Waiting time characteristics in cyclic queues

Waiting time characteristics in cyclic queues

Sanne R. Smits, Ivo Adan and Ton G. de Kok

April 16, 2003

Abstract

In this paper we study a single-server queue with FIFO service and cyclic interarrival and service times. An efficient approximative algorithm is developed for the first two moments of the waiting time. Numerical results are included to demonstrate that the algorithm yields accurate results. For the special case of exponential interarrival times we present a simple exact analysis.

1 Introduction

The present study concerns a multi-class queueing model with cyclic interarrival and service times. This model may be used, for example, when the inflow of customers depends on the day of the week, or on the hour of the day. More specifically, this model arises in the modelling of a manufacturing system producing replenishment orders for stock locations that are controlled by periodic order-up-to policies. In that situation, the interarrival times are typically deterministic, and the order sizes are location dependent.

Cyclic queueing models have been studied by Morrice and Gajulapalli (1994). They consider a model with cyclic exponential interarrival and service times, and derive bounds and exact results for the mean number in the system. Cohen (1996) also studies the cyclic model and presents functional equations for the stationary waiting time distributions. These equations formulate a Hilbert Boundary Value problem, which can be solved if the Laplace Stieltjes transforms (LST's) of all the interarrival time distributions or all the service time distributions are rational.

The central equation in this paper is Lindley's equation for the waiting times. It is used to derive an iterative method to compute approximations for the first two moments of the waiting time. This method extends the one developed by de Kok (1989); it is applicable under generally distributed interarrival and service times and produces accurate results. Lindley's equation is also used to exactly determine the moments of the waiting time in case of exponential (or Erlang) interarrival times.

The paper is organized as follows. Section 2 provides a description of the model and introduces some notation. In section 3 we present the moment-iteration method. In section 4 we treat the special case of exponential (or Erlang) interarrival times, and present an exact method for the computation of the moments of the waiting time. Numerical results are presented in Section 6, where we compare the approximations produced by the moment-iteration method with the exact results for the model with Erlang interarrival and service times, and with simulation results for uniform and discrete interarrival time distributions. Finally, Section 5 is devoted to an application of the moment-iteration method.

2 Model description

We consider a single-server queue providing FIFO service to N types of customers, numbered $1, \ldots, N$. The customers arrive in a cyclic pattern: first a type 1 customer, then one of type 2, then 3 until N and then the cycle repeats. Define for $1 \leq i \leq N$ and $k \geq 1$,

$$\begin{aligned}
A_{i,k} &= \text{the time between the arrival of the } k\text{th type } i \text{ customer} \\
&\quad\ \text{and the previous arrival,} \\
B_{i,k} &= \text{the service time of the } k\text{th type } i \text{ customer,} \\
W_{i,k} &= \text{the waiting time of the } k\text{th type } i \text{ customer,} \\
S_{i,k} &= \text{the sojourn time of the } k\text{th type } i \text{ customer } = W_{i,k} + B_{i,k}.
\end{aligned}$$

For each i both $\{A_{i,k}\}_{k\geq 1}$ and $\{B_{i,k}\}_{k\geq 1}$ are sequences of independent identically distributed (i.i.d.) random variables. For stability we assume that the traffic intensity ρ is less than 1:

$$\rho = \frac{\sum_{i=1}^{N} E[B_i]}{\sum_{i=1}^{N} E[A_i]} < 1, \tag{1}$$

where the generic random variables A_i and B_i have the same distribution as $A_{i,k}$ and $B_{i,k}$, respectively. It is readily verified that for $k \geq 1$,

$$\begin{aligned}
W_{1,k} &= (S_{N,k-1} - A_{1,k})^+, \\
W_{i,k} &= (S_{i-1,k} - A_{i,k})^+, \qquad i = 2, \ldots, N,
\end{aligned} \tag{2}$$

where $(x)^+ = \max(x, 0)$. These equations are the starting point for the iterative method presented in the next section. This method approximates the first two moments of the stationary waiting times,

$$E[W_i] = \lim_{k\to\infty} E[W_{i,k}], \qquad E[W_i^2] = \lim_{k\to\infty} E[W_{i,k}^2], \qquad i = 1, \ldots, N;$$

the limits exist by virtue of stability condition (1).

3 Moment-iteration method

Equation (2) relates the waiting time of a customer to the sojourn time of the previous customer. From this equation we get the following expression for the nth moments of $W_{i,k}$:

$$E[W_{1,k}^n] = \int_0^\infty \int_z^\infty (x-z)^n dF_{S_{N,k-1}}(x) dF_{A_{1,k}}(z),$$

$$E[W_{i,k}^n] = \int_0^\infty \int_z^\infty (x-z)^n dF_{S_{i-1,k}}(x) dF_{A_{i,k}}(z), \qquad i = 2, \ldots, N. \quad (3)$$

Here we concentrate on the first two moments (so $n = 1, 2$). If the first two moments of the sojourn time of the previous customer are known and we fit a tractable distribution to these two moments, then the above expressions with the fitted distribution can be used to compute an approximation for the first two moments of the waiting (and sojourn) time of the present customer. For the two moment fit we may use a mixed Erlang or hyperexponential distribution, depending on whether the coefficient of variation is less or greater than 1 (see, e.g., Tijms (1994)). This procedure is then repeated for the next customer and so on. The resulting iteration scheme is presented below.

Iteration scheme

1. Initially set $E[W_{i,0}] = E[W_{i,0}^2] = 0$ for $i = 1, \ldots, N$ and set $i = k = 1$;
2. Fit a tractable distribution to the first two moments of the sojourn time of the previous customer:
   Compute the first two moments and coefficient of variation of $S_{N,k-1}$ if $i = 1$, and of $S_{i-1,k}$ if $i > 1$. The fitted distribution $\tilde{F}_{S_{N,k-1}}(\cdot)$ or $\tilde{F}_{S_{i-1,k}}(\cdot)$ is a mixture of two Erlang distributions with the same scale parameter if the coefficient of variation is less than 1, and otherwise it is a hyperexponential distribution with Gamma normalization.
3. Compute $E[W_{i,k}]$ and $E[W_{i,k}^2]$ according to (3), with $F_{S_{N,k-1}}(\cdot)$ and $F_{S_{i-1,k}}(\cdot)$ replaced by the fitted distributions $\tilde{F}_{S_{N,k-1}}(\cdot)$ and $\tilde{F}_{S_{i-1,k}}(\cdot)$, respectively.

4. If $i < N$, then set $i = i + 1$ and go to step 2; if $i = N$, compute the two sums $\sum_{j=1}^{N} |E[W_{j,k-1}] - E[W_{j,k}]|$ and $\sum_{j=1}^{N} |E[W_{j,k-1}^2] - E[W_{j,k}^2]|$. If both are sufficiently small, then stop and use $E[W_{i,k}]$ and $E[W_{i,k}^2]$ as approximation for $E[W_i]$ and $E[W_i^2]$ for $i = 1, \ldots, N$; otherwise set $k = k + 1$ and $i = 1$ and go to step 2.

4 Exponential interarrival times

In this section we consider the special case that the interarrival time A_i is exponentially distributed with mean $1/\lambda_i$, $i = 1, \ldots, N$. Let W_i and S_i be the waiting time and the sojourn time in steady state of a type i customer, with Laplace-Stieltjes transforms $W_i(s)$ and $S_i(s)$, respectively. Letting $k \to \infty$ in (2) it follows that

$$W_i = (S_{i-1} - A_i)^+, \qquad i = 1, \ldots, N,$$

where, by convention, a type 0 customer is the same as a type N customer. From these equations we get for the transforms

$$\begin{aligned}
W_i(s) &= E(e^{-s(S_{i-1}-A_i)^+}) \\
&= P(S_{i-1} - A_i < 0) + E(e^{-s(S_{i-1}-A_i)} 1_{[S_{i-1}-A_i \geq 0]}) \\
&= P(S_{i-1} - A_i < 0) + E(e^{-s(S_{i-1}-A_i)}) - E(e^{-s(S_{i-1}-A_i)} 1_{[S_{i-1}-A_i < 0]}) \\
&= P(A_i > S_{i-1}) + \frac{\lambda_i}{\lambda_i - s} S_{i-1}(\lambda_i) - E(e^{s(A_i - S_{i-1})} | A_i > S_{i-1}) P(A_i > S_{i-1}).
\end{aligned}$$

