(t)\mathrm{d}t\right] \tag{2}$$

$$\text{s.t.} \sum_{i=f,v,s}^{I} Z_i(t) \le N, \quad t \ge 0;$$

$$0 \le Z_i(t) \le X_i^\pi(t), \quad i = f,v,s, t \ge 0,$$

where Ω denotes the set of admissible controls. Note that the superscript π in $X^\pi(t)$ and $Q^\pi(t)$ emphasizes the dependency of the processes on policy π.

The objective function includes an aggregation of the long-run profit from the three classes. In particular, it includes the profit from each completed service minus the holding cost of the waiting patients minus the return penalty γ for each returning patient.

The first constraint states that the total allocated number of servers cannot exceed the total service capacity N. The second constraint implies that the number of servers allocated to Class i cannot exceed the number of Class i customers.

This profit maximization problem is an MDP. The curse of dimensionality (Papadimitriou and Tsitsiklis 1999) – a large (infinite) state-space and policy-space – makes it prohibitively hard to solve and characterize the optimal scheduling policy. To gain structural insights into the optimal scheduling policy and capacity allocation, we take a deterministic fluid approach. Fluid models are known to provide good approximation of the first-order mean dynamics of stochastic systems, and are thus useful for a variety of applications related to service operations management (Zychlinski 2022). Such models are usually derived as limits through the Functional Law of Large Numbers. In this paper, we apply the conventional heavy traffic regime (Whitt 2002). In this regime, the arrival rates and service rates are scaled up (this is equivalent to scaling up time), while the number of servers is held fixed.

2.1. The Fluid Model

In the fluid model, deterministic continuous rates replace the stochastic processes. We use lowercase x_i, q_i and z_i, $i = f,v,s$, to denote the steady-state fluid content in the system, the queue length and the service capacity assigned to Class i, respectively. The decision variables z_i's can be thought of as the level of service capacity that is allocated in the long run to Class i. For a given capacity allocation, z_i, $i = f,v,s$, such that $z_f + z_v + z_s \le N$, and $0 \le z_i \le x_i$, the system dynamics under the fluid model are characterized by the following set of differential equations:

$$\begin{cases} \dot{q}_i(t) = \lambda_i - \mu_i z_i(t) - \theta_i q_i(t), \quad i = f, v; & \text{(first-time visitors)} \\ \dot{q}_s(t) = p_s \mu_v z_v(t) - \mu_s z_s(t) - \theta_s q_s(t); & \text{(second-time visitors)} \\ q_i(t), z_i(t) \geq 0, \quad i = f, v, s, \quad t \geq 0. \end{cases} \tag{3}$$

The first equation for the first face-to-face or virtual visit describes the rate of change in the corresponding queue length, which includes the arrival rate minus the departure rate. The latter includes the service completion rate and the abandonment rate from the queue. The second equation is for the supplementary in-person visit that may be needed after the virtual visit. Here, the arrival rate is the departure rate from the virtual service, $\mu_v z_v(t)$, multiplied by the return probability. Note that the virtual service is the feeding source of the supplementary visit. That is, if no capacity is allocated to the virtual service, no patients will require a supplementary visit.

The fluid analog for the long-run profit maximization problem is, therefore, the following infinite dimensional linear program:

$$\begin{aligned} \max_{q,z} \quad & \liminf_{T\to\infty} \frac{1}{T} \int_0^T \left[\sum_{i=f,v,s} [r_i \mu_i z_i(t) - c_i q_i(t)] - \gamma \mu_v p_s z_v(t) \right] \mathrm{d}t \\ \text{s.t.} \quad & \dot{q}_i(t) = \lambda_i - \mu_i z_i(t) - \theta_i q_i(t), \quad i = f, v; \quad \text{(first-time visitors)} \\ & \dot{q}_s(t) = p_s \mu_v z_v(t) - \mu_s z_s(t) - \theta_s q_s(t); \quad \text{(second-time visitors)} \\ & \sum_{i=f,v,s} z_i(t) \leq N, \quad t \geq 0; \\ & q_i(t), z_i(t) \geq 0, \quad i = f, v, s, \quad t \geq 0, \end{aligned} \tag{4}$$

where the system dynamics constraints are as in (3). In addition to reaching scheduling decisions for the three classes, we also wish to derive the optimal capacity allocation for the three service channels. We are therefore interested in finding the equilibrium point to which the fluid approximation converges. Theorems 1 and 2 allow us to rewrite the problem as a *finite* dimensional linear program. Specifically, if the fluid approximation converges to an equilibrium point $(\bar{q}, \bar{z})$ as $t \to \infty$, then maximizing the long-run average profit can be achieved by finding the optimal equilibrium point:

$$
\begin{aligned}
&\max_{\bar{q},\bar{z}} \sum_{i=f,v,s} [r_i \mu_i \bar{z}_i - c_i \bar{q}_i] - \gamma \mu_v p_s \bar{z}_v \\
&\text{s.t.} \quad \lambda_i = \mu_i \bar{z}_i + \theta_i \bar{q}_i, \quad i = f, v; \quad \text{(first-time visitors)} \\
&\qquad p_s \mu_v \bar{z}_v = \mu_s \bar{z}_s + \theta_s \bar{q}_s; \qquad \text{(second-time visitors)} \\
&\qquad \sum_{i=f,v,s} \bar{z}_i \leq N; \\
&\qquad \bar{q}_i, \bar{z}_i \geq 0, \quad i = f, v, s.
\end{aligned} \tag{5}
$$

The first two constraints are derived by imposing $\dot{q}_i(t) = 0$, $i = f, v, s$, in the first two fluid dynamic constraints in (4), and replacing the functions $q_i(t)$ and $z_i(t)$ by their equilibrium $\bar{q}_i$ and $\bar{z}_i$, respectively. Note that according to the first two constraints, the rate of visit loss is $\lambda_i - \mu_i \bar{z}_i = \theta_i \bar{q}_i$, for classes $i = f, v$, and $p_s \mu_v \bar{z}_v - \mu_s \bar{z}_s = \theta_s \bar{q}_s$ for Class s.

Rearranging (5), by substituting

$$
\bar{q}_i = (\lambda_i - \mu_i \bar{z}_i)/\theta_i, \ i = f, v, \quad \text{and} \quad \bar{q}_s = (p_s \mu_v \bar{z}_v - \mu_s \bar{z}_s)/\theta_s,
$$

in the objective function and omitting the constants that do not affect the optimal solution yields the following equivalent problem:

$$
\begin{aligned}
&\max_{\bar{z}} \sum_{i=f,v,s} \mathcal{R}_i \bar{z}_i \\
&\text{s.t.} \quad 0 \leq \bar{z}_i \leq \lambda_i/\mu_i, \quad i = f, v; \\
&\qquad 0 \leq \bar{z}_s \leq p_s \mu_v \bar{z}_v/\mu_s; \\
&\qquad \sum_{i=f,v,s} \bar{z}_i \leq N,
\end{aligned} \tag{6}
$$

where the $\mathcal{R}$ indexes are:

$$
\begin{aligned}
&\mathcal{R}_f = \mu_f (r_f + c_f/\theta_f), \\
&\mathcal{R}_v = \mu_v \left(r_v + c_v/\theta_v - p_s \left(\gamma + c_s/\theta_s\right)\right), \\
&\mathcal{R}_s = \mu_s (r_s + c_s/\theta_s).
\end{aligned} \tag{7}
$$

The first constraint in (6) states that at most λ_i/μ_i, $i=f,v$, service capacity is needed to handle the face-to-face and virtual customers. The second constraint states that at most $p_s\mu_v\bar{z}_v/\mu_s$ service capacity is needed to handle the returning patients.

Note that when $r_i=0$, $i=f,v,s$, and $p_s=0$, we retrieve the $c\mu/\theta$ indexes according to which the $c\mu/\theta$ rule prioritizes the classes (Atar et al. 2010). Note also that return probability p_s and return penalty γ reduce the virtual customers' index. That is, as the virtual service becomes less effective (i.e., associated with a higher return probability), the optimal policy will tend to utilize the virtual services less. This result is intuitive in the sense that prioritizing virtual patients will lead to excessive costs when some of them will be returning for supplementary service. In the extreme case where the return probability and penalty are very high, the optimal decision might be to cancel virtual services at that particular clinic entirely.

In addition to (7), we also introduce the index $\mathcal{R}_{v,s}$, which is a weighted average of the $\mathcal{R}_v$ and $\mathcal{R}_s$ indexes; namely,

$$\mathcal{R}_{v,s}=\frac{\mu_s}{\mu_s+p_s\mu_v}\mathcal{R}_v+\frac{p_s\mu_v}{\mu_s+p_s\mu_v}\mathcal{R}_s.$$

Note that when $p_s=0$, we have $\mathcal{R}_{v,s}=\mathcal{R}_v$. In Section 3, we will see that in some cases the scheduling and capacity allocation decisions rely on this integrated index. A further intuition on the integrated $\mathcal{R}_{v,s}$ index is provided in Remark 1.

Throughout the paper, we make the technical assumption that the $\mathcal{R}$ indexes are all distinct. That is, $\mathcal{R}_i\neq\mathcal{R}_j$ for $i\neq j$. Non-unique indexes could complicate our analysis in Sections 3 and 5 by adding many more cases to consider.

Lastly, we consider an autonomous differential equation:

$$\dot{q}(t)=f(q(t)) \quad \text{with} \quad q(0)=q_0. \tag{8}$$

Suppose there exists an equilibrium point $\bar{q}$ so that $f(\bar{q})=0$. Then, $\bar{q}$ is *globally asymptotically stable* if for any initial condition q_0, $\lim_{t\to\infty}||q(t)-\bar{q}||=0$, where $||\cdot||$ is the Euclidean norm.

3. The $\mathcal{R}$-Index Policy

Because of the dependency between virtual and returning patients, the optimal solution to (6) is not necessarily to straightforwardly assign larger values of $\bar{z}$ to the ones with the larger $\mathcal{R}$ index. Before we characterize the optimal solution to (6), we introduce two index-based policies, as well as the $\mathcal{R}$ rule, which combines the two.

DEFINITION 1 (THE NAIVE $\mathcal{R}$ RULE). Assign priority to Class i, $i = f, v, s$, having the higher $\mathcal{R}_i$ index.

For the following definition, we combine the virtual and supplementary classes together to form a joint (artificial) class $\{v, s\}$, which is associated with the $\mathcal{R}_{v,s}$ index. In the first step, we set the priority between Class f and the joint class $\{v, s\}$. Then in the second step, we set the priority within the joint class (for Classes v and s).

DEFINITION 2 (THE TWO-STEP $\mathcal{R}$ RULE). Assign priority to Class i, $i = f, \{v, s\}$, with the higher $\mathcal{R}_i$ index. Then, within the joint class, assign priority to Class i, $i = v, s$, with the higher $\mathcal{R}_i$ index.

Note that when $p_s = 0$, both the naive $\mathcal{R}$ rule and the two-step $\mathcal{R}$ rule retrieve the $c\mu/\theta$ rule (Atar et al. 2010).

DEFINITION 3. (The $\mathcal{R}$ rule)

Case 1. When $\mathcal{R}_s < \mathcal{R}_v$, prioritize the classes according to the naive $\mathcal{R}$ rule.

Case 2. When $\mathcal{R}_s > \mathcal{R}_v$, prioritize the classes according to the two-step $\mathcal{R}$ rule.

Next, we prove that the optimal solution to the fluid optimization problem (6)–(7) is a globally asymptotically stable equilibrium under the $\mathcal{R}$ index rule. Moreover, we characterize the capacity allocation in equilibrium for each service channel and for different parameter regimes. Theorem 1, which we prove in Appendix A, formalizes this result.

THEOREM 1 **(globally asymptotically stable equilibria)**. *When following the $\mathcal{R}$ rule for the system dynamics described in (4) from any initial condition, and for $\theta_i > 0$, $i = f, v, s$, the globally asymptotically stable equilibria, $\bar{z} = (\bar{z}_f, \bar{z}_v, \bar{z}_s)$ and $\bar{q} = (\bar{q}_f, \bar{q}_v, \bar{q}_s)$, are as shown in Table 1.*

The equilibrium queue lengths are then given by

$$\bar{q}_i = (\lambda_i - \mu_i \bar{z}_i)/\theta_i, \ \ i = f, v, \quad \text{and} \quad \bar{q}_s = (p_s \mu_v \bar{z}_v - \mu_s \bar{z}_s)/\theta_s. \tag{9}$$

Note that under Case 2, $\bar{z}_s = p_s \mu_v \bar{z}_v / \mu_s$, so that $\bar{q}_s = 0$. This makes sense since in this case, Class s is prioritized over Class v. If $\bar{q}_s > 0$, we could get an improvement by shifting some capacity from Class v to Class s. Therefore, in equilibrium, we need to make sure that just enough capacity is allocated to Class v to assure that $\bar{q}_s = 0$.

