Bounds and comparisons of the loss ratio in queues driven by an M/M/∞ source.

N.G. Duffield* and D.J. Daley †

June 27, 1994

Abstract

We obtain upper bounds for the loss probability in a queue driven by an M/M/∞ source. The bound is compared with exact numerical results, and with bounds for two related arrivals models: superposed two state Markov fluids, and the Ornstein-Uhlenbeck process. The bounds are shown to behave continuously through approximation procedures relating the models.

Keywords: Martingales, fluid flow models, heavy traffic limits

AMS 1991 Classifications: Primary 60K25; Secondary 68M20, 90B22, 68M20

DIAS-APG-94-25

*School of Mathematical Sciences, Dublin City University, Dublin 9, Ireland; Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. E-mail: duffieldn@dcu.ie

†Stochastic Analysis Group, Australian National University, Canberra ACT 2600, Australia E-mail: daryl@orac.anu.edu.au

1 Introduction.

There has been much interest recently in obtaining bounds for queue length distributions in ATM multiplexers driven by Markovian traffic. Two fundamental arrival processes have been investigated: superposed two-state Markov arrivals [5]; and the Ornstein–Uhlenbeck process [11]. In both cases, exponential bounds of the form

$$P[\text{queue} > b] \leq \Phi e^{-\delta b} \tag{1}$$

are obtained. The latter treatment was motivated in part as a heavy traffic approximation to the queue driven by an M/M/∞ process, which in turn approximates superpositions of Markov fluid sources. Queues with Markov fluid arrivals were studied in [1]. The queue with M/M/∞ arrivals has been investigated numerically in [6]. We mention also other recent work on bounds for queues with Markovian arrivals [2, 7].

The purpose of this note is to give explicit upper bounds of the form (1) directly for the M/M/∞ arrival process. Moreover, the bounds can be compared with existing results in a number of directions. Firstly, by bounding the mean queue length we can compare with the exact results of [6]. Secondly, we can compare the bounds with those for queues with finitely superposed Markov fluid arrivals, and queues with Ornstein–Uhlenbeck arrivals. The M/M/∞ process can be regarded as intermediate between these: the arrival process of L Markov fluid sources each with activity proportional to L^{-1} converges as $L \to \infty$ to the M/M/∞ process, which in turn converges in a rescaled heavy traffic limit to the Ornstein–Uhlenbeck process. We show that the bounds we obtain for M/M/∞ converge in the same manner. This is useful for the following reason. The prefactor Φ of (1) found in [5] for L-fold superpositions of bursty sources is of the form $\Phi = \phi^L$ for some $\phi < 1$ depending only on the load of the system and the parameters of a single source. This demonstrates the *economy of scale* which is to be obtained by statistical multiplexing. Our comparison of the bounds shows how this economy behaves through the approximation procedures described.

2 The M/M/∞ process: bounds

Let $X = (X_t)_{t \in \mathbf{R}_+}$ denote the stationary M/M/∞ process. X is the birth-death process with Poissonian births at some rate λ, each of which dies after an i.i.d. time which is exponentially distributed with mean μ^{-1}. More precisely, X has sample paths in $D_{\mathbf{Z}_+}[0, \infty)$, the space of non-negative integer valued paths which are right continuous and have left limits. The stationary distribution is Poissonian with mean λ/μ. The corresponding Markov semigroup on ℓ_∞ (the space of real sequences topologized with the supremum norm) has generator G

corresponding to the rate matrix $(G_{xy})_{x,y\in\mathbf{Z}_+}$ where

$$G_{xy} = \begin{cases} x\mu & (y = x-1), \\ -(\lambda + x\mu) & (x = y), \\ \lambda & (y = x+1), \\ 0 & (\text{otherwise}). \end{cases} \tag{2}$$

The existence of such a closed G generating a Markov semigroup follows from standard conditions (see e.g. section IV.4 of [8]).

The queueing process is as follows. X_t is the number of sources active at time $-t$. Each active source empties fluid into the buffer of a queue at rate a. Fluid is drained at rate s from the buffer. Defining the workload

$$W_t = \int_0^t (aX_{t'} - s)\, dt' \qquad (t \geq 0), \tag{3}$$

the fluid remaining in the buffer at time 0 is (see e.g. §§6–9 of [4])

$$Q = \sup_{t\geq 0} W_t. \tag{4}$$

The offered load is $\rho = E[aX_0]/s = a\lambda/(s\mu)$.

Let $\mathbb{I}$ denote the identity on ℓ_∞ and $\mathcal{N}$ the number operator, densely defined in ℓ_∞ by $(\mathcal{N}v)(x) = xv(x)$. For $\theta \in \mathbf{R}$ let $\omega(\theta)$ and $v(\cdot;\theta)$ be the maximal eigenvalue and corresponding eigenvector of the densely defined operator $G(\theta) = G + \theta(a\mathcal{N} - s\mathbb{I})$. Define $\delta = \sup\{\theta \mid \omega(\theta) = 0\}$, and abbreviate $v(\cdot;\delta)$ by $v(\cdot)$. Finally, let $\mathcal{F}$ be the canonical filtration generated by X.

Lemma 1 *$M_t := e^{\delta W_t} v(X_t)$ is an $\mathcal{F}$-martingale.*

Proof: Consider the joint Markov process $(W_t, X_t)_{t\geq 0}$ taking values in $\mathbf{R}\times\mathbf{Z}_+$. Its generator $\hat{G}$ is densely defined on $\mathcal{C}(\mathbf{R})\otimes\ell_\infty$ by

$$(\hat{G}g)(w,x) \frac{d}{dw} \otimes \mathbb{I} (\otimes\mathcal{N} G) g s (\mathbb{I}) g x (w + x (. \tag{5}$$

Letting $f(w,x) = e^{\delta w}v(x)$ then $\hat{G}f = (\mathbb{I}\otimes G(\delta))f = 0$, since $G(\delta)v = 0$. Consequently, by Dynkin's Theorem (see e.g. [10]), $f(W_t, X_t)$ is an $\mathcal{F}$-martingale. □

Let us find δ and v. The eigenvector equation $G(\delta)v = 0$ becomes

$$x\left(\mu v(x-1) + (\delta a - \mu)v(x)\right) + \left(\lambda v(x+1) - (\lambda + \delta s)v(x)\right) = 0 \tag{6}$$

for $x \in \mathbf{N}$. This is solved by setting

$$r = \frac{v(x+1)}{v(x)} = \frac{\lambda + \delta s}{\lambda} = \frac{\mu}{\mu - \delta a} \tag{7}$$

for all $n \in \mathbb{N}$. The second equality in (7) yields (for $\rho \leq 1$)

$$\delta = \frac{s\mu - a\lambda}{sa} = (1-\rho)\mu/a \quad \text{and hence} \quad r = \rho^{-1}. \tag{8}$$

Theorem 1 *When $\rho < 1$ and $b > 0$,*

$$P[Q \geq b] \leq \rho^{s/a} e^{(1-\rho)(s/a - \mu b/a)}. \tag{9}$$

Proof: For any $b > 0$ define the $\mathcal{F}$-stopping time $\tau = \inf\{t \geq 0 \mid W_t > b\}$. Note $\{Q > b\} = \{\tau < \infty\}$. By a mostly familiar argument involving Doob's Optional Sampling Theorem (see e.g. [13]).

$$E[M_0] = E[M_{\tau \wedge k}] \geq E[M_\tau; \tau < k] \tag{10}$$

and so by bounded convergence as $k \to \infty$,

$$E[M_0] \geq E[M_\tau \mid \tau < \infty] P[\tau < \infty]. \tag{11}$$

But if $\tau < \infty$, then $M_\tau \geq e^{\delta b} \inf_{x: x \geq s/a} v(x)$, because $e^{\delta W_\tau} = e^{\delta b}$ and $aX_\tau \geq s$. (Since X is right-continuous, W is continuous and hence $W_\tau = b$. Similarly, if $aX_\tau < s$, then $W_t < b$ on some interval $(\tau, \tau + \varepsilon)$, a contradiction). Hence

$$P[Q > b] \leq \frac{e^{-\delta b} E[M_0]}{\inf_{x: x \geq s/a} v(x)} = \frac{e^{-\delta b} E[\rho^{-X_0}]}{\inf_{x: x \geq s/a} \rho^{-x}} = \rho^{s/a} e^{(1-\rho)(s/a - \mu b/a)}. \tag{12}$$

□

3 Comparisons with finite superpositions.

The M/M/∞ process is itself a limit of a superposition of (rescaled) two state Markov fluid arrivals. This latter model was introduced in into queueing theory by Anick, Mitra and Sondhi [1]. Specifically, consider the continuous time Markov chain with state space $\{0, 1\}$ with transitions $0 \to 1$ occurring at rate λ_L and the reverse transition occurring at rate μ_L. The Markov chain represents a fluid source: in state 1 fluid arrives at rate a, in state 0 no fluid arrives. In an L-fold superposition, let X_t^L denote the number of sources active at time $-t$. X_t is the stationary Markov chain with generator corresponding to the rate matrix G^L where for $0 \leq x, y \leq L$

$$G_{xy}^L = \begin{cases} x\mu_L & (y = x - 1), \\ -((L-x)\lambda_L + x\mu_L) & (x = y), \\ (L-x)\lambda_L & (y = x+1), \\ 0 & (\text{otherwise}). \end{cases} \tag{13}$$

