Adv. Appl. Prob. **47,** 450–475 (2015)
Printed in Northern Ireland

STAFFING MANY-SERVER SYSTEMS WITH ADMISSION CONTROL AND RETRIALS

A. J. E. M. JANSSEN,* ** *Eindhoven University of Technology and Eurandom*
J. S. H. VAN LEEUWAARDEN,* *** *Eindhoven University of Technology*

Abstract

In many-server systems it is crucial to staff the right number of servers so that targeted service levels are met. These staffing problems typically lead to constraint satisfaction problems that are difficult to solve. During the last decade, a powerful many-server asymptotic theory has been developed to solve such problems and optimal staffing rules are known to obey the square-root staffing principle. In this paper we develop many-server asymptotics in the so-called quality and efficiency driven regime, and present refinements to many-server asymptotics and square-root staffing for a Markovian queueing model with admission control and retrials.

Keywords: Many-server systems; QED regime; Halfin–Whitt regime; heavy traffic; diffusion limits; admission control; square-root staffing; optimality gap; asymptotic dimensioning

2010 Mathematics Subject Classification: Primary 60K25; 68M10; 41A60

1. Introduction

In this paper we consider a Markovian many-server system with admission control in the quality and efficiency driven (QED) regime, or Halfin–Whitt regime [8], in which the number of servers s and the offered workload λ are related according to a square-root principle $\lambda = s - \gamma\sqrt{s}$, for some constant γ, and s (and λ) taken to infinity. We consider an admission control policy that lets an arriving customer enter the system according to a probability depending on the queue length. In particular, a customer meeting upon arrival k other customers is admitted with probability p_k.

Since certain customers are rejected upon arrival, it seems natural to extend the model with the feature that allows rejected customers to reattempt. The modeling of reattempts or retrials is known to be challenging [4], [6], which is why one often resorts to computational approaches [1]. These numerical approaches face increasing numerical difficulties when the number of servers becomes large, which is precisely the regime we are interested in. Therefore, we combine the QED regime with a limiting regime for slow retrials, meaning that rejected customers reattempt after a relatively (compared to the time scale of the system) long time. The combination of these two asymptotic regimes leads to a tractable model with a closed-form solution, for which we are able to derive QED approximations for some of the relevant performance measures.

Received 5 July 2012; revision received 30 January 2014.
* Postal address: Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.
** Email address: a.j.e.m.janssen@tue.nl
*** Email address: j.s.h.v.leeuwaarden@tue.nl

We leverage these QED approximations to obtain results for staffing problems. The core problem of staffing in many-server systems is to determine the right tradeoff between quality and capacity. Quality is formulated in terms of some targeted service level. Take as an example the delay probability $D_F(s, \lambda)$ (see (6)). A large delay probability is perceived as negative, and the targeted service level could be to keep the delay probability below some value ε. The smaller the ε, the better the offered service. Once the targeted service level is set, the objective from the system's perspective is to determine the largest load λ (or the lowest staffing level s) such that the target $D_F(s, \lambda) \leq \varepsilon$ is met. Because the delay probability is a continuous and monotone increasing function in λ, the constraint satisfaction problem is equivalent to finding the λ_{opt} such that $D_F(s, \lambda_{\text{opt}}) = \varepsilon$. To solve this inverse problem, we shall invoke the theory of asymptotic dimensioning introduced in [3] and extended in the first part of this paper to admission control and retrials. This theory exploits the QED regime for large systems, in a way that reduces considerably the complexity of the inverse problem. That is, in the QED regime (with $\lambda = s - \gamma\sqrt{s}$, $s \to \infty$, γ not scaling with s), the performance measures in our model can be approximated by their limiting counterparts. For instance, $D_F(s, \lambda)$ can be approximated by some function $D_*(\gamma)$ that only depends on γ (and no longer on s or λ). Hence, the inverse problem can then be approximatively solved by searching for the γ_* such that $D_*(\gamma_*) = \varepsilon$, and then setting the load according to $\lambda_* = s - \gamma_*\sqrt{s}$. We refer to this rule as *conventional square-root staffing*. In this asymptotic approach, one expects that the better the approximation $D_F(s, \lambda) \approx D_*(\gamma)$, the smaller the error $|\lambda_{\text{opt}} - \lambda_*|$. Based on the QED regime, one also expects the approximation λ_* to be accurate for large values of s. Indeed, we prove that $|\lambda_{\text{opt}} - \lambda_*| = O(1)$, where a function $f(\lambda) = O(g(\lambda))$ if $\limsup_{\lambda\to\infty} |f(\lambda)/g(\lambda)| < \infty$.

We shall also derive *refined staffing rules*, for which we first develop refined QED approximations for the objective function, and then characterize the approximate solutions to the constraint satisfaction problems. The refined staffing rules are of the form

$$\lambda_\bullet = s - \gamma_*\sqrt{s} + r_\bullet, \tag{1}$$

with $r_\bullet$ a simple function of γ_*, s, and ε. We shall uniquely identify $r_\bullet$ (for each considered constraint satisfaction problem), and prove that the refined staffing level in (1) satisfies

$$\lambda_{\text{opt}} - \lambda_\bullet = O(s^{-1/2}).$$

We refer to the order term that expresses the difference between the exact optimal staffing level and the approximate staffing level as the *optimality gap*. Hence, the optimality gap of $\lambda_\bullet$ is $O(s^{-1/2})$, which suggests that the staffing level $\lambda_\bullet$ becomes more accurate as s increases. Note that $\lambda_\bullet = \lambda_* + \gamma_\bullet$. Since the optimality gap of the conventional staffing level λ_* equals $O(1)$, we can expect that $\lambda_\bullet$ should be a more accurate prescription than λ_*. In addition, because $\gamma_\bullet$ in fact describes the optimality gap of λ_*, or more precisely, $\lambda_{\text{opt}} - \lambda_* = \gamma_\bullet + O(s^{-1/2})$, it allows us to perform an analytical assessment of the accuracy of both conventional and refined square-root staffing. Theorems 5 and 6 formally establish the necessary results for both conventional and refined staffing, in terms of explicit expressions of $\gamma_\bullet$ in terms of γ_*. Numerical experiments show that, unlike in the classical Erlang C model [3], the refinement $\gamma_\bullet$ is significant in many cases, particularly when the admission control becomes more lenient, and, hence, on average more customers are admitted.

The analysis in this paper for a model with retrials and admission control is performed in a similar spirit as was performed for the Erlang C model [8], [12], the Erlang B model [11], and the Erlang A model (with abandonments) [14]. In the recent paper [2] on a many-server

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

system with retrials, the admission control policy that rejects all delayed customers (loss system) was analyzed. This paper builds on several results obtained in [2]. Compared to earlier studies [2], [8], [11], [12], and [14], the system in this paper brings about additional mathematical challenges, because of the effects of rejection and reattempts. In short, we make the following contributions:

(i) We consider two stationary performance measures: the probability that an arriving customer finds all servers occupied $D_F(s, \lambda)$, and the probability that an arriving customer is rejected $D_F^R(s, \lambda)$. For both performance measures we derive in Theorems 2 and 3 the limiting expressions in the QED regime, for the cases with and without retrials.

(ii) We next consider dimensioning problems of the following type: for fixed s, find λ for which the rejection probability $D_F^R(s, \lambda)$ has some prescribed value. We solve this inverse problem both for finite s and for $s \to \infty$ in the QED regime. We derive refined staffing rules and characterize the optimality gaps; see Theorems 5 and 6.

(iii) We then consider dimensioning problems of the following type: for fixed s, find λ for which the carried traffic $\lambda(1 - D_F^R(s, \lambda))$ has a prescribed value. These dimensioning problems are shown to have typically two solutions: a small λ leads to a large probability of being admitted $1 - D_F^R(s, \lambda)$, while a large λ leads to a small admittance probability; see Theorems 7 and 9. This phenomenon of two solutions has also been observed in certain loss networks with alternative two-link routing [7].

In Section 2 we describe the model. In Section 3 we present the limiting expressions for the performance measures in the QED regime, and we use these results to deal with the dimensioning problems in Section 4. Finally, in Section 5 we present results for dimensioning problems with multiple solutions. The proofs of the main results are given in Sections 6–8, while the proofs of supporting results are deferred to Appendix A.

2. Description of the model

The basic model described in this section stems from [5]. Consider a system with s parallel servers to which customers arrive according to a Poisson process with rate λ. The service times of customers are exponentially distributed with mean 1. An *admission policy* dictates whether a customer is admitted to the system or rejected. A customer that finds a free server upon arrival is immediately assigned to that server, and leaves the system after service. A customer that finds k other customers in the system, $k \geq s$, upon arrival is allowed to join the queue with probability p_k and is rejected with probability $1 - p_k$. In this way, the sequence $(p_k)_{k \geq s}$ defines the admission policy. Since we are interested in large many-server systems working at critical load, and, hence, serving many customers, the parameter p_k should be interpreted as the fraction of customers admitted, instead of the probability that determines the fate of just one single customer. For the results presented in this paper we impose only mild conditions on the sequence $(p_k)_{k \geq s}$, allowing for a wide range of admission policies to be considered.

Under the above Markovian assumptions, and assuming that all interarrival times and service times are mutually independent, the system can be described as a birth–death process $(C(t))_{t \geq 0}$ with $C(t)$ the number of customers in the system at time t. The birth and death rates from state k are λp_k (with $p_k = 1$ for $k = 0, \ldots, s - 1$) and $\min(k, s)$, respectively. Assuming that the stationary distribution exists, with $\pi_k = \lim_{t \to \infty} \mathbb{P}(C(t) = k)$, it readily follows from solving

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

the detailed balance equations that

$$\pi_k = \begin{cases} \pi_0 \dfrac{\lambda^k}{k!}, & 1 \le k \le s, \\ \pi_0 \dfrac{\lambda^k}{s! s^{k-s}} \displaystyle\prod_{j=s}^{k-1} p_j, & k \ge s+1, \end{cases}$$

with

$$\pi_0^{-1} = \sum_{k=0}^{s} \frac{\lambda^k}{k!} + \frac{\lambda^s}{s!} F\left(\frac{\lambda}{s}\right) \tag{2}$$

and

$$F(x) = \sum_{n=0}^{\infty} p_s \cdots p_{s+n} x^{n+1}. \tag{3}$$

From (2) it can be seen that the stationary distribution exists when $F(\lambda/s) < \infty$. Since $p_{s+n} \in [0, 1]$, we have $F(\lambda/s) < \infty$ when $0 \le \lambda < s$. When $\lambda \ge s$ we need to be more careful. The radius of convergence of the power series $F(x)$ is given by $1/P$ with

$$P := \limsup_{n\to\infty} (p_s \cdot \ldots \cdot p_{s+n})^{1/(n+1)} \in [0, 1].$$

In a major portion of the main text, we assume the following condition:

$$P \in [0, 1), \qquad F\left(\frac{1}{P} - 0\right) = \lim_{x \uparrow 1/P} F(x) = \infty. \tag{4}$$

Under this condition it can be easily observed from (3) that the stability condition for our system becomes

$$\lambda \in [0, \lambda_P) \quad \text{with } \lambda_P = \frac{s}{P}, \tag{5}$$

where $\lambda_P = \infty$ when $P = 0$. The condition (4) is certainly not as general as possible to develop the theory, but it excludes cases that need separate consideration, thereby distracting attention from the bottom line of the exposition. Stability is guaranteed when $\lambda < s$. Also, when $\lim_{k\to\infty} p_k = 0$, we have $P = 0$ and, thus, stability for all $\lambda \ge 0$. Condition (4) is also satisfied for the case $p_k = p \in (0, 1)$ for all $k \ge s$, where $F(x) = px/(1 - px)$ so that $F(1/P - 0) = \infty$, with $P = p$. We exclude at this point the case $P = 1$, which would, for example, occur in the case $p_k = 1$, $k \ge s$, and the cases that $F(1/P - 0) < \infty$. However, in Section 6 and Appendix A, the results are proved under general conditions.