The $\bar{z}$'s in Theorem 1 can be interpreted as the long-run capacity allocated to each service channel. Specifically, when $\mathcal{R}_s < \mathcal{R}_v$, capacity is allocated to each service channel separately according

Table 1 Globally asymptotically stable equilibria

Case	$\bar{z}_f$	$\bar{z}_v$	$\bar{z}_s$
1. The naive $\mathcal{R}$ rule ($\mathcal{R}_s < \mathcal{R}_v$)			
1a. $\mathcal{R}_f < \mathcal{R}_s < \mathcal{R}_v$	$\frac{\lambda_f}{\mu_f} \wedge (N - \bar{z}_v - \bar{z}_s)$	$\frac{\lambda_v}{\mu_v} \wedge N$	$\frac{p_s\mu_v}{\mu_s}\bar{z}_v \wedge (N - \bar{z}_v)$
1b. $\mathcal{R}_s < \mathcal{R}_f < \mathcal{R}_v$	$\frac{\lambda_f}{\mu_f} \wedge (N - \bar{z}_v)$	$\frac{\lambda_v}{\mu_v} \wedge N$	$\frac{p_s\mu_v}{\mu_s}\bar{z}_v \wedge (N - \bar{z}_f - \bar{z}_v)$
1c. $\mathcal{R}_s < \mathcal{R}_v < \mathcal{R}_f$	$\frac{\lambda_f}{\mu_f} \wedge N$	$\frac{\lambda_v}{\mu_v} \wedge (N - \bar{z}_f)$	$\frac{p_s\mu_v}{\mu_s}\bar{z}_v \wedge (N - \bar{z}_f - \bar{z}_v)$
2. The two-step $\mathcal{R}$ rule ($\mathcal{R}_v < \mathcal{R}_s$)			
2a. $\mathcal{R}_f < \mathcal{R}_{v,s}$	$\frac{\lambda_f}{\mu_f} \wedge (N - \bar{z}_v - \bar{z}_s)$	$\frac{\lambda_v}{\mu_v} \wedge \frac{\mu_s N}{\mu_s + p_s\mu_v}$	$\frac{p_s\mu_v}{\mu_s}\bar{z}_v$
2b. $\mathcal{R}_{v,s} < \mathcal{R}_f$	$\frac{\lambda_f}{\mu_f} \wedge N$	$\frac{\lambda_v}{\mu_v} \wedge \frac{\mu_s(N-\bar{z}_f)}{\mu_s + p_s\mu_v}$	$\frac{p_s\mu_v}{\mu_s}\bar{z}_v$

$x \wedge y = \min(x, y)$.

to its $\mathcal{R}$ index. In this case, returning patients are treated as any other class (except for the fact that the class demand is determined by the capacity allocation to the virtual channel).
When $\mathcal{R}_s > \mathcal{R}_v$, however, capacity allocation of virtual and returning patients must be coordinated to assure that enough capacity is allocated to the virtual channel that feeds the supplementary channel. In this case, Classes v and s are treated jointly, and the relevant relation then becomes the face-to-face channel versus the joint channel of virtual and supplementary visits.

We are now ready for Theorem 2, which establishes the optimality of the $\mathcal{R}$ rule for the long-run profit maximization problem (6). The proof of the theorem, which is provided in Appendix B, is based on Theorem 1 where we guarantee that the fluid system converges to the equilibrium point under the $\mathcal{R}$ rule.

THEOREM 2 **(optimality of the $\mathcal{R}$ rule)**. *For the long-run profit maximization problem (6), with $\theta_i > 0$, $i = f, v, s$, and any initial condition, the $\mathcal{R}$ rule is optimal.*

The suggested $\mathcal{R}$ rule is an index-based policy; such policies often exhibit many desirable properties such as being simple to implement and achieving good (if not optimal) performance. Because of the dependency between virtual and returning patients, there is a need to combine the naive and two-step $\mathcal{R}$ rules. In Section 4, we demonstrate through extensive numerical experiments, the effectiveness and robustness of the $\mathcal{R}$ rule. Specifically, we show that the policy performs very close to optimal and much better than other known policies in various settings and under different system loads.

Note that $\mathcal{R}_v = \mu_v\left(r_v + c_v/\theta_v - p_s\left(\gamma + c_s/\theta_s\right)\right)$ is decreasing in p_s, so the switching point between Case 1 and Case 2 is

$$\tilde{p_s} = \frac{\mu_v\left(r_v + c_v/\theta_v\right) - \mu_s\left(r_s + c_s/\theta_s\right)}{\mu_v\left(\gamma + c_s/\theta_s\right)} = \frac{\mathcal{R}_v|_{p_s=0} - \mathcal{R}_s}{\mu_v\left(\gamma + c_s/\theta_s\right)}.$$

Specifically, if $p_s < \tilde{p_s}$, the naive $\mathcal{R}$ rule is optimal, and if $p_s > \tilde{p_s}$, the two-step $\mathcal{R}$ rule is optimal. In particular,

- When $\mathcal{R}_v|_{p_s=0} < \mathcal{R}_s$, the two-step $\mathcal{R}$ rule is optimal for every $p_s \in [0,1]$;
- When $\mathcal{R}_s < \mathcal{R}_v|_{p_s=1} = \mu_v\left(r_v + c_v/\theta_v - (\gamma + \frac{c_s}{\theta_s})\right)$, the naive $\mathcal{R}$ rule is optimal for every $p_s \in [0,1]$.

Remark 1. What is the motivation for the joint index $\mathcal{R}_{vs}$?

Intuitively, when $\mathcal{R}_s > \mathcal{R}_v$, we would tend to prioritize returning patients over virtual ones. Since the latter are the feeding source of the former, enough capacity needs to be allocated to virtual patients to assure that $\bar{z}_s = p_s\mu_v\bar{z}_v/\mu_s$.

By substituting $\bar{z}_s$ in the objection function in (6), we get

$$\max_{\bar{z}_f, \bar{z}_v} \ \mathcal{R}_f\bar{z}_f + \left(\mathcal{R}_v + \frac{p_s\mu_v}{\mu_s}\mathcal{R}_s\right)\bar{z}_v. \tag{10}$$

When the capacity constraint is active (i.e., $\bar{z}_f + \bar{z}_v + \bar{z}_s = N$), we have

$$\bar{z}_v = \frac{\mu_s}{\mu_s + p_s\mu_v}\left(N - \bar{z}_f\right),$$

which in turn is plugged in back into (10), and gives the following objective function

$$\max_{\bar{z}_f} \mathcal{R}_f\bar{z}_f + \mathcal{R}_{v,s}\left(N - \bar{z}_f\right).$$

Now it is clear that capacity needs to be allocated to Class f when $\mathcal{R}_f > \mathcal{R}_{v,s}$, and jointly to Classes v and s, otherwise.

Remark 2. In urgent care centers, which constitute our main motivating application, the queue regime is often not First Come First Served (FCFS) (e.g., Hu et al. 2022, Zychlinski et al. 2022), but some other merit that takes into account the severity of patients' conditions and service time. For example, the supplementary channel might be prioritized over the first-time in-person channel if the former has much shorter service times, since the patient will have already been diagnosed/treated remotely, or if the patient's condition has deteriorated (in our model this translates into a higher holding cost). Waiting patients might perceive this as being unfair when patients arriving after them start their service before them. To overcome this, when arriving at such centers, patients must usually sign in at a (self-service) registration stand and receive a number. Announcements and display screens in the waiting room show which number should enter each physician office. In this way, the prioritization can be done through the system without being too obvious.

3.1. Optimal System Design

The $\mathcal{R}$ rule addresses the operational question of how to schedule/prioritize the three classes or whom to admit when a physician becomes available. Additionally, the $\mathcal{R}$ rule addresses an even more basic question that comes to mind when considering such a hybrid setting; specifically, it focuses on the design question of how to allocate capacity among the three services channels. We find that in some extreme cases, it is better to utilize only one or two service channels.

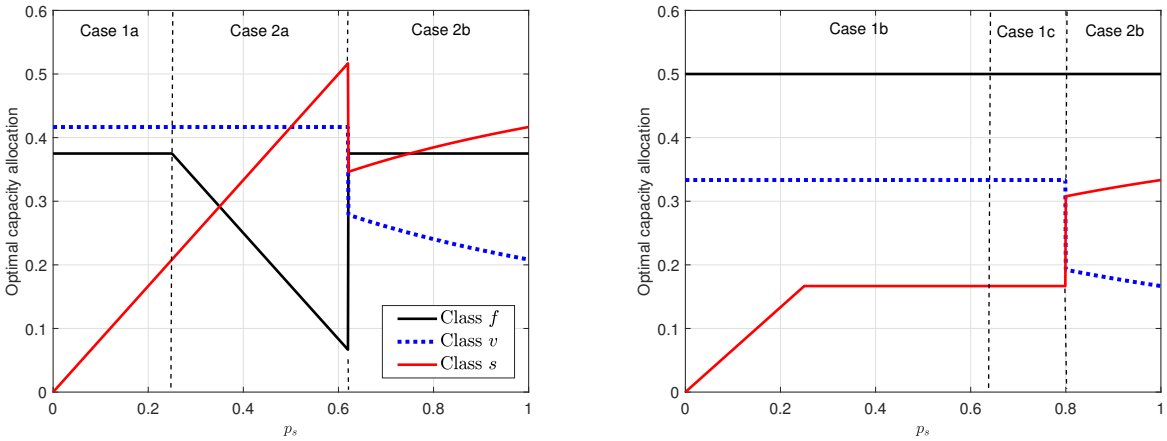
Another way of looking at the optimal system design according to the $\mathcal{R}$ rule is that it chooses which channel(s) to utilize, which channel(s) not to utilize, and at most one channel to partially utilize. Under the naive $\mathcal{R}$ rule, this intuition is straightforward. Under the two-step $\mathcal{R}$ rule, this observation is true due to the careful balancing of capacity allocated to the virtual and supplementary channels.

Table 1 describes the optimal long-run capacity allocation to the three service channels. As the need for supplementary service increases (i.e., the return probability increases), the virtual service becomes less effective. This is because many patients' issues cannot be resolved through the virtual channel. Consequently, more capacity will be allocated to the in-person service.

Figure 2 illustrates different structures of the optimal capacity allocation as a function of return probability p_s. For example, in the left plot, the policy switches from the naive $\mathcal{R}$ rule (Case 1a) to the two-step $\mathcal{R}$ rule (Case 2a) when $p_s = 0.24$. Then, at $p_s = 0.62$, the policy switches within the two-step $\mathcal{R}$ index to Case 2b, where the entire service capacity is allocated to the face-to-face channel. Indeed, when the virtual service is less effective, it is better to focus mainly on the face-to-face channel. Note that it is possible to have the same capacity allocation for different cases (see the right plot, when switching from Case 1b to Case 1c). Moreover, there could be extreme cases under heavy load where the optimal solution would tend not to utilize one or two service channels. For example, when the virtual service is associated with a very high return rate, it might be best to focus only on the face-to-face channel. Under such heavy loads when not all channels are utilized, it might also be beneficial to optimize staffing levels by considering the trade-off between service completion reward (including abandonment penalty) and staffing costs.

Translation of the optimal solution back to the stochastic system. In terms of scheduling, the translation relies on the priorities determined by the $\mathcal{R}$ rule when a service provider becomes available. In Section 4 we provide numerical examples demonstrating that the policy is effective when implemented in the stochastic system. The translation of the long-term capacity allocation is in terms of system design. That is, we determine how much capacity needs to be allocated on average to each service channel. If, for example, the healthcare facility follows a non-sharing policy

Figure 2 Optimal capacity allocation for different values of $\boldsymbol{p_s}$ $(\boldsymbol{N=1})$. In the left plot $\boldsymbol{\lambda_f=1.5}$, $\boldsymbol{\lambda_v=2.5}$, for classes $[\boldsymbol{f,v,s}]$: $\boldsymbol{r=[7,6,0]}$, $\boldsymbol{c=[1,0.2,1]}$, $\boldsymbol{\mu=[4,6,3]}$, $\boldsymbol{\theta=[0.12,0.01,0.03]}$ and $\boldsymbol{\gamma=5}$. In the right plot, $\boldsymbol{\lambda_f=2}$, $\boldsymbol{\lambda_v=2}$, $\boldsymbol{c=[1,0.6,0.8]}$, $\boldsymbol{\mu=[4,6,3]}$, $\boldsymbol{\theta=[0.08,0.03,0.04]}$ and $\boldsymbol{\gamma=0}$.



of physicians to different service channels (e.g., in a certain shift, physicians cannot switch between service channels), then the capacity allocation is the average amount of physicians' time that needs to be assigned to each service channel.