The stationary distribution of X^L is binomial with mean $L\lambda_L/(\Lambda_L+\mu_L)$. The superposition is served at constant rate s, and so the offered load is $\rho_L = aE[X_0^L]/s = aL\lambda_L/(s(\lambda_L+\mu_L))$. Let Q^L denote the corresponding queue length at time 0. By repeating the steps of Theorem 1 making the appropriate changes one proves:

Theorem 2 *When $\rho_L < 1$ and $b > 0$,*

$$P[Q^L \geq b] \leq \frac{e^{-\delta_L b} E[r_L^{X_0}]}{\inf_{x:x\geq s/a} r_L^x} = e^{-\delta_L b} r_L^{-s/a} \left(\frac{\lambda_L r_L + \mu_L}{\lambda_L + \mu_L}\right)^L , \tag{14}$$

where

$$\delta_L = \frac{(1-\rho_L)(\lambda_L+\mu_L)}{(a - s/L)} \quad \text{and} \quad r_L = \rho_L^{-1}\frac{aL - s\rho_L}{aL - s} > 1. \tag{15}$$

By standard methods it is proved that under the scaling limit

$$L \to \infty, \qquad L\lambda_L \to \lambda, \qquad \mu_L \to \mu \tag{16}$$

X^L converges to X in distribution. (See e.g. [10]). Furthermore, under (16)

$$\delta_L \to \delta, \qquad \rho_L \to \rho, \qquad r_L \to r \tag{17}$$

so that the bound in Theorem 2 converges to that of Theorem 1.

Daley and Ott [6] have performed exact numerical calculations of the mean buffer occupations in both cases. In Table 1 we present a comparison of the bounds for heavy traffic where we choose $\lambda_L = \lambda/L$, $\mu_L = \mu$ take $\mu/a = 1$ and vary $\kappa = s/a$. Here we use the fact that since $Q \geq 0$, $P[Q > b] \leq \phi e^{-\delta b}$ for $b > 0$ implies that $E[Q] \leq \phi/\delta$. In Table 2 we investigate the accuracy of the prefactor to the exponential bound, using the fact that $P[Q > b] \leq \phi e^{-\delta b}$ implies $P[Q = 0] \geq 1 - \phi$. In both tables, $L = \infty$ corresponds to M/M/∞ arrivals.

Table 1
Stationary mean buffer content: bounds and exact evaluations

		$\kappa = 6$			$\kappa = 12$			$\kappa = 24$		
L	ρ_L	.96	.98	.99	.96	.98	.99	.96	.98	.99
50	Exact	16.34	35.62	74.30	11.33	25.65	54.47	4.60	11.23	24.68
	Bound	19.36	38.77	77.52	14.44	28.97	57.90	6.75	13.64	27.23
100	Exact	18.66	40.65	84.79	15.28	34.48	73.12	10.20	24.39	53.14
	Bound	22.03	44.18	88.39	19.25	38.72	77.49	14.25	28.88	57.85
250	Exact	20.12	43.83	91.41	17.94	40.42	85.64	14.67	34.77	75.47
	Bound	23.72	47.59	95.25	22.47	45.25	90.62	20.08	40.73	81.70
∞	Exact	21.12	46.02	95.97	19.84	44.64	94.54	18.11	42.73	92.54
	Bound	24.88	49.94	99.97	24.75	49.88	99.94	24.51	49.76	99.88

Table 2
$P[Q^L = 0]$: *bounds and exact evaluations*

		$\kappa = 6$			$\kappa = 12$			$\kappa = 24$		
	ρ_L	.96	.98	.99	.96	.98	.99	.96	.98	.99
L										
50	Exact	.8876	.9432	.9714	.8415	.9194	.9593	.7498	.8706	.9343
	Bound	.9944	.9986	.9997	.9872	.9968	.9992	.9637	.9908	.9977
100	Exact	.8905	.9447	.9722	.8508	.9243	.9618	.7865	.8905	.9446
	Bound	.9948	.9987	.9997	.9889	.9972	.9993	.9746	.9936	.9984
∞	Exact	.8931	.9460	.9729	.8584	.9282	.9639	.8097	.9029	.9510
	Bound	.9951	.9988	.9997	.9902	.9976	.9994	.9805	.9951	.9988