By the memoryless property the overshoot $A_i - S_{i-1} | A_i > S_{i-1}$ is again exponential with parameter λ_i, so

$$\begin{aligned}
W_i(s) &= P(A_i > S_{i-1}) + \frac{\lambda_i}{\lambda_i - s} S_{i-1}(\lambda_i) - \frac{\lambda_i}{\lambda_i - s} P(A_i > S_{i-1}) \\
&= \frac{\lambda_i}{\lambda_i - s} S_{i-1}(\lambda_i) - \frac{s}{\lambda_i - s} P(A_i > S_{i-1}).
\end{aligned}$$

Note that $P(A_i > S_{i-1})$ is the probability that a type i customer does not have to wait. Let us write

$$P(A_i > S_{i-1}) = 1 - \Pi_i$$

(so Π_i is the probability of waiting). Further, using $S_i(s) = W_i(s)B_i(s)$ where $B_i(s)$ is the LST of B_i, the equations for the transforms $W_i(s)$ can be written in the form

$$(s-\lambda_i)W_i(s) + \lambda_i B_{i-1}(s)W_{i-1}(s) = (1-\Pi_i)s, \qquad i = 1, 2, \ldots, N. \quad (4)$$

From these equations we may solve $W_N(s)$ yielding

$$W_N(s) = \frac{\sum_{i=1}^{N}(1-\Pi_i)s/\lambda_i \prod_{j=1}^{i-1}(1-s/\lambda_j)\prod_{j=i}^{N-1}B_j(s)}{\prod_{i=1}^{N}B_i(s) - \prod_{i=1}^{N}(1-s/\lambda_i)}. \quad (5)$$

Of course, the other transforms $W_i(s)$ are given by similar (symmetrical) expressions. To determine the unknown probabilities Π_i we proceed as follows. First, we have to satisfy

$$W_N(0) = 1. \quad (6)$$

For the denominator in (5) it holds that

$$\prod_{i=1}^{N} B_i(s) - \prod_{i=1}^{N}(1-\frac{s}{\lambda_i}) = \left(\sum_{i=1}^{N}\frac{1}{\lambda_i} - \sum_{i=1}^{N}E[B_i]\right)s + O(s^2), \qquad (s \to 0)$$

and for the numerator

$$\sum_{i=1}^{N}\frac{(1-\Pi_i)s}{\lambda_i}\prod_{j=1}^{i-1}(1-\frac{s}{\lambda_j})\prod_{j=i}^{N-1}B_j(s) = s \cdot \sum_{i=1}^{n}\frac{1-\Pi_i}{\lambda_i} + O(s^2), \qquad (s \to 0).$$

Hence, from (6) we get

$$\sum_{i=1}^{N}\frac{1-\Pi_i}{\lambda_i} = \sum_{i=1}^{N}\frac{1}{\lambda_i} - \sum_{i=1}^{N}E[B_i]. \quad (7)$$

Further, since $W_N(s)$ is well defined for $\mathrm{Re}(s) \geq 0$, it follows that whenever the denominator in (5) vanishes for some s with $\mathrm{Re}(s) \geq 0$, the numerator should also vanish. In the appendix we will prove that the denominator has exactly N zeros s with $\mathrm{Re}(s) \geq 0$, say $s_0(=0), s_1, \ldots, s_{N-1}$. Since the numerator of (5) must also vanish at $s = s_1, \ldots, s_{N-1}$, we obtain the following equations.

$$\sum_{i=1}^{N}\frac{(1-\Pi_i)s_k}{\lambda_i}\prod_{j=1}^{i-1}(1-\frac{s_k}{\lambda_j})\prod_{j=i}^{N-1}B_j(s_k) = 0, \qquad k = 1, \ldots, N-1.$$

Together with (7), this forms a set of N equations for N waiting probabilities $\Pi_1, \ldots, \Pi_N$; it has a unique solution, since under the condition of stability (1), there is a unique stationary waiting time distribution and thus also a unique solution $W_i(s)$. This completes the determination of the transforms $W_i(s)$, as given by (5).

