Relationship between Delay and System Contents in Multiserver Queues with Geometric Service Times

P. GAO, S. WITTEVRONGEL and H. BRUNEEL
SMACS* Research Group
Department of Telecommunications and Information Processing
Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Gent
BELGIUM

Abstract: - In this paper, we consider a discrete-time multiserver queueing system with infinite buffer size, geometric service times and a FCFS (first-come-first-served) queueing discipline. A relationship between the probability distributions of the system contents and the packet delay is established. The relationship is general in the sense that it doesn’t require knowledge of the exact nature of the arrival process. By means of the relationship, results for the characteristics of the system contents for various discrete-time queueing models can be transformed into corresponding results for the delay characteristics, thus making a separate delay analysis superfluous.

Key-Words: - Discrete-time queueing, multiserver system, geometric service times, analytical techniques, generating functions

1 Introduction

Discrete-time queueing models play an important role in the performance evaluation of packet-based telecommunication networks, where buffers are used for the temporary storage of information packets which cannot be transmitted to their destination immediately. In discrete-time queueing models, time is divided into fixed-length slots and the service (transmission) of packets can start and end at slot boundaries only. Usually, the performance of a queueing system is expressed in terms of such quantities as the system contents (i.e., the total number of packets present in the queueing system) and the delay of a packet (i.e., the time (in slots) spent by a packet in the system), and in many cases, a separate full analysis is done for each of these quantities. However, during the past decades several researchers have tried to establish relationships between the probability distributions of the delay and the system contents in discrete-time queues. By means of such relationships, results for the characteristics of the system contents for a wide variety of discrete-time queueing models can be transformed into corresponding results for the delay characteristics, thus making a separate delay analysis superfluous.

The best known general relationship between system contents and delay is Little’s law ([1], [2]). It is valid for any arrival process and for any service process, but only deals with the first moment of both variables. Where researchers have tried to relate higher-order moments or other performance measures, restrictions were put on the arrival process or on the service process or on both, see for instance [3]-[9]. In this paper, we will derive a relationship between the probability distributions of the system contents and the packet delay for a multiserver queue with geometrically distributed service times. The relationship is general in the sense that it does not depend explicitly on the arrival process. It can be seen as a generalization of the results of [7]-[9], where similar relationships have been derived for the case of deterministic service times (of one slot), for a single-server queue ([7]) as well as a queue with multiple servers ([8], [9]), and also under the as-

*SMACS: Stochastic Modeling and Analysis of Communication Systems
sumption of a very general, possibly correlated, arrival process.

The paper is organized as follows. In Section 2, we describe the class of discrete-time queueing systems under study and introduce some notations. For the considered class of queueing systems, we establish a relationship between the steady-state probability generating functions (pgf’s) of the system contents and the packet delay in Section 3. The applicability of the result is discussed in Section 4.

2 System Description and Notations

We consider a discrete-time queueing system with \( c (c \geq 1) \) servers (output channels). Time is divided into fixed-length slots. Packets arrive at the input of the system according to a general (possibly correlated) arrival process and are queued in a buffer, until they can be transmitted via one of the \( c \) output channels based on a FCFS discipline. The buffer has an infinite storage capacity for packets. The service (or transmission) of a packet can start or end at slot boundaries only. The service times of the packets are assumed to be i.i.d. (independent and identically distributed) random variables and geometrically distributed with parameter \( 1 - \mu \) (0 < \( \mu \) ≤ 1), i.e., with probability mass function (pmf)

\[
g(n) = \text{Prob}[\text{service of a packet takes } n \text{ slots}] = \mu (1 - \mu)^{n-1}, \quad n \geq 1,
\]

and corresponding pgf

\[
G(z) = \frac{\mu z}{1 - (1 - \mu) z}, \quad (1)
\]

Moreover, the service and arrival processes are assumed to be mutually independent. Finally, it is assumed that the mean number of packet arrivals (\( \sigma \)) during an arbitrary slot is strictly less than the mean number of packets that can be transmitted from the buffer per slot (\( c \mu \)), so that the system can reach a steady state.

Let us denote by \( v_k \) the system contents (i.e., the total number of packets in the buffer system, including the packets under transmission, if any) at the beginning of slot \( k \), and by \( e_k \) the number of packet arrivals during slot \( k \). Then, in view of the above modeling assumptions, we have the following system equation:

\[
v_{k+1} = v_k - t_k + e_k, \quad (2)
\]

where

\[
t_k = \sum_{j=1}^{(v_k,c)} t_{k,j}, \quad (3)
\]

with \((.,.)^{-} \triangleq \min(.,.)\). The random variable \( t_k \) denotes the total number of departures at the end of slot \( k \), and \( t_{k,j} \) corresponds to the number of departures at the end of slot \( k \) via the \( j \)-th output channel when there is a packet being transmitted over this output channel. In view of the geometric distribution of the packet service times, \( t_{k,j} \) is a Bernoulli distributed random variable with parameter \( \mu \) (i.e., \( t_{k,j} = 1 \) with probability \( \mu \) and \( t_{k,j} = 0 \) with probability \( 1 - \mu \)). We define the conditional pgf \( T_i(z) \) as

\[
T_i(z) \triangleq E[z^{tk} | (v_k,c)^{-} = i], \quad i = 0, 1, ..., c,
\]

where \( E[.] \) denotes the expected value of the expression between the square brackets. Then, from (3), we have

\[
T_i(z) = (1 - \mu + \mu z)^i.
\]

Also, we denote by \( p(i,j) \) the steady-state joint probability

\[
p(i,j) \triangleq \text{Prob}[v = i, e = j] = \lim_{k \to \infty} \text{Prob}[v_k = i, e_k = j], \quad (4)
\]

and by \( P(z,x) \) the corresponding joint pgf, i.e.,

\[
P(z, x) \triangleq \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p(i,j) x^i z^j. \quad (5)
\]

Then the pgf \( V(z) \) of the system contents \( v \) at the beginning of an arbitrary slot in the steady state can be expressed as

\[
V(z) = P(z,1) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p(i,j) z^j. \quad (6)
\]
3 Relationship between System Contents and Delay

The delay of a packet is defined as the total number of slots between the end of the slot during which the packet arrived in the system and the end of the slot where the packet’s transmission finishes and the packet leaves the system. The purpose of this section is to derive a relationship between the steady-state probability distributions of the system contents at the start of an arbitrary slot and the delay of an arbitrary packet. In order to do so, let us consider an arbitrary packet P (referred to as the tagged packet), that arrives in the queueing system during some slot J in the steady state. Let \( d \) with pgf \( D(z) \) be the delay that P experiences. Also define the waiting time of a packet as the number of slots between the end of the packet’s arrival slot and the beginning of the slot when the packet’s transmission starts. Then it is clear that the delay of a packet is equal to the sum of the packet waiting time and the packet transmission time. Hence, the pgf \( D(z) \) is easily obtained as

\[
D(z) = W(z)G(z), \quad (7)
\]

where \( W(z) \) denotes the pgf of the waiting time \( w \) of P.

In the rest of this section, we will concentrate on the derivation of the pgf \( W(z) \). For this purpose, let us observe how the tagged packet P is served. Upon arrival, the packet P will find a number of other packets in the queueing system. Due to the FCFS discipline, just after the end of slot \( J \), all packets that arrived before slot \( J \), but have not been taken into service yet, and all packets that arrived in slot \( J \), but before the packet P, are waiting in the queue in front of P. Whenever there are servers available at the beginning of a slot, the packet at the head of the queue is selected for service, until eventually the tagged packet P itself gets into service. Let \( q \) be the number of packets staying in the system at the arrival instant of P (except P itself and the packets that leave the system at the end of the arrival slot of P, if such packets exist), i.e., \( q \) denotes the number of packets present in the system right after slot \( J \), that have been or will be selected for service before P. In order to derive \( W(z) \), we will now first establish a relationship between \( W(z) \) and the pgf \( Q(z) \) of \( q \), and next, we will express the pgf \( Q(z) \) in terms of the pgf \( V(z) \) of the system contents \( v \) at the beginning of an arbitrary slot in the steady state. Combination of these results with equation (7) will then yield the envisaged relationship between the system-contents and delay distributions.

3.1 Relation between \( W(z) \) and \( Q(z) \)

Let us observe the slots following the arrival slot \( J \). During slot \( J + 1 \), there will be either \( \tilde{t}_{J+1} (= \sum_{j=1}^{c} t_{J+1,j} ) \) packets leaving the system at the end of slot \( J + 1 \), if \( q \geq c \) (i.e., just after the end of that slot or at the beginning of slot \( J + 2 \), there will be \( q - \tilde{t}_{J+1} \) packets in the system with priority over the tagged packet P to be taken into service), or P will get into service, if \( q < c \) (i.e., P stops waiting). In case P didn’t get into service, an analogous reasoning holds for slot \( J + 2 \): either there will be \( q - \tilde{t}_{J+1} - \tilde{t}_{J+2} \) packets with service priority over P at the end of slot \( J + 2 \), if \( q - \tilde{t}_{J+1} \geq c \), or the packet P will get into service, if \( q - \tilde{t}_{J+1} < c \). Eventually, we see that the tagged packet will still be waiting for service during the \((J + i + 1)\)-th slot only if \( q - \tilde{t}_{J+1} - \tilde{t}_{J+2} - \ldots - \tilde{t}_{J+i} \geq c \). This leads to the following relationship between the waiting time \( w \) of P and the random variable \( q \):

\[
w > i \iff q - \sum_{r=1}^{i} \tilde{t}_{J+r} \geq c,
\]

or equivalently

\[
w > i \iff q \geq s_i, \quad (8)
\]

where the random variables \( s_i \) are defined as

\[
s_0 \triangleq c; \quad s_i \triangleq \sum_{r=1}^{i} \tilde{t}_{J+r} + c, \quad i \geq 1.
\]

Note that since the variables \( \tilde{t}_{J+r} \) in the above definition are i.i.d. random variables, the pgf of \( s_i \) can be expressed as

\[
S_i(z) \triangleq E[z^{s_i}] = z^c T_c(z)^i. \quad (9)
\]

The next step is now to transform the relationship (8) between the random variables \( w \) and
Specifically, considering the argument by its partial fraction expansion.

\[
\frac{W(z) - 1}{z - 1} = \sum_{i=0}^{\infty} z^i \text{Prob}[w > i]
\]

where \( q \) is a constant and \( x \) the variable of interest, we have

\[
\text{Prob}[s_i = n] = \frac{1}{n!} \frac{d^n}{dx^n} S_i(x) \bigg|_{x=0},
\]

and working out the sum over \( i \), we then find

\[
\frac{W(z) - 1}{z - 1} = \sum_{j=0}^{\infty} \text{Prob}[q = j]
\]

\[
\cdot \sum_{n=c}^{j} \frac{1}{n!} \frac{d^n}{dx^n} \left( 1 - zT_e(x) \right) \bigg|_{x=0}.
\]

Remark that working out the sum over \( i \) in (10) requires that \( |zT_e(x)| < 1 \) in the neighborhood of \( x = 0 \). This condition will always be fulfilled for \( |z| \leq 1 \), since \( |T_e(x)| < 1 \) for \( |x| < 1 \). To derive the partial derivatives in (11), we replace the argument by its partial fraction expansion. Specifically, considering \( z \) to be a constant and \( x \) the variable of interest, we have

\[
\frac{N(x)}{1 - zT_e(x)} = A + \sum_{p=0}^{c-1} \frac{-N(x_p)}{zT_e'(x_p)(x - x_p)},
\]

where \( N(x) \) is a polynomial of degree \( m (m \leq c) \), \( A \) is a constant (\( A = 0 \) when \( m < c \)) and the \( x_p \)'s \( (p = 0, 1, \ldots, c - 1) \) are the \( c \) solutions for \( x \) in terms of \( z \) of the equation

\[
1 - zT_e(x) = 0,
\]

which we assume to be distinct. (Although the \( x_p \)'s are functions of \( z \), we write \( x_p \) instead of \( x_p(z) \) to ease the notation.) Substitution of the expansion (12) in equation (11), with some further mathematical manipulations, leads to

\[
\frac{W(z) - 1}{z - 1} = \sum_{p=0}^{c-1} \frac{-x_p^c}{\sum_{j=0}^{\infty} zT_e'(x_p)(1 - x_p)}
\]

\[
\cdot \sum_{j=0}^{\infty} \text{Prob}[q = j](x_p^{-c} - x_p^{-1-j}),
\]

or, if we use the definition of the pgf \( Q(z) \),

\[
\frac{W(z) - 1}{z - 1} = \sum_{p=0}^{c-1} \frac{-1}{zT_e'(x_p)(1 - x_p)}
\]

\[
+ \sum_{p=0}^{c-1} \frac{x_p^{-c} Q(1/x_p)}{zT_e'(x_p)(1 - x_p)}
\]

\[
+ \sum_{j=0}^{\infty} \text{Prob}[q = j] \sum_{p=0}^{c-1} \frac{1 - x_p^{-c-j-1}}{zT_e'(x_p)(1 - x_p)}.
\]

Finally, again using the partial fraction expansion (12), now at \( x = 1 \), we find the following expression for the pgf of the packet waiting time:

\[
W(z) = (z - 1) \sum_{p=0}^{c-1} \frac{x_p^{-c} Q(1/x_p)}{zT_e'(x_p)(1 - x_p)}.
\]

**3.2 Relation between \( Q(z) \) and \( V(z) \)**

This subsection deals with the derivation of a relationship between \( Q(z) \) and \( V(z) \). Let us denote by \( v^* \) the number of packets present in the queuing system at the beginning of the arrival slot of \( P \) (slot \( J \)), by \( f \) the number of packets arriving during slot \( J \) but before \( P \), and by \( t \) the number of departures at the end of slot \( J \). Then \( q \) can be expressed as

\[
q = v^* - t + f.
\]

In order to derive the pgf \( Q(z) \) of \( q \), we need an expression for the joint distribution of the random variables \( v^* \) and \( f \). This distribution can be determined by conditioning on the value of the random variable \( e^* \), as follows:

\[
\text{Prob}[v^* = i, f = j]
\]

\[
= \sum_{l=j+1}^{\infty} \text{Prob}[v^* = i, f = j, e^* = l]
\]

\[
= \sum_{l=j+1}^{\infty} \text{Prob}[f = j|v^* = i, e^* = l]
\]

\[
\cdot \text{Prob}[v^* = i, e^* = l],
\]
where $e^*$ is the number of packet arrivals during slot $J$. Since $f$ is the number of packets that arrived in the system during the same slot as the tagged packet but before it, and the tagged packet has been chosen randomly from all arriving packets, we have

$$\text{Prob}[f = j | v^* = i, e^* = l] = \frac{1}{l}, \quad 0 \leq j \leq l - 1.$$  

Furthermore, the joint probability $\text{Prob}[v^* = i, e^* = l]$ corresponds to the fraction of packets that arrive in a slot where there are $l$ packet arrivals and the system contents at the beginning of the slot is $i$, owing to the fact that $P$ is an arbitrary packet. Since each such slot contains $l$ arrivals and the tagged packet could be any of these, it is clear that (see e.g. [10])

$$\text{Prob}[v^* = i, e^* = l] = \frac{lp(i,l)}{\sigma}.$$  

Combining the three previous expressions, we obtain

$$\text{Prob}[v^* = i, f = j] = \frac{1}{\sigma} \sum_{l=j+1}^{\infty} p(i, l). \quad (16)$$

Let us now define $M(z, x)$ as the joint pgf of the couple $(v^*, f)$, i.e.,

$$M(z, x) \triangleq E[z^{v^*} x^f]$$

$$= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \text{Prob}[v^* = i, f = j] z^i x^j. \quad (17)$$

From equation (16), we then have

$$M(z, x) = \frac{1}{\sigma} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{l=j+1}^{\infty} p(i, l) z^i x^j.$$  

Changing the order of the summations, working out the sum over $j$, and then using equations (5) and (6), we obtain the pgf $M(z, x)$ as

$$M(z, x) = \frac{P(z, 1) - P(z, x)}{\sigma(1-x)}. \quad (18)$$

Let us now return to the equation (15) for $q$. By means of equations (16)-(17) and some standard $z$-transform techniques, equation (15) can be transformed into the following expression for the pgf $Q(z)$ of $q$:

$$Q(z) = T_c(z) M(z, x)$$

$$+ \frac{1}{\sigma} \sum_{i=0}^{c-1} \sum_{j=0}^{\infty} (1 \leftarrow T_c(z)) \sum_{l=j+1}^{\infty} p(i, l) z^i x^j. \quad (19)$$

Changing the order of the summations in (19), working out the sum over $j$, and using equations (4)-(6) and (18), we then get

$$Q(z) = \frac{1}{\sigma(1-z)} \left\{ \sum_{i=0}^{c-1} z^i \left[ T_i(z) - T_c(z) \right] v(i) + T_c(z) P(z, 1) - T_c(z) P(z, x) \right. - \left. \sum_{i=0}^{c-1} \sum_{j=0}^{\infty} z^{i+j} T_i(z) - T_c(z) p(i, l) \right\}, \quad (20)$$

where $v(i) = \text{Prob}[v = i]$. On the other hand, from the system equation (2) and the definition of the steady-state joint pgf $P(z, x)$, we have

$$P(z, 1) = T_c(z) P(z, x)$$

$$+ \sum_{i=0}^{c-1} \sum_{j=0}^{\infty} z^{i+j} T_i(z) - T_c(z) p(i, j). \quad (21)$$

From (20) and (21), the pgf $Q(z)$ can finally be expressed as

$$Q(z) = \frac{1}{\sigma(1-z)} \left\{ \left[ T_c(z) - 1 \right] V(z) + \sum_{i=0}^{c-1} z^i \left[ T_i(z) - T_c(z) \right] v(i) \right\}. \quad (22)$$

### 4 Applicability

Combination of equations (7), (14) and (22) leads to a relationship between the steady-state pgf’s $V(z)$ and $D(z)$ of the system contents at the start of an arbitrary slot and the delay of an arbitrary packet, respectively. It has been verified that our result is in correspondence with those of [7]-[9] for the special case of constant packet service times of one slot. By means of the relationship, not only the pgf but also several other delay characteristics, such as moments and tail probabilities, can
be determined quite directly once the pgf of the system contents has been obtained. Specifically, the moments of the delay can be calculated by evaluating the derivatives of the pgf $D(z)$ with respect to $z$ at $z = 1$. The tail probabilities of the delay (i.e., the probabilities to exceed a given threshold) can be found for a sufficiently large (delay) threshold, by means of a method presented in [10].

The established relationship between system contents and delay is applicable to any discrete-time multiple-server queueing system as long as the service times of the packets are geometrically distributed and the queueing discipline is FCFS. The exact nature of the arrival process (such as the nature of the correlation, the nature of the source(s) that generate the packets) is not relevant. This means that although the statistics of the system contents and the delay may heavily depend on the specific nature of the arrival process, knowledge of the arrival process is not needed for the transformation from system contents to delay.

For a further illustration of the usefulness of our result we refer to [11]. In that paper, a discrete-time multiple-server queueing system with geometric service times and a batch renewal arrival process is studied. The batch renewal process is a correlated arrival process characterized in terms of general independent batch interarrival times and general independent batch sizes. Both the pgf of the system contents at the beginning of an arbitrary slot and the pgf of the delay of an arbitrary packet were derived. By means of the relationship we developed, the pgf of the delay could have been derived immediately from the pgf of the system contents, hence making a separate delay analysis superfluous.

References: