Operations Research Letters 41 (2013) 27–31

Contents lists available at SciVerse ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Accuracy of fluid approximations for queueing systems with congestion-sensitive demand and implications for capacity sizing

R.S. Randhawa

Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 1 August 2012
Received in revised form
24 October 2012
Accepted 31 October 2012
Available online 14 November 2012

Keywords:
Queueing
Fluid limits
Congestion-dependent demand
Optimal capacity selection
Order-1 accuracy

ABSTRACT

We study the accuracy of fluid approximations in single- and many-server queueing systems in which the arrival rate depends on the congestion in the system. If the potential demand rate exceeds the system's capacity, then the fluid approximations are found to exhibit $O(1)$-accuracy—their error does not increase with system size. These fluid approximations are used to solve two capacity sizing problems: minimizing total system cost and maximizing social welfare. We find that the solutions to both these problems exhibit interesting differences, and further that under some conditions, the fluid prescriptions exhibit $o(1)$-optimality, that is, their optimality gap is asymptotically zero.

1. Introduction

This paper studies the accuracy of fluid approximations in queueing systems in which customers decide to join the system based on the congestion in the system, and if they do join the system, then they wait for their service without abandoning or reneging. Each customer is endowed with a valuation for the service and joins the system only if her valuation exceeds the expected cost of obtaining service, which includes the access fee and waiting costs. Such demand models have been used in [5,7,4,2,3], among others.

We find that if the maximum potential arrival rate to the system exceeds the capacity, then the fluid approximations to the expected steady-state number-in-system and customer loss rate are $O(1)$-accurate, that is, as the potential arrival rate increases without bound, the error in the fluid approximation remains bounded. We characterize the magnitude of this $O(1)$-error and find that it depends both on the maximum offered load and the valuation distribution. Further, we show that while the accuracy of the fluid approximation deteriorates as the maximum offered load approaches one, it need not be monotone in the offered load.

We use these fluid approximations to obtain simple solutions to capacity sizing problems, both from the firm's and the social planner's perspective. We first consider the firm's cost minimization problem of sizing capacity to minimize the sum total of capacity, holding, and penalty costs for lost customers, and then we consider the social planner's objective of maximizing the total social welfare, which consists of the utility of the joining customers minus the firm's capacity costs. For the firm's cost minimization problem, we find that the fluid optimal solution either satisfies all the potential demand or none of it. The reason for this is that deciding to serve the customers now only hinges on a comparison of the per-unit capacity and penalty costs. If the penalty cost is higher, the firm serves all demand, otherwise the firm is better off simply setting a zero capacity level to "close shop" and only incurring the penalty cost for lost customers. The customer's holding cost is immaterial in the solution to the fluid problem. This cost minimization problem can also be construed as the firm's profit maximization problem in which customers pay a fixed fee for access to the firm's services (the penalty cost plays the role of this fixed fee).

The social planner's problem of maximizing the social welfare exhibits a very different nature. For the social welfare to be maximized, serving none of the customers can become a costly proposition, and so the social planner could invest in some capacity to satisfy the demand partially. A sufficient condition for this to occur is that the hazard rate of the valuation distribution is decreasing. In this case, the fluid solution turns out to be $o(1)$-optimal. That is, the difference between the performance of the fluid prescription and the optimal performance is asymptotically zero.

The social welfare problem studied in this paper is similar to cost minimization in a model with customer abandonment, where customers wait in the system until their impatience clock runs out, and thus in a sense they "transfer" their cost of not being served to

E-mail address: ramandeep.randhawa@marshall.usc.edu.

0167-6377/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2012.10.009

the system through holding costs. This abandonment model was studied in [1], and fluid-based prescriptions were found to be $O(1)$-optimal.

2. Accuracy of fluid approximations

Consider a single server queueing system to which customers arrive in the form of a Poisson process. The maximum arrival rate possible is denoted by λ, and we assume that customers have service requirements that are exponentially distributed with unit mean and are i.i.d. across customers. We denote the service rate of the server by μ_λ and consider the case in which $\mu_\lambda = \lambda/r$, where $r > 1$ denotes the maximum offered load. Each customer is endowed with a maximum time that she is willing to wait, and if this time is less than the expected steady-state sojourn time, then she prefers not to join the system. This willingness-to-wait is i.i.d. across customers, and the c.d.f. of its distribution is denoted by F. We assume that F is associated with a density f. Equivalently, we can consider a setting in which each customer is endowed with a valuation for the service, V, incurs a cost of h per unit of time spent waiting in the system and pays a fee of p to obtain service so that she joins the system only if the expected sojourn time is less than $(V-p)/h$, which can be interpreted as the willingness-to-wait and is distributed according to F.

