The restricted Erlang-R queue: Finite-size effects in service systems with returning customers

Galit B. Yom-Tov[1†], Fiona Sloothaak[2†], Johan S.H. van Leeuwaarden[3†]

[1]Faculty of Data and Decision Sciences, Technion — Israel Institute of Technology, Haifa, Israel.
[2]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands.
[3*]Tilburg School of Economics and Management, Tilburg University, Tilburg, Netherlands.

Contributing authors: gality@technion.ac.il; f.sloothaak@tue.nl; J.S.H.vanLeeuwaarden@tilburguniversity.edu;
[†]These authors contributed equally to this work.

Abstract

We study a restricted version of the Erlang-R queueing model that has both server and space constraints, in which the space restriction is administered by two types of admission policies: blocking or holding in a pre-entrant queue. This model captures a wide range of service environments, such as medical units, contact centers, and courts. We study the interplay between the two resource constraints and the re-entrant queueing structure. We analyze the connection between staffing, space, and policy decisions and the system performance as a tool for optimizing resource utilization and in determining stability and performance. We develop product form and many-server approximations of performance measures when the blocking policy applies, and then utilize these results to determine approximations for the holding policy as well. We show that system dynamics are captured by a single measure of the fraction time being "needy" in the network, and that returning customers should be explicitly accounted for both in steady-state and transient environments. We show that capacity allocation of both resource constraints (servers and space) should be considered jointly, and derive methodology to determine it. We show that applying a holding policy requires more resources

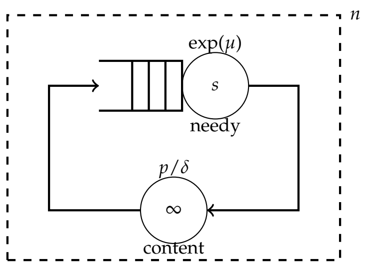
than a blocking policy and demonstrate the application of our policies in two case studies of resource allocation in hospitals.

Keywords: QED approximation, healthcare, Staffing

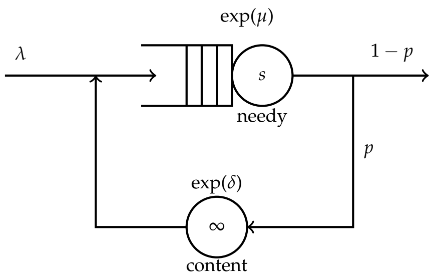
MSC Classification: MSC90B22 , MSC90B50 , MSC60F05

1 Introduction

Because service systems are inherently stochastic, it is common practice to use queueing theory for performance analysis and workforce planning. Traditionally, systems are modeled as a single-station queue, such as the $M/M/s$ (Erlang-C), $M/M/s/s$ (Erlang-B) or $M/M/s+M$ (Erlang-A) models, with fluid and diffusion approximations applied to provide insights into the process dynamics. However, single-station models often fail to capture the intricate dynamics of complex service networks. In a healthcare context, think of flows of patients in a hospital from one Medical Unit (MU) to another [1], routing within an Emergency Department (ED) between different stages of treatment [2, 3], or transitions between medical facilities [4, 5]. Queueing networks can effectively capture the dependencies among multiple service stages and resources. More specifically, we are interested in the practically relevant feature, that customers may require a specific resource *multiple times* during their stay in the system. Examples include nurses treating patients several times during an MU [6] or ED visit [7], or a service agent exchanging multiple text messages with a customer in a contact center [8]. To explicitly capture such situations, two key models have been introduced in the literature: the closed-ward membership model by de Véricourt and Jennings (see Figure 1a) and the Erlang-R model by Yom-Tov and Mandelbaum, where 'R' stands for *reentrant* customers (see Figure 1b).



(a) The closed ward membership model



(b) The Erlang-R model

Fig. 1 Two classic models with multiple services

A feature of service networks that is often ignored yet essential concerns the restriction on the number of customers that can simultaneously reside in the system.

For instance, in both contact centers and EDs, customers frequently return for multiple services; however, the number of customers a single server can handle at one time is restricted due to human cognitive limitations [10–12]. Similarly, in healthcare facilities, nurses treat patients repeatedly, while the physical number of beds poses a strict constraint on the number of patients who can be admitted to an MU at any given time. In such situations, two or more critical resources (e.g., personnel and beds) are simultaneously restricted. In this paper, we investigate the influence of such multiple restrictions on the open Erlang-R network dynamics and the required staffing policies, thereby extending both of the above mentioned models as explained next.

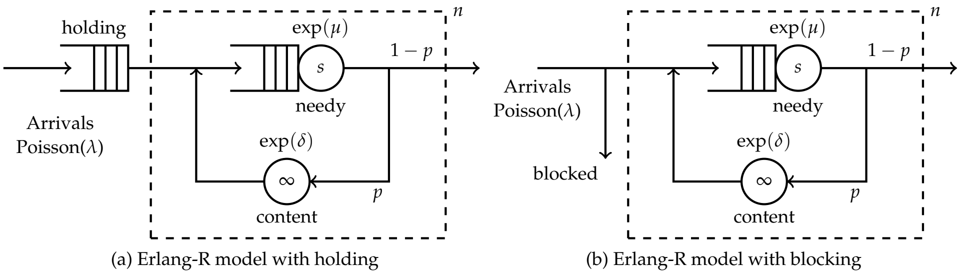


Fig. 2 Restricted Erlang-R models with maximally n customers in system, where waiting or blocking is allowed

The restricted Erlang-R model. The Erlang-R model [7] (Figure 1b) captures the reentrant dynamics that occur when each customer undergoes multiple stages of service during their stay. In this work, we extend the Erlang-R model by enforcing a constraint on the maximum number of available spaces inside the service facility while preserving the open network structure. Our model hence incorporates *two* types of resource constraints: servers that provide the actual service and spaces limiting the maximum number of concurrent customers inside the service system. These constraints affect the system in a highly interdependent way. We study two variants of this model, presented in Figure 2. In both variants, we assume there are s servers and a maximum capacity of n concurrent spaces. Customers arrive according to a Poisson process with rate λ. If a new arrival finds n or more customers already present in the system, we consider two options—either they wait outside the service facility in a holding queue until a vacant space becomes available (Figure 2a), or they are blocked (Figure 2b), as is the case when patients are diverted to an alternative MU when their primary MU is full [13]. Once customers are admitted, they require assistance from one of the s servers for an exponentially distributed duration with mean $1/\mu$. Then, customers leave the system with probability $1-p$, or, with probability p, return for another service iteration after an exponentially distributed time with mean $1/\delta$. Service is non-preemptive; instead, returning customers wait in an internal queue at the needy station until a server becomes available. Following de Véricourt and Jennings [6] and Yom-Tov and Mandelbaum [7], we call patients *needy* when they require attention from one of the servers and *content* when they are in the delayed return phase. Additionally, we define customers as *holding* when they are waiting outside the facility for an available space. We refer to the wait to enter the restricted area of the system from the holding state as the *holding time*, and the wait for a server to become available during a recurrent service stage at the needy station as the *delay time*. We assume that the arrival process, needy times, and content times are mutually independent. Note that both variants of our model extend the closed ward membership model [6, 9] by eliminating the constraint that strictly fixes the total number of customers in the system at n. This extension operates in two ways. First, it allows the number of customers in the system to fall below n, meaning the system can be underloaded at times. Second, when the number of customers reaches n, the system either blocks new arrivals or allows them to wait in a holding queue, depending on the model variant. Crucially, the restriction on the maximum capacity n can create scenarios where customers are forced to wait in the holding queue even while a server inside the facility sits idle. In both the holding queue and the internal needy queue, we apply a First-Come-First-Served (FCFS) discipline.

Examples of restricted Erlang-R. We now present examples from different domains to underscore the practical relevance of a restricted Erlang-R model: an MU within a hospital, aiming to capture both intra-unit and inter-unit dynamics; an ED operating under strict patient-to-physician ratio restrictions; and a text-based contact center.

The first example is an MU in a hospital. Such units specialize in specific types of illnesses (e.g., cardiac or oncology) and operate with limited resources, such as nurses and beds. Nursing services within the MU are administered multiple times throughout a patient's LOS. If the MU is full, new patients may wait for an available bed in the ED—a phenomenon known as boarding [14]—or be allocated to an alternative MU, a phenomenon known as off-placement or blocking [13]. The first option corresponds to the model variant with a holding queue, whereas the second corresponds to the variant with blocking. Both policies are costly and problematic, as they prolong a patient's LOS and reduce the quality of care [13–15]; personnel in an alternative unit (or the boarding section of the ED) may be less specialized, and delaying treatment has been shown to increase mortality. Moreover, ED boarding time can reduce the available capacity for treating other ED patients [16, 17], thereby endangering both the boarded patient and others. While both the number of personnel (nurses and physicians) and the number of beds impact service dynamics and quality of care, prior research has studied the capacity allocation of these resources separately. For instance, Green and Yankovic [18] and de Véricourt and Jennings [9] considered nurse staffing in medical units, while de Bruin et al. [19] considered bed allocation. The unified model proposed here enables us to capture the interdependencies between these two decisions and evaluate their combined impact on off-placement probabilities across the hospital.

Our second example is an ED operating in the US. Yom-Tov and Mandelbaum [7] demonstrated that the Erlang-R model can be used to determine staffing in an Israeli ED under both stable and time-varying conditions. They defined the physician as the server (represented in the needy state), while other medical processes, such as diagnostic lab tests, were captured using the content state. Nevertheless, empirical studies report that some countries, including the US, employ a different operational mode that applies strict restrictions on ED entrance [12]. In a typical US ED, a patient will not begin treatment until both a bed and a physician are assigned to them. In other scenarios, access to the ED itself may be blocked—a phenomenon known as Ambulance Diversion [20]. These entry constraints can stem from physical limitations (such as bed capacity) or managerial policies, such as the imposition of strict patient-to-physician ratios or ambulance diversion protocols.

A third example comes from the emerging domain of text-based customer service chat (CSC), which utilizes instant messaging or chat channels. This mode of service delivery is growing rapidly and increasingly replacing the traditional phone-based call centers. In a CSC contact center, each agent serves a limited number of customers in parallel, thereby efficiently utilizing the idle time spent waiting for a customer to reply. This managerial limitation is known as the agent's concurrency level and needs to be optimized [21, 22]. A text-based conversation can be viewed as a reentrant process where each reply from the agent constitutes a service iteration that is repeated several times during the customer's LOS [8]. The time elapsed while waiting for the customer to respond represents the content state, during which the server may attend to other concurrent conversations. Finally, similar dynamics appear in managing Large Language Model (LLM) inference sessions; here, each user session involves multiple sequential queries bound together, and physical memory limitations impose a limit to the number of concurrent sessions a server may process in parallel. In both of these examples, the model variant featuring a holding queue serves as the appropriate framework.

As mentioned, we consider two versions of the finite-capacity constraint, capturing the two most commonly used modes of operation: holding or blocking of new patients. The first version is called *Erlang-R with holding*, in which customers wait for an available space inside the system. The second version is called *Erlang-R with blocking*, in which customers who encounter a full system are blocked. Naturally, intermediate scenarios can be constructed where a certain a proportion of arriving customers leaves upon finding a full system, while the remainder joins the holding queue. While this paper focuses on these two extreme cases, straightforward adaptations can accommodate such intermediate scenarios; however, they are omitted here due to space constraints. These two extreme cases are represent two alternative ways of implementing the space constraint: blocking is a strict constraint, whereas holding is a softer constraint that allows customers to wait until the system load reduces. Moreover, as shown later, the two models are mathematically connected; specifically, developing approximations for the holding model variant requires utilizing the mathematical approximations of the strictly constrained blocking variant.

Our main goal is to provide two-fold staffing policies for the restricted Erlang-R models that ensure high resource utilization, while maintaining good quality-of-care. This goal relates to the philosophy of the Quality-and-Efficiency-Driven (QED) regime known in many-server asymptotic theory. We discuss the main ideas behind this regime further in §2. This is a challenge due to the complex structure of the networks. The relatively simple one is the blocking model and the much more complex one is the holding model. Specifically, the Erlang-R model with blocking possesses a product-form steady-state solution, while the Erlang-R model with holding is governed by a Quasi-Birth-Death (QBD) process that lacks such an elegant analytical solution. Yet, these models are closely related, and we rely on the results of the blocking variant to develop approximations for the holding variant. Therefore, the importance of developing these asymptotics goes beyond the practical value of the simple-to-use formulas; it enables us to better understand system dynamics of both models and the importance of several modeling issues. Hence, the second part of this paper is dedicated to answering the following questions: How do space restrictions influence capacity, utilization, and LOS? What is the difference between strict (blocking) and flexible (holding) constraints? How important is it to account for reentrant customer behavior when determining staffing and space allocation? Finally, can we identify specific conditions under which considering this behavior becomes more or less important?

1.1 Summary of results and contributions

In this paper, we first obtain exact product-form and asymptotic results for the Erlang-R model with blocking (§4.2). Following de Véricourt and Jennings [9], we employ a two-fold QED staffing policy: $s = R_1 + \beta\sqrt{R_1}$ for the number of servers and $n = R_1/r + \gamma\sqrt{R_1/r}$ for the maximal number of customers in the system, where β and γ are operational constants, $R_1 = \lambda/((1-p)\mu)$ is the offered load of the servers (defined as the expected number of customers in the needy station in steady state under infinite-server conditions), and r is the fraction of time a customer spends in the needy station when infinite servers are available (i.e., $r = \delta/(\delta + p\mu)$). We establish limiting expressions for performance measures, such as the probability of delay in the needy station and the probability of blocking, in the form of explicit functions that depend solely on β and γ. In deriving these limit results, we leverage the available product-form solution for the stationary distribution.

Similarly, we pursue QED approximations of performance measures for the Erlang-R model with holding. However, a direct analytic approach is obstructed by the absence of a product-form solution. To address this, we provide two solutions for establishing QED behavior. First, we provide stochastic performance bounds that stay meaningful in the QED regime (§3.3), demonstrating the non-degenerate behavior of the two-fold scaling in the large-system limit. Second, we develop a heuristic method that quantifies the difference between the scalable holding model and the blocking model (§4.3). This method is based on the novel fixed-point approach: customers blocked in the blocking model are seen as if they reattempt to get access after some delay. The behavior of this retrial model then resembles the Erlang-R model with blocking without retrials, albeit with an inflated arrival rate that emerges as the solution to a fixed-point equation. Using our results on the asymptotic behavior of the blocking model in the QED regime, we then obtain approximate QED performance measures for the holding model. This approach suggests that the results of the more tractable semi-open queueing network can serve as a foundation for developing approximations—rather than just bounds—for the non-tractable one. Finally, we leverage these QED results to develop algorithms for dimensioning and time-varying staffing (§5.1).

The above-mentioned approximations provide an efficient, flexible, and easy-to-implement methodology to jointly determine resource allocation of both servers (personnel) and spaces (beds) in a way that avoids waste, ensuring both resources operate in the QED regime.

Using the approximations developed, numerical analysis, and simulation, we provide the following operational implications:

- We show that applying restrictions on the number of customers in the system while allowing a holding queue may increase the instability region compared to the Erlang-R model. Therefore, higher capacity is needed when such flexible restrictions are applied. Alternatively, applying strict restrictions in which customers are blocked ensures the system is always stable under all parameters, , making it viable for situations with lower capacity.
- Extending the line of inquiry in Yom-Tov and Mandelbaum [7], we examine the importance of explicitly modeling returning customer dynamics. We conclude that modeling *reentrant* behavior is significantly more critical in a restricted Erlang-R network than in an unrestricted one. In the unrestricted Erlang-R model, returning customers need to be accounted for *only* in time-varying environments and mostly when arrival rate changes substantially during a customer's LOS [7]. In contrast, the restricted Erlang-R model requires explicit consideration of reentering customers under stationary conditions as well.
- We show that the influence of the network structure on the system dynamics crucially depends on the fraction of time a customer spends being needy state during their stay in the system. We then explore how this fraction (r) influences operational decisions.
- Combining our theoretical results, we explore the implications of managerial decisions when designing an MU (§5) and an ED (§6.4). In §5, we show that enabling patients to wait in an ED (a holding queue) before entering the MU (the reentrant system) necessitates greater resource allocation both in the MU and the ED. This comes from the fact that holding (as opposed to blocking) allows a higher throughput in the MU but also requires the patients to consume care resources while waiting in the ED. In Section 6.4, we compare the pros and cons of imposing strict constraints on entering an ED. We find that capacity restrictions have the ability to improve the quality-of-service of the processes within the facility at the expense of a slight increase in holding times and server efficiency.
- Finally, we show that restricting the number of admitted customers protects those customers with complicated service demands consisting of relatively high number of reentrant service stages (§6.3).

When dealing with time-varying arrival rate ($\lambda(t)$), modifications to the stationary QED staffing rules are required to obtain stable performance over time. To address this,, we transform our two-fold staffing policy into a time-varying policy using the Modified Offered Load (MOL) method (§6.4). This method approximates the time-dependent offered load at the needy station by analyzing a corresponding system operating with ample resources. Recognizing that the this infinite-server system coincides with the Erlang-R model with infinite servers, we adopt the modified offered load approximations given in [7]. Our numerical experiments validate the accuracy of this approach, which we subsequently apply in our case study §6.4.

2 Literature review

Capacity Allocation in Healthcare Systems. Due to the increasing demand of an aging population, tightening healthcare budgets, and rising turnover within the medical workforce, there is a growing need for efficient workforce management [23]. Personnel (nurse and physician) expenditure is one of the largest drivers of hospital costs [23], and inadequate nursing levels have been identified as a significant factor in medical errors and ED overcrowding. To establish appropriate nursing staffing levels, a policy must asses a wide range of variables, such as varying clinical expertise and patient acuity during the day. Traditional methods, such as the fixed nurse-to-patient ratios, fail to capture these dynamic conditions and are often too inflexible [6]. This highlights the need for time-varying staffing policies that can accommodate the complex and evolving nature of healthcare delivery [24]. Workforce management in healthcare systems has been studied extensively; see [24–29] for overviews. In recent years it has become apparent that queueing models are highly effective tools for developing staffing and routing recommendations, not just for large-scale service systems but also for the smaller, highly complex healthcare units [29].

The first to adopt this type of queueing-model approach were Green and Savin [30], who used the stationary, single-station Erlang-C model to set staffing levels in EDs and panel sizes for clinics. Using a similar approach, Bekker and de Bruin [31] applied the Erlang-B model to determine bed allocations for medical wards. The first to observe the significant impact of repeating services in a healthcare setting were de Véricourt and Jennings [9]. Motivated by the need to set nurse-to-patient ratios for MUs, they considered a closed queueing system with s nurses and n beds, which we will refer to as the *closed ward* membership model. This model is conceptually an Erlang-C model under the additional restriction that a finite, fixed population of n patients requires care. In their model, all beds are continuously occupied, and patients alternate between two phases: a needy phase, where patients require attention from a nurse, and a content phase, where they recover independently (see Figure 1a). The system dynamics of our restricted Erlang-R model are equivalent to those of the closed ward model if the holding queue were never empty.

Campello et al. [32] analyzed the concurrency decision in healthcare operations, referred to as ED case management, which determines the maximal number of patients a physician should manage in parallel. Using queueing networks, they analyzed the system stationary distribution. Note that in practice, such operational decisions are shaped not only by performance metrics like waiting times but also by psychological and cognitive constraints that limit a physician's capability to multitask. Empirically, Diwas [10] demonstrated that clinical performance degrades when physicians treat more than 6–7 patients simultaneously. Consequently, many hospitals in the US restrict entrance to the EDs when physicians are overloaded, even if physical beds are available. Similar findings show the need of limiting concurrency levels in text-based contact centers as well [11]. We too consider such constraints, and analyze their impact on performance. However, we take a different approach than Campello et al. [32]; instead of numerically analyzing steady-state distributions and system bounds, we develop many-server approximations that produce insight into the system dynamics, and can be incorporated into time-varying staffing procedures (see §6.4).

The model in [6] was developed to capture the internal dynamics of an MU. However, in an ED, beds are not constantly occupied, and the utilization levels depend on the time-varying flow of patients arriving from outside the system. To account for the transient nature of the patient arrival process while capturing these multi-stage service dynamics, Yom-Tov and Mandelbaum [7] introduced the Erlang-R model (see Figure 1b). Using a simulator tailored to an Israeli ED, they demonstrated that the complicated small ED dynamics can be captured using the relatively simple Erlang-R model, and hence, its recommendations can be implemented in ED workforce management. We note, however, that routing decisions within the ED necessitate more granular models; in particular, those that incorporate patient clinical complexity and severity alongside re-entrant dynamics [2, 33, 34]. Although the feature of multiple services is present in many systems (e.g. in CSC), it is particularly important for modeling EDs, because the duration of the needy state is typically much longer than the time a patient requires from medical personnel [7]. In contrast, Tezcan and Zhang [35] and Luo and Zhang [21] analyzed staffing for CSC omitting the reentering effect entirely. This capacity to capture lengthy, intermittent service cycles explains why the Erlang-R model is considered a canonical framework for healthcare systems [28, 29, 36]. The restricted Erlang-R model with holding or blocking proposed here directly extends the classical Erlang-R model by incorporating finite capacity constraints which, much like reentering services, exert a decisive impact on system performance.

Quality-and-Efficiency-Driven regime. The Quality-and-Efficiency driven (QED) regime for many-server systems, also known as the Halfin-Whitt regime, adheres to a square-root staffing rule, which is best explained for the classical Erlang-C model. This system can be completely characterized by the staffing level s and the offered load $R = \lambda/\mu$, which represents the average workload pressure per time unit on the system. Moreover, let $\rho = R/s$ denote the server utilization level. In the QED regime [37, 38], the server utilization level is driven to unity according to $(1-\rho)\sqrt{s} \to \beta$ as $s \to \infty$, for some fixed parameter $\beta > 0$. This gives rise to the square-root dimensioning rule $s = R + \beta\sqrt{R}$, which prescribes that the number of servers s should exceed the minimally required offered load R, but only by a relatively small amount $\beta\sqrt{R}$. As s grows large, the probability of delay converges to a non-degenerate function that depends uniquely on the parameter β. This function is strictly decreasing in $\beta > 0$ with a range of $(0, 1)$. Consequently, any targeted delay probability can be achieved by adjusting β. Moreover, the mean delay is of order $1/\sqrt{s}$ and is hence asymptotically negligible. This type of asymptotic approximation is extensively used in the analysis of service and computer systems. It has been successfully implemented in single-station queues [39, 40], multi-station networks [9], and non-Markovian queues [41]. For a tutorial and review of QED approximations, we refer the reader to [42].

In this paper, we take the same approach but jointly determine the capacities for both servers (nurses) and places (beds) such that the probability of waiting for a server in the needy station (denoted the probability of delay) and the probability of waiting for a space in the system (bed) (denoted the probability of holding) are non-degenerate. Moreover, the utilization levels of both servers and spaces converge to unity as the system size approaches infinity. While the QED regime gives precise mathematical limits as the system size (s and n in our case) goes to infinity, it is well established that this asymptotic behavior manifests rapidly. As a result, QED limits serve as sharp approximations even for smaller, finite-capacity systems. For theoretical support regarding this rapid convergence, we refer the reader to [43–45].

Some might question the relevance of the QED regime in some of the motivating examples we propose. The QED regime is natural for MUs and CSCs, where service times are an order of magnitude longer than waiting times (days vs. hours in MUs and minutes vs. seconds in CSCs). However, its applicability is less obvious for EDs, where waiting times can be substantial. Nonetheless, QED approximations are well known to be very robust, covering accurately both the Quality Driven and the Efficiency Driven regimes. Therefore, they are extremely useful when considering staffing for systems that transition between these regimes throughout the day, as an ED state tends to do.

One heuristic approach to developing approximations for the performance measures of queueing systems with finite-capacity constraints under the QED regime involves the use of fixed-point equations. This methodology was suggested by van Leeuwaarden et al. [46] and van Leeuwaarden et al. [47] and successfully implemented for Erlang-A model and retrial queues [40]. This approach characterizes the asymptotic dynamics of analytically intractable, finite-capacity systems through more tractable ones. We adopt this approach and extend it to develop the approximations for our queueing *network* in the holding setting in §4.3.

3 Models and performance measures

3.1 Three-dimensional Markov process

The restricted Erlang-R model assumes that the arrival process is Poisson, and that service and content times are independent and exponential. Hence, the system can be characterized in terms of a Markov process. Let $Q(t) = (H(t), Q_1(t), Q_2(t))$ represent the number of patients in the *holding*, *needy* and *content* state at time t, respectively. In both variants, n is the maximal number of customers admitted to the system; we have $Q_1(t) + Q_2(t) \leq n$ for all $t \geq 0$. Due to the absence of holding patients in the Erlang-R model with blocking, $H(t) = 0$ is enforced in this case, whereas $H(t)$ has unbounded support in the model with holding. This distinction requires us to explore the stationary distribution of the two variants separately.

