The snowball effect of customer slowdown in critical many-server systems

Jori Selen*†, Ivo J.B.F. Adan*†, Vidyadhar G. Kulkarni‡, Johan S.H. van Leeuwaarden†

September 27, 2018

Abstract

Customer slowdown describes the phenomenon that a customer's service requirement increases with experienced delay. In healthcare settings, there is substantial empirical evidence for slowdown, particularly when a patient's delay exceeds a certain threshold. For such threshold slowdown situations, we design and analyze a many-server system that leads to a two-dimensional Markov process. Analysis of this system leads to insights into the potentially detrimental effects of slowdown, especially in heavy-traffic conditions. We quantify the consequences of underprovisioning due to neglecting slowdown, demonstrate the presence of a subtle bistable system behavior, and discuss in detail the snowball effect: A delayed customer has an increased service requirement, causing longer delays for other customers, who in turn due to slowdown might require longer service times.

1 Introduction

The phenomenon of customer slowdown describes the fact that a customer's service requirement increases with the customer's experienced delay. While the operations management literature is largely built on the assumption that service times are independent of delay, a growing number of empirical studies, predominantly in healthcare settings, provide evidence for situations where slowdown occurs. This empirical evidence calls for the development of stochastic models that take into account slowdown, in order to not only assess its impact on the performance of service operations, but also to gain understanding of the fundamental changes that slowdown brings to system behavior.

A large body within the healthcare operations literature investigates the impact of workload on service times of patients. A canonical example in this domain is the admission of patients to the intensive care unit (ICU). There is substantial empirical evidence for slowdown in such settings: delays in receiving appropriate care can result in adverse effects such as an increased length of stay in the ICU [7, 8, 9, 22, 23, 24]. Since ICUs are typically heavily used and subject to unforeseen circumstances, delays in admitting patients are the rule rather than the exception, which makes the slowdown effect potentially threatening. A delayed patient that requires a longer service time will increase the overall workload of the system, therefore causing longer delays for other patients, who in turn due to slowdown might require longer service. This triggers a *snowball* effect, with an impact that is hard to assess without having

*Department of Mechanical Engineering, Eindhoven University of Technology, The Netherlands
†Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands
‡Department of Statistics and Operations Research, University of North Carolina, Chapel Hill
E-mail address: j.selen@tue.nl

a detailed understanding of the global system behavior that takes into account the subtle dependencies among customers due to slowdown. Particularly when a system like an ICU is designed to operate under heavy-traffic conditions, the neglect of slowdown might lead to underprovisioning and severe performance degradation.

Critical care systems such as an ICU are typically modeled as multi-server systems that operate in heavy-traffic regimes [1, 15]. The patients are the customers, the beds are the servers, and the performance analysis of the multi-server systems gives insight into the patient flow. We shall consider a Markovian multi-server system with the additional feature of slowdown. A detailed analysis of this system gives insight into the key features of slowdown, in particular when compared against multi-server systems without slowdown.

1.1 A threshold slowdown system

Slowdown can be modeled as a non-increasing function $\mu(\cdot)$ that describes the rate of service as a function of the queue length seen upon arrival. That is, a customer meeting n customers upon arrival will receive service with rate $\mu(n)$, regardless of arrivals and departures after the customer has joined the system. Note that, since the number of customers seen on arrival can be translated into an expected delay, the service rate can also be interpreted as a non-increasing function of the expected delay. Assuming a service rate that is a function of the state of the system, leads to a so-called state-dependent queueing system.

The majority of the empirical studies on slowdown has focused on a threshold slowdown: if a patient's delay surpasses a certain threshold, he will receive a longer service time and otherwise he receives a service of regular length. In terms of the slowdown function, this means that $\mu(n) = \mu_{\mathrm{H}}$ if $n \leq N$ and $\mu(n) = \mu_{\mathrm{L}}$ if $n > N$ with $\mu_{\mathrm{H}} > \mu_{\mathrm{L}}$. Note that the expected delay is translated to a number of customers n met on arrival and compared against the threshold N. The definition of the threshold varies across different medical conditions and situations. In [7] it is argued that a critically ill patient awaiting transfer from the emergency department to the ICU is labeled as *delayed* if the patient has waited longer than 6 hours. Delayed patients on average have an ICU length of stay that is 1 full day longer than the non-delayed average length of stay. Similar conclusions are drawn in [23] for the same situation in different hospitals. However, [23] uses a threshold of 8 hours. Both studies [7, 23] establish a strong correlation between the delay a patient experiences in receiving an assigned bed and the ICU length of stay. Depending on the medical condition, the delay threshold can be in the order of minutes, such as for cardiac arrest patients [9], hours, as seen in [7, 23], or even days, such as the 2 day delay in receiving surgery [24]; or a 3 day threshold of delay for pneumonia patients [22]. An encompassing study is performed in [8], where it is empirically verified that the slowdown effect is prevalent across multiple hospitals and patient conditions.

We shall adopt the model in [8], which is a multi-server model with a threshold service rate function $\mu(\cdot)$. Customers arrive according to a Poisson process with rate λ, have an exponential service requirement, and are served by s servers. Due to the threshold, we then distinguish between two types of customers: those who were taken into service immediately upon arrival (non-delayed) and those who have experienced delay (delayed). We set the threshold to $N = s$ so that non-delayed customers are served with a high service rate μ_{H} and delayed customers are served with a low service rate μ_{L} with $\mu_{\mathrm{H}} > \mu_{\mathrm{L}}$. Indeed, in that case, delays cause a longer service time. Define the two-dimensional Markov process $(X(t), Y(t))$, with $X(t) \in \mathbb{N}_0$ the total number of customers in the system at time t, and $Y(t) \in \{0, \ldots, s\}$ the number of non-delayed customers in service at time t. Denote $\rho_{\mathrm{H}} = \lambda/(s\mu_{\mathrm{H}})$ and $\rho_{\mathrm{L}} = \lambda/(s\mu_{\mathrm{L}})$. The system is stable when $\rho_{\mathrm{L}} < 1$. When $\mu_{\mathrm{H}} = \mu_{\mathrm{L}}$ the model reduces to the standard $M/M/s$ system. The load of the slowdown system is described by $\rho = (1 - \mathbb{P}(W > 0))\rho_{\mathrm{H}} + \mathbb{P}(W > 0)\rho_{\mathrm{L}}$, where W is the

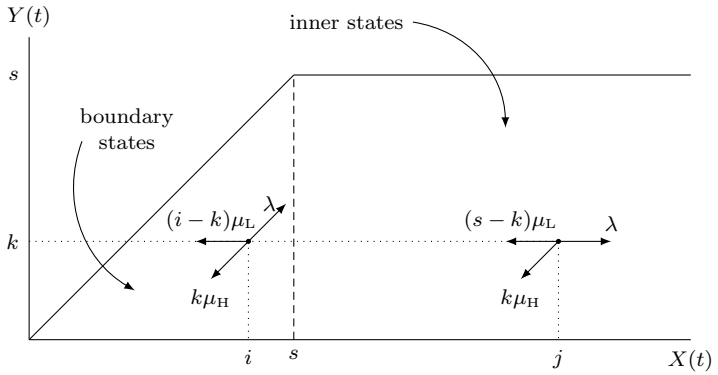


Figure 1: Transition rate diagram and state space of the threshold slowdown system.

stationary waiting time, so that $\mathbb{P}(W > 0)$ is the delay probability. Figure 1 displays the state space and the transition rate diagram. In [8], approximations are derived for key performance indicators that give insight into the slowdown effect. Based on parameter values calibrated from real ICU dataflows, the approximations in [8] indicate that the slowdown effect can be substantial, and should not be ignored in critical care systems that operate in heavy traffic.

This two-dimensional Markov process can be used to investigate the impact of slowdown, both qualitatively and quantitatively, in particular in comparison with the widely applied $M/M/s$ system (which neglects slowdown). While the focus in [8] lies on approximations for small to moderate-sized systems (ICUs of 6 and 15 beds), we focus on exact and asymptotic results, both for finite s and the regime $s \to \infty$ and $\rho \uparrow 1$. With exact results we refer to determining the stationary distribution of the Markov process using a numerically stable algorithm. This algorithm allows us to compute the exact two-dimensional stationary distribution for not only small but also large systems. Asymptotic results give rise to accurate approximations for the dimensioning of large systems in heavy traffic. Particularly, we are interested in the effect of slowdown in the Quality-and-Efficiency driven (QED) regime [16]. As it turns out, the way to establish non-degenerate limiting behavior for a multi-server system with slowdown in a QED-type regime is by letting ρ_{L} approach 1, and μ_{H} approach μ_{L} as $s \to \infty$. We find that this scaling window is such that the probability of delay converges to a value that lies strictly in the interval $(0, 1)$, which is a manifestation of non-degenerate limiting behavior. As pointed out in [8], deriving exact results becomes mathematically challenging because determining the stationary distribution of the Markov process involves high-dimensional matrix inversion. To relieve this computational burden of a large state space (particularly for large s), we exploit the fact that the Markov process has a block diagonal structure in the *inner* states (states with more than s customers in the system), which allows for an exact solution using matrix-analytic techniques. This technique typically relies on iterative algorithms that solve a non-linear matrix equation. For our model, we are able to find an exact solution for this matrix equation, which then immediately renders the problem of computing the stationary probabilities of the inner states computationally tractable, also for large s, see Section 4.1. What remains is the computation of the stationary probabilities of the *boundary* states (states with s or less customers in the system). We introduce a novel approach that computes the exact stationary probabilities of the boundary states by exploiting the transition structure and by introducing first-passage probabilities. A detailed description of this approach can be found in Section 4.2.

