OPERATIONS RESEARCH
DOI 10.1287/opre.1070.0464ec
pp. ec1–ec17

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Dimensioning Large-Scale Membership Services" by Francis de Véricourt and Otis B. Jennings, *Operations Research*, DOI 10.1287/opre.1070.0464.

Online Companion For: "Dimensioning Large-Scale Membership Services"

Francis de Véricourt
fdv1@duke.edu

Otis B. Jennings
otisj@duke.edu

Fuqua School of Business
Duke University
Durham, NC 27708,
U.S.A

This online companion is to accompany de Véricourt and Jennings, 2006, "Dimensioning Large Scale Services". With the exception of Theorem 9, all proofs of the paper are provided in this document. We also present numerical results for the model with cost. We conclude with some extensions of the model.

Appendix B: Proofs

B.1. Open vs. closed queueing system models

Proof of Proposition 2: From $(1-\rho_{s_n})\sqrt{s_n} \to \beta$ and Proposition 1 of Halfin and Whitt (1981), the limiting probability of delay of the open systems is equal to $\alpha_O(\beta)$. Note also that $(1-\rho_{s_n})\sqrt{s_n} \to \beta$ implies that $\rho_{s_n} \to 1$ which leads to $s_n/n \to r$. We can now determine the corresponding service grade for the closed systems by taking the following limit,

$$\lim_{n\to\infty}\left(\frac{s_n}{n}-r\right)\sqrt{n} = \lim_{n\to\infty} r\left(\frac{1}{\rho_{s_n}}-1\right)\sqrt{s_n}\sqrt{\frac{n}{s_n}} = \sqrt{r}\beta,$$

where the first equality follows from $\lambda_O^n = rn$ for all n, and the second equality follows from $(1-\rho_{s_n})\sqrt{s_n} \to \beta$. It follows from Proposition 1 that the limit of the approximate probability of delay is equal to $\alpha(\sqrt{r}\beta)$ for the closed systems. To obtain the last part of the result, note that

$$\begin{aligned}\alpha(\sqrt{r}\beta) < \alpha_O(\beta) &\Leftrightarrow \sqrt{r}e^{-\frac{\beta^2}{2r}}\frac{\Phi\left(\frac{\beta}{\sqrt{\bar{r}}}\right)}{\Phi\left(\frac{-\beta}{\sqrt{r\bar{r}}}\right)} > e^{\beta^2/2}\beta\sqrt{2\pi}\Phi(\beta)\\ &\Leftrightarrow \sqrt{r}h\left(\frac{\beta}{\sqrt{r\bar{r}}}\right)h(-\beta) > \beta\, h\left(\frac{-\beta}{\sqrt{\bar{r}}}\right),\end{aligned}$$

where $h(x) = \phi(x)/\Phi(-x)$ is the hazard rate of the normal distribution (with ϕ is the density of the standard normal distribution). The function $h(\cdot)$ is strictly increasing for all x (see Barlow and Proschan (1965)) and $h(x) > x$ for all x since $h'(x) = h(x)(h(x)-x) > 0$ (where the equality is obtained using $\phi'(x) = -x\phi(x)$). It follows that $h(-\beta) > h\left(\frac{-\beta}{\sqrt{\bar{r}}}\right)$ and $\sqrt{r}h\left(\frac{\beta}{\sqrt{r\bar{r}}}\right) > \frac{\beta}{\sqrt{\bar{r}}} > \beta$ from which we obtain the result. □

B.2. Convergence of distributions

In the following we provide proofs of Proposition 1 and Theorems 3 and 6 as well as some preliminary results.

For the purpose of this section, define for each real x and integer $n \geq 1$,

$$A_n(x) \equiv \sum_{k=0}^{\lfloor rn+x\sqrt{n}\rfloor} \binom{n}{k} \rho^k, \qquad A_n \equiv \sum_{k=0}^{s_n-1} \binom{n}{k} \rho^k = A_n((s_n - rn - 1)/\sqrt{n}),$$

$$B_n(x) \equiv \sum_{k=\lceil rn+x\sqrt{n}\rceil}^{n} \binom{n}{k} \frac{k!}{s_n!} s_n^{s_n-k} \rho^k \qquad \text{and} \qquad B_n \equiv \sum_{k=s_n}^{n} \binom{n}{k} \frac{k!}{s_n!} s_n^{s_n-k} \rho^k = B_n((s_n - rn)/\sqrt{n}).$$

Note that $A_n(x)$ and $B_n(x)$ can be rewritten as

$$A_n(x) = (1+\rho)^n \sum_{k=0}^{\lfloor rn+x\sqrt{n}\rfloor} \binom{n}{k} r^k \bar{r}^{n-k} = (1+\rho)^n P(X \leq rn + x\sqrt{n})$$

and

$$\begin{aligned} B_n(x) &= \frac{n!}{s_n!} \frac{\rho^n}{s_n^{n-s_n}} e^{s_n/\rho} \sum_{k=\lceil rn+x\sqrt{n}\rceil}^{n} \frac{1}{(n-k)!} \left(\frac{s_n}{\rho}\right)^{n-k} e^{-s_n/\rho} \\ &= \frac{n!}{s_n!} \frac{\rho^n}{s_n^{n-s_n}} e^{s_n/\rho} P(Y \leq \bar{r}n - x\sqrt{n}), \end{aligned}$$

respectively, where X is a binomial random variable with parameters n and r and Y is a Poisson random variable with rate s_n/ρ, i.e. $X \sim Bi(n, r)$ and $Y \sim P(s_n/\rho)$. Finally, applying the central limit theorem and (4) we have for any x,

$$P(X \leq rn + x\sqrt{n}) = P\left(\frac{X - rn}{\sqrt{nr\bar{r}}} \leq \frac{x}{\sqrt{r\bar{r}}}\right) \to \Phi\left(\frac{x}{\sqrt{r\bar{r}}}\right) \tag{25}$$

and

$$P(Y \leq \bar{r}n - x\sqrt{n}) = P\left(\frac{Y - s_n/\rho}{\sqrt{s_n/\rho}} \leq \frac{\bar{r}(rn - s_n)/\sqrt{n} - rx}{\sqrt{\bar{r}rs_n/n}}\right) \to \Phi\left(\frac{-rx - \bar{r}\beta}{r\sqrt{\bar{r}}}\right), \tag{26}$$

as $n \to \infty$.

The following lemma gives the limit of the condition steady state distributions under QED staffing. Let $\mathbf{N}_n(\infty)$ be defined as in Theorem 3 and $\mathbf{N}(\infty)$ be the steady state of the limiting diffusion process of Theorem 2.

