The Lagged PSA for Estimating Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates

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

We propose using a modification of the *simple peak hour approximation* (SPHA) for estimating peak congestion in multiserver queueing systems with exponential service times and time-varying periodic Poisson arrivals. This *lagged pointwise stationary approximation* (lagged PSA) is obtained by first estimating the time of the actual peak congestion by the time of peak congestion in an infinite server model and then substituting the arrival rate at this time in the corresponding stationary finite server model. We show that the lagged PSA is always more accurate than the SPHA and results in dramatically smaller errors when average service times are greater than a half an hour (based on a 24 hour period). More importantly, the lagged PSA reliably identifies proper staffing levels to meet targeted performance levels to keep congestion low.

(*Queues; Nonstationarity; Approximations*)

1. Introduction

There has been a longstanding practice among analysts of service systems with cyclic customer arrival processes to estimate peak congestion by using peak arrival rates in stationary queueing models. Here we identify the conditions under which this intuitive procedure which we call the *simple peak hour approximation* (SPHA) makes sense. This work is part of our continuing research in identifying simple approximations for queues with cyclic arrivals which are found in many real contexts (see e.g. Edie 1954, Segal 1974, Koopman 1972, Kolesar et al. 1975, Kolesar 1984, Holloran and Byrne 1986, Green and Kolesar 1989), but are generally too complex to solve analytically.

In an earlier paper (Green and Kolesar 1994), we tested the accuracy of the SPHA for multiple server Markovian queues with sinusoidal arrival rates. We found that the service rate μ is the major determinant of the accuracy of the SPHA and that for most practical purposes the SPHA is good enough whenever $\mu \geq 2$ in systems with a period of 24 hours. In a more recent paper (Green and Kolesar 1996) we have shown that for the $M_t/G/\infty$ queue with sinusoidal customer arrivals, the SPHA for the number of customers in the system is very good whenever $\mu \geq 1$.

In this paper we show that a relatively simple modification of the SPHA extends its utility for finite server systems well into the range of $\mu < 2$. Our approach is based on the fact that the epoch of peak congestion in a cyclic queueing system lags the epoch of peak customer arrivals and that estimating this lag is the key to improving the SPHA.

The accuracy of the SPHA is related to the accuracy of what we have called the *pointwise stationary approximation* or PSA (Green and Kolesar 1991). The PSA models the behavior of the system at each point in time using a stationary model with the arrival rate at that epoch. The peak of the PSA curve is what we have called the *simple peak epoch approximation* (SPEA). When the PSA curve is close to the actual system performance curve, the SPHA and will be close to the actual peak hour performance and the

0025-1909/97/4301/0080$01.25

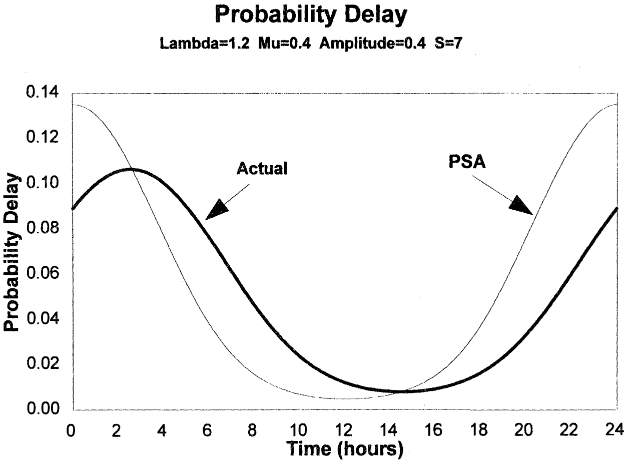
SPEA will be close to the actual peak epoch performance.

