Interdeparture times from a queueing system with preemptive resume priority

Philippe Nain

INRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(3) 954 90 20

Rapports de Recherche

N° 248

INTERDEPARTURE TIMES FROM A QUEUEING SYSTEM WITH PREEMPTIVE RESUME PRIORITY

Philippe NAIN

Octobre 1983

INTERDEPARTURE TIMES FROM A QUEUEING SYSTEM WITH PREEMPTIVE RESUME PRIORITY

Philippe NAIN
INRIA
Domaine de Voluceau - Rocquencourt
B.P. 153- 78153 LE CHESNAY CEDEX

Résumé

Nous considérons une file d'attente à capacité illimitée et à un seul serveur, qui reçoit des clients ayant des degrés de priorité différents.

Parmi les clients en file ayant la plus forte priorité, le serveur traite le plus ancien. Ce traitement est interrompu par l'arrivée d'un client plus prioritaire ; le service acquis est cependant conservé ("preemptive resume priority").

Dans l'hypothèse où les flux d'arrivée sont Poisson, nous calculons à l'état stationnaire et pour chaque classe de clients, la transformée de Laplace-Stieltjes de la loi des interdéparts.

Des résultats numériques permettent d'illustrer l'influence de cette discipline de service sur les processus de sortie. Une application aux réseaux de files d'attente est aussi donnée.

INTERDEPARTURE TIMES FROM A QUEUEING SYSTEM WITH PREEMPTIVE RESUME PRIORITY

Philippe NAIN INRIA, domaine de Voluceau, Rocquencourt, B.P. 153
78153 Le Chesnay Cédex

Keywords Queueing Systems, Output Process, Diffusion Process, Preemptive Resume Priority.

Abstract

We derive the Laplace-Stieljes transform of the limiting interdeparture times distribution for each class of customers of a queueing system with preemptive resume priority, Poisson inputs and general service times. Numerical results and an application to queueing networks are also given.

INTRODUCTION

Independently of the theoretical interest, the study of output processes has a strong practical motivation since the behavior (performance) of a queueing system is often expressed in terms of throughput.

Output processes have received much attention since the work of Burke [1] who showed that the departures from a M/M/s queueing system in equilibrium form a Poisson process. The reader will find mostly references on output processes in the papers of DALEY [2] and PACK [14].

In this study, we are concerned with the output processes of a queueing system under a preemptive resume priority.

More precisely we consider the following model : customers arrive at a service facility at r priority levels. At each priority level the input process is Poisson and these processes are mutually independent. The service times have an arbitrary distribution function which depends upon the priority level. A single server serves under a preemptive resume discipline. Results are obtained which characterize, for each class of customers, the asymptotic distribution of the interdeparture times (Section 2).

The influence of this service discipline on the interdeparture times distribution of the nonpriority customers is shown, through comparative numerical results (Section 3).

An application is given, concerning the analysis of queueing networks by means of diffusion approximation (Section 4).

This work uses previous results of WELCH [17] who studied the above processes. His study is based on a reduction of these processes to comparable processes in simple generalizations of M/G/1 queues. The results he obtained concern in particular the transient and asymptotic behavior of the size of the kth queue (k : 1,...,r) just after a departure from this queue. Such queues have also been studied by JAISWAL [9] , MILLER [11] , WHITE and CHRISTIE [18] , STEPHAN [15].

1. THE MODEL AND NOTATION

Customers arrive at a single server from r different sources. Customers arriving from source 1 have the highest priority and those arriving from source r the lowest. From each source or equivalently at each priority level, arrivals form homogeneous Poisson process with intensity λ_k and these processes are mutually independent. For $k = 1,\ldots,r$ customers with priority level k will be called k-customers. We define for $n = 1,2,\ldots$, $k = 1,\ldots,r$:

$\tau_{k,n}$ = arrival time of the nth arriving k-customers.

$\tau_{k,o} = 0$.

We assume that for fixed k, the service times process is a renewal process with an arbitrary renewal distribution which depends upon the priority level k. All the processes are supposed to be mutually independent.

We define for

$\mathrm{Re}(s) \geq 0$, $k = 1,\ldots,r$:

α_k = average service time of the k-customers (which is supposed finite).

$\phi_k(s)$ = Laplace-Stieltjes transform (LST) of the service times distribution of the k-customers.