1.2 On the relation with operator slowdown

Slowdown can refer to *customer* slowdown and *operator* slowdown. Customer slowdown refers to an increase of a customer's service requirement, caused by the delay experienced by that customer. Operator slowdown refers to a service rate that decreases with the workload present in the system. Operator slowdown usually occurs in large service systems, such as call centers, due to fatigued operators [10]. However, it is also common in medical applications under high workload, where care providers have to multitask and share (now crowded) central resources such as computer terminals [3]. The key difference is that customer slowdown starts with an individual delayed customer, and affects all customers behind this customer, while a decreased service rate in operator slowdown affects all customers that are in service. Customer slowdown therefore typically requires a more detailed state description, making it harder to analyze than operator slowdown. In this paper we indeed focus on customer slowdown, but we make comparisons with operator slowdown in several places.

Operator slowdown under Markovian assumptions leads to a one-dimensional Markov process which is more tractable than our two-dimensional process and is amenable to fluid analysis. In [10] an $M/M/s$-type model with operator slowdown is investigated. Additional properties in [10] are customer abandonments and state-dependent service rates. We make a comparison with [10] by extending our base model to also include customer abandonments in Section 2.2. Both slowdown models exhibit a bistable behavior in which the models alternate between two dominant regions. For the customer slowdown model, however, this behavior is more subtle than for the operator slowdown model (see Section 2.2).

Both customer and operator slowdown fall into the broad category of queueing systems with state-dependent service rates, like for instance an $M/G/1$ system with state-dependent service rates [17]. In [4], the optimal admission policy is studied for an $M/G/1$ system with service rates that increase with the workload below a certain threshold and decrease with the workload above this threshold. State-dependent queueing systems also arise when arrival and/or service rates are dynamically controlled to minimize average cost per time unit, see e.g. [2, 13, 28]. All these examples concern operator slowdown.

1.3 Structure of the paper

The paper is structured as follows. Based on a detailed analysis of the two-dimensional Markov process in Figure 1, we identify three key features of threshold slowdown systems: severe performance degradation due to the snowball effect; a subtle bistable system behavior; and the existence of non-degenerate limiting behavior in a QED-type heavy-traffic regime. We discuss these three features in Section 2. The first two features were identified by using the stationary distribution of the two-dimensional Markov process. Section 3 introduces the model in greater detail and Section 4 describes how we solve for its stationary distribution using matrix-analytic methods and some properties of regenerative processes. The QED-type heavy-traffic regime is outlined in Section 5. We conclude in Section 6 and present some supporting results in the appendix.

2 Key features of threshold slowdown systems

Unless stated otherwise, we assume a stable system, i.e. $\rho_{\mathrm{L}} < 1$.

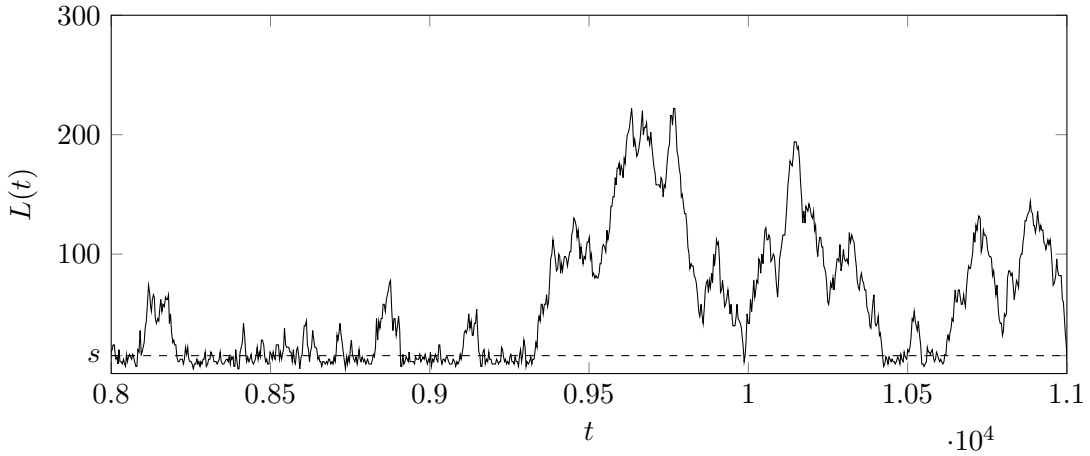


Figure 2: Sample path of the total number of customers in the system $L(t)$ as a function of time t, for $s = 15$, $\lambda = s$ and loads $\rho_{\mathrm{H}} = 0.7$ and $\rho_{\mathrm{L}} = 0.98$.

2.1 Performance degradation due to the snowball effect

We first present a detailed description of the snowball effect caused by slowdown and then assess the adverse effects for system performance.

For explanation purposes, we refer with *busy* periods and *idle* periods to the excursions of the process $X(\cdot)$ above and at level s, and below level s, respectively. Hence, during busy periods, newly arriving customers will experience delay and are thus subject to slowdown. The snowball effect sets in each time a new busy period starts. An example sample path of idle and busy periods is given in Figure 2, where we plot the total number of customers in the system $L(t)$ at time t. Compared with an $M/M/s$ system without slowdown (with a high service rate μ_{H}), the busy period in the time interval $(9300, 10000)$ is relatively long, due to the slowdown of delayed customers that reinforces, through other delayed customers, the persistence of the busy period. Such busy periods are essentially equivalent to busy periods in an $M/M/s$ system with a low service rate μ_{L}. These excursions during which congestion levels are high occur relatively frequently due to the snowball effect that triggers them, and this leads to severe performance degradation, particularly in heavy traffic.

This performance degradation is visible in Figure 3, which displays for the same parameter values as in Figure 2 the stationary distribution of the total number of customers in the system L of the threshold slowdown system. This stationary distribution is calculated using the numerical scheme that will be discussed in Section 4. We also plot the stationary distribution of an $M/M/s$ system with uniform service rate μ_{H} (the *fast* system) and with uniform service rate μ_{L} (the *slow* system). We append the subscripts H, or L to random variables to indicate that they belong to the $M/M/s$ system with high service rate, or low service rate, respectively. We see that the distribution of the slowdown system peaks around the same point as the fast system, but that the tail behavior of the slowdown system is more comparable to the slow system (which can be attributed to the snowball effect and long busy periods). Such a fat tail obviously has severe consequences for performance, and in Figure 3 we see for instance that the mean number of customers in the system increases considerably due to slowdown.

The neglect of slowdown might lead to underprovisioning. Table 1 provides an example in which we search for the number of servers s that are required to achieve a certain delay probability. Naturally, the threshold slowdown model requires equally many or more servers as required by the fast system with uniform service rate μ_{H}. In particular, differences between

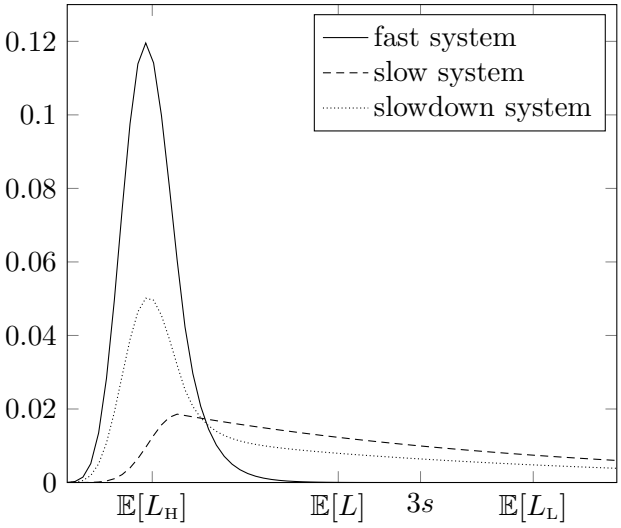


Figure 3: Stationary distributions of the total number of customers in the system L for systems with and without slowdown. Parameter values are $s = 15$, $\lambda = s$, $\rho_{\mathrm{H}} = 0.7$ and $\rho_{\mathrm{L}} = 0.98$. The expected number of customers in the systems are $\mathbb{E}[L_{\mathrm{H}}] \approx 10.8$, $\mathbb{E}[L_{\mathrm{L}}] \approx 59.4$ and $\mathbb{E}[L] \approx 34.5$.

μ_{H}	μ_{L}	λ	$\mathbb{P}(W > 0)$	s^*_{H}	s^*	$\mathbb{P}(W > 0)$	s^*_{H}	s^*
1	0.9	10	0.1	16	16	0.5	12	13
		12		18	18		14	15
		15		22	22		18	19
		20		27	28		23	24
1	0.7	10	0.1	16	17	0.5	12	15
		12		18	19		14	18
		15		22	23		18	22
		20		27	30		23	29

Table 1: Minimal number of servers s^*_{H} (fast system) and s^* (threshold slowdown system) required to achieve a certain $\mathbb{P}(W > 0)$.

the required number of servers in the fast model and the threshold slowdown model seem to increase with the delay probability, with the ratio $\mu_{\mathrm{H}}/\mu_{\mathrm{L}}$ and with the arrival rate λ.

Another indicator for substantial slowdown effect is the difference $\rho - \rho_{\mathrm{H}}$, as we will show in the next subsection. This difference is the increase in load caused by the slowdown effect with respect to the load of the fast system.

2.2 A subtle bistable behavior

The threshold slowdown system behaves as the fast system below the threshold, and as the slow system above the threshold. However, for many relevant parameter settings, neither the fast nor the slow system provides a good approximation for the slowdown system. The reason is that the slowdown system in fact is a rather intricate mixture of both system as will be explained in this subsection.

We start by examining the two-dimensional stationary distribution, which typically consists of two dominant regions: region 1 with only non-delayed customers and no customers waiting, and region 2 with delayed (slowdown) customers in service only and many waiting customers.

(a) $\rho_{\mathrm{H}} = 0.6$, $\rho - \rho_{\mathrm{H}} = 0.123$ and $\mathbb{P}(W > 0) = 0.32$.

(b) $\rho_{\mathrm{H}} = 0.95$, $\rho - \rho_{\mathrm{H}} = 0.027$ and $\mathbb{P}(W > 0) = 0.90$.