During our earlier empirical studies, we observed that the PSA and actual curves intersect at the epoch of actual peak congestion; see Figure 1 for an illustration. This occurs for the probability of delay, the expected delay and the expected number of customers in the system. Indeed, this result was proven to be true for the infinite server model with exponential service times (Eick et al. 1993a). Thus, if we knew the point of peak congestion (or equivalently the lag between the peak arrival rate and the peak congestion), we would only need to substitute the value of the arrival rate at that point into the PSA model, i.e. the simple stationary model, to get an exact solution for peak epoch behavior. Since, as shown in our earlier work, peak hour performance measures are almost identical to peak epoch measures for sinusoidal arrival rates (or any other arrival rate function which is relatively flat around the peak), this would also result in an excellent approximation for the peak hour or any other short interval of time. Clearly, this time lag is unknown but estimates can be obtained from results for the infinite server model developed by Eick, Massey and Whitt (1993a, 1993b). In this paper we show that this approach, which we call the *lagged PSA*, is simple to calculate and is always better than the SPHA when $\mu \leq 2$.

The proposed lagged PSA method is closely related to the *modified offered load* (MOL) approximation proposed by Jagerman (1975) to estimate blocking probabilities in the nonstationary Erlang loss model (see also Massey and Whitt 1994 and references therein). The MOL method estimates performance in the nonstationary system at time t by substituting the expected number of busy servers at time t obtained from the nonstationary infinite server model for the offered load (arrival rate times mean service time) at time t in the corresponding stationary finite server model. MOL has been shown to be quite accurate in estimating blocking probabilities in nonstationary Erlang loss models (Davis et al. 1995).

Jennings et al. (1996) suggest that the MOL could be used to estimate delay probabilities for determining server staffing in nonstationary finite server queues.

Figure 1 Probability Delay



The peak value produced by MOL in this case coincides with the lagged PSA estimate. Since MOL calculates congestion over time, it is more general than the lagged PSA but consequently requires more computation. In addition, for our focus of approximating peak congestion, the lagged PSA gives us insight as to how the accuracy of the approximation will be affected as a function of the system parameters by exploiting our knowledge about the behavior of the lag (see §§3 and 4). In a very recent paper, Massey and Whitt (1995) empirically test the MOL as a function of the number of servers for a specific example in which the arrival rate is changing slowly. We discuss their findings in light of our own in §4.

In §2, we describe the model and methodology we use in our analysis. In §3, we discuss the factors that affect lags in multiserver systems and the use of the infinite server model for estimating these lags. Section 4 contains numerical results on the accuracy of the lagged PSA for estimating peak epoch probability of delay and its usefulness for identifying appropriate server staffing levels under a broad range of conditions. As explained above, our results imply that this approach will also be good in these cases for the peak hour or any other short interval of time. We compare these results to those obtained using the PSA and another approximation based on the normal distribution recently proposed by Jennings et al. (1996). We end with a brief summary in §5.

2. Model and Methodology

Our analysis is based on $M(t)/M/s$ systems with $\lambda(t)$, the arrival rate at time t given by

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

where $\bar{\lambda}$ is the average arrival rate over the period T and A (>0) is the amplitude. The other model parameters are μ, the service rate and s, the number of servers. We assume that $\bar{\lambda} < s\mu$ and so the system will develop a periodic steady-state behavior (see Heyman and Whitt 1984 and Koopman 1972). Without loss of generality, we will assume that the period $T = 24$ hours.

Let $p_n(t)$ be the periodic steady-state probability that n customers are in the system at time t. These functions are the foundation of our results and are obtained by numerically solving the following standard set of differential equations that describe the system, see Gross and Harris (1985):

$$\begin{aligned} p_0'(t) &= -\lambda(t)p_0(t) + \mu p_1(t), \\ p_n'(t) &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) \\ &\quad - (\lambda(t) + n\mu)p_n(t), \quad 1 \leqq n < s, \\ p_n'(t) &= \lambda(t)p_{n-1}(t) + s\mu p_{n+1}(t) \\ &\quad - (\lambda(t) + s\mu)p_n(t), \quad n \geqq s. \end{aligned} \tag{2}$$

In this paper we focus on the probability of delay. Let $p_D(t)$ be the instantaneous probability that a customer arriving at time t is delayed. This is also the probability that all servers are busy at epoch t and is given by

$$p_D(t) = 1 - \sum_{n=0}^{s-1} p_n(t). \tag{3}$$

The peak epoch probability of delay is:

$$\text{peak } p_D = \max_{0 \le t \le 24} p_D(t). \tag{4}$$

The conclusions that follow are based on the examination of computational results for 169 model instances. We confine our study to systems in which the maximum traffic intensity is strictly less than one, that is, when

$$\rho_{\max} \equiv \max_t \frac{\lambda(t)}{s\mu} < 1. \tag{5}$$

We adopt this constraint because neither the SPHA nor the PSA are generally defined when $\rho_{\max}$ (rhomax) is greater or equal to one and because of computational difficulties which arise in solving (2) when system congestion is very high. Note, however, that the PSA for p_D is defined for any value of ρ (see Green and Kolesar 1991) and therefore the lagged PSA approach for estimating peak probability of delay can be used more generally. We also restrict our choice of parameter values so that the relative amplitude, $\text{RA} = A/\bar{\lambda} \le 1$, which makes $\lambda(t) \ge 0$ for all t. Since our previous research revealed that the SPHA is good for systems with service rates greater than 2, our choices of experimental models in this study focused on low service rates, i.e. $\mu < 2$, where the SPHA is not a useful approximation. We consider models with service rates as low as 0.125, that is with average service times as long as 8 hours, and with a broad range of average arrival rates. For each μ and λ, we varied the number of servers from the minimum needed to satisfy (5) to the number which resulted in a peak probability of delay less than or equal to 0.01. Otherwise, our choice of parameters was limited only by computational feasibility. More details on our experimental strategy are given in §4.

3. Estimating the Lag in the Peak

In Green and Kolesar (1994) we showed that for multiserver systems, the magnitude of the lag in peak congestion relative to the peak in the arrival rate depends on several factors. The primary determinant is the event frequency, i.e. the average number of arrivals and service completions per period. As the event frequency increases, the lag decreases. This is consistent with Whitt's (1991) result that the PSA is asymptotically correct as the arrival and service rates increase.

The lagged PSA uses the lag from the infinite server model to estimate the lag in the finite server system. It is important to note that in an infinite server model there is no queue and the usual performance measure is the expected number of busy servers. This measure most closely corresponds to the expected number of customers in system for finite server models. However, we propose the use of the infinite model lag for the expected number of busy servers to estimate the lags of probability of delay and expected delay as well as expected number in system in the finite server system. Though, as we showed in Green and Kolesar (1994), each of the several performance curves peaks at a somewhat different time, these differences are small particularly when peak probability of delay is not high. And, our principle aim is to identify staffing levels for such systems so that delays are not high.

For the infinite server model, the lag is solely a function of the service rate. The time of the peak of the mean number of busy servers for the case of sinusoidal arrival rates is given by (Eick et al. 1993b):

$$t_m = t_\lambda + \cot^{-1}(\mu/\gamma)/\gamma, \tag{6}$$

where $\gamma = 2\pi/24$ and t_λ is the time of the peak arrival rate. Thus the time lag of the peak is given (in hours) by

$$t_{\text{lag}} = (\cot^{-1}(\mu/\gamma))/\gamma. \tag{7}$$

In finite server systems, the time lag also increases as the peak probability of delay increases. So, for a given service rate, the lag predicted by the infinite server model which has no delays underestimates the actual lag in the finite server system. Therefore, the accuracy of Eq. (7) as a predictor of the lag in the finite system decreases as the peak probability of delay increases. This can be seen from the last column of Table 1 which illustrates the case of $\mu = 0.25$ for which the actual infinite model lag is 3.08 hours.

The lag can also be estimated for systems with nonsinusoidal arrival rates. Eick et al. (1993a) show that if $\lambda(t)$ is approximately quadratic before the peak, then the lag will be approximately equal to the expected service time when the service time is exponential. They also develop explicit formulas for $m(t)$ when $\lambda(t)$ is polynomial or a step function and propose several approximations for general arrival rate functions. Using simple calculus, these can be used to obtain estimates for the time of the extreme value and thus, the lag.

4. The Accuracy of the Lagged PSA

Our method consists of estimating t_m using (6), calculating the value of λ at t_m using (1) and approximating the peak p_D by using $\lambda(t_m)$ in the stationary $M/M/s$ equation. Our results are based on a study of four values of μ: 0.125, 0.25, 0.5, and 1, with λ and s varying over a broad range of values and RA fixed at a "worst case" value of 1. In all of these cases, the p_D computed from this lagged PSA approach is an upper bound to the actual peak probability of delay. This is because the predicted lag is a lower bound of the true lag and the true lag is less than 12 hours. (The maximum actual lag we observed was 4.5 hours. This occurred, of course, for the smallest μ we examined, $\mu = 0.125$ which is a mean service time of 8 hours.) Since the $\lambda(t)$ curve decreases for 12 hours after the peak, the value of $\lambda(t)$ that we use in the approximation is an upper bound of $\lambda(t_m)$, the arrival rate at the point where the PSA and actual curves intersect. (Of course, if μ is sufficiently small, the lag may be greater than 12 hours and thus the lagged PSA might not result in an upper bound. Given our results, it appears that this is unlikely to occur for expected service times shorter than 24 hours and thus is not an issue for most real systems.) Because of convergence problems with our numerical solution algorithm, we were unable to explore systems with such long expected service times.

How good is the lagged PSA? We consider 2 criteria: estimating system performance and identifying the minimum staffing levels needed to achieve a performance target. For each of these, we compare the accuracy of the lagged PSA with alternative approximations, particularly the SPHA.

4.1. Estimating Performance

Table 1 displays our results for $\mu = 0.25$. The relative errors for this value of μ are representative of the accuracy of the lagged PSA for all of the values of μ which we considered. (Of course, for larger values of μ, the lag approaches zero and thus the errors for the lagged PSA will go to zero.)

Since the lag is never negative, the lagged PSA is always smaller than the peak PSA (also called the *simple peak epoch approximation* or SPEA) and is thus a better approximation. This is illustrated in Table 1 for $\mu = 0.25$. Here, the SPEA is clearly awful with relatives errors starting at about 35% and going up to 545%. In contrast, the largest relative errors for the lagged PSA are about 27% and in many cases fall below the 10% level. Notice that for systems with small p_D the lagged PSA is quite good. This is consistent with the observation made by Massey and Whitt (1995) that the MOL improves as the number of servers increases. For all the 169 cases we examined across the four cited values of μ, the relative error of the lagged PSA was always below 35%.

Table 1 Comparison of *pD* Approximations

					Mu = 0.25					
					pD Approximations			Errors (%)		Lag (Hrs.)
		Rho		Lagged		Infinite	Lagged		Infinite	
Lambda	*s*	Max	Actual *pD*	PSA	SPEA	Normal	PSA	SPEA	Normal	
0.0625	1	0.50	0.372	0.423	0.500	0.859	13.53	34.30	130.68	3.50
	2	0.25	0.070	0.074	0.100	0.060	4.81	42.14	−14.36	3.25
	3	0.17	0.009	0.010	0.015	0.001	1.89	60.90	−91.96	3.17
	4	0.13	0.001	0.001	0.002	0.000	0.83	86.85	−99.88	3.08
0.125	2	0.50	0.223	0.251	0.333	0.363	12.63	49.50	62.96	3.42
	3	0.33	0.057	0.060	0.091	0.044	5.43	60.21	−23.27	3.25
	4	0.25	0.011	0.012	0.020	0.002	2.54	80.78	−81.06	3.17
	5	0.20	0.002	0.002	0.004	0.000	1.36	108.89	−97.97	3.17
0.25	3	0.67	0.262	0.309	0.444	0.419	18.07	69.64	59.79	3.50
	4	0.50	0.098	0.107	0.174	0.106	9.19	77.63	8.29	3.33
	5	0.40	0.030	0.032	0.060	0.018	4.88	95.86	−41.46	3.25
	6	0.33	0.008	0.008	0.018	0.002	2.73	122.42	−77.01	3.17
	7	0.29	0.002	0.002	0.005	0.000	1.66	156.86	−93.84	3.17
0.5	5	0.80	0.277	0.341	0.554	0.428	23.20	100.20	54.60	3.50
	6	0.67	0.137	0.156	0.285	0.169	13.82	107.37	23.20	3.33
	7	0.57	0.060	0.065	0.135	0.055	8.39	124.17	−8.38	3.25
	8	0.50	0.024	0.025	0.059	0.014	5.24	149.60	−38.84	3.17
	9	0.44	0.008	0.009	0.024	0.003	3.35	183.37	−64.52	3.17
1	9	0.89	0.263	0.333	0.653	0.390	26.44	148.10	47.99	3.42
	10	0.80	0.159	0.187	0.409	0.203	18.06	157.79	28.03	3.33
	11	0.73	0.089	0.100	0.245	0.097	12.43	175.97	8.77	3.25
	12	0.67	0.046	0.050	0.140	0.041	8.66	202.52	−10.54	3.25
	13	0.62	0.023	0.024	0.076	0.016	6.11	237.74	−29.69	3.17
	14	0.57	0.010	0.011	0.039	0.005	4.36	282.32	−47.78	3.17
	15	0.53	0.004	0.005	0.019	0.002	3.17	337.37	−63.68	3.17
2	17	0.94	0.222	0.282	0.737	0.310	26.70	231.68	39.55	3.33
	18	0.89	0.152	0.183	0.531	0.193	20.32	248.99	26.85	3.25
	19	0.84	0.100	0.115	0.374	0.115	15.50	273.99	14.80	3.25
	20	0.80	0.063	0.070	0.256	0.065	11.90	307.39	2.98	3.25
	21	0.76	0.038	0.041	0.171	0.035	9.18	349.95	−8.86	3.17
	22	0.73	0.022	0.024	0.111	0.017	7.11	402.76	−20.76	3.17
	24	0.67	0.007	0.007	0.043	0.004	4.39	545.49	−44.03	3.17

We also considered the infinite server approximation used in Jennings et al. (1996). The probability of delay at t is estimated by

$$p_D(t) \approx 1 - P(Q(t) \le s - 1) \qquad (8)$$

where $Q(t)$ is the number of busy servers at t in the infinite server model. For the case of exponential service times, the distribution of $Q(t)$ at the peak is approximated by a normal distribution with mean $m(t)$ and variance $v(t)$ given by

$$v(t) = m(t) = \frac{\bar{\lambda}}{\mu} + \frac{A}{\mu}\frac{1}{\sqrt{1 + (\gamma/\mu)^2}}. \qquad (9)$$

Thus, incorporating the standard continuity correction of +0.5, the peak probability of delay may be estimated from

Table 2 Staffing Requirements to Meet Target Peak *pD*

			Mu = 0.125			Mu = 0.25			Mu = 0.5	
	Target		Lagged			Lagged			Lagged	
Lambda	*pD*	Actual	PSA	SPEA	Actual	PSA	SPEA	Actual	PSA	SPEA
0.125	0.2	3	4	4	3	3	3	2	2	2
	0.1	4	4	5	3	3	3	2	2	2
	0.05	5	5	6	4	4	4	3	3	3
	0.01	6	6	7	4	4	5	3	3	4
0.25	0.2	5	5	7	4	4	4	3	3	3
	0.1	6	6	8	4	5	5	3	3	3
	0.05	7	7	9	5	5	6	4	4	4
	0.01	8	8	10	6	6	7	5	5	5
0.5	0.2	9	9	12	6	6	7	4	4	4
	0.1	10	10	13	7	7	8	5	5	5
	0.05	11	11	14	8	8	9	5	5	6
	0.01	13	13	16	9	9	10	6	6	7
1	0.2	17	17	21	10	10	12	7	7	7
	0.1	17	17	23	11	11	13	7	8	8
	0.05	18	18	24	12	12	14	8	8	9
	0.01	20	21	27	14	14	16	10	10	10

$$1 - \Phi\left(\frac{s - 1 + 0.5 - m(t)}{\sqrt{m(t)}}\right) \qquad (10)$$

where $\Phi(\ )$ is the standard normal probability function. A better approximation can then be obtained by a refinement suggested in §4 of Jennings et al. (1995) to account for using an infinite server approximation for a finite server system.

The results, contained in the column labeled "Infinite Normal" in Table 1, indicate that this method is generally far less accurate than the lagged PSA, particularly for staffing levels which result in a low probability of delay.

The maximum traffic intensity, rhomax, and the number of servers are the dominant factors affecting the size of the errors from using the lagged PSA. Using the standard of a relative error of 10% or less, p_D is satisfactorily approximated for all cases in our set that have a rhomax of less than or equal to 0.5. For larger numbers of servers ($s > 2$), the 10% standard holds for higher rhomax.

Although our work has focused on systems with relative amplitudes of 1, which we consider a worst case and a realistic model for many systems, we also looked at the effects of relative amplitude. Like the SPHA, the lagged PSA is significantly better for smaller values of RA. As an example, the errors for $\mu = 1$ and RA $= 1$ range up to 30% while for RA $= .5$ the maximum error is just above 5%.

4.2. Capacity Planning

Queueing models are often used to help identify system capacity or staffing levels necessary to achieve desirable performance. Much of our work in nonstationary systems has been motivated by our efforts in supporting such decisions for managing emergency services such as police patrol, firefighting and ambulances (e.g. Kolesar et al. 1975 and Green and Kolesar 1984). In these systems, it is desirable to keep peak p_D low. This goal of keeping delays small is, of course, becoming more prevalent in many service systems due to competitive pressures. We consider target peak p_D levels of 20%, 10%, 5% and 1%.

Our database again consists of all the 169 cases we examined for service rates ranging from 0.125 to 2. For each case examined, we compared the actual number of servers needed to meet each of the four target values with the number suggested by using the lagged PSA. This staffing level was determined by taking the smallest number of servers that resulted in a peak p_D (rounded to 2 decimal places) that was less than or equal to the target. In almost every case, the lagged PSA identified the same number of servers as the actual model—that is, made the correct resource allocation. In the few cases in which the lagged PSA was off, it was off by only one server. These results are illustrated in Table 2 for $\mu = 0.125$, 0.25 and 0.5 which also shows the values suggested by the SPEA. As can be easily seen, the SPEA often suggests the wrong number of servers and is consistently off for small μ such as $\mu = 0.125$. For such a low service rate, the error in the SPEA can be significant as shown by the case of $\lambda = 1$ and a targeted peak p_D of 0.1. Here, the actual model and the lagged PSA both identify 17 servers as being needed while the SPEA suggests 23.

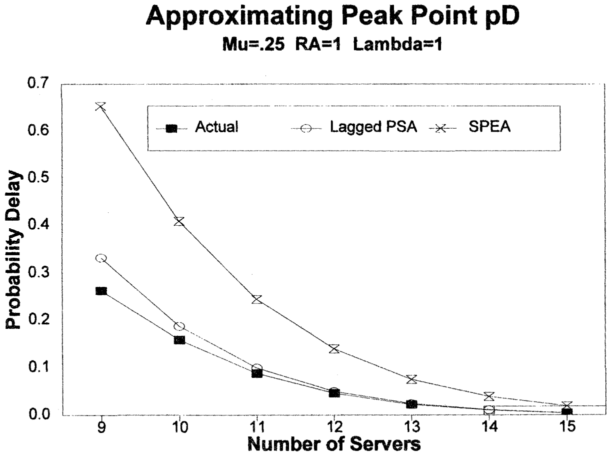
Figure 2 plots the lagged PSA, the SPEA and the actual peak p_D versus the number of servers for one of the cases in Table 1. The accuracy of the lagged PSA for low p_D is apparent as is its superiority to the SPEA.

Although we have used the lagged PSA to estimate peak epoch performance, it is likely to be as accurate for estimating peak hour performance as well—as long as the arrival rate is relatively flat around its peak. As shown in our previous work (Green and Kolesar 1994), the difference in the magnitude of the peak epoch delay and peak hour delay is very small for models with sinusoidal arrival rates. Also, the relative errors using the SPHA for estimating peak hour delays are almost identical to those using the SPEA for estimating peak epoch delays. Thus, the above results clearly indicate that the lagged PSA will always be better than the SPHA when service rates are low.

5. Conclusion

Previous work has shown that the SPEA and SPHA are good approximations for peak congestion in finite server systems for large service rates, but can be very inaccurate when $\mu < 2$. The work reported in this paper shows that the lagged PSA is far more accurate than the SPEA and the SPHA in these cases and is very reliable in identifying staffing levels to meet a targeted small probability of delay in the range of $0.125 \le \mu \le 2$. Furthermore, because at $\mu = 0.125$ the lag occurs at the steepest part of the $\lambda(t)$ curve and hence the error in estimating $\lambda(t_m)$ is most sensitive to an error in estimating t_{lag}, the errors are likely to be greatest for such values of μ. Hence, this method is likely to be good for smaller service rates as well.

Figure 2 **Approximating Peak Point** ***pD***



From a practical perspective, the lagged PSA is a simple modification of the often used SPHA or SPEA and thus is easy to understand and use. We propose that it be used for supporting capacity decisions for achieving targeted low peak probability of delay when mean service times are longer than half an hour.

References

Davis, J. L., W. A. Massey, and W. Whitt, "Sensitivity to the Service-Time Distribution in the Nonstationary Erlang Loss Model," *Management Sci.*, 41 (1995), 1107–1116.

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

Eick, S. G., W. A. Massey, and W. Whitt, "The Physics of the $M_t/G/\infty$ Queue," *Oper. Res.*, 41 (1993), 731–742.

——, —— and ——, "$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates," *Management Sci.*, 39 (1993), 241–252.

Green, L. V. and P. J. Kolesar, "The Feasibility of One-Officer Patrol in New York City," *Management Sci.*, 20 (1984), 964–981.

—— and ——, "Testing the Validity of a Queueing Model of Police Patrol," *Management Sci.*, 35 (1989), 127–148.

—— and ——, "The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals," *Management Sci.*, 37 (1991), 84–97.

Green, L. V. and P. J. Kolesar, "Simple Approximations of Peak Congestion in $M_t/G/\infty$ Queues with Sinusoidal Arrivals," Working Paper, Columbia University, New York, 1996.

—— and ——, "On the Accuracy of the Simple Peak Hour Approximation for Markovian Queues," *Management Sci.*, 41 (1995), 1353–1370.

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

Heyman, D. P. and W. Whitt, "The Asymptotic Behavior of Queues with Time-Varying Arrival Rates," *J. Appl. Prob.*, 21 (1984), 143–156.

Holloran, T. J. and J. E. Byrn, "United Airlines Station Manpower Planning System," *Interfaces*, 16 (1986), 39–50.

Jagerman, D. L., "Nonstationary Blocking in Telephone Traffic," *Bell System Tech. J.*, 54 (1975), 625–661.

Jennings, O. B., A. Mandelbaum, W. A. Massey, and W. Whitt, "Server Staffing to Meet Time-Varying Demand," *Management Science.*, 42 (1996), 1383–1394.

Kolesar, P. J., "Stalking the Endangered CAT: A Queueing Analysis of Congestion at Automated Teller Machines," *Interfaces*, 14 (1984), 16–26.

——, K. L. Rider, T. B. Craybill, and W. W. Walker, "A Queueing-Linear Programming Approach to Scheduling Police Patrol Cars," *Oper. Res.*, 23 (1975), 1045–1062.

Koopman, B. O., "Air-Terminal Queues under Time-Dependent Conditions," *Oper. Res.*, 20 (1972), 1089–1114.

Massey, W. A. and W. Whitt, "An Analysis of Modified Offered Load Approximation for the Nonstationary Erlang Loss Model," *Ann. Appl. Prob.*, 4 (1994), 145–153.

—— and ——, "Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates," AT&T Bell Laboratories, Murray Hill, NJ, 1995.

Segal, M., "The Operator Scheduling Problem: A Network Flow Approach," *Operations Res.*, 22 (1974), 808–823.

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

Accepted by Gabriel Bitran; received April 16, 1995. *This paper has been with the authors* 2 *months for* 1 *revision.*