The restricted Erlang-R queue

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

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

Johan S.H. van Leeuwaarden*   Britt W.J. Mathijsen*   Fiona Sloothaak*

Galit B. Yom-Tov †

December 22, 2016

Abstract

Motivated by health care systems with repeated services that have both personnel (nurse and physician) and space (beds) constraints, we study a restricted version of the Erlang-R model. The space restriction policies we consider are blocking or holding in a pre-entrant queue. We develop many-server approximations for the system performance measures when either policy applies, and explore the connection between them. We show that capacity allocation of both resources should be determined simultaneously, and derive the methodology to determine it explicitly. We show that the system dynamics is captured by the fraction of needy time in the network, and that returning customers should be accounted for both in steady-state and time-varying conditions. We demonstrate the application of our policies in two case studies of resource allocation in hospitals.

Keywords: Many-server approximations, QED regime, fixed-point analysis, returns, service systems, queueing models, health care operations, nurse staffing, beds allocation

1 Introduction

Because service systems are stochastic in nature, it is common practice to use queueing theory for performance analyses and workforce planning. Traditionally, systems are modeled after 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, and fluid and diffusion approximations are used to provide insights into the process dynamics. However, single station models often fail to capture the more intricate dynamics of service networks. In a health care context, think of flows of patients in a hospital from one medical ward to another [2], within the Emergency Department (ED) between different stages of treatment [20], or between medical facilities [40]. Queueing networks can capture the dependency between several service stages and several resources. More specifically, we are interested in the ubiquitous feature, particularly present in health care environments, that customers during their stay in the system might require a specific resource multiple times, e.g. physicians and nurses who treat patients several times during their stay in the medical wards [25] or the ED [38], while multiple resources are limited (e.g. physicians, nurses and beds).

An often ignored yet essential feature of service networks concerns the restriction on the number of customers that can reside in the system simultaneously. Call centers, for instance, only have a finite number of trunk lines [28], whereas in health care facilities, the number of beds poses a constraint on the number of patients that can be admitted at the same time to a medical unit. In such situations, two or more resources (personnel, trunk lines, beds) are restricted. In this

---

*Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands ({j.s.h.v.leeuwaarden,b.w.j.mathijsen,f.sloothaak}@tue.nl)

†Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, 32000 Haifa, Israel, (gality@ie.technion.ac.il)

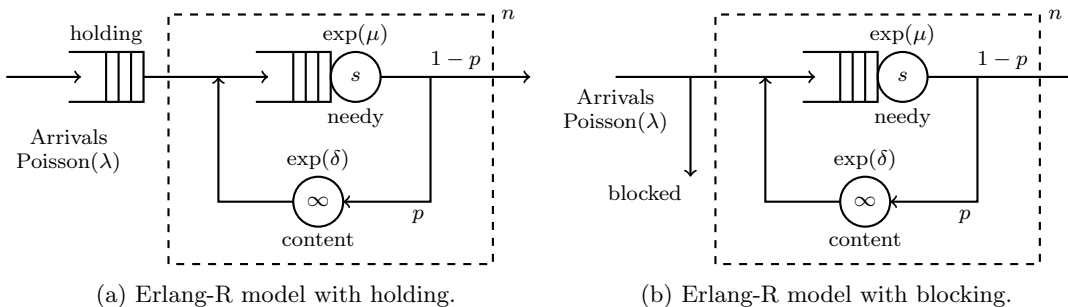(a) Erlang-R model with holding.　　　(b) Erlang-R model with blocking.

Figure 1: Restricted Erlang-R models with maximally n customers in system.

paper, we investigate the influence of such multiple restrictions on the network dynamics and the required staffing policies.

The restricted Erlang-R model. The canonical model for service networks with returns is the Erlang-R model [38] in which customers, during their stay in the system receive a random number of services from the same pool of servers. Yom-Tov & Mandelbaum [38] showed that such a simple network model can be used to determine staffing in an Israeli ED both in stable and time-varying conditions. Nevertheless, empirical studies report that some countries, such as the US, use a different operational mode that apply strict restrictions on entering the ED [33]. In typical US EDs, a patient will not enter the ED until both a bed and a physician are available to treat her. Those restrictions can be either physical (beds) restrictions or managerial ones—for instance by imposing a patient-to-physician ratio. In this work, we extend the Erlang-R model by enforcing a constraint on the maximum number of available places inside the service facility. Our model hence incorporates two kinds of resource constraints: servers that provide the actual service and the maximum available places inside the service system. Both affect the system in a highly interdependent way. The model, presented in Figure 1, assumes s servers and a maximum capacity of n concurrent places. We assume that customers arrive according to a Poisson process with rate $\lambda(t)$. In case a new arrival finds n or more customers already present, we consider two options—either she waits outside the service facility in a holding queue until a vacant space becomes available (Figure 1a) or she is blocked (Figure 1b), such as is the case when patients are sent to an alternative facility. Once customers are admitted, they require assistance from one of the s servers for an exponentially distributed duration with mean $1/\mu$. Then, with probability $1 - p$, customers leave the system or, with probability p, return to service again after an exponentially distributed time with mean $1/\delta$. Following Jennings & de Véricourt [25] and Yom-Tov & Mandelbaum [38], we call patients *needy* when they require attention from one of the servers and *content* when they are in the delayed return phase. In addition, we call customers *holding* when they are waiting outside the facility for an available space. We assume that the arrival process, the needy times and content times are mutually independent. In the holding queue and the needy queue, we apply the First-Come-First-Served (FCFS) discipline.

As mentioned, we consider two versions of the finite-capacity constraint. The first version is called *Erlang-R with holding*, in which customers wait for an available space in the system. The second version is called *Erlang-R with blocking*, in which customers meeting a full system are blocked. Naturally, intermediate scenarios can be constructed in which a proportion of the total arrival volume of customers indeed leaves upon finding a full system, while the rest joins the holding room. While this paper focuses on the two extreme cases, straightforward adaptions can fit these intermediate scenarios.

Examples of restricted Erlang-R. As noted before, an ED operated in the US can be modeled using a restricted Erlang-R model. Another health care example are Medical Units (MU) in a hospital. Such units specialize in specific types of illnesses (cadriatric, oncology, etc.) and have

limited resources such as nurses and beds. If the unit is full, new patients are either allocated to an alternative medical unit, i.e. blocked, or wait for an available bed. Both policies are problematic in terms of quality-of-care, because the personnel in the alternative unit (or the ED) may be less knowledgable and waiting in the ED was shown to increase mortality. Moreover, ED waiting may reduce available capacity for treating ED patients [8, 4], hence endangering both the delayed patient as well as others. Both the number of personnel (nurses and physicians) and the number of beds impact service dynamics and quality-of-care. Research so far looked at the capacity allocation of those resources separately. Green & Yankovic [14] and Jennings & de Véricourt [24] looked at nurse staffing in medical units, while Bekker et al. [10] looked into bed allocation. The unified model we suggest enables us to capture the dependency between those two decisions, and its impact on other medical units in the hospital. At the same time, we capture the two most commonly used modes of operation—blocking and holding of new patients.

1.1 Contributions

Our main goal is to provide staffing policies for the restricted Erlang-R models that ensures high resource utilization, while at the same time maintains a 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. In this paper, we obtain asymptotic results for the Erlang-R model with blocking in the QED regime (§4.2). Following [24], 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 number of customers in the system, where β and γ are constants, R_1 is the offered load of the servers and r is the fraction of time a customer spends in the needy state. We establish limiting expressions for performance measures, such as the probability of delay and blocking, in the form of explicit functions that depend solely on β and γ. In deriving these limit results, we use the available product-form solution for the stationary distribution.

Likewise, we pursue QED performance for the Erlang-R model with holding. However, a direct analytic approach is obstructed by the absence of product-form solutions. We provide two solutions for establishing QED behavior. First, we provide stochastic performance bounds that stay meaningful in the QED regime (§3.3), which demonstrate 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 is based on the following unique approach: initially blocked patients in the blocking model are seen as if they reattempt to get access after a some delay. The behavior of this retrial model then resembles the Erlang-R model with blocking without retrials, yet with an increased arrival rate. The increase in arrival rate turns out to be the solution of a fixed-point equation. Using our results on the asymptotic behavior of the model with blocking in the QED regime, we then obtain approximative QED performance measures for the model with holding. Finally, we use these QED results to develop algorithms for dimensioning and time-varying staffing (§5.1).

Using the approximations developed, numerical analysis and simulation we provide the following managerial insights:

- We show that all resource allocation of personnel and beds should be synchronized in order to avoid waste, and that the QED scaling provides an efficient, flexible, and easy to implement methodology to do so.

- We conclude that *reentrant* customers in a restricted network are more significant than in an open network (Erlang-R). In contrast to the open model, in which returning customers need to be accounted for only in time-varying systems, the restricted Erlang-R model requires explicit consideration of returning customers under stationary conditions as well.

- We show that the influence of the network structure on the system dynamics crucially depend on the fraction of time a patient spends being needy during her stay in the system. We then explore the influence of r on operational decisions.

3

- Combining the theoretical results, we explore the implication of managerial decisions in designing an MU (§5) and an ED (§6.4). In §5, we show that enabling customers to hold in ED before entering an MU, requires more resources both in the MU and the ED. In Section 6.4, we compare the pros and cons of imposing strict constraints on entering an ED. We find that size restrictions have the ability to improve the quality-of-service of the processes within the facility, at the expense of a slight increase in pre-entrant wait and server efficiency levels.

- Finally, we show that restricting the number of admitted customers protects those customers with complicated demand consisting of relatively many retrials/interruptions (§6.3).

When dealing with time-varying patient demands, one requires some modifications in the QED results in order to obtain stable performance at every moment in time. We transform the two-fold staffing policy into a time-varying one based on the Modified Offered Load (MOL) method (§6.4). This method approximates the offered load at the needy station at each point in time via a corresponding system with ample resources. By noting that the latter system coincides with the Erlang-R model with ample servers, we adopt the offered load approximations given in [38]. Numerical experiments justify this method and we use this approach in our case study §6.4.

2 Literature review

Due to increasing demand and tightening budgets in health care, there is a growing need for efficient workforce management [16]. Personnel (nurse and physician) expenditure is one of the biggest factors in hospital costs [27], and inadequate nursing levels have been mentioned as a significant factor in medical errors and ED overcrowding. In order to establish appropriate nursing levels, a staffing policy requires assessment of a wide range of variables, such as differing nurse expertise and patient acuity during the day. Current methods, such as the minimum nurse-to-patient ratios, are often too inflexible to capture those varying conditions. The American Hospital Association (AHA) and others call for dynamic staffing policies that can deal with the complex and evolving nature of health care [1]. Workforce management in health care systems has been studied extensively; see [11, 18, 19] for overviews. In recent years it has become apparent that queueing models can be helpful in developing staffing and routing recommendations, not just for large-scale service systems, but also for the small and complicated health care systems.

The first to try such an approach through queueing models were Green & Savin [15, 16] who used the single station stationary Erlang-C model to set staffing levels in EDs and panel sizes for clinics. Using a similar approach, [3] used Erlang-B model to determine bed allocation for medical wards. The first to observe the significant impact of interrupted services in a health care setting were Jennings & de Véricourt [24, 25]. Motivated by the need to set nurse-to-patient ratios for internal wards, they considered a closed queueing system with s nurses and n beds, whuch we will refer to as the *closed ward* model. This is essentially the Erlang-C model with the additional restriction that a finite population of the n patients requires care. In their model, all beds are always occupied, and patients alternate between two phases: the needy phase where patients require service of a nurse and the content phase where they do not; see Figure 2a. The system dynamics of restricted Erlang-R model are equivalent to those of the closed ward model of [24] if the holding queue would never be empty.

Campello et al. [7] analyzed a similar operational decision, referred to as ED case management, which determines the maximal number of patients a physician should handle in parallel. They also used queueing networks and analyzed the stationary distribution. Note that in practice such decision is not only affected by operational measurements such as waiting times, but also by psychological constraints that limit physician capability to manage multiple tasks (patients) in parallel. Diwas [12] provided empirical evidence that physicians should not treat more than 6-7 patients at the same time. Therefore, many hospitals in the US restrict entrance to EDs even if beds are available if physicians are overloaded. We too consider such constraints, and analyze their impact on performance. We take a different approach than [7]; instead of analyzing numerically

(a) The closed ward model.
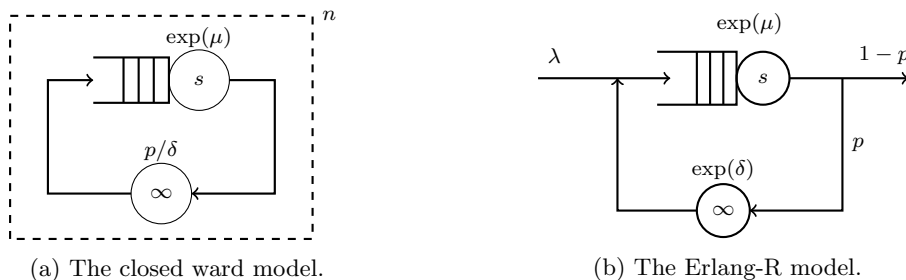
(b) The Erlang-R model.

Figure 2: Two basic models with interrupted services.

steady-state distributions, we develop many-server approximations that can produce insight into the system dynamics, and can be incorporated into time-varying staffing procedures (see §6.4).

