Synchronized and Gated Queueing Models

Athens University of Economics and Business
Department of Statistics

Dimitra Pinotsi

April 11, 2022

CONTENTS

ABSTRACT

This thesis deals with a number of related problems in Queueing Theory dealing with gated systems with an infinite number of servers, and systems with synchronized arrivals.

Gated systems with an infinite number of servers have been studied as models of industrial processes and communications systems. Due to the gating mechanism, an exact analysis is typically not available and numerical methods (or simulation) is often used for such systems. Interestingly, despite the fact that there is an infinite number of servers, the presence of the gating mechanism makes even the question of stability of the system an interesting question. In Chapter 1 the stability of the gated $M/G/\infty$ queue is investigated using a Foster–Lyapunov drift criterion and it is established that the finiteness of the first moment of the service time distribution is the necessary and sufficient condition for the positive recurrence of the system. This result was published in *Probability in Engineering and Information Sciences*.

In Chapter 2 we examine a system consisting of m independent exponential servers in parallel, driven by the same deterministic input. This is a modification of the Flatto-Hahn-Wright model of queueing theory which, unlike the original, turns out to be analytically tractable. We focus on the time-stationary distribution of the number of customers which is obtained using a Markov embedding approach together with the Palm inversion formula. This result was published in *Letters of Operations Research.*

In Chapter 3 a detailed analysis of the busy period of the gated $M/G/\infty$ system is carried out. The stage length Markov chain is analyzed and a series solution is given in the light traffic case. Similarly, a gated $GI/M/\infty$ system is considered and again the Markov chain analyzing the number of customers in each service stage is analyzed in light traffic. In both cases the final results depend on the solution of infinite linear systems.

The existence and uniqueness of solutions and a result establishing that the sequence of the solutions of finite linear systems resulting from truncation of the original infinite systems approximate its solution is presented in the Appendix.

In Chapter 4 we consider the problem of m servers in parallel fed by a single Poisson process which causes simultaneous arrivals in all systems. Under the assumption that service times in the servers are stochastically ordered we are able to obtain the joint workload distribution for the m systems using the rate conservation principle. We also study the joint statistics of the busy periods of the m systems.

1. STABILITY OF GATED INFINITE SERVER SYSTEMS

The question of stability for the M/G/∞ queue with gated service is investigated using a Foster–Lyapunov drift criterion. The necessary and sufficient condition for positive recurrence is shown to be the finiteness of the first moment of the service time distribution, thus weakening the stability condition given in Browne et al. [11].

1.1 The Gated, Exhaustive, Parallel Service System

We consider the model described in Browne et al. [11]. Customers arrive at an infinite server system in a Poisson stream at rate λ per unit time. Their service requirements are assumed to be i.i.d. random variables $\{\sigma_i\}_{i=1,2,\ldots}$, with common distribution $G(x) = P(\sigma_1 \leq x)$. The servers, working in parallel, provide gated and exhaustive service to the queue. Service is *gated* in that customers are served in stages as follows: At the beginning of a service period which we shall call the *stage*, the gate opens to admit waiting customers who are transferred from the queue to the servers, and then the gate closes. The servers then begin serving all these customers in parallel (we assume that there is an infinite number of servers available). The stage ends when the service of all customers admitted is complete. A new stage begins immediately by admitting to service all customers that have accumulated during the first stage in front of the gate. In case where there where no arrivals during the first stage we assume that the system waits until the first customer arrives. At this point the gate opens and the customer is admitted immediately to service, thus initiating a service stage with a single customer. In a possible variation to the above model, if at the end of a service stage there are no customers waiting in front of the gate, then the servers take a vacation (of random length) after which they return and a new service stage starts by admitting the customers that have meanwhile accumulated in front of the gate. (Service is *exhaustive* in that no vacation period ever begins as long as there are customers waiting in line.)

Denote the number of customers served in the nth stage by K_n and the duration of the nth stage by Y_n. Then it is not hard to see [11] that $(K_n)_{n\in\mathbb{N}}$ is a Markov Chain with discrete state space (the non-negative integers) while $(Y_n)_{n\in\mathbb{N}}$ also constitutes a Markov Chain with continuous state space (the non-negative reals). An analysis of these Markov Chains is given in [11], [12], together with approximations to the stationary distribution of these chains in a number of cases. A heavy traffic analysis of such a system has been carried in Tan and Knessl [29]. In the above papers however the stability issue for this system has not been settled completely. Other queueing systems with gated mechanisms have been considered in [10], [24], and [41].

Since the model in question has an infinite number of servers one expects that there would be no need for a stability condition analogous to that of the single server queue. If fact if we assume that the service time distribution has bounded support and set $T := \inf\{x : G(x) = 1\} < \infty$, then the duration of all service stages is bounded above by T and this guarantees stability. If the support of the service time distribution however is not bounded, then arbitrarily large service times are possible and the stability of the system is no longer obvious. If this were the ordinary $M/G/\infty$ queue then the necessary and sufficient condition for the existence of steady state would be the finiteness of the mean service time: $E\sigma < \infty$. In fact, since the performance of this system in terms of delays is worse than that of the corresponding $M/G/\infty$ queue because of the existence of the gating mechanism, one expects that the finiteness of the expected service time is a necessary condition for stability. In this chapter we will show using a Foster–Lyapunov drift criterion that it is both necessary and sufficient.

To see that this is not a trivial question let us envision the following scenario described in different terms in Browne et al. [11]. Set $M_k := \max(\sigma_1, \sigma_2, \ldots, \sigma_k)$ and let G^k denote the distribution of M_k where we have of course that $G^k(x) = (G(x))^k$ for all $x \geq 0$. The duration of a service stage is the maximum of all service times of customers served during this stage and thus, if we assume that the first stage serves k_1 customers, its duration, Y_1, has distribution G^{k_1}. If during Y_1 an unusually large number of customers, $k_2 >> k_1$ arrives as a result of statistical fluctuations, the second stage, Y_2, will be distributed according to G^{k_2} and thus be very likely much larger than Y_1. One can imagine such a process escalating, particularly if G is heavy–tailed. Browne et al. [11] have shown that *this escalation is not possible, provided that the service time distribution G has finite second moment* thus establishing stability under this condition. We show that the finiteness of the first moment of G is in fact not only necessary but also sufficient in order to establish the stability of this system.

1.2 The Markov Chain of service stages for lattice service time distributions

In order to establish the stability of the system we will examine the Markov chain of service stage lengths, $(Y_n)_{n=1,2,\ldots}$. In fact in this section we will simplify our task by assuming that the service time distribution G is lattice with lattice size $\delta > 0$, i.e. that all service times are integer multiples of δ. In particular we will set $g_j := P(\sigma = j\delta)$, $j = 1, 2, \ldots,$. An immediate consequence of this assumption is of course that the duration of the nth stage, Y_n, is also an integer multiple of δ. Thus (Y_n) also becomes a discrete state space Markov chain. For simplicity we will examine instead (Φ_n) where $\Phi_n := Y_n/\delta$, a Markov chain with state space the support of σ/δ i.e. the subset of $\mathbb{N}$ defined as $\mathbb{S} = \{j \in \mathbb{N} : g_j > 0\}$.

It is easy to see that (Φ_n) is an irreducible chain on $\mathbb{S}$. Thus if $\mathbb{S}$ is finite it must be positive recurrent and hence the system must be stable since both the stage lengths and the number of customers waiting in the queue are tight. Thus the only case of interest is when $\mathbb{S}$ is not finite.

The transition matrix for (Φ_n) can be evaluated by first computing the conditional probability that the $(n+1)$th stage serves $K_{n+1} = m$ customers, given the length of the nth stage. This amounts to computing the probability of m Poisson arrivals, taking also into consideration that in the case where we have 0 arrivals during the service stage we wait until the first arrival. Thus

$$P(K_{n+1} = m \mid Y_n = i\delta) = \begin{cases} \frac{1}{m!}(\lambda i\delta)^m e^{-\lambda i\delta}, & \text{for } m = 2, 3, \ldots, \\ \\ (\lambda i\delta + 1)e^{-\lambda i\delta} & \text{for } m = 1. \end{cases}$$

Also, recalling that $g_j := P(\sigma = j\delta)$ and $G_j := P(\sigma \le j\delta)$ we have

$$\begin{aligned} P(Y_{n+1} = j\delta \mid K_{n+1} = m) &= m g_j G_{j-1}^{m-1} + \frac{(m-1)m}{2} g_j^2 G_{j-1}^{m-2} + \cdots + \frac{m!}{m!} g_j^m \\ &= \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l}. \end{aligned}$$

Thus, if we denote the transition probability matrix for the Markov Chain (Φ_n) by $P_{ij} := P(\Phi_{n+1} = j \mid \Phi_n = i)$, combining the above we obtain

$$P_{ij} = (\lambda i\delta + 1)e^{-\lambda i\delta} g_j + \sum_{m=2}^{\infty} \frac{(\lambda i\delta)^m}{m!} e^{-\lambda i\delta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l}.$$

Setting $\beta := \lambda\delta$, the above expression can also be written as

$$\begin{aligned} P_{ij} &= (i\beta + 1)e^{-i\beta} g_j + \sum_{m=2}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} \\ &= e^{-i\beta} g_j + \sum_{m=1}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=1}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} \\ &= e^{-i\beta} g_j + \sum_{m=0}^{\infty} \frac{(i\beta)^m}{m!} e^{-i\beta} \sum_{l=0}^{m} \binom{m}{l} g_j^l G_{j-1}^{m-l} - \sum_{m=0}^{\infty} \frac{(i\beta)^m}{m!} G_{j-1}^m e^{-i\beta}. \quad (1.1) \end{aligned}$$

The double sum above can be simplified considerably as follows

$$\begin{aligned}\sum_{m=0}^{\infty}\frac{(i\beta)^m}{m!}e^{-i\beta}\sum_{l=0}^{m}\binom{m}{l}g_j^l G_{j-1}^{m-l} &= \sum_{l=0}^{m}\left\{\sum_{m=l}^{\infty}G_{j-1}^{m-l}(i\beta)^m\frac{m!}{l!(m-l)!}\frac{1}{m!}\right\}g_j^l e^{-i\beta}\\ &= \sum_{l=0}^{m}\frac{(i\beta)^l}{l!}g_j^l e^{-i\beta}e^{i\beta G_{j-1}}\\ &= e^{i\beta g_j - i\beta + i\beta G_{j-1}}\\ &= e^{i\beta G_j - i\beta} \qquad (1.2)\end{aligned}$$

where, in the last equation we have used the fact that $G_j = (G_{j-1} + g_j)$. Thus, from (1.1) and (1.2), using the notation $\overline{G}_j := 1 - G_j = \sum_{k=j+1}^{\infty} g_k$, the transition probability matrix can be written as

$$P_{ij} = e^{-i\beta\overline{G}_j} - e^{-i\beta\overline{G}_{j-1}} + e^{-i\beta}g_j, \qquad i,j = 1,2,3,\ldots \qquad (1.3)$$

We also point out for future reference the following inequality satisfied by the transition probabilities:

$$P_{ij} \;=\; e^{-i\beta\overline{G}_j} - e^{-i\beta\overline{G}_{j-1}} + e^{-i\beta}g_j \;\leq\; i\beta e^{-i\beta\overline{G}_j}g_j + e^{-i\beta}g_j. \qquad (1.4)$$

The above is easy to see, if we note that, from the mean value theorem we have

$$e^{-i\beta\overline{G}_j} - e^{-i\beta\overline{G}_{j-1}} = \int_{i\beta\overline{G}_j}^{i\beta\overline{G}_{j-1}} e^{-x}dx \leq i\beta\left(\overline{G}_{j-1} - \overline{G}_j\right)e^{-i\beta\overline{G}_j}$$

and remember that $\overline{G}_{j-1} - \overline{G}_j = g_j$.

1.3 Positive recurrence, finite mean stage length, and stability

In order to establish the stability of the system we have first to show the positive recurrence of the Markov chain (Φ_n). This problem is not easy to tackle directly since the stationary equations for (Φ_n),

$$\pi_j = \sum_{i=1}^{\infty}\pi_i P_{ij}, \qquad j = 1,2,\ldots, \qquad (1.5)$$

turn out to be intractable. It is however possible to address this problem using the classical Foster criterion for positive recurrence (see Asmussen [1]). Nonetheless, this is not sufficient for our purposes as the positive recurrence of (Φ_n) (and hence of (Y_n)) is a necessary but not a sufficient condition for the stability of our system.

To clarify this point consider an analogous situation in a simpler setting. Suppose that (X_n) is a zero drift simple random walk, (with steps ± 1) on $\mathbb{Z}$. Suppose that $X_0 = 0$ w.p. 1 and define $T_0 = 0$, $T_n = \inf\{k > T_{n-1} : X_k = 0\}$. Thus (T_n) are the corresponding epochs of return to 0. If we denote by $Y_n := T_n - T_{n-1}$, the corresponding durations of the excursions from state 0, then clearly (Y_n) is an i.i.d. sequence or random variables (with infinite mean). Viewed as a Markov Chain, (Y_n) *is positive recurrent* even though (X_n) is not.

Returning to our system, let us denote by $X_t = (U_t, V_t)$ the process whose first component is the number of customers waiting behind the gate and the second component the number of customers in service at time t. Let also $\{T_n\}$ denote the point process of the epochs when the gate opens and K_n the number of customers admitted into the system at epoch T_n. Then it should be clear that, (X_t) is a semi–regenerative process (e.g. see [1]) with respect to the Markov–Renewal process (T_n, K_n). Thus it is easy to see that the system will be stable (in the sense that it will possess a stationary distribution) if and only if in the *synchronous version* of the process (where time $t = 0$ coincides with a typical point T_0 of the Markov–Renewal Process in the sense that K_0 is distributed according to the stationary distribution of the embedded Markov chain), $ET_1 - T_0 < \infty$. This condition however translates to the requirement of positive recurrence for the Markov Chain (Φ_n) and the finiteness of the mean of its stationary distribution (1.5), i.e.

$$\sum_{i=1}^{\infty} i\pi_i < \infty. \tag{1.6}$$

1.4 The stability condition

The basic tool for establishing not only the positive recurrence of (Φ_n) but also the finiteness of the first moment of the stationary distribution (1.6) is the following generalization of the classic criterion of Foster. (This is Proposition 2.9 of Tweedie [52]. See also Meyn and Tweedie [36] for a comprehensive account of Foster–Lyapunov drift criteria in general state space Markov Chains.)

Theorem 1 (Foster–Lyapunov Criterion)**.** *Suppose that a Markov Chain with countable state space, say $\mathbb{Z}_+$, and transition probability matrix P_{ij} is irreducible and let N be a given natural number. If $V : \mathbb{Z}_+ \to \mathbb{R}_+$, is a non-negative and $f : \mathbb{Z}_+ \to \mathbb{R}_+$ a strictly positive function on the state space such that*

$$\sum_{j\in E} P_{ij}V(j) \leq b \qquad i \leq N, \tag{1.7}$$

$$\sum_{j\in E} P_{ij}V(j) \leq V(i) - f(i), \qquad i > N, \tag{1.8}$$

where $b > 0$, then the Markov Chain is positive recurrent and its (unique, due to its irreducibility) stationary distribution, π, satisfies $\sum_j \pi(j)f(j) < \infty$. Furthermore we have $\sum_j P_{ij}^n f(j) \to \sum_j \pi_j f(j)$ as $n \to \infty$.

We are now ready to state the following

Theorem 2. *The Markov Chain (Φ_n) is positive recurrent with stationary probability π_i that has finite mean $\sum_{i=1}^{\infty} \pi_i i < \infty$ provided that $E\sigma := \delta \sum_{j=1}^{\infty} j g_j < \infty$.*

Proof: We will assume that $g_j > 0$ for infinitely many j's, i.e. $\{g_j\}$ has an infinite tail, otherwise the Markov Chain (Φ_n) has finite state space and it is obviously positive recurrent. Furthermore, in this case the mean expected stage length in steady state is also finite and thus the system is stable.

We will use the Foster–Lyapunov drift criterion of theorem 1 with $f(i) = i$, $V(i) := ci$, where $c > 1$. With this choice (1.8) is equivalent to

$$\frac{1}{i} \sum_{j=1}^{\infty} P_{ij} j \ \leq \ 1 - c^{-1}. \tag{1.9}$$

Taking into account the inequality (1.4) we see that (1.9) is satisfied provided that the following inequality holds:

$$\sum_{j=1}^{\infty} \beta e^{-i\beta \overline{G}_j} g_j j \ + \ \frac{1}{i} \sum_{j=1}^{\infty} e^{-i\beta} g_j j \ \ \leq \ \ 1 - c^{-1}. \tag{1.10}$$

By the Dominated Convergence Theorem,

$$\lim_{i \to \infty} \sum_{j=0}^{\infty} \beta e^{-i\beta \overline{G}_j} g_j j = \sum_{j=0}^{\infty} \lim_{i \to \infty} \beta e^{-i\beta \overline{G}_j} g_j j = 0 \tag{1.11}$$

where we have used the dominating function jg_j. (This is a valid dominating function since $\sum_{j=0}^{\infty} jg_j < \infty$. We have also used the fact that $\overline{G}_j > 0$ for all $j \in \mathbb{N}$ since by assumption $\{g_j\}$ has an infinite tail.) Using again the Dominated Convergence Theorem on the second term of the right hand side of (1.11) with the same dominating function we have

$$\lim_{i \to \infty} \sum_{j=0}^{\infty} \frac{e^{-i\beta}}{i} g_j j = \sum_{j=0}^{\infty} \lim_{i \to \infty} \frac{e^{-i\beta}}{i} g_j j = 0. \tag{1.12}$$

Clearly (1.11) and (1.12) imply that (1.10) and hence (1.9) hold for all i greater than some $N \in \mathbb{N}$.

The finiteness of $\sum_j g_j j$ also immediately guarantees that (1.7) holds for some $b > 0$ and all $i \leq N$. Thus, appealing to theorem 1 we conclude that the Markov Chain (Φ_n) is positive recurrent when $E\sigma < \infty$ with stationary distribution π and furthermore that $\sum_{i=1}^{\infty} \pi_i i < \infty$. This finiteness of the first moment of the stationary distribution guarantees that the mean stage length is finite in steady state and thus that there exists a proper stationary regime for this system. ∎

Remark 3. *Note that we do not claim that when $E\sigma = \infty$, (Φ_n) cannot be positive recurrent.*

Thus, in view of the discussion in section 3 we have also shown the following

Corollary 4. *The condition $\sum_j g_j j < \infty$ is necessary and sufficient for the existence of a stationary distribution of the semi-regenerative process (X_t).*

Proof: The sufficiency of the moment condition follows directly from the above analysis. As for the necessity, it is enough to argue that the mean stage length is the maximum of a number of service times and, as such, stochastically greater than a service time. Thus, if $\sum_j g_j j$ diverges, the mean stage length must be infinite which implies that a stationary regime cannot exist for the process (X_t) by standard results regarding semi–regenerative processes. ∎

1.5 Non-lattice service time distributions

When the service time distribution G is not lattice the above analysis fails. However, a similar analysis could be carried out using drift criteria for Markov Chains on general state spaces (see [36]). Alternatively one could focus on the Markov Chain (K_n) of the number of customers served in each stage, which in every case has a denumerable state space. Instead of adopting any of these two approaches however we will sketch a simple argument based on stochastic ordering.

If F, G, are two distribution functions on $\mathbb{R}$ we say that F stochastically dominates G if $F(x) \leq G(x)$ for all $x \in \mathbb{R}$. This stochastic order (also known as strong order) will be denoted by $F \geq_{\text{st}} G$. For further background on stochastic ordering we refer the reader to Stoyan (1984). We begin with the following obvious

Proposition 5. *If G, $\tilde{G}$, are two service time distributions such that $G \leq_{\text{st}} \tilde{G}$ and if (Y_n), $(\tilde{Y}_n)$ are the corresponding service stage Markov Chains, then $Y_n \leq_{\text{st}} \tilde{Y}_n$ for all $n \in \mathbb{N}$.*

Proof: It is immediate by a appropriately constructing both chains on the same probability space. ∎

A distribution F on $\mathbb{R}^+$ is called lattice if there exists $\delta > 0$ such that $F\{\delta\mathbb{N}\} = 1$. Given $\delta > 0$ and a distribution function G on $\mathbb{R}^+$, we define a lattice distribution, G_δ by means of

$$G_\delta(x) = \sum_{n=0}^{\infty} G(n\delta)\mathbf{1}(n\delta \leq x < (n+1)\delta).$$

Then $G_\delta \geq_{\mathsf{st}} G$ and, by virtue of proposition 5, we have $Y_n^\delta \geq_{\mathsf{st}} Y_n$ for each $n \in \mathbb{N}$ for the corresponding service stage Markov Chains. Furthermore, if $\int_0^\infty xG(dx) < \infty$ then $\int_0^\infty xG_\delta(dx) < \infty$. Thus, the finiteness of the first moment of the service time distribution implies the positive recurrence of the countable state space Markov Chain (Y_n^δ) and the finiteness of the first moment of *its* stationary distribution which in turn implies the stability of the original system.

2. THE FLATTO-HAHN-WRIGHT MODEL

We consider m independent exponential servers in parallel, driven by the same deterministic input. This is a modification of the Flatto-Hahn-Wright model which turns out to be easily tractable. We focus on the time-stationary distribution of the number of customers which is obtained using the Palm inversion formula.

2.1 Introduction

Synchronized (or fork-join) queues have been an object of study over the last three decades as models of parallel processing in computer systems and assembly operations in manufacturing. In the model we examine here, the service facility consists of m single servers in parallel, each with its own queue. The m buffers have infinite capacity and individually operate according to the FIFO discipline. Upon arrival to the service facility, each customer splits in m parts, each part joining the corresponding queue. While each station viewed separately is an ordinary single server queue, the object is the determination of the *joint statistics* of the m queues which is in general hard to obtain.

The above system when customers arrive according to a Poisson process and the service requirements for the parts are independent, exponential random variables with rate depending on the type of part, is known as the FHW (Flatto–Hahn–Wright) model (see [15], [16], [55]). Flatto and Hahn [16] and Flatto [15], have studied the system (for the case $m = 2$) using complex analysis techniques. The waiting time in such systems has been studied by Zhang [57]. The probability distribution of the join queue length has been obtained by Li and Zhao [31]. Also, interpolation approximations for symmetric fork-join queues are given by Varma and Makowski [53].

The FHW model is of course a special case of a two–dimensional random walk on the positive quadrant. There is a rich theory connecting this problem to boundary value problems and the multidimensional extension of the Wiener–Hopf factorization. The reader is referred to Fayolle, Iasnogorodski, and Malyshev [19] both for an overview and for a state-of-the-art treatment of these issues.

Fork-join systems consisting of two queues with Poisson arrivals and service requirements which are i.i.d. sequences of exchangeable pairs of random variables have been studied in Baccelli [4]. We also mention the Taylor series approach used by Ayhan and Baccelli [3], where the assumption of exponential service times is relaxed, and Baccelli, Makowski, and Swhartz [6], where bounds for the performance of more general fork-join queues are obtained by means of stochastic ordering arguments (see also Li and Xu [30]). Because of the intractability of the FHW model most of the explicit results are asymptotic in nature. These include both asymptotics based on generating functions obtained by complex analysis techniques (e.g. [15], [16]) and results obtained using large deviation techniques [42]. We also mention the diffusion approximation of [29] and the related problem of fork-join fluid queues studied in [28].

A related line or research that studies queueing networks with signals and concurrent movements examines the FHW model in the framework of markovian queueing networks. We refer the interested reader to [14] and [13].

The model examined in this Chapter, unlike the classical FHW model, is tractable by means of elementary tools. In fact, due to the deterministic nature of arrivals and the independence of the service processes in the m queues, the *customer-stationary* (Palm with respect to the arrival processes) queue lengths are independent and thus the system (under the Palm probability measure) can be viewed as m independent $D/M/1$ queues. The situation becomes more complicated when we turn our attention to the *(time-) stationary version of the process* and this is the main focus here.

Section 2 gives a more detailed description of the model while in section 3 the Palm inversion formula in conjunction with an argument based on generating functions is used in order to derive the joint distribution of the stationary number of customers in the system. Section 4 provides an illustration of the above results by examining in more detail the system with two stations ($m = 2$). An expression is obtained for the stationary distribution of the workload, and the deterministic model is compared to the classical FHW model with Poisson arrivals in terms of the correlation coefficient of the stationary queue sizes.

2.2 Synchronized queues with deterministic arrivals

In the system considered here customers, each consisting of m parts, arrive to the service facility according to a deterministic process with constant interarrival times, equal to a. Upon arrival to the system, each customer splits into its constituent parts which join the corresponding queues. From that point on the parts move independently even though, for some applications, it may be useful to think that, after service completion, the parts of a customer that finish first wait in a “staging area” for their counterparts and, once all parts have completed their processing they are assembled into a finished unit. In a manufacturing context this could describe an assembly operation. The point process of arrival epochs to the system will be denoted by $\{T_n; n \in \mathbb{Z}\}$ where $T_{n+1} = T_n + a$. Service requirements for each queue are independent, exponential random variables with rate μ_k for the kth station. Clearly the system is stable iff $a \min\{\mu_1, \ldots, \mu_m\} > 1$. Denote by $\{X_t^k; t \in \mathbb{R}\}$ the number of customers in station k, $(k = 1, \ldots, m)$ and let $\mathbf{X}_t := (X_t^1, \ldots, X_t^m)$ denote the number of customers in the m queues. We will assume that this process has *right-continuous sample paths* with probability 1. In particular $(X^1_{T_n-}, \ldots, X^m_{T_n-})$ is the number of customers in the m queues as seen by an arrival, right before the arrival epoch. Suppose now that a stationary version of this process has been constructed on the probability space $(\Omega, \mathcal{F}, P)$ and let P^0 denote the Palm transformation of P under the point process $\{T_n; n \in \mathbb{Z}\}$. We will denote by E^0 the expectation with respect to P^0. Intuitively, P^0 is the probability measure conditioned on the event that the origin coincides with a typical arrival point, which by convention is denoted by T_0. Thus $P^0(T_0 = 0) = 1$. We refer the reader to Baccelli and Brémaud [5] for formal definitions and the mathematical framework. Since arrivals are deterministic and service times are independent in the m queues it is easy to see that, under P^0, the m queue-length processes $\{X_t^k; t \in \mathbb{R}\}$, $k = 1, 2, \ldots, m$, are independent. Thus the Palm version of the process can be readily analyzed by studying m independent $D/M/1$ systems. In particular

$$P^0(X^1_{0-} = n_1, \ldots, X^m_{0-} = n_m) = \prod_{k=1}^{m} (1 - \sigma_k)\sigma_k^{n_k}, \qquad n_k = 0, 1, 2, \ldots. \tag{2.1}$$

where σ_k be the unique solution of the equation

$$x = e^{-a\mu_k(1-x)}, \qquad k = 1, \ldots, m, \tag{2.2}$$

that is less than one. Indeed, besides the obvious solution, $x = 1$, it is clear from a convexity argument that the above equation has one more solution which, as is well known (see [1]), belongs to the interval $(0, 1)$ provided that the stability condition $a\mu_k > 1$ holds.

2.3 The stationary number of customers in the system

We now turn to the stationary version of the process. It is clear that the m queue–length processes are no longer independent. From standard results concerning the $GI/M/1$ queue (e.g. see [1, p. 280]) it follows that the marginal distribution for the stationary number of customers in each queue is a modified geometric distribution given by

$$\begin{aligned} P(X_0^k = n) &= \rho_k(1-\sigma_k)\sigma_k^{n-1}, \quad n = 1, 2, \ldots \\ P(X_0^k = 0) &= 1 - \rho_k, \end{aligned} \tag{2.3}$$

with $\rho_k = (a\mu_k)^{-1}$ for $k = 1, \ldots, m$. The corresponding p.g.f. (probability generating function) is given by

$$\varphi(z) := 1 - \rho_k + z\rho_k \frac{1-\sigma_k}{1-z\sigma_k}. \tag{2.4}$$

On the other hand, the joint distribution of the queue–lengths under the stationary probability measure P is harder to find. As we will see next it can be obtained from the Palm inversion formula using a conditioning argument.

We start with the following elementary lemma where, as usual, x^+ denotes the positive part of the real number x.

Lemma 6. *Let Y be a geometric random variable with p.g.f. $Ez^Y = \frac{(1-\sigma)z}{1-z\sigma}$, where $\sigma \in (0,1)$, and N a Poisson random variable, independent of Y, with mean β. Then*

$$Ez^{(Y-N)^+} = 1 - \frac{1-z}{1-z\sigma} e^{-\beta(1-\sigma)}.$$

Proof: Condition on N to obtain

$$E[z^{(Y-N)^+}|N] = \sum_{k=1}^{N}(1-\sigma)\sigma^{k-1} + \sum_{k=N+1}^{\infty}(1-\sigma)\sigma^{k-1}z^{k-N} = 1 - \sigma^N \frac{1-z}{1-z\sigma}.$$

Taking expectation with respect to N completes the proof. □

Denote by

$$\varphi(z_1, \ldots, z_m) = E\prod_{k=1}^{m} z_k^{X_0^k}$$

the probability generating function of the stationary number of customers in the system. Let us also denote by $\mathcal{A}_r$ the class of all subsets of the set $S_m := \{1, 2, \ldots, m\}$ containing exactly r elements. In particular we have of course that $|\mathcal{A}_r| = \binom{m}{r}$ where, as usual, $|B|$ denotes the cardinality of the set B. Also, for any $\vec{n} = (n_1, \ldots, n_m) \in \mathbb{N}_0^m$ define $\Phi_{\vec{n}} := \{k : n_k \geq 1\} \subseteq S_m$, the set of all indices corresponding to non-zero components of the vector $\vec{n}$. We are ready to state our main result.

Theorem 7. *The probability generating function of the stationary number of customers in the system is given by*

$$\varphi(z_1, \ldots, z_m) = 1 + \sum_{r=1}^{m}(-1)^r \sum_{A\in\mathcal{A}_r} C_A \prod_{k\in A} \frac{1-z_k}{1-z_k\sigma_k} \tag{2.5}$$

where

$$C_A := \frac{1-\prod_{k\in A}\sigma_k}{\sum_{k\in A}\rho_k^{-1}(1-\sigma_k)}, \tag{2.6}$$

the constants being indexed by the subsets $A \subseteq \{1,2,\ldots,m\}$. The corresponding probability distribution is given by

$$P(X_0^1=n_1,\ldots,X_0^m=n_m) = \Gamma_{\vec{n}} \prod_{k\in\Phi_{\vec{n}}}(1-\sigma_k)\sigma_k^{n_k-1} \tag{2.7}$$

where

$$\Gamma_{\vec{n}} := \sum_{\{A:A\supseteq\Phi_{\vec{n}}\}}(-1)^{|A|-|\Phi_{\vec{n}}|}C_A = \sum_{r=|\Phi_{\vec{n}}|}^{m}(-1)^{r-|\Phi_{\vec{n}}|}\sum_{A\in\mathcal{A}_r} C_A\mathbf{1}(\Phi_{\vec{n}}\subseteq A). \tag{2.8}$$

Remark: The expression (2.7) for $\Phi_{\vec{n}}=\varnothing$ (i.e. for $\vec{n}=(0,0,\ldots,0)$) becomes

$$P(X_0^1=0,\ldots,X_0^m=0) = 1+\sum_{r=1}^{m}(-1)^r\sum_{A\in\mathcal{A}_r}C_A. \tag{2.9}$$

Also, for $\Phi_{\vec{n}}=S_m$, i.e. when $n_k\geq 1$ for all k,

$$\Gamma_{\vec{n}} = C_{S_m} = \frac{1-\prod_{k=1}^{m}\sigma_k}{\sum_{k=1}^{m}\rho_k^{-1}(1-\sigma_k)}.$$

Proof: A straight-forward application of the Palm inversion formula (see [5]) gives

$$\varphi(z_1,\ldots,z_m) = a^{-1}E^0\int_0^a\left(\prod_{k=1}^{m} z_k^{X_t^k}\right)dt = a^{-1}E^0\int_0^a\left(\prod_{k=1}^{m} z_k^{(X_0^k-N_t^k)^+}\right)dt. \tag{2.10}$$

In the above expression $\{(N_t^1,\ldots,N_t^m); t\geq 0\}$ are m independent Poisson processes with rates μ_k, $k=1,\ldots,m$, representing the service processes in the m exponential servers. Furthermore, these Poisson processes are independent of the vector of queue lengths at time 0, $(X_0^1,\ldots,X_0^m)$. Finally, under the probability measure P^0, and since the sample paths are right-continuous,

$$P^0(X_0^1=n_1,\ldots,X_0^m=n_m) = \prod_{k=1}^{m}(1-\sigma_k)\sigma_k^{n_k-1}, \quad n_k=1,2,\ldots.$$

(In the above expression the customer arriving at $t = 0$ has been taken into account– cf. (2.1)). Thus, appealing to the Fubini theorem, and using the independence of the X_0^k under P^0 and Lemma 1, we can write the right-hand side of (2.10) as

$$a^{-1}\int_0^a \prod_{k=1}^m E^0 z_k^{(X_0^k - N_t^k)^+} dt = a^{-1}\int_0^a \prod_{k=1}^m \left(1 - \frac{1-z_k}{1-z_k\sigma_k} e^{-\mu_k t(1-\sigma_k)}\right) dt. \tag{2.11}$$

The product inside the integral on the right hand side of the above expression can be written as

$$\prod_{k=1}^m \left(1 - \frac{1-z_k}{1-z_k\sigma_k} e^{-\mu_k t(1-\sigma_k)}\right) = 1 + \sum_{r=1}^m (-1)^r \sum_{A\in\mathcal{A}_r} e^{-t\sum_{k\in A}\mu_k(1-\sigma_k)} \prod_{k\in A} \frac{1-z_k}{1-z_k\sigma_k}$$

and thus the right hand side of (2.11) becomes

$$1 + \sum_{r=1}^m (-1)^r \sum_{A\in\mathcal{A}_r} \left(\prod_{k\in A} \frac{1-z_k}{1-z_k\sigma_k}\right) a^{-1}\int_0^a e^{-t\sum_{k\in A}\mu_k(1-\sigma_k)} dt. \tag{2.12}$$

However,

$$a^{-1}\int_0^a e^{-t\sum_{k\in A}\mu_k(1-\sigma_k)} dt = \frac{1 - e^{-a\sum_{k\in A}\mu_k(1-\sigma_k)}}{a\sum_{k\in A}\mu_k(1-\sigma_k)} = \frac{1-\prod_{k\in A}\sigma_k}{\sum_{k\in A}\rho_k^{-1}(1-\sigma_k)}$$

where, in the last equation we have made use of the defining relation for the σ_k, the definition of ρ_k, and (2.2). From the above the joint generating function of the stationary number of customers in the m queues becomes

$$\varphi(z_1,\ldots,z_m) = 1 + \sum_{r=1}^m (-1)^r \sum_{A\in\mathcal{A}_r} \frac{1-\prod_{k\in A}\sigma_k}{\sum_{k\in A}\rho_k^{-1}(1-\sigma_k)} \prod_{k\in A} \frac{1-z_k}{1-z_k\sigma_k}.$$

This establishes (2.5). Since

$$\frac{1-z}{1-z\sigma} = 1 - \frac{z(1-\sigma)}{1-z\sigma} = 1 - \sum_{n=1}^\infty (1-\sigma)\sigma^{n-1}z^n,$$

we have

$$\varphi(z_1,\ldots,z_m) = 1 + \sum_{r=1}^m (-1)^r \sum_{A\in\mathcal{A}_r} C_A \prod_{k\in A}\left(1 - \frac{z_k(1-\sigma_k)}{1-z_k\sigma_k}\right)$$

or

$$\varphi(z_1,\ldots,z_m) = 1 + \sum_{r=1}^m (-1)^r \sum_{A\in\mathcal{A}_r} C_A \prod_{k\in A}\left(1 - \sum_{n_k=1}^\infty (1-\sigma_k)\sigma_k^{n_k-1} z_k^{n_k}\right). \tag{2.13}$$

We can now imagine the process of collecting terms from the above expression. We begin with an example: When $\Phi_{\vec{n}} = S_m$, i.e. when $n_k \geq 1$ for all $k = 1, 2, \ldots, m$, only the product

$$\prod_{k \in S_m} \left(1 - \sum_{n_k=1}^{\infty} (1-\sigma_k)\sigma_k^{n_k-1} z_k^{n_k} \right)$$

in (2.13) contains the term $z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}$ and the corresponding coefficient is

$$(-1)^m \prod_{k=1}^{m} (1-\sigma_k)\sigma_k^{n_k-1}.$$

From (2.13) we see that this term is multiplied by $(-1)^m C_{S_m}$ and thus the coefficient of the term $z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}$ in the expansion of $\varphi(z_1, \ldots, z_m)$ is equal to

$$(-1)^{m+m} C_{S_m} \prod_{k=1}^{m} (1-\sigma_k)\sigma_k^{n_k-1} = \frac{1 - \prod_{k=1}^{m} \sigma_k}{\sum_{k=1}^{m} \rho_k^{-1}(1-\sigma_k)} \prod_{k=1}^{m} (1-\sigma_k)\sigma_k^{n_k-1}.$$

(cf. Remark 1.) In the general case, the product

$$\prod_{k \in A} \left(1 - \sum_{n_k=1}^{\infty} (1-\sigma_k)\sigma_k^{n_k-1} z_k^{n_k} \right)$$

indexed by the set A contains the term $z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}$ if and only if $\Phi_{\vec{n}} \subseteq A$. The coefficient of this term when we expand this product is

$$1^{|A|-|\Phi_{\vec{n}}|} \cdot (-1)^{|\Phi_{\vec{n}}|} \prod_{k \in \Phi_{\vec{n}}} (1-\sigma_k)\sigma_k^{n_k-1}.$$

In order to find the coefficient of $z_1^{n_1} z_2^{n_2} \cdots z_m^{n_m}$ in the expansion of $\varphi(z_1, \ldots, z_m)$ it suffices to multiply this term by $(-1)^{|A|} C_A$ and then to sum over all sets $A \supseteq \Phi_{\vec{n}}$. We thus have

$$P(X_0^1 = n_1, \ldots, X_0^m = n_m) = \sum_{\Phi_{\vec{n}} \subseteq A} (-1)^{|A|-|\Phi_{\vec{n}}|} C_A \prod_{k \in \Phi_{\vec{n}}} (1-\sigma_k)\sigma_k^{n_k-1}. \tag{2.14}$$

The expression (2.7) is a restatement of the above. In the second expression for $\Gamma_{\vec{n}}$ in (2.8) we have split the sum according to the cardinality of the index set A. Finally (2.9) is the special case where $\Phi_{\vec{n}} = \varnothing$ and this completes the proof. □

Corollary 8. *In the symmetric case, where the service rates in all stations are equal to μ, the probability generating function of the stationary number of customers in the system is given by*

$$\varphi(z_1, \ldots, z_m) = 1 + \sum_{r=1}^{m} (-1)^r \rho \frac{1-\sigma^r}{r(1-\sigma)} \sum_{A \in \mathcal{A}_r} \prod_{k \in A} \frac{1-z_k}{1-z_k\sigma}. \tag{2.15}$$

The corresponding probability distribution is given by

$$P(X_0^1 = n_1, \ldots, X_0^m = n_m) \quad (2.16)$$
$$= (1-\sigma)^{|\Phi_{\vec{n}}|} \sigma^{\left(\sum_{k=1}^m n_k\right) - |\Phi_{\vec{n}}|} \sum_{r=|\Phi_{\vec{n}}|}^{m} (-1)^{r-|\Phi_{\vec{n}}|} \rho \frac{1-\sigma^r}{r(1-\sigma)} \binom{m-|\Phi_{\vec{n}}|}{r-|\Phi_{\vec{n}}|}$$

for $\Phi_{\vec{n}} \neq \varnothing$*, and*

$$P(X_0^1 = 0, \ldots, X_0^m = 0) = 1 + \sum_{r=1}^{m} (-1)^r \binom{m}{r} \rho \frac{1-\sigma^r}{r(1-\sigma)}. \quad (2.17)$$

Proof: Since all the service rates are the same we also have $\rho_k = \rho$ and $\sigma_k = \sigma$ for $k = 1, 2, \ldots, m$. Also note that

$$C_A = \rho \frac{1-\sigma^r}{r(1-\sigma)} \quad \text{for all} \quad A \in \mathcal{A}_r. \quad (2.18)$$

The characteristic function (2.15) follows by using (2.18) in (2.5). In order to derive (2.16) it suffices to use (2.18) in (2.7) to obtain

$$P(X_0^1 = n_1, \ldots, X_0^m = n_m) = \sum_{r=1}^{m} (-1)^{r+|\Phi_{\vec{n}}|} \rho \frac{1-\sigma^r}{r(1-\sigma)} \sum_{A \in \mathcal{A}_r} \mathbf{1}(\Phi_{\vec{n}} \subseteq A) \prod_{k \in \Phi_{\vec{n}}} (1-\sigma)\sigma^{n_k - 1}.$$

An elementary combinatorial argument gives

$$\sum_{A \in \mathcal{A}_r} \mathbf{1}(\Phi_{\vec{n}} \subseteq A) = \binom{m-|\Phi_{\vec{n}}|}{r-|\Phi_{\vec{n}}|}$$

and hence (2.16) follows. From these considerations, and the fact that $\Phi_{\vec{n}} = \varnothing$ for $\vec{n} = (0, 0, \ldots, 0)$, (2.17) is also obtained. □

2.4 The two-server system and further performance measures

To illustrate the above results we will apply them to a system with two servers ($m = 2$).

Proposition 9. *The stationary number of customers in a synchronized D/M/1 system with two stations is given by*

$$\begin{aligned}
P(X_0^1 = n_1, X_0^2 = n_2) &= C_{\{1,2\}}(1-\sigma_1)\sigma_1^{n_1-1}(1-\sigma_2)\sigma_2^{n_2-1} && n_1 \geq 1, n_2 \geq 1, \\
P(X_0^1 = 0, X_0^2 = n_2) &= (1-\sigma_2)\sigma_2^{n_2-1}\left(\rho_2 - C_{\{1,2\}}\right) && n_2 \geq 1, \\
P(X_0^1 = n_1, X_0^2 = 0) &= (1-\sigma_1)\sigma_1^{n_1-1}\left(\rho_1 - C_{\{1,2\}}\right) && n_1 \geq 1, \\
P(X_0^1 = 0, X_0^2 = 0) &= 1 - \rho_1 - \rho_2 + C_{\{1,2\}},
\end{aligned}$$

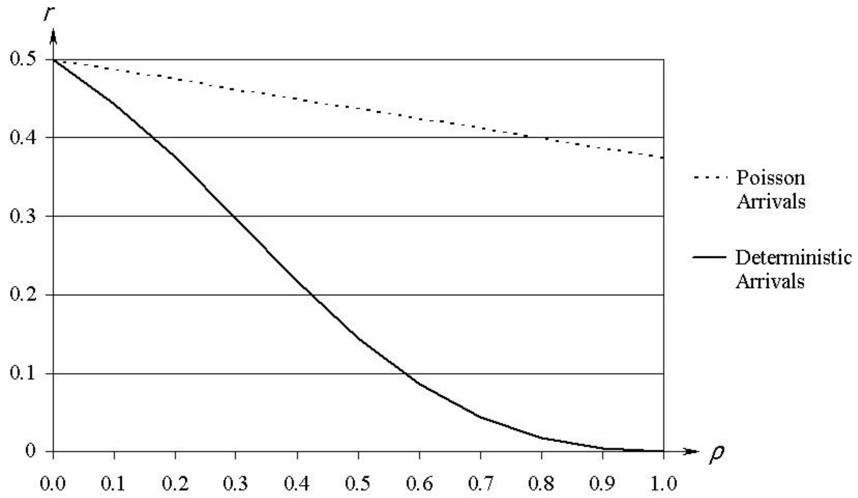

Fig. 2.1: The correlation coefficient, r, between the queue sizes in the two queues in the symmetric case as a function of the utilization ρ. Two plots are given, one for the model with deterministic arrivals and the other for the classical model with Poisson arrivals.

where

$$C_{\{1,2\}} := \frac{1-\sigma_1\sigma_2}{\rho_1^{-1}(1-\sigma_1)+\rho_2^{-1}(1-\sigma_2)}. \tag{2.19}$$

Proof: We apply the general result of theorem 1 noting that

$$C_{\{1\}} = \frac{1-\sigma_1}{\rho_1^{-1}(1-\sigma_1)} = \rho_1$$

and, similarly $C_{\{2\}} = \rho_2$. □

The correlation coefficient for the stationary number of customers in the two queues can be computed easily from the above stationary distribution and is given by

$$r = \sqrt{\frac{\rho_1\rho_2}{(1+\sigma_1-\rho_1)(1+\sigma_2-\rho_2)}}\left(\frac{(\rho_1\rho_2)^{-1}(1-\sigma_1\sigma_2)}{(1-\sigma_1)\rho_1^{-1}+(1-\sigma_2)\rho_2^{-1}}-1\right).$$

For the symmetric case, i.e. when $\mu_1 = \mu_2$ and hence $\rho_1 = \rho_2 = \rho$ and $\sigma_1 = \sigma_2 = \sigma$ we have

$$r = \frac{1}{2}\left(1-\frac{\rho}{1+\sigma-\rho}\right). \tag{2.20}$$

A plot of the correlation coefficient r as a function of ρ is given in figure 1.

It is of some interest to compare the correlation coefficient in the symmetric case to that of the classic FHW model. In fact the correlation coefficient in this case is

$$r = \frac{1}{2} - \frac{\rho}{8} \tag{2.21}$$

(see theorem 6.2 of [16]). As expected, the case with Poisson arrivals exhibits higher correlation between the two queues. More interesting perhaps is the heavy traffic behavior. As expected, in the case of deterministic arrivals examined in this chapter, the correlation between the two queues goes to zero as $\rho \to 1$. On the other hand, in the classic FHW model with Poisson arrivals the correlation coefficient goes to 3/8 as $\rho \to 1$. Of course, despite the fact that $\sigma = \rho$ in the case of Poisson arrivals, it would be mistaken to expect (2.20) to reduce to (2.21) in that case since the whole analysis leading to (2.20) is based on the assumption that arrivals are deterministic.

Finally we can use the results of theorem 1 together with the memoryless property of the exponential distribution in order to obtain the statistics for the workload process. For the sake of simplicity we present it for the case of a two–server station. The extension to the general m server model is obvious. If (W_t^1, W_t^2) is the workload vector at time t then we have the following

Proposition 10. *The stationary joint distribution of the workload in the two queues, $F(x_1, x_2) := P(W_0^1 \le x_1, W_0^2 \le x_2)$ is given by*

$$\begin{aligned} F(x_1, x_2) &= 1 - \rho_1 e^{-\mu_1(1-\sigma_1)x_1} - \rho_2 e^{-\mu_2(1-\sigma_2)x_2} \\ &\quad + \tfrac{1-\sigma_1\sigma_2}{(1-\sigma_1)\rho_1^{-1}+(1-\sigma_2)\rho_2^{-1}} e^{-\mu_1(1-\sigma_1)x_1-\mu_2(1-\sigma_2)x_2}. \end{aligned}$$

Proof: Start by conditioning on the number of customers present in the system at time 0, under the stationary probability measure P. Then

$$E\left[e^{-s_1 W_0^1 - s_2 W_0^2} \middle| X_0^1 = n_1, X_0^2 = n_2 \right] = \left(\frac{\mu_1}{s_1 + \mu_1} \right)^{n_1} \left(\frac{\mu_2}{s_2 + \mu_2} \right)^{n_2} \quad \text{for all } n_1, n_2 = 0, 1, 2, \ldots .$$

Taking into account the expression for the stationary distribution of the number of customers in the two queues we obtain, after some simplifications, the following expression for the joint Laplace transform of the stationary workload

$$\begin{aligned} Ee^{-s_1 W_0^1 - s_2 W_0^2} &= C_{\{1,2\}} \frac{\mu_1(1-\sigma_1)}{s_1 + \mu_1(1-\sigma_1)} \frac{\mu_2(1-\sigma_2)}{s_2 + \mu_2(1-\sigma_2)} + \frac{\mu_1(1-\sigma_1)}{s_1 + \mu_1(1-\sigma_1)} \left(\rho_1 - C_{\{1,2\}} \right) \\ &\quad + \frac{\mu_2(1-\sigma_2)}{s_2 + \mu_2(1-\sigma_2)} \left(\rho_2 - C_{\{1,2\}} \right) + 1 - \rho_1 - \rho_2 + C_{\{1,2\}} \end{aligned}$$

where $C_{\{1,2\}}$ is the constant given in (2.19). Straight-forward inversion of this transform completes the proof.

□

3. GATED QUEUES WITH AN INFINITE NUMBER OF SERVERS - BUSY PERIOD LENGTH AND NUMBER SERVED IN EACH STAGE

Here we return to the type of model discussed in Chapter 1. We consider gated, infinite server queues consisting of a waiting area and a service facility and operating under the following service protocol. Suppose initially that the service facility is empty and that there customers present in the waiting area. The gate opens admitting all waiting customers to the service facility, then closes again instantly and a *service stage* begins. Any customers arriving during this service stage remain in the waiting area. The service facility has an unlimited number of servers and thus all customers admitted for service are served in parallel. When all customers in the stage have completed service, they depart as a completed batch and the service stage ends. The gate opens admitting all customers in the waiting area and a new service stage begins immediately. If no customers are present, then the gate remains open until the arrival of the next customer, at which point a service stage (consisting of a single customer) is initiated.

Consecutive busy period lengths constitute a Markov chain in discrete time with continuous state space. The transition probability density is obtained. While the stationary density is not obtained in closed form, its moments are shown to satisfy an infinite linear system. An approximate solution is obtained by truncation. Similarly, the consecutive number of customers served in each busy period constitute a discrete time Markov chain with discrete (countable) state space. The transition probability matrix is obtained. Again the analysis results in a infinite system of equations and unknowns.

3.1 The Gated $M/GI/\infty$ System - Stage Length Density

Arrivals are Poisson with rate λ while the service requirements are assumed to be i.i.d. random variables with common distribution $G(x) = P\{\sigma_1 \leqslant x\}$. Let Y_n denote the duration of the nth stage. The sequence $\{Y_n\}$ constitutes a Discrete Time Continuous Space Markov chain. The transition density, of this Markov chain,

$f(y|x) := \mathbb{P}(Y_{n+1} \in dy \mid Y_n = x)$ is then given by

$$\begin{aligned} f(y|x) &= \sum_{k=1}^{\infty} k g(y) G^{k-1}(y) \frac{(\lambda x)^k}{k!} e^{-\lambda x} + e^{-\lambda x} g(y) \\ &= g(y)\lambda x e^{-\lambda x} \left\{ \sum_{k=1}^{\infty} \frac{(G(y)\lambda x)^{k-1}}{(k-1)!} \right\} + e^{-\lambda x} g(y) \end{aligned}$$

which, using the notation $\overline{G}(y) := 1 - G(y)$, gives

$$f(y|x) = \lambda x g(y) e^{-\lambda x \overline{G}(y)} + e^{-\lambda x} g(y). \quad (3.1)$$

The invariant density f of $\{Y_n\}$ satisfies the relationship

$$f(y) = \int_0^{\infty} f(x) f(y|x) dx. \quad (3.2)$$

Equation (3.2) is not easy to solve. We will obtain here a light-traffic solution in the form of series in λ as follows. We begin by expressing $f(y|x)$ as a power series in λ: From (3.1),

$$f(y|x) = \lambda x g(y) \sum_{k=0}^{\infty} (-1)^k \frac{(\lambda x)^k}{k!} \overline{G}(y)^k + g(y) \sum_{k=0}^{\infty} (-1)^k \frac{(\lambda x)^k}{k!}$$

Thus

$$\begin{aligned} f(y|x) = g(y) \Bigg[1 &+ (\lambda x)^2 \left(\frac{1}{2} - \overline{G}(y) \right) + \frac{(\lambda x)^3}{2!} \left(\frac{1}{3} - \overline{G}(y)^2 \right) \\ &+ \frac{(\lambda x)^4}{3!} \left(\frac{1}{4} - \overline{G}(y)^3 \right) + \cdots \Bigg]. \end{aligned} \quad (3.3)$$

Proposition 11. *Let f denote the invariant density of the stationary stage length, and suppose that $\beta_k := \int_0^{\infty} x^k f(x) dx$, $k = 1, 2, \ldots$, denote its moments. Suppose that σ_i are i.i.d. random variables with density g and define the quantities*

$$\gamma_{m,k} := \mathbb{E}[\min(\sigma_1, \sigma_2, \ldots, \sigma_k)^m] \quad (3.4)$$

Then the moments $\{\beta_i\}$, $i = 2, \ldots,$ satisfy the infinite linear system

$$\beta_i = \gamma_{i,1} + \frac{\lambda^2 \beta_2}{2!} (\gamma_{i,1} - \gamma_{i,2}) + \frac{\lambda^3 \beta_3}{3!} (\gamma_{i,1} - \gamma_{i,3}) + \cdots + \frac{\lambda^j \beta_j}{j!} (\gamma_{i,1} - \gamma_{i,j}) + \cdots \quad i = 2, 3, 4, \ldots \quad (3.5)$$

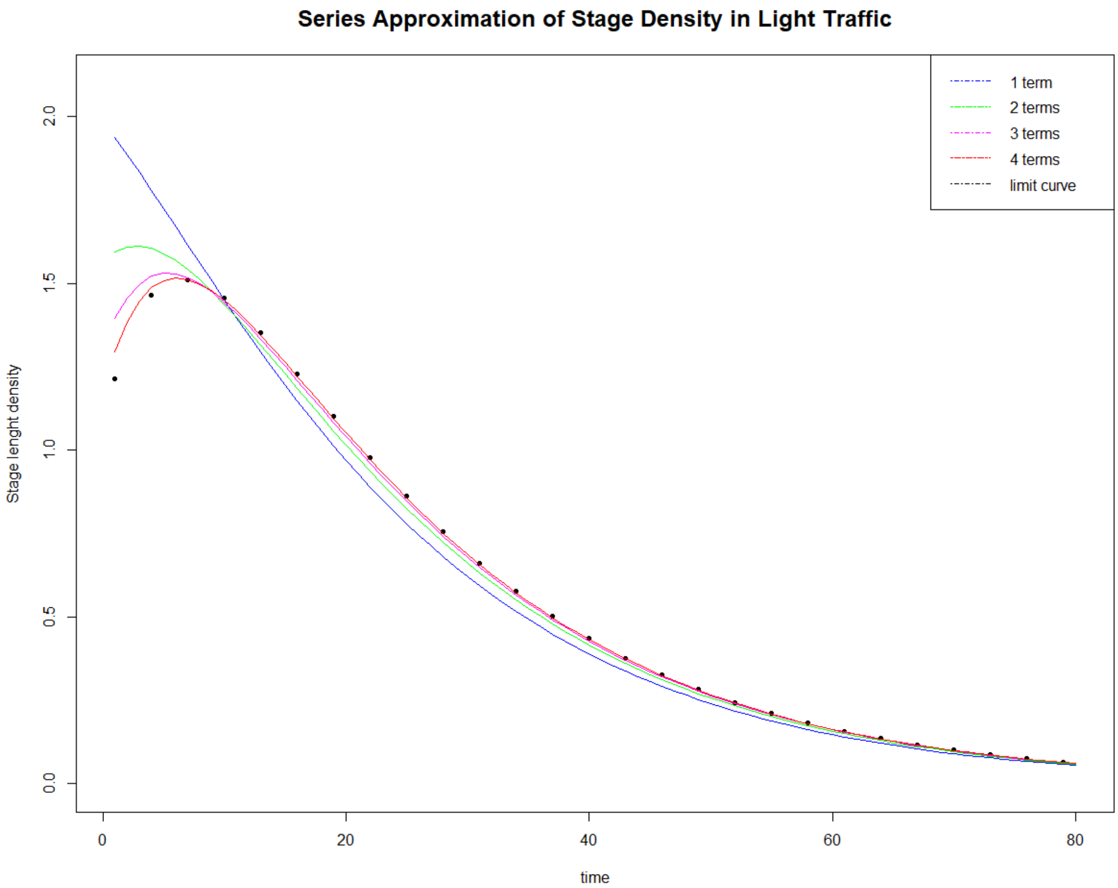

Fig. 3.1: A series representation for the stationary density of the length of a stage in a gated $M/G/\infty$ queue in light traffic.

The invariant density f of the stage length can be expressed in terms of the moments $\{\beta_i\}$ and the density g and complementary distribution function $\overline{G}$ of the service time by the series

$$f(y) = g(y)\left[1 + \frac{\lambda^2\beta_2}{2!}\left(1 - 2\overline{G}(y)\right) + \frac{\lambda^3\beta_3}{3!}\left(1 - 3\overline{G}(y)^2\right) + \frac{\lambda^4\beta_4}{4!}\left(1 - 4\overline{G}(y)^3\right) + \cdots\right]. \quad (3.6)$$

Proof. We begin with the relationship

$$\int_0^\infty y^m g(y)\left(1 - k\overline{G}(y)^{k-1}\right) dy = \gamma_{m,1} - \gamma_{m,k} \quad (3.7)$$

which holds because $kg(y)\overline{G}(y)^{k-1}$ is the density of $\min(\sigma_1, \sigma_2, \ldots, \sigma_k)$. From

$$\beta_i = \int_0^\infty y^i f(y) dy = \int_0^\infty \left(\int_0^\infty y^i f(y|x) dy \right) f(x) dx$$

taking into account (3.2) and (3.3) we have

$$\beta_i = \int_0^\infty y^i \left(\int_0^\infty \left[1 + \sum_{k=2}^\infty \frac{(\lambda x)^k}{k!} \left(1 - k\overline{G}(y)^{k-1}\right) \right] f(x) dx \right) g(y)\, dy$$

or

$$\beta_i = \int_0^\infty y^i \left[1 + \sum_{k=2}^\infty \frac{\lambda^k \beta_k}{k!} \left(1 - k\overline{G}(y)^{k-1}\right) \right] g(y)\, dy.$$

$$\beta_i = \int_0^\infty y^i g(y)\, dy + \sum_{k=2}^\infty \frac{\lambda^k \beta_k}{k!} \int_0^\infty y^i g(y) \left(1 - k\overline{G}(y)^{k-1}\right) dy.$$

Taking into account (3.7), we obtain (3.5). □

Written in extensive form (3.5)is an infinite system of linear equations for all the moments of the stage length of order 2 and above:

$$\begin{aligned}
\beta_2 &= \gamma_{2,1} + \frac{\lambda^2 \beta_2}{2!} (\gamma_{2,1} - \gamma_{2,2}) + \frac{\lambda^3 \beta_3}{3!} (\gamma_{2,1} - \gamma_{2,3}) + \frac{\lambda^4 \beta_4}{4!} (\gamma_{2,1} - \gamma_{2,4}) + \cdots \\
\beta_3 &= \gamma_{3,1} + \frac{\lambda^2 \beta_2}{2!} (\gamma_{3,1} - \gamma_{3,2}) + \frac{\lambda^3 \beta_3}{3!} (\gamma_{3,1} - \gamma_{3,3}) + \frac{\lambda^4 \beta_4}{4!} (\gamma_{3,1} - \gamma_{3,4}) + \cdots \\
\beta_4 &= \gamma_{4,1} + \frac{\lambda^2 \beta_2}{2!} (\gamma_{4,1} - \gamma_{4,2}) + \frac{\lambda^3 \beta_3}{3!} (\gamma_{4,1} - \gamma_{4,3}) + \frac{\lambda^4 \beta_4}{4!} (\gamma_{4,1} - \gamma_{4,4}) + \cdots \\
&\vdots
\end{aligned} \tag{3.8}$$

Assuming that the moment sequence can be determined from the above system, the invariant density can be obtained from the series (3.6). The discussion regarding the existence and uniqueness of the solution of the above infinite system, as well as its approximation by considering a truncated version of (3.8) and the convergence of the infinite system (3.6) will be discussed in the sequel.

3.1.1 Exponential Service Times

We continue the above discussion in the special case where the service time distribution is exponential, i.e. $g(x) := \mu e^{-\mu x}$, $x \geq 0$. Then, since the minimum of j independent exponential random variables $\min(\sigma_1, \ldots, \sigma_j)$ with $\sigma_l \sim \exp(\mu)$, is also exponential with rate $j\mu$, we have

$$\gamma_{i,1} = \frac{i!}{\mu^i}, \qquad \gamma_{i,j} = \frac{i!}{j^i \mu^i} \ .$$

Taking this into account the infinite system of linear equations (3.5) is written as

$$\beta_i \ = \ \frac{i!}{\mu^i} + \frac{\lambda^2\beta_2}{2!}\left(\frac{i!}{\mu^i} - \frac{i!}{2^i\mu^i}\right) + \frac{\lambda^3\beta_3}{3!}\left(\frac{i!}{\mu^i} - \frac{i!}{3^i\mu^i}\right) + \cdots + \frac{\lambda^j\beta_j}{j!}\left(\frac{i!}{\mu^i} - \frac{i!}{j^i\mu^i}\right) + \cdots \qquad i = 2, 3, \ldots \tag{3.9}$$

Let us next define the quantities

$$y_i := \frac{\lambda^i \beta_i}{i!}, \qquad i = 2, 3, \ldots. \tag{3.10}$$

Multiplying both sides of (3.9) by $\frac{\lambda^i}{i!}$, setting $\rho := \lambda/\mu$, and using the definition (3.10) we obtain the system

$$\rho^{-i} y_i \ = \ 1 + y_2\left(1 - 2^{-i}\right) + y_3\left(1 - 3^{-i}\right) + \cdots + y_j\left(1 - j^{-i}\right) + \cdots, \qquad i = 2, 3, 4, \ldots. \tag{3.11}$$

We will show that the solution of the finite linear system approximates the solution of the infinite system [47]. Then, the invariance stage length density is given by

$$f(x) \ = \ \mu e^{-\mu x}\left(1 + \sum_{k=2}^{\infty} y_k\left(1 - k e^{-\mu k x}\right)\right), \quad x > 0. \tag{3.12}$$

The mean stage length in this case is given by

$$\int_0^\infty x f(x) dx \ = \ \frac{1}{\mu}\left(1 + \sum_{k=2}^{\infty} y_k \frac{k-1}{k}\right) \ . \tag{3.13}$$

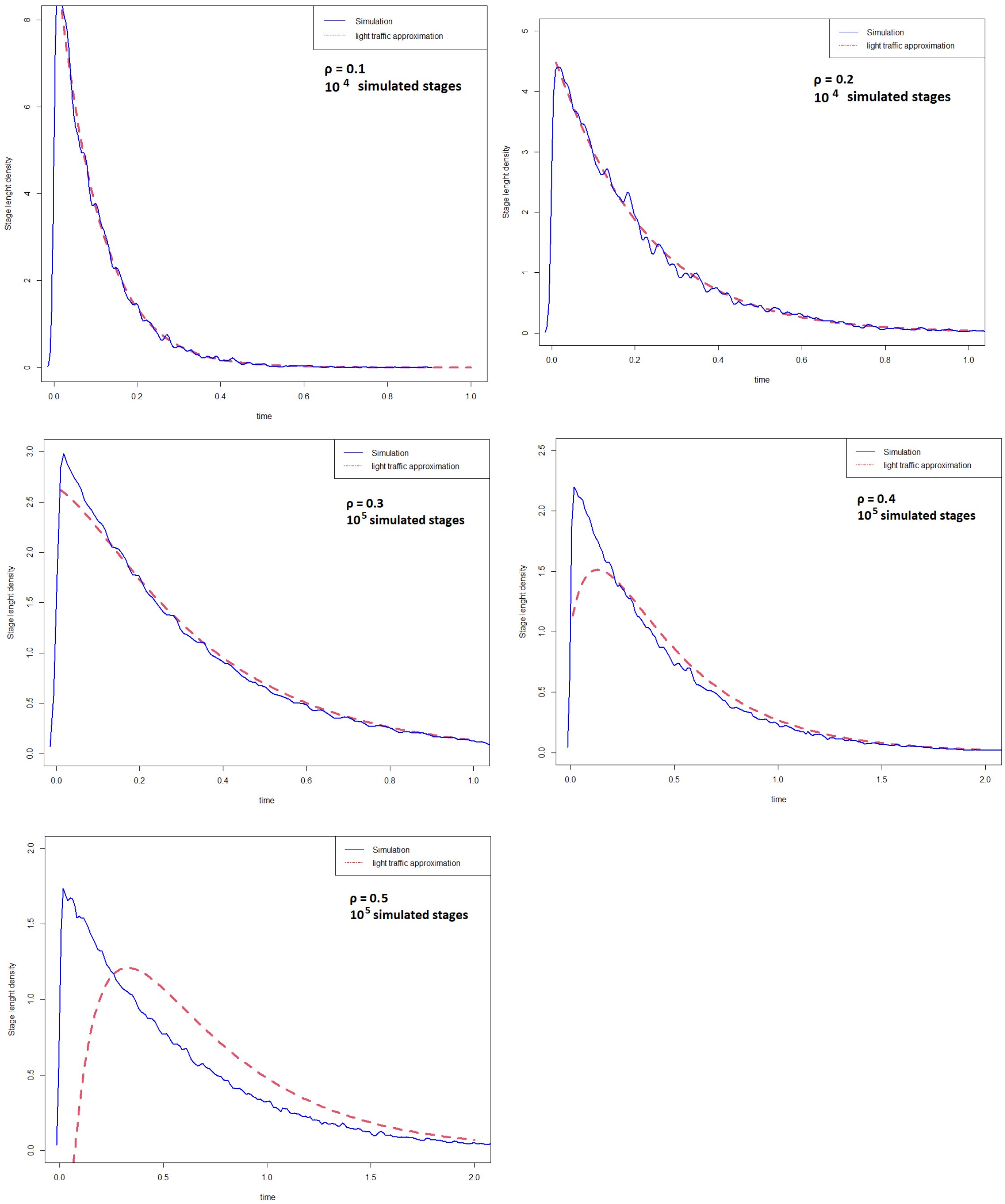

Fig. 3.2: A comparison of the actual density for the invariant density of the stage length on an $M/M/\infty$ queue in light traffic, as obtained by simulation with the solution given by (3.12). Notice the deterioration and, eventually, the invalidity of the quality of the light traffic approximation when ρ becomes larger.

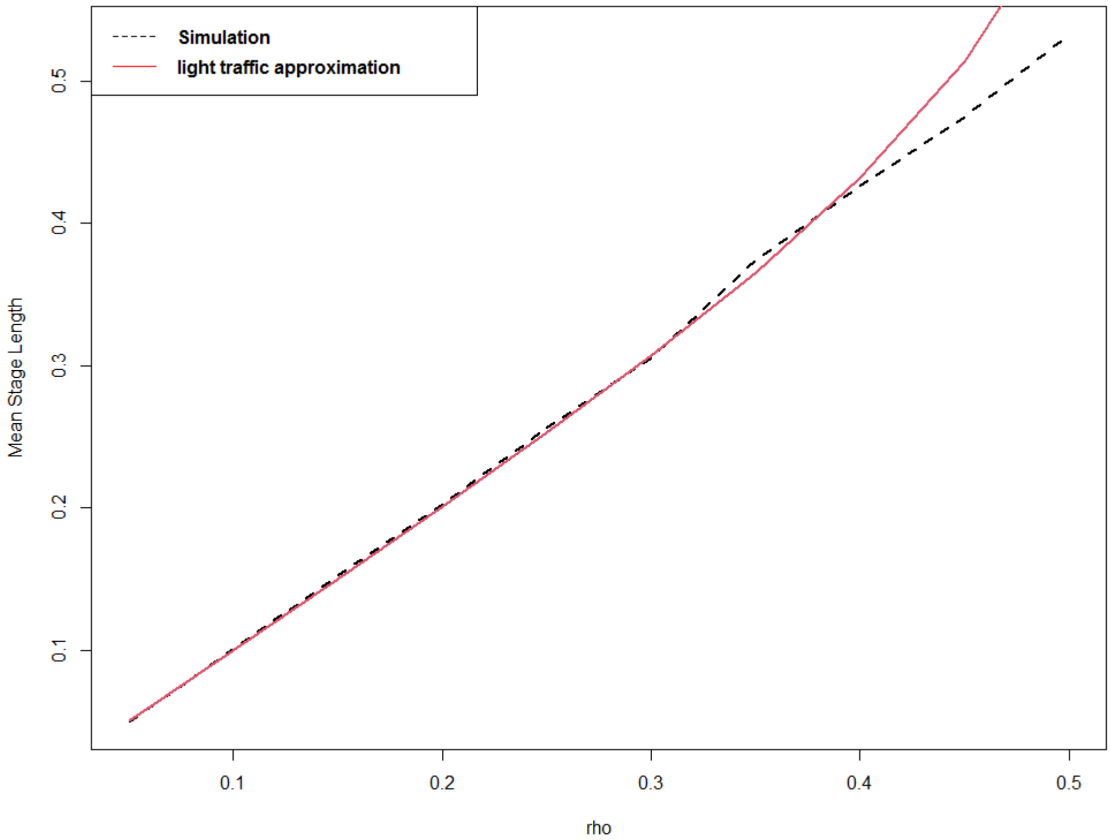

Fig. 3.3: The mean stage length in the Gated M/M/∞. The dotted line is obtained by simulating 10^4 stages. The red line, gives the light traffic approximation (equation 3.13). As the traffic intensity ρ increases, the quality of the light traffic approach deteriorates.

3.2 The Synchronized Gated $GI/M/\infty$ System - Number of Customers Served in a Stage

We now turn to an infinite server system closely related to that of section 3.1, in which customers arrive according to a renewal process $\{T_n\}$ and service times are independent exponential with rate μ. The gate mechanism operates essentially in the same fashion as in the system of section 3.1 with the exception that when the customers in a stage (who are again served in parallel) complete service and leave, the gate remains closed until the next arrival epoch. At this point, the customers that were in the waiting area, together with the newly arrived customer, are all admitted for service.

Suppose that $T_0 = 0$ and that the interarrival times $\tau_i := T_i - T_{i-1}$, $i = 1, 2, \ldots$, are i.i.d. random variables with common distribution G. The service is carried out in stages as follows: The beginning of a service stage always coincide with an arrival epoch of a customer. Thus service stage initiation epochs form a point process that is a subset of the arrival point process $\{T_n\}$.

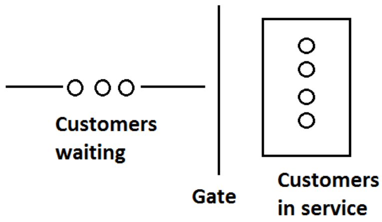

Fig. 3.4: The gated $GI/M/\infty$ queue

Let $\{X_n\}$ denote the number of customers served in the nth stage. Suppose also that $\{\sigma_i\}$ are independent, exponential random variables with rate μ. It is easy to see that $\{X_n\}$ is a Markov chain with state space $\{1, 2, 3, \ldots\}$ and transition probabilities given by

$$\begin{aligned} P(X_n = j \mid X_{n-1} = i) &= P\left(T_{j-1} < \max(\sigma_1, \sigma_2, \ldots, \sigma_i) \le T_j\right) \\ &= P\left(\max(\sigma_1, \sigma_2, \ldots, \sigma_i) \le T_j\right) - P\left(\max(\sigma_1, \sigma_2, \ldots, \sigma_i) \le T_{j-1}\right) \\ &= E(1 - e^{\mu T_j})^i - E(1 - e^{-\mu T_{j-1}})^i \\ &= \sum_{k=0}^{i} \binom{i}{k} (-1)^k E\left[e^{-k\mu T_j}\right] - \sum_{k=0}^{i} \binom{i}{k} (-1)^k E\left[e^{-k\mu T_{j-1}}\right] \end{aligned} \tag{3.14}$$

Since the arrivals are renewal, $T_j := \tau_1 + \cdots + \tau_j$ and $E[e^{-sT_j}] = \hat{G}(s)^j$ where $\hat{G}(s) = \int_0^\infty e^{-sx} dG(x)$ is the Laplace transform of the distribution G. Thus, from (3.14), with $P_{ij} := P(X_n = j \mid X_{n-1} = i)$,

$$P_{ij} = \sum_{k=0}^{i} \binom{i}{k} (-1)^k \hat{G}(k\mu)^{j-1} \left(\hat{G}(k\mu) - 1\right) \tag{3.15}$$

Proposition 12. *Suppose that the interarrival times have finite second moment, i.e.* $E\tau^2 < \infty$. *Then the Markov chain* $\{X_n\}$ *defined above is positive recurrent and its invariant distribution* $\{\pi_i\}$ *satisfies the equations*

$$\begin{aligned} \pi_i &= \sum_{j=1}^{\infty} \pi_j P_{ji} \qquad i = 1, 2, 3, \ldots \\ 1 &= \sum_{i=1}^{\infty} \pi_i \end{aligned} \tag{3.16}$$

Proof. To establish the positive recurrence we shall use again the Foster-Liapunov criterion of (Theorem 1 of Chapter 1). To this end we shall choose once more the functions

$$f(i) = i \quad \text{and} \quad V(i) = C \cdot i, \;\; i = 1, 2, \ldots, \tag{3.17}$$

where $C > 1$. Suppose that U is the renewal function associated with the arrival process, i.e. $U(t) := \sum_{k=1}^{\infty} P(T_k \leq t)$ for $t \geq 0$. If $\{Y_n\}$ denotes the sequence of stage lengths, as in the previous section, then

$$E[X_{n+1}|X_n, Y_n] \;=\; U(Y_n). \tag{3.18}$$

In a renewal process with increments having finite second moment the renewal function satisfies Lorden's inequality (see [1, p. 160]), namely

$$U(t) \;\leq\; \frac{t}{E\tau} + \frac{E[\tau^2]}{(E\tau)^2}.$$

Using Lorden's inequality we obtain the following inequality from (3.18)

$$E[X_{n+1}|X_n, Y_n] \;\leq\; \frac{Y_n}{E\tau} + \frac{E[\tau^2]}{(E\tau)^2}.$$

Given X_n, Y_n is the maximum of n independent exponential random variables with rate μ and thus

$$E[X_{n+1}|X_n] \;\leq\; \frac{1}{\mu E\tau}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) + \frac{E[\tau^2]}{(E\tau)^2}. \tag{3.19}$$

Let then $\rho := (\mu E\tau)^{-1}$, $b_0 := \frac{E[\tau^2]}{(E\tau)^2}$, and suppose $C > 1$. With these definitions

$$\sum_{j=1}^{\infty} P_{ij} C j \;=\; \frac{C}{\rho}\left(1 + \frac{1}{2} + \cdots + \frac{1}{i}\right) + C b_0 \quad \text{for all } i \in \mathbb{N}. \tag{3.20}$$

Define the function $h : \mathbb{N} \to \mathbb{R}$ via

$$h(i) \;:=\; \frac{C-1}{C}\, i - \frac{1}{\rho}\left(1 + \frac{1}{2} + \cdots + \frac{1}{i_0}\right) - b_0. \tag{3.21}$$

Since $h(i+1) - h(i) = \frac{C-1}{C} - \frac{1}{\rho\, i} > 0$ when i is large enough it follows that there exists N such that $i > N$ implies that $h(i) > 0$ or equivalently, from (3.21),

$$\frac{C}{\rho}\left(1 + \frac{1}{2} + \cdots + \frac{1}{i}\right) + C b_0 < (C-1)\, i, \qquad \text{for } i > N. \tag{3.22}$$

From (3.20), (3.22),

$$\sum_{j=1}^{\infty} P_{ij} C j \;=\; \frac{C}{\rho}\left(1 + \frac{1}{2} + \cdots + \frac{1}{i}\right) + b_0 < Ci - i. \tag{3.23}$$

This is indeed equation (1.8) with the choice (3.17). Taking $b := \frac{C}{\rho}\left(1+\frac{1}{2}+\cdots+\frac{1}{N}\right)+ Cb_0$ we also see from (3.20) that

$$\sum_{j=1}^{\infty} P_{ij}Cj \;=\; \frac{C}{\rho}\left(1+\frac{1}{2}+\cdots+\frac{1}{i}\right)+b_0 \;\leq\; b \qquad \text{for } i \leq N. \tag{3.24}$$

Therefore, by Theorem 1, we conclude that the Markov Chain $\{X_n\}$ is positive recurrent with invariant distribution given by (3.16). □

Suppose that $\{X_n\}$ is a stationary version of the Markov chain and let

$$f(z) \;=\; \sum_{i=1}^{\infty} \pi_i z^i \;=\; E[z^{X_0}].$$

denote the corresponding probability generating function. Instead of attempting to find a solution to the system (3.16) we will concentrate on the probability generating function f. Recall that the descending factorial moments, and the derivatives of f at evaluated at $z = 1$ are related via

$$\sum_{n=1}^{\infty} n(n-1)(n-2)\cdots(n-k+1)\pi_n \;=\; \frac{1}{k!}f^{(k)}(1)$$

We will obtain an infinite system which is satisfied by the descending factorial moments $f^{(m)}(1)$, $m = 1, 2, \ldots$ as described in the following

Proposition 13. *Define the quantities*

$$x_m := \frac{f^{(m)}(1)}{m!}, \quad \text{and} \quad a_{mk} := \frac{\hat{G}(k\mu)^{m-1}}{(1-\hat{G}(k\mu))^m}, \quad k, m = 1, 2, \ldots. \tag{3.25}$$

Then

$$x_m = \sum_{k=1}^{\infty} x_k (-1)^{k-1} a_{mk}, \qquad m = 1, 2, \ldots \tag{3.26}$$

Proof. Multiplying (3.15) by z^j and summing over j we obtain

$$\begin{aligned} E[z^{X_1} \mid X_0 = i] \;&=\; \sum_{j=1}^{\infty} z^j \sum_{k=1}^{\infty} \binom{i}{k} (-1)^k \hat{G}^{j-1}(k\mu)[\hat{G}(k\mu)-1] \\ &=\; \sum_{k=1}^{\infty} \binom{i}{k} (-1)^k [\hat{G}(k\mu)-1] \sum_{j=1}^{\infty} z^j \hat{G}^{j-1}(k\mu). \end{aligned} \tag{3.27}$$

Since the geometric series in the last sum converges (at least for $|z| < 1$) (3.27) becomes

$$E[z^{X_1} \mid X_0] = \sum_{k=1}^{\infty} \binom{X_0}{k} (-1)^k \frac{\hat{G}(k\mu)-1}{1-\hat{G}(k\mu)z} z$$

Taking expectation with respect to X_0, in the above equation (and interchanging the summation and the expectation) we obtain the following expression for the generating function of the stage duration:

$$E[z^{X_0}] \;=\; \sum_{k=1}^{\infty} E[X_0(X_0-1)\cdots(X_0-k+1)]\frac{(-1)^k}{k!}\frac{\hat{G}(k\mu)-1}{1-\hat{G}(k\mu)z}z. \tag{3.28}$$

In the above we have used the fact that the binomial moments, the descending factorial moments, and the derivatives of f at evaluated at $z=1$ are related via

$$E\binom{X_0}{k} \;=\; \frac{1}{k!}E[X_0(X_0-1)\cdots(X_0-k+1)] \;=\; \frac{1}{k!}f^{(k)}(1).$$

We have also used the fact that $Ez^{X_0} = Ez^{X_1}$ (by stationarity). Thus, we can rewrite (3.28) as

$$f(z) \;=\; z\sum_{k=1}^{\infty} f^{(k)}(1)\frac{(-1)^{k-1}}{k!}\frac{1-\hat{G}(k\mu)}{1-\hat{G}(k\mu)z}. \tag{3.29}$$

This last expression represents the probability generating function of the steady-state number of customers served in a stage (with the exception of the customer who initiates the stage) as a linear combination of p.g.f.'s of geometric random variables, the kth of which has probability of success $1-\hat{G}(k\mu)$. Of course (3.29) involves the factorial moments of the unknown distribution on the right hand side. Note that, if

$$g(z) := z\frac{1-q}{1-qz},$$

then using Leibniz' rule

$$D^m g(z) = z\frac{m!(1-q)q^m}{(1-qz)^{m+1}} + m\frac{(m-1)!(1-q)q^{m-1}}{(1-qz)^m} = \frac{m!(1-q)q^{m-1}}{(1-qz)^{m+1}}$$

and

$$D^m g(1) = \frac{m!q^{m-1}}{(1-q)^m}.$$

Thus differentiating (3.29) m times with respect to z term by term and evaluating at $z=1$ gives

$$f^{(m)}(1) = \sum_{k=1}^{\infty} f^{(k)}(1)\frac{(-1)^{k-1}}{k!}\frac{m!\hat{G}(k\mu)^{m-1}}{(1-\hat{G}(k\mu))^m} \qquad m=1,2,\ldots \tag{3.30}$$

Dividing the above equation and using the definitions (3.25) we obtain the system (3.26). □

There remains of course the question of the solution of the system (3.26). Assuming that the sequence $\{x_m\}$ has been determined note from (3.29) that

$$\begin{aligned}\sum_{i=1}^{\infty}\pi_i z^i &= \sum_{k=1}^{\infty} x_k(-1)^{k-1} z\,\frac{1-\hat{G}(k\mu)}{1-\hat{G}(k\mu)z}\\ &= \sum_{k=1}^{\infty} x_k(-1)^{k-1}\sum_{i=1}^{\infty} z^i\left(1-\hat{G}(k\mu)\right)\hat{G}(k\mu)^{i-1}\\ &= \sum_{i=1}^{\infty} z^i\sum_{k=1}^{\infty} x_k(-1)^{k-1}\left(1-\hat{G}(k\mu)\right)\hat{G}(k\mu)^{i-1}. \qquad (3.31)\end{aligned}$$

Thus we have the following expression for the stationary distribution (contingent upon the solution of the system (3.26)

$$\pi_i = \sum_{k=1}^{\infty} x_k(-1)^{k-1}\left(1-\hat{G}(k\mu)\right)\hat{G}(k\mu)^{i-1}, \quad i=1,2,\ldots. \qquad (3.32)$$

Clearly, from (3.25), the quantities (x_k) are positive and hence (3.32) gives the stationary distribution as an alternating sum of geometric probabilities.

3.2.1 The light traffic case

The infinite system of equations which is satisfied by the x_m must be complemented by an additional condition that will give a non-homogeneous system. The approach we follow provides a solution in the light traffic case, which we define in this context by means of the condition

$$\hat{G}(\mu) < \frac{1}{2}. \qquad (3.33)$$

In particular, if the arrival process is Poisson (λ) and if we set $\rho := \lambda/\mu$ then $\hat{G}(\mu) = \frac{\lambda}{\lambda+\mu} = \frac{\rho}{1+\rho} < \frac{1}{2}$ or, equivalently, $\rho < 1$.

This condition results from the requirement that the infinite sums $\sum_{m=0}^{\infty} 2^m\hat{G}(k\mu)^m$ converge for each k which is equivalent to the statement that the power series for $f(z)$ around the point $z=1$ has radius of convergence at least one. Hence in the power series

$$f(z) = \sum_{m=0}^{\infty}\frac{(z-1)^m}{m!}f^{(m)}(1)$$

we may take $z=0$. Clearly $f(0)=0$ (since a service stage consists of at least one customer). Thus in our notation $0 = 1+\sum_{m=1}^{\infty} x_m(-1)^m$ (because $\frac{f^{(0)}(1)}{0!} = f(1) = 1$)

and hence

$$-1 \;=\; \sum_{m=1}^{\infty} x_m(-1)^m \tag{3.34}$$

For $m = 1$ in (3.26), taking into account (3.25), we have

$$x_1 \;=\; \sum_{k=1}^{\infty} x_k(-1)^{k-1}\frac{1}{1-\hat{G}(k\mu)} \;=\; -\sum_{k=1}^{\infty} x_k(-1)^k + \sum_{k=1}^{\infty} x_k(-1)^k \frac{\hat{G}(k\mu)}{1-\hat{G}(k\mu)}$$

and hence we obtain the system

$$\begin{aligned} -1 \;&=\; x_1\frac{1-2\hat{G}(\mu)}{1-\hat{G}(\mu)} + \sum_{k=2}^{\infty} x_k(-1)^{k-1}\frac{\hat{G}(k\mu)}{1-\hat{G}(k\mu)} \qquad (3.35)\\ 0 \;&=\; x_m\left(\frac{(-1)^{m-1}\hat{G}(m\mu)^{m-1}}{(1-\hat{G}(m\mu))^m} - 1\right) + \sum_{\substack{k=1\\k\neq m}}^{\infty} x_k(-1)^{k-1}\frac{\hat{G}(k\mu)^{m-1}}{(1-\hat{G}(k\mu))^m},\\ &\qquad m = 2, 3, \ldots. \end{aligned}$$

In particular, when the arrival process is Poisson (λ), with $\hat{G}(s) = \frac{\lambda}{\lambda+s}$,

$$\begin{aligned} -1 \;&=\; x_1(1-\rho) + \sum_{k=2}^{\infty} x_k(-1)^{k-1}\frac{\rho}{k}\\ 0 \;&=\; x_m\left(\left(1+\frac{\rho}{m}\right)\left(-\frac{\rho}{m}\right)^{m-1} - 1\right) + \sum_{\substack{k=1\\k\neq m}}^{\infty} x_k(-1)^{k-1}\left(1+\frac{\rho}{k}\right)\left(\frac{\rho}{k}\right)^{m-1},\\ &\qquad m = 2, 3, \ldots. \qquad (3.36) \end{aligned}$$

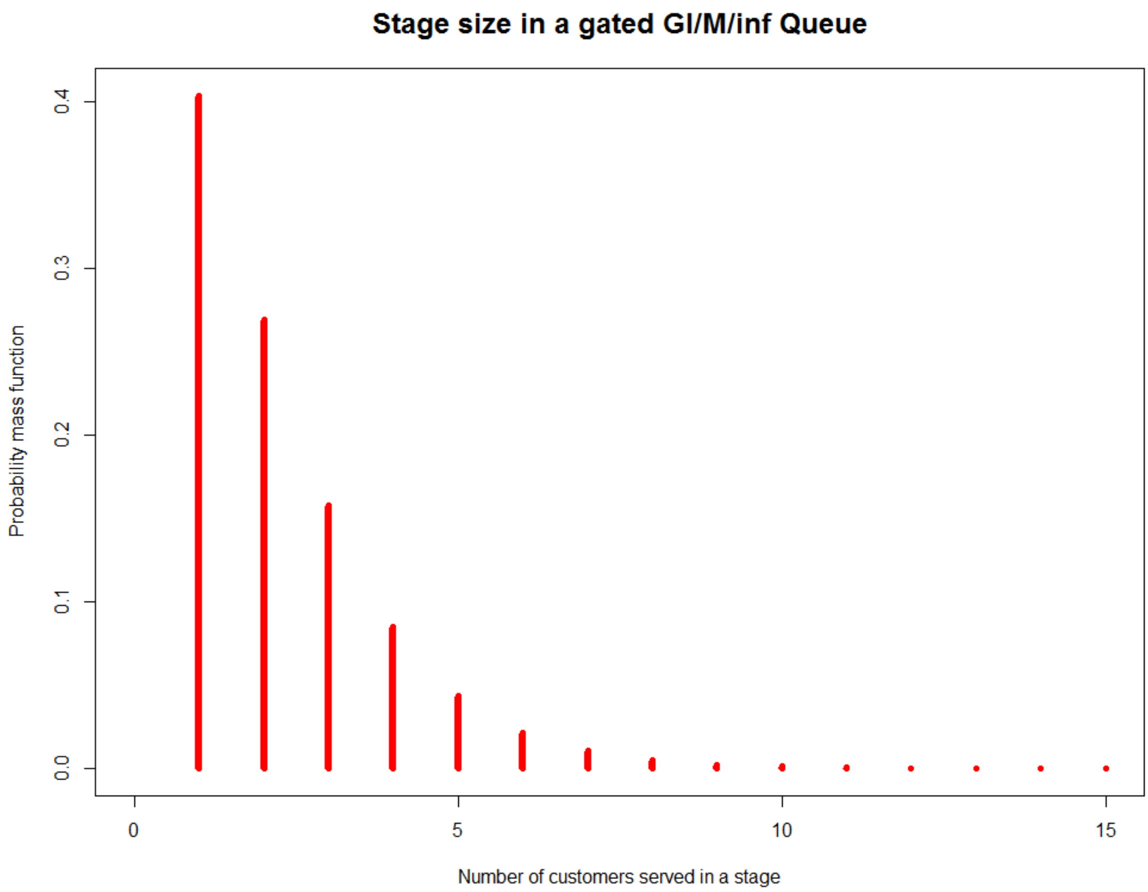

Fig. 3.5: Arrivals are Poisson (λ) and $\rho := \lambda/\mu = 0.85$. The system (3.26) is truncated at $N = 100$ and similarly 100 terms are taken in the series (3.32).

3.3 $D/M/\infty$ Queues

In this section we analyze the D/M/∞ queueing system. While the Markov chain analysis presented is well known at least since the early 60's (see Takacs [50]) the expression of the steady state distribution in terms of q–series (see seems to be new. Suppose that customers arrive singly at deterministic times $T_n = na$, $n \in \mathbb{Z}$, where $a > 0$ is the fixed interarrival time. Each customer, independent of anything else remains in the system for an exponentially distributed random variable with rate μ and then departs. Then it is easy to see that the number of customers in the system just prior to each arrival, $\{X_n; n \in \mathbb{Z}\}$ is an irreducible Markov Chain on the state space $\{0, 1, 2, \ldots, \}$ with

Transition Probability Matrix

$$\begin{bmatrix} p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ p^2 & 2pq & q^2 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ p^3 & 3p^2q & 3pq^2 & q^3 & 0 & 0 & 0 & 0 & 0 & \ldots \\ p^4 & 4p^3q & 6p^2q^2 & 4pq^3 & q^4 & 0 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & & \\ p^n & \binom{n}{1} p^{n-1}q & \binom{n}{2} p^{n-2}q^2 & \cdots & \binom{n}{k} p^{n-k}q^k & \cdots & \binom{n}{n-1} pq^{n-1} & q^n & 0 & \ldots \\ \vdots & & & & & & & & & \end{bmatrix}$$

The equilibrium equations can be expressed as

$$\pi_k = \sum_{n=0}^{\infty} \pi_{k-1+n} \binom{k+n}{n} q^k p^n, \qquad k = 1, 2, 3, \ldots. \tag{3.37}$$

Thus, if we denote by $\Pi(z) := \sum_{k=0}^{\infty} \pi_k z^k$ multiplying both sides of the above equation by z^k and summing term by term we obtain

$$\Pi(z) - \pi_0 \;=\; \sum_{k=1}^{\infty} z^k \sum_{n=0}^{\infty} \pi_{k-1+n} \binom{k+n}{n} q^k p^n \;=\; \sum_{m=0}^{\infty} \pi_m \left((p+qz)^{m+1} - p^{m+1} \right)$$

In the above string of equalities we have used the substitutions $l = k-1$ and $m = n+l$. We thus have

$$\Pi(z) - \pi_0 = (p+qz)\Pi(p+qz) - p\Pi(p).$$

Setting $z = 1$ and using the fact that $\Pi(1) = 1$ we see that $\pi_0 = p\Pi(p)$, a relation that can also be obtained from the equilibrium equations. Hence we see that the probability generating function of the stationary distribution satisfies the equation

$$\Pi(z) = (p+qz)\Pi(p+qz). \tag{3.38}$$

Using (3.38) recursively we obtain

$$\begin{aligned}\Pi(z) &= (p+qz)(p+qp+q^2z)\Pi(p+qp+q^2z) = \cdots \\ &= (p+qz)(p+qp+q^2z)\cdots(p+qp+\cdots+pq^{n-1}+q^nz) \\ &\qquad\qquad \times \Pi(p+qp+\cdots+pq^{n-1}+q^nz).\end{aligned}$$

Taking into account the fact that $p+qp+\cdots+pq^{k-1} = 1-q^k$, the above expression becomes

$$\Pi(z) = \Pi\left(1-q^n(1-z)\right)\prod_{k=1}^{n}\left(1-q^k(1-z)\right).$$

Letting $n \to \infty$, in view of the fact that

$$\lim_{n\to\infty} \Pi(1-q^n(1-z)) = \Pi(1) = 1$$

we obtain

$$\Pi(z) = \prod_{k=1}^{\infty}\left(1-q^k(1-z)\right). \qquad (3.39)$$

However it is known (see [39], p.9) that the function

$$F(z) = (1-qz)(1-q^2z)(1-q^3z)\cdots, \qquad |q|<1,$$

can be expanded in a power series

$$F(z) = A_0 + A_1z + A_2z^2 + A_3z^3 + \cdots,$$

where the coefficients A_n satisfy the recursive relationship $A_n(q^n-1) = A_{n-1}q^n$, $n = 1,2,3,\ldots$, $A_0 = 1$, which yields

$$A_n = (-1)^n\frac{q^{\frac{n(n+1)}{2}}}{(1-q)(1-q^2)\cdots(1-q^n)}, \qquad n = 1,2,3,\ldots \qquad (3.40)$$

Hence, with the above notation it follows that

$$\Pi(z) = \sum_{n=0}^{\infty} A_n(1-z)^n = \sum_{n=0}^{\infty}\frac{q^{\frac{n(n+1)}{2}}}{(1-q)(1-q^2)\cdots(1-q^n)}(z-1)^n. \qquad (3.41)$$

where we have used the convention that the empty product is equal to 1. From the above expression we obtain readily factorial moments.

The probability generating function of the number of departures between two arrivals is given by

$$\chi(z) = (q+pz)\Pi(q+pz) = (1-p(1-z))\prod_{k=1}^{\infty}(1-q^k(1-q-pz)) = \prod_{k=0}^{\infty}(1-pq^k(1-z))$$

or

$$\chi(z) \;=\; 1 + \sum_{n=1}^{\infty} \frac{p^n q^{\frac{n(n+1)}{2}}}{(1-q)(1-q^2)\cdots(1-q^n)}(z-1)^n. \qquad (3.42)$$

4. SYNCHRONIZED QUEUES WITH ORDERED SERVICE TIMES

4.1 *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 $c = 2$ and 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 [17], [16], [55]). 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 [16], and Flatto [15] have studied this system using complex analysis techniques. See also Fayolle and Iasnogorodsky [18] and Fayolle, Iasnogorodsky, and Malyshev [19]. Asymptotic results regarding this model have also been obtained using large deviation techniques by Weiss and Shwartz [42], [49]. See also the more recent papers by Badial et al. [7] and Kella and Boxma [27].

The fork–join queue when $c = 2$, arrivals are Poisson, and service requirements form an i.i.d sequence of exchangeable pairs of random variables has been studied in Baccelli [4]. We also mention the Taylor series expansion used in Ayhan and Baccelli [3] where the assumption of exponential service times is relaxed, and Baccelli, Makowski, and Swhartz [6]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 [5], [34]) 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. We focus on the case where the service times of parts are strongly ordered.

4.2 The rate conservation principle

On the probability space $(\Omega, \mathcal{F}, P)$ 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.

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 14. *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*

$$\phi(s_1, \ldots, s_c) = \frac{\sum_{i=1}^c s_i \psi_i(\ldots, s_{i-1}, s_{i+1}, \ldots)}{\sum_{i=1}^c s_i - \lambda\left(1 - \beta(s_1, \ldots, s_c)\right)}. \tag{4.1}$$

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*

$$\psi_i(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_c) := E[\mathbf{1}(W_0^i = 0) e^{-\sum_{j \neq i} s_j W_0^j}]. \tag{4.2}$$

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 $\{X_t; t \in \mathbb{R}\}$, where

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

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

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

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

$$\psi_i(s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_c)=E[\mathbf{1}(W_0^i=0)e^{-\sum_{j\neq i}s_jW_0^j}]. \qquad (4.3)$$

Thus from the above we obtain (4.1). ∎

Note that in the ordinary $M/G/1$ queue, $c=1$ and (4.1), (4.2), 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.

4.3 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 (4.2) 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. Thus (4.1) is written as

$$\phi(s_1,\ldots,s_c)=\frac{s_1\psi_1+\ \sum_{i=2}^{c}s_i\psi_i(s_1,\ldots,s_{i-1})}{\sum_{i=1}^{c}s_i-\lambda\left(1-\beta(s_1,\ldots,s_c)\right)}. \qquad (4.4)$$

If we set $\rho_i:=\lambda E^0\sigma_0^i$ then clearly

$$\psi_1=E[\mathbf{1}(W_0^1=0)]=1-\rho_1.$$

We are thus left with the problem of determining the $c-1$ unknown functions

$$\psi_i(s_1,\ldots,s_{i-1}):=E\left[\mathbf{1}(W_0^i=0)\,e^{\sum_{j=1}^{i-1}s_jW_0^j}\right],\qquad i=2,3,\ldots,c. \qquad (4.5)$$

It is clear that the problem of obtaining an expression for the unknown functions ψ_i hinges upon expressing the conditional expectations that define them in a more convenient form. As it turns out, the following lemma facilitates greatly this.

Lemma 15. *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*

$$\psi_i(s_1, \ldots, s_{i-1}) = (1 - \rho_i) E_i^0 e^{-\sum_{j=1}^{i-1} s_j W_0^j}. \tag{4.6}$$

Proof: If $\mathcal{F}_t^{S^i}$ is the internal history of the point process $\{S_n^i\}$ (see [9] for a definition) and $\mathcal{F}_t^{W^i} = \sigma - \{W_u^i; u \leq t\}$ the history of the process W^i, define the filtration $\mathbb{F}^i := \{\mathcal{F}_t^i; t \in R\}$ via $\mathcal{F}_t = \mathcal{F}_t^{S^i} \bigvee \mathcal{F}_t^{W^i}$. Then the $\mathbb{F}^i$–stochastic intensity of $\{S_n^i\}$ is given by

$$\alpha_t^i = \lambda \mathbf{1}(W_t^i = 0). \tag{4.7}$$

We now apply Papangelou's theorem (see [5], [40]): 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,

$$E_i^0 e^{-\sum_{j=1}^{i-1} s_j W_0^j} = \frac{E[\alpha_0^i \, e^{-\sum_{j=1}^{i-1} s_j W_0^j}]}{E\alpha_0^i}. \tag{4.8}$$

In view of the expression for the stochastic intensity in (4.7) the right hand side of the above equation becomes

$$E^i e^{-\sum_{j=1}^{i-1} s_j W_0^j} = \frac{E[\mathbf{1}(W_0^i = 0) \, e^{-\sum_{j=1}^{i-1} s_j W_0^j}]}{E[\mathbf{1}(W_0^i = 0)]} \tag{4.9}$$

and hence, from (4.6) and (4.9) we obtain (4.6). ∎

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 16. *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 $\{_n\}$ where $_n := (v_n^1, \ldots, v_n^{i-1})$ are i.i.d. vectors with joint Laplace transform $\widetilde{\beta}_{i-1}(s_1, \ldots, s_{i-1})$ which satisfies the equation*

$$\widetilde{\beta}_{i-1}(s_1, \ldots, s_{i-1}) = \beta_i(s_1, s_2, \ldots, s_{i-1}, \lambda(1 - \widetilde{\beta}_{i-1}(s_1, \ldots, s_{i-1})) - \textstyle\sum_{j=1}^{i-1} s_j). \quad (4.10)$$

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

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

Taking expectation, first with respect to K given σ_0^i and then with respect to σ_0^i we obtain (4.10). ∎

The question of whether equation (4.10) defines uniquely the multidimensional Laplace transform $\widetilde{\beta}(s_1, \ldots, s_{n-1})$ is addressed in detail in the appendix using a multidimensional analog of the classical Lagrange inversion theorem.

We are now ready to determine the functions ψ_i. To this end define

$$\phi_{i-1}^{(1)}(s_1, \ldots, s_{i-1}) := \frac{\psi_i(s_1, \ldots, s_{i-1})}{1 - \rho_i}, \qquad i = 2, 3, \ldots, c. \quad (4.11)$$

With this definition, (4.4) becomes

$$\phi(s_1, \ldots, s_c) = \frac{(1 - \rho_1)s_1 + \sum_{i=2}^{c} (1 - \rho_i) s_i \phi_i^{(1)}(s_1, \ldots, s_{i-1})}{\sum_{i=1}^{c} s_i - \lambda\left(1 - \beta(s_1, \ldots, s_c)\right)}. \quad (4.12)$$

Theorem 17. *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 where* $m = 1, 2, \ldots, c$

$$\phi_n^{(m)}(s_1, \ldots, s_n) = \frac{(1-\rho_1^{(m)})s_1 + \sum_{j=2}^n (1-\rho_j^{(m)})\phi_j^{(m+1)}(s_1, \ldots, s_{j-1})}{\sum_{j=1}^n s_j - \lambda(1-\beta^{(m)}(s_1, \ldots, s_n))}, \quad n = 1, 2, \ldots, c-m, \tag{4.13}$$

$$\beta_k^{(m+1)}(s_1, \ldots, s_k) = \beta_{k+1}^{(m)}\left(s_1, s_2, \ldots, s_k, \lambda(1-\beta_k^{(m+1)}(s_1, \ldots, s_k)) - \sum_{j=1}^k s_j\right). \tag{4.14}$$

4.4 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 18. *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 is given by*

$$E[e^{-s_1 W_0^1 - s_2 W_0^2}] = \frac{s_1(1-\rho_1)}{s_1 - \lambda + \lambda\widetilde{\beta}_1(s_1)}\left(1 - \lambda\frac{\beta(s_1, s_2) - \widetilde{\beta}(s_1)}{s_1 + s_2 - \lambda + \lambda\beta(s_1, s_2)}\right) \tag{4.15}$$

where $\widetilde{\beta}_1$ *is the unique solution of the equation*

$$\widetilde{\beta}_1(s_1) = \beta(s_1, \lambda(1-\widetilde{\beta}_1(s_1)) - s_1). \tag{4.16}$$

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

$$\phi(s_1, s_2) = \frac{s_1\psi_1 + s_2\psi_2(s_1)}{s_1 + s_2 - \lambda(1-\beta(s_1, s_2))}$$

where

$$\psi_2(s_1) := E\left[\mathbf{1}(W_0^2 = 0)\, e^{-s_1 W_0^1}\right],$$
$$\psi_1 := E\left[\mathbf{1}(W_0^1 = 0)\, e^{-s_2 W_0^2}\right] = E[\mathbf{1}(W_0^1 = 0)].$$

Based on the results of the previous section, using the same notation we have

$$\phi(s_1, s_2) = \frac{s_1(1-\rho_1) + s_2(1-\rho_2)\phi_2(s_1)}{s_1 + s_2 - \lambda(1-\beta(s_1, s_2))} \tag{4.17}$$

where

$$\phi_2(s_1) = \frac{1-\widetilde{\rho}_1}{1-\widetilde{\rho}_1 \, \frac{1-\widetilde{\beta}_1(s_1)}{s_1 \widetilde{m}_1}} \tag{4.18}$$

and $\widetilde{\beta}_1$ is the unique solution of the equation

$$\widetilde{\beta}_1(s_1) = \beta\left(s_1, \lambda(1-\widetilde{\beta}_1(s_1)) - s_1\right). \tag{4.19}$$

This implicit equation determines $\widetilde{\beta}_1(s)$. Also, $\widetilde{m}_1 = -\widetilde{\beta}'(0)$ and $\widetilde{\rho}_1 = \lambda\widetilde{m}_1$. In particular,

$$\widetilde{m}_1 := \frac{m_1 - m_2}{1-\lambda m_2}$$

and

$$\widetilde{\rho}_1 = \frac{\rho_1 - \rho_2}{1-\rho_2}. \tag{4.20}$$

Thus we have

$$\phi(s_1, s_2) = \frac{s_1(1-\rho_1)}{s_1 + s_2 - \lambda(1-\beta(s_1, s_2))} \frac{s_1 + s_2 - \lambda(1-\widetilde{\beta}(s_1))}{s_1 - \lambda(1-\widetilde{\beta}(s_1))}$$

5. APPENDIX

5.1 Infinite Linear Systems and Strictly Diagonally Dominant Matrices

Consider the infinite system of linear equations

$$\sum_{j=1}^{\infty} a_{ij}x_j \;=\; b_i, \qquad i = 1, 2, \ldots. \tag{5.1}$$

We suppose that $\{b_i\}$, $i = 1, 2, \ldots$, is a *bounded* sequence of real numbers. We will discuss sufficient conditions under which this system has a *unique, bounded solution* $\{x_i\}$, $i = 1, 2, \ldots$ following the results in [47] (see also [48]). Therein it is shown that, under these conditions, if one considers the sequence of *the truncated linear systems* indexed by N,

$$\sum_{j=1}^{N} a_{ij}x_j^N \;=\; b_i, \qquad i = 1, 2, \ldots, N, \tag{5.2}$$

then their solutions converge to a solution of (5.1). This result, besides being used to prove the existence and uniqueness of a bounded solution of (5.1), can be used in practice to provide an approximate solution of the system.

Our interest in this result here stems from the occurrence of the infinite systems (3.11) and (3.36) which we have solved numerically by truncation. In this Appendix we justify this procedure.

An infinite matrix $A := [a_{ij}]$, $i, j = 1, 2, 3, \ldots$, is *strictly diagonally dominant* if it satisfies the conditions

$$\sigma_i|a_{ii}| \;=\; \sum_{\substack{j=1\\ j\neq i}}^{\infty} |a_{ij}|, \qquad \text{with } 0 \le \sigma_i < 1, \;\; i = 1, 2, \ldots. \tag{5.3}$$

Consider also the following three conditions.

$$\sum_{i=1}^{\infty} \frac{1}{|a_{ii}|} < \infty, \tag{5.4}$$

$$\sum_{\substack{j=1 \\ j\neq i}}^{\infty} |a_{ij}| \leq M, \qquad \text{for some } M \text{ and all } i. \tag{5.5}$$

$$\sum_{i=1}^{\infty} |a_{ij}| < \infty \quad \text{for each } j. \tag{5.6}$$

The following theorem summarizes the results in [47] which are relevant in our treatment of the systems (3.11) and (3.36).

Theorem 19. *Suppose that the system (5.1) with bounded right hand side has a strictly diagonally dominant matrix A which in addition satisfies conditions (5.4) and (5.5). Then, for each N the truncated system (5.2) has a unique solution (x_i^N), $i = 1, 2, \dots, N$. For each fixed j, the sequence $\{x_j^N\}$, $N = j, j+1, \dots$ is Cauchy and thus converges to a limit x_j. The sequence $\{x_j\}$ is a bounded solution of (5.1). If, in addition, condition (5.6) is also satisfied then this bounded solution is unique.*

(Even when conditions (5.3)–(5.6) hold and thus the infinite system (5.1) has a unique bounded condition, it may also admit unbounded solutions as is shown in [47]. These solutions however do not arise as limits of the solution sequence of the truncated systems.)

5.2 The infinite linear systems of Chapter 3

5.2.1 The system (3.11)

We will begin by recasting the infinite system (3.11) in an equivalent form that will be amenable to treatment in the framework of the previous section. We begin by noting that

$$s := y_2 + y_3 + \cdots$$

is a finite quantity. Indeed,

$$E[e^{\lambda Y}] \;=\; E\sum_{n=0}^{\infty} \frac{\lambda^n Y^n}{n!} \;=\; \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}\beta_n \;=\; 1 + \lambda\beta_1 + s$$

and it can be seen that for sufficiently small λ the above infinite sum is finite. ($\lambda < \mu$ implies that $Ee^{\lambda Y} < \infty$ because we may compare the system in question with one where customers within a stage are served not in parallel but in series, one after another, in other words by comparing the system we study with a gated $M/M/1$ queue.)

Write now the system as

$$y_i \;=\; \rho^i + \rho^i \sum_{j=2}^{\infty} y_j \left(1 - j^{-i}\right) \;=\; \rho^i + \rho^i s - \sum_{j=2}^{\infty} y_j \left(\frac{\rho}{j}\right)^i, \quad i = 2, 3, \ldots \tag{5.7}$$

and add the above equations term by term for all i to obtain

$$s \;=\; \frac{\rho^2}{1-\rho}(1+s) - \sum_{i=2}^{\infty}\sum_{j=2}^{\infty} y_j \left(\frac{\rho}{j}\right)^i \;=\; \frac{\rho^2}{1-\rho}(1+s) - \sum_{j=2}^{\infty} y_j \frac{\rho^2}{j(j-\rho)}.$$

This equation, together with the system (5.7), results in a new system. Set $x_1 := s$ and $x_i := y_i$ for $i = 2, 3, \ldots$. Then we have the new system

$$\begin{aligned}\frac{1-\rho-\rho^2}{1-\rho} x_1 + \sum_{j=2}^{\infty} x_j \frac{\rho^2}{j(j-\rho)} \;&=\; \frac{\rho^2}{1-\rho},\\ \rho^{-i} x_i - x_1 + \sum_{\substack{j=2\\ j\neq i}}^{\infty} x_j \left(\frac{1}{j}\right)^i \;&=\; 1, \quad i = 2, 3, \ldots.\end{aligned} \tag{5.8}$$

We will use the results of section 5.1 to show that system (5.8), and thus the equivalent system (3.11), has a unique solution and that, furthermore, the solutions of the truncated systems resulting by keeping the first N equations and the first N unknowns, constitutes a Cauchy sequence converging to this solution. This justifies the elementary truncation approximation.

We first check the diagonal dominance condition (5.3). For $i = 1$ we must have

$$\frac{1-\rho-\rho^2}{1-\rho} \;>\; \sum_{j=2}^{\infty} \frac{\rho^2}{j(j-\rho)}. \tag{5.9}$$

Since $0 < \rho < 1$,

$$\sum_{j=2}^{\infty} \frac{\rho^2}{j(j-\rho)} < \rho^2 \sum_{j=2}^{\infty} \frac{1}{j(j-1)} = \frac{\rho^2}{2}$$

because the second series is a well known telescopic series. Therefore, if

$$\frac{1-\rho-\rho^2}{1-\rho} > \frac{\rho^2}{2}$$

then (5.9) holds. Equivalently if $\rho \in (0, \rho_0)$ where $\rho_0 \approx 0.34$ is the smallest positive root of the equation $1 - r - \frac{3}{2}r^2 + \frac{1}{2}r^3 = 0$ then (5.9) holds.

For $i \geq 2$ we must have

$$\rho^{-i} > 1 + \sum_{\substack{j=2 \\ j \neq i}}^{\infty} \left(\frac{1}{j}\right)^i \tag{5.10}$$

The series above always converges since $i \geq 2$. Also

$$\sum_{\substack{j=2 \\ j \neq i}}^{\infty} \left(\frac{1}{j}\right)^i - \frac{1}{i^i} \leq \left(\frac{\pi^2}{6} - 1\right) - \frac{1}{i^i}$$

with equality holding if $i = 2$. Hence (5.10) is satisfied for all $i \geq 2$ since $\rho^{-i} \geq \rho^{-2} > \frac{\pi^2}{6} \approx 1.64$ when $\rho < 0.34$.

Turning to (5.4)

$$\sum_{i=1}^{\infty} \frac{1}{|a_{ii}|} = \frac{1-\rho}{1-\rho-\rho^2} + \sum_{i=2}^{\infty} \rho^i < \infty$$

(since $\rho < 1$). Finally, (5.5) can also be seen to hold using very similar computations and inequalities. Hence Theorem 19 shows that, if $\rho \in (0, \rho_0)$ the unique solution of the system can be approximated by the solutions of the sequence of truncated finite systems. (In practice we have noticed experimentally that, as long as $\rho < 1$ the approximation procedure converges. There is no contradiction here. Theorem Appendix gives sufficient conditions for the approximation procedure to hold.)

5.2.2 The system (3.36)

Consider now the system (3.36) encountered in the analysis of the gated GI/M/∞ system. We will transform it in order to apply the theorem. Set $w_i := i{\cdot}x_i$, $i = 1, 2, \ldots$. Then

$$\begin{aligned} -1 &= w_1(1-\rho) + \sum_{i=2}^{\infty} w_j (-1)^{j-1} \frac{\rho}{j^2} \\ 0 &= w_i \left(\left(1 + \frac{\rho}{i}\right) \frac{(-1)^{i-1}\rho^{-1}}{i^i} - \frac{\rho^{-i}}{i}\right) + \sum_{\substack{j=1 \\ j \neq i}}^{\infty} w_j (-1)^{j-1} \left(1 + \frac{\rho}{j}\right) \frac{\rho^{-1}}{j^i}, \\ & \qquad i = 2, 3, \ldots. \end{aligned} \tag{5.11}$$

We will first establish (5.3) for the above system. For $i = 1$ it suffices to show that $1 - \rho > \sum_{j=2}^{\infty} \rho j^{-2}$. However this sum is equal to $\rho\left(\frac{\pi^2}{6} - 1\right)$. Therefore, in order for the strict diagonal dominance to hold we should have $1 - \rho > \rho\left(\frac{\pi^2}{6} - 1\right)$ or

$\rho < \frac{6}{\pi^2} \approx 0.608$. The following inequality holds

$$\frac{\rho^{-i}}{i} > \sum_{j=1}^{\infty} \left(1 + \frac{\rho}{j}\right) \frac{\rho^{-1}}{j^i} \tag{5.12}$$

Indeed

$$\begin{aligned} \sum_{j=1}^{\infty} \left(1 + \frac{\rho}{j}\right) \frac{\rho^{-1}}{j^i} &\geq \rho^{-1} \left(1 + \int_1^{\infty} x^{-i} dx\right) + 1 + \int_1^{\infty} x^{-(i+1)} dx \\ &= \rho^{-1} \frac{i}{i-1} + \frac{i+1}{i}. \end{aligned}$$

We can then show that $\frac{\rho^{-i}}{i} > \rho^{-1}\frac{i}{i-1} + \frac{i+1}{i}$ for all $i \geq 2$ provided that $\rho \in (0, 0.21)$. It also holds that

$$\sum_{i=1}^{\infty} \frac{1}{\left|\left(1 + \frac{\rho}{i}\right)(-1)^{i-1}\frac{\rho^{-1}}{i^i} - \frac{\rho^{-i}}{i}\right|} \leq \sum_{i=1}^{\infty} K\rho^i < \infty.$$

$$\sum_{j=1}^{\infty} |a_{ij}| = \sum_{j=1}^{\infty} \left(1 + \frac{\rho}{j}\right) \frac{\rho^{-1}}{j^i} < \infty.$$

$$\sum_{i=1}^{\infty} |a_{ij}| = \sum_{i=1}^{\infty} \left(1 + \frac{\rho}{j}\right) \frac{\rho^{-1}}{j^i} < \infty.$$

This then establishes (5.4), (5.5), and (5.6) and hence Theorem 19 holds.

BIBLIOGRAPHY

[1] S. Asmussen (2003). *Applied Probability and Queues,* 2nd edition, John Wiley.

[2] Adan, I.J.B.F., J. Wessels, and W.H.M. Zijm, (1993). Analysing multiprogramming queues by generating functions. *SIAM J. Appl. Math.*, **53,** 4, 1123-1131.

[3] H. Ayhan and F. Baccelli (2001). Expansions for joint Laplace transform of stationary waiting times in (max,+)-linear systems with Poisson input. *Queueing Systems,* **37**, 291-328.

[4] Baccelli, F. (1985). Two parallel queues created by arrivals with two demands: The M/G/2 symmetric case. Technical Report 426, INRIA-Rocquencourt.

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

[6] F. Baccelli, 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**, 629-660.

[7] E.S. Badila, O.J. Boxma, J.A.C. Resing, and E.M.M. Winands (2014). Queues and Risk Models with Simultaneous Arrivals. *Advances in Applied Probability,* **46**, 812-831.

[8] Boxma, O., G. Koole, and Z. Liu (1994). Queueing–Theoretic Solution Methods for Models of Parallel and Distributed Systems. In: *Performance Evaluation of Parallel and Distributed Systems,* eds. O.J. Boxma, G.M. Koole, CWI Tract 105, Amsterdam, pp. 1-24.

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

[10] Avi-Itzhak, B. and S. Halfin (1989). Response Times in Gated M/G/1 Queues: The Processor-Sharing Case. *Queueing Systems* **4,** 3, 263-279.

[11] Browne, S., E.G. Coffman, Jr., E.N. Gilbert, and P.E. Wright (1992). Gated, Exhaustive, Parallel Service, *Probab. Engrg. Inform. Sci.* **6**, 217-239.

[12] Browne, S., E.G. Coffman, Jr., E.N. Gilbert, and P.E. Wright (1992). The Gated Infinite-Server Queue: Uniform Service Times. *SIAM Journal on Applied Mathematics* **52,** 6, 1751-1762.

[13] X. Chao and M. Miyazawa (2000). Queueing networks with instantaneous movements: a unified approach by quasi–reversibility. *Advances in Applied Probability*, **32**, 284-313.

[14] X. Chao, M. Miyazawa, and M. Pinedo (1999). *Queueing Networks: Customers, Signals, and Product Form Solutions.* Wiley.

[15] L. Flatto (1985). Two Parallel Queues Created by Arrivals with two Demands II, *SIAM J. Appl. Math.* **45**, 5, 861-878.

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

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

[18] G. Fayolle and R. Iasnogorodski (1979). Two coupled processors: The reduction to a Riemann-Hilbert problem. *Z. Wahrsch. Verw. Gebiete*, **47**, 325–351.

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

[20] J. Gallien and L.M. Wein (2001). A simple and effective component procurement policy for stochastic assembly systems. *Queueing Systems* **38,** 221-248.

[21] I.J. Good. (1960). Generalization to several variables of Lagrange's expansion, with applications to stochastic processes. *Proc. Cambridge Philos. Soc.,* **56,** 367-380.

[22] D.F. Holman, M.L. Chaudhry and B.R.K. Kashyap (1982). On the number in the system $GI^X/M/\infty$, *Sankhyā Ser. A* **44** Pt. 1 294-297.

[23] D.F. Holman, M.L. Chaudhry and B.R.K. Kashyap (1983). On the service system $M^X/G/\infty$, *European J. Oper. Res.* **13** 142-145.

[24] Jagerman, D.L. and B. Sengupta (1989). A functional equation arising in a queue with a gating mechanism, *Probab. Engrg. Inform. Sci.* **3**, 417-433.

[25] L. V. Kantorovich, G. P. Akilov (1982). *Functional Analysis* Pergamon Press.

[26] De Klein, S.J. (1988). *Fredholm Integral Equations in Queueing Analysis.* Ph.D. thesis, University of Utrecht.

[27] Offer Kella and Onno Boxma (2020). Synchronized Lévy Queues. *Journal of Applied Probability*, **57**, 1222-1233.

[28] O. Kella and M. Miyazawa (2001). Parallel fluid queues with constant inflows and simulaneous random reductions, *Journal of Applied Probability*, **38**, 609-620.

[29] C. Knessl (1991). On the diffusion approximation to a fork and join queueing model, *SIAM Journal of Applied Mathematics,* **51**, 1, 160-171.

[30] H. Li and S.H. Xu (2000). On the dependence structure and bounds of correlated parallel queues and their application to synchronized stochastic systems, *Journal of Applied Probability* **37,** 1020-1043.

[31] Jun Li and Yiqiang Q. Zhao, (2010). On The Probability Distribution Of Join Queue Length In A Fork-Join Model, *Probability in the Engineering and Informational Sciences*, **24,** 473–483.

[32] L. Liu, B.R.K. Kashyap and J.G.C. Templeton (1990). On the $GI^X/G/\infty$ system, *J. Appl. Probab.* **27** 671-683.

[33] L. Liu and J.G.C. Templeton (1991). The $GR_n^X/G_n/\infty$ system: System size, *Queueing Systems* **8**, 323-356.

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

[35] H. Masuyama and T. Takine (2002). Analysis of an Infinite-Server Queue with Batch Markovian Arrival Streams, *Queueing Systems* **42,** 269-296.

[36] Meyn, S.P. and R.L. Tweedie (1993). *Markov Chains and Stochastic Stability,* Springer Verlag.

[37] Pinotsi, D. and M.A. Zazanis (2004). Stability Conditions for Gated $M/G/\infty$ Queues, *Probability in the Engineering and Informational Sciences*, 18(1) 103-110.

[38] Pinotsi, D. and M.A. Zazanis (2005). Synchronized queues with deterministic arrivals, *Operations Research Letters*, 33 (6) 560-566.

[39] Pólya, G. and G. Szegö (1972). *Problems and Theorems in Analysis I,* Springer Verlag.

[40] 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.

[41] Rege, K.M. and B. Sengupta (1989). A Single Server Queue with Gated Processor-Sharing Discipline. *Queueing Systems,* **4,** 3, 249-261.

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

[43] Stoyan, D. (1984). *Stochastic Ordering,* John Wiley.

[44] R.A. Sack (1965). Interpretation of Lagrange's expansion and its generalization to several variables as integration formulas. *J. SIAM Appl. Math.* **13,** 47-59.

[45] R.A. Sack (1965). Generalizations of Lagrange's expansion for functions of several implicitly defined variables. *J. SIAM Appl. Math.* **13,** 913-926.

[46] R.A. Sack (1966). Factorization of Lagrange's expansion by means of exponential generating functions. *J. SIAM Appl. Math.* **14,** 1, 1-15.

[47] P. N. Shivakumar and R. Wong (1973). Linear Equations in Infinite Matrices, *Linear Algebra and Its Applications* 7, 53-62.

[48] P.N. Shivakumar, K.C. Sivakumar, and Yang Zhang (2016). *Infinite Matrices and Their Recent Applications*, Springer, 2016.

[49] Shwartz, A. and A. Weiss, (1995). *Large Deviations for Performance Analysis.* Chapman and Hall.

[50] L. Takács (1962). *Introduction to the theory of queues,* Oxford University Press.

[51] Tan X. and C. Knessl (1994). Heavy Traffic Asymptotics for a Gated, Infinite-Server Queue with Uniform Service Times. *SIAM Journal of Applied Mathematics,* **54**, 6, 1768-1779.

[52] R.L. Tweedie. (2000) Markov Chains: Structure and Applications in *Handbook of Statistics* **19,** ed. D.N. Shanbhag and C.R. Rao, Elsevier Amsterdam, 817-851.

[53] Subir Varma and Armand M. Makowski. (1994). Interpolation approximations for symmetric fork-join queues. *Performance Evaluation*, **20**, 245-265.

[54] P.D. Welch. (1965). On the busy period of a facility which serves customers of several types. *Journal of the Royal Statistical Society. Series B*, **27,** 2, 361-370.

[55] P.E. Wright (1992). Two parallel processors with coupled inputs. *Advanced in Applied Probability,* **24**, 986-1007.

[56] Zazanis, M.A. and R. Suri (1994). Perturbation Analysis of the GI/G/1 queue, *Queueing Systems*, **18**, 199-248.

[57] Zhenseng Zhang, (1990). Analytical results for waiting time and system size distributions in two parallel queueing systems, *SIAM J. Appl. Math.*, **50,** 4, 1176-1193.