Figure 4: For parameter settings that are mild or extreme the slowdown system resembles either the fast or the slow system. The horizontal axis plots the total number of customers in the system and the vertical axis indicates the number of non-delayed customers in service. The contour plot shows where the probability mass is located (darker colour means more mass). Parameter values are $s = 15$ and $\rho_{\mathrm{L}} = 0.98$.

Region 1 thus complies with the fast system and region 2 with the slow system. An important parameter that determines whether region 1 or region 2 is dominant is ρ_{H}. A low to moderate ρ_{H} makes region 1 dominant, which suggests using the fast system as a proxy. A high load ρ_{H} makes region 2 more important, and in fact, when ρ_{H} approach 1, the slow system will be a good approximation. See the two examples in Figure 4. Notice here that for a system with a high delay probability, i.e. Figure 4(b), the increase in load $\rho - \rho_{\mathrm{H}}$ due to the slowdown effect is small, since both loads ρ_{H} and ρ_{L} are large and comparable. In contrast, the increase in load in Figure 4(a) is much larger.

Arguably the most natural scenario, when ρ_{H} is high but not extremely high, say $\rho_{\mathrm{H}} \in (0.7, 0.9)$, gives a less clean picture. Then the slowdown system is a mixture of the fast and slow systems, under the right condition that ρ_{L} is decisively larger than ρ_{H}. A good example is $\rho_{\mathrm{H}} = 0.8$ and $\rho_{\mathrm{L}} = 0.98$, as can be seen in Figure 5(b). By increasing the load ρ_{L} busy periods become longer, causing the shift in probability mass towards region 2 and increasing the severity of the slowdown effect in terms of $\rho - \rho_{\mathrm{H}}$ as is witnessed in Figures 5(a)-(b).

Our slowdown system thus has a subtle bistable behavior, which rises to the surface when both ρ_{H} and $\rho_{\mathrm{L}} - \rho_{\mathrm{H}}$ are substantial but not extreme. A more extreme bistability effect would occur when ρ_{L} could become larger than 1. We therefore next discuss two extensions of our model that allow for $\rho_{\mathrm{L}} \geq 1$:

(i) A threshold slowdown system with a finite waiting room;

(ii) A threshold slowdown system with customer abandonments.

System (i) can have at most N customers in the system and is therefore inherently stable. When $\rho_{\mathrm{L}} \geq 1$ and ρ_{H} is sufficiently small the system will alternate between periods during which

(a) $\rho_{\mathrm{L}} = 0.9$, $\rho - \rho_{\mathrm{H}} = 0.047$ and $\mathbb{P}(W > 0) = 0.46$.

(b) $\rho_{\mathrm{L}} = 0.98$, $\rho - \rho_{\mathrm{H}} = 0.144$ and $\mathbb{P}(W > 0) = 0.80$.

Figure 5: Two dominant regions in the stationary distribution become visible when the load of the delayed customers ρ_{L} increases. Parameter values are $s = 15$ and $\rho_{\mathrm{H}} = 0.8$.

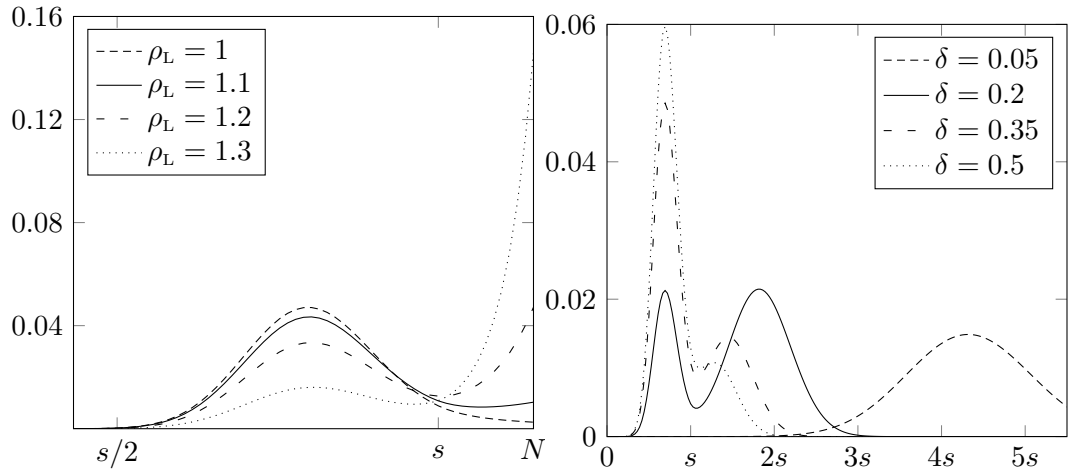
the process settles in the high-occupancy states around N, and periods in low-occupancy states below s. This gives rise to bistable behavior, and for some parameter ranges even leads to a bimodal distribution as seen in Figure 6(a). This bimodality can be explained by the fact that for $\rho_{\mathrm{L}} \geq 1$, the process has two clear points of attraction: the state N and the state $\rho_{\mathrm{H}} s$ where the rate of arriving and departing customers is equal. Note that our original slowdown system has only one point of attraction, because $\rho_{\mathrm{H}} < \rho_{\mathrm{L}} < 1$.

System (ii) assumes that waiting customers abandon the system after an exponential time with mean $1/\delta$. Because the total abandonment rate scales linearly with the number of waiting customers, also this system is inherently stable. For $\rho_{\mathrm{L}} \geq 1$ it has two points of attraction: one below s, and one above s precisely where the total rate of arriving customers equals the rate of departing (abandoning and served) customers. For $\rho_{\mathrm{L}} \geq 1$ this process alternates between the two points of attraction as is shown in Figure 6(b). This system is closely related to the operator slowdown system with abandoning customers considered in [10]. In [10], the bistability effect was also observed, where the two points of attraction were identified explicitly. Explicitly characterizing the two points of attraction in the customer slowdown model is more difficult due to the two-dimensional nature of the system.

2.3 Scaling limits

So far we increase either the number of servers or the arrival rate. We continue by examining a scaling of both parameters at the same time. It is well known in the literature that for $G/M/s$ queues, one should scale the arrival rate or the number of servers such that the load $\rho^{(s)} \sim 1 - \beta/\sqrt{s}$ as $s \to \infty$ to achieve QED performance [16]. In terms of our model parameters, the scaling is then as follows:

$$\lambda^{(s)} = s\mu_{\mathrm{L}}(1 - \beta/\sqrt{s}), \tag{2.1}$$



(a) A finite buffer that limits the total number of customers in the system to N. Here, $s = 81$, $N = 93$ and $\rho_{\mathrm{H}} = 0.8$.

(b) Allowing customers to abandon the queue with rate δ. Here, $s = 36$, $\rho_{\mathrm{H}} = 0.7$ and $\rho_{\mathrm{L}} = 1.2$.

Figure 6: Two extensions of the multi-server system with slowdown that exhibit, for a narrow range of parameter values, a bistability effect that is visible in the bimodal marginal distribution of the total number of customers in the system.

with constant $\beta > 0$ and $s > \beta^2$ to guarantee a positive arrival rate $\lambda^{(s)}$. By applying (2.1) to our multi-server system with slowdown one finds that we establish so-called Quality-Driven (QD) performance. QD performance refers to a very high quality of service, e.g. the probability of delay goes to 0 and many servers are idle. This might be undesirable in view of unnecessary operational costs (overdimensioning). The reason for QD performance is that since $\mu_{\mathrm{H}} > \mu_{\mathrm{L}}$ we have $\rho_{\mathrm{H}} < 1$ in the limit for $s \to \infty$. This ensures that the system stabilizes around a state with relatively low occupancy and with only non-delayed customers in service and no customers waiting in the system. To obtain QED system behavior we set the high service rate according to

$$\mu_{\mathrm{H}}^{(s)} = \mu_{\mathrm{L}}(1 + \gamma/\sqrt{s}), \tag{2.2}$$

with constant $\gamma > 0$. Note that now $\mu_{\mathrm{H}}^{(s)}/\mu_{\mathrm{L}} \to 1$ for $s \to \infty$ and thus $\rho_{\mathrm{H}}^{(s)}$ also goes to 1. We refer to the combination of (2.1) and (2.2) as a QED-type regime. The reason for this choice of scaling becomes clear when we examine the load of the slowdown system with s servers

$$\rho^{(s)} = \Big(1 - \frac{\beta}{\sqrt{s}}\Big)\frac{1 + \mathbb{P}(W^{(s)} > 0)\frac{\gamma}{\sqrt{s}}}{1 + \frac{\gamma}{\sqrt{s}}}, \tag{2.3}$$

which shows that $\rho^{(s)} \uparrow 1$ as $s \to \infty$. Compared to the standard scaling of the load in $G/M/s$ queues, the load in the customer slowdown model approaches 1 slower as it is multiplied by the second term in (2.3). Figure 7 depicts the probability of delay as a function of s, which indeed shows that the probability of delay converges to a value in $(0, 1)$.

Using stochastic coupling techniques, we related the two-dimensional process to two one-dimensional processes that serve as a stochastic lower and upper bound (at the process level in terms of stochastic domination). These two related processes are the fast and the slow system introduced earlier. The bounding processes provide sharp approximations for the two-dimensional process. For both bounding processes, we show that in the QED-type regime, the

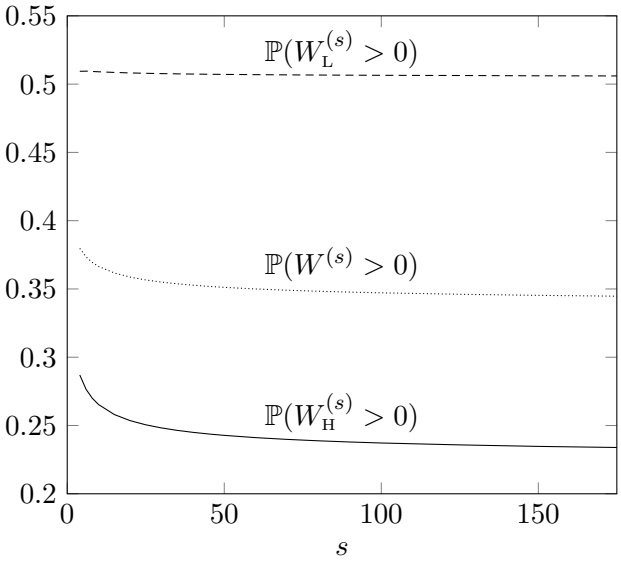


Figure 7: Probability to wait for the fast, slow, and customer slowdown system. For all three systems, the scaling (2.1) and (2.2) is used with $(\beta, \gamma) = (0.5, 0.5)$.

probability of delay converges to a value strictly in between 0 and 1, and this then also holds for the two-dimensional process. Hence, this provides strong evidence for the existence of a non-trivial stochastic-process limit. Formally establishing the existence and characterizing this stochastic-process limit is a challenging open problem, because the limiting process is likely to be two-dimensional as well, and classical techniques to prove stochastic-process limits [29] do not seem to carry over easily.

