Slowdown Services: Staffing Service Systems with Load-Dependent Service Rate

Jing Dong
Department of Industrial Engineering and Operations Research, Columbia University, NY, USA jd2736@columbia.edu

Pnina Feldman
Haas School of Business, University of California, Berkeley, CA, USA, feldman@haas.berkeley.edu

Galit Yom-Tov
Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel 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 service performance. We modify the Erlang-A model to account for service slowdowns and perform an asymptotic 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 regime performance levels. To stabilize and improve system performance in this case, we propose two solutions: to add staffing on a permanent basis or to implement a threshold admission control policy where customers are occasionally blocked or rerouted to other service systems. The former is a static solution that provides a Quality Driven regime performance, while the latter is a real-time dynamic solution that leads to QED performance.

Key words: Service systems; Halfin-Whitt regime; State dependent queues; Limit theorem: fluid and diffusion analysis; Equilibrium analysis: bi-stability

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 suggests that in many service systems the two are correlated (see for example Batt and Terwiesch (2012), Gerla and Kleinrock (1980) and KC and Terwiesch (2009)). Depending on the service environment, heavily loaded systems may experience service speedups or slowdowns.

In this paper 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 routing. While both speedup and slowdown effects are present in service systems, ignoring other quality issues, service speedups do not worsen performance $^1$. Therefore, in this paper, we focus on systems which exhibit slowdowns and show that slowdown effects may not only decrease the quality of customer service, but also increase agents’ workload dramatically.
Several factors may cause service slowdowns when the system is congested. High congestion levels may induce pressure on agents, which according to the psychology literature may impact human perception, information processing and decision making (see for example Bertrand and van Ooijen (2002)). 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, thus causing fatigue. Empirical studies have provided evidence that fatigue leads to deterioration of productivity (see for example KC and Terwiesch (2009), Caldwell (2001)). 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)). Customers may also demand a longer and more personalized service following a long wait.

So how should a firm design service systems that exhibit slowdowns? Generally, there are two objectives that play opposing roles in the design of service systems. On one hand, to increase efficiency and reduce operational costs, system designers aim to increase resource utilization. On the other hand, high utilization leads to increased level of delay and abandonment, thereby reducing quality of service. A common approach to design a service system is to balance 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 a many-server asymptotic analysis suggests a Square-Root Staffing (SRS) rule to balance this tradeoff. According to the SRS rule the number of servers, s, is set such that 
\[ s = R + \beta \sqrt{R}, \]
where \( R = \lambda / \mu \) is the offered load of the system, and \( \beta \), the SRS parameter, is a multiplier set according to the queueing model under consideration. For the SRS rule in a multi-server queueing model with abandonments (commonly referred to as the Erlang-A model), \( \beta \) 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 Efficiency-Driven (ED) regime and 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 \( s = R - \alpha R \) for \( 0 < \alpha < 1 \), where \( \alpha \) is typically selected in the range 0.1 – 0.25 (Whitt 2004). This results in 100% occupancy, probability of waiting close to 1 and 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. In this case, the staffing level is set to \( s = R + \alpha R \)
for $\alpha > 0$, where the typical range of $\alpha$ 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 Erlang-A model when staffing according the SRS rule. We use the term load sensitivity to describe the level of service rate deterioration as a response to increased workload. We show that staffing to operate in the QED regime may not be a good enough solution in some systems with load-sensitive service times. Depending on model parameters, we observe that systems designed to operate in the QED regime may have unstable performance, alternating between being overloaded and underloaded, 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 staffing rule, or even a Quality Driven staffing rule, may result in an undesirable performance, typically found when using Efficiency Driven staffing rules. We therefore propose alternative staffing rules and admission control policies that can be applied in the presence of service slowdowns.

We make the following key contributions:

- 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.

- When the load sensitivity is relatively low (i.e., service time does not decrease significantly with the load placed on the system), the SRS rule is an appropriate staffing rule, but the parameter $\beta$ is determined by modifying the Garnett functions to account for the service rate decrease. We develop new approximation functions and show that QED performance is achievable (§§6.1). To derive these approximations it is not necessary to estimate the entire service rate function - it is sufficient to accurately estimate the service rate function around zero.

- When the load sensitivity is high, the system alternates between two equilibria (bi-stability): one provides QED performance and the other provides ED performance (§4). Consequently, applying the SRS rule in this case does not consistently result in QED performance levels. We investigate how load sensitivity and consumer patience influence the occurrence of bi-stability, and the frequency of transitions between the two equilibria (§5). We show that sensitivity and patience have opposing effects on system performance. While a higher load sensitivity increases the occurrence of ED performance, a higher abandonment rate decreases such occurrences. This suggests that when systems are load sensitive, encouraging customers to patiently wait for service may be operationally undesirable.
• To overcome the bi-stability phenomenon, we propose two operational solutions: increasing staffing and admission control (§§6.2). In order to avoid bi-stability one option is to increase staffing by a factor of order \( n \) (between 3 – 20%), which may result in QD regime performance. Since a permanent increase of staffing is inefficient, we propose an alternative solution where instead of increasing the supply, we limit incoming demand according to an admission control rule. We identify a threshold level on the number of consumers in the system, that depends on the offered-load, the load-sensitivity, and customer patience values. We block or reroute customers from the system when the number of customers in the system reaches that threshold. We develop diffusion approximations for systems that operate under such an admission control rule, and show that the proportion of blocked customers is \( O(1/\sqrt{n}) \).

• Our results therefore suggest that if the sensitivity level is below a threshold, an increased staffing solution works well and if the sensitivity level is above the threshold, increased staffing may be wasteful and an admission control policy is more preferable.

2. Literature Review

Palm (1957) introduced the Erlang-A \((M/M/s + M)\) model to incorporate abandonment in the traditional many server Erlang-C \((M/M/s)\) queue. Mandelbaum and Zeltyn (2007) showed that abandonment is a significant factor that needs to be considered when modeling service systems and making staffing decisions. Garnett et al. (2002) conducted an asymptotic analysis of the Erlang-A model. Based on the asymptotic behavior, they derived approximations for the probabilities of waiting and abandonment and provided guidance for the design of large service systems. These approximations are the basis of our analysis. In this paper, we study a modified Erlang-A model that accounts for a load-dependent service rate. A new feature in our model is the bi-stability phenomenon. A similar phenomenon is studied in Chan et al. (2012) in the context of ICU flow and in Gibbens et al. (1990) in the context of communication networks.

A few papers consider the load-dependent service rate but their analysis is in the single server queue setting without abandonment. Whitt (1990) and Boxma and Vlasiou (2007) use a generalized version of the Lindley recursion to 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. They developed stability conditions for the systems. 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.

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 the long-run average throughput of the system and they show that under certain
conditions a threshold policy is optimal. Likewise, in Subsection 6.2.2, we also consider a threshold admission control policy, but our objective is to maintain a certain performance level. Since our system has multiple servers, we also consider a policy that adjusts the staffing level of the system.

This paper is also related to the literature on service control, where the service rate or staffing level can be adjusted based on the workload of the system. Weber and JR. (1987) study an optimal control of the service rate in a network of single server queues to minimize the costs associated with waiting and service speed. Chan et al. (2012) studied the case of intensive care units where speedup can be implemented by an early discharge of current patients. Hence, both papers assume that management can control or adjust the service rate. In contrast, in this paper we assume that the service rate is exogenous, and cannot be changed by system managers. However, we do consider adjusting the staffing level as a possible solution to stabilize system performance, a problem that arises with time-varying demand. To cope with time-varying arrivals, a few papers considered dynamic staffing (e.g., Green et al. (2007), Yom-Tov and Mandelbaum (2011)). 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 can not propose a predetermined policy whereby the staffing levels change in a predictable fashion. Instead, we propose policies that mitigate the effects of unpredictable system behavior.

3. Model Setup

3.1. Load dependent Erlang-A model

We analyze a modification of the Erlang-A ($M/M/s+M$) model which incorporates the dependence of service rate on workload through queue length. Specifically, we consider an $M/M_Q/s+M$ queue, with $s$ identical servers. Each server can serve only one customer at a time. Customers arrive to the system according to a Poisson process with rate $\lambda$. 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 assume customers have finite patience. The patience time of each arriving customer is exponentially distributed with rate $\theta$, 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 load in the system. We measure the load placed on the system by the proportion of people in the queue, i.e., the service
rate is $\mu((Q - s)^+/s)$. This scaling preserves the load sensitivity of service rate irrespective of the size of the system. This is essential when considering scaling for approximations.

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 $\mu'(0) := \lim_{x \to 0^+} \mu'(x)$ and $\mu''(0) := \lim_{x \to 0^+} \mu''(x)$. We further assume that the service rate function exhibits diminishing decreasing rate until it reaches a minimum positive level. 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.

### 3.2. The QED heavy-traffic regime

To balance the quality of service with system efficiency, we aim to operate the queue in the QED regime. For an $M/M/s + M$ queue, the QED regime is obtained by holding the service rate and abandonment rate fixed while letting the aggregate arrival rate $\lambda$ and number of servers $s$ grow to infinity in such a way that the utilization rate $\rho = \lambda/(s\mu)$ approaches 1 with a certain rate. Specifically, consider a sequence of systems indexed by $n$. Let $\lambda_n$ denote the arrival rate of the $n$-th system, $s_n$ denote the number of servers in the $n$-th system, and $\rho_n := \lambda_n/(s_n\mu)$. It is assumed that

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

for some $\beta \in \mathbb{R}$, or equivalently, that the number of servers is set by the square root formula – $s_n = R_n + \beta \sqrt{R_n}$, 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 scaled queue length process converges to a diffusion limit. Based on that limit they develop the approximations for the performance measures, which is exact in the asymptotic sense. Let $h(\cdot)$ denote the hazard rate function of a standard normal random variable, then

$$P(W) := \lim_{n \to \infty} P_n(W) = \left(1 + \frac{h\left(\beta \sqrt{\mu/\theta}\right)}{\sqrt{\mu/\theta}h(-\beta)}\right)^{-1},$$

$$\lim_{n \to \infty} \sqrt{n}P_n(\text{Ab}) = \lim_{n \to \infty} \sqrt{n} \left(1 - \frac{h\left(\beta \sqrt{\mu/\theta}\right)}{h\left(\beta \sqrt{\mu/\theta} + \sqrt{\theta/(s_n\mu)}\right)}\right) P_n(W)$$

$$= (\sqrt{\theta/\mu}h\left(\beta \sqrt{\mu/\theta} - \beta\right) P(W).$$
The probability of waiting, $P(W)$, is non-degenerate and the probability of abandonment, $P(\text{Ab})$, converges to zero at rate $1/\sqrt{n}$. Thus, systems that operate in this regime achieve both good performance and high efficiency. However, as Figures 1 & 2 illustrate, we observe that in the modified Erlang-A model, using the square root formula does not guarantee similar performance.