The model in [24, 25] was developed for modeling internal dynamics within an internal ward. However, in the ED, beds are not constantly occupied and the utilization level depends on the flow of patients that arrive from outside the system. Yom-Tov & Mandelbaum [38] highlight the interrupted services while accounting for the transient nature of patient's arrival process, and introduced the Erlang-R model as a model for an ED. The Erlang-R model is an open two-station queueing network that has the same layout as the restricted Erlang-R model, except that all patients find a bed available upon arrival, see Figure 2b. In both models patients experience the interrupted services, but the Erlang-R model has no further restrictions on the bed capacity, hence neglecting the finite-size effects. Yom-Tov & Mandelbaum [38] showed, using a simulator tailored to an Israeli ED, 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. Although the feature of interrupted services is present in many systems, it is particularly important for modeling EDs, because the duration of the interruption is typically much longer than the time patients require care from a nurse. This explains why the Erlang-R model is considered to be the canonical model for EDs. The restricted Erlang-R model with holding/blocking thus extends the Erlang-R model with finite-size constraints which, like interrupted services, are expected to have a decisive impact on 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 Erlang-C model. This system can be characterized completely by the staffing level s and the offered load $R = \lambda/\mu$, which is the average workload pressure per time unit on the system. Moreover, let $\rho = R/s$ denote the server utilization level. Since exact analysis provides little qualitative insight in performance with respect to the system parameters, we resort to asymptotic analysis. This in turn provides approximations for the true system behavior. In the QED regime [6, 17], the utilization level is driven to unity in accordance 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 nurses s exceeds the minimally required offered load R, but only by a relatively small amount $\beta\sqrt{R}$. As s grows large, the probability of delay tends to a non-degenerate function that only depends on the parameter β. This function is strictly decreasing in $\beta > 0$ with range $(0, 1)$. Consequently, any targeted delay probability can be achieved by adjusting β. Moreover, the mean delay is of order $1/\sqrt{s}$ and hence asymptotically negligible. In this paper, we take the same approach, but determine capacity for nurses and beds simultaneously in such a way that both the probability to wait for a nurse and the probability to wait for a bed are non-degenerate. Moreover, the utilization of both nurses and beds goes to unity as the size of the system increases. While the QED regime gives precise limits when the system size (s and n in our case) goes to infinity, it is by now well known that the asymptotic behavior kicks in quickly, so that QED limits serve as sharp approximations, already for small systems. See [23, 39, 34, 35, 13, 32] for various works that provide theoretical support for this fast relaxation.
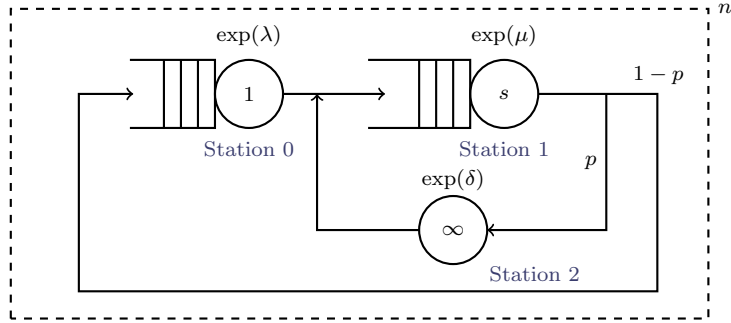
5

Figure 3: The Erlang-R model with blocking, viewed as a closed Jackson network.

In recent years, a new approach for developing approximations of performance measures of queueing systems with finite-size constraints in the QED regime was suggested by van Leeuwaarden et al. [36, 37]. This approach characterizes the asymptotic dynamics of these analytically intractable systems with finite-size restrictions through more tractable ones. We adopt this approach for developing the approximations for the holding model in §4.3.

3 Models and performance measures

3.1 Three-dimensional Markov process

Since in the restricted Erlang-R model described the arrival process is taken Poisson, and all service and content times are assumed independent and exponential, 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 maximum number of patients admitted to 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 can be interpreted as the offered workload brought towards the needy queue, and the content (infinite-server) queue, respectively. Furthermore, we define

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

which is the fraction of time a patient spends in the needy state (in case she experienced no wait during her sojourn).

3.1.1 Erlang-R model with blocking.

In case of the blocking model, $Q(t)$ 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 in Figure 3 is an $M/M/s$ queue with service rate μ, modeling the number of needy patients, $Q_1(t)$. Station 2 models the number of content patients, $Q_2(t)$, and can therefore be represented as an infinite-server queue with service rate δ. A patient can enter the unit 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 in case its queue length is positive, i.e. if $n - Q_1(t) - Q_2(t) > 0$.

6

Observe that because patients finding a full network are blocked, the number of patients in the system cannot grow beyond n. Hence, the system is stable for all parameter settings, and hence 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. Let $\pi_b(j, k)$ denote the steady-state probabilities of having j needy and k content patients 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 Erlang-R with holding.

The Erlang-R model with holding 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 patients in the system (including the holding queue). Using the restriction $Q_1(t) + Q_2(t) \leq n$ together with the fact that no bed is left vacant if a patient is waiting in the holding queue, this yields

$$H(t) = (N(t) - n)^+, \quad t \geq 0,$$

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 model with holding 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

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

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

Contrary to the model with blocking, the system with holding *can* become unstable in case capacity is insufficient to satisfy patient demand.

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

$$\frac{\lambda}{(1-p)\mu s} < \frac{\sum_{i=0}^s \frac{i}{s}\binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^n \binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i}{\sum_{i=0}^s \binom{n}{i}\left(\frac{\delta}{p\mu}\right)^i + \sum_{i=s+1}^n \binom{n}{i}\frac{i!}{s!}s^{s-i}\left(\frac{\delta}{p\mu}\right)^i} =: \rho_{\max}(s, n). \tag{4}$$

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.

(a) $R_{\max}$ as a function of s.   (b) $R_{\max}$ as a function of n.
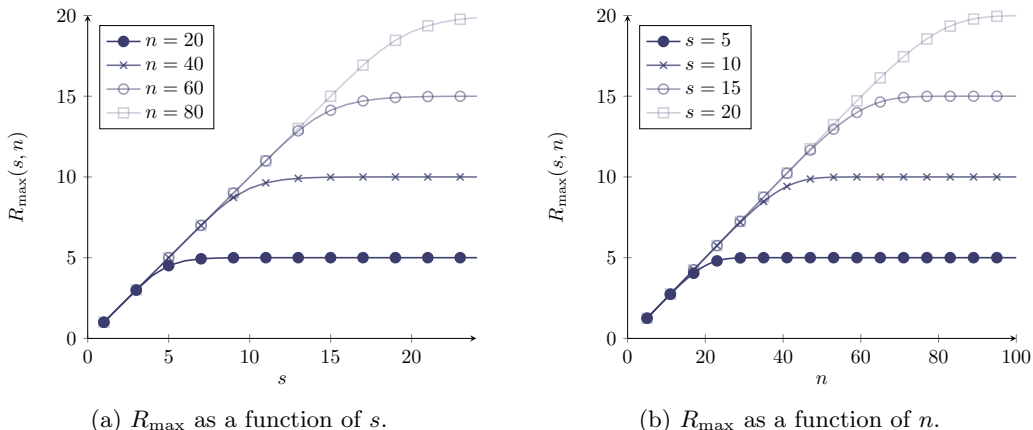
Figure 4: The maximum achievable workload in the restricted Erlang-R model with holding for $r = 0.25$.

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. Figure 4 shows the influence of both s and n on the maximum feasible workload in case $r = 0.25$. 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 s fixed. A logical practical consequence is that a larger number of beds allows for a larger patient volume to enter the ED with the same number of nurses. Moreover, $R_{\max}(s,n)$ is increasing in s, but as in Figure 4a, adding an extra nurse does not increase the stability region in case n is too tight. Conversely, adding extra beds does not increase $R_{\max}(s,n)$ if the number of 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 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 synchronized QED capacity recommendation for both servers and beds in §4.

Managerial insight 1. Both the number of nurses s and the number of beds n play a restricting role on the maximum demand the system can handle. Therefore, they should be balanced together.

Provided that the system is stable, the stationary distribution of the QBD process X can be obtained numerically by the matrix geometric method [31]. 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) at entering the system, the probability of delay at the needy queue, expected waiting time for a nurse, utilization of nurses and utilization of beds:

$$\mathbb{P}(\text{hold}) = \sum_{i=0}^{\infty} \sum_{j=0}^{n} \pi(i, j, n-j), \qquad \mathbb{P}(\text{delay}) \approx \sum_{i=0}^{\infty} \sum_{j=s}^{n} \sum_{k=0}^{n-j} \pi(i, j, k), \qquad (5)$$

8

$$\mathbb{E}[W] \approx \sum_{i=0}^{\infty} \sum_{j=s}^{n} \sum_{k=0}^{n-j} \frac{\max\{0, j-s+1\}}{\mu} \pi(i,j,k), \tag{6}$$

$$\rho_s = \frac{1}{s} \sum_{i=0}^{\infty} \sum_{j=0}^{n} \sum_{k=0}^{n-j} \min\{j,s\} \pi(i,j,k), \qquad \rho_n = \frac{1}{n} \sum_{i=0}^{\infty} \sum_{j=0}^{n} \sum_{k=0}^{n-j} \min\{i,n\} \pi(i,j,k). \tag{7}$$

It should be stressed that the above expression for the delay probability and the expected waiting time for a nurse is not exact. For the blocking model one can use the Arrival Theorem [9], 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 content patients 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 patients in the ED, in the following result.

Proposition 2. *Let Q_1^b, Q_2^b, Q_1^h, Q_2^h denote the nurse 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 $Q_2^b(0) = Q_2^b(0)$. For all $t \geq 0$,*

$$Q_1^b(t) + Q_2^b(t) \preceq_{\mathrm{st}} Q_1^h(t) + Q_2^h(t) \preceq_{\mathrm{st}} n, \tag{8}$$

$$Q_2^b(t) \preceq_{\mathrm{st}} Q_2^h(t), \tag{9}$$

$$Q_1^b(t) \preceq_{\mathrm{st}} Q_1^h(t) + H(t), \tag{10}$$

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

The proof of Proposition 2 uses sample path coupling and can be found in Appendix B. Note that as an immediate consequence, we have

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

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

Managerial insight 2. Under similar offered load and capacity constraints, utilization levels for the nurses in the Erlang-R model with blocking are lower than in the Erlang-R model with holding. Moreover, the total number of waiting patients 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

We do not want to waste capacity of either servers or beds without getting significant advantage in term of performance. 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 system behavior, i.e. high occupancy levels of both nurses and beds 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 [24] and [28], 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, let us 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 amount of works.

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 interrupted services. It follows that the appropriate staffing rule for the nurses in the QED regime remains $s = R_1 + \beta\sqrt{R_1}$ for some constant $\beta > 0$.

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

$$
\begin{aligned}
s &= R_1 + \beta\sqrt{R_1} + o(\sqrt{R_1}) \\
n &= \frac{R_1}{r} + \gamma\sqrt{\frac{R_1}{r}} + o(\sqrt{R_1})
\end{aligned}
\tag{11}
$$

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

Recall that we saw in Figure 4 that resources seem efficiently utilized if $s/n \approx r$. Scaling (11) 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 [24], a similar scaling regime is considered, 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 (11) 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 (11) indeed leads to 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 patient becomes negligible as $R_1 \to \infty$. From these empirical findings we deduce that under scaling (11) the restricted Erlang-R model exhibits QED behavior on two levels: Outside the facility while waiting for an available bed, and inside the facility while waiting for attention of a nurse.

We also check how the Erlang-R model with blocking or holding and the closed ward model of [24] relate under scaling (11). 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 (11), and the delay probability of the model with holding lies in between the other two. Second, note that the expected waiting time for a nurse 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

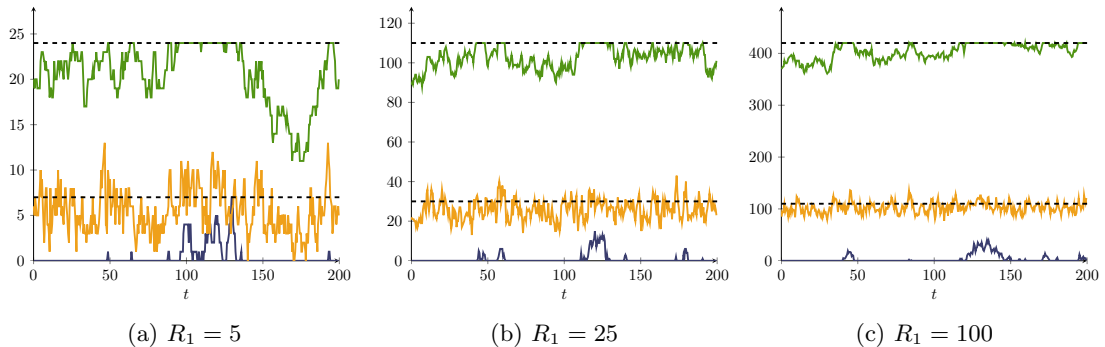(a) $R_1 = 5$ (b) $R_1 = 25$ (c) $R_1 = 100$

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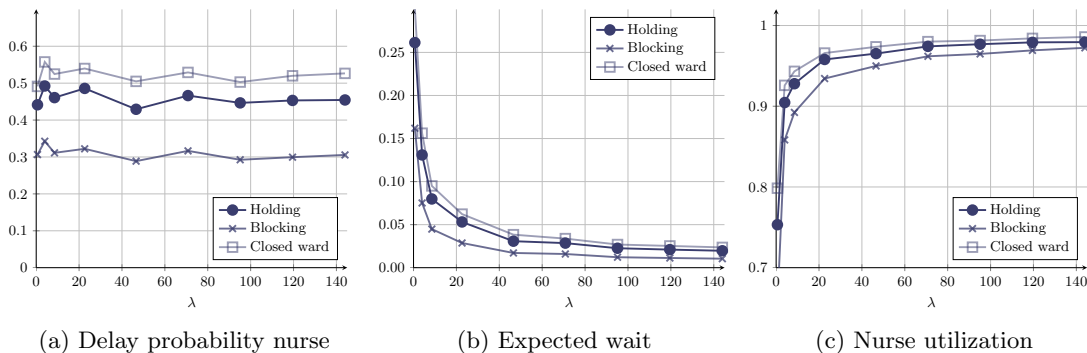
Figure 6: Asymptotic behavior of the restricted Erlang-R model with holding and blocking, and the Closed ward model for $\mu = 1$, $\delta = 0.2$, $p = 0.8$ and $\beta = \gamma = 0.5$.

rate of convergence seems again comparable. Finally, and most importantly, we notice an ordering between the three models. Namely, in all performance measures 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 patients in the Erlang-R model with blocking, holding and the closed ward, respectively. Then,

$$Q_1^b(\infty) \preceq_{\mathrm{st}} Q_1^h(\infty) \preceq_{\mathrm{st}} Q_1^c(\infty). \tag{12}$$

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 (11). 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 performance measures defined in §3.2 in terms of β and γ.