2.4 Insights

Here we summarize the insights obtained in this section.

Customer slowdown of the threshold type leads to severe performance degradation, particularly in heavy traffic. Compared to a system without slowdown, the busy periods are relatively long due to the slowdown of delayed customers that reinforces the persistence of the busy period. We refer to this effect as the snowball effect, which describes the correlated service times when customers are delayed. Further, for a relatively high load ρ_{H}, we find that the threshold slowdown system is a mixture of the fast and slow systems. This mixture effect is visible in the two-dimensional stationary distribution, where it manifests itself as two dominant regions in terms of probability mass – a subtle bistable effect. Finally, by using a QED-type scaling for the arrival rate and the fast service rate μ_{H}, we have shown that a non-degenerate limit behavior occurs as the number of servers increases.

3 Model description

Recall that $X(t) \in \mathbb{N}_0$ is the total number of customers in the system at time t and $Y(t) \in \{0, 1, \ldots, s\}$ is the number of non-delayed customers in the system at time t. Note that $X(t) \geq Y(t)$. Then, $\{(X(t), Y(t)),\ t \geq 0\}$ is an irreducible continuous-time Markov chain with discrete state space $V \cup W$, given by

$$V = \{(i,j) \mid 0 \leq i < s,\ 0 \leq j \leq i\}, \quad W = \{(i,j) \mid i \geq s,\ 0 \leq j \leq s\}. \tag{3.1}$$

Recall that we refer to the states with s or less customers in the system, as the *boundary* states. With *inner* states we refer to the states with more than s customers in the system. For an inner state (i, j) with $i > s$, we have the following transition rates:

- From (i, j) to $(i + 1, j)$ with rate λ, $0 \leq j \leq s$;
- From (i, j) to $(i - 1, j)$ with rate $(s - j)\mu_{\mathrm{L}}$, $0 \leq j \leq s$;
- From (i, j) to $(i - 1, j - 1)$ with rate $j\mu_{\mathrm{H}}$, $1 \leq j \leq s$.

The transition rate diagram of the continuous-time Markov chain is shown in Figure 1.

Define level i as the set of all states with a total of i customers in the system. Now we have the following alternative description of the transition rates. The matrices Λ_k contain the transition rates from level i to level $i + k$ with $i > s$. Let I be the identity matrix of size $s + 1$. Then the matrices Λ_k are given by $\Lambda_1 = \lambda I$,

$$\Lambda_0 = -\begin{pmatrix} \lambda + s\mu_{\mathrm{L}} & & & \\ & \lambda + (s-1)\mu_{\mathrm{L}} + \mu_{\mathrm{H}} & & \\ & & \ddots & \\ & & & \lambda + s\mu_{\mathrm{H}} \end{pmatrix}, \tag{3.2}$$

and

$$\Lambda_{-1} = \begin{pmatrix} s\mu_{\mathrm{L}} & & & & \\ \mu_{\mathrm{H}} & (s-1)\mu_{\mathrm{L}} & & & \\ & 2\mu_{\mathrm{H}} & \ddots & & \\ & & \ddots & \mu_{\mathrm{L}} & \\ & & & s\mu_{\mathrm{H}} & 0 \end{pmatrix}. \tag{3.3}$$

By assumption $\rho_{\mathrm{H}} < \rho_{\mathrm{L}}$, and we have the following condition for ergodicity of the Markov process.

Lemma 3.1. *The Markov process is ergodic if and only if*

$$\rho_{\mathrm{L}} < 1. \tag{3.4}$$

Proof. We require that the mean drift in the negative direction is larger than the mean drift in the positive direction; see Neuts' mean drift condition [21, Theorem 1.7.1]. This gives

$$\boldsymbol{\pi}\Lambda_1\mathbf{1} < \boldsymbol{\pi}\Lambda_{-1}\mathbf{1}, \tag{3.5}$$

where $\mathbf{1}$ is a column vector of ones of size $s + 1$, $\boldsymbol{\pi}$ is the solution of $\boldsymbol{\pi}\sum_{k=-1}^{1}\Lambda_k = 0$ with $\boldsymbol{\pi}\mathbf{1} = 1$. We clearly have $\boldsymbol{\pi} = (1, 0, \ldots, 0)$ and thus the result follows. □

4 Obtaining the stationary distribution

Assume that (3.4) holds and define the stationary probabilities

$$p(i, j) \coloneqq \lim_{t\to\infty} \mathbb{P}(X(t) = i, Y(t) = j), \quad (i, j) \in V \cup W. \tag{4.1}$$

The balance equations for the inner states are obtained by equating the rate out of and into an inner state (i,j), yielding, for $i > s,\ 0 \le j \le s$,

$$\begin{aligned}(\lambda + j\mu_{\rm H} + (s-j)\mu_{\rm L})p(i,j) = \lambda p(i-1,j) + (s-j)\mu_{\rm L}p(i+1,j)\\ + (j+1)\mu_{\rm H}p(i+1,j+1),\end{aligned} \tag{4.2}$$

where by convention $p(i,s+1) = 0$. Equations (4.2) are referred to as the inner equations. The balance equations for states with $i \le s$ are called the boundary equations.

The stationary probabilities of the inner states are determined using matrix-analytic methods that search for the solution to a non-linear matrix equation. Exploiting structural properties of the Markov process, we derive explicit solutions for these matrices. For similar explicit results using the matrix-analytic methods, see [25, 26, 27]. Next, we solve the boundary equations. Since we want to be able to solve the stationary distribution also for large s, solving the $(s+1)(s+2)/2$ boundary equations using Gaussian elimination might become computationally cumbersome. We therefore present a more sophisticated algorithm that exploits the structure of the state space and the explicit matrix solution.

4.1 Inner equations

Let $\mathbf{p}_i = (p(i,0), p(i,1), \ldots, p(i,s))$, and rewrite the inner balance equations as

$$\mathbf{p}_{i-1}\Lambda_1 + \mathbf{p}_i\Lambda_0 + \mathbf{p}_{i+1}\Lambda_{-1} = 0, \quad i > s. \tag{4.3}$$

The rate matrix R is defined as the minimal non-negative solution of the non-linear matrix equation [21, Theorem 3.1.1]

$$\Lambda_1 + R\Lambda_0 + R^2\Lambda_{-1} = 0. \tag{4.4}$$

It can be shown that the stationary probabilities satisfy

$$\mathbf{p}_i = \mathbf{p}_s R^{i-s}, \quad i \ge s. \tag{4.5}$$

Since the transition matrices are all lower triangular, so is the rate matrix R. Denote

$$R = \begin{pmatrix} R_{0,0} & & & \\ R_{1,0} & R_{1,1} & & \\ \vdots & & \ddots & \\ R_{s,0} & \cdots & & R_{s,s} \end{pmatrix} \tag{4.6}$$

and note that R^2 is again a lower triangular matrix with elements $(R^2)_{i,j} = \sum_{k=j}^{i} R_{i,k}R_{k,j}$ for $i \ge j$.

Equation (4.4) consists of $(s+1)^2$ separate equations. For the diagonal elements we have

$$\lambda - (\lambda + (s-j)\mu_{\rm L} + j\mu_{\rm H})R_{j,j} + (s-j)\mu_{\rm L}R_{j,j}^2 = 0, \quad 0 \le j < s, \tag{4.7}$$

$$\lambda - (\lambda + s\mu_{\rm H})R_{s,s} = 0, \quad j = s, \tag{4.8}$$

where $R_{j,j}$ in (4.7) is obtained as the minimal non-negative solution. The minimal non-negative solution of (4.7) is contained in the interval $(0,1)$, because for $R_{j,j} = 0$ the left-hand side of (4.7) is positive, for $R_{j,j} = 1$ the left-hand side of (4.7) is negative, and we are dealing with a continuous function. Interestingly, $R_{0,0} = \rho_{\rm L}$ and $R_{j,j}$ is monotonically decreasing in j. For each element on the subdiagonals we have a linear equation with solution

$$R_{i,j} = \frac{\sum_{k=j+1}^{i-1} R_{i,k}R_{k,j}(s-j)\mu_{\rm L} + \sum_{k=j+1}^{i} R_{i,k}R_{k,j+1}(j+1)\mu_{\rm H}}{\lambda + (s-j)\mu_{\rm L} + j\mu_{\rm H} - \big(R_{i,i} + R_{j,j}\big)(s-j)\mu_{\rm L}}, \tag{4.9}$$

for $j = i - h,\ h \leq i \leq s$ and $h = 1, 2, \ldots, s$. In (4.9) we use the convention that $\sum_{i=i_0}^{i_1} f(i) = 0$ if $i_1 < i_0$. Equations (4.7)-(4.9) fully describe the rate matrix R.

