Stochastically Ordered Synchronized Queues

Michael A. Zazanis*
Department of Statistics
Athens University of Economics and Business
76 Patision str. Athens, 104 34 Greece

Abstract

We consider a synchronized queueing system in which customers arrive according to a Poisson process to a station consisting of c parallel servers, each with its own queue. Upon arrival, customers split into parts and each part joins the corresponding queue. Under the assumption that the service requirements of the c parts of each customer are strongly ordered we obtain the joint distribution of the workloads.

KEYWORDS: FORK-JOIN QUEUES, STOCHASTICALLY ORDERED QUEUES, RATE CONSERVATION PRINCIPLE

1 Introduction

Synchronized (or fork–join) queues have been an object of study over the last three decades as models of parallel processing. The simplest model consists of c parallel processors, each with its own queue. Customers, upon arrival, break into c sub-entities which we will call parts. Part i requires service from server i, where $i = 1, 2, \ldots, c$, and, if necessary, joins the corresponding queue which is assumed to have unlimited capacity and operate under a FIFO discipline. While each station viewed separately is an ordinary single server queue, the joint statistics of the c queues are typically not easy to obtain.

The above system when service requirements for the parts are independent, exponential random variables, identically distributed for each type of part, is known as the Flatto-Hahn-Wright model (see [5], [4], [11]). In this case, while each queue considered separately is an ordinary M/M/1 queue, determining the joint distribution is far from easy. Flatto and Hahn [4] have studied this system (for the case $c = 2$) using complex analysis techniques. See also Fayolle, Iasnogorodsky, and Malyshev [6]. Asymptotic results regarding this model have been obtained using large deviation techniques by Weiss and Shwartz [10]. We also mention

*email: zazanis@aueb.gr

Baccelli, Makowski, and Swhartz [2] where bounds for the performance of more general fork–join queues are obtained by means of stochastic ordering arguments.

Our approach to this problem makes use of Miyazawa's Rate Conservation Principle (see [1], [8]) in order to obtain effortlessly an expression for the joint Laplace transform of the stationary workload. This expression depends on unknown functions which, in general, are not easily determined. In this paper we examine the case where the service times of parts are strongly ordered.

2 The rate conservation principle

On the probability space $(\Omega, \mathscr{F}, P)$ a point process $\{T_n\}$ has been defined which we will assume to be a stationary Poisson process with rate λ. We will denote by P^0 the Palm transformation of P with respect to $\{T_n\}$ and by E^0 expectation with respect to P^0 as usual. For background on Palm theory we refer the reader to [1].

The Poisson process $\{T_n\}$ is assumed to feed c queues in parallel. Each arriving customer splits into c parts. The service requirements of the c parts of the nth customer are denoted by $\boldsymbol{\sigma}_n = (\sigma_n^1, \ldots, \sigma_n^c)$. We assume $\{\boldsymbol{\sigma}_n\}$ to be an i.i.d. sequence of random vectors with given joint distribution $G(x_1, \ldots, x_c) := P^0(\sigma_0^1 \leq x_1, \ldots, \sigma_0^c \leq x_c)$ and corresponding joint Laplace transform

$$\beta(s_1, \cdots, s_c) := E^0 e^{-\sum_{i=1}^c s_i \sigma_0^i}.$$

Theorem 1. *If we denote the joint Laplace transform of the workload process in steady state by* $\phi(s_1, \ldots, s_c) := E e^{-\sum_{i=1}^c s_i W_0^i}$ *then*

$$\phi(s_1, \ldots, s_c) = \frac{\sum_{i=1}^c s_i \psi_i(\ldots, s_{i-1}, s_{i+1}, \ldots)}{\sum_{i=1}^c s_i - \lambda\,(1 - \beta(s_1, \ldots, s_c))}. \tag{1}$$

The numerator in the above equation depends on c *unknown functions* $\psi_i : \mathbb{C}^{c-1} \mapsto \mathbb{C}$, $i = 1, 2, \ldots, c$ *where*

$$\psi_i(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_c) := E[\mathbf{1}(W_0^i = 0) e^{-\sum_{j \neq i} s_j W_0^j}]. \tag{2}$$

Proof: We examine the behavior of the workload vector $(W_t^1, \ldots, W_t^c)$. If we apply the Miyazawa Rate Conservation Principle on the process $\{X_t; t \in \mathbb{R}\}$, defined by

$$X_t := e^{-\sum_{i=1}^c s_i W_t^i}\ ,$$

we obtain

$$\lambda E^0 \left[e^{-\sum_{i=1}^c s_i (W_0^i + \sigma_0^i)} - e^{-\sum_{i=1}^c s_i W_0^i} \right] + E\left[\frac{d}{dt} e^{-\sum_{i=1}^c s_i W_t^i} \right] = 0$$

or

$$\lambda\,(\beta(s_1, \ldots, s_c) - 1)\, \phi(s_1, \ldots, s_c) + E\left[e^{-\sum_{i=1}^c s_i W_t^i} \sum_{i=1}^c s_i \mathbf{1}(W_0^i > 0) \right] = 0\ .$$

Hence

$$\lambda\left(1-\beta(s_1,\ldots,s_c)\right)\phi(s_1,\ldots,s_c)=\sum_{i=1}^{c}s_i\left(Ee^{-\sum_{i=1}^{c}s_iW^i}-E\left[\mathbf{1}(W_0^i=0)e^{-\sum_{j\neq i}s_iW_0^i}\right]\right)$$

or

$$\phi(s_1,\ldots,s_c)\left(\sum_{i=1}^{c}s_i-\lambda\left(1-\beta(s_1,\ldots,s_c)\right)\right)=\sum_{i=1}^{c}s_i\psi_i(\ldots,s_{i-1},s_{i+1},\ldots)$$

where

$$\psi_i(s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_c)=E[\mathbf{1}(W_0^i=0)e^{-\sum_{j\neq i}s_jW_0^j}]. \tag{3}$$

Thus from the above we obtain (1). ∎

Note that in the ordinary $M/G/1$ queue, $c=1$ and (1), (2), imply that the solution depends on one unknown constant which is easily determined from the requirement that $\phi(0)=1$ by an application of de l'Hôpital's rule.

3 Stochastically ordered service times

Suppose now that the service requirements for parts of different types are strongly ordered, i.e., for all $n\in\mathbb{Z}$ $\sigma_n^1\geq\sigma_n^2\geq\cdots\geq\sigma_n^c$ P^0–a.s. Then, it is easy to see that P–a.s. $W_0^1\geq W_0^2\geq\cdots\geq W_0^c$. We thus have the inequalities $\mathbf{1}(W_0^1=0)\leq\mathbf{1}(W_0^2=0)\leq\cdots\leq\mathbf{1}(W_0^c=0)$ holding P-a.s. and hence, from (2) it becomes clear that in this case ψ_i *depends only on the* $i-1$ *variables* $s_1,s_2,\ldots,s_{i-1}$, $i=2,3,\ldots,c-1$, while ψ_1 is a constant. Indeed, if we set $\rho_i:=\lambda E^0\sigma_0^i$ then clearly $\psi_1=E[\mathbf{1}(W_0^1=0)]=1-\rho_1$. Also, set

$$\chi_i(s_1,s_2,\ldots,s_{i-1})=\frac{\psi_i(s_1,s_2,\ldots,s_{i-1})}{1-\rho_i}=E\left[e^{-\sum_{j=1}^{i-1}s_jW_0^j}\mid W_0^i=0\right], \tag{4}$$

$i=2,3,\ldots,c$. Thus (1) is written as

$$\phi(s_1,\ldots,s_c)=\frac{s_1(1-\rho_1)+\sum_{i=2}^{c}s_i(1-\rho_i)\chi_i(s_1,\ldots,s_{i-1})}{\sum_{i=1}^{c}s_i-\lambda\left(1-\beta(s_1,\ldots,s_c)\right)}. \tag{5}$$

There remains the problem of determining the $c-1$ unknown functions $\chi_i(s_1,\ldots,s_{i-1})$, $i=2,3,\ldots,c$. This problem hinges upon expressing the conditional expectations that define them in a more convenient form and, as it turns out, the following lemma facilitates greatly this.

Lemma 1. *Let $\{S_n^i\}$ denote the point process defined by the beginnings of busy periods for station i. If we denote by P_i^0 the Palm transformation of P with respect to this point process, and by E_i^0 the corresponding Palm expectation, then*

$$\chi_i(s_1,\ldots,s_{i-1})=E_i^0e^{-\sum_{j=1}^{i-1}s_jW_0^j}. \tag{6}$$

Proof: If $\mathscr{F}_t^{S^i}$ is the internal history of the point process $\{S_n^i\}$ (see [3] for a definition) and $\mathscr{F}_t^{W^i} = \sigma-\{W_u^i; u \leq t\}$ the history of the process W^i, define the filtration $\mathbb{F}^i := \{\mathscr{F}_t^i; t \in R\}$ via $\mathscr{F}_t = \mathscr{F}_t^{S^i} \bigvee \mathscr{F}_t^{W^i}$. Then the $\mathbb{F}^i$–stochastic intensity of $\{S_n^i\}$ is given by

$$\alpha_t^i = \lambda \mathbf{1}(W_t^i = 0). \tag{7}$$

We now apply Papangelou's theorem (see [1], [9]): Since $\{W_t^i\}$, $\{S_n^i\}$, are jointly stationary and the processes $\{W_t^j\}$ have left–continuous sample paths with probability 1 and thus are predictable,

$$E_i^0 e^{-\sum_{j=1}^{i-1} s_j W_0^j} = \frac{E[\alpha_0^i \; e^{-\sum_{j=1}^{i-1} s_j W_0^j}]}{E\alpha_0^i}. \tag{8}$$

In view of the expression for the stochastic intensity in (7) the right hand side of the above equation becomes

$$E^i e^{-\sum_{j=1}^{i-1} s_j W_0^j} = \frac{E[\mathbf{1}(W_0^i = 0) \; e^{-\sum_{j=1}^{i-1} s_j W_0^j}]}{E[\mathbf{1}(W_0^i = 0)]} \tag{9}$$

and hence, from (6) and (9) we obtain (6). ∎

Consider now a smaller fork–join system with the following characteristics: The system consists of $i-1$ stations in parallel and the customers (who arrive again according to Poisson process with rate λ) now consist of $i-1$ parts. The service vector for the nth customer is again $\boldsymbol{\sigma}_n := (\sigma_n^1, \ldots, \sigma_n^{i-1})$, this time however we split it into a sum of two parts,

$$\boldsymbol{\sigma}_n = (\sigma_n^i, \sigma_n^i, \ldots, \sigma_n^i) + (\sigma_n^{i-1} - \sigma_n^i, \sigma_n^{i-2} - \sigma_n^i, \ldots, \sigma_n^1 - \sigma_n^i).$$

The first vector on the right hand side of the above equation represents work that has preemptive priority over the lower priory work represented by the second vector. (The second vector is of course always non-negative because of our strong ordering assumption.) Thus each customer brings to all stations the same amount of high-priority work and a varying amount of lower priority work. Clearly, the amount of high priority work is precisely the amount of work in the ith station of the original system. Also, the epochs of busy period initiation for high priority work are precisely the points $\{S_n^i\}$, and thus in order to obtain an expression for $\psi_i(s_1, \ldots, s_{i-1})$ it suffices to study the workload vector of lower priority work at these epochs.

In the sequel we will use the notation $\beta_i(s_1, \ldots, s_i) := \beta(s_1, s_2, \ldots, s_i, 0, \ldots, 0)$. We begin with the following

Lemma 2. *In the preemptive priority fork-join system with $i-1$ stations described above the steady-state workload vector of lower priority work considered at the epochs of busy period initiation for high priority work is equal to the workload vector in a fork-join system with Poisson arrivals with the same arrival rate and with service requirement vector sequence $\{\mathbf{v}_n\}$ where $\mathbf{v}_n := (v_n^1, \ldots, v_n^{i-1})$ are i.i.d. vectors with joint Laplace transform $\gamma_i(s_1, \ldots, s_{i-1})$ which is the unique solution with modulus less than one which satisfies the equation*

$$\gamma_i(s_1, \ldots, s_{i-1}) = \beta_i(s_1, s_2, \ldots, s_{i-1}, \lambda(1 - \gamma_i(s_1, \ldots, s_{i-1})) - \textstyle\sum_{j=1}^{i-1} s_j). \tag{10}$$

Proof: It is obvious that secondary work is performed only during the idle periods of high priority work and these are exponentially distributed with rate λ. Thus the lower priority workload vector at the end of the idle periods of high priority work is that of a *modified fork–join* system where customers arrive according to a Poisson process with rate λ and bring service requirement vector equal to the vector of secondary work accumulated during a high–priority busy period. To determine the new service requirement vector we will use an argument based on a sub-busy period decomposition. Let $(\sigma_0^1, \sigma_0^2, \ldots, \sigma_0^i)$ the service requirement vector that initiates the typical busy period of station i. If there are K Poisson arrivals during the service time σ_0^i then the random vector of service requirements for the modified fork–join system, $(Y_0^1, Y_0^2, \ldots, Y_0^{i-1})$, satisfies the relationship

$$(Y_0^1, Y_0^2, \ldots, Y_0^{i-1}) = \left(\sigma_0^1 - \sigma_0^i, \sigma_0^2 - \sigma_0^i, \ldots, \sigma_0^{i-1} - \sigma_0^i\right) + \sum_{k=1}^{K} \left(Y_k^1, Y_k^2, \ldots, Y_k^i\right),$$

where $\mathbf{Y}_k$, $k = 1, 2, \ldots, K$ are independent random vectors with the same distribution as $\mathbf{Y}_0$. Conditioning on σ_0^i and K we have

$$E_i^0[e^{-s_1 Y_0^1 - \cdots - s_{i-1} Y_0^{i-1}} \mid \sigma_0^i, K] = E_i^0[e^{-s_1 \sigma_0^1 - \cdots - s_{i-1} \sigma_0^{i-1}} \mid \sigma_0^i] \, e^{\sigma_0^i \sum_{j=1}^{i-1} s_j} \, \left(\gamma_i(s_1, \ldots, s_{i-1})\right)^K.$$

Taking expectation, first with respect to K given σ_0^i, and then with respect to σ_0^i, we obtain (10). ∎

Using lemma 2, one can recursively determine the unknown functions χ_i in (5) and hence $\phi(s_1, \ldots, s_c)$ itself. We then have the following

Theorem 2. *The joint Laplace transform of the stationary workload in the fork–join system with strongly ordered service requirements is given by the following recursive relations. Define*

$$\rho_j^{(i)} = \frac{\rho_j - \rho_i}{1 - \rho_i}, \qquad j = 1, 2, \ldots, i-1. \tag{11}$$

$$\gamma_i(s_1, \ldots, s_{i-1}) = \beta_i \left(s_1, s_2, \ldots, s_{i-1}, \lambda(1 - \gamma_i(s_1, \ldots, s_{i-1})) - \textstyle\sum_{j=1}^{i-1} s_j\right) \tag{12}$$

$$\chi_i(s_1, \ldots, s_{i-1}) = \frac{(1 - \rho_1^{(i)}) s_1 + \sum_{j=2}^{i-1} (1 - \rho_j^{(i)}) s_j \chi_j(s_1, \ldots, s_{j-1})}{\sum_{j=1}^{i-1} s_j - \lambda(1 - \gamma_i(s_1, \ldots, s_{i-1}))}, \tag{13}$$

where $i = 2, \ldots, c$. With these definitions, $\phi(s_1, \ldots, s_c)$ is given by (5).

4 An explicit expression when $c = 2$

Here we examine in more detail the case where $c = 2$ and we give an explicit expression for the joint Laplace transform of the equilibrium workload under the hypothesis that the service requirements are strongly ordered.

Proposition 1. *If the joint Laplace transform of the service requirements is* $\beta(s_1, s_2) := E^0 e^{-s_1\sigma^1 - s_2\sigma^2}$ *where* $\sigma^1 \geq \sigma^2$ *w.p. 1 then the joint Laplace transform of the workload in the two queues in steady state,* $\phi(s_1, s_2) := E[e^{-s_1 W_0^1 - s_2 W_0^2}]$ *is given by*

$$\phi(s_1, s_2) = \frac{s_1(1-\rho_1)}{s_1 - \lambda\left(1 - \lambda\gamma_2(s_1)\right)} \frac{s_1 + s_2 - \lambda\left(1 - \gamma_2(s_1)\right)}{s_1 + s_2 - \lambda\left(1 - \lambda\beta(s_1, s_2)\right)} \tag{14}$$

where $\gamma_2(s_1)$ *is the unique solution of the equation*

$$\gamma_2(s_1) = \beta(s_1, \lambda(1 - \gamma_2(s_1)) - s_1). \tag{15}$$

Proof: Specializing the general situation to the case $c = 2$ we have

$$\phi(s_1, s_2) = \frac{s_1(1-\rho_1) + s_2(1-\rho_2)\chi_2(s_1)}{s_1 + s_2 - \lambda(1 - \beta(s_1, s_2))} \tag{16}$$

where $\chi_2(s_1) := E\left[e^{-s_1 W_0^1} \mid W_0^2 = 0\right]$ is given by

$$\chi_2(s_1) = \frac{(1 - \rho_1^{(2)})s_1}{s_1 - \lambda\left(1 - \gamma_2(s_1)\right)} \tag{17}$$

and $\gamma_2(s_1)$ is the unique solution of the equation (15). Substituting (17) into (16) completes the proof. ∎

References

[1] F. Baccelli and P. Brémaud. *Elements of Queueing*, 2nd edition, Springer–Verlag, New York (2003).

[2] Baccelli, F., A.M. Makowski, and A. Shwartz. (1989). The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability*, **21**, 629-642.

[3] P. Brémaud (1981). *Point Processes and Queues. Martingale Dynamics.* Springer-Verlag, New York.

[4] Flatto, L. and S. Hahn, (1984). Two Parallel Queues Created by Arrivals with Two Demands I, *SIAM J. Appl. Math.* **44**, 5, 1041-1053.

[5] Flatto, L. and H. P. McKean, (1977). Two queues in parallel, *Comm. Pure Appl. Math.*, **30**, 255–263.

[6] G. Fayolle, R. Iasnogorodski, V. Malyshev (1999) *Random Walks in the Quarter-Plane: Algebraic Methods, Boundary Value Problems and Applications*, Springer Verlag.

[7] Knessl, C. (1991). On the diffusion approximation to a fork and join queueing model, *SIAM J. Appl. Math.* **51**, 1, 160-171.

[8] Miyazawa, M. (1979). “A formal approach to queueing processes in the steady state and their applications,” *J. Appl. Prob.* **16,** 332–346.

[9] Papangelou, F. (1974). On the Palm probabilities of processes of points and processes of lines. In: *Stochastic Geometry*, E.F. Harding and D.G. Kendall Eds., 114-147, Wiley.

[10] A. Shwartz and A. Weiss (1993). Induced rare events: analysis via large deviations and time reversal, *Journal of Applied Prob.* **25**, 667-689.

[11] P.E. Wright (1992). Two parallel processors with coupled inputs. *Advances in Applied Probability*, **24**, 986–1007.