**Figure 1**  Sample paths of the number of people in the system for $M/MQ/s + M$ queues with different load sensitivity parameter values, $b$ ($s = 500$, $\lambda = 480$, $\mu = 0.6 + 0.4 \exp(-b(q-s)^+ / s)$ and $\theta = 0.5$)

(a) $b = 1$  
(b) $b = 2$

**Figure 2**  Histogram of the number of people in the system for $M/MQ/s + M$ queues with different load sensitivity parameter values, $b$ ($s = 500$, $\lambda = 490$, $\mu = 0.6 + 0.4 \exp(-b(q-s)^+ / s)$ and $\theta = 0.5$)

(a) $b = 1$  
(b) $b = 2$

In the absence of workload sensitivity, (i.e., $b = 0$), the system with the same parameters as in Figure 1 operates in the QED regime, with $\beta = 0.45$, $P(W) = 0.3831$ and $P(\text{Ab}) = 0.0073$. Figure 1 illustrates that this is not necessarily the case when systems exhibit load sensitivity. In the first case ($b = 1$), the number of people in the system is stable, and the system operates in the QED regime with low probability of waiting $P(W) = 0.4708$ and abandonment $P(\text{Ab}) = 0.0142$. Nevertheless, the performance is worse than the one obtained without sensitivity. In the second case ($b = 2$), we observe a phenomenon of bi-stability. The trajectory moves between two levels: a low level where the number of people in the system is small and the performance is good ($P(W) \approx 0.45$ and $P(\text{Ab}) \approx 0.015$), and a high level where the number of people in the system is large and the service
level is poor \((P(W) = 1 \text{ and } P(\text{Ab}) \approx 0.25)\). The average performance yields \(P(W) = 0.9152\) and \(P(\text{Ab}) = 0.1602\).

Figure 3 illustrates how the probability of waiting and the probability of abandonment change as functions of the load sensitivity parameter, \(b\). We observe that the effect of load sensitivity is non linear and that 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 3).

**Figure 3** Performance measures for \(M/M_Q/s + M\) queues as a function of the load sensitivity parameter, \(b\) \((s = 500, \lambda = 490, \mu = 0.6 + 0.4\exp(-b(q - s)^+/s)\) and \(\theta = 0.5\))

These demonstrations imply that it is not straightforward to achieve QED performance in slowdown service systems. We use the many-server asymptotic analysis approach to analyze the dynamics of such systems when using the SQS rule.

### 4. Fluid analysis

In this section, we establish the fluid limit of the queue length process of the load-sensitive 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.

#### 4.1. Fluid approximation

To develop the fluid limit, we consider a sequence of systems indexed by \(n\), where both the arrival rate and the number of servers grow proportionally to \(n\). For the \(n\)-th system, we denote \(Q_n \equiv \{Q_n(t) : t \geq 0\}\) as the queue length process (number of people in the system). We denote the arrival rate as \(\lambda_n\) and the number of servers as \(s_n\). 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. As we are interested in the QED asymptotic regime, we assume that there exists a \(\beta\) such that \(\lim_{n \to \infty} \sqrt{n}(1 - \lambda_n/(s_n\mu(0))) = \beta\).
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:

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

where \((x)^+ = \max(0, x)\) and \((x \wedge y) = \min(x, y)\). We define the fluid-scaled process

\[
\bar{Q}_n(t) = \frac{Q_n(t)}{n}
\]

**Theorem 1.** Assume \( \lambda_n/n \to \lambda \), \( s_n/n \to s \) as \( n \to \infty \). 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

\[
\bar{Q}(t) = \bar{Q}(0) + \lambda t - \int_0^t \mu \left( \frac{(\bar{Q}(u) - s)^+}{s} \right) (\bar{Q}(u) \wedge s) \, du - \int_0^t \theta (\bar{Q}(u) - s)^+ \, du.
\]

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 \( f(q) = \lambda - \mu((q - s)^+/(q \wedge s)) - \theta(q - s)^+ \). Then we can write \( \bar{Q}(t) \) as the solution to the following autonomous differential equation with initial value \( \bar{Q}(0) \):

\[
\dot{\bar{Q}} = f(\bar{Q})
\]

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

The load sensitive Erlang-A model is positive recurrent, i.e., it does not grow to infinity almost surely, because the abandonment rate is positive.

### 4.2. Equilibrium analysis

Next, we analyze the long term 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 \( q_0 \).

**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 a 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 \( q_0 \neq \bar{q} \). To study how the system behaves in the long-run, we analyze the stability of the equilibrium points.
**Definition 2 (Stability of equilibrium).** Let $\bar{q}$ be an equilibrium point 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 any $t \geq 0$. Otherwise, $\bar{q}$ is **unstable**. If $\delta$ can be chosen such that not only $\bar{q}$ is 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 starting close enough to $\bar{q}$, trajectories defined by (2) converge to $\bar{q}$ as $t \to \infty$. An equilibrium may also be **semistable**. In 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.

To characterize the equilibria of the fluid model in (2), we analyze the function $f(q)$. When $q \leq s$, $f'(q) = -\mu(0) < 0$, and $f(q)$ is a linearly decreasing function that starts at $f(0) = \lambda > 0$ and ends at $f(s) = \lambda - \mu(0)s = 0$. When $q \geq s$, $f'(q) = -\mu'((q - s)/s) + \theta$ and $f''(q) = -1/s\mu''((q - s)/s) \leq 0$ as $\mu''(x) \geq 0$ for $x \geq 0$. Therefore, $f(q)$ is concave on $[s, \infty)$. Let $\hat{q} = \arg \max_{q \in [s, \infty)} f(q)$. We refer to $\hat{q}$ as the critical point of the system. Depending on the actual form of $f(q)$, we distinguish between the following two cases (as shown in Figure 4):

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

**Figure 4** Flow rate function under two cases

(a) Case I – Low sensitivity  
(b) Case II – High sensitivity

**Theorem 2.**  
(i) If $-\mu'(0) \leq 0$ (Low Sensitivity), there is a unique equilibrium, $\bar{q}$, with $\bar{q} = s$. Furthermore, $\bar{q}$ is asymptotically stable.

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

The equilibrium value is the solution to $f(q) = 0$. In the low sensitivity case, the equilibrium is unique and is given by $\bar{q} = s$. Therefore, in this case the fluid model will converge to that value. In the high sensitivity case, there are two solutions to $f(q) = 0$, $\bar{q}_1$ and $\bar{q}_2$, and the fluid model may
converge to either one, depending on the starting point. In the stochastic level, the trajectory of the queue length process fluctuates around \( \bar{q} = s \) under Low Sensitivity. Under High Sensitivity, the queue length process may alternate between the two equilibria. This drives the bi-stability phenomenon observed in Figure 1b.

Figure 5 demonstrates that the bi-stability phenomenon in the high sensitivity case is not always clearly observable in a simulation run of the stochastic system. The frequency with which the trajectory alternates between the two equilibria and the proportion of time it spends around each depends on the system parameters.

We observe in Figure 5 that when the load sensitivity is relatively low, e.g., \( b_1 = 1.25 \), the queue length process fluctuates around the lower equilibrium. Even if the process starts at a very high congestion level, it will reach the lower equilibrium level fairly quickly. For moderate load sensitivity values, e.g., \( b_1 = 1.5 \), the queue length process alternates between the two equilibria, fluctuating around each one for some time. When the load sensitivity level is relatively high, e.g., \( b_1 = 1.75 \), the queue length process fluctuates around the higher equilibrium. This unpredictability in behavior will be analytically explored in Section 5.

5. Bi-stability analysis

In this section, we analyze factors that affect the proportion of time the system spends around each equilibrium in the long run as well as the corresponding performance under High Sensitivity. As we are interested in the fluctuation of the system in the stochastic level, we analyze the model in the original scale. Specifically, we model the process as a Birth and Death (B&D) process and use stochastic comparison as in Muller and Stoyan (2002) to compare the performance of different systems. Asymptotic analysis still plays a role in this section, so we keep referring to the sequence of systems indexed by \( n \).
As before, $Q_n$ is a B&D process with birth rate $\lambda_n$ and state-dependent death rate $\mu((q-s_n)^+/s_n)(q \wedge s_n) + \theta(q-s_n)^+$ for all state $q \in R^+$. Motivated by the fluid analysis, we look at the following flow rate function:

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

When $f_n(q) > 0$ the birth rate surpasses the death rate and the queue is more likely to build up. When $f_n(q) < 0$, the death rate surpasses the birth rate and the queue is more likely to decrease. Thus, the $q$ that solves $f_n(q) = 0$ divides the B&D process into regions in which $f_n(q)$ is either positive or negative.

We observe that when applying SRS with $\beta < 0$, $s_n < \lambda_n/\mu(0)$. Hence, $f_n(q) = 0$ has only one root $\bar{q}_n$ (see Figure 6a). In this cases, for all $q < \bar{q}_n$, $f_n(q) > 0$ and for all $q > \bar{q}_n$, $f_n(q) < 0$. Thus, the queue length process $Q_n(\cdot)$ will fluctuate around $\bar{q}_n$ and the bi-stability phenomenon will not arise. On the other hand, for large enough systems, when using SRS with $\beta > 0$, $s_n > \lambda_n/\mu(0)$. In this case $f_n(q) = 0$ has three roots (see Figure 6b), and bi-stability occurs.

