A Note on Approximating Peak Congestion in $M_t / G / \infty$ Queues with Sinusoidal Arrivals

Linda V. Green • Peter J. Kolesar
Columbia Business School, Columbia University, New York, New York 10027-6902

We study the $M_t / G / \infty$ queue where customers arrive according to a sinusoidal function $\lambda_t = \lambda + A \sin(2\pi t / T)$ and the service rate is μ. We show that the expected number of customers in the system during peak congestion can be closely approximated by $(\lambda + A) / \mu$ for service distributions with coefficient of variation between 0 and 1. Motivated by a result derived by Eick, Massey, and Whitt that the time lag of the peak congestion from the peak of the customer arrivals is $1 / 2\mu$ for models with deterministic service times, we show that the time lag for exponential service times is closely approximated by $1 / \mu$. Based on a cycle length of 24 hours and regardless of the values of other system parameters, these approximations are of the order of 1% accuracy for $\mu = 1$, and the accuracy increases rapidly with increasing μ.
(*Queues; Nonstationarity; Approximations*)

1. Introduction

Customer arrivals at many service facilities follow periodic patterns. Examples include: aircraft landings (Koopman 1972), post-office mail receipts (Oliver and Samuel 1962), emergency calls for police assistance (Larson 1972; Green and Kolesar 1984, 1989), fire alarms (Walker et al. 1979), telephone traffic (Segal 1974), transactions at automatic teller machines (Kolesar 1984), telemarketing (Andrews and Parsons 1989), and automobiles at bridge toll plazas (Edie 1954). Unfortunately, most queuing models that include time-varying arrivals are too complex to solve analytically. This has motivated our interest in simple approximate solutions for queues with cyclic arrivals (Green and Kolesar 1991, 1995; Green et al. 1991).

Eick, Massey, and Whitt (1993a) (EMW), building on earlier work of Palm and Khintchine, have derived closed-form expressions for the time-varying expected number of customers in an $M_t / G / \infty$ queue with sinusoidal arrivals. Using these exact results, we show that the magnitude and the time lag of peak congestion in an $M_t / G / \infty$ queue may, in many cases, be approximated very accurately by simple "back of the envelope" expressions. This is important for several reasons. First, the approximations provide obvious computational convenience for practitioners. This is particularly useful for cases other than exponential or deterministic service time distributions where EMW's exact results may be quite difficult for practitioners to evaluate. (See section 2 for the case of Erlang 2 service times.) Moreover, as our numerical results indicate, the approximations we propose are so accurate that in many practical situations there would be no advantage in having the exact values. Second, these approximations may also be important for use in other models where a simple structure is needed, as, for example, in resource allocation or optimization (Harel 1988). Indeed, a paper by a team that includes Whitt and Massey (Jennings et al. 1996) use their own more complex approximations for the exact $M_t / G / \infty$ results in this way. Third, and perhaps most importantly, our own experience has shown that such simple expressions are very useful for providing insight to practitioners in evaluating alternative operating designs (Kolesar and Green 1998).

From the practical point of view, the $M_t / G / \infty$ queue is, itself, almost always an approximation—real systems never have an infinite number of servers. An infinite server model may, however, give a reasonable

0025-1909/98/4411/S137$05.00

description of congestion in actual operations if the number of servers available is large relative to the demand. Consequently, using an approximate solution for the $M_t/G/\infty$ queue can be quite natural in an applied context. (For further justification and discussion of the use of infinite server models in finite server environments see Whitt (1991) and Jennings et al. (1994).)

2. The Model and Main Result

We study a queuing system in which customers arrive according to a sinusoidal Poisson process with arrival rate given by

$$\lambda_t = \lambda + A \sin(2\pi t/T). \tag{1}$$

Here, λ is the overall mean arrival rate, A is the amplitude of the arrival function and T is its period. (Motivated by the many practical cases in which a daily cycle is evident, we take $T = 24$ hours as standard.) We denote the relative amplitude by $RA = A/\lambda$. An infinite number of servers are available who commence serving each customer immediately upon its arrival. Service times are independent identically distributed random variables with a general service time distribution with cumulative distribution $G(t)$ and mean $1/\mu$. We restrict our attention to service time distributions with coefficients of variation between 0 and 1. This range—from constant to exponentially distributed service times—covers a very broad class of applications and is also convenient to explore computationally.

Let L_t denote the periodic stationary expected number of customers in the system at time t. Motivated by the many applications in which system capacity is fixed over the period and therefore must be sufficient to provide good performance when demand is highest, we focus on the "peak-epoch" and "peak-hour" values of L_t. We denote the peak-epoch, or instantaneous maximum value of L_t, by L_{peak} which is

$$L_{\text{peak}} = \max_{0 \le t \le T} L_t. \tag{2}$$

The average value of L_t during the peak hour of the day, which we denote by $L_{\text{peak-hour}}$, is

$$L_{\text{peak-hour}} = \max_{0 \le t \le T} \int_t^{t+1} L_v \, dv. \tag{3}$$

Transforming the results of EMW which are stated in terms of systems with $\mu = 1$ to our frame of reference which is systems with $T = 24$, we get for the case of exponential service times (EMW Equation 18) that L_{peak} is

$$L_{\text{peak}} = \frac{\lambda}{\mu} + \frac{A}{\mu}\left(\frac{1}{\sqrt{1 + (\gamma/\mu)^2}}\right), \tag{4}$$

where the arrival frequency $\gamma = 2\pi/24$. With sinusoidal arrivals on a 24 hour period, customer arrivals peak at $t = 6$ hours. However, the peak congestion occurs at a time lag of t_{lag} hours after the peak of the customer arrival rate where (EMW eq. 16)

$$t_{\text{lag}} = \frac{\cot^{-1}(\mu/\gamma)}{\gamma}. \tag{5}$$

For the model with deterministic service times, we get (EMW eq. 27)

$$L_{\text{peak}} = \frac{\lambda}{\mu} + \frac{A}{\mu}\left[\frac{2\sin(\gamma/2\mu)\mu}{\gamma}\right], \tag{6}$$

while the peak congestion occurs at a lag of (EMW eq. 26)

$$t_{\text{lag}} = 1/2\mu. \tag{7}$$

We also analyze an Erlang 2 service time model which has a coefficient of variation of $1/\sqrt{2}$ as an "intermediary" case between exponential service times with coefficient of variation of 1 and deterministic service times with coefficient of variation of 0. The Erlang 2 model with mean $1/\mu$ has density function

$$f(x) = (2\mu)^2 x e^{-2\mu x}, \qquad x \ge 0. \tag{8}$$

Using eq. 12 in EMW we derive

$$L_{\text{peak}} = \frac{\lambda}{\mu} + \frac{A}{\mu}\sqrt{D}, \tag{9}$$

where D is

$$D = \left(\frac{4\mu^2}{2(4\mu^2 + \gamma^2)} + \frac{4\mu^2(4\mu^2 - \gamma^2)}{2(4\mu^2 + \gamma^2)^2}\right)^2 + \left(\frac{2\mu\gamma}{2(4\mu^2 + \gamma^2)} + \frac{8\mu^3\gamma}{(4\mu^2 + \gamma^2)^2}\right)^2. \tag{10}$$

We define the Simple Peak Epoch Approximation (SPEA) for the expected number of customers in the system at the epoch of peak congestion as

$$L_{\text{SPEA}} = \frac{\lambda + A}{\mu}. \tag{11}$$

We will use equation (11) to approximate equations (4), (6), and (9).

Noting that the arrival rate of a sinusoid on a 24 hour period peaks at $t = 6$, we define

$$\lambda_{\text{peak-hour}} \equiv \int_{5.5}^{6.5} \lambda_t \, dt, \tag{12}$$

and, using $\lambda_{\text{peak-hour}}$, the Simple Peak Hour Approximation (SPHA) for expected number of customers in the system during the hour of peak congestion is defined as

$$L_{\text{SPHA}} = \frac{\lambda_{\text{peak-hour}}}{\mu}. \tag{13}$$

Note that L_{SPEA} is the long run mean number in the system for a stationary $M/G/\infty$ queue operating steadily with the peak epoch arrival rate. Similarly, L_{SPHA} is the mean number in the system for a stationary $M/G/\infty$ queue operating steadily with an arrival rate that is the average over the peak hour. In Green and Kolesar (1995), we showed that such approximations were accurate in finite server systems for large service rate values, i.e., where $\mu > 20$. We show here that for infinite server systems, the approximations are good for much smaller service rates as well, i.e., when $\mu \geq 1$.

Our approximations are based on a concept that we have named the "Pointwise Stationary Approximation," or PSA for short (Green and Kolesar (1991). The PSA idea is to approximate the behavior of the actual cyclic queue at each point in time by a stationary version of the system model that uses the parameter values that apply at that instant. Therefore, the PSA for the expected number in the system at the peak epoch in the $M_t/G/\infty$ queue is just L_{SPEA}. This PSA approach assumes that the process is, in effect, dynamically in equilibrium at its time-dependent parameter values. Intuitively, the PSA will be a better approximation of systems where customers clear quickly and therefore act relatively independently at each time epoch, i.e., when the service rate is high. Indeed, Whitt (1991) proved that for finite server systems the PSA is asymptotically correct as the arrival and service rates increase. Similarly, infinite server models should be better approximated by the PSA approach than finite server models since each customer's residence in the system is independent of that of other customers and therefore of the past.

3. Approximating Peak-Epoch Congestion

Table 1 and Table 2 show relative errors for L_{SPEA} for the exponential and deterministic models, respectively. Relative error is defined by

$$\text{RelErr} = \frac{L_{\text{SPEA}} - L_{\text{peak}}}{L_{\text{peak}}} \times 100. \tag{14}$$

Consistent with our earlier findings for approximating peak-epoch behavior of finite server systems, two factors determine the accuracy of the approximation. The first and more important is the service rate. The tables show that relative errors decrease dramatically as the service rate increases. This is consistent with EMW's result that L_t converges uniformly in t to the PSA as the service rate increases. The second factor is the relative amplitude. Not surprisingly, relative errors decrease with decreasing relative amplitudes, although somewhat more slowly than with the service rate. These conclusions are true for both service time models. Note that, except for an anomaly that occurs for deterministic service times with $\mu = 1/24$ when all customers stay in the system for exactly 24 hours, the deterministic service time model is better approximated than the exponential service time model.

Note that the parameter ranges in the tables are broad: Values of μ range from a low of $1/24$, where a service takes 24 hours to complete on average, to a high of 512, where a service takes about 6 seconds on average. Relative amplitudes range from 0.1 to 1.0 in steps of 0.1, and there is also a relative amplitude of 0.01. For the exponential model with $\mu = 1.0$, which is an average service time of an hour, the relative errors are of the order of a percent for the higher relative amplitudes (Table 1). For the deterministic model with $\mu = 1.0$ the relative errors are of the order of a tenth of a percent for the higher relative amplitudes (Table 2).

Our results for the Erlang 2 case (not tabled), as expected, fall between the exponential and deterministic values. For example, when $\mu = 1.0$ and relative amplitude $= 0.5$ the Erlang 2 has a relative error of

Table 1 SPEA for the Exponential Model: Relative Errors (Percent)

	Relative Amplitude										
Mu	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
0.042	72.834	66.454	59.895	53.150	46.211	39.071	31.719	24.146	16.343	8.298	0.842
0.083	53.455	49.256	44.852	40.227	35.365	30.247	24.852	19.157	13.137	6.762	0.695
0.167	30.121	28.090	25.907	23.554	21.009	18.249	15.244	11.962	8.361	4.394	0.460
0.333	11.955	11.255	10.487	9.642	8.706	7.664	6.498	5.184	3.691	1.980	0.212
0.500	6.050	5.713	5.342	4.929	4.470	3.953	3.370	2.704	1.938	1.048	0.113
0.667	3.584	3.389	3.173	2.933	2.664	2.361	2.017	1.623	1.167	0.633	0.069
1	1.657	1.569	1.470	1.361	1.238	1.099	0.940	0.758	0.546	0.297	0.032
1.333	0.946	0.895	0.840	0.778	0.708	0.628	0.538	0.434	0.313	0.171	0.019
2	0.425	0.402	0.377	0.350	0.318	0.283	0.242	0.196	0.141	0.077	0.008
4	0.107	0.101	0.095	0.088	0.080	0.071	0.061	0.049	0.036	0.019	0.002
8	0.027	0.025	0.024	0.022	0.020	0.018	0.015	0.012	0.009	0.005	0.001
16	0.007	0.006	0.006	0.006	0.005	0.004	0.004	0.003	0.002	0.001	0.000
32	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000
64	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
128	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
256	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
512	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

0.494% while the relative errors for the exponential and deterministic cases are 1.099% and 0.107%, respectively. (The relative error for the Erlang 2 model is approximately 44% of the relative error for the exponential case whenever $\mu > 1.0$, regardless of the relative amplitude.)

Table 2 SPEA for the Deterministic Model: Relative Errors (Percent)

	Relative Amplitude										
Mu	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
0.042	100.000	90.000	80.000	70.000	60.000	50.000	40.000	30.000	20.000	10.000	1.000
0.083	22.203	20.792	19.261	17.595	15.777	13.782	11.585	9.153	6.447	3.416	0.361
0.167	5.246	4.956	4.636	4.280	3.883	3.437	2.932	2.355	1.689	0.915	0.099
0.333	1.292	1.223	1.147	1.061	0.966	0.857	0.734	0.592	0.427	0.232	0.025
0.500	0.572	0.542	0.509	0.471	0.429	0.381	0.326	0.263	0.190	0.104	0.011
0.667	0.322	0.305	0.286	0.265	0.241	0.214	0.184	0.148	0.107	0.058	0.006
1	0.143	0.135	0.127	0.118	0.107	0.095	0.082	0.066	0.048	0.026	0.003
1.333	0.080	0.076	0.071	0.066	0.060	0.054	0.046	0.037	0.027	0.015	0.002
2	0.036	0.034	0.032	0.029	0.027	0.024	0.020	0.016	0.012	0.006	0.001
4	0.009	0.008	0.008	0.007	0.007	0.006	0.005	0.004	0.003	0.002	0.000
8	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000
16	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
32	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
64	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
128	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
256	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
512	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Our concept of a "good" approximation can now be made explicit. By good, we mean relative errors on the order of a percent and one can see from the tables that all models with $\mu > 1.333$ satisfy this definition.

4. Peak-Hour Approximations

Having demonstrated that the PSA approach works well for peak epochs, we now examine its accuracy for peak hours. Since, as demonstrated in section 3, the PSA approach is worst for the exponential service time model it will be sufficient to demonstrate the utility of L_{SPHA} by an analysis of the exponential service time case. In the following we make use of the exact solution for L_t in the exponential case (EMW eq. 15.) Since L_t is symmetric about its peak epoch, $L_{\text{peak-hour}}$, the peak-hour average number of customers in the system, is obtained by

$$\begin{aligned} L_{\text{peak-hour}} &= \int_{\alpha}^{\alpha+1} L_t\, dt = \int_{\alpha}^{\alpha+1} \lambda/\mu \\ &\quad + A/\mu\left[\frac{\sin(\gamma t) - \gamma/\mu \cos(\gamma t)}{1 + (\gamma/\mu)^2}\right] dt \\ &= \frac{\lambda}{\mu} + \frac{A/\mu}{1 + (\gamma/\mu)^2} \\ &\quad \times \left[\frac{\sin(\gamma\alpha) - \sin(\gamma(\alpha+1))}{\mu}\right. \\ &\qquad \left. + \frac{\cos(\gamma\alpha) - \cos(\gamma(\alpha+1))}{\gamma}\right], \end{aligned} \tag{15}$$

where

$$\alpha = 6 + \frac{\cot^{-1}(\mu/\gamma)}{\gamma} - 1/2 \tag{16}$$

is the beginning of the peak hour.

From equations (12) and (13) we have

$$\begin{aligned} L_{\mathrm{SPHA}} &= \frac{\lambda}{\mu} + \frac{A}{\mu\gamma}\left[\cos\left(\frac{11\pi}{24}\right) - \cos\left(\frac{13\pi}{24}\right)\right] \\ &= \frac{\lambda + 0.997A}{\mu}. \end{aligned} \tag{17}$$

We note that L_{SPEA} and L_{SPHA} are nearly identical because of the flatness of the sine function in the hour surrounding the peak. Table 3 contains our comparisons of L_{SPHA} to the exact peak hour average expected number in the system for the case of exponential service times with $RA = 1$ (a worst case) and the same range of service rates as in Table 1. Table 3 shows that L_{SPHA} is a very good approximation for the average peak hour expected number in the system whenever μ is greater than one. Furthermore, because of the flatness of the exact solution in the hour surrounding its peak, there is little difference between the peak-epoch and peak-hour values of either the approximate or exact solutions. Thus, for all practical purposes, the simpler $L_{\mathrm{SPEA}} = (\lambda + A)/\mu$ can be used to approximate either the peak-epoch or peak-hour congestion.

5. Time-Lags of the Peak-Epoch Congestion

In queuing systems where customers arrive in a periodic pattern, congestion builds to the peak epoch which occurs some time *after* the peak in the arrival rate function, itself. In the $M_t/G/\infty$ system, with sinusoidal arrivals, the time lag from the peak of the arrival rate to the epoch of peak congestion depends only on the service rate—decreasing rapidly in μ for both the exponential and deterministic service time models. This is shown by equations (5) and (7), and also in Table 4. The time lags are generally less for the deterministic service time model. As is demonstrated in Table 4, for $\mu > 1$ one can approximate the time lag for exponential service quite closely by twice the time lag for deterministic service, that is, simply by $1/\mu$. The error is about 2% at $\mu = 1$ and decreases rapidly as μ increases.

6. Closing Remarks

We feel confident that our conclusions that the L_{SPEA} and L_{SPHA} are excellent approximations apply to a very broad range of practical situations. Our own experience has been with applications of queuing models to telemarketing, police patrol, fire fighting, hospitals, copy machine repairs, and automatic teller machine operations. In all these situations service times were close to exponential, relative amplitudes were close to one, cycle lengths were 24 hours, and service rates were in the range of say 0.5 (copy machine repairs) to 100 (automatic teller machine transactions). The arrival rate

Table 3 Comparing SPEA and SPHA to Exact Peak Epoch and Peak Hour Congestion

	Exponential Service Times, RA = 1							
	Peak Epoch Congestion			Peak Hour Congestion			Actual Peak Point-Hour	
Mu	Exact	SPEA	RelErr	Exact	SPHA	RelErr	Diff	Pct Diff
0.042	111.089	192.000	72.834	111.046	191.726	72.655	0.043	0.039
0.083	62.559	96.000	53.455	62.518	95.863	53.338	0.042	0.066
0.167	36.889	48.000	30.121	36.852	47.932	30.065	0.037	0.100
0.333	21.437	24.000	11.955	21.410	23.966	11.935	0.027	0.126
0.500	15.087	16.000	6.050	15.067	15.977	6.041	0.020	0.134
0.667	11.585	12.000	3.584	11.569	11.983	3.579	0.016	0.138
1	7.870	8.000	1.657	7.859	7.989	1.655	0.011	0.140
1.333	5.944	6.000	0.946	5.935	5.991	0.944	0.008	0.141
2	3.983	4.000	0.425	3.977	3.994	0.424	0.006	0.142
4	1.998	2.000	0.107	1.995	1.997	0.107	0.003	0.143
8	1.000	1.000	0.027	0.998	0.999	0.027	0.001	0.143
16	0.500	0.500	0.007	0.499	0.499	0.007	0.001	0.143
32	0.250	0.250	0.002	0.250	0.250	0.002	0.000	0.143
64	0.125	0.125	0.000	0.125	0.125	0.000	0.000	0.143
128	0.062	0.063	0.000	0.062	0.062	0.000	0.000	0.143
256	0.031	0.031	0.000	0.031	0.031	0.000	0.000	0.143
512	0.016	0.016	0.000	0.016	0.016	0.000	0.000	0.143

Table 4 Time Lags for Peak Congestion (Hours)

	Exact Time Lags		Approximation for Exponential	
Mu	Deterministic	Exponential	$(1/\mu)$	Percent Error
0.042	12.0000	5.3971	24.0000	344.68
0.083	6.0000	4.8229	12.0000	148.81
0.167	3.0000	3.8346	6.0000	56.47
0.333	1.5000	2.5431	3.0000	17.97
0.500	1.0000	1.8424	2.0000	8.55
0.667	0.7500	1.4293	1.5000	4.94
1	0.5000	0.9780	1.0000	2.24
1.333	0.3750	0.7406	0.7500	1.27
2	0.2500	0.4972	0.5000	0.57
4	0.1250	0.2496	0.2500	0.14
8	0.0625	0.1250	0.1250	0.04
16	0.0313	0.0625	0.0625	0.01
32	0.0156	0.0312	0.0313	0.00
64	0.0078	0.0156	0.0156	0.00
128	0.0039	0.0078	0.0078	0.00
256	0.0020	0.0039	0.0039	0.00
512	0.0010	0.0020	0.0020	0.00

process in each instance was smooth enough and flat enough at the peak that our conclusions about the accuracy of the approximation should apply. Furthermore, in many of these applications there was an operating policy to keep customer delays close to zero—a situation consistent with the use of an infinite server model.

Our numerical results for the three service time distributions—deterministic, gamma (2) and exponential—illustrate that the difference between the approximate and exact solution increases as the coefficient of variation increases. Thus, the reader is advised that, for coefficients of variation greater than 1, the L_{SPEA} and L_{SPHA} will be worse than Tables 1, 2 and 3 indicate.

References

Andrews, B. H., H. L. Parsons 1989. L. L. Bean chooses a telephone agent scheduling system. *Interfaces* **19**, 6, 1–9.

Edie, L. C. 1954. Traffic Delays at Toll Booths. *Oper. Res.* **2** 107–138.

Eick, S. G., W. A. Massey, W. Whitt 1993a. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

——, ——, – 1993b. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

Green, L., P. Kolesar 1984. The feasibility of one-officer patrol in New York City. *Management Sci.* **20** 964–981.

——, – 1989. Testing the validity of a queuing model of police patrol. *Management Sci.* **35** 127–148.

——, – 1991, The pointwise stationary approximations for queues with nonstationary arrivals. *Management Sci.* **37** 84–97.

——, – 1995. On the accuracy of the simple peak-hour approximation for Markovian queues. *Management Sci.* **41** 1353–1370.

——, – A. Svoronos 1991. Some effects of nonstationarity on multiserver Markovian queuing systems. *Oper. Res.* **39** 502–511.

Gross, D., C. M. Harris 1985. *Fundamentals of Queuing Theory*, 2nd ed., John Wiley & Sons, New York.

Harel, A. 1988. Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Management Sci.* **34** 959–972; Erratum **34** (1988), p. 1277.

Holloron, T. J., J. E. Byrne 1986. United Airlines station manpower planning system. *Interfaces* **16** 39–50.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt 1996. Server staffing to meet nonstationary demand. *Management Sci.* **42** 1383–1394.

Khintchine, A. Y. 1960. *Mathematical Methods of Queuing*. Charles Griffen & Company, London (translation from the Russian book of 1955).

Kolesar, P. 1984. Stalking the endangered CAT: a queuing analysis of congestion at automated teller machines. *Interfaces* **14** 16–26.

——, L. Green 1998. Generating insight on service system design via a normal approximation to Erlang's delay formula. *Production and Oper. Management* **7** 282–293.

-, K. L. Rider, T. B. Craybill, W. W. Walker 1975. A queuing-linear programming approach to scheduling police patrol cars. *Oper. Res.* **23** 1045–1062.

Koopman, B. O. 1972. Air-terminal queues under time-dependent conditions. *Operations Res.* **20** 1089–1114.

Larson, R. C. 1972. *Urban Police Patrol Analysis*. MIT Press, Cambridge.

Oliver, R. M., A. H. Samuel 1962. Reducing letter delays in post offices. *Oper. Res.* **10** 839–892.

Palm, C. 1988. *Intensity Variations in Telephone Traffic*. North-Holland, Amsterdam (translation of a 1943 article in *Ericsson Technics* **44** 1–189.)

Segal, M. 1974. The operator scheduling problem: a network flow approach. *Oper. Res.* **22** 808–823.

Walker, J. M. Chaiken, E. J. Ignall, eds. 1979. *Fire Department Deployment Analysis*. North-Holland, New York.

Whitt, W. 1991. The Pointwise Stationary Approximation for $M_t/M_t/S$ Queues is Asymptotically Correct as the Rates Increase. *Management Sci.* **37** 307–314.

Accepted by G. Bitran; received November 22, 1994. This paper has been with the authors for 7 months for 2 revisions.