11

Theorem 2. *Let s and n scale as in (11) with $-\infty < \beta < \infty$, $\gamma > 0$ as $\lambda \to \infty$. Then, if $\beta \neq 0$,*

$$g^b(\beta, \gamma) := \lim_{\lambda \to \infty} \mathbb{P}^b(\text{delay}) = \left(1 + \frac{\beta \int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\beta^2 + \eta^2})e^{\frac{1}{2}\omega^2}\Phi(\omega)}\right)^{-1}, \qquad (13)$$

$$f^b(\beta, \gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block}) = \frac{\sqrt{r}\phi(\gamma)\Phi(-\omega\sqrt{r}) + \phi(\sqrt{\beta^2 + \eta^2})e^{\frac{1}{2}\omega^2}\Phi(\omega)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)}, \qquad (14)$$

$$h^b(\beta, \gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{E}[W] = \frac{\frac{\phi(\beta)\Phi(\eta)}{\beta^2} + \left(\frac{\beta}{r} - \frac{\gamma}{\sqrt{r}} - \frac{1}{\beta}\right)\frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega) - \sqrt{\frac{1-r}{r}}\frac{\phi(\beta)\phi(\eta)}{\beta}}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)}, \qquad (15)$$

and if $\beta = 0$,

$$g_0^b(\gamma) := \lim_{\lambda \to \infty} \mathbb{P}^b(\text{delay}) = \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta) + \phi(\eta)\right)}\right)^{-1}, \qquad (16)$$

$$f_0^b(\gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block}) = \frac{\sqrt{r}\,\phi(\gamma)\Phi(-\omega\sqrt{r}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta) + \phi(\eta)\right)}, \qquad (17)$$

$$h_0^b(\gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{E}[W] = \frac{1}{2\mu} \frac{\left(\gamma^2/r + 1\right)\Phi(\eta) + \eta\phi(\eta)}{\frac{r}{1-r}\sqrt{2\pi}\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{r}{1-r}}\left(\eta\Phi(\eta) + \phi(\eta)\right)}, \qquad (18)$$

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

The proof of Theorem 2 can be found in Appendix C.

Theorem 2 proves that the scaling (11) results in QED behavior: the probability of waiting in Equations (13) and (16) converges to a limit that is strictly between 0 and 1. Notice that all limits in Theorem 2 are functions of three parameters: β and γ, which are decision variables, and the fraction of needy time r, which is dictated by the physics of the system.

Theorem 2 also shows that the probability of blocking (Equations (14) and (17)) 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 immediately to be 40%. At the same time, the probability of getting immediately a bed is 97%. Thus, our QED policy puts more emphasis on preventing blocking than waiting.

Theorem 2 further shows that the expected waiting (Equations (15) and (18)) is of order $1/\sqrt{R_1}$ too and hence vanishes in the large-system limit.

We see from Theorem 2 that achieving target service levels is always an interplay between 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 (11) automatically captures the right dimensioning ratio as the system scales up, Theorem 2 shows that the parameters β and γ provide a means to fine-tune the performance. Figure 7b confirms how adding nurses, i.e. increasing β, does not improve the blocking probability if the number of 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 2 as approximations in a realistic setting, we plot in Figure 8 the exact probability of delay and blocking for an Erlang-R model with
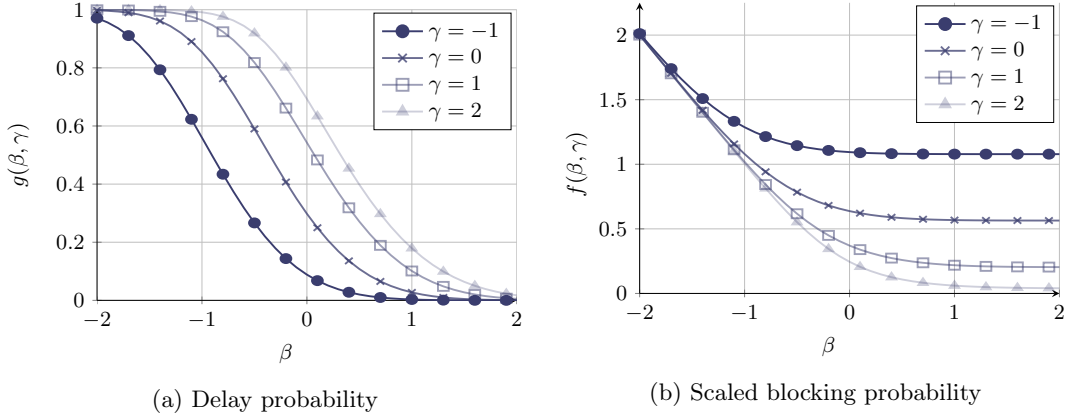
(a) Delay probability

(b) Scaled blocking probability

Figure 7: Asymptotic delay and scaled blocking probability for $r = 0.5$ based on Theorem 2.

(a) Delay probability
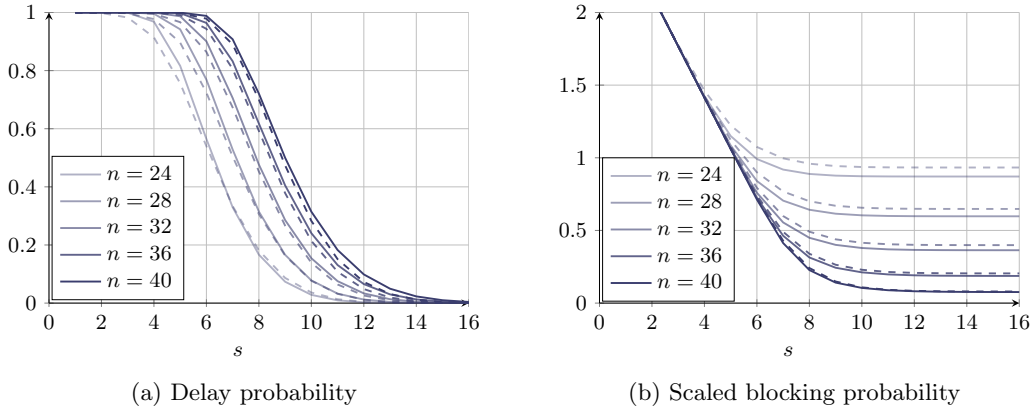
(b) Scaled blocking probability

Figure 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 $p = 0.75$.

$R = 8$ and $r = 0.25$, as a function of s. The exact probabilities are given by Equation (5), and their respective asymptotic approximations are based on Theorem 2. Despite the realistic moderate size of the system ($R = 8$), 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 [6, 23].

In Appendix E.1, we furthermore compare the asymptotic delay and blocking probability in three additional scenarios. In Tables 2–4 we compute the exact probabilities of delay and blocking through the explicit forms in (5) for various values of the offered-load R_1, which are omitted here due to space constraints. 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. Naturally, fluctuations occur for relatively small values of R_1, because s and n need to be rounded to an integer.

4.3 QED limits for Erlang-R with holding

As explained in §4, 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 (11), the upper bound on the utilization level of the nurses

needed to achieve stability in the model with holding, as given in Proposition 1, converges to unity as $R \to \infty$. 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* (11). *Then, for $\lambda \to \infty$,*

$$\rho_{\max}(s, n) \to 1.$$

The proof can be found in Appendix D. Combining Proposition 3 with Proposition 1 shows that indeed the scaling we use results in a highly utilized system.

As observed before, the nature of the two variants of the model is similar up to the fact that a fraction of the patients is deferred on arrive 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 room. In fact, Theorem 2 shows that the blocking probability is of order $1/\sqrt{R_1}$, yielding a volume of blocked patients of order $\sqrt{R_1}$ in setting with blocking. Accordingly, if $R^b = R_1$ and R^h denote the nominal load arriving to the nurses in the model with blocking and holding, respectively, we can argue that

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

for some $\alpha > 0$. Notice that this additional load is of the same order as the safety staffing in the blocking model staffing rule (11). As s and n remain unchanged, we rewrite (11) in terms of R^h,

$$s = R^h + (\beta - \alpha)\sqrt{R^h} + o(\sqrt{R^h}),$$
$$n = \frac{R^h}{r} + \left(\gamma - \alpha/\sqrt{r}\right)\sqrt{\frac{R^h}{r}} + o(\sqrt{R^h}), \tag{19}$$

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 (11) 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 patients 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 (19), this volume asymptotically equals $R^h \cdot \mathbb{P}^b(\text{block})$, with parameters $\beta - \alpha$ and $\gamma - \alpha/\sqrt{r}$, which by Theorem 2 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

$$\alpha = f^h\left(\beta - \alpha, \gamma - \alpha/\sqrt{r}\right), \tag{20}$$

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

$$\mathbb{P}^h(\text{delay}) \approx g^b(\beta - \alpha, \gamma - \alpha/\sqrt{r}) =: g^h(\beta, \gamma). \tag{21}$$

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

$$\sqrt{R_1} \cdot \mathbb{E}[W] \approx h^b(\beta - \alpha, \gamma - \alpha/\sqrt{r}) =: h^h(\beta, \gamma). \tag{22}$$

This also implies that the holding queue is $O(\sqrt{R_1})$. Subsequently, we argue that the expected holding time (pre-entering wait) under the QED policy is $O(1/\sqrt{R_1})$ and hence asymptotically negligible. We justify this claim numerically in §6.

Remark 3. Notice that in the reasoning leading to (20), we implicitly assumed that the additional volume $\alpha\sqrt{R_1}$ is an independent Poisson process, which is obviously not the case. Therefore, (21)–(22) are approximations for pre-limit systems that are not asymptotically correct as $\lambda \to \infty$. Nevertheless, our heuristic approach seems to performs well as we confirm numerically next.

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 decision. 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.

To see how this heuristic approach performs under different settings, and particularly if $R_1 \to \infty$, we compare in Appendix E.2 the approximated delay probability in the Erlang-R model with holding as solution of the fixed-point procedure to the outcomes of simulation experiments. We omit the tables here due to space constraints. We conclude from these tables 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.

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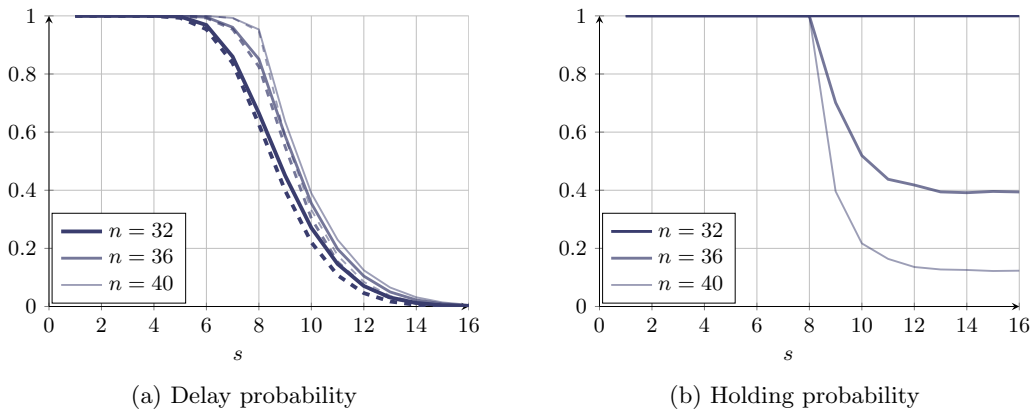
Figure 9: Comparison of simulated delay probability (solid) against asymptotic approximations (dashed) with $\beta = (s - R_1)/\sqrt{R_1}$ and $\gamma = (n - R_1/r)/\sqrt{R_1/r}$ for $\lambda = 2$, $\mu = 1$, $\delta = 0.25$ and $p = 0.75$.

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 nurses s and the number of beds 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 at the pre-entrant queue 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 2 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 number of beds and 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.

15

We illustrate the resource allocation decisions in an MU setting, using data originated from two articles by Lundgren & Segesten [29] and Green & Yankovic [14]. Green & Yankovic describe an MU that has 42 beds, with average occupancy level 78%, and Average Length of Stay (ALOS) 4.3 days. Lundgren & 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 is 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 parameters 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 nominal offered load $R_1 = 3.2$ and $R_1/r = 34.4$.
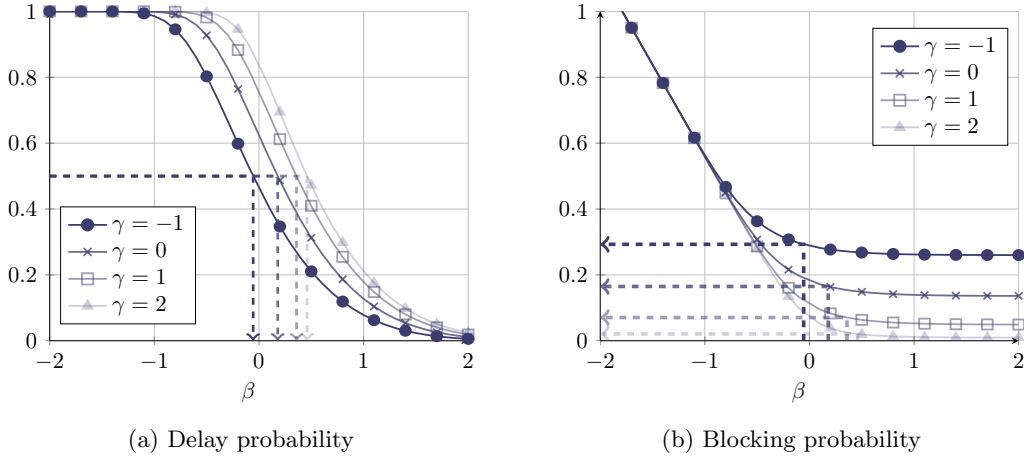

(a) Delay probability (b) Blocking probability

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

Figure 10 visualizes the dimensioning procedure for this particular MU. The hospital management can find a pair of n and s to meet certain criteria, for example 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. According 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 probability of blocking of more than 10 percent 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 $n_i = \lceil R_1/r + \gamma \sqrt{R_1/r} \rceil$, 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 of number of beds.

