Benes's Formula for M/G/1-FIFO 'Explained' by Preemptive-Resume LIFO

Robert B. Cooper; Shun-Chen Niu

Journal of Applied Probability, Vol. 23, No. 2 (Jun., 1986), 550-554.

Stable URL:
http://links.jstor.org/sici?sici=0021-9002%28198606%2923%3A2%3C550%3ABFFM%27B%3E2.0.CO%3B2-J

Journal of Applied Probability is currently published by Applied Probability Trust.

J. Appl. Prob. **23**, 550–554 (1986)
Printed in Israel

BENEŠ'S FORMULA FOR $M/G/1$–FIFO 'EXPLAINED' BY PREEMPTIVE-RESUME LIFO

ROBERT B. COOPER,* *Florida Atlantic University*
SHUN-CHEN NIU,** *The University of Texas at Dallas*

Abstract

We provide a term-by-term interpretation of Beneš's well-known but mysterious inversion of the Pollaczek–Khintchine formula. The strategy is to recognize the equality of waiting time in $M/G/1$–FIFO with remaining work in $M/G/1$–LIFO–preemptive resume. In the process, we give a new and simple derivation of some known results for $M/G/1$–LIFO–preemptive resume.

POLLACZEK-KHINTCHINE FORMULA

Introduction

Beneš (1957) has shown that the celebrated Pollaczek–Khintchine formula, which gives the Laplace–Stieltjes transform of the equilibrium waiting-time distribution function $W(\cdot)$ for the $M/G/1$ queue with service in order of arrival, can be inverted to give the explicit formula

$$W(t)=\sum_{j=0}^{\infty}(1-\rho)\rho^{j}\tilde{H}^{*j}(t), \tag{1}$$

where $\rho<1$ is the utilization of the server and $\tilde{H}^{*j}(\cdot)$ is the j-fold self-convolution of the distribution function $\tilde{H}(\cdot)$,

$$\tilde{H}(t)=\frac{1}{\tau}\int_{0}^{t}(1-H(x))dx, \tag{2}$$

where $H(\cdot)$ is the service-time distribution function and τ is the mean service time. $\tilde{H}(\cdot)$ can be interpreted as describing the forward recurrence time of the random variable whose distribution function is $H(\cdot)$; if we denote their Laplace–Stieltjes transforms by $\tilde{\eta}(\cdot)$ and $\eta(\cdot)$, respectively, then (2) is

Received 6 August 1984; revision received 3 May 1985.

* Postal address: College of Business and Public Administration, Computer and Information Systems Department, Florida Atlantic University, P.O. Box 3091, Boca Raton, FL 33431–0991, USA.

** Postal address: School of Management and Administration. The University of Texas at Dallas, Box 830688, Richardson, TX 75083–0688, USA.

equivalent to

(3) $$\tilde{\eta}(s)=\frac{1}{\tau}\frac{1-\eta(s)}{s}.$$

It follows from the theorem of total probability that

(4) $$W(t)=\sum_{j=0}^{\infty}\Pr\{N=j\}\Pr\{W\leqq t \mid N=j\},$$

where N is the number of customers present when an arbitrary customer (the *test customer*, whose viewpoint we adopt) arrives, and W is the test customer's waiting time. When the service times are exponentially distributed ($M/M/1$), then

(5) $$\Pr\{N=j\}=(1-\rho)\rho^j \qquad (j=0,1,2,\cdots)$$

and

(6) $$\Pr\{W\leqq t \mid N=j\}=\tilde{H}^{*j}(t).$$

However, in the general case ($M/G/1$), neither (5) nor (6) remains valid, although (1), of course, does.

These facts are all well known, and are discussed in detail in the textbooks by Kleinrock (1975) and Cooper (1981). Kleinrock (1975), p. 201, calls the result (1) 'most intriguing', and goes on to say, 'Tempting as it is to try to give a physical explanation for the simplicity of this result [for $M/G/1$] and its relation to $M/M/1$, no satisfactory, intuitive explanation has been found to explain this dramatic form.' Cooper (1981), p. 219, (if one may quote oneself) calls (1) 'mysterious', and observes that '... despite its simple form, [Equation (1)] does not seem to yield to similar [to $M/M/1$] term-by-term interpretation in the general case. This provides a dramatic counterexample to the folk theorem (often true) that simple results have simple explanations.' Gross and Harris (1985), p. 275, also comment on the surprising form of (1).

The purpose of this note is to provide a term-by-term interpretation of (1); in the process we provide a new and simple derivation of some known facts about the equilibrium preemptive-resume $M/G/1$ queue with service in reverse order of arrival (an arriving customer preempts the customer, if any, in service; when the preempted customer reenters service, his service time continues from the point of interruption).

Strategy

Let us call the ordinary $M/G/1$ queue with service in order of arrival, $M/G/1$–FIFO; and the preemptive-resume $M/G/1$ queue with service in reverse order of arrival, $M/G/1$–LIFO–PRes. Imagine that a 'copy' is made of the arrival times and service times that characterize the customers in the system $M/G/1$–FIFO; and suppose this copy is submitted to the system $M/G/1$–LIFO–PRes. Then, both systems have identical busy periods. Define the *remaining work* at time t_0 as the total amount of elapsed time from t_0 until there would be no more work for the server, assuming that any customer who arrives at $t \geqq t_0$ were assigned zero service time. Therefore, if t_0 is an arrival epoch of a test customer, the waiting time of the test customer in $M/G/1$–FIFO is identical with the remaining work at t_0 in $M/G/1$–LIFO–PRes. Our strategy, then, is to calculate the remainining work in $M/G/1$–LIFO–PRes at an arrival epoch t_0, and to show that its distribution function (1) follows directly from (4), (5), and (6), where W is now interpreted as remaining work. That is, we show that in $M/G/1$–LIFO–PRes (i) the equilibirum distribution of the number of customers present just prior to an arrival epoch t_0 is given by (5), and (ii) the remaining service times of the customers present at t_0^- are mutually independent, each having the distribution function (2).

Proofs of assertions (i) and (ii) can be found buried in the literature on insensitive generalized semi-Markov processes (see, e.g., Schassberger (1978), Burman (1981), and Niu (1983)). Kelly (1976) has obtained these results by an argument similar to ours, but considerably more complicated; his primary interest was in proving that the output process from the equilibrium system $M/G/1$–LIFO–PRes is a Poisson process. (However, see Exercise 7, p. 81 of Kelly (1979), where the interpretation of Beneš's formula is stated explicitly.) Prabhu (1980) has interpreted Beneš's formula in terms of ladder processes. And Fakinos (1981) has shown how these results can be extended to $G/G/1$, including a similar term-by-term interpretation of the analogue of (1). Our contribution is to give a simple, direct argument for $M/G/1$.

Analysis of $M/G/1$–LIFO–PRes

Observe that in the system $M/G/1$–LIFO–PRes, no customer is affected in any way by any customer who arrives before him. Also, every customer who is waiting in the queue has necessarily been preempted from the server at least once; and, with respect to remaining service times, the only relevant information is the number of times a customer has been preempted. Therefore, conditional on a test customer arriving at time t_0 and finding $N = j > 0$ other customers present, the customers who are waiting in the queue at t_0^+ all have service times whose remaining durations are identically distributed and independent of each other and the queue size, with distribution function $\hat{H}(\cdot)$, say, with Laplace–Stieltjes transform $\hat{\eta}(\cdot)$.

Let $\Pi_j = \Pr\{N = j\}$ be the equilibrium probability that there are j customers present just prior to an arbitrary arrival epoch; and let Π_j^* be the equilibrium probability that there are j customers present just after an arbitrary departure epoch. Then, if the test customer (who enters service immediately upon arrival at t_0 but requires zero service time) finds j other customers present, the previous event was either (a) an arrival who found $j-1$ other customers present, in which case the test customer arrives during this previous arrival's (ordinary) service time, or (b) a departure who left behind j other customers, in which case the test customer arrives during the remaining service time of the customer who reentered service at this departure epoch. That is,

$$\Pi_j = \Pi_{j-1} \int_0^\infty (1-\exp(-\lambda x))dH(x) + \Pi_j^* \int_0^\infty (1-\exp(-\lambda x))d\hat{H}(x) \qquad (j=1,2,\cdots), \tag{7}$$

where λ is the arrival rate.

Since each departure leaves behind him exactly those customers who were present when he arrived, therefore

$$\Pi_j = \Pi_j^* \qquad (j=0,1,2,\cdots). \tag{8}$$

Substitution of (8) into (7) yields

$$\Pi_j = \frac{1-\eta(\lambda)}{\hat{\eta}(\lambda)} \Pi_{j-1} \qquad (j=1,2,\cdots), \tag{9}$$

where $\eta(\cdot)$ and $\hat{\eta}(\cdot)$ are the Laplace–Stieltjes transforms characterizing an ordinary and a remaining service time, respectively. It follows from (9) that

$$\Pi_j = \left(\frac{1-\eta(\lambda)}{\hat{\eta}(\lambda)}\right)^j \Pi_0 \qquad (j=1,2,\cdots). \tag{10}$$

But Π_0 equals the fraction of time that the server is idle (because Poisson arrivals see time averages — see Wolff (1982)); hence

$$\Pi_0 = 1-\rho. \tag{11}$$

Equations (10) and (11), together with the normalization requirement $\Sigma \Pi_j = 1$, imply

$$\frac{1-\eta(\lambda)}{\hat{\eta}(\lambda)} = \rho; \tag{12}$$

and (10), (11), and (12) imply (5).

Since $\rho = \lambda\tau$, it follows from (12) that

$$\hat{\eta}(\lambda) = \frac{1}{\tau}\frac{1-\eta(\lambda)}{\lambda} \tag{13}$$

for $0 < \lambda < 1/\tau$, and therefore, by analytic continuation, for all $0 \leqq \lambda < \infty$. Comparison of (13) and (3) shows that $\hat{\eta}(\cdot) = \tilde{\eta}(\cdot)$. Hence, $\hat{H}(\cdot) = \tilde{H}(\cdot)$, given by (2), and the proof is complete.

Comments

Equation (5) shows that, remarkably, the equilibrium distribution of N in $M/G/1$–LIFO–PRes is insensitive to the form of the service-time distribution function, depending only on its mean, and is identical with the equilibrium distribution of N in $M/M/1$. It follows that mean queue length (and mean waiting time) is smaller (greater) in $M/G/1$–FIFO than in $M/G/1$–LIFO–PRes when service times are less (more) variable than the exponential distribution. It is also true (and easy to show) that the mean response time (total time spent waiting and in service) for a customer in $M/G/1$–LIFO–PRes is proportional to the amount of service time required by the customer. These properties are shared by the $M/G/1$ processor-sharing model. (See, for example Yashkov (1983) and Ott (1984).)

References

BENEŠ, V. E. (1957) On queues with Poisson arrivals. *Ann. Math. Statist.* **28**, 670–677.

BURMAN, D. Y. (1981) Insensitivity in queueing systems. *Adv. Appl. Prob.* **13**, 846–859.

COOPER, R. B. (1981) *Introduction to Queueing Theory*, 2nd edn. Elsevier North-Holland, New York.

FAKINOS, D. (1981) The $G/G/1$ queueing system with a particular queue discipline. *J. R. Statist. Soc.* B **43**, 190–196.

GROSS, D. AND HARRIS, C. M. (1985) *Fundamentals of Queueing Theory*, 2nd edn. Wiley, New York.

KELLY, F. P. (1976) The departure process from a queueing system. *Math. Proc. Camb. Phil. Soc.* **80**, 283–285.

KELLY, F. P. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.

KLEINROCK, L. (1975) *Queueing Systems*, Vol. 1. Wiley, New York.

NIU, S. C. (1983) Equilibrium analysis of insensitive generalized semi-Markov processes: a uniformization approach. Unpublished.

OTT, T. J. (1984) The sojourn-time distribution in the $M/G/1$ queue with processor sharing. *J. Appl. Prob.* **21**, 360–378.

PRABHU, N. U. (1980) *Stochastic Storage Processes: Queues, Insurance Risk, and Dams.* Springer-Verlag, New York.

SCHASSBERGER, R. (1978) Insensitivity of steady-state distribution of generalized semi-Markov processes with speeds. *Adv. Appl. Prob.* **10**, 836–851.

WOLFF, R. W. (1982) Poisson arrivals see time averages. *Operat. Res.* **30**, 223–231.

YASHKOV, S. F. (1983) A derivation of response time distribution for an $M/G/1$ processor-sharing queue. *Prob. Control. Inf. Theory* **12**, 133–148.