Denoting the expected steady-state sojourn time by w_λ, it follows that the actual arrival rate of customers to the system equals $\lambda\bar{F}(w_\lambda)$, where $\bar{F} = 1 - F$. Using the expression for the expected steady-state sojourn time in an $M/M/1$ queue, it follows that w_λ solves the following equilibrium condition, the validity of which follows from [5]:

$$w_\lambda = \frac{1}{\mu_\lambda - \lambda\bar{F}(w_\lambda)}. \tag{1}$$

We now analyze the fluid version of the problem. We will argue informally to identify the limit and its accuracy, and then formally state and prove the result in Proposition 1 (the proof appears in the Appendix). Let $\bar{w}$ denote the sojourn time of customers in the fluid model. That is, all arriving customers experience a fixed delay of $\bar{w}$ before entering service. This suggests that the customers who now arrive to the system in the form of a fluid flow should do so at the rate $\lambda\bar{F}(\bar{w})$ and will be processed at a fixed rate μ_λ. One expects the inflow rate to equal the outflow rate, which suggests that the fluid limit should satisfy the following relation:

$$\lambda\bar{F}(\bar{w}) = \mu_\lambda \Leftrightarrow r\bar{F}(\bar{w}) = 1. \tag{2}$$

We now use the sojourn time of customers in the fluid model $\bar{w}$ as an approximation to the pre-limit system to see how well it performs. Noting that the expected sojourn time in the pre-limit system w_λ must exceed the fluid value $\bar{w}$ (otherwise the queue would be unstable), we can write $w_\lambda = \bar{w} + \Delta_\lambda$, where Δ_λ is a small term that should shrink to zero as λ grows without bound. We next use the Taylor series expansion for $\bar{F}(w_\lambda)$ around $\bar{w}$. In order to do so, we make the following assumption on F.

Assumption 1. If $r > 1$, then for $\bar{w}$ that solves (2), the density of the customer's willingness-to-wait distribution is strictly positive at $\bar{w}$ and is continuously differentiable on $[\bar{w}-\epsilon, \bar{w}+\epsilon]$ for some $\epsilon > 0$.

Using the Taylor series expansion in (1), we obtain

$$\bar{w} + \Delta_\lambda = \frac{1}{\lambda f(\bar{w})\Delta_\lambda + O(\lambda\Delta_\lambda^2)}. \tag{3}$$

It follows that

$$\Delta_\lambda = \frac{1}{\lambda\bar{w}f(\bar{w})} + o(1/\lambda).$$

Now, consider the expected steady-state number of customers in the system. Applying Little's law, we have

$$\begin{aligned}\mathbb{E}Q_\lambda &= \lambda\bar{F}(w_\lambda)w_\lambda = \lambda(\bar{F}(\bar{w}) - f(\bar{w})\Delta_\lambda)(\bar{w}+\Delta_\lambda) + o(1)\\ &= \lambda\bar{F}(\bar{w})\bar{w} + \left(\frac{1}{\bar{w}H(\bar{w})} - 1\right) + o(1),\end{aligned} \tag{4}$$

where $H(w) = f(w)/\bar{F}(w)$ for $w \geq 0$ denotes the hazard rate of the willingness-to-wait distribution. Denoting the fluid approximation by $\lambda\bar{q}$, where

$$\bar{q} = \bar{F}(\bar{w})\bar{w}, \tag{5}$$

we have the following result.

Proposition 1. *If the maximum offered load exceeds unity and Assumption 1 holds, then the fluid approximation to the expected steady-state number of customers in the system is $O(1)$-accurate. In particular, if $r > 1$, then $(\mathbb{E}Q_\lambda - \lambda\bar{q}) \to \left(\frac{1}{\bar{w}H(\bar{w})} - 1\right)$ as $\lambda \to \infty$.*