Before doing so, we introduce some additional notation. We define

$$R_1 := \frac{\lambda}{(1-p)\mu} \qquad R_2 := \frac{p\lambda}{(1-p)\delta}, \tag{1}$$

where R_1 and R_2 are the offered load at the needy and content stations, respectively. They are obtained by solving the balance equations of the network to quantify the effective steady-state arrival to each station and multiplying with the expected service time [7]. They can be interpreted as the offered workload brought towards each station or as the average number of people in the station when no restriction on capacity exist (i.e., s and n are infinite) [48]. Furthermore, we define

$$r := \frac{\delta}{\delta + p\mu}, \tag{2}$$

which is the fraction of time a customer spends in the needy state if ample capacity is available (i.e., without waiting).

3.1.1 The Erlang-R with blocking

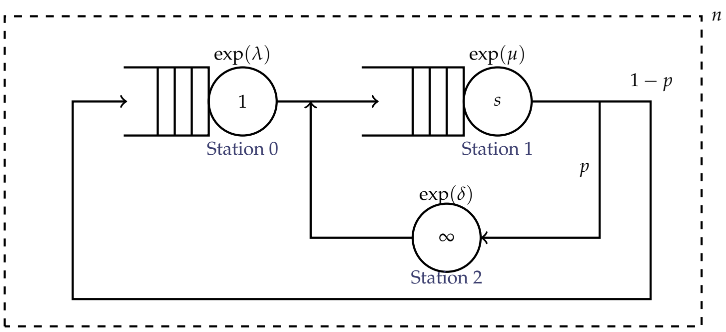


Fig. 3 The Erlang-R with blocking, viewed as a closed Jackson network.

In the blocking variant (Figure 2b), $Q(t)$ reduces to a finite-state Markov process $Q(t) = (Q_1(t), Q_2(t))$, where $Q_1(t) + Q_2(t) \leq n$ for all $t \geq 0$. In fact, this is equivalent to the closed Jackson network depicted in Figure 3 with finite population n. Station 1 is an $M/M/s$ queue with service rate μ, modeling the number of needy customers, $Q_1(t)$. Station 2 models the number of content customers, $Q_2(t)$, and can therefore be represented as an infinite-server queue with service rate δ. A customer can enter the system only if $Q_1(t) + Q_2(t) < n$. Station 0—a single-server queue—moderates this as it only produces output at rate λ when its queue length is positive, i.e. if $n - Q_1(t) - Q_2(t) > 0$.

Observe that because customers finding a full network are blocked, the number of customers in the system cannot grow beyond n. Hence, the system is stable for all parameter settings, and a steady-state distribution exists. Moreover, the simplification of the model with blocking allows us to express the steady-state distribution of the system in explicit product-form [49, 50]. Let $\pi_b(j,k)$ denote the steady-state probabilities of having j needy and k content customers in the system. Then,

$$\pi_b(j,k) = \begin{cases} \pi_0 \frac{1}{\kappa(j)} \frac{1}{k!} \cdot R_1^j \cdot R_2^k, & \text{if } j+k \leq n, \\ 0, & \text{else,} \end{cases} \tag{3}$$

where

$$\kappa(j) := \begin{cases} j!, & \text{if } j \leq s, \\ s!\, s^{j-s}, & \text{else,} \end{cases}$$

and $\pi_0^{-1} = \sum_{j+k\leq n} \frac{1}{\kappa(j)} \frac{1}{k!} \cdot R_1^j \cdot R_2^k$.

3.1.2 The Erlang-R with holding

The Erlang-R with holding variant does not lead to a Jackson network with an elegant product-form solution for the steady-state distribution, because the holding queue cannot be modeled as a station that is independent from the other queues in the system. However, we are able to describe the system as a two-dimensional Markov process without loss of information. To see this, define $N := \{N(t),\, t \geq 0\}$ with $N(t) := H(t) + Q_1(t) + Q_2(t)$, the total number of customers in the system (including the holding queue). Using the restriction $Q_1(t) + Q_2(t) \leq n$ together with the fact that no space (bed) is left vacant if a customer (patient) is waiting in the holding queue, yields $H(t) = (N(t) - n)^+$, $t \geq 0$, where $(\cdot)^+ := \max\{0, \cdot\}$. For the same reason, $Q_2(t) = N(t) - Q_1(t)$ if $H(t) = 0$, and $Q_2(t) = n - Q_1(t)$ otherwise. In other words,

$$Q_2(t) = \min\{N(t), n\} - Q_1(t), \quad t \geq 0.$$

Therefore, we can express the state of all three queues in the Erlang-R with holding model using a two-dimensional Markov process $X := \{X(t),\, t \geq 0\}$, where $X(t) := (N(t), Q_1(t))$. The process X lives on the semi-infinite strip

$$X(t) \in \{\, (i,j) \,|\, j \leq \min\{i,n\}, i \in \mathbb{N}_0, j \in \{0,1,\ldots,n\} \,\},$$

and belongs to the class of Quasi-Birth-Death (QBD) processes. The reader is referred to Appendix A for a detailed description of this process, in terms of its transition diagram and generator matrix.

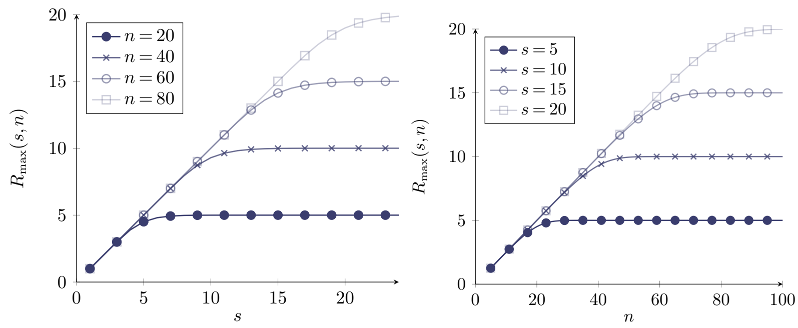
In contrast with the model with blocking, the system with holding *can* become unstable if the capacity is insufficient to satisfy customer demand.

Proposition 1 *The Erlang-R model with holding is stable if and only if*

$$\frac{\lambda}{(1-p)\mu s} < \frac{\sum_{i=0}^{s} \frac{i}{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i}{\sum_{i=0}^{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n}\binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i} =: \rho_{\max}(s,n).$$

The proof is given in Appendix A.2 and follows from the general theory for QBD processes.

Observe that $\rho_{\max}(s,n)$ poses an upper bound on the occupancy level of the servers in the holding model, which is clearly smaller than 1 for all s and n. In addition, this implies that the maximum workload $R_{\max}(s,n) := s \cdot \rho_{\max}(s,n)$ the system is able to handle is strictly less than s. If we compare this to the open Erlang-R model, in which the maximal attainable workload equals s, we observe the effect of finite-size constraints on operational performance. A system with space restrictions and holding has a larger instability region than a system without holding, and therefore, can handle lower maximal workload. Therefore, it requires a higher capacity to avoid instability. Intuitively, this happens because the agent may be idle while customers wait in the holding queue, which reduces the agent utilization. On the other hand, we will show next that a holding queue allows for a higher effective volume to be served compared to the blocking option. Figure 4 shows the influence of both s and n



(a) $R_{\max}$ as a function of s (b) $R_{\max}$ as a function of n

Fig. 4 The maximum achievable workload in the restricted Erlang-R model with holding for $r = 0.25$.

on the maximum feasible workload in case $r = 0.25$ (the fraction of needy time, from Eq. (2)). From these graphs, note that if $s \ll rn$, $R_{\max}$ grows almost linearly with s. Furthermore, $R_{\max}(s,n)$ is increasing in n for fixed s. A logical practical consequence is that a larger number of space (beds) allows for a larger customer volume to enter the system with the same number of servers. Moreover, $R_{\max}(s,n)$ is increasing in s, but as in Figure 4a, adding an extra server does not increase the stability region if n is too tight. Conversely, adding extra space (bed) does not increase $R_{\max}(s,n)$ if the number of servers (nurses) does not allow for an increase in offered load, see Figure 4b. Additionally, it is easily verified that $R_{\max}(s,n)$ is upper bounded by both s and $R_{\max}(n,n) = rn$. Therefore, a careful balance is called for between servers (nurses) and spaces (beds), so that resources will be efficiently utilized. We observe that when the ratio $s/n \approx r$, the system is better balanced. We will propose an appropriate balance between resources by defining a joined QED capacity recommendation for both servers and spaces in §4.

Provided that the system is stable, the stationary distribution of the QBD process X can be obtained numerically by the matrix geometric method [51]. Subsequently, we can derive the stationary distribution of the original $Q(t)$, denoted by $\pi_h(\cdot,\cdot,\cdot)$.

3.2 Performance measures

In this work, we concentrate on five performance measures that are central to our analysis. In the definitions that follow, we present expressions for these measures in terms of a general three-dimensional measure π, which one can replace by either π_b or π_h, depending on the scenario considered. In the remainder of this work, we will augment the measures related to the Erlang-R model with blocking and holding by the superscript b and h, respectively[1].

As relevant performance measures, we consider the probability of holding (blocking) when entering the system ($\mathbb{P}(\text{hold})$ and $\mathbb{P}(\text{block})$, respectively), the probability of a non-zero delay time at the needy queue ($\mathbb{P}(\text{delay})$), the per visit expected waiting time for a server at the needy station ($\mathbb{E}[W]$), the utilization of a server at the needy station (ρ_s) and the utilization of spaces (beds) in the system (ρ_n):

$$\mathbb{P}(\text{hold}) = \sum_{i=0}^{\infty}\sum_{j=0}^{n} \pi(i,j,n-j), \qquad \mathbb{P}(\text{delay}) \approx \sum_{i=0}^{\infty}\sum_{j=s}^{n}\sum_{k=0}^{n-j} \pi(i,j,k), \tag{4}$$

$$\mathbb{E}[W] \approx \sum_{i=0}^{\infty}\sum_{j=s}^{n}\sum_{k=0}^{n-j} \frac{\max\{0, j-s+1\}}{\mu}\,\pi(i,j,k),$$

$$\rho_s = \frac{1}{s}\sum_{i=0}^{\infty}\sum_{j=0}^{n}\sum_{k=0}^{n-j} \min\{j,s\}\pi(i,j,k), \qquad \rho_n = \frac{1}{n}\sum_{i=0}^{\infty}\sum_{j=0}^{n}\sum_{k=0}^{n-j} \min\{i,n\}\pi(i,j,k).$$

It should be stressed that the above expressions for the delay probability and the expected waiting time for a server are not exact. For the blocking model one can use the Arrival Theorem [52], whereby the exact expression uses $n-1$ instead of n; for the holding model, the arrival process to the needy queue, which consists of both external arrivals and the content customers becoming needy, is not Poisson. Therefore, we cannot use the PASTA argument for the holding model. However, for both models, as we will be studying the system as s and n become large, this approximation error will become negligible.

3.3 Stochastic bounds

Although the two variants of the Erlang-R model differ with respect to the admission policy, and require different mathematical treatment, we would like to be able to capture their relative performance. We substantiate the intuition that the holding room leads to more customers in the system, in the following result.

---

[1] In line with $H(t) = 0$, we use $\pi_b(i,j,k) = \pi_b(j,k)$ if $i = 0$, with $\pi_b(j,k)$ as in (3), and $\pi_b(i,j,k) = 0$ otherwise, when considering the model with blocking.

Proposition 2 *Let $Q_1^b, Q_2^b, Q_1^h, Q_2^h$ denote the needy and content queue length processes in the Erlang-R model with blocking and holding, respectively. Let $H(0) = 0$, $Q_1^b(0) = Q_1^h(0)$ and $Q_2^b(0) = Q_2^h(0)$. For all $t \geq 0$,*

$$\begin{aligned} Q_1^b(t) + Q_2^b(t) &\preceq_{\text{st}} Q_1^h(t) + Q_2^h(t) \preceq_{\text{st}} n, \\ Q_2^b(t) &\preceq_{\text{st}} Q_2^h(t), \\ Q_1^b(t) &\preceq_{\text{st}} Q_1^h(t) + H(t), \end{aligned}$$

where $X \preceq_{\text{st}} Y$ implies $\mathbb{P}(X \geq k) \leq \mathbb{P}(Y \geq k)$ for all $k \geq 0$.

The proof of Proposition 2 is based on basic coupling arguments and can be found in Appendix B. Note that as an immediate consequence, we have

$$\mathbb{P}^b(\text{block}) = \lim_{t\to\infty} \mathbb{P}\left(Q_1^b(t) + Q_2^b(t) \geq n\right) \leq \lim_{t\to\infty} \mathbb{P}\left(Q_1^h(t) + Q_2^h(t) \geq n\right) = \mathbb{P}^h(\text{hold})$$

and by similar reasoning $\rho_n^b \leq \rho_n^h$.

Proposition 2 suggests that under similar offered load and capacity constraints, utilization levels for the servers in the Erlang-R model with blocking are lower than in the Erlang-R model with holding. Moreover, the total number of waiting customers in the setting with holding is stochastically larger than in the setting with blocking, and in the open Erlang-R model. We further discuss the differences between both models in §5 and §6.

4 Two-fold QED regime

Wasting capacity of either servers or beds without getting a significant advantage in terms of performance is undesirable. We therefore take an asymptotic approach that lets the external arrival rate λ grow to infinity, while scaling s and n accordingly. In doing so, we intend to establish QED-type behavior, i.e. high occupancy levels of both servers and places and good quality-of-service.

4.1 Two-fold scaling rule

In order to identify the scaling of s and n as $\lambda \to \infty$, we draw inspiration from the two-fold scaling rule in [9] and [53], which follows the celebrated square-root staffing principle discussed in §2. This principle suggests that, in the most general setting, capacity should be equal to the expected offered load entering the system, say R, plus an additional variability hedge that is proportional to $\sqrt{R}$. In the restricted Erlang-R model, we have two capacity sources, namely s and n, which experience different relevant quantities of work.

The offered load the servers in the needy queue experience is given by R_1, as in the regular Erlang-R model; it does not change due to the finite-size effects, since all patients are served eventually. Hence, we only need to account for the reentering services. It follows that the appropriate staffing rule for the nurses (servers) in the QED regime remains $s = R_1 + \beta\sqrt{R_1}$ for some constant $\beta > 0$.

To establish the bed (space) capacity level, we need to reflect on the load offered in relation to the beds. Observe that beds remain occupied both in needy and content states. This suggests that $R_T := R_1 + R_2 = R_1/r$, with R_1 and R_2 as in Eq. (1) and r the expected fraction of time a patient spends at the nurse station, defined in Eq. (2). As a result, the appropriate staffing rule is $n = R_T + \gamma\sqrt{R_T}$ for some constant $\gamma > 0$. In conclusion, the two-fold QED scaling rule is given by

$$\begin{aligned} s &= R_1 + \beta\sqrt{R_1} + o(\sqrt{R_1}) \\ n &= \frac{R_1}{r} + \gamma\sqrt{\frac{R_1}{r}} + o(\sqrt{R_1}) \end{aligned} \tag{5}$$

with $\beta, \gamma > 0$ constants, $R_1 := \lambda/((1-p)\mu)$, and $r = \delta/(\delta + p\mu)$.

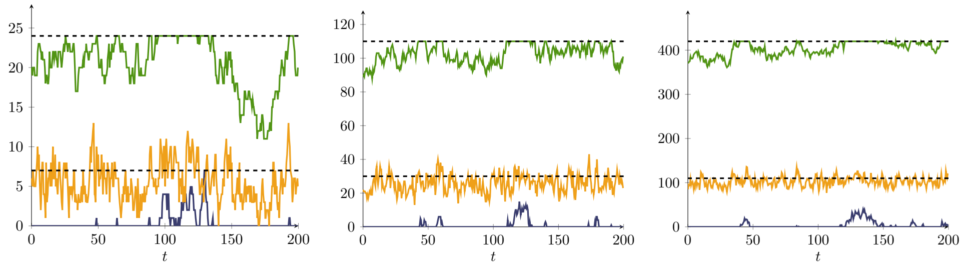
Recall that we saw in Figure 4 that resources seem efficiently utilized if $s/n \approx r$. Scaling (5) is in line with this reasoning since

$$\frac{s}{n} = r\left(1 + \frac{\beta - \gamma\sqrt{r}}{\sqrt{R_1}} + O(1/R_1)\right).$$

Remark 1 In [9], a similar scaling regime is considered for the closed ward membership model, which only relates s and n through a square-root scaling, namely the regime $s = rn + \hat{\gamma}\sqrt{n}$, which is equivalent to the second relation in (5) if $\hat{\gamma} = \beta\sqrt{r} - \gamma r$. Due to the absence of external arrivals in this closed system, they let the number of beds n approach infinity as opposed to λ in our settings. Nevertheless, this results in the same asymptotic regime.

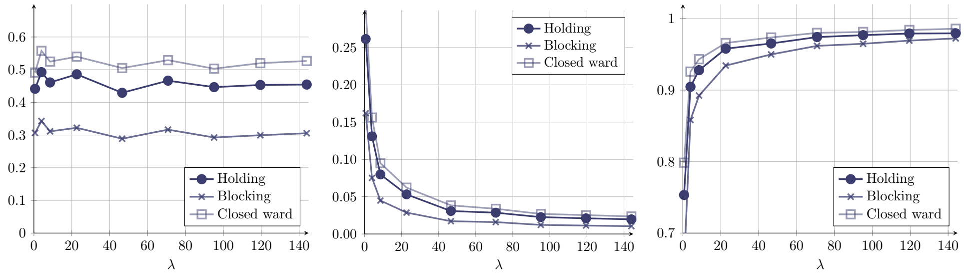
Before turning to asymptotic expressions for the performance measures concerning the Erlang-R model with blocking/holding, we conduct a few numerical experiments to confirm that the scaling in (5) indeed leads to the desired QED behavior. In Figure 5 we plotted the sample paths of the three-dimensional queue length process of the holding model in which β and γ are fixed, and R_1 is increased. Observe that the needy queue length $Q_1(t)$, plotted in orange in Figure 5, fluctuates around the values s, and stabilizes for larger values of R_1. This naturally implies that the server (nurses) utilization approaches 100%, while the number of patients waiting is $O(\sqrt{R_1})$. Furthermore, we see that the percentage of occupied beds also tends to 100%, while the holding queue length remains small. The holding queue is of much smaller order than R_1, which implies that the holding time of a customer becomes negligible as $R_1 \to \infty$. From these empirical findings we deduce that under scaling (5) the restricted Erlang-R model exhibits QED behavior on two levels: Outside the facility while waiting for an available bed (space), and inside the facility while waiting for the attention of a server.

We also check how the Erlang-R model with blocking or holding and the closed ward model of de Véricourt and Jennings [9] relate under scaling (5). In Figure 6, we plot the performance measures, obtained through simulation, for the three models in which we fix $\beta = \gamma = 0.5$ and vary the arrival rate λ. First, we see that $\mathbb{P}(\text{delay})$ stabilizes as $\lambda \to \infty$ in all three models under scaling (5), and the delay probability of the model with holding lies in between the other two. Second, note that the expected



(a) $R_1 = 5$ (b) $R_1 = 25$ (c) $R_1 = 100$

Fig. 5 Sample paths of $H(t)$ (blue), $Q_1(t)$ (orange) and $Q_1(t) + Q_2(t)$ (green) of the Erlang-R model with holding with parameters $\mu = 1$, $\delta = 0.25$, $p = 0.75$ and $\beta = \gamma = 1$. The staffing levels s and n are depicted by the dashed lines.



(a) Delay probability nurse (b) Expected wait (c) Nurse utilization

Fig. 6 Asymptotic behavior of the restricted Erlang-R model with holding and blocking, and the Closed ward model for $\mu = 1$, $\delta = 0.2$, $p = 0.8$ and $\beta = \gamma = 0.5$.

waiting time for a server in all models converges to 0 as λ increases. In fact, the rate of decay is similar in all three models. We observe that ρ_s approaches unity in all three models, and the rate of convergence seems again comparable. Finally, and most importantly, we notice an ordering between the three models. Namely, in all performances of the restricted Erlang-R model considered in Figure 6, Erlang-R with holding appears to be upper bounded by the closed ward and lower bounded by the Erlang-R with blocking. In a multitude of parameter settings of (β, γ), we have seen the same ordering, leading to the following conjecture:

Conjecture 1 *Let $Q_1^b(\infty)$, $Q_1^h(\infty)$ and $Q_1^c(\infty)$ denote the stationary number of needy customers in the Erlang-R model with blocking, holding and the closed ward, respectively. Then,*

$$Q_1^b(\infty) \preceq_{\text{st}} Q_1^h(\infty) \preceq_{\text{st}} Q_1^c(\infty).$$

Observe that Conjecture 1 poses a stronger statement than the third assertion in Proposition 2.

4.2 QED limits for Erlang-R with blocking

We now continue our analysis by examining the limiting behavior under scaling (5). We first derive QED limits for some performance measures of the Erlang-R model with blocking. Using the explicit expressions for the blocking model in (3), we derive the limiting values of the relevant performances of the restricted Erlang-R model defined in §3.2 in terms of β and γ.

Theorem 1 *Let s and n scale as in* (5) *with* $-\infty < \beta < \infty$, $\gamma > 0$ *as* $\lambda \to \infty$. *Then, if* $\beta \neq 0$,

$$g^b(\beta,\gamma) := \lim_{\lambda\to\infty} \mathbb{P}^b(\text{delay}) = \left(1 + \frac{\beta \int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\beta^2+\eta^2})\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}\right)^{-1}, \tag{6}$$

$$\begin{aligned} f^b(\beta,\gamma) &:= \lim_{\lambda\to\infty} \sqrt{R_1}\cdot\mathbb{P}^b(\text{block}) \\ &= \frac{\sqrt{r}\phi(\gamma)\Phi(-\omega\sqrt{r}) + \phi(\sqrt{\beta^2+\eta^2})\,\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}{\int_{-\infty}^{\beta}\Phi\left(\frac{\gamma-t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}, \end{aligned} \tag{7}$$

$$\begin{aligned} h^b(\beta,\gamma) &:= \lim_{\lambda\to\infty} \sqrt{R_1}\cdot\mathbb{E}[W] \\ &= \frac{\frac{\phi(\beta)\Phi(\eta)}{\beta^2} + \left(\frac{\beta}{r} - \frac{\gamma}{\sqrt{r}} - \frac{1}{\beta}\right)\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega) - \sqrt{\frac{1-r}{r}}\frac{\phi(\beta)\phi(\eta)}{\beta}}{\int_{-\infty}^{\beta}\Phi\left(\frac{\gamma-t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}\mathrm{e}^{\frac{1}{2}\omega^2}\Phi(\omega)}, \end{aligned} \tag{8}$$

and if $\beta = 0$,

$$g_0^b(\gamma) := \lim_{\lambda\to\infty} \mathbb{P}^b(\text{delay}) = \left(1 + \frac{\int_{-\infty}^{0}\Phi\left(\frac{\gamma-t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta)+\phi(\eta)\right)}\right)^{-1}, \tag{9}$$

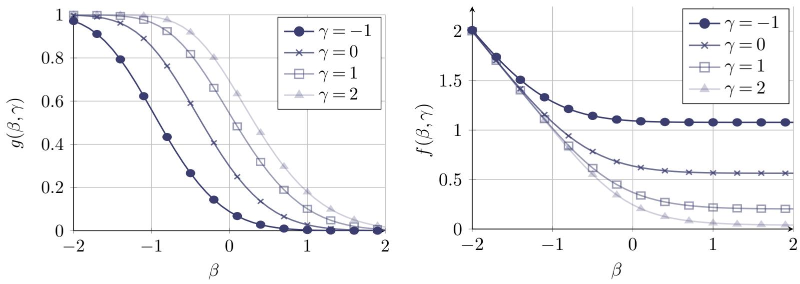
$$f_0^b(\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1}\cdot\mathbb{P}^b(\text{block}) = \frac{\sqrt{r}\,\phi(\gamma)\Phi(-\omega\sqrt{r}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{\beta}\Phi\left(\frac{\gamma-t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta)+\phi(\eta)\right)}, \tag{10}$$

$$h_0^b(\gamma) := \lim_{\lambda\to\infty} \sqrt{R_1}\cdot\mathbb{E}[W] = \frac{1}{2\mu}\,\frac{\left(\gamma^2/r+1\right)\Phi(\eta)+\eta\phi(\eta)}{\frac{r}{1-r}\sqrt{2\pi}\int_{-\infty}^{0}\Phi\left(\frac{\gamma-t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{r}{1-r}}\left(\eta\Phi(\eta)+\phi(\eta)\right)}, \tag{11}$$

where $\eta = \frac{\gamma-\beta\sqrt{r}}{\sqrt{1-r}}$, $\omega := \frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}$, *and* ϕ *and* Φ *are the PDF and CDF of a standard Normal distribution, respectively.*

The proof of Theorem 1 is provided in Appendix C. This theorem underscores the favorable performance of the QED regime: the blocking probability vanishes at a rate of $O(1/\sqrt{R_1})$, implying that, asymptotically, all patients are admitted to the system. Meanwhile, the probability of waiting for an available nurse converges to a non-degenerate limit strictly between $(0,1)$, and even when a patient needs to wait, their expected waiting time vanishes at order $O(1/\sqrt{R_1})$.

More specifically, Theorem 1 proves that the scaling (5) the probability of delay in Equations (6) and (9) converges to a limit that is strictly between 0 and 1. Notice that all limits in Theorem 1 are functions of three parameters: β (the QED slack parameter associated with server capacity) and γ (the QED slack parameter associated with space capacity), which are decision variables, and the fraction of needy time r, which is dictated by the physics of the system. Theorem 1 also shows that the probability of blocking (Equations (7) and (10)) is of order $1/\sqrt{R_1}$. For example, assume that the fraction of needy time r is 0.5 and the system is large (100 servers). Using Figure 7, we observe that, by choosing the pair $\gamma = 1$ and $\beta = 0.245$, we actually aim at a probability of getting served at the needy queue immediately to be 40%. At the same time, the probability of immediately getting a bed is 97%. Thus, our QED policy puts more emphasis on preventing blocking than waiting.



(a) Delay probability (b) Scaled blocking probability

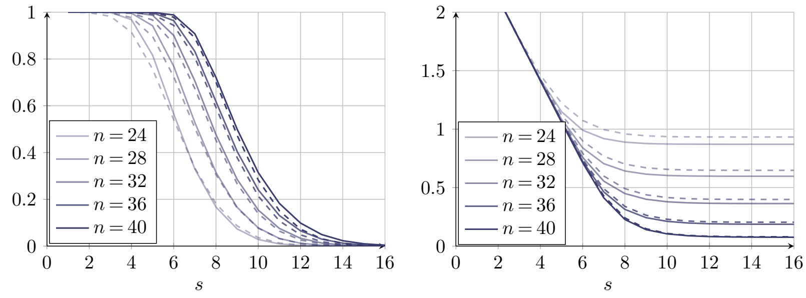
Fig. 7 Asymptotic delay and scaled blocking probability for $r = 0.5$ based on Theorem 1.

Theorem 1 further shows that the expected waiting (Equations (8) and (11)) is of order $1/\sqrt{R_1}$ too and hence vanishes in the large-system limit. We see from Theorem 1 that achieving target service levels is always an interplay between the slack parameters β and γ. Figure 7a shows for instance that in order to keep $\mathbb{P}(\text{delay}) \in (0.25, 0.75)$, choosing $\gamma = -1$ requires β to stay within the range $[-1.4, -0.5]$, while $\gamma = 1$ corresponds to values of β in $[-0.4, 0.5]$.

While the two-fold scaling rule in (5) automatically captures the right dimensioning ratio as the system scales up, Theorem 1 shows that the parameters β and γ provide a means to fine-tune the performance. Figure 7b confirms how adding servers (nurses), i.e. increasing β does not improve the blocking probability if the number of spaces (beds), i.e. γ is too tight. This is in accordance with our previous observations in Figure 4 for the exact steady-state distribution.

To test the accuracy of the asymptotic results in Theorem 1 as approximations in a realistic setting, we plot in Figure 8 the exact probability of delay and blocking for an Erlang-R model with $R_1 = 8$ and $r = 0.25$, as a function of s. The exact probabilities are given by Equation (4), and their respective asymptotic approximations are based on Theorem 1. Despite the realistic moderate size of the system ($R_1 = 8$), we see that the QED approximations are remarkably accurate for many settings (s, n). This fast relaxation is in line with observations made earlier in the QED literature [37, 43]. In

Appendix E, we further compare the asymptotic delay and blocking probability in several scenarios. The numerical results show that $g^b(\beta, \gamma)$, $f^b(\beta, \gamma)$ and $h^b(\beta, \gamma)$ provide accurate approximations to $\mathbb{P}(\text{delay})$, $\sqrt{R_1}\mathbb{P}(\text{block})$ and $\sqrt{R_1}\,\mathbb{E}[W]$ in pre-limit systems. The quality of the approximations increases with R_1. Fluctuations occur for relatively small values of R_1, since s and n need to be rounded to an integer.



(a) Delay probability (b) Scaled blocking probability

Fig. 8 Comparison of exact performance measures (solid) against asymptotic approximations (dashed) with $\beta = (s - R_1)/\sqrt{R_1}$ and $\gamma = (n - R_1/r)/\sqrt{R_1/r}$ for $\lambda = 2$, $\mu = 1$, $\delta = 0.25$ and $p = 0.75$.

4.3 QED approximations for Erlang-R with holding

As explained in Section 3.1.2, the model with holding has no product-form steady-state distribution, which makes it hard (if not impossible) to obtain QED limits. Instead, we derive QED approximations by exploiting a connection with the blocking model. We first prove that under scaling (5), the upper bound on the utilization level of the severs needed to achieve stability in the model with holding, as given in Proposition 1, converges to unity as $R_1 \to \infty$. This facilitates high utilization levels of both nurses and beds, a key characteristic of the QED regime.

Proposition 3 *Let s and n scale with $R_1 \to \infty$ as in (5). Then, for $\lambda \to \infty$, $\rho_{\max}(s, n) \to 1$.*

The proof can be found in Appendix D.

As observed before, the nature of the two variants of the model is similar except for the fact that a fraction of the patients is blocked on arrival in the setting with blocking, whereas all the arriving patients are eventually admitted into the system in the holding model. This implies that, given s and n, the nurses face an increased workload in case of a holding mode. In fact, Theorem 1 shows that the blocking probability is of order $1/\sqrt{R_1}$, yielding a volume of blocked patients of order $\sqrt{R_1}$ in the setting with blocking. Accordingly, if $R^b = R_1$ and R^h denotes the nominal load arriving to the nurses in the model with blocking and holding, respectively, we can argue that

$$R^h = R^b + \alpha\sqrt{R^b} + o(\sqrt{R^b}), \tag{12}$$

for some $\alpha > 0$. Notice that this additional load is of the same order as the safety staffing in the blocking model staffing rule (5). As s and n remain unchanged, we rewrite (5) in terms of R^h,

$$s = R^h + (\beta - \alpha)\sqrt{R^h} + o(\sqrt{R^h}),$$
$$n = \frac{R^h}{r} + \left(\gamma - \frac{\alpha}{\sqrt{r}}\right)\sqrt{\frac{R^h}{r}} + o(\sqrt{R^h}), \tag{13}$$

where we have used $R^b = O(R^h)$. Observe that the square-root principle prevails also after this substitution, albeit with different hedging parameters. We therefore heuristically argue that the holding model under scaling (5) with parameters β and γ mimics the blocking model with parameters $\beta - \alpha$ and $\gamma - \alpha/\sqrt{r}$, respectively.

Observe, however, that we have not yet specified the value of α. By definition, $\alpha\sqrt{R^b}$ is the expected volume of customers that would be rejected in the model with blocking, that is, R^h times the probability of not being admitted to the system directly. By the construction in (13), this volume asymptotically equals $R^h \cdot \mathbb{P}^b(\text{block})$, with parameters $\beta - \alpha$ and $\gamma - \alpha/\sqrt{r}$, which by Theorem 1 is approximated by

$$f^b\left(\beta - \alpha, \gamma - \alpha/\sqrt{r}\right)/\sqrt{R^h},$$

as R^h grows large. In conclusion, α is characterized as the solution of the fixed-point equation

$$\alpha = f^b\left(\beta - \alpha, \gamma - \alpha/\sqrt{r}\right), \tag{14}$$

and as a result, we are able to approximate the delay probability in the holding model as

$$\mathbb{P}^h(\text{delay}) \approx g^b(\beta - \alpha, \gamma - \alpha/\sqrt{r}) =: g^h(\beta, \gamma). \tag{15}$$

Likewise, the scaled mean waiting time for a server can be approximated by

$$\sqrt{R_1} \cdot \mathbb{E}[W] \approx h^b(\beta - \alpha, \gamma - \alpha/\sqrt{r}) =: h^h(\beta, \gamma). \tag{16}$$

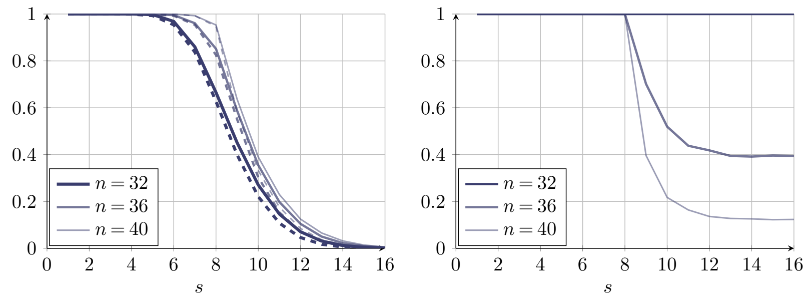
Note that (12) implies that the volume of customers that are blocked in the blocking system, but successfully retained in the holding system, is of order $O(\sqrt{R_1})$. Because these retained customers are precisely the drivers of congestion within the holding queue, and since QED fluctuations inherently occur on the $\sqrt{R_1}$ scale, the holding queue length itself is of order $O(\sqrt{R_1})$ as well. Subsequently, we argue that the expected holding time under the QED policy is $O(1/\sqrt{R_1})$, and is therefore asymptotically negligible. We justify this claim numerically in §6.

Remark 2 Notice that in the reasoning leading to (14), we implicitly assumed that the additional volume $\alpha\sqrt{R_1}$ is an independent Poisson process, which is obviously not the case. Therefore, (15)–(16) provide approximations for finite systems that are not asymptotically exact as $\lambda \to \infty$.

The approximation should be viewed as a heuristic fixed-point approximation in the QED regime, in the spirit of [40, 46, 47]. The key idea is to replace the original analytically intractable system by a tractable surrogate whose parameters are determined self-consistently through a fixed-point equation. We follow the same philosophy here for the holding network.

Nevertheless, the approximation performs well in all numerical experiments reported here and in Appendix E. Moreover, in all experiments the fixed-point equation (14) admitted a unique solution for α, although we currently do not have a proof of this property.

In Figure 9, we repeat the numerical experiments of Figure 8 for the model with holding. Since the heuristic does not provide an approximation for the holding probability, Figure 9b only plots the simulated holding probabilities. Those are provided to better understand the implication of operational decisions. Recall that the holding system is only stable (i.e. $\mathbb{P}(\text{hold}) < 1$) if both $s > R_1 = 8$ and $n > R_1/r = 32$. We nevertheless included the boundary case $n = 32$ for illustrative purposes. The graphs in Figure 9 show that the heuristic captures the congestion levels well, even for this moderate-size system.



(a) Delay probability (b) Holding probability

Fig. 9 Comparison of simulated delay probability (solid) against asymptotic approximations (dashed) with $\beta = (s - R_1)/\sqrt{R_1}$ and $\gamma = (n - R_1/r)/\sqrt{R_1/r}$ for $\lambda = 2$, $\mu = 1$, $\delta = 0.25$ and $p = 0.75$.

We also checked how the performance of the heuristic approximation under different settings, and particularly, when $R_1 \to \infty$. In these experiments, we have compared the approximated delay probability in the Erlang-R model with holding as a solution of the fixed-point procedure to the outcomes of simulation experiments. We conclude that the approximation is good. As R_1 increases, the simulated values move closer to the heuristic approximation. Extensive numerical experiments suggest that load is slightly underestimated in the limit. The best results in terms of accuracy are attained for small r. This suggests that the quality of the heuristic method improves as r gets smaller. These are exactly the parameter settings for which this model is relevant.

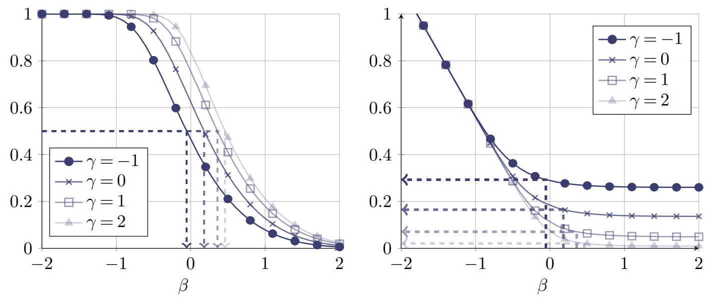
5 Dimensioning

We will now use the accurate asymptotic approximations of the previous section to define a procedure that determines resource capacity in the restricted Erlang-R models. That is, we aim to set the number of servers s and the number of places n, such that a preset performance level is achieved. We take the probability of delay at the needy queue and the probability of blocking/holding as the target performance objectives.

5.1 Capacity setting for Erlang-R with blocking

In the setting with blocking, we can readily use the asymptotic results of Theorem 1 to (numerically) find a pair of parameters (β^*, γ^*) to meet the performance requirements. For instance, given that we want the delay probability to be at most ε, we first solve the equation $g^b(\beta^*, \gamma^*) = \varepsilon$ and then assign $s = \lceil R_1 + \beta^* \sqrt{R_1} \rceil$ and $n = \lceil R_1/r + \gamma^* \sqrt{R_1/r} \rceil$. Note that there could be multiple solutions to that problem, i.e. there could be multiple combinations of the number of beds and the number of nurses that can result in the same value of a single performance level. The system manager can ultimately decide which of these feasible solutions fits the environment best, for instance taking into account space and cost constraints.

We illustrate the resource allocation decisions in an MU setting, using data from two articles: Lundgren and Segesten [54] and Green and Yankovic [18]. Green and Yankovic describe an MU that has 42 beds, with an average occupancy level of 78%, and an Average LOS of 4.3 days. Lundgren and Segesten studied nurses' service times in a medical-surgical ward. They found that the average service time in their unit was 15.3 minutes per service, and that the average demand rate for each patient was 0.38 requests per hour. Therefore, we take an average service time of 15 minutes and assume that there are 0.4 requests per hour from each patient. Fitting this data to our model results in the following parameter values: $\lambda = 0.32, \mu = 4, \delta = 0.4$, $p = 0.975$ and the fraction of needy time is then approximately $r = 0.09$. This yields a nominal offered load of $R_1 = 3.2$ and $R_1/r = 34.4$. Figure 10 visualizes



(a) Delay probability (b) Blocking probability

Fig. 10 Approximate performance of restricted Erlang-R with blocking for $r \approx 0.09$ and $R_1 = 3.2$, as functions of β.

the dimensioning procedure for this particular MU. The hospital management can find a pair of n and s to meet certain criteria, e.g. to achieve target delay probability $\varepsilon = 0.5$ with reasonable blocking probability. Figure 10a indicates that this target can be achieved by a variety of pairs, for instance $(\beta_1, \gamma_1) = (-0.06, -1)$,

$(\beta_2, \gamma_2) = (0.16, 0)$, $(\beta_3, \gamma_3) = (0.36, 1)$ or $(\beta_4, \gamma_4) = (0.46, 2)$, among infinitely many others. From to Figure 10b, the pairs named above lead to blocking probabilities 0.293, 0.165, 0.071 and 0.021, respectively. If the manager decides that the probability of blocking of more than 10% is not acceptable, this leaves the choices $(\beta_3, \gamma_3) = (0.36, 1)$ or $(\beta_4, \gamma_4) = (0.46, 2)$ as candidate parameter pairs. Using the two-fold square-root staffing rule $s_i = \lceil R_1 + \beta_i \sqrt{R_1} \rceil$ and $n_i = [R_1/r + \gamma\sqrt{R_1/r}]$, this yields feasible staffing levels $(s_3, n_3) = (4, 40)$ and $(s_4, n_4) = (5, 46)$. The ultimate decision to apply any of these solutions can be based on external factors, such as operational costs or space limitations in the unit.

5.2 Capacity setting for Erlang-R with holding

For the holding model, we need a more sophisticated approach, exploiting the asymptotic approximation with the fixed-point equation in (14). We propose the following dimensioning procedure to achieve a preset target delay probability at the needy queue.

Algorithm 1 Stationary dimensioning algorithm

Require: Target delay probability ε. Parameters λ, μ, δ and p.

1: **procedure** FINDSTAFFING(s, n)
2: Set $R_1 := \frac{\lambda}{(1-p)\mu}$ and $r := \frac{\delta}{\delta + p\mu}$
3: Determine parameters (β^*, γ^*) such that $g^b(\beta^*, \gamma^*) = \varepsilon$
4: Set $\beta = \beta^* + f^b(\beta^*, \gamma^*)$ and $\gamma = \gamma^* + f^b(\beta^*, \gamma^*)/\sqrt{r}$
5: **return** $s = \lceil R_1 + \beta\sqrt{R_1} \rceil$ and $n = \lfloor R_1/r + \gamma\sqrt{R_1/r} \rfloor$
6: **end procedure**

Remark 3 In Step 3 of Algorithm 1 infinitely many pairs (β^*, γ^*) satisfy the delay probability equation. For practical purposes, it is convenient to fix either β^* or γ^* beforehand, and then solve $g^b(\beta^*, \gamma^*) = \varepsilon$ for the remaining unknown. The preset value should however be chosen with care, since $g^b(\beta^*, \gamma^*)$ is upper bounded by the Halfin-Whitt delay probability formula $g_{\mathrm{HW}}(\beta^*) = (1 + \beta^*\Phi(\beta^*)/\phi(\beta^*))^{-1}$. In particular, if $\varepsilon > g_{\mathrm{HW}}(\beta^*)$, then no feasible solution to $g^b(\beta^*, \gamma^*) = \varepsilon$ exists. Furthermore, it is evident from Step 4 that the final values (β, γ) are always larger than (β^*, γ^*).

Moreover, we point out that in Step 5, we round the value of s up, while rounding the value of n down. This is related to the empirical findings of the previous section, in which we saw that the heuristic slightly underestimates the delay probability. Since increasing n increases $\mathbb{P}(\text{delay})$, just as reducing s increases $\mathbb{P}(\text{delay})$, we hope to diminish this discrepancy. Finally, the adjustments needed to adapt Algorithm 1 to time-varying environments are given in Section 6.4. Using that procedure stabilize the probability of delay in a similar way to [7].

We now use the same example as in §5.1 to demonstrate capacity allocation decisions for the model with holding. This can be viewed as the additional capacity the MU needs in terms of nurses and beds, in order to account for the fact that patients are waiting in the ED to be admitted, into the preferred MU, instead of being blocked and transferred to a less preferred unit. Observe that the holding model leaves less flexibility for management in choosing system parameters due to stability constraints. For example, the policy with $n = 30$ ($\gamma = -0.75$) is infeasible in the holding model. For similar reasons, only nurse staffing levels with $\beta > 0$, or $s > R_1 = 3.2$ are feasible.

Targeting a delay probability of 0.5 with $n = 40$, Figure 11 shows that operating an MU with holding room requires $\beta = 0.475$ or $s = 5$. Recall that under the blocking policy, only $s = 4$ nurses were needed to achieve a delay probability of 0.5. This example hence shows how the managerial decision to have a holding room, rather than deferring patients to less preferred medical units, requires additional staffing in that unit (as well as the ED). This example also shows that the facility with a holding room is able to treat fewer patients simultaneously than under blocking constraints, in line with the bounds in §3.3 and Conjecture 1.

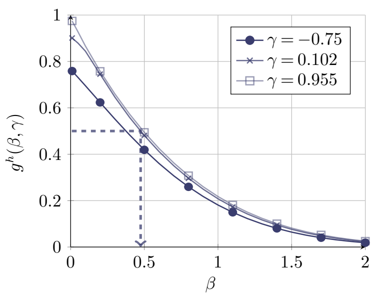


Fig. 11 Approximate delay probability of restricted Erlang-R system with holding for $r \approx 0.09$ and $R_1 = 3.2$.

6 Model analysis and operational implications

In this section, we use the analysis and algorithms developed in earlier sections to gain insights into the importance of the capacity restrictions and customer returns in a restricted Erlang-R system by drawing a comparison to related models studied in the literature.

6.1 The influence of customer returns or the role of r

Here we study how the parameter r affects the service level in the restricted Erlang-R model with blocking, on the basis of the asymptotic expressions in Theorem 1.

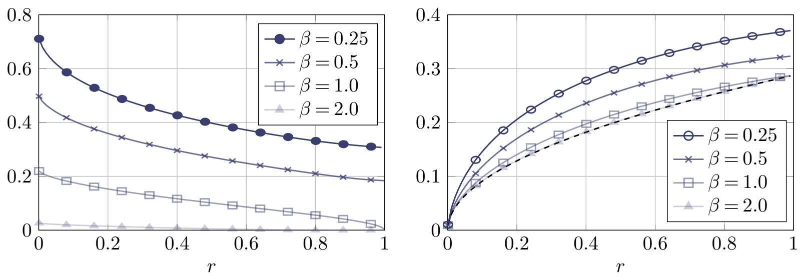
To better understand the connection with the single-station model and the importance of returns we examine the role of r. Recall the interpretation of r as the fraction of time a patient is needy during his stay within the system in the idealized scenario with infinite capacity ($r \in (0, 1)$). The case $r = 1$ corresponds to the setting wherepatients are needy all the time, in which case customers get service consecutively. When $r = 1$ the infinite-server queue, describing the number of content patients, disappears from the queueing system and we end up with a standard loss model—$M/M/s/n$ queue—in which capacity is scaled as $s = R_1 + \beta\sqrt{R_1}$, $n = R_1 + \gamma\sqrt{R_1}$. This staffing rule only makes sense when $\beta < \gamma$, since no delay is experienced if $n \leq s$. If indeed $\gamma > \beta$, then the asymptotic delay and the scaled blocking probabilities are given by [55],

$$g_B(\beta, \gamma) = \frac{1 - e^{-\beta(\gamma-\beta)}}{1 - e^{-\beta(\gamma-\beta)} + \beta\Phi(\beta)/\phi(\beta)}, \qquad f_B(\beta, \gamma) = \frac{\beta e^{-\beta(\gamma-\beta)}}{1 - e^{-\beta(\gamma-\beta)} + \beta\Phi(\beta)/\phi(\beta)}.$$

We can see that blocking probability approaches a lower bound function (dash line) as β increases (Figure 12b). To understand this, observe that as β grows, delays at the needy queue vanish (Figure 12a). Then the sojourn time of an admitted patient only consists of a geometric number of needy and content periods with mean $(1/\mu + p/\delta)/(1-p) = 1/r\mu(1-p)$. The blocking model can in this case be modeled as an $M/G/n/n$ queue, with offered load $\lambda/(r\mu(1-p)) = R_1/r$, in which the scaled blocking probability is,

$$\sqrt{R_1}\,\mathbb{P}(\text{block}) = \sqrt{R_1}\,\frac{(R_1/r)^n/n!}{\sum_{k=0}^{n}(R_1/r)^k/k!} \to \sqrt{r}\,\frac{\phi(\gamma)}{\Phi(\gamma)},$$

as $R_1 \to \infty$, see [56]. This function of r is plotted in Figure 12b as the dashed line.



(a) Delay probability (b) Scaled blocking probability

Fig. 12 Asymptotic performance measures as a function of r in the restricted Erlang-R model with blocking for $\gamma = 1$.

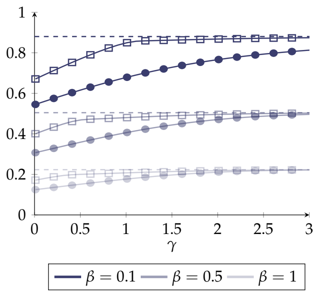
We observe that in general the probability of blocking increases with r, regardless of the capacity constraints on the needy station. We can explain this by observing that r influences only n in the QED staffing rule. When n reduces, more customers are blocked. Therefore, if customers spend relatively more time in the needy state, which usually indicates services that have fewer stages (number of returns is lower), blocking will increase. Delays, on the other hand, will decrease in such situations—the minimal delay possible can be achieved if service is aggregated into one round ($r = 1$). Returns or interruptions increase delays significantly under QED staffing.

Operational implications

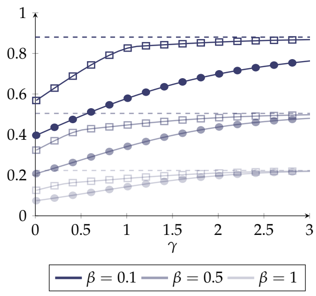
1. Returning customers should be explicitly accounted for in determining staffing in a system with space constraints both in steady-state and transient conditions.
2. The above becomes more important as the proportion of time spent in the needy state becomes small, since then the number of customers contributing to the space constraints increases.
3. As returns become more spread over the patient's length-of-stay (r decreases), delay increases and blocking decreases.

