Completion Time of Service Unit Interrupted by PH-Markov Renewal Customers and its Applications

Fumiaki MACHIHARA

NTT Communication Switching Laboratories, Hashida Research Laboratory
3-9-11 Midoricho, Musashino-shi, Tokyo, 180 Japan

An analysis of completion times of service with PH-Markov renewal preemption is presented for several preemptive disciplines. Using the results and the theory of the queue with semi-Markov service times, the PH-MR, M/G₁, G₂/1 queue with PH-Markov renewal preemption is analyzed and the effect of preemption is discussed. The non-preemptive queue and FIFO queue are also analyzed.

1. INTRODUCTION

Completion time of service unit is defined as the length of a period that begins from the instant a service starts and ends at the instant the server becomes free [1]. It has an important role in analyzing preemptive priority queueing models. Gaver [21], Keilson [3] and Avi-Itzhak - Naor [4] have studied the completion time for the case in which the service time is interrupted by Poisson customers with general service times. In this paper, we extend their results to the case in which the service time is interrupted by PH-Markov renewal customers. In a PH-Markov renewal process, the Laplace-Stieltjes transform of each element for the semi-Markov kernel of this process is represented by a rational function.

At first, we analyze the busy period of the PH-MR/G/1 queue, where PH-MR signifies PH-Markov renewal arrivals. The busy period is a special case of the completion time for the preemptive resume discipline. Next, general completion time distributions for the preemptive resume and preemptive repeat disciplines are represented using the busy period distribution mentioned above. The completion time distributions have matrix structures and the Laplace-Stieltjes transforms of these distribution matrices are explicitly provided. Finally, these results are applied to the analyses of the PH-MR, M/G₁, G₂/1 queues with preemptive discipline in which the priority and non-priority customer arrival processes are PH-Markov renewal and Poissonian. Mean waiting times for the priority and non-priority customers are explicitly given. The formulas for the priority queue with non-preemption and the queue with FIFO discipline are also provided. These results make it possible to analyze the packet multiplexers for voice and data. Several service disciplines such as voice packet preemption priority, voice packet non-preemption priority and voice-and-data packet FIFO may be considered. These systems are generalized models of the PH-MR, M/G/1 with the FIFO discipline analyzed by Sriram [5] and Heffes [6]. Approximating the regular interarrival by a PH-interarrival distribution such as Erlang type one, we can analyze the preemptive clocked schedules model [7,8,9].

2. PH-MARKOV RENEWAL PROCESS

We consider a Markov process on the state [1,2,...,m,m+1,...m+n] with infinitesimal generator
where elements of the \( m \times m \) matrix \( T \) satisfy \( T_{ii} < 0 \) for \( i = 1, \ldots, m \) and \( T_{ij} > 0 \) for \( i \neq j \). Elements of the \( m \times n \) matrix \( T^0 \) satisfy \( T^0_{ij} > 0 \) for all \( i \) and \( j \).

Also, \( (T + T^0)\mathbf{e} = 0 \) (\( \mathbf{e} \) is a column vector with zero elements) and the initial probability matrix \( Q \) is given by \((q, Q)\), where \( q \) is \( n \times m \) matrix also \( q \leq 1 \). This process has \( n \) absorbing states \([m+1, \ldots, m+n] \). Hereafter, \( X \) and \( \bar{X} \) signify matrix and vector representations, respectively.

**Lemma 2.1** The probability density \( f(x) \) of the time until absorption in a certain state of \([m+1, \ldots, m+n] \), corresponding to the initial probability matrix \( Q \), is given by

\[
f(x) = \mathbf{e}^T \exp(Tx) T^0. \tag{2.2}
\]

We further consider a Markov process with the infinitesimal generator

\[
A^0 = T + T^0. \tag{2.1}
\]

In this process, we see that the successive visits to the instantaneous states \([m+1, \ldots, m+n] \) form the Markov renewal process with underlying density \( f(x) \). When the density function has the phase structure, we call this Markov renewal process a PH-Markov renewal process (PH-MRP).

## 3. BUSY PERIOD DISTRIBUTION OF PH-MR/G/1 QUEUE

Consider the PH-MR/G/1 queue, where the arrival process is PH-Markov renewal with underlying density \( f_1(x) = \mathbf{e}^T \exp(T_1x) T_1^0 \). When the arrival process is PH-renewal, the interarrival time pdf has the scalar representation \( \mathbf{e}^T \exp(T_1x) T_1^0 \) [5]. The service time distribution, \( G_1(x) \), is general and the Laplace-Stieltjes transform (LST) is given by \( g_1(s) \).

Let us consider the busy period of the PH-MR/G/1. The busy period is the length of time between the arrival of a customer at an empty queue and the first epoch thereafter in which the queue becomes empty again. In the PH-MR/G/1 queue, the joint density of the busy period length and the number of customers served during this period, \( b(x,n) = (b_{ij}(x,n)) \), has the following structure. Let \( D \) denote the arrival epoch of a customer at an empty queue and let \( T \) denote the length of the busy period. \( b_{ij}(x,n) \) is defined to be

\[
b_{ij}(x,n) = P[x < T < x+dx, A(T) = n, J(T) = j / K(0) = i],
\]

where \( J(T) \) and \( K(0) \) have the values on the states \([1, \ldots, m]\) and \([m+1, \ldots, m+n]\), respectively. \( A(T) \) denotes the number of customers served during \([0, T]\).

Define the double transform of \( b(x,n) \) as:

\[
b_1^*(s,z) = \int_0^\infty e^{-sx} \int_0^z b_{ij}(x,t) dt dx \quad (i = 1, \ldots, n, \quad j = 1, \ldots, m)
\]

**Theorem 3.1** The double transform \( b_1^*(s,z) \) is given by the minimal non-negative solution of

\[
b_1^*(s,z) = g_1^*(s) - T^0 - T_1^0 b_1^*(s,z), \tag{3.1}
\]

where \( I \) is the identity matrix and

\[
g_1^*(z) = \int_0^\infty \exp(-xz) dG_1(x). \tag{3.2}
\]

**Corollary 3.2** The arrival phase states probability vector, \( \mathbf{P} \), at the busy period completion epochs is given by

\[
\mathbf{P} = \frac{1}{2} T_1^{-1} T^0 \mathbf{b}_1^*(0,1) = \mathbf{P} \quad \text{and} \quad \mathbf{P} \mathbf{e} = 1.
\]

In particular, when the arrival process is PH-renewal, the arrival phase state probability vector is given by \( \mathbf{P} = b_1^*(0,1) \).
4. COMPLETION TIME

Consider the service time of a non-priority customer underlying the distribution \( G_2(x) \). We study the completion time of the service which is preempted by the priority customers whose interarrival time density and service time distribution are \( f_1(x) \) and \( G_1(x) \), respectively. Suppose that the service process of a non-priority customer begins at 0 and ends at epoch \( T \) in \((x, x+dx)\). The completion time density, \( c(x) \), consists of elements

\[
C_{ij}(x)dx := \{0 < T < x + dx, J(T) = j / J(0) = i, i, j = 1, \ldots, m. \}
\]

**Theorem 4.1** The Laplace transform \( c^*(s) \) of the completion time density \( c(t) \) is presented by

\[
c^*(s) = \left\{
\begin{array}{ll}
g^*_2(s) - T_2 - T_1 - T_1 b^*_2(s, 1) & \text{(Preemptive Resume)}, \\
[0, [I - (s I - T_1)]^{-1}(1 - \exp(-s I - T_1)) T_1 b^*_1(s, 1)]^{-1} \\
\cdot \exp(-s I - T_1) dG_2(z) & \text{(Preemptive Repeat Identical)}, \\
[I - (s I - T_1)]^{-1} T_1 b^*_1(s, 1)^{-1} \\
\cdot g^*_2(s I - T_1) & \text{(Preemptive Repeat Different)},
\end{array}
\right.
\]

where \( b^*_1(s, 1) \) has been given in the previous section.

5. APPLICATION: PRIORITY QUEUE WITH PH-MARKOV RENEWAL PREEMPTION AND ITS RELATED MODELS

We study the model in which priority PH-Markov renewal customers preempt the non-priority Poissonian customer services. Let \( f_1(x) = g_1 \exp(T_1(x)) \) and \( f_2(x) = \lambda_2 \exp(-\lambda_2 x) \) denote the interarrival time densities of PH-Markov renewal priority and non-priority Poissonian customers, respectively. Let \( G_1(x) \) and \( G_2(x) \) denote the service time distributions of PH-Markov renewal and Poissonian customers.

Since the arrival and service processes of the priority customers are not influenced by the non-priority customer ones, characteristics of the priority customers can be analyzed by the N/G/1 queue theory [11].

We study now characteristics of the non-priority customers. Let \( Z_1, Z_2, \ldots \) denote the successive service completion epochs of the Poissonian customers. We define \( N_k \) to be the number of Poissonian customers at epoch \( Z_k + 0 \) immediately after \( Z_k \). Utilization factor \( \sigma \) is given by

\[
\sigma = g_2^*(-T_1)^{-1} e / u_1 + \lambda_2 / u_2 ,
\]

where \( g \) is the invariant probability vector for \( g_1^*(-T_1)^{-1} e \) and \( u_1 \) and \( u_2 \) are the service rates of priority and non-priority customers, respectively. For \( \sigma < 1 \), we define the stationary probability vector and probability-generating function vector to be

\[
\pi_1 := \lim_{k \to \infty} P(N_k = 1) \quad \text{and} \quad g(z) := \sum_{i=0}^{\infty} z^i \pi_1^i .
\]

**Theorem 5.1** The probability-generating function vector \( g(z) \) is given by

\[
g(z)(z I - c^*(\lambda_2 - \lambda_2 z)) = \pi_1 [\lambda_2 I - T_2 - T_1 b^*_2(1 - z) ]^{-1}
\cdot (\lambda_2 z I - T_2 - T_1 b^*_2(1 - z) ) c^*(\lambda_2 (1 - z)) .
\]

It remains only to find \( \pi_0 \). Suppose that 0 is a service completion epoch of a Poisson customer who finds the system empty. Consider the first passage time

\[
T_f = \inf[Z_k N^*(Z_k + 0) = 0 / N^*(0) = 0],
\]

where \( N^*(x) \) denote the number of Poisson customers at \( x \).

Let us consider the joint density \( g(x, n) \) of \( T_f \) and the number, \( N_p \), of service completions of Poissonian customers during \((0, T_f)\). That is,
\[ g(x,n)dx := P[x < T_f < x+dx, N_p = n, J(T_f) = j/J(0) = i] \quad (1 \leq i, j \leq m). \]

**Lemma 5.2** The double transform

\[ a^*(s,z) = \int_0^\infty e^{-sx} \sum_{n=0}^\infty a(x,n)dx \]

is given by

\[ a^*(s,z) = \left[ \lambda_2 z - T_1 - T_1 b^*(\lambda_2,1) \right]^{-1} \left[ \lambda_2 b^*_S(m,s,z) \right] + \int_0^\infty e^{-sx} \left[ (\lambda_2^n) \exp(-\lambda_2 x/n!) b^*_S(\lambda_2,1)dx \right] b^*_S(m,s,z), \]

(5.2)

where

\[ b^*_S(m,s,z) = z \sum_{n=0}^\infty e^{-sx} \left[ (\lambda_2^n) \exp(-\lambda_2 x/n!) c(x)dx \right] b^*_S(n,s,z). \]

(5.3)

This \( b^*_S(s,z) \) is the double transform of joint density of the busy period length and the number of customers served during this period for a semi-Markovian service system \( M/SM/1 \) in which the Poissonian arrival rate is \( \lambda_2 \) and the service time density is \( c(x) \).

\( \pi_0 \) is determined by \( a^*(s,z) \) as follows:

\[ \pi_0 = g_0 (\text{da}^*(1,z)/dz \big|_{z=1})^d, \]

where \( g_0 \) is the invariant probability vector of \( a^*(0,1) \).

\( g_0 \) is equivalent to the stationary arrival phase state probability vector at a service completion epoch of a non-priority customer who finds the system empty. As mentioned by Neuts and his colleagues [10,12], \( a^*(s,z) \) and \( da^*(s,z)/dz \) may be computed by (modified) successive substitutions. Under the preemptive resume discipline, in particular, \( a^*(s,z) \) can be transformed to more suitable form for computing. Under this discipline, the server is indifferent as to whether or not the PH-Markov renewal customers have priority. Therefore, the joint distribution of the busy period length and the number of service completions during this period are independent of the service order. This implies that \( a(x,n) \) is independent of the service order, and the following lemma is satisfied.

**Lemma 5.3**

\[ a^*(s,z) = \left[ \lambda_2 z - T_1 - T_1 b^*(\lambda_2,1) \right]^{-1} \left[ \lambda_2 z - b^*_S(m,s,z) \right] + \int_0^\infty e^{-sx} \left[ (\lambda_2^n) \exp(-\lambda_2 x/n!) c(x)dx \right] b^*_S(n,s,z), \]

(5.5)

where \( b^*_S(s,z) \) is the Laplace-Stieltjes transform of the M/G/1 busy period length (the arrival rate is \( \lambda_2 \) and the service time distribution is \( G_1(x) \)). And \( b^*_S(s,z) \) is the double transform of the PH-MR/G/1 busy period length and the number of customers served during this period, where the transforms of the interarrival time and service time distributions are \( f^*_1(s) \) and \( g^*_1(s) := g^*_1(s + \lambda_2 - \lambda_2 b^*_2(s,1)). \)

Now, we propose the computational method for \( a^*(0,1) \) and \( da^*(0,z)/dz \big|_{z=1} \). Hereafter, it is assumed that both \( b^*_S(0,1) \) and \( T_1^0 b^*_S(0,1) \) are diagonalizable. When the order of \( T_1 \), \( m \), is 2, this assumption is true. Even when \( m > 2 \), it is assured to be true for many computational experiments.

**Lemma 5.4** \( b^*_S(0,1) \) can be obtained by the following iterative substitutions:

\[ b^*_S(0,1) = g^*_1 b^*(\lambda_1,1), \]

\[ b^*_S(n+1)(0,1) = g^*_1 b^*(\lambda_1,1) - T_1 b^*_S(n,0,1), \quad n = 0, 1, \ldots, \]

where

\[ g^*_1 b^*(\lambda_1,1) = \int_0^\infty \text{diag}[\exp(-\alpha_1(n)), \ldots, \exp(-\alpha_m(n))] \left( P(n) - 1 \right) dG_1(x) \]

\[ = \sum_{n=1}^\infty \text{diag}[g^*_1(\alpha_1(n)), \ldots, g^*_m(\alpha_m(n))] P(n-1) \]

5A.5.4
for \( P(n-1) \langle -T_{-1} - T_{-1} b_{c}^{0}(0,1) \rangle P(n) = \text{diag}[\alpha_{1}(n), \ldots, \alpha_{m}(n)] \).

For sufficiently large \( n = N \), \( b_{c}^{0}(0,1) \) can be determined as
\[
b_{c}^{0}(0,1) = b_{c}^{0}(N)(0,1) \cdot P(N) \cdot \text{diag}[\alpha_{1}(N), \ldots, \alpha_{m}(N)] \cdot P(N)^{-1}. \tag{5.6}
\]
Without loss of generality, \( \alpha_{1} = \alpha_{1}(N) \) and \( P = P(N) \) may be written. Then the following lemma is satisfied.

Lemma 5.5
\[
\begin{align*}
b_{c}^{0}(0,1) & \sim b_{c}(0,1) \\
\text{satisfies} \quad & \text{Lemma 5.5}.
\end{align*}
\tag{5.7}
\]

\[ db^{(0, z)}/dz \bigg|_{z=1} \equiv (I - g_{1}P \text{diag}[-g_{c}^{(0,1)}, \alpha_{2}^{-1}(1-g_{c}(\alpha_{2})), \ldots, \alpha_{m}^{-1}(1-g_{c}(\alpha_{m}))] P^{-1} b_{c}(0,1) \tag{5.8}
\]

and
\[ db^{(0, z)}/dz \bigg|_{z=1} \equiv (I - g_{1}P \text{diag}[-g_{c}^{(0,1)}, \alpha_{2}^{-1}(1-g_{c}(\alpha_{2})), \ldots, \alpha_{m}^{-1}(1-g_{c}(\alpha_{m}))] P^{-1} b_{c}(0,1) \tag{5.9}
\]

where
\[-g_{c}^{(0,1)} = 1/u_{1}(1 - \alpha_{2}) \quad \text{and} \quad -b_{c}^{(0,1)} = 1/u_{2}(1 - \alpha_{2}) \quad \text{for} \quad \alpha_{2} = \lambda_{2}/u_{2}.
\]

Considering \( dz_{1}/dz \bigg|_{z=1} = 1/(1 - \alpha_{2}) \) and \( dz_{2}/dz \bigg|_{z=1} = \lambda_{2}/u_{1}(1 - \alpha_{2}), \)
\((\alpha_{1} = \lambda_{1}/u_{1}) \) from Lemma 5.3 through (5.5), we can compute \( s^{*}(0,1) \) and
\[ da^{*}(0,1)/dz \bigg|_{z=1} = (\lambda_{2}/(1 - \alpha_{2}))[\lambda_{2}I - T_{-1} - \lambda_{2}b_{c}(\lambda_{2}, 1)]^{-1}
\]
\[ + [I + (\lambda_{2}/u_{2})b_{c}(0,1)/dz \bigg|_{z=1} + u_{1}^{-1} db_{c}(0,1)/dz \bigg|_{z=1} \equiv e.
\]

Let \( w_{N}^{*}(s) \) denote the LST of the distribution vector of the waiting time of an arbitrary non-priority customer waiting in the queue for service for the first time. The LST of the sojourn time distribution of this customer is given by \( h_{N}^{*}(s) = w_{N}^{*}(s) e(s) \). The distribution of the number of non-priority Poissonian customers left behind by this customer has the probability generating function \( g(z) \) given by (5.1). If non-priority customer service is in arrival order, the customers behind this arbitrary customer must all have arrived during his sojourn time. We, therefore, obtain
\[ g(z) = h_{N}^{*}(1 - z) = w_{N}^{*}(1 - z) e.(\lambda_{2}(1 - z)) e.
\]

The mean waiting time of non-priority customers, \( E[W_{N}] \), is given by
\[ E[W_{N}] = w_{N}^{*}(0) e = g'(1) e/\lambda_{2} + w_{N}^{*}(0) e = (0, e).
\]

\[ w_{N}^{*}(0) e \] is the probability vector satisfying
\[ g(1) = w_{N}^{*}(0) e \quad \text{and} \quad w^{*}(0) e = 1.
\]

Theorem 5.6
\[ g(1) \quad \text{and} \quad g'(1) \quad \text{are obtained as:}
\]
\[ g(1) = [\pi + \pi_{0}(\lambda_{2}I - T_{-1} - \lambda_{2}b_{c}(\lambda_{2}, 1))^{-1}(T_{-1} + \lambda_{2}b_{c}(0,1)) e^{*}(0)]
\]
\[ \cdot (I - c^{*}(0) + G)^{-1}, \tag{5.10}
\]

where \( \pi \) is the invariant probability vector of \( c^{*}(0) \) and \( G \) is a matrix with \( m \) identical rows equal to \( \pi \).
\[ g'(1) = \lambda_{2}g(1)u'(0) + \pi_{0}(\lambda_{2}I - T_{-1} - \lambda_{2}b_{c}(\lambda_{2}, 1))^{-1}
\]
\[ \cdot (Y'(1) e - \lambda_{2}Y(1) u'(0)), \]

where \( Y(z) = d(\lambda_{2}(1 - z)) \)
\[ \cdot (\lambda_{2}I - \lambda_{2}I + T_{-1} + \lambda_{2}b_{c}(\lambda_{2}(1 - z), 1))/z - d(\lambda_{2}(1 - z)).
\]

\( d(s) \) is the Perron-Frobenius eigenvalue [14] of \( c^{*}(s) \) and \( u(s) \) is the corresponding right eigenvector satisfying \( u(0) = e.\)
The explicit expression for \( \hat{y}(1) \) and \( \hat{y}'(1) \) are given as:
\[
\hat{y}(1) = \lambda_2 \left( I - \left( z_1 I d_1 b^*(0,1) \right) - d'(0) \left( T^0_{11} + T^0_{12} b^*(0,1) \right) \right) / (1 + \lambda_2 d'(0))
\]
and
\[
\hat{y}'(1) = \left( \lambda_2^2 / (1 + \lambda_2 d'(0)) \right) \left( 1 - \left( \lambda_2 T^0_{11} b^*(0,1) \right) \right) - \left[ 2d'(0) - \lambda_2 d''(0) / (1 + \lambda_2 d'(0)) \right] \left( I - \left( z_1 I d_1 b^*(0,1) \right) \right) e,
\]
where \( u'(0) = \pi (-c^*(0) e) e - (I - c^*(0) + G)^{-1} (-c^*(0) e) \),
\[
d'(0) = \pi e^*(0) e,
\]
and
\[
d''(0) = \pi e^*(0) e + 2\pi (c^*(0) - d'(0)) \left( I - c^*(0) + G \right)^{-1} (c^*(0) - d'(0)) e.
\]

Figure 1 shows the mean sojourn time of non-priority Poissonian customers for the PH-MR, M/D1, D2/1, where both of priority and non-priority customer service times have constant, \( d_1 \) and \( d_2 \), respectively. The arrival process of the priority customers is a second order Markov modulated Poisson process (2-MMPP), often called a switched Poisson process. This is very important as a model of packet multiplexers for voice and data [5,6]. The 2-MMPP is governed by
\[
\begin{pmatrix}
0_1 \\
0_1
\end{pmatrix}
= \begin{pmatrix}
T_{11} & T_{12} \\
T_{21} & T_{22}
\end{pmatrix}
\begin{pmatrix}
0_1 \\
0_1
\end{pmatrix}
\quad \text{and} \quad
\begin{pmatrix}
T^0_{11} \\
T^0_{12}
\end{pmatrix}
= \begin{pmatrix}
T^0_{11} & 0 \\
0 & T^0_{22}
\end{pmatrix},
\]
where \( T_{11} = -T_{12} \) and \( T_{21} = -T_{22} \).

The four parameters, \( T_{11}, T_{21}, T^0_{11} \) and \( T^0_{22} \) are determined by the arrival rate \( \lambda_1 \), the squared coefficient of variation \( c_2 \), the balanced condition and the autocorrelation coefficient of the sequence of interarrival times \( \Theta \) in a same manner as [13]. Poisson arrival rate \( \lambda_2 \) and mean service time \( d_2 \) are fixed as \( \lambda_2 = 0.2 \) and \( d_2 = 1 \).

The effect of \( \Theta \) is very noticeable. This implies that it is not good to approximate the voice packet arrival process with a renewal process. The effect of service time \( d_1 \) is not negligible. Therefore, it is uncertain whether the priority PH-MR, M/G1, G2/1 can be well-approximated with the more suitable model PH-MP, M/G/1, in which the service time distribution is an appropriate mixture of the service times of two classes of customers such as FIFO queue [6].

Our results can be directly applied to the queue where the PH-MR customers have non-preemptive priority, because the waiting time of the non-priority Poissonian customers for the non-preemptive queue is identical to that for the preemptive queue. Therefore, the LST, \( \hat{y}_N(s) \), of the sojourn time distribution of non-priority customers is given by \( \hat{y}_N(s) e \). In particular, the mean sojourn time is given by
\[
E[\hat{y}_N(s)] = E[w_N] + d_2 = E[h_N] - (c^*(0) - c^*(0)) e - d_2.
\]

Figure 2 shows the mean completion time \( w^*(0) (-c^*(0)) e \). For sufficiently small \( c_1 \equiv 0 \) and sufficiently large \( c_1 \equiv 1 \), it is satisfied that
\[
w^*(0) (-c^*(0)) e = d_2 / (1 - c_1).
\]
This is clear for \( c_1 \equiv 0 \). Since \( \lim_{c_1 \to 0} w^*(0) = \pi \), where \( \pi \) is the invariant probability vector of \( c^*(0) \), (5.12) is satisfied. It is easily proved that
\[
\pi (-c^*(0)) e = d_2 / (1 - c_1).
\]
Let \( E[h_p] \) and \( E[\hat{y}_p(s)] \) denote the mean sojourn times of priority customers under preemptive resume discipline and non-preemptive discipline. Since the weighted sum of conditional waiting times, \( c_1 E[h_p] + c_2 E[h_N] \), is independent of the queueing discipline [15], it is satisfied that
\[
E[\hat{y}_p(s)] = E[h_p] + c_2 (E[h_N] - E[h_N]) / c_1.
\]
For the FIFO discipline, the individual mean waiting times can be derived in a similar manner.

Fig. 1 Mean Sojourn Times

Fig. 2 Mean Completion Time

REFERENCES