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 (20). We propose the following dimensioning procedure to achieve a preset target delay probability at the needy queue.

Algorithm 1. STATIONARY DIMENSIONING ALGORITHM.
Input: Target delay probability ε. Parameters λ, μ, δ and p.
Output: Staffing levels s and n.

1. Set $R_1 := \frac{\lambda}{(1-p)\mu}$ and $r = \frac{\delta}{\delta + p\mu}$.
2. Determine parameters (β^*, γ^*) such that $g^b(\beta^*, \gamma^*) = \varepsilon$.

16

3. Set $\beta = \beta^* + f^b(\beta^*, \gamma^*)$ and $\gamma = \gamma^* + f^b(\beta^*, \gamma^*)/\sqrt{r}$.
4. Return $s = \lceil R_1 + \beta\sqrt{R_1} \rceil$ and $n = \lfloor R_1/r + \gamma\sqrt{R_1/r} \rfloor$.

Remark 4. In Step 2 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

$$g_{\mathrm{HW}}(\beta^*) = \left(1 + \frac{\beta^* \Phi(\beta^*)}{\phi(\beta^*)} \right)^{-1}.$$

Hence, if $\varepsilon > g_{\mathrm{HW}}(\beta^*)$, then no feasible solution to $g^b(\beta^*, \gamma^*) = \varepsilon$ exists. This should be considered when choosing β^*. Furthermore, it is evident from Step 3 that the final values (β, γ) are always larger than (β^*, γ^*).

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 workforce in that unit (as well as the ED). This example also shows that the facility with holding room is able to treat fewer patients simultaneously than under blocking constraints, in line with the bounds in §3.3 and Conjecture 1.
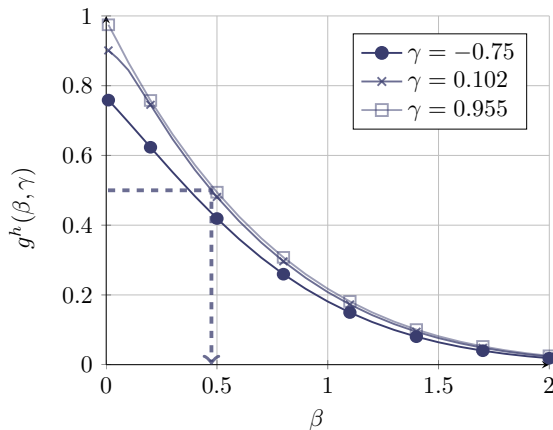
Figure 11: Approximate delay probability of restricted Erlang-R system with holding for $r \approx 0.09$ and $R_1 = 3.2$

6 Model analysis and managerial 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.
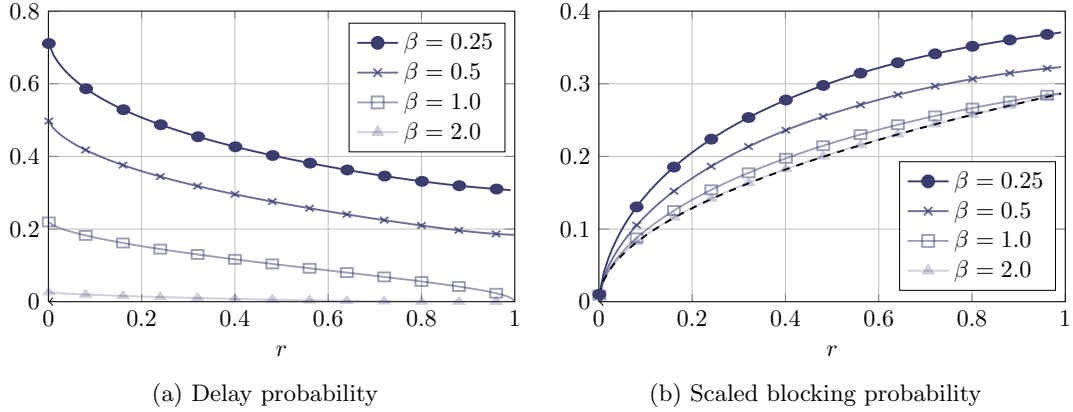
(a) Delay probability (b) Scaled blocking probability

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

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 2.

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, i.e. for $r \in (0,1)$. The case $r = 1$ corresponds to the setting in which patients are needy all the time, in this case customers get service in one time. 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

$$s = R_1 + \beta\sqrt{R_1}, \qquad n = R_1 + \gamma\sqrt{R_1}.$$

This staffing rule only makes sense in case $\beta < \gamma$, since no delay is experienced if $n \leq s$. If indeed $\gamma > \beta$, then the asymptotic delay probability and scaled blocking probability are given by [30],

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

We can see that $f^b(\beta, \gamma)$ of increasing β approaches a lower bound that is a function of r. To see this, observe that as β grows, delays at the nurse queue vanish. 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 known to be, see [22],

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

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

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 patients are blocked. Therefore, if customers spend relatively more time in needy state, which usually indicates services that are less interrupted, blocking will increase. Delays, on the other hand, will decrease in such situations—the minimal delay possible can be achieved if service is given in one time ($r = 1$). Returns or interruptions increase delays significantly under QED staffing.
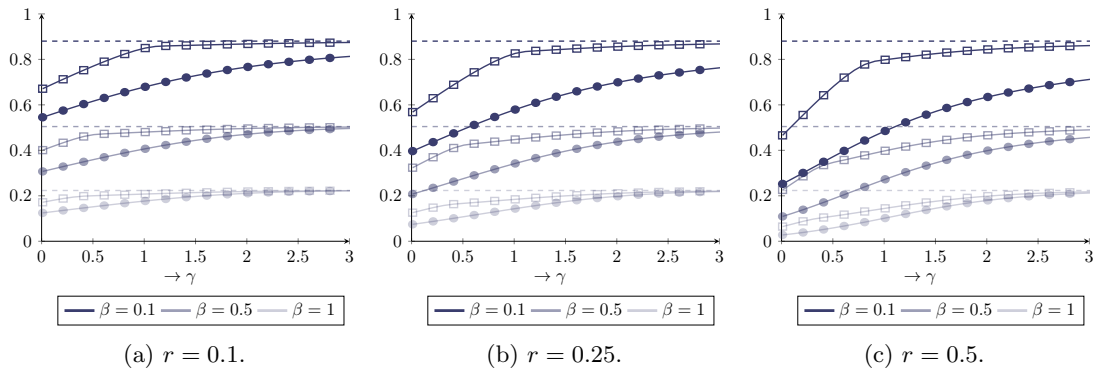
Figure 13: Asymptotic delay probability in open Erlang-R (dashed), restricted Erlang-R with blocking ($\bullet$) and restricted Erlang-R with holding ($\square$), as function of γ.

Managerial insight 3.

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 spend in needy state becomes small, since then the number of customer contributing to the space constrains 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 ($g_{\mathrm{HW}}(\beta)$) models, as functions of γ, while keeping β fixed, for three values of r. We make a couple of observations. Notice that

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

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 conversion is fast—one can provide probability of waiting close to that of the open model with small values of γ. Indeed, the fact that when using QED staffing not much of excessive delay results from the beds restriction is important by itself. Also, we observe that the difference between delay probabilities increases with r.

6.3 The impact of visit number

We next reflect on the impact of operational capacity decisions on different customer populations. We measure patient's complexity by the number of times she needs to see 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 health care environment.

Figure 14 shows the waiting time in the needy and pre-entring 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-entring queue decreases with n, while the needy waiting time increases. For patients who require a relative 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.

Remark 5. From a different perspective, note that in communication queueing systems, the partition of a job to sizable quantities and scheduling those jobs in a similar dynamic to the Erlang-R model became a popular way for increasing throughput. This is because this effectively schedule jobs by their size even though the total job requirements are uncertain. This in fact creates a shortest-job-first policy without prior knowledge of job size [5]. Considering that perspective we note that the Erlang-R model actually prioritize simple jobs over complex ones. But without restrictions, when load is too high, such procedures may lead to very long LOS of long jobs. The capacity restriction we analyze in this paper, in both of its versions, limits such delays. Hence, even in cases in which the returns themselves are created by a managerial decision, imposing the additional managerial restriction on entering the system has benefits.

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

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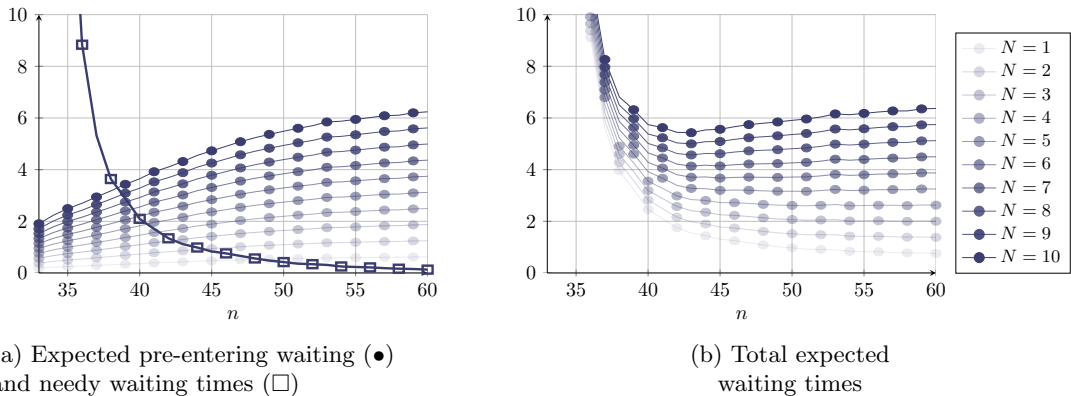
Figure 14: Expected waiting times as a function of n given the number of visits N in the Erlang-R model with holding with $\lambda = 2$ $\mu = 1$, $\delta = 0.25$, $p = 0.75$ and $s = 9$.

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 [38]). In order to set time-varying staffing levels $s(t)$ and $n(t)$, we adopt the MOL approximation of the demand process of [26]. This approach initially presumes infinite capacity to obtain the number of customers $R(t)$ in the queueing system as a function of time. This offered load function then replaces (constant) value of R in the stationary dimensioning scheme under consideration, 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 $R_1(t)$

(a) Dynamic arrival rate and load functions

(b) Capacity functions for $\beta = \gamma = 0.5$
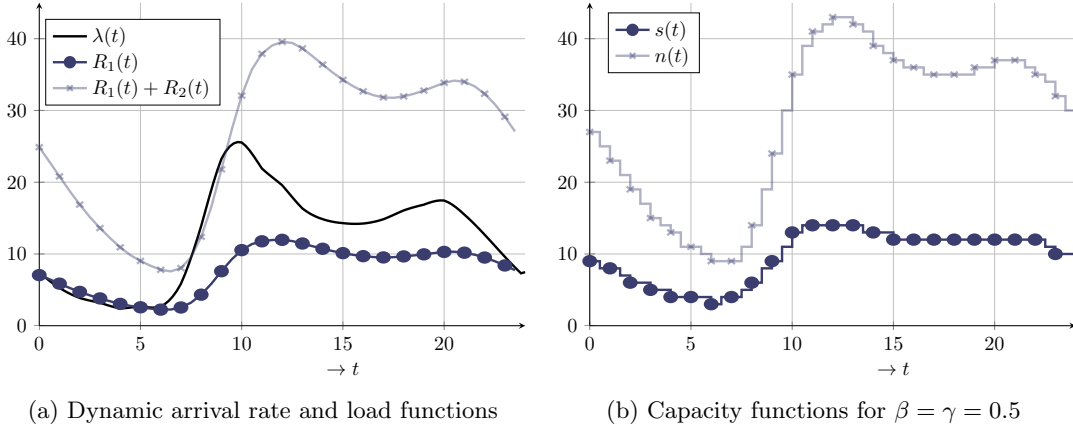
Figure 15: Empirical arrival rate and offered load functions $R_1(t)$ and $R_1(t) + R_2(t)$ in Israeli ED and corresponding capacity functions

and the offered load function for the beds $R_1(t) + R_2(t)$ as the solution of the system of ODEs,

$$\frac{d}{dt}R_1(t) = \lambda(t) + \delta R_2(t) - \mu R_1(t), \tag{23}$$

$$\frac{d}{dt}R_2(t) = p\mu R_1(t) - \delta R_2(t), \tag{24}$$

see [38, 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 more quiet 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 [33, 7]. 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 to recommend an appropriate one.

Extrapolating Algorithm 1 to the time-varying case, Step 4 is replaced by

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

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 these we investigate, using simulation, the differences in the time-varying performance indicators between the policy with blocking and holding.

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 with holding. 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 $O(1/\sqrt{R_1})$.

Finally note that the three parameter settings lead to different nurse-to-patient ratios. Clearly, larger β leads to small nurse-to-patient ratios (due do larger staffing). Figure 16c demonstrates

(a) $\mathbb{P}$(delay)   (b) $\mathbb{P}$(block) or $\mathbb{P}$(hold)   (c) Nurse-to-patient ratio.
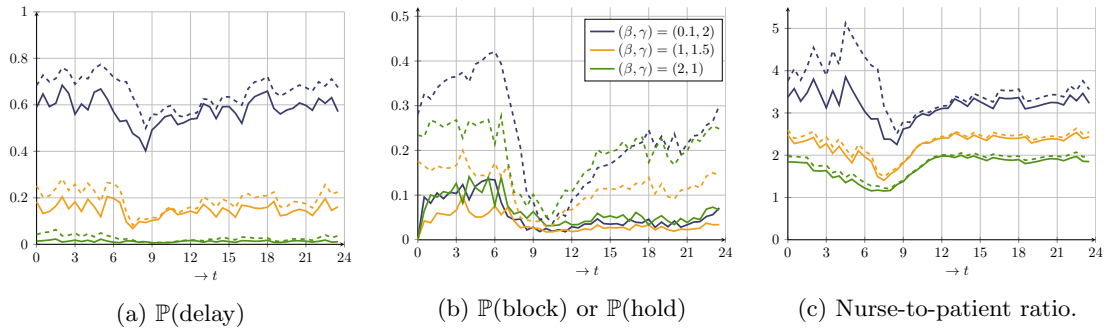
Figure 16: Simulation results for case study. Solid and dashed lines represent time-varying performance in the blocking and holding model, respectively.

(a) Mean holding queue length   (b) Mean service queue length
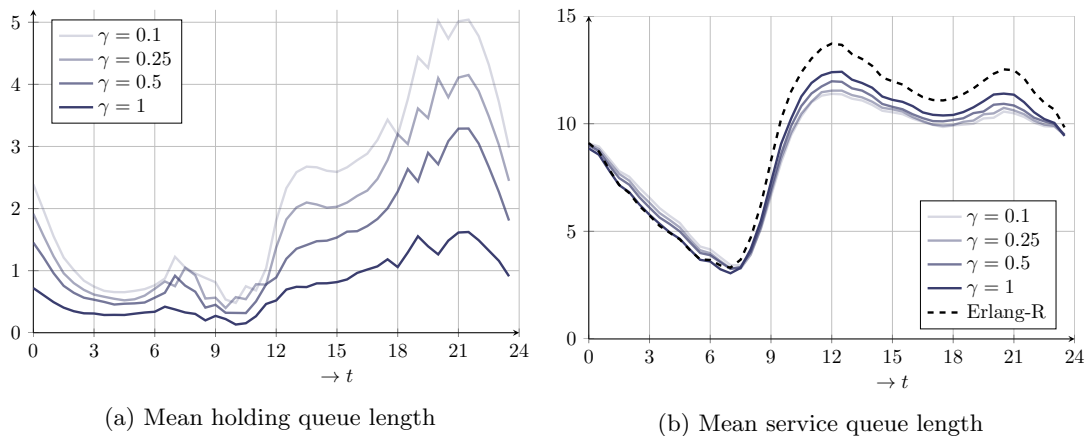
Figure 17: Simulated queue length of holding model with different values of γ.

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 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 latter 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 service queue lengths in the holding model as a function of the parameter γ in Figure 17. Note that for all γ considered, the holding queue length are, as expected, of a smaller order than the number of patients admitted. Also, the holding queue length decreases as we increase γ. The service queue lengths naturally approach the expected queue lengths 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 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.

References

[1] American Hospital Association. Survey of hospital leaders, 2007.

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

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

[4] R. Bennidor and S.H. Israelit. Emergency department intermediate stay unit - a failed model. Technion working paper, 2015.

[5] T. Bonald and C. Comté. Networks of multi-server queues with parallel processing. Unpublished manuscript, 2016.

[6] S.C. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, January-February 2004.

[7] F. Campello, A. Ingolfsson, and R.A. Shumsky. Queueing models of case managers. *Management Science*, 2016. Articles in advance.

[8] R. Carmen and I. van Nieuwenhuyse. How inpatient boarding impacts ED performance: A queueing analysis. Working paper, KU Leuven, 2016.

[9] H. Chen and D.D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization.* Springer, 2001.

[10] A.M. de Bruin, R. Bekker, L. van Zanten, and G.M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2009.

[11] B.T. Denton, editor. *Handbook of healthcare operations management: Methods and applications.* North Holland, New York, 2nd edition, 2013.

[12] KC S. Diwas. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2014.

[13] D. Gamarnik and D.A. Goldberg. 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.

[14] L. Green and N. Yankovic. Identifying good nursing levels: A queuing approach. *Operations Research*, 59(4):942–955, 2011.

[15] L.V. Green. Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine*, 13:61–68, 2006.

[16] L.V. Green and S. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[17] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[18] R.W. Hall, editor. *Patient Flow: Reducing Delay in Healthcare Delivery.* Springer, 2006.

[19] R.W. Hall, editor. *Handbook of Healthcare System Scheduling*. Springer, 2012.

[20] Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

[21] J.R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

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

[23] A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and B. Zwart. Refining square-root safety staffing by expanding Erlang C. *Operations Research*, 59(6):1512–1522, 2011.

[24] O.B. Jennings and F. de Véricourt. Dimensioning large-scale memberschip services. *Operations Research*, 55(1):173–187, 2008.

[25] O.B. Jennings and F. de Véricourt. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[26] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[27] G. Kazahaya. Harnessing technology to redesign labor cost management reports. *Healthcare Financial Management*, 59(4):94–100, 2005.

[28] P. Khudyakov, P.D. Feigin, and A. Mandelbaum. Designing a call center with an IVR (Interactive Voice Response). *Queueing Systems*, 66(3):215–237, 2010.

[29] S. Lundgren and K. Segesten. Nurses use of time in a medical-surgical ward with all-rn staffing. *Journal of Nursing Management*, 9(1):13–20, 2001.

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

[31] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. The John Hopkins University Press, Baltimore, 1981.

[32] J. Sanders, S.C. Borst, A.J.E.M. Janssen, and J.S.H. vvan Leeuwaarden. Optimality gaps in asymptotic dimensioning of many-server systems. *Operations Research Letters*, 44(3):369–365, 2016.

[33] H. Song, A.L. Tucker, and K.L. Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053, 2015.

[34] J.S.H. van Leeuwaarden and C. Knessl. Transient behavior of the halfinwhitt diffusion. *Stochastic Processes and their Applications*, pages 1524–1545, 2011.

[35] J.S.H. van Leeuwaarden and C. Knessl. Spectral gap of the Erlang-A model in the Halfin-Whitt regime. *Stochastic Systems*, 2(1):149–207, 2012.

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

[37] J.S.H. van Leeuwaarden, B.W.J. Mathijsen, and F. Sloothaak. Cloud provisioning in the QED regime. *Proceedings of the 9th EAI International Conference on Performance Evaluation Methodologies and Tools*, pages 180–187, 2016.

[38] G.B Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

[39] B. Zhang, J.S.H. van Leeuwaarden, and B. Zwart. Refining square-root staffing for call centers with impatient customers. *Operations Research*, 60:461–474, 2012.

[40] N. Zychlinski, A. Mandelbaum, P. Momčilović, and I. Cohen. Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. Under review, 2016.

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* $(i,0), 1 \leq i < n$*:* 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* $(i,0), i \geq n$*:* Again, the only possible transitions arises 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. She 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

25

Figure 18: Illustration of state space and the transitions for the Erlang-R model with holding.

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 and transition rates of the Erlang-R model with holding are illustrated in Figure 18.

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

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

26

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

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

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

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 [31] follows that the Markov process $\{N(t), Q_1(t)\}$ is ergodic (stable) if and only if

$$
\pi A_0 e < \pi A_2 e, \tag{25}
$$

27

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

$$\pi A^* = 0, \quad \pi e = 1, \tag{26}$$

and

$$A^* = \begin{pmatrix} -n\delta & n\delta \\ p\mu & -(p\mu + (n-1)\delta) & (n-1)\delta \\ & \ddots & \ddots & \ddots \\ & & sp\mu & -(ps\mu + (n-s)\delta) & (n-s)\delta \\ & & & \ddots & \ddots & \ddots \\ & & & & ps\mu & -(ps\mu + \delta) & \delta \\ & & & & & ps\mu & -ps\mu \end{pmatrix}.$$

Then π must satisfy the balance equations

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

with $\nu(j) = \min\{j, s\}\mu$, and the normalization condition

$$\sum_{i=0}^{n} \pi_i = 1.$$

It is readily verified that

$$\pi_i = \begin{cases} \pi_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \le i \le s, \\ \pi_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s + 1 \le i \le n \end{cases} \tag{27}$$

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 2a is given by*

$$\hat{\pi}_i = \begin{cases} \hat{\pi}_0 \binom{n}{i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } 0 \le i \le s, \\ \hat{\pi}_0 \binom{n}{i} \frac{i!}{s!} s^{s-i} \left(\frac{\delta}{p\mu}\right)^i & \text{for } s + 1 \le i \le n \end{cases} \tag{28}$$

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 patient at queue i, so $r_1(m) = \min\{m, s\}$ and $r_2(m) = m$. The throughput vector $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}_{\not\vdash}{}^2$ must satisfy $\gamma = \gamma P$ and we find that $\gamma = (p, 1)$ suffices. From the general theory of Jackson networks, see [21], 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,

$$g_1(i) = \begin{cases} \frac{1}{i! \mu^i} & \text{for } 0 \le i \le s, \\ \frac{1}{s! s^{i-s} \mu^i} & \text{for } s + 1 \le i \le n, \end{cases}$$

$$g_2(n - i) = \frac{1}{(n - i)!} \left(\frac{p}{\delta}\right)^n \left(\frac{\delta}{p}\right)^i,$$

and rewriting the expressions yields (28). $\qquad\square$

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

$$A_0 + G A_1 + G^2 A_2 = 0. \tag{29}$$

The balance equations can be written as

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

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

$$\pi_n G^{i-n-1} (A_0 + G A_1 + G A_2) = 0, \quad i = n + 1, n + 2, ...$$

Moreover, we have the boundary equations

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

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 (29), it is easy to solve numerically by using the following algorithm (matrix-geometric method). Rewriting (29) 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 [31]. Once $||G_{k+1} - G_k||_2$ is below a certain preset threshold, we approximate G by G_{k+1}.

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 $Q_1^h(t) + Q_2^h(t) \le n = Q_1^J(t) + Q_2^J(t)$ 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 $q_1^b + q_2^b < n$,
$$(h, q_1^h + 1, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{30}$$

2. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b < n$,
$$(h + 1, q_1^h, q_2^h, q_1^b + 1, q_2^b) \qquad \text{with rate } \lambda, \tag{31}$$

3. if $q_1^h + q_2^h < n$ and $q_1^b + q_2^b = n$,
$$(h, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{32}$$

4. if $q_1^h + q_2^h = n$ and $q_1^b + q_2^b = n$,
$$(h + 1, q_1^h + 1, q_2^h, q_1^b, q_2^b) \qquad \text{with rate } \lambda, \tag{33}$$

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 $h > 0$

$$\begin{cases} (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (h-1, q_1^h, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1-p)\mu. \end{cases} \quad (34)$$

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

$$\begin{cases} (h-1, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \quad (35)$$

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

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^b \wedge s)(1-p)\mu, \\ (0, q_1^h - 1, q_2^h, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)](1-p)\mu. \end{cases} \quad (36)$$

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

$$\begin{cases} (0, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } (q_1^h \wedge s)(1-p)\mu, \\ (0, q_1^h, q_2^h, q_1^b - 1, q_2^b) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)](1-p)\mu. \end{cases} \quad (37)$$

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

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

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^b \wedge s)p\mu, \\ (h, q_1^h - 1, q_2^h + 1, q_1^b, q_2^b) & \text{with rate } [(q_1^h \wedge s) - (q_1^b \wedge s)]p\mu. \end{cases} \quad (38)$$

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

$$\begin{cases} (h, q_1^h - 1, q_2^h + 1, q_1^b - 1, q_2^b + 1) & \text{with rate } (q_1^h \wedge s)p\mu, \\ (h, q_1^h, q_2^h, q_1^b - 1, q_2^b + 1) & \text{with rate } [(q_1^b \wedge s) - (q_1^h \wedge s)]p\mu. \end{cases} \quad (39)$$

Become needy

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

$$\begin{cases} (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^b \delta, \\ (h, q_1^h + 1, q_2^h - 1, q_1^b, q_2^b) & \text{with rate } (q_2^h - q_2^b)\delta, \end{cases} \quad (40)$$

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

$$\begin{cases} (h, q_1^h + 1, q_2^h - 1, q_1^b + 1, q_2^b - 1) & \text{with rate } q_2^h \delta, \\ (h, q_1^h, q_2^h, q_1^b + 1, q_2^b - 1) & \text{with rate } (q_2^b - q_2^h)\delta, \end{cases} \quad (41)$$

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 $\hat{X}^h(t) \overset{d}{=} X^h(t)$.

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 $\hat{X}^b(t) \overset{d}{=} X^h(t)$.

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 (30)–(33). 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 $q_1^b + q_2^b = n$, 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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \overset{(i)}{\leq} q_1^h + q_2^h + 1 = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

    (iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\leq} q_1^h + 1 = \hat{Q}_1^h(t) = \hat{Q}_1^h(t) + \hat{H}(t)$.

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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b + 1 \le n = q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b + 1 \overset{(iii)}{\le} q_1^h + 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\le} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\le} q_1^h + h < q_1^h + h + 1 = \hat{H}(t)$.

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 $q_1^b \le q_1^h$, which excludes the state $(0, q_1^h, q_2^h, q_1^b, q_2^b)$ in (37) 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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 \overset{(i)}{\le} q_1^h + q_2^h - 1 < q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\le} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

2. If $Z(t) = (h-1, q_1^h, q_2^h, q_1^b, q_2^b)$, then $h > 0$ and $q_1^h \ge q_1^b$ (*).

(i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\le} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b \overset{(*)}{\le} q_1^h - 1 \le q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b - 1 < q_1^b + q_2^b \overset{(i)}{\le} q_1^h + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b - 1 < q_1^b \overset{(*)}{\le} q_1^h + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

4. If $Z(t) = (h, q_1^h - 1, q_2^h, q_1^b - 1, q_2^b)$, then $h = 0$.

(i) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = (q_1^b - 1) + q_2^b - 1 \overset{(i)}{<\le} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\le} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

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) $\hat{Q}_1^b(t) + \hat{Q}_2^b(t) = q_1^b + q_2^b \overset{(i)}{\le} (q_1^h - 1) + q_2^b \overset{(ii)}{\le} (q_1^h - 1) + q_2^h = \hat{Q}_1^h(t) + \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = qy \overset{(*)}{\le} q_1^h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

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 (38)–(39), 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) $\hat{Q}_2^b(t) = q_2^b + 1 \overset{(ii)}{\le} q_2^h + 1 = \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b - 1 \overset{(iii)}{\le} q_1^h + h - 1 = \hat{Q}_1^h(t) + \hat{H}(t)$.

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) $\hat{Q}_2^b(t) = q_2^b \overset{(ii)}{\le} q_2^h < q_2^h + 1 = \hat{Q}_2^h(t)$.

(iii) $\hat{Q}_1^b(t) = q_1^b \overset{(iii)}{\le} q_1^h + h < q_1^h + 1 + h = \hat{Q}_1^h(t) + \hat{H}(t)$.

33

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 (41). 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

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

The other two orderings follow similarly.

Remark 6. 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 $\hat{Q}_1^h(t) < \hat{Q}_1^b(t)$.

C Proof of Theorem 2 - Performance measures of Erlang-R with blocking

For convenience we state the theorem here again.

Theorem 7. *Let s and n scale as in* (11) *with $\beta, \gamma > 0$ as $\lambda \to \infty$. Then, if $\beta \neq 0$*

$$g^b(\beta, \gamma) := \lim_{\lambda \to \infty} \mathbb{P}^b(\text{delay}) = \left(1 + \frac{\beta \int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\phi(\beta)\Phi(\eta) - \phi(\sqrt{\beta^2 + \eta^2})e^{\frac{1}{2}\omega^2}\Phi(\omega)}\right)^{-1}, \qquad (42)$$

$$f^b(\beta, \gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block}) = \frac{\sqrt{r}\phi(\gamma)\Phi(-\omega\sqrt{r}) + \phi(\sqrt{\beta^2 + \eta^2})e^{\frac{1}{2}\omega^2}\Phi(\omega)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2 + \eta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)}, \qquad (43)$$

$$h^b(\beta, \gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{E}[W] = \frac{\frac{\phi(\beta)\Phi(\eta)}{\beta^2} + \left(\frac{\beta}{r} - \frac{\gamma}{\sqrt{r}} - \frac{1}{\beta}\right)\frac{\phi(\sqrt{\eta^2 + \beta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega) - \sqrt{\frac{1-r}{r}}\frac{\phi(\beta)\phi(\eta)}{\beta}}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2 + \eta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)}, \qquad (44)$$

and if $\beta = 0$,

$$g_0^b(\gamma) := \lim_{\lambda \to \infty} \mathbb{P}^b(\text{delay}) \left(1 + \frac{\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t)}{\sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta) + \phi(\eta)\right)}\right)^{-1} \qquad (45)$$

$$f_0^b(\gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{P}^b(\text{block}) = \frac{\sqrt{r}\,\phi(\gamma)\Phi(-\omega\sqrt{r}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{\beta} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{1-r}{r}}\frac{1}{\sqrt{2\pi}}\left(\eta\,\Phi(\eta) + \phi(\eta)\right)}, \qquad (46)$$

$$h_0^b(\gamma) := \lim_{\lambda \to \infty} \sqrt{R_1} \cdot \mathbb{E}[W] = \frac{1}{2\mu} \frac{\left(\gamma^2/r + 1\right)\Phi(\eta) + \eta\phi(\eta)}{\frac{r}{1-r}\sqrt{2\pi}\int_{-\infty}^{0} \Phi\left(\frac{\gamma - t\sqrt{r}}{\sqrt{1-r}}\right) d\Phi(t) + \sqrt{\frac{r}{1-r}}\left(\eta\Phi(\eta) + \phi(\eta)\right)}, \qquad (47)$$

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

In this appendix we prove the heavy-traffic approximations of the system-measures introduced in Theorem 2. 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* (11) *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* (11) *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 (11). Define A as the expression*

$$A = \sum_{\substack{i,j|i \le s, \\ i+j \le n-1}} \frac{1}{i!j!} R_1^i R_2^j e^{-(R_1+R_2)}. \tag{48}$$

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 (11) 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

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

C.1 Proof of Lemma 1

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

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

$$= \frac{e^{s(1-\rho)}}{\sqrt{2\pi}\beta} \rho^s P(X_\lambda \le n-s-1)$$

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

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

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 $\lambda \to \infty$.

$$\lim_{\lambda \to \infty} \frac{n-s-R_2}{\sqrt{R_2}} = \lim_{\lambda \to \infty} \frac{n - \frac{R_1}{r} - s + R_1}{\sqrt{\frac{R_1}{r} - R_1}} = \frac{\gamma\sqrt{\frac{R_1}{r}} - \beta\sqrt{R_1}}{\sqrt{\frac{R_1}{r} - R_1}} = \frac{\gamma - \beta\sqrt{r}}{\sqrt{1-r}}. \tag{49}$$

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

$$\mathbb{P}(X_\lambda \le n-s-1) \to \mathbb{P}(N(0,1) \le \eta) = \Phi(\eta), \quad as \ \lambda \to \infty \tag{50}$$

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

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

one obtains

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

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. $\qquad\square$

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:

$$B_2 \approx \frac{e^{s-R_1-R_2}}{\sqrt{2\pi s}}\frac{\rho^n}{1-\rho}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^l = \frac{e^{s(1-\rho)-R_2}}{\sqrt{2\pi s}}\frac{\sqrt{s}\rho^n}{\beta}e^{\frac{R_2}{\rho}}\sum_{l=0}^{n-s-1}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^l e^{-\frac{R_2}{\rho}}$$

$$= \frac{e^{s(1-\rho)+R_2\left(\frac{1-\rho}{\rho}\right)}}{\sqrt{2\pi\beta}}\rho^n\mathbb{P}(Y_\lambda \le n-s-1),$$

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).

$$\lim_{\lambda\to\infty}\frac{n-s-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}} = \lim_{\lambda\to\infty}\frac{\eta\sqrt{R_2}+R_2-\frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}} = \lim_{\lambda\to\infty}\eta\sqrt{\rho}+\frac{\sqrt{R_2}(\rho-1)}{\sqrt{\rho}}$$

$$= \eta - \lim_{\lambda\to\infty}\sqrt{\frac{sp\mu}{\delta}}(1-\rho) = \eta - \sqrt{\frac{p\mu}{\delta}}\beta.$$

$$(51)$$

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

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

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

$$s(1-\rho) + R_2\left(\frac{1-\rho}{\rho}\right) + n\ln\rho = s(1-\rho) + R_2\left(\frac{1-\rho}{\rho}\right) - n\left(1-\rho+\frac{(1-\rho)^2}{2}\right)$$

$$= -\left(n-s-\frac{R_2}{\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2} \approx -\left(\eta\sqrt{R_2}+R_2-\frac{R_2}{\rho}\right)(1-\rho) - \frac{n(1-\rho)^2}{2}$$

$$= \left(\frac{R_2}{\rho}-\frac{n}{2}\right)(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho) \approx \left(\frac{R_2}{\rho}-\frac{1}{2}\left(\frac{R_1}{r}+\gamma\sqrt{\frac{R_1}{r}}\right)\right)(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho)$$

$$= \left(\frac{R_2}{\rho}-\frac{R_1}{2r}\right)(1-\rho)^2 - \frac{1}{2}\gamma\sqrt{\frac{R_1}{r}}(1-\rho)^2 - \eta\sqrt{R_2}(1-\rho)$$

$$= \left(\frac{R_2}{\rho}-\frac{R_1}{2r}\right)\frac{\beta^2}{R_!} - \frac{1}{2}\gamma\sqrt{\frac{R_1}{r}}\frac{\beta^2}{R_1} - \eta\sqrt{R_2}\beta\sqrt{\frac{1}{R_1}}$$

$$\approx \frac{p\mu}{\delta\rho}\beta^2 - \frac{1}{2}\left(\frac{p\mu}{\delta}+1\right)\beta^2 - \eta\beta\sqrt{\frac{p\mu}{\delta}} = -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}\omega^2.$$

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 $[0,s]$.

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

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 (48). Let us consider a lower estimate for A given by the following sum:

$$A \geq A_1 = \sum_{h=0}^{\ell}\sum_{i=s_{h+1}}^{s_h}\frac{1}{i!}R_1^i e^{-R_1} \cdot \sum_{j=0}^{n-s_h-1}\frac{1}{j!}R_2^j e^{-R_2}$$

$$= \sum_{h=0}^{\ell}\sum_{i=s_{h+1}}^{s_h}\frac{1}{i!}R_1^i e^{-R}\mathbb{P}(Y_n \leq n-s_h-1) \quad (52)$$

$$= \sum_{h=0}^{\ell}\mathbb{P}(s_{h+1} \leq X_n \leq s_h)\mathbb{P}(Y_n \leq n-s_h-1)$$

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:

$$\begin{aligned}
\mathbb{P}(s_{h+1} \leq X_n \leq s_h) &= \Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon), \ h = 0,..,\ell-1 \\
\mathbb{P}(0 \leq X_n \leq s_\ell) &= \Phi(\beta - \ell\epsilon).
\end{aligned} \tag{53}$$

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

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

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

$$\lim_{\lambda \to \infty} \frac{n - s_h - R_2}{\sqrt{R_2}} = \lim_{\lambda \to \infty} \frac{n - s - h\epsilon\sqrt{R_1} - R_2}{\sqrt{R_2}} = \lim_{\lambda \to \infty} \frac{R_2 + \eta\sqrt{R_2} - h\epsilon\sqrt{R_1} - R_2}{\sqrt{R_2}}$$

$$= \eta - h\epsilon\frac{\sqrt{R_1}}{\sqrt{R_2}} = \eta - h\epsilon\sqrt{\frac{\delta}{p\mu}},$$

we obtain:

$$\mathbb{P}(Y_n \leq n - s_h) = \Phi\left(\eta - h\epsilon\sqrt{\frac{\delta}{p\mu}}\right), \ h = 0,..,\ell \tag{54}$$

It follows from (52), (53), and (54) that

$$\lim_{\lambda \to \infty} A \geq \sum_{h=0}^{\ell-1}(\Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon))\Phi\left(\eta - h\epsilon\sqrt{\frac{\delta}{p\mu}}\right) + \Phi(\beta - \ell\epsilon)\Phi\left(\eta - \ell\epsilon\sqrt{\frac{\delta}{p\mu}}\right) \tag{55}$$

which is the lower Riemann-Stieltjes sum of the integral

$$-\int_0^\infty \Phi\left(\eta + x\sqrt{\frac{\delta}{p\mu}}\right) d\Phi(\beta - x) = \int_{-\infty}^\beta \Phi\left(\eta + (\beta - t)\sqrt{\frac{\delta}{p\mu}}\right) d\Phi(t) \tag{56}$$

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:

$$A \leq A_2 = \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} R_1^i e^{-R_1} \sum_{j=0}^{n-s_{h+1}-1} \frac{1}{j!} R_2^j e^{-R_2}$$

$$= \sum_{h=0}^{\ell} \sum_{i=s_{h+1}}^{s_h} \frac{1}{i!} R_1^i e^{-R} \mathbb{P}(Y_n \leq n - s_{h+1} - 1) \tag{57}$$

$$= \sum_{h=0}^{\ell} \mathbb{P}(s_{h+1} \leq X_n \leq s_h) \mathbb{P}(Y_n \leq n - s_{h+1} - 1)$$

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

$$\lim_{\lambda \to \infty} A \leq \sum_{h=0}^{\ell-1} (\Phi(\beta - h\epsilon) - \Phi(\beta - (h+1)\epsilon)) \Phi\left(\eta - (h+1)\epsilon\sqrt{\frac{\delta}{p\mu}}\right) + \Phi(\beta - \ell\epsilon) \tag{58}$$

which is the upper Riemann-Stieltjes sum for the integral (56). When $\epsilon \to 0$ the boundaries (55) and (58) 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

Proof. First, we will rewrite Equation (48):

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

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,

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

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

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

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

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

$$\approx e^{s-R} \frac{1}{\sqrt{2\pi s}} \rho^s \left((n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2} + \frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} \right)$$

$$= \frac{1}{\sqrt{2\pi s}} \left((n-s-R_2) \sum_{l=0}^{n-s-1} \frac{1}{l!} R_2^l e^{-R_2} + \frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} \right).$$

As seen in Equation (50)

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

By using Stirling's formula:

$$\frac{R_2^{n-s} e^{-R_2}}{(n-s-1)!} = \frac{(n-s) R_2^{n-s} e^{-R_2}}{(n-s)!} \approx \frac{(n-s) e^{n-s-R_2}}{\sqrt{2\pi(n-s)}} \left(\frac{R_2}{n-s} \right)^{n-s}$$

$$= \frac{(n-s) e^{n-s-R_2+(n-s)\ln\left(\frac{R_2}{n-s}\right)}}{\sqrt{2\pi(n-s)}} = \sqrt{\frac{n-s}{2\pi}} e^{(n-s)\left(1 - \frac{R_2}{n-s} + \ln\left(\frac{R_2}{n-s}\right)\right)}.$$

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

$$(n-s) \left(1 - \frac{R_2}{n-s} + \ln\left(\frac{R_2}{n-s}\right) \right) = (n-s) \left(1 - \frac{R_2}{n-s} - \left(1 - \frac{R_2}{n-s}\right) - \frac{1}{2} \left(1 - \frac{R_2}{n-s}\right)^2 \right)$$

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

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

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

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. $\qquad\square$

C.5 Approximation of the Probability of Delay

The first approximation will be for the measure: the probability of waiting or the probability of delay.

Theorem 8. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (11). Then if $\beta \neq 0$*

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

and if $\beta = 0$,

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

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

Proof. Proof: By the Arrival theorem for closed networks,

$$
\mathbb{P}_n(W > 0) = \mathbb{P}_{n-1}(Q_1(\infty) \geq s) = \sum_{m=s}^{n-1} \sum_{i=s}^{m} \pi_{n-1}^b(i, m - i)
$$

$$
= \pi_0^{n-1} \sum_{m=s}^{n-1} \sum_{i=s}^{m} \frac{1}{s! s^{i-s}} R_1^i \frac{1}{(m-i)!} R_2^{m-i},
$$

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

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

42

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

$$= \sum_{j=0}^{n-s-1}\sum_{i=0}^{n-s-j-1}\frac{1}{s!s^i}\frac{1}{j!}R_1^{i+s}R_2^j e^{-(R_1+R_2)} = \frac{1}{s!}R_1^s e^{-(R_1+R_2)}\sum_{j=0}^{n-s-1}\frac{1}{j!}R_2^j\sum_{i=0}^{n-s-j-1}\rho^i, \tag{59}$$

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:

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

$$= \frac{1}{s!}R_1^s e^{-(R_1+R_2)}\frac{1}{1-\rho}\sum_{j=0}^{n-s-1}\frac{1}{j!}R_2^j - \frac{1}{s!}R_1^s e^{-(R_1+R_2)}\frac{\rho^{n-s}}{1-\rho}\sum_{j=0}^{n-s-1}\frac{1}{j!}\left(\frac{R_2}{\rho}\right)^j = B_1 - B_2. \tag{60}$$

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

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

This proves Theorem 8. □

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 9. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (11) with $\beta \neq 0$. Then*

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

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

Proof. Proof: It follows from (6) that the expectation of the waiting time is given by

$$\mathbb{E}[W] = \int_0^\infty p_n(s;t)dt = \frac{1}{\mu s}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}(i,m-i)(i-s+1)$$

$$= \frac{1}{\mu s}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}(i,m-i)(i-s) + \frac{1}{\mu s}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}(i,m-i) \tag{61}$$

$$= \frac{1}{\mu s}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}(i,m-i)(i-s) + \frac{1}{\mu s}\mathbb{P}(W>0) = C + D,$$

43

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 (48) and (59), respectively, and C is given by,
$$C = \frac{1}{\mu s}\sum_{m=s}^{n-1}\sum_{i=s}^{m}\pi_{n-1}(i, m-i)\,(i-s)$$
$$= \frac{1}{\mu s}\pi_0\sum_{m=s}^{n-1}\sum_{i=s}^{m}\frac{R_1^i}{s!s^{i-s}}\frac{R_2^{m-i}}{(m-i)!}\,(i-s) = \frac{1}{\mu s}\frac{Ge^{-(R_1+R_2)}}{A+B}.$$

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
$$\sum_{l=0}^{M}l\,\rho^l = \rho\Big(\sum_{l=0}^{M}\rho^l\Big)' = \rho\Big(\frac{1-\rho^{M+1}}{1-\rho}\Big)' = -M\frac{\rho^{M+1}}{1-\rho} + \rho\frac{1-\rho^M}{(1-\rho)^2}, \tag{62}$$

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

44

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

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

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

$$= e^{R_1+R_2} \frac{1}{1-\rho} (\rho B_1 - B_2),$$

with B_1 as in Lemma 1. Note that

$$n - s - \frac{R_2}{\rho} - 1 = R_2 - \eta\sqrt{R_2} - \frac{R_2}{\rho} - 1 = R_2\left(1 - 1/\rho\right) + \eta\sqrt{R_2} - 1$$

$$= R_2\left(\frac{R_1 - s}{\sqrt{R_1}}\right) + \eta\sqrt{R_2} - 1 = R_2 \cdot -\frac{\beta\sqrt{R_1}}{R_1} + \eta\sqrt{R_2} - 1 \approx \sqrt{R_1}\left(\eta\sqrt{\xi} - \beta\xi\right)$$

$$\approx \sqrt{s}\left(\eta/\sqrt{\xi} - \beta/\xi\right),$$

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

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

$$= \frac{n-s}{1-\rho} e^{R_1+R_2} \, \mathbb{P}\left(\mathrm{Pois}(R_1) = s\right) \mathbb{P}\left(\mathrm{Pois}(R_2) = n-s\right)$$

$$= \frac{R_2 + \eta\sqrt{R_2}}{\beta/\sqrt{R_1}} e^{R_1+R_2} \mathbb{P}\left(\mathrm{Pois}(R_1) = s\right) \mathbb{P}\left(\mathrm{Pois}(R_2) = n-s\right)$$

$$= \sqrt{R_2} \, e^{R_1+R_2} \frac{1}{\beta} \sqrt{R_1} \mathbb{P}\left(\mathrm{Pois}(R_1) = s\right) \cdot \sqrt{R_2}\mathbb{P}\left(\mathrm{Pois}(R_2) = n-s\right))$$

$$\approx \sqrt{R_1} \, e^{R_1+R_2} \frac{1}{\sqrt{\xi}\beta} \, \phi(\beta)\phi(\eta),$$

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

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

and for G_2

$$\frac{1}{\sqrt{s}} e^{-(R_1+R_2)} G_2 = \frac{1}{\sqrt{s}} \frac{1}{1-\rho} (\rho B_1 - B_2) \approx \frac{1}{\beta} (\rho B_1 - B_2)$$

$$\to \frac{\phi(\beta)\Phi(\eta)}{\beta^2} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta^2} e^{\frac{1}{2}\omega^2} \Phi(\omega),$$

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

$$\frac{1}{\sqrt{s}} e^{-(R_1+R_2)} G \to \left(\frac{1}{\xi} - \frac{\eta}{\beta\sqrt{\xi}} - \frac{1}{\beta^2}\right) \phi(\sqrt{\beta^2+\eta^2}) e^{\frac{1}{2}\omega^2} \Phi(\omega) + \frac{\phi(\beta)\Phi(\eta)}{\beta^2} - \frac{\phi(\beta)\phi(\eta)}{\sqrt{\xi}\beta}. \quad (63)$$

We can conclude,

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

$$\to \frac{1}{\mu} \frac{\left(\frac{1}{\xi} - \frac{\eta}{\beta\sqrt{\xi}} - \frac{1}{\beta^2}\right) \phi(\sqrt{\beta^2+\eta^2}) e^{\frac{1}{2}\omega^2} \Phi(\omega) + \frac{\phi(\beta)\Phi(\eta)}{\beta^2} - \frac{\phi(\beta)\phi(\eta)}{\sqrt{\xi}\beta}}{\int_{-\infty}^{\beta} \Phi(\eta + (\beta-t)\sqrt{\xi}) \, d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\beta^2+\eta^2})}{\beta} e^{\frac{1}{2}\omega^2} \Phi(\omega)},$$

as $s \to \infty$. $\square$

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

Theorem 10. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in* (11) *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

$$\mathbb{E}[W] = \frac{1}{\mu s} \frac{Ge^{-(R_1 + R_2)} + B}{A + B}. \tag{64}$$

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,

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

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

$$= \eta\sqrt{R_2}\,\mathbb{P}(\text{Pois}(R_2) \leq n-s-1) + (n-s-1)\mathbb{P}(\text{Pois}(R_2) = n-s)$$

$$= \eta\sqrt{R_2}\,\mathbb{P}(\text{Pois}(R_2) \leq n-s) + (R_2-1)\mathbb{P}(\text{Pois}(R_2) = n-s)$$

$$\approx \sqrt{R_2}\left(\eta\,\Phi(\eta) + \phi(\eta)\right) = \sqrt{\xi^{-1}}\sqrt{R_1}\left(\eta\,\Phi(\eta) + \phi(\eta)\right). \tag{65}$$

Furthermore

$$n-s-1-R_2 = \eta\sqrt{R_2} - 1 \approx \eta\sqrt{\xi^{-1}}\sqrt{R_1} \tag{66}$$

and

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

$$\approx \sqrt{R_1}\xi^{-1}\phi(0)\Phi(\eta). \tag{67}$$

46

Combining (65)-(67), we find

$$e^{-R_1-R_2}G \approx \frac{1}{2}\xi^{-1}\eta\sqrt{R_1}\,\mathbb{P}(\text{Pois}(R_1) = s)\,(\eta\Phi(\eta) + \phi(\eta)) + \frac{1}{2}\sqrt{R_1}\phi(0)\Phi(\eta)$$

$$\approx \frac{1}{2}\xi^{-1}\eta\sqrt{R_1}\,(\eta\,\Phi(\eta) + \phi(\eta)) + \frac{1}{2}\xi^{-1}\sqrt{R_1}\,\phi(0)\Phi(\eta)$$

$$= \frac{1}{2}\sqrt{R_1}\,\xi^{-1}\phi(0)\left[(\eta^2 + 1)\Phi(\eta) + \eta\,\phi(\eta)\right],$$

for R_1 large. Hence, we can conclude

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

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

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 2 of the probability of blocking. For convenience we repeat here the relevant theorem.

Theorem 11. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (11) with $\beta \neq 0$. Then*

$$\lim_{\lambda\to\infty}\sqrt{s}\mathbb{P}(block) = \frac{\sqrt{r}\phi(\gamma)\Phi(-\omega\sqrt{r}) + \phi(\sqrt{\eta^2 + \beta^2})e^{\frac{\omega^2}{2}}\Phi(\omega)}{\int_{-\infty}^{\beta}\Phi\left(\eta + (\beta - t)\sqrt{\xi}\right)d\Phi(t) + \frac{\phi(\beta)\Phi(\eta)}{\beta} - \frac{\phi(\sqrt{\eta^2 + \beta^2})}{\beta}e^{\frac{1}{2}\omega^2}\Phi(\omega)} \tag{68}$$

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 (5) 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

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

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

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

$$\tilde{B}_1 = \frac{1}{s!}R_1^s\frac{1}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}R_2^le^{-(R_1+R_2)}$$

$$\tilde{B}_2 = \frac{1}{s!}R_1^s\frac{\rho^{n-s+1}}{1-\rho}\sum_{l=0}^{n-s}\frac{1}{l!}\left(\frac{R_2}{\rho}\right)^le^{-(R_1+R_2)}.$$

Note that by Lemmas 1,2 and 3

$$\lim_{\lambda\to\infty} \tilde{B}_1 = \lim_{\lambda\to\infty} B_1 = \frac{\phi(\beta)\Phi(\eta)}{\beta},$$

$$\lim_{\lambda\to\infty} \tilde{B}_2 = \lim_{\lambda\to\infty} B_2 = \frac{\phi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1),$$

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

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

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

$$= \frac{1}{n!}e^{-(R_1+R_2)}(R_1+R_2)^n \sum_{i=0}^{s} \mathbb{P}(X_\lambda = (i, n-i))$$

$$= \mathbb{P}(Y_\lambda = n)\sum_{i=0}^{s} \mathbb{P}(X_\lambda = (i, n-i)) = \mathbb{P}(Y_\lambda = n)\mathbb{P}(X_\lambda^1 \le s)$$

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 51

$$\mathbb{P}(X_\lambda^1 \le s) = \Phi\left(\frac{s - np_i}{\sqrt{np_i(1-p_i)}}\right) = \Phi\left(\frac{s - n\frac{R_1}{R_1+R_2}}{\sqrt{n\frac{R_1}{R_1+R_2}(1 - \frac{R_1}{R_1+R_2})}}\right)$$

$$= \Phi\left(\frac{s - n\frac{R_1}{R_1+R_2}}{\sqrt{n\frac{R_1}{R_1+R_2}\frac{R_2}{R_1+R_2}}}\right) = \Phi\left(\frac{s(R_1+R_2) - nR_1}{\sqrt{nR_1R_2}}\right)$$

$$= \Phi\left(\sqrt{\frac{R_1}{R_2}}\frac{s\frac{R_1+R_2}{R_1} - n}{\sqrt{n}}\right) = \Phi\left(\sqrt{\frac{R_1}{R_2}}\frac{s(1 + \frac{R_2}{R_1}) - n}{\sqrt{n}}\right)$$

$$= \Phi\left(\sqrt{\frac{R_1}{R_2}}\frac{s + \frac{R_2}{\rho} - n}{\sqrt{n}}\right) = \Phi\left(-\sqrt{\frac{R_1}{R_2}}\sqrt{\frac{R_2}{n\rho}} \cdot \frac{n - s - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right)$$

$$= \Phi\left(-\sqrt{\frac{s}{n}} \cdot \frac{n - s - \frac{R_2}{\rho}}{\sqrt{\frac{R_2}{\rho}}}\right) \approx \Phi\left(-\sqrt{\frac{R_1}{R_1+R_2}} \cdot \left(\eta - \beta\frac{R_2}{R_1}\right)\right)$$

$$= \Phi\left(\frac{\beta\frac{R_2}{R_1} - \eta}{\sqrt{1 + \frac{R_2}{R_1}}}\right) = \Phi\left(\frac{\beta\sqrt{\frac{p\mu}{\delta}} - \eta}{\sqrt{1 + \frac{p\mu}{\delta}}}\right). \tag{69}$$