LEMMA 1. *Under (4), for any $x < \beta$,*

$$\lim_{n\to\infty} P(N_n(\infty) \leq rn + x\sqrt{n} \,|\, N_n(\infty) < s_n) = \Phi\left(\frac{x}{\sqrt{r\bar{r}}}\right) \Big/ \Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right) \tag{27}$$

and for any $x \geq \beta$,

$$\lim_{n\to\infty} P(N_n(\infty) \leq rn + x\sqrt{n} \,|\, N_n(\infty) \geq s_n) = \left[\Phi\left(\frac{rx + \bar{r}\beta}{r\sqrt{\bar{r}}}\right) - \Phi\left(\frac{\beta}{r\sqrt{\bar{r}}}\right)\right] \Big/ \Phi\left(-\frac{\beta}{r\sqrt{\bar{r}}}\right). \tag{28}$$

Proof: Assume $x < \beta$. For sufficiently large n we have $rn + s\sqrt{n} < s_n$. Using the steady state distribution provided in (1), it follows that

$$P(N_n(\infty) \leq rn + x\sqrt{n} \,|\, N_n(\infty) < s_n) = \frac{A_n(x)}{A_n} = \frac{P(X \leq rn + x\sqrt{n})}{P(X \leq s_n - 1)}.$$

Equation (27) follows immediately from (4) and (25).

Now assume that $x \geq \beta$. Again using the distribution in (1), we have

$$P(N_n(\infty) \leq rn + x\sqrt{n} \,|\, N_n(\infty) \geq s_n) \approx \frac{B_n - B_n(x)}{B_n} = \frac{P(Y \leq n - s_n) - P(Y \leq \bar{r}n - x\sqrt{n})}{P(Y \leq n - s_n)}.$$

Equation (28) follows immediately from (4) and (26).□

We now verify that the approximate delay probability has a nondegenerate limit under the QED staffing assumption.

Proof of Proposition 1: Using the distribution from (1),

$$P(N_n \geq s_n) = \left(1 + \frac{A_n}{B_n}\right)^{-1}.$$

The approximate probability of delay can be written as

$$P(N_n \geq s_n) = \left(1 + C_n \frac{P(X \leq s_n - 1)}{P(Y \leq n - s_n)}\right)^{-1}, \tag{29}$$

where

$$C_n \equiv \frac{s_n!}{n!} \left(\frac{s_n}{\rho}\right)^n \frac{(1+\rho)^n}{s_n^{s_n}} e^{-s_n/\rho}.$$

To study the limit of C_n as $n \to \infty$, we apply the Stirling's formula $n! \sim (2\pi n)^{1/2} n^n e^{-n}$ twice, to n and to s_n, yielding

$$C_n \sim \sqrt{r} \left(1 + \frac{\beta}{r\sqrt{n}}\right)^{n+1/2} e^{n - s_n/r}. \tag{30}$$

Since $(1 + \beta/(r\sqrt{n}))^{n+1/2} \sim e^{\beta\sqrt{n}/r - \beta^2/2r^2}$ and $n - s_n/r = -\beta\sqrt{n}/r$ from (4), (25), (26), (29) and (30), we deduce that $C_n \to e^{-\beta^2/2r^2}\sqrt{r}$ and

$$P(N_n \geq s_n) \to f(\beta) = \left(1 + e^{-\beta^2/2r^2}\sqrt{r} \frac{\Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right)}{\Phi\left(\frac{-\beta}{r\sqrt{\bar{r}}}\right)}\right)^{-1},$$

a strictly decreasing function mapping the reals to $(0, 1)$.

Suppose that a_n has a limit $\alpha \in (0, 1)$ and that $\beta \neq f^{-1}(\alpha)$ is a (possibly infinite) limit point of $\{(s_n/n - r)\sqrt{n}\}$. Assume for now that $\beta > f^{-1}(\alpha)$. Construct a sequence $\{s'_n\}$ such that $s'_n \leq s_n$ and $(s'_n/n - r)\sqrt{n} \to \beta' = \min((\beta + f^{-1}(\alpha))/2, f^{-1}(\alpha) + 1)$, as $n \to \infty$. Notice that $f^{-1}(\alpha) < \beta' < \infty$, which implies $\alpha > f(\beta')$. Let N'_n denote the number of users in the nth system with s'_n servers.

Since $s'_n \leq s_n$, $P(N'_n \geq s'_n) \geq P(N_n \geq s_n)$. However, taking the limit of both sides yields $f(\beta') \geq \alpha$, a contradiction. A similar argument shows that $\beta < f^{-1}(\alpha)$ is also impossible. Hence, the convergence $a_n \to \alpha \in (0,1)$ implies $\{(s_n/n - r)\sqrt{n}\}$ has a unique finite limit as well.□

Proof of Theorem 3: The proof of this theorem hinges on the first convergence in (13), which follows from Proposition 1 and Lemma 1 above. The remaining results are consequences of (13) and the continuous mapping theorem. □

Proof of Theorem 6: The limiting value of the approximate delay probability follows from Proposition 1. As in the proof of Theorem 3 above, one only needs to prove that $\mathbf{N}_n(\infty) \Rightarrow \mathbf{N}(\infty) \sim Normal(0, \bar{r}\bar{s}/r)$. Recall that in the ED regime $\bar{s} < r$. For sufficiently large n (such that $s_n < n(1 - \bar{r}\bar{s}/r)$), one can express the distribution of $\mathbf{N}_n(\infty)$ as

$$\begin{aligned} P(\mathbf{N}_n(\infty) \leq x) &= P(N_n(\infty) \leq n(1 - \bar{r}\bar{s}/r) + x\sqrt{n}) \\ &= P(N_n(\infty) < s_n) + \left(\frac{P(s_n \leq N_n(\infty) \leq n - \bar{r}\bar{s}n/r + x\sqrt{n})}{P(N_n(\infty) \geq s_n)} \right) P(N_n(\infty) \geq s_n) \\ &= 1 - \alpha_n + \left(\frac{P(Y \leq n - s_n) - P(Y \leq \bar{r}\bar{s}n/r - n\sqrt{n})}{P(Y \leq n - s_n)} \right) \alpha_n, \end{aligned}$$

for any real x. By Proposition 1 and the fact that $\bar{s} < r$, $\alpha_n \to 1$. The condition $\bar{s} < r$ and (26) imply $P(Y \leq n - s_n) \to 1$. Finally, by the central limit theorem and the fact that $\sqrt{n}|\bar{s} - s_n/n| \to 0$ as $n \to \infty$,

$$P(Y \leq \bar{r}\bar{s}n/r - n\sqrt{n}) = P\left(\frac{Y - s_n/\rho}{\sqrt{s_n/\rho}} \leq \frac{(\bar{r}\bar{s}n/r - s_n/\rho)/\sqrt{n} - x}{\sqrt{\frac{r}{r}\frac{s_n}{n}}} \right) \to \Phi\left(\frac{x}{\sqrt{\frac{\bar{r}\bar{s}}{r}}} \right)$$

and, hence, $P(N_n(\infty) \leq x) \to \Phi(x/\sqrt{\frac{\bar{r}\bar{s}}{r}})$ as well. This concludes the proof. □

Proof of Theorem 8: The limiting value of the approximate delay probability follows from Proposition 1. To complete the proof, one only needs to prove that $\mathbf{N}_n(\infty) \Rightarrow \mathbf{N}(\infty) \sim Normal(0, r\bar{r})$. Recall that in the QD regime $\bar{s} > r$. One can express the distribution of $\mathbf{N}_n(\infty)$ as

$$\begin{aligned} P(\mathbf{N}_n(\infty) \leq x) &= P(N_n(\infty) \leq rn + x\sqrt{n}) \\ &= \left(\frac{P(N_n(\infty) \leq rn + x\sqrt{n})}{P(N_n(\infty) \leq s_n)} \right) P(N_n(\infty) \leq s_n) \\ &= \left(\frac{P(X \leq rn + x\sqrt{n})}{P(X \leq s_n)} \right) (1 - \alpha_n), \end{aligned}$$

for any x and sufficiently large n so that $s_n > rn + x\sqrt{n}$. From Proposition 1 and (25) it follows that $P(\mathbf{N}_n(\infty) \leq x) \to \Phi(x/\sqrt{r\bar{r}})$. □

B.3. Convergence of sequences of processes

Proof of Theorem 1: We follow the framework of Browne and Whitt (1995) which summarizes a useful theorem of Stone (1963). For each integer $n \geq 1$, the process N_n is a birth-death process with state space $\{0, 1, \ldots, n\}$, state-dependent arrival rate $\lambda_n(j) = (n-j)\lambda$, and state-dependent service rate $\mu_n(j) = (j \wedge s_n)\mu$. The fluid scaled process $\bar{\mathbf{N}}_n$ has the state space $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$ and drift and diffusion functions

$$\bar{m}_n(x) = \frac{\lambda_n(\lfloor nx \rfloor)}{n} - \frac{\mu_n(\lfloor nx \rfloor)}{n} = \frac{(n - \lfloor nx \rfloor)\lambda}{n} - \frac{(\lfloor nx \rfloor \wedge s_n)\mu}{n}$$

and

$$\bar{\sigma}_n^2(x) = \frac{\lambda_n(\lfloor nx \rfloor)}{n^2} + \frac{\mu_n(\lfloor nx \rfloor)}{n^2} = \frac{(n - \lfloor nx \rfloor)\lambda}{n^2} + \frac{(\lfloor nx \rfloor \wedge s_n)\mu}{n^2},$$

respectively, where $\lfloor x \rfloor$ is the largest integer no more than x. We take the limit of both quantities to obtain the infinitesimal mean and variance of our limiting diffusion:

$$\bar{m}(x) \equiv \lim_{n \to \infty} \bar{m}_n(x) = \lambda - (\lambda + \mu)x + \mu(x - \bar{s})^+ \tag{31}$$

and

$$\bar{\sigma}(x) \equiv \lim_{n \to \infty} \bar{\sigma}_n(x) = 0.$$

The limiting infinitesimal variance is zero, implying that the limiting diffusion is degenerate. Moreover, the deterministic initial point yields a deterministic path for our limiting process. The drift of the limiting process is (31), which coincides with (7).

For investigating the limiting value $\bar{\mathbf{b}}(\infty)$ we consider three cases, corresponding with the three limiting staffing regimes, separately.

Case 1: $\bar{s} = r$**.** If $\bar{\mathbf{b}}(0) < r$, then, as long as $\bar{\mathbf{b}}(t) < r$, we have

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda - (\lambda + \mu)\bar{\mathbf{b}}(t).$$

Solving for $\bar{\mathbf{b}}$ yields

$$\bar{\mathbf{b}}(t) = r + (\bar{\mathbf{b}}(0) - r)e^{-(\lambda+\mu)t},$$

a function that is strictly increasing, always less than r for finite t, and asymptotically equal to r as $t \to \infty$. Likewise, if $\bar{\mathbf{b}}(0) > r$, then

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t)) - \mu r = \lambda r - \lambda\bar{\mathbf{b}}(t),$$

the solution of which,

$$\bar{\mathbf{b}}(t) = r + (\bar{\mathbf{b}}(0) - r)e^{-\lambda t},$$

is strictly decreasing and asymptotically equal to r. Lastly, if $\bar{\mathbf{b}}(0) = r$, then $\bar{\mathbf{b}}(t) = r$ for all $t \geq 0$.

Case 2: $\bar{s} < r$**.** Suppose $\bar{\mathbf{b}}(0) < \bar{s}$. Then, for sufficiently small $t \geq 0$,

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t)) - \mu\bar{\mathbf{b}}(t) = \lambda - (\lambda + \mu)\bar{\mathbf{b}}(t),$$

whose solution is

$$\bar{\mathbf{b}}(t) = r + (\bar{\mathbf{b}}(0) - r)e^{-(\lambda+\mu)t}, \quad t \leq (\lambda + \mu)^{-1}\tau \equiv \ln\left(\frac{r - \bar{\mathbf{b}}(0)}{r - \bar{s}}\right).$$

At time τ, we have $\bar{\mathbf{b}}(\tau) = \bar{s}$. We know then that, for any initial value $\bar{\mathbf{b}}(0)$, there is some finite $t_0 \geq 0$ such that $\bar{\mathbf{b}}(t_0) \in [\bar{s}, \infty)$. In fact, because the drift at $\bar{s}$ of $\bar{\mathbf{b}}$ is positive, $\bar{\mathbf{b}}(t) \in [\bar{s}, \infty)$ and

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t)) - \mu\bar{s} = \lambda - \mu\bar{s} - \lambda\bar{\mathbf{b}}(t)$$

for all $t \geq t_0$. The solution to the above ODE is,

$$\bar{\mathbf{b}}(t) = \left(1 - \frac{\bar{r}\bar{s}}{r}\right) + \left(\bar{\mathbf{b}}(t_0) + \frac{\bar{r}\bar{s}}{r} - 1\right)e^{-\lambda(t-t_0)}, \quad t \geq t_0.$$

The function is monotonic and asymptotically approaches $\bar{\mathbf{b}}(\infty) = 1 - \frac{\bar{r}\bar{s}}{r}$. The trajectory is strictly monotonic if $\bar{\mathbf{b}}(0) \neq \bar{\mathbf{b}}(\infty)$.