Table 1 displays some numerical experiments to illustrate Proposition 1. In these experiments, we fix the customer's willingness-to-wait distribution to be exponential with unit mean and vary the maximum potential arrival rate λ and the maximum offered load r. In all cases, the $O(1)$ behavior of the approximation is apparent. For maximum offered loads close to 1, as expected, the fluid approximation is not very accurate. However, as the maximum offered load increases, the accuracy increases. Another trend to observe in the table is that the approximation error increases with λ and converges to $[\bar{w}H(\bar{w})]^{-1} - 1$ for large λ (the case $\lambda = \infty$ displays this constant in the table). Note that the accuracy of the fluid approximation depends on the product $\bar{w}H(\bar{w})$. If this product is high, then the fluid approximation will be quite accurate. For the exponential distribution, the hazard rate is a constant (in this case unity), so the accuracy directly depends on $\bar{w}$. As r approaches 1, $\bar{w}$ approaches 0, and we obtain a loss in accuracy. As the maximum offered load r increases, the fluid sojourn time $\bar{w}$ increases, and thus the accuracy improves. When $\bar{w}$ equals 1 (which occurs for $r = e$), this *error becomes zero* (or more precisely $o(1)$). As the maximum offered load increases further, $\bar{w}$ increases beyond 1 and the fluid approximation now *exceeds* the expected steady-state number-in-system (see the cases $r = 3$ and 5 in the table). The increase in $\bar{w}$ beyond 1 now leads to a decrease in accuracy, and the fluid approximation moves away from the actual quantity as demonstrated (the error is greater for $r = 5$ compared with $r = 3$ in the table). However, this gap is bounded below by -1.

To demonstrate the effect of the willingness-to-wait distribution on the accuracy of the fluid approximation, we consider another example, that of a normal distribution with unit mean and a standard deviation of 0.2 that is truncated to be non-negative. Table 2 displays the results. In this case, we only considered maximum offered loads close to 1. We find that the fluid approximation works very well in this case, and it is only as the maximum arrival rate becomes very close to the capacity ($r = 1.01$) that the errors become substantive. This example illustrates that the accuracy of the fluid approximation is driven not by the offered load r by itself, but rather through the term $[\bar{w}H(\bar{w})]^{-1} - 1$.

We end this discussion by considering the rate of customer loss from the system, i.e., the rate of customers not joining the system. This rate is given by $\lambda F(w_\lambda)$. Arguing as before, we obtain

$$\begin{aligned}\lambda F(w_\lambda) &= \lambda F(\bar{w}+\Delta_\lambda) = \lambda F(\bar{w}) + \lambda f(\bar{w})\Delta_\lambda + O(\lambda\Delta_\lambda^2)\\ &= \lambda F(\bar{w}) + \frac{1}{\bar{w}} + o(1).\end{aligned} \tag{6}$$

In this case, the approximation error does not depend on the hazard rate of the willingness-to-wait distribution. The following proposition states the formal result.

Table 1
The performance of the fluid approximation for the expected steady-state number-in-system for an exponential distribution $F(x) = 1 - \exp(-x)$ for different maximum offered loads.

λ	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$
	$r = 1.1$			$r = 1.2$			$r = 1.3$		
25	5.18	2.17	3.01	6.01	3.80	2.19	6.71	5.05	1.66
50	8.43	4.33	4.10	10.42	7.60	2.82	12.13	10.09	2.04
100	13.96	8.66	5.30	18.57	15.20	3.37	22.52	20.18	2.34
200	23.81	17.33	6.48	34.19	30.39	3.80	42.91	40.36	2.55
∞	–	–	9.5	–	–	4.5	–	–	2.8
	$r = 1.5$			$r = 3$			$r = 5$		
25	7.77	6.76	1.01	9.03	9.16	−0.13	7.7	8.0	−0.4
50	14.71	13.52	1.19	18.20	18.31	−0.11	15.7	16.1	−0.4
100	28.34	27.03	1.31	36.62	36.52	−0.10	31.8	32.2	−0.4
200	55.45	54.06	1.39	73.14	73.24	−0.10	64.0	64.4	−0.4
∞	–	–	1.5	–	–	–0.1	–	–	−0.4

Table 2
The performance of the fluid approximation for the expected steady-state number-in-system for a normal distribution with mean 1 and standard deviation 0.2.

λ	$r = 1.01$			$r = 1.05$			$r = 1.1$		
	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$	$\mathbb{E}Q_\lambda$	$\lambda\bar{q}$	$\mathbb{E}Q_\lambda - \lambda\bar{q}$
25	16.33	13.22	3.12	16.74	15.86	0.88	16.85	16.66	0.19
50	31.24	26.43	4.81	32.95	31.73	1.22	33.64	33.32	0.32
100	59.61	52.87	6.74	64.95	63.46	1.49	67.04	66.63	0.41
200	114.4	105.7	8.66	128.58	126.92	1.66	133.73	133.27	0.46
∞	–	–	13.0	–	–	1.9	–	–	0.5

Proposition 2. *If the maximum offered load exceeds unity and Assumption 1 holds, then the fluid approximation to the rate of customer loss is $O(1)$-accurate. In particular, if $r > 1$, then $(\lambda F(w_\lambda) - \lambda F(\bar{w})) \to 1/\bar{w}$ as $\lambda \to \infty$.*

Notice that in this case the approximation is not as accurate as in the case of abandonments in which the approximation error in fact decreases with system size (cf. [1]).

Remark 1 (*Relaxing* Assumption 1)**.** Consider the case in which the density of the willingness-to-wait distribution is zero at $\bar{w}$ and is fairly smooth as well in the vicinity of $\bar{w}$. That is, $f^{(\ell)}(\bar{w}) := \frac{\partial^\ell f(\bar{w})}{\partial w^\ell} = 0$ for all $\ell < k$ for some $k > 0$. In this case, the Taylor series expansion used in (3) yields $\Delta_\lambda = \left(\frac{(k+1)!}{\lambda \bar{w} f^{(k)}(\bar{w})}\right)^{\frac{1}{k+1}} = O(\lambda^{-\frac{1}{k+1}})$. The error in the fluid approximation for the customer loss rate can then be computed as $(\lambda F(w_\lambda) - \lambda F(\bar{w})) = 1/\bar{w} + o(1)$, which remains unchanged. However, the approximation error for the expected steady-state number-in-system is now $(\mathbb{E}Q_\lambda - \lambda\bar{q}) = \lambda^{\frac{k}{k+1}} \bar{F}(\bar{w}) \left(\frac{(k+1)!}{\bar{w} f^{(k)}(\bar{w})}\right)^{\frac{1}{k+1}} o(\lambda^{\frac{k}{k+1}})$, and thus the accuracy of this approximation is $O(\lambda^{\frac{k}{k+1}})$. Note that as the willingness-to-wait distribution becomes flatter near $\bar{w}$, this approximation error deteriorates further and can be arbitrary close to $O(\lambda)$. These observations are similar to those for systems with abandonments in which the error of the fluid approximation for the net rate of abandonments is independent of distributional assumptions, whereas that of the expected number-in-system depends on the behavior of the willingness-to-wait distribution in the vicinity of the fluid limit.

Remark 2 (*Generally Distributed Service Requirements*)**.** Suppose customer service requirements have unit mean and coefficient of variation c_s. Then, we can still characterize the accuracy of the fluid approximations using the $M/G/1$ steady-state formulas. Defining $\sigma^2 = (1 + c_s^2)/2$, we can write the analog of (1) as

$$w_\lambda = \frac{\lambda \bar{F}(w_\lambda)}{\mu_\lambda} \frac{\sigma^2}{(\mu_\lambda - \lambda \bar{F}(w_\lambda))} + \frac{1}{\mu_\lambda}.$$

Thus, if Assumption 1 holds, then we can argue as in the exponentially distributed service time case to compute $\Delta_\lambda = \sigma^2/(\lambda \bar{w} f(\bar{w})) + o(1/\lambda)$, and we obtain

$$\mathbb{E}Q_\lambda = \lambda \bar{F}(\bar{w})\bar{w} + \sigma^2 \left([\bar{w}H(\bar{w})]^{-1} - 1\right) + o(1),$$

$$\lambda F(w_\lambda) = \lambda F(\bar{w}) + \frac{\sigma^2}{\bar{w}} + o(1).$$

So, we find that the accuracy of the fluid approximation decreases as σ^2 increases, i.e., as service times become more variable.

3. Capacity sizing

In this section, we consider the capacity sizing problem of selecting the service rate of the single server. Throughout this section, we set the cost per unit of service rate to c per unit time and assume that the distribution F has a strictly positive density on $[0, \infty)$ that is continuously differentiable (so that Assumption 1 holds).

3.1. Capacity sizing to minimize firm costs

The system incurs customer holding costs at rate h per customer per unit of time spent in system, and there is a penalty of p per customer who does not join the system. The objective is to select capacity μ to minimize the sum total of holding, penalty, and capacity costs—i.e., $h\mathbb{E}Q_\lambda + pb_\lambda + c\mu$. Note that this cost minimization problem can equivalently be cast as a profit maximization problem, with p interpreted as the access fee customers pay for obtaining service.

