Delay against system contents in discrete-time G/Geom/c queue

P. Gao, S. Wittevorgel and H. Bruneel

A discrete-time multiserver queueing system with infinite buffer size, geometric service times and a first-come-first-served (FCFS) queueing discipline is considered. A relationship between the probability distributions of the system contents and the packet delay is established.

Introduction: General relationships between the probability distributions of the system contents during an arbitrary slot and the packet delay of an arbitrary customer have been derived in [1] for a stable discrete-time single-server queue, and in [2] and [3] for a queue with multiple servers, both for the case of deterministic service times (of one slot). The relationships are general in the sense that they do not require knowledge of the exact nature of the arrival process. In this Letter, we further extend these relationships to the case of a multi-server queue with geometric service times.

System description and notations: We consider a discrete-time queueing system with \( c (c \geq 1) \) servers and an infinite storage capacity. Time is divided into fixed-length slots. Customers (or ‘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 served based on a first-come-first-served (FCFS) discipline. The service of a packet can start or end at slot boundaries only. The service times of the packets are assumed to be independent geometrically distributed random variables with common probability generating function (pgf) \( G(z) = \frac{\mu}{1 - (1 - \mu)z}, \ 0 < \mu \leq 1 \). Moreover, the service and arrival processes are mutually independent. Finally, we assume that the system can reach a steady state.

Let us denote by \( v \) the system contents (i.e. the total number of packets in the buffer system, including the packets under service, if any) at the beginning of slot \( k \), by \( e \) the number of packet arrivals during slot \( k \), and by \( t \) the total number of departures at the end of slot \( k \). Then, we have the following system equations

\[
v_{k+1} = v_k - t_k + e_k;
\]

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

where in view of the geometric service-time distribution, the \( t_{k,j} \)'s are independent Bernoulli random variables with parameter \( \mu \). This implies

\[
T_k(z) := E[z^q] | \text{min}(v_k, c) = j] = (1 - \mu + \mu z^j), \quad i = 0, 1, \ldots, c
\]

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

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

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

\[
P(z, x) := \sum_{i=0}^{c} \sum_{j=0}^{c} p(i, j)z^i x^j
\]

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 \( V(z) = P(z, 1) \).

Relationship between system contents and delay: The delay of a packet is defined as the total number of slots between the end of the packet's arrival slot and the end of the slot where the service of the packet finishes and the packet leaves the system. We focus on an arbitrary packet \( P \), that arrives in the system during some slot \( j \) in the steady state. Let \( d \) with pgf \( D(z) \) be the delay of \( P \). Since the delay of a packet is equal to the sum of its waiting time and its service time, \( D(z) \) is obtained as

\[
D(z) = W(z)G(z)
\]

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

In the rest of this Section, we concentrate on the derivation of \( W(z) \).

First, we establish a relationship between \( W(z) \) and the pgf \( Q(z) \) of \( q \), the number of packets present in the system right after slot \( J \), that are selected for service before \( P \). Next, we express \( Q(z) \) in terms of the pgf \( V(z) \) of \( v \). Combination of these results with (4) then yields the envisaged relationship between the system-contents and delay distributions.

(a) Relationship between \( W(z) \) and \( Q(z) \): Let us observe the slots following slot \( J \). In slot \( J+i \), there will be either \( I_{J+i} = \sum_{j=1}^{c} \) packets leaving the system at the end of the slot, if \( q \geq c \) (i.e. at the beginning of slot \( J+2 \), there will be \( q-I_{J+1} \) packets in the system with priority over \( P \) to be taken into service), or \( P \) will get into service, if \( q < c \). In case \( P \) did not get into service, an analogous reasoning holds for slot \( J+2 \): either there will be \( q-I_{J+1} - I_{J+2} \) packets with service priority over \( P \) at the end of slot \( J+2 \), if \( q-I_{J+1} \geq c \), or \( P \) will get into service, if \( q-I_{J+1} < c \). Eventually, we see that \( P \) will still be waiting for service during slot \( J+i+1 \) only if \( q-I_{J+1} - I_{J+2} - \cdots - I_{J+c} \geq c \). This leads to

\[
w > i \Leftrightarrow q \geq s_i
\]

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

\[
s_0 := c; \quad s_i := 1 + \sum_{j=1}^{c} I_{J+i} + c, \quad i \geq 1
\]

Since the variables \( I_{J+i} \) are independent and identically distributed, the pgf of \( s_i \) can be expressed as

\[
S_i(z) := E[z^{s_i}] = z^c T_i(c)
\]

The next step is to transform the relationship (5) between \( W(z) \) and \( q \) into a relationship between the pgf's. To this end, we proceed as follows:

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

\[
= \sum_{i=0}^{c-1} z^i \text{Prob}[q = i] \sum_{j=0}^{c} z^j = \frac{1}{z} \sum_{i=0}^{c} z^i d^{E} w_i = \frac{1}{z} \sum_{i=0}^{c} z^i d^{E} w_i
\]

where in the second last step, we have used the statistical independence of \( q \) and the \( s_i \)'s, and the probability generating property of pgf's. Remark that working out the sum over \( i \) in the last step requires that \( |zT_i(c)| < 1 \) in the neighbourhood of \( z = 0 \). This condition will always be fulfilled for \( |z| \leq 1 \), since \( |T_i(c)| < 1 \) for \( |s| < 1 \). To derive the partial derivatives in (7), we replace the argument by its partial fraction expansion. Specifically, considering \( z \) to be a constant and \( x \) the variable of interest, we have

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

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 solutions for \( x \) in terms of \( z \) of the equation \( 1-zT_i(x) = 0 \), which we assume to be distinct. Substituting (8) in (7), calculating the partial derivatives, working out the sum over \( n \) and then again using the (8), now at \( x = 1 \), we finally find

\[
W(z) = \left( z - 1 \right) \sum_{p=0}^{c-1} \frac{x^{-1}Q(1/x_p)}{zT_i(x_p)(1-x_p)}
\]

(b) Relationship between \( Q(z) \) and \( V(z) \): Let us denote by \( v^* \) the number of packets in the queueing system at the beginning of slot \( J \) by \( f \) the number of packets arriving during slot \( J \) but before \( P \), and by \( r \) the number of departures at the end of slot \( J \). Then \( q \) can be expressed as

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

To derive the pgf \( Q(z) \) of \( q \), we need an expression for the joint distribution of \( v^* \) and \( f \). This can be determined by conditioning on the value of the random variable \( v^* \), the number of packet arrivals during slot \( J \), as follows:

\[
\text{Prob}[v^* = i, f = j] = \sum_{k=1}^{c} \text{Prob}[f = j | v^* = i, \ e^* = k]
\]

\[
ELECTRONICS LETTERS 21st August 2003 Vol. 39 No. 17
\]
Here, the joint probability \( \text{Prob}[v = i, e = l] \) corresponds to the fraction of packets that arrive in a slot where there are \( i \) packet arrivals and the system contents at the beginning of the slot is \( l \), because \( P \) is an arbitrary packet. Since each such slot contains \( l \) packet arrivals and \( P \) could be any of these, it is clear that

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

where \( \sigma \) denotes the mean number of packet arrivals during an arbitrary slot. Together with the fact that \( P \) has been chosen randomly from all arriving packets, this gives

\[
\text{Prob}[v = i, e = l] = \frac{1}{l} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p(i, l)
\]

The joint pgf \( M(z, x) \) of \( v \) and \( f \) can be easily obtained from this as

\[
M(z, x) = E[z^v x^f] = \frac{P(z, 1) - P(z, x)}{\sigma(1 - x)}
\]

Let us now return to (10) for \( q \). By means of (11) and some standard \( z \)-transform techniques, (10) can be transformed into

\[
Q(z) = T(\frac{1}{z}) M(z, z) + \sum_{l=0}^{1} \frac{\sum_{i=0}^{\infty} \Phi_1(z) \sum_{j=0}^{\infty} p(i, l) z^j}{\sigma(1 - z)}
\]

where \( \Phi_1(z) \equiv [T(1/z) - T(1/z)]z^l \). Changing the order of the summations in (13), working out the sum over \( j \), and using (2), (3) and (12), we then obtain

\[
Q(z) = T(\frac{1}{z}) [P(z, 1) - P(z, z)] + \sum_{l=0}^{1} \frac{\sum_{i=0}^{\infty} z^i \Phi_1(z) p(i, l)}{\sigma(1 - z)}
\]

where \( \Phi_1(z) = \text{Prob}[v = l] \). Conversely, from (1) and (3), we have

\[
P(z, 1) = T(\frac{1}{z}) P(z, z) + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} z^j \Phi_1(z) p(i, j)
\]

From (14) and (15), we finally get

\[
Q(z) = \frac{\sum_{i=0}^{\infty} \Phi_1(z) p(i) - \Phi_2(z) V(z)}{\sigma(1 - z)}
\]

\textbf{Applicability:} Combination of (4), (9) and (16) leads to a relationship between the steady-state pgf’s \( \Phi(z) \) and \( D(z) \). It has been verified that our result is in correspondence with [1–3] for the special case of constant 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, thus making a separate delay analysis superfluous. 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 based on the dominant pole of \( D(z) \) [4].

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 queuing 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.

© IEE 2003

Electronics Letters Online No: 20030796

DOI: 10.1049/el:20030796

6 June 2003

P. Gao, S. Wittevrongel and H. Bruneel (SMACS Research Group, Department TELIN, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium)

References