In the remainder of this section we show how the moments of the waiting times may be determined. As starting point we take the equations (4). Substituting the Taylor series

$$W_i(s) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} E[W_i^k] s^k$$

and

$$B_{i-1}(s) W_{i-1}(s) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \sum_{j=0}^{k} \binom{k}{j} E[W_{i-1}^j] E[B_{i-1}^{k-j}] s^k$$

we obtain

$$\sum_{k=0}^{\infty} \frac{(-1)^{k-1}}{k!} \left(k E[W_i^{k-1}] + \lambda_i E[W_i^k] - \lambda_i \sum_{j=0}^{k} \binom{k}{j} E[W_{i-1}^j] E[b_{i-1}^{k-j}] \right) s^k = (1 - \Pi_i) s.$$

Equating the coefficients of s^k on the left- and right-hand side yields

$$1 + \lambda_i (E[W_i] - E[w_{i-1}]) - \lambda_i E[B_{i-1}] = (1 - \Pi_i), \qquad i = 1, \ldots, N, \tag{8}$$

and for $k > 1$,

$$k E[W_i^{k-1}] + \lambda_i (E[W_i^k] - E[W_{i-1}^k]) - \lambda_i \sum_{j=0}^{k-1} \binom{k}{j} E[W_{i-1}^j] E[B_{i-1}^{k-j}] = 0, \qquad i = 1, \ldots, N. \tag{9}$$

Adding (9) over all $i = 1, \ldots, N$ gives the equation

$$\sum_{i=1}^{N} k E[W_i^{k-1}] - \sum_{i=1}^{N} \lambda_{i+1} \sum_{j=0}^{k-1} \binom{k}{j} E[W_i^j] E[B_i^{k-j}] = 0.$$

This can be rewritten as (replace k by $k+1$)

$$\sum_{i=1}^{N} E[W_i^k](k+1)(1-\lambda_{i+1}E[B_i]) = \sum_{i=1}^{n} \lambda_{i+1} \sum_{j=0}^{k-1} \binom{k+1}{j} E[W_i^j]E[B_i^{k+1-j}], \tag{10}$$

which is valid for $k \geq 1$. Using the equations (8)-(10), all moments $E[W_i^k]$ can now be computed recursively. First note that addition of (8) over all i gives an identity; so we may omit one of these equations. Then together with (10) for $k = 1$, we have a set of N equations, from which the first moments $E[W_i]$, $i = 1, \ldots, N$, can be computed. To find the second moments we use the equations (9) and (10) for $k = 2$, where in (9) we may again omit one equation. Also note that the lower (first) moments occurring in these equations are now known. The third moments follow from the equations (9) and (10) for $k = 3$, and so on. It will be clear that we can repeat to use (9) and (10) to successively compute all moments (if they exist).

Remark 4.1 *Erlang interarrival times*
If the interarrival time A_i is Erlang-$r(\lambda_i)$ distributed, then we can think of the interarrival time as consisting of r sub-interarrivals. By associating with each subarrival an arrival of a customer, where the first $r-1$ customers have zero service times and the last, r-th one has a service time B_i, we can analyze this case along the same lines as the exponential case.

5 Numerical results

In this section we validate the approximative moment-iteration method by simulation and exact results (if available). To do so we will consider models with deterministic, uniform, exponential or (mixed) Erlang distributions for the inter-arrival and service times. The models are indicated with the well-known three letter code $A/B/1$; it means that all customer types have the same type of interarrival time distribution, indicated by A, and the same type of service time distribution, indicated by B, but of course, the parameters of the distributions depend on the customer type. For each case of inter-arrival and service time distributions we randomly generate 10^3 settings of the parameters, and each setting will be evaluated for 2, 5 and 25 customer types. If only mean and coefficient of variation are specified, we fit a distribution according to the recipe described in section (3). Table 1 gives an overview of the parameter settings, where $U(a, b)$ denotes the uniform distribution on (a, b) and $U(S)$ the (discrete) uniform distribution on the points in the set S.

Model	Parameter	Value
$D/G/1$	$E[B_i]$	$U(0.2, 0.99)$
	$C^2_{B_i}$	$U(0.2, 2)$
	A_i	1
$U/G/1$	$E[B_i]$	$U(0.3, 0.99)$
	$C^2_{B_i}$	$U(0.2, 2)$
	$\min A_i$	0.7
	$\max A_i$	1.3
$M/M/1$	$E[B_i]$	$U(0.3, 0.99)$
	$E[A_i]$	1
$E_k/M/1$	$E[B_i]$	$U(0.3, 0.99)$
	$E[A_i]$	1
	k	$U(\{1, 2\})$
$E_k/E_l/1$	$E[B_i]$	$U(0.3, 0.99)$
	l	$U(\{(1, 2, 3, 4\})$
	$E[A_i]$	1
	k	$U(\{1, 2\})$

Table 1: Parameter settings

Note that, according to table 1, we will generate settings with different traffic intensities. We divide the settings into three categories: low load ($0.4 < \rho < 0.6$), medium load ($0.6 \leq \rho < 0.8$) and high load ($\rho \geq 0.8$). We compare the estimates produced by the moment-iteration method with the exact results of section 4 in case of exponential or Erlang interarrival times, and with simulation results otherwise. The simulation results are based on 10 independent replicas of $6 \cdot 10^6$ arrivals. In table 2 and 3 we display percentage errors; $\bar{\delta}^{E[W]}$ and $\max \delta^{E[W]}$ denote the *average* and *maximal* percentage error of the average of the waiting time over all customer types and all settings generated in a category and $\bar{\delta}^{\sigma(W)}$ and $\max \delta^{\sigma(W)}$ denote the *average* and *maximal* percentage error of the standard deviation of the waiting time over all customer types and all settings generated in a category

Load	Av. Error	N	Model				
			$D/G/1$	$U/G/1$	$M/M/1$	$E_k/M/1$	$E_k/E_l/1$
$0.4 \le \rho < 0.6$	$\bar{\delta}^{E[W]}$	2	2.99	1.99	2.19	3.59	5.78
		5	4.14	3.97	0.93	1.54	3.31
		25	5.89	5.43	0.91		
	$\bar{\delta}^{\sigma(W)}$	2	6.30	6.45	2.09	3.95	7.02
		5	7.25	6.83	1.29	3.46	8.27
		25	9.39	8.85	0.34		
$0.6 \le \rho < 0.8$	$\bar{\delta}^{E[W]}$	2	1.93	2.38	1.30	1.93	3.60
		5	3.04	3.36	0.78	0.86	3.04
		25	3.93	3.56	1.27		
	$\bar{\delta}^{\sigma(W)}$	2	4.14	3.11	1.04	1.81	5.00
		5	5.87	5.97	0.69	1.10	7.08
		25	5.60	4.95	0.73		
$0.8 \le \rho$	$\bar{\delta}^{E[W]}$	2	1.25	1.48	1.80	2.31	1.69
		5	1.13	0.97	0.81	0.77	1.13
		25	2.13	1.21	0.95		
	$\bar{\delta}^{\sigma(W)}$	2	3.73	4.77	2.84	0.52	3.04
		5	3.61	5.89	0.76	0.76	3.09
		25	3.41	3.40	0.64		

Table 2: Average percentage errors of the moment-iteration method.

We may conclude that the moment-iteration algorithm produces accurate results, especially for high loads; typically the average of the waiting time is more accurately estimated than the standard deviation. For low loads the percentage error of the moments of the waiting time may be big; the absolute error, however, will be modest (in comparison with the service times). It further seems that the type of distribution of the interarrival times does not influence the quality of the estimates produced by the moment-iteration method; the traffic intensity and the coefficients of variation of the service times are more crucial.