4. Numerical Experiments

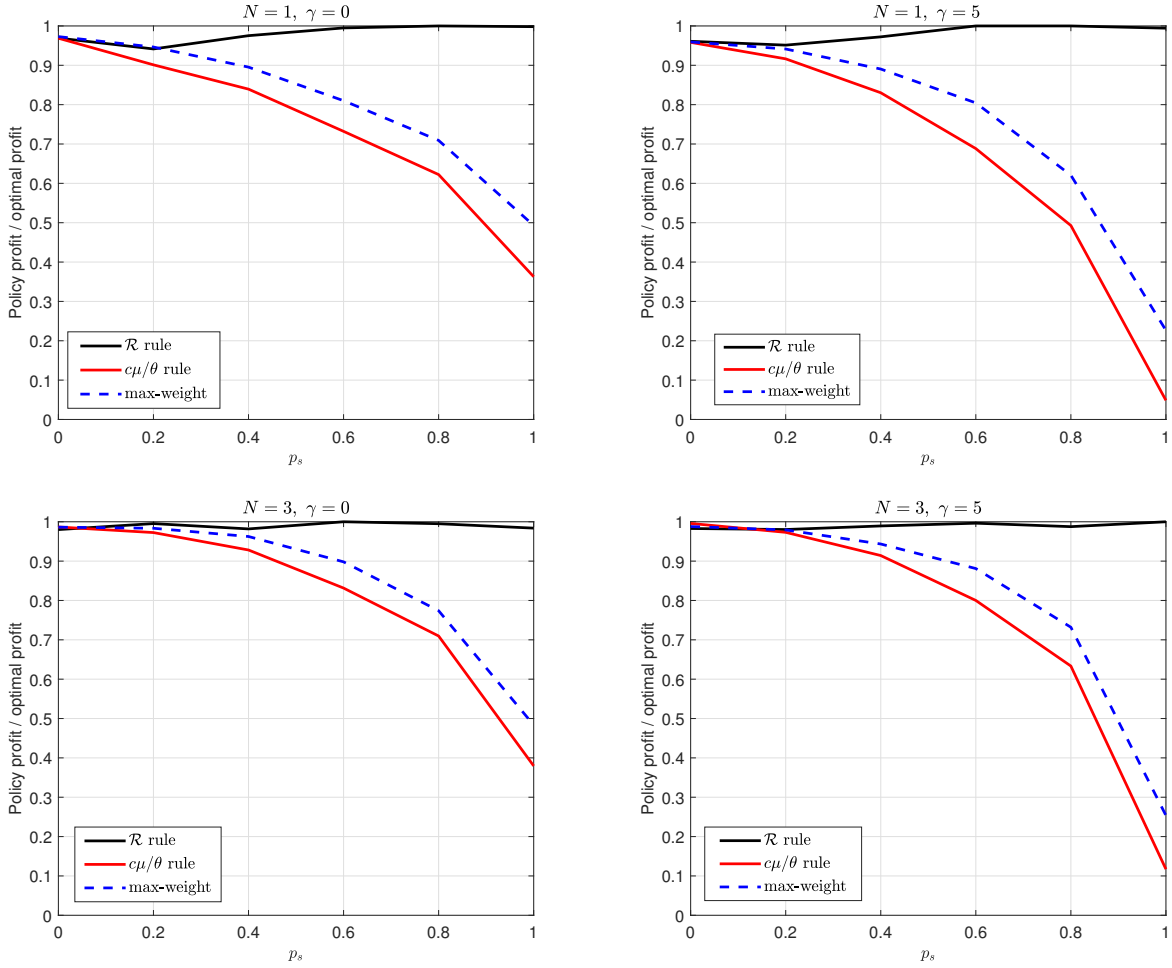
In this section we examine the performance of the $\mathcal{R}$ rule in the original stochastic system using simulation. When the number of servers is very small, we can solve the MDP numerically and then compare it to the performance of the $\mathcal{R}$ rule. We also compare two other well-known policies: the $c\mu/\theta$ rule (Atar et al. 2010) and the max-weight policy (Stolyar 2004, Dai and Lin 2005). We modify the $c\mu/\theta$ rule to include the reward from each service completion; specifically, the index of Class i is now $(r_i + c_i/\theta_i)\,\mu_i$. We also modify the max-weight policy to include the reward from each service completion and the abandonment rate, as follows: at each time t, given $X(t)=x$ and $Q(t)=q$, the server allocation, $z=(z_f,z_v,z_s)$, under the max-weight policy is the solution to the following IP:

$$\begin{aligned} &\max_{z} \sum_{i=1}^{I} (r_i + c_i/\theta_i)\,\mu_i z_i x_i \\ &\text{s.t.} \quad \sum_{i=1}^{I} z_i \le N, \quad 0 \le z_i \le x_i, \quad z_i \in \mathbb{N}_0, \ i=f,v,s. \end{aligned} \tag{11}$$

We note that the numerical experiments for the stochastic system assume preemption. The fluid analysis, however, applies for both preemptive and non-preemptive regimes.

Figure 3 presents the ratio between each policy's long-run average profit and the optimal profit achieved by explicitly solving the MDP. We present the ratios for different values of the return probability p_s in four scenarios. We observe that the $\mathcal{R}$ rule performs very well in all scenarios and all values of p_s. The $c\mu/\theta$ and the max-weight policy perform reasonably well under small return probability. Their respective performance, however, deteriorates in comparison with the optimal policy as the return probability increases. This deterioration prevails even when there is no penalty associated with patient return (i.e., $\gamma = 0$).

Figure 3 Profit ratio of each policy to the optimal policy for different values of $\boldsymbol{p_s}$. $\boldsymbol{\lambda_f = 1.1375N}$, $\boldsymbol{\lambda_v = 1.925N}$, for classes $[\boldsymbol{f, v, s}]$: $\boldsymbol{r = [17.5, 15, 0]}$, $\boldsymbol{c = [2.5, 0.2, 1]}$, $\boldsymbol{\mu = [4, 6, 3]}$, $\boldsymbol{\theta = [0.12, 0.01, 0.03]}$.



Recall that the $\mathcal{R}$ rule was derived from a fluid approximation model that can arise as a limit through the functional law of large numbers under the conventional heavy traffic regime. We,

therefore, wish to examine the policies' performances under different system loads. To this end, we define the traffic intensity as follows:

$$\rho := \frac{1}{N}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s}\right].$$

Table 2 compares the long-run average profit of the three policies to the optimal profit achieved by solving the MDP. We vary ρ by proportionally scaling up the arrival rates. For each case and policy we present (in parentheses) the ratio between the policy's profit and the optimal profit. We observe that the $\mathcal{R}$ rule performs very well in all cases. The performance under moderate and high traffic intensities, when effective scheduling policies are most needed, is very close to optimal. This is due to the fact that the $\mathcal{R}$ rule was derived under conventional heavy traffic, which tends to be more accurate as traffic intensity increases. Moreover, we see that the $c\mu/\theta$ and max-weight policies perform slightly better than the $\mathcal{R}$ rule under very low traffic intensity. Under high traffic intensity, however, their performance deteriorates; this deterioration does not happen under the $\mathcal{R}$ rule, which keeps performing very close to optimal.

Table 2 Comparison between the long-run average profit under different policies and traffic intensities. The numbers in parentheses are the profit ratios between each policy and the optimal one (MDP). The parameters are for classes $[\boldsymbol{f}, \boldsymbol{v}, \boldsymbol{s}]$ are: $r = [\mathbf{17.5, 15, 0}]$, $\boldsymbol{c} = [\mathbf{1, 0.2, 1}]$, $\boldsymbol{\mu} = [\mathbf{4, 6, 3}]$, $\boldsymbol{\theta} = [\mathbf{0.12, 0.01, 0.03}]$, $\boldsymbol{\gamma} = \mathbf{0}$, and $\boldsymbol{p_s} = \mathbf{0.7}$.

	Case	Traffic intensity	(λ_f, λ_v)	Long-run average profit			
				MDP	$\mathcal{R}$ rule	$c\mu/\theta$	max-weight
$N=1$	1	$\rho = 0.3$	$(0.325, 0.55)$	14.53	13.39 (0.922)	14.13 (0.973)	13.58 (0.935)
	2	$\rho = 0.45$	$(0.4875, 0.825)$	20.95	20.60 (0.983)	20.56 (0.981)	19.96 (0.952)
	3	$\rho = 0.6$	$(0.65, 1.1)$	29.18	29.02 (0.995)	27.05 (0.927)	27.78 (0.952)
	4	$\rho = 0.75$	$(0.8125, 1.375)$	35.64	35.24 (0.989)	31.72 (0.89)	32.55 (0.913)
	5	$\rho = 0.9$	$(0.975, 1.65)$	41.71	41.4 (0.993)	34.28 (0.822)	35.42 (0.849)
	6	1.05	$(1.1375, 1.925)$	40.47	40.25 (0.994)	33.25 (0.822)	24.85 (0.861)
	7	$\rho = 1.2$	$(1.3, 2.2)$	33.86	33.6 (0.992)	28.32 (0.836)	28.1 (0.83)
$N=3$	8	$\rho = 0.3$	$(0.975, 1.65)$	42.35	41.26 (0.974)	42.15 (0.995)	41.23 (0.973)
	9	$\rho = 0.45$	$(1.4625, 2.475)$	63.25	62.87 (0.994)	61.83 (0.978)	62.89 (0.994)
	10	$\rho = 0.6$	$(1.95, 3.3)$	88.84	88.6 (0.997)	86.33 (0.972)	86.93 (0.978)
	11	$\rho = 0.75$	$(2.4375, 4.125)$	109.6	109.4 (0.998)	104.38 (0.952)	105.62 (0.964)
	12	$\rho = 0.9$	$(2.925, 4.95)$	127.9	127.63 (0.998)	115.23 (0.901)	119.862 (0.935)
	13	$\rho = 1.05$	$(3.4125, 5.775)$	122.3	122.22 (0.999)	108.97 (0.891)	113.36 (0.972)
	14	$\rho = 1.2$	$(3.9, 6.6)$	101.8	101.67 (0.999)	87.62 (0.861)	86.63 (0.851)

Next, we examine the performance of the $\mathcal{R}$ rule for different system sizes. Figure 4 compares the suggested $\mathcal{R}$ rule to the $c\mu/\theta$ rule. Using simulation, we calculated the average capacity allocated to each channel (i.e., the average number of patients in service) and the average profit for each policy. The circles present the optimal allocation and long-run profit according to the fluid solution. The results are presented for different system sizes. As N increases, we scale up the arrival rates proportionally using an appropriate Poisson process. In Example 1, the $\mathcal{R}$ rule allocation (which coincides with the fluid solution) suggests allocating a similar amount of capacity to each service channel (the capacity ratio is $(0.375, 0.34, 0.285)$ for the face-to-face, virtual and supplementary channels, respectively). The $c\mu/m$ rule, however, assigns most of the capacity to the supplementary and virtual channels, and relatively little capacity to the face-to-face channel (the capacity ratio is $(0.115, 0.425, 0.46)$). The $\mathcal{R}$ rule in this example achieves a 78% higher long-run average profit than the $c\mu/\theta$ rule. In Example 2, the $\mathcal{R}$ rule allocation (which again coincides with the fluid solution) suggests allocating most of the capacity to the face-to-face channel, and the remainder to the virtual and supplementary channels. The $c\mu/m$ rule, however, assigns most of the capacity to the virtual channel and then to the supplementary one. It allocates almost no capacity to the face-to-face channel. In terms of system design, following the $c\mu/\theta$ rule will lead to utilizing only the virtual and supplementary channels, and eliminating the face-to-face channel. The $\mathcal{R}$ rule achieves a 21% higher long-run average profit than the $c\mu/\theta$ rule. Note that to facilitate a relatively fair comparison, we do not consider the return penalty in these experiments (i.e., $\gamma = 0$).

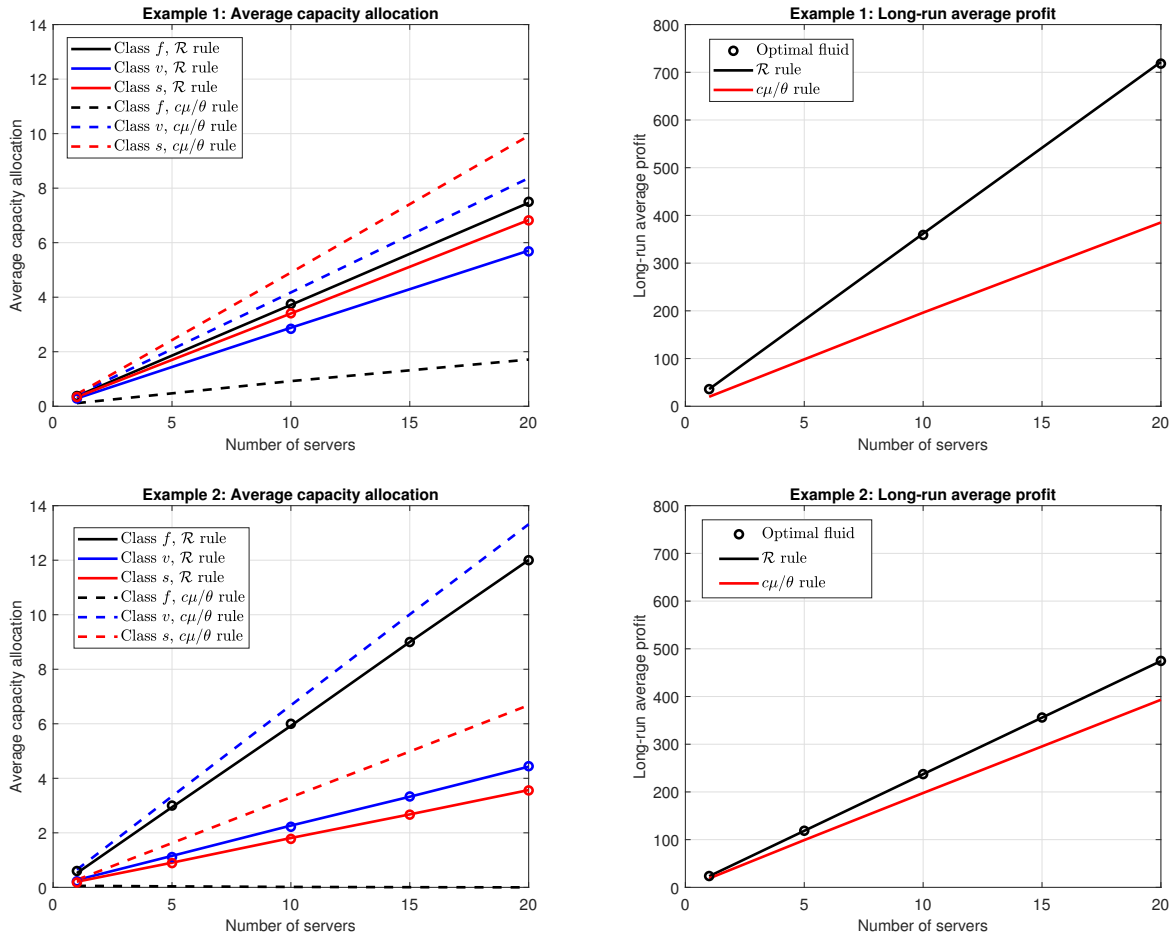
To demonstrate why the naive $\mathcal{R}$ rule is insufficient, especially in heavily loaded systems, we focus on a setting where the scheduling according to the $\mathcal{R}$ rule and the naive $\mathcal{R}$ rule are different. Specifically, let $N = 15$, $p_s = 0.8$, $\gamma = 0$, $\lambda_f = 45$, $\lambda_v = 90$; for classes $[f, v, s]$: $r = [12, 12, 8]$, $c = [2.5, 0.2, 1.5]$, $\mu = [4, 6, 6]$, $\theta = [0.8, 0.1, 0.3]$. In this example, the $\mathcal{R}$ rule allocates most of the capacity to the supplementary and virtual channels, while allocating very little capacity to the face-to-face channel. The naive $\mathcal{R}$ rule, however, allocates most of the capacity to the face-to-face channel. These two policies lead to a completely different system design. In this case, the long-run average profit under the $\mathcal{R}$ rule is 700, which is 15.7% higher than the 590 achieved by the naive $\mathcal{R}$ rule.

We can summarize these examples by first stating that the fluid approximation accurately describes the stochastic system. Second, we see that each policy can lead to a different prioritization and different system design. Third, we observe that the $\mathcal{R}$ rule achieves higher average profit compared to the other two policies for different system sizes.

4.1. Performance of the $\mathcal{R}$ Rule Under Transient Profit Maximization

In some settings, there could be random shocks that move the system far from its usual mode of operation. In these cases, improving the transient performance of the system becomes the main

Figure 4 Policy performance for different system sizes. The circles represent the optimal fluid solution. In Example 1, the parameters are: $\boldsymbol{\lambda_f = 1.5N}$, $\boldsymbol{\lambda_v = 2.5N}$, $\boldsymbol{r = [17.5, 15, 0]}$, $\boldsymbol{c = [2.5, 0.2, 1]}$, $\boldsymbol{\mu = [4, 6, 3]}$, $\boldsymbol{\theta = [0.12, 0.01, 0.03]}$ $\boldsymbol{p_s = 0.6}$, $\boldsymbol{\gamma = 0}$ and in Example 2: $\boldsymbol{\lambda_f = 2.4N}$, $\boldsymbol{\lambda_v = 4N}$, $\boldsymbol{r = [7, 6, 4]}$, $\boldsymbol{c = [2.5, 0.2, 1.2]}$, $\boldsymbol{\mu = [4, 6, 6]}$, $\boldsymbol{\theta = [0.8, 0.1, 0.3]}$ $\boldsymbol{p_s = 0.8}$ and $\boldsymbol{\gamma = 0}$.



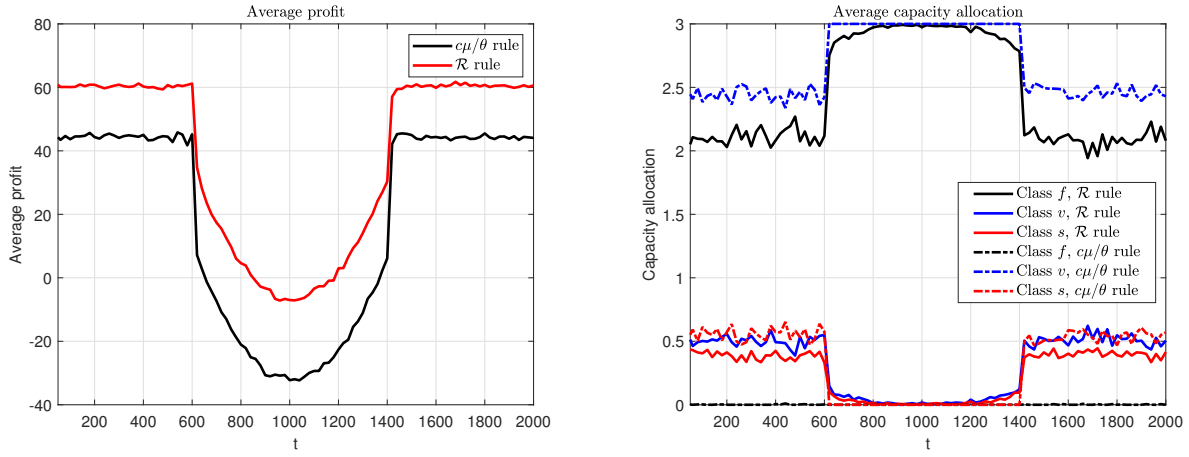
goal. That is, we want to find an effective scheduling policy to support demand surges. The Covid-19 pandemic, for example, caused sudden surges in demand on healthcare systems around the world. We consider the objective of maximizing the cumulative expected profit over a finite time horizon T. The fluid equivalent is, therefore,

$$\max_{z,q} \int_{t=0}^{T} \sum_{i=f,v,s} [r_i \mu_i z_i(t) - c_i q_i(t)] - \gamma p_s \mu_v z_v(t) \mathrm{d}t.$$

We observe through numerical experiments that even when the arrival rate is highly non-stationary, the $\mathcal{R}$ rule still performs very well in maximizing the transient profit compared to the $c\mu/\theta$ rule. Figure 5 presents a scenario in which at time $t = 600$, both classes experience a quadratic

surge in demand that lasts 800 time units. The left plot presents the instantaneous profit as a function of time over the time horizon $[0, 2000]$. The right plot presents the average number of patients in service for each class over the time horizon. During the surge in demand, both the $\mathcal{R}$ rule and the $c\mu/\theta$ rule become more extreme in their prioritization: the $\mathcal{R}$ rule allocates the entire capacity to the face-to-face channel, while the $c\mu/\theta$ allocates the entire capacity to the virtual channel, leaving almost no capacity for returning patients. In terms of the objective function, the $\mathcal{R}$ rule achieves twice the cumulative profit than the $c\mu/\theta$ rule. This result demonstrates the robustness of the $\mathcal{R}$ rule, even in maximizing the transient performance.

Figure 5 Policy's transient performance. The parameters are $\boldsymbol{N=3}$, $\boldsymbol{\lambda_f(t) = -3.88e^{-5}t^2 + 0.078t - 19.42}$ when $\boldsymbol{600 \le t \le 1400}$, and otherwise $\boldsymbol{\lambda_f(t) = 9}$. $\boldsymbol{\lambda_v(t) = -6.46e^{-5}t^2 + 0.13t - 32.37}$, when $\boldsymbol{600 \le t \le 1400}$, and otherwise $\boldsymbol{\lambda_f(t) = 15}$. $\boldsymbol{r = [7, 6, 4]}$, $\boldsymbol{c = [2.5, 0.2, 1]}$, $\boldsymbol{\mu = [4, 6, 6]}$, $\boldsymbol{\theta = [0.8, 0.1, 0.3]}$, $\boldsymbol{p_s = 0.8}$ and $\boldsymbol{\gamma = 0}$.



5. Model Extensions

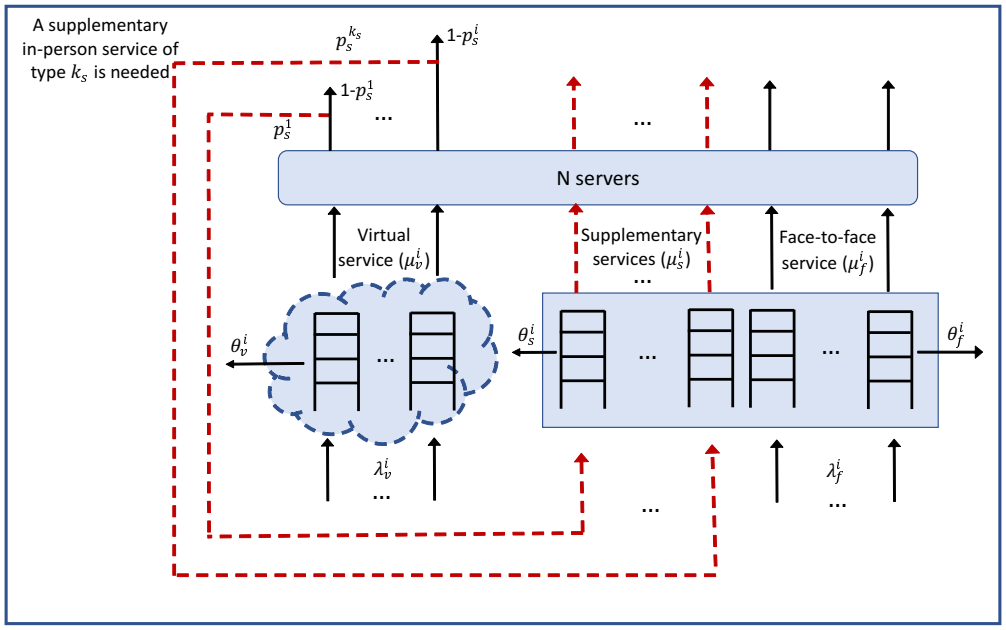
We now discuss two model extensions. The first considers multiple face-to-face, virtual and supplementary classes. The second extension refers to the virtual channel as a teletriage system that classifies patients according to the supplementary service they require (as opposed to a binary decision that we have considered thus far).

5.1. Multiple Classes in Each Service Channel

In this section we consider a more general setting with k_f face-to-face classes of patients, and k_v $(=k_s)$ virtual and supplementary classes, as illustrated in Figure 6. The different classes within each channel may represent different severity levels. Specifically, each face-to-face, virtual and

supplementary class is characterized by $\left(\lambda_f^i, \mu_f^i, \theta_f^i, c_f^i, r_f^i\right)$, $i = 1, \ldots, k_f$, $(\lambda_v^i, \mu_v^i, \theta_v^i, c_v^i, r_v^i, p_s^i, \gamma^i)$, $i = 1, \ldots, k_v$, and $(\mu_s^i, \theta_s^i, c_s^i, r_s^i)$, $i = 1, \ldots, k_v$, respectively. In total, the scheduling problem in this case needs to consider $k_f + 2k_v$ classes.

Figure 6 A hybrid system with multiple classes in each service channel.