We assume that the server is subject to a preemptive resume priority service discipline [9]. Hence, it is convenient to define the following stochastic variables for $n = 1,2,\ldots$, $k = 1,\ldots r$:

$\tau'_{k,n}$ = departure time of the nth departing k-customer.

$\xi_{k,n}$ = size of the k-queue at time $\tau'_{k,n} + 0$.

$e_{k,n}$ = beginning time of the service of the nth k-customer.

$C_{k,n} = \tau'_{k,n} - e_{k,n}$.

$C_{k,n}$ is called in the literature the "completion time" of the nth k-customer, which is defined as being the service-plus-interruption time of this customer.

$W_{k,n+1} = e_{k,n+1} - \tau_{k,n+1}$.

$W_{k,n+1}$ is the waiting time of the (n+1)th k-customer.

We first observe that because of the service discipline, the 1-customers behave exactly as customers of an M/G/1 queue with input parameter λ_1 and Laplace-Stieltjes transform of the service times distribution $\phi_1(s)$. The Laplace-Stieltjes transform of the limiting interdeparture times distribution of an M/G/1 queue can be found in TAKÁCS [16]. In the following, we will consider $k \geq 2$.

2. THE INTERDEPARTURE TIMES

In the proof of Theorem 1 we will use the following basic lemma due to WELCH [17].

LEMMA.

For $k = 2, \ldots, r$ the occupation time of the server with respect to the first $k-1$ queues at time t $(t > 0)$ is the same as the occupation time of the server at time t of an M/G/1 queue with input parameter $\Lambda_{k-1} \underline{\text{def}} \sum_{i=1}^{k-1} \lambda_i$ and Laplace-Stieltjes transform of the service times distribution $\Phi_{k-1}(s) \underline{\text{def}} \sum_{i=1}^{k-1} \lambda_i \phi_i(s) / \Lambda_{k-1}$ $(\text{Re}(s) \geq 0)$, given the same initial conditions in these two queueing systems. □

THEOREM 1

The LST for the limiting interdeparture times distribution is, for $k=2,\ldots,r$, $\mathrm{Re}(s) \geq 0$, $\sum_{i=1}^{k} \lambda_i \alpha_i < 1$:

$$\lim_{n \to \infty} E\{\exp(-s(\tau'_{k,n+1} - \tau'_{k,n}))\} = \phi_k (s + \Lambda_{k-1}(1 - \nu_{k-1}(s))).$$

$$\cdot\{1 - (\Lambda_k - \Lambda_{k-1}\nu_{k-1}(\lambda_k))(1 - \sum_{i=1}^{k} \lambda_i \alpha_i)(s + \Lambda_{k-1}(1 - \nu_{k-1}(s))) /$$

$$(\lambda_k(s + \Lambda_k - \Lambda_{k-1}\nu_{k-1}(\lambda_k + s)))\}$$

where $\nu_{k-1}(s)$ is the root with minimum absolute value of the equation,

$$z = \Phi_{k-1}(s + \Lambda_{k-1}(1-z)).$$

Proof. For $n \geq 1$, we have:

$$\tau'_{k,n+1} - \tau'_{k,n} = \begin{cases} C_{k,n+1} & \text{if } \xi_{k,n} \neq 0 \\ \theta_{k,n+1} + W_{k,n+1} + C_{k,n+1} & \text{if } \xi_{k,n} = 0 \end{cases} \qquad (1)$$

where $P(\theta_{k,n+1} \leq x) = (1 - \exp(-\lambda_k x))\, 1_{x \geq 0}$ where 1_A is the indicator function of A.

We first investigate the following LST,

$$f_{k,n+1}(s) \underline{\text{def}} E\{\exp(-s(W_{k,n+1} + \theta_{k,n+1}))\}. \qquad (2)$$

Hence,

$$f_{k,n+1}(s) = \int_0^\infty \lambda_k \exp(-(\lambda_k + s)x)\, E\{\exp(-s W_{k,n+1}) / \theta_{k,n+1} = x\}\, dx. \qquad (3)$$

Define :

$\rho_{k,n+1}$ = occupation time of the server with respect to the first k-1 queues at the time $\tau_{k,n+1}$ given $\xi_{k,n} = 0$.

Using the lemma, we have that the duration of $W_{k,n+1}$ given $\theta_{k,n+1} = x$ $(x \geq 0)$ is the duration of a busy period of an M/G/1 queue with input parameter Λ_{k-1}, with LST of the service times distribution $\Phi_{k-1}(s)$ and with an initial waiting time distribution whose the LST is $E\{\exp(-s\rho_{k,n+1})/\theta_{k,n+1} = x\}$.

TAKÁCS's result shows then that ([16], remark 4, page 63) :

$$E\{\exp(-s\,W_{k,n+1}) / \theta_{k,n+1} = x\} = E\{\exp(-(s + \Lambda_{k-1}(1 - \nu_{k-1}(s)))\,\rho_{k,n+1}) / \theta_{k,n+1} = x\} \quad (4)$$

with $\nu_{k-1}(s)$ the root with minimum absolute value of the equation, $z = \Phi_{k-1}(s + \Lambda_{k-1}(1-z))$.

Moreover, WELCH [17] shows that for $Re(\xi) \geq 0$,

$$\int_0^\infty E\{\exp(-\xi\rho_{k,n+1} / \theta_{k,n+1} = x)\}\, e^{-sx}dx = \frac{1 - \xi/(s+\Lambda_{k-1}(1-\nu_{k-1}(s)))}{s - \xi + \Lambda_{k-1}(1 - \Phi_{k-1}(\xi))} \cdot \quad (5)$$

From (2),(3),(4),(5), it follows that $\forall n \geq 1$, $\forall k \geq 2$, $\forall Re(s) \geq 0$,

$$f_{k,n+1}(s) = (\lambda_k + \Lambda_{k-1}(\nu_{k-1}(s) - \nu_{k-1}(\lambda_k+s))) / (\Lambda_k+s - \Lambda_{k-1}\nu_{k-1}(s+\lambda_k)) \cdot \quad (6)$$

We conclude the proof using the two following arguments :

i. $(\theta_{k,n+1} + W_{k,n+1})$ and $C_{k,n+1}$ are two independent stochastic variables because of the arrival Poisson Processes and the independence hypothesis of the input and service times renewal processes in queues $1,\ldots,k$.

ii. $\forall n \geq 1$, $\forall Re(s) \geq 0$, $\forall k = 2,\ldots,r$,

$$E\{\exp(-s\, C_{k,n+1})\} = \phi_k(s + \Lambda_{k-1}(1 - \nu_{k-1}(s))).$$

This result can be found in [9] for example.

Hence, from 1,2,6,i,ii, we obtain ,

$$\lim_{n\to\infty} E\{\exp(-s\ (\tau'_{k,n+1} - \tau'_{k,n})\} = \lim_{n\to\infty} P(\xi_{k,n} \neq 0)\ \phi_k(s + \Lambda_{k-1}(1 - \nu_{k-1}(s)) +$$

$$\lim_{n\to\infty} P(\xi_{k,n} = 0)\ \phi_k(s + \Lambda_{k-1}(1 - \nu_{k-1}(s)))\ (\lambda_k + \Lambda_{k-1}(\nu_{k-1}(s) - \nu_{k-1}(\lambda_k + s)))$$

$$/\ (\Lambda_k + s - \Lambda_{k-1}\ \nu_{k-1}\ (s + \lambda_k)). \qquad (7)$$

It remains to determine, $\lim_{n\to\infty} P(\xi_{k,n} = 0)$.

This can be done using the theorem in the Appendix A with $z = 0$.

We find,

$$\lim_{n\to\infty} P(\xi_{k,n} = 0) = (\Lambda_k - \Lambda_{k-1}\ \nu_{k-1}\ (\lambda_k))\ (1 - \sum_{i=1}^{k} \lambda_i \alpha_i)\ /\ \lambda_k.$$

The proof is then concluded introducing the above result in (7).

3. NUMERICAL RESULTS

We know that the mean interdeparture times of the k-customers is λ_k^{-1} for $k = 1,2,\ldots,r$. In what follows we are therefore only eoncerned with : quantities related to the variance of the interdeparture processes.

In order to get numerical results comparable with known analytic results, we consider the case where all the service times are exponentially

distributed. The queue with the 1-customers is then simply a M/M/1 queue with input parameter λ_1 and service rate α_1^{-1}.

Let V_2 be the variance of the stationary distribution of the inter-departure times of the nonpriority customers (the 2-customers).

Let s be the ratio $s = \left(\frac{CCV_D - CCV_A}{CCV_A}\right) \times 100$, where CCV_A and CCV_D are respectively the squared coefficients of variation of the interarrival and interdeparture times distributions of the 2-customers. Due to Poisson inputs $CCV_A = 1$ and s is then simply reduced to $s = (\lambda_2^2 V_2 - 1) \times 100$.

For three given levels of priority traffic ρ_1, Table 1 gives s versus the nonpriority traffic ρ_2

$$(\rho_k \overset{def}{=} \lambda_k \alpha_k, \quad k = 1,2)$$

For small values of ρ_1 and ρ_2, Table 1 shows that the departure process of the nonpriority customers remains "close" to a Poisson process (s small) with intensity λ_2. In other words, this indicates that under this priority rule and for small values of ρ_1, ρ_2, the queue of the 2-customers "almost" behaves as a M/M/1 queue with input parameter λ_2 and service rate α_2^{-1}, which would be exactly the case if $\lambda_1 = 0$ [1]. On the other hand, as soon as ρ_1 and/or ρ_2 increase then the departure process of the nonpriority customers is no longer close to a Poisson process (s increases).

4. APPLICATION TO QUEUEING NETWORKS AND LIMITATIONS

An application of Theorem 1 arises in the approximation of queueing networks by means of diffusion processes. First, let us briefly recall the method of diffusion approximations in queueing theory.

The idea is to approximate a process which is not time-continuous - for instance the number of jobs - by a time-continuous process - a diffusion process - according to the central limit theorem.

This technique has originated in queueing theory with the work of GAVER [3] and NEWELL [12] for a single server queue and generalised to queueing networks by KOBAYASHI [10] and GELENBE, PUJOLLE [6]. Many authors applied this approximation method for various queueing systems, for which exact results are unknown or not readily usable GELENBE [4], HEYMAN [8] , HALACHIMI and FRANTA [4] , NAIN [13] and many others.

The method involves the choice of two diffusion parameters b and α, respectively called the drift and instantaneous variance, and which characterise the diffusion process in a unique way, given the initial conditions. These diffusion parameters must be chosen in order that the corresponding diffusion process reflects the particular queueing system under consideration.

For a single server queue, b and α are functions of the mean and variance of the interarrival times and service times distributions (for a full treatment see GELENBE, MITRANI [5]). Thus, if we consider a queueing network (open or not) where node i_o consists of the queueing system investigated in this paper (all arrivals to node i_o - external or not - are supposed Poissonean), we know from Theorem 1, the mean and variance of the interarrival times distribution of the k-customers (k=1,...,r) arriving to node j if $p_{i_o,k,j,.} \neq 0$, (for $i_o \neq j$ or $i_o = j$ and k=1 and the service times of the 1-customers exponentially distributed) where $p_{i_o,k,j,.} \stackrel{def}{=}$ probability that a k-customer having terminated its service at node i_o enters node j.

Then, making the "usual" assumptions (in the context of diffusion approximation in queueing theory) that the departure processes from node i_o are all renewal processes and mutually independent, we can extend the method of diffusion approximations to queueing networks containing nodes of type i_o.

However this application is limited to particular queueing networks since a node of type i_o cannot belong to general closed or open queueing networks. For example, a node of type i_o cannot be visited more than one time by a job otherwise inputs into this node would not form a Poisson process. This application is valid for instance if all arrivals to nodes of type i_o are external and Poisson.

APPENDIX A

THEOREM. (WELCH [17])

For $k \geq 2$ and if $\sum_{i=1}^{k} \lambda_i \alpha_i < 1$, the Markov chain $\{\xi_{k,n} : n = 1,2,\ldots\}$ is ergodic and, independent of the initial distribution, we have :

$$\sum_{j=0}^{\infty} \lim_{n\to\infty} P(\xi_{k,n}=j)\, z^j = \emptyset\,(\lambda_k(1-z))\;(\lambda_k z - \Lambda_k + \Lambda_{k-1}\,\nu_{k-1}(\lambda_k(1-z)))\cdot$$

$$\cdot\,(1 - \sum_{i=1}^{k} \lambda_i\alpha_i) \;/\; (\lambda_k(z - \emptyset\,(\lambda_k(1-z)))) \text{ where}$$

$\emptyset(s) = \emptyset_k\,(s + \Lambda_{k-1}\,(1 - \nu_{k-1}(s)))$ and $\nu_{k-1}(s)$ is the root with minimum absolute value of the equation, $z = \Phi_{k-1}\,(s + \Lambda_{k-1}\,(1-z))$.

APPENDIX B

Let V_k be the variance of the interdeparture times distribution of the k-customers, k=2,...,r.

Then :

$$V_k = \frac{\partial^2}{\partial s^2} \lim_{n\to\infty} E\{\exp(-s(\tau'_{k,n+1} - \tau'_{k,n}))\}\Big|_{s=0} - 1/\lambda_k^2 \quad . \quad (\mathrm{Re}\ s \geq 0)$$

A straightforward but tedius computation yields :

$$V_k = F_k^2 - a_k(2\, F_k^1\, G_k^1 + G_k^2)/\lambda_k - 1/\lambda_k^2$$

where :

$$F_k^1 = \frac{-1}{\mu_k(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)} ,$$

$$F_k^2 = \frac{\sum_{i=1}^{k-1} \lambda_i\, E(S_i^2)}{\mu_k(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)^3} + \frac{E(S_k^2)}{(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)^2}$$

$$G_k^1 = \frac{1}{(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)A_k} ,$$

$$G_k^2 = \frac{-\sum_{i=1}^{k-1} \lambda_i\, E(S_i^2)}{(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)^3\, A_k} - 2\,\frac{(1 - \Lambda_{k-1}\nu'_{k-1}(\lambda_k))}{(1 - \sum_{i=1}^{k-1} \lambda_i\alpha_i)\, A_k^2} ,$$

$A_k = \Lambda_k - \Lambda_{k-1}\nu_{k-1}(\lambda_k)$ with $\nu_{k-1}(\lambda_k)$ given in Theorem 1,

$$a_k = A_k(1 - \sum_{i=1}^{k} \lambda_i\alpha_i) ,$$

$$\nu'_{k-1}(\lambda_k) = \frac{d}{ds}\,\nu_{k-1}(s)\Big|_{s=\lambda_k} = \frac{-\sum_{i=1}^{k-1} \lambda_i \int_o^\infty x\, e^{-A_k x}\, dB_i(x)}{\Lambda_{k-1}(1 - \sum_{i=1}^{k-1} \lambda_i \int_o^\infty x\, e^{-A_k x}\, dB_i(x))}$$

$B_i(.)$ = service times distribution of the i-customers ,

$$E(S_i^2) = \int_o^\infty x^2\, d\, B_i(x), \quad i=1,\ldots,k .$$

ρ_2 \ ρ_1		.1	.3	.5
.1	s =	0,6	5,37	24,0
.2		2,4	15,2	64,4
.3		4,8	28,6	108,8
.4		7,3	42,3	154
.5		10,2	56,5	
.6		13,3	71,4	
.7		16,1		
.8		19,2		

Table 1 : Comparison between the input process and the output process of the non-priority units ($\mu_1 = \mu_2 = 10$)

REFERENCES

[1] P. J. Burke, the Output of a Queueing System, Oper. Res. 4 (1956) 699-704.

[2] D. J. Daley, Notes on Queueing Output Processes, Proceedings of the Conference on Mathematical Methods in Queueing Theory at Western Michigan University (Springer Verlag, New York, 1973).

[3] D. P. Gaver, Diffusion approximation for certain congestion problems, J. Appl. Proba. 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] H. Kobayashi, Application of the diffusion approximation to queueing networks, Part 1. Equilibrium queue distributions, J. ACM 21 (2) (1974) 316-328.

[11] R. G. Miller, Priority Queues, Ann. Math. Statist. 31 (1960) 86-103.

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

[13] P. Nain, Queueing system with service interruptions : an approximation model, Performance Evaluation 3 (1983) 123-129.

[14] C. D. Pack, Output of Multiserver Queueing Systems, Oper. Res. 26 (1978) 492-509.

[15] F. F. Stephan, Two Queues under Preemptive Priority with Poisson Arrival and Service Rates, J. Oper. Res. Soc. Am. 6 (1958) 399-418.

[16] L. Takács, Introduction to the theory of queues (Oxford Univ. Press, 1962).

[17] P. D. Welch, On Preemptive Resume Priority Queues, Ann. Math. Statist. 35 (1964) 600-611.

[18] H. White and L. S. Christie, Queueing with Preemptive Priorities or with Breakdown, J. Oper. Res. Soc. Am. 6 (1958) 79-95.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique