Service Systems with Slowdowns: Potential Failures and Proposed Solutions

Jing Dong
Northwestern University, jing.dong@northwestern.edu

Pnina Feldman
University of California Berkeley, feldman@haas.berkeley.edu

Galit Yom-Tov
Technion – Israel Institute of Technology, gality@tx.technion.ac.il

Many service systems exhibit service slowdowns when the system is congested. Our goal in this paper is to investigate this phenomenon and its effect on system performance. We modify the Erlang-A model to account for service slowdowns and carry out the performance analysis in the Quality-and-Efficiency Driven (QED) regime. We find that when the load sensitivity is low, the system can achieve QED performance, but the square-root staffing parameter requires an adjustment to achieve the same performance as an ordinary Erlang-A queue. When the load sensitivity is high, the system alternates randomly between a QED and an Efficiency Driven (ED) regime performance levels, a phenomenon which we refer to as *bi-stability*. We analyze how the system scale and the model parameters affect the bi-stability phenomenon and propose an admission control policy to avoid ED performance.

Key words: Service systems; Halfin-Whitt regime; QED; Erlang-A model; State-dependent queues; Load-dependent queues; Bi-stability

History: Submitted August 2013, revised June 2014.

1. Introduction

A central assumption in the operations management literature is that service times are independent of the load of the system. However, empirical and anecdotal evidence suggest that in many service systems the two are correlated (see for example Batt and Terwiesch (2012), Gerla and Kleinrock (1980), KC and Terwiesch (2009) and Feldman et al. (2014)). Depending on the service environment, heavily-loaded systems may experience service speedups or slowdowns. While speedup was theoretically investigated in Chan et al. (2014), slowdown was so far been neglected.

Slowdown of service rate, when the system is congested, is a widely spread phenomenon, which is contributed to several psychological, physiological and technical reasons. High congestion levels may induce pressure on agents, which according to the psychology literature (see for example Bertrand and van Ooijen (2002)) may impact human perception, information processing and decision making. All of these aspects may influence operational performance. While a relatively low level of arousal may increase productivity, high levels of pressure hurt performance (Wickens et al. 2012). High congestion levels may also require individuals to conduct multiple tasks in parallel which involves

a cognitive switching cost (Batt and Terwiesch 2012). At the same time, high congestion levels may lead staff to work longer hours without proper rest, causing fatigue. Empirical studies provide evidence that fatigue leads to deterioration in productivity (e.g. KC and Terwiesch (2009), Caldwell (2001)). Service rate may also deteriorate due to external capacity limitations, for example, IT systems perform slower when heavily loaded, and hence, the service times of the workers who use them may increase (Batt and Terwiesch 2012). On the customer side, it is well established that patients' condition may deteriorate if treatment is delayed in health care facilities, causing a service slowdown (Chalfin et al. 2007, Chan et al. 2013). For example, it is shown in Chan et al. (2013) that one additional hour of delay, when transferring from the ER to the ICU, leads to an increase in the length of stay in the ICU by 6.5 to 23 hours. Lastly, another reason for slowdown services is that customers may demand a longer and more personalized service following a long wait. For example, agents might need to take some extra time to mollify irritated customers who experience long waits.

Motivated by these empirical findings, from both call-centers and healthcare facilities, we investigate how the dependence between service rate and workload affects the operational performance of the system, measured by delay and abandonment, and how service providers can cope with the consequences of this dependence by adjusting staffing or admission.

Generally, there are two objectives that play opposing roles in the design of service systems. On the one hand, to increase efficiency and reduce operational costs, system designers aim to increase resource utilization. On the other hand, high utilization rates lead to increased levels of delay and abandonment, thereby reducing quality of service. A common approach is to design a service system that balances the tradeoff between system performance, measured by the probability of waiting and the probability of abandonment experienced by customers, and resource utilization, measured by the fraction of time an agent or a resource is occupied. The Quality-and-Efficiency-Driven (QED) regime in the many-server asymptotic analysis suggests a Square-Root Staffing (SRS) rule to balance this tradeoff. According to the SRS rule the number of servers, n, is set such that $n = R + \beta\sqrt{R}$, where $R = \lambda/\mu$ is the offered load of the system, and β, the SRS parameter, is set to achieve certain performance measures. For the SRS rule in an exponential type multi-server queue with abandonment (commonly referred to as the *Erlang-A* model), β is determined using the Garnett functions (Garnett et al. 2002). Applying the SRS rule to the Erlang-A model implies that a significant proportion of customers (e.g. 30%–80%) gets served immediately upon arrival and the probability of abandonment is small (e.g. $< 5\%$) (Garnett et al. 2002). Other operating regimes considered in the literature include the Efficiency-Driven (ED) regime and the Quality-Driven (QD) regime, where the staffing level and the offered load grow in fixed proportion. ED staffing is used when the staffing cost is very high. In this case, the staffing level is set to $n = R - \alpha R$

for $0 < \alpha < 1$, where α is typically selected in the range 0.1–0.25 (Whitt 2004). This results in 100% occupancy, probability of waiting close to 1 and a very high abandonment rate (5%–30%) (Garnett et al. 2002). A QD regime is used when the system requires a very high level of service quality. In this case, the staffing level is set to $n = R + \alpha R$ for $\alpha > 0$, where the typical range of α is as in the ED regime. This staffing level results in very low abandonment (almost 0) and negligible waiting, but also in an agent occupancy which is far below 100% (Garnett et al. 2002).

In this paper, we modify the Erlang-A model to account for the slowdown effect and analyze the performance of the modified model when staffing according to the SRS rule. We use the term *load sensitivity* to describe the rate of service rate deterioration as a response to increased workload. We show that the SRS rule may not be a good enough solution in some systems with load-sensitive service rates. Depending on the model parameters, we observe that systems designed to operate in the QED regime may have undesirable performances, alternating between being heavily overloaded and moderately loaded, or even end up being constantly heavily overloaded. This results in a very high probability of waiting (close to 1) and a significant proportion of customer abandonment (e.g. 10%–20%). Hence, a QED regime staffing rule, or even a QD regime staffing rule, may result in unwanted performances, typically found when using ED regime staffing rules. We therefore propose to consider alternative staffing rules and admission control policies that can be applied in the presence of service slowdowns.

We make the following key contributions:

1) We show that the effect of load sensitivity on system performance is nonlinear. Systems with low sensitivity may exhibit only a modest deterioration in performance, whereas when the sensitivity increases beyond a threshold, the performance deteriorates drastically. We show that the threshold that separates the two cases is derived from the *relative* relation between the service rate sensitivity level around zero wait time and the abandonment rate (§4.2).

2) When the *load sensitivity* is relatively low (i.e., the service rate does not decrease significantly with the load placed on the system), the SRS rule leads to a QED performance. However, for a fixed square-root staffing parameter, β, the performance deteriorates with the load sensitivity level. We develop new approximation functions in the presence of load sensitivity, which can be used when making staffing decisions (§5). To derive these approximations, it is sufficient to accurately estimate the service rate function around zero.

3) When the *load sensitivity* is relatively high, the system alternates between two performance regions, a phenomenon we refer to as bi-stability: one provides a QED performance while the other has an ED performance (§4). Therefore, in such cases, applying the SRS rule does not consistently result in QED performance. We investigate how the system scale and other parameters influence the occurrence of bi-stability, and the proportion of time the system spends around each performance

regions (§6). To achieve traditional QED regime performance in this case, we propose implementing an admission control policy (§6.3). We show that while a higher load sensitivity increases the occurrence of ED performance, a higher abandonment rate decreases such occurrences. We also show that large systems converge to the ED performance with an exponential rate. Sensitivity increases the rate of convergence, and abandonment rate decreases this rate of convergence. Two interesting observations follow from our analysis. Firstly, under SRS, the performance of the modified Erlang-A queue, both the probability of waiting and the probability of abandonment, deteriorates with system scale. This is in contrast to the traditional Erlang-A model in the QED regime. Secondly, firms should encourage customers to abandon when having load-sensitive service rate. This can be done in real applications by, for example, providing delay announcements.

4) The model we analyze captures agent driven service slowdowns. In addition, we show, using numerical examples (§7), that this model is robust and the main insights carry over to a larger class of models. This includes settings in which the service rate deterioration is customer-driven (i.e., longer waiting results in longer service requirement for that specific customer), in which it is agent-driven (i.e., agents change their service rate according to queue length), or in situations where there is a delay in the slowdown effect on service rate (e.g., slowdown is caused by agent fatigue).

2. Literature Review

In this paper, we study a modified Erlang-A model that accounts for the load-dependent service rate. The Erlang-A ($M/M/n+M$) queue was first introduced by Palm (1957) to incorporate abandonments in the traditional Erlang-C ($M/M/n$) queue. Mandelbaum and Zeltyn (2007) showed that abandonment is a significant factor in modeling service systems and making staffing decisions. Garnett et al. (2002) conducted heavy traffic asymptotic analysis of the Erlang-A model in the QED regime. They derived approximations for the probability of waiting and abandonment and provided guidance for the design of large service systems. Our analysis differs from Garnett et al. (2002) because we do not assume the service rate to be constant but load-dependent.

A few papers consider state-dependent service rates but most of them are in the single server queue setting without abandonment. Whitt (1990) and Boxma and Vlasiou (2007) study the steady-state behavior of the delay process (waiting time distribution) of a $G/G/1$ queue, where both the service rate and the arrival rate depend linearly on the delay process. Mandelbaum and Pats (1998) derived the fluid and diffusion limits of a network of single server queues with state-dependent arrival rate, service rate and routing probability. Weerasinghe (2013) studies the fluid and diffusion approximations of $G/M/n+GI$ queues with state-dependent service rate, but they only analyze convergence over finite time intervals (i.e. intervals of the form $[0,T]$). Our work is also different

from Zohar et al. (2002) and Armony et al. (2009) who analyzed how delay announcements affect system performance by changing the strategic behavior of customers. This was done by combining game theory analysis with queueing models.

The bi-stability phenomenon is studied in different contexts: ICU flows (Chan et al. 2014), communication networks (Gibbens et al. 1990), multi-class stochastic networks (Antunes et al. 2009) and many-server systems with retrials (Janssen and van Leeuwaarden 2014). The phenomenon is also studied in statistical physics (e.g. Hollander (2004), Olivieri and Vares (2005)). The conjectured trajectory of the system under bi-stability is that it fluctuates within one stable region for a long time and then, due to some rare event it reaches the other stable region and remains there for a while (Antunes et al. 2009). In this paper, we study the bi-stability phenomenon through asymptotic analysis of the stationary distribution and sensitivity analysis of system parameters (§6). We impose exponential assumptions on the service time and patience time distributions for tractability reasons.

In terms of staffing and admission control policies, Bekker and Borst (2006) studied the optimal admission control of an $M/G/1$ queue with service rate that is first increasing and then decreasing as a function of the workload. Their objective is to optimize throughput and they show that under certain conditions a threshold policy is optimal. Likewise, in §6.3, we also consider a threshold admission control policy, but our objective is to maintain a certain performance level. Admission control in the QED regime has also been studied in Janssen et al. (2013), Daley et al. (2013) and references therein. We also consider staffing policies in face of load-dependent slowdown effect. A few papers considered dynamic staffing (e.g., Green et al. (2007), Yom-Tov and Mandelbaum (2014)) to cope with time-varying arrivals. They allow the staffing level to change over time according to a predictable offered load function. In our model, the fluctuations in performance arise because of the bi-stability phenomenon. The system alternates between two equilibria in an *unpredictable* stochastic way. Therefore, we cannot propose a predetermined policy whereby the staffing levels change in a predictable fashion. Instead, we propose static policies that mitigate the effect of the unpredictable system behavior.

3. Model Setup

3.1. The Load-dependent Erlang-A model

We analyze a modified Erlang-A ($M/M/n+M$) model which incorporates the dependence of service rate on workload through the queue length process. Specifically, we consider an $M/M_Q/n+M$ queue with the following assumptions: Customers arrive to the system according to a Poisson process with rate λ. The system has n identical servers; each server can serve only one customer at a time. If a customer arrives and finds a server free, she starts service with that server immediately.

Otherwise, she waits in the queue. Customers are served on a First-Come-First-Served basis. The service requirement is exponentially distributed with a state-dependent rate function $\mu(\cdot) \in C^2$ (we denote these facts in Kendall's notation by the second M_Q). We also assume that customers have finite patience. The patience time of each customer is exponentially distributed with a constant rate θ, which we refer to as the abandonment rate. If a customer does not get into service before her patience time expires, she abandons the queue.

We denote the queue length process by $Q \equiv \{Q(t) : t \geq 0\}$, where $Q(t)$ counts the number of customers in the system (waiting and in service) at time t. Motivated by the empirical findings on slowdowns, we assume that the service rate of each server is a function of the scaled queue length process, $\mu((Q-n)^+/n)$, where $(x)^+ = \max\{0, x\}$. This scaling makes the workload process $((Q(t)-n)^+/n)$ of the same order as the delay process (waiting time of an imaginary arrival at time t) (Whitt 2004). It is essential when considering scaling for approximations. From a practical point of view, one can also interpret this scaling as agent sensitivity to *individual* future load—queue length divided by number of servers.

We are interested in service systems in which the service rate deteriorates as the congestion level grows. We measure the level of load sensitivity by $\mu'(x)$ and let $\mu^{(i)}(0) := \lim_{x \to 0+} \mu^{(i)}(x)$ for $i = 1, 2$. We further assume that the service rate function maintains a minimum positive level, thus exhibits a diminishing decreasing rate. Formally:

ASSUMPTION 1. *$\mu'(x) \leq 0$ and $\mu''(x) \geq 0$ for all $x \geq 0$. $lim_{x\to\infty}\mu(x) = \mu(\infty) > 0$.*

In our numerical demonstrations, we use a specific form of the service rate function: $\mu(x) = c + a\exp(-bx)$ with parameters $a, b, c > 0$, which clearly satisfies Assumption 1. To demonstrate changes in load sensitivity, we change the values of b while keeping all other parameters fixed. We refer to b as the *load sensitivity* parameter.

Under our assumptions on the service rate function, $Q(t)$ is a Birth-and-Death (B&D) process with birth rate λ and state-dependent death rate $\mu((Q-n)^+/n)(Q \wedge n) + \theta(Q-n)^+$, where $x \wedge y = \min\{x, y\}$. As $\theta > 0$, $Q(t)$ admits a unique steady-state distribution. We denote the steady-state distribution, by $\pi(q)$, where

$$\pi(q) := P(Q(\infty) = q);$$

$\pi(q)$ measures the long run average proportion of time the system spends at q.

