MANAGEMENT SCIENCE
DOI 10.1287/mnsc.1090.1041ec

e-companion
ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Real-Time Delay Estimation in Overloaded Multiserver Queues with Abandonments" by Rouba Ibrahim and Ward Whitt, *Management Science*, DOI 10.1287/mnsc.1090.1041.

e-Companion

EC.1. Introduction

We present additional material in this e-companion. In §EC.2, we give the proof of Theorem 4. In §EC.3, we present detailed simulation results for the $M/M/s+GI$ model. In §EC.4, we present additional experimental results for non-exponential service-time distributions. In §EC.5, we present simulation results substantiating the heavy-traffic limits of §5, for the $GI/M/s+M$ model, with alternative interarrival-time distributions and alternative values of the abandonment rate α. We present additional experimental results in the supplement to the main paper, available on the authors' webpages, Ibrahim and Whitt (2008).

EC.2. Proof of Theorem 4

We now prove the convergence in distribution in (35). The proof follows the general approach used to prove Theorem 6.4 of Talreja and Whitt (2008), exploiting stochastic-process limits in order to obtain the desired one-dimensional limit in $\mathbb{R}$. As in (6.37) of Talreja and Whitt (2008), we use the continuous mapping theorem with the composition map to treat random time changes. We start with the joint convergence

$$(\hat{W}_{Q,s}(t), \hat{N}_s(t), \hat{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), N(0,\sigma_w^2)) \quad \text{in} \quad D^2 \times \mathbb{R} \tag{EC.1}$$

for the processes defined in (18), (33) and (27), where the limits are mutually independent.

For the $M/M/s+M$ model, we can obtain the joint convergence from the individual limits established above, because we can regard the component processes on the left as mutually independent. That requires some comment, however. First in time we have the waiting time for the last customer to enter service, which is distributed asymptotically the same as $W_s(\infty)$. Then we have the buildup of the queue behind this customer until this customer starts service, given by $\hat{N}_s(t)$. Finally, we have the remaining times between successive departures after the new arrival enters the system, as given by $\hat{W}_{Q,s}(t)$, which involves independent exponential random variables. These are mutually independent with reference to the designated arrival at one fixed time, for whom we are doing the estimation. The processes are well defined as independent random elements of D, but they only correctly apply to describe our system at a single time, as stated in the final one-dimensional limit in (35). (In the case of the $G/M/s+M$ model, we assume that the joint limit of $(\hat{N}_s(t), \hat{W}_s(\infty))$ is the same as if these were independent.)

Assuming the limit in (EC.1), since $\bar{N}_s$ converges to a deterministic limit, we can append the limit for $\bar{N}_s$ to get

$$(\hat{W}_{Q,s}(t), \hat{N}_s(t), \bar{N}_s(t), \hat{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), a(t), N(0,\sigma_w^2)) \quad \text{in} \quad D^3 \times \mathbb{R}. \tag{EC.2}$$

We can now apply the continuous mapping theorem with composition to perform a random time change with $\bar{N}_s$ to obtain the limit

$$\hat{W}_{Q,s}(\bar{N}_s(t)) \equiv \sqrt{s}\left(W_{Q,s}(N_s(t)) - c(\bar{N}_s(t))\right) \Rightarrow B(d(a(t))) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \tag{EC.3}$$

jointly with the limit in (EC.2), where B is the given standard Brownian motion and $a(t)$ is defined in (32). We can now apply a random-time-change argument one more time with $W_s(\infty)$ to obtain the limit

$$\hat{Z}_s \equiv \sqrt{s}\left(W_{Q,s}(N_s(W_s(\infty))) - c(\bar{N}_s(W_s(\infty)))\right) \Rightarrow B(d(a(w))) \stackrel{\mathrm{d}}{=} N(0, d(a(w))) \quad \text{in} \quad \mathbb{R} \tag{EC.4}$$

as $s \to \infty$, again jointly with the limit in (EC.2), where again the limit involves the same Brownian motion B.

We obtain the desired limit in (35) by writing

$$\hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s}\,(W_{LES,s}(W_s(\infty)) - W_s(\infty)) \equiv \hat{Z}_s + \hat{Y}_s \tag{EC.5}$$

for $\hat{Z}_s$ in (EC.4) and

$$\hat{Y}_s \equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \tag{EC.6}$$

and establishing a limit for $\hat{Y}_s$ in (EC.6) within the framework of the initial limits in (EC.2). In order to make a connection to the given limits for $(\hat{N}_s(t), W_s(\infty))$ in (EC.2), we exploit a Taylor series expansion for the functions $c(t)$ and $a(t)$ in (19) and (32). Note that

$$c(q) = w \equiv \frac{1}{\alpha}\ln(\rho), \quad a(w) = q \equiv \frac{\lambda - \mu}{\alpha} \quad \text{and} \quad d(q) = \frac{q}{\lambda\mu}. \tag{EC.7}$$

Hence, $d(a(w)) = d(q) = q/(\lambda\mu)$.

