Decomposition Formulas for Single Server Queues with Vacations : A Unified Approach by the Rate Conservation Law

Masakiyo Miyazawa
Department of Information Sciences
Science University of Tokyo, Noda, Chiba 278, Japan

Published, Stochastic Models 10, No. 2, 389-413, 1994

Abstract We study single server vacation models under general vacation policies. By using the rate conservation law we obtain a unified approach for deriving decomposition formulas for these vacation models. We consider the stationary distributions of the workload and queue length in $M/GI/1$ type vacation models and of the virtual workload in $M/GI/1$ vacation models. These extend the decomposition formulas obtained earlier. For the exhaustive and multiple vacation policies, we give an alternative proof for the decomposition of the virtual workload distribution for a stationary input. For $GI/GI/1$ vacation models, stochastic inequalities are obtained for the workload distribution when the interarrival time distributions belong to certain classes. We further discuss generalizations of decomposition formulas for the case of the Markovian arrival process (MAP).

Keywords: M/GI/1 type queue, GI/GI/1 queue, G/G/1 queue, server vacation, general vacation policy, workload, queue length, decomposition, stochastic ordering, NBUE(NWUE) distribution, Markovian arrival process.

1. INTRODUCTION

A single server queue with server vacations is the queueing model in which the server takes vacations, i.e., service is stopped for random durations, whose starting times and lengths may depend on the history of the system. The vacation model is useful in studying cyclic server, priority queues and others. In particular, the $M/GI/1$ queue with server vacations is of basic importance because, for a large class of vacation policies, the stationary distributions of its workload and queue length can be decomposed, respectively, into two independent components, one of which is calculated by using the corresponding model without server vacations. Doshi [7] provides a nice overview on this topic. He proved them by using level crossing arguments or by sample path arguments (see [6] and [8]). Bardhan and Sigman [1] recently gave a simple derivation of the decomposition formulas for the workload by using the rate conservation law. Miyazawa [21] discussed a certain generalization of the decomposition formulas for the $M/GI/1$ queue with two levels of service rates, which includes the vacation models as special cases.

Here we develop those new approaches. Our purpose is to derive decomposition formulas in single server queues with server vacations in a unified way. We consider the stationary distributions of the workload, virtual workload and queue length, where the virtual workload is the sum of the workload and the remaining vacation time. We are only concerned with stationary distributions, and the word "stationary" will be suppressed. For deriving decomposition formulas, we need additional assumptions, i.e. structures for models. Two typical assumptions have been used in the literature. The first assumption is Poisson arrivals of customers, and the second is the restriction of vacation policy to special cases such as the exhaustive and multiple vacation policy or Bernoulli schedules (see Section 2 for the details of those vacation policies). It has been shown that the first assumption can be removed for the decomposition of the virtual workload distribution if the second assumption is made (see e.g. [8], [12]). On the other hand, the restriction of vacation policy can be relaxed if the arrival process of customers during busy periods is Poisson (see e.g. [7]). This model will be called $M/GI/1$ type queue with general server vacations. Note that the arrival process of customers in idles periods may be arbitrary in the $M/GI/1$ type queue.

Those two generalizations have been studied by different approaches. We consider them by a single principle, the rate conservation law of Miyazawa

[18]. For this purpose, we begin with a general vacation model under stationary assumptions and apply the rate conservation law to it for deriving basic formulas used in the later sections. This approach not only provides unified proofs but also enables us to consider further generalizations of the decomposition formulas. This is not surprising because the rate conservation law holds under a very general situation and includes level crossing arguments, which have been widely used in the literature (see e.g. [9]). Refer to Miyazawa [22] for the details of the rate conservation law.

This paper has the following structure. In Section 2, we introduce a general single server model with server vacations and derive the basic formulas. Point processes and Palm probability measures concerning them will play important roles in those derivations. We consider four point processes generated by the arrival instants of customers in busy periods and in idle periods, by the departure instants of customers and by the starting instants of server vacations. We also note the PASTA and conditional PASTA properties.

In Sections 3 and 4, we derive the standard decomposition formulas known in the literature, and consider their extensions. Section 3 is concerned with the $M/GI/1$ type queues. We derive decomposition formulas for the workload, queue length and virtual workload distributions. The results due to Doshi [7] and Shanthikumar [23] are extended. For example, decomposition formulas for the queue length distribution are obtained under the assumption that service may be interrupted by server vacations, which is not allowed in [7] and [23]. Section 4 is concerned with a single server queue with stationary input process, which will be denoted by $G/G/1$. We here assume the exhaustive and multiple vacation policy. The vacation durations may be stationary but is assumed to be independent of the arrival process. Doshi [8] obtained the decomposition formula for the virtual workload in this model. This result is a generalization of [5] and [17], but his proof seems to have a problem (see Remark 4.2). Under stationary assumptions we give an alternative proof to his results, which provides us more detailed decomposition formulas.

In Sections 5 and 6, we consider the workload distributions of non-Poisson arrival models under general vacation policies. Our object is to see how the arrival process affects the decomposition formulas. Section 5 is concerned with the $GI/GI/1$ queue. By using basic formulas derived from the rate conservation law, we get stochastic inequalities on the decomposition when

the interarrival time distributions belong to NBUE or NWUE class, whose definition is given in Section 5. Section 6 is concerned with Markovian arrival processes, MAP for short, which was introduced by Lucantoni et al. [17]. We obtain vector versions of decomposition formulas for this case, which partially generalize the results of [17].

2. STATIONARY VACATION MODEL AND THE RATE CONSERVATION LAW

In this section we introduce a general framework for a single server queue with vacations and derive basic formulas by using the rate conservation law of Miyazawa [18]. Let us consider a single server queue with infinite waiting room and with server vacations under a general vacation policy. We fix a service discipline, which may be arbitrary as far as it satisfies work conserving principle, i.e. no additional work is produced or reduced by the server. We denote the arrival and service times of the n-th customer by τ_n and S_n, respectively, where $\cdots < \tau_0 \leq 0 < \tau_1 < \tau_2 < \cdots$. We will assume that not more than one customer arrives at once (see assumption (ii) below). Define a point process N_a by

$$N_a(B) = \#\{n|\tau_n \in B\} \qquad (B \in \mathcal{B}(R)) \ ,$$

where $\#$ denotes the number of elements and $\mathcal{B}(R)$ is the Borel σ-field on $R = (-\infty, +\infty)$. The sequence of random pairs $\Psi_a \equiv \{(\tau_n, S_n)\}_{n=-\infty}^{+\infty}$ is called a *marked point process* in the theory of point processes. We call N_a and Ψ arrival and input processes, respectively. Let $T_n = \tau_n - \tau_{n-1}$, which is the interarrival time between the n-th and $n-1$-th customers. We assume that

(i) Ψ_a is stationary under a probability measure P,

(ii) The arrival process N_a is a simple point process, i.e. $N_a(\{t\}) \leq 1$ for all $t \in R$, has a finite and positive intensity $\lambda \equiv E(N_a((0,1]))$, and the path-wise traffic intensity $\overline{\rho} < 1$ *a.s.* P, where

$$\overline{\rho} = \lim_{t\to\infty} \frac{1}{t} \sum_{n=1}^{N_a((0,t])} S_n \ .$$

A single server queue satisfying assumption (i) is denoted by $G/G/1$ (see e.g. [10]), while assumption (ii) ensures the existence of stationary processes

for various characteristics in the $G/G/1$ queue without server vacations (see Remark 2.1). On server vacations, we assume that

(iii) The vacation policy does not refer to future arrivals of customers and their service times.

The formal description of this condition will be given later. Bernoulli schedule, under which the server goes on vacations with a given probability every time when service is finished, is one of the simplest vacation policies. A typical vacation policy studied in the literature is that the server starts vacations only when the system becomes empty. This policy is called *exhaustive*, which clearly satisfies (iii). The exhaustive policy is called *multiple* if the server repeats vacations until he finds the system not empty. The exhaustive vacation policy is a special case of the policies that the server starts vacations by observing the workload or queue length. See Doshi [6] [7] and [9] for further examples of vacation policies, all of which satisfy (iii).

Under the above assumptions, we consider stationary processes of various characteristics of the queue. For this purpose, we usually need an additional stability condition, e.g. the assumption that the input rate is less than the rate of the service which is actually done. Such a condition greatly changes according to the vacation policy (see Remark 2.1 below for more discussions). So we here just assume the existence of basic stationary processes for simplicity. Let $W(t)$, $L(t)$ and $R_v(t)$ are the workload, queue length and remaining vacation time, respectively, at time $t \in R$, where the workload means the total remaining service time of customers in system, the queue length includes a customer in service and the remaining vacation time is assumed to be 0 if the server is busy. We assume that

(iv) Ψ_a, $\{W(t)\}_{t\in R}$, $\{L(t)\}_{t\in R}$ and $\{R_v(t)\}_{t\in R}$ are jointly stationary under probability measure P.

Note that (i) is included in (iv). Assumption (iv) is equivalent to assume the existence of a probability space $(\Omega, \mathcal{F}, P)$ satisfying the following assumptions.