3.2. The QED heavy-traffic regime

For a sequence of $M/M/n+M$ queues indexed by n, the QED regime is obtained by holding the service rate and abandonment rate fixed while letting the aggregate arrival rate, λ_n, and the

number of servers, n, grow to infinity such that the utilization rate $\rho_n := \lambda_n/(n\mu)$ approaches 1 at rate $1/\sqrt{n}$. Specifically, we assumed that

$$\sqrt{n}(1-\rho_n) \to \beta \text{ as } n \to \infty \tag{1}$$

for some $\beta \in \mathbb{R}$, or, equivalently, that the number of servers is set by the SRS formula—$n = R_n + \beta\sqrt{R_n}$, where $R_n = \lambda_n/\mu$.

Garnett et al. (2002) proved that when a sequence of Erlang-A systems satisfies Equation (1) (i.e., operates in the QED regime), the probability of waiting, $P(W)$, is non-degenerate and the probability of abandonment, $P(Ab)$, converges to zero at rate $1/\sqrt{n}$. Thus, systems that operate in this regime achieve both good performance and high efficiency. However, as Figure 1 illustrates, in the modified Erlang-A model, square-root staffing does not always guarantee similar performance. In the absence of workload sensitivity, (i.e., $b = 0$), a system with the following parameters: $n = 523$, $\lambda = 500$, $\mu(q) = 0.6 + 0.4\exp(-b(q-n)^+/n)$ and $\theta = 0.3$, operates in the QED regime with $P(W) = 0.1882$ and $P(Ab) = 0.0018$. The upper diagrams in Figure 1 show that when the load sensitivity is low ($b = 0.5$), the queue length process fluctuates around 500 and the system operates with low probabilities of waiting and abandonment ($P(W) = 0.2050$ and $P(Ab) = 0.0023$). However, the performance is worse than the one obtained without sensitivity. The lower diagrams in Figure 1, show that when the load sensitivity is high ($b = 1.5$), the queue length process can alternately fluctuate around two regions (one in which the number of customers in the system, Q, is around 500 and the other in which Q is around 920). The sample path in Figure 1c illustrates how the queue length process fluctuates in one region for a while before it "unexpectedly" moves to the other region and vise versa. The system performance is quite different in the two regions. The lower region results in the QED regime performance ($P(W) \approx 0.22$ and $P(Ab) \approx 0.02$) while the upper region leads to the ED regime performance ($P(W) \approx 1$ and $P(Ab) \approx 0.21$). The average performance is $P(W) = 0.9090$ and $P(Ab) = 0.2008$. We refer to this phenomenon as bi-stability (see Definition 3 for a more precise definition of bi-stability).

Figure 2 shows how the probabilities of waiting and abandonment change with the load sensitivity parameter, b. We observe that the effect of load sensitivity is nonlinear. The performance deteriorates drastically as the sensitivity parameter grows beyond a certain level (e.g., at around $b = 1.5$ for the parameters in Figure 2).

This implies that the SRS rule may not be an adequate policy to achieve a QED performance in service systems that have load-sensitive service rates. In the next section, we use the many-server heavy traffic analysis to characterize the dynamics of such systems.

Figure 1 **Sample path and approximated stationary distribution of the number of people in the system for $M/M_Q/n+M$ queues with different load sensitivity parameter values, b ($n=523$, $\lambda=500$, $\mu=0.6+0.4\exp(-b(q-n)^+/n)$ and $\theta=0.3$)**

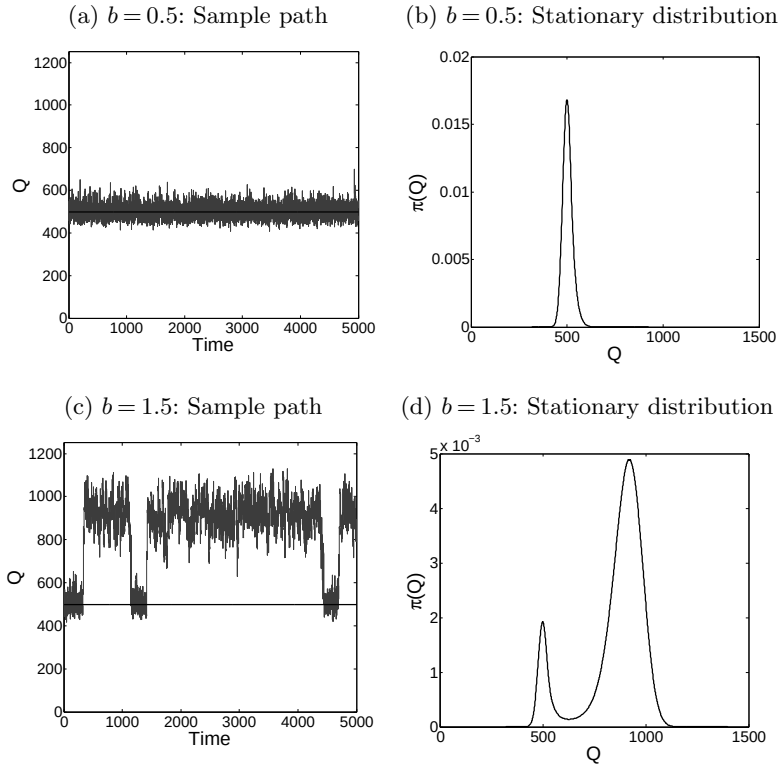
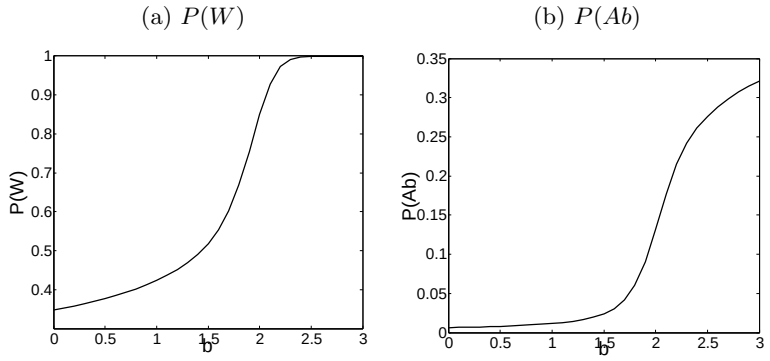


Figure 2 **Performance measures for $M/M_Q/n+M$ queues as a function of the load sensitivity parameter, b ($n=512$, $\lambda=500$, $\mu=0.6+0.4\exp(-b(q-n)^+/n)$ and $\theta=0.5$)**



4. Fluid Analysis

In this section, we establish the fluid limit of the queue length process of the modified Erlang-A model. This deterministic model serves as an approximation for the corresponding stochastic system when the system scale is large. We then conduct an equilibrium analysis of the fluid model to characterize the stationary performance of the modified Erlang-A model.

4.1. Fluid approximation

To develop the fluid limit, we consider a sequence of $M/M_Q/n+M$ queues indexed by n, where the arrival rate $\lambda_n \to \infty$ as $n \to \infty$. For the n-th system, we denote the queue length process (number of people in the system) by $Q_n \equiv \{Q_n(t) : t \geq 0\}$. The abandonment rate does not scale with n and the service rate function takes the same form when applied to the scaled queue length process $((Q_n(t)-n)^+/n)$. As we are interested in the QED asymptotic regime, we assume that there exists a β such that $\lim_{n\to\infty} \sqrt{n}(1-\lambda_n/(n\mu(0))) = \beta$.

Let $A \equiv \{A(t) : t \geq 0\}$, $S \equiv \{S(t) : t \geq 0\}$ and $R \equiv \{R(t) : t \geq 0\}$ be three independent Poisson processes, each with unit rate. A, S and R generate the arrival, service completion and abandonment processes, respectively. Then, the pathwise construction of Q_n is:

$$Q_n(t) = Q_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)\,du\right) - R\left(\int_0^t \theta(Q_n(u)-n)^+ du\right).$$

We define the fluid-scaled process as

$$\bar{Q}_n(t) = \frac{Q_n(t)}{n}.$$

Let $\mathcal{D} := D([0,\infty),\mathbb{R})$ denote the functional space of all right-continuous real-valued functions on the interval $[0,\infty)$ with left limit everywhere in $(0,\infty)$, endowed with Skorohod (J_1) topology.

THEOREM 1. *Assume* $\lim_{n\to\infty} \sqrt{n}(1-\lambda_n/(n\mu(0))) = \beta$. *If* $\bar{Q}_n(0) \Rightarrow \bar{Q}(0)$ *in* $\mathbb{R}^+$, *then* $\bar{Q}_n \Rightarrow \bar{Q}$ *in* $\mathcal{D}$ *as* $n \to \infty$. *The limit process* $\bar{Q}$ *is the unique solution satisfying the following integral equation*

$$\bar{Q}(t) = \bar{Q}(0) + \mu(0)t - \int_0^t \mu\left(\left(\bar{Q}(u)-1\right)^+\right)\left(\bar{Q}(u)\wedge 1\right)du - \int_0^t \theta\left(\bar{Q}(u)-1\right)^+ du.$$

The proof of Theorem 1 and all subsequent results can be found in Appendix A.

Let $f(q)$ be the flow rate function of the fluid system at state q. That is, $f(q) = \mu(0) - \mu((q-1)^+)(q\wedge 1) - \theta(q-1)^+$. Then we can write $\bar{Q}(t)$ as the solution to the following autonomous differential equation with initial value $\bar{Q}(0)$:

$$\dot{\bar{Q}} = f(\bar{Q}), \tag{2}$$

where $\dot{\bar{Q}}$ denotes the derivative of $\bar{Q}$ with respect to t.

4.2. Equilibrium analysis

Next, we analyze the limiting behavior of the fluid model, i.e., the state of the system as $t \to \infty$. To make the dependence of the flow, $\bar{Q}(t)$, on its initial value, $\bar{Q}(0)$, explicit, we write $\Phi(q_0,t) = \bar{Q}(t)$ with an initial value q_0. We start with the definitions of equilibrium and stability.

DEFINITION 1 (EQUILIBRIUM). A point $\bar{q}$ is an **equilibrium** of the dynamic system (2) if

$$\Phi(\bar{q},t) = \bar{q}, \text{ for all } t \geq 0.$$

By Definition 1, $\bar{q}$ is an equilibrium of the system if when the trajectory of the flow defined by (2) starts at $\bar{q}$, it stays there. In our model, $\bar{q}$ can be computed by solving $f(q) = 0$. However, it is unclear where the trajectories of the flow converge to, if the initial value $q_0 \neq \bar{q}$. We therefore analyze the stability of the equilibrium points.

DEFINITION 2 (STABILITY OF EQUILIBRIUM). Let $\bar{q}$ be an equilibrium of the dynamic system. $\bar{q}$ is said to be **stable** if for any $\epsilon > 0$, there exist $\delta > 0$, such that if $|q - \bar{q}| < \delta$, $|\Phi(q,t) - \bar{q}| < \epsilon$ for all $t \geq 0$. Otherwise, $\bar{q}$ is **unstable**. If δ can be chosen such that not only is $\bar{q}$ stable, but also $\lim_{t\to\infty} \Phi(q,t) = \bar{q}$ for $|q - \bar{q}| < \delta$, then $\bar{q}$ is said to be **asymptotically stable**.

By Definition 2, $\bar{q}$ is asymptotically stable if when starting close enough to $\bar{q}$, trajectories defined by (2) converge to $\bar{q}$ as $t \to \infty$. An equilibrium may also be **semistable**. For a semistable equilibrium, trajectories that start on one side of the equilibrium converge to it, whereas trajectories that start on the other side do not. Note that a semistable equilibrium is unstable by Definition 2.

We define bi-stability of a stochastic model based on its fluid limit (assuming it exists).

DEFINITION 3 (BI-STABILITY). A stochastic system is **bi-stable** if its corresponding fluid limit has two (semi-)stable equilibria.

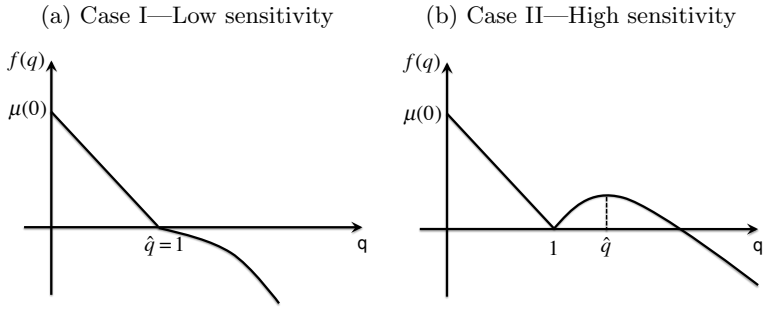
To characterize the equilibria of the fluid model in (2), we analyze the function $f(q)$ (as illustrated in Figure 3). When $q \leq 1$, $f(q) = \mu(0) - \mu(0)q$ is a linearly decreasing function that starts at $f(0) = \mu(0) > 0$ and ends at $f(1) = \mu(0) - \mu(0) = 0$. When $q \geq 1$, under Assumption 1, $f(q) = \mu(0) - \mu(q-1) - \theta(q-1)$, $f'(q) = -\mu'(q-1) - \theta$ and $f''(q) = -\mu''(q-1) \leq 0$. Therefore, $f(q)$ is concave on $[1, \infty)$. Let $\hat{q} = \arg\max_{q\in[1,\infty)} f(q)$. Depending on the actual form of $f(q)$, we distinguish between the following two cases (as shown in Figure 3):

Case I (Low Sensitivity): $-\mu'(0) \leq \theta$.

Case II (High Sensitivity): $-\mu'(0) > \theta$.

Under Case I, the case with low sensitivity, we have $\hat{q} = 1$ and under Case II, the case with high sensitivity, $\hat{q}$ is the root of $f'(q) = 0$ for $q > 1$. The following theorem summarizes the stability analysis of the equilibria for the two cases.

Figure 3 **Flow rate function under two cases**



THEOREM 2. *Under Assumption 1, the fluid approximation (2) has the following equilibria:*

i) If $-\mu'(0) \le \theta$ (Low Sensitivity), there is a unique equilibrium, $\bar{q}$, with $\bar{q} = 1$. Furthermore, $\bar{q}$ is asymptotically stable.

ii) If $-\mu'(0) > \theta$ (High Sensitivity), there are two equilibria, $\bar{q}_1$ and $\bar{q}_2$, with $\bar{q}_1 = 1$ and $\bar{q}_2 > \hat{q}$. Furthermore, $\bar{q}_1$ is a semistable equilibrium and $\bar{q}_2$ is an asymptotically stable equilibrium.

OBSERVATION 1. The dynamics of the load-sensitive Erlang-A model depend on the *relative* value of the load sensitivity around zero and the abandonment rate.

Following the results in Theorem 2, we expect different system dynamics under the two cases.

In the low sensitivity case, $\bar{q} = 1$ is the unique equilibrium of the fluid model. It is asymptotically stable. Therefore, the fluid model will converge to that value. In the original queueing system, we would expect to see the trajectory of the queue length process fluctuate around n. We analyze its performance in more detail in §5[1].

In the high sensitivity case, there are two equilibria, $\bar{q}_1$ and $\bar{q}_2$. The fluid model may converge to either one, depending on the starting point. In the original stochastic model, the queue length process may alternate between the two equilibria. This drives the bi-stability phenomenon observed in Figure 1c. However, $\bar{q}_1$ is a semistable equilibrium. Therefore, we expect the queue length process to eventually spend most of the time around the higher equilibrium level as the system scale grows large. We explore how the scale parameter n and other system parameters affect the bi-stability phenomenon under High Sensitivity in §6.

5. Performance Analysis under Low Sensitivity

In this section, we conduct asymptotic analysis for the modified Erlang-A model under low load sensitivity ($-\mu'(0) < \theta$). We establish closed-form approximations for the performance measures ($P(W)$ and $P(Ab)$), which can be used to determine the corresponding square-root staffing parameters. We then present numerical results to demonstrate the quality of the approximations.

Let Y_n denote the normalized steady-state queue length process. In particular,

$$Y_n = \frac{Q_n(\infty) - n}{\sqrt{n}}.$$

We then have the following result about the limiting distribution of Y_n.

THEOREM 3. *Under low sensitivity ($-\mu'(0) < \theta$) and SRS with parameter β, Y_n converges weakly to a random variable with the following probability density function*

$$g(y) = \begin{cases} \frac{C_1}{\sqrt{2\pi}} \exp\left(-\frac{(y+\beta)^2}{2}\right) & \text{if } y \le 0 \\ \frac{C_2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\beta\sigma^2)^2}{2\sigma^2}\right) & \text{if } y > 0, \end{cases}$$

[1] Section 5 does not analyze the boundary condition in which $-\mu'(0) = \theta$.

where

$$\sigma = \sqrt{\frac{\mu(0)}{\mu'(0)+\theta}}\ ,\ C_1 = \frac{h(\beta\sigma)}{\sigma\phi(\beta)}\left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\ ,\ C_2 = \frac{h(\beta\sigma)}{\phi(\beta\sigma)}\left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}$$

and $h(\cdot)$ *denotes the hazard rate function of the standard normal distribution. Specifically,* $h(z) = \phi(z)/\bar{\Phi}(z)$*, where* $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ *and* $\bar{\Phi}(z) = \int_z^\infty \phi(z)dz$ *is the complementary cumulative distribution function.*

Theorem 3 shows that the limiting distribution of the scaled process has normal tails but it is not symmetric around zero unless $(\mu'(0)+\theta)/\mu(0) = 1$, and the left tail decays slower as the sensitivity level $|\mu'(0)|$ increases.

From Theorem 3, we have the following asymptotic results about the performance measures.

COROLLARY 1. *Under low sensitivity* $(-\mu'(0) < \theta)$ *and SRS with parameter* β*,*

$$\lim_{n\to\infty} P_n(W) = \left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}$$

and

$$\lim_{n\to\infty} \sqrt{n}P_n(Ab) = \left(\frac{h(\beta\sigma)}{\sigma}-\beta\right)\left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu'(0)+\theta},$$

where $\sigma = \sqrt{\mu(0)/(\mu'(0)+\theta)}$.

Corollary 1 implies that the performance measures deteriorate with the load sensitivity level $\mu'(0)$, and it leads to the following approximations of the system performance measures:

$$P_n(W) \approx \left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1} \tag{3}$$

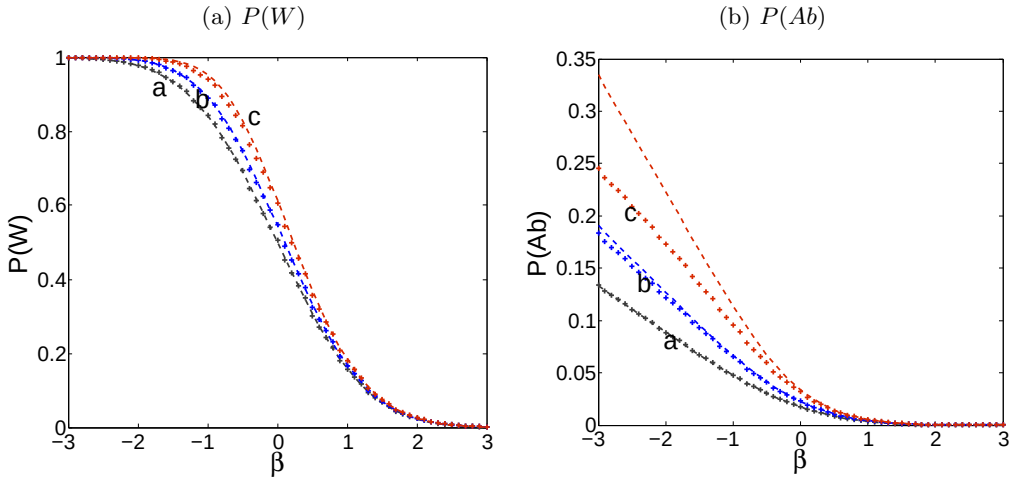
and

$$P_n(Ab) \approx \left(1-\frac{h(\beta\sigma)}{h(\beta\sigma+1/(\sigma\sqrt{n}))}\right)\left(1+\frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu'(0)+\theta}. \tag{4}$$

Figure 4 demonstrates the precision of these approximations (denoted by dashed lines) compared to the actual performance measures (marked by '+' signs), derived by simulation for different system parameters. Specifically, we choose three evenly spaced values of load sensitivity, measured by $\mu'(0)$, and the values of the square-root staffing parameter β between -3 to 3.

We observe that (3) provides a good approximation for $P(W)$ for a wide range of load sensitivity levels and β values. On the other hand, (4) provides a good approximation of $P(Ab)$ for only lower levels of load sensitivity ($|\mu'(0)| \le 0.3$). In other words, the precision of (4) deteriorates as the load sensitivity level approaches abandonment rate, $|\mu'(0)| \to \theta$; in that case the approximation tends to overestimate the system performance measures. Practically speaking, however, when applying QED staffing, people generally aim at achieving small abandonment rates (less than 10%). Restricting

Figure 4 **Approximations for $P(W)$ and $P(Ab)$ at three different load sensitivity levels: a: $\mu'(0)=0$, b: $\mu'(0)=-0.3$, c: $\mu'(0)=-0.6$**



attention to the range of β's which result in $P(Ab)<10\%$ (i.e. $\beta>-1$ when $\mu'(0)=-0.6$), (4) is a good approximation, with a maximum gap of only 0.025.

We also observe that for a fixed β, system performance ($P(W)$ and $P(Ab)$) deteriorates with the load sensitivity level $\mu'(0)$. Therefore, neglecting to account for load sensitivity would *underestimate* system performance. Put differently, fixing a target system performance, a load sensitive service system requires more staffing to achieve the same level of performance. One can use (3) and (4) to find the appropriate square-root staffing parameter to achieve certain performance measures in the QED regime.

OBSERVATION 2. We conclude this section by drawing some connections to the ordinary Erlang-A model. We notice that the limiting distribution of Y_n, in Theorem 3, is the same as the limiting distribution of the normalized queue length process of a sequence of ordinary Erlang-A model with the same arrival rate λ_n, constant service rate $\mu(0)$ and reduced abandonment rate $\mu'(0)+\theta$ in stationarity (Garnett et al. 2002). This is because, under the low sensitivity conditions, the two systems have the same arrival rate and very similar death rates. When $q<n$, the death rates of the two systems are equal; when $q>n$, the death rate of the modified Erlang-A queue is:

$$\begin{aligned}\mu\left(\frac{q-n}{n}\right)n+\theta(q-n)&=\mu(0)n+(\mu'(0)+\theta)(q-n)+\mu''(\eta)\frac{(q-n)^2}{n}\ \text{for some}\ \eta\in(0,(q-n)/n)\\&\approx\mu(0)n+(\mu'(0)+\theta)(q-n)\ \text{when}\ (q-n)^2/n\ \text{is small, i.e. when}\ q-n=O(\sqrt{n}).\end{aligned}$$

The reduced abandonment rate in the corresponding ordinary Erlang-A model suggests that the load sensitivity effectively lessens the stabilizing effect of abandonment. We also notice that as $\mu''(\cdot)\geq 0$, $\mu(\frac{q-n}{n})n+\theta(q-n)\geq\mu(0)n+(\mu'(0)+\theta)(q-n)$, the ordinary Erlang-A model with reduced abandonment rate is stochastically larger than the modified model (Lemma 3). Therefore,

the stationary queue length process, $Q_n(\infty)$, of our modified model is within $n \pm O(\sqrt{n})$ with high probability.

6. Bi-Stability Analysis: Performance Analysis under High Sensitivity

In this section, we analyze the system dynamics when sensitivity is high and, in particular, the factors that affect the bi-stability phenomenon. We start with the scale parameter n. We then keep n fixed and analyze the effect of other system parameters, specifically, the square-root staffing parameter β, the sensitivity of the service rate function and the abandonment rate θ. We also propose an admission control policy to eliminate the bi-stability effect and avoid ED performance.

6.1. The effect of the scale parameter n

We begin by characterizing the peaks (local maxima) of the stationary distribution. When bi-stability occurs there are two peaks, as was shown in Figure 1d. Naturally, there is a one-to-one correspondence between these peaks and the (semi-)stable equilibria of the fluid model. To see this, recall that $Q_n = \{Q_n(t) : t \geq 0\}$ is a B&D process with birth rate λ_n and state-dependent death rate $\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+$, where $Q_n(t) = q$. Let $\pi_n(\cdot)$ denote the steady-state probability density function of Q_n. Motivated by the fluid analysis, we also define the following flow rate function

$$f_n(q) = \lambda_n - \mu\left(\frac{(q-n)^+}{n}\right)(q \wedge n) - \theta(q-n)^+.$$

From the detailed balance equation $\lambda_n \pi_n(q) = \left(\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+\right)\pi_n(q+1)$, we get

$$\pi_n(q+1) - \pi_n(q) = \left(\frac{\lambda_n}{\mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+} - 1\right)\pi_n(q).$$

As a result, if $\lambda_n \geq \mu((q-n)^+/n)(q \wedge n) + \theta(q-n)^+$ $(f_n(q) > 0)$, then the stationary distribution is increasing in q, i.e. $\pi_n(q+1) \geq \pi_n(q)$; otherwise, if $f_n(q) < 0$, then the the stationary distribution is decreasing in q, i.e. $\pi_n(q+1) < \pi_n(q)$. Hence, we can find the values of peaks of $\pi_n(\cdot)$ by analyzing the function $f_n(q)$. We distinguish among three different cases, as illustrated in Figure 5, depending on the number of the roots of $f_n(q)$ (one to three). We analyze $f_n(\cdot)$ in two regions separately: $[0, n]$ and $[n, \infty)$.

Region $[0, n]$: $f_n(q)$ is linearly decreasing on $[0, n]$ with $f_n(0) = \lambda_n > 0$, $f_n(n) = \lambda_n - \mu(0)n$. When $\beta < 0$ (case (a) in Figure 5), $f_n(n) > 0$ and the function $f_n(\cdot)$ does not have a root in this region. When $\beta > 0$ (case (b) and (c) in Figure 5), $f_n(n) < 0$ and there exists a root $\bar{q}_{n,1} = \lambda_n/\mu(0) < n$, i.e. $f_n(\bar{q}_{n,1}) = 0$.

Region $[n, \infty)$: For $q > n$, if we let $x_n = (q-n)/n$, then we have

$$\frac{f_n(q)}{n} = \frac{\lambda_n}{n} - \mu(x_n) - \theta x_n.$$

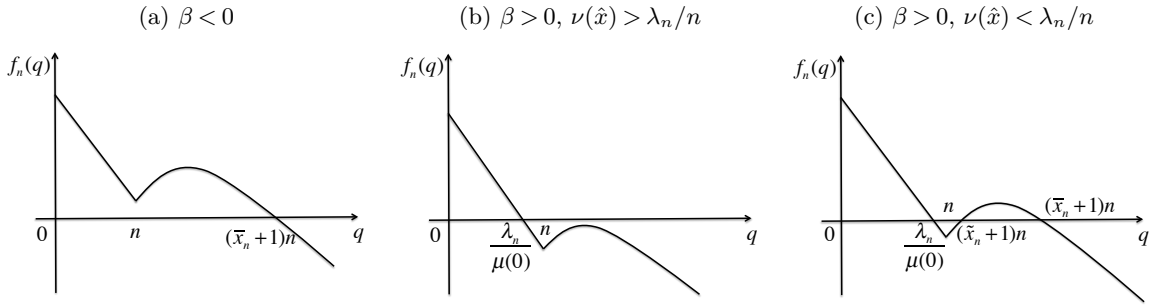
Let

$$\nu(x) := \mu(x) + \theta x$$

for $x \geq 0$. We denote $\hat{x}$ ($\hat{x} > 0$) as the root of $\nu'(x) = 0$. Under Assumption 1 and High Sensitivity, $\nu(\cdot)$ is convex and attains its minimum at $\hat{x}$. Thus $f_n(q)$ is increasing on $[n, n(\hat{x}+1)]$ and decreasing on $(n(\hat{x}+1), \infty)$. When $\beta < 0$ (case (a) in Figure 5), as $\nu(0) < \lambda_n/n$, there exists a unique $\bar{x}_n > \hat{x}$, such that $\lambda_n/n = \nu(\bar{x}_n)$. This implies that there is a root in this region: $(\bar{x}_n + 1)n$. When $\beta > 0$, since $\nu(0) > \lambda_n/n$, we have two cases: if $\nu(\hat{x}) > \lambda_n/n$ (case (b) in Figure 5), then $f_n(q) < 0$ for all $q > n$, hence, there is no root in this region; otherwise, $\nu(\hat{x}) < \lambda_n/n$ (case (c) in Figure 5) and there exists a unique $0 < \tilde{x}_n < \hat{x}$, such that $\lambda_n/n = \nu(\tilde{x}_n)$, and a unique $\bar{x}_n > \hat{x}$, such that $\lambda_n/n = \nu(\bar{x}_n)$. This implies that there are two roots in this region:$(\tilde{x}_n + 1)n$ and $(\bar{x}_n + 1)n$.

Based on the above analysis, we draw connections to the value of the peaks of the stationary distribution for the three cases accordingly.

Figure 5 $f_n(q)$ **with positive or negative** β**s**



(a) when $\beta < 0$, $f_n(q) \geq 0$ for $q \leq (\bar{x}_n + 1)n$, and $f_n(q) < 0$ for $q > (\bar{x}_n + 1)n$. Therefore, $\pi_n(\cdot)$ has only one peak, $\bar{q}_{n,2} = \lfloor (\bar{x}_n + 1)n \rfloor$;

(b) when $\beta > 0$ and $\nu(\hat{x}) > \lambda_n/n$, $f_n(q) > 0$ for $q \leq \lambda_n/\mu(0)$, and $f_n(q) < 0$ for $q > \lambda_n/\mu(0)$. Therefore, $\pi_n(\cdot)$ has only one peak, $\bar{q}_{n,1} = \lfloor \lambda_n/\mu(0) \rfloor$;

(c) when $\beta > 0$ and $\nu(\hat{x}) < \lambda_n/n$, $\pi_n(\cdot)$ has two peaks, one at $\bar{q}_{n,1} = \lfloor \lambda_n/\mu(0) \rfloor$ (because $f_n(q) \geq 0$ on $[0, \lambda_n/\mu(0)]$, $f_n(q) < 0$ on $(\lambda_n/\mu(0), (\hat{x}_n + 1)n)$) and the other at $\bar{q}_{n,2} = \lfloor (\bar{x}_n + 1)n \rfloor$ (because $f_n(q) \geq 0$ on $[(\hat{x}_n + 1)n, (\bar{x}_n + 1)n]$ and $f_n(q) \leq 0$ on $((\bar{x}_n + 1)n, \infty)$).

As $\lim_{n\to\infty} \lambda_n/n = \mu(0)$, when $\beta > 0$, the stationary distribution, $\pi_n(\cdot)$, may have a unique peak for small values of n, but will eventually have two peaks as n grows large.

Let $\bar{x}$ be the root of $\mu(0) - \nu(x) = 0$ on $(\hat{x}, \infty)$. The following lemma characterizes the value of $\tilde{x}_n$ and $\bar{x}_n$.

LEMMA 1. *Assume $\beta > 0$ and $\lambda_n/n > \nu(\hat{x})$ (this ensures the existence of $\tilde{x}_n$ and $\bar{x}_n$),*

$$\lim_{n\to\infty} \sqrt{n}\tilde{x}_n = -\frac{\beta\mu(0)}{\mu'(0) + \theta} \text{ and } \lim_{n\to\infty} \bar{x}_n = \bar{x}.$$

Lemma 1 implies that the distance between the $\lambda_n/\mu(0)$ and $(\tilde{x}_n+1)n$ is $O(\sqrt{n})$, which is why there are only two equilibria in the fluid limit.

We now characterize the relative *magnitude* of the two peaks ($\bar{q}_{n,1}$ and $\bar{q}_{n,2}$) as the system scale grows.

THEOREM 4. *Under High Sensitivity ($-\mu'(0)>\theta$) and SRS with $\beta>0$*

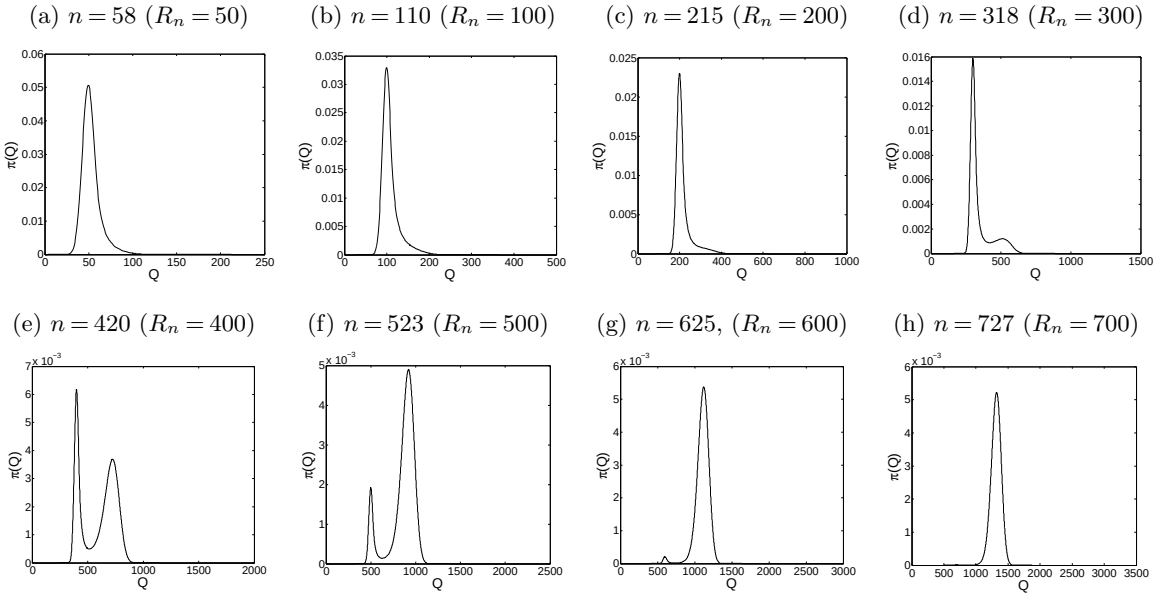
$$\lim_{n\to\infty}\frac{1}{n}\log\frac{\pi_n(\bar{q}_{n,2})}{\pi_n(\bar{q}_{n,1})}=I(\bar{x}),$$

where $I(\bar{x})=\int_0^{\bar{x}}\log\frac{\mu(0)}{\nu(x)}dx>0$.

We call $I(\cdot)$ the rate function. Theorem 4 indicates that $\pi_n(\bar{q}_{n,2})\approx\pi_n(\bar{q}_{n,1})\exp(nI(\bar{x}))$. This means that the difference in magnitude between the two peaks ($\pi_n(\bar{q}_{n,1})$ and $\pi_n(\bar{q}_{n,2})$) grows exponentially in n. Figure 6 demonstrates how the stationary distribution of the system with $\beta>0$ evolves with the scale parameter, n. For small values of n ($n\le 200$), $\pi_n(\cdot)$ has a unique peak ($\bar{q}_{n,1}$). As n increases, a "second peak" ($\bar{q}_{n,2}$) emerges and its magnitude compared to the first peak increases. For very large n, only $\bar{q}_{n,2}$ remains.

OBSERVATION 3. Theorem 4 suggests that unlike the traditional Erlang-A model that uses SRS, where a larger system provides better performance levels (e.g. smaller abandonment rate, shorter waiting times), the performance of our modified model with high service rate sensitivity deteriorates as the system scale grows.

Figure 6 Approximated stationary distribution of the number of people in the system for $M/M_Q/n+M$ queues with scale parameter values n ($n=\lceil R_n+\sqrt{R_n}\rceil$, $\mu=0.6+0.4\exp(-1.5(q-n)^+/n)$ and $\theta=0.3$).



We next analyze factors that affect the value of the rate function $I(\bar{x})$. To facilitate the comparison, we restrict our analysis to the following ordering of load sensitivity.

DEFINITION 4. For two service rate functions $\mu_i(\cdot)$ and $\mu_j(\cdot)$, with $\mu_i(0)=\mu_j(0)$, we say that μ_j is more load-sensitive than μ_i, if $\mu_j(x)\le\mu_i(x)$ for all $x>0$.

The next lemma examines the effect of the system parameters on the value of the higher level fluid equilibrium, $\bar{q}_2$, and the value of the rate function, $I(\bar{x})$.

LEMMA 2. *Under High Sensitivity ($-\mu'(0)>\theta$) and SRS with $\beta>0$,*

i) the more load-sensitive the service rate function μ is, the larger the value of the higher level fluid equilibrium $\bar{q}_2$ and the rate function $I(\bar{x})$;

ii) the larger the abandonment rate θ is, the smaller the value of the higher level fluid equilibrium $\bar{q}_2$ and the rate function $I(\bar{x})$.

OBSERVATION 4. Lemma 2 indicates that as the load sensitivity increases, the distance between the two equilibria increases, and with it the rate of convergence to the upper equilibria ($I(\bar{x})$). Hence, we observe bi-stability in smaller systems only. Abandonment rate has the opposite effect—as θ increases, the convergence to the upper equilibria is slower and, therefore, we observe bi-stability in larger systems.

6.2. The effect of other system parameters

For a fixed system scale parameter, n, in this section, we analyze the effect of the square-root staffing parameter, β, the service rate sensitivity and the abandonment rate, θ, on the bi-stability phenomenon. As the existence of the two peaks only arises for $\beta>0$ and large n, we restrict attention to these parameter ranges.

We begin by giving a formal definition for the time *around* the lower/upper equilibrium level. Define the threshold $\tilde{q}_n=\lfloor n(\tilde{x}_n+1)\rfloor$. Then, *the region around the lower equilibrium level* is $[0,\tilde{q}_n]$ and *the region around the upper equilibrium level* is $(\tilde{q}_n,\infty)$. As $\tilde{x}_n$ is the root of $\lambda_n/n-\nu(x)=0$ on $[0,\hat{x})$, $f_n(q)<0$ for $q\in(\bar{q}_{n,1},\tilde{q}_n)$ and $f_n(q)>0$ for $q\in(\tilde{q}_n,\bar{q}_{n,2})$. Thus, $\pi_n(q)$ is decreasing on $(\bar{q}_{n,1},\tilde{q}_n)$ and increasing on $(\tilde{q}_n,\bar{q}_{n,2})$, i.e. $\tilde{q}_n$ is the valley of $\pi_n(q)$ (see for example the valley around 600 in Figure 6f). Let P_{π_n} denote the steady-state distribution of Q_n. Then the proportion of time the system spends around the lower (or upper) equilibrium can be define as $P_{\pi_n}(Q_n\le\tilde{q}_n)$ ($P_{\pi_n}(Q_n>\tilde{q}_n)$).

The next lemma provides the basis for the main comparison results (Theorem 5) in this subsection.

LEMMA 3. *For two positive recurrent B&D processes, $Y^{(1)}$ and $Y^{(2)}$, defined on the same state space $\mathbb{Z}^+$, denote γ_i and $\xi_i(\cdot)$ as the birth rate and state-dependent death rate of $Y^{(i)}$, for $i=1,2$. If $\gamma_1=\gamma_2$ and $\xi_1(y)\ge\xi_2(y)$ for every $y\in\mathbb{Z}^+$, then*

$$P(Y^{(1)}(\infty)>y)\le P(Y^{(2)}(\infty)>y).$$

From Lemma 3 we have the following theorem that studies the effect of the system parameters on the proportion of time the system spends around each equilibrium.

THEOREM 5. *Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta > 0$,*

i) if $\mu(\infty) \geq \theta$, then the proportion of time the system spends around the upper equilibrium decreases with the square-root staffing parameter, β;

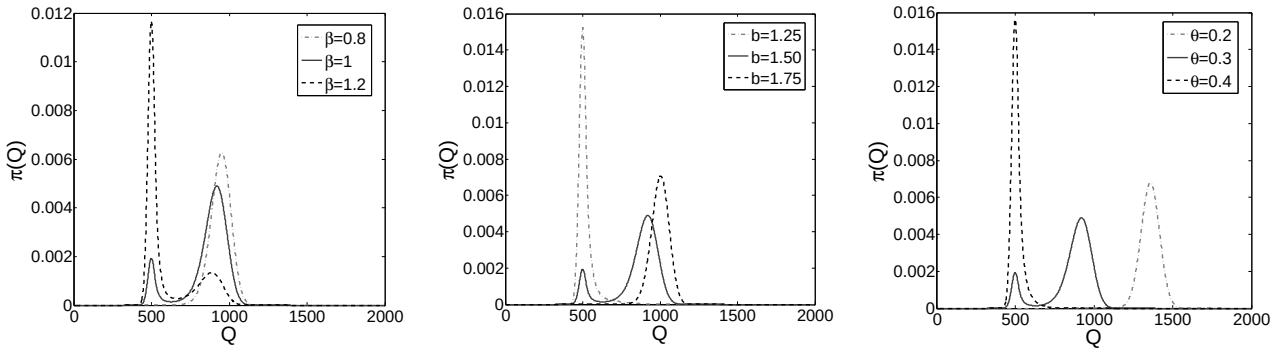
ii) under Definition 4, the more load-sensitive the service rate function is, the larger the proportion of time the system spends around the upper equilibrium;

iii) the proportion of time the system spends around the upper equilibrium decreases with the abandonment rate, θ.

Figure 7 demonstrates how the value of the peaks and proportion of time the system spends around each peak changes with the square-root staffing parameter β, the sensitivity parameter, b, and the abandonment rate, θ. We notice that the value of the second peak (the larger one) increases with the load sensitivity parameter b and decreases with the abandonment rate θ, as was proved in Lemma 2. In addition, we notice that the value of the second peak decreases with the square-root staffing parameter β, but the difference is much smaller when compared to the effect of b and θ. (The change associated with β is not apparent in the fluid level and, hence, less significant.)

Figure 7 **Approximated stationary distribution of the number of people in the system for $M/M_Q/n+M$ queues with different system parameters ($n=\lambda+\beta\sqrt{\lambda}$, $\lambda=500$, $\mu=0.6+0.4\exp(-b(q-s)^+/s)$ and θ).**

(a) β ($b=1.5$,$\theta=0.3$) (b) b ($\beta=1$, $\theta=0.3$) (c) θ ($\beta=1$, $b=1.5$)



The effect of β on the performance measures of our load-sensitive model is similar to that of the traditional/nonsensitive Erlang-A model; $P(W)$ and $P(Ab)$ both decrease with β. The effect of load sensitivity is also straightforward. The system with a less sensitive service rate function has on average a higher service rate. The performance measures hence improve. In contrast, the effect of the abandonment rate, θ, is quite counterintuitive. In the traditional Erlang-A model, it is well established that if customers are less patient (i.e., θ increases), $P(W)$ decreases but $P(Ab)$ increases (Garnett et al. 2002), while in our modified Erlang-A model with high sensitivity, both

the probability of waiting and the probability of abandonment decrease with θ. This is because the load-sensitive system reaches the high equilibrium less frequently as θ increases.

The analysis implies that the abandonment rate and the load sensitivity of the service rate function affect system performance differently when service rates exhibit slowdowns due to congestion. While a high load sensitivity level negatively affects system performance, a high abandonment rate may actually improve performance by alleviating the deterioration in service rate. Hence, managers are advised to *encourage* customers to abandon in a load-sensitive environment. This can be done, for example, by providing delay announcements when the system is loaded, as it was shown that announcements increase abandonment rate (Mandelbaum and Zeltyn 2007, Huang et al. 2014).

6.3. An admission control policy to avoid bi-stability under High Sensitivity

We next introduce an admission control policy to eliminate the bi-stability phenomenon and avoid ED regime performance under high sensitivity. If we want to eliminate the higher-level equilibrium by increasing the staffing level, then the new staffing level should be such that $\lambda_n/\bar{n} \leq \nu(\hat{x})$. Suppose we set $\bar{n} = \lceil \lambda_n/\nu(\hat{x}) \rceil$. Then we have

$$\begin{aligned} \frac{\bar{n}-n}{n} &= \left\lceil \frac{\lambda_n/n}{\nu(\hat{x})} \right\rceil - \left\lceil \frac{\lambda_n/n}{\mu(0)} - \frac{\beta}{\sqrt{n}} \sqrt{\frac{\lambda_n/n}{\mu(0)}} \right\rceil \\ &\to \left\lceil \frac{\lambda}{\nu(\hat{x})} \right\rceil - 1 \end{aligned}$$

This implies that we need to increase staffing by $O(n)$ servers, which may be very costly. Another potential drawback of this approach is that by raising the staffing level to $\bar{n}$, a service provider may "overstaff" the system to operate in the QD regime. We thus consider an alternative admission control policy. Specifically, we block the incoming arrivals or reroute them to other service facilities once a certain threshold, c, is reached, thereby preventing the system from reaching the higher level equilibrium. To implement this policy, the system provider needs to characterize the appropriate threshold level, and the cost that such a policy entails on the system in terms of the proportion of customers blocked/rerouted.

The "right" threshold could again be chosen based on our bi-stability analysis. Basically, any choice of c_n, satisfying $n < c_n \leq (\tilde{x}_n + 1)n$, eliminates bi-stability, but the choice presents a tradeoff between the level of performance and the proportion of customers blocked: Setting a small c_n improves performance ($P(W)$ and $P(Ab)$ are low), but increases the proportion of customers that are blocked ($P(Bl)$).

Under the admission control policy, the system becomes a multi-server queue with finite waiting room. To gain more insight on its performance, we conduct some asymptotic analysis on its performance. Specifically, we consider a sequence of $M/M_Q/n/c_n + M$ queues indexed by n. System n has

arrival rate λ_n, state-dependent service rate $\mu((q-n)^+/n)$, abandonment rate θ and a finite system capacity c_n, so that incoming customers are blocked once the number of customers in the system reaches c_n. We denote the queue length process of the n-th system by $Q_n^c(\cdot)$. We next develop a diffusion approximations for Q_n^c for $c_n \le (\tilde{x}_n+1)n$.

A pathwise construction of Q_n^c is

$$Q_n^c(t) = Q_n^c(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du\right) - R\left(\theta\int_0^t (Q_n^c(u)-n)^+ du\right) - L_n(t),$$

where $L_n(t) = \int_0^t 1\{Q_n^c(s) = c_n\} dA(\lambda_n t)$. L_n counts the number of arrivals that are blocked from the system in $[0,t]$. We define the diffusion-scaled process

$$\hat{Q}_n^c(t) := \frac{Q_n^c(t)-n}{\sqrt{n}}.$$

THEOREM 6. *Assume $\sqrt{n}(1-\rho_n) \to \beta$ as $n\to\infty$, where $\rho_n = \lambda_n/(n\mu(0))$ and $c_n/\sqrt{n} \to c \le -\beta\mu(0)s/(\mu'(0)+\theta)$ as $n\to\infty$. If $\hat{Q}_n^c(0) \Rightarrow \hat{Q}^c(0)$ in $\mathbb{R}$ as $n\to\infty$, then $\hat{Q}_n^c \Rightarrow \hat{Q}^c$ in $\mathcal{D}$ as $n\to\infty$. The limit process $\hat{Q}^c$ is the unique process satisfying the stochastic integral equation:*

$$\hat{Q}^c(t) = \hat{Q}^c(0) - \beta\mu(0)t + \sqrt{2\mu(0)}B(t) - \int_0^t \left[\mu(0)(\hat{Q}^c(u)\wedge 0) + (\mu'(0)+\theta)\hat{Q}^c(u)^+\right] du - \hat{L}(t), \quad (5)$$

where $\{B(t): t\ge 0\}$ is a standard Brownian motion. $\hat{L}$ is the unique nondecreasing nonnegative process in $\mathcal{D}$ satisfying equation (5) and $\int_0^\infty 1\left\{\hat{Q}^c(t) < c\right\} d\hat{L}(t) = 0$.

Q_n^c is an irreducible Markov chain with a finite state space. Thus, $\hat{Q}^c$ admits a unique stationary distribution, π. As $E_\pi[\hat{Q}^c(t)] = E_\pi[\hat{Q}^c(0)]$, by Theorem 6 and the Basic Adjoint Relation (Chen and Yao 2001),

$$E_\pi[\hat{L}(t)] = \left(-\beta\mu(0) - \mu(0)E_\pi[\hat{Q}^c(0)\wedge 0] - (\mu'(0)+\theta)E_\pi[\hat{Q}^c(0)^+]\right) t$$

and the proportion of customers that are blocked from the n-th system, $P_n(Bl)$, satisfies

$$\begin{aligned} P_n(Bl) &\approx \frac{\sqrt{n}E_\pi[\hat{L}(t)]}{\lambda_n t} \\ &= \frac{1}{\sqrt{n}} \frac{\left(-\beta\mu(0) - \mu(0)E_\pi[\hat{Q}^c(0)\wedge 0] - (\mu'(0)+\theta)E_\pi[\hat{Q}^c(0)^+]\right)}{\mu(0)}. \end{aligned}$$

The probability of blocking is of $O(1/\sqrt{n})$. This implies that for large systems, the proportion of customers blocked and the proportion of time the system is blocked are very small. As the system is restricted to fluctuate around the lower equilibrium $\bar{q}_1$, we expect QED regime performance for $P(W)$ and $P(Ab)$, i.e. non-degenerate probability of waiting and $O(1/\sqrt{n})$ probability of abandonment.

Table 1 compares the performance of a load sensitive queue with admission control and without (Base) for different load sensitive parameters (b). We observe that the admission control policy keeps the performance measures within the QED regime characteristics for all sensitivity parameters tested. The performance measures are much improved—the probability of waiting is reduced from 98–100% to 20–50%, the probability of abandonment is reduced from 20–40% to 0.1–0.9%, and the "cost" of such policy, measured by the proportion of customers blocked, is quite low (at most 1.51% for the parameters tested in Table 1). Note that the total proportion of lost customers, i.e. $P(Ab)+P(Bl)$, is significantly lower than the base line—reduced from 20–40% to 0.13–0.16%.

Table 1 Performance comparison of systems with different load sensitivity parameter, b. **($\mu(q)=0.6+0.4\exp(-b_i(q-n)^+/n)$, $\lambda=500$, $n=511$ and $\theta=0.3$)**

	Base		Admission control				
b	$P(W)$	$P(Ab)$	c_n	$P(W)$	$P(Ab)$	$P(Bl)$	$P(Ab)+P(Bl)$
1.25	0.9830	0.2021	579	0.4873	0.0090	0.0057	0.0147
1.75	1	0.3199	541	0.3426	0.0031	0.0107	0.0138
2.25	1	0.3562	530	0.2583	0.0016	0.0135	0.0151
2.75	1	0.3718	525	0.2056	0.0010	0.0151	0.0161

7. Extensions

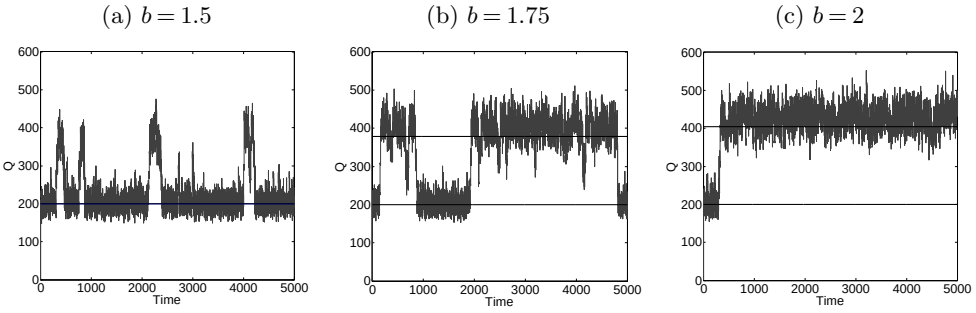
The model in Section 3 is the most befitting to explain agent-driven slowdowns, where the effect of the load of the system on service rates is applied instantly and to all agents simultaneously. Such a model fits situations where agents observe the current load and adjust their working rates accordingly. The exact same model cannot be directly applied to capture all various sources of the slowdown effect described in the introduction. In this section, we modify the base model to better fit other sources of slowdown effects. In particular, we study: a) Customer-driven slowdowns, where each customer's waiting time affects only her own service rate, and b) Agent-driven slowdowns with a time lag, which can explain slowdowns caused by fatigue and hence takes time to take effect. Utilizing numerical approaches, we find that the primary insights (the existence of the two equilibria and the stochastic fluctuations between them) from our original model remain.

7.1. Customer-driven slowdown

In this section, we assume that a customer's service time is positively correlated with his own waiting time. In particular, the service time of customer i is distributed as an exponential random variable with rate $\mu_i=\mu(w_i\lambda/n)$ where w_i is the waiting time of customer i (the total amount of time customer i waits before entering service). Note that in this setting, the service time does not change once the customer is taken into service. Chan et al. (2013) analyze a model similar in spirit by establishing a more tractable upper bound system. However, as their model doesn't allow abandonments, they must put a bound on the slowdown effect and bi-stability doesn't arise there.

We expect the customer-driven model to exhibit similar behavior to our modified Erlang-A model, because if we denote the number of people in the system when customer i enters service by q_i, then by Little's law, we have $E[(q_i - n)^+] = \lambda E[w_i]$. We test this intuition via simulation. Figure 8 plots the sample paths of the queue length process for the customer-driven slowdown model for different values of the load-sensitivity parameter b. The solid horizontal lines are the equilibrium levels predicted by the modified Erlang-A model with load-dependent service rate function $\mu((Q-n)^+/n)$. For $b = 1.5$, we only have one equilibrium point for the $M/M_Q/n$ queue (see Figure 5 (b)). We observe the bi-stability still exists in this case and as the load-sensitivity parameter b increases, the system spends more time around the upper equilibrium level. But we also notice that the bi-stability phenomenon occurs for smaller values of b than the corresponding $M/M_Q/n + M$ model (Figure 8 (a)). Moreover, while the lower equilibrium levels are about the same in both models, the upper equilibrium level is larger in the customer-driven slowdown model than in the corresponding $M/M_Q/n + M$ model (Figure 8 (b) and (c)).

Figure 8 **Sample paths of the number of people in the system with different sensitivity parameters,** b **(**$n = 214$**,** $\lambda = 200$**,** $\theta = 0.3$ **and** $\mu_i = 0.6 + 0.4\exp(-bw_i\lambda/n)$**.)**



7.2. Agent-driven slowdown with a lag

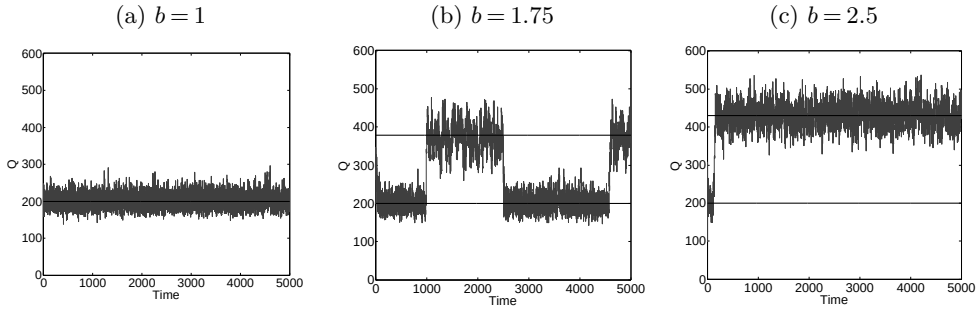
Another possible cause for the slowdown effect is due to fatigue of agents. Under high congestion levels, agents are working under pressure and without proper rest, which may eventually lead to deterioration in productivity. One way to model a slowdown effect caused by fatigue is to incorporate a time lag between the occurrence of high congestion levels and the deterioration in productivity. To capture this, we set the service rate as a function of the average queue length process over a time interval of length l, where l is of the same order as the service times. Specifically, the service rate at time t is

$$\mu\left(\int_{t-l}^{t} \frac{(Q(u)-n)^+}{n} du\right).$$

We observe that bi-stability still exists in this case. We find that the length of the time lag, l, affects the frequency at which the system moves between the two equilibria. When l is relatively

small, the system moves "easily" from one equilibrium to the other; as l increases, it becomes "harder" for the system to move between the two equilibria. Figure 9 illustrates how the bi-stability phenomenon evolves as the sensitivity parameter, b, increases, for a relatively small time lag (e.g., $l=5$). The solid horizontal lines are the equilibrium levels predicted by our modified Erlang-A model with service rate function $\mu((Q-n)^+/n)$. We observe that the proportion of time the system spends around the higher equilibrium grows with b and the equilibrium levels are about the same as the corresponding $M/M_Q/n+M$ model. In contrast, for a large time lag (e.g., $l=30$), the trajectory of the system depends largely on its initial position, and tends to stay around the initial equilibrium level for a very long period of time (potentially forever), regardless of the value of b. Figure 10 demonstrates the sample paths commonly observed in this case. If the system starts around the lower equilibrium, it keeps fluctuating around that level (Figure 10 (a)), whereas if the system starts around the higher equilibrium, then it stays there (Figure 10 (b)). These observations have operational implications—whereas systems with short lags will move from the upper equilibrium to the lower equilibrium on their own, systems with long lags may need external interventions to move from the upper equilibrium to the lower one.

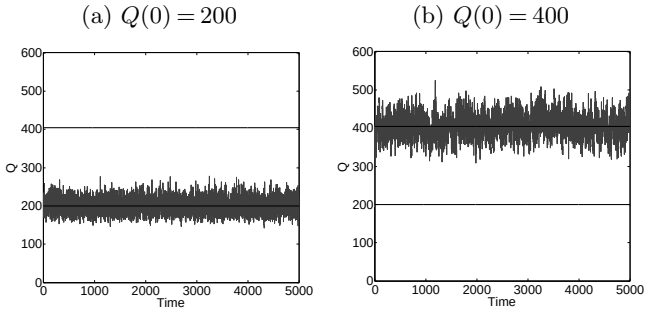
Figure 9 **Sample paths of the number of people in the system with time lag of length $l=5$ and different levels of the sensitivity parameter, b ($n=214$, $\lambda=200$, $\theta=0.3$ and $\mu\left(\int_{t-l}^{t}(Q(u)-n)^+/n\,du\right)=0.6+0.4\exp\left(-b\int_{t-l}^{t}(Q(u)-s)^+/s du\right)$.)**



8. Concluding Remarks

Motivated by empirical findings in service systems, we modified the Erlang-A model to account for the effect of workload-dependent service rates. When the load sensitivity is low relative to the abandonment rate, we observe a small gap between the performance of the standard Erlang-A model and the load-sensitive model, though the latter has lower quality of service. We show that this reduction in quality measures can be fixed by adjusting the square-root staffing rule parameter. When the load sensitivity is high, we observe a bi-stability phenomenon where the system alternates between two equilibria: one equilibrium results in QED performance and the other equilibrium results in ED performance. We conduct a sensitivity analysis on the proportion

Figure 10 **Sample paths of the number of people in the system with time lag of length $l = 30$ and different initial queue lengths ($s = 214$, $\lambda = 200$, $\theta = 0.3$ and $\mu\left(\int_{t-l}^{t}(Q(u)-s)^{+}/sdu\right) = 0.6 + 0.4\exp\left(-2\int_{t-l}^{t}(Q(u)-s)^{+}/sdu\right)$.)**



of time the system spends around each equilibrium and propose an admission control policy to achieve QED performance in this case. Lastly, we illustrate via numerical experiments that the bi-stability phenomenon persists in a broader class of load-sensitive service systems which extend to different sources of the slowdown effect.

We would like to conclude with some remarks regarding the construction of the model. First, for the sake of simplicity, throughout the manuscript, we assume a decreasing and convex service rate function and a constant abandonment rate. These assumptions lead to the bi-stability results. We notice that meta-stability (multiple (semi-)stable equilibria) can arise for more general forms of the service rate function and load-dependent abandonment rate. Most of the analyses in this paper (the fluid analysis and the asymptotic analysis of the stationary distribution) can be applied to the more general cases as well. Our second remark is on the practical estimation of the service rate function. From our analyses, it is apparent that to design service systems with a load-dependent slowdown effect, it is sufficient to accurately estimate the service rate function around zero, for most purposes. The derivative of the service rate function at zero is all that is needed to distinguish between the low and the high sensitivity cases, and to approximate the performance measures in the low sensitivity case. To implement the admission control policy in the high sensitivity case, it is sufficient to estimate the service rate function up to $O(1/\sqrt{n})$.

Acknowledgements

The authors would like to thank the area editor, associate editor and two referees for their insightful comments and suggestions. We also thank Noah Gans and Junfei Huang for providing helpful comments. Some of the third author's research was carried out while being a postdoctoral researcher at Columbia University. The hospitality of this institution is acknowledged and truly appreciated.

References

Antunes, N., C. Fricker, P. Robert, D. Tibi. 2009. Stochastic networks with multiple stable points. *The Annals of Probability* **36**(1) 255–278.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 68–81.

Batt, R.J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times. Working Paper.

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

Bertrand, J.W.M., H.P.G. van Ooijen. 2002. Workload based order release and productivity: a missing link. *Production Planning and Control* **12**(7) 665–678.

Boxma, O.J., M. Vlasiou. 2007. On queues with service and interarrival times depending on waiting times. *Queueing Systems* **56** 121–132.

Caldwell, J.A. 2001. The impact of fatigue in air medical and other types of operations: a review of fatigue facts and potential countermeasures. *Air Medical Journal* **20**(1) 25–32.

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

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

Chan, C.W., G.B. Yom-Tov, G. Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62**(2) 462–482.

Chen, H., D.D. Yao. 2001. *Fundamentals of queueing networks: performance, asymptotics and optimization*. Springer-Verlag.

Daley, D.J., J.S.H. van Leeuwaarden, Y. Nazarathy. 2013. BRAVO for many-server QED systems with finite buffers. Working paper.

Feldman, P., J. Li, G.B. Yom-Tov, E. Yom-Tov. 2014. Service time sensitivity to load: Who is to "blame"? Working Paper.

Fleming, P.J., A. Stolyar, B. Simon. 1994. Heavy traffic limit for a mobile phone system loss model. *Proc. 2nd Internat. Conf. Telecommunication Systems, Modeling, and Analysis*. Nashville, TN, 158–176.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208–227.

Gerla, M., L. Kleinrock. 1980. Flow control: A comparative survey. *IEEE Transactions on Communications* **28**(4) 553–574.

Gibbens, R.J., P.J. Hunt, F.P. Kelly. 1990. Bistability in communication networks. *Disorder in Physical Systems*. Oxford University Press, 113–128.

Green, L.V., P.J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **26**(1) 13–39.

Hollander, F. Den. 2004. Metastability under stochastic dynamics. *Stochastic Process. Appl.* **114**(1) 1–26.

Huang, J., A. Mandelbaum, H. Zhang, J. Zhang. 2014. Refined models for efficiency-driven queues with applications to delay announcements and staffing. Working Paper.

Janssen, A.J.E.M., J.S.H. van Leeuwaarden. 2014. Staffing many-server systems with admission control and retrials. Working paper.

Janssen, A.J.E.M., J.S.H. van Leeuwaarden, J. Sanders. 2013. Scaled control in the QED regime. Working paper.

KC, D., C. Terwiesch. 2009. Impact of work load on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

Kocaga, Y.L., M. Armony, A. Ward. 2014. Staffing call centers with uncertain arrival rate and co-sourcing. Working paper.

Mandelbaum, A., G. Pats. 1998. State-dependent stochastic networks. part I: Approximations and applications with continuous diffusion limits. *The Annals of Applied Probability* **8**(2) 569–646.

Mandelbaum, A., S. Zeltyn. 2007. Service engineering in action: The Palm/Erlang-A queue with applications to call centers. D. Spath, K. P. Fahnrich, eds., *Advances in Service Innovations*. Springer-Verlag, 17–48.

Olivieri, E., E.M. Vares. 2005. *Large Deviation and Metastability*. Cambridge Univ. Press.

Palm, C. 1957. Research on telephone traffic carried by full availability groups. *Tele* **1** 107.

Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193–267.

Weerasinghe, A. 2013. Diffusion approximation for G/M/n+GI queues with state-dependent service rate. *Mathematics of Operations Research* .

Whitt, W. 1990. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* **6** 335–352.

Whitt, W. 2004. Efficiency-driven heavy traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.

Wickens, C., J. Hollands, R. Parasuraman, S. Banbury. 2012. *Engineering Psychology and Human Performance*. 4th ed. Pearson.

Yom-Tov, G., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *M&SOM* **16**(2) 283–299.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science* **48** 566–583.

Appendix A: Proofs

Proof of Theorem 1. The proof follows from the method outlined in Pang et al. (2007). We write

$$\begin{aligned} Q_n(t) &= Q_n(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du\right) - R\left(\theta\int_0^t (Q_n(u)-n)^+du\right) \\ &= Q_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) \\ &\quad +\lambda_n t - \int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du - \theta\int_0^t (Q_n(u)-n)^+du \end{aligned}$$

where

$$\begin{aligned} M_{n,1}(t) &= A(\lambda_n t) - \lambda_n t \\ M_{n,2}(t) &= S\left(\int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du\right) - \int_0^t \mu\left(\frac{(Q_n(u)-n)^+}{n}\right)(Q_n(u)\wedge n)du \\ M_{n,3}(t) &= R\left(\theta\int_0^t (Q_n(u)-n)^+du\right) - \theta\int_0^t (Q_n(u)-n)^+du. \end{aligned}$$

Let $\bar{Q}_n(t) = Q_n(t)/n$ and $\bar{M}_{n,i} = M_{n,i}/n$ for $i = 1,2,3$. Then

$$\begin{aligned} \bar{Q}_n(t) &= \bar{Q}_n(0) + \bar{M}_{n,1}(t) - \bar{M}_{n,2}(t) - \bar{M}_{n,3}(t) \\ &\quad +\frac{\lambda_n}{n}t - \int_0^t \mu\left((\bar{Q}_n(u)-1)^+\right)\left(\bar{Q}_n(u)\wedge 1\right)du - \theta\int_0^t \left(\bar{Q}_n(u)-1\right)^+du. \end{aligned}$$

Let $d(q) = -\mu((q-1)^+)(q\wedge 1) - \theta(q-1)^+$. As $\mu'(\cdot)\le 0$ and $\mu''(\cdot)\ge 0$, $|\mu'(x)|\le|\mu'(0)|$. It is easy to check that

$$|d(q_1) - d(q_2)| \le \max\{\mu(0), |\mu'(0)| + \theta\}|q_1 - q_2|.$$

Thus $d(\cdot)$ is Lipschitz. This implies that

$$q(t) = b + x(t) + \int_0^t d(q(u))du$$

has a unique solution and constitutes a function $\phi : \mathcal{D}\times R \to \mathcal{D}$ that is continuous (see Theorem 4.1 in Pang et al. (2007)).

Let $\eta(t)\equiv 0$. We next show that $\bar{M}_{n,i}\to\eta$ in $\mathcal{D}$ w.p. 1 as $n\to\infty$ for $i = 1,2,3$.

Applying the Functional Strong Law of Large Numbers to Poisson processes, we have $\sup_{0\le t\le T}\left\{\frac{A(nt)}{n} - t\right\}\to 0$, $\sup_{0\le t\le T}\left\{\frac{S(nt)}{n} - t\right\}\to 0$ and $\sup_{0\le t\le T}\left\{\frac{R(nt)}{n} - t\right\}\to 0$ w.p. 1 as $n\to\infty$ for any $T>0$. We thus have

$$\bar{M}_{n,1}\to\eta \text{ in } \mathcal{D} \text{ w.p. 1 as } n\to\infty.$$

As $Q_n(t) < Q_n(0) + A(\lambda_n t)$, $\int_0^t Q_n(u)du \le t\left(Q_n(0) + A(\lambda_n t)\right)$. This implies that for any fixed $T>0$ there exists $\tau>0$, such that

$$P\left(\frac{\mu(0)}{n}\int_0^T Q_n(u)du > \tau\right)\to 0 \text{ as } n\to\infty.$$

Then,

$$P\left(\left\|\bar{M}_{n,2}\right\|_T > \epsilon\right) \le P\left(\frac{\mu(0)}{n}\int_0^T Q_n(u)du > \tau\right) + P\left(\left\|\frac{S(nt)}{n} - t\right\|_\tau > \frac{\epsilon}{2}\right).$$

This leads to

$$\bar{M}_{n,2} \to \eta \text{ in } \mathcal{D} \text{ w.p. } 1 \text{ as } n \to \infty.$$

Similarly we can show that

$$\bar{M}_{n,3} \to \eta \text{ in } \mathcal{D} \text{ w.p. } 1 \text{ as } n \to \infty.$$

By the Continuous Mapping Theorem (CMT) we have the fluid limit in Theorem 1. □

Proof of Theorem 2. We prove asymptotic stability by the Lyapunov method. Specifically, a function $V(q): \mathbb{R}^+ \to \mathbb{R}^+$ is called a Lyapunov function of (2) about its equilibrium $\bar{q}$ if $V(\bar{q}) = 0$ and $V(q) > 0$, $0 < |q - \bar{q}| < \delta$ for some $\delta > 0$. We denote $\dot{V}$ as the derivative of $V(\cdot)$ with respect to q. $\bar{q}$ is **locally asymptotically stable**, if there exists a Lyapunov function $V(q)$, such that $\dot{V}(q) < 0$ for all $0 < |q - \bar{q}| < \delta$ for some $\delta > 0$. $\bar{q}$ is **globally asymptotically stable**, if the locally asymptotically stable conditions hold for all $\delta \in \mathbb{R}^+$.

For the low sensitivity case, we use the following Lyapunov function

$$V(q) = |q - \bar{q}|,$$

where $\bar{q}$ is the specified equilibrium. Hence,

$$\dot{V}(q) = \text{sign}(q - \bar{q}) f(q).$$

Recall that $f(q) = \mu(0) - \mu((q-1)^+)(q \wedge 1) - \theta(q-1)^+$.

Under Assumption 1 and the assumptions of the low sensitivity case, $f(\cdot)$ is decreasing. $\bar{q} = \mu(0)/\mu(0) = 1$ and

$$\dot{V}(q) = \begin{cases} -\mu(0) + \mu(0)q < -\mu(0) + \mu(0)\bar{q} = 0, & q < \bar{q}; \\ \mu(0) - \mu(q-1) - \theta(q-1) < \mu(0) - \mu(0) = 0, & q > \bar{q}. \end{cases}$$

Therefore, $\bar{q}$ is a globally asymptotically stable equilibrium.

Under Assumption 1 and the assumptions of the high sensitivity case, $f(1) = 0$; thus $\bar{q}_1 = 1$. $f(q)$ is increasing on $[\bar{q}_1, \hat{q})$ and decreasing on $[\hat{q}, \infty)$. Since $f(\hat{q}) > 0$ and $\lim_{q\to\infty} f(q) = -\infty$, there exists $\bar{q}_2 > \hat{q}$ such that $f(\bar{q}_2) = 0$.

As $f(q) > 0$ for $q < 1$ and $f(q) > 0$ for $1 < q < \hat{q}$, $\bar{q}_1$ is semistable.

Let

$$V_2(q) = |q - \bar{q}_2|.$$

For $q \in (\bar{q}_1, \infty)$,

$$\dot{V}_2(q) = \begin{cases} -\mu(0) + \mu(q-1) + \theta(q-1) < -\mu(0) + \mu(0)\bar{q}_1 = 0, & \bar{q}_1 < q \le \hat{q}, \\ -\mu(0) + \mu(q-1) + \theta(q-1) < -\mu(0) + \mu(\bar{q}_2 - 1) + \theta(\bar{q}_2 - 1) = 0, & \hat{q} < q < \bar{q}_2, \\ \mu(0) - \mu(q-1) - \theta(q-1) < \mu(0) - \mu(\bar{q}_2 - 1) - \theta(\bar{q}_2 - 1) = 0, & q > \bar{q}_2. \end{cases}$$

Therefore, $\bar{q}_2$ is a locally asymptotically stable equilibrium. □

In order to prove Theorem 3, we start with the following lemma.

LEMMA 4. *Assume* $\sqrt{n}(1-\lambda_n/(n\mu(0))) \to \beta$ *as* $n \to \infty$. *For any* $0 < y_1 < y_2 < \infty$,

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} = -\int_{y_1}^{y_2} \beta + \frac{\mu'(0)+\theta}{\mu(0)} y \, dy$$

and

$$\lim_{n\to\infty} \log \frac{\pi_n(\lfloor n-\sqrt{n}y_1 \rfloor)}{\pi_n(\lfloor n-\sqrt{n}y_2 \rfloor)} = -\int_{-y_2}^{-y_1} \beta + y \, dy.$$

Proof of Lemma 4. From the detailed balance equation of the B&D process, we have

$$\frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} = \prod_{k=\lfloor n+\sqrt{n}y_1 \rfloor+1}^{\lfloor n+\sqrt{n}y_2 \rfloor} \frac{\lambda_n}{\mu\left((k-n)/n\right) n + \theta(k-n)}.$$

Then,

$$\begin{aligned}
\log \frac{\pi_n(\lfloor n+\sqrt{n}y_2 \rfloor)}{\pi_n(\lfloor n+\sqrt{n}y_1 \rfloor)} &= \lfloor (y_2-y_1)\sqrt{n} \rfloor \log \rho_n - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \log\left(1 + \frac{\mu(\frac{k}{n}) - \mu(0) + \theta\frac{k}{n}}{\mu(0)}\right) \\
&= -\lfloor (y_2-y_1)\sqrt{n} \rfloor (1-\rho_n) - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \frac{\mu'(0)+\theta}{\mu(0)} \frac{k}{\sqrt{n}} \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \\
&\to -(y_2-y_1)\beta - \int_{y_1}^{y_2} \frac{\mu'(0)+\theta}{\mu(0)} y dy.
\end{aligned}$$

Likewise,

$$\begin{aligned}
\log \frac{\pi_n(\lfloor n-\sqrt{n}y_1 \rfloor)}{\pi_n(\lfloor n-\sqrt{n}y_2 \rfloor)} &= \lfloor (y_2-y_1)\sqrt{n} \rfloor \log \rho_n - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} \log(1-\frac{k}{n}) \\
&= -\lfloor (y_2-y_1)\sqrt{n} \rfloor (1-\rho_n) - \sum_{k=\lfloor \sqrt{n}y_1 \rfloor+1}^{\lfloor \sqrt{n}y_2 \rfloor} -\frac{k}{\sqrt{n}} \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \\
&\to -(y_2-y_1)\beta - \int_{y_1}^{y_2} -y dy.
\end{aligned}$$

□

Proof of Theorem 3. The technique used in this proof follows from Fleming et al. (1994). We denote G_n as the cumulative distribution function (CDF) of the scaled process Y_n. We first prove the relative compactness of G_n by a sandwich argument using stochastic comparison.

Let $\{Q_n^l(t)\}$ and $\{Q_n^u(t)\}$ denote the queue length processes of two sequences of ordinary Erlang-A queues: both have n servers and arrival rate λ_n, which are the same as the original process $Q_n(t)$. We keep the service rate and the abandonment rate fixed regardless of the system scale. The service rates of both systems are fixed at $\mu(0)$. The abandonment rate of $Q_n^l(t)$ is θ whereas the abandonment rate of $Q_n^u(t)$ is $\theta+\mu'(0)$.

As

$$\mu\left(\frac{q-n}{n}\right) n + \theta(q-n) \le \mu(0)n + \theta(q-n)$$

and

$$\begin{aligned}
\mu\left(\frac{q-n}{n}\right) n + \theta(q-n) &= \mu(0)n + (\mu'(0)+\theta)(q-n) + \mu''(\eta)\frac{(q-n)^2}{n} \text{ for some } \eta \in (0,(q-n)/n) \\
&\ge \mu(0)n + (\mu'(0)+\theta)(q-n),
\end{aligned}$$

based on Lemma 3, we have

$$P(Q_n(\infty) > q) \geq P(Q_n^l(\infty) > q)$$

and

$$P(Q_n(\infty) > q) \leq P(Q_n^u(\infty) > q).$$

Following the definition of Y_n, we let $Y_n^l := (Q_n^l(\infty) - n)/\sqrt{n}$ and $Y_n^u := (Q_n^u(\infty) - n)/\sqrt{n}$. We also denote G_n^l and G_n^u as the CDFs of the scaled processes X_n^l and X_n^u, respectively. Then both G_n^u and G_n^l converge uniformly to some limiting distributions (Fleming et al. 1994). We denote their limits as G^l and G^u respectively. Since $G_n^u(y) \leq G_n(y) \leq G_n^l(y)$, we have that for any $\epsilon > 0$, there exists a small enough y such that

$$\limsup_{n\to\infty} G_n(y) \leq \lim_{n\to\infty} G^l(y) < \epsilon$$

and

$$1 \geq \lim_{y\to\infty} \liminf_{n\to\infty} G_n(y) \geq \lim_{y\to\infty} \lim_{n\to\infty} G_n^u(y) = 1.$$

Thus, G_n is relatively compact. The limit of G_n exists and is a well-defined CDF.

From Lemma 4, the distribution G is absolutely continuous with probability density function of the form

$$g(y) = \begin{cases} \frac{C_1}{\sqrt{2\pi}} \exp\left(-\frac{(y+\beta)^2}{2}\right) & \text{if } y < 0, \\ \frac{C_2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y+\beta\sigma^2)^2}{2\sigma^2}\right) & \text{if } y \geq 0, \end{cases}$$

where $\sigma = \sqrt{\mu(0)/(\mu'(0)+\theta)}$, and C_1 and C_2 are the normalizing constants. Using the fact that $\int_{-\infty}^{\infty} g(y)\,dy = 1$ and $g(y)$ is continuous at 0, we have

$$C_1 = \frac{h(\beta\sigma)}{\sigma\phi(\beta)}\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1},$$

and

$$C_2 = \frac{h(\beta\sigma)}{\phi(\beta\sigma)}\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}.$$

□

Proof of Corollary 1. As $P_n(W) = P(Q_n(\infty) \geq n) = P(Y_n \geq 0)$, and

$$\lim_{n\to\infty} P(Y_n \geq 0) = C_2\bar{\Phi}(\beta\sigma) = \left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1},$$

where $\sigma = \sqrt{\mu(0)/(\mu'(0)+\theta)}$. We thus have the desired limit for $P_n(W)$.

For $P_n(Ab)$, we have

$$\begin{aligned} \sqrt{n}P_n(Ab) &= E[(Q_n(\infty) - n)^+]\frac{\theta\sqrt{n}}{\lambda_n} \\ &= E[Y_n | Y_n \geq 0]P_n(W)\frac{\theta n}{\lambda_n}. \end{aligned}$$

As $\lim_{n\to\infty} E[Y_n | Y_n \geq 0] = \sigma h(\beta\sigma) - \beta\sigma^2$,

$$\begin{aligned} \lim_{n\to\infty} \sqrt{n}P_n(Ab) &= \left(\sigma h(\beta\sigma) - \beta\sigma^2\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu(0)} \\ &= \left(\frac{h(\beta\sigma)}{\sigma} - \beta\right)\left(1 + \frac{h(\beta\sigma)}{\sigma h(-\beta)}\right)^{-1}\frac{\theta}{\mu'(0)+\theta}. \end{aligned}$$

□

Proof of Lemma 1. For the first limit in the lemma, let $\psi_n(x) = \lambda_n/n - \mu(x) - \theta x$ for $x \geq 0$. Then $\tilde{x}_n$ is the unique root of $\psi_n(x) = 0$ on $[0, \hat{x}]$. Since $\lambda_n/n \to \mu(0)$ as $n \to \infty$ and $\psi_n(\cdot)$ is continuous and monotonically increasing on $[0, \hat{x}]$, $\tilde{x}_n \to 0$ as $n \to \infty$. Applying Taylor expansion to $\mu(\cdot)$, we have

$$\psi_n(\tilde{x}) = \lambda_n/n - \mu(0) - (\mu' + \theta)\tilde{x}_n + O(\tilde{x}_n^2) = 0.$$

Then

$$\frac{\mu(0) - \lambda_n/n}{\tilde{x}_n} \to -\mu'(0) + \theta \text{ as } n \to \infty.$$

As $\sqrt{n}(1 - \lambda_n/(n\mu(0))) \to \beta$, $\tilde{x}_n = O(1/\sqrt{n})$. We then have

$$\sqrt{n}\tilde{x}_n = -\frac{\sqrt{n}(\mu(0) - \lambda_n/n)}{\mu'(0) + \theta} + O\left(\frac{1}{\sqrt{n}}\right).$$

Thus, $\sqrt{n}\tilde{x}_n \to -\mu(0)\beta/(\mu'(0) + \theta)$ as $n \to \infty$.

For the second limit in the lemma, since $\lambda_n/n - \nu(x) \to \mu(0) - \nu(x)$ as $n \to \infty$ and $\nu(\cdot)$ is continuously increasing on $(\hat{x}, \infty)$, $\bar{x}_n \to \bar{x}$. It is also easy to check that $\bar{x} = \bar{q}_2 - 1$, i.e. $\bar{x}$ measures the distance between the two fluid equilibria. □

Proof of Theorem 4. We first establish some asymptotic results about the value of the peaks, $\bar{q}_{n,1}$ and $\bar{q}_{n,2}$.

$$\frac{n - \bar{q}_{n,1}}{\sqrt{n}} = \sqrt{n}\left(1 - \frac{\lambda_n}{n\mu(0)}\right) + O\left(\frac{1}{\sqrt{n}}\right) \to \beta \text{ as } n \to \infty.$$

As $\lim_{n\to\infty} \lambda_n/n = \mu(0)$ and $\nu(x)$ is continuously decreasing on $(\hat{x}, \infty)$,

$$\frac{\bar{q}_{n,2} - n}{n} = \bar{x}_n + O(\frac{1}{n}) \to \bar{x} \text{ as } n \to \infty.$$

Using the detailed balance equation of the B&D process, we have

$$\begin{aligned}
\pi_n(\bar{q}_{n,2}) &= \pi_n(\bar{q}_{n,1}) \prod_{k=\bar{q}_{n,1}+1}^{\bar{q}_{n,2}} \frac{\lambda_n}{\mu((k-n)^+/n)(k \wedge n) + \theta(k-n)^+} \\
&= \pi_n(\bar{q}_{n,1}) \exp\left((\bar{q}_{n,2} - \bar{q}_{n,1}) \log \frac{\lambda_n}{n} - \sum_{k=\bar{q}_{n,1}+1}^{n-1} \log\left(\mu(0)\frac{k}{n}\right) - \sum_{k=n}^{\bar{q}_{n,2}} \log \nu\left(\frac{q-n}{n}\right)\right).
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{1}{n} \log \frac{\pi_n(\bar{q}_{n,2})}{\pi_n(\bar{q}_{n,1})} &= \frac{\bar{q}_{n,2} - \bar{q}_{n,1}}{n} \log \frac{\lambda_n}{n} - \frac{n - \bar{q}_{n,1}}{n} \log \mu(0) + O\left(\frac{1}{\sqrt{n}}\right) - \frac{n}{n} \sum_{k=0}^{\bar{x}_n n} \log \nu(\frac{k}{n}) \frac{1}{n} \\
&\to \bar{x} \log \mu(0) - \int_0^{\bar{x}} \log \nu(x) dx.
\end{aligned}$$

As $\nu(x) < \mu(0)$ for $x \in (0, \bar{x})$, $\bar{x} \log \mu(0) - \int_0^{\bar{x}} \log \nu(x) dx > 0$.

□

Proof of Lemma 2. By Theorem 2, under high sensitivity conditions, $\bar{q}_2 > 1$ is the unique root of $f(q) = \mu(0) - \mu((q-1)^+) - \theta(q-1)^+$ on $(1, \infty)$. We establish the results of this lemma by comparing pairs of systems, (1) and (2); we denote the higher level equilibrium as $\bar{q}_2^{(1)}$ and $\bar{q}_2^{(2)}$ for the two systems, respectively. For each part of the lemma we differ the two systems by two values of a specific system parameter.

i) Keep all other system parameters equal and vary the service rate function $\mu(\cdot)$, such that $\mu_{(2)}(\cdot)$ is more sensitive than $\mu_{(1)}(\cdot)$. Then, we have

$$0=\mu_{(1)}(0)-\mu_{(1)}\left(\bar{q}_2^{(1)}-1\right)-\theta\left(\bar{q}_2^{(1)}-1\right)\leq\mu_{(2)}(0)-\mu_{(2)}\left(\bar{q}_2^{(1)}-1\right)-\theta\left(\bar{q}_2^{(1)}-1\right).$$

As $\mu_{(2)}(0)-\mu_{(2)}(q-1)-\theta(q-1)$ is nonnegative on $[1,\bar{q}_2^{(2)}]$ and strictly negative on $(\bar{q}_2^{(2)},\infty)$, $\bar{q}_2^{(2)}\geq\bar{q}_2^{(1)}$, which implies $\bar{x}^{(1)}\leq\bar{x}^{(2)}$. Then

$$I_{(1)}(\bar{x}^{(1)})=\int_0^{\bar{x}^{(1)}}\log\frac{\mu_{(1)}(0)}{\mu_{(1)}(x)+\theta x}\,dx\leq\int_0^{\bar{x}^{(2)}}\log\frac{\mu_{(2)}(0)}{\mu_{(2)}(x)+\theta x}\,dx=I_{(2)}(\bar{x}^{(2)})$$

ii) Keep all other system parameters equal and vary the abandonment rate θ, such that $\theta_{(1)}<\theta_{(2)}$. Then,

$$0=\mu(0)-\mu\left(\bar{q}_2^{(2)}-1\right)-\theta_{(2)}\left(\bar{q}_2^{(2)}-1\right)<\mu(0)-\mu\left(\bar{q}_2^{(2)}-1\right)-\theta_{(1)}\left(\bar{q}_2^{(2)}-1\right).$$

Following the same rationale as in part i), we have $\bar{q}_2^{(1)}>\bar{q}_2^{(2)}$, which implies $\bar{x}^{(1)}>\bar{x}^{(2)}$. Then

$$I_{(1)}(\bar{x}^{(1)})=\int_0^{\bar{x}^{(1)}}\log\frac{\mu(0)}{\mu(x)+\theta_{(1)}x}\,dx>\int_0^{\bar{x}^{(2)}}\log\frac{\mu(0)}{\mu(x)+\theta_{(2)}x}\,dx=I_{(2)}(\bar{x}^{(2)}).$$

□

Proof of Lemma 3. We prove the theorem by first introducing a coupling, under which the entire sample path of $Y^{(1)}$ and $Y^{(2)}$ are ordered, i.e.

$$P(Y^{(1)}(t)\leq Y^{(2)}(t)\text{ for all }t\geq 0)=1.$$

Fix $\tilde{Y}^{(1)}(0)=\tilde{Y}^{(2)}(0)=y_0$ for any $y_0\in\mathbb{Z}^+$. The coupling argument uses the thinning property of Poisson process and goes as follows. When $(\tilde{Y}^{(1)}(t),\tilde{Y}^{(2)}(t))=(y_1,y_2)$ We generate the next potential transition by an exponential random variable with rate $\gamma_1+\xi_1(y_1)\vee\xi_2(y_2)$. We then generate a uniform random variable independent of everything else. If $U\leq\gamma_1/(\gamma_1+\xi_1(y_1)\vee\xi_2(y_2))$, we treat it as an arrival to both $\tilde{Y}^{(1)}$ and $\tilde{Y}^{(2)}$; else if $U\leq(\gamma_1+\xi_1(y_1)\wedge\xi_1(y_2))/(\gamma_1+\xi_1(y_1)\vee\xi_2(y_2))$, we treat it as a departure for both processes; else we impose a departure on $\tilde{Y}^{(i)}$ with the larger departure rate only. As when $y_1=y_2$, we always have $\xi_1(y_1)\geq\xi_2(y_2)$, under this coupling $\tilde{Y}^{(1)}(t)\leq\tilde{Y}^{(2)}(t)$, for all $t\geq 0$, path by path. Let $P_{y_0}(\cdot):=P(\cdot|Y^{(1)}=y_0,Y^{(2)}=y_0)$. Then we have

$$P_{y_0}(Y^{(1)}(t)>y)=P_{y_0}(\tilde{Y}^{(1)}(t)>y,\tilde{Y}^{(1)}(t)<\tilde{Y}^{(2)}(t))\leq P_{y_0}(\tilde{Y}^{(2)}(t)>y)=P_{y_0}(Y^{(2)}(t)>y)$$

for any $t\geq 0$.

As $\lim_{t\to\infty}P_{y_0}(Y^{(i)}(t)>y)=P(Y^{(i)}(\infty)>y)$,$i=1,2$, for all $y_0\in\mathbb{Z}^+$, and $Y^{(1)}$ and $Y^{(2)}$ are defined on the same state space, $P(Y^{(1)}(\infty)>y)\leq P(Y^{(2)}(\infty)>y)$. □

Before we prove Theorem 5, we first prove the following lemma as a preparation.

LEMMA 5. *Under High Sensitivity and SRS with $\beta>0$,*

i) the larger the value of the SRS parameter β is, the larger the value of $\tilde{q}_n$;

ii) under Definition 4, the more load sensitive the service rate function is, the smaller the value of $\tilde{q}_n$;

iii) the larger the abandonment rate θ is, the smaller the value of $\tilde{q}_n$.

Proof of Lemma 5. The proof of Lemma 5 follows the same strategy as the proof of Lemma 2. Specifically, we compare pairs of systems, (1) and (2). For each part of the lemma, we differ the two system by two values of a specific system parameter.

i) Keep all other system parameters equal and vary the staffing parameter β, such that $\beta_{(1)} < \beta_{(2)}$. Denote $n_{(1)} = R_n + \beta_{(1)}\sqrt{R_n}$ and $n_{(2)} = R_n + \beta_{(2)}\sqrt{R_n}$. Then $n_{(1)} < n_{(2)}$ and

$$0 = \lambda_n/n_{(1)} - \mu(\tilde{x}_n^{(1)}) - \theta\tilde{x}_n^{(1)} > \lambda_n/n_{(2)} - \mu(\tilde{x}_n^{(1)}) - \theta\tilde{x}_n^{(1)}.$$

As $\lambda_n/n_{(2)} - \mu(x) - \theta x$ is increasing on $[0, \hat{x}]$ and is nonpositive on $[0, \tilde{x}_n^{(2)}]$, $\tilde{x}_n^{(2)} > \tilde{x}_n^{(1)}$. Thus, $\tilde{q}_n^{(2)} = \lfloor(\tilde{x}_n^{(2)} + 1)n_{(2)}\rfloor > \tilde{q}_n^{(1)} = \lfloor(\tilde{x}_n^{(1)} + 1)n_{(1)}\rfloor$

ii) Keep all other system parameters equal and vary the service rate function $\mu(\cdot)$, such that $\mu_{(2)}(\cdot)$ is more sensitive than $\mu_{(1)}(\cdot)$. Then, we have

$$0 = \frac{\lambda_n}{n} - \mu_{(2)}\left(\tilde{x}_n^{(2)}\right) - \theta\tilde{x}_n^{(2)} \geq \frac{\lambda_n}{n} - \mu_{(1)}\left(\tilde{x}_n^{(2)}\right) - \theta\tilde{x}_n^{(2)}.$$

Following the same rationale as in part i), we have $\tilde{q}_n^{(1)} \geq \tilde{q}_n^{(2)}$.

ii) Keep all other system parameters equal and vary the abandonment rate θ, such that $\theta_{(1)} < \theta_{(2)}$. Then,

$$0 = \frac{\lambda_n}{n} - \mu\left(\tilde{x}_n^{(1)}\right) - \theta_{(1)}\tilde{x}_n^{(1)} > \frac{\lambda_n}{n} - \mu\left(\tilde{x}_n^{(1)}\right) - \theta_{(2)}\tilde{x}_n^{(1)}.$$

Following the same rationale as in part i), we have $\tilde{q}_n^{(2)} > \tilde{q}_n^{(1)}$.

□

Proof of Theorem 5. We prove Theorem 5 by comparing the death rates of pairs of systems denoted by $Q^{(1)}$ and $Q^{(2)}$.

i) Keeping all other parameters equal, for $\beta_{(1)} < \beta_{(2)}$, we denote $n_{(1)} = R + \beta_{(1)}\sqrt{R}$, $n_{(2)} = R + \beta_{(2)}\sqrt{R}$ where $R = \lambda/\mu(0)$. Then when $q \leq n_{(1)}$, the death rates of the two systems are equal; when $n_{(1)} < q \leq n_{(2)}$,

$$\mu(0)q - \left(\mu\left(\frac{q}{n_{(1)}} - 1\right)n_{(1)} + \theta(q - n_{(1)})\right) \geq (\mu(0) - \theta)(q - n_{(1)}) \geq 0;$$

when $q > n_{(2)}$

$$\begin{aligned}
&\left(\mu\left(\frac{q}{n_{(2)}} - 1\right)n_{(2)} + \theta(q - n_{(2)})\right) - \left(\mu\left(\frac{q}{n_{(1)}} - 1\right)n_{(1)} + \theta(q - n_{(1)})\right) \\
&= \left(\mu\left(\frac{q}{n_{(2)}} - 1\right) - \mu\left(\frac{q}{n_{(1)}} - 1\right)\right)n_{(2)} + \mu\left(\frac{q}{n_{(1)}} - 1\right)(n_{(2)} - n_{(1)}) - \theta(n_{(2)} - n_{(1)}) \\
&\geq -\mu'\left(\frac{q}{n_{(1)}} - 1\right)\frac{(n_{(2)} - n_{(1)})q}{n_{(1)}} + (\mu(\infty) - \theta)(n_{(2)} - n_{(1)}) \geq 0.
\end{aligned}$$

Then

$$P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \geq P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \geq P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)}),$$

where the first inequality follows from Lemma 3 and the second inequality follows from Lemma 5.

ii) Keeping all other parameters equal, for system (2) more sensitive than system (1), we have $\mu_{(1)}((q - n)^+/n)(q \wedge n) + \theta(q - n)^+ \geq \mu_{(2)}((q - n)^+/n)(q \wedge n) + \theta(q - n)^+$ for all $q \geq 0$. Then $P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \leq P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \leq P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)})$.

iii) Keeping all other parameters equal, for $\theta_{(1)} < \theta_{(2)}$, we have $\mu((q-n)^+/n)(q\wedge n) + \theta_{(1)}(q-n)^+ \le \mu((q-n)^+/n)(q\wedge n) + \theta_{(2)}(q-n)^+$ for all $q \ge 0$. Then $P(Q^{(1)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(1)}) \ge P(Q^{(2)}(\infty) > \tilde{q}_n^{(2)})$.

□

Proof of Theorem 6. The proof of Theorem 6 also follows from the method outlined in Pang et al. (2007). We use both the Functional Central Limit Theorem (FCLT) and CMT. We again write

$$\begin{aligned} Q_n^c(t) &= Q_n^c(0) + A(\lambda_n t) - S\left(\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du\right) - R\left(\theta\int_0^t (Q_n^c(u)-n)^+du\right) - L_n(t) \\ &= Q_n^c(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) - L_n(t) \\ &\quad + \lambda_n t - \int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du - \theta\int_0^t (Q_n^c(u)-n)^+du \end{aligned}$$

where

$$\begin{aligned} M_{n,1}(t) &= A(\lambda_n t) - \lambda_n t \\ M_{n,2}(t) &= S\left(\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du\right) - \int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du \\ M_{n,3}(t) &= R\left(\theta\int_0^t (Q_n^c(u)-n)^+du\right) - \theta\int_0^t (Q_n^c(s)-n)^+du. \end{aligned}$$

Let $\hat{Q}_n^c(t) = (Q_n(t)-n)/\sqrt{n}$, $\hat{Y}_n(t) = Y_n(t)/\sqrt{n}$ and $\hat{M}_{n,i} = M_{n,i}/\sqrt{n}$ for $i = 1,2,3$. As $\hat{Q}_n^c(\cdot) < c_n$, $\hat{Q}_n^c(t) = O(\sqrt{n})$. Applying Taylor expansion, we have

$$\begin{aligned} \hat{Q}_n^c(t) &= \hat{Q}_n^c(0) + \hat{M}_{n,1}(t) - \hat{M}_{n,2}(t) - \hat{M}_{n,3}(t) - L_n(t) \\ &\quad + \frac{\lambda_n - \mu(0)n}{\sqrt{n}}t - \int_0^t \mu(0)(\hat{Q}_n^c(u)\wedge 0)du - \int_0^t \mu'(0)\hat{Q}_n^c(u)^+du - \int_0^t \theta\hat{Q}_n^c(u)^+du + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Let $d(q) = -\mu'(0)(q\wedge 0) - (\mu'(0)+\theta)q^+$. Consider the integral representation

$$q(t) = b + x(t) + \int_0^t d(q(s))ds - l(t), \tag{6}$$

where $l(t)$ is a nondecreasing nonnegative function in $\mathcal{D}$ such that (6) holds and $\int_0^\infty 1\{q(t) < c\}dl(t) = 0$. As $d(\cdot)$ is Lipschitz, the integration (6) has a unique solution (q,y) and it constitutes a Bonafide function $(\phi_1,\phi_2) : \mathcal{D}\times R \to \mathcal{D}\times\mathcal{D}$ mapping (b,x) into (q,y). Moreover (ϕ_1,ϕ_2) is continuous (see Theorem 7.3 in Pang et al. (2007)).

We notice that $\hat{M}_{n,i}$'s are square-integrable martingales with respect to the filtration

$$\mathcal{F}_{n,t} := \sigma\left\{Q_n(0), A(\lambda_n s), S\left(\int_0^s \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)(Q_n^c(u)\wedge n)du\right), R\left(\theta\int_0^t (Q_n^c(u)-n)^+du\right) : 0\le s\le t\right\}$$

augmented by including all null sets. Also

$$\begin{aligned} \langle M_{n,1}\rangle(t) &= \frac{\lambda_n t}{n} \\ \langle M_{n,2}\rangle(t) &= \int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)\frac{Q_n^c(u)\wedge n}{n}du \\ \langle M_{n,3}\rangle(t) &= \frac{\theta}{n}\int_0^t (Q_n^c(u)-n)^+du. \end{aligned}$$

As

$$\frac{\lambda_n t}{n} \to \mu(0)t \text{ as } n \to \infty \text{ w.p. } 1,$$

$\{\langle M_{n,1}\rangle\}$ is stochastically bounded. By the crude bound $Q_n^c(s) < Q_n^c(0) + A(\lambda_n t)$, we have

$$\int_0^t \mu\left(\frac{(Q_n^c(u)-n)^+}{n}\right)\frac{Q_n^c(u)\wedge n}{n}du \le \mu(0)t\left(\frac{Q_n^c(0)}{n} + \frac{A(\lambda_n t)}{n}\right).$$

Since $\{Q_n^c(0)/n\}$ and $\{A(\lambda_n t)/n\}$ are stochastically bounded, $\{\langle M_{n,2}\rangle\}$ is stochastically bounded.

Similarly, we can show that $\{\langle M_{n,3}\rangle\}$ is also stochastically bounded. This implies that $\{M_{n,i}\}$'s for $i = 1, 2, 3$ are stochastically bounded, which in turn implies the stochastic boundedness of $\{\hat{Q}_n^c\}$ in $\mathcal{D}$. Thus,

$$\hat{Q}_n^c/\sqrt{n} \Rightarrow \eta \text{ in } \mathcal{D} \text{ as } n \to \infty$$

where η is the zero function defined above.

By FCLT for Poisson processes and CMT with composition map, we have

$$(M_{n,1}, M_{n,2}, M_{n,3}) \Rightarrow (B_1 \circ \lambda\omega, B_2 \circ s\mu(0)\omega, B_3 \circ \eta)$$

where $\omega(t) \equiv 1$ for any t.

Finally, applying the CMT with the integral representation (6), we get the result in Theorem 6. □