![Figure 6](image_url)  

The flow rate function $f_n(q)$ with positive or negative $\beta$s

(a) $\beta < 0$  
(b) $\beta > 0$

The following theorem characterizes the value of $\bar{q}_n$ under SRS with $\beta < 0$.

**Theorem 3.** Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta < 0$, $f_n(q) = 0$ has only one root, $\bar{q}_n$, and $\bar{q}_n > \bar{q}_n$ where $\bar{q}_n$ is the root of $f_n'(0) = 0$. Assume $\lambda_n/n \to \lambda$ and $s_n/n \to s$ as $n \to \infty$. Let $\psi(x) := \lambda - \mu(x/s)s - \theta x$ and $\hat{x}$ be the root of $\psi'(x) = 0$. Then there exists $\eta > \hat{x}$ such that $\psi(\eta) = 0$ and

$$\lim_{n \to \infty} \frac{s_n}{\bar{q}_n - s_n} = \eta.$$  

Theorem 3 indicates that the root $\bar{q}_n$ is of order $n$ beyond $s_n$. Hence, the number of customers in the system is very high and the resulting performance is ED (the occupancy and probability of waiting are close to 1 and the probability of abandonment is high). We therefore recommend to use only positive values of $\beta$ when applying the SRS formula, if the service rate function is highly sensitive to load. We assume $\beta > 0$ in subsequent analysis of the high sensitivity case.
Lemma 1. Under High Sensitivity ($-\mu'(0) > \theta$) and SRS with $\beta > 0$, let $\bar{q}_n$ denote the root of $f_n(q)$ for $q > s_n$. $\bar{q}_n$ is the maximizer of $f_n(q)$ on $[s_n, \infty)$. Then

$$\lim_{n \to \infty} f_n(\bar{q}_n) > 0.$$ 

Lemma 1 implies that for large enough systems, we can define $0 < \bar{q}_{n,1} < \bar{q}_{n,2} < \bar{q}_{n,3}$ as the three roots of $f_n(q_n,i) = 0$ for $i = 1, 2, 3$, and

$$f_n(q) = \begin{cases} > 0 & \text{when } q \in [0, \bar{q}_{n,1}); \\ < 0 & \text{when } q \in (\bar{q}_{n,1}, \bar{q}_{n,2}); \\ > 0 & \text{when } q \in (\bar{q}_{n,2}, \bar{q}_{n,3}); \\ < 0 & \text{when } q \in (\bar{q}_{n,3}, \infty). \end{cases}$$

When $q < \bar{q}_{n,1}$, $f_n(q) > 0$ and the queue length process tends to increase. When $\bar{q}_{n,1} < q < \bar{q}_{n,2}$, $f_n(q) < 0$ and the queue length process tends to decrease. Thus we expect $Q_n$ to fluctuate around $\bar{q}_{n,1}$ in stationarity. Likewise, when $\bar{q}_{n,2} < q < \bar{q}_{n,3}$, $f_n(q) > 0$ and the queue length process tends to increase. When $q > \bar{q}_{n,3}$, $f_n(q) < 0$ and the queue length process tends to decrease. Therefore, we expect $Q_n$ to fluctuate around $\bar{q}_{n,3}$ in stationarity as well. The bi-stability arises by alternating between $\bar{q}_{n,1}$ and $\bar{q}_{n,3}$.

The following theorem characterizes the value of $\bar{q}_{n,i}$'s, $i = 1, 2, 3$, under SRS with $\beta > 0$.

Theorem 4. Under High Sensitivity ($-\mu'(0) < \theta$) and SRS with $\beta > 0$, if $\lambda_n/n \to \lambda$, $s_n/n \to s$ as $n \to \infty$, then we have

$$\lim_{n \to \infty} \frac{\bar{q}_{n,1} - s_n}{\sqrt{n}} = -\beta s;$$
$$\lim_{n \to \infty} \frac{\bar{q}_{n,2} - s_n}{\sqrt{n}} = -\frac{\beta \mu(0) s}{\mu'(0) + \theta};$$

For $\psi(x) = \lambda - \mu(x/s)s - \theta x$ and $\hat{x}$ being the root of $\psi'(x) = \mu'(x/s) + \theta$, there exists $\eta > \hat{x}$ such that $\psi(\eta) = 0$ and

$$\lim_{n \to \infty} \frac{\bar{q}_{n,3} - s_n}{n} = \eta.$$
Figure 7  Roots of the flow function \( f_n(q) \) for different values of \( b \) (\( s = 500, \lambda = 475, \mu = 0.6 + 0.4 \exp(-b(q-s)/s), \theta = 0.3 \)) \((\ast : \tilde{q}_2, + : \tilde{q}_3)\)

Figure 8  The flow rate function \( f_n(q) \) with \( b = 1.25 \) (\( s = 500, \lambda = 475, \mu = 0.6 + 0.4 \exp(-b(q-s)/s), \theta = 0.3 \))

**Definition 3.** For two service rate functions \( \mu_1(\cdot) \) and \( \mu_2(\cdot) \) with \( \mu_1(0) = \mu_2(0) \), we say that \( \mu_1 \) is more load-sensitive than \( \mu_2 \) if \( \mu_1(x) \leq \mu_2(x) \) for \( x \geq 0 \).

As we are comparing pairs of systems of the same scale, we suppress the dependence on the scale parameter \( n \).

**Lemma 2.** As the sensitivity of the service rate function increases, the corresponding value of \( \tilde{q}_2 \) decreases and the value of \( \tilde{q}_3 \) increases. The value of \( \tilde{q}_1 \) does not depend on the sensitivity of the service rate function.

Figure 7 illustrates how \( \tilde{q}_2 \) and \( \tilde{q}_3 \) change with \( b \) using a numerical example. Specifically, \( \tilde{q}_2 \) decreases in \( b \) while \( \tilde{q}_3 \) increases in \( b \). Observe that for \( \tilde{q}_2 \) and \( \tilde{q}_3 \) to exist, the sensitivity factor needs to be high enough. As illustrated in Figure 8, if \( b \) is small, \( f_n(q) \) becomes negative for all \( q \geq s_n = 500 \) and there is only one root (\( \tilde{q}_1 \)) that solves \( f_n(q) = 0 \).

As the load sensitivity increases, the distance between \( \tilde{q}_1 \) and \( \tilde{q}_2 \) becomes smaller. Intuitively, this suggests that it is easier for the queue length process, \( Q \), to move from the lower equilibrium, \( \tilde{q}_1 \), to the upper equilibrium, \( \tilde{q}_3 \) (through the threshold \( \tilde{q}_2 \)). At the same time, the distance between \( \tilde{q}_2 \) and \( \tilde{q}_3 \) increases. This suggests that it is harder for the queue length process to move from the upper equilibrium, \( \tilde{q}_3 \), to the lower equilibrium, \( \tilde{q}_1 \) (through the threshold \( \tilde{q}_2 \)). We formalize this intuition mathematically and show how load sensitivity affects the proportion of time the system spends around each equilibrium.
Let $Q_{\mu}^{(i)} := \{Q_{\mu}^{(i)}(t) : t \geq 0\}, i = 1, 2$, denote two queue length processes of the modified Erlang-A model with different load sensitivities where $Q_{\mu}^{(1)}$ is more sensitive than $Q_{\mu}^{(2)}$ following Definition 3 ($\mu_1(\cdot) < \mu_2(\cdot)$). Note that $Q_{\mu}^{(1)}$ and $Q_{\mu}^{(2)}$ are B&G processes with the same arrival rate, but the death rate of $Q_{\mu}^{(2)}$ is larger than the death rate of $Q_{\mu}^{(2)}$ state by state. Hence, there exists a coupling scheme such that $Q_{\mu}^{(1)}$ dominates $Q_{\mu}^{(2)}$ path by path with the same initial condition and $Q_{\mu}^{(1)}$ is stochastically larger than $Q_{\mu}^{(2)}$ in stationarity (Muller and Stoyan 2002). Taken together with Lemma 2, this stochastic comparison result gives two relations. The first compares the stationary probability for the queue length process to be around the upper equilibrium (greater than $\bar{q}_2$):

$$P(Q_{\mu}^{(1)}(\infty) > \bar{q}_2^{(1)}) \geq P(Q_{\mu}^{(1)}(\infty) > \bar{q}_2^{(2)}) \geq P(Q_{\mu}^{(2)}(\infty) > \bar{q}_2^{(2)})$$

where the first inequality follows from Lemma 2 and the second inequality follows from the stochastic comparison result. Similarly, the second relation identifies the stationary probability of the queue length process being around the lower equilibrium (less than $\bar{q}_2$):

$$P(Q_{\mu}^{(1)}(\infty) < \bar{q}_2^{(1)}) \leq P(Q_{\mu}^{(1)}(\infty) < \bar{q}_2^{(2)}) \leq P(Q_{\mu}^{(2)}(\infty) < \bar{q}_2^{(2)})$$

This implies that the more sensitive the service rate function is, the longer the system spends around $\bar{q}_3$. It confirms the phenomenon observed in Figure 5.

The next lemma identifies the effect of the abandonment rate $\theta$ on $\bar{q}_1$, $\bar{q}_2$ and $\bar{q}_3$.

**Lemma 3.** As the abandonment rate increases, the corresponding value of $\bar{q}_2$ increases and the value of $\bar{q}_3$ decreases. The value of $\bar{q}_1$ does not depend on the sensitivity of the service rate function.

Following the same line of analysis as before, Lemma 3 together with the stochastic comparison imply that as the abandonment rate increases, the system spends longer time around $\bar{q}_1$. Figure 9 illustrates how the values of $\bar{q}_1$ and $\bar{q}_2$ change with $\theta$ and Figure 10 compares the sample paths of three $M/MQ/s + M$ queues with different abandonment rates.

The analysis above implies that the abandonment rate and the sensitivity of the service rate function affect system performance differently. While higher sensitivity negatively affects performance,
Figure 10  Sample paths of the number of people in the system in $M/MQ/s + M$ queues with different abandonment rates, $\theta$ ($s = 500, \lambda = 475$ and $\mu((q-s)^+/s) = 0.6 + 0.4 \exp(-1.5(q-s)^+/s)$)

(a) $\theta = 0.25$  
(b) $\theta = 0.3$  
(c) $\theta = 0.35$

when service rates exhibit slowdowns due to congestion, a high abandonment rate may actually improve performance by alleviating the deterioration in service rate. This is in contrast to common practice, where service providers aim to minimize abandonment rates. In the traditional/non-sensitive Erlang-A model, if customers are less patient (i.e., $\theta$ increases), the probability of waiting decreases while the probability of abandonment increases; in contrast, in an Erlang-A model with high sensitivity, both the probability of waiting and the probability of abandonment decrease as $\theta$ increases, because the system reaches the high equilibrium less frequently.

6. Staffing Policies

The fluid analysis in Section 4 indicated that using regular square root staffing, without considering load dependency, may lead to poor and sometimes unstable performance measures. We propose two methods to overcome such problems. The first is to increase staffing to an appropriate level. The second is to set an admission control rule, according to which customers are re-routed elsewhere when the system is overloaded. We analyze staffing policies for the low sensitivity ($\S\S$6.1) and the high sensitivity ($\S\S$6.2) cases and show that increasing staffing is an appropriate solution for the low sensitivity case, whereas admission control is preferable in systems exhibiting high sensitivity.

6.1. Staffing policy under Low Sensitivity

We showed that when the load sensitivity is low, the system has a unique equilibrium given by $\bar{q} = s$ (Theorem 2a). The system thus fluctuates around this equilibrium and the fluctuation is within $O(\sqrt{s})$ with high probability (Gurvich et al. (2012)). This justifies the following analysis.

Let $\phi(q) := \mu((q-s)^+/s)(q \land s) + \theta(q-s)^+$ denotes the state-dependent death rate function of the queue length process. Using a Taylor expansion for the service rate function, $\mu(x) = \mu(0) + \mu'(0)x + O(x^2)$, we have

$$\phi(q) = \mu(0)(s \land q) + (\mu'(0) + \theta)(q-s)^+ + O(((q-s)^+)^2/s)).$$  (6)

We use equation (6) to approximate the dynamics of the stochastic process $Q$ by a birth and death process, $Q^T$, with birth rate $\lambda$ and state dependent death rate $\mu(0)(q \land s) + (\mu'(0) + \theta)(q-s)^+$
(the superscript $T$ stands for Taylor expansion). Observe that $Q^T$ has the same dynamics as an ordinary Erlang-A model with arrival rate $\lambda$ (the same as $Q$), a constant service rate $\mu(0)$ and a modified “abandonment” rate $\mu'(0) + \theta$ (compared to $\theta$ in $Q$). We therefore derive approximations for the probability of waiting and the probability of abandonment of the modified Erlang-A queue using approximations of the corresponding performance measures of $Q^T$. These approximations are based on the asymptotic results from Garnett et al. (2002), reviewed in Section 3.2.

The analogy between the probability of waiting of the two queues is straightforward. The probability of waiting is directly associated with the queue length process and the two queue length processes follow approximately the same dynamics. Therefore, $P(W) := P(Q(\infty) > s) \approx P(Q^T(\infty) > s) =: P(W^T)$. Using the results from Garnett et al. (2002), we can write:

$$P(W) \approx P(W^T) \approx \left(1 + \frac{h(\beta \sqrt{\mu(0)/(\mu'(0) + \theta)})}{\sqrt{\mu(0)/(\mu'(0) + \theta)} h(-\beta)}\right)^{-1}. \tag{7}$$

On the other hand, the analogy between the probability of abandonment between the two queues is not as straightforward. Since $Q^T$ has a reduced abandonment rate ($\mu'(0) < 0$), some of the transitions associated with abandonment in $Q^T$ actually reflect a slowdown of service completion in $Q$. Accordingly, the abandonment rate in $Q^T$ underestimates the actual abandonment rate in $Q$. To correct the discrepancy, we modify the Garnett approximation for the probability of abandonment. We substitute the balance equation, $(\mu'(0) + \theta)E[(Q^T(\infty) - s)^+] = \lambda P(Ab^T)$, into the balance equation, $\theta E[(Q(\infty) - s)^+] = \lambda P(Ab)$, by using the approximation $E[(Q(\infty) - s)^+] \approx E[(Q^T(\infty) - s)^+]$. This results in the following approximation:

$$P(Ab) \approx P(Ab^T) \frac{\lambda}{\mu'(0) + \theta} \times \frac{\theta}{\lambda} \approx P(W^T) \left(1 - \frac{h(\beta \sqrt{\mu(0)/(\mu'(0) + \theta)})}{h(\beta \sqrt{\mu(0)/(\mu'(0) + \theta)} + \sqrt{(\mu'(0) + \theta)/(s\mu(0))})}\right) \frac{\theta}{\mu'(0) + \theta}. \tag{8}$$

(For the special case where $\mu'(0) = -\theta$, Equations (7) and (8) are not well defined.)

Figures 11 demonstrates the accuracy of the approximations (denoted by solid lines) compared to probabilities obtained through Monte Carlo simulation (marked by ‘+’ signs). To illustrate the comparison for different parameter values, we choose evenly spaces values of the load sensitivity, measured by $\mu'(0)$, such that $-\mu'(0) \in [0, \theta)$, and the values of the SRS parameter $\beta$ between $-3$ to $3$.^3

We observe that the approximation of $P(W)$ is quite good for all load sensitivity levels, and $\beta$ values (the maximum gap is 0.05). For $P(Ab)$, the approximation is reliable for large values of $\beta$, but the accuracy deteriorates with the level of sensitivity (as $\mu'(0) \rightarrow -\theta$). For practical
Figure 11  Approximating \( P(W) \) and \( P(\text{Ab}) \) for four sensitivity levels: a: \( \mu'(0) = 0 \), b: \( \mu'(0) = -0.3 \), c: \( \mu'(0) = -0.6 \), d: \( \mu'(0) = -0.9 \). (\( \mu(0) = \theta = 1 \), \( \lambda = 500 \))

(a) Probability of waiting  (b) Probability of abandonment

purposes, however, as the QED regime aims for less than 10% abandonments, when evaluating the accuracy of the approximation, we can restrict attention to the range of \( \beta \)'s which result in \( P(\text{Ab}) < 10\% \) (corresponding to the horizontal line in Figure 11b). For example, when \( \mu'(0) = -0.9 \), the relevant range is \( \beta > -0.5 \). In that range the approximation works well (the maximum gap is 0.06). Therefore, we conclude that both approximations are reasonably accurate.

Figures 11 also illustrates that for fixed \( \beta \), system performance (both \( P(W) \) and \( P(\text{Ab}) \)) deteriorates in the presence of load sensitivity. 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. The appropriate staffing level can be determined by equations (7) and (8).

6.2. Policies to avoid bi-stability under High Sensitivity

Theorem 3 shows that service systems with high sensitivity and \( \beta < 0 \) always converge to the high equilibrium. This is because the lower equilibrium does not exist in the stochastic level in this case. The high equilibrium results in a high probability of waiting (\( P(W) = 1 \)) and a high abandonment rate (\( P(\text{Ab}) > 0.2 \)) and is associated with ED performance. To achieve a QED performance level, we therefore require \( \beta \) to be positive. But setting \( \beta > 0 \) is not sufficient to eliminate the bi-stability phenomenon and ensure QED performance.

We propose and analyze two policies to eliminate the bi-stability effect and avoid ED performance. The first solution (§§6.2.1) is to increase the staffing level sufficiently so that the system has only one (low) equilibrium. The second solution (§§6.2.2) is to control the incoming flow of arrivals by rerouting customers to other service facilities once a threshold is reached, thereby preventing
the system from reaching the high equilibrium. In the remaining of this section, we extend on each policy. We compare the performance of the two policies numerically in Subsection 6.2.3. We show that increased staffing leads to a QD staffing rule and performance while admission control can achieve QED performance. Therefore, for systems with high sensitivity, admission control is the recommended policy of the two.

6.2.1. Increase staffing. One approach to eliminate the upper equilibrium is to increase the fluid (limiting) staffing level such that \( f(\hat{q}) = \lambda - \mu((\hat{q} - s)/s)s - \theta(\hat{q} - s) = 0 \), where \( \hat{q} \) is the root of \( f'(q) = \mu'((q - s)/s) + \theta = 0 \). Letting \( x = (q - s)/s \), this is equivalent to finding \( \bar{s} \), such that

\[
\lambda/\bar{s} = \mu(\bar{x}) + \theta \bar{x},
\]

where \( \bar{x} \) is the root of \( \mu'(x) + \theta = 0 \).

We next investigate the additional staffing required to achieve that goal. As before, \( n \) denotes the system scale and \( s_n \) is the SRS level with parameter \( \beta \) (i.e. \( s_n = \lambda_n/\mu(0) + \beta\sqrt{\lambda_n/\mu(0)} \)). Let \( \bar{s}_n \) be the minimum staffing level required to eliminate bi-stability. It follows that \( \bar{s}_n = \lambda_n/(\mu(\bar{x}) + \theta \bar{x}) \).

Then the following holds as the system scale \((n)\) increases:

\[
\Delta s = \lim_{n \to \infty} \frac{\bar{s}_n - s_n}{n} = \lim_{n \to \infty} \frac{\lambda_n/n}{\mu(\bar{x}) + \theta \bar{x}} - \frac{\lambda_n/n}{\mu(0)} - \beta \sqrt{n} \frac{\lambda_n/n}{\mu(0)} = \frac{\lambda}{\mu(\bar{x}) + \theta \bar{x}} - \frac{\lambda}{\mu(0)} > 0,
\]

where the last inequality follows because \( \mu'(0) + \theta < 0 \) in the high sensitivity case.

This implies that to avoid the bi-stability phenomenon, we need to increase the staffing level by approximately \( n\Delta s \). The value of \( \Delta s \) can in some cases be quite small. For example, for the parameters used in Figure 5 - when \( b_1 = 1.25 \), \( \Delta s_1 = 3.89\% \); when \( b_2 = 1.5 \), \( \Delta s_2 = 6.54\% \); and when \( b_3 = 1.75 \), \( \Delta s_3 = 9.09\% \).

A potential drawback of this approach is that by raising the staffing level to \( \bar{s}_n \), a service provider may “overstaff” the system to operate in the QD regime, since \( \Delta s \) is positive and of order 1. An alternative might be to increase the staffing to a level lower than \( \bar{s}_n \). This, however, will not eliminate the bi-stability phenomenon. Consequently, the system may achieve a good average performance, but the variation in performance level may be very large. To illustrate this, consider a modified Erlang-A \((M/\mu_s/s + M)\) queue with arrival rate \( \lambda = 500 \), system scale \( n = 500 \), a service rate function \( \mu(q) = 0.6 + 0.4\exp(-1.75(q - s_n)^+/s_n) \), and abandonment rate \( \theta = 0.3 \). Applying the standard square root staffing rule with \( \beta = 1 \) yields \( s_n = 522 \) and the resulting performance measures (obtained from simulation) are \( P(W) \approx 0.9999 \) and \( P(Ab) \approx 0.2889 \); This is considered an ED performance level. If instead we increase the staffing level according to this policy, we get that \( s_n = 546 \), and the resulting performance measures are \( P(W) \approx 0.0299 \) and \( P(Ab) \approx 0.0003 \), which is considered a QD performance level. Consider now an intermediate increase of the staffing.
level to $s_n = 536$. At $s_n = 536$, $P(W) \approx 0.3069$ and $P(Ab) \approx 0.0563$. Although this performance is very close to QED performance, Figure 12 demonstrates that the average performance level is highly variable over time. Since our goal is to provide a stable performance over time, we exclude this and other similar solutions from the analysis.

**Figure 12** Histogram of performance measures

(a) $P(W)$  
(b) $P(Ab)$  
(c) Average queue length

### 6.2.2. Admission control

One solution to avoid bi-stability is to increase supply by increasing the staffing level on a permanent basis. We examined this solution in the previous section and showed that it results in a QD staffing rule which is inefficient. An alternative approach is to constrain demand. We examine this approach in this section. In particular, instead of increasing the staffing level, a service provider may block customers once a certain threshold is reached. This solution is motivated by the nature of the bi-stability phenomenon. By constraining the incoming flow of customers sufficiently, the system fluctuates only around the lower equilibrium (eliminating the higher equilibrium). To implement this policy, the system provider needs to characterize the appropriate threshold, and the cost that such policy entails on the system in terms of the proportion of customers blocked/rerouted.

We now consider a sequence of $M/M_q/s_n/c_n + M$ queues denoted by $\{Q_n^c(\cdot) : n \geq 1\}$. System $n$ has arrival rate $\lambda_n$, state dependent service rate $\mu((q - s_n)^+ / s_n)$, $s_n$ servers, abandonment rate $\theta$ and a finite system capacity $c_n$, so that incoming consumers are blocked once the number of customers in the system reaches $c_n$.

To find the appropriate value of $c_n$, we refer to the bi-stability analysis in Section 5. We showed that when $\tilde{q}_{n,2} < q < \tilde{q}_{n,3}$, the difference between the birth rate and the death rate, $f_n(q)$, is positive, and the queue length process tends to increase. This implies that if we set $\tilde{q}_{n,2} < c_n < \tilde{q}_{n,3}$, $c_n$ becomes a new equilibrium (instead of $\tilde{q}_{n,3}$), but the bi-stability phenomenon remains and as a consequence, we still expect relatively high $P(W)$ and $P(Ab)$. Moreover, in this case, a significant proportion of customers will be blocked.

A better choice is to set $c_n \leq \tilde{q}_{n,2}$. When $\tilde{q}_{n,1} < q < \tilde{q}_{n,2}$, $f_n(q) < 0$ and the queue length process tends to decrease. Any choice that satisfies $c_n \leq \tilde{q}_{n,2}$ will eliminate 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 $P(\text{Ab})$ are low), but increases the proportion of customers that are blocked ($P(\text{Bl})$). To find the optimal threshold within this range, a service provider should evaluate the costs associated with each performance measure and strike a balance between them. Since such costs differ greatly in different service environments, we provide the exact solution for the stationary performance measures as a function of $c_n$ and $\beta$ in Appendix B. Within this range, the unstable equilibrium, $\bar{q}_{n,2}$, is a natural candidate for $c_n$, as at $\bar{q}_{n,2}$ the proportion of blocked consumers is minimized. (Choosing a smaller threshold level ($c_n < \bar{q}_{n,2}$) is inefficient in the sense that it would waste the system’s own potential to revert to the lower equilibrium, $\bar{q}_{n,1}$.)

We therefore examine this threshold level in more detail. In particular, we set $c_n$ to

$$c_n = \lfloor \bar{q}_{n,2} \rfloor,$$

From Theorem 4, the distance between the threshold $c_n = \lfloor \bar{q}_{n,2} \rfloor$ and the number of servers $s_n$ follows

$$c_n - s_n \rightarrow -\frac{\beta \mu(0)s}{\mu'(0) + \theta} \text{ as } n \rightarrow \infty,$$

suggesting that the threshold $c_n$ is of order $\sqrt{n}$ above $s_n$.

To gain insight on the performance of this policy, we develop diffusion approximations for the new queue when $c_n = \lfloor \bar{q}_{n,2} \rfloor$. The pathwise construction of $Q_n^c$ is

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

where

$$Y_n(t) = \int_0^t 1\{Q_n^c(s) = c_n\} dA(\lambda_n t).$$

It counts the number of arrivals that are blocked from the system in $[0, t]$. We define the diffusion-scaled process

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

**Theorem 5.** Assume $\sqrt{n}(1 - \rho_n) \rightarrow \beta$ as $n \rightarrow \infty$ where $\rho_n = \lambda_n / (s_n \mu(0))$. If $\hat{Q}_n^c(0) \Rightarrow \hat{Q}^c(0)$ in $\mathbb{R}$ as $n \rightarrow \infty$, then $\hat{Q}_n^c \Rightarrow \hat{Q}^c$ in $\mathcal{D}$ as $n \rightarrow \infty$. The limit process $\hat{Q}^c$ is the unique process satisfying the stochastic integral equation:

$$\hat{Q}^c(t) = \hat{Q}^c(0) - \beta \mu(0)st + \sqrt{2\mu(0)s}B(t) - \int_0^t \left[ \mu(0)(\hat{Q}^c(u) \wedge 0) + (\mu'(0) + \theta)\hat{Q}^c(u)^+ \right] du - \hat{Y}(t) \tag{11}$$

where $\{B(t) : t \geq 0\}$ is a standard Brownian motion. $\hat{Y}$ is the unique nondecreasing nonnegative process in $\mathcal{D}$ satisfying equation (11) and

$$\int_0^\infty 1 \left\{ \hat{Q}^c(t) < -\frac{\beta \mu(0)s}{\mu'(0) + \theta} \right\} d\hat{Y}(t) = 0.$$
$Q^c_n$ is an irreducible Markov chain with finite state space. Thus, $\hat{Q}^c$ admits a unique stationary distribution, $\pi$. As $E_\pi[Q_n(t)] = E_\pi[Q_n(0)]$, by Theorem 5 and the Basic Adjoint Relation (Chen and Yao 2001),

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

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

$$P_n(Bl) = E\left[\frac{Y_n(t)}{A(\lambda_n t)}\right] \approx \frac{\sqrt{n}E_\pi[\hat{Y}(t)]}{\lambda_n t} = \frac{1}{\sqrt{n}} \left(-\beta \mu(0)s - \mu(0)E_\pi[\hat{Q}^c(0) \wedge 0] - (\mu'(0) + \theta)E_\pi[\hat{Q}^c(0)^+]\right).$$

The probability of blocking is of order $1/\sqrt{n}$. This implies that for large systems, the proportion of customers blocked and the proportion of time the system is blocked (PASTA) 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 $O(1/\sqrt{n})$ abandonments. We do not have explicit approximation formulas for these probabilities, but we demonstrate the performance using an example. Figure 13 plots the performance measures, $P(W)$, $P(Ab)$ and $P(Bl)$ and the standardized threshold level $(c_n - s_n)/\sqrt{n}$, as a function of $\beta$ under the admission control policy with different load sensitivity parameter values. As expected, $P(Bl)$ and $P(Ab)$ are of the same order ($O(1/\sqrt{n})$). Observe that $P(W)$ and $P(Ab)$ are not monotone in $\beta$ as opposed to the nonsensitive case ($b = 0$). This is because for small values of $\beta$, the threshold level is very small, which results in a relatively high probability of blocking and low $P(W)$ and $P(Ab)$. This non monotonicity implies that while reaching a QED performance level is guaranteed, it is not possible to achieve all range of desired QED performance levels. This is in contrast to systems with low sensitivity, for which the entire range of $P(W)$ is achievable.

When $\beta$ is large, the performance of all systems is similar, because despite the existence of a theoretical bi-stability, the probability to reach the high equilibrium is naturally small. Also, $P(Bl)$ is monotonically decreasing in $\beta$ and monotonically increasing in $b$ but it is in general quite small (less than 3.5%).

**6.2.3. Comparison of policies and managerial insights** In this section we test the performance of the two policies numerically. We compare how each policy improves the service level for different load sensitivity parameter values ($b$) and abandonment rates ($\theta$), and report on the implementation “cost” of each policy, i.e., the amount of added staffing or the proportion of customers blocked.

To demonstrate the effect of load sensitivity, we compare five modified Erlang-A models in which the service rate function, $\mu(q) = 0.6 + 0.4 \exp(-b_i(q - s)^+/s)$, varies only in the sensitivity
The performance measures, $P(W)$, $P(\Ab)$ and $P(\Bl)$ and the standardized threshold level $(c_n - s_n)/\sqrt{n}$ as a function of the SRS parameter $\beta$ ($\lambda = 500$, $\mu = 0.6 + 0.4 \exp(-b(q - s)^+ / s)$, $\theta = 0.3$ and assume $R = 500$)

![Figure 13](image)

Parameter $b_i$. The basic offered load for the five scenarios is $R = \lambda / \mu(0) = 500$. Table 1 reports on simulation results of average performance (using the method of batch means). In the Base Case columns we present the results of the load-sensitive system. In the Increase Staffing columns, we present the results that correspond to a staffing level of $s = R + R \Delta s$, where $\Delta s$ is determined by equation (9). Finally, in the Admission Control columns, we present the results of the system that operates under a threshold control policy with threshold level $c$, set by equation (10).

<table>
<thead>
<tr>
<th>$b$</th>
<th>$P(W)$</th>
<th>$P(\Ab)$</th>
<th>Staffing level ($\Delta s$)</th>
<th>$P(W)$</th>
<th>$P(\Ab)$</th>
<th>Threshold $c$</th>
<th>$P(W)$</th>
<th>$P(\Ab)$</th>
<th>$P(\Bl)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>0.5029</td>
<td>0.0120</td>
<td>-</td>
<td>0.0997</td>
<td>0.0233</td>
<td>579</td>
<td>0.4873</td>
<td>0.0090</td>
<td>0.0057</td>
</tr>
<tr>
<td>1.25</td>
<td>0.9820</td>
<td>0.520</td>
<td>520 (3.89%)</td>
<td>0.0293</td>
<td>0.0003</td>
<td>541</td>
<td>0.3426</td>
<td>0.0031</td>
<td>0.0107</td>
</tr>
<tr>
<td>1.75</td>
<td>1</td>
<td>0.3199</td>
<td>546 (9.09%)</td>
<td>0.0293</td>
<td>0.0003</td>
<td>541</td>
<td>0.3426</td>
<td>0.0031</td>
<td>0.0107</td>
</tr>
<tr>
<td>2.25</td>
<td>1</td>
<td>0.3563</td>
<td>569 (13.66%)</td>
<td>0.0015</td>
<td>0</td>
<td>530</td>
<td>0.2583</td>
<td>0.0016</td>
<td>0.0135</td>
</tr>
<tr>
<td>2.75</td>
<td>1</td>
<td>0.3718</td>
<td>588 (17.53%)</td>
<td>0.0001</td>
<td>0</td>
<td>525</td>
<td>0.2056</td>
<td>0.0010</td>
<td>0.0151</td>
</tr>
</tbody>
</table>
As expected, when the sensitivity increases above a certain level (e.g. $b \geq 1.25$), a load sensitive system exhibits bi-stability, which causes severe deterioration in system performance (high probabilities of waiting and abandonment).

As the sensitivity increases, to eliminate bi-stability, the additional amount of staffing required increases. In particular, the percentage of extra staff needed, $\Delta s$, increases from 3.89% (for $b = 1.25$) to 17.53% (for $b = 2.75$). As noted in Subsection 6.2.1, the percentage of increase is rather fixed regardless of system scale. These levels of extra staffing yield a QD regime performance for highly sensitive systems (e.g. $b \geq 1.75$). While the policy results in very good performance, adding these levels of extra staffing (as high as 17.53%) may lead to low worker utilization, and can potentially be a costly solution.

Alternatively, with the admission control policy, the proportion of consumers blocked is relatively low (at most 1.51% for the parameters in Table 1). $P(Bl)$ decreases further as the scale of the system, $n$, grows, since it is scaled by $1/\sqrt{n}$. The admission control policy keeps the service performance level within the QED regime characteristics for all sensitivity parameters tested. Hence, this policy achieves a good service level while keeping the staffing level according to the SRS rule.

From the simulation results presented in Table 1, it seems that the admission control policy is better than the increasing staffing policy. However, when the system scale is very small, the increasing staffing policy may be preferable. This is because when the system scale gets smaller, the extra staffing needed also decreases, as it is scaled by $n$, while the probability of blocking increases, as it is scaled by $1/\sqrt{n}$. For instance, if we reduce the scale of the system to 50, for $b = 4$, when we use the increase staffing policy, the required staffing level is 63, under which $P(W) = 0.1152$ and $P(Ab) = 0.0124$. The threshold level using the admission control policy when $s = 54$ is 57, under which $P(Bl) = 0.0484$, $P(W) = 0.0993$ and $P(Ab) = 0.0017$. While the resulting $P(W)$ is similar in both solutions, the increased staffing solution suggests adding 9 workers whereas the difference in customers lost through either blocking or abandonment is 3.77% (4.84% + 0.17% − 1.24%).

Similarly, we compare between the two policies for different abandonment rates while keeping sensitivity parameter constant. We provide this comparison in Appendix C.

7. Conclusion
Motivated by empirical findings in service systems, we modified the Erlang-A model to account for the effect of workload-dependent service rate. When the load sensitivity is low, we observe a gap between the performance of a standard Erlang-A model and the load-sensitive model. 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, using a modification of the Garnett function. When the load sensitivity is high, we observe a bi-stability phenomenon where the system fluctuates
between two equilibria: one equilibrium that results in QED performance and another equilibrium that results in ED performance. We conduct a sensitivity analysis of the frequency with which the system moves between the two equilibria and propose two operational solutions to avoid the occurrence of bi-stability. One solution involves a permanent increase of staffing and results in QD performance in general. The other solution involves a dynamic control of admission, where customers are blocked as soon as the queue length reaches a threshold. Such a solution requires a small percentage of blocking and results in QED performance. We summarize that to design service systems with workload-dependent service rates, it is not necessary to estimate the entire service rate function; it is sufficient to accurately estimate the service rate function from zero to $\sqrt{n}$. The value of the service rate function around zero is all that is needed to distinguish between low and high sensitivity cases. To choose the appropriate strategy in the high sensitivity case, it is sufficient to estimate the service rate function up to $\sqrt{n}$.

**Appendix A: Proofs**

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

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

$$= Q_n(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t)$$

$$+ \lambda_n t - \int_0^t \mu \left( \frac{(Q_n(u) - s_n)^+}{s_n} \right) (Q_n(u) \wedge s_n) du - \theta \int_0^t (Q_n(u) - s_n)^+ du$$

where

$$M_{n,1} = A(\lambda_n t) - \lambda_n t$$

$$M_{n,2} = S \left( \int_0^t \mu \left( \frac{(Q_n(u) - s_n)^+}{s_n} \right) (Q_n(u) \wedge s_n) du \right) - \int_0^t \mu \left( \frac{(Q_n(u) - s_n)^+}{s_n} \right) (Q_n(u) \wedge s_n) du$$

$$M_{n,3} = R \left( \theta \int_0^t (Q_n(u) - s_n)^+ du \right) - \theta \int_0^t (Q_n(u) - s_n)^+ du$$

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

$$\tilde{Q}_n(t) = \tilde{Q}_n(0) + \tilde{M}_{n,1}(t) - \tilde{M}_{n,2}(t) - \tilde{M}_{n,3}(t)$$

$$+ \frac{\lambda_n}{n} t - \int_0^t \mu \left( \frac{(Q_n(u) - s_n/n)^+}{s_n/n} \right) \left( \tilde{Q}_n(u) \wedge \frac{s_n}{n} \right) du - \theta \int_0^t \left( \tilde{Q}_n(u) - \frac{s_n}{n} \right)^+ du$$

Let $g(q) = -\mu(\frac{q - s}{s})(q \wedge s) - \theta(q - s)^+$. As $\mu^\prime(\cdot) \leq 0$ and $\mu^\prime(\cdot) \geq 0$, $|\mu^\prime(x)| < |\mu^\prime(0)|$. It is easy to check that

$$|g(q_1) - g(q_2)| \leq \max\{\mu(0), |\mu^\prime(0)| + \theta\} |p_1 - p_2|$$
Thus \( g(\cdot) \) is Lipschitz. This implies that
\[
q(t) = b + x(t) + \int_0^t g(q(u))du
\]
has a unique solution and it constitute a function \( \phi : D \times R \to D \) 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 \( D \) w.p. 1 as \( n \to \infty \) for \( i = 1, 2, 3 \).

Applying Functional Strong Law of Large Numbers to Poisson process, we have
\[
sup_{0 \leq t \leq T} \left\{ \frac{A(nt)}{n} - t \right\} \to 0, \sup_{0 \leq t \leq T} \left\{ \frac{S(nt)}{n} - t \right\} \to 0 \text{ and } \sup_{0 \leq t \leq T} \left\{ \frac{R(nt)}{n} - t \right\} \to 0 \text{ w.p. } 1 \text{ as } n \to \infty
\]
for any \( T > 0 \). We thus have
\[
\bar{M}_{n,1} \to \eta \text{ in } D \text{ w.p. } 1 \text{ as } n \to \infty.
\]

As \( Q_n(t) < Q_n(0) + A(\lambda_n t), \int_0^t Q_n(u)du \leq t (Q_n(0) + A(\lambda_n t)) \). 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,
\[
P \left( \left\| \bar{M}_{n,2} \right\|_T > \epsilon \right) \leq P \left( \frac{\mu(0)}{n} \int_0^T Q_n(u)du > \tau \right) + P \left( \left\| \frac{S(nt)}{n} - t \right\|_T > \frac{\epsilon}{2} \right).
\]

This leads to
\[
\bar{M}_{n,2} \to \eta \text{ in } D \text{ w.p. } 1 \text{ as } n \to \infty.
\]

Similarly we can show that
\[
\bar{M}_{n,3} \to \eta \text{ in } D \text{ w.p. } 1 \text{ as } n \to \infty.
\]

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

**Proof of theorem 2.** We prove asymptotic stability by the method of Lyapunov. Specifically, a function \( V(q) : R^+ \to R^+ \) 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 \( t \). \( \bar{q} \) is **locally asymptotically stable**, if there exist a Lyapunov function \( V(q) \) and \( \dot{V}(q) < 0 \), \( 0 < |q - \bar{q}| < \delta \) for some \( \delta > 0 \). \( \bar{q} \) is **globally asymptotically stable**, if the Lyapunov function and the stability condition hold for all \( \delta > 0 \).

For the low sensitivity case, we use the following Lyapunov function
\[
V(q) = |q - \bar{q}|
\]
where \( \bar{q} \) is the specified equilibrium. Hence,
\[
\dot{V}(q) = \text{sign}(q - \bar{q}) f(q).
\]
Under the assumptions of the low sensitivity case, \( \bar{q} = \lambda / \mu(0) = s \) and

\[
\dot{V}(q) = \begin{cases} 
-\lambda + \mu(0)q < -\lambda + \mu(0)\bar{q} = 0, & q < \bar{q}; \\
\lambda - \mu((q - s)/s) - \theta(q - s) < \lambda - \mu(0)s = 0, & q > \bar{q}.
\end{cases}
\]

Therefore, \( \bar{q} \) is globally asymptotically stable equilibrium.

Under the assumptions of the high sensitivity case, \( f(s) = 0 \), thus \( \bar{q}_1 = s \). \( f(q) \) is decreasing on \([\hat{q}, \infty)\) with \( f(\hat{q}) > 0 \) and \( \lim_{q \to \infty} f(q) = -\infty \), thus there exits \( \bar{q}_2 > \hat{q} \) such that \( f(\bar{q}_2) = 0 \).

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

Let

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

For \( q \in (\hat{q}_1, \infty) \),

\[
\dot{V}_2(q) = \begin{cases} 
-\lambda + \mu((q - s)/s)s + \theta(q - s) < -\lambda + \mu(0)\bar{q}_1 = 0, & \bar{q}_1 < q < \hat{q}, \\
\lambda - \mu((q - s)/s)s + \theta(q - s) < \lambda - \mu((\bar{q}_2 - s)/s) + \theta(\bar{q}_2 - s) = 0, & \hat{q} < q \leq \bar{q}_2, \\
\lambda - \mu((q - s)/s)s - \theta(q - s) < \lambda - \mu((\bar{q}_2 - s)/s)- \theta(\bar{q}_2 - s) = 0, & q > \bar{q}_2.
\end{cases}
\]

Therefore \( \bar{q}_2 \) is a locally asymptotically stable equilibrium. □

**Proof of Theorem 3.** \( f_n(q) \) is decreasing on \([0, s_n]\) with \( f_n(s_n) = \lambda_n - \mu(0)s_n > 0 \). By the concavity of \( f_n(q) \) on \([s_n, \infty)\) and the assumption of the high sensitivity case, \( f_n(q) \) is increasing on \([s_n, \bar{q}_n]\) and decreasing on \([\bar{q}_n, \infty)\) with \( f_n(\infty) = -\infty \). Hence, \( f_n(q) \) has only one root, \( \bar{q}_n \), and \( \bar{q}_n > \bar{q}_n \).

Let \( \psi_n(x) := \lambda_n/n - \mu(xn/s_n)s_n/n - \theta x \) and \( \hat{x}_n \) denote the point where \( \psi_n(x) \) attains its maximum on \((0, \infty)\). Note that \( \psi_n(\cdot) \) is the following transformation of \( f_n(\cdot) \): for \( q > s_n \) and \( x = (q - s_n)/n \), \( \psi_n(x) = f_n(q)/n \). Then there exist \( \bar{x}_n > \hat{x}_n \) such that \( \psi_n(\bar{x}_n) = 0 \). As \( \psi_n(\cdot) \to \psi(\cdot) \) on \([0, \infty)\) as \( n \to \infty \) and \( \eta \) is the only root of \( \psi(\cdot) \) on \([\hat{x}, \infty)\), \( \bar{x}_n, \hat{x}_n, \eta \to \infty \) as \( n \to \infty \). Using \( \bar{x}_n = (\bar{q}_n - s_n)/n \), we have \( \lim_{n \to \infty} \bar{q}_n/n = s + \eta \). □

**Proof of Lemma 1.** When \( q > s_n \), let \( \gamma_n(x) := f_n(q)/s_n \) with \( x = (q - s_n)/s_n \). Then we have

\[
\gamma_n(x) = \frac{\lambda_n}{s_n} - \mu(x) - \theta x
\]

Since \( \lambda_n/s_n \to \mu(0) \) as \( n \to \infty \), \( \gamma_n(x) \to \mu(0) - \mu(x) - \theta x \) as \( n \to \infty \) and we write \( \gamma(\cdot) := \mu(0) - \mu(0) - \theta x \). Let \( \hat{x} \) denote the root of \( \gamma'(\cdot) = 0 \), then \( (\bar{q}_n - s_n)/s_n \to \hat{x} \) as \( n \to \infty \). As \( \gamma'(0) = \mu(0) - \mu(0) = 0 \) and \( \gamma'(\hat{x}) = -\mu'(0) - \theta > 0 \) under the assumption of high sensitivity case, we have \( \gamma(\hat{x}) > 0 \). Hence \( \lim_{n \to \infty} f_n(\bar{q}_n) > 0 \). □

**Proof of Theorem 4.** We start with Equation (3). When \( \beta > 0, \lambda_n < \mu(0)s_n \). Hence, the first root of \( f_n(q) \), \( \bar{q}_{n,1} \), is in the range \([0, s_n]\), for which \( f_n(q) = \lambda_n - \mu(0)q \). Solving \( f_n(q) = 0, q < s_n \), we have \( \bar{q}_{n,1} = \lambda_n/\mu(0) \). Then

\[
\frac{\bar{q}_{n,1} - s_n}{\sqrt{n}} = -s\sqrt{n}\left(\frac{s_n}{sn} - \frac{\lambda_n}{s_n\mu(0)}\right) \to -\beta s \text{ as } n \to \infty.
\]
In order to prove Equation (4), we again let \( \psi_n(x) = \lambda_n/n - \mu(xn/s_n)s_n/n - \theta x \) and \( \hat{x}_n \) denote the point where \( \psi_n(x) \) attains its maximum on \((0, \infty)\). Recall that for \( q > s_n \) and \( x = (q - s_n)/n \), \( \psi_n(x) = f_n(q)/n \). By Assumption 1, \( \psi_n(x) \) is concave and continuously increasing on \([0, \hat{x}_n]\). Since \( \psi_n(\hat{x}_n) > 0 \) by Lemma 1 and \( \psi_n(0) < 0 \), there exists \( 0 < \bar{x}_{n,2} < \hat{x}_n \) such that \( \psi_n(\bar{x}_{n,2}) = 0 \). Under our limiting regime, \( \psi_n(\cdot) \to \psi(\cdot) \) on \([0, \infty)\) as \( n \to \infty \), where \( \psi(x) = \lambda - (\mu(x/s) + \theta x) \). Since \( \lambda/(s\mu(0)) = 1 \) and \( \psi(\cdot) \) is a continuously increasing function on \([0, \hat{x}]\), \( \psi(0) = 0 \) is the only root of \( \psi(\cdot) \) on \([0, \hat{x}]\). Therefore, \( \bar{x}_{n,2} \to 0 \) as \( n \to \infty \), meaning that the second root becomes very close to 0 as the system grows. Using Taylor expansion for the service rate function \( \mu(x) = \mu(0) + \mu'(0)x + O(x^2) \), and applying it to \( \psi_n(\bar{x}_{n,2}) = 0 \), we have

\[
\frac{\lambda_n}{n} - \mu(0)\frac{s_n}{n} - \mu'(0)\bar{x}_{n,2} - \theta \bar{x}_{n,2} + O(\bar{x}_{n,2}^2) = 0. \tag{12}
\]

Then

\[
\frac{\lambda_n/n - \mu(0)s_n/n}{\bar{x}_{n,2}} \to \mu'(0) + \theta \text{ as } n \to \infty. \tag{13}
\]

From the assumption that \( \sqrt{n}(1 - \rho_n) \to \beta \), we get that the numerator in (13) is order of \( 1/\sqrt{n} \), hence, \( \bar{x}_{n,2} = O(1/\sqrt{n}) \) too. It then follows from (12) that

\[
\sqrt{n}\bar{x}_{n,2} = \frac{\sqrt{n}(\lambda_n/n - \mu(0)s_n/n)}{\mu'(0) + \theta} + O\left(\frac{1}{\sqrt{n}}\right).
\]

Thus, \( (\bar{q}_{n,2} - s_n)/\sqrt{n} = \sqrt{n}\bar{x}_{n,2} \to -\beta\mu(0)s/(\mu'(0) + \theta) \) as \( n \to \infty \).

The proof of Equation (5) follows form the proof of Theorem 3. \( \square \)

**Proof of Lemma 2.** We consider two service rate functions \( \mu_1(\cdot) \) and \( \mu_2(\cdot) \). \( \mu_1 \) is more sensitive than \( \mu_2 \) according to Definition 3. Let \( 0 < \bar{q}_1^{(i)} < \bar{q}_2^{(i)} < \bar{q}_3^{(i)} \) denote the three equilibria of system \( i \) and \( \bar{q}^{(1)} \) be the critical point of system \( i \) for \( i = 1, 2 \). As \( \mu_1(0) = \mu_2(0) \), \( \bar{q}_1^{(1)} = \lambda/\mu_1(0) = \lambda/\mu_2(0) = \bar{q}_1^{(2)} \).

As \( \lambda < \mu_1((q - s)/s) + \theta(q - s) \leq \mu_2((q - s)/s) + \theta(q - s) \) for \( \bar{q}_1^{(1)} < q < \bar{q}_2^{(1)} \) and \( \lambda = \mu_1((\bar{q}_2^{(1)} - s)/s) + \theta(\bar{q}_2^{(1)} - s) \leq \mu_2((\bar{q}_2^{(1)} - s)/s) + \theta(\bar{q}_2^{(1)} - s) \), we have \( \bar{q}_2^{(2)} \geq \bar{q}_2^{(1)} \). Likewise as \( \lambda > \mu_2((\bar{q}_2^{(2)} - s)/s) + \theta(\bar{q}_2^{(2)} - s) \geq \mu_1((\bar{q}_2^{(2)} - s)/s) + \theta(\bar{q}_2^{(2)} - s) \) and \( \lambda = \mu_1((\bar{q}_3^{(1)} - s)/s) - \theta(\bar{q}_2^{(1)} - s) \leq \mu_1((\bar{q}_3^{(1)} - s)/s) - \theta(\bar{q}_2^{(1)} - s) \), we have \( \bar{q}_3^{(2)} \geq \bar{q}_3^{(1)} \). \( \square \)

**Proof of Theorem 5** The proof of Theorem 5 also follows from the method outlined in Pang et al. (2007). We use both Functional Central Limit Theorem (FCLT) and CMT.

We again write

\[
Q_n^c(t) = Q_n^c(0) + A(\lambda_n) - S\left(\int_0^t \mu\left(\frac{(Q_n^c(u) - s_n)^+}{s_n}\right) (Q_n^c(u) \wedge s_n) du\right) - R\left(\theta \int_0^t (Q_n^c(u) - s_n)^+ du\right) - Y_n(t)
\]

\[
= Q_n^c(0) + M_{n,1}(t) - M_{n,2}(t) - M_{n,3}(t) - Y_n(t)
\]

\[
+ \lambda_n t - \int_0^t \mu\left(\frac{(Q_n^c(u) - s_n)^+}{s_n}\right) (Q_n^c(u) \wedge s_n) du - \theta \int_0^t (Q_n^c(u) - s_n)^+ du
\]
where
\[
M_{n,1} = A(\lambda_n t) - \lambda_n t
\]
\[
M_{n,2} = S \left( \int_0^t \mu \left( \frac{(Q_n^c(u) - s_n)^+}{s_n} \right) (Q_n^c(u) \wedge s_n) du \right) - \int_0^t \mu \left( \frac{(Q_n^c(u) - s_n)^+}{s_n} \right) (Q_n^c(u) \wedge s_n) du
\]
\[
M_{n,3} = R \left( \theta \int_0^t (Q_n^c(u) - s_n)^+ du \right) - \theta \int_0^t (Q_n^c(s) - s_n)^+ du.
\]

Let \( \hat{Q}_n^c(t) = (Q_n(t) - s_n)/\sqrt{n}, \hat{Y}_n(t) = Y_n(t)/\sqrt{n} \) and \( \hat{M}_{n,i} = M_{n,i}/\sqrt{n} \) for \( i = 1, 2, 3 \). As \( \hat{Q}_n^c(\cdot) < c_n \), \( \hat{Q}_n(t) = O(\sqrt{n}) \). Applying Taylor expansion, we have
\[
\hat{Q}_n(t) = \hat{Q}_n(0) + \hat{M}_{n,1}(t) - \hat{M}_{n,2}(t) - \hat{M}_{n,3}(t) - Y_n(t)
\]
\[
+ \frac{\lambda_n - \mu(0)}{\sqrt{n}} t - \int_0^t \mu(0) (\hat{Q}_n^c(u) \wedge 0) du - \int_0^t \mu'(0) \hat{Q}_n^c(u)^+ du - \int_0^t \theta \hat{Q}_n^c(u)^+ du + O \left( \frac{1}{\sqrt{n}} \right).
\]

Let \( g(q) = -\mu'(0)(q \wedge 0) - (\mu'(0) + \theta) q^+ \). Consider the integral representation
\[
q(t) = b + x(t) + \int_0^t g(q(s)) ds - y(t), \quad (14)
\]
where \( y(t) \) is a nondecreasing nonnegative function in \( D \) such that (14) holds and \( \int_0^\infty 1 \{ q(t) < -\beta \mu(0)s/(\mu'(0) + \theta) \} dy(t) = 0 \). As \( g(\cdot) \) is Lipschitz, the integration (14) has a unique solution \( (q, y) \) and it constitutes a Bonafide function \( (\phi_1, \phi_2) : D \times R \rightarrow D \times D \) mapping \( (b, x) \) into \( (q, y) \). Moreover \( (\phi_1, \phi_2) \) is continuous (see Theorem 7.3 in Pang et al. (2007)).

\( \hat{M}_{n,i} \) are square-integrable martingales with respect to the filtration
\[
\mathcal{F}_{n,t} := \sigma \left\{ Q_n(0), A(\lambda_n s), S \left( \int_0^s \mu \left( \frac{(Q_n^c(u) - s_n)^+}{s_n} \right) (Q_n^c(u) \wedge s_n) du \right), R \left( \theta \int_0^t (Q_n^c(u) - s_n)^+ du \right) : 0 \leq s \leq t \}
\]
augmented by including all null sets. And
\[
\langle M_{n,1} \rangle(t) = \frac{\lambda_n t}{n}
\]
\[
\langle M_{n,2} \rangle(t) = \int_0^t \mu \left( \frac{(Q_n^c(u) - s_n)^+}{s_n} \right) \frac{Q_n^c(u) \wedge s_n}{n} du
\]
\[
\langle M_{n,3} \rangle(t) = \frac{\theta}{n} \int_0^t (Q_n^c(u) - s_n)^+ du.
\]

As
\[
\frac{\lambda_n t}{n} \rightarrow \lambda t \text{ as } n \rightarrow \infty \text{ w.p. 1},
\]
\( \{ \langle M_{n,1} \rangle \} \) is stochastically bounded. By the crude bound \( Q_n^c(s) < Q_n^c(0) + A(\lambda_n t) \), we have
\[
\int_0^t \mu \left( \frac{(Q_n^c(u) - s_n)^+}{s_n} \right) \frac{Q_n^c(u) \wedge s_n}{n} du \leq \mu(0) t \left( \frac{Q_n^c(0)}{n} + \frac{A(\lambda_n t)}{n} \right).
\]

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 \( D \). Thus,
\[ \frac{\hat{Q}_n^c}{\sqrt{n}} \to \eta \text{ in } \mathcal{D} \text{ as } n \to \infty \]

where \( \eta \) 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 (14), we get the result in Theorem 5. \( \square \)

**Appendix B: Stationary performance measures for \( M/M_q/s/c + M \) queue**

Let \( \pi(k) \) denote the stationary probability that the queue length process of our \( M/M_q/s/c + M \) is equal to \( k \). Using the balance equations for the birth and death process, we have

\[
\pi(k) = \begin{cases} 
\frac{\lambda^k}{k!} \frac{1}{\mu(0)^s} \pi(0) & 0 \leq k \leq s \\
\prod_{l=s+1}^{k} \frac{\lambda}{s \cdot \mu((l-s)/s) + (l-s) \cdot \theta} \left( \frac{\lambda}{s!} \right) \pi(0) & s < k \leq c
\end{cases}
\]

where

\[
\pi(0) = \left[ \sum_{k=0}^{s} \frac{\lambda^k}{k!} \frac{1}{\mu(0)^s} \pi(0) + \sum_{k=s+1}^{c} \prod_{l=s+1}^{k} \frac{\lambda}{s \cdot \mu((l-s)/s) + (l-s) \cdot \theta} \left( \frac{\lambda}{s!} \right) \pi(0) \right]^{-1}
\]

By PASTA we have

\[
P(W) = \sum_{k=s}^{c} \pi(k)
\]

and

\[
P(Bl) = \pi(c)
\]

For \( P(Ab) \), we use the balance equation

\[
\theta E[(Q - s)^+] = \lambda P(Ab)
\]

and have

\[
P(Ab) = \frac{\theta}{\lambda} \sum_{k=s+1}^{c} (k-s) \pi(k).
\]

We can also calculate other performance measures conditioning on the number of people in the queue. We refer to Appendix B in Garnett et al. (2002) for detailed derivation for a more general class of performance measures.
Appendix C: Comparison between Increase staffing and Admission control policy for High Sensitivity systems and different abandonment rates

Table 2 illustrates the policy comparison as a function of the abandonment rate, \( \theta \). In this case we vary \( \theta \), while keeping the load sensitivity parameter fixed at \( b = 2 \). As in Table 1, the admission control policy results in QED performance levels and requires blocking of only a small fraction of consumers. However, in accordance with our observation in Section 5, the abandonment rate and the sensitivity factor impact the system performance differently: a smaller abandonment rate increases congestion and hence requires an increase in additional staffing and blocking.

| \( \lambda = 500 \) (\( n = 500 \)), \( s = 511 \) (i.e., the SRS parameter \( \beta = 0.5 \)) and \( \mu(q) = 0.6 + 0.4 \exp(-2(q-s)^+/s) \) | Systems with different abandonment rates. |
|---|---|---|---|---|---|---|---|
| \( \theta \) | Base | Increase staffing | Admission control |
| \( P(W) \) | \( P(\text{Ab}) \) | Staffing level \( [\Delta S] \) | \( P(W) \) | \( P(\text{Ab}) \) | Threshold \( c \) | \( P(W) \) | \( P(\text{Ab}) \) | \( P(\text{Bl}) \) |
| 0.6 | 0.6207 | 0.0478 | - | - | - | - | - | - |
| 0.5 | 0.8904 | 0.1485 | 517 (3.36%) | 0.5665 | 0.0541 | 558 | 0.4274 | 0.0097 | 0.0080 |
| 0.4 | 0.9999 | 0.2816 | 533 (6.54%) | 0.1357 | 0.0057 | 542 | 0.3497 | 0.0044 | 0.0105 |
| 0.3 | 1 | 0.3422 | 558 (11.47%) | 0.0071 | 0.0001 | 534 | 0.2919 | 0.0021 | 0.0122 |
| 0.2 | 1 | 0.3762 | 597 (19.24%) | 0 | 0 | 530 | 0.2576 | 0.0011 | 0.0134 |

References


Boxma, O.J., M. Vlasiou. 2007. On queues with service and interarrival times depending on waiting times. *Queueing Systems* 56 121–132. 2


Chan, Carri W., Galit B. Yom-Tov, Gabriel Escobar. 2012. When to use speedup: an examination of intensive care units with readmissions. Working paper. 2


Gurvich, I., J. Huang, A. Mandelbaum. 2012. Excursion-based universal approximations for erlang-a queue in steady state. 6.1


Palm, C. 1957. Research on telephone traffic carried by full availability groups. *Tele* 1 107. 2


Whitt, W. 1990. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* 6 335–352. 2