Load	Max Error	N	Model				
			$D/G/1$	$U/G/1$	$M/M/1$	$E_k/M/1$	$E_k/E_l/1$
$0.4 \leq \rho < 0.6$	$\max \delta^{E[W]}$	2	12.61	14.52	4.62	4.35	9.04
		5	16.02	24.87	2.23	7.23	10.43
		25	19.42	26.91	2.86		
	$\max \delta^{\sigma(W)}$	2	18.08	18.73	5.34	7.05	18.82
		5	22.27	23.24	5.92	6.38	19.82
		25	23.26	25.55	3.45		
$0.6 \leq \rho < 0.8$	$\max \delta^{E[W]}$	2	8.87	13.29	3.90	4.42	8.42
		5	9.77	13.57	4.03	3.81	15.56
		25	12.24	12.66	2.76		
	$\max \delta^{\sigma(W)}$	2	14.71	18.77	5.27	4.08	11.56
		5	16.49	19.88	4.24	4.01	15.50
		25	19.35	22.10	4.30		
$0.8 \leq \rho$	$\max \delta^{E[W]}$	2	4.04	4.06	2.10	3.59	7.20
		5	7.54	6.88	2.37	1.85	8.40
		25	7.84	8.33	1.96		
	$\max \delta^{\sigma(W)}$	2	9.62	16.21	2.41	2.21	9.84
		5	11.46	17.24	2.70	1.59	8.33
		25	12.53	17.60	2.24		

Table 3: Maximal percentage errors of the moment-iteration method.

6 Application

In this section we apply the moment-iteration method to the manufacturing problem mentioned in the Introduction. We consider a manufacturing system with N stockpoints, one for each item, and one production facility, which produces all the items in a FIFO order. The objective is to determine the sojourn time of the replenishment orders placed by the stockpoints. Figure 1 shows a schematic representation of the model for $N = 4$.

The stockpoints are controlled by periodic order-up-to policies. The periodic order-up-to policy operates as follows: Every R_i periods of time the inventory position of stockpoint i is inspected and a replenishment order is placed at the production facility as to raise the inventory position up to the order-up-to level. The inventory position is defined as the physical inventory level plus the stock on order minus the backorders. We assume that the review period of each item is identical ($R_i = R$). The processing times of replenishment orders from stockpoint i are assumed to be i.i.d. random

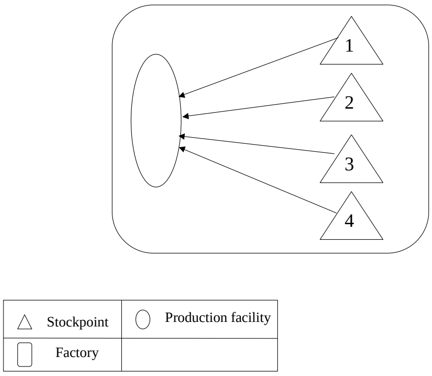

Figure 1: Schematic representation of the model.

variables with mean $E[B_i]$ and standard deviation $\sigma(B_i)$. The first time the inventory position is inspected is denoted by t_i^0 $(i = 1, ..., N)$. The values of R_i, $E[B_i]$, $\sigma(B_i)$ and t_i^0 $(i = 1, ..., N)$ are listed in Table 4.

i	1	2	3	4
R_i	105.33	105.33	105.33	105.33
$E[B_i]$	19.24	25.20	27.15	24.52
$\sigma(B_i)$	7.05	8.02	5.34	4.81
t_i^0	0	21.06	48.69	78.45

Table 4: Input parameters.

Now we are interested in the sojourn time or production lead time S_i of replenishment orders of stockpoint i; obviously the production lead time S_i is required for setting an appropriate stock level.

The time A_i between the arrival of an item i replenishment order and item $i-1$ replenishment order follows from t_i^0 and R_i $(i = 1, ..., N)$. Since A_i is deterministic, we are dealing with a cyclic $D/G/1$ queue. Given A_i and the first two moments of B_i we can determine the first two moments of W_i and S_i $(i = 1, ..., N)$ by using the moment-iteration method presented in Section 3. The results are presented in Table 5.

i	1	2	3	4
A_i	26.88	21.06	27.63	29.79
$E[W_i]$	5.42	5.76	6.11	5.81
$\sigma(W_i)$	6.92	7.72	8.42	7.59
$E[S_i]$	24.66	30.96	33.26	30.33
$\sigma(S_i)$	9.88	11.13	9.97	8.98

Table 5: Numerical results.