4 Comparison with Ornstein–Uhlenbeck arrivals.

Finally, we investigate the relation of the bound of Theorem 1 to those found for the Ornstein–Uhlenbeck arrival process in [11]. The Ornstein–Uhlenbeck arrival process can be seen as the last step in a chain of approximations: finite superpositions are approximated by M/M/∞ processes; the latter in turn by the Ornstein–Uhlenbeck process. To be precise we take the limit

$$\kappa \to \infty, \quad \text{with} \quad a = s\kappa^{-1}, \quad \mu = \beta\kappa^{-1/2} \quad \text{and} \quad \lambda = \beta(\kappa^{-1/2} - \nu) \tag{18}$$

for positive constants s, β and ν. This has the consequence that the load is $\rho = 1 - \nu\kappa^{-1/2}$: we are dealing with a heavy traffic limit. For a particular value of κ we denote the corresponding activity process X of section 2 by X^κ. Set $Z_t^\kappa = \kappa^{1/2} a (X^\kappa_{\kappa^{1/2} t} - \lambda/\mu)$. Then under the limits (18) Z^κ converges to Z, the stationary diffusion which is the solution of the stochastic differential equation

$$dZ_t = -\beta Z_t + s\sqrt{2\beta}\, dB_t. \tag{19}$$

where B is standard Brownian motion. (See [9] and [11]). In other words, Z is a stationary Ornstein–Uhlenbeck velocity process with mean 0 and variance s^2. The queue length for the process X^κ is

$$Q^\kappa = \sup_{t \geq 0} \int_0^t dt'(aX^\kappa_{t'} - s) = \sup_{t \geq 0} \int_0^t dt'(Z^\kappa_{t'} - \beta s \nu). \tag{20}$$

In [11] the bound for the queue $\tilde{Q} = \sup_{t\geq 0} \int_0^t (Z_{t'} - \beta s \nu) dt'$ driven by Z and served at rate $\beta s \nu$ is found though martingale methods to be

$$P[\tilde{Q} > b] \leq e^{-\nu^2/2} e^{-\beta \nu b / s}. \tag{21}$$

We remark now that the bound of Theorem 1 converges to the RHS of (21) under the limit (18). To see this note that under (18), $\delta = \nu\beta/s$ independent of κ, while the prefactor of Theorem 1 has the limit

$$\lim_{\kappa\to\infty}(\rho e^{1-\rho})^{s/a} = \lim_{\kappa\to\infty}((1-\nu\kappa^{-1/2})e^{-\nu\kappa^{-1/2}})^{\kappa} = e^{-\nu^2/2}. \qquad (22)$$

5 Conclusions.

The bounds of Theorem 2 for finite superpositions contain a prefactor which is exponential in L, the number of sources in the superposition. These determine the economies of scale which are to be found by multiplexing larger number of sources together at constant load, in the sense that they demonstrate how the usual effective bandwidth approximation $P[Q > b] \approx e^{-\delta b}$ overestimate the probability of loss. (See for example, [12] for discussion of the effective bandwidth approximation, and [3] for a further discussion on economies of scale). The bounds obtain by successive approximations in Theorem 1 and equation (21) and their convergence in these approximations shows how the remnant of this economy survives in the heavy traffic limit.

Acknowledgements. N.G.D. acknowledges a Visiting Fellowship from the Australian National University, where this work was started, and the hospitality of the Stochastic Analysis Group.

References

[1] D. Anick, D. Mitra and M.M. Sondhi (1982). Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61:1872–1894

[2] S. Asmussen and T. Rolski (1993). Risk theory in a periodic environment: the Cramer–Lundberg approximation and Lundberg's inequality. Preprint

[3] D.D. Botvich and N.G. Duffield (1994). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Preprint.

[4] A.A. Borovkov (1976). *Stochastic processes in queueing theory.* Springer, New York.

[5] E. Buffet and N.G. Duffield (1992). Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *J. Appl. Prob.*, to appear.

[6] D.J. Daley and T. Ott (1993). On a fluid model for a packet switch. Preprint.

[7] N.G. Duffield (1993). Exponential bounds for queues with Markovian arrivals. *Queueing Systems* , to appear.

[8] S.N. Ethier and T.G. Kurtz (1986) *Markov processes: characterization and convergence.* Wiley, New York.

[9] Iglehart, D.L. (1965) Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* 2:429–441.

[10] S. Karlin and H.M. Taylor (1981). *A second course in stochastic processes* Academic Press, New York.

[11] V. Kulkarni and T. Rolski (1993). Fluid model driven by an Ornstein–Uhlenbeck process. Preprint.

[12] W. Whitt (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunications Systems.* 2:71–107

[13] D. Williams (1991). *Probability with martingales.* CUP, Cambridge.