Let

$$B(s, \lambda) = \frac{\lambda^s/s!}{\sum_{k=0}^{s} \lambda^k/k!}$$

denote the Erlang B formula representing the stationary blocking probability in an M/M/s/s system. An important performance measure is the stationary probability $D_F(s, \lambda) = \sum_{k=s}^{\infty} \pi_k$ that an arriving customer finds all servers occupied, given by

$$D_F^{-1}(s, \lambda) = \frac{B^{-1}(s, \lambda) + F(\lambda/s)}{1 + F(\lambda/s)}, \tag{6}$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

where D_F^{-1} is shorthand notation for $(D_F)^{-1}$. Note that for $p_k = 0$, $k \geq s$, the term $F(\lambda/s)$ vanishes, and the probability $D_F(s, \lambda)$ reduces to $B(s, \lambda)$. Also, for $p_k = 1$, $k \geq s$, the probability $D_F(s, \lambda)$ reduces to the Erlang C formula given by

$$C(s, \lambda) = \left[\frac{\lambda^s}{(s-1)!(s-\lambda)}\right]\left[\sum_{k=0}^{s-1} \frac{\lambda^k}{k!} + \frac{\lambda^s}{(s-1)!(s-\lambda)}\right]^{-1},$$

representing the stationary delay probability in an M/M/s system.

Another relevant performance measure is the stationary probability

$$D_F^R(s, \lambda) = \sum_{k=s}^{\infty} \pi_k (1 - p_k)$$

of being rejected, given by

$$D_F^{-R}(s, \lambda) = \frac{B^{-1}(s, \lambda) + F(\lambda/s)}{1 + (1 - s/\lambda)F(\lambda/s)}. \tag{7}$$

It follows from results in Section 6 that

$$\max\left\{0, 1 - \frac{s}{\lambda}\right\} \leq D_F^R(s, \lambda) \leq B(s, \lambda) \leq D_F(s, \lambda) \leq 1$$

for $0 \leq \lambda < \lambda_P$. Moreover, as a consequence of condition (4) we have

$$D_F(s, \lambda_P - 0) = 1, \qquad D_F^R(s, \lambda_P - 0) = 1 - P.$$

The birth–death process $(C(t))_{t \geq 0}$ relies on the assumption that rejected customers are considered lost. Alternatively, we could assume that rejected customers reattempt to enter the system after some time. In that case, rejected customers start producing reattempts until they are allowed to enter. Assume that periods between successive reattempts of a rejected customer are exponentially distributed with rate μ, independent of interarrival and service times. The system can then be described as a two-dimensional process $(C(t), N(t))_{t \geq 0}$ with $C(t)$ the number of customers in the system and $N(t)$ the number of rejected customers at time t. Under the above assumptions this process is a continuous-time Markov chain on the lattice infinite strip $\{0, 1, \ldots, s\} \times \mathbb{Z}_+$.

Since the transition rates of this process clearly depend on the second coordinate, the process $\{(C(t), N(t)); t \geq 0\}$ is difficult to analyze. In fact, even deriving the stationary distribution poses analytical difficulties, and no closed-form solution seems to be available. We therefore make the following assumption: reattempts arrive to the system according to a Poisson process with rate Ω, independent of the Poisson process of customers that arrive to the system for the first time. This assumption is also known as the *retrials see time averages* (RTA) approximation. Under this assumption, the total flow of customers arriving to the system is a Poisson process with rate $\lambda + \Omega$. The unknown rate Ω should then be the solution to

$$\Omega = (\lambda + \Omega) D_F^R(s, \lambda + \Omega). \tag{8}$$

Equation (8) is intuitively clear as it equates two expressions for the rate of reattempts. In some cases the RTA approximation can be theoretically justified. Cohen [4] showed that the system

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

with $p_k = 0$ for $k \geq s$, in the limit as $\mu \downarrow 0$, behaves as an Erlang loss system, except with an increased arrival intensity. More specifically, as $\mu \downarrow 0$, the distribution of the number of busy servers converges to the corresponding distribution for the standard Erlang loss system M/M/s/s (which is a truncated Poisson distribution), but with increased arrival rate $\lambda + \Omega$, where Ω is defined as the solution to (8). Indeed, in the case of infinitely long retrial times, it is intuitive that the flow of reattempts is independent from the flow of primary customers. For retrial queues with finite retrial times, the RTA approximation has proved useful and accurate for many retrial systems. We shall refer to (8) as the *generalized Cohen equation*.

Theorem 1. (Unique solution to Cohen's equation.) *Under condition (4), there is a unique solution $\Omega_{s,F}(\gamma)$ of (8) for any $\lambda \in (0, s)$.*

A proof of Theorem 1 can be distilled from [5, Section 3]. In Appendix A we present a self-contained proof of Theorem 1 under general conditions.

3. QED limits

The QED regime for many-server systems refers to scaling of the arrival rate λ and the number of servers s such that, while both λ and s increase toward infinity, the traffic intensity $\rho = \lambda/s$ approaches unity and

$$(1 - \rho)\sqrt{s} \to \gamma, \tag{9}$$

where γ is a fixed constant. The scaling combines large capacity with high utilization. For the Erlang loss and delay systems, this kind of scaling leads to the classical results (see, e.g. [11, Section 5.2]), for $\gamma \in (-\infty, \sqrt{s})$ fixed,

$$\frac{1}{\sqrt{s}} B^{-1}(s, s - \gamma\sqrt{s}) = \frac{\Phi(\gamma)}{\varphi(\gamma)} + O\left(\frac{1}{\sqrt{s}}\right), \qquad s \to \infty, \tag{10}$$

and for $\gamma \in (0, \sqrt{s})$ fixed,

$$\lim_{s\to\infty} C(s, s - \gamma\sqrt{s}) = \left(1 + \gamma \frac{\Phi(\gamma)}{\phi(\gamma)}\right)^{-1}, \tag{11}$$

where $\Phi(x)$ and $\varphi(x)$ denote the standard normal cumulative distribution function and density, respectively.

The following result will prove useful in establishing QED limiting results.

Lemma 1. (Two decompositions.) *For $\lambda \in [0, \lambda_P)$,*

$$D_F^{-1}(s, \lambda) = (1 - q_\lambda) B^{-1}(s, \lambda) + q_\lambda C^{-1}(s, \lambda), \tag{12}$$

$$D_F^{-R}(s, \lambda) = B^{-1}(s, \lambda) + \frac{q_\lambda}{1 - q_\lambda} C^{-1}(s, \lambda), \tag{13}$$

where

$$q_\lambda = \frac{(s/\lambda) F(\lambda/s)}{1 + F(\lambda/s)}. \tag{14}$$

The proof of Lemma 1 is given in Appendix A (see also [13]). Note that the function $C(s, \lambda)$ is also defined for $\lambda > s$, while in the M/M/s queue the stability condition is $\lambda < s$. It is further shown in Appendix A that $0 \leq q_\lambda \leq 1$, with $q_\lambda = 1$ if and only if $p_k = 1$, $k \geq s$. Thus, for instance, D_F^{-R} always exceeds B^{-1} and the excess is given by the second term of the right-hand side of (13), which is the product of a factor entirely determined by the admission policy

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

and the Erlang C formula. Also, D_F^{-1} is a convex combination, with a γ-dependent convexity parameter $1-q_\lambda$, of the Erlang B and C formulae. When an admission policy is mild, implying that q_λ is close to 1, we have that the Erlang C formula is dominant. When an admission policy is strict, the Erlang B formula is dominant. Aside from these general comments, the variety of weight functions q_λ that can occur, see (14), is rather substantial. In the QED regime, though, the Erlang B formula is always dominant. Indeed, in the QED limit, we have $\lambda/s \to 1$, and so $q_\lambda \to F(1)/(1+F(1))$ which is a finite number $\in (0,1)$ by our condition (4). Now (10) and (11) show that B^{-1} grows like $\sqrt{s}$ while C^{-1} remains bounded as $s \to \infty$.

We now apply the scaling (9) to the system with admission policy. We henceforth keep working with the notation for the QED regime in (9), which is why we reformulate the stability condition (5) as

$$\gamma \in (\gamma_P, \sqrt{s}]$$

with

$$\gamma_P = -\frac{1-P}{P}\sqrt{s} \in (-\infty, 0).$$

Theorem 2. (QED limits without retrials.) *Under condition (4), for $\lambda = s - \gamma\sqrt{s}$, with $\gamma \in (-\infty, \sqrt{s}]$ fixed,*

$$\lim_{s\to\infty} \sqrt{s} D_F(s, s-\gamma\sqrt{s}) = (1+F(1))\frac{\phi(\gamma)}{\Phi(\gamma)}, \tag{15}$$

$$\lim_{s\to\infty} \sqrt{s} D_F^R(s, s-\gamma\sqrt{s}) = \frac{\phi(\gamma)}{\Phi(\gamma)}. \tag{16}$$

Proof. We have by continuity of $F(x)$ at $x=1$ that, for fixed $\gamma \in (-\infty, \sqrt{s}]$,

$$F\left(\frac{s-\gamma\sqrt{s}}{s}\right) = F\left(1-\frac{\gamma}{\sqrt{s}}\right) = F(1)+o(1), \qquad s\to\infty.$$

Therefore,

$$\frac{1}{\sqrt{s}} D_F^{-1}(s, s-\gamma\sqrt{s}) = \frac{1}{1+F(1)}\,\frac{\Phi(\gamma)}{\varphi(\gamma)} + o(1), \qquad s\to\infty,$$

and

$$\frac{1}{\sqrt{s}} D_F^{-R}(s, s-\gamma\sqrt{s}) = \frac{\Phi(\gamma)}{\varphi(\gamma)} + O\left(\frac{1}{\sqrt{s}}\right), \qquad s\to\infty.$$

This implies (15) and (16).

A first observation is that the limiting expressions (15) and (16) are similar as for the Erlang B formula (10), and the only difference between the limits of D_F and B is the factor $1+F(1)$, which incorporates all information about the admission policy.

Let $\bar{D}_F(s,\lambda) = D_F(s, \lambda+\Omega)$ and $\bar{D}_F^R(s,\lambda) = D_F^R(s,\lambda+\Omega)$ with Ω as in (8). Hence, $\bar{D}_F(s,\lambda)$ and $\bar{D}_F^R(s,\lambda)$ are the stationary probability that an arriving customer finds all servers occupied, and the stationary probability that an arriving customers is rejected, respectively, in the system *with* retrials using the RTA approximation.

Theorem 3. (QED limits with retrials.) *Under condition (4), for $\lambda = s-\gamma\sqrt{s}$, with $\gamma \in (-\infty, \sqrt{s}]$ fixed, and with Ω defined as in (8),*

$$\lim_{s\to\infty} \frac{\Omega}{\sqrt{s}} = a$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

with a the unique positive solution of the equation $a = \varphi(\gamma - a)/\Phi(\gamma - a)$. Furthermore,

$$\lim_{s\to\infty} \sqrt{s}\bar{D}_F(s, s - \gamma\sqrt{s}) = (1 + F(1))\frac{\phi(\gamma - a)}{\Phi(\gamma - a)},$$
$$\lim_{s\to\infty} \sqrt{s}\bar{D}^R_F(s, s - \gamma\sqrt{s}) = \frac{\phi(\gamma - a)}{\Phi(\gamma - a)}.$$

The proof of Theorem 3 is presented in Section 7. Theorem 3 shows that the additional load due to retrials Ω, for a system with many servers, is of the order $\sqrt{s}$. In particular, as the number of servers grows large, Ω is well approximated by $a\sqrt{s}$, where a is a constant that no longer depends on s. This also means that for the overall retrial system the arrival rate $\lambda + \Omega$ is approximately $s - (\gamma - a)\sqrt{s}$.

4. Dimensioning problems

First consider the situation without retrials, and the problem of finding the arrival rate λ such that the probability $D_F(s, \lambda)$ to find all servers occupied or the probability $D^R_F(s, \lambda)$ that service is denied altogether has a prescribed value. Here the number of servers and the admission policy, embodied by F, are assumed to be given.

Problem 1. For fixed s, ε, find γ such that

$$\sqrt{s}\, D(s, s - \gamma\sqrt{s}) = \varepsilon \quad \text{with } D = D_F \text{ or } D^R_F. \tag{17}$$

In Section 6 it will be shown that, under condition (4), $D_F(s, s - \gamma\sqrt{s})$ decreases strictly from 1 at $\gamma = \gamma_P$ to 0 at $\gamma = \sqrt{s}$. Unfortunately, such a result does not hold for D^R_F: There are policies F satisfying condition (4) such that $D^R_F(s, s - \gamma\sqrt{s})$ is not monotonic as a function of γ. These policies are, however, rather rare. Relevant policies for which $D^R_F(s, s - \gamma\sqrt{s})$ can be shown to be monotonic include $p_k = p \in (0, 1)$ for $k \geq s$, and, for some $N \geq s$, $p_k = 1$ for $s \leq k \leq N$ and 0 for $k \geq N + 1$, see [10, Propositions 15 and 16]. When $D^R_F(s, s - \gamma\sqrt{s})$ is monotonic, it decreases from $1 - P$ at $\gamma = \gamma_P$ to 0 at $\gamma = \sqrt{s}$.

We have the following result.

Theorem 4. (Unique solutions.) *Under condition (4),*

(i) *Equation (17) with $D = D_F$ has a unique solution $\gamma = \gamma_{s,F}(\varepsilon)$ for any $\varepsilon \in (0, \sqrt{s})$. Assuming further that $D^R_F(s, s - \gamma\sqrt{s})$ is monotonic in γ.*

(ii) *Equation (17) with $D = D^R_F$ has a unique solution $\gamma = \gamma^R_{s,F}(\varepsilon)$ for any $\varepsilon \in (0, (1 - P)\sqrt{s})$.*

We next consider Problem 1 in the QED regime, and first introduce some definitions. Let, for $\gamma \in (\gamma_P, \sqrt{s})$,

$$g_{s,F}(\gamma) := \sqrt{s}D_F(s, s - \gamma\sqrt{s}), \qquad g^R_{s,F}(\gamma) := \sqrt{s}D^R_F(s, s - \gamma\sqrt{s}). \tag{18}$$

Furthermore, define for $\gamma \in \mathbb{R}$, $g_\infty(\gamma) = \phi(\gamma)/\Phi(\gamma)$, and define $\gamma_{\infty,F}(\varepsilon)$ and $\gamma^R_\infty(\varepsilon)$ as the solutions of

$$(1 + F(1))g_\infty(\gamma) = \varepsilon \quad \text{and} \quad g_\infty(\gamma) = \varepsilon, \tag{19}$$

respectively. It is well-known, see [2, Subsection 4.1], that $g_\infty(\gamma)$ strictly decreases from $+\infty$ at $\gamma = -\infty$ to 0 at $\gamma = \infty$, and so both equations in (19) have unique solutions when $\varepsilon > 0$.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

In Theorem 5 below, we give a limit result for $\gamma_{s,F}(\varepsilon)$ and $\gamma^R_{s,F}(\varepsilon)$ as $s \to \infty$ that involves $\gamma_{\infty,F}(\varepsilon)$ and $\gamma^R_{\infty,F}(\varepsilon)$, respectively. For this result, the following observations are made. From (6) and (7) we obtain

$$g_{s,F}(\gamma) = \sqrt{s}B(s, s-\gamma\sqrt{s})\frac{1+F(1-\gamma/\sqrt{s})}{1+B(s,s-\gamma\sqrt{s})F(1-\gamma/\sqrt{s})}$$

and

$$g^R_{s,F}(\gamma) = \sqrt{s}B(s, s-\gamma\sqrt{s})\frac{1-\gamma/\sqrt{s}(1-\gamma/\sqrt{s})^{-1}F(1-\gamma/\sqrt{s})}{1+B(s,s-\gamma\sqrt{s})F(1-\gamma/\sqrt{s})}.$$

Now we have from [9, Theorem 14],

$$\sqrt{s}B(s, s-\gamma\sqrt{s}) = g_\infty(\gamma) + \frac{1}{\sqrt{s}}h_\infty(\gamma) + O(s^{-1}), \tag{20}$$

where

$$h_\infty(\gamma) = -\tfrac{1}{3}(\gamma^3 + (\gamma^2+2)g_\infty(\gamma))g_\infty(\gamma),$$

and the O in (20) holds uniformly in any bounded set of γs. Using (20), together with

$$F\left(1-\frac{\gamma}{\sqrt{s}}\right) = F(1) - \frac{\gamma}{\sqrt{s}}F'(1) + O(s^{-1}),$$

which holds because of condition (4), the following result is established upon computation.

Lemma 2. *Under condition (4),*

$$g_{s,F}(\gamma) = (1+F(1))g_\infty(\gamma) + \frac{1}{\sqrt{s}}h_{\infty,F}(\gamma) + O(s^{-1}), \tag{21}$$

$$g^R_{s,F}(\gamma) = g_\infty(\gamma) + \frac{1}{\sqrt{s}}h^R_{\infty,F}(\gamma) + O(s^{-1}), \tag{22}$$

where

$$h_{\infty,F}(\gamma) = (1+F(1))h_\infty(\gamma) - (\gamma F'(1) + (1+F(1))F(1))g_\infty(\gamma),$$
$$h^R_{\infty,F}(\gamma) = h_\infty(\gamma) - (\gamma + g_\infty(\gamma))g_\infty(\gamma)F(1).$$

The O in (21) and (22) holds uniformly in any bounded set of γs.

Theorem 5. (Asymptotic dimensioning without retrials.) *Under condition (4),*

$$\gamma_{s,F}(\varepsilon) = \gamma_{\infty,F}(\varepsilon) + \frac{1}{\sqrt{s}}\eta_{\infty,F}(\varepsilon) + O(s^{-1}), \tag{23}$$

$$\gamma^R_{s,F}(\varepsilon) = \gamma^R_\infty(\varepsilon) + \frac{1}{\sqrt{s}}\eta^R_{\infty,F}(\varepsilon) + O(s^{-1}), \tag{24}$$

with

$$\eta_{\infty,F}(\varepsilon) = -\frac{h_{\infty,F}(\gamma_{\infty,F}(\varepsilon))}{(1+F(1))g'_\infty(\gamma_{\infty,F}(\varepsilon))}, \tag{25}$$

$$\eta^R_{\infty,F}(\varepsilon) = -\frac{h^R_{\infty,F}(\gamma^R_\infty(\varepsilon))}{g'_\infty(\gamma^R_\infty(\varepsilon))}. \tag{26}$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Theorem 5 provides the limits of $\gamma_{s,F}(\varepsilon)$ and $\gamma^R_{s,F}(\varepsilon)$ as $s \to \infty$, together with first-order corrections $\eta_{\infty,F}(\varepsilon)$ and $\eta^R_{\infty,F}(\varepsilon)$ that are simple functions of the limits $\gamma_{\infty,F}(\varepsilon)$ and $\gamma^R_\infty(\varepsilon)$. Here it is instrumental to note that

$$g'_\infty(\gamma) = -g_\infty(\gamma)(\gamma + g_\infty(\gamma)).$$

The proof of Theorem 5 will be given in Section 8.

We next discuss the numerical experiments that we conducted in order to illustrate the analytical results. Remember that the objective of Problem 1 is to determine maximal sustainable load λ such that the rejection probability $D^R_F(s,\lambda)$ is below a threshold $\varepsilon_s := \varepsilon/\sqrt{s}$. In Section 1 we have denoted the true maximal load by λ_{opt}, and we have explained the concepts of square-root staffing and asymptotic dimensioning, in order to obtain accurate estimates of λ_{opt} that are asymptotically sharp in the QED regime.

The conventional square-root staffing rule is to use the QED approximation $\sqrt{s}D^R_F(s,\lambda) \approx D_*(\gamma)$, obtain the solution to $D_*(\gamma) = \varepsilon_s$, say γ_*, and then prescribe the load as $\lambda_* = s - \gamma_*\sqrt{s}$. Theorem 5 allows for refined square-root staffing based on a better QED approximation $\sqrt{s}D^R_F(s,\lambda) \approx D_\bullet(\gamma)$ and the solution $\gamma_\bullet$ to $D_\bullet(\gamma) = \varepsilon_s$.

For the asymptotic dimensioning sketched above and in Section 1, applied to Problem 1, we identify the following key functions and parameters:

$$D_*(\gamma) = \frac{\phi(\gamma)}{\Phi(\gamma)},$$

$$D_\bullet(\gamma) = D_*(\gamma) + \frac{1}{\sqrt{s}} h^R_{\infty,F}(\gamma),$$

$$\gamma_* = \gamma^R_\infty(\varepsilon), \qquad \gamma_\bullet = \gamma_* + \frac{1}{\sqrt{s}}\eta^R_{\infty,F}(\varepsilon),$$

and using the square-root rule $\lambda = s - \gamma\sqrt{s}$,

$$\lambda_\star = s - \gamma_*\sqrt{s}, \qquad \lambda_\bullet = s - \gamma_\bullet\sqrt{s} = \lambda_\star + r_\bullet,$$

with $r_\bullet = h^R_{\infty,F}(\gamma_*)/g'_\infty(\gamma_*)$. In Table 1 we present results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with $s = 100$ servers. Note that $|\lambda_{\text{opt}} - \lambda_\bullet|$ is always less than 0.1, and how the refinements $r_\bullet$ lead to much sharper estimates of the true optimal values.

When considering the same situation as in Table 1, except with $p_k = p = 0.5$, it appears that the estimates are less accurate. Based on this example, and many other numerical experiments not reported here, we conclude that square-root staffing becomes less accurate for systems with admission control policies that allow more customers to enter. Such systems really benefit from the refined staffing rules.

Let us now turn to the same dimensioning problem, but now for the system *with* retrials.

TABLE 1: Results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with $D = D^R_F$ and $s = 100$ servers.

ε	λ_{opt}	λ_*	$\lambda_\bullet$	$r_\bullet$	$\sqrt{s}D^R_F(s,\lambda_*)$	$\sqrt{s}D^R_F(s,\lambda_\bullet)$
0.010	75.324	72.836	75.409	2.573	0.004	0.010
0.020	77.554	75.504	77.621	2.117	0.011	0.020
0.050	80.999	79.519	81.045	1.525	0.034	0.051
0.100	84.157	83.088	84.190	1.102	0.080	0.101

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Problem 2. For fixed s, ε, find λ such that

$$\sqrt{s} D_F^R(s, \lambda + \Omega) = \varepsilon \tag{27}$$

with Ω defined as in (8).

We only consider the D_F^R case in (27), but results for D_F can be obtained in a similar manner. Let us first bring the generalized Cohen equation (8) in a form that is amenable for analysis in the QED regime. Write $\Omega = a\sqrt{s}$ so that (8) becomes

$$a\sqrt{s} = (s - (\gamma - a)\sqrt{s})\, D_F^R(s, s - (\gamma - a)\sqrt{s}). \tag{28}$$

With $f_{s,F}^R$ defined as

$$f_{s,F}^R(\gamma) := \left(1 - \frac{\gamma}{\sqrt{s}}\right) g_{s,F}^R(\gamma) = \sqrt{s}\left(1 - \frac{\gamma}{\sqrt{s}}\right) D_F^R(s, s - \gamma\sqrt{s}), \tag{29}$$

we can write (28) concisely as

$$a = f_{s,F}^R(\gamma - a) \tag{30}$$

in which γ is given and a is to be solved.

