The M/G/1 FIFO queue with several customer classes

The $M/G/1$ FIFO queue with several customer classes

Onno J. Boxma *† and Tetsuya Takine ‡

July 14, 2003

Abstract

In this note we present short derivations of the joint queue length distribution in the $M/G/1$ queue with several classes of customers and FIFO service discipline.

Keywords: single server queue, several customer classes, FIFO, joint queue length distribution.

1 Introduction

Takine [3] derives the joint queue length distribution in a class of FIFO single server queues with multiple, possibly correlated, non-Poissonian arrival streams, where the service time distributions of customers may be different for different streams. He observes (pp. 350-351): "When the arrival streams follow independent Poisson processes, the joint probability generating function of the stationary distribution of the number of customers from each stream is given by a simple formula, which should be found somewhere, even though we could not find it in the literature." The purpose of the present note is threefold: (i) point at an unpublished report of Wallström [4] which contains that simple formula for the generating function (GF) of the joint distribution of numbers of customers; (ii) present short and elementary proofs of that formula; (iii) correct a minor error in [3] which results in a wrong expression for this GF, thus properly documenting this basic result for a classical queue.

2 Model description

We consider a single server system with K customer classes, which are served without priorities according to the FIFO discipline. Customers of class i arrive according to a Poisson process with rate λ_i, and require a service time B_i with distribution $B_i(\cdot)$, having Laplace-Stieltjes transform (LST) $\beta_i\{\cdot\}$, with mean $\mathsf{E}B_i$. The total arrival rate is $\lambda := \sum_{i=1}^{K} \lambda_i$. All arrival intervals and service times are independent. The traffic load of class i is denoted by $\rho_i := \lambda_i \mathsf{E}B_i$, and the total traffic load by $\rho := \sum_{i=1}^{K} \rho_i$. The common buffer has infinite capacity. Hence $\rho < 1$ is a necessary and sufficient condition for the existence of steady-state

*Department of Mathematics & Computer Science and EURANDOM; Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands;

†CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

‡Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan.

distributions of performance measures like waiting times and queue lengths. Since the arrival processes are Poissonian, all customer classes have the same waiting time distribution. Denote the steady-state waiting time by W, the vector of the steady-state numbers of *waiting* customers by $(X_1, \ldots, X_K)$ and the vector of the steady-state numbers of customers in the system (including the one possibly in service) by $(Y_1, \ldots, Y_K)$.

3 The queue length distribution

In the sequel, let $L := \sum_{i=1}^{K} \lambda_i(1 - z_i)$ and $|z_j| \leq 1$, $j = 1, \ldots, K$. Wallström [4] proves the following result:

Theorem 3.1.

$$\mathsf{E}[z_1^{Y_1} \ldots z_K^{Y_K}] \; = \; (1 - \rho)\frac{\sum_{i=1}^{K} \lambda_i(1 - z_i)\beta_i\{L\}}{\sum_{i=1}^{K} \lambda_i[\beta_i\{L\} - z_i]}. \tag{3.1}$$

Wallström obtains this result by considering a busy period that starts with the service of a class-i customer. He distinguishes successive generations in this busy period: generation 0 is the first service time, and generation j consists of the service times of all customers arriving during generation $j-1$. He conditions the queue length vector at some arbitrary time t during the j-th generation of the busy period on the numbers of arrivals in the $(j - 1)$-st generation, and on the numbers of departures in the j-th generation prior to t. By deconditioning he obtains (3.1) after a lengthy calculation.

Below we present two elementary proofs of the above theorem. Starting point in both proofs is the observation that, if the total number X of *waiting* customers equals n, then the vector $(X_1, \ldots, X_K)$ of *waiting* customers is multinomially distributed with parameters $(n, \frac{\lambda_1}{\lambda}, \ldots, \frac{\lambda_K}{\lambda})$ (a similar statement for the vector of total numbers of customers is *not* true in general). Hence

$$\mathsf{E}[z_1^{X_1} \ldots z_K^{X_K} | X = n] \; = \; (\frac{\lambda_1}{\lambda} z_1 + \cdots + \frac{\lambda_K}{\lambda} z_K)^n \; = \; (1 - \frac{L}{\lambda})^n,$$

resulting in

$$\mathsf{E}[z_1^{X_1} \ldots z_K^{X_K}] \; = \; \mathsf{E}[(1 - \frac{L}{\lambda})^X].$$

An application of the distributional form of Little's law [2] for the total number of waiting customers in an $M/G/1$ queue, viz., $\mathsf{E}[z^X] = \mathsf{E}[\exp(-\lambda(1-z))W]$, yields, in combination with the Pollaczek-Khintchine formula for the LST of the waiting time distribution (cf. Cohen [1], p. 255):

$$\mathsf{E}[z_1^{X_1} \ldots z_K^{X_K}] \; = \; \mathsf{E}[\mathrm{e}^{-LW}] \; = \; (1 - \rho)\frac{L}{\sum_{i=1}^{K} \lambda_i[\beta_i\{L\} - z_i]}. \tag{3.2}$$

Proof 1. This proof is based on the concept of *attained waiting time*, viz., the time the customer presently in service (if any) has already spent in the system. Denote by $W_{A,i}$ the steady-state attained waiting time of a class-i customer in service. Since all waiting customers have arrived during the attained waiting time, an argument similar to the one leading to (3.2) yields:

$$\mathsf{E}[z_1^{Y_1} \ldots z_K^{Y_K}] \; = \; (1 - \rho) + \sum_{i=1}^{K} \rho_i z_i \mathsf{E}[\mathrm{e}^{-LW_{A,i}}]. \tag{3.3}$$

Observe that $W_{A,i}$ equals in distribution the sum of two independent terms: The steady-state waiting time and the past part of the service of the class-i customer. Hence

$$\mathsf{E}[z_1^{Y_1} \dots z_K^{Y_K}] \;=\; (1-\rho) + \sum_{i=1}^{K} \rho_i z_i \mathsf{E}[\mathrm{e}^{-LW}] \frac{1-\beta_i(L)}{L\mathsf{E}B_i}. \tag{3.4}$$

The theorem follows after substitution of (3.2) in (3.4).

Proof 2. The following proof is based on relations between queue lengths and numbers of waiting customers, at various epochs. Using (3.2) one may rewrite Theorem 3.1 as

$$\mathsf{E}[z_1^{Y_1} \dots z_K^{Y_k}] \;=\; \mathsf{E}[z_1^{X_1} \dots z_K^{X_k}] \frac{\sum_{i=1}^{K} \lambda_i (1-z_i) \beta_i\{L\}}{L}. \tag{3.5}$$

We sketch a direct proof of this formula. Indicating by ST (ST_i) the event that a service of any class (of class-i) has just started, by D (D_i) the event that a customer (of class i) has just left, and by A (A_i) the event that an arrival (of class i) is about to take place, we can write:

$$\begin{aligned} \mathsf{E}[z_1^{X_1} \dots z_K^{X_K}] \;&=^{(1)}\; \mathsf{E}[z_1^{X_1} \dots z_K^{X_K} | A] =^{(2)} \mathsf{E}[z_1^{X_1} \dots z_K^{X_K} | D] \\ &=^{(3)}\; \mathsf{E}[z_1^{X_1} \dots z_K^{X_K} | ST] =^{(4)} \mathsf{E}[z_1^{X_1} \dots z_K^{X_K} | ST_i]. \end{aligned} \tag{3.6}$$

(1) follows from PASTA. (2): Burke's level crossing argument holds for the *total* number of waiting customers right before an arrival and right after a departure, and then once more apply the multinomial argument for the vectors of numbers of waiting customers. (3) is trivial when a departure coincides with a service start; and if there were no waiting customers after a departure, then there are also no waiting customers right after the next service starts. (4) follows from the facts that all waiting customers just after the start of a class-i service arrived during the waiting time of the class-i customer and all customer classes have the same waiting time distribution. Finally notice that $\mathsf{E}[z_1^{Y_1} \dots z_K^{Y_k} | D_i] = \beta_i\{L\} \mathsf{E}[z_1^{X_1} \dots z_K^{X_k} | ST_i]$. Multiplying both sides of (3.5) by L, and applying PASTA (for an arbitrary customer or a class-i customer, that does not matter) to the lefthand side, it remains to prove:

$$\sum_{i=1}^{K} \lambda_i (1-z_i) \mathsf{E}[z_1^{Y_1} \dots z_K^{Y_k} | A_i] \;=\; \sum_{i=1}^{K} \lambda_i (1-z_i) \mathsf{E}[z_1^{Y_1} \dots z_K^{Y_k} | D_i]. \tag{3.7}$$

Ignoring terms in which indices j_m become negative, this formula states that

$$\begin{aligned} &\sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \dots, Y_i = j_i, \dots, Y_K = j_K | A_i) \\ &\qquad\qquad - \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \dots, Y_i = j_i - 1, \dots, Y_K = j_K | A_i) \\ &\quad = \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \dots, Y_i = j_i, \dots, Y_K = j_K | D_i) \\ &\qquad\qquad - \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \dots, Y_i = j_i - 1, \dots, Y_K = j_K | D_i), \end{aligned}$$

or

$$\sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \ldots, Y_i = j_i, \ldots, Y_K = j_K | A_i)$$

$$+ \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \ldots, Y_i = j_i - 1, \ldots, Y_K = j_K | D_i)$$

$$= \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \ldots, Y_i = j_i - 1, \ldots, Y_K = j_K | A_i)$$

$$+ \sum_{i=1}^{K} \lambda_i \mathsf{P}(Y_1 = j_1, \ldots, Y_i = j_i, \ldots, Y_K = j_K | D_i). \tag{3.8}$$

Observe that this formula states a global balance equation for the state $(j_1, \ldots, j_i, \ldots, j_K)$, equating rates to leave respectively enter that state.

Remark 3.1. Theorem 1 in [3] contains the same result as Theorem 3.1 for a much more general multiclass model. When specializing this theorem to the case of independent Poisson streams in Formula (21) of [3], a minor error occurs.

Remark 3.2. As observed by Wallström [4], classical $M/G/1$ results are obtained when one chooses either $z_j = 1$, $\forall j \neq i$, or $z_j = z$, $\forall j$.

Acknowledgment. The authors thank Sem Borst and Ho Woo Lee for interesting discussions.

References

[1] COHEN, J.W. *The Single Server Queue*. North-Holland, Amsterdam, 1982.

[2] KEILSON, J. AND L.D. SERVI. The distributional form of Little's law and the Fuhrmann-Cooper decomposition. *O.R. Letters 9* (1990), 239-247.

[3] TAKINE, T. Queue length distribution in a FIFO single-server queue with multiple arrival streams having different service time distributions. *Queueing Systems 39* (2001), 349-375.

[4] WALLSTRÖM, B. On the $M/G/1$ queue with several classes of customers having different service time distributions. Report 1-19, Lund Institute of Technology, 1980.