The equivalent problem to (5), for which the optimal equilibrium point would maximize the long-run average profit, is:

$$
\begin{aligned}
\max_{\bar{q},\bar{z}} \quad & \sum_{j=f,v,s} \sum_{i=1}^{k_j} \left[r_j^i \mu_j^i \bar{z}_j^i - c_j^i \bar{q}_j^i \right] - \sum_{i=1}^{k_v} \gamma^i \mu_v^i p_s^i \bar{z}_v^i \\
\text{s.t.} \quad & \lambda_j^i = \mu_j^i \bar{z}_j^i + \theta_j^i \bar{q}_j^i, \quad j = f, v, \quad i = 1, \ldots, k_j; \quad \text{(first-time visitors)} \\
& p_s^i \mu_v^i \bar{z}_v^i = \mu_s^i \bar{z}_s^i + \theta_s^i \bar{q}_s^i; \quad i = 1, \ldots, k_v; \quad \text{(second-time visitors)} \\
& \sum_{j=f,v,s} \sum_{i=1}^{k_j} \bar{z}_j^i \le N; \\
& \bar{q}_j^i, \bar{z}_j^i \ge 0, \quad j = f, v, s, \quad i = 1, \ldots, k_j.
\end{aligned}
\tag{12}
$$

Rearranging (12) and omitting the constants yields the following:

$$\begin{aligned}
&\max_{\bar{z}} \sum_{j=f,v,s} \sum_{i=1,\dots k_j} \mathcal{R}_j^i \bar{z}_j^i \\
&\text{s.t.} \quad 0 \le \bar{z}_j^i \le \lambda_j^i/\mu_j^i, \quad j=f,v, \quad i=1,\dots,k_j; \quad \text{(first-time visitors)} \\
&\qquad 0 \le \bar{z}_s^i \le p_s^i \mu_v^i \bar{z}_v^i/\mu_s^i, \quad i=1,\dots,k_v; \quad \text{(second-time visitors)} \\
&\qquad \sum_{j=f,v,s} \sum_{i=1,\dots k_j} \bar{z}_j^i \le N,
\end{aligned} \tag{13}$$

where the $\mathcal{R}$ indexes are:

$$\begin{aligned}
&\mathcal{R}_f^i = \mu_f^i \left(r_f^i + c_f^i/\theta_f^i\right), \quad i=1,\dots,k_f; \\
&\mathcal{R}_v^i = \mu_v^i \left(r_v^i + c_v^i/\theta_v^i - p_s^i \left(\gamma^i + c_s^i/\theta_s^i\right)\right), \quad i=1,\dots,k_v; \\
&\mathcal{R}_s^i = \mu_s^i \left(r_s^i + c_s^i/\theta_s^i\right), \quad i=1,\dots,k_s.
\end{aligned} \tag{14}$$

We also have the equivalent $\mathcal{R}_{v,s}$ for each virtual/supplementary class:

$$\mathcal{R}_{v,s}^i = \frac{\mu_s^i}{\mu_s^i + p_s^i \mu_v^i} \mathcal{R}_v^i + \frac{p_s^i \mu_v^i}{\mu_s^i + p_s^i \mu_v^i} \mathcal{R}_s^i, \quad i=1,\dots,k_v.$$

Proving the optimality of a generalized form of the $\mathcal{R}$ rule in this setting requires us to follow the line of analysis conducted in Section 3. That is, we must first prove that under the generalized $\mathcal{R}$ rule, the fluid approximation converges to an equilibrium point that is a globally asymptotically stable one (a generalization of Theorem 1). Then, we must prove that the optimal solution to (13)–(14) is the globally asymptotically stable equilibrium (a generalization of Theorem 2). Utilizing this approach will quickly become prohibitively tedious with too many scenarios to consider. We, therefore, provide an algorithm that extends the essence of the $\mathcal{R}$ rule for setting the prioritization among classes. Note that for a given set of parameters, this algorithm needs to be run *once*.

The following algorithm utilizes the sorted set $\mathcal{S}$, including the classes' $\mathcal{R}$ indexes according to which the priority among classes is set.

Algorithm 1 *(The generalized $\mathcal{R}$ rule for multiple classes)*

1. *Set* $\mathcal{S} \leftarrow \left\{\mathcal{R}_f^i, i = 1, \ldots k_f\right\}$
2. *For each Class* i, $i = 1, \ldots, k_v$
   (a) *If* $\mathcal{R}_s^i < \mathcal{R}_v^i$, *then* $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{R}_v^i, \mathcal{R}_s^i\}$
   (b) *Otherwise,* $\mathcal{S} \leftarrow \mathcal{S} \cup \left\{\mathcal{R}_{v,s}^i\right\}$
3. *Sort the set* $\mathcal{S}$ *in a decreasing order*
4. *Replace the* $\mathcal{R}_{v,s}^i$*'s in* $\mathcal{S}$ *with* $\{\mathcal{R}_s^i, \mathcal{R}_v^i\}$
5. *Return* $\mathcal{S}$

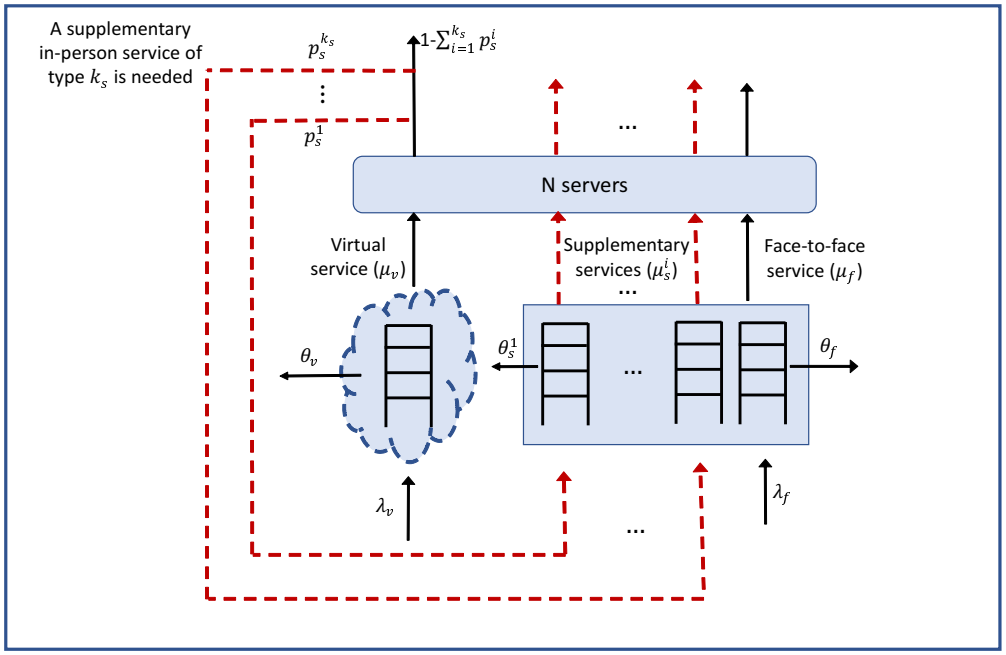
The prioritization of classes will be done according to their order in the sorted set $\mathcal{S}$. The algorithm shares the same principles as the $\mathcal{R}$ rule presented in Section 3: if the index of the virtual class is higher than the index of the supplementary class, then the prioritization is set according to the $\mathcal{R}$ indexes. If, however, the index of the supplementary class is higher than the index of the virtual class, we jointly prioritize the virtual and supplementary classes according to their integrated $\mathcal{R}_{v,s}$ index.

5.2. A Teletriage System with Multiple Supplementary Services

Thus far, we assumed that each virtual service can lead to one type of supplementary service. Nevertheless, one common strategy for controlling healthcare needs, which is called "forward triage", refers to the online channel as a sorting stage offered to patients. It allows them to be efficiently screened before being referred to a medical center. Respiratory symptoms, which may be early signs of Covid-19, for example, can commonly be evaluated using this approach (Hollander and Carr 2020).

Motivated by such a teletriage setting, we study a model extension in which, based on an initial virtual assessment, patients are classified according to the supplementary service they require. The supplementary services vary in their urgency, length and/or cost. To this end, we consider k_s optional supplementary services for each virtual service. Each supplementary service i occurs with probability p_s^i, $i = 1, \ldots, k_s$, and is associated with a supplementary class that is characterized by $(p_s^i, \mu_s^i, \theta_s^i, c_s^i, r_s^i, \gamma_i)$, as illustrated in Figure 7.

Figure 7 A hybrid system with multiple supplementary services.



The equivalent problem to (5), for which the optimal equilibrium point would maximize the long-run average profit, is:

$$
\begin{aligned}
\max_{\bar{q},\bar{z}} \ & \sum_{j=f,v} \left[r_j \mu_j \bar{z}_j - c_j \bar{q}_j \right] + \sum_{i=1}^{k_s} \left[r_s^i \mu_s^i \bar{z}_s^i - c_s^i \bar{q}_s^i - \gamma^i \mu_v p_s^i \bar{z}_v \right] \\
\text{s.t.} \ & \lambda_i = \mu_i \bar{z}_i + \theta_i \bar{q}_i, \quad i = f, v; \quad \text{(first-time visitors)} \\
& p_s^i \mu_v \bar{z}_v = \mu_s^i \bar{z}_s^i + \theta_s \bar{q}_s^i; \quad i = 1, \ldots, k_v; \quad \text{(second-time visitors)} \\
& \sum_{j=f,v} \bar{z}_j + \sum_{i=1}^{k_s} \bar{z}_s^i \le N; \\
& \bar{q}_j, \bar{z}_j, \bar{q}_s^i, \bar{z}_s^i \ge 0, \quad j = f, v, \quad i = 1, \ldots, k_v.
\end{aligned}
\tag{15}
$$

Rearranging (15) yields the following:

$$\begin{aligned}
&\max_{\bar{z}} \sum_{j=f,v} \mathcal{R}_j \bar{z}_j + \sum_{i=1}^{k_s} \mathcal{R}_s^i \bar{z}_s^i \\
&\text{s.t.} \quad 0 \le \bar{z}_j \le \lambda_j/\mu_j, \quad j = f, v; \quad \text{(first-time visitors)} \\
&\qquad 0 \le \bar{z}_s^i \le p_s^i \mu_v \bar{z}_v/\mu_s^i, \quad i = 1, \ldots, k_v; \quad \text{(second-time visitors)} \\
&\qquad \sum_{j=f,v} \bar{z}_j + \sum_{i=1}^{k_s} \mathcal{R}_s^i \bar{z}_s^i \le N,
\end{aligned} \tag{16}$$

where the $\mathcal{R}$ indexes are:

$$\begin{aligned}
&\mathcal{R}_f = \mu_f \left(r_f + c_f/\theta_f\right); \\
&\mathcal{R}_v = \mu_v \left(r_v + c_v/\theta_v - \sum_{i=1}^{k_s} p_s^i \left(\gamma^i + c_s^i/\theta_s^i\right)\right); \\
&\mathcal{R}_s^i = \mu_s^i \left(r_s^i + c_s^i/\theta_s^i\right), \quad i = 1, \ldots, k_s.
\end{aligned} \tag{17}$$

We denote the joint virtual/supplementary index for a set of supplementary classes, $\mathcal{B}$:

$$\mathcal{R}_{v,s}^{\mathcal{B}} = \frac{\sum_{i\in\mathcal{B}} \mu_s^i}{\sum_{i\in\mathcal{B}} \mu_s^i + \mu_v \sum_{i\in\mathcal{B}} p_s^i} \mathcal{R}_v + \sum_{i\in\mathcal{B}} \frac{p_s^i \mu_v}{\sum_{i\in\mathcal{B}} \mu_s^i + \mu_v \sum_{i\in\mathcal{B}} p_s^i} \mathcal{R}_s^i. \tag{18}$$

The $\mathcal{R}_{v,s}^{\mathcal{B}}$ index is an extension of the $\mathcal{R}_{v,s}$ index that was used when each virtual service could lead to a single type of supplementary service. The interpretation of the index remains the same – it is the weighted average of the virtual and supplementary $\mathcal{R}$ indexes. When $\mathcal{B}$ includes one supplementary service, we retrieve the original $\mathcal{R}_{v,s}$.

As stated in Section 5.1, proving the optimality of a generalized form of the $\mathcal{R}$ rule in this setting requires us to follow the line of analysis conducted in Section 3. Utilizing this approach will quickly become prohibitively tedious. Therefore, we provide a heuristic algorithm in the spirit of the $\mathcal{R}$ rule for setting the prioritization among classes. Note that for a given set of parameters, this algorithm needs to be run *once*.

The following algorithm uses two sorted sets: $\mathcal{S}$ and $\mathcal{B}$. The set $\mathcal{S}$ includes the classes' $\mathcal{R}$ indexes according to which the priority among classes is set. The set $\mathcal{B}$ includes the classes for which their $\mathcal{R}_s^i$ index is larger than $\mathcal{R}_v$.

Algorithm 2 *(The generalized $\mathcal{R}$ rule for multiple supplementary services)*

1. *Set* $\mathcal{S} \leftarrow \{\mathcal{R}_f\}$ *and* $\mathcal{B} = \{v\}$
2. *For each Class* i, $i = 1, \ldots, k_s$
   (a) *If* $\mathcal{R}_s^i < \mathcal{R}_v$, *then* $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathcal{R}_s^i\}$
   (b) *Otherwise,* $\mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
3. *Calculate* $\mathcal{R}_{v,s}^{\mathcal{B}}$ *according to (18)*
4. *Sort in a decreasing order the set* $\mathcal{S} \cup \mathcal{R}_{v,s}^{\mathcal{B}}$ *and the set* $\mathcal{B}$ *according to the* $\mathcal{R}$ *indexes*
5. *Replace the index* $\mathcal{R}_{v,s}^{\mathcal{B}}$ *in* $\mathcal{S}$ *with the* $\mathcal{R}$ *indexes of the classes in* $\mathcal{B}$
6. *Return* $\mathcal{S}$

The prioritization of classes will be done according to their order in the sorted set $\mathcal{S}$. The ideas behind the algorithm are the same as for the $\mathcal{R}$ rule presented in Section 3: The supplementary classes whose $\mathcal{R}$ index is smaller than $\mathcal{R}_v$ can be prioritized as any other class. The supplementary classes whose $\mathcal{R}$ index is larger than $\mathcal{R}_v$ need to be considered jointly with the virtual class and the other supplementary classes by using the joint $\mathcal{R}_{v,s}^{\mathcal{B}}$ index.

6. Concluding Remarks and Future Directions

Motivated by healthcare provision trends, in this paper we study the optimal scheduling and capacity allocation for multi-server queues where patients may return for supplementary service. The main motivating example for this work is a hybrid healthcare setting that provides three service channels: face-to-face, virtual and in-person supplementary services that some patients require following virtual service. The strong dependency between the virtual and returning patients (i.e., the former constitute the feeding source for the latter) imposes additional constraints when scheduling and allocating capacity. Using a fluid relaxation approach, we derive and prove the optimality of the $\mathcal{R}$ index rule for maximizing the long-run average profit. From an operational point of view, the $\mathcal{R}$ rule helps prioritize classes (i.e., which class to admit when a service provider becomes available). From a design perspective, the $\mathcal{R}$ rule allocates capacity for each service channel. We show that the $\mathcal{R}$ rule performs very close to optimal, and significantly better than other known policies, in various settings and under different system loads. Lastly, we show that even

though the $\mathcal{R}$ rule is designed to maximize long-run average profit, it also performs well in a transient time-horizon and non-stationary arrival scenario.

We identify a few interesting future research directions. The first is to consider non-stationary systems with time-varying arrival rates. Indeed, urgent care centers often experience such arrival patterns and peak hours where the demand is much higher than at other times of the day (Armony et al. 2015). Deriving an effective robust scheduling policy under arbitrary time-varying arrival rates is challenging, since the optimal policy may depend on these time-varying arrival rates as well as on the system's state. Moreover, it is plausible that the arrival rates to the virtual and in-person channels are not synchronized: at some hour during the day, patients may prefer the virtual channel, while at other hours, patients may prefer the in-person channel; these patterns, in turn, also affect the supplementary channel.

The second direction is related to improved continuity of care. That is, allow patients to see the same physician in their virtual visit and supplementary in-person visit, when needed. Specifically, there would be two types of queues to consider: The first is a joint queue for all physicians and first-time visitors. The second type of queue, for second-time visitors, would be separate for each physicians. The queue management and scheduling policy would have to take into account both queue types.

Another interesting direction is to incorporate strategic behavior. On the one hand, the patient chooses which service channel to use: face-to-face or virtual, according to the expected waiting cost and return probability. On the other hand, by considering patients' behavior, the healthcare provider chooses how to allocate capacity among the three services in order to maximize its profits. Another interesting direction is to focus on reimbursement policies (e.g., fee-for-service vs. bundled payment) for the different service channels. Through these reimbursement policies, virtual services can be encouraged or discouraged in order to optimize system performance and healthcare provision.

References

Ahmed, S., K. Sanghvi, D. Yeo. 2020. Telemedicine takes centre stage during COVID-19 pandemic. *BMJ Innovations* **6**(4). 2

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194. 29

Ashwood, J.S., A. Mehrotra, D. Cowling, L. Uscher-Pines. 2017. Direct-to-consumer telehealth may increase access to care but does not decrease spending. *Health Affairs* **36**(3) 485–491. 2

Atar, R., C. Giat, N. Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439. 3, 6, 12, 13, 17

Atar, R., A. Mandelbaum, M.I. Reiman. 2004. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **14**(3) 1084–1134. 6

Bavafa, H., L.M. Hitt, C. Terwiesch. 2018. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* **64**(12) 5461–5480. 2

Bavafa, H., S. Savin, C. Terwiesch. 2021. Customizing primary care delivery using e-visits. *Production and Operations Management* **30**(11) 4306–4327. 5

Bokolo, A.J. 2020. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. *Journal of Medical Systems* **44**(7) 1–9. 1

Çakıcı, Ö.E., A.F. Mills. 2021. On the role of teletriage in healthcare demand management. *Manufacturing & Service Operations Management* **23**(6) 1483–1504. 2, 5

Chauhan, V., S. Galwankar, B. Arquilla, M. Garg, S. Di Somma, A. El-Menyar, V. Krishnan, J. Gerber, R. Holland, S.P. Stawicki. 2020. Novel coronavirus (COVID-19): Leveraging telemedicine to optimize care while minimizing exposures and viral transmission. *Journal of Emergencies, Trauma, and Shock* **13**(1) 20. 2

Cox, D.R., W.L. Smith. 1961. Queues. *Methuen, London* . 6

Dai, J.G., W. Lin. 2005. Maximum pressure policies in stochastic processing networks. *Operations Research* **53**(2) 197–218. 17

Doshi, A., Y. Platt, J.R. Dressen, B.K. Mathews, J.C. Siy. 2020. Keep calm and log on: Telemedicine for COVID-19 pandemic response. *Journal of Hospital Medicine* **15**(5) 302–304. 2

Harrison, J.M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research* **52**(2) 243–257. 6

Hollander, J.E., B.G. Carr. 2020. Virtually perfect? Telemedicine for COVID-19. *New England Journal of Medicine* **382**(18) 1679–1681. 2, 25

Hu, Y., C.W. Chan, J. Dong. 2022. Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science, forthcoming* . 6, 8, 15, 33

Huang, J., B. Carmeli, A. Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908. 6

Hur, J., M.C. Chang. 2020. Usefulness of an online preliminary questionnaire under the COVID-19 pandemic. *Journal of Medical Systems* **44** 1–2. 1

Kadir, M.A. 2020. Role of telemedicine in healthcare during COVID-19 pandemic in developing countries. *Telehealth and Medicine Today* . 1

Long, Z., N. Shimkin, H. Zhang, J. Zhang. 2020. Dynamic scheduling of multiclass many-server queues with abandonment: The generalized cμ/h rule. *Operations Research* **68**(4) 1218–1230. 6

Mandelbaum, A., A.L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855. 6

McConnochie, K.M., S.D. Ronis, N.E. Wood, P.K. Ng. 2015. Effectiveness and safety of acute care telemedicine for children with regular and special healthcare needs. *Telemedicine and e-Health* **21**(8) 611–621. 2

Monaghesh, E., A. Hajizadeh. 2020. The role of telehealth during COVID-19 outbreak: A systematic review based on current evidence. *BMC Public Health* **20**(1) 1–9. 1

O'Reilly-Jacob, M., P. Mohr, M. Ellen, C. Petersen, C. Sarkisian, S. Attipoe, E. Rich. 2021. Digital health & low-value care. *Healthcare*, vol. 9. Elsevier, 100533. 2

Papadimitriou, C.H., J.N. Tsitsiklis. 1999. The complexity of optimal queuing network control. *Mathematics of Operations Research* **24**(2) 293–305. 3, 9

Puha, A.L., A.R. Ward. 2019. Scheduling an overloaded multiclass many-server queue with impatient customers. *Operations Research & Management Science in the Age of Analytics*. INFORMS, 189–217. 6

Rajan, B., T. Tezcan, A. Seidmann. 2019. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* **65**(3) 1236–1267. 5

Shi, Z., A. Mehrotra, C.A. Gidengil, S.J. Poon, L. Uscher-Pines, K.N Ray. 2018. Quality of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. *Health Affairs* **37**(12) 2014–2023. 2

Stolyar, A. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* **14**(1) 1–53. 17

Uscher-Pines, L., R. Malsberger, L. Burgette, A. Mulcahy, A. Mehrotra. 2016. Effect of teledermatology on access to dermatology care among medicaid enrollees. *JAMA dermatology* **152**(8) 905–912. 2

Van Mieghem, J.A. 1995. Dynamic scheduling with convex delay costs: The generalized cμ rule. *The Annals of Applied Probability* 809–833. 6

Whitt, W. 2002. Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues. *Space* **500** 391–426. 9

Wong, M.Y.Z., D.V. Gunasekeran, S. Nusinovici, C. Sabanayagam, K. K. Yeo, C.-Y. Cheng, Y.-C. Tham. 2021. Telehealth demand trends during the COVID-19 pandemic in the top 50 most affected countries: Infodemiological evaluation. *JMIR public health and surveillance* **7**(2) e24445. 1

Zychlinski, N., C.W. Chan, J. Dong. 2020. Scheduling queues with simultaneous and heterogeneous requirements from multiple types of servers. *2020 Winter Simulation Conference (WSC)*. IEEE, 2365–2376. 6

Zychlinski, N., C.W Chan, J. Dong. 2022. Managing queues with different resource requirements. *Operations Research, forthcoming* . 6, 15

Zychlinski, Noa. 2022. Applications of fluid models in service operations management. *Queueing Systems* 1–25. 9

Appendix A: Proof of Theorem 1.

The proof follows a similar line of arguments for each case in Table 1. Therefore, we only present the proof for Case 1a as representative of Cases 1b and 1c, as part of the naive $\mathcal{R}$ rule, and Case 2a as representative of Case 2b, as part of the two-step $\mathcal{R}$ rule. Since the rest of the cases follow similarly, we omitted them. Our proof, which is based on the construction of a Lyapunov function, resembles the proof of Theorem 4 in Hu et al. (2022).

A.1. Case 1a: $\mathcal{R}_f < \mathcal{R}_s < \mathcal{R}_v$

In this case we consider the four sub-cases described in Table 3. For each sub-case we prove that the globally asymptotically stable equilibrium $\bar{q}_f, \bar{q}_v, \bar{q}_s$ is as it appears in the table. In this case, the $\mathcal{R}$ rule gives strict priority to Class v, then to Class r, and finally to Class f.

Table 3 Globally asymptotically stable equilibria – Case 1a.

Sub-case	$\bar{q}_f$	$\bar{q}_v$	$\bar{q}_s$
I. $\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s} N \leq N$	0	0	0
II. $\frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} \leq N < \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s} N$	$\frac{\mu_f}{\theta_f}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s} - N\right]^+$	0	0
III. $\frac{\lambda_v}{\mu_v} \leq N < \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s}$	$\frac{\lambda_f}{\theta_f}$	0	$\frac{1}{\theta_s}\left(\lambda_v p_s - \mu_s\left(N - \frac{\lambda_f}{\mu_v}\right)\right)$
IV. $N < \frac{\lambda_v}{\mu_v}$.	$\frac{\lambda_f}{\theta_f}$	$\frac{\lambda_v - \mu_v N}{\theta_v}$	$\frac{p_s\mu_v N}{\theta_s}$

$x^+ = \max(x, 0)$.

- **Sub-case 1a-I.** $\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s} N \leq N$. We consider the Lyapunov function

$$V(q) = \sum_{i=f,v,s} \frac{1}{\mu_i} |q_i - \bar{q}_i|,$$

where the equilibrium point $\bar{q} = (0, 0, 0)$, and show its asymptotic stability. To this end, we first verify that $V(\bar{q}) = 0$ and $V(q) \to \infty$ as $\|q\| \to \infty$. Then, we show that $\nabla_q V(q)^T f(q) < 0$ for $q \neq \bar{q}$, where $\dot{q}(t) = f(q(t))$, as defined in (8).

— **When $q_v(t) > 0$**, all capacity is allocated to Class v. Specifically, the system dynamics in (4) are as follows:

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v N - \theta_v q_v(t); \\ \dot{q}_s(t) = p_s\mu_v N - \theta_s q_s(t); \\ \dot{q}_f(t) = \lambda_f - \theta_f q_f(t). \end{cases}$$

We have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_v}\left(\lambda_v - \mu_v N - \theta_v q_v(t)\right) + \frac{1}{\mu_s}\left(p_s\mu_v N - \theta_s q_s(t)\right) + \frac{1}{\mu_f}\left(\lambda_f - \theta_f q_f(t)\right) \\ &= \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - N\left(1 - \frac{p_s\mu_v}{\mu_s}\right) - \frac{\theta_v q_v(t)}{\mu_v} - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f} < 0,\end{aligned}$$

where the inequality comes from the sub-case's condition and the assumption that $\theta > 0$.

— **When $q_v(t) = 0$ and $q_s(t) > 0$**, the required capacity for Class v is allocated, and any leftover capacity is allocated to Class r. The system dynamics are therefore:

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v \tilde{z}_v; \\ \dot{q}_s(t) = p_s\mu_v\tilde{z}_v - \mu_s\left(N - \tilde{z}_v\right) - \theta_s q_s(t); \\ \dot{q}_f(t) = \lambda_f - \theta_f q_f(t), \end{cases}$$

where $\tilde{z}_v = \left(\frac{\lambda_v}{\mu_v} \wedge N\right)$.

We have,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_v}\left(\lambda_v - \mu_v\tilde{z}_v\right) + \frac{1}{\mu_s}\left(p_s\mu_v\tilde{z}_v - \mu_s\left(N - \tilde{z}_v\right) - \theta_s q_s(t)\right) + \frac{1}{\mu_f}\left(\lambda_f - \theta_f q_f(t)\right) \\ &= \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - N + \frac{\mu_v p_s}{\mu_s}\tilde{z}_v - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f} \\ &\leq \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - N + \frac{\mu_v p_s}{\mu_s}N - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f} < 0,\end{aligned}$$

where the first inequality comes from the fact that $\tilde{z}_v = (\lambda_v/\mu_v \wedge N) \leq N$; the last inequality comes from the sub-case's condition and the assumption that $\theta > 0$.

— **When $q_v(t) = 0$, $q_s(t) = 0$ and $q_f(t) > 0$**, the required capacity to Class v and then r is allocated; any other left capacity is allocated to Class f. The system dynamics are as follows:

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v\tilde{z}_v; \\ \dot{q}_s(t) = p_s\mu_v\tilde{z}_v - \mu_s\left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge \left(N - \tilde{z}_v\right)\right); \\ \dot{q}_f(t) = \lambda_f - \mu_f\left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge \left(N - \tilde{z}_v\right)\right)\right)\right) - \theta_f q_f(t); \end{cases}$$

where as before, $\tilde{z}_v = \left(\frac{\lambda_v}{\mu_v} \wedge N\right)$.

We have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_v}\left(\lambda_v - \mu_v \tilde{z}_v\right) + \frac{1}{\mu_s}\left(p_s\mu_v\tilde{z}_v - \mu_s\left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right) \\ &\quad + \frac{1}{\mu_f}\left(\lambda_f - \mu_f\left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right) - \theta_f q_f(t)\right) \\ &= \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - \tilde{z}_v + \frac{p_s\mu_v}{\mu_s}\tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right) \\ &\quad - \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right) - \frac{\theta_f q_f(t)}{\mu_f}.\end{aligned}$$

If $\tilde{z}_v = \lambda_v/\mu_v$, we have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{\lambda_f}{\mu_f} - \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \frac{\lambda_v}{\mu_v} - \frac{p_s\lambda_v}{\mu_s}\right)\right) - \frac{\theta_f q_f(t)}{\mu_f} \\ &= \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} - N\right]^+ - \frac{\theta_f q_f(t)}{\mu_f} \\ &\le \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s}N - N\right]^+ - \frac{\theta_f q_f(t)}{\mu_f} < 0.\end{aligned}$$

The first equality comes from the fact that when $\lambda_v \le N\mu_v$, we have $p_s\lambda_v/\mu_s \le N - \lambda_v/\mu_v$. This is because $\lambda_v \le N\mu_v\mu_s\left(p_s\mu_v + \mu_s\right)$, which is equivalent to $\lambda_v\left(p_s/\mu_s + 1/\mu_v\right) \le N$. The inequality comes from the sub-case's condition and the assumption that $\theta > 0$.

If $\tilde{z}_v = N$, we have

$$\nabla_q V(q)^T f(q) = \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - N + \frac{p_s\mu_v}{\mu_s}N - \frac{\theta_f q_f(t)}{\mu_f} < 0,$$

where the inequality come from the sub-case's condition and the assumption that $\theta > 0$.

- **Sub-case 1a-II.** $\frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} \le N < \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s}N$. We consider the Lyapunov function

$$V(q) = \sum_{i=f,v,s}\frac{1}{\mu_i}|q_i - \bar{q}_i|,$$

where the equilibrium point $\bar{q} = \left(\frac{\mu_f}{\theta_f}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s} - N\right]^+,\ 0,\ 0\right)$, and show its asymptotic stability.

—**When $q_v(t) \ge \bar{q}_v$, $q_s(t) \ge \bar{q}_s$, $q_f(t) \ge \bar{q}_f$, and $q(t) \ne \bar{q}$,**

The system dynamics are as follows:

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v\tilde{z}_v - \theta_v q_v(t); \\ \dot{q}_s(t) = p_s\mu_v\tilde{z}_v - \mu_s\left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right) - \theta_s q_s(t); \\ \dot{q}_f(t) = \lambda_f - \mu_f\left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right) - \theta_f q_f(t). \end{cases}$$

We have,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_v}\left(\lambda_v - \mu_v\tilde{z}_v - \theta_v q_v(t)\right) + \frac{1}{\mu_s}\left(p_s\mu_v\tilde{z}_v - \mu_s\left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right) - \theta_s q_s(t)\right)\\ &\quad + \frac{1}{\mu_f}\left(\lambda_f - \mu_f\left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right) - \theta_f q_f(t)\right)\\ &= \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} - \tilde{z}_v + \frac{p_s\mu_v}{\mu_s}\tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right) - \frac{\theta_f q_v(t)}{\mu_v} - \frac{\theta_s q_s(t)}{\mu_f} - \frac{\theta_f q_f(t)}{\mu_f}\\ &\quad - \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right).\end{aligned}$$

If $\tilde{z}_v = \lambda_v/\mu_v$, we have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{\lambda_f}{\mu_f} - \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \frac{\lambda_v}{\mu_v} - \frac{p_s\lambda_v}{\mu_s}\right)\right) - \frac{\theta_v q_v(t)}{\mu_v} - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f}\\ &= \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} - N\right]^+ - \frac{\theta_v q_v(t)}{\mu_v} - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f}\\ &< \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} - N\right]^+ - \frac{\theta_v \bar{q}_v}{\mu_v} - \frac{\theta_s \bar{q}_s}{\mu_s} - \frac{\theta_f \bar{q}_f}{\mu_f} = 0,\end{aligned}$$

where the first inequality comes from the fact that $q_v(t) \geq \bar{q}_v,\ q_s(t) \geq \bar{q}_s,\ q_f(t) \geq \bar{q}_f$, and $q(t) \neq \bar{q}$.

If $\tilde{z}_v = N$, we have

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s}N - N - \frac{\theta_v q_v(t)}{\mu_v} - \frac{\theta_s q_s(t)}{\mu_s} - \frac{\theta_f q_f(t)}{\mu_f}\\ &< \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} - N - \frac{\theta_v \bar{q}_v}{\mu_v} - \frac{\theta_s \bar{q}_s}{\mu_s} - \frac{\theta_f \bar{q}_f}{\mu_f} = 0,\end{aligned}$$

where the inequality comes first from the fact that when $\tilde{z}_v = N$, $N\mu_v < \lambda_c$; then from the fact that $q_v(t) \geq \bar{q}_v,\ q_s(t) \geq \bar{q}_s,\ q_f(t) \geq \bar{q}_f$, and $q(t) \neq \bar{q}$.

— **When $q_v(t) < \bar{q}_v,\ q_s(t) < \bar{q}_s,\ q_f(t) < \bar{q}_f$, and $q(t) \neq \bar{q}$**, due to the absolute value in the Lyapunov function, we get the same $\nabla_q V(q)^T f(q)$ as in the previous case only with a negative sign and, therefore,

$$\begin{aligned}\nabla_q V(q)^T f(q) < &-\frac{\lambda_f}{\mu_f} - \frac{\lambda_v}{\mu_v} + \tilde{z}_v - \frac{p_s\mu_v}{\mu_s}\tilde{z}_v + \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right) + \frac{\theta_f \bar{q}_v}{\mu_v} + \frac{\theta_s \bar{q}_s}{\mu_f} + \frac{\theta_f \bar{q}_f}{\mu_f}\\ &+ \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \tilde{z}_v - \left(\frac{p_s\mu_v}{\mu_s}\tilde{z}_v \wedge (N-\tilde{z}_v)\right)\right)\right).\end{aligned}$$

From here the proof follows the same line of arguments as in previous case. The other four options for the different relations between $q_i(t)$ and $\bar{q}_i$ are handled in the exact same way and, therefore, are omitted.

- **Sub-case 1a-III.** $\frac{\lambda_v}{\mu_v} \leq N < \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s}$. We consider the Lyapunov function

$$V(q) = \sum_{i=f,v,s} |q_i - \bar{q}_i|,$$

where the equilibrium point $\bar{q} = (\bar{q}_v,\ \bar{q}_s,\ \bar{q}_f) = (0,\ (p_s\lambda_v - \mu_s\left(N - \lambda_v/\mu_v\right))/\theta_s,\ \lambda_f/\theta_f)$, and show its asymptotic stability.

—**When $q_v(t) \geq \bar{q}_v$, $q_s(t) \geq \bar{q}_s$, $q_f(t) \geq \bar{q}_f$, and $q(t) \neq \bar{q}$**, the system dynamics are:

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v\frac{\lambda_v}{\mu_v} - \theta_v q_v(t) = -\theta_v q_v(t); \\ \dot{q}_s(t) = p_s\lambda_v - \mu_s\left(N - \frac{\lambda_v}{\mu_v}\right) - \theta_s q_s(t); \\ \dot{q}_f(t) = \lambda_f + p_s\lambda_v - \theta_f q_f(t). \end{cases}$$

We, therefore, have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \lambda_f + p_s\lambda_v - \mu_s\left(N - \frac{\lambda_v}{\mu_v}\right) - \theta_v q_v(t) - \theta_s q_s(t) - \theta_f q_f(t) \\ &< \lambda_f + p_s\lambda_v - \mu_s\left(N - \frac{\lambda_v}{\mu_v}\right) - \theta_v\bar{q}_v - \theta_s\bar{q}_s - \theta_f\bar{q}_f = 0, \end{aligned}$$

where the inequality comes from the fact that $q_v(t) \geq \bar{q}_v,\ q_s(t) \geq \bar{q}_s,\ q_f(t) \geq \bar{q}_f$, and $q(t) \neq \bar{q}$.

—**When $q_v(t) < \bar{q}_v$, $q_s(t) < \bar{q}_s$, $q_f(t) < \bar{q}_f$, and $q(t) \neq \bar{q}$**, we get the same $\nabla_q V(q)^T f(q)$ as in the previous case with a negative sign; namely,

$$\begin{aligned} \nabla_q V(q)^T f(q) &= -\lambda_f - p_s\lambda_v + \mu_s\left(N - \frac{\lambda_v}{\mu_v}\right) + \theta_v q_v(t) + \theta_s q_s(t) + \theta_f q_f(t) \\ &< -\lambda_f - p_s\lambda_v + \mu_s\left(N - \frac{\lambda_v}{\mu_v}\right) + \theta_v\bar{q}_v + \theta_s\bar{q}_s + \theta_f\bar{q}_f = 0, \end{aligned}$$

where the inequality comes from the fact that $q_v(t) > \bar{q}_v,\ q_s(t) > \bar{q}_s,\ q_f(t) > \bar{q}_f$, and $q(t) \neq \bar{q}$. The other four options for the different relations between $q_i(t)$ and $\bar{q}_i$ are handled in the exact same way and, therefore, are omitted.

- **Sub-case 1a-IV.** $N < \frac{\lambda_v}{\mu_v}$. We consider the Lyapunov function

$$V(q) = \sum_{i=f,v,s} |q_i - \bar{q}_i|,$$

where the equilibrium point $\bar{q} = (\bar{q}_v,\ \bar{q}_s,\ \bar{q}_f) = ((\lambda_v - \mu_v N)/\theta_v,\ p_s\mu_v N/\theta_s,\ \lambda_f/\theta_f)$, and show its asymptotic stability. Since the conditions $V(\bar{q}) = 0$ and $V(q) \to \infty$ as $\|q\| \to \infty$ can easily be verified, we focus on showing that $\nabla_q V(q)^T f(q) < 0$ for $q \neq \bar{q}$.

— **When $q_v(t) \geq \bar{q}_v,\ q_s(t) \geq \bar{q}_s,\ q_f(t) \geq \bar{q}_f,$ and $q(t) \neq \bar{q}$**, all capacity is allocated to Class v. The system dynamics are therefore,

$$\begin{cases} \dot{q}_v(t) = \lambda_v - \mu_v N - \theta_v q_v(t); \\ \dot{q}_s(t) = p_s \mu_v N - \theta_s q_s(t); \\ \dot{q}_f(t) = \lambda_f - \theta_f q_f(t); \end{cases}$$

We have

$$\begin{aligned} \nabla_q V(q)^T f(q) &= \lambda_v - \mu_v N - \theta_v q_v(t) + p_s \mu_v N - \theta_s q_s(t) + \lambda_f - \theta_f q_f(t) \\ &= \lambda_f + \lambda_v - \mu_v N + p_s \mu_v N - \theta_v q_v(t) - \theta_s q_s(t) - \theta_f q_f(t) \\ &< \lambda_f + \lambda_v - \mu_v N + p_s \mu_v N - \theta_v \bar{q}_v - \theta_s \bar{q}_s - \theta_f \bar{q}_f = 0, \end{aligned}$$

where the inequality comes from the fact that $q_v(t) \geq \bar{q}_v,\ q_s(t) \geq \bar{q}_s,\ q_f(t) \geq \bar{q}_f$, and $q(t) \neq \bar{q}$.

— **When $q_v(t) < \bar{q}_v,\ q_s(t) < \bar{q}_s,\ q_f(t) < \bar{q}_f,$ and $q(t) \neq \bar{q}$**, we have,

$$\begin{aligned} \nabla_q V(q)^T f(q) &= -\left(\lambda_v - \mu_v N - \theta_v q_v(t)\right) - \left(p_s \mu_v N - \theta_s q_s(t)\right) - \left(\lambda_f - \theta_f q_f(t)\right) \\ &= -\lambda_f - \lambda_v + \mu_v N - p_s \mu_v N + \theta_v q_v(t) + \theta_s q_s(t) + \theta_f q_f(t) \\ &< -\lambda_f - \lambda_v + \mu_v N - p_s \mu_v N + \theta_v \bar{q}_v + \theta_s \bar{q}_s + \theta_f \bar{q}_f = 0, \end{aligned}$$

where the inequality comes from the fact that $q_v(t) < \bar{q}_v,\ q_s(t) < \bar{q}_s,\ q_f(t) < \bar{q}_f$.

The other four options for the different relations between $q_i(t)$ and $\bar{q}_i$ are handled in the exact same way and, therefore, are omitted.

A.2. Case 2a: $\mathcal{R}_f < \mathcal{R}_{v,s}$ ($R_v < R_s$)

In this case we consider the four sub-cases described in Table 4. For each sub-case we prove that the globally asymptotically stable equilibrium $\bar{q}_f, \bar{q}_v, \bar{q}_s$ is as it appears in the table.

Table 4 Globally asymptotically stable equilibria – Case 2a.

Sub-case	$\bar{q}_f$	$\bar{q}_v$	$\bar{q}_s$
I. $\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s \lambda_v}{\mu_s} \leq N$	0	0	0
II. $\frac{\lambda_v}{\mu_v} + \frac{p_s \lambda_v}{\mu_s} \leq N < \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s \lambda_v}{\mu_s}$	$\frac{\mu_f}{\theta_f}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s} - N\right]^+$	0	0
III. $N < \frac{\lambda_v}{\mu_v} + \frac{p_s \lambda_v}{\mu_s}$	$\frac{\lambda_f}{\theta_f}$	$\frac{1}{\theta_v}\left(\lambda_v - \frac{\mu_v \mu_s}{p_s \mu_v + \mu_s} N\right)$	0

Next, we construct a Lyapunov function for each case and show the globally asymptotically stability of the equilibrium point. Since the line of arguments is the same for all cases, we provide the proof for Case 2a.I and omit the others.

- **Sub-case 2a-I.** $\frac{\lambda_f}{\mu_f}+\frac{\lambda_v}{\mu_v}+\frac{p_s\lambda_v}{\mu_s}\le N$. We consider the Lyapunov function

$$V(q)=\frac{1}{\mu_f}|q_f-\bar{q}_f|+\frac{p_s\mu_v+\mu_s}{\mu_v\mu_s}|q_v-\bar{q}_v|+\frac{1}{\mu_s}|q_s-\bar{q}_s|,$$

where the equilibrium point $\bar{q}=(0,0,0)$, and show its asymptotic stability.

— **When $q_v(t)>0$ and $q_s(t)=0$**, all capacity is allocated to Classes s and v. Specifically, the system dynamics in (4) are as follows:

$$\begin{cases}\dot{q}_v(t)=\lambda_v-\frac{\mu_v\mu_s}{p_s\mu_v+\mu_s}N-\theta_v q_v(t);\\ \dot{q}_s(t)=\frac{p_s\mu_v\mu_s}{p_s\mu_v+\mu_s}N-\frac{\mu_s p_s\mu_v}{p_s\mu_v+\mu_s}N=0;\\ \dot{q}_f(t)=\lambda_f-\theta_f q_f(t);\end{cases}$$

We have

$$\begin{aligned}\nabla_q V(q)^T f(q)&=\frac{p_s\mu_v+\mu_s}{\mu_v\mu_s}\left(\lambda_v-\frac{\mu_v\mu_s}{p_s\mu_v+\mu_s}N-\theta_v q_v(t)\right)+\frac{1}{\mu_f}\left(\lambda_f-\theta_f q_f(t)\right)\\&=\frac{\lambda_f}{\mu_f}+\frac{\lambda_v(p_s\mu_v+\mu_s)-\mu_v\mu_s N}{\mu_v\mu_s}-\frac{\theta_v q_v(t)}{\mu_v\mu_s}-\frac{\theta_f q_f(t)}{\mu_f}\\&=\frac{\lambda_f}{\mu_f}+\frac{\lambda_v}{\mu_v}+\frac{p_s\lambda_v}{\mu_s}-N-\frac{\theta_v q_v(t)}{\mu_v\mu_s}-\frac{\theta_f q_f(t)}{\mu_f}<0,\end{aligned}$$

where the inequality comes from the sub-case's condition and the assumption that $\theta>0$.

— **When $q_v(t)=0$, $q_s(t)=0$ and $q_f(t)>0$**, the required capacity is allocated to Class v and then s is allocated: λ_v/μ_v to Class v and $p_s\lambda_v/\mu_s$; any leftover capacity is allocated to Class f. The system dynamics are therefore,

$$\begin{cases}\dot{q}_v(t)=\lambda_v-\mu_v\frac{\lambda_v}{\mu_v}=0;\\ \dot{q}_s(t)=p_s\lambda_v-p_s\lambda_v=0;\\ \dot{q}_f(t)=\lambda_f-\mu_f\left(\frac{\lambda_f}{\mu_f}\wedge\left(N-\frac{\lambda_v}{\mu_v}-\frac{p_s\lambda_v}{\mu_s}\right)\right)-\theta_f q_f(t).\end{cases}$$

We have

$$\nabla_q V(q)^T f(q)=\frac{1}{\mu_f}\left(\lambda_f-\mu_f\left(\frac{\lambda_f}{\mu_f}\wedge\left(N-\frac{\lambda_v}{\mu_v}-\frac{p_s\lambda_v}{\mu_s}\right)\right)-\theta_f q_f(t)\right)$$

$$
\begin{aligned}
&= \frac{\lambda_f}{\mu_f} - \left(\frac{\lambda_f}{\mu_f} \wedge \left(N - \frac{\lambda_v}{\mu_v} - \frac{p_s \lambda_v}{\mu_s} \right) \right) - \frac{\theta_f q_f(t)}{\mu_f} \\
&= \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s \lambda_v}{\mu_s} - N \right]^+ - \frac{\theta_f q_f(t)}{\mu_f} \\
&\leq \left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s \mu_v}{\mu_s} N - N \right]^+ - \frac{\theta_f q_f(t)}{\mu_f} < 0,
\end{aligned}
$$

where the first inequality comes from the fact that when $\lambda_v \leq N\mu_v$. The second inequality comes from the sub-case's condition and the assumption that $\theta > 0$.

The development for Case 2a.II and Case 2a.III follows similarly and is thus omitted. The Lyapunov functions we use are

$$
V(q) = \frac{1}{\mu_f} |q_f - \bar{q}_f| + \frac{p_s \mu_v + \mu_s}{\mu_v \mu_s} |q_v - \bar{q}_v| + \frac{1}{\mu_s} |q_s - \bar{q}_s|,
$$

for Case 2a.II, and

$$
V(q) = \sum_{i=f,v,s} |q_i - \bar{q}_i|,
$$

for Case 2a.III. The equilibrium queue lengths are given in Table 4. Q.E.D.

Appendix B: Proof of Theorem 2.

Recall the long-run profit maximization problem (6)–(7). Let $\bar{z}^* = (\bar{z}_f^*, \bar{z}_v^*, \bar{z}_s^*)$ and $\bar{q}^* = (\bar{q}_f^*, \bar{q}_v^*, \bar{q}_s^*)$ denote its solution (i.e., long-run average capacity allocation and corresponding queue length for each class). To prove the optimality of the $\mathcal{R}$ rule, it suffices to show that $\bar{z}^*$ and $\bar{q}^*$ constitute the globally asymptotically stable equilibrium established in Theorem 1. We, therefore, consider the same cases as in Theorem 1, and present the optimal solution $\bar{z}^*$ and $\bar{q}^*$. As before, we present the results for Case Ia and Cases 2b, and omit the other cases that follow the same line of arguments.

B.1. Case 1a: $\mathcal{R}_f < \mathcal{R}_s < \mathcal{R}_v$

We consider the four sub-cases described in Table 5. In this case, the $\mathcal{R}$ rule gives strict priority to Class v, then to Class r, and, finally, to Class f.

Except for sub-case I, where there is enough capacity to serve all customers, the other sub-cases priorities are: Class v, then Class s, and lastly Class f. This is in line with the $\mathcal{R}$ rule prioritization in this case.

Table 5 Optimal solution – Case 1a.

Sub-case	$(\bar{z}_f^*, \bar{z}_v^*, \bar{z}_s^*)$	$(\bar{q}_f^*, \bar{q}_v^*, \bar{q}_s^*)$
I. $\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s} N \leq N$	$\left(\frac{\lambda_f}{\mu_f}, \frac{\lambda_v}{\mu_v}, \frac{p_s\lambda_v}{\mu_s}\right)$	$(0,0,0)$
II. $\frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} \leq N < \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\mu_v}{\mu_s} N$	$\left(N - \frac{\lambda_v}{\mu_v} - \frac{p_s\lambda_v}{\mu_s}, \frac{\lambda_v}{\mu_v}, \frac{p_s\lambda_v}{\mu_s}\right)$	$\left(\frac{\mu_f}{\theta_f}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s} - N\right]^+, 0, 0\right)$
III. $\frac{\lambda_v}{\mu_v} \leq N < \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s}$	$\left(0, \frac{\lambda_v}{\mu_v}, N - \frac{\lambda_v}{\mu_v}\right)$	$\left(\frac{\lambda_f}{\theta_f}, 0, \frac{1}{\theta_s}\left(\lambda_v p_s - \mu_s\left(N - \frac{\lambda_f}{\mu_v}\right)\right)\right)$
IV. $N < \frac{\lambda_v}{\mu_v}$.	$(0, N, 0)$	$\left(\frac{\lambda_f}{\theta_f}, \frac{\lambda_v - \mu_v N}{\theta_v}, \frac{p_s\mu_v}{\theta_s} N\right)$

Table 6 Optimal solution – Case 2a.

Sub-case	$(\bar{z}_f^*, \bar{z}_v^*, \bar{z}_s^*)$	$(\bar{q}_f^*, \bar{q}_v^*, \bar{q}_s^*)$
I. $\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} \leq N$	$\left(\frac{\lambda_f}{\mu_f}, \frac{\lambda_v}{\mu_v}, \frac{p_s\lambda_v}{\mu_s}\right)$	$(0,0,0)$
II. $\frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s} \leq N < \frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s}$	$\left(N - \frac{\lambda_v}{\mu_v} - \frac{p_s\lambda_v}{\mu_s}, \frac{\lambda_v}{\mu_v}, \frac{p_s\lambda_v}{\mu_s}\right)$	$\left(\frac{\mu_f}{\theta_f}\left[\frac{\lambda_f}{\mu_f} + \frac{\lambda_v}{\mu_v} + \frac{\lambda_v p_s}{\mu_s} - N\right]^+, 0, 0\right)$
III. $N < \frac{\lambda_v}{\mu_v} + \frac{p_s\lambda_v}{\mu_s}$.	$\left(0, \frac{\mu_s}{p_s\mu_v+\mu_s} N, \frac{p_s\mu_v}{p_s\mu_v+\mu_s} N\right)$	$\left(\frac{\lambda_f}{\theta_f}, \frac{1}{\theta_v}\left(\lambda_v - \mu_v \frac{\mu_s}{p_s\mu_v+\mu_s} N\right), 0\right)$

B.2. Case 2a: $\mathcal{R}_{v,f} < \mathcal{R}_f$ ($\mathcal{R}_v < \mathcal{R}_s$)

We consider the three sub-cases described in Table 6. In this case, the $\mathcal{R}$ rule prioritizes Class s and Class v and then Class f.

Except for sub-case I, where there is enough capacity to serve all customers, the other sub-cases allocate capacity to Classes v and s while keeping a constant ratio between the capacities. Lastly, if some capacity remains, it is allocated to Class f. This is in line with the $\mathcal{R}$ rule in this case (i.e., the two-step $\mathcal{R}$ rule). Q.E.D.