OPERATIONS RESEARCH
Vol. 58, No. 5, September–October 2010, pp. 1398–1413
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5805 | 1398

informs

DOI 10.1287/opre.1100.0815

On the Accuracy of Fluid Models for Capacity Sizing in Queueing Systems with Impatient Customers

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, a-bassamboo@northwestern.edu

Ramandeep S. Randhawa
Marshall School of Business, University of Southern California, Los Angeles, California 90089, ramandeep.randhawa@marshall.usc.edu

We consider queueing systems in which customers arrive according to a Poisson process and have exponentially distributed service requirements. The customers are impatient and may abandon the system while waiting for service after a generally distributed amount of time. The system incurs customer-related costs that consist of waiting and abandonment penalty costs. We study capacity sizing in such systems to minimize the sum of the long-term average customer-related costs and capacity costs. We use fluid models to derive prescriptions that are asymptotically optimal for large customer arrival rates. Although these prescriptions are easy to characterize, they depend intricately upon the distribution of the customers' time to abandon and may prescribe operating in a regime with offered load (the ratio of the arrival rate to the capacity) greater than 1. In such cases, we demonstrate that the fluid prescription is optimal up to $O(1)$. That is, as the customer arrival rate increases, the optimality gap of the prescription remains bounded.

Subject classifications: queues: balking and reneging; approximations; limit theorems; capacity sizing; impatient customers; order-1 accuracy; fluid approximation.
Area of review: Stochastic Models.
History: Received May 2009; revision received November 2009; accepted December 2009. Published online in *Articles in Advance* August 17, 2010.

1. Introduction

Customers in queueing systems are typically impatient and may leave without obtaining service. A canonical example of such a system is a call center, where customers regularly hang up when they feel that they have waited long enough. This phenomenon of *abandonments* needs to be carefully incorporated into capacity-planning decisions in such systems, and this has motivated significant research in recent years. (See Gans et al. 2003 and Aksin et al. 2007 for a survey of such research in call centers.)

This paper addresses the classical management problem of selecting capacity in a queueing system with impatient customers. This capacity selection is to balance the cost of personnel (capacity) on the one hand, and customer-related costs on the other hand. The customer-related costs in such systems are twofold: there are costs associated with customer delays and also with customers who abandon, i.e., leave the system without obtaining service. This is a complex problem that is not amenable to exact optimization. The goal of this paper is to develop simple capacity prescriptions that are "extremely" accurate and study the impact of customer patience distribution on the firm's capacity decision.

We use approximations of the underlying system that treat incoming work as "fluid" that arrives at a constant rate and is processed deterministically with the installed capacity. These fluid approximations utilize only key characteristics of the arrival process, and the service and patience distributions. In particular, only the mean arrival rate and service time are utilized. However, these do require the entire patience distribution when the arrival rate exceeds the capacity of the system (this is referred to as the overloaded regime).

Extant literature that analyzes queueing systems with large arrival rates suggests that such fluid approximations should have an accuracy gap that increases with the arrival rate. The basis for this observation is the fact that the fluid model is derived by applying the strong law of large numbers and that it ignores the stochastic fluctuations suggested by the central limit theorem, which are on the order of the square root of the system scale. These observations have been made for critically loaded systems, which have an offered load approximately equal to 1. In the overloaded regime, we find that the stochastic fluctuations ignored by the fluid model are better characterized by the large deviations theory rather than the central limit theorem, which leads to considerably better performance of the fluid prescription. In this case, we prove that the accuracy gap of the fluid approximations for the expected steady-state queue length and the net rate of customer abandonments *does not* increase with the arrival rate. This $O(1)$-accuracy is particularly surprising as the performance metrics themselves increase with the arrival rate. Further, the accuracy gap of

the fluid approximation for the net rate of customer abandonment *decreases* with the arrival rate.

Using these "extremely accurate" fluid approximations, we compute the capacity level that best balances the tradeoff between the customer-related and capacity costs. We find that the key element of the customer patience distribution in determining the firm's capacity is its hazard rate. In fact, our capacity prescription is obtained by equating this hazard rate to a function of the cost parameters. For patience distributions with increasing hazard rates (and for the exponential distribution), we prove that it is asymptotically optimal for the firm to operate in the critically loaded regime with capacity installed approximately equal to the customer arrival rate. This formally provides support to the extensive performance analysis of this regime undertaken in the literature. However, when the patience distribution has a *decreasing* hazard rate, the firm benefits from reducing its capacity and operating in the overloaded regime. In this case, using the $O(1)$-accuracy of the fluid, we find that our simple fluid-based capacity prescription is in fact $O(1)$-optimal—the optimality gap of our prescription remains *bounded* with the arrival rate, even though the actual cost increases with the arrival rate.

1.1. Relevant Literature

A large body of literature has analyzed queueing systems with impatient customers. (See Boxma and de Waal 1994 for an overview of such models.) However, system design in the presence of abandoning customers is fairly recent. Garnett et al. (2002) is among the first to analyze systems with exponential patience distributions and uses the many-server asymptotic analysis introduced in Halfin and Whitt (1981) to study multiserver systems.

In this paper, we use the exact expressions for expected queue length and abandonment rate derived for $M/M/n+G$ systems in Mandelbaum and Zeltyn (2005) (which are based on work by Baccelli and Hebuterne 1981). We use these expressions to derive the fluid-based approximations and to study their accuracy. These fluid approximations are also derived in Mandelbaum and Zeltyn (2005) and (in a discrete-time framework) in Whitt (2006). However, our contribution is that we explicitly prove that these approximations are $O(1)$-accurate in the overloaded regime under some regularity assumptions on the customer patience distribution.

Our work relates to extant literature on capacity planning, including Maglaras and Zeevi (2003) and Borst et al. (2004), which study the optimal economic regime in systems without abandonments; and Harrison and Zeevi (2005) and Bassamboo et al. (2006, 2010), which study optimal capacity sizing in systems with parameter uncertainty (see Gans et al. 2003 and Aksin et al. 2007 for additional references). Our work also relates to Mandelbaum and Zeltyn (2009), which studies staffing in a system operating in a many-server configuration with general customer patience distributions with the objective of satisfying a quality-of-service constraint. Although this is a related problem, we focus on identifying the optimal economic regime when the objective is to minimize system costs rather than to meet a constraint. Another related paper is Ward and Glynn (2005), which studies a system operating in a single-server configuration with generally distributed interarrival and service requirements in addition to generally distributed patience times. In that paper, the authors study diffusion-based approximations for the expected steady-state queue length and net abandonment rate, and these can be used to refine our prescriptions in the single-server configuration.

1.2. Organization of the Paper

The next section begins with a description of the basic system model and the capacity-sizing problem. Section 3 then discusses the fluid approximations for the two performance measures of interest: expected steady-state queue length and net customer abandonment rate. We show that under some regularity conditions for the customer patience distribution, we obtain $O(1)$-accuracy of the fluid approximations. In §4, we use these fluid approximations to solve the capacity-sizing problem. We show that if the hazard rate of the patience distribution is decreasing (along with some regularity conditions), the fluid-based prescription yields an optimality gap that does not increase with the system size. In §5 we consider the single-server configuration, in which the capacity sizing involves selecting the service rate of the server. We show that this capacity-sizing problem is similar to that for the many-server configuration and that analogous results hold. We conclude in §6 by discussing the key findings of the paper and future research. All proofs are relegated to the appendices.

2. Model

We consider a system in which customers arrive according to a Poisson process with rate λ and have service requirements that are independent and exponentially distributed with unit mean. The customers are served in the order that they arrive to the system, and they wait in a queue if all servers are busy when they arrive. The customers are impatient and abandon the system after a random amount of time (which we refer to as their patience time), if their service has not commenced. This patience time is independent and identically distributed across customers, and we denote its distribution by G and the corresponding density function by g.

The system manager's objective is to select a capacity level to minimize the system cost, which is the sum of the capacity costs and customer-related costs. We consider two customer-related costs: (1) a delay cost of h incurred per customer per unit of time spent waiting for a server, and (2) a penalty cost of p incurred for each customer who

abandons the queue (does not obtain service). We consider two system configurations: (1) a many-server configuration with multiple servers, each operating at the fixed rate μ, and (2) a single-server configuration with a single server. In the many-server configuration, the capacity decision is to select the number of servers (each working at the fixed rate μ), whereas in the single-server configuration it is to select the service rate of the single server. We focus for the most part on the many-server configuration. The results for the single-server configuration are analogous and are discussed in §5.

Considering the many-server configuration, we denote the cost per server by c, the steady-state queue length and the net customer abandonment rate as a function of the number of servers n and the arrival rate λ by $Q_\lambda(n)$ and $\alpha_\lambda(n)$, respectively. Thus, the optimization problem is given by

$$\min_{n\in\mathbb{Z}_+}\{\Pi_\lambda(n) := cn + p\alpha_\lambda(n) + h\mathbb{E}Q_\lambda(n)\}. \tag{1}$$

For this $M/M/n+G$ system, Mandelbaum and Zeltyn (2005) derive exact expressions for the expected queue length and net abandonment rate (we reproduce the expressions in Appendix A). However, these expressions are fairly complicated and are not amenable to exact analysis. Instead, we use the fluid approximations for these performance measures to derive near-optimal prescriptions. These prescriptions are easy to derive and, as we will demonstrate, are extremely accurate. We will formally characterize the optimality gap of these prescriptions. To do so, we first study the fluid approximations for the expected steady-state queue length and the net abandonment rate and their accuracy, setting aside the capacity decision. Then, in §4, we formally consider the capacity-sizing problem and discuss the prescriptions that one obtains using the fluid approximation and their performance.

3. The Fluid Approximation

In this section, we study the fluid approximations for the expected steady-state queue length and the net abandonment rate. To do so, we study the system as the customer arrival rate λ increases without bound and the processing capacity of the system also increases proportional to λ. That is, we fix the offered load at $\rho > 0$ so that the number of servers in a system with arrival rate λ is $n_\lambda = \lambda/\rho$. (Note that the service and patience distributions do not scale with λ.) Section 4 proves that such a regime is indeed asymptotically optimal for high customer arrival rates. (The optimality gap will also be characterized there.) The single-server case is treated in §5. (In that case, we have $n = 1$, and the capacity of the single server increases with λ as $\mu_\lambda = \lambda/\rho$.)

Mandelbaum and Zeltyn (2005) derives the fluid approximations for the expected steady-state queue length ($\mathbb{E}Q_\lambda$) and net abandonment rate (α_λ). (For convenience we remove the explicit dependence of the performance measures on the number of servers.) Our focus will be to characterize the accuracy of these approximations. For this, it will be useful to understand the fluid approximation; therefore, we begin by providing an intuitive derivation of the fluid estimates. First, consider the net abandonment rate. In the underlying fluid model of the system, customers arrive at the fixed rate λ and get processed deterministically at the rate $n_\lambda\mu$. Thus, the net rate of abandonments considering the rate balance should equal $(\lambda - n_\lambda\mu)^+$. That is,

$$\alpha_\lambda \approx (\lambda - n_\lambda\mu)^+ = \lambda(1 - 1/\rho)^+.$$

Thus, the approximate fraction of abandonments is $(1-1/\rho)^+$. Denoting the mean waiting time in the queue by $\bar{w}$, we expect customers to abandon if their time to abandon is less than $\bar{w}$. That is, the fraction of customers abandoning equals $G(\bar{w})$, which gives us the following characterization of the mean waiting time:

$$G(\bar{w}) = \left(1 - \frac{1}{\rho}\right)^+. \tag{2}$$

To characterize the queue length, let us consider the customers in the queue. In this fluid model, the number of customers whose current time in the queue lies in the interval $[w, w+dw]$ for $w < \bar{w}$ equals $\lambda\bar{G}(w)dw$. (This relation holds as the number of customers arriving in the dw interval is λdw, and $G(w)$ fraction of these customers would have abandoned.) Summing up all these customers gives us the approximation

$$\mathbb{E}Q_\lambda \approx \lambda\int_0^{\bar{w}} \bar{G}(w)\,dw. \tag{3}$$

For convenience, we denote $\bar{q} = \int_0^{\bar{w}} \bar{G}(w)\,dw$. The following proposition then makes the fluid characterization precise:

Proposition 1 (Mandelbaum and Zeltyn 2005, Theorem 6.1). *As λ increases without bound, we obtain $\mathbb{E}Q_\lambda/\lambda \to \int_0^{\bar{w}} \bar{G}(w)\,dw = \bar{q}$, where $\bar{w}$ is given in* (2) *and $\alpha_\lambda/\lambda \to (1-1/\rho)^+$.*

This result implies that we can approximate $\mathbb{E}Q_\lambda \approx \lambda\bar{q}$ and $\alpha_\lambda \approx \lambda(1-1/\rho)^+$. Note that this approximation is also consistent with Whitt (2006), which develops fluid approximations for $GI/GI/n+G$ systems.

We next focus on the error in these approximations. It will be convenient to make the following assumption on the density of the customer patience distribution g:

Assumption 1. *For $\bar{w}$ that solves* (2)*, there exists some $\Delta > 0$ such that the density of the patience distribution g is strictly positive and continuously differentiable on $[\bar{w}-\Delta, \bar{w}+\Delta]$.*

Note that an implication of this assumption is that it ensures that (2) has a unique solution.

For expositional ease, it will be useful to define the following convention. For any two real-valued nonnegative functions f, g, we say that (1) $f(\lambda) = O(g(\lambda))$ if there exists a positive finite constant K such that $f(\lambda) \leqslant Kg(\lambda)$ for all $\lambda > 0$, (2) $f(\lambda) = o(g(\lambda))$ if $\limsup_{\lambda\to\infty} f(\lambda)/g(\lambda) = 0$, and (3) $f(\lambda) = \Theta(g(\lambda))$ if there exist positive finite constants K_1, K_2 such that $K_1 g(\lambda) \leqslant f(\lambda) \leqslant K_2 g(\lambda)$ for all $\lambda > 0$.

Table 1. Comparison of the fluid approximation for the expected steady-state queue-length $\lambda\bar{q}$ with the actual value $\mathbb{E}Q_\lambda$ computed using exact formulas for exponential, Pareto, and hyperexponential patience distributions.

	$\rho = 1.1$			$\rho = 1.2$			$\rho = 1.5$		
λ	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$
				Exponential distribution					
25	3.30	2.27	1.03	4.69	4.17	0.52	8.40	8.33	0.07
50	5.62	4.55	1.07	8.71	8.33	0.37	16.68	16.67	0.01
100	10.04	9.09	0.94	16.84	16.67	0.17	33.33	33.33	4.42×10^{-4}
200	18.81	18.18	0.62	33.37	33.33	0.03	66.67	66.67	6.25×10^{-7}
				Pareto distribution					
25	1.99	1.16	0.82	2.76	2.18	0.58	4.91	4.59	0.32
50	3.26	2.33	0.93	4.91	4.36	0.56	9.45	9.18	0.28
100	5.61	4.65	0.96	9.17	8.71	0.46	18.60	18.35	0.25
200	10.16	9.31	0.85	17.76	17.43	0.34	36.95	36.70	0.25
				Hyperexponential distribution					
25	1.68	0.93	0.75	2.33	1.75	0.59	4.17	3.72	0.45
50	2.72	1.86	0.85	4.07	3.49	0.58	7.86	7.44	0.42
100	4.62	3.72	0.90	7.49	6.98	0.51	15.27	14.87	0.40
200	8.28	7.45	0.83	14.36	13.96	0.40	30.14	29.74	0.39

Note. The approximation error does not increase with the customer arrival rate (system size).

3.1. Accuracy of the Fluid Approximation for the Expected Queue Length

We begin by conducting a numerical study to examine the error in the fluid approximation. We consider three patience distributions with unit mean: (1) exponential distribution; (2) Pareto distribution with shape parameter 2—i.e., $\bar{G}(x) = 1/(1+x)^2$ for $x \geqslant 0$; and (3) hyperexponential distribution that is distributed as exponential with mean 1/4 with probability 4/7, and exponential with mean 2 with probability 3/7. We set the service distribution to be a unit mean exponential distribution and the capacity of each server $\mu = 1$. Table 1 displays the numerically computed expected steady-state queue length using the formulas in Appendix A, the fluid approximation $\lambda\bar{q}$, and the error in the fluid approximation $|\mathbb{E}Q_\lambda - \lambda\bar{q}|$ for different values of arrival rate λ and offered load ρ. Observe that as the arrival rate increases, the error of the approximation *does not increase*. In fact, the maximum error observed is 1.07. Note that as the offered load ρ increases, the accuracy of the prescription improves further. The performance of the approximation for the exponential distribution is even better. As the arrival rate increases, the error in the approximation seems to *decrease*.

These observations are in stark contrast to the fact that the error in fluid approximations typically increases with the arrival rate. In fact, if we focus on the critically loaded case $\rho = 1$, then we do obtain an error that is proportional to the square root of the arrival rate. Table 2 displays the results for this case. Here, the fluid approximation is precisely zero, and hence the error in the approximation is equal to the expected queue length. For this critically loaded setting, we can further refine the fluid approximation (which is based on functional strong law) using the functional central limit theorem, which measures deviations on the square root level (see Garnett et al. 2002 for details). Thus, the gap between the actual performance and fluid approximation grows as the square root of the system size. However, when the system is overloaded with $\rho > 1$, the fluid approximation is nontrivial, and in this setting the deviations are better explained using large deviations theory rather than the central limit theorem.

To understand this result for the overloaded regime, consider the exponential patience distribution with unit mean, and set $\mu = 1$. In this case, the number of customers in the system in steady state has a Poisson distribution with mean λ. (Because the mean patience and service times are identical, the number of customers in the system is identical to that in an $M/M/\infty$ queue with arrival rate λ and unit

Table 2. The expected steady-state queue-length $\mathbb{E}Q_\lambda$ computed using exact formulas for exponential, Pareto, and hyperexponential patience distributions for the critically loaded case with $\rho = 1$.

	$\rho = 1$		
	Exponential	Pareto	Hyperexponential
λ	$\mathbb{E}Q_\lambda$	$\mathbb{E}Q_\lambda$	$\mathbb{E}Q_\lambda$
25	1.99	1.25	1.06
50	2.82	1.73	1.47
100	3.99	2.42	2.04
200	5.64	3.39	2.85

Notes. The fluid approximation is zero in this case, and thus the accuracy error equals $\mathbb{E}Q_\lambda$. The approximation error increases proportional to the square root of the arrival rate λ.

service rate.) Thus, we have

$$\mathbb{E}Q_\lambda = \mathbb{E}[Poisson(\lambda) - n_\lambda]^+,$$

where $Poisson(\lambda)$ is a Poisson random variable with mean λ. For large λ, we can approximate the Poisson distribution by a normal distribution with mean λ and standard deviation $\sqrt{\lambda}$ to obtain

$$\begin{aligned}\mathbb{E}Q_\lambda &\approx \mathbb{E}[Normal(\lambda,\lambda) - n_\lambda]^+ \\ &= (\lambda - n_\lambda)^+ + K_1\sqrt{\lambda}\exp(-K_2(\lambda - n_\lambda)^2) \\ &= \lambda(1-1/\rho)^+ + K_1\sqrt{\lambda}\exp(-K_2\lambda^2(1-1/\rho)^2),\end{aligned}$$

for some finite positive constants K_1 and K_2. If the number of servers is such that $\rho = 1$, then we obtain

$$\mathbb{E}Q_\lambda \approx K_1\sqrt{\lambda},$$

and the error in the fluid approximation increases with system size. If the number of servers is such that $\rho > 1$, then the expected queue length is

$$\mathbb{E}Q_\lambda \approx \lambda(1-1/\rho)^+ + o(1).$$

Thus, in this case, the error in the fluid approximation does not increase with system size. The following result formalizes this intuition and proves that if Assumption 1 holds, the fluid approximation is accurate up to a constant. (The exponential patience distribution leads to an even better performance of the fluid approximation, and this is described in Proposition 2.)

THEOREM 1 ($O(1)$-ACCURACY). *If the system is overloaded ($\rho > 1$) and Assumption 1 holds, then the fluid approximation for the expected queue length is accurate up to $O(1)$. That is,*

$$|\mathbb{E}Q_\lambda - \lambda\bar{q}| \leqslant C$$

for some finite constant $C > 0$ and all $\lambda > 0$.

This result says that if the density of the patience distribution possesses some regularity conditions in the vicinity of $\bar{w}$, then the fluid approximation is $O(1)$-accurate, i.e., the approximation error does not grow with system size. This is in contrast with the critically loaded case of $\rho = 1$, for which, even when Assumption 1 holds, the fluid approximation is accurate only up to $O(\sqrt{\lambda})$. In fact, if further regularity conditions hold, we can get even better performance for the overloaded regime. On the other hand, if Assumption 1 does not hold, we can obtain errors that increase with system size even for the overloaded regime. The following discussion provides an illustration of such extremely accurate and poor performances.

An Example of Approximation Error Decreasing with System Size. Theorem 1 proves that the fluid approximation has an $O(1)$ error for patience distributions that satisfy Assumption 1. The following result proves that for the exponential patience distribution, the approximation is accurate up to $o(1)$, i.e., the error *decreases* in the arrival rate.

PROPOSITION 2 ($o(1)$-ACCURACY OF THE FLUID APPROXIMATION). *If the patience distribution is exponential, the error in the fluid approximation decreases as the system size increases, that is, $|\mathbb{E}Q_\lambda - \lambda\bar{q}| \to 0$ as $\lambda \to \infty$.*

Examples of Arbitrarily High Approximation Error. We now identify cases in which the approximation is not very accurate. In particular, if the patience distribution is not well behaved in the vicinity of $\bar{w}$, the fluid prescription may not be very accurate. The accuracy of the fluid approximation depends on the rate of convergence of the expected queue length to the fluid limit (given by (3)). In fact, for continuous distributions, we can rewrite (2) as $\bar{w} = G^{-1}(1-1/\rho)$. Note that the actual (exact) offered wait also involves computing the inverse of the patience distribution. Therefore, the accuracy of the fluid approximation is related to the precision with which this inverse can be calculated in the vicinity of $G(\bar{w})$. This precision is related to the difference $G(x) - G(\bar{w})$ for x in the vicinity of $\bar{w}$. In particular, the higher this gap, the closer $\bar{w}$ is to the exact offered wait, and thus the faster the convergence.

Using a Taylor's series expansion, and assuming sufficient differentiability of G, we obtain

$$\begin{aligned}G(x) - G(\bar{w}) \approx g(\bar{w})(x-\bar{w}) &+ \sum_{i=2}^{n}\frac{g^{(i)}(\bar{w})(x-\bar{w})^i}{i!} \\ &+ g^{(n+1)}(\psi(x))\frac{(x-\bar{w})^{n+1}}{(n+1)!}\end{aligned}$$

for $\psi(x)$ between x and $\bar{w}$, where $g^{(k)}$ denotes the k^{th} derivative of g. For $g(\bar{w}) > 0$ (as required by Assumption 1), this gap is linear. However, if $g(\bar{w}) = 0$, and if higher derivatives of the density are zero at $\bar{w}$ as well, this gap decreases, and thus reduces the rate of convergence and the accuracy of the fluid approximation. The following example demonstrates that this phenomenon can lead to the arbitrarily poor accuracy of $\Theta(\lambda^{1-\epsilon})$ for arbitrarily small $\epsilon > 0$.

PROPOSITION 3 ($\Theta(\lambda^{1-\epsilon})$-ACCURACY OF FLUID APPROXIMATION). *For the patience distribution with density $g(x) = 1$ for $x < 1/2$, $g(x) = 2^m(m+1)(x-1/2)^m$ for $1/2 \leqslant x \leqslant 1$ and $m > 0$, and $g(x) = 0$ for $x > 1$, when $\rho = 2$, the error in the fluid approximation $|\mathbb{E}Q_\lambda - \lambda\bar{q}| = \Theta(\lambda^{(m+1)/(m+2)})$. That is, for any $0 < \epsilon < 1/2$, there exists $m > 0$ such that the error in the fluid approximation for this patience distribution is $\Theta(\lambda^{1-\epsilon})$.*

A similar observation has been made for the critically loaded regime in Mandelbaum and Zeltyn (2005, Internet Supplement, Theorem 6.2).

To conclude the discussion on the accuracy of the fluid approximation for the queue length, we would like to point

out that although it may seem that the fluid approximation is a lower bound for $\mathbb{E}Q_\lambda$ (as per observations in Table 1), this is not true in general, as demonstrated in the following remark.

Remark 1 (Fluid Approximation Is Not Always a Lower Bound for the Expected Queue Length). In many systems, the fluid model provides a lower bound to the system's performance. For instance, in queueing systems without abandonments, the fluid limit for the expected queue length equals zero; see Whitt (2002). However, in the $M/M/n+G$ system the fluid approximation $\lambda\bar{q}$ need *not* be a lower bound for the expected queue length. This is demonstrated in Table 3, where the expected queue length is smaller than the fluid approximation for the Erlang distribution. This observation is consistent with the simulation results in Whitt (2006).

3.2. Accuracy of the Fluid Approximation for the Net Rate of Abandonment

We now turn to the fluid approximation for the net rate of customer abandonment, which is given by $\lambda(1-1/\rho)$. The following result proves that this approximation is even more accurate than that for the expected steady-state queue length. In fact, this approximation is accurate up to $o(1)$, and its error decreases with the arrival rate.

Theorem 2. *If the system is overloaded with $\rho > 1$, then the fluid approximation for the net rate of customer abandonment is accurate up to $o(1)$. In particular,*

$$\lambda(1-1/\rho) \leqslant \alpha_\lambda \leqslant \lambda(1-1/\rho) + C_1 e^{-C_2\lambda}$$

for finite constants $C_1, C_2 > 0$.

The results show that the fluid approximation is extremely accurate for the net rate of customer abandonment. In fact, the error decays exponentially in the arrival rate. Note that the only regularity condition that is needed for this result to hold is that the patience distribution has a density.

Table 4 numerically demonstrates the accuracy of the fluid approximations for the cases described in §3.1. We see that these approximations are indeed very accurate when the offered load exceeds 1.

We end this section by making two observations.

Remark 2. The fluid approximation for the net abandonment rate indeed serves as a lower bound. This is intuitive because $\lambda(1-1/\rho)$ measures the net rate of abandonment when the servers work all the time, and thus, accounting for idle time, the actual number will only be greater. This should be contrasted with our observation in Remark 1, where we noted that the fluid approximation is not necessarily a lower bound for the expected queue length. In the latter case, due to abandonments, the queue length may increase or decrease when one accounts for the idle time.

Remark 3 (Special Properties of $M/M/n+M$ Systems). When the patience distribution is exponential, the expected queue length can be expressed in terms of the net rate of customer abandonment. In particular, we have $\alpha_\lambda = \gamma\mathbb{E}Q_\lambda$, where $1/\gamma$ is the mean patience time. Combining this observation with Theorem 2, we obtain that the fluid approximation for the expected queue length in these systems (1) is a lower bound and (2) has an accuracy of $o(1)$ (the error in fact decreases exponentially fast). However, these properties need not hold for general patience distributions, as noted earlier.

4. Capacity Sizing to Minimize Costs

In this section, we use the fluid approximations from the previous section to develop a near-optimal solution to the capacity-sizing problem. We first use the approximation to obtain a capacity prescription. Then, we study the optimality gap for the fluid-based prescription. Recall that the objective is to choose a capacity level that minimizes the sum of personnel cost and customer-related cost. Thus, the optimization problem is

$$\min_{n\in\mathbb{Z}_+}\{\Pi_\lambda(n) = cn + p\alpha_\lambda(n) + h\mathbb{E}Q_\lambda(n)\}. \tag{4}$$

We denote the optimal solution to this problem by $\Pi^\star_\lambda$. We use the fluid approximations for these performance metrics derived in the previous section. For convenience and to rule out pathological cases, we make the following assumption in this section, which encompasses Assumption 1:

Assumption 2. *The density of the patience distribution g is differentiable and strictly positive on $[0,\infty)$.*

We first derive the fluid-based prescription, and then we discuss its optimality gap. To compute the prescrip-

Table 3. Accuracy of the fluid approximation for the Erlang patience distribution with mean 1 and shape parameter 2.

	$\rho = 1.1$			$\rho = 1.2$			$\rho = 1.5$		
λ	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$
					Erlang distribution				
25	5.78	6.07	−0.29	7.93	8.57	−0.64	12.45	12.86	−0.41
50	11.16	12.14	−0.98	16.15	17.13	−0.98	25.32	25.72	−0.40
100	22.51	24.29	−1.78	33.19	34.26	−1.07	51.07	51.44	−0.37
200	46.36	48.58	−2.22	67.52	68.53	−1.01	102.51	102.88	−0.36

Note. The fluid approximation is not a lower bound for the expected queue length.

Table 4. Comparison of the fluid approximation for the net rate of abandonment $\lambda\bar{\alpha} = \lambda(1 - 1/\rho)$ with the actual value α_λ computed using exact formulas for exponential, Pareto, and hyperexponential patience distributions; the fluid approximation is identical for all distributions.

	$\rho = 1.1$			$\rho = 1.2$			$\rho = 1.5$		
λ	α_λ	$\lambda\bar{\alpha}$	$\alpha_\lambda - \lambda\bar{\alpha}$	α_λ	$\lambda\bar{\alpha}$	$\alpha_\lambda - \lambda\bar{\alpha}$	α_λ	$\lambda\bar{\alpha}$	$\alpha_\lambda - \lambda\bar{\alpha}$
				Exponential distribution					
25	3.30	2.27	1.03	4.69	4.17	0.52	8.40	8.33	0.07
50	5.62	4.55	1.07	8.71	8.33	0.37	16.68	16.67	0.01
100	10.04	9.09	0.94	16.84	16.67	0.17	33.33	33.33	4.4×10^{-4}
200	18.81	18.18	0.62	33.37	33.33	0.03	66.67	66.67	6.3×10^{-7}
				Pareto distribution					
25	3.61	2.27	1.34	4.96	4.17	0.79	8.52	8.33	0.18
50	6.06	4.55	1.51	9.03	8.33	0.69	16.74	16.67	0.07
100	10.61	9.09	1.52	17.13	16.67	0.46	33.34	33.33	0.01
200	19.46	18.18	1.27	33.53	33.33	0.19	66.67	66.67	2.8×10^{-4}
				Hyperexponential distribution					
25	3.71	2.27	1.44	5.05	4.17	0.88	8.56	8.33	0.23
50	6.21	4.55	1.66	9.15	8.33	0.81	16.77	16.67	0.10
100	10.82	9.09	1.73	17.26	16.67	0.59	33.36	33.33	0.02
200	19.71	18.18	1.52	33.62	33.33	0.29	66.67	66.67	1.1×10^{-3}

Note. The approximation error decreases with the customer arrival rate λ.

tion, as in the previous section, we set the capacity level to $n_\lambda = \lambda/\rho\mu$, where $\rho > 0$ is the offered load (which does not depend on λ). Then, we obtain $\alpha_\lambda(n_\lambda) \approx [\lambda - n_\lambda\mu]^+$ and $\mathbb{E}Q_\lambda(n_\lambda) \approx \lambda\bar{q} = \lambda\int_0^{\bar{w}} \bar{G}(u)\,du$, where $\bar{w} = G^{-1}((1 - n_\lambda\mu/\lambda)^+) = G^{-1}((1 - 1/\rho)^+)$. This leads us to the following optimization problem for the underlying fluid model:

$$\min_{n\in\mathbb{Z}_+} cn + p[\lambda - n\mu]^+ + h\lambda\int_0^{\bar{w}} \bar{G}(u)\,du \tag{5}$$

$$= \lambda \min_{\rho\geqslant 0}\left\{\frac{c}{\mu\rho} + p\left[1 - \frac{1}{\rho}\right]^+ + h\int_0^{G^{-1}((1-1/\rho)^+)} \bar{G}(u)\,du\right\}. \tag{6}$$

We denote an optimizer of (6) by $\rho^\star$. It is easy to see that $\rho^\star \geqslant 1$ because when $\rho^\star = 1$, the queueing and abandonment costs are zero, and thus increasing the capacity further (which is equivalent to decreasing ρ) will only increase the capacity cost without lowering customer-related costs. Thus, the optimization problem reduces to

$$\min_{\rho\geqslant 1} \frac{c}{\mu\rho} + p(1 - 1/\rho) + h\int_0^{G^{-1}(1-1/\rho)} \bar{G}(u)\,du. \tag{7}$$

To obtain a simpler form, we will equivalently optimize on the parameter $w = G^{-1}(1 - 1/\rho)$ to obtain

$$\min_{w\geqslant 0}\left\{\bar{\Pi}(w) := p + \left(\frac{c}{\mu} - p\right)\bar{G}(w) + h\int_0^w \bar{G}(u)\,du\right\}. \tag{8}$$

As the fluid model provides an accurate approximation of the actual performance metrics, we expect the corresponding prescription to perform quite well. This is formalized in the following result.

THEOREM 3. *If Assumption* 2 *holds and* (8) *has a unique solution* $w^\star$, *then*

1. *The capacity level* $n_\lambda^\star = \lambda\bar{G}(w^\star)/\mu$ *is asymptotically optimal in the sense that*

$$\frac{\Pi_\lambda^\star}{\Pi_\lambda(n_\lambda^\star)} \to 1, \text{ as } \lambda \to \infty.$$

2. *Further, if* $w^\star > 0$, *i.e.,* $n_\lambda^\star < \lambda/\mu$*—then the capacity level* $\lambda^\star$ *is* $O(1)$*-optimal, i.e.,*

$$\Pi_\lambda(n_\lambda^\star) \leqslant \Pi_\lambda^\star + C$$

for some constant $C > 0$ *and all* $\lambda > 0$.

Note that if (8) has multiple optima, then the optimality gap will be $O(1)$ if all the minimizers are strictly positive.

Characterizing the Fluid-Based Prescription. We now take a closer look at the optimization problem (8) to identify the optimizer $w^\star$. Note that we have

$$\bar{\Pi}'(w) = h\bar{G}(w)\left[1 - \frac{c/\mu - p}{h}h_a(w)\right], \tag{9}$$

where h_a is the hazard rate function, and the first-order optimality condition is given by:

$$h_a(w) = \frac{g(w)}{\bar{G}(w)} = \frac{h}{c/\mu - p}. \tag{10}$$

Thus, the optimal solution $w^\star$ either solves (10) or is a corner solution, i.e., $w^\star = 0$ or ∞. The case $w^\star = 0$ corresponds to the critically loaded regime with capacity equal to the arrival rate. The case $w^\star > 0$ corresponds on the

overloaded regime where the capacity is set lower than the arrival rate. Note that the case $w^\star = \infty$ corresponds to the trivial case in which the customers have an infinite offered wait, or equivalently there is zero investment in capacity. We next provide sufficient conditions for the overloaded and critically loaded regime to be asymptotically optimal.

