e-Companion

to

REAL-TIME DELAY ESTIMATION BASED ON DELAY HISTORY

by
Rouba Ibrahim and Ward Whitt

IEOR Department
Columbia University
{rei2101, ww2040}@columbia.edu

A. Introduction

We present additional material in this e-companion. First, in §B we present additional experimental results; we present many more in an online supplement available on the authors' web pages. Next, in §C we establish steady-state heavy-traffic limits for these estimators. At the end of the section, we show that the bad performance of the LCS estimator for large s can be explained in part by its behavior in the QED many-server heavy-traffic limiting regime. Unlike the LES, HOL and RCS delay estimators, the LCS delay estimator is *not* asymptotically consistent in this limiting regime. Finally, in §D we present a cautionary example showing the possible pitfalls of the LES and HOL delay estimators.

B. Additional Tables and Figures

Paralleling Table 1 in §3, which displays the ASE's for seven different estimators in the $GI/M/100$ model for the M, D and H_2 arrival processes, we display the corresponding estimated ASE's for the same estimators for the $GI/M/s$ models with $s = 10$ and $s = 1$ in Tables 6 and 7 below. The estimator LCS fares better as s decreases. The ASE's of LCS and RCS do not differ greatly for $s = 10$ and are identical for $s = 1$.

Paralleling Figure 1 in §3, where we display plots of the relative average squared errors (RASE's) for several of the estimators in the $D/M/100$ model, we display the RASE's for the $M/M/100$ and $H_2/M/100$ models in Figures 2 and 3. Again we see linear or near-linear performance as a function of ρ. The advantage of QL over LES increases as c_a^2 increases. Again the HOL and LES values fall on top of each other, so we only show LES.

Paralleling Table 3 in §4, where we compare the approximations for the MSE's of the three estimators θ_{HOL}^{d}, θ_{HOL}^{ar} and θ_{HOL}^{sr} in the $H_2/M/s$ model with $s = 100$ and $s = 1$, we show

Estimated ASE in units of 10^{-1}

$M/M/s$ *model with* $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	4.95	10.1	10.1	10.8	10.9	11.9	257.2
	±0.23	±0.42	±0.41	±0.41	±0.42	±0.41	±48.1
0.95	1.98	4.16	4.17	4.83	4.94	5.87	39.61
	±0.025	±0.040	±0.042	±0.039	±0.041	±0.041	±2.3
0.93	1.42	3.03	3.05	3.67	3.77	4.62	20.01
	±0.013	±0.032	±0.037	±0.036	±0.033	±0.036	±0.66
0.9	1.00	2.19	2.20	2.79	2.88	3.63	10.10
	±0.017	±0.033	±0.042	±0.036	±0.035	±0.036	±0.49
0.85	0.661	1.50	1.53	2.04	2.11	2.69	4.41
	±0.0032	±0.0076	±0.012	±0.0092	±0.0085	±0.0097	±0.083

$D/M/s$ *model with* $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	2.49	2.63	2.63	2.99	3.05	3.57	59.3
	±0.084	±0.083	±0.086	±0.085	±0.086	±0.086	±10.2
0.95	1.01	1.16	1.16	1.50	1.55	2.00	10.1
	±0.018	±0.018	±0.020	±0.019	±0.019	±0.019	±0.83
0.93	0.730	0.876	0.877	1.21	1.26	1.66	5.24
	±0.010	±0.011	±0.013	±0.012	±0.011	±0.012	±0.29
0.9	0.518	0.663	0.663	0.977	1.02	1.37	2.66
	±0.0058	±0.0057	±0.0091	±0.0077	±0.0066	±0.0078	±0.12
0.85	0.352	0.494	0.494	0.779	0.814	1.06	1.24
	±0.0025	±0.0026	±0.0057	±0.0047	±0.0028	±0.0047	±0.0053

$H_2/M/s$ *model with* $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	12.8	62.6	62.6	64.4	65.1	67.3	1594
	±0.69	±4.0	±4.1	±4.1	±4.1	±5.6	±258
0.95	4.81	22.3	22.3	23.9	24.6	26.5	229
	±0.081	±0.47	±0.48	±0.47	±0.47	±0.81	±9.1
0.93	3.42	15.4	15.4	17.0	17.5	19.4	115
	±0.069	±0.35	±0.37	±0.35	±0.35	±0.35	6.8
0.9	2.34	10.1	10.1	11.6	11.8	13.7	54.4
	±0.036	±0.18	±0.20	±0.19	±0.18	±0.18	±2.9
0.85	1.50	6.00	6.02	7.25	7.50	8.97	22.8
	±0.022	±0.12	±0.13	±0.12	±0.13	±0.076	±1.37

Table 6: A comparison of the efficiency of different real-time delay estimators for the $GI/M/10$ queue as a function of the traffic intensity ρ and the interarrival-time distribution (M, D and H_2). Only the direct estimators are considered. Estimates of the average squared error ASE are shown together with the half width of the 95% confidence interval. The units are 10^{-1} throughout.

Estimated ASE

$M/M/s$ *model with* $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	20.1	42.2	42.4	44.1	44.1	44.1	405.0
	±0.42	±0.77	±0.79	±0.78	±0.78	±0.78	±23.4
0.93	14.4	30.6	30.7	32.4	32.4	32.4	207.5
	±0.19	±0.37	±0.39	±0.37	±0.37	±0.37	±10.4
0.9	9.99	21.8	22.0	23.5	23.5	23.5	100.6
	±0.084	±0.19	±0.21	±0.19	±0.19	±0.19	±3.4
0.85	6.68	15.1	15.4	16.6	16.6	16.6	44.9
	±0.043	±0.093	±0.095	±0.010	±0.010	±0.010	±0.88

$D/M/s$ *model with* $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	10.1	11.6	11.6	12.6	12.6	12.6	101.1
	±0.15	±0.15	±0.16	±0.15	±0.15	±0.15	±7.2
0.93	7.32	8.79	8.79	9.73	9.73	9.73	52.7
	±0.081	±0.078	±0.086	±0.080	±0.080	±0.080	±2.4
0.9	5.19	6.64	6.65	7.56	7.56	7.56	26.8
	±0.038	±0.037	±0.041	±0.040	±0.040	±0.040	±0.94
0.85	3.53	4.96	4.95	5.82	5.82	5.82	12.4
	±0.018	±0.018	±0.020	±0.020	±0.021	±0.020	±0.36

$H_2/M/s$ *model with* $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	48.7	226.4	226.5	231.1	231.1	231.1	2339
	±1.13	±5.14	±5.23	±5.15	±5.15	±5.15	±425
0.93	34.3	154.4	154.4	158.9	158.9	158.9	1151
	±0.63	±2.9	±2.9	±3.0	±3.0	±3.0	±181
0.9	23.48	101.3	101.4	105.5	105.5	105.5	552.9
	±0.37	±2.3	±2.4	±2.4	±2.4	±2.4	±103
0.85	14.95	60.0	60.2	63.9	63.9	63.9	224.4
	±0.104	±0.52	±0.53	±0.51	±0.51	±0.51	±6.2

Table 7: A comparison of the efficiency of different real-time delay estimators for the $GI/M/1$ queue as a function of the traffic intensity ρ and the interarrival-time distribution (M, D and H_2). Only the direct estimators are considered. Estimates of the average squared error ASE are shown together with the half width of the 95% confidence interval.

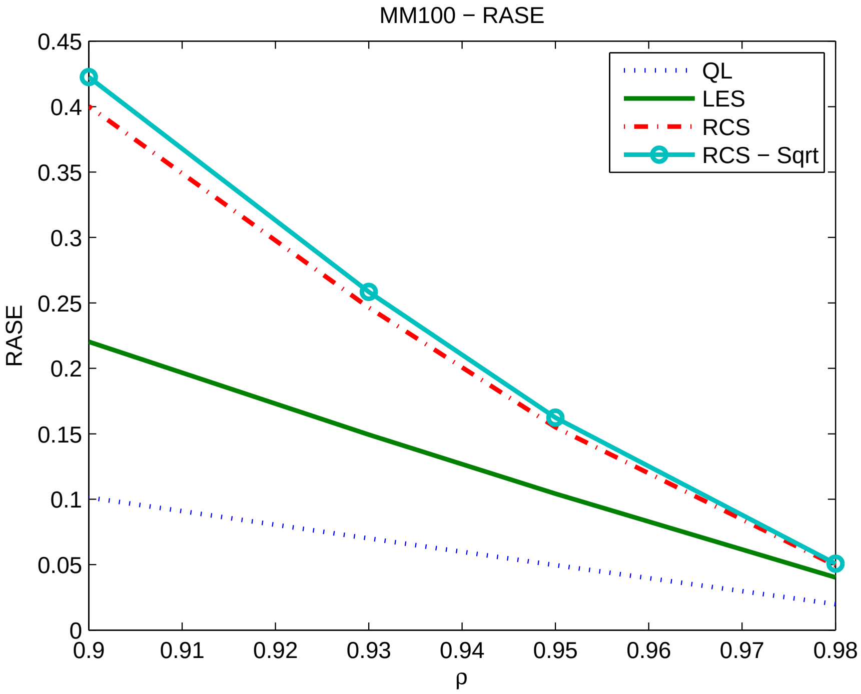

Figure 2: The relative average squared error ($RASE$) for the $M/M/100$ model.

corresponding results for the $M/M/s$ model with $s = 100$ and $s = 1$ in Table 8. We have used simulation to estimate all quantities here, even though we could compute them analytically. This case thus provides a crosscheck on both our analytic formulas and the simulations.

Finally, we present one table illustrating our study of the number of past customers we need to consider for RCS, as discussed at the end of §2. Table 9 present simulation results for the $H_2/M/100$ model as a function of ρ. These results support the conclusion that $RCS - c\sqrt{s}$ is virtually identical to RCS itself when $c = 4$, and that small errors are observed when $c = 2$ and $s = 1$. These conclusions held uniformly over all interarrival-time distributions and all $s \geq 100$.

Evaluating the alternative HOL estimators

Approximations in the $M/M/s$ model for $s = 100$ and $s = 1$

ρ	0.85	0.90	0.93	0.95	0.98	0.99
$E[W\|W > 0]$	0.0666	0.0993	0.1435	0.2012	0.500	0.901
conf. int.	±0.0018	±0.0027	±0.0018	±0.0019	±0.037	±0.059
$E[W^2\|W > 0]$	0.0089	0.0196	0.0414	0.0811	0.500	1.53
conf. int.	±0.0006	±0.0012	±0.0016	±0.0026	±0.097	±0.24
$MSE(\theta^d)$	0.00153	0.00219	0.00307	0.00422	0.01020	0.01823
term 1	0.00020	0.00020	0.00020	0.00020	0.00020	0.00015
term 2	0.00073	0.00139	0.00227	0.00342	0.00940	0.01748
term 3	0.00060	0.00060	0.00060	0.00060	0.00060	0.00060
$MSE(\theta^{sr})$	0.00173	0.00239	0.00327	0.00442	0.01040	0.01844
term 1	0.00113	0.00179	0.00267	0.00382	0.00980	0.01784
term 2	0.00060	0.00060	0.00060	0.00060	0.00060	0.00060
$MSE(\theta^{ar})$	0.00133	0.00199	0.00287	0.00402	0.01000	0.01804
term 1	0.00113	0.00179	0.00267	0.00382	0.00980	0.01784
term 2	0.00020	0.00020	0.00020	0.00020	0.00020	0.00020

Approximations in the $M/M/1$ model

ρ	0.80	0.85	0.90	0.95	0.96	0.98
$E[W\|W > 0]$	5.01	6.68	9.98	20.04	24.80	50.70
conf. int.	±0.03	±0.04	±0.08	±0.36	±0.33	±2.4
$E[W^2\|W > 0]$	50.3	89.6	200.3	806.6	1211	5290
conf. int.	±0.69	±1.36	±5.1	±37.4	±45	640
$MSE(\theta^d)$	12.02	15.36	21.98	42.08	51.58	103.4
term 1	2.01	2.01	2.00	2.02	1.94	2.11
term 2	4.01	7.35	13.98	34.07	43.64	95.25
term 3	6.00	6.00	6.00	6.00	6.00	6.00
$MSE(\theta^{sr})$	14.02	17.35	23.97	44.07	53.61	105.31
term 1	8.02	11.35	17.97	38.07	47.61	99.31
term 2	6.00	6.00	6.00	6.00	6.00	6.00
$MSE(\theta^{ar})$	10.02	13.35	19.97	40.07	49.61	101.31
term 1	8.02	11.35	19.97	38.07	47.61	99.31
term 2	2.00	2.00	2.00	2.00	2.00	2.00

Table 8: Evaluation of the MSE approximations for the estimators θ^d_{HOL}, θ^{sr}_{HOL}, and θ^{ar}_{HOL} in steady-state using (4.11), (4.9) and (4.10) together with simulation estimates of the first two moments of the conditional delay $E[W_\infty | W_\infty > 0]$. The $M/M/s$ model is considered as a function of the traffic intensity ρ for $s = 100$ and $s = 1$.