Case 3: $\bar{s} > r$**.** Suppose $\bar{\mathbf{b}}(0) > \bar{s}$. Then, for sufficiently small $t \geq 0$,

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t)) - \mu\bar{s} = \lambda - \mu\bar{s} - \lambda\bar{\mathbf{b}}(t),$$

whose solution is

$$\bar{\mathbf{b}}(t) = \left(1 - \frac{\bar{r}\bar{s}}{r}\right) + \left(\bar{\mathbf{b}}(0) + \frac{\bar{r}\bar{s}}{r} - 1\right)e^{-\lambda t}, \quad t \leq \tau \equiv \lambda^{-1}\ln\left(\frac{r\bar{\mathbf{b}}(0) + \bar{r}\bar{s}}{\bar{s}}\right).$$

The process $\bar{\mathbf{b}}$ is decreasing since $\bar{s} > r$ and $\bar{\mathbf{b}}(0) > r$. At time $t = \tau$, we have $\bar{\mathbf{b}}(t) = \bar{s}$. It follows then that, for any initial value $\bar{\mathbf{b}}(0)$, there exists a finite $t_0 \geq 0$ such that $\bar{\mathbf{b}}(t_0) \in [0, \bar{s}]$. The drift at $\bar{s}$ of $\bar{\mathbf{b}}$ is negative. Hence, $\bar{\mathbf{b}}(t) \in [0, \bar{s}]$ and

$$\frac{d\bar{\mathbf{b}}}{dt}(t) = \lambda(1 - \bar{\mathbf{b}}(t)) - \mu\bar{\mathbf{b}}(t) = \lambda - (\lambda + \mu)\bar{\mathbf{b}}(t)$$

for all $t \geq t_0$. The solution to the above ODE is

$$\bar{\mathbf{b}}(t) = r + \left(\bar{\mathbf{b}}(t_0) - r\right)e^{-(\lambda+\mu)(t-t_0)}, \quad t \geq t_0.$$

This function is monotonic and asymptotically approaches r. The trajectory is strictly monotonic if $\bar{\mathbf{b}}(0) \neq r$ and constant otherwise. □

Proof of Theorem 2: We proceed in the same fashion as in the proof of Theorem 1. The difference here is that the limiting diffusion will not be degenerate. The state space of the scaled process $\mathbf{N}_n$ is $\{\frac{-rn}{\sqrt{n}}, \frac{1-rn}{\sqrt{n}}, \ldots, \frac{\bar{r}n-1}{\sqrt{n}}, \frac{\bar{r}n}{\sqrt{n}}\}$. The drift and the diffusion functions of $\mathbf{N}_n$ are

$$m_n(x) = \frac{(n - \lfloor rn + x\sqrt{n} \rfloor)\lambda}{\sqrt{n}} - \frac{(\lfloor rn + x\sqrt{n} \rfloor \wedge s_n)\mu}{\sqrt{n}}$$

and

$$\sigma_n^2(x) = \frac{(n - \lfloor rn + x\sqrt{n} \rfloor)\lambda}{n} + \frac{(\lfloor rn + x\sqrt{n} \rfloor \wedge s_n)\mu}{n},$$

respectively. The limiting drift function is piecewise linear, with two parts. On the interval $(-\infty, \beta)$ the drift is $m(x) = -(\lambda + \mu)x$. On the interval $[\beta, \infty)$ the drift is $m(x) = -\lambda x + \mu\beta$. The limiting diffusion function is a constant, $2\mu r$. By Stone's Theorem, the limiting process is a piecewise linear diffusion, with instantaneous drift given in (11) and infinitesimal variance equal to $2\mu r$; see, e.g. Browne and Whitt (1995). Also, by Browne and Whitt (1995), this process has a limiting distribution given by (12). Joint convergence of the scaled queue length and idle time process follows from the continuous mapping theorem; see, for example, Whitt (2002). □

Proof of Corollary 1: Convergence of the first three components follow from Theorem 1 and the continuous mapping theorem; see, e.g., Whitt (2002). The convergence of W_n follows from the convergence of $\bar{\mathbf{Q}}_n$, the weak law of large numbers, and the fact that when the number of activate users exceeds the number of servers, all s_n servers must be processing users.

We only sketch the proof that D_n converges in probability to a constant. As noted in the paper, the distribution of D_n is a change of measure of the conditional distribution of W_n, involving the distribution of $N_n(\infty)$ and the state-dependent collective activation rate $\lambda(n - N_n)$. The following asymptotic relations hold: $W_n(\infty) = \frac{(b-\bar{s})^+}{\mu\bar{s}} + o(1)$, $N_n(\infty) = nb + o(n)$, and hence, $\lambda(n - N_n(\infty)) = \lambda n(1-b) + o(n)$. One can take advantage of these relations and (3) to show the result. □

Proof of Theorem 4: Our proof is similar to the proof of Theorem 3 in Garnett et al. (2002) and relies on a corollary by Puhalskii (1994). Introduce the processes D_n and A_n which track the cumulative number of processed users activated users (including the initial active users at time 0), respectively, as functions of time. Let X_n and Y_n be fluid-scaled versions: $X_n(t) = (1/n)D_n$ and $Y_n \equiv (1/n)(A_n - s_n + 1)$. The virtual waiting time can be expressed $W_n(t) = (Z_n(t) - t)^+$, where

$$Z_n(t) = \inf\{s \geq 0 : D_n(s) \geq A_n(t) - (s_n - 1)\}.$$

It is instructive to express the cumulative number of activated users as $A_n(t) = N_n(0) + A_n^0(t)$, where $A_n^0(t)$ counts the number of users activated *after* time 0. By the law of large numbers, we

have $X_n \Rightarrow X$, $Y_n \Rightarrow Y$, and $Z_n \Rightarrow Z$, as $n \to \infty$, where $X(t) = Y(t) = \mu r t$, for all $t \geq 0$, and $Z(t) = t$. By Theorem 2, we have $\mathbf{N}_n \Rightarrow \mathbf{N}$, as $n \to \infty$. It follows that, as $n \to \infty$, $U_n \equiv \sqrt{n}(X_n - X) \Rightarrow U$ and $V_n \equiv \sqrt{n}(Y_n - Y) \Rightarrow V$, where

$$U(t) = \sqrt{\mu r} B_d(t) - \mu \int_0^t (\mathbf{N}(s) \wedge \beta)\, ds,$$

$$V(t) = \mathbf{N}(0) - \beta + \sqrt{\mu r} B_a(t) - \lambda \int_0^t \mathbf{N}(s)\, ds,$$

The processes B_d and B_a are independent, standard Brownian motions that together constitute the martingale component of $\mathbf{N}$. Using the corollary in Puhalskii (1994) we have, as $n \to \infty$, $\sqrt{n}(Z_n - Z) \Rightarrow \widetilde{Z}$, jointly with the convergence $\mathbf{N}_n \Rightarrow \mathbf{N}$, where

$$\widetilde{Z}(t) \equiv \frac{V(t) - U(Z(t))}{X'(Z(t))} = \frac{\mathbf{N} - \beta}{\mu r}.$$

The convergence in (15) follows from the continuous mapping theorem. □

Proof of Theorem 5: The state space of the scaled process $\mathbf{N}_n$ is $\{\frac{-bn}{\sqrt{n}}, \frac{1-bn}{\sqrt{n}}, \ldots, \frac{n-bn-1}{\sqrt{n}}, \frac{n-bn}{\sqrt{n}}\}$, where $b = 1 - \frac{\bar{r}\bar{s}}{r}$. The drift and the diffusion functions of $\mathbf{N}_n$ are

$$m_n(x) = \frac{(n - \lfloor bn + x\sqrt{n} \rfloor)\lambda}{\sqrt{n}} - \frac{(\lfloor bn + x\sqrt{n} \rfloor \wedge s_n)\mu}{\sqrt{n}}$$

and

$$\sigma_n^2(x) = \frac{(n - \lfloor bn + x\sqrt{n} \rfloor)\lambda}{n} + \frac{(\lfloor bn + x\sqrt{n} \rfloor \wedge s_n)\mu}{n},$$

respectively. Because $\sqrt{n}(s_n/n - \bar{s}) \to 0$, the limiting drift function is $\lim_{n\to\infty} m_n(x) = \lambda x, \forall x$. Whereas in the QED case the limiting drift function is piecewise defined, here the limiting function is simply linear. The limiting diffusion function is a constant, equal to $\sigma^2 = 2\mu\bar{s}$. By Stone's Theorem, the limiting process is an OU process. The steady state distribution of this process is Normal with mean zero and variance $\bar{r}\bar{s}/r$. Again, joint convergence of the scaled queue length and idle time process follows from the continuous mapping theorem. □

Proof of Theorem 7: The proof is similar to that of Theorem 5. The linear drift functions under ED and QD cases are both determined by which term ultimately dominates, either $bn + \sqrt{n}x$ or s_n. In the ED case, the first quantity is the larger for sufficiently large n, whereas in the QD case the second term is eventually larger. Further, when s_n is the larger of the two, there are always excess servers. As a result, an increase in x has a corresponding increase in the service rate; hence the μ term in the limiting drift function under QD staffing. Likewise when s_n is smaller, the servers are always busy, additional jobs are not accompanied with an increase in the service rate, and thus, under ED staffing, the limiting drift is absent the μ term. The limiting diffusion is similarly affected, because its value is $\sigma^2(x) = 2\mu(r \wedge \bar{s})$. □

B.4. Decreasing Probability of Delay

PROPOSITION 4. *$P_{n,s}(D_n > 0)$ is decreasing in s.*

Proof: From (2), the probability of delay is equal to $B(s)/(A(s)+B(s))$ where

$$A(s) \equiv \sum_{k=0}^{s}(n-k)\binom{n}{k}\rho^k = n\sum_{k=0}^{s}\binom{n-1}{k}\rho^k$$

and

$$B(s) \equiv \sum_{k>s}(n-k)\binom{n}{k}\frac{k!}{s!}s^{s-k}\rho^k = n\sum_{k>s}\binom{n-1}{k}\frac{k!}{s!}s^{s-k}\rho^k$$

with $A(s+1) > A(s)$ and $B(s) > n\sum_{k>s+1}\binom{n-1}{k}\frac{k!}{s!}s^{s-k}\rho^k > B(s+1)$. It follows that $A(s+1)B(s) - A(s)B(s+1) > 0$ which implies the result. □

B.5. Asymptotic PASTA

Proof of Theorem 9: Fix q. If $\beta = \infty$ then, from the proof of Proposition 1, we have that $\alpha = 0$. Likewise, $\beta = -\infty$ implies that $\alpha = 1$. Under both of these degenerate cases, we have $\alpha(q) = \alpha$, regardless of the size of q, provided it is finite. Now suppose that β is finite. By Theorems 2 and 3, we have that (24) holds. To see this, one only needs to integrate the density of the limiting diffusion process, provided in (12), over the appropriate intervals.

The quantity $\gamma_n(q)$ can be expressed as

$$\gamma_n(q) = \lim_{t\to\infty}\frac{\int_0^t 1_{(N_n(s)\geq s_n+q\sqrt{n})}dA_n(s)}{A_n(t)},$$

where A_n is the activation process. We can divide the numerator and denominator each by t and handle them separately. We have the equivalence:

$$\int_0^t 1_{(N_n(s)\geq s_n+q\sqrt{n})}dA_n(s) = B_n\left(\lambda\int_0^t (n-N_n(s))\,1_{(N_n(s)\geq s_n+q\sqrt{n})}\,ds\right), \tag{32}$$

where B_n is a rate 1 Poisson process. By elementary Markov chain results, we have $B_n(t)/t \to 1$ almost surely and

$$(1/t)\int_0^t (n-N_n(s))\,1_{(N_n(s)\geq s_n+q\sqrt{n})}\,ds \Rightarrow nP(N_n(\infty)\geq s_n+q\sqrt{n}) - \mathbf{E}[N_n(\infty)1_{(N_n(\infty)\geq s_n+q\sqrt{n})}]$$

as $t\to\infty$; see, e.g. Karlin and Taylor (1975). It follows that

$$\gamma_n(q) = \frac{n\alpha_n(q) - \mathbf{E}[N_n(\infty)1_{(N_n(\infty)\geq s_n+q\sqrt{n})}]}{n-\mathbf{E}[N_n(\infty)]}.$$

Now we again handle the numerator and denominator separately, dividing each by n. By Corollary 1, $N_n(0)/n \to b$ implies $N_n(t)/n \to b$, for each $t \geq 0$. By the same argument in the proof of Theorem 3, $N_n(\infty)/n \to b$, as well. By the bounded convergence theorem, $(1/n)\mathbf{E}[N_n(\infty)] \to b$ and $(1/n)\mathbb{E}[\mathbf{N}_n(\infty)1_{(N_n(\infty)\geq s_n+q\sqrt{n})}] \to b\alpha(q)$. We can conclude that (23) holds. □

Appendix C: Numerical Study for the Cost Minimization Problem

Table 1 $s_n^* - s_n^C$ as a function of n and r for $a/c = 0.1$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
0.1	0	0	-1	0	0	0	0	0	0	0
0.5	0	0	1	2	2	3	3	2	-2	-14
0.9	0	1	-1	-1	-2	-4	-5	-20	-45	-98

Table 2 $s_n^* - s_n^C$ as a function of n and r for $a/c = 1$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
0.1	0	0	0	0	0	0	0	0	0	1
0.5	0	0	0	0	0	0	0	0	0	0
0.9	0	0	1	1	0	1	1	1	1	1

Table 3 $s_n^* - s_n^C$ as a function of n and r for $a/c = 10$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
0.1	0	1	0	1	0	0	0	0	0	1
0.2	-1	0	-1	0	0	0	0	0	0	0
0.9	-1	-1	-1	-1	-1	-1	0	0	0	0

In this section, we evaluate the heuristic for the cost minimization problem from Section 6. Consider cost structure (a) and suppose $r \leq \sqrt{a/(c\mu)}$. Given the cost ratio a/c we consider the staffing rule $s_n^C = rn + \beta^* \sqrt{n}$, where β^* is the solution of (22) (we determine β^* by performing a simple search along the real line). In all our experiments, we fix $\mu = 1$. The results are summarized in Tables $1-3$ which present the difference between the optimal staffing level s_n^* and s_n^C for different values of n, r and a/c. For cases with $a/c = 0.1$ and r is either 0.5 or 0.9, then $r > \sqrt{a/c}$ and our heuristic does not perform well, as predicted. On the other hand, we have $r \leq \sqrt{a/(c\mu)}$ for all the other cases and the corresponding errors $|s_n^* - s_n^C|$ are all smaller than or equal to one.

It is also interesting to compare how a heuristic designed for an open queueing system would perform on a closed system with the same offered load. To address this issue, we again equate the offered load of the open and closed systems and we consider the heuristic proposed by Borst et al. (2004): $s_n^O = rn + y^*(a/c)\sqrt{rn}$ where $y^*(a/c)$ is defined by (40) (in Borst et al. 2004) and which performs extremely well for open queues. On the other hand, Tables 4–6 show that, not surprisingly, the heuristic does not perform well when applied to close queues, unless the values

of r and a/c are small. The difference $s_n^* - s_n^C$ can be significantly large otherwise, reaching 32 for instance when $a/c = 0.1$ and $r = 0.9$. The difference $s_n^* - s_n^C$ is actually negative, suggesting that when staffing the system in order to minimize cost, failing to recognize a closed system results in overstaffing. This is consistent with Proposition 2.

Table 4 $s_n^* - s_n^O$ as a function of n and r for $a/c = 0.1$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
0.1	0	0	0	0	0	1	0	1	1	1
0.5	-1	-2	-2	-2	-3	-4	-5	-11	-21	-41
0.9	-2	-3	-5	-6	-7	-10	-13	-32	-61	-122

Table 5 $s_n^* - s_n^O$ as a function of n and r for $a/c = 1$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
	0	1	0	1	1	1	1	2	2	4
	0	-1	-1	-1	-1	-1	-1	-3	-3	-5
	-1	-2	-2	-2	-3	-3	-4	-6	-9	-14

Table 6 $s_n^* - s_n^O$ as a function of n and r for $a/c = 10$

r	$n=2$	$n=4$	$n=6$	$n=8$	$n=10$	$n=15$	$n=20$	$n=50$	$n=100$	$n=200$
0.1	1	1	1	1	1	1	1	2	2	4
0.5	-1	-1	-1	-1	-1	-2	-2	-2	-3	-4
0.9	-3	-3	-4	-4	-5	-5	-6	-9	-11	-17

Appendix D: Extensions

D.1. Abandonment

Abandonment occurs when members waiting in-queue become impatient and decide to leave. When a member abandons the queue, one can envision one of two things taking place next: either the abandoning member reverts to the inactive state or they enter some type of *retrial* state, from which they will attempt to join the queue later. In the former case, one can model the retrial times as a sequence of i.i.d. random variables that are (keeping in line with the remaining random variables) exponentially distributed with mean $1/\nu > 0$. For simplicity, the quantity ν, referred to as the *retrial rate*, will be assumed to be equal to the activation rate λ. As such, the retrial and inactive members are effectively indistinguishable (at least with regards to their contribution to the collection of active users) and there is no substantive difference between the retrial and inactive states. In the following we consider what happens to the results of our model when abandonment is modelled. Analogs to some of the results of Sections 3 and 4 are provided without proof.

Consider the sequence of *membership services models with abandonment*, indexed by n. Let N_n^a denote the number of active users in the nth system. Assume that for a newly activated user, the time to potential abandonment (which also can be thought of as the associated member's patience) is exponentially distributed with mean $1/\theta > 0$. If the member's queueing time reaches the potential abandonment time, then the member indeed leaves the queue and reverts to the inactive state. Including both the service completion process and the abandonment process, the total rate at which members become inactive is $\mu(N_n^a \wedge s_n) + \theta(N_n^a - s_n)^+$.

The staffing regime boundaries are not affected by the addition of abandonment. Moreover, the fluid behavior under abandonment is only affected in the ED regime. This can be seen in the analog to Theorem 1 below (for which the proof can be replicated). Let $\bar{\mathbf{N}}_n^a \equiv N_n^a/n$ be the fluid-scaled number of active members in the membership model with abandonment and define $\gamma \equiv \lambda/(\lambda+\theta)$.

THEOREM 10. *(fluid limit for the $M/M/s_n + M/\infty/n$ membership model) If $\bar{\mathbf{N}}_n^a(0) \Rightarrow \bar{\mathbf{b}}^a(0)$, where $\bar{\mathbf{b}}^a(0)$ is a positive deterministic constant, then*

$$\bar{\mathbf{N}}_n^a \Rightarrow \bar{\mathbf{b}}^a \quad in \quad \mathcal{D} \quad as \quad n \to \infty,$$

where $\bar{\mathbf{b}}^a$ obeys the following ODE

$$\frac{d\bar{\mathbf{b}}^a}{dt}(t) = \lambda - (\lambda+\mu)\bar{\mathbf{b}}^a(t) + (\mu-\theta)\left(\bar{\mathbf{b}}^a(t) - \bar{s}\right)^+,$$

and has the steady state value

$$b^a \equiv \bar{\mathbf{b}}^a(\infty) = \begin{cases} \gamma\left(1 - \frac{\bar{s}}{r}\right) + \bar{s}, & \bar{s} < r \\ r, & \bar{s} \geq r. \end{cases} \tag{33}$$

By defining $q^a \equiv (b^a - \bar{s})^+ = \gamma\left(1 - \frac{\bar{s}}{r}\right)^+$ as the limiting fraction of members in-queue and comparing this quantity to its abandonment-free analog $q = (1 - \frac{\bar{s}}{r})^+$, we see that abandonment decreases the queue length by a factor γ: $q^a = \gamma q$.

In Section 4 we proved diffusion limits for the sequence of membership models. The first key step in this process is appropriately centering the processes. The abandonment analog of (9) is

$$\mathbf{N}_n^a(t) \equiv \frac{N_n^a(t) - b^a n}{\sqrt{n}}, \quad t \geq 0,$$

for each $n \geq 0$. We can immediately provide analogs to Theorems 2 and 5 by replicating their proofs. No analog to Theorem 7 is needed because in the limit as $n \to \infty$, no members in a sequence of systems staffed in the QD limiting regime abandon. For simplicity we state the theorems exclusively in terms of the scaled active member process.

THEOREM 11. *(stochastic-process limit for the $M/M/s_n + M/\infty/n$ queue in the QED regime) Consider the sequence of membership models with abandonment operating in the QED staffing regime, i.e. satisfying (4). If $\mathbf{N}_n^a(0) \Rightarrow \mathbf{N}^a(0)$ as $n \to \infty$, then*

$$\mathbf{N}_n^a \Rightarrow \mathbf{N}^a \quad in \quad \mathcal{D}, \quad as \quad n \to \infty,$$

where $\mathbf{N}^a$ is a diffusion process with infinitesimal mean

$$m^a(x) = \begin{cases} -(\lambda + \mu)x & x < \beta, \\ -(\lambda + \theta)x + (\theta - \mu)\beta & x \geq \beta \end{cases}$$

and (constant) infinitesimal variance $(\sigma^a)^2 = 2\mu r$. The steady state, $\mathbf{N}^a(\infty)$, has probability density

$$f_{QED}^a(x) = \begin{cases} \frac{1-\alpha^a}{\sqrt{r\bar{r}}}\, \phi\left(\frac{x}{\sqrt{r\bar{r}}}\right) \Big/ \Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right) & x < \beta, \\ \frac{\alpha^a}{\sqrt{\gamma\bar{r}}}\, \phi\left(\frac{rx + \bar{r}\beta}{r\sqrt{\gamma\bar{r}}}\right) \Big/ \Phi\left(\frac{-\beta}{r\sqrt{\gamma\bar{r}}}\right) & x \geq \beta, \end{cases}$$

where

$$\alpha^a \equiv \left(1 + \sqrt{\frac{\gamma}{r}} \frac{h\left(\frac{\beta}{r\sqrt{\gamma\bar{r}}}\right)}{h\left(-\frac{\beta}{\sqrt{r\bar{r}}}\right)}\right)^{-1}. \tag{34}$$

THEOREM 12. *(stochastic-process limit for the $M/M/s_n + M/\infty/n$ queue in the ED regime) Consider the sequence of membership models with abandonment operating in the ED staffing regime ($\bar{s} < r$). If $\mathbf{N}_n^a(0) \Rightarrow \mathbf{N}^a(0)$ as $n \to \infty$ and*

$$\sqrt{n}|\bar{s} - s_n/n| \to 0,$$

both as $n \to \infty$, then

$$\mathbf{N}_n^a \Rightarrow \mathbf{N}^a \quad in \quad \mathcal{D}, \quad as \quad n \to \infty,$$

where and $\mathbf{N}^a$ *is an OU process with infinitesimal mean* $m^a(x) = -(\lambda+\theta)x$ *and infinitesimal variance*

$$(\sigma^a)^2 = 2\mu\bar{s} + \theta q^a$$

and has steady state distribution $\mathbf{N}^a(\infty) = Normal\left(0, \gamma\left(\frac{\bar{r}\bar{s}}{r} + \frac{\theta q^a}{2\lambda}\right)\right)$.

As analogs to their non-abandonment versions, Theorems 11 and 12 can be also be used for staffing decisions. For instance, (34) provides the asymptotic probability of delay as a function of the service grade β. It makes sense that for a given staffing level, the presence of abandonment reduces the probability of delay. Thus fewer servers are required to produce the same probability of delay. Likewise, the ED regime result (Theorem 12) suggests a method for staffing to meet an expected delay constraint.

D.2. Balking

Balking occurs when newly activated members see the queue and decide not to join. As was the case in our modeling of abandonment, after balking, members enter into a retrial state from which they emerge at rate ν back to the active state. For the last result of this section, we will allow for any arbitrary value of ν. For now though, assume that the retrial rate is the same as the activation rate: $\nu = \lambda$. In this case, the retrial and inactive members are effectively indistinguishable.

Members balk from the queue if it is longer than a certain threshold. However, their choice not to join is likely based on their anticipation of the wait that is associated with this queue length. It makes sense then, given that the staffing level increases roughly linear with n, that their tolerance for queue length grows linearly with n as well. We provide a slightly more general approach. Suppose that newly activated members in the nth system balk from the queue if it already has $\ell_0 n + \ell\sqrt{n}$ others in it.

Let N_n^b denote the active user process when users balk from the queue and $\mathbf{N}_n^b$ be the diffusion scaled version of the process, where for each $n \geq 1$,

$$\mathbf{N}_n^b(t) \equiv \frac{N_n^b(t) - bn}{\sqrt{n}}, \quad t \geq 0.$$

The superscript 'b' stands for balking and is not a parameter. Conversely, the centering constant bn is the same as in (9).

It should be clear that staffing policies in the QD regime are unaffected by the balking phenomenon; no one queues in the QD regime. Moreover, in the QED regime, the queue length is $o(n)$. So balking takes place in the QED regime only if $\ell_0 = 0$. This assumption leads to the following analog of Theorem 2.

THEOREM 13. *(stochastic-process limit for the $M/M/s_n/\ell\sqrt{n}/n$ queue in the QED regime) Consider the sequence of membership models with balking operating in the QED staffing regime, i.e. satisfying (4). Assume that $\ell > 0$. If $\mathbf{N}_n^b(0) \Rightarrow \mathbf{N}^b(0) \leq \beta + \ell$ as $n \to \infty$, then*

$$\mathbf{N}_n^b \Rightarrow \mathbf{N}^b \quad in \quad \mathcal{D}, \quad as \quad n \to \infty,$$

where $\mathbf{N}^b$ is a diffusion process with infinitesimal mean

$$m^b(x) = \begin{cases} -(\lambda+\mu)x & x < \beta, \\ -\lambda x - \mu\beta & \beta \leq x < \beta + \ell, \end{cases}$$

infinitesimal variance $(\sigma^b)^2 = 2\mu r$, and reflected above at $\beta + \ell$. The steady state, $\mathbf{N}^b(\infty)$, has probability density

$$f_{QED}^b(x) = \begin{cases} \frac{1-\alpha^b}{\sqrt{r\bar{r}}}\, \phi\left(\frac{x}{\sqrt{r\bar{r}}}\right) \Big/ \Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right) & x < \beta, \\ \frac{\alpha^b}{\sqrt{\gamma\bar{r}}}\, \phi\left(\frac{rx+\bar{r}\beta}{r\sqrt{\gamma\bar{r}}}\right) \Big/ \left[\Phi\left(\frac{\beta+\ell}{r\sqrt{\gamma\bar{r}}}\right) - \Phi\left(\frac{\beta}{r\sqrt{\gamma\bar{r}}}\right)\right] & \beta \leq x < \beta + \ell, \end{cases}$$

where

$$\alpha^b \equiv \left(1 + e^{-\frac{\beta^2}{2r^2}} \sqrt{r} \frac{\Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right)}{\Phi\left(\frac{\beta+\ell}{r\sqrt{\bar{r}}}\right) - \Phi\left(\frac{\beta}{r\sqrt{\bar{r}}}\right)}\right)^{-1}.$$

The quantity α^b above is the limiting approximate probability of delay. If one allows for $\ell = 0$, then the resulting model has blocking. This case was considered by Randhawa and Kumar (2005).

Now suppose we are staffing in the ED regime. Because $\nu = \lambda$, for every $\bar{s}$ there is one natural choice of ℓ_0, namely $b - \bar{s} = 1 - \frac{\bar{s}}{r}$ from (8). If instead we have $\ell_0 > b$ then, in the limit as $n \to \infty$, no members balk. If $\ell_0 < b$ then the limiting diffusion is trivial. In particular, the number of users in queue is equal to $\ell_0 n + \ell\sqrt{n} + o(\sqrt{n})$. That is, the deviation of the number of active users from the maximum that the system will accommodate is small relative to $\sqrt{n}$. We will return to this scenario at the conclusion of this section.

Suppose for now that $\ell_0 = b - \bar{s}$ and that ℓ is any real number. The following is a balking analog of Theorem 5.

THEOREM 14. *(stochastic-process limit for the $M/M/s_n/(b-\bar{s})n + \ell\sqrt{n}/n$ queue in the ED regime) Consider the sequence of membership models with balking operating in the ED staffing regime ($\bar{s} < r$). If $\mathbf{N}_n^b(0) \Rightarrow \mathbf{N}^b(0) \leq \ell$ as $n \to \infty$ and*

$$\sqrt{n}|\bar{s} - s_n/n| \to 0,$$

both as $n \to \infty$, then

$$\mathbf{N}_n^b \Rightarrow \mathbf{N}^b \quad in \quad \mathcal{D}, \quad as \quad n \to \infty,$$

where and $\mathbf{N}^b$ *is an OU process, truncated above at* ℓ*, with infinitesimal mean* $m^b(x) = -\lambda x$ *and infinitesimal variance,*

$$(\sigma^b)^2 = 2\mu\bar{s}$$

and whose steady state distribution has the following density

$$f^b_{ED}(x) = \sqrt{\frac{r}{\bar{r}\bar{s}}}\phi\left(x\sqrt{\frac{r}{\bar{r}\bar{s}}}\right) \Big/ \Phi\left(\ell\sqrt{\frac{r}{\bar{r}\bar{s}}}\right) \quad x \le \ell.$$

To conclude this section, we allow for the retrial rate and the activation rate to be different ($\nu \neq \lambda$), provided that staffing is in the ED regime and the order n balking constant is less than the deterministic limit of the sequence of fluid-scaled queue length processes ($\ell_0 < b - \bar{s}$). In addition to the active user process N^b_n, we also track the number of retrial users R^b_n and the number of inactive users M^b_n as a function of time, for each $n \ge 1$. Our remaining arguments are strictly intuition-based. We will assume, without loss of generality that $\ell = 0$.

Suppose we define the following diffusion scaled processes:

$$\mathbf{N}^b_n \equiv \frac{N^b_n - n(\ell_0 + \bar{s})}{\sqrt{n}}, \qquad \mathbf{M}^b_n \equiv \frac{M^b_n - n(1-b)}{\sqrt{n}}, \qquad \text{and} \qquad \mathbf{R}^b_n \equiv \frac{R^b_n - n(b - \bar{s} - \ell_0)}{\sqrt{n}}, \qquad n \ge 1.$$

Because $\mathbf{N}^b_n + \mathbf{M}^b_n + \mathbf{R}^b_n = 0$, the system is two-dimensional. As $n \to \infty$, we anticipate that the limiting scaled active user process converges to a degenerate process. Specifically, this process converges to the constant 0, provided that $\mathbf{N}^b_n(0) \Rightarrow 0$, $\mathbf{M}^b_n(0) \Rightarrow \mathbf{M}^b(0)$ and $\mathbf{R}^b_n(0) \Rightarrow -\mathbf{M}^b(0)$, where $\mathbf{M}^b(0)$ is a proper random variable. It follows that this relation between $\mathbf{M}^b$ and $\mathbf{R}^b$ will hold for all time t so that the limiting system is one-dimensional. The intuition behind why the limit reduces in dimension is straightforward. For sufficiently large systems, the number of active users is always greater than $n\bar{s}$. So the output rate is μs_n. However, because the number of active users is at most $n\ell_0 + s_n$ the input rate exceeds μs_n by an order n quantity. This means that users balk at an order n rate and further that the queue never drops below its maximum amount.

In essence, a relatively fixed number, $n\ell_0 + s_n$, of users are always active; thus $\mathbf{N}^b_n \Rightarrow 0$. This reduction in the state space is a type of state space collapse. Although analyzing the limiting process $\mathbf{R}^b$ is equivalent to analyzing the limiting process $\mathbf{M}^b$, it turns out the latter is conceptually easier. Because $\mathbf{N}^b$ is trivial, the input process to the collection of inactive users is constant at rate μs_n. However, the output (activation) rate (regardless of whether the activated users join the queue or balk) depends on the number of inactive users, exclusively; the activation rate is proportional to the number of inactive users. It follows then the limiting process is the same as the one in Theorem 5. Namely, the limiting process is an OU process whose steady state is $Normal(0, \bar{r}\bar{s}/r)$.

One curiosity with the above conjecture is that the retrial rate ν has no effect on the limiting diffusion-scaled retrial user process. This is not to say that retrial rate does not affect individual users. The faster they retry, the sooner they emerge from the retrial state. However, the faster the retrial rate, the more likely that inactive users will balk on their initial attempt to join the queue.