Asymptotic Optimality of the Overloaded Regime. For patience distributions with monotone decreasing hazard rates, using (9), we note that $\overline{\Pi}$ is quasi-convex. This implies that if there is a strictly positive solution $w^\star$ to (10), then $\overline{\Pi}$ is minimized at $w^\star$. Thus, we obtain a sufficient condition for the asymptotic optimality of the overloaded regime.

PROPOSITION 4. *If Assumption* 2 *holds, the hazard rate of the customer patience distribution is monotone decreasing, and there exists a solution* $w^\star > 0$ *to* (10); *then the overloaded regime with capacity* $n_\lambda^\star = \lambda\bar{G}(w^\star)/\mu$ *is* $O(1)$*-optimal. That is,* $\Pi_\lambda(n_\lambda^\star) \leqslant \Pi_\lambda^\star + C$ *for some constant* $C > 0$ *and all* $\lambda > 0$.

This proposition covers patience distributions such as the Weibull distribution with shape parameter less than one, and the Pareto distribution. As an illustration, let us consider the Pareto distribution with $\bar{G}(x) = 1/(1+x)^2$ for $x \geqslant 0$. The hazard rate of this distribution $h_a(x) = 2/(1+x)$ is decreasing. If the problem parameters are such that $h/(c/\mu - p) < 2$, then the unique solution to (10) is $w^\star = 2((c/\mu - p)/h) - 1$. Thus, applying Proposition 4, we obtain that the capacity prescription $n_\lambda^\star = \lambda\bar{G}(w^\star)/\mu$ is $O(1)$-optimal.

Asymptotic Optimality of the Critically Loaded Regime. Analogous to the previous case, we note that $\overline{\Pi}$ is quasi-concave for patience distributions with monotone increasing hazard rates. Thus, for such distributions, the optimization problem (8) has a corner solution. Alternatively, if capacity costs are sufficiently low, one expects the fluid problem to prescribe ample capacity so that $w^\star$ is zero. The following result formalizes these findings into sufficient conditions for the critically loaded regime to be asymptotically optimal.

PROPOSITION 5. *If Assumption* 2 *and either of the following conditions holds, then we have the following*:

(a) *the hazard rate of the patience distribution is monotone increasing and* $c/\mu < p + h/\gamma$;

(b) *capacity is sufficiently inexpensive,* $c/\mu < p$,

then the critically loaded regime with capacity $n_\lambda^\star = \lambda/\mu$ *is asymptotically optimal. That is,* $\Pi_\lambda^\star/\Pi_\lambda(n_\lambda^\star) \to 1$, *as* $\lambda \to \infty$.

Thus, the critically loaded regime is asymptotically optimal for patience distributions with increasing hazard rates such as the normal and uniform distributions. The exponential distribution is a special case with a constant hazard rate. However, in this case it is easy to see that the optimal solution is critically loaded, and the fluid prescription would have an optimality gap that increases with the arrival rate. It is interesting to note that if there is high uncertainty in the arrival rate at the time of capacity selection, then Bassamboo et al. (2010) demonstrate that the corresponding fluid prescription results in $O(1)$-optimality.

We next use a numerical study to demonstrate the accuracy of the fluid prescription.

4.1. A Numerical Study

We consider systems with arrival rates $\lambda = 25, 50, 100,$ and 200, service rate $\mu = 1$, and mean patience time $1/\gamma = 1$. The cost parameters are as follows: the cost of a server $c = 1$ per server per unit time; holding cost $h = 1$ per customer per unit time; and penalty cost $p = 0.45$ per abandoned customer. For the patience distributions, we consider the same three unit mean distributions that we considered in §3.1: (1) $M(1)$: exponential distribution; (2) Pareto(2): Pareto distribution with shape parameter 2, i.e., $\bar{G}(x) = 1/(1+x)^2$ for $x \geqslant 0$; and (3) $H_2(1)$: a hyperexponential distribution that is distributed as exponential with mean 1/4 with probability 4/7, and exponential with mean 2 with probability 3/7.

For each patience distribution, we compute the fluid prescription by solving (8) and setting the capacity $\bar{n}_\lambda = \bar{G}(w^\star)\lambda/\mu$. We then compare the performance of this prescription with the numerically computed optimal values. The results are displayed in Table 5. For the exponential distribution, the fluid prescription sets the offered load $\rho^\star = 1$; for the Pareto distribution, it sets the offered load $\rho^\star = 1.2$; whereas for the hyperexponential distribution, it prescribes $\rho^\star = 1.67$. Thus, noting that Assumption 1 holds for the Pareto and hyperexponential distributions, we obtain $O(1)$-optimality of these prescriptions. Indeed, we observe that the optimality gap of the prescription in absolute terms does not increase with the arrival rate for these distributions (see Table 5). For the exponential distribution, the optimality gap does grow with the system size (proportional to $\sqrt{\lambda}$); however, even in this case the error as a fraction of the true cost is not significant. Thus, Table 5 demonstrates that when the fluid prescription leads to an overloaded regime, its performance is excellent, and even when the prescription leads to the critically loaded regime, the performance is very good.

5. The Single-Server Configuration

In this section, we discuss the capacity-sizing problem when the system operates in a single-server configuration. As before, we first consider the fluid approximations for the performance measures. For this, we study the system as the arrival rate increases without bound and the capacity of the server μ_λ increases with the arrival rate λ as $\mu_\lambda = \lambda/\rho$ for fixed offered load $\rho > 0$. It is easy to see that the underlying fluid model for the single-server configuration is the

Table 5. Performance of the fluid prescription compared with the numerically computed optimal.

Arrival rate	Optimal		Prescription		Optimality gap	
λ	Capacity	Cost	Capacity	Cost	Absolute	Percentage (%)
			Exponential distribution			
25	22	27.51	25	27.88	0.37	1.34
50	46	53.57	50	54.08	0.51	0.95
100	95	105.07	100	105.78	0.71	0.67
200	193	207.19	200	208.18	0.99	0.48
			Pareto distribution			
25	17	25.41	20	25.67	0.26	1.02
50	36	50.26	41	50.55	0.29	0.58
100	76	99.97	82	100.12	0.16	0.15
200	160	199.43	165	199.48	0.05	0.03
			Hyperexponential distribution			
25	15	24.63	15	24.63	0	0
50	31	48.68	30	48.69	0.01	0.02
100	62	96.87	61	96.88	0.01	0.01
200	124	193.31	122	193.32	0.01	0.004

Notes. For the exponential distribution, although the error is theoretically increasing with system size, the performance is still very good. For the Pareto and hyperexponential distributions, the results demonstrate the $O(1)$-optimality proved in Theorem 3.

same as that for the many-server configuration. That is, the net rate of customer abandonment $\alpha_\lambda \approx \lambda(1-1/\rho)^+$, and the expected steady-state queue length $\mathbb{E}Q_\lambda \approx \lambda\bar{q} = \lambda\int_0^{\bar{w}} \bar{G}(w)dw$, where w solves $G(\bar{w}) = (1-1/\rho)^+$.

Analogous to Theorems 1 and 2 for the many-server configuration, we obtain the following $O(1)$-accuracy of the fluid approximation in the single-server configuration as well.

Theorem 4 (Accuracy of Fluid Approximations for the Single-Server Configuration). *Suppose that the system is overloaded* ($\rho > 1$); *then,*

1. *If Assumption* 1 *holds, then the fluid approximation for the expected queue length is* $O(1)$*-accurate. That is,* $|\mathbb{E}Q_\lambda - \lambda\bar{q}| \leqslant C$ *for* $\lambda > 0$ *and some finite constant* $C > 0$.

2. *The fluid approximation for the net abandonment rate is accurate up to* $o(1)$. *In particular, we have* $\lambda(1-1/\rho) \leqslant \alpha_\lambda \leqslant \lambda(1-1/\rho) + C_1 e^{-C_2\lambda}$ *for finite constants* $C_1, C_2 > 0$.

We now turn to the capacity-sizing problem. The cost structure is the same as that for the many-server configuration, with the exception that we now denote the capacity cost per unit rate of capacity by $\bar{c}$. Denoting the steady-state queue length and the net customer abandonment rate as a function of the capacity μ and arrival rate λ by $Q_\lambda(\mu)$ and $\alpha_\lambda(\mu)$, respectively, the optimization problem is

$$\min_{\mu\geqslant 0}\{\Pi_\lambda(\mu) := \bar{c}\mu + p\alpha_\lambda(\mu) + h\mathbb{E}Q_\lambda(\mu)\}. \tag{11}$$

Comparing this optimization problem with that for the many-server configuration (4), we find the only difference is that in the many-server case the capacity cost was incurred per server, whereas here it is incurred for each unit of service rate. Thus, the analysis in §4 follows unaltered if we replace the quantity c/μ in that section (in (6)) with the unit capacity rate $\bar{c}$ for the single-server case. Based on Theorem 4, we then obtain exactly same result as in Theorem 3. For brevity, we do not reproduce the result.

6. Discussion

This paper studies the capacity-sizing problem in queueing systems with Poisson arrivals, exponentially distributed service times, and impatient customers with generally distributed time to abandon. We consider the problem of selecting the number of servers in many-server systems and the service rate in single-server systems. We propose fluid-based capacity prescriptions (that are identical for both systems). Unlike models without abandonments, the systems we study here are self-stabilizing, and hence the fluid-based prescription is implementable without further refinements. This prescription depends intricately on the entire customer patience distribution (and not just on its mean). Further, it does not always lead the system into a critically loaded regime. We identify conditions on the hazard rate of the patience distribution for which the fluid model prescribes an overloaded regime. In such cases, we explicitly prove that this prescription is $O(1)$-optimal, i.e., the optimality gap of the prescription does not increase with system size. This should be contrasted with the typical error of fluid prescriptions that increases with system size.

We perform numerical studies for exponential, Pareto, and hyperexponential patience distributions, and our results demonstrate the $O(1)$-optimality for the Pareto and hyperexponential patience distributions. For the exponential patience distribution, which always has a critically loaded fluid prescription even though the prescription is not $O(1)$-optimal, we obtain very good performance. This suggests that even when the system is critically loaded, the fluid prescription can provide a good approximation to the true optimal.

Our results demonstrate that it may be economical for a system to operate overloaded. This regime is conceptually similar to the efficiency-driven regime in queueing systems without abandonments, which is heavier than the conventional square-root regime. In this spirit, the observations in this paper are similar to those in Kumar and Randhawa (2010), where the authors study queueing systems without abandonments and demonstrate that for convex delay costs, the system is "very heavily" loaded, with the system utilization rapidly approaching one as the system size grows.

We would also like to point out that although we studied many-server and single-server configurations in this paper, our analysis is equivalently applicable to a capacity selection problem in which the number of servers is

fixed up-front and the optimal service rate must be selected to minimize system costs. This equivalence is a result of the fact that at the fluid scale, the performance estimates depend upon the total processing rate, rather than on the division of capacity.

In this paper, we focus on the case of exponential interarrival and service times. Based on recent work by Whitt (2006), we expect the fluid-based capacity prescriptions to remain unchanged even for generally distributed interarrival and service time. However, formally analyzing this system and proving the $O(1)$-optimality is a subject for future work. A related generalization possible is to use the extremely accurate fluid approximations to solve capacity sizing and routing problems in multiclass queueing systems with multiple server pools. Another avenue for future work would be to incorporate delay announcements (as in Armony et al. 2009) and study the implications for capacity sizing.

Organization of the Appendix

Appendix A provides a summary of the exact expressions for the expected steady-state queue length and net customer abandonment rate. Then, Appendix B contains a generalization of Theorems 1 and 2, along with its proof. Finally, Appendix C contains a proof of all the results presented in the paper.

Appendix A: Exact Expressions for Expected Queue Length and Waiting Time in an $M/M/n+G$ System

Mandelbaum and Zeltyn (2005, Internet Supplement, pp. 2–3) obtains the following exact expressions for the expected steady-state queue length, offered wait, and overall abandonment rate in an $M/M/n+G$ system:

$$\mathbb{E}Q_\lambda(n) = \frac{\lambda^2 J_{H,\lambda}}{\mathscr{E}_\lambda + \lambda J_\lambda} \tag{12}$$

$$\alpha_\lambda(n) = \lambda\left(\frac{1+(\lambda-n\mu)J_\lambda}{\mathscr{E}_\lambda + \lambda J_\lambda}\right), \tag{13}$$

where defining $H(x) = \int_0^x \bar{G}(y)\,dy$, we have

$$\mathscr{E}_\lambda(n) = \int_0^\infty e^{-t}\left(1+t\frac{\mu}{\lambda}\right)^{n-1} dt \tag{14}$$

$$J_\lambda(n) = \int_0^\infty \exp\left(\lambda H(x) - n\mu x\right) dx \tag{15}$$

$$J_{H,\lambda}(n) = \int_0^\infty H(x)\exp\left(\lambda H(x) - n\mu x\right) dx. \tag{16}$$

Appendix B: Generalization of Theorems 1 and 2 with Proof

We present a general version of Theorems 1 and 2 for a sequence of systems in which the offered load may vary with the arrival rate. In particular, consider a sequence of $M/M/n+G$ systems indexed by their arrival rate λ. Let ρ_λ denote the offered load in the system with arrival rate λ, i.e., $\rho_\lambda = \lambda/n_\lambda\mu$. Then, the following result holds:

THEOREM 5. *For a sequence of systems with offered load ρ_λ, if $\rho_\lambda \to \rho > 1$ as $\lambda \to \infty$, then for $\bar{w} := \bar{G}^{-1}(1/\rho)$:*

1. *If Assumption 1 holds at $\bar{w}$, then the fluid approximation to the steady-state expected queue length is $O(1)$-accurate. In particular,*

$$\limsup_{\lambda\to\infty}\left|\mathbb{E}Q_\lambda - \lambda\int_0^{\bar{G}^{-1}(1/\rho_\lambda)} \bar{G}(w)\,dw\right| \leqslant \sqrt{g(\bar{w})}\left(\frac{3|g'(\bar{w})|}{\rho g(\bar{w})^2} + \frac{1}{2}\right).$$

2. *The fluid approximation to the net rate of abandonments is a lower bound, i.e.,*

$$\alpha_\lambda \geqslant \lambda(1-1/\rho_\lambda).$$

Further, the approximation is $o(1)$-accurate, and we have

$$\limsup_{\lambda\to\infty}\frac{\log(\alpha_\lambda - \lambda(1-1/\rho_\lambda))}{\lambda} \leqslant -(H(\bar{w}) - \bar{w}/\rho) < 0. \tag{17}$$

B.1. Proof of Theorem 5

We begin by defining the function $L_\lambda(x) := (H(x) - x/\rho_\lambda)$. Note that $L_\lambda''(x) = -g(x) \leqslant 0$, and hence L_λ is a concave function. Let $\bar{w}_\lambda$ denote the maximizer of L_λ. Then, we have $L_\lambda'(\bar{w}_\lambda) = 0$, or $\bar{w}_\lambda = \bar{G}^{-1}(1/\rho_\lambda)$. Further, noting that $L_\lambda''(\bar{w}) < 0$, we obtain that $\bar{w}_\lambda$ is the unique maximizer of L. This property will be useful in proving our main result. Also, note that Assumption 1 implies that $\bar{w}_\lambda \to \bar{w} = \bar{G}^{-1}(1/\rho)$. This implies that for sufficiently large λ,

$$|\bar{w}_\lambda - \bar{w}| \leqslant \delta := \Delta/2, \tag{18}$$

where Δ is obtained from Assumption 1. Thus, using Assumption 1, we obtain that g is strictly positive and continuously differentiable on $[\bar{w}_\lambda - \delta,\ \bar{w}_\lambda + \delta]$ for all λ sufficiently large. For convenience, we will denote $\mathscr{E}_\lambda(n_\lambda)$, $J_\lambda(n_\lambda)$, and $J_{H,\lambda}(n_\lambda)$ by $\mathscr{E}_\lambda$, J_λ, and $J_{H,\lambda}$, respectively.

Proof Roadmap. We first derive some asymptotic properties of $\mathscr{E}_\lambda$, J_λ, and $J_{H,\lambda}$ in Lemmas 1–3. Then we prove that the expected steady-state queue length given in (12) equals $\lambda J_{H,\lambda}/J_\lambda + o(1)$ (this follows from Lemmas 1 and 2 below). Because both the terms $J_{H,\lambda}$ and J_λ integrate an exponential term that increases with λ, the ratio is characterized by the point that maximizes the exponent, which is $\bar{w}$ (see Lemma 2). Thus, using Lemma 3, which focuses on a small neighborhood of $\bar{w}$ to show that $\lambda J_{H,\lambda}/J_\lambda = \lambda H(\bar{w}) + O(1)$, part 1 of the result follows. Part 2 follows by using the exact expression for the net rate of abandonments in (13), and applying Lemma 1 and the fact that J_λ increases exponentially fast in λ (see (31) below).

LEMMA 1. *We have* $\mathscr{E}_\lambda \leqslant \rho_\lambda/(\rho_\lambda - 1)$ *and* $\mathscr{E}_\lambda \to \rho/(\rho-1)$ *as* $\lambda \to \infty$.

PROOF OF LEMMA 1. We have

$$\left(1+t\frac{\mu}{\lambda}\right)^{n_\lambda - 1} \leqslant \left(1+t\frac{\mu}{\lambda}\right)^{n_\lambda} \leqslant e^{t/\rho_\lambda}, \tag{19}$$

where we use the fact that $(1+y/x)^x \leqslant e^y$ for $x, y \geqslant 0$ (which follows from $\log(1+z) \leqslant z$ for $z \geqslant 0$) for the second inequality. Multiplying both sides by e^{-t} and integrating, we obtain the bound $\mathscr{E}_\lambda \leqslant \rho_\lambda/(\rho_\lambda - 1)$. Further, noting that $(1+t(\mu/\lambda))^{n_\lambda} \to e^{t/\rho}$ as $\lambda \to \infty$ and applying the dominated convergence theorem, we obtain $\mathscr{E}_\lambda \to \rho/(\rho-1)$. Q.E.D.

LEMMA 2. *As* $\lambda \to \infty$, *we have*

$$\frac{J_\lambda}{\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(\lambda L_\lambda(x))\,dx} \to 1, \tag{20}$$

$$\cdot \frac{1}{J_\lambda}\left(\int_0^{\bar{w}_\lambda-\delta} (H(x)-H(\bar{w}_\lambda))\exp(\lambda L_\lambda(x))\,dx + \int_{\bar{w}_\lambda+\delta}^{\infty} (H(x)-H(\bar{w}_\lambda))\exp(\lambda L_\lambda(x))\,dx\right) \to 0, \tag{21}$$

$$\frac{\mathscr{E}_\lambda}{J_\lambda} \to 0. \tag{22}$$

PROOF OF LEMMA 2. We first prove (20). Using the Taylor series expansion of $L_\lambda(x)$ around $\bar{w}_\lambda$, we can write $L_\lambda(x) = L_\lambda(\bar{w}_\lambda) - g(\psi(x))(x-\bar{w}_\lambda)^2/2$, where $\psi(x)$ lies between x and $\bar{w}_\lambda$. Thus, we can write the denominator as

$$\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(\lambda L_\lambda(x))\,dx = \exp(\lambda L_\lambda(\bar{w}_\lambda)) \int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(-\lambda g(\psi(x))(x-\bar{w}_\lambda)^2/2)\,dx. \tag{23}$$

Noting that g is continuous and thus bounded on $[\bar{w}_\lambda-\delta, \bar{w}_\lambda+\delta]$, we obtain the bound

$$\begin{aligned}\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(-\lambda g(\psi(x))(x-\bar{w}_\lambda)^2/2)\,dx &\geqslant \int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(-\lambda K_1(x-\bar{w}_\lambda)^2/2)\,dx \\ &= \sqrt{\frac{2\pi}{\lambda K_1}}\,\mathrm{Erf}(\delta\sqrt{\lambda K_1/2}),\end{aligned} \tag{24}$$

where $K_1 = \sup_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]} g(x) > 0$ and $\mathrm{Erf}(x) = 2/\sqrt{\pi}\int_0^x e^{-t^2}\,dt$ is the standard error function. Using this in (23), we obtain

$$\liminf_{\lambda\to\infty} \sqrt{\lambda}\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx \geqslant \sqrt{\frac{2\pi}{K_1}}. \tag{25}$$

Turning to the numerator, note that we can write

$$\begin{aligned}&\exp(-\lambda L_\lambda(\bar{w}_\lambda))\left(J_\lambda - \int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(\lambda L_\lambda(x))\,dx\right) \\ &\quad= \int_0^{\bar{w}_\lambda-\delta} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx + \int_{\bar{w}_\lambda+\delta}^{\infty} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx.\end{aligned} \tag{26}$$

Next, using the unimodality of $L_\lambda(x)$, we note that $L_\lambda(x)$ is weakly increasing on $[0, \bar{w}_\lambda-\delta]$, and thus

$$\int_0^{\bar{w}_\lambda-\delta} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx \leqslant \bar{w}_\lambda \exp(\lambda(L_\lambda(\bar{w}_\lambda-\delta)-L_\lambda(\bar{w}_\lambda))). \tag{27}$$

Note that $L_\lambda(\bar{w}_\lambda-\delta) - L_\lambda(\bar{w}_\lambda) = H(\bar{w}_\lambda-\delta) - H(\bar{w}_\lambda) + \delta/\rho_\lambda$. Thus, noting that $\rho_\lambda \to \rho$ and $\bar{w}_\lambda \to \bar{w}$ as $\lambda \to \infty$, we obtain $L_\lambda(\bar{w}_\lambda-\delta) - L_\lambda(\bar{w}_\lambda) \to H(\bar{w}-\delta) - H(\bar{w}) + \delta/\rho$ as $\lambda \to \infty$. Further, noting that $H(x) - x/\rho$ is maximized at $\bar{w}$ (because $H'(\bar{w}) - 1/\rho = 0$, $H''(w) \leqslant 0$ with $H''(\bar{w}) < 0$), we obtain $\lim_{\lambda\to\infty} L_\lambda(\bar{w}_\lambda-\delta) - L_\lambda(\bar{w}_\lambda) < 0$. Hence, the term on the right-hand side of (27) can be bounded above by $e^{-\epsilon\lambda}$ for some $\epsilon > 0$ and sufficiently large λ, which gives us

$$\limsup_{\lambda\to\infty} \sqrt{\lambda}\int_0^{\bar{w}_\lambda-\delta} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx = 0. \tag{28}$$

We also have

$$\begin{aligned}&\int_{\bar{w}_\lambda+\delta}^{\infty} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx \\ &\quad= \int_{\bar{w}_\lambda+\delta}^{\max\{2\rho_\lambda/\gamma,\,\bar{w}_\lambda+\delta\}} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx + \int_{\max\{2\rho_\lambda/\gamma,\,\bar{w}_\lambda+\delta\}}^{\infty} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx \\ &\quad\overset{(a)}{\leqslant} (2\rho_\lambda/\gamma - (\bar{w}_\lambda+\delta))^+ \exp(\lambda(L_\lambda(\bar{w}_\lambda+\delta)-L_\lambda(\bar{w}_\lambda))) + \int_{2\rho_\lambda/\gamma}^{\infty} \exp(-\lambda x/(2\rho_\lambda))\,dx \\ &\quad\leqslant (2\rho_\lambda/\gamma - (\bar{w}_\lambda+\delta))^+ \exp(\lambda(L_\lambda(\bar{w}_\lambda+\delta)-L_\lambda(\bar{w}_\lambda))) + 2\rho_\lambda/\lambda e^{-\lambda/\gamma},\end{aligned}$$

where (a) follows by noting that L_λ is decreasing on $[\bar{w}_\lambda+\delta, \max\{2\rho_\lambda/\gamma, \bar{w}_\lambda+\delta\}]$ and using $L_\lambda(x) = H(x) - x/\rho_\lambda \leqslant 1/\gamma - x/\rho_\lambda < -x/(2\rho_\lambda)$ for $x > 2\rho_\lambda/\gamma$. Thus, arguing as before, we obtain

$$\limsup_{\lambda\to\infty} \sqrt{\lambda}\int_{\bar{w}_\lambda+\delta}^{\infty} \exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx = 0. \tag{29}$$

Thus, using (28) and (29) in (26), we obtain

$$\lim_{\lambda\to\infty} \sqrt{\lambda}\exp(-\lambda L_\lambda(\bar{w}_\lambda))\left(J_\lambda - \int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta} \exp(\lambda L_\lambda(x))\,dx\right) = 0. \tag{30}$$

Combining the above with (25), we obtain

$$\lim_{\lambda\to\infty}\frac{J_\lambda-\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\exp(\lambda L_\lambda(x))\,dx}{\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\exp(\lambda L_\lambda(x))\,dx}=0,$$

which proves (20).

We now turn to (21). Using the fact that $0\leqslant H(x)\leqslant 1/\gamma$, arguing as before, we obtain

$$\limsup_{\lambda\to\infty}\sqrt{\lambda}\exp(-\lambda L_\lambda(\bar w_\lambda))\cdot\left(J_{H,\lambda}-\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}H(x)\exp(\lambda L_\lambda(x))\,dx\right)=0.$$

Multiplying (30) by $H(\bar w_\lambda)$ and subtracting from the above, we obtain the bound

$$\limsup_{\lambda\to\infty}\sqrt{\lambda}\exp(-\lambda L_\lambda(\bar w_\lambda))\cdot\left(\int_0^\infty(H(x)-H(\bar w_\lambda))\exp(\lambda L_\lambda(x))\,dx-\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}(H(x)-H(\bar w_\lambda))\exp(\lambda L_\lambda(x))\,dx\right)=0.$$

Thus, (21) follows by using (20) and (25).

Finally, we turn to (22). Using (25) and the definition of J_λ, we have for sufficiently large λ,

$$J_\lambda\geqslant\sqrt{\frac{2\pi}{K_1\lambda}}\exp(\lambda L_\lambda(\bar w_\lambda)).$$

Defining $L(x)=H(x)-x/\rho$, we have $\sup_{0\leqslant x\leqslant T}|L_\lambda(x)-L(x)|\to 0$ as $\lambda\to\infty$ for any $T>0$. Thus, we have $L_\lambda(\bar w_\lambda)\to L(\bar w)$, where $\bar w=\bar G^{-1}(1/\rho)$. Noting that L is a concave function with $L(0)=0$ and $L'(0)=1-1/\rho>0$, and that $L'(\bar w)=0$ so that $\bar w$ is the unique maximizer of L, we obtain that $\bar w>0$, and further, $L(\bar w)>0$. Thus, using the fact that L_λ converges uniformly to L, we obtain $\lim_{\lambda\to\infty}L_\lambda(\bar w_\lambda)=L(\bar w)>0$. Hence, choosing any $0<\epsilon<L(\bar w)$, we have

$$J_\lambda\geqslant\sqrt{\frac{2\pi}{K_1\lambda}}\exp(\lambda(L(\bar w)-\epsilon)) \tag{31}$$

for all λ sufficiently large. The result then follows by noting that $\mathscr{C}_\lambda\to\rho/(\rho-1)<\infty$ (cf. Lemma 1). Q.E.D.

Lemma 3. *If Assumptions* 1 *holds, then*

$$\limsup_{\lambda\to\infty}\left|\frac{\lambda J_{H,\lambda}}{J_\lambda}-\lambda H(\bar w_\lambda)\right|\leqslant\sqrt{g(\bar w)}\left(\frac{3|g'(\bar w)|}{\rho g(\bar w)^2}+\frac{1}{2}\right).$$

Proof of Lemma 3. Applying Lemma 2 and multiplying the numerator and denominator by $\exp(-\lambda L_\lambda(\bar w_\lambda))$, the result follows if we prove

$$\limsup_{\lambda\to\infty}\left|\frac{\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\lambda(H(x)-H(\bar w_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx}{\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx}\right|\leqslant\sqrt{g(\bar w)}\left(\frac{3|g'(\bar w)|}{\rho g(\bar w)^2}+\frac{1}{2}\right). \tag{32}$$

Arguing as in the proof of Lemma 2, we have

$$\liminf_{\lambda\to\infty}\sqrt{\lambda}\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\geqslant\sqrt{\frac{2\pi}{K_1}}, \tag{33}$$

where $K_1=\sup_{x\in[\bar w-\Delta,\,\bar w+\Delta]}g(x)>0$.

Define $\epsilon_\lambda=\lambda^{-1/2+\kappa}$ for $0<\kappa<1/6$. Noting that for large λ, $\epsilon_\lambda<\delta$, we can write the numerator as

$$\begin{aligned}&\left|\int_{\bar w_\lambda-\delta}^{\bar w_\lambda+\delta}\lambda(H(x)-H(\bar w_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|\\&\leqslant\left|\int_{\bar w_\lambda-\delta}^{\bar w_\lambda-\epsilon_\lambda}\lambda(H(x)-H(\bar w_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|\\&\quad+\left|\int_{\bar w_\lambda-\epsilon_\lambda}^{\bar w_\lambda+\epsilon_\lambda}\lambda(H(x)-H(\bar w_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|\\&\quad+\left|\int_{\bar w_\lambda+\epsilon_\lambda}^{\bar w_\lambda+\delta}\lambda(H(x)-H(\bar w_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|.\end{aligned} \tag{34}$$

We now prove that the second term on the right-hand side above is $O(1/\sqrt{\lambda})$, whereas the other terms are $o(1/\sqrt{\lambda})$.

Using the Taylor series expansion of $L_\lambda(x)$ around $\bar w_\lambda$, we can write $L_\lambda(x)=L_\lambda(\bar w_\lambda)-g(\psi(x))(x-\bar w_\lambda)^2/2$, where $\psi(x)$ lies between x and $\bar w_\lambda$. Then, the first term on the right-hand side of the above relation can be expressed as:

$$\begin{aligned}&\left|\int_{\bar w_\lambda-\delta}^{\bar w_\lambda-\epsilon_\lambda}\lambda(H(x)-H(\bar w_\lambda))\exp\left(-\lambda g(\psi(x))\frac{(x-\bar w_\lambda)^2}{2}\right)dx\right|\\&\overset{(a)}{\leqslant}\int_{\bar w_\lambda-\delta}^{\bar w_\lambda-\epsilon_\lambda}\frac{\lambda}{\gamma}\exp\left(-\lambda K_2\frac{(x-\bar w_\lambda)^2}{2}\right)dx\\&\leqslant\int_{-\infty}^{\bar w_\lambda-\epsilon_\lambda}\frac{\lambda}{\gamma}\exp\left(-\lambda K_2\frac{(x-\bar w_\lambda)^2}{2}\right)dx\\&\overset{(b)}{\leqslant}\frac{1}{\gamma}\sqrt{\frac{2\pi}{K_2}}\sqrt{\lambda}\exp(-\lambda\epsilon_\lambda^2K_2/2),\end{aligned}$$

where $K_2=\inf_{x\in[\bar w-\Delta,\,\bar w+\Delta]}g(x)$ and (a) follows as $g(x)\geqslant K_2>0$ and $x\in[\bar w_\lambda-\delta,\bar w_\lambda]$ along with the fact $|H(x)-H(\bar w_\lambda)|\leqslant 1/\gamma$, and (b) follows from the Chernoff bound for the normal distribution. Thus, using $\epsilon_\lambda=\lambda^{-1/2+\kappa}$ for $0<\kappa<1/6$, we obtain

$$\lim_{\lambda\to\infty}\sqrt{\lambda}\left|\int_{\bar w_\lambda-\delta}^{\bar w_\lambda-\epsilon_\lambda}\lambda(H(x)-H(\bar w_\lambda))\cdot\exp(-\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|=0. \tag{35}$$

Similarly, for the third term on the right-hand side of (34), we have:

$$\lim_{\lambda\to\infty}\sqrt{\lambda}\left|\int_{\bar w_\lambda+\epsilon_\lambda}^{\bar w_\lambda+\delta}\lambda(H(x)-H(\bar w_\lambda))\cdot\exp(-\lambda(L_\lambda(x)-L_\lambda(\bar w_\lambda)))\,dx\right|=0. \tag{36}$$

We next turn to the middle term on the right-hand side of (34). It will be useful to note the Taylor series expansion of $L_\lambda(x)$:

$$L_\lambda(x) = L_\lambda(\bar{w}_\lambda) - g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2} - g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}, \tag{37}$$

where $\zeta(x)$ lies between x and $\bar{w}_\lambda$. Using this expansion, the middle term on the right-hand side of (34) can be written as

$$\begin{aligned}
&\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(H(x)-H(\bar{w}_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx\right| \\
&\quad= \left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(H(x)-H(\bar{w}_\lambda))\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right.\right. \\
&\qquad\qquad \left.\left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx\right| \\
&\quad\leqslant \frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right.\right. \\
&\qquad\qquad \left.\left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx\right| \\
&\qquad + \frac{K_1}{2}\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)^2\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right. \\
&\qquad\qquad \left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx, \qquad (38)
\end{aligned}$$

where we use the Taylor series expansion for $H(x)$ around $H(\bar{w}_\lambda)$ to obtain $H(x) = H(\bar{w}_\lambda) + H'(\bar{w}_\lambda)(x-\bar{w}_\lambda) + H''(\nu(x))(x-\bar{w}_\lambda)^2/2$, where $\nu(x)$ lies between x and $\bar{w}_\lambda$. We further use $H'(\bar{w}_\lambda) = \bar{G}(\bar{w}_\lambda) = 1/\rho_\lambda$ and $|H''(x)|/2 = g(x)/2 \leqslant \sup_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]} g(x)/2 = K_1/2$.

We now consider the first term in the right-hand side of (38).

$$\begin{aligned}
&\frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right.\right. \\
&\qquad\qquad \left.\left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx\right| \\
&\overset{(a)}{\leqslant} \frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right.\right. \\
&\qquad\qquad \left.\left. -\lambda M_2(x-\bar{w}_\lambda)^3\right)dx\right| \\
&\overset{(b)}{=} \frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)(1-\lambda M_2(x-\bar{w}_\lambda)^3 e^{-\lambda M_2(\xi(x)-\bar{w}_\lambda)^3})\right. \\
&\qquad\qquad \left.\cdot\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx\right| \\
&\overset{(c)}{\leqslant} \frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx\right| \\
&\quad + \frac{2}{\rho_\lambda}\lambda^2|M_2|e^{\lambda|M_2|\epsilon_\lambda^3}\int_{\bar{w}_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda}(x-\bar{w}_\lambda)^4 \\
&\qquad\qquad \cdot\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx \\
&\overset{(d)}{=} \frac{3|M_2|}{\rho_\lambda}\sqrt{\frac{2\pi}{g(\bar{w}_\lambda)^5}}e^{\lambda|M_2|\epsilon_\lambda^3}\frac{\mathrm{Erf}(\sqrt{g(\bar{w}_\lambda)}\sqrt{\lambda}\epsilon_\lambda/\sqrt{2})}{\sqrt{\lambda}},
\end{aligned}$$

where $\xi(x)$ lies between x and $\bar{w}_\lambda$, $\mathrm{Erf}(x) = (2/\sqrt{\pi})\int_0^x e^{-t^2}\,dt$ is the standard error function, the inequality (a) follows by defining $M_2 := \inf_{x\in[\bar{w}-\Delta,\bar{w}+\Delta]} g'(x)/6$, which is finite because g' is continuous, which in turn implies that $|g'(x)| < \infty$ for $x \in [\bar{w}-\Delta, \bar{w}+\Delta]$; (b) follows by applying Taylor's formula to $\exp(-\lambda M_2(x-\bar{w}_\lambda)^3)$ around $\bar{w}_\lambda$. The relation (c) follows by noting that $|\xi(x)-\bar{w}_\lambda| \leqslant \epsilon_\lambda$, and (d) follows by integrating the terms and noting that the first integral is zero.

Thus, taking the limit as $\lambda\to\infty$ and using $\epsilon_\lambda = \lambda^{-1/2+\kappa}$ for $0 < \kappa < 1/6$, we obtain

$$\begin{aligned}
&\limsup_{\lambda\to\infty}\sqrt{\lambda}\frac{1}{\rho_\lambda}\left|\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right.\right. \\
&\qquad\qquad \left.\left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx\right| \\
&\quad\leqslant \frac{3|M_2|}{\rho}\sqrt{\frac{2\pi}{g(\bar{w})^5}}. \qquad (39)
\end{aligned}$$

We now turn to the second term on the right-hand side of (38).

$$\begin{aligned}
&\frac{K_1}{2}\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)^2\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right. \\
&\qquad\qquad \left. -\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx \\
&\overset{(a)}{\leqslant} \frac{K_1}{2}\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda} \lambda(x-\bar{w}_\lambda)^2\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right. \\
&\qquad\qquad \left. -\lambda M_3(x-\bar{w}_\lambda)^3\right)dx \\
&\quad + \frac{K_1}{2}\int_{\bar{w}_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)^2\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right. \\
&\qquad\qquad \left. -\lambda M_2(x-\bar{w}_\lambda)^3\right)dx \\
&\overset{(b)}{=} \frac{K_1}{2}\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda} \lambda(x-\bar{w}_\lambda)^2(1-\lambda M_3(x-\bar{w}_\lambda)^3 e^{-\lambda M_3(\xi(x)-\bar{w}_\lambda)^3}) \\
&\qquad \cdot\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx \\
&\quad + \frac{K_1}{2}\int_{\bar{w}_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda} \lambda(x-\bar{w}_\lambda)^2(1-\lambda M_2(x-\bar{w}_\lambda)^3 e^{-\lambda M_2(\xi(x)-\bar{w}_\lambda)^3}) \\
&\qquad \cdot\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx
\end{aligned}$$

$$\overset{(c)}{\leqslant} K_1 \int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda} \lambda(x-\bar{w}_\lambda)^2 \exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right) dx$$
$$+\lambda^2(|M_2|e^{\lambda|M_2|\epsilon_\lambda^3}+|M_3|e^{\lambda|M_3|\epsilon_\lambda^3})\frac{K_1}{2}\int_{\bar{w}_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda}(x-\bar{w}_\lambda)^5$$
$$\cdot\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}\right)dx$$
$$\overset{(d)}{=} K_1\sqrt{\frac{\pi}{2\lambda g(\bar{w}_\lambda)^3}}\,\mathrm{Erf}(\epsilon_\lambda\sqrt{\lambda g(\bar{w}_\lambda)/2})+o(1/\sqrt{\lambda}),$$

where the inequality (a) follows by noting that $M_2 = \inf_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]} g'(x)/6$ and defining $M_3 := \sup_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]} g'(x)/6$. Note that because g' is continuous, we have $|g'(x)|<\infty$ for $x\in[\bar{w}-\Delta,\bar{w}+\Delta]$. The relation (b) follows by applying Taylor's formula to $\exp(-\lambda M_2(x-\bar{w}_\lambda)^3)$ and $\exp(-\lambda M_3(x-\bar{w}_\lambda)^3)$ around $\bar{w}_\lambda$, (c) follows by noting that $|\xi(x)-\bar{w}_\lambda|\leqslant\epsilon_\lambda$ and using the symmetry of the terms, and (d) follows by integrating the terms and noting that the second integral is $o(1/\sqrt{\lambda})$ for $\epsilon_\lambda=\lambda^{-1/2+\kappa}$ for $0<\kappa<1/6$. Thus, taking the limit as $\lambda\to\infty$, we obtain

$$\limsup_{\lambda\to\infty}\sqrt{\lambda}\frac{K_1}{2}\int_{\bar{w}_\lambda-\epsilon_\lambda}^{\bar{w}_\lambda+\epsilon_\lambda}\lambda(x-\bar{w}_\lambda)^2\exp\left(-\lambda g(\bar{w}_\lambda)\frac{(x-\bar{w}_\lambda)^2}{2}-\lambda g'(\zeta(x))\frac{(x-\bar{w}_\lambda)^3}{6}\right)dx$$
$$\leqslant K_1\sqrt{\frac{\pi}{2g(\bar{w})^3}}. \tag{40}$$

Thus, combining (39) and (40) in (38), and using (33), we obtain

$$\limsup_{\lambda\to\infty}\left|\frac{\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta}\lambda(H(x)-H(\bar{w}_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx}{\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta}\exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx}\right|$$
$$\leqslant K_1\left(\frac{3|M_2|}{\rho}\sqrt{\frac{1}{g(\bar{w})^5}}+\frac{K_1}{2}\sqrt{\frac{1}{g(\bar{w})^3}}\right),$$

where $K_1=\sup_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]}g(x)$ and $M_2=\inf_{x\in[\bar{w}-\Delta,\,\bar{w}+\Delta]}g'(x)/6$. Noting that g is continuously differentiable on $[\bar{w}-\Delta,\bar{w}+\Delta]$, and that we can choose Δ arbitrarily small, we obtain:

$$\limsup_{\lambda\infty}\left|\frac{\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta}\lambda(H(x)-H(\bar{w}_\lambda))\exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx}{\int_{\bar{w}_\lambda-\delta}^{\bar{w}_\lambda+\delta}\exp(\lambda(L_\lambda(x)-L_\lambda(\bar{w}_\lambda)))\,dx}\right|$$
$$\leqslant\liminf_{\Delta 0}K_1\left(\frac{3|M_2|}{\rho}\sqrt{\frac{1}{g(\bar{w})^5}}+\frac{K_1}{2}\sqrt{\frac{1}{g(\bar{w})^3}}\right)$$
$$=\left(\frac{3|g'(\bar{w})|}{\rho}\sqrt{\frac{1}{g(\bar{w})^3}}+\frac{1}{2}\sqrt{g(\bar{w})}\right),$$

which completes the proof. Q.E.D.

Proof of Theorem 5. We first prove part 1. We have

$$|\mathbb{E}Q_\lambda-\lambda\bar{q}| = \left|\frac{\lambda^2 J_{H,\lambda}}{\mathscr{E}_\lambda+\lambda J_\lambda}-\lambda H(\bar{w}_\lambda)\right|$$
$$=\left|\frac{\lambda^2 J_{H,\lambda}-\lambda H(\bar{w}_\lambda)\mathscr{E}_\lambda-\lambda^2 J_\lambda H(\bar{w}_\lambda)}{\mathscr{E}_\lambda+\lambda J_\lambda}\right|$$
$$\leqslant\left|\frac{\lambda J_{H,\lambda}-\lambda J_\lambda H(\bar{w}_\lambda)}{J_\lambda(1+\mathscr{E}_\lambda/(\lambda J_\lambda))}\right| + H(\bar{w}_\lambda)\frac{\mathscr{E}_\lambda}{J_\lambda(1+\mathscr{E}_\lambda/(\lambda J_\lambda))}. \tag{41}$$

Applying Lemmas 2 and 3, we obtain

$$\limsup_{\lambda\to\infty}|\mathbb{E}Q_\lambda-\lambda\bar{q}|\leqslant\limsup_{\lambda\to\infty}\left|\frac{\lambda J_{H,\lambda}}{J_\lambda}-\lambda H(\bar{w}_\lambda)\right|$$
$$\leqslant\sqrt{g(\bar{w})}\left(\frac{3|g'(\bar{w})|}{\rho g(\bar{w})^2}+\frac{1}{2}\right).$$

This completes the proof of part 1.

Turning to part 2, using the exact expression for the net rate of abandonments (13), we obtain

$$\alpha_\lambda-\lambda(1-1/\rho_\lambda)=\lambda\frac{1+\lambda(1-1/\rho_\lambda)J_\lambda}{\mathscr{E}_\lambda+\lambda J_\lambda}-\lambda(1-1/\rho_\lambda)$$
$$=\frac{1-\mathscr{E}_\lambda(1-1/\rho_\lambda)}{J_\lambda+\mathscr{E}_\lambda/\lambda}. \tag{42}$$

Lemma 1 proves that $\mathscr{E}_\lambda\leqslant\rho_\lambda/(\rho_\lambda-1)$, which implies that $1-\mathscr{E}_\lambda(1-1/\rho_\lambda)\geqslant 0$. Thus, we obtain $\alpha_\lambda\geqslant\lambda(1-1/\rho_\lambda)$. Taking logarithms on both sides of (42), dividing by λ, taking the limit as $\lambda\to\infty$, and using (31), we obtain

$$\limsup_{\lambda\to\infty}\frac{\log(\alpha_\lambda-\lambda(1-1/\rho_\lambda))}{\lambda}\leqslant-(L(\bar{w})-\epsilon)$$

for all $0<\epsilon<L(\bar{w})$, and the result follows. Q.E.D.

Appendix C: Proof of Results

C.1. Proof of Theorems

Theorems 1 and 2 follow from the more general Theorem 5.

Proof of Theorem 3. Consider any sequence of capacity levels $\{n_\lambda\}$. If the corresponding offered loads $\{\rho_\lambda\}$ satisfy $\rho_\lambda\to\rho>1$ as $\lambda\to\infty$, then applying Theorem 5, for λ sufficiently large, we obtain

$$|\Pi_\lambda(n_\lambda)-\lambda\overline{\Pi}(\bar{w}_\lambda)|\leqslant K, \tag{43}$$

where $K>0$ is a constant, and $\bar{w}_\lambda=\bar{G}^{-1}(1/\rho_\lambda)$. Combining this with $\bar{w}_\lambda\to\bar{w}$ as $\lambda\to\infty$ and the continuity of $\overline{\Pi}(\cdot)$, we obtain

$$\lim_{\lambda\to\infty}\frac{\Pi_\lambda(n_\lambda)}{\lambda\overline{\Pi}(\bar{w})}=1. \tag{44}$$

In the case that $\rho_\lambda \to 1$, noting the fact that the expected steady-state queue length and net rate of abandonment are both nonnegative, we obtain the lower bound

$$\Pi_\lambda(n_\lambda) \geqslant cn_\lambda.$$

Noting that $n_\lambda/\lambda \to 1/\mu$ as $\lambda \to \infty$ and $\overline{\Pi}(0) = c/\mu$, we have

$$\liminf_{\lambda\to\infty} \frac{\Pi_\lambda(n_\lambda)}{\lambda\overline{\Pi}(0)} \geqslant 1. \tag{45}$$

Further, by the optimality of $w^\star$, we have $\overline{\Pi}(w) \geqslant \overline{\Pi}(w^\star)$ for all $w \geqslant 0$. Using this fact along with (44) and (45), we obtain

$$\liminf_{\lambda\to\infty} \frac{\Pi_\lambda(n_\lambda)}{\lambda\overline{\Pi}(w^\star)} \geqslant 1.$$

Turning to the prescription, we have

$$\lim_{\lambda\to\infty} \frac{\Pi_\lambda(n_\lambda^\star)}{\lambda\overline{\Pi}(w^\star)} = 1,$$

which follows from Theorem 5 for the case $w^\star > 0$, and from Theorem 4.1 in Mandelbaum and Zeltyn (2005) for the case $w^\star = 0$. Thus, we have $\Pi_\lambda^\star/\Pi_\lambda(n_\lambda^\star) \to 1$ as $\lambda \to \infty$.

For the case $n_\lambda^\star < \lambda/\mu$, applying Theorem 5, we further have $\Pi_\lambda(n_\lambda^\star) - \lambda\overline{\Pi}(w^\star) \leqslant K_1$ for a constant $K_1 \geqslant 0$. Let $\{n_\lambda^{opt}\}$ denote the sequence of capacity levels that minimize the actual cost, i.e., solve (4). Then, the corresponding offered loads $\{\rho_\lambda^{opt}\}$ must satisfy $\rho_\lambda^{opt} \to \rho^\star = 1/\bar{G}(w^\star)$, and thus, applying Theorem 5, we obtain

$$\lambda\overline{\Pi}(w_\lambda^{opt}) - \Pi_\lambda(n_\lambda^{opt}) \leqslant K_2,$$

where $w_\lambda^{opt} = \bar{G}^{-1}(1/\rho_\lambda^{opt})$ and $K_2 \geqslant 0$ is a constant. Thus, we have

$$\begin{aligned} 0 \leqslant \Pi_\lambda(n_\lambda^\star) - \Pi_\lambda(n_\lambda^{opt}) &\leqslant \lambda(\overline{\Pi}(w^\star) - \overline{\Pi}(w_\lambda^{opt})) \\ &\quad + (K_1 + K_2) \qquad (46) \\ &\leqslant (K_1 + K_2), \qquad (47) \end{aligned}$$

because $w^\star$ minimizes $\overline{\Pi}$, and the $O(1)$-optimality follows. Q.E.D.

Proof of Theorem 4. This proof follows exactly as the proof of Theorem 5, with the only exception being that we obtain $\mathscr{E}_\lambda = 1$ (using $n = 1$ in (14)). We omit the details. Q.E.D.

C.2. Proof of Propositions

Proof of Proposition 2. In this case, the expected queue length can be computed exactly by using formula (12). The error in the fluid approximation can then be shown to be $o(1)$. Alternatively, note that in an $M/M/n+M$ queueing system the expected queue length times the rate of individual customer abandonment equals the net rate at which customers abandon. That is, denoting the mean time to abandon by $1/\gamma$, we have

$$\alpha_\lambda = \gamma\mathbb{E}Q_\lambda,$$

for any capacity level (see, for instance, Garnett et al. 2002). Combining this with Theorem 2, the result immediately follows. Q.E.D.

Proof of Proposition 3. The expected steady-state queue length is given by (12). As $\rho = 2$, using (2), we obtain $\bar{w} = 1/2$. The expected steady-state queue length $\mathbb{E}Q_\lambda = \lambda^2 J_{H,\lambda}/(\mathscr{E}_\lambda + \lambda J_\lambda)$, and the corresponding fluid limit is $\lambda H(\bar{w})$. Thus, we can write the approximation error as

$$\mathbb{E}Q_\lambda - \lambda H(\bar{w}) = \frac{\lambda^2(J_{H,\lambda} - H(\bar{w})J_\lambda) - \lambda\mathscr{E}_\lambda H(\bar{w})}{\mathscr{E}_\lambda + \lambda J_\lambda}.$$

We multiply both the numerator and denominator by $e^{-\lambda L(\bar{w})}$, where $L(x) = H(x) - x/\rho$, to obtain

$$\begin{aligned} &\mathbb{E}Q_\lambda - \lambda H(\bar{w}) \\ &\quad = \frac{\lambda^2(J_{H,\lambda} - H(\bar{w})J_\lambda)e^{-\lambda L(\bar{w})} - \lambda\mathscr{E}_\lambda H(\bar{w})e^{-\lambda L(\bar{w})}}{\mathscr{E}_\lambda e^{-\lambda L(\bar{w})} + \lambda J_\lambda e^{-\lambda L(\bar{w})}}. \qquad (48) \end{aligned}$$

The first term in the numerator is given by

$$\begin{aligned} &\lambda^2(J_{H,\lambda} - H(\bar{w})J_\lambda)e^{-\lambda L(\bar{w})} \\ &\quad = \int_0^\infty \lambda^2 (H(x) - H(\bar{w})) \exp(\lambda(L(x) - L(\bar{w}))) \, dx \\ &\quad = \int_0^{\bar{w}} \lambda^2 (H(x) - H(\bar{w})) \exp(\lambda(L(x) - L(\bar{w}))) \, dx \\ &\qquad + \int_{\bar{w}}^1 \lambda^2 (H(x) - H(\bar{w})) \exp(\lambda(L(x) - L(\bar{w}))) \, dx \\ &\qquad + \int_1^\infty \lambda^2 (H(x) - H(\bar{w})) \exp(\lambda(L(x) - L(\bar{w}))) \, dx. \end{aligned}$$

Next we note that for $x \in [\bar{w}, 1]$,

$$C_1(x - \bar{w}) \leqslant H(x) - H(\bar{w}) \leqslant C_2(x - \bar{w}), \tag{49}$$

where $C_1, C_2 > 0$ are finite constants independent of λ. This relation holds because H is a concave nondecreasing function, and thus $H(x) \leqslant H(\bar{w}) + \bar{G}(\bar{w})(x - \bar{w})$, and also

$$\begin{aligned} H(x) &\geqslant \frac{(x - \bar{w})}{(1 - \bar{w})} H(1) + \frac{(1 - x)}{(1 - \bar{w})} H(\bar{w}) \\ &= H(\bar{w}) + \frac{H(1) - H(\bar{w})}{1 - \bar{w}}(x - \bar{w}). \end{aligned}$$

Thus, we obtain

$$\begin{aligned} &\lambda^2(J_{H,\lambda} - H(\bar{w})J_\lambda)e^{-\lambda L(\bar{w})} \\ &\quad \leqslant \int_0^{\bar{w}} \lambda^2 (H(x) - H(\bar{w})) \exp(\lambda(L(x) - L(\bar{w}))) \, dx \end{aligned}$$

$$
\begin{aligned}
&+\int_{\bar{w}}^{1}\lambda^2 C_2(x-\bar{w})\exp(\lambda(L(x)-L(\bar{w})))\,dx\\
&+\int_{1}^{\infty}\lambda^2(H(x)-H(\bar{w}))\exp(\lambda(L(x)-L(\bar{w})))\,dx\\
&=\left[\frac{1}{4}\left(3e^{-\lambda/8}\lambda-2\lambda-\sqrt{2\pi\lambda}\,\mathrm{Erf}\left(\frac{\sqrt{\lambda}}{2\sqrt{2}}\right)\right)\right]\\
&\quad+\left[\frac{C_2}{4}\frac{1}{2+m}((8+4m)^{2/(2+m)}\lambda^{2(m+1)/(m+2)}\right.\\
&\quad\left.\cdot(\Gamma(2/(2+m))-\Gamma(2/(2+m),\lambda/(8+4m)))\right]\\
&\quad+\left[\frac{1}{2}\frac{(1+m)\lambda\exp(-\lambda/(8+4m))}{2+m}\right]\\
&\leqslant C_3\lambda^{2(m+1)/(m+2)},
\end{aligned}
$$

where $\mathrm{Erf}(x)=2/\sqrt{\pi}\int_0^x e^{-t^2}dt$ is the standard error function, $\Gamma(x)$ and $\Gamma(x,y)$ denote the complete and incomplete Gamma functions, respectively, and $C_3>0$ is a finite constant. The last inequality follows because the other nonnegative terms are exponentially decreasing in λ and $2(m+1)/(m+2)>1$ for $m>0$.

For the lower bound, we repeat the argument using the lower bound in (49) to obtain $e^{-\lambda L(\bar{w})}\lambda^2(J_{H,\lambda}-H(\bar{w})J_\lambda)=\Theta(\lambda^{2(m+1)/(m+2)})$. Further, for the second term in the numerator of (48), we note that as $L(\bar{w})>0$ and $\mathscr{E}_\lambda\leqslant\rho/(\rho-1)=2$ (cf. Lemma 1), we have $e^{-\lambda L(\bar{w})}\lambda\mathscr{E}_\lambda H(\bar{w})=o(1)$.

The denominator of (48) can be expressed as follows

$$
\begin{aligned}
&e^{-\lambda L(\bar{w})}\mathscr{E}_\lambda+\lambda\int_0^\infty\exp(\lambda(L(x)-L(\bar{w})))\,dx\\
&=e^{-\lambda L(\bar{w})}\mathscr{E}_\lambda+\left[\sqrt{\lambda}\sqrt{\frac{\pi}{2}}\mathrm{Erf}\left(\frac{\sqrt{\lambda}}{2\sqrt{2}}\right)\right.\\
&\quad\left.+\lambda^{(m+1)/(m+2)}2^{-m/(2+m)}(2+m)^{1/(2+m)}\Gamma\left(1+\frac{1}{2+m}\right)\right]\\
&=\Theta(\lambda^{(m+1)/(m+2)}).
\end{aligned}
$$

Combining the bounds for the numerator and denominator of (48), we obtain $\mathbb{E}Q_\lambda-\lambda H(\bar{w})=\Theta(\lambda^{(m+1)/(m+2)})$. This completes the proof. Q.E.D.

Proof of Proposition 4. Because $\overline{\Pi}$ is strictly quasi-convex for decreasing hazard rate patience distributions, $w^\star$ is the unique minimizer of (8). The result then follows by applying Theorem 3 (part 2). Q.E.D.

Proof of Proposition 5. Observe that if conditions (a) or (b) hold, we cannot have an interior solution to (8), and thus we must have $w^\star=0$ or $w^\star=\infty$. This follows because if the patience distribution has an increasing hazard rate, then $\overline{\Pi}$ is quasi-concave, and if $c/\mu<p$, then the first-order optimality condition (10) cannot hold because the hazard rate is always nonnegative. For $c/\mu<p+h/\gamma$, we have

$$\overline{\Pi}(0)=c/\mu<p+h/\gamma=\overline{\Pi}(\infty),$$

and thus $w^\star=0$ is the optimal solution to (8). The result then follows by applying Theorem 3 (part 1). Q.E.D.

Endnotes

1. See Appendix B for a general version in which the offered load varies with the arrival rate.
2. Note that the exponential distribution has a constant hazard rate, whereas the other two distributions have decreasing hazard rates. This property will be used in the numerical experiments in §4.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* **57**(1) 66–81.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. F. J. Kylstra, ed. *Performance'*81. North-Holland, Amsterdam, 159–179.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3) 419–435.

Bassamboo, A., R. S. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* **56**(10) 1668–1686.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Boxma, O. J., P. R. de Waal. 1994. Multiserver queues with impatient customers. J. Labetoulle, J. W. Roberts, eds. *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks*. Elsevier, Amsterdam, 743–756.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

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

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1) 20–36.

Kumar, S., R. S. Randhawa. 2010. Exploiting market size in service systems. *Manufacturing Service Oper. Management* **12**(3) 511–526.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Mandelbaum, A., S. Zeltyn. 2005. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* **51**(3-4) 361–402.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* **57**(5) 1189–1205.

Ward, A., P. Glynn. 2005. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* **50**(4) 371–400.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.