6.2 Comparing restricted and unrestricted Erlang-R models

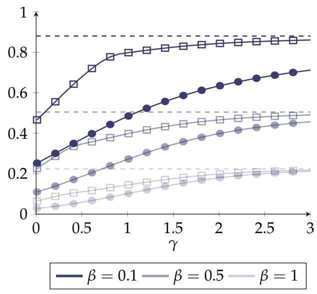
Given the expressions for the asymptotic delay probability in the open Erlang-R model, and its restricted versions with blocking and holding, we compare the three policies for various values of β, γ and r. Figure 13 plots the delay probability for blocking ($g^b(\beta, \gamma)$), holding ($g^h(\beta, \gamma)$) and Erlang-R ($g_{\text{HW}}(\beta)$) models, as functions of γ, while keeping β fixed, for three values of r.



(a) $r = 0.1$



(b) $r = 0.25$



(c) $r = 0.5$

Fig. 13 Asymptotic delay probability in open Erlang-R (dashed), restricted Erlang-R with blocking (•) and restricted Erlang-R with holding (□), as a function of γ.

We make a couple of observations. Notice that

$$g^b(\beta, \gamma) \leq g^h(\beta, \gamma) \leq g_{\text{HW}}(\beta)$$

for all $\beta, \gamma > 0$ and r. In that sense, the holding model is an interpolation between the blocking and the open model. As expected, the delay probabilities in the restricted models converge to those of the open Erlang-R model, because increasing γ is tantamount to lifting the stringent constraints on the system size. Note that the rate of convergence is fast—one can provide probability of delay close to that of the open model with small values of γ. Indeed it is important to note that when using QED staffing not much of the excessive delay results from the space restriction. Also, we observe that the difference between delay probabilities increases with r. Hence, taking into account the re-entrant structure of the network is most important if $r << 1$ and more so if applying blocking (compared to the holding policy).

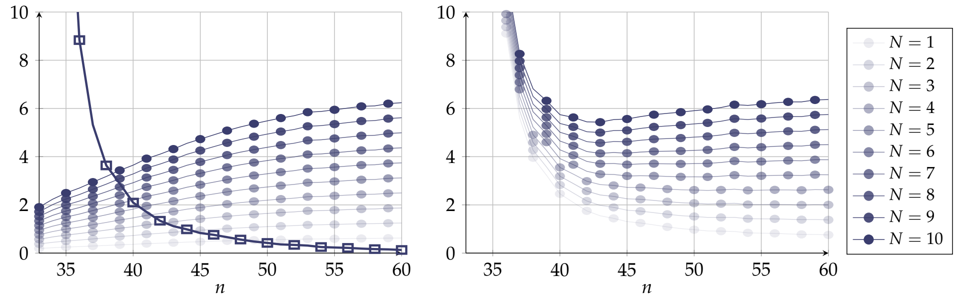
6.3 The impact of customer complexity

We next reflect on the impact of operational capacity decisions on different customer populations. We measure patient's complexity by the number of times he requires service by the nurse or the physician during her stay. In the ED context, simple-to-treat patients will need to see the physician once, while complex ones will need multiple visits. Hence, we divide the patients into complexity groups by the number of visits in the needy station. Since the number of visits is geometrically distributed, we have a higher proportion of simple patients than complex ones; that fits well the healthcare environment. Bayati et al. [17], Schull et al. [57] showed how low-complexity patients impact complex ones. We suggest here that the re-entrant structure is part of the cause as complex patients need to wait repeatedly in the inner queue for servers. In such situations, creating a holding queue may limit that phenomena.

Figure 14 shows the waiting time in the needy and pre-entering queues, and the total waiting time, as a function n (number of beds), for each complexity group. Obviously, the expected waiting time in the pre-entering queue decreases with n, while the needy waiting time increases. For patients who require a relatively large number of visits of the physician, in this case more than 6, the total needy wait is the dominant part of the total waiting time. Therefore, as n grows, the total waiting time first decreases and then increases. In fact, Figure 14b suggests that there is an optimal number of beds n that minimizes the total wait for each complexity type. Thus, size restrictions reduce the length-of-stay of patients with complex health conditions (given that the constraint is not too tight). On the other hand, this figure also shows that no such n exists for patients who only require little assistance. Hence, there is no n that improves the sojourn time of all patients in the ED simultaneously. This leaves the decision to the hospital manager to weigh the importance of patients of different complexity levels [17, 57].

6.4 Case study: Comparison of operational decisions

We now illustrate how the managerial decision to operate under a specific operational regime affects ED performance in terms of efficiency and quality-of-care, through a case study. The practical environment we investigate is the ED of a moderately-sized hospital, which faces the arrival pattern $\lambda(t)$ plotted in Figure 15a on a typical workday. Other parameters of the model are estimated to be $\mu = 6.67$, $\delta = 2.18$ and $p = 0.76$, so that $r = 0.301$. (Parameters were taken from [7]). In order to set time-varying staffing levels $s(t)$ and $n(t)$, we adopt the modified offered load (MOL) fluid approximation of the demand process of [58]. This approach initially presumes infinite capacity to obtain the expected number of customers $R(t)$ in the queueing system as a function of time. This offered load function then replaces



(a) Expected pre-entering waiting (□) and needy waiting times (•)

(b) Total expected waiting times

Fig. 14 Expected waiting times as a function of n given the number of visits N in the Erlang-R model with holding with $\lambda = 2$ $\mu = 1$, $\delta = 0.25$, $p = 0.75$ and $s = 9$.

the (constant) value of R in the stationary dimensioning scheme under consideration, in order to determine the adequate number of servers at each point in time. Following this idea in our two-dimensional queueing system, we find the offered load function for the nurses $R_1(t)$ and the offered load function for the beds $R_1(t) + R_2(t)$ as the solution of the system of ODEs,

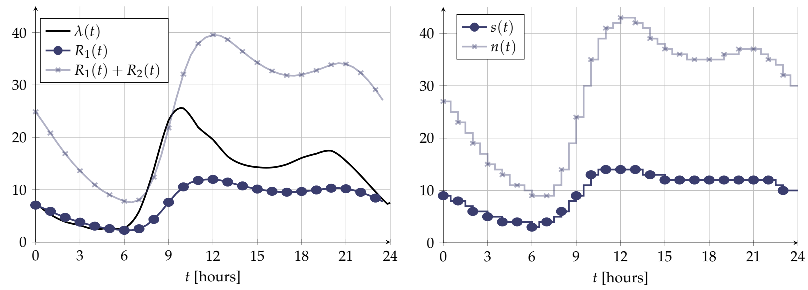
$$\frac{d}{dt}R_1(t) = \lambda(t) + \delta R_2(t) - \mu R_1(t), \tag{17}$$
$$\frac{d}{dt}R_2(t) = p\mu R_1(t) - \delta R_2(t),$$

see Yom-Tov and Mandelbaum [7, Thm. 2] for details. For this case study's parameters, these offered load functions are also plotted in Figure 15a. While the number of nurses can be adjusted in a relatively flexible manner, the value of n, which echoes a hard restriction on the ED capacity, is naturally less amenable to fluctuations. The reason is that the maximum ED capacity is to a large extent determined by its hardware, such as beds and medical equipment. However, the ED manager might deliberately consider reducing n during quieter periods of the day, e.g. during the night, by imposing bed-to-physician constraints. This is done, for example, when setting a case management constraint [12, 32]. Therefore, we consider the scenario in which both s and n are time-dependent but we do not force a constant case management quantity, rather let our new methodology recommend an appropriate time-varying one.

xtick = 0,3,6,9,12,15,18,21,24,

Extrapolating Algorithm 1 to the time-varying case, Step 2 is replaced by calculating $R_1(t)$ and $R_2(t)$ from Eq. (17) and Step 5 is replaced by

$$s(t) = R_1(t) + \beta\sqrt{R_1(t)},$$
$$n(t) = R_1(t) + R_2(t) + \gamma\sqrt{R_1(t) + R_2(t)},$$



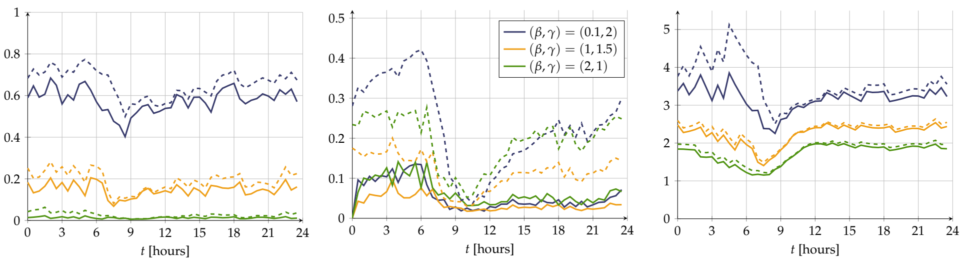
(a) Time-varying arrival rate and load functions (b) Capacity functions for $\beta = \gamma = 0.5$

Fig. 15 Empirical arrival rate and offered load functions in Israeli ED and corresponding capacity functions.

for some $\beta, \gamma > 0$. Since $R_1(t)$ and $R_2(t)$ are given, the QED staffing problem again reduces to finding the pair (β, γ).

Figure 15b plots the capacity functions for $\beta = 0.5$ and $\gamma = 0.5$, assuming capacity can only be adjusted every 30 minutes. In this case study, we consider three pairs of parameters (β, γ). For each of the pairs, we investigate (using simulation) the differences between the policy with blocking and holding by showing the time-varying performance indicators.

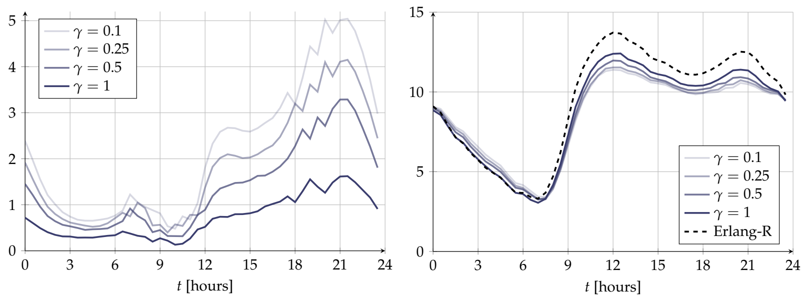
The simulation results are presented in Figure 16. Figure 16a shows that the MOL approach for capacity allocation roughly stabilizes the delay probability. Figure 16b shows that the fraction of patients not entering the ED on arrival in the blocking model is reasonable for all parameter pairs considered and are ordered according to γ. We also see a significant difference comparing between blocking and holding, that does not replicate itself in needy delays. This suggests an advantage for the holding policy, although it comes at the “cost" of some holding queues (Figure 17). Observe also that the holding probability drops in the period 8–13, which is exactly the period when the system is experiencing peak offered load. Hence, this temporary reduction is in line with our asymptotic findings that the probability of blocking/holding is $O(1/\sqrt{R_1})$.



(a) $\mathbb{P}(\text{delay})$ (b) $\mathbb{P}(\text{block})$ or $\mathbb{P}(\text{hold})$ (c) Nurse-to-patient ratio

Fig. 16 Simulation results for case study. Solid and dashed lines represent time-varying performance in the blocking and holding model, respectively.

Finally note that the three parameter settings lead to different nurse-to-patient ratio levels. Clearly, larger β leads to small nurse-to-patient ratios (due do larger staffing). Figure 16c demonstrates that for $(\beta, \gamma) = (1, 1.5)$ and $(\beta, \gamma) = (2, 1)$ the difference between the holding policy and the blocking policy is small. However, for $(\beta, \gamma) = (0.1, 2)$ we see a significant increase in the nurse-to-patient ratio during night hours. This may be due to the tight nurse schedule, that causes the holding queue to build up just before midnight. This queue then empties later on, causing an increase in workload per nurse in the period 24–7. To see the direct effect of the size restriction on the queue lengths, we plotted the mean holding and needy queue lengths in the holding model as a function of the parameter γ in Figure 17.



(a) Mean holding queue length (b) Mean needy queue length

Fig. 17 Simulated queue length of holding model with different values of γ.

Note that for all γ considered, the holding queue length is, as expected, of a smaller order than the number of patients admitted. Also, the holding queue length decreases as we increase γ. The needy queue length naturally approaches the expected queue length in the Erlang-R model as γ is increased.

7 Conclusions and future research

In this research we developed and analyzed a queueing network with repeated services that has capacity restrictions both on service and accessibility. We not only developed approximations for the performance of the system with a blocking policy, but also constructed a new innovative way to use those results to approximate performance of the intractable model with holding. Those approximations were based on a fixed-point analysis, and enabled us to take the first step towards characterizing the pre-entering queue behavior in the QED regime. In addition, the unified analysis enabled us to connect the two policies and to deepen the understanding of the practical application of the two. We showed in two realistic case studies the application of those two models to a medical unit resource allocation. We discussed the implication of holding patients both in terms of the range of possible combinations that results from stability constraints, and in terms of its impact on waiting in the system. Finally, we showed that the approximations are accurate for a wide range of parameters.

We point out that the our models could be generalized by incorporating features that are naturally supported by product-form queueing networks, such as multiple customer classes with different routing probabilities and classes with a deterministic number of visits to the needy queue. In fact, this could lead to rigorous analytical expressions for the setting as considered in §6.3. We believe this is an interesting avenue for future research.

Declarations

All authors contributed to the study conception and design. Analysis were performed by Galit Yom-Tov, Fiona Sloothaak and Johan S.H. van Leeuwaarden. The first draft of the manuscript was written by Galit Yom-Tov and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. No funding was received for conducting this study.

Appendix A Description of the QBD process

A.1 The QBD process

We consider the QBD process $X = \{N, Q_1\}$ in stationarity. Let $\nu(i) = \min\{i, s\}\mu$. To determine the (outgoing) transition rates of the process X we distinguish between the following cases:

- *Transitions from* $(0, 0)$*:* There are no patients in the Emergency Department and thus the only possible occurrence is when a new patient arrives. This results in a transition to $(1, 1)$ and occurs with rate λ.
- *Transitions from* $(i, 0), 1 \le i < n$*:* There are exactly i patients assigned to a bed of which none are seen by a nurse. Then either one of those patients becomes needy, or a new patient arrives at the Emergency Department that can immediately be seen by a nurse. The first results in a transition to $(i, 1)$ and occurs at rate $i\delta$, and the second results in a transition to $(i + 1, 1)$ and occurs with rate λ.
- *Transitions from* $(i, 0), i \ge n$*:* Again, the only possible transitions arise from either a newly-arrived patient or a patient assigned to a bed becoming needy. However, a newly-arrived patient finds all beds occupied and needs to wait. Thus, with rate λ we have a transition to $(i+1, 0)$ and with rate $n\delta$ a transition to $(i, 1)$.
- *Transitions from* $(i, i), i < n$*:* In this case all patients assigned to a bed are in need of service. With rate λ a new patient arrives at the Emergency Department. He joins the (possible) queue to be seen by a nurse immediately, so this results in a transition to $(i + 1, i + 1)$. Moreover, since there are only $s < n$ nurses, a service completion occurs with rate $\nu(i)$. With probability p the patient turns to the holding phase, so in total we still have i patients with one patient less in queue for a nurse. With probability $1 - p$ the patient leaves the Emergency Department, decreasing both N and Q_1 by one. In other words, with rate $p\nu(i)$ we have a transition to $(i, i - 1)$ and with rate $(1 - p)\nu(i)$ we have a transition to $(i - 1, i - 1)$.

- *Transitions from* (n, n)*:* Similar to the previous case, we have a transition to $(n, n-1)$ with rate $ps\mu$ and with rate $(1-p)s\mu$ we have a transition to $(n-1, n-1)$. In this case however, a newly-arrived patient finds all beds occupied, resulting in a transition to $(n+1, n)$ with rate λ.
- *Transitions from* $(i, n), i > n$*:* We have a transition to $(i+1, n)$ with rate λ and a transition to $(i, n-1)$ with rate $ps\mu$. In case that a patient leaves the Emergency Department there are $i - n > 0$ patients in the holding room waiting for an available bed. Thus, one of the $i - n$ patients in the holding room is assigned to the available bed in need of service. That is, with rate $(1-p)s\mu$ we have a transition to $(i-1, n)$.
- *Transitions from* $(i, j), 1 \leq j < i < n$*:* There are four possible transitions. First, with rate λ there is a new arrival which results in a transition to $(i+1, j+1)$. Second, with rate $(i-j)\delta$ a patient in one of the beds becomes needy, which results in a transition to $(i, j+1)$. Third, with rate $p\nu(j)$ a patient turns to the content state after service completion, which results in a transition to $(i, j-1)$. Last, with rate $(1-p)\nu(j)$ a patient leaves the Emergency Department after service completion, which results in a transition to $(i-1, j-1)$.
- *Transitions from* $(n, j), 1 \leq j < n$*:* This case is similar to the previous one. The only difference arises when a new patient arrives, since all n beds are already occupied. Thus, with rate λ we have a transition to $(n+1, j)$.
- *Transitions from* $(i, j), i > n, 1 \leq j \leq n$*:* This case is the previous one, except when a patient leaves the Emergency Department after service completion. Then one of the $(i-n)$ patients in the holding room will be assigned to a bed in need of service. This results in a transition to $(i-1, j)$ with rate $(1-p)\nu(j)$.

The state space can be partitioned according to its levels, where level i corresponds to a total queue length $N = i$ patients. This results in an infinite-sized matrix consisting of blocks, where each block corresponds to the transition flow from one level to another. Since the only transitions allowed are within the same level or between two adjacent levels in a QBD-process, we obtain a tridiagonal block structure. Each block consists of elements representing the transition rate of one state to another, and therefore each block is a matrix of size at most $(n+1) \times (n+1)$.

For the Erlang-R model with holding this gives the following result. Let P denote the transition matrix of the process $\{N(t), Q_1(t)\}$. We have the boundary levels $\{1, 2, ..., n\}$ and P is of the form

$$P = \begin{pmatrix} B_{00} & B_{01} & & & & & & & \\ B_{10} & B_{11} & B_{12} & & & & & & \\ & B_{21} & B_{22} & B_{23} & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & & B_{n\,n-1} & & & & \\ & & & B_{n-1\,n} & B_{nn} & A_0 & & & \\ & & & & A_2 & A_1 & A_0 & & \\ & & & & & A_2 & A_1 & A_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where $B_{ii} \in \mathbb{R}_{\mathbb{1}}^{(i+1)\times(i+1)}$, $B_{i\,i-1} \in \mathbb{R}_{\mathbb{1}}^{(i+1)\times i}$, $B_{i-1\,i} \in \mathbb{R}_{\mathbb{1}}^{i\times(i+1)}$, and $A_0, A_1, A_2 \in \mathbb{R}_{\mathbb{1}}^{(n+1)\times(n+1)}$. The matrices of transition rates for the boundary states are given by

$$B_{00} = (-\lambda),\ B_{i-1\,i} = \begin{pmatrix} 0 & \lambda & & \\ & \ddots & \lambda & \\ & & \ddots & \ddots \\ & & 0 & \lambda \end{pmatrix},$$

$$B_{i\,i-1} = \begin{pmatrix} 0 & & \\ (1-p)\mu & 0 & \\ & (1-p)\nu(2) & \ddots \\ & & \ddots & 0 \\ & & & (1-p)\nu(i) \end{pmatrix},$$

and

$$B_{ii} = \begin{pmatrix} -(\lambda + i\delta) & i\delta & & \\ p\mu & -(\lambda + \mu + (i-1)\delta) & (i-1)\delta & \\ \ddots & \ddots & \ddots & \\ & p\nu(i-1) & -(\lambda + \nu(i-1) + \delta) & \delta \\ & & & p\nu(i) - (\lambda + \nu(i)) \end{pmatrix}.$$

Moreover, the transition rates are given by

$$A_0 = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \ddots & \\ & & & \lambda \end{pmatrix},\ A_2 = \begin{pmatrix} 0 & & & & & & \\ & (1-p)\mu & & & & & \\ & & 2(1-p)\mu & & & & \\ & & & \ddots & & & \\ & & & & s(1-p)\mu & & \\ & & & & & \ddots & \\ & & & & & & s(1-p)\mu \end{pmatrix},$$

and

$$A_1 = \begin{pmatrix} -(\lambda+n\delta) & n\delta & & & & & \\ p\mu & -(\lambda+\mu+(n-1)\delta) & (n-1)\delta & & & & \\ \ddots & \ddots & \ddots & & & & \\ & sp\mu & -(\lambda+s\mu+(n-s)\delta) & (n-s)\delta & & & \\ & \ddots & \ddots & \ddots & & & \\ & & sp\mu & -(\lambda+s\mu+\delta) & \delta \\ & & & sp\mu & -(\lambda+s\mu) \end{pmatrix}.$$

A.2 Stability condition

From the general theory of QBD processes [51], it follows that the Markov process $\{N(t), Q_1(t)\}$ is ergodic (stable) if and only if

$$\pi A_0 e < \pi A_2 e, \tag{A1}$$

where e is the all-one column vector and $\pi = (\pi_0, ..., \pi_n)$ is the equilibrium distribution of the Markov process with generator $A_0 + A_1 + A_2$. In other words, π is such that

$$\pi(A_0 + A_1 + A_2) = 0,\ \pi e = 1, \tag{A2}$$

and

$$A_0 + A_1 + A_2 = \begin{pmatrix} -n\delta & n\delta & & & & & \\ p\mu & -(p\mu+(n-1)\delta) & (n-1)\delta & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & sp\mu & -(ps\mu+(n-s)\delta) & (n-s)\delta & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & ps\mu & -(ps\mu+\delta) & \delta \\ & & & & & ps\mu & -ps\mu \end{pmatrix}.$$

Then π must satisfy the balance equations

$$\begin{aligned} -n\delta\pi_0 + p\mu\pi_1 &= 0, \\ (n-j+1)\delta\pi_{j-1} - (p\nu(j) + (n-j)\delta)\pi_j + p\nu(j+1)\pi_{j+1} &= 0, \\ \delta\pi_{n-1} - ps\mu\pi_n &= 0, \end{aligned}$$

with $\nu(j) = \min\{j, s\}\mu$, and the normalization condition $\sum_{i=0}^{n} \pi_i = 1$. It is readily verified that

$$\pi_i = \begin{cases} \pi_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \leq i \leq s, \\ \pi_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1 \leq i \leq n \end{cases} \tag{A3}$$

with

$$\pi_0 = \left(\sum_{i=0}^{s} \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n} \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i \right)^{-1}$$

satisfies the balance equations and the normalization condition.

Proposition 4 *The distribution of the closed two-node Jackson network illustrated in Figure 1a is given by*

$$\hat{\pi}_i = \begin{cases} \hat{\pi}_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \leq i \leq s, \\ \hat{\pi}_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s+1 \leq i \leq n \end{cases} \tag{A4}$$

with

$$\hat{\pi}_0 = \left[\sum_{i=0}^{s} \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^{n} \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i \right]^{-1}.$$

Proof: We have a two-node closed Jackson network, with probability transition matrix

$$P = \begin{pmatrix} 1-p & p \\ 1 & 0 \end{pmatrix}.$$

Let $r_i(m)$ denote the rate of service when there are m patients at queue i, so $r_1(m) = \min\{m, s\}$ and $r_2(m) = m$. The throughput vector $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}_{\mathbb{1}}{}^2$ must satisfy $\gamma = \gamma P$ and we find that $\gamma = (p, 1)$ suffices. From the general theory of Jackson networks, see [59], it follows that the stationary distribution is given by

$$\pi_i = G^{-1} g_1(i) g_2(n-i)$$

with

$$g_1(i) = \frac{(\gamma_1/\mu)^i}{\prod_{m=1}^{i} r_1(m)}, \quad g_2(n-i) = \frac{(\gamma_2/\delta)^{n-i}}{\prod_{m=1}^{n-i} r_2(m)},$$

and normalization constant $G = \sum_{i=0}^{n} g_1(i) g_2(n-i)$. Then,

$$g_1(i) = \begin{cases} \frac{1}{i!\mu^i} & \text{for } 0 \leq i \leq s, \\ \frac{1}{s!s^{i-s}\mu^i} & \text{for } s+1 \leq i \leq n, \end{cases} \qquad g_2(n-i) = \frac{1}{(n-i)!} \left(\frac{p}{\delta}\right)^n \left(\frac{\delta}{p}\right)^i,$$

and rewriting the expressions yields (A4).

A.3 Stationary distribution

Assuming that the stability condition is satisfied, we can determine the unique stationary distribution of the Markov process $\{N(t), Q_1(t)\}$. The vector π_i can be written as $\pi_{n+i} = \pi_n G^i$ for $i = 0, 1, ...$, where G is the minimal nonnegative solution of the non-linear matrix equation

$$A_0 + GA_1 + G^2 A_2 = 0. \tag{A5}$$

The balance equations can be written as

$$\pi_{i-1}A_0 + \pi_i A_1 + \pi_{i+1}A_2 = 0,\ i = n+1, n+2, ...$$

and using $\pi_{n+i} = \pi_n G^{i-n}$ for $i = 0, 1, ...$, we find that

$$\pi_n G^{i-n-1}\left(A_0 + GA_1 + GA_2\right) = 0,\ i = n+1, n+2, ...$$

Moreover, we have the boundary equations

$$\begin{aligned}
\pi_0 B_{00} + \pi_1 B_{10} &= 0\\
\pi_0 B_{01} + \pi_1 B_{11} + \pi_2 B_{21} &= 0\\
\pi_1 B_{12} + \pi_1 B_{22} + \pi_2 B_{32} &= 0\\
&\vdots\\
\pi_{n-2} B_{n-2\,n-1} + \pi_{n-1} B_{n-1\,n-1} + \pi_n B_{n\,n-1} &= 0\\
\pi_{n-1} B_{n-1\,n} + \pi_n B_{nn} + \pi_{n+1} A_2 &= 0,
\end{aligned}$$

along with the normalization equation

$$1 = \sum_{i=0}^{\infty} \pi_i e = \sum_{i=0}^{n-1} \pi_i e + \pi_n (I - G)^{-1} e,$$

where we slightly abuse notation by using e as the all- ones vector of appropriate size. We note that the matrix G has a spectral radius less than one and therefore $(I - G)$ is invertible.

These equations provide the tools for finding the equilibrium probabilities. Although it is hard to solve G analytically from Equation (A5), it is easy to solve numerically by using the following algorithm (matrix-geometric method). Rewriting (A5) gives

$$G = -(A_0 + G^2 A_2) A_1^{-1},$$

where A_1 is invertible, since it is a transient generator matrix. Let

$$G_{k+1} = -(A_0 + G_k^2 A_2) A_1^{-1},$$

starting with $G_k = 0$. We note that $G_k \uparrow G$ as k grows to infinity [51]. Once $||G_{k+1} - G_k||_2$ is below a certain preset threshold, we approximate G by G_{k+1}.

Appendix B Proof of Proposition 2

First, note that by definition of the Erlang-R model with holding, in which no more that n patients can be admitted in the ED simultaneously, that $Q_1^h(t) + Q_2^h(t) \leq n = Q_1^J(t) + Q_2^J(t)$ follows directly. Therefore, we only consider the relation between the states in the blocking and holding variants Erlang-R model.

As noted Section 3.1, the model with holding can be characterized as a three-dimensional Markov chain $X^h(t) = (H(t), Q_1^h(t), Q_2^h(t))$ in which the components denote the number of holding, needy and content patients respectively. The Erlang-R model with blocking similarly admits a Markov process description, but with two dimensions, namely $X^b(t) = (Q_1^b(t), Q_1^b(t))$.

We prove the result by constructing a coupling between the Markov processes X^h and X^b. Let $Z(t) := \left(\hat{X}^h(t), \hat{X}^b(t)\right) = \left(\hat{H}(t), \hat{Q}_1^h(t), \hat{Q}_2^h(t), \hat{Q}_1^b(t), \hat{Q}_2^b(t)\right)$.

We first define the transition rates of this five-dimensional Markov process, which naturally only depend on the current state of the system. After that we show that the transition rates relevant to $\hat{X}^h(t)$ and $X^h(t)$ coincide with those of either $X^h(t)$ or $\hat{X}^b(t)$, respectively. The latter implies that the marginal transitions of $\hat{X}^h(t)$ and $X^b(t)$ (and $\hat{X}^b(t)$ and $X^h(t)$) are equal, and hence so are their probability distribution of the Markov processes.

Let $Z(t) = (h, q_1^h, q_2^h, q_1^b, q_2^b)$. While defining the reachable states from this state and associated transition rates, we distinguish four transition types, and further differentiate the transition rates depending on the current state.

Arrival. Arrivals to occur in both models simultaneously, but are handled differently according to the current queue lengths.

1. If $q_1^h + q_2^h < n$ and $q_1^b + q_2^b < n$,
$$(h, q_1^h + 1, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{B6}$$
2. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b < n$,
$$(h + 1, q_1^h, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{B7}$$
3. if $q_1^h + q_2^h < n$ and $q_1^b + q_2^b = n$,
$$(h, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{B8}$$
4. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b = n$,
$$(h + 1, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{B9}$$

Departure. Basically, we align service completions in the two models, but allow a completion occurring solely in either of one of the two models, only if the queue length in this model is strictly larger than in the other one.

1. If $q_1^h \geq q_1^b$ and $h > 0$

$$\begin{cases} (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (h-1, q_1^h, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1-p)\mu. \end{cases} \tag{B10}$$

2. If $q_1^h < q_1^b$ and $h > 0$

$$\begin{cases} (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \tag{B11}$$

3. If $q_1^h \geq q_1^b$ and $h = 0$

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (0, q_1^h - 1, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1-p)\mu. \end{cases} \tag{B12}$$

4. If $q_1^h < q_1^b$ and $h = 0$

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \tag{B13}$$

Become content The differentiation between transitions is similar to those in the *departure* transition type.

1. If $q_1^h \geq q_1^b$,

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^b \wedge s)p\mu, \\ (h, q_1^h - 1, q_2^h + 1, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)]p\mu. \end{cases} \tag{B14}$$

2. If $q_1^h < q_1^b$,

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b + 1) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)]p\mu. \end{cases} \tag{B15}$$

Become needy

1. If $q_2^h \geq q_2^b$,

$$\begin{cases} (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^b\delta, \\ (h, q_1^h + 1, q_2^h - 1, q_1^b, q_2^b) & \text{with rate } (q_2^h - q_2^b)\delta, \end{cases} \tag{B16}$$

2. If $q_2^h < q_2^b$,

$$\begin{cases} (h, q_1^h+1, q_2^h-1, q_1^b+1, q_2^b-1) & \text{with rate } q_2^h\delta, \\ (h, q_1^h, q_2^h, q_1^b+1, q_2^b-1) & \text{with rate } (q_2^b - q_2^h)\delta, \end{cases} \tag{B17}$$

This set of transitions defines the dynamics of the Markov process $Z(t) = (\hat{X}^h(t), \hat{X}^b(t))$. Let us now restrict our attention to the transitions in which (at least one of the) first three coordinates of $Z(t)$ changes, that is, the marginal transitions of the process $\hat{X}^h$. Let $\hat{X}^h(t) = (h, q_1^h, q_2^h)$, then according to the transition scheme above, $\hat{X}^h$ moves to state

1. If $q_1^h + q_2^h < n$ (and hence necessarily $h = 0$),

$$\begin{cases} (0, q_1^h+1, q_2^h) & \text{with rate } \lambda, \\ (0, q_1^h-1, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h-1, q_2^h+1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (0, q_1^h+1, q_2^h-1) & \text{with rate } q_2^h\delta. \end{cases}$$

2. if $q_1^h + q_2^h = n$ and $h = 0$,

$$\begin{cases} (1, q_1^h, q_2^h) & \text{with rate } \lambda, \\ (0, q_1^h, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h-1, q_2^h+1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (0, q_1^h+1, q_2^h-1) & \text{with rate } q_2^h\delta. \end{cases}$$

3. if $h > 0$ (and hence necessarily $q_1^h + q_2^h = n$),

$$\begin{cases} (h+1, q_1^h, q_2^h) & \text{with rate } \lambda, \\ (h-1, q_1^h, q_2^h) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (h, q_1^h-1, q_2^h+1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (h, q_1^h+1, q_2^h-1) & \text{with rate } q_2^h\delta. \end{cases}$$

One can check that these transitions indeed coincide with the transitions in the original holding model, hence $\hat{X}^h(t) \stackrel{d}{=} X^h(t)$.

Similarly, when the focusing on transitions of $Z(t)$ that are relevant for $\hat{X}^b(t)$, we deduce the following transition scheme. If $\hat{X}^b(t) = (q_1^b, q_2^b)$, then the next move according to the transitions of $Z(t)$ is

$$\begin{cases} (q_1^b + 1_{\{q_1^b+q_2^b<n\}}, q_2^b) & \text{with rate } \lambda, \\ (q_1^b-1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (q_1^b-1, q_2^b+1) & \text{with rate } (q_1^b \wedge s)p\mu, \\ (q_1^b+1, q_2^b-1) & \text{with rate } q_2^b\delta. \end{cases}$$

These transition rates clearly coincide with the original Erlang-R model with blocking, and also hence $\hat{X}^b(t) \stackrel{d}{=} X^h(t)$.

Next, we show that under this coupling scheme we have that if $\hat{H}(0) = 0$, $\hat{Q}_1^h(0) = \hat{Q}_1^b(0)$ and $\hat{Q}_1^h(0) = \hat{Q}^b(0)$ then for all $t \geq 0$, $Z(t)$ satisfies the hypothesis:

(i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) \leq \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$,
(ii) $\hat{Q}_2^b(t) \leq \hat{Q}_2^h(t)$,
(iii) $\hat{Q}_1^b(t) \leq \hat{Q}_1^h(t) + H(t)$.

We do so by induction on the next state reached after a transition of the joint Markov process $Z = (\hat{X}^h, \hat{X}^b)$. First of all, $Z(0)$ clearly satisfies (i)-(iii). Next, assume $Z(t^-) = (h, q_1^h, q_2^h, q_1^b, q_2^b)$ satisfies the hypothesis and a transition occurs at t. We show that under the specified coupling scheme, the state reached after the next transition, $Z(t)$ must satisfy (i)-(iii) as well. To do so, we differentiate between the four types of transitions that could occur: arrival, departure, become content and become needy.
Arrival.
Recall that under our coupling scheme an arrival always occurs in both the holding and blocking model simultaneously, see (B6)–(B9). Furthermore, q_2^h and q_2^b are unchanged during this transition, rendering (ii) trivial.

By hypothesis $q_1^b + q_2^b \leq q_1^h + q_2^h$, hence the event $q_1^h + q_2^h < n$ and $q_1^h + q_2^b = n$, with resulting state $(0, q_1^h + 1, q_2^h, q_1^b, q_2^b)$, can be excluded from our analysis We check the conditions for the remaining three cases.

1. If $Z(t) = (0, q_1^h + 1, q_2^h, q_1^b + 1, q_2^b)$, then $q_1^b + q_2^b < n$ and $q_1^h + q_2^h < n$.
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \stackrel{(i)}{\leq} q_1^h + q_2^h + 1 = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \stackrel{(iii)}{\leq} q_1^h + 1 = \hat{Q}_1^h(t) = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h + 1, q_1^h, q_2^h, q_1^b + 1, q_2^b)$, then $q_1^b + q_2^b < n$ and $q_1^h + q_2^h = n$.
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \leq n = q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \stackrel{(iii)}{\leq} q_1^h + 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
3. If $Z(t) = (h + 1, q_1^h, q_2^h, q_1^b, q_2^b)$, then $q_1^b + q_2^b = q_1^h + q_2^h = n$.
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \stackrel{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b \stackrel{(iii)}{\leq} q_1^h + h < q_1^h + h + 1 = \hat{H}(t)$.

Departure. By carefully examining the possible state transitions of $Z(t)$ following a departure, we list six reachable states. However, by (iii), we have that if $h = 0$, then $q_1^b \leq q_1^h$, which excludes the state $(0, q_1^h, q_2^h, q_1^b, q_2^b)$ in (B13) from the reachability graph. We check the remaining states for conditions (i)–(iii). Again, during a departure, q_2^b and q_2^h are unchanged, so (ii) is automatically satisfied by the induction hypothesis.

1. If $Z(t) = (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b)$, then $h > 0$.
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 \overset{(i)}{\leq} q_1^h + q_2^h - 1 < q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h-1, q_1^h, q_2^h, q_1^b, q_2^b)$, then $h > 0$ and $q_1^h \geq q_1^b$ (*).
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(*)}{\leq} q_1^h - 1 \leq q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
3. If $Z(t) = (h, q_1^h, q_2^h, q_1^b - 1, q_2^b)$, then $h > 0$ and $q_1^h < q_1^b$ (*).
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 < q_1^b + q_2^b \overset{(i)}{\leq} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 < q_1^b \overset{(*)}{\leq} q_1^h + h = \hat{Q}_1^h(t) + \hat{H}(t)$.
4. If $Z(t) = (h, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b)$, then $h = 0$.
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = (q_1^b - 1) + q_2^b - 1 < \overset{(i)}{\leq} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
5. If $Z(t) = (0, q_1^h - 1, q_2^h, q_1^b, q_2^b)$, then $h = 0$ and $q_1^h > q_1^b$ (*).
   (i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\leq} (q_1^h - 1) + q_2^b \overset{(ii)}{\leq} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = qy \overset{(*)}{\leq} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

Content start. On the event of a patient becoming content, it is clear that the sums $\hat{Q}_1^h(t) + \hat{Q}_2^h(t)$ and $\hat{Q}_1^b(t) + \hat{Q}_2^b(t)$ and $H(t)$ are unaffected. This means that (i) is directly satisfied by the induction hypothesis. According to (B14)–(B15), three states can be reached.

1. If $Z(t) = (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1)$,
   (ii) $\hat{Q}_2^b(t) = q_2^b + 1 \overset{(ii)}{\leq} q_2^h + 1 = \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\leq} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h, q_1^h - 1, q_2^h + 1, q_1^b, q_2^b)$, then $q_1^h > q_1^b$ (*),
   (ii) $\hat{Q}_2^b(t) = q_2^b \overset{(ii)}{\leq} q_2^h < q_2^h + 1 = \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\leq} q_1^h + h < q_1^h + 1 + h = \hat{Q}_1^h(t) + \hat{H}(t)$.
3. If $Z(t) = (h, q_1^h, q_2^h, q_1^b - 1, q_2^b + 1)$, then $q_1^b > q_1^h$ and hence by (iii) $h > 0$. The latter is only possible if $q_1^h + q_2^h = n$ (*),
   (ii) $\hat{Q}_2^b(t) = q_2^b + 1 \leq n - q_1^b + 1 = (q_1^h + q_2^h) - q_1^b + 1 \overset{(*)}{\leq} q_2^h = \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b - 1 < q_1^h + h - 1 \overset{(*)}{\leq} q_1^h + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

Become needy.
Just as in the event of content start, the sums $\hat{Q}_1^h(t)+\hat{Q}_2^h(t)$ and $\hat{Q}_1^b(t)+\hat{Q}_2^b(t)$ and $H(t)$ are unaffected, whereby (i) is directly satisfied by the induction hypothesis. By (ii), we have $q_2^h \geq q_2^b$. This excludes the state $(h, q_1^h, q_2^h, q_1^b+1, q_2^b-1)$ from being reached, see (B17). We check the remaining to possibilities.

1. If $Z(t) = (h, q_1^h+1, q_2^h-1, q_1^b+1, q_2^b-1)$.
   (ii) $\hat{Q}_2^b(t) = q_2^b - 1 \overset{(ii)}{\leq} q_2^h - 1 = \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\leq} q_1^h + h + 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.
2. If $Z(t) = (h, q_1^h+1, q_2^h-1, q_1^b, q_2^b)$, then $q_2^h > q_2^b$ (*).
   (ii) $\hat{Q}_2^b(t) = q_2^b \overset{(*)}{\leq} q_2^h - 1 = \hat{Q}_2^h(t)$.
   (iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\leq} q_1^h + h < q_1^h + 1 + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

Hence, the state reached after any feasible transition under the coupling scheme satisfies the conditions (i)–(iii). Thus we conclude that the joint process $(\hat{H}(t), \hat{Q}_1^h(t), \hat{Q}_2^h(t), \hat{Q}_1^b(t), \hat{Q}_2^b(t))$ adheres to (i)–(iii) for all t. Consequently, we have that (i) implies

$$
\begin{aligned}
\mathbb{P}\left(Q_1^b(t)+Q_2^b(t) \geq k\right) &= \mathbb{P}\left(Q_1^b(t)+Q_2^b(t) \geq k\right) \\
&= \sum_{j=0}^{n} \mathbb{P}\left(\hat{Q}_1^b(t)+\hat{Q}_2^b(t) \geq k, \hat{Q}_1^h(t)+\hat{Q}_2^h(t) = j\right) \\
&= \sum_{j=k}^{n} \mathbb{P}\left(\hat{Q}_1^b(t)+\hat{Q}_2^b(t) \geq k, \hat{Q}_1^h(t)+\hat{Q}_2^h(t) = j\right) \\
&\leq \sum_{j=h}^{n} \mathbb{P}\left(\hat{Q}_1^h(t)+\hat{Q}_2^h(t) = j\right) \\
&= \mathbb{P}\left(Q_1^h(t)+Q_2^h(t) \geq k\right) = \mathbb{P}\left(Q_1^h(t)+Q_2^h(t) \geq k\right).
\end{aligned}
$$

The other two orderings follow similarly.

Remark 4 Note that under this coupling scheme we cannot get the ordering $\hat{Q}_1^h(t)(t) \geq \hat{Q}_1^b(t)(t)$ for all $t \geq 0$. A minimal counter example occurs for $s = n = 1$. Let $Z(0) = ((0,0,0),(0,0))$. First, two arrivals occur, such that state $((1,1,0),(1,0))$ is reached, followed by a departure transition, yielding $((0,1,0),(0,0))$. Next, the one patient left in the model with holding system becomes content, so that we obtain $((0,0,1),(0,0))$. At this stage, if an arrival occurs, the arriving patient will be put in the holding queue in the model with holding, and admitted to nurse queue in the model with blocking. Hence we end up in state $((1,0,1),(1,0))$, in which $\hat{Q}_1^h(t) < \hat{Q}_1^b(t)$.

Appendix C Proof of Theorem 1 — Performance measures of Erlang-R with blocking

In this appendix we prove the heavy-traffic approximations of the system-measures introduced in Theorem 1. As a first stage we present and prove four lemmas.

Lemma 1 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta \neq 0$. Define B_1 as the expression*

$$B_1 = \frac{e^{-R}}{s!} R_1^s \frac{1}{1-\rho} \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2}.$$

Then

$$\lim_{\lambda\to\infty} B_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}.$$

Lemma 2 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta \neq 0$. Define B_2 as the expression*

$$B_2 = \frac{e^{-(R_1+R_2)}}{s!} R_1^s \frac{\rho^{n-s}}{1-\rho} \sum_{l=0}^{n-s-1} \frac{1}{l!} \left(\frac{R_2}{\rho}\right)^l.$$

Then

$$\lim_{\lambda\to\infty} B_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta} e^{\frac{1}{2}\omega^2} \Phi(\omega).$$

Lemma 3 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5). Define A as the expression*

$$A = \sum_{\substack{i,j|i\le s,\\ i+j\le n-1}} \frac{1}{i!j!} R_1^i R_2^j e^{-(R_1+R_2)}. \tag{C18}$$

Then

$$\lim_{\lambda\to\infty} A = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta}{\mu p}}\right) d\Phi(t).$$

Lemma 4 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta = 0$. Define B as the expression*

$$B = e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j \sum_{i=0}^{n-s-j-1} \rho^i.$$

Then

$$\lim_{\lambda\to\infty} B = \sqrt{\frac{\mu p}{\delta}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta)\right).$$

C.1 Proof of Lemma 1

By using Stirling's formula $\left(s! \approx \sqrt{2\pi s}\left(\frac{s}{e}\right)^s\right)$, and the QED assumption that $\sqrt{s}(1 - R_1/s) \to \beta$ as $\lambda \to \infty$, one obtains for B_1:

$$B_1 \approx \frac{e^{s-R}}{\sqrt{2\pi s}}\rho^s \frac{\sqrt{s}}{\beta} \sum_{l=0}^{n-s-1} \frac{1}{l!}\left(R_2\right)^l e^{-R_2} = \frac{e^{s-R}\rho^s}{\sqrt{2\pi}\beta} \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2}$$
$$= \frac{e^{s(1-\rho)}}{\sqrt{2\pi}\beta}\rho^s P(X_\lambda \leq n-s-1)$$

where $\rho = \frac{\lambda}{(1-p)s\mu}$, and X_λ is a random variable with the Poisson distribution with parameter R_2. When $\lambda \to \infty$, $R_2 \to \infty$ too, since p, and δ are fixed. Note that

$$\mathbb{P}(X_\lambda \leq n-s-1) = \mathbb{P}\left(\frac{X_\lambda - R_2}{\sqrt{R_2}} \leq \frac{n-s-1-R_2}{\sqrt{R_2}}\right).$$

Now we need to find the limit for the following fraction $\frac{n-s-R_2}{\sqrt{R_2}}$ as $\lambda \to \infty$ using assumption that $\sqrt{s}(1 - R_1/s) \to \beta$ as $\lambda \to \infty$.

$$\lim_{\lambda\to\infty} \frac{n-s-R_2}{\sqrt{R_2}} = \lim_{\lambda\to\infty} \frac{n - \frac{R_1}{r} - s + R_1}{\sqrt{\frac{R_1}{r} - R_1}} = \frac{\gamma\sqrt{\frac{R_1}{r}} - \beta\sqrt{R_1}}{\sqrt{\frac{R_1}{r} - R_1}} = \frac{\gamma - \beta\sqrt{r}}{\sqrt{1-r}}. \tag{C19}$$

Hence, define $\eta = \frac{\gamma-\beta\sqrt{r}}{\sqrt{1-r}}$. Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\left(\frac{X_\lambda - R_2}{\sqrt{R_2}}\right) \Rightarrow N(0,1)$$

and due to assumption QED (i) of the lemma we get

$$\mathbb{P}(X_\lambda \leq n-s-1) \to \mathbb{P}(N(0,1) \leq \eta) = \Phi(\eta), \;\; as\; \lambda \to \infty \tag{C20}$$

where $N(0,1)$ is a standard normal random variable with distribution function $\Phi(\cdot)$. It follows thus that

$$B_1 \approx \frac{e^{s(1-\rho)}}{\sqrt{2\pi}\beta}\rho^s\Phi(\eta) = \frac{e^{s(1-\rho+\ln\rho)}}{\sqrt{2\pi}\beta}\Phi(\eta).$$

Making use of the expansion $\ln \rho = \ln(1-(1-\rho)) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2$, $(\rho \to 1)$ one obtains

$$B_1 \approx \frac{e^{s(1-\rho-(1-\rho)-\frac{(1-\rho)^2}{2})}}{\sqrt{2\pi}\beta}\Phi(\eta) = \frac{e^{-\frac{s(1-\rho)^2}{2}}}{\sqrt{2\pi}\beta}\Phi(\eta)$$

by QED assumption that $\sqrt{s}(1-R_1/s) \to \beta$, it is clear that $s(1-\rho)^2 \to \beta^2$, when $\lambda \to \infty$. This implies

$$\lim_{\lambda\to\infty} B_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta}$$

where $\phi(\cdot)$ is the standard normal density function, and $\Phi(\cdot)$ is the standard normal distribution function. This proves Lemma 1.

C.2 Proof of Lemma 2

Again according to Stirling's formula, and the QED assumption that $\frac{n-s-R_1-R_2}{\sqrt{R_1+R_2}} \to \eta$ as $\lambda \to \infty$, one obtains for B_2:

$$\begin{aligned} B_2 &\approx \frac{e^{s-R_1-R_2}}{\sqrt{2\pi s}}\frac{\rho^n}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^l = \frac{e^{s(1-\rho)-R_2}}{\sqrt{2\pi s}}\frac{\sqrt{s}\rho^n}{\beta}e^{\frac{R_2}{\rho}}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^l e^{-\frac{R_2}{\rho}} \\ &= \frac{e^{s(1-\rho)+R_2\left(\frac{1-\rho}{\rho}\right)}}{\sqrt{2\pi}\beta}\rho^n\mathbb{P}(Y_\lambda \le n-s-1), \end{aligned}$$

where $\rho = \frac{\lambda}{(1-p)s\mu}$, and Y_λ is a random variable with the Poisson distribution with parameter $\frac{R_2}{\rho}$. Note that

$$\mathbb{P}(Y_\lambda \le n-s-1) = \mathbb{P}\left(\frac{Y_\lambda - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}} \le \frac{n-s-1-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right).$$

Now we need to find the limit for the following fraction $\frac{n-s-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}$ as $\lambda \to \infty$ using assumption QED (i).

$$\begin{aligned} &\lim_{\lambda\to\infty}\frac{n-s-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}} = \lim_{\lambda\to\infty}\frac{\eta\sqrt{R_2}+R_2-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}} = \lim_{\lambda\to\infty}\eta\sqrt{\rho}+\frac{\sqrt{R_2}(\rho-1)}{\sqrt{\rho}} \\ &= \eta - \lim_{\lambda\to\infty}\sqrt{\frac{sp\mu}{\delta}}(1-\rho) = \eta - \sqrt{\frac{p\mu}{\delta}}\beta. \end{aligned} \tag{C21}$$

Denote $\omega = \eta - \beta\sqrt{\frac{p\mu}{\delta}}$. Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\left(\frac{Y_\lambda - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right) \Rightarrow N(0,1)$$

and

$$\mathbb{P}(Y_\lambda \le n - s - 1) \to \mathbb{P}(N(0,1) \le \omega) = \Phi(\omega), \text{ as } \lambda \to \infty,$$

where $N(0,1)$ is a standard normal random variable with distribution function Φ. It follows thus that

$$B_2 \approx \frac{e^{s(1-\rho)+R_2\left(\frac{1-\rho}{\rho}\right)}}{\sqrt{2\pi}\beta}\rho^n\Phi(\omega) = \frac{e^{s(1-\rho)+R_2\left(\frac{1-\rho}{\rho}\right)+n\ln\rho}}{\sqrt{2\pi}\beta}\Phi(\omega).$$

Making use of the expansion

$$\ln\rho = \ln(1-(1-\rho)) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2, \;\; (\rho \to 1)$$

and using our assumptions that as $\lambda \to \infty$: $\rho \to 1$, $s \approx R_1 + \beta\sqrt{R_1}$, $n \approx \frac{R_1}{r} + \gamma\sqrt{\frac{R_1}{r}}$, and $n - s \approx R_2 + \eta\sqrt{R_2}$, one obtains

$$\begin{aligned}
& s(1-\rho) + R_2\left(\frac{1-\rho}{\rho}\right) + n\ln\rho = s(1-\rho) + R_2\left(\frac{1-\rho}{\rho}\right) - n\left(1-\rho+\frac{(1-\rho)^2}{2}\right) \\
&= -\left(n - s - \frac{R_2}{\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2} \approx -\left(\eta\sqrt{R_2} + R_2 - \frac{R_2}{\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2} \\
&= \left(\frac{R_2}{\rho} - \frac{n}{2}\right)(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho) \\
&\approx \left(\frac{R_2}{\rho} - \frac{1}{2}\left(\frac{R_1}{r} + \gamma\sqrt{\frac{R_1}{r}}\right)\right)(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho) \\
&= \left(\frac{R_2}{\rho} - \frac{R_1}{2r}\right)(1-\rho)^2 - \frac{1}{2}\gamma\sqrt{\frac{R_1}{r}}(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho) \\
&= \left(\frac{R_2}{\rho} - \frac{R_1}{2r}\right)\frac{\beta^2}{R_!} - \frac{1}{2}\gamma\sqrt{\frac{R_1}{r}}\frac{\beta^2}{R_1} - \eta\sqrt{R_2}\beta\sqrt{\frac{1}{R_1}} \\
&\approx \frac{p\mu}{\delta\rho}\beta^2 - \frac{1}{2}\left(\frac{p\mu}{\delta} + 1\right)\beta^2 - \eta\beta\sqrt{\frac{p\mu}{\delta}} = -\frac{1}{2}(\eta^2 + \beta^2) + \frac{1}{2}\omega^2.
\end{aligned}$$

Therefore,

$$\lim_{\lambda\to\infty} B_2 \approx \frac{e^{s(1-\rho)+R_2\left(\frac{1-\rho}{\rho}\right)+n\ln\rho}}{\sqrt{2\pi}\beta}\Phi(\omega) \approx \frac{e^{-\frac{1}{2}(\eta^2+\beta^2)+\frac{1}{2}\omega^2}}{\sqrt{2\pi}\beta}\Phi(\omega)$$
$$= \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega).$$

This proves Lemma 2.

C.3 Proof of Lemma 3

We will find the asymptotic behavior of A by finding its lower and upper bounds. Let us consider a partition $\{s_h\}_{h=0}^{l}$ of the interval $[0,s]$.

$$s_h = s - h\tau, \;\; h = 0,1,...,\ell; \;\; s_{\ell+1} = 0$$

where $\tau = \left[\epsilon\sqrt{R_1}\right]$, ϵ is an arbitrary non-negative real and ℓ is a positive integer.

If λ and s tend to infinity and satisfy the QED assumption that $\frac{n-s-R_1-R_2}{\sqrt{R_1+R_2}} \to \eta$ as $\lambda \to \infty$, then $\ell < \frac{s}{\tau}$ for λ big enough and all the s_h belong to $[0,s]$; $h = 0,1,...,\ell$. Emphasize that the length τ of every interval $[s_{h-1}, s_h]$ depends on λ. The variable A is given by the formula (C18). Let us consider a lower estimate for A given by the following sum:

$$\begin{aligned} A \geq A_1 &= \sum_{h=0}^{\ell}\sum_{i=s_{h+1}}^{s_h}\frac{1}{i!}R_1^i e^{-R_1}\cdot\sum_{j=0}^{n-s_h-1}\frac{1}{j!}R_2^j e^{-R_2} \\ &= \sum_{h=0}^{\ell}\sum_{i=s_{h+1}}^{s_h}\frac{1}{i!}R_1^i e^{-R}\mathbb{P}(Y_n \leq n-s_h-1) \qquad \text{(C22)} \\ &= \sum_{h=0}^{\ell}\mathbb{P}(s_{h+1}\leq X_n \leq s_h)\mathbb{P}(Y_n \leq n-s_h-1) \end{aligned}$$

where X_n and Y_n are independent Poisson random variables with parameters R_1 and R_2, respectively.

If $\lambda \to \infty$ then $R_1 \to \infty$, since p and μ are fixed. Note that

$$\mathbb{P}(s_{h+1}\leq X_n \leq s_h) = \mathbb{P}\left(\frac{s_{h+1}-R_1}{\sqrt{R_1}} \leq \frac{X_n-R_1}{\sqrt{R_1}} \leq \frac{s_h-R_1}{\sqrt{R_1}}\right).$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{X_n-R}{\sqrt{R_1}} \Rightarrow N(0,1).$$

Since

$$\lim_{\lambda\to\infty}\frac{s_h-R_1}{\sqrt{R_1}}=\lim_{\lambda\to\infty}\frac{s-h\epsilon\sqrt{R_1}-R_1}{\sqrt{R_1}}=\lim_{\lambda\to\infty}\frac{R_1-\beta\sqrt{R_1}-h\epsilon\sqrt{R_1}-R_1}{\sqrt{R_1}}=\beta-h\epsilon$$

we obtain:

$$\begin{aligned}&\mathbb{P}(s_{h+1}\le X_n\le s_h)=\Phi(\beta-h\epsilon)-\Phi(\beta-(h+1)\epsilon),\ h=0,..,\ell-1\\&\mathbb{P}(0\le X_n\le s_\ell)=\Phi(\beta-\ell\epsilon).\end{aligned}\tag{C23}$$

Similarly, if $\lambda\to\infty$ then $R_2\to\infty$, since p,δ and γ are fixed. Note that

$$\mathbb{P}(Y_n\le n-s_h)=\mathbb{P}\left(\frac{Y_n-R_2}{\sqrt{R_2}}\le\frac{n-s_h-R_2}{\sqrt{R_2}}\right).$$

Thus, when $\lambda\to\infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{Y_n-R_2}{\sqrt{R_2}}\Rightarrow N(0,1).$$

Since

$$\begin{aligned}\lim_{\lambda\to\infty}\frac{n-s_h-R_2}{\sqrt{R_2}}&=\lim_{\lambda\to\infty}\frac{n-s-h\epsilon\sqrt{R_1}-R_2}{\sqrt{R_2}}=\lim_{\lambda\to\infty}\frac{R_2+\eta\sqrt{R_2}-h\epsilon\sqrt{R_1}-R_2}{\sqrt{R_2}}\\&=\eta-h\epsilon\frac{\sqrt{R_1}}{\sqrt{R_2}}=\eta-h\epsilon\sqrt{\frac{\delta}{p\mu}},\end{aligned}$$

we obtain:

$$\mathbb{P}(Y_n\le n-s_h)=\Phi\left(\eta-h\epsilon\sqrt{\frac{\delta}{p\mu}}\right),\ h=0,..,\ell\tag{C24}$$

It follows from (C22), (C23), and (C24) that

$$\lim_{\lambda\to\infty}A\ge\sum_{h=0}^{\ell-1}(\Phi(\beta-h\epsilon)-\Phi(\beta-(h+1)\epsilon))\Phi\left(\eta-h\epsilon\sqrt{\frac{\delta}{p\mu}}\right)+\Phi(\beta-\ell\epsilon)\Phi\left(\eta-\ell\epsilon\sqrt{\frac{\delta}{p\mu}}\right)\tag{C25}$$

which is the lower Riemann-Stieltjes sum of the integral

$$-\int_0^\infty\Phi\left(\eta+x\sqrt{\frac{\delta}{p\mu}}\right)d\Phi(\beta-x)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta}{p\mu}}\right)d\Phi(t)\tag{C26}$$

corresponding to the partition $\{\beta - h\epsilon\}_{h=0}^{\ell}$ of the semi axis $(-\infty, \beta)$. Similarly, let us take the upper estimate for A as the following sum:

$$\begin{aligned} A \leq A_2 &= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} R_1^i e^{-R_1} \sum_{j=0}^{n-s_{h+1}-1} \frac{1}{j!} R_2^j e^{-R_2} \\ &= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} R_1^i e^{-R} \mathbb{P}(Y_n \leq n - s_{h+1} - 1) \\ &= \sum_{h=0}^{\ell} \mathbb{P}(s_{h+1} \leq X_n \leq s_h) \mathbb{P}(Y_n \leq n - s_{h+1} - 1) \end{aligned}$$

where X_n and Y_n are the same random variable as before. Using the same calculation that were computed for the upper boundary we obtain

$$\lim_{\lambda \to \infty} A \leq \sum_{h=0}^{\ell-1} \left(\Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon)\right) \Phi\left(\eta - (h+1)\epsilon\sqrt{\frac{\delta}{p\mu}}\right) + \Phi(\beta - \ell\epsilon) \tag{C27}$$

which is the upper Riemann-Stieltjes sum for the integral (C26). When $\epsilon \to 0$ the boundaries (C25) and (C27) lead to the following equality

$$\lim_{\lambda \to \infty} A = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta}{p\mu}}\right) d\Phi(t) = \int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)$$

This proves Lemma 3.

C.4 Proof of Lemma 4

First, we will rewrite Equation (C18):

$$B = e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j \sum_{i=0}^{n-s-j-1} \rho^i = e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j \frac{1-\rho^{n-s-j}}{1-\rho}.$$

When $\beta = 0$, as $\lambda \to \infty$, by the QED assumption that $\sqrt{s}(1-\rho) \to \beta$, $\rho = 1$. Therefore,

$$\sum_{i=0}^{n-s-j-1} \rho^i = n - s - j.$$

When $\beta \to 0$, $\rho \to 1$ but still $\rho \neq 1$, the expression $\frac{1-\rho^{n-s-j}}{1-\rho}$ can be approximated by $\frac{1-\rho^i}{1-\rho} \approx i$. Thus,

$$\lim_{\rho \to 1} \sum_{i=0}^{n-s-j-1} \rho^i = n - s - j,$$

which is the same phrase as when $\rho = 1$. Hence,

$$\begin{aligned}
B &\approx e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j (n-s-j) \\
&= e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \left((n-s) \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j - \sum_{j=0}^{n-s-1} \frac{1}{j!} {R_2}^j - \sum_{j=0}^{n-s-1} \frac{j}{j!} {R_2}^j \right) \\
&= e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \left((n-s) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l - \sum_{j=0}^{n-s-2} \frac{1}{j!} {R_2}^j - R_2 \sum_{j=0}^{n-s-2} \frac{1}{j!} {R_2}^j \right) \\
&= e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \left((n-s) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l - \sum_{l=0}^{n-s-2} \frac{1}{l!} R_2^l - R_2 \sum_{l=0}^{n-s-2} \frac{1}{l!} R_2^l \right) \\
&= e^{-(R_1+R_2)} \frac{1}{s!} {R_1}^s \left((n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l + \frac{R_2^{n-s}}{(n-s-1)!} \right) \\
&\approx e^{s-R} \frac{1}{\sqrt{2\pi s}} \rho^s \left((n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2} + \frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} \right) \\
&= \frac{1}{\sqrt{2\pi s}} \left((n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2} + \frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} \right).
\end{aligned}$$

As seen in Equation (C20)

$$(n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2} \approx \eta \sqrt{R_2} \Phi(\eta).$$

By using Stirling's formula:

$$\begin{aligned}
&\frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} = \frac{(n-s) R_2^{n-s} e^{-R_2}}{(n-s)!} \approx \frac{(n-s) e^{n-s-R_2}}{\sqrt{2\pi (n-s)}} \left(\frac{R_2}{n-s} \right)^{n-s} \\
&= \frac{(n-s) e^{n-s-R_2+(n-s)\ln\left(\frac{R_2}{n-s}\right)}}{\sqrt{2\pi(n-s)}} = \sqrt{\frac{n-s}{2\pi}} e^{(n-s)\left(1-\frac{R_2}{n-s}+\ln\left(\frac{R_2}{n-s}\right)\right)}.
\end{aligned}$$

By assuming that $\frac{n-s-R_1-R_2}{\sqrt{R_1+R_2}} \to \eta$ when $\lambda \to \infty$

$$\begin{aligned}
&(n-s)\left(1-\frac{R_2}{n-s}+\ln\left(\frac{R_2}{n-s}\right)\right)\\
&=(n-s)\left(1-\frac{R_2}{n-s}-\left(1-\frac{R_2}{n-s}\right)-\frac{1}{2}\left(1-\frac{R_2}{n-s}\right)^2\right)\\
&=-\frac{n-s}{2}\left(1-\frac{R_2}{n-s}\right)^2=-\frac{1}{2}\frac{(n-s-R_2)^2}{n-s}\approx-\frac{1}{2}\frac{\left(\eta\sqrt{R_2}\right)^2}{\eta\sqrt{R_2}+R_2}\\
&\approx-\frac{1}{2}\frac{\left(\eta\sqrt{R_2}\right)^2}{\eta\sqrt{R_2}+R_2}\approx-\frac{\eta^2}{2}.
\end{aligned}$$

Therefore, by the QED assumption that $\frac{n-s-R_1-R_2}{\sqrt{R_1+R_2}} \to \eta$.

$$\begin{aligned}
\sqrt{\frac{n-s}{2\pi}}e^{(n-s)\left(1-\frac{R_2}{n-s}+\ln\left(\frac{R_2}{n-s}\right)\right)} &\approx \sqrt{\frac{\eta\sqrt{R_2}+R_2}{2\pi}}e^{-\frac{\eta^2}{2}}\\
&=\sqrt{\eta\sqrt{R_2}+R_2}\phi(\eta)\approx\sqrt{R_2}\phi(\eta).
\end{aligned}$$

Combining the above approximations and the assumption that $\beta = 0$ and therefore $s = R_1$ yields

$$\begin{aligned}
B &\approx \frac{1}{\sqrt{2\pi s}}\left((n-s-R_2)\sum_{l=0}^{n-s-1}\frac{1}{l!}R_2^l e^{-R_2}+\frac{R_2^{n-s}e^{-R_2}}{(n-s-1)!}\right)\\
&\approx \frac{1}{\sqrt{2\pi s}}\left(\eta\sqrt{R_2}\Phi(\eta)+\sqrt{R_2}\phi(\eta)\right)=\frac{\sqrt{R_2}}{\sqrt{2\pi s}}\left(\eta\Phi(\eta)+\phi(\eta)\right)\\
&=\frac{\sqrt{R_2}}{\sqrt{2\pi R}}\left(\eta\Phi(\eta)+\phi(\eta)\right)=\sqrt{\frac{p\mu}{\delta}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right)\\
&=\sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right).
\end{aligned}$$

This proves Lemma 4.

C.5 Approximation of the probability of delay

Proof: By the Arrival theorem for closed networks,

$$\begin{aligned}
\mathbb{P}_n(W>0)=\mathbb{P}_{n-1}(Q_1(\infty)\geq s)&=\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}^b(i,m-i)\\
&=\pi_0^{n-1}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\frac{1}{s!s^{i-s}}R_1^i\frac{1}{(m-i)!}R_2^{m-i},
\end{aligned}$$

where

$$\pi_0^{n-1} = \left(\sum_{l=0}^{n-1} \frac{1}{l!} (R_1+R_2)^l + \sum_{m=s}^{n-1} \sum_{i=s}^{m} \left(\frac{1}{s!s^{i-s}} - \frac{1}{i!} \right) \frac{1}{(m-i)!} R_1^i R_2^{m-i} \right)^{-1}.$$

Thus,

$$\mathbb{P}_n(W>0) = \left(1+\frac{A}{B}\right)^{-1},$$

where

$$\begin{aligned} A &= \sum_{l=0}^{n-1} \frac{1}{l!} (R_1+R_2)^l e^{-(R_1+R_2)} - \sum_{m=s}^{n-1} \sum_{i=s}^{m} \frac{1}{i!(m-i)!} R_1^i R_2^{m-i} e^{-(R_1+R_2)} \\ &= \sum_{\substack{i,j|i\le s,\\ i+j\le n-1}} \frac{1}{i!j!} R_1^i R_2^j e^{-(R_1+R_2)}, \end{aligned}$$

$$\begin{aligned} B &= \sum_{m=s}^{n-1} \sum_{i=s}^{m} \frac{1}{s!s^{i-s}} R_1^i \frac{1}{(m-i)!} R_2^{m-i} e^{-(R_1+R_2)} \\ &= \sum_{j=0}^{n-s-1} \sum_{i=s}^{n-j-1} \frac{1}{s!s^{i-s}} R_1^i \frac{1}{j!} R_2^j e^{-(R_1+R_2)} \\ &= \sum_{j=0}^{n-s-1} \sum_{i=0}^{n-s-j-1} \frac{1}{s!s^i} \frac{1}{j!} R_1^{i+s} R_2^j e^{-(R_1+R_2)} = \frac{1}{s!} R_1^s e^{-(R_1+R_2)} \sum_{j=0}^{n-s-1} \frac{1}{j!} R_2^j \sum_{i=0}^{n-s-j-1} \rho^i, \end{aligned} \tag{C28}$$

and $\rho = \frac{\lambda}{(1-p)s\mu} = \frac{R_1}{s}$.

Then under the QED assumption that $\sqrt{s}(1-\rho) \to \beta,\ -\infty < \beta < \infty$, if $\beta \neq 0$ as $\lambda \to \infty$, we can rewrite the right-hand side in the following way:

$$\begin{aligned} B &= \frac{1}{s!} R_1^s e^{-(R_1+R_2)} \sum_{j=0}^{n-s-1} \frac{1}{j!} R_2^j \frac{1-\rho^{n-s-j}}{1-\rho} = \frac{1}{s!} R_1^s e^{-(R_1+R_2)} \frac{1}{1-\rho} \sum_{j=0}^{n-s-1} \frac{1}{j!} R_2^j \\ &- \frac{1}{s!} R_1^s e^{-(R_1+R_2)} \frac{\rho^{n-s}}{1-\rho} \sum_{j=0}^{n-s-1} \frac{1}{j!} \left(\frac{R_2}{\rho}\right)^j = B_1 - B_2. \end{aligned} \tag{C29}$$

Applying Lemmas 1,2, and 3 if $\beta \neq 0$ we get

$$\lim_{\lambda\to\infty} \mathbb{P}(W>0) = \left(1 + \frac{\beta \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta-t)\sqrt{\frac{\delta}{p\mu}}\right) d\Phi(t)}{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\eta^2+\beta^2}) e^{\frac{1}{2}\omega^2} \Phi(\omega)}\right)^{-1}.$$

Applying Lemma 3 and 4 when $\beta = 0$ we get

$$\lim_{\lambda\to\infty} \mathbb{P}(W > 0) = \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\sqrt{\frac{1-r}{r}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta)\right)}\right)^{-1}.$$

C.6 Approximation of the expected waiting time

In this appendix we will prove the approximation for the expected waiting time, stated in Section 4.2. The exact measure was defined in Section 3.2. For convenience we state the theorem here again. The first theorem gives the approximation for the case where $\beta \neq 0$.

Theorem 2 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta \neq 0$. Then*

$$\lim_{\lambda\to\infty} \sqrt{s}\mathbb{E}[W] = \frac{1}{\mu} \frac{\left(\frac{1}{\xi} - \frac{\eta}{\beta\sqrt{\xi}} - \frac{1}{\beta^2}\right) \phi(\sqrt{\beta^2+\eta^2})\, \mathrm{e}^{\frac{1}{2}\omega^2}\, \Phi(\omega) + \frac{\phi(\beta)\Phi(\eta)}{\beta^2} - \frac{\phi(\beta)\phi(\eta)}{\sqrt{\xi}\beta}}{\int_{-\infty}^{\beta} \Phi(\eta + (\beta - t)\sqrt{\xi})\, d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}\, \mathrm{e}^{\frac{1}{2}\omega^2}\, \Phi(\omega)},$$

where $\xi = \frac{R_1}{R_2} = \frac{\delta}{p\mu}$, $\omega = \eta - \beta\sqrt{\xi^{-1}}$.

Proof Proof: It follows from (4) that the expectation of the waiting time is given by

$$\begin{aligned}
\mathbb{E}[W] &= \int_0^\infty p_n(s;t)dt = \frac{1}{\mu s} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}(i, m-i)(i - s + 1) \\
&= \frac{1}{\mu s} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}(i, m-i)(i-s) + \frac{1}{\mu s} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}(i, m-i) \\
&= \frac{1}{\mu s} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}(i, m-i)(i-s) + \frac{1}{\mu s}\mathbb{P}(W > 0) = C + D,
\end{aligned}$$

where D is given by

$$D = \frac{1}{\mu s}\mathbb{P}(W > 0) = \frac{1}{\mu s} \frac{B}{A+B}$$

and A and B were defined in (C18) and (C28), respectively, and C is given by,

$$\begin{aligned}
C &= \frac{1}{\mu s} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}(i, m-i)\,(i-s) \\
&= \frac{1}{\mu s} \pi_0 \sum_{m=s}^{n-1} \sum_{i=s}^{m} \frac{R_1^i}{s!s^{i-s}} \frac{R_2^{m-i}}{(m-i)!} (i-s) = \frac{1}{\mu s} \frac{Ge^{-(R_1+R_2)}}{A+B}.
\end{aligned}$$

We will rewrite G in the following way:

$$G = \sum_{m=s}^{n-1} \sum_{i=s}^{m} \frac{R_1^i}{s!s^{i-s}} \frac{R_2^{m-i}}{(m-i)!} (i-s) = \sum_{j=0}^{n-s-1} \sum_{i=s}^{n-j-1} \frac{R_1^i}{s!s^{i-s}} \frac{R_2^j}{j!} (i-s)$$

$$= \sum_{j=0}^{n-s-1} \sum_{i=0}^{n-s-j-1} \frac{R_1^{i+s}}{s!s^i} \frac{R_2^j}{j!}\, i = \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \sum_{i=0}^{n-s-j-1} i\,\rho^i\, \frac{R_2^j}{j!}$$

Using the formula

$$\sum_{l=0}^{M} l\,\rho^l = \rho\Big(\sum_{l=0}^{M}\rho^l\Big)' = \rho\left(\frac{1-\rho^{M+1}}{1-\rho}\right)' = -M\frac{\rho^{M+1}}{1-\rho} + \rho\,\frac{1-\rho^M}{(1-\rho)^2},$$

we can rewrite G as a sum $G = G_1 + G_2$, where

$$G_1 = -\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} (n-s-j-1)\,\frac{\rho^{n-s-j}}{1-\rho}\,\frac{R_2^j}{j!},$$

and

$$G_2 = \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \rho\,\frac{1-\rho^{n-s-j-1}}{(1-\rho)^2}\,\frac{R_2^j}{j!}.$$