			ASE *in the* $H_2/M/s$ *model with* $s = 100$			
ρ	$ASE(RCS)$	$ASE(RCS - s)$	$ASE(RCS - 4\sqrt(s))$	$ASE(RCS - 2\sqrt(s))$	$ASE(RCS - \sqrt(s))$	$ASE(RCS - log(s))$
0.98	2.439×10^{-2} $\pm 4.84 \times 10^{-4}$	2.439×10^{-2} $\pm 4.84 \times 10^{-4}$	2.439×10^{-2} $\pm 4.84 \times 10^{-4}$	$2.442 \times 10^{-2}(\mathbf{0.123})$ $\pm 4.88 \times 10^{-4}$	$2.511 \times 10^{-2}(\mathbf{2.95})$ $\pm 4.92 \times 10^{-4}$	$3.724 \times 10^{-2}(\mathbf{52.7})$ $\pm 6.81 \times 10^{-4}$
0.97	2.229×10^{-2} $\pm 4.70 \times 10^{-4}$	2.229×10^{-2} $\pm 4.70 \times 10^{-4}$	2.229×10^{-2} $\pm 4.70 \times 10^{-4}$	$2.229 \times 10^{-2}(\mathbf{0.141})$ $\pm 4.73 \times 10^{-4}$	$2.367 \times 10^{-2}(\mathbf{3.28})$ $\pm 4.73 \times 10^{-4}$	$3.566 \times 10^{-2}(\mathbf{55.6})$ $\pm 5.80 \times 10^{-4}$
0.95	1.989×10^{-2} $\pm 3.67 \times 10^{-4}$	1.989×10^{-2} $\pm 3.67 \times 10^{-4}$	1.989×10^{-2} $\pm 3.67 \times 10^{-4}$	$1.992 \times 10^{-2}(\mathbf{0.136})$ $\pm 3.67 \times 10^{-4}$	$2.058 \times 10^{-2}(\mathbf{3.48})$ $\pm 3.64 \times 10^{-4}$	$3.175 \times 10^{-2}(\mathbf{59.6})$ $\pm 5.45 \times 10^{-4}$
0.93	1.715×10^{-2} $\pm 3.56 \times 10^{-4}$	1.715×10^{-2} $\pm 3.56 \times 10^{-4}$	1.715×10^{-2} $\pm 3.56 \times 10^{-4}$	$1.718 \times 10^{-2}(\mathbf{0.150})$ $\pm 3.54 \times 10^{-4}$	$1.780 \times 10^{-2}(\mathbf{3.78})$ $\pm 3.60 \times 10^{-4}$	$2.800 \times 10^{-2}(\mathbf{63.2})$ $\pm 5.89 \times 10^{-4}$
0.90	1.344×10^{-2} $\pm 4.90 \times 10^{-4}$	1.344×10^{-2} $\pm 4.90 \times 10^{-4}$	1.344×10^{-2} $\pm 4.90 \times 10^{-4}$	$1.347 \times 10^{-2}(\mathbf{0.182})$ $\pm 4.89 \times 10^{-4}$	$1.399 \times 10^{-2}(\mathbf{4.06})$ $\pm 4.99 \times 10^{-4}$	$2.233 \times 10^{-2}(\mathbf{66.3})$ $\pm 8.61 \times 10^{-4}$

Table 9: A comparison of the efficiency of the candidate RCS-$f(s)$ delay estimators for the $H_2/M/s$ queue with $s = 100$ as a function of the traffic intensity ρ. Below each point estimates for ASE is shown with the half width of the 95-percent confidence interval. Also included in parentheses are the values of the relative percent difference.

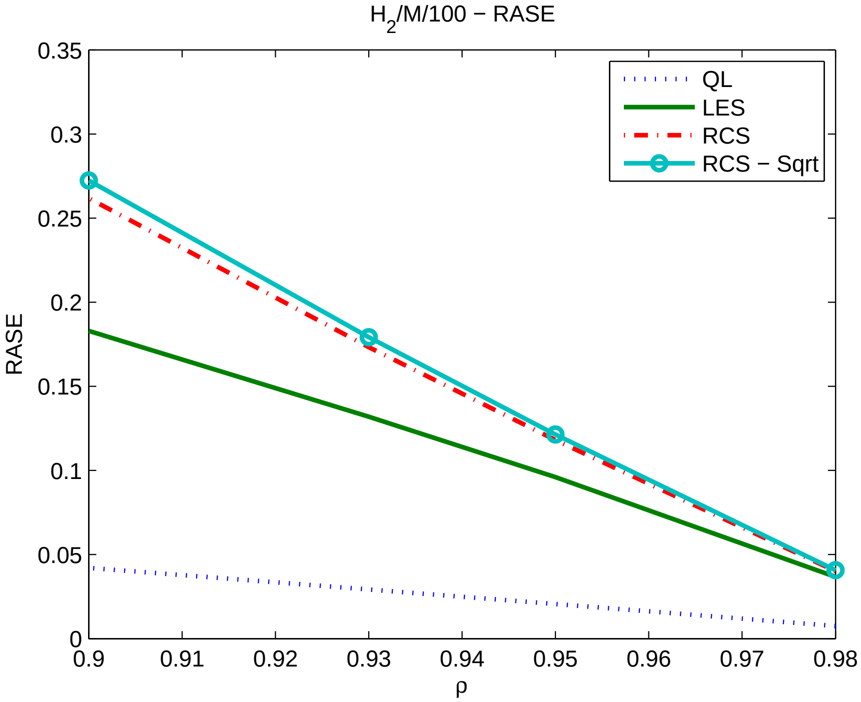

Figure 3: The relative average squared error ($RASE$) for the $H_2/M/100$ model.

C. Heavy-Traffic Limits

In this section we present additional heavy-traffic limits, extending the discussion in §6. We start by establishing heavy-traffic limits for the steady-state random variables. We see what happens "on average" to the random variable $W_{HOL,s,\rho}(w)$. We consider both the classical heavy-traffic regime in which $\rho \uparrow 1$ for fixed s and the QED (many-server heavy-traffic limiting) regime in which both $\rho \uparrow 1$ and $s \to \infty$ with $((1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$; see Chapters 5, 9 and 10 of Whitt (2002) for background. For more on the QED regime for $GI/G/s$ queues, see Halfin and Whitt (1981), Puhalskii and Reiman (2000), Jelenkovic et al. (2004) and Whitt (2004b, 2005).

The Classical Heavy-Traffic Regime. We start with the classic heavy-traffic (HT) regime in which $\rho \uparrow 1$ with fixed s. We look at the distribution of $W_{HOL,s}(w)$, assuming that the observed waiting time w experienced by the customer at the head of the line is a random variable $W^h_{\infty,s,\rho}$, assumed to be the steady-state delay in model (s,ρ) experienced by a customer at the head of the line at an arrival epoch, conditional on there being at least one customer in the queue. Thus let $W_{HOL,s,\rho}(W^h_{\infty,s,\rho})$ denote a random variable with the distribution

$$P(W_{HOL,s,\rho}(W^h_{\infty,s,\rho}) \le x) \equiv \int_0^\infty P(W_{HOL,s,\rho}(w) \le x)\, dP(W^h_{\infty,s,\rho} \le w) \ , \qquad (3.1)$$

in model (s,ρ), where in this subsection s is held fixed. This means that $E[W_{HOL,s,\rho}(W^h_{\infty,s,\rho})] \equiv E[E[W_{HOL,s,\rho}(W^h_{\infty,s,\rho})|W^h_{\infty,s,\rho}]]$. The random variable $W^h_{\infty,s,\rho}$ is not quite distributed as the steady-state waiting time at the arrival epoch, $W_{\infty,s,\rho}$, or the conditional steady-state waiting time, $(W_{\infty,s,\rho}|W_{\infty,s,\rho} > 0)$, but it is asymptotically equivalent to both of these in the heavy-traffic limit.

In order to relate the HOL and QL estimators, it is important to exploit the joint convergence of the steady-state queue length and waiting time. Such joint convergence is discussed extensively for the single-server queue in Chapter 9 of Whitt (2002); it was also used in Iglehart and Whitt (1970), which treated more general models. Let $(Q_{\infty,s,\rho}, W_{\infty,s,\rho})$ be a random vector with the limiting steady-state distribution of $(Q_{k,s,\rho}, W_{k,s,\rho})$, where $Q_{k,s,\rho}$ is the queue length and $W_{k,s,\rho}$ is the delay just before $A_{k,s,\rho}$, where $A_{k,s,\rho}$ is the k^{th} arrival epoch, all in model (s,ρ).

Here we will use the following established steady-state heavy-traffic limit:

$$(1-\rho)(Q_{\infty,s,\rho}, W_{\infty,s,\rho}) \Rightarrow (L, L/s) \quad \text{as} \quad \rho \uparrow 1 \ , \qquad (3.2)$$

where $L \overset{\mathrm{d}}{=} Exp(c_a^2 + 1)/2$ with $Exp(m)$ denoting a random variable having an exponential distribution with mean m. We give a detailed proof in a subsection below starting from the known steady-state distribution for $Q_{\infty,s,\rho}$. The joint convergence follows from the limit for $Q_{\infty,s,\rho}$ and the law of large numbers, using the representation

$$(Q_{\infty,s,\rho}, W_{\infty,s,\rho}) = \left(Q_{\infty,s,\rho}, (Q_{\infty,s,\rho}+1)\left(\left[\sum_{i=1}^{Q_{\infty,s,\rho}+1} (V_i/s)\right]/(Q_{\infty,s,\rho}+1)\right)\right) . \tag{3.3}$$

We can apply (3.2) and previous results to get the following limits for our estimators. Let RMSE $\equiv MSE/Mean^2$ be the relative mean squared error. Let $c^2_{W_{Q,s,\rho}}(Q_{\infty,s,\rho})$ be the random variable assuming the value $c^2_{W_{Q,s,\rho}}(n)$ with probability $P(Q_{\infty,s,\rho} = n)$ for $n \geq 0$. Let other random variables involving c^2 and RMSE be defined analogously. We prove the following theorem in a subsection below.

Theorem C.1. (*classical heavy-traffic limit*) *If $\rho \uparrow 1$ in the family of $GI/M/s$ models indexed by (s, ρ) with fixed s, then*

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{E[W_{Q,s,\rho}(Q_{\infty,s,\rho})|Q_{\infty,s,\rho}]} = \frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{(Q_{\infty,s,\rho}+1)/s} \Rightarrow 1 \ , \tag{3.4}$$

$$\frac{W_{\infty,s,\rho}}{W^h_{\infty,s,\rho}} \Rightarrow 1 \quad \text{and} \quad \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} \Rightarrow 1 \ , \tag{3.5}$$

from which we can deduce that

$$(1-\rho)(Q_{\infty,s,\rho}, W_{\infty,s,\rho}, W^h_{\infty,s,\rho}, W_{Q,s,\rho}(Q_{\infty,s,\rho}), W_{HOL,s,\rho}(W^h_{\infty,s,\rho})) \Rightarrow (L, L/s, L/s, L/s, L/s) \tag{3.6}$$

and

$$(1-\rho)^{-1}(c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}, c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}, RMSE(W^h_{\infty,s,\rho})) \Rightarrow (1/L, (c_a^2+1)/L, (c_a^2+1)/L) \tag{3.7}$$

where $L \overset{\mathrm{d}}{=} Exp((c_a^2+1)/2)$ as above, so that

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W_{Q,s,\rho}(Q_{\infty,s,\rho})} \Rightarrow 1, \quad \frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2 + 1 \ , \tag{3.8}$$

$$\frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}} \Rightarrow 1 \quad \text{and} \quad \frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2 + 1 \ . \tag{3.9}$$

The limits in (3.4) and (3.5) show that the direct QL and HOL estimators are (weakly) relatively consistent in the classical heavy-traffic limit, while the limits in (3.7)–(3.9) compare the asymptotic efficiency of the different estimators. In this heavy traffic limit, the direct and refined HOL estimators have asymptotically the same efficiency, while the QL estimator is asymptotically more efficient by the constant factor $c_a^2 + 1$.

We conjecture (but have not yet proved) that there is appropriate uniform integrability, so that the moments of these random variables converge as well as distributions, see p. 31 of Billingsley (1999). Then from (3.7) and (3.8) we obtain associated convergence of the moments:

$$E\left[\frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}}\right] \to c_a^2 + 1 \quad \text{and} \quad \frac{E[c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}]}{E[c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}]} \to c_a^2 + 1 \ , \tag{3.10}$$

and similarly for the direct estimator. These limits supplement the previous limits, implying that the QL delay estimator is asymptotically more efficient than the HOL and LES delay estimators by the constant factor $c_a^2 + 1$ in the classical heavy-traffic limit.

The QED Many-Server Heavy-Traffic Regime. We now consider the QED HT regime, in which both $\rho \uparrow 1$ and $s \uparrow \infty$ with $(1-\rho)\sqrt{s} \to \beta$ for some positive constant β.

This alternative QED regime is appealing because, unlike the classical HT regime, the probability that a customer is delayed approaches a nondegenerate limit, strictly between 0 and 1:

$$P(W_{\infty,s,\rho} > 0) \to \alpha \quad \text{and} \quad P(Q_{\infty,s,\rho} > 0) \to \alpha, \quad 0 < \alpha < 1 \ , \tag{3.11}$$

where $\alpha \equiv \alpha(\beta/\sqrt{c_a^2+1})$ for $\alpha(x) \equiv [1 + x\Phi(x)/\phi(x)]^{-1}$, where ϕ is the cdf and ϕ is the probability density function (pdf) of the standard normal $N(0,1)$; see (1.1) of Whitt (2004b).

With minor modifications, the story is the same as for the classical HT regime, so we will be brief. A major difference is that the queue length is of order $O(\sqrt{s}) = O(1/(1-\rho))$, while the waiting time is of order $O(1/\sqrt{s}) = O((1-\rho))$. As before, the ratio $W_{\infty,s,\rho}/Q_{\infty,s,\rho}$ is of order $O(1/s)$, but now $s \to \infty$.

Paralleling (3.2), we have the joint limit

$$(Q_{\infty,s,\rho}/\sqrt{s}, (1-\rho)Q_{\infty,s,\rho}, \sqrt{s}W_{\infty,s,\rho}, W_{\infty,s,\rho}/(1-\rho)) \Rightarrow (Z, \beta Z, Z, Z/\beta) \ , \tag{3.12}$$

where $P(Z > 0) = \alpha$ for the same $\alpha \equiv \alpha(\beta/\sqrt{c_a^2+1})$ defined above and $(Z|Z > 0) \overset{\mathrm{d}}{=} L \overset{\mathrm{d}}{=} Exp((c_a^2+1)/2)$. The limit for $Q_{\infty,s,\rho}$ was established by Halfin and Whitt (1981), but Whitt (2004b) corrects an error in the expression for α when the arrival process is non-Poisson. The joint limit with $W_{\infty,s,\rho}$ can be established as in (3.3). Paralleling (3.39), here we have

$$\begin{aligned} &((1-\rho)(Q_{\infty,s,\rho}|Q_{\infty,s,\rho} > 0), (W_{\infty,s,\rho}|W_{\infty,s,\rho} > 0)/(1-\rho), W^h_{\infty,s,\rho}/(1-\rho), (1-\rho)A(W^h_{\infty,s,\rho})) \\ \Rightarrow \quad &(\beta L, L/\beta, L/\beta, \beta L) \ , \end{aligned} \tag{3.13}$$

where again $L \stackrel{\mathrm{d}}{=} (Z|Z > 0) \stackrel{\mathrm{d}}{=} Exp(c_a^2 + 1)/2$; as before, the important point is that the same random variable L appears in all four components on the right.

We now state the theorem, omitting the proof.

Theorem C.2. (*QED heavy-traffic limit*) *If $\rho \uparrow 1$ and $s \uparrow \infty$ so that $(1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$ in the family of $GI/M/s$ models indexed by ρ and s, then*

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{(Q_{\infty,s,\rho}+1)/s} \Rightarrow 1 \quad and \quad \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} \Rightarrow 1 \ . \tag{3.14}$$

$$(1-\rho)^{-1}(W_{Q,s,\rho}(Q_{\infty,s,\rho}), W_{HOL,s,\rho}(W^h_{\infty,s,\rho})) \Rightarrow (L/\beta, L/\beta) \tag{3.15}$$

and

$$(1-\rho)^{-1}(c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}, c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}, RMSE(W^h_{\infty,\rho,s})) \Rightarrow (1/\beta L, (c_a^2+1)/\beta L, (c_a^2+1)/\beta L) \tag{3.16}$$

where $L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$ as above, so that

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W_{Q,s,\rho}(Q_{\infty,s,\rho})} \Rightarrow 1, \quad \frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2 + 1 \ . \tag{3.17}$$

$$\frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}} \Rightarrow 1 \quad and \quad \frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2 + 1 \ . \tag{3.18}$$

Just as in the classical HT regime, we conjecture that there is appropriate uniform integrability, so that the moments converge as well as distributions. Then we will obtain associated convergence of the moments, just as in (3.10).

Heavy-Traffic Detail: Proof of (3.2). In this section we prove the classical heavy-traffic limit for the steady-state joint distribution of the queue length and waiting time at arrival epochs stated in (3.2):

$$(1-\rho)(Q_{\infty,\rho}, W_{\infty,\rho}) \Rightarrow (L, L/s) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{3.19}$$

where $L \stackrel{\mathrm{d}}{=} Exp(c_a^2 + 1)/2$ with $Exp(m)$ denoting a random variable that is exponentially distributed with mean m. We consider this a known result, but we cannot point to a place where a proof is given.

We draw on well-known properties of the steady-state distribution of the $GI/M/s$ queue. The key initial result is the fact that the conditional distribution of the queue length at an arrival epoch, given that the arrival must wait, is a geometric distribution, i.e.,

$$P(Q_{\infty,\rho} = j | W_{\infty,\rho} > 0) = (1-\omega)\omega^j, \quad j \geq 0 \ , \tag{3.20}$$

where the single parameter ω in (3.20) is the unique root of the equation

$$\omega = \int_0^\infty e^{-(1-\omega)sx}\, dF(x) \equiv \hat{f}((1-\omega)s) \ , \tag{3.21}$$

where $\hat{f}$ is the Laplace-Stieltjes transform of the cdf F, i.e.,

$$\hat{f}(z) \equiv \int_0^\infty e^{-zx}\, dF(x) \ ; \tag{3.22}$$

see (14.10), (14.11), (14.12) and (14.19) of Cooper (1982). This property was used in the proof of Theorem 4.3.

The key then is the way that the root $\omega \equiv \omega(\rho)$ depends on the traffic intensity ρ as $\rho \uparrow 1$. Anticipating that we should have $\omega(\rho) \uparrow 1$ as $\rho \uparrow 1$, we see that the argument of the Laplace-Stieltjes transform should approach 0 in the limit. It should thus come as no surprise that we can rigorously establish the desired result by expanding the Laplace transform $\hat{f}(z)$ in a Taylor series about $z = 0$; see p. 435 of Feller (1971) for supporting theory. As was first observed by Smith (1953, p. 461), it follows that

$$\frac{1-\omega(\rho)}{1-\rho} \to \frac{2}{c_a^2+1} \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.23}$$

The expansion appears in a more general context in formula (17) of Abate and Whitt (1994). In the special case of the $GI/M/s$ queue, equation (7) there reduces to equation (3.21) here. An alternative approach involving upper and lower bounds is given in Whitt (1984); that focuses on the more elementary $GI/M/1$ model, but the key root has the same structure. The equation differs only by the constant factor s appearing in the equation (3.21). Additional theoretical results about characterizing roots for queues appears in Neuts (1986), Choudhury and Whitt (1994) and Glynn and Whitt (1994).

It is well known – see pages 1-2 of Feller (1971) – that if X_m is a random variable with a geometric distribution having mean m, then

$$\frac{X_m}{cm} \Rightarrow Exp(1/c) \quad \text{as} \quad m \to \infty \ . \tag{3.24}$$

By (3.20), $(Q_{\infty,\rho}|W_{\infty,\rho} > 0)$ has a geometric distribution with mean $1/(1-\omega(\rho))$. Thus we can combine (3.20), (3.23) and (3.24) to obtain

$$(1-\rho)(Q_{\infty,\rho}|W_{\infty,\rho} > 0) \Rightarrow Exp((c_a^2+1)/2) \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.25}$$

It is also known that

$$P(W_{\infty,\rho} > 0) = \frac{A}{1-\omega} \quad \text{where} \quad A = \left[\frac{1}{1-\omega} + X\right]^{-1} \ , \tag{3.26}$$

with $X \equiv X(\rho) \to X(1)$, $0 < X(1) < \infty$, as $\rho \uparrow 1$; see (14.14)–(14.17) of Cooper (1982). Hence

$$P(W_{\infty,\rho} > 0) = [1 + (1 - \omega(\rho))X(\rho)]^{-1} \to 1 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.27}$$

Combining (3.25) and (3.27), we obtain the first part of (3.19):

$$(1 - \rho)Q_{\infty,\rho} \Rightarrow L \overset{\mathrm{d}}{=} Exp((c_a^2 + 1)/2) \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.28}$$

Given that

$$W_{\infty,\rho} \overset{\mathrm{d}}{=} \sum_{i=1}^{Q_{\infty,\rho}+1} (V_i/s) \ , \tag{3.29}$$

we have

$$\frac{W_{\infty,\rho}}{Q_{\infty,\rho} + 1} \Rightarrow \frac{1}{s} \quad \text{as} \quad \rho \uparrow 1 \tag{3.30}$$

by the weak law of large numbers, since $Q_{\infty,\rho} \Rightarrow \infty$ as a consequence of (3.28). We then apply Theorem 11.4.5 of Whitt (2002) to write the joint limit

$$((1 - \rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho} + 1)) \Rightarrow (L, (1/s)) \ . \tag{3.31}$$

We then can apply the continuous mapping theorem with the function $h : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $h(x, y) = (x, xy)$ to get

$$h(((1 - \rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho} + 1)) \Rightarrow h(L, (1/s)) = (L, L/s) \ , \tag{3.32}$$

but

$$h(((1 - \rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho} + 1)) = \left((1 - \rho)Q_{\infty,\rho}, (1 - \rho)W_{\infty,\rho}\frac{Q_{\infty,\rho}}{Q_{\infty,\rho} + 1}\right) \ . \tag{3.33}$$

Since $Q_{\infty,\rho} \Rightarrow \infty$,

$$\frac{Q_{\infty,\rho}}{Q_{\infty,\rho} + 1} \Rightarrow 1 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.34}$$

Hence,

$$|h(((1 - \rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho} + 1)) - (1 - \rho)(Q_{\infty,\rho}, W_{\infty,\rho})| \Rightarrow 0 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{3.35}$$

Thus we can combine (3.32), (3.35) and the convergence-together theorem, Theorem 11.4.7 of Whitt (2002), to complete the proof of (3.19).

Proof of Theorem C.1. First we show that $W^h_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. As a consequence of the limit in (3.2), we must have $W_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. Suppose that we do *not* have $W^h_{\infty,s,\rho} \Rightarrow \infty$. Then there must exist a subsequence $\{\rho_k\}$ with $\rho_k \uparrow 1$ as $k \to \infty$, a constant K and a positive constant $\epsilon > 0$ such that $P(W^h_{\infty,s,\rho_k} > K) > \epsilon$ for all k. Since

$$W_{\infty,s,\rho} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{A(W^h_{\infty,s,\rho})+2} (V_i/s) \ , \tag{3.36}$$

conditional on $W_{\infty,s,\rho} > 0$, which holds with probability 1 in the limit, there must exist a new constant K' such that $P(W_{\infty,s,\rho_k} > K') > \epsilon/2$ for all k as well, but that contradicts the established limit $W_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. Hence we must have $W^h_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$, as claimed above.

Given that $\rho \uparrow 1$ and $W^h_{\infty,s,\rho} \Rightarrow \infty$, we get $A(W^h_{\infty,s,\rho})/W^h_{\infty,s,\rho} \Rightarrow s$ and

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} = \left(\frac{\sum_{i=1}^{A(W^h_{\infty,s,\rho})+2}(V_i/s)}{A(W^h_{\infty,s,\rho})+2}\right)\left(\frac{A(W^h_{\infty,s,\rho})+2}{W^h_{\infty,s,\rho}}\right) \Rightarrow (1/s) \times s = 1 \ , \tag{3.37}$$

by the law of large numbers for partial sums and renewal processes. Similarly, by (3.2), we also have $Q_{\infty,s,\rho} \Rightarrow \infty$, so that

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{Q_{\infty,s,\rho}+1} = \frac{\sum_{i=1}^{Q_{\infty,s,\rho}+1}(V_i/s)}{Q_{\infty,s,\rho}+1} \Rightarrow 1/s \ . \tag{3.38}$$

The limits (3.37) and (3.38) imply (3.4) and (3.5).

Since the limits in (3.37) and (3.38) are deterministic, we can apply Theorem 11.4.5 of Whitt (2002) to obtain joint convergence of all these with the limits in (3.2):

$$\left((1-\rho)Q_{\infty,s,\rho}, (1-\rho)W_{\infty,s,\rho}, (1-\rho)W^h_{\infty,s,\rho}, \frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{Q_{\infty,s,\rho}+1}, \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}}\right)$$
$$\Rightarrow \left(L, \frac{L}{s}, \frac{L}{s}, \frac{1}{s}, 1\right) \ . \tag{3.39}$$

We next apply the continuous mapping theorem, see Section 3.4 of Whitt (2002), with the function $h : \mathbb{R}^5 \to \mathbb{R}^5$ defined by $h(v, w, x, y, z) = (v, w, x, vy, xz)$ to get (3.6) from (3.39).

To continue, we next consider the random variable $c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}$. Starting from the limit in (3.6), we can apply the Skorohod representation theorem, Theorem 3.2.2 on p. 78 of Whitt (2002), to get random variables $\tilde{W}^h_{\infty,s,\rho}$ with the same probability law as $W^h_{\infty,s,\rho}$ but for which we have the convergence $(1-\rho)\tilde{W}^h_{\infty,s,\rho} \to \tilde{L}/s$ as $\rho \uparrow 1$ w.p.1, where $\tilde{L} \stackrel{\mathrm{d}}{=} L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$. Next note that $c^2_{W_{HOL,s,\rho}(w)}/c^2_{W_{HOL,s,1}(w)} \to 1$ w.p.1 as $\rho \uparrow 1$ and $w \to \infty$ in any order. Then, by (4.13),

$$\frac{c^2_{W_{HOL,s,\rho}(\tilde{W}^h_{\infty,s,\rho})}}{1-\rho} = \left(\frac{c^2_{W_{HOL,s,\rho}(\tilde{W}^h_{\infty,s,\rho})}}{c^2_{W_{HOL,s,1}(\tilde{W}^h_{\infty,s,\rho})}}\right)\left(\frac{\tilde{W}^h_{\infty,s,\rho}c^2_{W_{HOL,s,1}(\tilde{W}^h_{\infty,s,\rho})}}{(1-\rho)\tilde{W}^h_{\infty,s,\rho}}\right) \rightarrow \frac{(c_a^2+1)/s}{\tilde{L}/s} \tag{3.40}$$

as $\rho \uparrow 1$ w.p.1. Essentially the same reasoning applies to the random variable RMSE $(W^h_{\infty,s,\rho})$, giving the same limit. The equality in distribution then implies the associated convergence in distribution for the last two components of the original random vector in (3.7). We now treat the first component. Since $(Q_{\infty,s,\rho}+1)c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})} = 1$, a deterministic quantity, by (2.2), we can apply (4.13) to get

$$\begin{aligned} \frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} &= \left(\frac{Q_{\infty,s,\rho}+1}{W^h_{\infty,s,\rho}}\right)\left(\frac{W^h_{\infty,s,\rho}c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{(Q_{\infty,s,\rho}+1)c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}}\right) \\ &= \left(\frac{Q_{\infty,s,\rho}+1}{W^h_{\infty,s,\rho}}\right)W^h_{\infty,s,\rho}c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})} \Rightarrow s \times \frac{c_a^2+1}{s} = c_a^2+1 \end{aligned} \tag{3.41}$$

We then reason as before in establishing (3.39), first to express this limit jointly with the last two components of (3.7) and then to apply the continuous mapping theorem to complete the proof of (3.7) itself. Finally, (3.8) and (3.9) follow from the previous results. ■

Customers Who Have Completed Service. In this final subsection, supplementing the application of the snapshot principle in §6, we consider the estimators based on the delays experienced by previous customers to *complete* service. Unlike for the LES and HOL estimators, we find that the LCS estimator behaves very differently in the classical and QED HT regimes. The way to see this is to observe that the LCS customer completed service a full service time in the past. That LCS customer arrived a waiting time plus a service time in the past.

In both heavy-traffic regimes, the service time is an exponential random variable with mean 1. In the classical HT regime, the waiting times are exploding in heavy traffic, so that a service time is negligible compared to the waiting time. Thus we see that LCS will be asymptotically equivalent to LES and HOL in the classical HT regime, for any fixed number of servers. The LCS estimator will be consistent as well in the classical heavy-traffic regime.

However, the story is very different in the QED HT regime. The service times remain unchanged, but now the waiting times become smaller, being of order $O(1/\sqrt{s})$. Now the service time is the same order as the time scaling. The stochastic-process limit in (6.2) describes the waiting time experience of each customer, but for the last customer to complete service at time t, we have a different limit. Let $A^L_{s,\rho}(t)$ denote the arrival time of the last customer to complete service at time t in model (s, ρ). The relevant limit now will be

$$\sqrt{s} W_{s,\rho}(A^L_{s,\rho}(t)) \Rightarrow Y(t - V) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{3.42}$$

where $Y(t)$ is the limit process in (6.2) and V is a service time, an exponential random variable with mean 1. In other words, the waiting time at time t is approximately $Y(t)/\sqrt{s}$, while the waiting time of the last customer to complete service immediately prior to time t is approximately $Y(t - V)/\sqrt{s}$. Thus, in the QED HT limit the LCS estimator is *not* consistent. The effectiveness of the LCS estimator depends on the difference between $Y(t - V)$ and $Y(t)$. However, we do not attempt to do further analysis; here we are content to observe that the LCS estimator has inferior asymptotic performance in the QED HT regime. That is consistent with our simulation results, which show that the LCS estimator performs poorly for large s.

Fortunately, there is better information that we can obtain from customers who have already completed service in the QED HT regime. Other customers who have completed service are very likely to have arrived much more recently than the last customer to complete service. The minimum service time among the last m customers to complete service is $1/m$. Since the waiting times are of order $1/\sqrt{s}$, it is natural to consider $m = O(\sqrt{s})$; then the minimum service time among these customers also will be of order $O(1/\sqrt{s})$.

As a bound, first consider the customer among the last $c\sqrt{s}$ customers to complete service with the minimum service time. That customer's service time is exponentially distributed with mean $1/c\sqrt{s} = O(1/\sqrt{s})$. By (6.2), the customer's waiting time is also of order $O(1/\sqrt{s})$. Since the times between successive service completions are i.i.d. exponential random variables with mean $1/s$, the last $c\sqrt{s}$ service completions occur over a time interval having mean $c/\sqrt{s} = O(1/\sqrt{s})$. Hence this customer arrived $O(1/\sqrt{s})$ in the past. Hence we deduce that if we consider the customer among the last $c\sqrt{s}$ customers to complete service with the minimum service time, then that delay estimator is consistent in the QED HT regime.

Even better will be the RCS and RCS-$c\sqrt{s}$ estimators, because those customers necessarily arrive at least as recently. We summarize these conclusions in the following theorem. To state the theorem, let $W^{RCS}_{\infty,s,\rho}$ and $W^{RCS-c\sqrt{s}}_{\infty,s,\rho}$ be the steady-state RCS and RCS-$c\sqrt{s}$ delays in model (s, ρ); and let $W_{RCS,s,\rho}(w)$ and $W_{RCS-c\sqrt{s},s,\rho}(w)$ be the associated random variables having the conditional distribution of the delay to be estimated given the observed RCS and RCS-$c\sqrt{s}$ delays.

Theorem C.3. (*performance of LCS, RCS and RCS-$c\sqrt{s}$ in the QED HT regime*) *If $\rho \uparrow 1$ and $s \uparrow \infty$ so that $(1 - \rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$ in the family of $GI/M/s$ models indexed by*

s and ρ, then the RCS and RCS-$c\sqrt{s}$ estimators are relatively consistent, i.e.,

$$\frac{W_{RCS,s,\rho}(W_{\infty,s,\rho}^{RCS})}{W_{\infty,s,\rho}^{RCS}} \Rightarrow 1 \quad \mathit{and} \quad \frac{W_{RCS-c\sqrt{s},s,\rho}(W_{\infty,s,\rho}^{RCS-c\sqrt{s}})}{W_{\infty,s,\rho}^{RCS-c\sqrt{s}}} \Rightarrow 1 \ , \tag{3.43}$$

but the LCS estimator is not relatively consistent.

In this relatively crude sense, the estimators LES, HOL, RCS and RCS-$c\sqrt{s}$ are all asymptotically equivalent in the QED regime, but LCS is not. However, it remains to describe the asymptotic efficiency of RCS and RCS-$c\sqrt{s}$, paralleling the results for the HOL (and LES) estimator SCV's in (3.16) and (3.17).

D. A Pathological Example for LES

We have drawn very positive conclusions about the LES delay estimator $W_{LES}(w)$ in the $GI/M/s$ queue. To provide some balancing perspective, in this section we demonstrate potential weaknesses of the estimator $W_{LES}(w)$ for other service-time distributions. To illustrate the possible deficiencies of the LES estimator, we consider a specific stable $D/G/1$ queueing model with non-exponential service-time distribution in light traffic. Let the arrival process be deterministic with interarrival times 1.

We deliberately choose a difficult service-time distribution: let the service-time distribution be a two-point probability distribution, which usually assumes a very small value ϵ, but occasionally takes a very large value M; specifically, let

$$P(V = M >> 1) = \delta = 1 - P(V = \epsilon << 1) \ , \tag{4.1}$$

where the traffic intensity

$$\rho \equiv E[V]/E[U] = E[V] = \delta M + (1-\delta)\epsilon \ . \tag{4.2}$$

We suppose that δ is very small, so that ρ itself is very small and the service time is only equal to the large value M very rarely. If δ is sufficiently small, relatively few customers will have to wait in queue before starting service, but occasionally a customer will have one of the very long service times.

To see the deficiencies of the LES estimator, we will consider an epoch at which a customer with service time M arrives at an empty system. If δ is small enough, then with high probability the customer with the large service time M will not have to wait before starting service, but he will remain in service for a long time, precisely M. Thus the following M customers will all have to wait before starting service. For each of them, however, the last served customer to have entered service – the customer with service time M – will have not had to wait at all.

To quantify the effect, let us call the customer with service time M customer 0. Then, assuming that these following M customers themselves all have ϵ service times (which has high probability), customer k will have to wait precisely $M - k + (k-1)\epsilon$ before starting service. Customer number M will have to wait only $(M-1)\epsilon$. But, for all M customers with positive waiting times, the last served customer will have waited 0 before starting service.

To go further, suppose that ϵ is very small, so that $(M-1)\epsilon$ is itself less than 1. Then customer M will have to wait less than 1 before starting service, so that $M+1$ will not have to wait at all before starting service. We thus have the strange estimation phenomenon: *The delay of the last served customer is* 0 *for all customers that themselves experience positive delays.* Thus, whenever an estimation needs to be made (because the customer must wait in queue), the estimated delay will be 0. Moreover, the actual delays of these customers who have to wait may be quite large: as large as $M-1$ and averaging about $M/2$ for all customers forced to wait. This example allows arbitrarily large M, but after choosing M, we must choose ϵ and δ suitably small.

We have only described one possible scenario. The story we have described breaks down when two or more customers with large service time M interact, but by choosing δ sufficiently small, this deviation from the story can be made to occur relatively rarely. Thus the phenomenon we have described will hold for the vast majority of the customers that are delayed.

We can make the situation described above apply w.p.1 if we abandon the condition of i.i.d. service times. If we instead assume that customers $2kM$ have service times M, while all other customers have service time ϵ with $\epsilon < 1/M$ (e.g., $\epsilon = 0$), then we obtain the scenario above w.p.1. In addition, the average delay is approximately $M/4$, so the average delay can be made arbitrarily large by choosing M large. Thus this scenario does not only apply in very light traffic. Nevertheless, we regard this example as pathological. We are thinking of situations in which the delay of a new arrival should not be too different from the delay of the last customer to enter service.

For this example, the HOL estimator would fare somewhat better, but it would not do so great either. Given the scenario described above, when the customer at the head of the line has waited $w = k$, the random variable $W_{HOL}(w)$ depicting the delay of this new arrival is very likely to take the value $M - k + (k-1)\epsilon$ instead of w.

References

Abate, J. and W. Whitt. 1994. A heavy-traffic expansion for asymptotic decay rates of tail probabilities in multichannel queues. *Operations Res. Letters* 15, 223–230.

Choudhury, G. L. and W. Whitt. 1994. Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 Queue. *Stochastic Models* 10, 453–498.

Cooper, R. B. 1981. *Introduction to Queueing Theory*, second edition, North-Holland, New York.

Glynn, P. W. and W. Whitt. 1994. Logarithmic Asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31A, 131–156 (also called *Studies in Applied Probability, Papers in Honour of Lajos Takcs*, J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England).

Halfin, S. and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29, 567–588.

Iglehart, D. L. and W. Whitt. 1970. Multiple channel queues in heavy traffic II: sequences, networks, and batches. *Advances in Applied Probability* 2, 355–369.

Jelenkovic P., A. Mandelbaum A. and P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* 47, 53–69.

Neuts, M. F. 1986. The caudal characteristic curve of queues. *Adv. Appl. Probab.* 18, 221–254.

Smith, W. L. 1953. On the distribution of queueing times. *Proc. Camb. Phil. Soc.* 49, 449–461.

Whitt, W. 1984. On approximations for queues, I: extremal distributions. *AT&T Bell Lab. Tech. J.* 63, 115–138.

Whitt, W. 2004b. A diffusion approximation for the G/GI/n/m queue. *Operations Research* 52, 922–941.

Whitt, W. 2005. Heavy-traffic limits for the G/H2*/n/m queue. *Math. Oper. Res.* 30, 1–27.