We write

$$\hat{Y}_s \equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3}, \quad \text{(EC.8)}$$

where

$$\hat{Y}_{s,1} \equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty)) - c(a(W_s(\infty)))\right),$$
$$\hat{Y}_{s,2} \equiv \sqrt{s}\left(c(a(W_s(\infty))) - c(a(w))\right),$$
$$\hat{Y}_{s,3} \equiv \sqrt{s}\left(c(a(w)) - W_s(\infty)\right), \quad \text{(EC.9)}$$

Using a Taylor series expansion of c, we see that

$$\hat{Y}_{s,1} - c'(a(w))\sqrt{s}\left(\bar{N}_s(W_s(\infty)) - a(W_s(\infty))\right) \Rightarrow 0, \quad \text{(EC.10)}$$

where $c'(a(w)) = 1/\lambda$. By Theorem 3,

$$\hat{Y}_{s,1} \Rightarrow \frac{1}{\lambda}\hat{G}(w) \stackrel{\mathrm{d}}{=} N(0, \sigma_n^2(w)/\lambda^2) \quad \text{as} \quad s \to \infty. \quad \text{(EC.11)}$$

Using a Taylor series expansion of $c \circ a$, noting that $a'(w) = \mu$, we get

$$\hat{Y}_{s,2} - c'(a(w))a'(w)\sqrt{s}\left(W_s(\infty) - w\right) \Rightarrow 0, \quad \text{(EC.12)}$$

so that, by Theorem 2,

$$\hat{Y}_{s,2} \Rightarrow \frac{\mu}{\lambda}N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty. \quad \text{(EC.13)}$$

Similarly, using the relation $c(a(w)) = c(q) = w$ and replacing $c(a(w))$ by w, we get

$$\hat{Y}_{s,3} - \sqrt{s}\left(w - W_s(\infty)\right) \Rightarrow 0, \quad \text{(EC.14)}$$

so that, by Theorem 2 again,

$$\hat{Y}_{s,3} \Rightarrow N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty, \quad \text{(EC.15)}$$

where the limiting random variables $N(0, \sigma_w^2)$ in (EC.13) and (EC.15) are identical. By these constructions, we obtain convergence of the vector $(\hat{Y}_{s,1}, \hat{Y}_{s,2}, \hat{Y}_{s,3})$ jointly with the initial limits in (EC.2) and thus also jointly with $\hat{Z}_s$ in (EC.4). The processes $\hat{Y}_{s,i}$ are each asymptotically equivalent to processes that are simple functions of the processes in the original limit (EC.2).

Hence,

$$\hat{Y}_s \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3} \Rightarrow N\left(0, \frac{\sigma_n^2(w)}{\lambda^2} + \frac{(\lambda-\mu)^2\sigma_w^2}{\lambda^2}\right). \quad \text{(EC.16)}$$

We can thus obtain the limit from (EC.4)–(EC.6), (EC.8), (EC.9) and (EC.16) by adding the normal components.

EC.3. Simulation Results for the $M/M/s+GI$ Model

In this section, we present tables of simulation results (point and 95% confidence interval estimates) quantifying the performance of the alternative delay estimators in the $M/M/s+GI$ model. The corresponding plots are shown and discussed in §6. For the abandonment-time distribution, we consider M, H_2, and E_{10} distributions. We consider alternative values of $s \geq 100$, and vary the arrival rate, λ, to keep the traffic intensity, ρ, fixed for alternative values of s ($\rho = 1.4$). We let the abandonment rate, α, be equal to 1.

With exponential abandonments, Table EC.1 shows that, consistent with theory, QL_m is the best possible delay estimator, under the MSE criterion. The QL_r^m and QL_r estimators are nearly identical, with QL_r^m slightly outperforming QL_r. They are both nearly as efficient as QL_m. Consistent with (38), the LES estimator performs worse than QL_m, but not greatly so: The relative error between the simulation estimates for ASE(LES)/ASE(QL_m) and the numerical value, 2, given by (38) is less than 1% throughout. Consistent with (29), the NI estimator is less efficient than QL_m: The relative error between the simulation estimates for ASE(NI)/ASE(QL_m) and the numerical value, 3.5, given by (29) is less than 1% throughout. The QL estimator performs significantly worse than all the other estimators, particularly for large values of s. The ratio ASE(QL)/ASE(QL_m) ranges from about 3 when $s = 100$ to nearly 16 when $s = 1000$.

With hyperexponential abandonments, Table EC.2 shows that QL_{ap} is the best delay estimator. The QL_r estimator performs nearly the same as QL_{ap} and is only slightly outperformed. The QL_m estimator, which is optimal for the $GI/M/s+M$ model, is now outperformed by QL_r, particularly

Efficiency of the estimators in the $M/M/s+M$ model with $\rho = 1.4$ and $\alpha = 1.0$

s	ASE[θ_{QL_m}]	ASE[$\theta_{QL_r^m}$]	ASE[θ_{QL_r}]	ASE[θ_{QL}]	ASE[θ_{LES}]	ASE[θ_{NI}]
100	2.867×10^{-3}	2.869×10^{-3}	3.130×10^{-3}	8.693×10^{-3}	5.772×10^{-3}	1.00×10^{-2}
	$\pm 1.76 \times 10^{-5}$	$\pm 1.78 \times 10^{-5}$	$\pm 1.89 \times 10^{-5}$	$\pm 3.20 \times 10^{-5}$	$\pm 2.79 \times 10^{-5}$	$\pm 5.97 \times 10^{-5}$
300	9.587×10^{-4}	9.601×10^{-4}	1.039×10^{-3}	5.602×10^{-3}	1.922×10^{-3}	3.351×10^{-3}
	$\pm 6.86 \times 10^{-6}$	6.92×10^{-6}	$\pm 6.41 \times 10^{-6}$	$\pm 2.64 \times 10^{-5}$	$\pm 1.50 \times 10^{-5}$	$\pm 6.03 \times 10^{-5}$
500	5.761×10^{-4}	5.661×10^{-4}	6.224×10^{-4}	5.017×10^{-3}	1.153×10^{-3}	2.038×10^{-3}
	$\pm 1.94 \times 10^{-6}$	$\pm 3.86 \times 10^{-6}$	$\pm 2.94 \times 10^{-6}$	$\pm 2.41 \times 10^{-5}$	$\pm 9.99 \times 10^{-6}$	$\pm 2.26 \times 10^{-5}$
700	4.104×10^{-4}	4.201×10^{-4}	4.440×10^{-4}	4.682×10^{-3}	8.166×10^{-4}	1.441×10^{-3}
	$\pm 1.82 \times 10^{-6}$	2.839×10^{-4}	$\pm 2.71 \times 10^{-6}$	$\pm 2.40 \times 10^{-5}$	$\pm 5.78 \times 10^{-6}$	$\pm 1.57 \times 10^{-5}$
1000	2.892×10^{-4}	2.839×10^{-4}	3.136×10^{-4}	4.492×10^{-3}	5.752×10^{-4}	1.019×10^{-3}
	$\pm 3.48 \times 10^{-6}$	$\pm 3.86 \times 10^{-6}$	$\pm 3.09 \times 10^{-6}$	$\pm 1.54 \times 10^{-5}$	$\pm 6.91 \times 10^{-6}$	$\pm 3.00 \times 10^{-5}$

Table EC.1 Point and confidence interval estimates of the ASEs - average square errors - of the estimators

when s is large (e.g., ASE(QL$_m$)/ASE(QL$_{ap}$) is close to 2 when $s = 1000$). The LES estimator performs worse than QL$_m$ when $s = 100$, but is nearly identical to QL$_m$ when $s = 1000$. The NI estimator performs worse than LES, but not as bad as QL. Once more, QL is the least efficient delay estimator: The ratio ASE(QL)/ASE(QL$_{ap}$) ranges from about 2 when $s = 100$ to about 10 when $s = 1000$.

With Erlang abandonments, Table EC.3 shows that QL$_{ap}$ is, once more, the best possible delay estimator, except when s is very large (e.g., $s = 700$ or $s = 1000$). The QL$_r$ estimator performs worse than QL$_{ap}$ for relatively small values of s, but slightly outperforms QL$_{ap}$ for relatively large values of s. The NI estimator is more competitive in this model, than in the previous two models. It is nearly as efficient as QL$_{ap}$, particularly when s is large. The LES estimator also fares well, but is slightly outperformed by QL$_{ap}$, QL$_r$ and NI. The QL$_m$ estimator performs significantly worse than QL$_{ap}$ when s is large (e.g., when $s = 1000$, ASE(QL$_m$)/ASE(QL$_{ap}$) ≈ 9). Finally, QL is yet again the least effective estimator for this model. The ratio ASE(QL)/ASE(QL$_{ap}$) ranges from about 15 when $s = 100$ to about 93 when $s = 1000$.

Efficiency of the estimators in the $M/M/s+H_2$ model with $\rho=1.4$ and $\alpha=1.0$

s	$ASE[\theta_{QL_{ap}}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.859×10^{-3}	2.100×10^{-3}	2.032×10^{-3}	4.662×10^{-3}	3.866×10^{-3}	6.503×10^{-3}
	$\pm6.52\times10^{-6}$	$\pm5.54\times10^{-6}$	$\pm6.31\times10^{-6}$	$\pm1.83\times10^{-5}$	$\pm8.10\times10^{-6}$	$\pm3.85\times10^{-5}$
300	6.116×10^{-4}	7.933×10^{-4}	6.599×10^{-4}	2.593×10^{-3}	1.236×10^{-3}	2.165×10^{-3}
	$\pm4.64\times10^{-6}$	$\pm7.62\times10^{-6}$	$\pm8.82\times10^{-6}$	$\pm2.25\times10^{-5}$	$\pm1.76\times10^{-5}$	$\pm2.09\times10^{-5}$
500	3.695×10^{-4}	5.367×10^{-4}	3.921×10^{-4}	2.205×10^{-3}	7.331×10^{-4}	1.311×10^{-3}
	$\pm2.19\times10^{-6}$	$\pm2.12\times10^{-6}$	$\pm2.47\times10^{-6}$	$\pm9.97\times10^{-6}$	$\pm5.41\times10^{-6}$	$\pm1.03\times10^{-5}$
700	2.630×10^{-4}	4.257×10^{-4}	2.802×10^{-4}	2.024×10^{-3}	5.250×10^{-4}	9.378×10^{-4}
	$\pm1.43\times10^{-6}$	$\pm1.89\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm2.52\times10^{-6}$	$\pm1.07\times10^{-5}$
1000	1.833×10^{-4}	3.474×10^{-4}	1.978×10^{-4}	1.900×10^{-3}	3.691×10^{-4}	6.533×10^{-4}
	$\pm1.55\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm6.90\times10^{-7}$	$\pm5.93\times10^{-6}$	$\pm3.00\times10^{-6}$	1.14×10^{-5}

Table EC.2 **Point and confidence interval estimates of the ASEs - average square errors - of the estimators**

Efficiency of the estimators in the $M/M/s+E_{10}$ model with $\rho=1.4$ and $\alpha=1.0$

s	$ASE[\theta_{QL_{ap}}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	5.388×10^{-3}	9.400×10^{-3}	6.317×10^{-3}	8.097×10^{-2}	8.810×10^{-3}	6.077×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm3.48\times10^{-5}$	4.51×10^{-5}	$\pm2.47\times10^{-4}$	$\pm3.91\times10^{-5}$	$\pm2.63\times10^{-5}$
300	1.955×10^{-3}	7.211×10^{-3}	2.139×10^{-3}	7.211×10^{-2}	2.933×10^{-3}	2.040×10^{-3}
	$\pm5.13\times10^{-6}$	$\pm3.86\times10^{-5}$	$\pm1.83\times10^{-5}$	$\pm3.301\times10^{-4}$	$\pm3.22\times10^{-5}$	$\pm2.23\times10^{-5}$
500	1.244×10^{-3}	6.746×10^{-3}	1.293×10^{-3}	7.049×10^{-2}	1.760×10^{-3}	1.288×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm2.68\times10^{-5}$	$\pm1.35\times10^{-5}$	$\pm2.48\times10^{-4}$	$\pm2.44\times10^{-5}$	$\pm2.61\times10^{-5}$
700	9.572×10^{-4}	6.584×10^{-3}	9.319×10^{-4}	6.975×10^{-2}	1.241×10^{-3}	9.966×10^{-4}
	$\pm8.31\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm1.30\times10^{-5}$
1000	7.369×10^{-4}	6.454×10^{-3}	6.694×10^{-4}	6.902×10^{-2}	8.830×10^{-4}	8.242×10^{-4}
	$\pm1.96\times10^{-5}$	$\pm1.70\times10^{-5}$	$\pm1.13\times10^{-5}$	$\pm1.68\times10^{-4}$	$\pm1.28\times10^{-5}$	$\pm1.17\times10^{-5}$

Table EC.3 **Point and confidence interval estimates of the ASEs - average square errors - of the estimators**

EC.4. Simulation Results for the $M/GI/s+M$ Model

In this section we present simulation results quantifying the performance of the alternative delay estimators with non-exponential service-time distributions; i.e., we consider the $M/GI/s+M$ model. In this model, QL$_{ap}$ coincides with QL$_m$, so we do not include separate results for it. For the service-time distribution, we consider D, E_{10}, and $LN(1,1)$ (lognormal with mean and variance

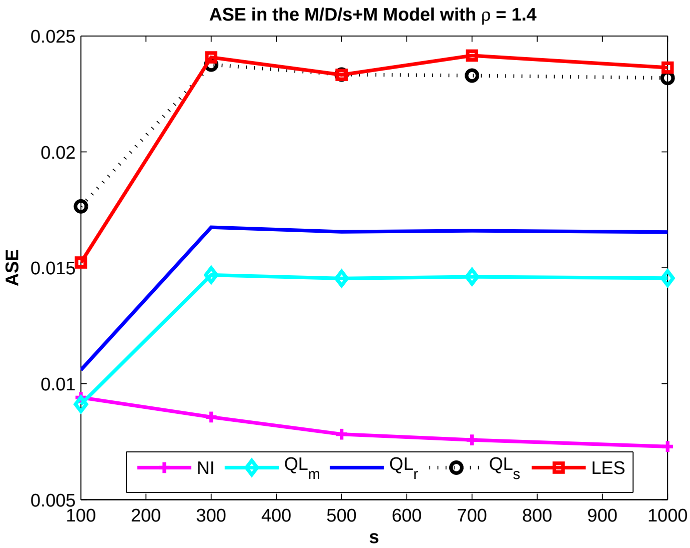

Figure EC.1

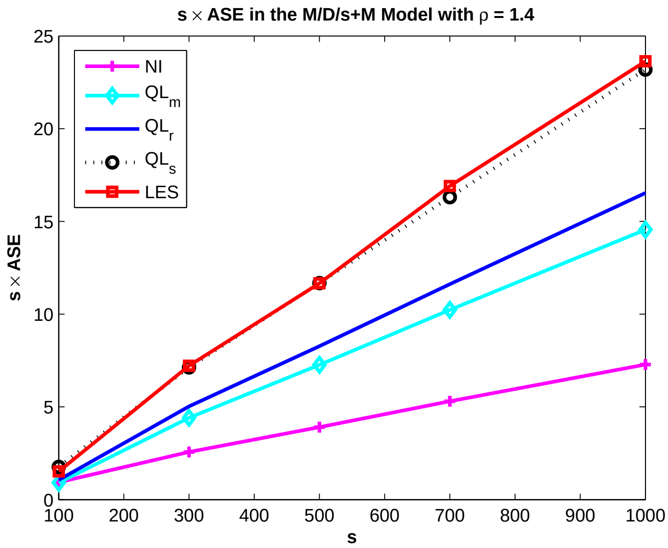

Figure EC.2

equal to 1) distributions. We let $\mu = \alpha = 1.0$, and vary λ, for alternative values of s, to keep $\rho = 1.4$. Corresponding tables with estimates of the 95% confidence intervals, and additional simulation results for the $M/GI/s+M$ model, are presented in the supplement, Ibrahim and Whitt (2008).

EC.4.1. Results for the $M/D/s+M$ model

Figures EC.1 and EC.2 show that all delay estimators do not perform well in this model. The NI estimator, which uses no information at all beyond the model, is the most effective delay estimator, when $s \geq 300$. (For $s = 100$, QL_m slightly outperforms NI.) But even the NI estimator is not very accurate: The RRASE for NI is roughly equal to 25% for all values of s considered. This suggests that our procedures for estimating delays perform relatively poorly when the service times are deterministic. The ASE's for QL_m, QL_r, QL, and LES do not vary much in this model; e.g., $\mathrm{ASE}(\mathrm{QL}_m)$ varies little about 0.01, for all values of s considered. Figure EC.2 shows that, unlike previous models, the accuracy of the estimators does not improve as the number of servers increases. Alternative delay estimation procedures, appropriate for deterministic service times, remain to be investigated.

