Queueing Systems with Service Interruptions: An Approximation Model

Philippe Nain

Queueing Systems with Service Interruptions: An Approximation Model

Philippe Nain
INRIA, Domaine de Voluceau, 78153 Rocquencourt, Le Chesnay Cédex, France

Received 8 July 1982

We present a single queueing model which can be used to analyse queueing systems with service interruptions. The model is based on a diffusion approximation using an instantaneous return process which reflects the particular queueing system under consideration. Applications to queueing systems with preemptive resume priority and breakdowns are given in this paper as well as numerical results showing the accuracy of the approximation.

Keywords: Queueing Systems, Diffusion Process, Breakdowns, Preemptive Resume Priority.

Philippe Nain was born in Paris. He received a Master's degree in Mathematics in 1978, a Diplome d'Etude Approfondie (D.E.A) in Statistics in 1979 and a Doctorat de 3ème cycle specializing in Modelling Computer Systems in 1981 at the University of Paris XI at Orsay. He is currently a researcher in Computer Science at the I.N.R.I.A (France). His research interests include modelling and queueing theory.

1. Introduction

This paper is an extension of Fischer's paper [2] in which the author proposed a model to approximate the distribution function of the steady state load of a given class of customers, for various systems with service interruptions. We pursue and generalize this work, in order to approximate the stationary queue length distribution in such systems, in the case where the arrival processes are not necessarily all Poisson processes. We use the method of the diffusion approximation which is based, according to the central limit theorem, on the approximation of a process which is not time-continuous by a time-continuous process. This technique has originated in the queueing theory with the work of Gaver [3] and Newell [14] for a single-server queue and generalised to queueing networks by Kobayashi [11,12]. Gelenbe [4] investigated the instantaneous return process which permits of not requiring the introduction of results from queueing theory to obtain the boundary conditions.

The diffusion approximation is used in queueing theory to obtain good approximations [15] of steady state [6,7,8] as well as time-dependent characteristics [3] of queueing systems whose corresponding exact results are unknown or not readily usable.

In particular, queueing systems with service interruptions are often such systems [13].

In this paper we use a diffusion approximation which reflects the particular system under consideration.

The structure of the approximation is given in Section 2. Applications to queueing systems with preemptive resume priority and breakdowns are proposed in Section 3.

Numerical comparisons with the exact results or with results obtained through simulation models are given in Section 4.

2. The approximation

In what follows, we consider a GI/GI/1 queue with interarrival times distribution of mean $1/\lambda$ and variance V_a and with service times distribution of mean $1/\mu$ and variance V_s. We assume that $\rho \triangleq \lambda/\mu < 1$.

Let $\{P(k)\}_{k\in\mathbb{N}}$ be the stationary queue length distribution. If the server continually serves customers (ρ close to 1), then Gelenbe [4] used a diffusion approximation with an instantaneous return process to show that $P(k)$ $(k = 0, 1, 2, \dots)$ is given by

$$P(k) \sim \begin{cases} 1-\rho & \text{if } k = 0, \\ \rho(1-\hat{\rho})\hat{\rho}^{k-1} & \text{if } k = 1, 2, \dots \end{cases} \tag{0}$$

where $\hat{\rho} = \exp\{-2\beta/\alpha\}$, $\beta = \lambda - \mu$, $\alpha = \lambda^3 V_a + \mu^3 V_s$.

β and α are called the infinitesimal mean and variance of the diffusion process approximating the number of customers in the queue at time t. The mean queue length at steady state provided by this approximation is then given by

$$N \sim \frac{\rho}{1-\hat{\rho}}.$$

We now consider a GI/GI/1 queue with service interruptions. That is, for periods of time, the system behaves as a standard GI/GI/1 queue, and for periods of time as a system with arrivals and no departures.

The idea is to modify the infinitesimal mean and variance, β and α, in order to take into account the service interruptions.

Let $I(x)$ be the distribution function for a length of time during which the server is available and $B(x)$ the distribution function for a length of time during which the server is unavailable.

We assume that $I(x)$ and $B(x)$ have finite means, $E(I)$ and $E(B)$, and variances, V_I and V_B. We further assume that all random variables are mutually independent.

Let $N(t)$ be the queue size at time t, $A(t)$ and $D(t)$ the number of arrivals and departures in $(0, t]$, respectively; then

$$N(t) = N(0) + A(t) - D(t), \quad t > 0. \tag{1}$$

For any renewal process $\{M(t);\ t \geqslant 0\}$ where the inter-event times have mean m and variance V, for large values of t,

$$E(M(t)) \sim t/m \tag{2}$$

and

$$\operatorname{var}(M(t)) \sim tV/m^3 \quad \text{(see Cox [1])}. \tag{3}$$

By hypothesis $\{A(t);\ t \geqslant 0\}$ is a renewal process, so from (2) and (3) we obtain for large values of t

$$E(A(t)) \sim \lambda t, \tag{4}$$

$$\operatorname{var}(A(t)) \sim \lambda^3 V_a t. \tag{5}$$

The process $\{D(t);\ t \geqslant 0\}$ is not a renewal process because of the idle periods and of the unavailability periods of the server. However, in heavy traffic (see at the end of Section 2), the server will be occupied most of the time during its availability periods, so we approximate $D(t)$ by $\tilde{D}(t)$, where

$$E(\tilde{D}(t)) = \mu y(t), \tag{6}$$

$$\operatorname{var}(\tilde{D}(t)) = \mu^3 V_s y(t) \tag{7}$$

and where $y(t)$ is a random variable taking its values in $[0, t]$ and which is the amount of time in $[0, t]$ during which the server is available ($y(t) = t$ for a standard GI/GI/1 queue).

Using the following well-known formulas,

$$E(\tilde{D}(t)) = E\big(E(\tilde{D}(t)/y(t))\big)$$

and

$$\operatorname{var}(\tilde{D}(t)) = \operatorname{var}\big(E(\tilde{D}(t)/y(t))\big) + E\big(\operatorname{var}(\tilde{D}(t)/y(t))\big),$$

we obtain from relations (6) and (7)

$$E(\tilde{D}(t)) = \mu\, E(y(t)), \tag{8}$$

$$\operatorname{var}(\tilde{D}(t)) = \mu^2 \operatorname{var}(y(t)) + \mu^3 V_s E(y(t)). \tag{9}$$

We must now determine $E(y(t))$ and $\operatorname{var}(y(t))$ to obtain an approximation of $E(D(t))$ and of $\operatorname{var}(D(t))$.

Using the definition of $y(t)$ and renewal theory (see Cox [1, pp. 91–101]), we have that

$$E(y(t)) \sim \frac{E(I)}{E(I)+E(B)}\, t \tag{10}$$

and

$$\operatorname{var}(y(t)) \sim \frac{E(I)^2 V_B + E(B)^2 V_I}{(E(I)+E(B))^3}\, t \tag{11}$$

for large values of t.

From relations (1), (4), (5), (8), (9), (10) and (11) we obtain the approximate results

$$\lim_{t\to\infty} \frac{E(N(t))}{t} \sim \lambda - \mu \frac{E(I)}{E(I)+E(B)}$$

and

$$\lim_{t\to\infty}\frac{\operatorname{var}(N(t))}{t}\sim\lambda^3V_a+\mu^3V_s\frac{E(I)}{E(I)+E(B)}+\mu^2\frac{E(I)^2V_B+E(B)^2V_I}{(E(I)+E(B))^3},$$

which suggests that we approximate the jump process $\{N(t);\ t\geqslant 0\}$ by a diffusion process $\{\tilde{N}(t);\ t\geqslant 0\}$, with infinitesimal mean and variance given by

$$\tilde{\beta}=\lambda-\mu\frac{E(I)}{E(I)+E(B)}, \tag{12}$$

$$\tilde{\alpha}=\lambda^3V_a+\mu^3V_s\frac{E(I)}{E(I)+E(B)}+\mu^2\frac{E(I)^2V_B+E(B)^2V_I}{(E(I)+E(B))^3}. \tag{13}$$

Let $\{Q(k)\}_{k\in\mathbb{N}}$ be the stationary queue length distribution of this GI/GI/1 queue with service interruptions.

Using a diffusion approximation with an instantaneous return process [5], $Q(k)$ $(k=0,1,\dots)$ is then given by (see (0))

$$Q(k)\sim\begin{cases}1-\rho\dfrac{E(I)+E(B)}{E(I)} & \text{if } k=0,\\ \rho(1-\hat{q})\hat{q}^{k-1}\dfrac{E(I)+E(B)}{E(I)} & \text{if } k=1,2,\dots\end{cases} \tag{14}$$

where $\hat{q}=\exp\{-2\tilde{\beta}/\tilde{\alpha}\}$ with $\tilde{\beta}$ and $\tilde{\alpha}$ given by (12) and (13).

The hypothesis of heavy traffic takes the form of

$$\rho\frac{E(I)+E(B)}{E(I)}\quad\text{close to 1,}$$

and the mean queue length at steady state provided by this approximation is

$$\tilde{N}\sim\frac{\rho(E(I)+E(B))}{E(I)(1-\hat{q})}. \tag{15}$$

3. Applications

3.1. Queueing system with preemptive resume priority

We assume that customers arrive at a service facility at r priority levels ($r\geqslant 2$). A single server serves under a preemptive resume discipline (see Jaiswal [9]) where level 1 has the highest priority and level r, the lowest. Those customers arriving at priority level i will be called *i-customers* ($i=1,\dots,r$).

For each priority level $1,\dots,r-1$, the arrivals form a Poisson process with parameter $\lambda_1,\dots,\lambda_{r-1}$ and these renewal processes are mutually independent. For the lowest priority level r, the arrival process has an arbitrary renewal distribution function with finite mean $1/\lambda_r$ and variance V_{a_r} and this process is independent of the arrival processes for priority levels $1,\dots,r-1$.

The service times have an arbitrary distribution function which depends upon the priority level. Let $S_i(x)$ be the service times distribution function of the i-customers ($i=1,\dots,r$).

Further, we assume that the service times processes are mutually independent and independent of the arrival processes and, that the necessary equilibrium condition $\sum_{i=1}^{r}\rho_i<1$ is verified ($\rho_i\triangleq\lambda_i/\mu_i$) [9]. For the i-customers let $1/\mu_i$ be the mean service time and V_{s_i} the variance of the service times distribution ($i=1,\dots,r$).

Case 1. Let $r=2$.

For the 1-customers, the system behaves as a standard M/G/1 queueing system, and no approximation is required.

We use the approximation to describe the behavior of the 2-customers and particularly the queue size distribution at stationarity.

First, one must determine $B(x)$ and $I(x)$ in order to reflect the priority rule under consideration. The length of time the 2-customers have control of the server is the idle period of the 1-customers and so

$$I(x)=1-e^{-\lambda_1 x}\quad(x\geqslant 0).$$

Hence

$$E(I)=1/\lambda_1, \tag{16}$$

$$V_I=1/\lambda_1^2. \tag{17}$$

The length of time the 2-customers do not have

control of the server is the busy period of the 1-customers and so

$$E(B)=\frac{1}{\mu_1(1-\rho_1)}, \tag{18}$$

$$V_B=\frac{E(S_1^2)}{(1-\rho_1)^3}-E(B)^2 \tag{19}$$

where $E(S_1^2)$ is the second moment of the 1-customers service times distribution (see Kleinrock [10]).

Let $\{P_2(k)\}_{k\in\mathbb{N}}$ be the stationary queue size distribution of the 2-customers.

Using (12), (13), (14), (16), (17), (18), (19) we obtain

$$P_2(k)\sim\begin{cases}1-\dfrac{\rho_2}{1-\rho_1} & \text{if } k=0,\\ \dfrac{\rho_2}{1-\rho_1}(1-\hat\rho_2)\hat\rho_2^{k-1} & \text{if } k=1,2,\ldots\end{cases} \tag{20}$$

where

$$\hat\rho_2=\exp\{-2\beta_2/\alpha_2\},\qquad \beta_2=\lambda_2-\mu_2(1-\rho_1),$$
$$\alpha_2=\lambda_2^3V_{a_2}+\lambda_1\mu_2^2E(S_1^2)+\mu_2^3V_{s_2}(1-\rho_1).$$

From (15) we find the mean queue size at steady state,

$$N_2\sim\frac{\rho_2}{(1-\rho_1)(1-\hat\rho_2)}. \tag{21}$$

Case 2. Let $r>2$ and consider the r-customers.

The analysis proceeds through reductions of the processes of interest to corresponding processes in a simple generalization of an M/G/1 queue.

Indeed, regarding to the r-customers, the customers of priority level $1,\ldots,r-1$ have the same behavior as the behavior of customers of an M/G/1 queueing system with mean interarrival times $1/\Lambda_{r-1}$ where $\Lambda_{r-1}=\sum_{i=1}^{r-1}\lambda_i$ and with service times distribution function

$$\sum_{i=1}^{r-1}\lambda_iS_i(x)/\Lambda_{r-1}\quad(x\geqslant 0)$$

(see Welch [16]).

Thus, the problem is reduced to Case 1 (the '$r=2$' case). Let $\{P_r(k)\}_{k\in\mathbb{N}}$ be the stationary-queue size distribution of the r-customers.

Applying the same method as in Case 1 where the above M/G/1 queueing system is substituted for the M/G/1 queueing system with mean interarrival times $1/\lambda_1$ and with service times distribution function $S_1(x)$, we easily obtain from (20)

$$P_r(k)\sim\begin{cases}1-\dfrac{\rho_r}{1-\sum_{i=1}^{r-1}\rho_i} & \text{if } k=0,\\ \dfrac{\rho_r}{1-\sum_{i=1}^{r-1}\rho_i}(1-\hat\rho_r)\hat\rho_r^{k-1} & \text{if } k=1,2,\ldots\end{cases}$$

where $\hat\rho_r=\exp\{-2\beta_r/\alpha_r\}$, $\beta_r=\lambda_r-(1-\sum_{i=1}^{r-1}\rho_i)\mu_r$,

$$\alpha_r=\lambda_r^3V_{a_r}+\mu_r^3V_{s_r}\left(1-\sum_{i=1}^{r-1}\rho_i\right)+\mu_r^2\sum_{i=1}^{r-1}\lambda_iE(S_i^2).$$

The mean queue length at steady state provided by the above approximation, for the r-customers, is then, from (21),

$$N_r\sim\frac{\rho_r}{(1-\sum_{i=1}^{r-1}\rho_i)(1-\hat\rho_r)}.$$

Numerical comparisons between the approximate mean queue length at steady state (for $r=2$) and exact results or results obtained through a simulation model are given in Section 4, for different values of the traffic load.

3.2. Queueing systems with random or scheduled breakdowns

We consider a GI/GI/1 queue with two types of breakdowns, random or scheduled. Random breakdowns mean that the length of time the server is available has an exponential distribution and that the length of time the server is not available also has an exponential distribution, independent of the first.

Scheduled breakdowns mean that the availability and the unavailability periods of the server are deterministic. For either case, let i (b) be the expected length of the on (off) period.

So, for random breakdowns

$$I(x)=1-\exp\{-x/i\},$$
$$B(x)=1-\exp\{-x/b\},\quad x\geqslant 0$$

and for scheduled breakdowns

$$I(x)=\begin{cases}1 & \text{if } x\geqslant i,\\ 0 & \text{if } x<i\end{cases}$$

and

$$B(x)=\begin{cases}1 & \text{if } x\geqslant b,\\ 0 & \text{if } x<b.\end{cases}$$

Let N_R (N_S) be the mean queue size at stationarity

of the GI/GI/1 queue with random (scheduled) breakdowns.

Using the approximation described in Section 2, we obtain (see (14) and (15)):

$$N_R \sim \frac{\rho(i+b)}{i(1-\hat{\rho}_R)} \tag{22}$$

where $\hat{\rho}_R = \exp\{-2\beta_R/\alpha_R\}$, $\beta_R = \lambda - \mu i/(i+b)$,

$$\alpha_R = \lambda^3 V_a + \mu^3 V_s \frac{i}{i+b} + 2\mu^2 \frac{(ib)^2}{(i+b)^3}$$

for random breakdowns, and

$$N_S \sim \frac{\rho(i+b)}{i(1-\hat{\rho}_S)} \tag{23}$$

where $\hat{\rho}_S = \exp\{-2\beta_S/\alpha_S\}$, $\beta_S = \beta_R$,

$$\alpha_S = \lambda^3 V_a + \mu^3 V_s \frac{i}{i+b} \quad \text{for scheduled breakdowns.}$$

In both cases the equilibrium condition is $\rho < c \triangleq i/(i+b)$. No exact result is known for the GI/GI/1 queue with random or scheduled breakdowns. Exact results are only known for the corresponding M/G/1 queue (see Mitrany and Avi-Itzhak [13]).

Numerical comparisons between N_R (N_S) and the corresponding result obtained through a simulation model are given in Section 4, for different values of the traffic load.

4. Numerical comparisons and conclusions

In this section we present numerical comparisons between the expected queue length (exact results or results obtained through a simulation) and the approximation given in Section 3.

The exact results are known, for the queueing systems described in Section 3, only if all the arrival processes are Poisson processes. In the other cases we used simulation models.

For the preemptive resume priority, Figs. 1 to 3 give families of curves (each representing a given level of priority traffic) of the expected nonpriority queue length as a function of its traffic load. The exact results can be found in Jaiswal [9].

The approximation is very good *for any level of high priority traffic* in the Markovian case (see Fig. 1) (if interarrival times and service times have exponential distribution functions) and the accuracy of the approximation decreases as the coefficient of variation of the service times distribution decreases from 1 (see Figs. 2 and 3).

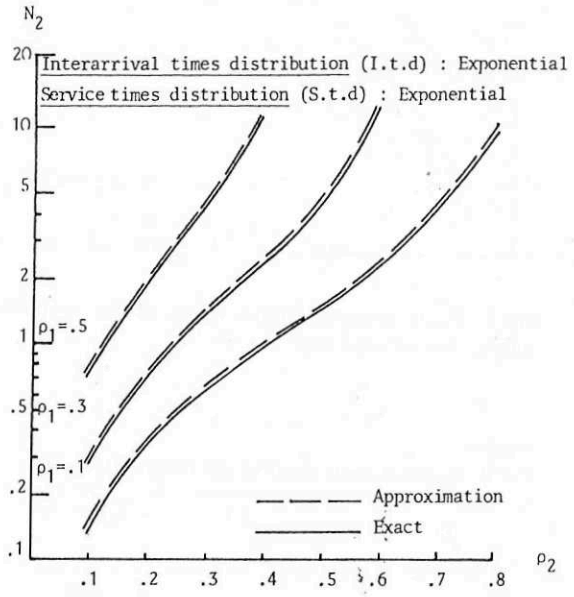

Fig. 1. Preemptive resume priority (P.R.P): Expected queue length of the nonpriority customers in the Markovian case.

Table 1 shows comparative results between the mean queue length at steady state obtained through a simulation model and the approximation described in Section 3, in the case where the arrival

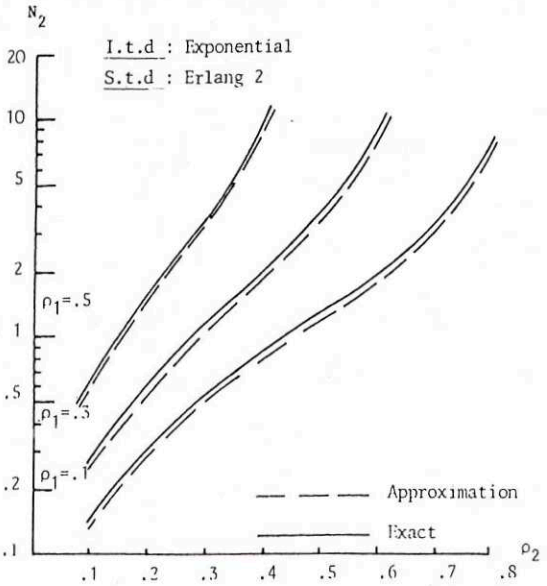

Fig. 2. P.R.P: Expected queue length of the nonpriority customers in a non-Markovian case.

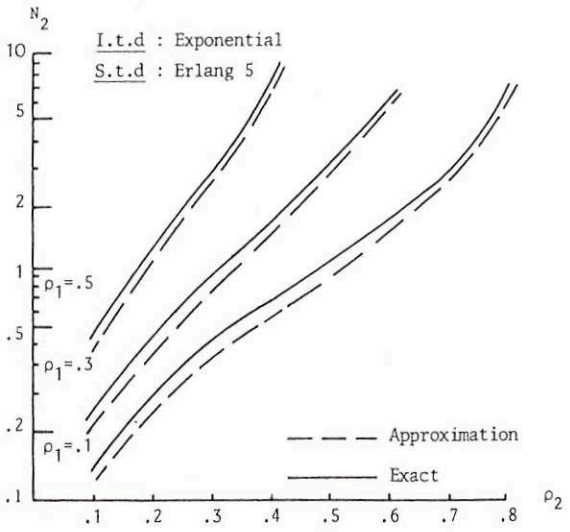

Fig. 3. P.R.P: Expected queue length of the nonpriority customers in a non-Markovian case.

process of the 2-customers has an arbitrary renewal distribution function. The approximation always remains in the confidence interval of the result obtained through a simulation model.

Tables 2 and 3 give the corresponding results for a GI/GI/1 queueing system with random or scheduled breakdowns, and comments similar to the ones presented for Table 1, can also be made as to the behavior of the approximation.

We have presented a diffusion approximation which gives accurate results for queueing systems with service interruptions. These results are very useful for such queueing systems especially when the arrival processes are not all Poisson processes, since the exact results are unknown.

The accuracy of the approximation can be summarized as follows: As the coefficients of variation of arrival processes and of service times processes become close to 1, the approximation becomes very good.

Table 1
P.R.P: Expected queue length of the nonpriority customers (I.t.d and S.t.d of the 2-customers are Erlang 2)

ρ_1 \ ρ_2		0.1	0.2	0.3	0.4	0.5	0.6	0.8
0.1	Approximation		0.275		0.682		1.531	5.568
	Simulation		0.286		0.706		1.497	5.555
			∓0.02		∓0.03		∓0.07	∓0.2
0.3			0.546	0.984		3.051	7.080	
			0.559	0.982		2.810	7.470	
			∓0.03	∓0.05		∓0.14	∓0.37	
0.5		0.557	1.444	3.160	8.200			
		0.568	1.425	3.050	8.014			
		∓0.03	∓0.07	∓0.15	∓0.4			

Table 2
Er 2/Er 2/1 queue with random breakdowns: Expected queue length ($c \overset{\text{def}}{=} i/(i+b)$)

c \ ρ		0.1	0.2	0.3	0.4	0.5	0.6
0.9	Approximation	0.115	0.241	0.386	0.567	0.817	1.214
	Simulation	0.107	0.251	0.374	0.603	0.793	1.401
		∓0.008	∓0.02	∓0.03	∓0.05	∓0.07	+0.26
0.66		0.160	0.346	0.591	0.980	1.788	
		0.161	0.333	0.633	0.975	1.915	
		∓0.01	∓0.03	∓0.07	∓0.13	∓0.39	
0.5		0.222	0.514	1.018	2.246		
		0.204	0.538	1.037	2.307		
		∓0.02	∓0.06	∓0.12	∓0.46		

Table 3
Er 2/Er 2/1 queue with scheduled breakdowns: Expected queue length

c	ρ	0.1	0.2	0.3	0.4	0.5	0.6
0.9	Approximation	0.115	0.241	0.385	0.566	0.815	1.210
	Simulation	0.110	0.233	0.416	0.538	0.813	1.233
		∓0.01	∓0.02	∓0.04	∓0.05	∓0.07	∓0.15
0.66		0.158	0.339	0.576	0.949	1.723	
		0.151	0.322	0.604	0.934	1.559	
		∓0.01	∓0.02	∓0.05	∓0.18	∓0.28	
0.5		0.214	0.487	0.949	2.229		
		0.203	0.468	0.921	1.889		
		∓0.02	∓0.04	∓0.10	∓0.25		

Moreover, the accuracy of the approximation is independent of the traffic load (see Figs. 1, 2 and 3).

The diffusion approximation has been applied to queueing systems with preemptive resume priority and random or scheduled breakdowns.

References

[1] D.R. Cox, Renewal Theory (Methuen and Co., London, 1962).

[2] M.J. Fischer, Approximation to queueing systems with interruptions, Management Sci. 24 (3) (1977) 338–344.

[3] D.P. Gaver, Diffusion approximation for certain congestion problems, J. Appl. Probab. 5 (1968) 607–623.

[4] E. Gelenbe, On approximate computer system models, J. ACM 22 (1975) 261–269.

[5] E. Gelenbe and I. Mitrani, Analysis and Synthesis of Computer Systems (Academic Press, New York, 1980).

[6] E. Gelenbe and G. Pujolle, The behaviour of a single queue in a general queueing network, Acta Inform. 7 (1976) 123–136.

[7] B. Halachimi and W.R. Franta, A diffusion approximate solution to the $G/G/k$ queueing systems, Comput. Oper. Res. 4 (1977) 37–46.

[8] D.P. Heyman, A diffusion model approximation for the GI/GI/1 queue in heavy traffic, Bell. Syst. Tech. J. (1975) 1637–1640.

[9] N.K. Jaiswal, Priority Queues (Academic Press, New York, 1968).

[10] L. Kleinrock, Queueing Systems Vol. 1 (Wiley, New York, 1975).

[11] H. Kobayashi, Application of the diffusion approximation to queueing networks, Part 1. Equilibrium queue distributions, J. ACM 21 (2) (1974) 316–328.

[12] H. Kobayashi, Application of the diffusion approximation to queueing networks, Part 2. Non-equilibrium distributions and computer modelling, J. ACM 21 (1974) 459–469.

[13] I.L. Mitrany and B. Avi-Itzhak, A many-server queue with service interruptions, Oper. Res. 16 (3) (1968) 628–638.

[14] G.F. Newell, Application of Queueing Theory (Chapman and Hall, London, 1971).

[15] M. Reiser and H. Kobayashi, Accuracy of a diffusion approximation for some queueing networks, IBM J. Res. Develop. 18 (1974) 110–124.

[16] P.D. Welch, On preemptive resume priority queues, Ann. Math. Statist. 35 (1964) 600–611.