Denoting w_λ as the expected steady-state sojourn time, we can write $\mathbb{E}Q_\lambda = \lambda \bar{F}(w_\lambda) w_\lambda$, and $b_\lambda = \lambda F(w_\lambda)$. Thus, the cost minimization problem can be written as

$$\begin{aligned} &\min_{\mu>0, w_\lambda>0} \Pi_\lambda(\mu) := h\lambda \bar{F}(w_\lambda) w_\lambda + p\lambda F(w_\lambda) + c\mu, \\ &\text{s.t. } w_\lambda = \frac{1}{\mu - \lambda \bar{F}(w_\lambda)}. \end{aligned} \tag{7}$$

Note that if $\mu = 0$, then none of the customers join the system, and we obtain $\Pi_\lambda(0) = p\lambda$. Let Π_λ^* denote the optimal total cost.

The fluid analog of the above optimization problem is obtained by replacing w_λ by its fluid approximation, and is given as follows:

$$\min_{\mu\geq 0} h\lambda\bar{F}(\bar{w})\bar{w} + p\lambda F(\bar{w}) + c\mu,$$
$$\text{s.t. } \lambda\bar{F}(\bar{w}) = \mu.$$

This optimization problem is equivalent to:

$$\min_{\bar{w}\geq 0} \hat{\Pi}(\bar{w}) := \bar{F}(\bar{w})(h\bar{w} + (c-p)). \tag{8}$$

The solution to this problem can be obtained by comparing the cost of capacity with the penalty cost of lost customers. If the per unit penalty cost exceeds the per unit capacity cost, i.e., $p \geq c$, then because $c - p \leq 0$, we obtain $\hat{\Pi}(\bar{w}) \geq \hat{\Pi}(0) = (c-p)$ for all $\bar{w} \geq 0$ so that it would be optimal to set $\bar{w} = 0$ and serve all customers. On the other hand, if $c > p$, then $\hat{\Pi}(\bar{w}) > 0$ for all $0 \leq \bar{w} < \infty$, and $\hat{\Pi}(\infty) = 0$ so that it would be optimal to set $\bar{w} = \infty$, which is equivalent to setting $\mu = 0$, and we would obtain the trivial solution in which no capacity is invested in. The following proposition formally states this result.

Proposition 3. *The solution to (8) is either to meet all demand or to meet none. If the per unit penalty cost exceeds the per unit cost of capacity, then the fluid capacity prescription meets all demand, i.e., $\bar{\mu}^*_\lambda = \lambda$, otherwise the fluid optimal decision is to invest in zero capacity, i.e., $\bar{\mu}^*_\lambda = 0$. Further, the fluid capacity prescription is asymptotically optimal in the sense that $\Pi^*_\lambda = \Pi_\lambda(\bar{\mu}^*_\lambda) + o(\lambda)$.*

3.2. Capacity sizing to maximize social welfare

The optimization in the previous section considered the perspective of the firm and showed that serving customers is an all-or-nothing proposition. In this section, we investigate whether the same holds for the firm and customers put together, that is, we optimize the social welfare of the system. We now interpret p as the price the customers pay for service and the customers' willingness-to-wait as their willingness-to-pay or valuation. We assume that customers have i.i.d. valuations for the service; the corresponding distribution is denoted by F and has finite mean. So, customers join the system if their valuation exceeds the cost of obtaining service, which includes the fee p and the cost of waiting.

For a system with steady-state expected sojourn time of w_λ, customers with valuation exceeding $p + hw_\lambda$ join the system. Thus, the social welfare at a capacity level μ comprises the firm's capacity costs $c\mu$ and the value gained by the customers (ignoring the fees paid to the firm), which equals $(v - hw_\lambda)$ for a customer with valuation v, and thus $\lambda \int_{p+hw_\lambda}^{\infty} (v - hw_\lambda) f(v)dv$ in expectation over all customers. It is worth noting that now customers may not join due to both the price and the waiting time. So, we modify our terminology of maximum offered load to denote the ratio of the maximum potential arrival rate at the given price to the system capacity.

The system manager's objective is to select capacity to maximize the social welfare. The following is the optimization problem.

$$\max_{\mu>0, w_\lambda>0} \Pi_\lambda(\mu) := \lambda \int_{p+hw_\lambda}^{\infty} (v - hw_\lambda) f(v)dv - c\mu, \tag{9}$$
$$\text{s.t. } w_\lambda = \frac{1}{\mu - \lambda\bar{F}(p + hw_\lambda)}. \tag{10}$$

Note that if $\mu = 0$, then none of the customers join the system, and the social welfare equals 0, i.e., $\Pi_\lambda(0) = 0$. Let Π^*_λ denote the optimal social welfare.

The fluid analog of the above optimization problem is as follows:

$$\min_{\mu\geq 0} \bar{\Pi}_s(\mu) := \int_{p+h\bar{w}(\mu)}^{\infty} (v - h\bar{w}(\mu)) f(v)dv - c\mu,$$

where $\bar{w}(\mu)$ solves $\bar{F}(p + h\bar{w}) = \mu$. This optimization problem can be recast in the following simpler form:

$$\min_{\bar{w}\geq 0} \hat{\Pi}_s(\bar{w}) := \int_{p+h\bar{w}}^{\infty} (v - h\bar{w}) f(v)dv - c\bar{F}(p + h\bar{w}). \tag{11}$$

Note that we have $\hat{\Pi}'_s(\bar{w}) = h\bar{F}(p + h\bar{w})\Big[(c-p)H(p + h\bar{w}) - 1\Big]$. Thus, the first-order optimality condition is

$$H(p + hw_i) = \frac{1}{c-p}. \tag{12}$$

It follows that (11) either has corner solutions, 0 or ∞, or interior solutions, w_i, that satisfy (12) (there may be multiple optima). Let $\bar{w}^*$ denote the largest maximizer of (11). When $\bar{w}^* = 0$, then the fluid prescription satisfies all the potential demand and we obtain $o(\lambda)$ accuracy of the prescription. However, an interior solution leads to an "overloaded regime" in which the fluid approximations are extremely accurate and in this case interestingly we obtain $o(1)$-accuracy of the fluid prescription. That is, the difference between the actual optimal social welfare (solution to (9)) and the social welfare obtained by using the fluid prescription is asymptotically zero. This is interesting because even though the fluid approximation to the actual objective is accurate only up to $O(1)$, the corresponding prescription when applied to the system exhibits a higher $o(1)$ accuracy (see [6] for a general version of this property).

To see when the fluid optimization problem would have an interior solution, consider $\hat{\Pi}'_s$ and observe that if the hazard rate of the valuation distribution is monotone, then $\hat{\Pi}_s$ is unimodal. In case the hazard rate is decreasing and a solution w_i to (12) exists, then it would be the unique (interior) maximizer of $\hat{\Pi}_s$. On the other hand, if the hazard rate is increasing, then the fluid problem can only have corner solutions. This is formalized in the following result.

Proposition 4. *We have*

1. *The capacity prescription $\bar{\mu}^*_\lambda = \lambda\bar{F}(p + h\bar{w}^*)$ is asymptotically optimal in the sense that $\Pi^*_\lambda = \Pi_\lambda(\bar{\mu}^*_\lambda) + o(\lambda)$.*
2. *If $0 < \bar{w}^* < \infty$, then the capacity prescription is $o(1)$-optimal—i.e., $(\Pi^*_\lambda - \Pi_\lambda(\bar{\mu}^*_\lambda)) \to 0$ as $\lambda \to \infty$.*
3. *If the hazard rate of the valuation distribution is decreasing and there exists a solution $w_i \in (0, \infty)$ to (12), then $\bar{w}^* = w_i$ and the fluid capacity prescription is $o(1)$-optimal and leads to a regime with maximum offered load greater than 1.*
4. *If the hazard rate of the valuation distribution is increasing, then $\bar{w}^* = 0$ or ∞—i.e., fluid capacity prescription equals $\lambda\bar{F}(p)$ or 0, respectively.*

Remark 3 (*The Case of Endogenous Prices*)**.** If the system manager can set prices in addition to capacity, one expects the fluid solution to always be critically loaded, that is, to have maximum offered load equal to unity. To see this, we compute $\frac{\partial\hat{\Pi}_s(\bar{w})}{\partial p} = (c-p)f(p + h\bar{w})$. It follows that $\hat{\Pi}_s$ is maximized at $p = c$. Now, we note that for $p = c$, $\frac{\partial\hat{\Pi}_s(\bar{w})}{\partial\bar{w}} = -h\bar{F}(c + h\bar{w}) < 0$, and thus we obtain $\bar{w}^* = 0$. That is, the optimal solution is critically loaded.

4. Many-server configuration

The analysis of the previous sections applies to the many-server configuration as well. In this case, the system consists of multiple identical servers, each with a service rate denoted by μ. We set the number of servers $s_\lambda = \frac{\lambda}{r\mu}$ for some $r > 1$ (we will ignore integrality considerations for convenience). Analogous to (1), we can write the expected steady-state time in the queue as

$$w_\lambda = \frac{\alpha(\lambda\bar{F}(w_\lambda), s_\lambda)}{s_\lambda\mu - \lambda\bar{F}(w_\lambda)}, \tag{13}$$

where $\alpha(\ell, s) = \frac{\frac{(s\rho)^s}{s!(1-\rho)}}{\frac{(s\rho)^s}{s!(1-\rho)} + \sum_{k=0}^{s-1}\frac{(s\rho)^k}{k!}}$, where $\rho = \frac{\ell}{s\mu}$. As in the single server case, we observe that if the maximum offered load exceeds 1, i.e., $r > 1$, then we must have $\liminf_{\lambda\to\infty} w_\lambda > 0$. This implies that we must have $\alpha(\lambda\bar{F}(w_\lambda), s_\lambda) \to 1$ as $\lambda \to \infty$. Using this in (13) and defining $\bar{w}$ as the solution to $r\bar{F}(\bar{w}) = 1$, we obtain that $w_\lambda = \bar{w} + \frac{1}{\lambda\bar{w}f(\bar{w})} + o(1/\lambda)$ exactly as in the single server case. It follows that, if $\bar{w} > 0$, then

$$\mathbb{E}Q_\lambda = \lambda\bar{F}(\bar{w})\bar{w} + [\bar{w}H(\bar{w})]^{-1} - 1 + o(1),$$
$$\lambda F(w_\lambda) = \lambda F(\bar{w}) + \frac{1}{\bar{w}} + o(1).$$

Analogous to Section 3, we can consider the system manager's problem of selecting capacity to either minimize costs or maximize social welfare. The fluid optimization problem here is identical to that for the single server case and one can establish an analog of Proposition 4. For brevity we omit the details.

Appendix. Proofs

We omit the proof of Proposition 3 because it is analogous to that of Proposition 4.

Proof of Propositions 1 and 2. We begin by observing that $w_\lambda \to \bar{w}$ as $\lambda \to \infty$. This follows because we must have $\liminf_{\lambda\to\infty} w_\lambda \geq \bar{w}$ for the stability of the queue, and then noting that if along a subsequence indexed by λ_k, we have $w_{\lambda_k} \geq \bar{w} + \epsilon$ for some $\epsilon > 0$ and all k large enough, then (1) would not hold along this subsequence. Thus, using the continuity of F, we obtain $\lambda F(w_\lambda) = \lambda F(\bar{w}) + o(\lambda)$, and applying Little's law, $\mathbb{E}Q_\lambda = \lambda\bar{q} + o(\lambda)$. Now, consider the case $1 < r < \infty$. Because $w_\lambda \to \bar{w}$, we can write $w_\lambda = \bar{w} + \Delta_\lambda$, where $\Delta_\lambda = o(1)$. Then, the result follows by using the Taylor series expansion of $\bar{F}(\bar{w}_\lambda)$ around $\bar{w}$ to derive (4) and (6). □

Proof of Proposition 4. We prove parts 1 and 2; the other parts follow by using these parts in conjunction with the discussion preceding the statement of Proposition 4. The following lemma will be useful in proving this result.

Lemma 1. *For an $M/M/1$ queue with potential arrival rate λ, service rate μ_λ, and customer valuation distribution F, in which customers incur an access fee of p and a per unit waiting cost of h, if $\lambda\bar{F}(p)/\mu_\lambda \to r \geq 1$, then we have $\Pi_\lambda(\mu_\lambda) = \lambda\bar{\Pi}_s(\mu_\lambda/\lambda) + o(\lambda)$. Further, if $1 < r < \infty$, then we have*

$$\limsup_{\lambda\to\infty}(\Pi_\lambda(\mu_\lambda) - \lambda\bar{\Pi}_s(\mu_\lambda/\lambda)) = C(\bar{w}) := -\frac{h}{\bar{w}}\left(p + \frac{1}{H(p+h\bar{w})}\right),$$

where $\bar{w}$ solves $r\bar{F}(p+h\bar{w}) = \bar{F}(p)$.

Proof of Lemma 1. Define $\bar{w}_\lambda$ as the solution to $\lambda\bar{F}(p+h\bar{w}_\lambda) = \mu_\lambda$. Then, for the stability of the queue, we must have $w_\lambda > \bar{w}_\lambda$. We further observe that $\lim_{\lambda\to\infty}(w_\lambda - \bar{w}_\lambda) = 0$. To see this, suppose that there exists a subsequence indexed by λ_k such that $w_{\lambda_k} - \bar{w}_{\lambda_k} \geq \epsilon$ for some $\epsilon > 0$ and all k large enough, then (10) would not hold along this subsequence. Thus, we can write $w_\lambda = \bar{w}_\lambda + \Delta_\lambda$, where $\Delta_\lambda = o(1)$. Note that the continuity of F implies that $\bar{w}_\lambda \to \bar{w}$ as $\lambda \to \infty$. Using this, we obtain $\lim_{\lambda\to\infty}\Pi_\lambda(\mu_\lambda)/\lambda = \lim_{\lambda\to\infty}\bar{\Pi}_s(\mu_\lambda/\lambda) = \hat{\Pi}_s(\bar{w})$, and this proves the first part of the result.

Next, consider the case $1 < r < \infty$. Applying the Taylor series expansion to $\bar{F}(p + hw_\lambda)$ around $\bar{w}_\lambda$ and arguing as in Section 2, we obtain $\Delta_\lambda = \frac{1}{\lambda\bar{w}_\lambda f(p+h\bar{w}_\lambda)} + o(1/\lambda)$. Using this and the Taylor series expansion of $\int_{p+hw_\lambda}^{\infty}(v - hw_\lambda)f(v)dv$ around $\bar{w}_\lambda$, we obtain $\Pi_\lambda(\mu_\lambda) = \lambda\bar{\Pi}_s(\mu_\lambda/\lambda) - \frac{h}{\bar{w}_\lambda}\left(p + \frac{1}{H(p+h\bar{w}_\lambda)}\right) + o(1)$. The result then follows by taking the limit and using $\bar{w}_\lambda \to \bar{w}$ as $\lambda \to \infty$. □

Let μ_λ^* and w_λ^* denote any optimizer of (9)–(10) and the corresponding expected steady-state sojourn time, respectively. Consider the sequence $\{\mu_\lambda^*\}$, and pick any convergent subsequence, i.e., $\mu_{\lambda_k}^*/\lambda_k \to \bar{\mu}$ as $k \to \infty$. Such a sequence must exist because we can bound $\mu_\lambda^*/\lambda \leq K$ for all λ large enough and some finite constant $K > 1$. Applying Lemma 1, we obtain that $\lim_{k\to\infty}\Pi_{\lambda_k}^*/\lambda_k = \bar{\Pi}_s(\bar{\mu}) = \hat{\Pi}_s(\bar{w})$, where $\bar{w}$ solves $\bar{F}(p + h\bar{w}) = \bar{\mu}$. Now, using the fact that $\bar{w}^*$ maximizes $\hat{\Pi}_s$, it follows that $\hat{\Pi}_s(\bar{w}) \leq \hat{\Pi}_s(\bar{w}^*)$. Thus, we have $\lim_{k\to\infty}\Pi_{\lambda_k}^*/\lambda_k \leq \hat{\Pi}_s(\bar{w}^*)$. Further noting that

$$\hat{\Pi}_s(\bar{w}^*) = \lim_{k\to\infty}\Pi_{\lambda_k}(\lambda_k\bar{F}(p+\bar{w}^*))/\lambda_k \leq \lim_{k\to\infty}\Pi_{\lambda_k}^*/\lambda_k = \hat{\Pi}_s(\bar{w}),$$

we obtain $\hat{\Pi}_s(\bar{w}) = \hat{\Pi}_s(\bar{w}^*)$, and thus, $\Pi_\lambda^* = \Pi_\lambda(\lambda\bar{F}(p+h\bar{w}^*)) + o(\lambda)$. For the case $0 < \bar{w}^* < \infty$, the proof of part 2 follows by combining Lemma 1 with Theorem 1 of [6]. □

References

[1] A. Bassamboo, R. Randhawa, On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers, Operations Research 58 (5) (2010) 1398–1413.
[2] I. Gurvich, M. Armony, C. Maglaras, Cross-selling in a call center with a heterogeneous customer population, Operations Research 57 (2) (2009) 299–313.
[3] S. Kumar, R. Randhawa, Exploiting market size in service systems, Manufacturing & Service Operations Management 12 (3) (2010) 511–526.
[4] C. Maglaras, A. Zeevi, Pricing and capacity sizing for systems with shared resources: approximate solutions and scaling relations, Management Science 49 (8) (2003) 1018–1038.
[5] H. Mendelson, S. Whang, Optimal incentive-compatible priority pricing for the $M/M/1$ queue, Operations Research 38 (5) (1990) 870–883.
[6] R.S. Randhawa, The optimality gap of asymptotically-derived prescriptions with applications to queueing systems, Working Paper.
[7] W. Whitt, How multiserver queues scale with growing congestion-dependent demand, Operations Research 51 (4) (2003) 531–542.