EC.4.2. Results for the $M/E_{10}/s+M$ model

Simulation results with an E_{10} distribution (SCV $= 0.1$) for the service times, suggest that the proposed delay estimators remain effective, even with very low variability in the service times.

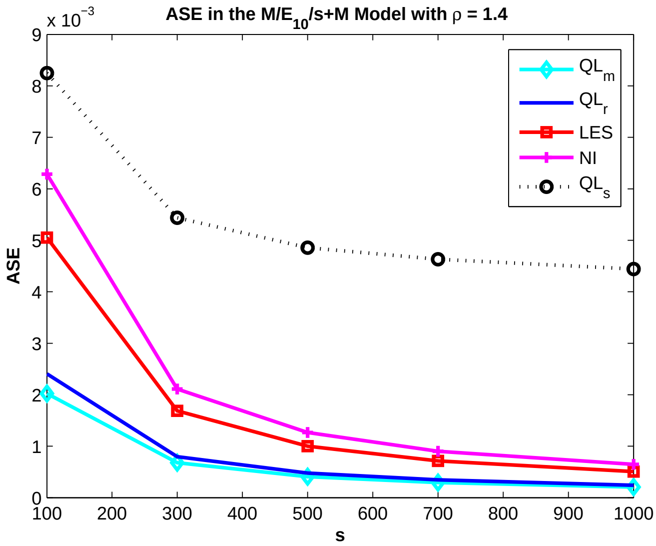

Figure EC.3

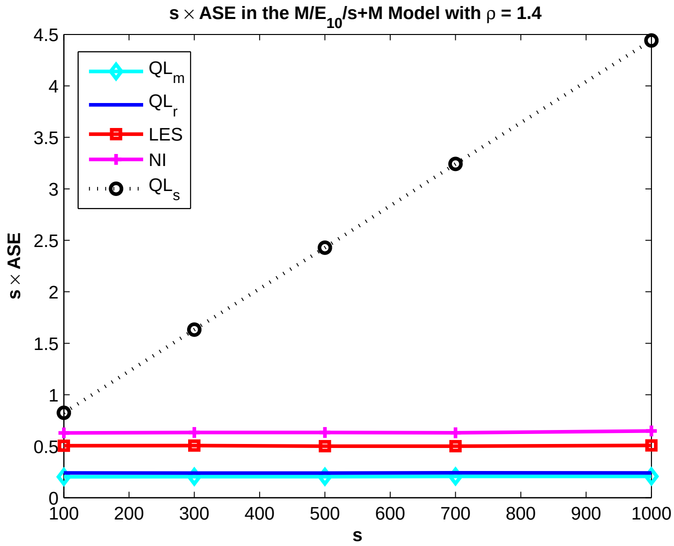

Figure EC.4

Figures EC.3 and EC.4 show that QL_m is the most effective delay estimator for this model. The QL_r estimator is nearly identical to QL_m, particularly when s is large enough ($s \geq 300$). Once more, the relative accuracy of the delay estimators improves as s increases. The RRASE for QL_m ranges from approximately 13% when $s = 100$ to approximately 4% when $s = 1000$. The LES estimator is relatively accurate as well: The RRASE of LES ranges from approximately 21% when $s = 100$ to approximately 7% when $s = 1000$. The NI estimator does not perform as well as LES, nor as bad as QL. The QL estimator is the least efficient estimator: The ratio $\text{ASE}(\text{QL})/\text{ASE}(\text{QL}_m)$ ranges from approximately 4 when $s = 100$ to approximately 22 when $s = 1000$. Consistent with §5, Figure EC.4 shows that all estimators, except QL, have an ASE which is inversely proportional to the number of servers, but mathematical support for the estimators has yet to be provided with non-exponential service-time distributions.

EC.4.3. Results for the $M/LN(1,1)/s+M$ model

We consider the lognormal distribution for the service times because there is empirical evidence suggesting a remarkable fit of the service-time distribution to the lognormal distribution; e.g., see Brown et al. (2005). Table EC.4 shows that QL_m is the most effective delay estimator for this model. The RRASE for QL_m ranges from approximately 14% when $s = 100$ to approximately 5% when

Efficiency of the estimators in the $M/LN(1,1)/s+M$ model with $\rho = 1.4$ and $\alpha = 1.0$

s	ASE[θ_{QL_m}]	ASE[θ_{QL_r}]	ASE[θ_{QL}]	ASE[θ_{LES}]	ASE[θ_{NI}]
100	2.359×10^{-3}	2.596×10^{-3}	8.207×10^{-3}	5.248×10^{-3}	9.089×10^{-3}
	$\pm 7.00 \times 10^{-6}$	$\pm 9.02 \times 10^{-6}$	$\pm 4.45 \times 10^{-5}$	$\pm 2.37 \times 10^{-5}$	$\pm 4.80 \times 10^{-5}$
300	7.810×10^{-4}	8.506×10^{-4}	5.394×10^{-3}	1.716×10^{-3}	3.032×10^{-3}
	$\pm 5.14 \times 10^{-6}$	$\pm 5.68 \times 10^{-6}$	3.36×10^{-5}	$\pm 1.25 \times 10^{-5}$	$\pm 5.30 \times 10^{-5}$
500	4.663×10^{-4}	5.0685×10^{-4}	4.836×10^{-3}	1.029×10^{-3}	1.826×10^{-3}
	$\pm 2.04 \times 10^{-6}$	$\pm 2.12 \times 10^{-6}$	$\pm 2.085 \times 10^{-5}$	$\pm 7.29 \times 10^{-6}$	$\pm 8.10 \times 10^{-6}$
700	3.346×10^{-4}	3.635×10^{-4}	4.615×10^{-3}	7.438×10^{-4}	1.290×10^{-3}
	$\pm 2.71 \times 10^{-6}$	$\pm 3.37 \times 10^{-6}$	$\pm 1.77 \times 10^{-5}$	$\pm 6.47 \times 10^{-6}$	$\pm 1.12 \times 10^{-5}$
1000	2.340×10^{-4}	2.548×10^{-4}	4.443×10^{-3}	5.290×10^{-4}	8.942×10^{-4}
	$\pm 1.84 \times 10^{-6}$	$\pm 2.81 \times 10^{-6}$	$\pm 2.54 \times 10^{-5}$	$\pm 5.90 \times 10^{-6}$	$\pm 2.46 \times 10^{-5}$

Table EC.4 Point and confidence interval estimates of the ASEs - average square errors - of the estimators

$s = 1000$. The QL$_r$ estimator is slightly less efficient than QL$_m$: The ratio ASE(QL$_r$)/ASE(QL$_m$) ranges from approximately 1.1 when $s = 100$ to approximately 1.08 when $s = 1000$. The LES estimator is relatively accurate as well: The RRASE of LES ranges from approximately 26% when $s = 100$ to approximately 7% when $s = 1000$. The NI estimator does not perform as well as LES, nor as bad as QL. The QL estimator is the least efficient estimator: the ratio ASE(QL)/ASE(QL$_m$) ranges from approximately 4 when $s = 100$ to approximately 19 when $s = 1000$.

EC.5. Simulations Results for the $GI/M/s+M$ Model

In this section, we present simulation results quantifying the performance of the alternative delay estimators with non-exponential interarrival-time distributions; i.e., we consider the $GI/M/s+M$ model. For the interarrival-time distribution, we consider D and H_2 distributions.

We also consider different abandonment rates; specifically we let $\alpha = 0.2$ and $\alpha = 5.0$. As indicated by Formulas (3) and (7), the queue length and delay tend to be inversely proportional to α. Thus, changing α from 1.0 to 0.2 or 5.0 tends to change congestion by a factor of 5. The system is very heavily overloaded when $\alpha = 0.2$, but relatively lightly loaded when $\alpha = 5.0$.

We consider the same values of s as before and we let $\mu = 1$. We vary λ to get a fixed value of ρ

($\rho = 1.4$), for alternative values of s. Additional simulation results for the $GI/M/s+M$ model are presented in the supplement, Ibrahim and Whitt (2008).

EC.5.1. Results for the $D/M/s+M$ model with $\alpha = 0.2$

Table EC.5 compares the efficiencies of the alternative delay estimators in the $D/M/s+M$ model with $\alpha = 0.2$. Consistent with theory, QL_m is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QL_m ranges from approximately 35% when $s = 100$ to approximately 11% when $s = 1000$. The QL_r estimator is slightly less efficient than QL_m: $\mathrm{ASE}(\mathrm{QL}_r)/\mathrm{ASE}(\mathrm{QL}_m)$ is less than 1.05 for all values of s considered. The LES estimator is slightly less accurate, with an RRASE ranging from approximately 40% when $s = 100$ to approximately 13% when $s = 1000$. The NI estimator is less accurate than LES, but not as bad as QL. The QL estimator is, once more, the least effective estimator: The ratio $\mathrm{ASE}(\mathrm{QL})/\mathrm{ASE}(\mathrm{QL}_m)$ ranges from approximately 8 when $s = 100$ to approximately 71 when $s = 1000$.

Tables EC.6 and EC.7 substantiate (39) and (29) of §5, that compare the performances of QL_m, LES and NI in the $D/M/s+M$ model. Consistent with (39), Table EC.6 shows that the performance of LES is close to that of QL_m, when the arrival process is deterministic. The simulation estimates of $\mathrm{ASE}(\mathrm{LES})/\mathrm{ASE}(\mathrm{QL}_m)$, for alternative values of s, are remarkably close to the numerical value, approximately 1.286, predicted by (39); the relative error (RE) observed is less than 1% for all values of s considered. Consistent with (29), Table EC.7 shows that the performance of NI is worse than that of LES and QL_m. The simulation estimates of $\mathrm{ASE}(\mathrm{NI})/\mathrm{ASE}(\mathrm{QL}_m)$ are also remarkably close to the numerical value, 2.25, predicted by (29); the RE observed is less than 4% for all values of s considered.

EC.5.2. Results for the $H_2/M/s+M$ model

Table EC.8 compares the efficiencies of the alternative delay estimators in the $H_2/M/s+M$ model with $\alpha = 5.0$, which makes the model more lightly loaded. Consistent with theory, QL_m is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QL_m ranges from approximately 8% when $s = 100$ to approximately 2% when $s = 1000$.

Efficiency of the estimators in the $D/M/s+M$ model with $\rho = 1.4$ and $\alpha = 0.2$

s	ASE[θ_{QL_m}]	ASE[θ_{QL_r}]	ASE[θ_{QL}]	ASE[θ_{LES}]	ASE[θ_{NI}]
100	1.436×10^{-2}	1.492×10^{-2}	1.192×10^{-1}	1.863×10^{-2}	3.266×10^{-2}
	$\pm 9.78 \times 10^{-5}$	$\pm 9.40 \times 10^{-5}$	$\pm 1.57 \times 10^{-4}$	$\pm 1.64 \times 10^{-4}$	$\pm 5.33 \times 10^{-4}$
300	4.798×10^{-3}	5.005×10^{-3}	1.071×10^{-1}	6.172×10^{-3}	1.056×10^{-2}
	$\pm 5.99 \times 10^{-5}$	$\pm 6.08 \times 10^{-5}$	$\pm 1.41 \times 10^{-4}$	$\pm 7.45 \times 10^{-5}$	$\pm 1.92 \times 10^{-4}$
500	2.865×10^{-3}	2.966×10^{-3}	1.044×10^{-1}	3.672×10^{-3}	6.641×10^{-3}
	$\pm 5.43 \times 10^{-5}$	$\pm 5.24 \times 10^{-5}$	$\pm 1.071 \times 10^{-4}$	$\pm 6.67 \times 10^{-5}$	$\pm 2.933 \times 10^{-4}$
700	2.091×10^{-3}	2.170×10^{-3}	1.033×10^{-1}	2.691×10^{-3}	4.802×10^{-3}
	$\pm 2.39 \times 10^{-5}$	$\pm 1.90 \times 10^{-5}$	$\pm 1.53803 \times 10^{-4}$	$\pm 3.23 \times 10^{-5}$	$\pm 2.26 \times 10^{-4}$
1000	1.435×10^{-3}	1.507×10^{-3}	1.026×10^{-1}	1.859×10^{-3}	3.030×10^{-3}
	$\pm 1.15 \times 10^{-5}$	$\pm 1.52 \times 10^{-5}$	$\pm 1.20 \times 10^{-4}$	$\pm 2.06 \times 10^{-5}$	$\pm 1.05 \times 10^{-4}$

Table EC.5 **Point and confidence interval estimates of the ASEs - average square errors - of the estimators**

Comparison of the efficiency of LES and QL$_m$ in the $D/M/s+M$ model with $\rho = 1.4$ and $\alpha = 0.2$

s	ASE[θ_{QL_m}]	ASE[θ_{LES}]	ASE[θ_{LES}]/ASE[θ_{QL_m}]	Predicted ratio by (38)	RE (%)
100	1.436×10^{-2}	1.863×10^{-2}	1.297	1.286	0.885
	$\pm 9.78 \times 10^{-5}$	$\pm 1.642 \times 10^{-4}$			
300	4.798×10^{-3}	6.172×10^{-3}	1.286	1.286	0.0421
	$\pm 5.99 \times 10^{-5}$	$\pm 7.45 \times 10^{-5}$			
500	2.865×10^{-3}	3.672×10^{-3}	1.281	1.286	-0.329
	$\pm 5.43 \times 10^{-5}$	$\pm 6.67 \times 10^{-5}$			
700	2.091×10^{-3}	2.691×10^{-3}	1.287	1.286	0.107
	$\pm 2.39 \times 10^{-5}$	$\pm 3.23 \times 10^{-5}$			
1000	1.435×10^{-3}	1.859×10^{-3}	1.296	1.286	0.765
	$\pm 1.15 \times 10^{-5}$	$\pm 2.05 \times 10^{-5}$			

Table EC.6

In this more lightly loaded setting, the ASE's of all the estimators are relatively low, being smaller than for the $M/M/s+M$ model with $\alpha = 1.0$ in Table EC.1 by a factor of about 4, despite having $c_a^2 = 4.0$ instead of $c_a^2 = 1.0$. However, the lighter loading makes the ED heavy-traffic approximations less appropriate.

The QL$_r$ estimator is less efficient than QL$_m$: ASE(QL$_r$)/ASE(QL$_m$) ranges from approximately

Comparison of the efficiency of NI and QL$_m$ in the $D/M/s+M$ model with $\rho = 1.4$ and $\alpha = 0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{NI}]$	ASE$[\theta_{NI}]$/ASE$[\theta_{QL_m}]$	Predicted ratio by (28)	RE (%)
100	1.436×10^{-2} $\pm 9.78 \times 10^{-5}$	3.266×10^{-2} $\pm 5.33 \times 10^{-4}$	2.275	2.25	1.09
300	4.798×10^{-3} $\pm 5.99 \times 10^{-5}$	1.056×10^{-2} $\pm 1.92 \times 10^{-4}$	2.201	2.25	-2.18
500	2.865×10^{-3} $\pm 5.43 \times 10^{-5}$	6.641×10^{-3} $\pm 2.933 \times 10^{-4}$	2.318	2.25	3.01
700	2.091×10^{-3} $\pm 2.39 \times 10^{-5}$	4.802×10^{-3} $\pm 2.26 \times 10^{-4}$	2.297	2.25	2.08
1000	1.435×10^{-3} $\pm 1.15 \times 10^{-5}$	3.130×10^{-3} $\pm 1.05 \times 10^{-4}$	2.111	2.25	-3.08

Table EC.7

Efficiency of the estimators in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\alpha = 5.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	7.193×10^{-4} $\pm 2.63 \times 10^{-6}$	1.059×10^{-3} $\pm 4.47 \times 10^{-6}$	2.217×10^{-3} $\pm 1.01 \times 10^{-5}$	2.393×10^{-3} $\pm 6.72 \times 10^{-6}$	3.101×10^{-3} $\pm 1.42 \times 10^{-5}$
300	2.008×10^{-4} $\pm 7.85 \times 10^{-7}$	2.675×10^{-4} $\pm 1.28 \times 10^{-6}$	7.240×10^{-4} $\pm 2.63 \times 10^{-6}$	7.569×10^{-4} $\pm 2.70 \times 10^{-6}$	1.169×10^{-3} $\pm 5.82 \times 10^{-6}$
500	1.167×10^{-4} $\pm 7.05 \times 10^{-7}$	1.495×10^{-4} $\pm 8.78 \times 10^{-7}$	4.792×10^{-4} $\pm 2.68 \times 10^{-6}$	4.540×10^{-4} $\pm 1.71 \times 10^{-6}$	7.624×10^{-4} $\pm 6.07 \times 10^{-6}$
700	8.277×10^{-5} $\pm 4.12 \times 10^{-7}$	1.042×10^{-4} $\pm 6.52 \times 10^{-7}$	3.856×10^{-4} $\pm 2.50 \times 10^{-6}$	3.280×10^{-4} $\pm 1.27 \times 10^{-6}$	5.714×10^{-4} $\pm 4.72 \times 10^{-6}$
1000	5.733×10^{-5} $\pm 2.48 \times 10^{-7}$	7.141×10^{-5} $\pm 2.44 \times 10^{-7}$	3.184×10^{-4} $\pm 1.34 \times 10^{-6}$	2.302×10^{-4} $\pm 1.19 \times 10^{-6}$	4.0951×10^{-4} $\pm 4.15 \times 10^{-6}$

Table EC.8 Point and confidence interval estimates of the ASEs - average square errors - of the estimators

1.5 when $s = 100$ to approximately 1.25 when $s = 1000$. The LES estimator is less accurate, with an RRASE ranging from approximately 14% when $s = 100$ to approximately 4% when $s = 1000$. The QL estimator performs slightly worse than LES: The ratio ASE(QL)/ASE(QL$_m$) ranges from about 3 when $s = 100$ to about 5 when $s = 1000$. The NI estimator is the least efficient estimator for this model.

Comparison of the efficiency of LES and $\mathbf{QL}_m$ in the $H_2/M/s+M$ model with $\rho=1.4$ and $\alpha=5.0$

s	$\text{ASE}[\theta_{QL_m}]$	$\text{ASE}[\theta_{LES}]$	$\text{ASE}[\theta_{LES}]/\text{ASE}[\theta_{QL_m}]$	Predicted by (39)	RE (%)
100	7.193×10^{-4} $\pm2.63\times10^{-6}$	2.393×10^{-3} $\pm6.72\times10^{-6}$	3.326	4.143	-19.7
300	2.008×10^{-4} $\pm7.85\times10^{-7}$	7.569×10^{-4} $\pm2.70\times10^{-6}$	3.769	4.143	-9.03
500	1.167×10^{-4} $\pm7.05\times10^{-7}$	4.540×10^{-4} $\pm1.71\times10^{-6}$	3.891	4.143	-6.09
700	8.277×10^{-5} $\pm4.12\times10^{-7}$	3.280×10^{-4} $\pm1.27\times10^{-6}$	3.962	4.143	-4.36
1000	5.733×10^{-5} $\pm2.48\times10^{-7}$	2.302×10^{-4} $\pm1.19\times10^{-6}$	4.014	4.143	-3.10

Table EC.9

Comparison of the efficiency of NI and $\mathbf{QL}_m$ in the $H_2/M/s+M$ model with $\rho=1.4$ and $\alpha=5.0$

s	$\text{ASE}[\theta_{QL_m}]$	$\text{ASE}[\theta_{NI}]$	$\text{ASE}[\theta_{NI}]/\text{ASE}[\theta_{QL_m}]$	Predicted ratio by (28)	RE (%)
100	7.193×10^{-4} $\pm2.63\times10^{-6}$	3.101×10^{-3} $\pm1.42\times10^{-5}$	4.310	7.25	-40.5
300	2.008×10^{-4} $\pm7.85\times10^{-7}$	1.169×10^{-3} $\pm5.82\times10^{-6}$	5.821	7.25	-19.7
500	1.167×10^{-4} $\pm7.05\times10^{-7}$	7.624×10^{-4} $\pm6.07\times10^{-6}$	6.533	7.25	-9.89
700	8.277×10^{-5} $\pm4.12\times10^{-7}$	5.714×10^{-4} $\pm4.72\times10^{-6}$	6.904	7.25	-4.78
1000	5.733×10^{-5} $\pm2.48\times10^{-7}$	4.0951×10^{-4} $\pm4.15\times10^{-6}$	7.143	7.25	-1.48

Table EC.10

Tables EC.9 and EC.10 substantiate (40) and (29) of §5, that compare the performances of QL_m, LES and NI in the $H_2/M/s+M$ model. Consistent with (40), Table EC.9 shows that the performance of LES is significantly worse than that of QL_m, when the arrival process is highly variable. The simulation estimates of $\text{ASE(LES)}/\text{ASE}(\text{QL}_m)$, for alternative values of s, are close to the numerical value, approximately 4.143, predicted by (40), especially for large values of s; the

RE observed ranges from approximately −20% for $s = 100$ to approximately −3% when $s = 1000$. We observe a relatively poor performance of the approximation in (40) when the number of servers is small. That is understandable because the system is not very heavily loaded when $\alpha = 5.0$. Consistent with (29), Table EC.10 shows that the performance of NI is much worse than that of QL_m, when the arrival process is highly variable. The approximation in (29) performs poorly when $s = 100$ (RE $\approx -40\%$) but becomes remarkably accurate when $s = 1000$ (RE $\approx -1.5\%$).

References

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. J. Amer. Statist. Assoc. 100: 3650.

Ibrahim, R. and W. Whitt. 2008. Supplement to “Real-time delay estimation in overloaded multiserver queues with abandonments” IEOR Department, Columbia University, New York, NY. Available at http://columbia.edu/∼ww2040.