We thus have $D_F^R(s, \lambda + \Omega) = \varepsilon/\sqrt{s}$ if and only if

$$g_{s,F}^R(\gamma - a_{s,F}(\gamma)) = \varepsilon, \tag{31}$$

where $a = a_{s,F}(\gamma)$ solves (30) for $\gamma > 0$.

Using $f_{s,F}^R(\delta) = (1 - \delta/\sqrt{s})\, g_{s,F}^R(\delta)$, the equation in (31) takes the form

$$\frac{a}{1 - (\gamma - a)/\sqrt{s}} = \varepsilon, \tag{32}$$

where $a = a_{s,F}(\gamma)$. It follows that

$$a = \frac{1 - \gamma/\sqrt{s}}{1 - \varepsilon/\sqrt{s}}\varepsilon, \tag{33}$$

and, therefore,

$$\gamma - a = \frac{\gamma - \varepsilon}{1 - \varepsilon/\sqrt{s}}. \tag{34}$$

Hence, we can solve Problem 2 by first solving $\delta = \delta_{s,F}^R(\varepsilon)$ from

$$g_{s,F}^R(\delta) = \varepsilon, \tag{35}$$

see (31), and then setting $(\gamma - \varepsilon)/(1 - \varepsilon\sqrt{s}) = \delta$, i.e.

$$\gamma = \gamma_{s,F}^R(\varepsilon) = \delta + \varepsilon - \frac{\delta\varepsilon}{\sqrt{s}}, \qquad \delta = \delta_{s,F}^R(\varepsilon). \tag{36}$$

As to (35) we operate under the assumption of monotonicity of D_F^R as made in the beginning of this section.

Denote by $\delta = \delta_\infty(\varepsilon)$ the solution of $g_\infty(\delta) = \varepsilon$. Observe that $\delta_\infty(\varepsilon) = \gamma_\infty^R(\varepsilon)$, see (19). Then we have by Theorem 5, (33), and (34) that

$$\delta_{s,F}^R(\varepsilon) = \delta_\infty(\varepsilon) - \frac{1}{\sqrt{s}} \frac{h_{\infty,F}^R(\delta_\infty(\varepsilon))}{g'_\infty(\delta_\infty(\varepsilon))} + O(s^{-1}).$$

Using this in (36), we arrive at the following result.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

TABLE 2: Results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 2 with retrials and $s = 100$ servers.

ε	λ_{opt}	λ_*	$\lambda_\bullet$	$r_\bullet$	$\sqrt{s}D_F^R(s,\lambda_*)$	$\sqrt{s}D_F^R(s,\lambda_\bullet)$
0.010	75.249	72.736	75.336	2.600	0.004	0.010
0.020	77.399	75.304	77.470	2.166	0.010	0.020
0.050	80.594	79.019	80.647	1.628	0.034	0.051
0.100	83.315	82.088	83.359	1.271	0.077	0.101

Theorem 6. (Asymptotic dimensioning with retrials.) *Under condition (4), and assuming that* $D_F^R(s, s-\gamma\sqrt{s})$ *is monotonic in* γ,

$$\gamma_{s,F}^R(\varepsilon) = \varepsilon + \delta_\infty(\varepsilon) + \frac{1}{\sqrt{s}}\theta_{\infty,F}^R(\varepsilon) + O(s^{-1}),$$

with

$$\theta_{\infty,F}^R(\varepsilon) = -\delta_\infty(\varepsilon)\varepsilon - \frac{h_{\infty,F}^R(\delta_\infty(\varepsilon))}{g'_\infty(\delta_\infty(\varepsilon))}.$$

For the asymptotic dimensioning scheme, applied to Problem 1 *with retrials*, we identify the following key functions and parameters:

$$\gamma_* = \varepsilon + \delta_\infty(\varepsilon), \qquad \gamma_\bullet = \gamma_* + \frac{1}{\sqrt{s}}\theta_{\infty,F}^R(\varepsilon),$$

and using the square-root rule $\lambda = s - \gamma\sqrt{s}$,

$$\lambda_\star = s - \gamma_*\sqrt{s}, \qquad \lambda_\bullet = s - \gamma_\bullet\sqrt{s} = \lambda_\star + r_\bullet,$$

with

$$r_\bullet = (\gamma_* - \varepsilon)\varepsilon + \frac{h_{\infty,F}^R(\gamma_* - \varepsilon)}{g'_\infty(\gamma_* - \varepsilon)}.$$

In Table 2 we present results for the admission policy $p_k = p = 0.1$, $k \geq s$, and for Problem 1 with retrials and $s = 100$ servers.

5. Bistability

In Section 4 we considered dimensioning problems that were formulated directly in terms of the probabilities D_F and D_F^R, and we have seen that these problems admit generically one solution when the prescribed value for D_F or D_F^R is in the appropriate range. The situation is different when dimensioning problems of a more complicated nature are considered. In this section we consider carried traffic quantities $\lambda(1 - D(s,\lambda))$ with $D = D_F$ or D_F^R, with and without retrials, and we will see that the corresponding dimensioning problems may have two solutions. One could refer to this situation as *bistability*, which has also been observed in certain loss networks with alternative two-link routing [7]. The bistability stresses the fact that sample paths of the Markov process tend to be concentrated around two relatively stable points of the system.

Note that $\lambda(1 - D_F(s,\lambda))$ is the rate of arrivals that enter the system in one attempt, and $\lambda(1 - D_F^R(s,\lambda))$ is the rate of arrivals that pass the system, possibly after having to wait in the queue that arises when all servers are occupied.

We first consider the system without retrials.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Problem 3. For fixed s, ε, find λ such that

$$\lambda(1 - D(s, \lambda)) = \varepsilon \quad \text{with } D = D_F \text{ or } D_F^R. \tag{37}$$

In Section 6 we prove the following result.

Theorem 7. ((Non-)uniqueness of solutions.) *Under condition (4),*

(i) *Equation (37) with $D = D_F$ has at least two solutions $\lambda \in (0, \lambda_P)$ when $\varepsilon > 0$ is sufficiently small.*

(ii) *Equation (37) with $D = D_F^R$ has a unique solution $\lambda \in (0, \lambda_P)$ when $\varepsilon \in (0, s)$.*

Theorem 7(i) is intuitively clear since $\lambda(1 - D_F(s, \lambda))$ is positive for all $\lambda \in (0, \lambda_P)$ while it vanishes when $\lambda \downarrow 0$ or $\lambda \uparrow \lambda_P$. The result of Theorem 7(ii) is less obvious and depends on the monotonicity of $\lambda(1 - D_F^R(s, \lambda))$ as a function of $\lambda \in (0, \lambda_P)$. For the case $p_k = p \in (0, 1),\ k \geq s$, it follows from (53) and [10, Proposition 15(iv)] that $\lambda(1-D_F(s, \lambda))$ is strictly concave, and so (37) with $D = D_F$ has *exactly* two solutions when $\varepsilon > 0$ is sufficiently small.

We next consider carried traffic quantities in the case of retrials. For convenience, we shall restrict ourselves to the choice $D = D_F^R$ in (37).

Problem 4. For fixed s, ε, find λ such that

$$\lambda(1 - D_F^R(s, \lambda + \Omega)) = \varepsilon \tag{38}$$

with Ω defined as in (8).

Observe that $\lambda(1 - D_F^R(s, \lambda + \Omega)) = \varepsilon$ if and only if

$$(s - \gamma\sqrt{s})\left(1 - \frac{1}{\sqrt{s}} g_{s,F}^R(\gamma - a_{s,F}(\gamma))\right) = \varepsilon. \tag{39}$$

Using $f_{s,F}^R(\delta) = (1 - \delta/\sqrt{s})\, g_{s,F}^R(\delta)$ and (30) in (39), we can write (38) as

$$(s - \gamma\sqrt{s})\left(1 - \frac{1}{\sqrt{s}} \frac{a}{1 - (\gamma - a)/\sqrt{s}}\right) = \sqrt{s}\, \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a} = \varepsilon, \tag{40}$$

where $a = a_{s,F}(\gamma)$. It is now not possible to simply eliminate a, and a study of the function

$$L_{s,F}(\gamma) = \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a_{s,F}(\gamma)}$$

is required.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

We now give a detailed presentation of what can be achieved analytically for the case that $F \equiv 0$ ($p_k = 0$ for all $k \geq s$). It is a challenging problem to explore in what respect this detailed analytic result can be extended to more general admission policies. We thus consider the carried-traffic problem (39) for the case that $F \equiv 0$ and study the function

$$L_s(\gamma) = L_{s,F}(\gamma) = \frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a_s(\gamma)}, \qquad 0 < \gamma < \sqrt{s}, \tag{41}$$

where $a_s(\gamma)$ is the solution of the Cohen equation $a = f_s(\gamma - a)$ with $f_s = f_{s,F\equiv 0}$. Both f_s and a_s have been studied in great detail [2]. Using the results of [2] and extensions thereof, the following is shown in Appendix B.

Theorem 8. *There holds*

$$L_s(\gamma) \in (0, \sqrt{s} - \gamma), \qquad \gamma \in (0, \sqrt{s}),$$
$$L_s(\gamma) = \gamma\, s(1 + O(\gamma\sqrt{s})), \quad \gamma \downarrow 0,$$
$$L_s(\gamma) = (\sqrt{s} - \gamma)\left(1 + O\left(\frac{e^s}{\sqrt{s}}\left(1 - \frac{\gamma}{\sqrt{s}}\right)^s\right)\right), \qquad \gamma \uparrow \sqrt{s}.$$

Furthermore, $L_s(\gamma)$ is unimodal on $(0, \sqrt{s})$, and the maximum of $L_s(\gamma)$ is assumed at the unique solution $\gamma = \hat{\gamma}_s$ of the equation

$$\gamma\, a_s(\gamma) = \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right),$$

and

$$L_s(\hat{\gamma}_s) = \frac{\hat{\gamma}_s}{\hat{\gamma}_s + (1/2\sqrt{s})}(\sqrt{s} - \hat{\gamma}_s). \tag{42}$$

From the detailed information provided by Theorem 8, it is seen that the equation

$$\frac{(\sqrt{s} - \gamma)^2}{\sqrt{s} - \gamma + a} = \frac{\varepsilon}{\sqrt{s}},$$

see (40), has two, one, or zero solutions according as $\varepsilon <, =$ or $> \sqrt{s}\, L_s(\hat{\gamma}_s)$ with $L_s(\hat{\gamma}_s)$ given in (42). Furthermore, it is seen that

$$\sqrt{s}\, L_s(\hat{\gamma}_s) = s + O(\sqrt{s}).$$

For further properties of the maximizer $\hat{\gamma}_s$, see [10, Theorem 10 and Appendix B].

Theorem 9. (Bistability.) *Problem 4 has two, one, or zero solutions according as $\varepsilon <, =$ or $> \sqrt{s}\, L_s(\hat{\gamma}_s)$ with $L_s(\hat{\gamma}_s)$ given in (42).*

In Figure 1 we display

$$\frac{1}{\sqrt{s}} L_s(\delta\sqrt{s}), \qquad 0 < \delta < 1, \text{ for } s = 1, 5, 10, 50, 100.$$

It is observed that the graphs approximate the graph of the function $1 - \delta$, $0 < \delta < 1$, when s becomes large.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

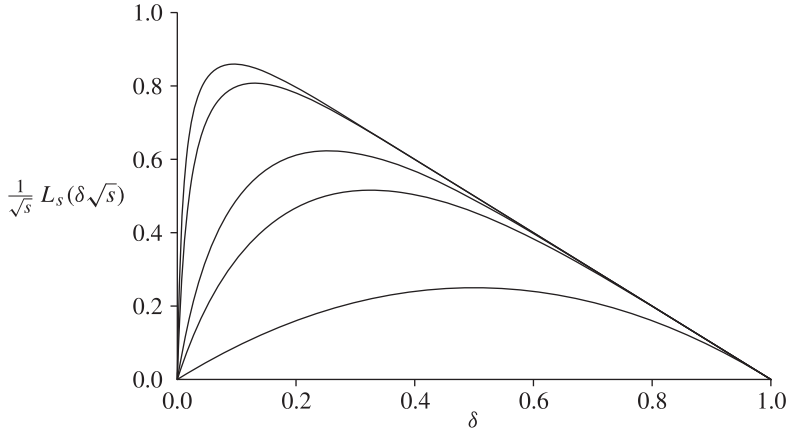

FIGURE 1: Plot of $(1/\sqrt{s})L_s(\delta\sqrt{s})$, $0 < \delta < 1$, for the cases $s = 1, 5, 10, 50, 100$.

6. Proof of the results on dimensioning

In this section we drop condition (4) so that both $P = 1$ and $F(\lambda_P - 0) < \infty$ are allowed. The equations in (17) and in (37) take the form

$$g(\gamma) = \varepsilon, \qquad g = g_{s,F} \text{ or } g^R_{s,F}, \tag{43}$$

and

$$\gamma + f(\gamma) = \frac{s - \varepsilon}{\sqrt{s}}, \qquad f = f_{s,F} \text{ or } f^R_{s,F}, \tag{44}$$

respectively, where $g^R_{s,F}$ and $f^R_{s,F}$ as in (18) and (29), and we define for $\gamma \in (\gamma_P, \sqrt{s}]$,

$$f_{s,F}(\gamma) := \left(1 - \frac{\gamma}{\sqrt{s}}\right) g_{s,F}(\gamma) = \sqrt{s}\left(1 - \frac{\gamma}{\sqrt{s}}\right) D_F(s, s - \gamma\sqrt{s}). \tag{45}$$

To get an insight and pertinent results about when the equations in (43) and (44) have (unique) solutions and how to find these solutions, we relate the functions $f_{s,F}$ and $f^R_{s,F}$ to the functions f_s and g_s given by

$$f_s(\gamma) = \left(1 - \frac{\gamma}{\sqrt{s}}\right) g_s(\gamma) = \sqrt{s}\left(1 - \frac{\gamma}{\sqrt{s}}\right) B(s, s - \gamma\sqrt{s}), \qquad \gamma \le \sqrt{s}. \tag{46}$$

The latter functions have been studied in considerable detail [2], in particular with respect to monotonicity properties. Some of these properties are collected at the beginning of Appendix A.

Lemma 3. *For* $\gamma \in (\gamma_P, \sqrt{s}]$,

$$f_{s,F}(\gamma) = f_s(\gamma) \frac{1 + (1 - \gamma/\sqrt{s}) H_s(\gamma)}{1 + 1/\sqrt{s} f_s(\gamma)\, H_s(\gamma)}, \tag{47}$$

and

$$f^R_{s,F}(\gamma) = f_s(\gamma) \frac{1 - \gamma/\sqrt{s} H_s(\gamma)}{1 + 1/\sqrt{s} f_s(\gamma) H_s(\gamma)}, \tag{48}$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

with

$$H_s(\gamma) = \sum_{n=0}^{\infty} p_s \cdot \cdots \cdot p_{s+n} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n. \tag{49}$$

Proof. This follows in a straightforward manner from (6) and (7) and the definition of H_s, where we note that $F(1 - \gamma/\sqrt{s}) = (1 - \gamma/\sqrt{s})\, H_s(\gamma)$.

Lemma 3 shows that $f_{s,F}$ and $f^R_{s,F}$ factorize into an admission policy independent part $f_s(\gamma)$ and a part comprising the admission policy via H_s. By dividing either side of (47) and (48) by $(1 - \gamma/\sqrt{s})$, it is seen that a similar result as Lemma 3 holds for $g_{s,F}$ and $g^R_{s,F}$, with the same factors comprising the admission policy as in (47) and (48).

The following result gives a global picture for $f_{s,F}$ and $f^R_{s,F}$ in terms of inequalities. We observe that $F(\lambda_P - 0) = \infty \iff H_s(\gamma_P + 0) = \infty$.

Proposition 1. (i) *For* $\gamma_P < \gamma < \sqrt{s}$,

$$\max\{0, -\gamma\} \le f^R_{s,F}(\gamma) \le f_s(\gamma) \le f_{s,F}(\gamma) \le \min\left\{\sqrt{s} - \gamma, \frac{\sqrt{s}\, f_s(\gamma)}{\gamma + f_s(\gamma)}\right\}. \tag{50}$$

(ii) *There is equality in the first inequality in (50) if and only if* $\gamma \in (0, \sqrt{s})$ *and* $p_k = 1$ *for all* $k \ge s$, *and in that case*

$$H(\gamma) = \frac{\sqrt{s}}{\gamma}, \qquad \gamma_P = 0 < \gamma < \sqrt{s}.$$

(iii) *There is equality in the second inequality in (50) for any* $\gamma \in (\gamma_P, \sqrt{s})$ *if and only if* $p_k = 0$ *for all* $k \ge s$. *There is equality in the third inequality in (50) for any* $\gamma \in (\gamma_P, \sqrt{s})$ *if and only if* $p_k = 0$ *for all* $k \ge s$.

(iv) *There is equality in the fourth inequality in (50) if and only if* $\gamma \in (0, \sqrt{s})$ *and* $p_k = 1$ *for all* $k \ge s$.

(v) *For* $\gamma = \sqrt{s}$,

$$f^R_{s,F}(\gamma) = f_s(\gamma) = f_{s,F}(\gamma) = 0.$$

(vi) $f^R_{s,F}(\gamma_P + 0) = -\gamma_P$ *if and only if* $H_s(\gamma_P + 0) = \infty$.

(vii) $f_{s,F}(\gamma_P + 0) = \sqrt{s} - \gamma_P$ *if and only if* $H_s(\gamma_P + 0) = \infty$.

The proof of Proposition 1 is given in Appendix A and uses the representations (47) and (48).

In general, monotonicity properties for the functions $f_{s,F}$, $g^R_{s,F}$ and of the functions $\gamma + f_{s,F}(\gamma)$ and $\gamma + f^R_{s,F}(\gamma)$, see (43) and (44) are not easy to establish or manifestly not true. There is the following result.

Proposition 2. (i) $f_{s,F}$ *and* $g_{s,F}$ *are strictly decreasing.*

(ii) $\gamma + f^R_{s,F}(\gamma)$, $\gamma_P < \gamma \le \sqrt{s}$, *is strictly increasing.*

The proof of Proposition 2 is given in Appendix A. We now show how the various results given in this section can be used to prove Theorem 4 and Theorem 7.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

6.1. Proof of Theorem 4

Under condition (4) we have $F(\lambda_P - 0) = H_s(\gamma_P + 0) = \infty$. It follows from (45) and (47) that

$$g_{s,F}(\gamma_P + 0) = \frac{f_{s,F}(\gamma_P + 0)}{1 - \gamma_P/\sqrt{s}} = \frac{f_s(\gamma_P)}{1 - \gamma_P/\sqrt{s}} \frac{1 - \gamma_P/\sqrt{s}}{f_s(\gamma_P)/\sqrt{s}} = \sqrt{s}. \tag{51}$$

In a similar fashion it follows from (18), (29), and (48), using $\gamma_P = -(1-P)P^{-1}\sqrt{s}$, that

$$g^R_{s,F}(\gamma_P + 0) = (1-P)\sqrt{s}.$$

Furthermore,

$$D_F(s,\lambda) = \sqrt{s} g_{s,F}(\gamma), \qquad D^R_F(s,\lambda) = \sqrt{s} g^R_{s,F}(\gamma) \tag{52}$$

when $\lambda = s - \gamma\sqrt{s}$ while $g_{s,F}(\sqrt{s}) = g^R_{s,F}(\sqrt{s}) = 0$ as $B(s,0) = 0$. Then Theorem 4(i) follows from Proposition 2(i) while Theorem 4(ii) holds by the monotonicity assumption made in the discussion preceding Theorem 4.

6.2. Proof of Theorem 7

The function $\lambda(1 - D_F(s,\lambda))$ depends continuously on $\lambda \in (0, \lambda_P)$, is positive for $\lambda \in (0, \lambda_P)$, and satisfies

$$\lim_{\lambda \downarrow 0} \lambda(1 - D_F(s,\lambda)) = 0 = \lim_{\lambda \uparrow \lambda_P} \lambda(1 - D_F(s,\lambda)),$$

see (51) and (52). This yields assertion (i).

A computation, using (45), shows that

$$\lambda(1 - D_F(s,\lambda)) = s - \sqrt{s}(\gamma + f_{s,F}(\gamma)) \tag{53}$$

when $\lambda = s - \gamma\sqrt{s}$. Hence, by Proposition 2(ii), we have that $\lambda(1 - D^R_F(s,\lambda))$ is strictly increasing in $\lambda \in (0, \lambda_P)$. Furthermore,

$$\lim_{\lambda \downarrow 0} \lambda(1 - D^R_F(s,\lambda)) = 0, \qquad \lim_{\lambda \uparrow \lambda_P} \lambda(1 - D^R_F(s,\lambda)) = s$$

by (51) and (52). This yields assertion (ii).

7. Proof of Theorem 3

We prove Theorem 3 under condition (4), where we assume that $P \in (0,1)$ (the $P = 0$ case requires only minor modification of the analysis below). Under this assumption, we have that $H_s(\gamma) \le H_s(0) < \infty$ for $\gamma \ge 0$, see (49). By Theorem 1, the generalized Cohen equation (8), written in QED coordinates as in (30), has a unique solution $a_{s,F}(\gamma)$ for any $\gamma \in (0, \sqrt{s})$.

Theorem 10. *For any $\gamma > 0$,*

$$a_\infty(\gamma) \ge a_{s,F}(\gamma) = a_\infty(\gamma) + O\left(\frac{1}{\sqrt{s}}\right), \qquad s \to \infty,$$

where the O holds uniformly in any compact set of $\gamma \in (0,\infty)$, and where $a_\infty(\gamma)$ is the unique solution of, see [2, Section 3.3],

$$a = f_\infty(\gamma - a) = \frac{\varphi(\gamma - a)}{\Phi(\gamma - a)}.$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Proof. We have $f^R_{s,F}(\delta) \le f_s(\delta) < f_\infty(\delta)$ and $f^R_{s,F}(\delta) = f_\infty(\delta) + O(1/\sqrt{s})$ uniformly in any compact set of $\delta \in \mathbb{R}$, see (48), Proposition 1(i), and [2, Propositions 5 and 6]. Let $\gamma > 0$. We prove below that $a_{s,F}(\gamma) \le a_\infty(\gamma)$. Assuming this, there is an $M > 0$ such that

$$f_\infty(\gamma - a) - \frac{M}{\sqrt{s}} \le f^R_{s,F}(\gamma - a) \le f_\infty(\gamma - a)$$

holds for $a \in [0, a_\infty(\gamma)]$ and all $s \ge 1$. Since $a + f_\infty(\gamma - a)$ is convex in a and $f'_\infty(\delta) > -1$, $\delta \in \mathbb{R}$, see [2, Equation (21)], it follows from the mean value theorem that there is an $\eta > 0$ such that

$$f_\infty(\gamma - a) - a \ge \eta(a_\infty(\gamma) - a), \qquad a \in [0, a_\infty(\gamma)],$$

where it is observed that $f_\infty(\gamma - a) - a = 0$ at $a = a_\infty(\gamma)$. Then

$$\begin{aligned} 0 = f^R_{s,F}(\gamma - a_{s,F}(\gamma)) - a_{s,F}(\gamma) &\ge f_\infty(\gamma - a_{s,F}(\gamma)) - a_{s,F}(\gamma) - \frac{M}{\sqrt{s}} \\ &\ge \eta(a_\infty(\gamma) - a_{s,F}(\gamma)) - \frac{M}{\sqrt{s}}, \end{aligned}$$

and so

$$a_{s,F}(\gamma) \ge a_\infty(\gamma) - \frac{M}{\eta\sqrt{s}}.$$

In the above argument, M and η can be chosen independently of γ in any compact subset of $(0, \infty)$, and so the result follows.

We still have to show that $a_{s,F}(\gamma) \le a_\infty(\gamma)$. We have from $f^R_{s,F}(\delta) \le f_s(\delta)$ that

$$f_\infty(\gamma - a_{s,F}(\gamma)) \ge f^R_{s,F}(\gamma - a_{s,F}(\gamma)) = a_{s,F}(\gamma).$$

Therefore,

$$\gamma - a_\infty(\gamma) + f_\infty(\gamma - a_\infty(\gamma)) = \gamma \ge \gamma - a_{s,F}(\gamma) + f_\infty(\gamma - a_{s,F}(\gamma)). \tag{54}$$

Now $\delta + f_\infty(\delta)$ is increasing, see [2, Equation (21)], and so it follows from (54) that $\gamma - a_\infty(\gamma) \ge \gamma - a_{s,F}(\gamma)$, i.e. $a_{s,F}(\gamma) \le a_\infty(\gamma)$.

The limiting behavior of $D_F(s, \lambda + \Omega)$ and $D^R_F(s, \lambda + \Omega)$ is now easily found. We have, see (45) and (21), for $\gamma > 0$,

$$\begin{aligned} \frac{1}{\sqrt{s}} D_F^{-1}(s, s - (\gamma - a_{s,F}(\gamma))) &= \frac{1 - (\gamma - a_{s,F}(\gamma))/\sqrt{s}}{f_{s,F}(\gamma - a_{s,F}(\gamma))} \\ &\to \left(1 + H(0)\frac{\phi(\gamma - a_\infty(\gamma))}{\Phi(\gamma - a_\infty(\gamma))}\right)^{-1} \\ &= ((1 + H(0))\, a_\infty(\gamma))^{-1}, \qquad s \to \infty. \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{1}{\sqrt{s}} D_F^{-R}(s, s - (\gamma - a_{s,F}(\gamma))) &= \frac{1 - (\gamma - a_{s,F}(\gamma))/\sqrt{s}}{f^R_{s,F}(\gamma - a_{s,F}(\gamma))} \\ &= \frac{1}{a_{s,F}(\gamma)}\left(1 - \frac{\gamma - a_{s,F}(\gamma)}{\sqrt{s}}\right) \\ &\to a_\infty^{-1}(\gamma), \qquad s \to \infty, \end{aligned}$$

where it also has been used that $a = a_{s,F}(\gamma)$ satisfies (30).

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

8. Proof of Theorem 5

We give the proofs of (23) and (25) in detail; the proofs of (24) and (26) being quite similar. Take any $\gamma_1, \gamma_2 \in \mathbb{R}$ with

$$\gamma_1 < \gamma_{\infty,F}(\varepsilon) < \gamma_2. \tag{55}$$

From the strict decreasingness of $g_\infty(\gamma)$ and the definition of $\gamma_{\infty,F}(\varepsilon)$, we have

$$(1 + F(1))g_\infty(\gamma_2) < (1 + F(1))g_\infty(\gamma_{\infty,F}(\varepsilon)) = \varepsilon < (1 + F(1))g_\infty(\gamma_1). \tag{56}$$

Because $g_{s,F}(\gamma) \to (1 + F(1))g_\infty(\gamma)$ as $s \to \infty$, we have from (56) that $g_{s,F}(\gamma_2) < \varepsilon < g_{s,F}(\gamma_1)$ when s is large. By monotonicity of $g_{s,F}$, this implies that $\gamma_{s,F}(\varepsilon) \in [\gamma_1, \gamma_2]$ for large s. Since γ_1 and γ_2 in (55) are arbitrary, it follows that $\gamma_{s,F}(\varepsilon) \to \gamma_{\infty,F}(\varepsilon)$ as $s \to \infty$. In particular, $\gamma_{s,F}(\varepsilon)$ is bounded in $s \geq 1$.

We next show from the weakened form

$$g_{s,F}(\gamma) = (1 + F(1))g_\infty(\gamma) + O(s^{-1/2}) \tag{57}$$

of (21) that

$$\gamma_{s,F}(\varepsilon) = \gamma_{\infty,F}(\varepsilon) + O(s^{-1/2}). \tag{58}$$

To that end, we consider (57) with $\gamma = \gamma_{s,F}(\varepsilon)$ and write

$$\begin{aligned} g_\infty(\gamma_{s,F}(\varepsilon)) = g_\infty(\gamma_{\infty,F}(\varepsilon)) + (\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_\infty(\gamma_{\infty,F}(\varepsilon)) \\ + O((\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))^2). \end{aligned} \tag{59}$$

Using this, together with

$$g_{s,F}(\gamma_{s,F}(\varepsilon)) = \varepsilon = (1 + F(1))g_\infty(\gamma_{\infty,F}(\varepsilon)), \tag{60}$$

in (57) we obtain

$$\begin{aligned} \varepsilon = \varepsilon + (1 + F(1))(\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_\infty(\gamma_{\infty,F}(\varepsilon)) \\ + O((\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))^2) + O(s^{-1/2}). \end{aligned} \tag{61}$$

It is well known, see [2, Equation (21)], that $g'_\infty(\gamma)$ is negative and bounded away from 0 when γ is in a bounded set. From (61) we, therefore, obtain (58).

We finally show (23), and for this we repeat the argument for showing (58), but now using the full strength of (21) with $\gamma = \gamma_{s,F}(\varepsilon)$. Using (58) in (59) yields

$$g_\infty(\gamma_{s,F}(\varepsilon)) = g_\infty(\gamma_{\infty,F}(\varepsilon)) + (\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_\infty(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1}).$$

Furthermore, again by (58),

$$h_{\infty,F}(\gamma_{s,F}(\varepsilon)) = h_{\infty,F}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1/2}).$$

When we use this in (21) with $\gamma = \gamma_{s,F}(\varepsilon)$ together with (60), we obtain

$$\begin{aligned} \varepsilon = \varepsilon + (1 + F(1))(\gamma_{s,F}(\varepsilon) - \gamma_{\infty,F}(\varepsilon))g'_\infty(\gamma_{\infty,F}(\varepsilon)) \\ + O(s^{-1}) + \frac{1}{\sqrt{s}} h_{\infty,F}(\gamma_{\infty,F}(\varepsilon)) + O(s^{-1}). \end{aligned}$$

From this (23) and (25) follow at once.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Appendix A. Remaining proofs, except the proof of Theorem 8

We start by recalling some basic properties, shown in [2], of the functions f_s and g_s given in (46). We have

(a) $f_s(\gamma)$ is strictly convex and decreases strictly in $-\infty < \gamma \le \sqrt{s}$ from $+\infty$ to 0, and $\gamma + f_s(\gamma)$ increases strictly in $-\infty < \gamma \le \sqrt{s}$ from 0 to $\sqrt{s}$. Furthermore, $\gamma + f_s(\gamma) = O(1/\gamma)$, $\gamma \to -\infty$.

(b) $g_s(\gamma)$ decreases strictly in $-\infty < \gamma \le \sqrt{s}$ from $\sqrt{s}$ to 0.

(c) $f_s(\gamma)$ and $g_s(\gamma)$ increase in $s \ge 1$ to $f_\infty(\gamma) = g_\infty(\gamma) = \varphi(\gamma)/\Phi(\gamma)$ uniformly in any compact set of $\gamma \in \mathbb{R}$.

Proof of Proposition 1. We have, for $0 < \gamma < \sqrt{s}$,

$$H_s(\gamma) \le \sum_{n=0}^{\infty} \left(1 - \frac{\gamma}{\sqrt{s}}\right)^n = \frac{\sqrt{s}}{\gamma}, \tag{62}$$

with equality if and only if $p_{s+n} = 1$, $n = 0, 1, \ldots$. Hence, $1 - \gamma\, H_s(\gamma)/\sqrt{s} \ge 0$ for $0 < \gamma < \sqrt{s}$, and, evidently, $1 - \gamma\, H_s(\gamma)/\sqrt{s} \ge 0$ for $\gamma_P < \gamma \le 0$. Therefore, from (48), $f^R_{s,F}(\gamma) \ge 0$, $\gamma_P < \gamma < \sqrt{s}$, with equality for any γ if and only if $p_{s+n} = 1$, $n = 0, 1, \ldots$. Next, we write (48) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f^R_{s,F}(\gamma) = -\gamma + \frac{\gamma + f_s(\gamma)}{1 + (1/\sqrt{s}) f_s(\gamma)\, H_s(\gamma)}, \tag{63}$$

and then it follows from $\gamma + f_s(\gamma) > 0$ that $f^R_{s,F}(\gamma) > -\gamma$. This proves the first inequality in (50).

Next, we write (48) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f^R_{s,F}(\gamma) = f_s(\gamma) - \frac{1}{\sqrt{s}} f_s(\gamma)\, H_s(\gamma) \frac{\gamma + f_s(\gamma)}{1 + (1/\sqrt{s}) f_s(\gamma)\, H_s(\gamma)}, \tag{64}$$

and it follows from $\gamma + f_s(\gamma) > 0$ that $f^R_{s,F}(\gamma) \le f_s(\gamma)$, with equality if and only if $H_s(\gamma) = 0$ if and only if $p_{s+n} = 0$, $n = 0, 1, \ldots$. This proves the second inequality in (50).

Next, we write (47) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}(\gamma) = f_s(\gamma) + f_s(\gamma) H_s(\gamma) \frac{1 - (\gamma + f_s(\gamma))/\sqrt{s}}{1 + f_s(\gamma) H_s(\gamma)/\sqrt{s}}, \tag{65}$$

and it follows from $\gamma + f_s(\gamma) < \sqrt{s}$ that $f_{s,F}(\gamma) \ge f_s(\gamma)$, with equality if and only if $H_s(\gamma) = 0$ if and only if $p_{s+n} = 0$, $n = 0, 1, \ldots$. This proves the third inequality in (50).

Next, we write (47) for $\gamma_P < \gamma < \sqrt{s}$ as

$$f_{s,F}(\gamma) = \sqrt{s} - \gamma - \sqrt{s}\, \frac{\sqrt{s} - \gamma - f_s(\gamma)}{\sqrt{s} + f_s(\gamma)\, H_s(\gamma)}, \tag{66}$$

and it follows from $\gamma + f_s(\gamma) < \sqrt{s}$ that $f_{s,F}(\gamma) < \sqrt{s} - \gamma$. Since $\sqrt{s} - \gamma > f_s(\gamma)$ for $\gamma < \sqrt{s}$, we have that the function $x \ge 0 \mapsto (1 + x(1 - \gamma/\sqrt{s}))/(1 + x\, f_s(\gamma)/\sqrt{s})$ is strictly increasing. From (47) and (62) it then follows (with $x = H_s(\gamma) \le \sqrt{s}/\gamma$) that

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

$f_{s,F}(\gamma) \le \sqrt{s}\, f_s(\gamma)/(\gamma + f_s(\gamma))$ for $0 < \gamma < \sqrt{s}$, with equality if and only if $p_{s+n} = 0$, $n = 0, 1, \ldots$. Furthermore, for $\gamma \le 0$, we have

$$\sqrt{s} - \gamma - \frac{\sqrt{s}\, f_s(\gamma)}{\gamma + f_s(\gamma)} = \frac{\gamma(\sqrt{s} - \gamma - f_s(\gamma))}{\gamma + f_s(\gamma)} \le 0,$$

with equality if and only if $\gamma = 0$. This proves the fourth inequality in (50).

The cases of equality in the inequalities in (50) have been indicated already along with their proofs, and this settles Proposition 1(ii)–(iv).

Proposition 1(v) follows from Proposition 1(i) and the fact that $f_s(\sqrt{s}) = 0$.

Proposition 1(vi)–(vii) follow from the representations (48) and (47) and the fact that $H_s(\gamma)$ increases to $H_s(\gamma_P + 0)$ as γ decreases to γ_P by nonnegativity of all p_k. This completes the proof of Proposition 1.

Note. We have the following consequences of (64), (65), and $\gamma + f_s(\gamma) = O(1/\gamma)$, $\gamma \to -\infty$:

(a) $f^R_{s,F}(\gamma) = f_s(\gamma) + O(1/\gamma)$, $\gamma_P < \gamma < 0$,

(b) $f_{s,F}(\gamma) = f_s(\gamma) + \sqrt{s} + O(1/\gamma)$, $\gamma_P < \gamma < 0$, when $F \not\equiv 0$,

(c) $g^R_{s,F}(\gamma), g_{s,F}(\gamma) = \sqrt{s} + O(1/\gamma)$, $\gamma_P < \gamma < 0$.

These results are in particular relevant when $P = 0$ so that $\gamma_P = -\infty$.

Proof of Proposition 2. (i) Since $f_{s,F}(\gamma) = (1 - \gamma/\sqrt{s})\, g_{s,F}(\gamma)$, it is sufficient to show that $g_{s,F}(\gamma)$ is strictly decreasing. We have from (66) that

$$g_{s,F}(\gamma) = \frac{f_{s,F}(\gamma)}{1 - \gamma/\sqrt{s}} = \sqrt{s}\left(1 - \frac{1 - (1/\sqrt{s}) g_s(\gamma)}{1 + (1/\sqrt{s}) f_s(\gamma)\, H_s(\gamma)}\right).$$

Now $g_s(\gamma)$ strictly decreases in $-\infty < \gamma \le \sqrt{s}$ from $\sqrt{s}$ to 0, and $f_s(\gamma)\, H_s(\gamma)$ is nonnegative and decreasing in $\gamma_P < \gamma \le \sqrt{s}$. It follows that

$$\frac{1 - (1/\sqrt{s}) g_s(\gamma)}{1 + (1/\sqrt{s}) f_s(\gamma)\, H_s(\gamma)}$$

is nonnegative and strictly increasing in $\gamma_P < \gamma \le \sqrt{s}$, and the proof is complete.

(ii) We have that $\gamma + f_s(\gamma)$ is positive and strictly increasing in $-\infty < \gamma \le \sqrt{s}$, and $f_s(\gamma)\, H_s(\gamma)$ is nonnegative and decreasing in $\gamma_P < \gamma \le \sqrt{s}$. It follows that

$$\frac{\gamma + f_s(\gamma)}{1 + (1/\sqrt{s}) f_s(\gamma)\, H_s(\gamma)}$$

is strictly increasing in $\gamma_P < \gamma \le \sqrt{s}$. Then it follows from the representation (63) of $f^R_{s,F}$ that $\gamma + f^R_{s,F}(\gamma)$ is strictly increasing in $\gamma_P < \gamma \le \sqrt{s}$.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Proof of Lemma 1. We have from (46) and (47) that

$$\begin{aligned}\frac{1}{\sqrt{s}} D_F^{-1}(s, s-\gamma\sqrt{s}) &= \frac{1-\gamma/\sqrt{s}}{f_{s,F}(\gamma)} \\ &= \frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} \frac{1+(1/\sqrt{s}) f_s(\gamma) H_s(\gamma)}{1+(1-\gamma/\sqrt{s})\, H_s(\gamma)} \\ &= \frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} \frac{1-(1/\sqrt{s})\gamma H_s(\gamma)+(1/\sqrt{s}) H_s(\gamma)(f_s(\gamma)+\gamma)}{1+(1-\gamma/\sqrt{s})\, H_s(\gamma)} \\ &= (1-p(\gamma)) \frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} + p(\gamma) \frac{1}{\sqrt{s}} \frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} (f_s(\gamma)+\gamma), \quad (67)\end{aligned}$$

with $p(\gamma) = q_\lambda$ where q_λ is given in (14). Now, by (46),

$$\frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} = \frac{1}{\sqrt{s}} B^{-1}(s, s-\gamma\sqrt{s}), \quad (68)$$

and

$$\frac{1-\gamma/\sqrt{s}}{f_s(\gamma)} (f_s(\gamma)+\gamma) = 1 - \frac{\gamma}{\sqrt{s}} + \frac{\gamma}{\sqrt{s}} B^{-1}(s, s-\gamma\sqrt{s}) = C^{-1}(s, s-\gamma\sqrt{s}). \quad (69)$$

Then (12) follows from (67)–(69). The proof of (13) is similar.

We observe that $0 \le p(\gamma) \le 1$, which follows from (62).

Proof of Theorem 1. From [5, Section 3] one can extract a proof of Theorem 1. This proof depends on basic properties of birth–death processes. The proof that we give here is presented in QED coordinates and uses the analytic properties of $f^R_{s,F}$, such as given in Propositions 1 and 2. The proof is given under general conditions, so that also $P = 1$ and $F(1/P - 0)$, $H_s(\gamma_P + 0) < \infty$ is allowed.

We distinguish the following cases:

(a) $P = 0$. Then $\gamma_P = -\infty$, and both $H_s(\gamma)$ and $f^R_{s,F}(\gamma)$ are well-defined and analytic in $-\infty < \gamma \le \sqrt{s}$. Set $\gamma_{P,F} = 0$.

(b) $P \in (0, 1]$. Then $\gamma_P = -(1-P)\sqrt{s}/P \in (-\infty, 0]$, and both $H_s(\gamma)$ and $f^R_{s,F}(\gamma)$ are well-defined and analytic in $\gamma_P < \gamma \le \sqrt{s}$. In this case b, we distinguish the subcases:

(b1) $H_s(\gamma_P + 0) < \infty$. Then, by Abel's theorem, $H_s(\gamma)$ is continuous in $\gamma_P \le \gamma \le \sqrt{s}$, and so is $f^R_{s,F}(\gamma)$;

(b2) $H_s(\gamma_P + 0) = \infty$.

In these cases we have from (48) and (63)

$$f^R_{s,F}(\gamma_P + 0) = -\gamma_P + \frac{\gamma_P + f_s(\gamma_P)}{1 + (1/\sqrt{s}) f_s(\gamma_P) H_s(\gamma_P + 0)}.$$

Hence, $f^R_{s,F}(\gamma_P + 0) > -\gamma_P$ in (b1) while $f^R_{s,F}(\gamma_P + 0) = -\gamma_P$ in (b2). Set $\gamma_{P,F} = f^R_{s,F}(\gamma_P + 0) + \gamma_P$.

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

In (a), the $P = 0$ case, we have by Propositions 1 and 2 that

$$f^R_{s,F}(\gamma) > -\gamma, \qquad (f^R_{s,F})'(\gamma) > -1, \tag{70}$$

for $\gamma < \sqrt{s}$, and it follows as in the proof of [2, Theorem 8, Section 4.3], that for any $\gamma \in (0, \sqrt{s})$ there is a unique solution a of the equation $a = f^R_{s,F}(\gamma - a)$. This solution, $a_{s,F}(\gamma)$, satisfies $a_{s,F}(\gamma) \to +\infty$ as $\gamma \downarrow 0$. For if $b := \liminf_{\gamma\downarrow 0} a_{s,F}(\gamma) < \infty$, we would have $b = \liminf_{\gamma\downarrow 0} f(\gamma - a_{s,F}(\gamma)) = f(-b)$, contradicting the first item in (70). Similarly, it can be shown, compare the beginning of the proof of [2, Theorem 8, Section 4.8], by considering $c := \limsup_{\gamma\uparrow\sqrt{s}} a_{s,F}(\gamma)$, that $a_{s,F}(\gamma) \to 0$ as $\gamma \uparrow \sqrt{s}$. Assume now that $P \in (0, 1]$, in which we exclude the case that $p_k = 1$, $k \geq s$. Now (70) holds for $\gamma_P < \gamma < \sqrt{s}$. Let $\gamma \in (\gamma_{P,F}, \sqrt{s})$. We have

$$f^R_{s,F}(\gamma - 0) > 0, \tag{71}$$

while

$$f^R_{s,F}(\gamma - (\gamma - \gamma_P) + 0) = \gamma_{P,F} - \gamma_P < \gamma - \gamma_P,$$

and so

$$f^R_{s,F}(\gamma - \delta) < \delta \tag{72}$$

when δ is less than but close to $\gamma - \gamma_P > 0$. By continuity, it follows from (71) and (72) that the equation $a = f^R_{s,F}(\gamma - a)$ has a solution a, and this solution is unique by (70). To show that this solution, $a_{s,F}(\gamma)$, satisfies $a_{s,F}(\gamma) \to f^R_{s,F}(\gamma_P + 0)$ as $\gamma \downarrow \gamma_{P,F}$, we need the following lemma, whose proof is an exercise in basic analysis of functions of one real variable.

Lemma 4. *Let $c < 0 < d$ and let $h\colon (c, d) \to \mathbb{R}$ be smooth and such that $h(a) > 0$, $h'(a) < 1$ for $c < a < d$ while $h(a) \to d$ when $a \uparrow d$. Then for any ε, $0 < \varepsilon < -c$, there is a unique solution $a(\varepsilon)$ of the equation $a = h(a - \varepsilon)$, and $a(\varepsilon) \to d$ as $\varepsilon \downarrow 0$.*

Taking in the lemma $h(a) = f^R_{s,F}(\gamma_{P,F} - a)$ and $c := -\sqrt{s} + \gamma_{P,F} < a < f^R_{s,F}(\gamma_P + 0) =: d$, it follows that $a_{s,F}(\gamma) \to f^R_{s,F}(\gamma_P + 0)$ as $\gamma \downarrow \gamma_{P,F}$. Finally, $a_{s,F}(\gamma) \to 0$ as $\gamma \uparrow \sqrt{s}$ can be shown by using the same argument as in the proof of [2, Theorem 4, Subsection 5.8], for showing that $a(\gamma) \to 0$ as $\gamma \uparrow \sqrt{s}$. This completes the proof.

Appendix B. Proof of Theorem 8

In this appendix we present the proof of Theorem 8 on the function L_s in (41), given in terms of the function $a_s(\gamma)$ that solves Cohen's equation $a = f_s(\gamma - a)$. The proofs rely heavily on (extensions of) the results in [2]. In particular, we use

$$\gamma a_s(\gamma) = 1 - \frac{2\gamma}{\sqrt{s}} - \left(1 - \frac{2}{s}\right)\gamma^2 + 4\left(1 - \frac{1}{s}\right)\frac{\gamma^3}{\sqrt{s}} + O(\gamma^4), \qquad \gamma \downarrow 0, \tag{73}$$

which is a sharpening of [2, Theorem 3]. This sharpening can be obtained by the method to prove [2, Theorem 3] where, as an intermediate step, [2, Proposition 2] should be sharpened to

$$f_s(\delta) = -\delta - \frac{1}{\delta} - \frac{2}{\delta^2\sqrt{s}} + \left(2 - \frac{6}{s}\right)\frac{1}{\delta^3} + \left(16 - \frac{24}{s}\right)\frac{1}{\delta^4\sqrt{s}} + O\left(\frac{1}{\delta^5}\right), \qquad \delta \to -\infty,$$

using the methods of [2, Section 4.1]. We shall also use and sharpen [2, Proposition 1],

$$1 - \frac{2}{\sqrt{s}}\gamma - \gamma^2 < \gamma a_s(\gamma) < 1 - \frac{1}{\sqrt{s}}\gamma, \qquad 0 < \gamma < \sqrt{s}. \tag{74}$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Proof of $0 < L_s(\gamma) < \sqrt{s} - \gamma$, $0 < \gamma < \sqrt{s}$. This follows from the definition in (41) and $a_s(\gamma) > 0$.

Proof of $L_s(\gamma) = \gamma s(1 + O(\gamma\sqrt{s}))$, $\gamma \downarrow 0$. This follows from the definition in (41) and (74).

Proof of $L_s(\gamma) = (\sqrt{s} - \gamma)(1 + O(s^{-1/2}e^s(1 - \gamma/\sqrt{s})^s))$, $\gamma \uparrow \sqrt{s}$. This follows from the proof of [2, Theorem 4, Subsection 5.8] and Stirling's formula.

Proposition 3. *For* $0 < \gamma < \sqrt{s}$,

$$L_s'(\gamma) = 0 \Longleftrightarrow \gamma a_s(\gamma) = \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right).$$

Proof. We have, from the definition of L_s in (41),

$$L_s'(\gamma) = 0 \Longleftrightarrow \sqrt{s} - \gamma + 2a_s(\gamma) + (\sqrt{s} - \gamma)\, a_s'(\gamma) = 0. \tag{75}$$

By implicit differentiation in [2, Equation (14)] and the expression in [2, Subsection 4.3] for f_s' in terms of f_s we have

$$a_s'(\gamma) = \frac{-a_s(\gamma)(\gamma + 1/\sqrt{s})}{1 - \gamma/\sqrt{s} - \gamma\, a_s(\gamma)}, \qquad 0 < \gamma < \sqrt{s}. \tag{76}$$

Using this in (75) with the facts that $\gamma > 0$ and $1 - \gamma/\sqrt{s} - \gamma\, a_s(\gamma) > 0$, we have, for $0 < \gamma < \sqrt{s}$,

$$L_s'(\gamma) = 0 \Longleftrightarrow (\gamma a_s(\gamma))^2 + \left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(\gamma\sqrt{s} - \frac{1}{2}\right)\gamma a_s(\gamma) - \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right)^2 \gamma\sqrt{s} = 0. \tag{77}$$

The quadratic in $\gamma\, a_s(\gamma)$ occurring in the second proposition in (77) has the roots

$$\gamma a_s(\gamma) = -\frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(\gamma\sqrt{s} - \frac{1}{2}\right) \pm \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right)\left(\gamma\sqrt{s} + \frac{1}{2}\right). \tag{78}$$

Since $\gamma\, a_s(\gamma) > 0$, only the root in (78) with the $+$-sign needs to be considered. The latter root equals $\frac{1}{2}(1 - \gamma/\sqrt{s})$, and this completes the proof.

To show unimodality of L_s, we should consider the function $\gamma\, a_s(\gamma)/(1 - \gamma/\sqrt{s})$, $0 < \gamma < \sqrt{s}$. This function assumes the values 1 and 0 at $\gamma = 0+$ and $\gamma = \sqrt{s} - 0$, and so, by Proposition 3, it is sufficient to show that this function is strictly decreasing in $0 < \gamma < \sqrt{s}$. The result we show below is somewhat stronger.

Proposition 4. *It holds that* $\gamma\, a_s(\gamma)/(1 - \gamma/\sqrt{s})^2$ *decreases strictly in* $0 < \gamma < \sqrt{s}$ *when* $s > 1$.

Proof. We compute

$$\left(\left(1 - \frac{\gamma}{\sqrt{s}}\right)^{-2} \gamma a_s(\gamma)\right)'$$
$$= \left(1 - \frac{\gamma}{\sqrt{s}}\right)^{-2}\left[\frac{2}{\sqrt{s}}\left(1 - \frac{\gamma}{\sqrt{s}}\right)^{-1} \gamma\, a_s(\gamma) + a_s(\gamma) - \gamma\, a_s'(\gamma)\right].$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

Using (76), we, thus, see that for $0 < \gamma < \sqrt{s}$

$$\begin{aligned}
&\left(\left(1-\frac{\gamma}{\sqrt{s}}\right)^{-2}\gamma\, a_s(\gamma)\right)' < 0\\
&\quad\Longleftrightarrow \left(1-\gamma\, a_s(\gamma)-\frac{\gamma}{\sqrt{s}}\right)\left(\frac{2\gamma}{\sqrt{s}}+1-\frac{\gamma}{\sqrt{s}}\right)-\gamma\left(\gamma+\frac{1}{\sqrt{s}}\right)\left(1-\frac{\gamma}{\sqrt{s}}\right)<0\\
&\quad\Longleftrightarrow \left(1-\frac{\gamma}{\sqrt{s}}\right)(1-\gamma^2)-\gamma a_s(\gamma)\left(1+\frac{\gamma}{\sqrt{s}}\right)<0.
\end{aligned}$$

It is therefore sufficient to show that, for $0 < \gamma < \sqrt{s}$,

$$\gamma a_s(\gamma) > (1-\gamma^2)\,\frac{1-\gamma/\sqrt{s}}{1+\gamma/\sqrt{s}}. \tag{79}$$

We compute

$$\begin{aligned}
&\frac{1-\gamma/\sqrt{s}}{1+\gamma/\sqrt{s}}(1-\gamma^2)\\
&\qquad = 1-\frac{2\gamma}{\sqrt{s}}-\left(1-\frac{2}{s}\right)\gamma^2+2\left(1-\frac{1}{s}\right)\frac{\gamma^3}{\sqrt{s}}+O(\gamma^4), \qquad \gamma\downarrow 0,
\end{aligned} \tag{80}$$

and so, by (73), we see that (79) holds for small positive γ. Now suppose that γ, $0 < \gamma < \sqrt{s}$, is such that

$$\gamma a_s(\gamma) = (1-\gamma^2)\frac{1-\gamma/\sqrt{s}}{1+\gamma/\sqrt{s}}. \tag{81}$$

At such a γ we compute, using (76) and (80) twice,

$$\begin{aligned}
(\gamma a_s(\gamma))' &= a_s(\gamma)-\frac{\gamma\, a_s(\gamma)(\gamma+1/\sqrt{s})}{1-\gamma/\sqrt{s}-\gamma\, a_s(\gamma)}\\
&= a_s(\gamma)-\frac{1+\gamma/\sqrt{s}}{1-\gamma/\sqrt{s}}a_s(\gamma)\\
&= -\frac{2}{\sqrt{s}}\,\frac{1-\gamma^2}{1+\gamma/\sqrt{s}}.
\end{aligned}$$

At the same time, we compute

$$\left((1-\gamma^2)\frac{1-\gamma/\sqrt{s}}{1+\gamma/\sqrt{s}}\right)' = -\frac{2}{\sqrt{s}}\,\frac{1+\gamma\sqrt{s}-\gamma^2-\gamma^3/\sqrt{s}}{(1+\gamma/\sqrt{s})^2}.$$

Since $s > 1$, we have for $0 < \gamma < \sqrt{s}$

$$1+\gamma\sqrt{s}-\gamma^2-\frac{\gamma^3}{\sqrt{s}} > \left(1+\frac{\gamma}{\sqrt{s}}\right)(1-\gamma^2) = 1+\frac{\gamma}{\sqrt{s}}-\gamma^2-\frac{\gamma^3}{\sqrt{s}}.$$

Hence, at a $\gamma \in (0, \sqrt{s})$ where (81) holds, we have

$$(\gamma\, a_s(\gamma))' > \left((1-\gamma^2)\,\frac{1-\gamma/\sqrt{s}}{1+\gamma/\sqrt{s}}\right)'. \tag{82}$$

https://doi.org/10.1239/aap/1435236983 Published online by Cambridge University Press

This is in particular so for

$$\gamma_0 := \inf\{0 < \gamma < \sqrt{s} \mid (81) \text{ holds}\}. \tag{83}$$

This $\gamma_0 \in (0, \sqrt{s})$ since (79) holds for small positive γ and since we have assumed that there is a $\gamma \in (0, \sqrt{s})$ such that (81) holds. However, the validity of (81) and (82) for $\gamma = \gamma_0$ implies that

$$\gamma a_s(\gamma) < (1 - \gamma^2)\,\frac{1 - \gamma/\sqrt{s}}{1 + \gamma/\sqrt{s}}$$

holds for γs close to but less than γ_0. However, (79) holds for γs close to 0, and so there is, by continuity, a $\gamma_1 < \gamma_0$ such that (81) holds. Contradiction, see (83). This proves that (79) holds for all $\gamma \in (0, \sqrt{s})$, and the proof is complete.

We have from Proposition 4 that there is a unique root $\gamma = \hat{\gamma}_s \in (0, \sqrt{s})$ of

$$\gamma a_s(\gamma) = \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{s}}\right). \tag{84}$$

From the behavior of $L_s(\gamma)$ near $\gamma = 0$ and $\gamma = \sqrt{s}$, it, thus, follows that L_s is unimodal, with unique maximum at $\gamma = \hat{\gamma}_s$. The value of $L_s(\gamma)$ at $\gamma = \hat{\gamma}_s$, see (42), is easily obtained by inserting (84) into the definition of L_s in (41).

References

[1] Artalejo, J. R. and Gómez-Corral, A. (2008). *Retrial Queueing Systems*. Springer, Berlin.

[2] Avram, F., Janssen, A. J. E. M. and van Leeuwaarden, J. S. H. (2013). Loss systems with slow retrials in the Halfin–Whitt regime. *Adv. Appl. Prob.* **45,** 274–294.

[3] Borst, S., Mandelbaum, A. and Reiman, M. (2004). Dimensioning large call centers. *Operat. Res.* **52,** 17–34.

[4] Cohen, J. W. (1957). Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Rev.* **18,** 49–100.

[5] Falin, G. I. and Artalejo, J. R. (1995). Approximations for multiserver queues with balking/retrial discipline. *OR Spektrum* **17,** 239–244.

[6] Falin, G. I. and Templeton, J. G. C. (1997). *Retrial Queues*. Chapman & Hall, London.

[7] Gibbens, R. J., Hunt, P. J. and Kelly, F. P. (1990). Bistability in communication networks. In *Disorder in Physical Systems*. Oxford University Press, pp. 113–127.

[8] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

[9] Jagerman, D. L. (1974). Some properties of the Erlang loss function. *Bell System Tech. J.* **53,** 525–551.

[10] Janssen, A. J. E. M. and van Leeuwaarden, J. S. H. (2013). Staffing many-server systems with admission control and retrials (report version). Available at http://arxiv.org/abs/1302.3006.

[11] Janssen, A. J. E. M., van Leeuwaarden, J. S. H. and Zwart, B. (2008). Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. Appl. Prob.* **40,** 122–143.

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

[13] Nesenbergs, M. (1979). A hybrid of Erlang B and C formulas and its applications. *IEEE Trans. Commun.* **27,** 59–68.

[14] Zhang, B., van Leeuwaarden, J. S. H. and Zwart, B. (2012). Staffing call centers with impatient customers: refinements to many-server asymptotics. *Operat. Res.* **60,** 461–474.