Note that the traffic intensity of the cyclic $D/G/1$ queue describing the manufacturing system is 0.91. Hence Table 2 indicates that in this case (with $N = 4$) we may expect errors in the mean waiting time close to 1 percent and in its standard deviation close to 4 percent; this seems sufficiently accurate for practical purposes. From Table 4, it follows that $c_{B_1} = 0.37$, $c_{B_2} = 0.32$, $c_{B_3} = 0.20$ and $c_{B_4} = 0.20$. In Table 5 we see that replenishment orders from stockpoint 1 clearly benefit from the smaller variation in processing times of replenishment orders from stockpoints 3 and 4.

References

Cohen, J.W. (1996) *Periodic Pollaczek wiating time processes.* C.C. Heyde et al., Athens Conference on Applied Probability and Time Series. Springer, NY-1996

Kok, A.G. de (1989) *A Moment-iteration method for approximating the waiting-time characteristics of the GI/G/1 queue.* Probability in the Engineering and Informational Sciences, vol 3, p 273-287

Morrice, D.J. and Gajulapalli, R.S. (1994) *A single server queue with cyclically indexed arrival and services rates.* Queueing Systems, vol 15, p 165-198

Tijms, H.C. (1994) *Stochastic models: an algorithmic approach.* John Wiley & Sons, Chichester

7 Appendix

In this appendix we will prove that the denominator of (5) has N zeros s with $\mathrm{Re}(s) \geq 0$. To do so we will use Rouche's theorem, which reads as follows. Let $f(z)$ and $g(z)$ be analytic functions inside and on a smooth contour C, and suppose that $|g(z)| < |f(z)|$ on C. Then $f(z)$ and $f(z) + g(z)$ have the same number of zeros inside C. Of course, we may also replace $f(z) + g(z)$ by $g(z) - f(z)$ in this formulation, and actually, that is the form we will use.

We first assume that there is some $\epsilon > 0$ such that the transforms $B_i(s), i = 1, \ldots, N$, are analytic for all s with $\mathrm{Re}(s) > -\epsilon$. This assumption holds, for example, for service time distributions with a finite support or an exponential tail. Now take

$$f(z) = \prod_{i=1}^{N} (1 - \frac{s}{\lambda_i})$$

and

$$g(z) = \prod_{i=1}^{N} B_i(s).$$

As contour we take the circle C with center $\max_i \lambda_i$ and radius $\delta + \max_i \lambda_i$ with $0 < \delta < \epsilon$. It is easily verified that for all z on C,

$$|g(z)| \leq \prod_{i=1}^{N} B_i(\mathrm{Re}(z)) \leq \prod_{i=1}^{N} B_i(-\delta) = g(-\delta), \qquad f(-\delta) \leq |f(z)|. \tag{11}$$

Note that

$$g'(0) = -\sum_{i=1}^{N} E[B_i], \qquad f'(0) = -\sum_{i=1}^{N} \frac{1}{\lambda_i}.$$

Hence the stability condition (1) states that $g'(0) > f'(0)$. Thus, for sufficiently small $\delta > 0$, it holds that $g(-\delta) < f(-\delta)$, which implies together with (11) that $|g(z)| < |f(z)|$ for all z on C. Rouche's theorem now guarantees that $f(z)$ and $g(z) - f(z)$ (which is the denominator in (5)) have the same number of zeros inside C. Since $f(z)$ has N zeros inside C, the same holds for the denominator of (5).

To complete the proof we have to remove the initial assumption that for some $\epsilon > 0$ the transforms $B_i(s)$ are analytic for all s with $\mathrm{Re}(s) > -\epsilon$. To this end, first consider, instead of B_i, the truncated service times $\min(B_i, K)$ where $K > 0$ is some constant. For these truncated service times the claim for the zeros of the denominator (5) holds; by letting K tend to infinity, the claim also follows for the original service time distributions.

Remark 7.1 In fact, we did not only show that the existence of N zeros s with $\mathrm{Re}(s) \geq 0$, but also that they are located inside or on the circle with center $\max_i \lambda_i$ and radius $\max_i \lambda_i$; this maybe useful for the numerical calculation of the zeros.