By the normal approximation of the Poisson distribution:

$$\mathbb{P}(Y_\lambda = n) \approx \frac{1}{\sqrt{R_1+R_2}}\phi\left(\frac{n - (R_1+R_2)}{\sqrt{R_1+R_2}}\right) \approx \frac{1}{\sqrt{s}\sqrt{1 + \frac{p\mu}{\delta}}}\phi\left(\frac{\eta\sqrt{\frac{p\mu}{\delta}} + \beta}{\sqrt{1 + \frac{p\mu}{\delta}}}\right). \tag{70}$$

We based on the following equivalences (as λ tends to ∞) to develop Equations 69 and 70:

$$R_1 + R_2 = \frac{\lambda}{(1-p)\mu} + \frac{p\lambda}{(1-p)\delta} \approx s + (1-p)s\mu \left(\frac{p}{(1-p)\delta} \right) = s \left(1 + \frac{\mu p}{\delta} \right) = s \left(1 + \frac{R_2}{R_1} \right);$$

$$n - (R_1 + R_2) \approx s + R_2 + \eta\sqrt{R_2} - (R_1 + R_2) = s + \eta\sqrt{R_2} - R_1 \approx s + \eta\sqrt{R_2} - s + \beta\sqrt{s}$$

$$\approx \eta\sqrt{\frac{s\mu p}{\delta}} + \beta\sqrt{s};$$

$$\frac{n - (R_1 + R_2)}{\sqrt{R_1 + R_2}} \approx \frac{\eta\sqrt{\frac{\mu p}{\delta}} + \beta}{\sqrt{1 + \frac{\mu p}{\delta}}}.$$

Following Equations 69 and 70 we get

$$\delta_1 = \mathbb{P}(Y_\lambda = n)\mathbb{P}(X_\lambda^1 \le s) \approx \frac{1}{\sqrt{s}\sqrt{1 + \frac{\mu p}{\delta}}} \phi \left(\frac{\eta\sqrt{\frac{\mu p}{\delta}} + \beta}{\sqrt{1 + \frac{\mu p}{\delta}}} \right) \Phi \left(\frac{\beta\sqrt{\frac{\mu p}{\delta}} - \eta}{\sqrt{1 + \frac{\mu p}{\delta}}} \right). \qquad (71)$$

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

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

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

When comparing δ_2 to B_2 form Equation 60, 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 11. $\qquad \square$

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

Theorem 12. *Let the variables λ, s and n tend to ∞ simultaneously and satisfy the QED scaling conditions in (11) with $\beta = 0$. Define and $\xi = \frac{R_1}{R_2} = \frac{\delta}{p\mu}$, then*

$$\lim_{\lambda \to \infty} \sqrt{s}\mathbb{P}(block) = \frac{\sqrt{r}\phi(\gamma)\Phi(-\frac{\gamma\sqrt{r}}{\sqrt{1-r}}) + \frac{1}{\sqrt{2\pi}}\Phi(\eta)}{\int_{-\infty}^{0} \Phi\left(\eta - t\sqrt{\xi} \right) d\Phi(t) + \sqrt{\frac{1-r}{r}} \frac{1}{\sqrt{2\pi}} \left(\eta\Phi(\eta) + \phi(\eta) \right)}, \qquad (72)$$

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

Proof. Proof: It follows from (5) that the probability of blocking is given by

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

where

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

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

$$\tilde{A} = \sum_{\substack{i,j \mid i \leq s, \\ i+j \leq n}} \frac{1}{i! j!} R_1^i R_2^j e^{-(R_1+R_2)}$$

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

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 11.

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

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

This proves Theorem 12.

$\square$

D   Proof of Proposition 3

Define

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

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 (11) is equivalent to showing that

$$1 - \rho_{\max}(s,n) = \frac{C(s,n) - A(s,n)}{C(s,n) + B(s,n)} = \frac{(1+b)^{-n}[C(s,n) - A(s,n)]}{(1+b)^{-n}[C(s,n) + B(s,n)]} \to 0. \tag{73}$$

First, we rewrite

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

$$= \frac{n}{s}\left(\frac{b}{1+b}\right)\sum_{k=0}^{s-1}\binom{n-1}{k}\left(\frac{b}{1+b}\right)^k\left(\frac{1}{1+b}\right)^{n-1-k}$$

$$= \frac{rn}{s}\sum_{k=0}^{s-1}\binom{n-1}{k}r^k(1-r)^{n-1-k} = \frac{rn}{s}\mathbb{P}(\mathrm{Bin}(n-1,r)\leq s-1)$$

$$= \frac{rn}{s}\mathbb{P}\left(\frac{\mathrm{Bin}(n-1,r)-(n-1)r}{\sqrt{nr(1-r)}}\leq\frac{s-1-(n-1)r}{\sqrt{nr(1-r)}}\right)\to\Phi\left(\frac{\beta-\gamma\sqrt{r}}{\sqrt{1-r}}\right),$$

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

$$(1+b)^{-n}C(s,n) = \sum_{k=0}^{s}\binom{n}{k}\left(\frac{b}{1+b}\right)^k\left(\frac{1}{1+b}\right)^{n-k} = \sum_{k=0}^{s}\binom{n}{k}r^k(1-r)^{n-k}$$

$$= \mathbb{P}(\mathrm{Bin}(n,r)\leq s)\to\Phi\left(\frac{\beta-\gamma\sqrt{r}}{\sqrt{1-r}}\right).$$

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

$$(1+b)^{-n}B(s,n) = (1+b)^{-n}\sum_{k=s+1}^{n}\binom{n}{k}\frac{k!}{s!}s^{s-k}b^k = (1+b)^{-n}\frac{n!}{s!}s^s\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{-k}$$

$$= (1+b)^{-n}\frac{n!}{s!}s^s\left(\frac{b}{s}\right)^n\sum_{k=s+1}^{n}\frac{1}{(n-k)!}\left(\frac{s}{b}\right)^{n-k} = r^n\frac{n!}{s!}s^{s-n}\sum_{m=0}^{n-s-1}\frac{1}{m!}\left(\frac{s}{b}\right)^m$$

$$= \left(\frac{r}{s}\right)^n\frac{n!}{s!}s^s\,\mathrm{e}^{s/b}\,\mathbb{P}(\mathrm{Pois}(s/b)\leq n-s-1),$$

in which

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

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 (73) converges to a constant value as λ grows, and hence the $1 - \rho_{\max}(s,n) \to 0$ as $\lambda \to \infty$.

E   Numerical results on accuracy of the asymptotic approximations

In this section, we examine to accuracy of the asymptotic approximations of Theorem 2 and the heuristic method in Section 4.3. We perform our numerical experiments for three case instances, with parameter settings as in Table 1.

	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

Table 1: Parameter settings for numerical experiments.

E.1   Restricted Erlang-R model with blocking

	$\beta=1,\ \gamma=1$			$\beta=1,\ \gamma=2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1270	0.0900	0.2283	0.1553	0.0212	0.1085
10	0.1340	0.0910	0.1919	0.1628	0.0206	0.1205
25	0.1981	0.0945	0.1614	0.2356	0.0216	0.2145
50	0.1513	0.0963	0.1588	0.1830	0.0205	0.1496
100	0.1880	0.0956	0.1532	0.2231	0.0224	0.2055
250	0.1797	0.0971	0.1399	0.2143	0.0219	0.2057
	0.1767	*0.0981*	*0.1437*	*0.2108*	*0.0217*	*0.1947*

	$\beta=2,\ \gamma=1$			$\beta=2,\ \gamma=2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0237	0.0868	0.0282	0.0322	0.0192	0.0391
10	0.0206	0.0872	0.0188	0.0278	0.0183	0.0264
25	0.0277	0.0876	0.0123	0.0363	0.0174	0.0174
50	0.0185	0.0913	0.0116	0.0249	0.0175	0.0166
100	0.0232	0.0888	0.0103	0.0303	0.0183	0.0145
250	0.0203	0.0905	0.0079	0.0267	0.0179	0.0109
	0.0188	*0.0914*	*0.0084*	*0.0247*	*0.0177*	*0.0118*

Table 2: Numerical results for Erlang-R model with blocking for Case 1.

	$\beta = 1,\ \gamma = 1$			$\beta = 1,\ \gamma = 2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0911	0.1538	0.0479	0.1431	0.0345	0.0909
10	0.1010	0.1498	0.0560	0.1520	0.0326	0.1025
25	0.1594	0.1509	0.1058	0.2192	0.0405	0.1785
50	0.1201	0.1506	0.0726	0.1697	0.0381	0.1248
100	0.1514	0.1539	0.1001	0.2088	0.0398	0.1704
250	0.1459	0.1524	0.0957	0.2003	0.0397	0.1618
	0.1429	*0.1569*	*0.0940*	*0.1976*	*0.0391*	*0.1617*

	$\beta = 2,\ \gamma = 1$			$\beta = 2,\ \gamma = 2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0130	0.1484	0.0044	0.0277	0.0294	0.0109
10	0.0121	0.1432	0.0042	0.0244	0.0267	0.0098
25	0.0182	0.1383	0.0070	0.0319	0.0295	0.0141
50	0.0119	0.1415	0.0043	0.0216	0.0301	0.0090
100	0.0154	0.1413	0.0059	0.0270	0.0290	0.0119
250	0.0136	0.1403	0.0051	0.0236	0.0291	0.0103
	0.0126	*0.1445*	*0.0048*	*0.0220*	*0.0284*	*0.0097*

Table 3: Numerical results for Erlang-R model with blocking for Case 2.

E.2   Restricted Erlang-R model with holding

	$\beta = 1,\ \gamma = 1$			$\beta = 1,\ \gamma = 2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0547	0.1945	0.0221	0.1181	0.0604	0.0617
10	0.0579	0.2158	0.0237	0.1325	0.0526	0.0746
25	0.1113	0.2086	0.0544	0.1959	0.0641	0.1311
50	0.0813	0.2050	0.0363	0.1523	0.0562	0.0933
100	0.1060	0.2146	0.0509	0.1873	0.0632	0.1250
250	0.1006	0.2179	0.0475	0.1820	0.0596	0.1214
	0.1011	*0.2185*	*0.0478*	*0.1792*	*0.0605*	*0.1199*

	$\beta = 2,\ \gamma = 1$			$\beta = 2,\ \gamma = 2$		
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{P}(b)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0034	0.1888	0.0009	0.0175	0.0510	0.0057
10	0.0030	0.2093	0.0008	0.0172	0.0416	0.0058
25	0.0070	0.1937	0.0020	0.0243	0.0440	0.0089
50	0.0043	0.1946	0.0011	0.0163	0.0414	0.0056
100	0.0061	0.1999	0.0017	0.0207	0.0431	0.0076
250	0.0052	0.2037	0.0014	0.0185	0.0401	0.0067
	0.0052	*0.2039*	*0.0014*	*0.0173*	*0.0404*	*0.0063*

Table 4: Numerical results for Erlang-R model with blocking for Case 3.

	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1532	0.1031	0.1628	0.1216
10	0.1622	0.1272	0.1697	0.1331
25	0.2340	0.2116	0.2413	0.2342
50	0.1817	0.1468	0.1890	0.1678
100	0.2199	0.1931	0.2304	0.2269
250	0.2110	0.1852	0.2176	0.2230
	0.2076	*0.1777*	*0.2187*	*0.2050*

	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 1$	
R_1	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0310	0.0121	0.0344	0.0148
10	0.0267	0.0123	0.0292	0.0128
25	0.0348	0.0171	0.0373	0.0184
50	0.0240	0.0108	0.0258	0.0125
100	0.0293	0.0143	0.0317	0.0163
250	0.0256	0.0120	0.0276	0.0145
	0.0229	*0.0104*	*0.0257*	*0.0124*

Table 5: Simulated and approximated probability of delay in Erlang-R model with holding for Case 1.

R_1	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.1327	0.0740	0.1620	0.1096
10	0.1446	0.0894	0.1683	0.1207
25	0.2204	0.1631	0.2442	0.2203
50	0.1694	0.1122	0.1888	0.1507
100	0.2098	0.1524	0.2322	0.2111
250	0.2033	0.1534	0.2190	0.1979
	0.1840	*0.1277*	*0.2109*	*0.1759*

R_1	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 1$	
	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0219	0.0079	0.0322	0.0137
10	0.0199	0.0073	0.0284	0.0115
25	0.0283	0.0128	0.0375	0.0163
50	0.0190	0.0078	0.0255	0.0107
100	0.0244	0.0097	0.0314	0.0151
250	0.0214	0.0083	0.0272	0.0134
	0.0169	*0.0066*	*0.0234*	*0.0104*

Table 6: Simulated and approximated probability of delay in Erlang-R model with holding for Case 2.

R_1	$\beta = 1,\ \gamma = 1$		$\beta = 1,\ \gamma = 2$	
	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0977	0.0413	0.1521	0.0851
10	0.1070	0.0469	0.1648	0.1028
25	0.1926	0.1076	0.2421	0.1874
50	0.1431	0.0727	0.1876	0.1342
100	0.1855	0.1012	0.2282	0.1714
250	0.1775	0.0963	0.2217	0.1765
	0.1442	*0.0711*	*0.1981*	*0.1354*

R_1	$\beta = 2,\ \gamma = 1$		$\beta = 2,\ \gamma = 1$	
	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$	$\mathbb{P}(d)$	$\sqrt{R_1}\mathbb{E}[W]$
5	0.0072	0.0019	0.0250	0.0081
10	0.0067	0.0018	0.0235	0.0082
25	0.0148	0.0043	0.0325	0.0133
50	0.0092	0.0025	0.0217	0.0081
100	0.0132	0.0038	0.0277	0.0105
250	0.0114	0.0033	0.0246	0.0099
	0.0078	*0.0022*	*0.0188*	*0.0069*

Table 7: Simulated and approximated probability of delay in Erlang-R model with holding for Case 3.