Therefore,

$$\begin{aligned}
G_1 &= -\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} (n-s-j-1)\,\frac{\rho^{n-s-j}}{1-\rho}\,\frac{R_2^j}{j!} \\
&= -(n-s-1)\,\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \frac{\rho^{n-s-j}}{1-\rho}\,\frac{R_2^j}{j!} + \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} j\,\frac{\rho^{n-s-j}}{1-\rho}\,\frac{R_2^j}{j!} \\
&= -(n-s-1)\,\frac{\rho^{n-s}}{1-\rho}\,\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \frac{(R_2/\rho)^j}{j!} + \frac{R_1^s}{s!}\,\frac{\rho^{n-s}}{1-\rho} \sum_{j=0}^{n-s-1} j\,\frac{(R_2/\rho)^j}{j!} \\
&= -(n-s-1)\,\frac{\rho^{n-s}}{1-\rho}\,\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \frac{(R_2/\rho)^j}{j!} + \frac{R_1^s}{s!}\,\frac{R_2}{\rho}\,\frac{\rho^{n-s}}{1-\rho} \sum_{j=0}^{n-s-2} \frac{(R_2/\rho)^j}{j!} \\
&= -\,(n-s-R_2/\rho-1)\,\frac{\rho^{n-s}}{1-\rho}\,\frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \frac{(R_2/\rho)^j}{j!} - \frac{R_1^s}{s!}\,\frac{1}{1-\rho}\,\frac{R_2^{n-s}}{(n-s-1)!} \\
&= -\,(n-s-R_2/\rho-1)\,\mathrm{e}^{R_1+R_2} B_2 - \frac{1}{1-\rho}\,\frac{R_1^s}{s!}\,\frac{R_2^{n-s}}{(n-s-1)!},
\end{aligned}$$

where B_2 was defined in Lemma 2. For G_2 we have,

$$\begin{aligned}
G_2 &= \frac{R_1^s}{s!}\,\frac{\rho}{(1-\rho)^2} \sum_{j=0}^{n-s-1} \left(1-\rho^{n-s-j-1}\right) \frac{R_2^j}{j!} = \frac{R_1^s}{s!}\,\frac{\rho}{(1-\rho)^2} \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} - \frac{R_1^s}{s!}\,\frac{\rho^{n-s}}{(1-\rho)^2} \sum_{j=0}^{n-s-1} \frac{(R_2/\rho)^j}{j!} \\
&= \mathrm{e}^{R_1+R_2}\,\frac{1}{1-\rho}\,(\rho B_1 - B_2),
\end{aligned}$$

with B_1 as in Lemma 1. Note that

$$\begin{aligned}
&n-s-\frac{R_2}{\rho}-1 = R_2 - \eta\sqrt{R_2} - \frac{R_2}{\rho} - 1 = R_2\,(1-1/\rho) + \eta\sqrt{R_2} - 1 \\
&= R_2\left(\frac{R_1-s}{\sqrt{R_1}}\right) + \eta\sqrt{R_2} - 1 = R_2 \cdot -\frac{\beta\sqrt{R_1}}{R_1} + \eta\sqrt{R_2} - 1 \approx \sqrt{R_1}\left(\eta\sqrt{\xi} - \beta\xi\right) \approx \sqrt{s}\left(\eta/\sqrt{\xi} - \beta/\xi\right),
\end{aligned}$$

for s large, where $\xi = R_1/R_2 = \delta/p\mu$. Furthermore,

$$\frac{1}{1-\rho}\,\frac{R_1^s}{s!}\,\frac{R_2^{n-s}}{(n-s-1)!} = \frac{n-s}{1-\rho}\,\frac{R_1^s}{s!}\,\frac{R_2^{n-s}}{(n-s)!}$$

$$
\begin{aligned}
&= \frac{n-s}{1-\rho}\,\mathrm{e}^{R_1+R_2}\,\mathbb{P}\left(\mathrm{Pois}(R_1)=s\right)\,\mathbb{P}\left(\mathrm{Pois}(R_2)=n-s\right)\\
&= \frac{R_2+\eta\sqrt{R_2}}{\beta/\sqrt{R_1}}\mathrm{e}^{R_1+R_2}\mathbb{P}\left(\mathrm{Pois}(R_1)=s\right)\,\mathbb{P}\left(\mathrm{Pois}(R_2)=n-s\right)\\
&= \sqrt{R_2}\,\mathrm{e}^{R_1+R_2}\,\frac{1}{\beta}\,\sqrt{R_1}\mathbb{P}\left(\mathrm{Pois}(R_1)=s\right)\cdot\sqrt{R_2}\mathbb{P}\left(\mathrm{Pois}(R_2)=n-s\right))\\
&\approx \sqrt{R_1}\,\mathrm{e}^{R_1+R_2}\,\frac{1}{\sqrt{\xi}\beta}\,\phi(\beta)\phi(\eta),
\end{aligned}
$$

where we used that $\sqrt{R}\,\mathbb{P}(\mathrm{Pois}(R)=R+x\sqrt{R})\to\phi(x)$ as $R\to\infty$. Hence, we have for G_1

$$
\frac{1}{\sqrt{s}}\mathrm{e}^{-(R_1+R_2)}\,G_1\to-\left(\eta/\sqrt{\xi}-\beta/\xi\right)\cdot\frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega)-\frac{1}{\sqrt{\xi}\beta}\,\phi(\beta)\phi(\eta)
$$

and for G_2

$$
\begin{aligned}
\frac{1}{\sqrt{s}}\mathrm{e}^{-(R_1+R_2)}\,G_2 &= \frac{1}{\sqrt{s}}\frac{1}{1-\rho}\,(\rho B_1-B_2)\approx\frac{1}{\beta}\,(\rho B_1-B_2)\\
&\to\frac{\phi(\beta)\Phi(\eta)}{\beta^2}-\frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta^2}\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega),
\end{aligned}
$$

as $s\to\infty$. In total, this gives

$$
\frac{1}{\sqrt{s}}\mathrm{e}^{-(R_1+R_2)}\,G\to\left(\frac{1}{\xi}-\frac{\eta}{\beta\sqrt{\xi}}-\frac{1}{\beta^2}\right)\phi(\sqrt{\beta^2+\eta^2})\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega)+\frac{\phi(\beta)\Phi(\eta)}{\beta^2}-\frac{\phi(\beta)\phi(\eta)}{\sqrt{\xi}\beta}.
$$

We can conclude,

$$
\begin{aligned}
\sqrt{R_1}\,\mathbb{E}[W] &= \frac{\sqrt{R_1}}{\mu s}\frac{Ge^{-(R_1+R_2)}}{A+B}+\frac{\sqrt{s}}{\mu s}\,\mathbb{P}(W>0)=\frac{\sqrt{s}}{\mu s}\,\frac{Ge^{-(R_1+R_2)}}{A+B}+\frac{1}{\mu\sqrt{s}}\,\mathbb{P}(W>0)\\
&\to\frac{1}{\mu}\,\frac{\left(\frac{1}{\xi}-\frac{\eta}{\beta\sqrt{\xi}}-\frac{1}{\beta^2}\right)\phi(\sqrt{\beta^2+\eta^2})\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega)+\frac{\phi(\beta)\Phi(\eta)}{\beta^2}-\frac{\phi(\beta)\phi(\eta)}{\sqrt{\xi}\beta}}{\int_{-\infty}^{\beta}\Phi(\eta+(\beta-t)\sqrt{\xi})\,d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}\,\mathrm{e}^{\frac{1}{2}\omega^2}\,\Phi(\omega)},
\end{aligned}
$$

as $s\to\infty$. □

The second theorem gives the approximation for the case where $\beta=0$.

Theorem 3 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta=0$. Then*

$$
\lim_{\lambda\to\infty}\sqrt{s}\mathbb{E}[W]=\frac{1}{2\mu}\frac{\xi^{-1}\left((\eta^2+1)\Phi(\eta)+\eta\phi(\eta)\right)}{\sqrt{2\pi}\int_{-\infty}^{0}\Phi\left(\eta-t\sqrt{\xi}\right)d\Phi(t)+\sqrt{\xi^{-1}}\left(\eta\Phi(\eta)+\phi(\eta)\right)},
$$

where $\xi=\frac{R_1}{R_2}=\frac{\delta}{p\mu}$, $\omega=\eta-\beta\sqrt{\xi^{-1}}$.

Proof Proof: As before

$$
\mathbb{E}[W]=\frac{1}{\mu s}\frac{Ge^{-(R_1+R_2)}+B}{A+B}.
$$

Since $\beta = 0$, we have $\rho = 1$ so that

$$G = \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} \sum_{i=0}^{n-s-j-1} i \, \frac{R_2^j}{j!} = \frac{1}{2} \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} (n-s-j)(n-s-j-1) \, \frac{R_2^j}{j!}$$

$$= \frac{1}{2} \frac{R_1^s}{s!} (n-s-1) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) - \frac{1}{2} \frac{R_1^s}{s!} \sum_{j=0}^{n-s-1} j \, \frac{R_2^j}{j!} (n-s-j)$$

$$= \frac{1}{2} \frac{R_1^s}{s!} (n-s-1) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) - \frac{1}{2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-2} \frac{R_2^j}{j!} (n-s-j-1)$$

$$= \frac{1}{2} \frac{R_1^s}{s!} (n-s-1) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) - \frac{1}{2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j-1)$$

$$= \frac{1}{2} \frac{R_1^s}{s!} (n-s-1) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) - \frac{1}{2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) + \frac{1}{2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!}$$

$$= \frac{1}{2} \frac{R_1^s}{s!} (n-s-R_2-1) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) + \frac{1}{2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!}.$$

Here,

$$\begin{aligned} \mathrm{e}^{-R_2} \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} (n-s-j) &= \mathrm{e}^{-R_2} (n-s) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} - \mathrm{e}^{-R_2} R_2 \sum_{j=0}^{n-s-2} \frac{R_2^j}{j!} \\ &= \mathrm{e}^{-R_2} (n-s-R_2) \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} + \mathrm{e}^{-R_2} (n-s) \frac{R_2^{n-s}}{(n-s)!} \\ &= \eta \sqrt{R_2} \, \mathbb{P}(\mathrm{Pois}(R_2) \le n-s-1) + (n-s-1) \mathbb{P}(\mathrm{Pois}(R_2) = n-s) \\ &= \eta \sqrt{R_2} \, \mathbb{P}(\mathrm{Pois}(R_2) \le n-s) + (R_2-1) \mathbb{P}(\mathrm{Pois}(R_2) = n-s) \\ &\approx \sqrt{R_2} \, (\eta \, \Phi(\eta) + \phi(\eta)) = \sqrt{\xi^{-1}} \sqrt{R_1} \, (\eta \, \Phi(\eta) + \phi(\eta)) \, . \end{aligned} \tag{C30}$$

Furthermore

$$n-s-1-R_2 = \eta \sqrt{R_2} - 1 \approx \eta \sqrt{\xi^{-1}} \sqrt{R_1} \tag{C31}$$

and

$$\mathrm{e}^{-R_1-R_2} \frac{R_1^s}{s!} R_2 \sum_{j=0}^{n-s-1} \frac{R_2^j}{j!} = \xi^{-1} \sqrt{R_1} \sqrt{R_1} \mathbb{P}(\mathrm{Pois}(R_1) = s) \, \mathbb{P}(\mathrm{Pois}(R_2 \le n-s-1) \tag{C32}$$

$$\approx \sqrt{R_1} \xi^{-1} \, \phi(0) \Phi(\eta). \tag{C33}$$

Combining (C30)–(C33), we find

$$\begin{aligned} \mathrm{e}^{-R_1-R_2} G &\approx \frac{1}{2} \xi^{-1} \eta \sqrt{R_1} \, \mathbb{P}(\mathrm{Pois}(R_1) = s) \ \ (\eta \Phi(\eta) + \phi(\eta)) + \frac{1}{2} \sqrt{R_1} \phi(0) \Phi(\eta) \\ &\approx \frac{1}{2} \xi^{-1} \eta \sqrt{R_1} \, (\eta \, \Phi(\eta) + \phi(\eta)) + \frac{1}{2} \xi^{-1} \sqrt{R_1} \, \phi(0) \Phi(\eta) \\ &= \frac{1}{2} \sqrt{R_1} \, \xi^{-1} \phi(0) \left[(\eta^2+1) \Phi(\eta) + \eta \, \phi(\eta) \right] , \end{aligned}$$

for R_1 large. Hence, we can conclude

$$\sqrt{R_1} \, \mathbb{E}[W] = \frac{\sqrt{R_1}}{\mu s} \frac{G e^{-R_1-R_2}}{A+B} + \frac{\sqrt{R_1}}{\mu s} \mathbb{P}(W > 0) \approx \frac{1}{\sqrt{R_1} \mu} \frac{\mathrm{e}^{-R_1-R_2} G}{A+B} + \frac{1}{\sqrt{R_1} \mu} \mathbb{P}(W > 0)$$

$$\to \frac{1}{2\mu} \frac{\xi^{-1}\left[(\eta^2+1)\Phi(\eta)+\eta\,\phi(\eta)\right]}{\sqrt{2\pi}\int_{-\infty}^{0}\Phi(\eta-t\sqrt{\xi})d\Phi(t)+\sqrt{\xi^{-1}}(\eta\Phi(\eta)+\phi(\eta))},$$

where we used that $\phi(0)=1/\sqrt{2\pi}$, and $\eta=\frac{\gamma-\beta\sqrt{r}}{\sqrt{1-r}}$.

C.7 Approximation of the blocking probability

We prove here the approximations given in Theorem 1 of the probability of blocking. For convenience we repeat here the relevant theorem.

Theorem 4 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta \neq 0$. Then*

$$\lim_{\lambda\to\infty}\sqrt{s}\mathbb{P}(block)=\frac{\sqrt{r}\phi(\gamma)\Phi(-\omega\sqrt{r})+\phi(\sqrt{\eta^2+\beta^2})e^{\frac{\omega^2}{2}}\Phi(\omega)}{\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\xi}\right)d\Phi(t)+\frac{\phi(\beta)\Phi(\eta)}{\beta}-\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)}$$

where $\eta=\frac{\gamma-\beta\sqrt{r}}{\sqrt{1-r}}$, and $\omega=\eta-\frac{\beta\sqrt{1-r}}{\sqrt{r}}=\frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}$.

Proof Proof: It follows from (4) that the probability of blocking is given by

$$\mathbb{P}_n=\pi_0\left(\sum_{i=0}^{s}\frac{1}{i!}R_1^i\frac{1}{(n-i)!}R_2^{n-i}+\sum_{i=s+1}^{n}\frac{1}{s!s^{i-s}}R_1^i\frac{1}{(n-i)!}R_2^{n-i}\right)=\frac{\delta_1+\delta_2}{\tilde{\xi}+\tilde{\zeta}_1-\tilde{\zeta}_2},$$

where

$$\delta_1=\sum_{i=0}^{s}\frac{1}{i!}R_1^i\frac{1}{(n-i)!}R_2^{n-i}e^{-(R_1+R_2)}$$

$$\delta_2=\sum_{i=s+1}^{n}\frac{1}{s!s^{i-s}}R_1^i\frac{1}{(n-i)!}R_2^{n-i}e^{-(R_1+R_2)}$$

$$\tilde{A}=\sum_{\substack{i,j|i\le s,\\ i+j\le n}}\frac{1}{i!j!}R_1^iR_2^je^{-(R_1+R_2)}$$

$$\tilde{B_1}=\frac{1}{s!}R_1^s\frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}R_2^le^{-(R_1+R_2)}$$

$$\tilde{B_2}=\frac{1}{s!}R_1^s\frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^le^{-(R_1+R_2)}.$$

Note that by Lemmas 1,2 and 3

$$\lim_{\lambda\to\infty}\tilde{B_1}=\lim_{\lambda\to\infty}B_1=\frac{\phi(\beta)\Phi(\eta)}{\beta},$$

$$\lim_{\lambda\to\infty}\tilde{B_2}=\lim_{\lambda\to\infty}B_2=\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1),$$

$$\lim_{\lambda\to\infty}\tilde{A}=\lim_{\lambda\to\infty}A=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{\delta}{p\mu}}\right)d\Phi(t),$$

$$
\begin{aligned}
\delta_1 &= \sum_{i=0}^{s} \frac{1}{i!} R_1^i \frac{1}{(n-i)!} R_2^{n-i} e^{-(R_1+R_2)} = \frac{1}{n!} e^{-(R_1+R_2)} \sum_{i=0}^{s} \frac{n!}{i!(n-i)!} R_1^i R_2^{n-i} \\
&= \frac{1}{n!} e^{-(R_1+R_2)} (R_1+R_2)^n \sum_{i=0}^{s} \frac{n!}{i!(n-i)!} \left(\frac{R_1}{R_1+R_2}\right)^i \left(\frac{R_2}{R_1+R_2}\right)^{n-i} \\
&= \frac{1}{n!} e^{-(R_1+R_2)} (R_1+R_2)^n \sum_{i=0}^{s} \mathbb{P}(X_\lambda = (i, n-i)) \\
&= \mathbb{P}(Y_\lambda = n) \sum_{i=0}^{s} \mathbb{P}(X_\lambda = (i, n-i)) = \mathbb{P}(Y_\lambda = n)\mathbb{P}(X_\lambda^1 \le s)
\end{aligned}
$$

where X_λ is a random variable with Multinomial distribution with parameters (n, p_i, p_j), $p_i = \frac{R_1}{R_1+R_2}$, $p_j = \frac{R_2}{R_1+R_2}$, Y_λ is a random variable with Poisson distribution with parameter $R_1 + R_2$, and X_λ^1 is a random variable with Binomial distribution with parameters (n, p_i). By the CLT and the use of (C21)

$$
\begin{aligned}
\mathbb{P}(X_\lambda^1 \le s) &= \Phi\left(\frac{s - np_i}{\sqrt{np_i(1-p_i)}}\right) = \Phi\left(\frac{s - n\frac{R_1}{R_1+R_2}}{\sqrt{n\frac{R_1}{R_1+R_2}(1-\frac{R_1}{R_1+R_2})}}\right) = \Phi\left(\frac{s - n\frac{R_1}{R_1+R_2}}{\sqrt{n\frac{R_1}{R_1+R_2}\frac{R_2}{R_1+R_2}}}\right) \\
&= \Phi\left(\frac{s(R_1+R_2) - nR_1}{\sqrt{nR_1R_2}}\right) = \Phi\left(\sqrt{\frac{R_1}{R_2}} \frac{s\frac{R_1+R_2}{R_1} - n}{\sqrt{n}}\right) = \Phi\left(\sqrt{\frac{R_1}{R_2}} \frac{s(1+\frac{R_2}{R_1}) - n}{\sqrt{n}}\right) \\
&= \Phi\left(\sqrt{\frac{R_1}{R_2}} \frac{s + \frac{R_2}{\rho} - n}{\sqrt{n}}\right) = \Phi\left(-\sqrt{\frac{R_1}{R_2}}\sqrt{\frac{R_2}{n\rho}} \cdot \frac{n - s - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right) = \Phi\left(-\sqrt{\frac{s}{n}} \cdot \frac{n - s - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right) \\
&\approx \Phi\left(-\sqrt{\frac{R_1}{R_1+R_2}} \cdot \left(\eta - \beta\frac{R_2}{R_1}\right)\right) = \Phi\left(\frac{\beta\frac{R_2}{R_1} - \eta}{\sqrt{1+\frac{R_2}{R_1}}}\right) = \Phi\left(\frac{\beta\sqrt{\frac{p\mu}{\delta}} - \eta}{\sqrt{1+\frac{p\mu}{\delta}}}\right).
\end{aligned}
\tag{C34}
$$

By the normal approximation of the Poisson distribution:

$$
\mathbb{P}(Y_\lambda = n) \approx \frac{1}{\sqrt{R_1+R_2}}\phi\left(\frac{n-(R_1+R_2)}{\sqrt{R_1+R_2}}\right) \approx \frac{1}{\sqrt{s}\sqrt{1+\frac{p\mu}{\delta}}}\phi\left(\frac{\eta\sqrt{\frac{p\mu}{\delta}}+\beta}{\sqrt{1+\frac{p\mu}{\delta}}}\right). \tag{C35}
$$

We based on the following equivalences (as λ tends to ∞) to develop Equations (C34) and (C35):

$$
\begin{aligned}
&R_1 + R_2 = \frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} \approx s + (1-p)s\mu\left(\frac{p}{(1-p)\delta}\right) = s\left(1+\frac{\mu p}{\delta}\right) = s\left(1+\frac{R_2}{R_1}\right); \\
&n - (R_1+R_2) \approx s + R_2 + \eta\sqrt{R_2} - (R_1+R_2) = s + \eta\sqrt{R_2} - R_1 \approx s + \eta\sqrt{R_2} - s + \beta\sqrt{s} \\
&\approx \eta\sqrt{\frac{s\mu p}{\delta}} + \beta\sqrt{s}; \\
&\frac{n-(R_1+R_2)}{\sqrt{R_1+R_2}} \approx \frac{\eta\sqrt{\frac{\mu p}{\delta}}+\beta}{\sqrt{1+\frac{\mu p}{\delta}}}.
\end{aligned}
$$

Following Equations (C34) and (C35) we get

$$\delta_1 = \mathbb{P}(Y_\lambda = n)\mathbb{P}(X_\lambda^1 \le s) \approx \frac{1}{\sqrt{s}\sqrt{1+\frac{\mu p}{\delta}}}\phi\left(\frac{\eta\sqrt{\frac{\mu p}{\delta}}+\beta}{\sqrt{1+\frac{\mu p}{\delta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{\mu p}{\delta}}-\eta}{\sqrt{1+\frac{\mu p}{\delta}}}\right).$$

Now lets find an approximation for $\tilde{\delta_2}$.

$$\begin{aligned}\delta_2 &= \sum_{i=s+1}^{n}\frac{1}{s!s^{i-s}}R_1^i\frac{1}{(n-i)!}R_2^{n-i}e^{-(R_1+R_2)} = \frac{e^{-(R_1+R_2)}}{s!s^{-s}}\sum_{i=s+1}^{n}\rho^i\frac{1}{(n-i)!}R_2^{n-i}\\ &= \frac{e^{-(R_1+R_2)}}{s!s^{-s}}R_2^n\sum_{j=0}^{n-s-1}\frac{1}{j!}\left(\frac{\rho}{R_2}\right)^{n-j} = \frac{e^{-(R_1+R_2)}}{s!s^{-s}}\rho^n\sum_{j=0}^{n-s-1}\frac{1}{j!}\left(\frac{R_2}{\rho}\right)^j.\end{aligned}$$

When comparing δ_2 to B_2 form Equation (C29), we observe that

$$\delta_2 = (1-\rho)B_2 \approx \frac{\beta}{\sqrt{s}}B_2.$$

Therefore, based on the approximation of B_2 from Lemma 2 we get

$$lim_{\lambda\to\infty}\delta_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\sqrt{s}}e^{\frac{1}{2}\omega^2}\Phi(\omega).$$

This proves Theorem 4. □

The next theorem gives the approximation for the case where $\beta = 0$.

Theorem 5 *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (5) with $\beta = 0$. Define and $\xi = \frac{R_1}{R_2} = \frac{\delta}{p\mu}$, then*

$$\lim_{\lambda\to\infty}\sqrt{s}\mathbb{P}(block) = \frac{\sqrt{r}\phi(\gamma)\Phi(-\frac{\gamma\sqrt{r}}{\sqrt{1-r}}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0}\Phi\left(\eta - t\sqrt{\xi}\right)d\Phi(t) + \sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\Phi(\eta)+\phi(\eta)\right)},$$

where $\eta = \frac{\gamma-\beta\sqrt{r}}{\sqrt{1-r}}$.

Proof Proof: It follows from (4) that the probability of blocking is given by

$$\mathbb{P}_n = \frac{\delta_1+\delta_2}{\tilde{A}+\tilde{B}},$$

where

$$\begin{aligned}\delta_1 &= \sum_{i=0}^{s}\frac{1}{i!}R_1^i\frac{1}{i!}R_2^{n-i}e^{-(R_1+R_2)}\\ \delta_2 &= \sum_{i=s+1}^{n}\frac{1}{s!s^{i-s}}R_1^i\frac{1}{(n-i)!}R_2^{n-i}e^{-(R_1+R_2)}\\ \tilde{A} &= \sum_{\substack{i,j|i\le s,\\ i+j\le n}}\frac{1}{i!j!}R_1^iR_2^je^{-(R_1+R_2)}\\ \tilde{B} &= \frac{1}{s!}R_1^s\frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}R_2^le^{-(R_1+R_2)} - \frac{1}{s!}R_1^s\frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^le^{-(R_1+R_2)}.\end{aligned}$$

Note that by Lemmas 3 and 4

$$\lim_{\lambda\to\infty} \tilde{A} = \lim_{\lambda\to\infty} A = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta}{\mu p}}\right) d\Phi(t)$$

$$\lim_{\lambda\to\infty} \tilde{B} = \lim_{\lambda\to\infty} B = \sqrt{\frac{\mu p}{\delta}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta)\right).$$

In addition, the approximations for δ_1 and δ_2 are the same as the proof of Theorem 4.

$$\lim_{\beta\to 0}\lim_{\lambda\to\infty} \sqrt{s}\delta_2 = \frac{1}{\sqrt{2\pi}}\Phi(\eta)$$

$$\lim_{\beta\to 0}\lim_{\lambda\to\infty} \sqrt{s}\delta_1 = \frac{1}{\sqrt{1+\frac{\mu p}{\delta}}}\phi\left(\frac{\eta}{\sqrt{1+\frac{\delta}{\mu p}}}\right)\Phi\left(-\frac{\eta}{\sqrt{1+\frac{\mu p}{\delta}}}\right).$$

This proves Theorem 5.

□

Appendix D Proof of Proposition 3

Define

$$A(s,n) = \sum_{k=0}^{s} \frac{k}{s}\binom{n}{k} b^k, \quad B(s,n) = \sum_{k=s+1}^{n} \frac{k!}{s!}\binom{n}{k} s^{s-k} b^k, \quad C(s,n) = \sum_{k=0}^{s}\binom{n}{k} b^k,$$

where $b = \delta/p\mu = r/(1-r)$. Then

$$\rho_J(s,n) = \frac{A(s,n) + B(s,n)}{C(s,n) + B(s,n)}.$$

Proving that $\rho_{\max}(s,n) \to 1$ as $R \to \infty$ with s and n as in (5) is equivalent to showing that

$$1 - \rho_{\max}(s,n) = \frac{C(s,n) - A(s,n)}{C(s,n) + B(s,n)} = \frac{(1+b)^{-n}[C(s,n) - A(s,n)]}{(1+b)^{-n}[C(s,n) + B(s,n)]} \to 0. \tag{D36}$$

First, we rewrite

$$\begin{aligned}
(1+b)^{-n} A(s,n) &= (1+b)^{-n} \sum_{k=1}^{s} \frac{n}{s}\binom{n-1}{k-1} b^k \\
&= \frac{n}{s}\left(\frac{b}{1+b}\right)\sum_{k=0}^{s-1}\binom{n-1}{k}\left(\frac{b}{1+b}\right)^k\left(\frac{1}{1+b}\right)^{n-1-k} \\
&= \frac{rn}{s}\sum_{k=0}^{s-1}\binom{n-1}{k} r^k (1-r)^{n-1-k} = \frac{rn}{s}\mathbb{P}(\mathrm{Bin}(n-1,r) \le s-1) \\
&= \frac{rn}{s}\mathbb{P}\left(\frac{\mathrm{Bin}(n-1,r) - (n-1)r}{\sqrt{nr(1-r)}} \le \frac{s-1-(n-1)r}{\sqrt{nr(1-r)}}\right) \to \Phi\left(\frac{\beta - \gamma\sqrt{r}}{\sqrt{1-r}}\right),
\end{aligned}$$

since $nr/s = 1 + O(1/\sqrt{R_1})$. Also,

$$\begin{aligned}
(1+b)^{-n} C(s,n) &= \sum_{k=0}^{s}\binom{n}{k}\left(\frac{b}{1+b}\right)^k\left(\frac{1}{1+b}\right)^{n-k} \\
&= \sum_{k=0}^{s}\binom{n}{k} r^k (1-r)^{n-k} = \mathbb{P}(\mathrm{Bin}(n,r) \le s) \to \Phi\left(\frac{\beta - \gamma\sqrt{r}}{\sqrt{1-r}}\right).
\end{aligned}$$

Therefore, we have $(1+b)^{-n}[C(s,n)-A(s,n)] \to 0$ as $\lambda \to \infty$. For the remaining term,

$$(1+b)^{-n}B(s,n) = (1+b)^{-n}\sum_{k=s+1}^{n}\binom{n}{k}\frac{k!}{s!}s^{s-k}b^k = (1+b)^{-n}\frac{n!}{s!}s^s\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{-k}$$
$$= (1+b)^{-n}\frac{n!}{s!}s^s\left(\frac{b}{s}\right)^n\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{n-k} = r^n\frac{n!}{s!}s^{s-n}\sum_{m=0}^{n-s-1}\frac{1}{m!}\left(\frac{s}{b}\right)^m$$
$$= \left(\frac{r}{s}\right)^n\frac{n!}{s!}s^s\,\mathrm{e}^{s/b}\,\mathbb{P}(\mathrm{Pois}(s/b)\le n-s-1),$$

in which

$$\mathbb{P}(\mathrm{Pois}(s/b)\le n-s-1) = \mathbb{P}\left(\frac{\mathrm{Pois}(s/b)-s/b}{\sqrt{s/b}} \le \frac{n-s-1-s/b}{\sqrt{s/b}}\right) \to \Phi\left(\frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}\right),$$

as $\lambda \to \infty$. By Stirling's approximation,

$$\left(\frac{r}{s}\right)^n\frac{n!}{s!}s^s\,\mathrm{e}^{s/b} \sim \left(\frac{r}{s}\right)^n\sqrt{\frac{n}{s}}\,\frac{n^n\mathrm{e}^{-n}}{s^s\mathrm{e}^{-s}}\,s^s\,\mathrm{e}^{s/b} = \left(\frac{rn}{s}\right)^n\sqrt{\frac{n}{s}}\mathrm{e}^{-n+s+s/b}$$
$$= \left(\frac{rn}{s}\right)^n\sqrt{\frac{n}{s}}\mathrm{e}^{-n+s/r}.$$

Since,

$$\frac{rn}{s} = 1 + \frac{\gamma\sqrt{r}-\beta}{\sqrt{R_1}} + O(1/R_1),$$

we find $\sqrt{n/s} = 1/\sqrt{r} + O(1/\sqrt{R_1})$ and

$$\log\left[\left(\frac{rn}{s}\right)^n\sqrt{\frac{n}{s}}\mathrm{e}^{-n+s/r}\right] = n\log\left[\frac{rn}{s}\right] - n + \frac{s}{r}$$
$$= -n\left[\left(1-\frac{rn}{s}\right)+\frac{1}{2}\left(1-\frac{rn}{s}\right)^2+O(1/R^{3/2})\right]+\frac{s}{r}\left(1-\frac{rn}{s}\right)$$
$$= \frac{s}{r}\left(1-\frac{rn}{s}\right)^2-\frac{n}{2}\left(1-\frac{rn}{s}\right)^2+O(1/\sqrt{R_1}) = \frac{(\gamma\sqrt{r}-\beta)^2}{2r}+O(1/\sqrt{R_1}),$$

as $R \to \infty$ and hence,

$$(1+b)^{-n}B(s,n) \to \phi\left(\frac{\gamma\sqrt{r}-\beta}{\sqrt{r}}\right)\Phi\left(\frac{\gamma-\beta/\sqrt{r}}{\sqrt{1-r}}\right).$$

Hence, we conclude that the denominator of (D36) converges to a constant value as λ grows, and hence the $1-\rho_{\max}(s,n) \to 0$ as $\lambda \to \infty$.

Appendix E Numerical results on accuracy of the asymptotic approximations

To see how this heuristic approach performs under different settings, and particularly if $R_1 \to \infty$, we compare here the approximated delay probability in the Erlang-R model with holding as solution of the fixed-point procedure to the outcomes of simulation experiments.

In this section, we examine to accuracy of the asymptotic approximations of Theorem 1 and the heuristic method in Section 4.3. We perform our numerical experiments for three case instances, with parameter settings as in Table 1.

Table E1 Parameter settings for numerical experiments

	μ	δ	p	r
Case 1	1	0.10	0.90	0.10
Case 2	1	0.25	0.75	0.25
Case 3	1	0.50	0.50	0.50

E.1 Restricted Erlang-R model with blocking

Table E2 Numerical results and limiting approximation for restricted Erlang-R model with blocking for Case 1

	$\beta = 1,\ \gamma = 1$			$\beta = 1,\ \gamma = 2$			$\beta = 2,\ \gamma = 1$		
R	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1270	0.0900	0.2283	0.1553	0.0212	0.1085	0.0237	0.0868	0.0282
10	0.1340	0.0910	0.1919	0.1628	0.0206	0.1205	0.0206	0.0872	0.0188
25	0.1981	0.0945	0.1614	0.2356	0.0216	0.2145	0.0277	0.0876	0.0123
50	0.1513	0.0963	0.1588	0.1830	0.0205	0.1496	0.0185	0.0913	0.0116
100	0.1880	0.0956	0.1532	0.2231	0.0224	0.2055	0.0232	0.0888	0.0103
250	0.1797	0.0971	0.1399	0.2143	0.0219	0.2057	0.0203	0.0905	0.0093
App	*0.1767*	*0.0981*	*0.1437*	*0.2108*	*0.0217*	*0.1947*	*0.0188*	*0.0914*	*0.0084*

Table E3 Numerical results and limiting approximation (App) for restricted Erlang-R model with blocking for Case 2

	$\beta = 1,\ \gamma = 1$			$\beta = 1,\ \gamma = 2$			$\beta = 2,\ \gamma = 1$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0911	0.1538	0.0479	0.1431	0.0345	0.0909	0.0130	0.1484	0.0044
10	0.1010	0.1498	0.0560	0.1520	0.0326	0.1025	0.0121	0.1432	0.0042
25	0.1594	0.1509	0.1058	0.2192	0.0405	0.1785	0.0182	0.1383	0.0070
50	0.1201	0.1506	0.0726	0.1697	0.0381	0.1248	0.0119	0.1415	0.0043
100	0.1514	0.1539	0.1001	0.2088	0.0398	0.1704	0.0154	0.1413	0.0059
250	0.1459	0.1524	0.0957	0.2003	0.0397	0.1618	0.0136	0.1403	0.0051
App	*0.1429*	*0.1569*	*0.0940*	*0.1976*	*0.0391*	*0.1617*	*0.0126*	*0.1445*	*0.0048*

Table E4 Numerical results and limiting approximation for restricted Erlang-R model with blocking for Case 3

	$\beta=1,\ \gamma=1$			$\beta=1,\ \gamma=2$			$\beta=2,\ \gamma=1$		
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{P}(\mathrm{b})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0547	0.1945	0.0221	0.1181	0.0604	0.0617	0.0034	0.1888	0.0009
10	0.0579	0.2158	0.0237	0.1325	0.0526	0.0746	0.0030	0.2093	0.0008
25	0.1113	0.2086	0.0544	0.1959	0.0641	0.1311	0.0070	0.1937	0.0020
50	0.0813	0.2050	0.0363	0.1523	0.0562	0.0933	0.0043	0.1946	0.0011
100	0.1060	0.2146	0.0509	0.1873	0.0632	0.1250	0.0061	0.1999	0.0017
250	0.1006	0.2179	0.0475	0.1820	0.0596	0.1214	0.0052	0.2037	0.0014
App	*0.1011*	*0.2185*	*0.0478*	*0.1792*	*0.0605*	*0.1199*	*0.0052*	*0.2039*	*0.0014*

E.2 Restricted Erlang-R model with holding

Table E5 Simulated and approx. probability of delay in restricted Erlang-R model with holding for Case 1

	$\beta=1,\ \gamma=1$		$\beta=1,\ \gamma=2$		$\beta=2,\ \gamma=1$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1532	0.1031	0.1628	0.1216	0.0310	0.0121
10	0.1622	0.1272	0.1697	0.1331	0.0267	0.0123
25	0.2340	0.2116	0.2413	0.2342	0.0348	0.0171
50	0.1817	0.1468	0.1890	0.1678	0.0240	0.0108
100	0.2199	0.1931	0.2304	0.2269	0.0293	0.0143
250	0.2110	0.1852	0.2176	0.2230	0.0256	0.0120
App	*0.2076*	*0.1777*	*0.2187*	*0.2050*	*0.0229*	*0.0104*

Table E6 Simulated and approx. probability of delay in restricted Erlang-R model with holding for Case 2

	$\beta=1,\ \gamma=1$		$\beta=1,\ \gamma=2$		$\beta=2,\ \gamma=1$	
R_1	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1327	0.0740	0.1620	0.1096	0.0219	0.0079
10	0.1446	0.0894	0.1683	0.1207	0.0199	0.0073
25	0.2204	0.1631	0.2442	0.2203	0.0283	0.0128
50	0.1694	0.1122	0.1888	0.1507	0.0190	0.0078
100	0.2098	0.1524	0.2322	0.2111	0.0244	0.0097
250	0.2033	0.1534	0.2190	0.1979	0.0214	0.0083
App	*0.1840*	*0.1277*	*0.2109*	*0.1759*	*0.0169*	*0.0066*

Table E7 Simulated and approx. probability of delay in restricted Erlang-R model with holding for Case 3

App	$\beta = 1, \gamma = 1$		$\beta = 1, \gamma = 2$		$\beta = 2, \gamma = 1$	
R	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(\mathrm{d})$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0977	0.0413	0.1521	0.0851	0.0072	0.0019
10	0.1070	0.0469	0.1648	0.1028	0.0067	0.0018
25	0.1926	0.1076	0.2421	0.1874	0.0148	0.0043
50	0.1431	0.0727	0.1876	0.1342	0.0092	0.0025
100	0.1855	0.1012	0.2282	0.1714	0.0132	0.0038
250	0.1775	0.0963	0.2217	0.1765	0.0114	0.0033
App	*0.1442*	*0.0711*	*0.1981*	*0.1354*	*0.0078*	*0.0022*

References

[1] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y.N., Tseytlin, Y., Yom-Tov, G.B.: On patient flow in hospitals: A data-based queueing-science perspective. Stochastic Systems **5**(1), 146–194 (2015) https://doi.org/10.1214/14-SSY153

[2] Huang, J., Carmeli, B., Mandelbaum, A.: Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Operations Research **63**(4), 892–908 (2015) https://doi.org/10.1287/opre.2015.1389

[3] Carmeli, N., Yom-Tov, G.B., Boxma, O.J.: State-dependent estimation of delay distributions in fork-join networks. Manufacturing & Service Operations Management **25**(3), 1081–1098 (2023) https://doi.org/10.1287/msom.2022.1167

[4] Zychlinski, N., Mandelbaum, A., Momčilović, P., Cohen, I.: Bed blocking in hospitals due to scarce capacity in geriatric institutionsâĂŤcost minimization via fluid models. Manufacturing & Service Operations Management **22**(2), 396–411 (2020) https://doi.org/10.1287/opre.2015.1389

[5] Armony, M., Yom-Tov, G.B.: Hospital vs. home care: Trading off predischarge and postdischarge infection and mortality risks. Manufacturing & Service Operations Management **0**(0), 1–19 (2025) https://doi.org/10.1287/msom.2021.0189

[6] Véricourt, F., Jennings, O.B.: Nurse staffing in medical units: A queueing perspective. Operations Research **59**(6), 1320–1331 (2011) https://doi.org/10.1287/opre.1110.0968

[7] Yom-Tov, G.B., Mandelbaum, A.: Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. Manufacturing & Service Operations Management **16**(2), 283–299 (2014) https://doi.org/10.1287/msom.2013.0474

[8] Daw, A., Castellanos, A., Yom-Tov, G.B., Pender, J., Gruendlinger, L.: The co-production of service: Modeling services in contact centers using hawkes

processes. Management Science **71**(3), 2635–2656 (2025) https://doi.org/10.1287/mnsc.2021.04060

[9] Véricourt, F., Jennings, O.B.: Dimensioning large-scale memberschip services. Operations Research **55**(1), 173–187 (2008) https://doi.org/10.1287/opre.1070.0464

[10] Diwas, K.S.: Does multitasking improve performance? Evidence from the emergency department. Manufacturing & Service Operations Management **16**(2), 168–183 (2014) https://doi.org/10.1287/msom.2013.0464

[11] Goes, P.B., Ilk, N., Lin, M., Zhao, L.J.: When more is less: Field evidence on unintended consequences of multitasking. Management Science **64**(7), 2973–3468 (2018) https://doi.org/10.1287/mnsc.2017.2763

[12] Song, H., Tucker, A.L., Murrell, K.L.: The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. Management Science **61**(12), 3032–3053 (2015) https://doi.org/10.1287/mnsc.2014.2118

[13] Song, H., Tucker, A.L., Graue, R., Moravick, S., Yang, J.J.: Capacity pooling in hospitals: The hidden consequences of off-service placement. Management Science **66**(9), 3825–3842 (2020) https://doi.org/10.1287/mnsc.2019.3395

[14] Chan, C.W., Farias, V.F., Escobar, G.J.: The impact of delays on service times in the intensive care unit. Management Science **63**(7), 2049–2072 (2017) https://doi.org/10.1287/mnsc.2016.2441

[15] Carmen, R., Yom-Tov, G.B., Van Nieuwenhuyse, I., Foubert, B., Ofran, Y.: The role of specialized hospital units in infection and mortality risk reduction among patients with hematological cancers. PLoS ONE **14**(3), 0211694 https://doi.org/10.1371/journal.pone.0211694

[16] Carmen, R., van Nieuwenhuyse, I., van Houdt, B.: Inpatient boarding in emergency departments: Impact on patient delays and system capacity. European Journal of Operational Research **271**(3), 953–967 (2018) https://doi.org/10.1016/j.ejor.2018.06.018

[17] Bayati, M., Kwasnick, S., Luo, D., Plambeck, E.L.: Low-Acuity Patients Delay High-Acuity Patients in an Emergency Department. SSRN (2025). https://doi.org/10.2139/ssrn.3095039

[18] Green, L., Yankovic, N.: Identifying good nursing levels: A queuing approach. Operations Research **59**(4), 942–955 (2011) https://doi.org/10.1287/opre.1110.0943

[19] Bruin, A.M., Bekker, R., van Zanten, L., Koole, G.M.: Dimensioning hospital wards using the Erlang loss model. Annals of Operations Research **178**(1), 23–43

(2009) https://doi.org/10.1007/s10479-009-0647-8

[20] Deo, S., Gurvich, I.: Centralized vs. decentralized ambulance diversion: A network perspective. Management Science **57**(7), 1300–1319 (2011) https://doi.org/10.1287/mnsc.1110.1342

[21] Luo, J., Zhang, J.: Staffing and control of instant messaging contact centers. Operations Research **61**(2), 328–343 (2013) https://doi.org/10.1287/opre.1120.1151

[22] Daw, A., Yom-Tov, G.B.: Asymmetries of service: Interdependence and synchronicity. Operations Research **0**(0), 0–0 (2026) https://doi.org/10.1287/opre.2024.0843

[23] American Hospital Association: 2025 healthcare workforce scan (2025). https://www.aha.org/system/files/media/file/2024/11/2025-Health-Care-Workforce-Scan.pdf Accessed 18 Nov 2025

[24] Zhu, Z., Zheng, W., Tang, N., Zhong, W.: Review of manpower management in healthcare system: Strategies, challenges, and innovations. Journal of Multidisciplinary Healthcar **17**, 5341–5351 (2024) https://doi.org/10.2147/JMDH.S497932

[25] Denton, B.T. (ed.): Handbook of Healthcare Operations Management: Methods and Applications, 2nd edn. North Holland, New York, NY (2013)

[26] Hall, R.W. (ed.): Patient Flow: Reducing Delay in Healthcare Delivery. Springer, New York, NY (2006)

[27] Hall, R.W. (ed.): Handbook of Healthcare System Scheduling. Springer, New York, NY (2012)

[28] Keskinocak, P., Savva, N.: A review of the healthcare-management (modeling) literature published in manufacturing & service operations management. Manufacturing & Service Operations Management **22**(1), 59–72 (2020) https://doi.org/10.1287/msom.2019.0817

[29] Elalouf, A., Wachtel, G.: Queueing problems in emergency departments: A review of practical approaches and research methodologies. Operations Research Forum **3**, 2 (2022) https://doi.org/10.1007/s43069-021-00114-8

[30] Green, L.V., Savin, S.: Reducing delays for medical appointments: A queueing approach. Operations Research **56**(6), 1526–1538 (2008) https://doi.org/10.1287/opre.1080.0575

[31] Bekker, R., Bruin, A.M.: Time-dependent analysis for refused admissions in clinical wards. Annals of Operations Research **178**(1), 45–65 (2009) https://doi.org/

10.1007/s10479-009-0570-z

[32] Campello, F., Ingolfsson, A., Shumsky, R.A.: Queueing models of case managers. Management Science **63**(3), 882–900 (2017) https://doi.org/10.1287/mnsc.2015.2368

[33] Saghafian, S., Wallace, J.H., Van Oyen, M.P., Desmond, J.S., Kronick, S.L.: Patient streaming as a mechanism for improving responsiveness in emergency departments. Operations Research **60**(5), 1080–1097 (2012) https://doi.org/10.1287/opre.1120.1096

[34] Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., Kronick, S.L.: Complexity-augmented triage: A tool for improving patient safety and operational efficiency. Manufacturing & Service Operations Management **16**(3), 329–345 (2014) https://doi.org/10.1287/msom.2014.0487

[35] Tezcan, T., Zhang, J.: Routing and staffing in customer service chat systems with impatient customers. Operations Research **62**(4), 943–956 (2014) https://doi.org/10.1287/opre.2014.1284

[36] Chan, C.W., Yom-Tov, G.B., Escobar, G.: When to use speedup: An examination of service systems with returns. Operations Research **62**(2), 462–482 (2014) https://doi.org/10.1287/opre.2014.1258

[37] Borst, S.C., Mandelbaum, A., Reiman, M.I.: Dimensioning large call centers. Operations Research **52**(1), 17–34 (2004) https://doi.org/10.1287/opre.1030.0081

[38] Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Operations Research **29**(3), 567–588 (1981) https://doi.org/10.1287/opre.29.3.567

[39] Mandelbaum, A., Zeltyn, S.: Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. Operations Research **57**(5), 1189–1205 (2009) https://doi.org/10.1287/opre.1080.0651

[40] Avram, F., Janssen, A., van Leeuwaarden, J.: Loss systems with slow retrials in the Halfin-Whitt regime. Advances in Applied Probability **45**, 274–294 (2013) https://doi.org/10.1239/aap/1363354111

[41] Reed, J.: The $G/GI/N$ queue in the Halfin-Whitt regime. Ann. Appl. Probab. **19**, 2211–2269 (2009) https://doi.org/10.1214/09-AAP609

[42] van Leeuwaarden, J.S., Mathijsen, B.W., Zwart, B.: Economies-of-scale in many-server queueing systems: Tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime. SIAM Review **61**(3), 403–440 (2019) https://doi.org/10.1137/17M1133944

[43] Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, B.: Refining square-root safety staffing by expanding Erlang C. Operations Research **59**(6), 1512–1522 (2011) https://doi.org/10.1287/opre.1110.0991

[44] Zhang, B., van Leeuwaarden, J.S.H., Zwart, B.: Staffing call centers with impatient customers: Refinements to many-server asymptotics. Operations Research **60**(2), 461–474 (2012) https://doi.org/10.1287/opre.1110.1016

[45] Gamarnik, D., Goldberg, D.A.: On the rate of convergence to stationarity of the $M/M/N$ queue in the Halfin-Whitt regime. The Annals of Applied Probability **23**(5), 1879–1912 (2013) https://doi.org/10.1214/12-AAP889

[46] van Leeuwaarden, J.S.H., Mathijsen, B.W.J., Sloothaak, F.: Delayed workload shifting in many-server systems. ACM SIGMETRICS Performance Evaluation Review **43**(2), 10–12 (2015) https://doi.org/10.1145/2825236.2825240

[47] van Leeuwaarden, J., Mathijsen, B., Sloothaak, F.: Cloud provisioning in the QED regime. In: Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools, pp. 180–187. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL (2016). https://doi.org/10.4108/eai.14-12-2015.2262649

[48] Whitt, W.: Offered load analysis for staffing. Manufacturing & Service Operations Management **2**(15), 166–169 (2013) https://doi.org/10.1287/msom.1120.0428

[49] Gordon, W.J., Newell, G.F.: Closed queuing systems with exponential servers. Operations Research **2**(15), 254–265 (1967) https://doi.org/10.1287/opre.15.2.254

[50] Serfozo, R.: Introduction to Stochastic Networks. Springer, New York, NY (1999)

[51] Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, p. 352. The John Hopkins University Press, Baltimore (1981)

[52] Chen, H., Yao, D.D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer, New York, NY (2001)

[53] Khudyakov, P., Feigin, P.D., Mandelbaum, A.: Designing a call center with an IVR (Interactive Voice Response). Queueing Systems **66**(3), 215–237 (2010) https://doi.org/10.1007/s11134-010-9193-y

[54] Lundgren, S., Segesten, K.: Nurses use of time in a medical-surgical ward with all-RN staffing. Journal of Nursing Management **9**(1), 13–20 (2001) https://doi.org/10.1046/j.1365-2834.2001.00192.x

[55] Massey, W.A., Wallace, R.B.: An asymptotically optimal design of the $M/M/c/k$ queue. Unpublished report (2004)

[56] Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, B.: Gaussian expansions and bounds for the Poisson distribution applied to the Erlang-B formula. Advances in Applied Probability **40**(1), 122–143 (2008) https://doi.org/10.1239/aap/1208358889

[57] Schull, M.J., Kiss, A., Szalai, J.-P.: The effect of low-complexity patients on emergency department waiting times. Annals of Emergency Medicine **49**(3), 257–264 (2007) https://doi.org/10.1016/j.annemergmed.2006.06.027

[58] Jennings, O.B., Mandelbaum, A., Massey, W.A., Whitt, W.: Server staffing to meet time-varying demand. Management Science **42**(10), 1383–1394 (1996) https://doi.org/10.1287/mnsc.42.10.1383

[59] Jackson, J.R.: Jobshop-like queueing systems. Management Science **10**(1), 131–142 (1963) https://doi.org/10.1287/mnsc.10.1.131