Recall that a lower triangular matrix is non-singular if it has all non-zero elements on the diagonal. Thus, the matrix R is non-singular and also $I - R$ is non-singular. The inverse of $I - R$ is required to compute the stationary probabilities, as the normalization of the stationary distribution partially follows from $\mathbf{p}_s(I + R + R^2 + \cdots)\mathbf{1} = \mathbf{p}_s(I - R)^{-1}\mathbf{1}$. The elements of the inverse are given by

$$((I-R)^{-1})_{j,j} = \frac{1}{(I-R)_{j,j}}, \quad 0 \leq j \leq s, \tag{4.10}$$

$$((I-R)^{-1})_{i,j} = \frac{-\sum_{k=j}^{i-1}(I-R)_{i,k}((I-R)^{-1})_{k,j}}{(I-R)_{i,i}}, \quad 0 \leq j < i \leq s. \tag{4.11}$$

Instead of searching for R, one can also search for the matrix G, defined as the minimal non-negative solution of the non-linear matrix equation

$$\Lambda_{-1} + \Lambda_0 G + \Lambda_1 G^2 = 0. \tag{4.12}$$

The matrices R and G are related as $\Lambda_1 G = R\Lambda_{-1}$ and thus $G = \Lambda_1^{-1} R \Lambda_{-1}$, which exists since Λ_1 is a diagonal matrix.

4.2 Boundary equations

The boundary equations can be represented as a set of $(s+1)(s+2)/2$ linear equations, which can be solved using Gaussian elimination with an arithmetic complexity of $\mathrm{O}(s^6)$ [12, Chapter 10]. By exploiting the structure of the boundary equations one can reduce the arithmetic complexity to $\mathrm{O}(s^4)$. In short, we wish to embed the Markov process on level s for which we need the return probabilities when jumping to a higher level (the matrix G), combined with the return probabilities when jumping to a level below (yet to be determined). This allows us to compute the non-normalized stationary probabilities of the states in level s. Then, we recursively compute the remaining boundary probabilities starting from level $s-1$, working our way down to level 0, leading to a non-normalized stationary distribution. Finally, the normalized stationary distribution follows by summing over all states and dividing the non-normalized version of the stationary distribution by the resulting sum.

To this end we introduce two first passage probabilities. To aid the derivation of these probabilities we introduce subsets of V, indexed by a state $(k, l) \in V$. We identify the triangular set of states $T_{(k,l)}$ to the South-West of the state (k, l), specifically, $T_{(k,l)} \coloneqq \{(i, j) \mid k - l \leq i \leq k - 1,\ 0 \leq j \leq i - (k - l)\}$, see Figure 8.

Let $\theta_k(i, j)$ be the first passage probability to phase $j + 1$ in state $(i + 1 - k, j + 1)$, where the Markov process starts in state $(i, j) \in T_{(s,s-k)}$. Note that by phase j we refer to the set of states with $Y(t) = j$. By one-step analysis we obtain

$$\theta_0(i,j) = \frac{\lambda \cdot 1 + (i-j)\mu_{\mathrm{L}} \cdot 0 + j\mu_{\mathrm{H}}\theta_0(i-1,j-1)\theta_0(i,j)}{\lambda + (i-j)\mu_{\mathrm{L}} + j\mu_{\mathrm{H}}}, \quad (i,j) \in T_{(s,s)}, \tag{4.13}$$

$$\theta_k(i,j) = \frac{\lambda \cdot 0 + (i-j)\mu_{\mathrm{L}}\theta_{k-1}(i-1,j) + j\mu_{\mathrm{H}}\sum_{l=0}^{k}\theta_l(i-1,j-1)\theta_{k-l}(i-l,j)}{\lambda + (i-j)\mu_{\mathrm{L}} + j\mu_{\mathrm{H}}}, \quad (i,j) \in T_{(s,s-k)},\ k > 0, \tag{4.14}$$

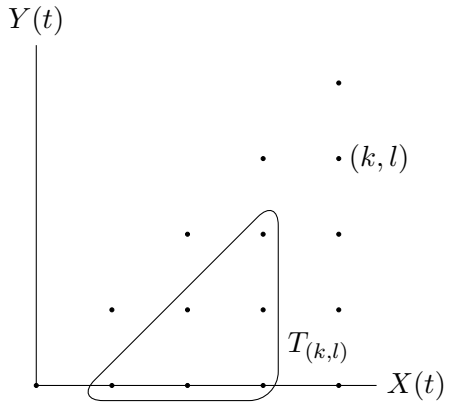


Figure 8: A visual clarification of the triangular set of states $T_{(k,l)}$.

from which the following expressions can be readily derived,

$$\theta_0(i,j) = \frac{\lambda}{\lambda + (i-j)\mu_{\rm L} + j\mu_{\rm H}(1-\theta_0(i-1,j-1))}, \quad (i,j) \in T_{(s,s)}, \tag{4.15}$$

$$\theta_k(i,j) = \frac{(i-j)\mu_{\rm L}\theta_{k-1}(i-1,j) + j\mu_{\rm H}\sum_{l=1}^{k}\theta_l(i-1,j-1)\theta_{k-l}(i-l,j)}{\lambda + (i-j)\mu_{\rm L} + j\mu_{\rm H}(1-\theta_0(i-1,j-1))}, \quad (i,j) \in T_{(s,s-k)},\ k>0. \tag{4.16}$$

Note that $\theta_0(i,i) = 1$, which means that if the Markov process reaches a state on the main diagonal, it eventually always reaches state (s,s).

Let $\psi_{(k,l)}(i,j)$ be the first passage probability to level k in state (k,l), where the Markov process starts from state $(i,j) \in T_{(k,l)}$. We express these first passage probabilities in terms of $\theta_k(i,j)$ as follows

$$\psi_{(k,l)}(i,j) = \sum_{m=j+1}^{i+1} \theta_{i+1-m}(i,j)\psi_{(k,l)}(m,j+1), \quad (i,j) \in T_{(k,l)}. \tag{4.17}$$

The computation of the first passage probabilities ψ is the most time consuming step in the derivation of the boundary probabilities. Equation (4.17) is evaluated for a total of $s(s+1)(s+2)(s+3)/24$ combinations of (i,j) and (k,l), leading to an arithmetic complexity of $\mathrm{O}(s^4)$.

Let Ψ be the matrix of elements $\psi_{(s,k)}(s-1,j)$ where j is the row number and k the column number. The balance equations of the Markov process embedded at level s are

$$\mathbf{p}_s\big(\Lambda_{-1}\Psi + \Lambda_0 + \Lambda_1 G\big) = 0. \tag{4.18}$$

One solves this homogeneous set of equations using the numerically stable Grassmann version of the Gaussian elimination algorithm [14] to obtain the stationary probabilities at level s. This solution is not normalized.

In order to obtain the remaining boundary probabilities, one embeds the Markov process on the levels $i, i+1, \ldots, s, \ldots$ with $i < s$, for which we derive the following balance equations:

$$\begin{aligned} &p(i,j)\big(\lambda + (i-j)\mu_{\rm L} + j\mu_{\rm H}(1-\theta_0(i-1,j-1))\big) \\ &= p(i+1,j)(i+1-j)\mu_{\rm L} + p(i+1,j+1)(j+1)\mu_{\rm H} \\ &\quad + \sum_{k=0}^{j-1} p(i,k)(i-k)\mu_{\rm L}\psi_{(i,j)}(i-1,k) + \sum_{k=1}^{j-1} p(i,k)k\mu_{\rm H}\psi_{(i,j)}(i-1,k-1), \quad j \le i. \end{aligned} \tag{4.19}$$

One recursively solves (4.19) by first computing $p(i,0)$, followed by $p(i,1)$, et cetera, until $p(i,i)$ is computed. Then, the remaining boundary probabilities follow by solving (4.19) for $i = s-1, s-2, \ldots, 1$. Finally, the probability of having an empty system is found by examining the balance equation in state $(0,0)$, so that

$$p(0,0) = (p(1,0)\mu_{\mathrm{L}} + p(1,1)\mu_{\mathrm{H}})/\lambda. \tag{4.20}$$

Recall that the stationary probabilities of level s are not normalized. Thus, the obtained stationary distribution of the boundary and inner states (obtained through (4.5)) are yet to be normalized. The normalized stationary distribution follows by dividing each non-normalized stationary probability by the sum over all states $\sum_{(i,j)\in V} p(i,j) + \mathbf{p}_s(I-R)^{-1}\mathbf{1}$.

Using the stationary distribution, one can now obtain performance measures such as the delay probability

$$\mathbb{P}(W > 0) = \sum_{i=0}^{\infty} \mathbf{p}_{s+i}\mathbf{1} = \mathbf{p}_s(I-R)^{-1}\mathbf{1}, \tag{4.21}$$

or the mean queue length

$$\mathbb{E}[Q] = \sum_{i=0}^{\infty} i\mathbf{p}_{s+i}\mathbf{1} = \mathbf{p}_s R(I-R)^{-2}\mathbf{1}. \tag{4.22}$$

4.3 Extensions

We next describe how to obtain the stationary distribution of the slowdown model with (i) a finite buffer; or (ii) customer abandonments.

4.3.1 Finite buffer

The transition rate diagram of the slowdown system with a finite buffer is identical to the one shown in Figure 1 but truncated at level N. Recall that we defined the matrices Λ_k to contain the transition rates from level i to level $i+k$; since we now introduced the finite buffer, we restrict i as $s < i \leq N$. At level N, the matrix containing the transitions rates to level $N-1$ remains unchanged and is still Λ_{-1}. The only difference is that there are no jumps in the positive direction and thus Λ_1 does not exist and therefore the matrix Λ_0 changes at level N, now indexed by an additional subscript N and given by $\Lambda_{0,N} = \Lambda_0 + \Lambda_1$.

The equilibrium equations in vector-matrix form are given by

$$\mathbf{p}_{i-1}\Lambda_1 + \mathbf{p}_i\Lambda_0 + \mathbf{p}_{i+1}\Lambda_{-1} = 0, \quad s < i < N, \tag{4.23}$$

$$\mathbf{p}_{N-1}\Lambda_1 + \mathbf{p}_N\Lambda_{0,N} = 0. \tag{4.24}$$

We now have the following relation, see [11, Section 2.2],

$$\mathbf{p}_i = \mathbf{p}_{i-1}R_i, \quad s < i \leq N, \tag{4.25}$$

where R_i is a level-dependent rate matrix with identical interpretation as the standard rate matrix of the matrix-geometric approach. One can now solve for the rate matrix R_N by manipulating (4.24) as follows

$$\mathbf{p}_N = -\mathbf{p}_{N-1}\Lambda_1\big(\Lambda_{0,N}\big)^{-1} = \mathbf{p}_{N-1}R_N. \tag{4.26}$$

Note that $\Lambda_{0,N}$ is a diagonal matrix with negative elements and is therefore indeed non-singular. The remaining rate matrices are found from (4.23) as

$$\mathbf{p}_i = -\mathbf{p}_{i-1}\Lambda_1\Big(\Lambda_0 + R_{i+1}\Lambda_{-1}\Big)^{-1} = \mathbf{p}_{i-1}R_i, \quad s < i < N. \tag{4.27}$$

The matrix $\Lambda_0 + R_{i+1}\Lambda_{-1}$ is lower triangular with negative elements on the diagonal and is therefore non-singular; for the proof of this statement, see [5, p. 519].

This leaves us to compute $\mathbf{p}_s$ and the equilibrium probabilities of the boundary states. We do so with the approach we have derived earlier for the slowdown system with an infinite buffer. The missing ingredients are the first passage probabilities from level $s+1$ to level s, which are found through the relation

$$G_i = \Lambda_1^{-1}R_i\Lambda_{-1}, \quad i > s. \tag{4.28}$$

Note that the auxiliary matrices G_i are level-dependent and have the same interpretation as the standard auxiliary matrix in the matrix-analytic approach. Thus, we substitute the level-dependent matrix G_{s+1} for G in (4.18) and are able to compute the complete stationary distribution.

4.3.2 Customer abandonments

The base model is appended by adding transitions with rate $l\delta$ from state $(s+l, j)$ to state $(s+l-1, j)$ for $l > 0$. These transitions model a waiting customer abandoning the queue. By appending the base model with these transitions a level-dependent QBD (LDQBD) process is created. We use solution techniques for LDQBD processes as presented in [5, 20] to compute the stationary distribution of the semi-infinite strip of states and once again use the earlier derived technique to compute the equilibrium distribution of the boundary states. We briefly sketch the solution approach here.

The aggregated abandonment rate depends on the number of customers waiting in the queue. This leads to level-dependent transition rate matrices which we label with an additional subscript l, such that $\Lambda_{k,l}$ describes the transition rates from level $s+l$ to level $s+l-k$ with $l > 0$. The transition rate matrices are given by $\Lambda_{1,l} = \Lambda_1$, $\Lambda_{0,l} = \Lambda_0 - l\delta I$ and $\Lambda_{-1,l} = \Lambda_{-1} + l\delta I$.

The solution approach is based on the same premise as for the finite QBD process case, namely

$$\mathbf{p}_i = \mathbf{p}_{i-1}R_i, \quad i > s, \tag{4.29}$$

where R_i is a level-dependent rate matrix with identical interpretation as the standard rate matrix of the matrix-geometric approach.

The following is explained in greater detail in [5]. Since generally only numerical solutions can be found for the R_i matrices of LDQBD processes, one resorts to computing the sequence $\{R_i\}_{s<i\leq N^*}$, where N^* is chosen "large enough". By [5, Lemma 1] we have the explicit expression

$$R_i = \sum_{j=0}^{\infty} U_i^j \prod_{k=0}^{j-1} D_{i+2^{j-k}}^{j-1-k}, \quad i > s, \tag{4.30}$$

where U_i^j and D_i^j are matrices defined recursively and are a function of the level-dependent transition matrices. Truncating the infinite sum in (4.30), one computes R_{N^*}. The remaining rate matrices then follow from the standard relation

$$R_i = -\Lambda_{1,i-s-1}\big(\Lambda_{0,i-s} + R_{i+1}\Lambda_{-1,i-s+1}\big)^{-1}, \quad i > s. \tag{4.31}$$

Note that the inverse exists.

As in the finite buffer case, this leaves us to compute $\mathbf{p}_s$ and the equilibrium probabilities of the boundary states. Once again, the first passage probabilities G_{s+1} are needed and follow from (4.28). Then, we substitute the level-dependent matrix G_{s+1} for G in (4.18) and are able to compute the complete stationary distribution.

5 A QED-type regime

We next analyze the behavior of the multi-server queueing system incorporating slowdown for large s and $\rho_{\mathrm{L}} \to 1$ by considering a sequence of queues, indexed by s. We write $(X(t), Y(t))) = (X^{(s)}(t), Y^{(s)}(t))$, $\lambda = \lambda^{(s)}$, $\mu_{\mathrm{H}} = \mu_{\mathrm{H}}^{(s)}$, and $\rho_{\mathrm{L}} = \rho_{\mathrm{L}}^{(s)}$. Without loss of generality we keep μ_{L} constant and assume throughout that $\mu_{\mathrm{H}}^{(s)} > \mu_{\mathrm{L}}$.

Let $\mathbb{P}(W^{(s)} > 0)$ denote the probability that a customer has to wait in a slowdown system with s servers. We will identify a regime in which $\mathbb{P}(W^{(s)} > 0) \to \mathbb{P}(W > 0) \in (0, 1)$ so that the limiting system displays non-degenerate behavior, as in the classical QED regime. In order to do so, we introduce a random variable $X_{\mathrm{H}}^{(s)}(t)$ that represents the total number of customers at time t in an $M/M/s$ queue where all customers are served with the high service rate $\mu_{\mathrm{H}}^{(s)}$. As we have done before, we refer to this queueing system as the *fast* system. Let the random variable $X_{\mathrm{L}}^{(s)}(t)$ represent the total number of customers at time t in an $M/M/s$ queue where all customers are served with the low service rate μ_{L}. We refer to this queueing system as the *slow* system.

For two real-valued random variables A and B, we say that A stochastically dominates B if

$$\mathbb{P}(A \leq x) \leq \mathbb{P}(B \leq x), \tag{5.1}$$

and we denote this as $A \succeq B$. The following result is proved in Appendix A.

Lemma 5.1. $X_{\mathrm{L}}^{(s)}(t) \succeq X^{(s)}(t) \succeq X_{\mathrm{H}}^{(s)}(t)$.

We next introduce the scaling

$$\lambda^{(s)} = s\mu_{\mathrm{L}} - \beta\mu_{\mathrm{L}}\sqrt{s}, \tag{5.2}$$

$$\mu_{\mathrm{H}}^{(s)} = \mu_{\mathrm{L}}(1 + \gamma/\sqrt{s}), \tag{5.3}$$

with constants $\mu_{\mathrm{L}}, \beta, \gamma > 0$ and $s \geq \beta^2$. Note that $\mu_{\mathrm{H}}^{(s)}/\mu_{\mathrm{L}} \to 1$ for $s \to \infty$. We refer to the scaling (5.2) and (5.3) as a QED-type scaling regime. We introduce the scaled random variables

$$D^{(s)}(t) \coloneqq \frac{X^{(s)}(t) - s}{\sqrt{s}},\ D_{\mathrm{H}}^{(s)}(t) \coloneqq \frac{X_{\mathrm{H}}^{(s)}(t) - s}{\sqrt{s}},\ D_{\mathrm{L}}^{(s)}(t) \coloneqq \frac{X_{\mathrm{L}}^{(s)}(t) - s}{\sqrt{s}}. \tag{5.4}$$

Note that Lemma 5.1 also holds for these scaled random variables, i.e. $D_{\mathrm{L}}^{(s)}(t) \succeq D^{(s)}(t) \succeq D_{\mathrm{H}}^{(s)}(t)$. The following lemma is proved in Appendix B.

Lemma 5.2. *If $D_{\mathrm{H}}^{(s)}(0) = d_{\mathrm{H}}^{(s)}$ and $D_{\mathrm{H}}(0) = d_{\mathrm{H}}$ a.s. with $d_{\mathrm{H}}^{(s)} \to d_{\mathrm{H}}$, and $D_{\mathrm{L}}^{(s)}(0) = d_{\mathrm{L}}^{(s)}$ and $D_{\mathrm{L}}(0) = d_{\mathrm{L}}$ a.s. with $d_{\mathrm{L}}^{(s)} \to d_{\mathrm{L}}$, then for $s \to \infty$, and for every $t \geq 0$, the scaled random variables $D_{\mathrm{H}}^{(s)}(t) \xrightarrow{d} D_{\mathrm{H}}(t)$ and $D_{\mathrm{L}}^{(s)}(t) \xrightarrow{d} D_{\mathrm{L}}(t)$, where the infinitesimal means of the diffusion processes are given by*

$$m_{\mathrm{H}}(x) = \begin{cases} \mu_{\mathrm{L}}(-\beta - \gamma - x), & x \leq 0, \\ \mu_{\mathrm{L}}(-\beta - \gamma), & x > 0, \end{cases} \quad m_{\mathrm{L}}(x) = \begin{cases} \mu_{\mathrm{L}}(-\beta - x), & x \leq 0, \\ -\beta\mu_{\mathrm{L}}, & x > 0, \end{cases} \tag{5.5}$$

and constant infinitesimal variances $\sigma_{\mathrm{H}}^2(x) = \sigma_{\mathrm{L}}^2(x) = 2\mu_{\mathrm{L}}$.

Remark 5.3. Both processes $D_{\mathrm{H}}(\cdot)$ and $D_{\mathrm{L}}(\cdot)$ behave as an Ornstein-Uhlenbeck process below level zero and a reflected Brownian motion above level zero.

Corollary 5.4. *The probability density functions of $D_{\mathrm{H}}(\infty)$ and $D_{\mathrm{L}}(\infty)$ are given by*

$$f_{D_{\mathrm{H}}}(x) = \begin{cases} C_{\mathrm{H}} \frac{\phi(x+\beta+\gamma)}{\Phi(\beta+\gamma)}, & x \leq 0, \\ (1-C_{\mathrm{H}})(\beta+\gamma)\mathrm{e}^{-(\beta+\gamma)x}, & x > 0, \end{cases}, \quad f_{D_{\mathrm{L}}}(x) = \begin{cases} C_{\mathrm{L}} \frac{\phi(x+\beta)}{\Phi(\beta)}, & x \leq 0, \\ (1-C_{\mathrm{L}})\beta\mathrm{e}^{-\beta x}, & x > 0, \end{cases} \tag{5.6}$$

with

$$C_{\mathrm{H}} = \frac{\beta+\gamma}{\beta+\gamma+\frac{\phi(\beta+\gamma)}{\Phi(\beta+\gamma)}}, \quad C_{\mathrm{L}} = \frac{\beta}{\beta+\frac{\phi(\beta)}{\Phi(\beta)}}, \tag{5.7}$$

and $\phi(x)$ and $\Phi(x)$ the probability density function and cumulative density function of the standard normal distribution.

Proof. Since we are dealing with piecewise-linear diffusion processes, one computes the probability density functions using [6, Sections 1 and 4]. □

Remark 5.5. The stationary distribution of the diffusion process related to the fast system is equal to the distribution of the limiting random variable of the sequence $(X_{\mathrm{H}}^{(s)}(\infty) - s)/\sqrt{s}$ as shown in [16, Corollary 2], which establishes that an interchange of limits is allowed. Thus, one can use $X_{\mathrm{H}}^{(s)}(\infty) \approx s + D_{\mathrm{H}}(\infty)\sqrt{s}$. The same applies for the slow system.

Corollary 5.6. *The limiting probability of delay in the slowdown system $\mathbb{P}(W^{(s)} > 0) \to \mathbb{P}(W > 0) \in (0,1)$ for $s \to \infty$ and can be bounded as follows*

$$\left(1+\beta\frac{\Phi(\beta)}{\phi(\beta)}\right)^{-1} = \mathbb{P}(W_{\mathrm{H}} > 0) \geq \mathbb{P}(W > 0) \geq \mathbb{P}(W_{\mathrm{L}} > 0) = \left(1+(\beta+\gamma)\frac{\Phi(\beta+\gamma)}{\phi(\beta+\gamma)}\right)^{-1}. \tag{5.8}$$

Proof. The limiting probability of delay in the fast system is computed from the distribution of $D_{\mathrm{H}}(\infty)$ as

$$\mathbb{P}(W_{\mathrm{H}} > 0) = \int_0^\infty f_{D_{\mathrm{H}}}(x)\,\mathrm{d}x = 1 - C_{\mathrm{H}} \in (0,1) \tag{5.9}$$

and identically for the slow system to get $\mathbb{P}(W_{\mathrm{L}} > 0) = 1 - C_{\mathrm{L}} \in (0,1)$. Using Lemma 5.1 we find that these are lower and upper bounds on the limiting probability to wait $\mathbb{P}(W > 0)$, respectively. □

6 Conclusion

We have studied a threshold slowdown system in a Markovian setting. The threshold slowdown system incorporates a slowdown effect in which customers that are delayed require a longer service time. A delayed customer requiring a longer service time will increase the overall workload in the system, therefore causing longer delays for other customers, who in turn due to slowdown also require a longer service time. We refer to this phenomenon as the snowball effect. The snowball effect has been shown to be the leading cause of a severe performance degradation and the neglect of slowdown might lead to underprovisioning. A subtle bistable behavior is witnessed for slowdown systems with relevant parameter settings: the slowdown system either has only non-delayed customers and no customers waiting, or only delayed customers with many customers waiting, switching between configurations over time. We have introduced a QED-type regime for the slowdown system with many-servers and established non-degenerate limiting behavior. The snowball effect has been shown to persist in this QED-type regime.

A Proof of Lemma 5.1

The proof is based on a coupling argument and follows the same reasoning as [8, Appendix B]. We distinguish between two cases: (i) $X_{\mathrm{L}}^{(s)}(t) \succeq X^{(s)}(t)$; and (ii) $X^{(s)}(t) \succeq X_{\mathrm{H}}^{(s)}(t)$. Recall that for two real-valued random variables A and B, we say that A first-order stochastically dominates B if

$$\mathbb{P}(A \leq x) \leq \mathbb{P}(B \leq x). \tag{A.1}$$

(i) Assume that both queues see a common arrival process. Let the service time for the i-th arriving customer in the slow system be $B_{\mathrm{L}}(i)$; the corresponding service time in the slowdown model is then either $B(i) = B_{\mathrm{L}}(i)$ or $B(i) = \mu_{\mathrm{L}}/\mu_{\mathrm{H}}^{(s)} B_{\mathrm{L}}(i)$ depending on whether the slowdown model has high (s or more customers in the system) or low congestion (less than s customers in the system) upon arrival of the i-th customer. Finally, we assume that both systems start empty. Let $W(i)$ and $W_{\mathrm{L}}(i)$ denote the waiting time of the i-th arriving customer before beginning service in the slowdown and slow system, respectively. We have the following result.

Lemma A.1. *$W_{\mathrm{L}}(i) \geq W(i)$ a.s. for all i. Moreover, $X_{\mathrm{L}}^{(s)}(t) \geq X^{(s)}(t)$ a.s. for all t.*

Proof. Let us prove the first statement using induction and fix an arbitrary event in the sample space $\omega \in \Omega$ leading to a sample path of the process. We append the argument ω to the variables to indicate that we are studying a fixed sample path. Since we start with an empty system, observe that for the first customer we have $W_{\mathrm{L}}(1,\omega) = W(1,\omega) = 0$. Assume that the statement is true for the j-th arriving customer and consider the $(j+1)$-th arriving customer. For the sake of contradiction assume $W_{\mathrm{L}}(j+1,\omega) < W(j+1,\omega)$. When customer $j+1$ starts service in the slow system:

- There are at most $s-1$ customers among the first j arriving customers present in the slow system.
- At least s customers from among the first j arriving customers are still present in the slowdown system since customer $j+1$ has not yet started service in the slowdown system.

From these two observations we conclude that there is a customer among the first j arriving customers that finished service strictly earlier in the slow system than in the slowdown system. However, due to the coupling we have $B(i,\omega) \leq B_{\mathrm{L}}(i,\omega),\ i = 1,\ldots,j$ and thus we have a contradiction. We have consequently established that $W_{\mathrm{L}}(i,\omega) \geq W(i,\omega)$ for all i. Recall that we fixed an arbitrary event and thus it holds for all $\omega \in \Omega$. The latter statement of the proposition follows immediately. □

Lemma A.1 indeed shows that $X_{\mathrm{L}}^{(s)}(t) \succeq X^{(s)}(t)$.

(ii) This case follows using the exact same reasoning as for case (i) and we thus omit the proof.

B Proof of Lemma 5.2

The following proof is based on the proof in [19, Proposition 3.2]. We first describe convergence in distribution for a general sequence of birth–death processes and apply these results to our processes of interest.

Define $A^{(s)}(\cdot)$ as a continuous-time birth–death process with state space $E^{(s)} = \{a^{(s)}(i) \mid 0 \leq i < \infty\}$, where $a^{(s)}(i)$ is increasing in i and $\lim_{i\to\infty} a^{(s)}(i) = a^{(s)}(\infty)$. Then let

$$e^{(s)}(x) = \underset{i\in\mathbb{N}_0}{\arg\sup}\{a^{(s)}(i) : a^{(s)}(i) \leq x\}, \quad x \in [a^{(s)}(0), a^{(s)}(\infty)). \tag{B.1}$$

Denote by $\lambda^{(s)}(a^{(s)}(i))$ and $\mu^{(s)}(a^{(s)}(i))$ the birth–death parameters associated with state $a^{(s)}(i)$. The infinitesimal mean and variance of the process $A^{(s)}(\cdot)$ are given by

$$\begin{aligned} m^{(s)}(x) &= \lambda^{(s)}(e^{(s)}(x))\Big(a^{(s)}(e^{(s)}(x)+1) - a^{(s)}(e^{(s)}(x))\Big) \\ &\quad - \mu^{(s)}(e^{(s)}(x))\Big(a^{(s)}(e^{(s)}(x)) - a^{(s)}(e^{(s)}(x)-1)\Big), \end{aligned} \tag{B.2}$$

$$\begin{aligned} \big(\sigma^2\big)^{(s)}(x) &= \lambda^{(s)}(e^{(s)}(x))\Big(a^{(s)}(e^{(s)}(x)+1) - a^{(s)}(e^{(s)}(x))\Big)^2 \\ &\quad + \mu^{(s)}(e^{(s)}(x))\Big(a^{(s)}(e^{(s)}(x)) - a^{(s)}(e^{(s)}(x)-1)\Big)^2, \end{aligned} \tag{B.3}$$

whenever $x \in [a^{(s)}(0), a^{(s)}(\infty))$.

Stone's theorem [18, Theorem 5.1] then establishes convergence in distribution of the sequence of Markov processes to a limiting diffusion process.

Theorem B.1. (Stone) *Let $A^{(s)}(0) = a^{(s)}$ and $A(0) = a$ a.s., with $a^{(s)} \to a$. Then the following two conditions are sufficient for $A^{(s)} \xrightarrow{d} A$ as elements of $D[0,\infty)$:*

(i) *$E^{(s)}$ becomes dense in $\mathbb{R}$ as $s \to \infty$;*

(ii) *For every compact subinterval U of $\mathbb{R}$*

$$\lim_{s\to\infty} m^{(s)}(x) = m(x), \quad \lim_{s\to\infty} \big(\sigma^2\big)^{(s)}(x) = \sigma^2(x), \tag{B.4}$$

uniformly for $x \in U$.

We first focus on the fast system with scaled process $D_{\mathrm{H}}^{(s)}(\cdot)$ and state space $E_{\mathrm{H}}^{(s)} = \{(i-s)/\sqrt{s} \mid i \in \mathbb{N}_0\}$. Naturally, $\lim_{s\to\infty} E_{\mathrm{H}}^{(s)}$ is dense in $\mathbb{R}$, by which we mean that $\forall x \in \mathbb{R}, \forall \epsilon > 0, \exists s > 0$ such that $\inf_{y\in E_{\mathrm{H}}^{(s)}} |x-y| < \epsilon$, thus satisfying condition (i) of Theorem B.1. Note that for this state space $e^{(s)}(x) = \lfloor s + x\sqrt{s} \rfloor,\ x \in [-\sqrt{s}, \infty)$.

Now, use that for the fast (and slow) system the expression $a^{(s)}(e^{(s)}(x)+1) - a^{(s)}(e^{(s)}(x))$ in (B.2) and (B.3) is equal to $1/\sqrt{s}$ to obtain the infinitesimal mean and variance of the process $D_{\mathrm{H}}^{(s)}(\cdot)$ as

$$m_{\mathrm{H}}^{(s)}(x) = \begin{cases} \frac{1}{\sqrt{s}}\big(\lambda^{(s)} - \lfloor s + x\sqrt{s} \rfloor \mu_{\mathrm{H}}^{(s)}\big), & x \le 0, \\ \frac{1}{\sqrt{s}}\big(\lambda^{(s)} - s\mu_{\mathrm{H}}^{(s)}\big), & x > 0, \end{cases} \tag{B.5}$$

$$\big(\sigma_{\mathrm{H}}^2\big)^{(s)}(x) = \begin{cases} \frac{1}{s}\big(\lambda^{(s)} + \lfloor s + x\sqrt{s} \rfloor \mu_{\mathrm{H}}^{(s)}\big), & x \le 0, \\ \frac{1}{s}\big(\lambda^{(s)} + s\mu_{\mathrm{H}}^{(s)}\big), & x > 0. \end{cases} \tag{B.6}$$

We use the scaling (5.2) and (5.3) to obtain

$$m_{\mathrm{H}}^{(s)}(x) = \begin{cases} \mu_{\mathrm{L}}\big(-\beta - \frac{\lfloor s+x\sqrt{s} \rfloor}{s}\gamma + \frac{s - \lfloor s+x\sqrt{s} \rfloor}{\sqrt{s}}\big), & x \le 0, \\ \mu_{\mathrm{L}}\big(-\beta - \gamma\big), & x > 0, \end{cases} \tag{B.7}$$

$$\big(\sigma_{\mathrm{H}}^2\big)^{(s)}(x) = \begin{cases} \mu_{\mathrm{L}}\big(1 + \frac{\lfloor s+x\sqrt{s} \rfloor}{s} - \frac{\beta}{\sqrt{s}} + \frac{\lfloor s+x\sqrt{s} \rfloor}{s}\frac{\gamma}{\sqrt{s}}\big), & x \le 0, \\ \mu_{\mathrm{L}}\big(2 - \frac{\beta-\gamma}{\sqrt{s}}\big), & x > 0. \end{cases} \tag{B.8}$$

By requiring that $\lim_{s\to\infty} m_{\mathrm{H}}^{(s)}(x)$ and $\lim_{s\to\infty}\big(\sigma_{\mathrm{H}}^2\big)^{(s)}(x)$ are finite, we indeed find that β and γ can be any value larger than 0. For every compact subinterval U of $\mathbb{R}$, $\lim_{s\to\infty} m_{\mathrm{H}}^{(s)}(x) = m_{\mathrm{H}}(x)$

and $\lim_{s\to\infty}\big(\sigma^2_{\mathrm{H}}\big)^{(s)}(x) = \sigma^2_{\mathrm{H}}(x)$ uniformly for $x \in U$. Thus, condition (ii) of Theorem B.1 is satisfied and $D^{(s)}_{\mathrm{H}}(t) \xrightarrow{d} D_{\mathrm{H}}(t)$.

Next we turn to the slow system with the associated scaled process $D^{(s)}_{\mathrm{L}}(\cdot)$. Its state space is equal to $E^{(s)}_{\mathrm{H}}$ and thus in the limit for $s \to \infty$ also dense in $\mathbb{R}$. Using the scaling as proposed in (5.2) and (5.3), the infinitesimal mean and variance of the process $D^{(s)}_{\mathrm{L}}(\cdot)$ are

$$m^{(s)}_{\mathrm{L}}(x) = \begin{cases} \mu_{\mathrm{L}}\big(-\beta + \frac{s-\lfloor s+x\sqrt{s}\rfloor}{\sqrt{s}}\big), & x \le 0, \\ -\beta\mu_{\mathrm{L}}, & x > 0, \end{cases} \tag{B.9}$$

$$\big(\sigma^2_{\mathrm{L}}\big)^{(s)}(x) = \begin{cases} \mu_{\mathrm{L}}\big(1 + \frac{\lfloor s+x\sqrt{s}\rfloor}{s} - \frac{\beta}{\sqrt{s}}\big), & x \le 0, \\ \mu_{\mathrm{L}}\big(2 - \frac{\beta}{\sqrt{s}}\big), & x > 0. \end{cases} \tag{B.10}$$

For every compact subinterval U of $\mathbb{R}$, $\lim_{s\to\infty} m^{(s)}_{\mathrm{L}}(x) = m_{\mathrm{L}}(x)$ and $\lim_{s\to\infty}\big(\sigma^2_{\mathrm{L}}\big)^{(s)}(x) = \sigma^2_{\mathrm{L}}(x)$ uniformly for $x \in U$. Again, conditions (i) and (ii) of Theorem B.1 are satisfied and $D^{(s)}_{\mathrm{L}}(t) \xrightarrow{d} D_{\mathrm{L}}(t)$.

Acknowledgement

This work was supported by a free competition grant from NWO and an ERC starting grant.

References

[1] 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*, 5:1–49, 2015.

[2] B. Ata and S. Shneorson. Dynamic control of an $M/M/1$ service system with adjustable arrival and service rates. *Management Science*, 52(11):1778–1791, 2006.

[3] R.J. Batt and C. Terwiesch. Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, 2012.

[4] R. Bekker and S.C. Borst. Optimal admission control in queues with workload-dependent service rates. *Probability in the Engineering and Informational Sciences*, 20(04):543–570, 2006.

[5] L. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth–and–death processes. *Stochastic Models*, 11(3):497–525, 1995.

[6] S. Browne and W. Whitt. Piecewise-linear diffusion processes. In J.H. Dshalalow, editor, *Advances in Queueing: Theory, Methods, and Open Problems*, volume 4, chapter 18, pages 463–480. CRC Press, Boca Raton, FL, 1995.

[7] D.B. Chalfin, S. Trzeciak, A. Likourezos, B.M. Baumann, and R.P. Dellinger. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine*, 35(6):1477–1483, 2007.

[8] C.W. Chan, V.F. Farias, and G. Escobar. The impact of delays on service times in the intensive care unit. Working paper, 2013.

[9] P.S. Chan, H.M. Krumholz, G. Nichol, and B.K. Nallamothu. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine*, 358(1):9–17, 2008.

[10] J. Dong, P. Feldman, and G. Yom-Tov. Service system with slowdowns: Potential failures and proposed solutions. *Operations Research*, 62(2):305–324, 2015.

[11] E.H. Elhafsi and M. Molle. On the solution to QBD processes with finite state space. *Stochastic Analysis and Applications*, 25(4):763–779, 2007.

[12] J.B. Fraleigh and R.A. Beauregard. *Linear Algebra*. Addison-Wesley, 1995.

[13] J.M. George and J.M. Harrison. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731, 2001.

[14] W.K. Grassmann, M.I. Taksar, and D.P. Heyman. Regenerative analysis and steady state distributions for Markov chains. *Operations Research*, 33(5):1107–1116, 1985.

[15] L. Green. Queueing analysis in healthcare. In R. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 206, chapter 10, pages 281–307. Springer, New York, NY, 2006.

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

[17] C.M. Harris. Queues with state-dependent stochastic service rates. *Operations Research*, 15(1):117–130, 1967.

[18] D.L. Iglehart. Weak convergence in applied probability. *Stochastic Processes and Their Applications*, 2(3):211–241, 1974.

[19] A.J.E.M. Janssen, J.S.H. van Leeuwaarden, and J. Sanders. Scaled control in the QED regime. *Performance Evaluation*, 70(10):750–769, 2013.

[20] J.P. Kharoufeh. Level-dependent quasi-birth–and–death processes. In J.J. Cochran, L.A. Cox, P. Keskinocak, J.P. Kharoufeh, and J.C. Smith, editors, *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, New York, NY, 2011.

[21] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover Publications, Mineola, NY, 1994.

[22] B. Renaud, A. Santin, E. Coma, N. Camus, D. van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, C. Brun-Buisson, and J. Labarère. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine*, 37(11):2867–2874, 2009.

[23] D.B. Richardson. The access-block effect: Relationship between delay to reaching an inpatient bed and inpatient length of stay. *The Medical Journal of Australia*, 177(9):492–495, 2002.

[24] A.W. Siegmeth, K. Gurusamy, and M.J. Parker. Delay to surgery prolongs hospital stay in patients with fractures of the proximal femur. *Journal of Bone & Joint Surgery, British Volume*, 87(8):1123–1126, 2005.

[25] B. Van Houdt and J.S.H. van Leeuwaarden. Triangular $M/G/1$-type and tree-like quasi-birth-death Markov chains. *INFORMS Journal on Computing*, 23(1):165–171, 2011.

[26] J.S.H. van Leeuwaarden, M.S. Squillante, and E.M.M. Winands. Quasi-birth–and–death processes, lattice path counting, and hypergeometric functions. *Journal of Applied Probability*, 46(2):507–520, 2009.

[27] J.S.H. van Leeuwaarden and E.M.M. Winands. Quasi-birth–and–death processes with an explicit rate matrix. *Stochastic models*, 22(1):77–98, 2006.

[28] R.R. Weber and S. Stidham Jr. Optimal control of service rates in networks of queues. *Advances in Applied Probability*, 19(1):202–218, 1987.

[29] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, NY, 2002.