(iv-1) There exists a measurable operator group $\{\theta_t\}_{t\in R}$ on Ω, i.e., θ_t is a measurable mapping from Ω to Ω for each t satisfying $\theta_s \circ \theta_t = \theta_{s+t}$ for all $s, t \in R$,

(iv-2) $\{\theta_t\}$ is stationary with respect to P, i.e. $P(D) = P(\theta_t^{-1}(D))$ for all $t \in R$ and all $D \in \mathcal{F}$,

(iv-3) Ψ_a, $\{W(t)\}$, $\{L(t)\}$ and $\{R_v(t)\}$ are consistent with $\{\theta_t\}$,

where $f \circ \theta_t(\omega) = f(\theta_t(\omega))$ $(\omega \in \Omega)$ for a function f defined on Ω, the marked point process Ψ_a is said to be consistent to $\{\theta_t\}$ if

$$\Psi_a \circ \theta_t = \{(\tau_n - t, S_n)\} \qquad \text{for all } t \in R\ ,$$

and a stochastic process $\{X(t)\}_{t \in R}$ does so if

$$X(s) \circ \theta_t = X(s+t) \qquad \text{for all } s, t \in R\ .$$

In the following, all random variables and stochastic processes are assumed to be defined on $(\Omega, \mathcal{F}, P)$.

Remark 2.1 Under assumptions (i) and (ii) there exist unique processes $\{W(t)\}$ and $\{L(t)\}$ which are jointly stationary with Ψ_a for the single server queue without vacation. This can be verified by using Loynes' arguments (see e.g. [10], [19]), in which stationary processes are obtained by letting the starting time of the system tend to $-\infty$. On the other hand, for the vacation model satisfying (i)-(iii), we can not apply Loynes' arguments to verify (iv) except for special cases such as Bernoulli schedules because its dynamics interacts with the input process in a complicated way. However, our main concerns are with the $M/GI/1$ type or $GI/GI/1$ queues for a general vacation policy and with the $G/G/1$ queues for the exhaustive and multiple vacation policy. For those models, in many cases, we can verify (iv) through constructing appropriate Markov or regenerative processes (see e.g. [6]). In particular, we will prove the existence of the stationary processes for the exhaustive and multiple vacation models in Section 4.

The stationary framework by $\{\theta_t\}$ is the basis of our approach. In particular, by (ii), we can define a probability distribution P_a on $(\Omega, \mathcal{F})$ by

$$P_a(D) = \lambda^{-1} E\left(\int_0^1 1_D \circ \theta_u N_a(du)\right) \qquad (D \in \mathcal{F})\ , \tag{2.1}$$

where 1_D is the indicator function of a set D. This probability measure is called a Palm probability measure of P concerning N_a. It represents the conditional probability measure of P given that a customer arrives at time

0 (see [10] and [18] for its details). Define the traffic intensity $\rho = E(\overline{\rho})$. Then we get, by the ergodic theorem,

$$\rho = E\left(\int_0^1 S_1 \circ \theta_u N_a(du)\right) = \lambda E_a(S_1) \tag{2.2}$$

where E_a denotes the expectation concerning P_a.

Remark 2.2 Assumption (i) can be replaced by its synchronous version, i.e. by the assumption that $\{(T_n, S_n)\}$ is a stationary sequence. In this case, we begin with a Palm probability measure and define a time-stationary probability measure P by using the inversion formula of point processes (see e.g. [10]). Thus, for example, the $GI/GI/1$ queue falls into our framework in this sense.

In addition to the above processes, we will consider the processes of the following characteristics defined for each time t.

$\alpha(t) = -W'(t)$, where $W'(t)$ is the right-hand derivative of $W(t)$,

$R_s(t)$: the remaining service time of a customer in service or of a waiting customer who has been interrupted by server vacation,

$V(t) = W(t) + R_v(t)$: the virtual workload.

Note that the server is busy (idle) at time t if and only if $\alpha(t) = 1 (= 0)$. So, the time intervals in which $\alpha(t) = 1$ are called busy periods while those of $\alpha(t) = 0$ are called idle periods. Without loss of generality, we can assume that all processes $\{W(t)\}$, $\{\alpha(t)\}$, $\{R_s(t)\}$, $\{R_v(t)\}$, $\{V(t)\}$ and $\{L(t)\}$ are right-continuous and have left-hand limits for all $\omega \in \Omega$. By (iv-1)-(iv-3), all of them are consistent with $\{\theta_t\}$ and therefore stationary under P.

For each $t \in R$, let $\mathcal{F}_t$ be a sub-σ field of $\mathcal{F}$ so that it is increasing in t and includes all events generated by $\{(\tau_n, S_n)\}_{\tau_n \le t}$ but does not include those of additional information by $\{(\tau_n, S_n)\}_{\tau_n > t}$. In the literature (see e.g. [3]), $\{\mathcal{F}_t\}_{t \in R}$ is called a filtration. Denote the n-th starting time of server vacations and its length by γ_n and C_n, respectively, where $\cdots < \gamma_0 \le 0 < \gamma_1 < \gamma_2 < \cdots$. Then, assumption (iii) is expressed by

(iii') For each integer n, γ_n and $\gamma_n + C_n$ are stopping times with respect to the filtration $\{\mathcal{F}_t\}$.

Note that $\{W(u)\}_{u\le t}$, $\{L(u)\}_{u\le t}$ and $\{\alpha(u)\}_{u\le t}$ are determined by $\{\gamma_n\}_{\gamma_n\le t}$, $\{\gamma_n+C_n\}_{\gamma_n+C_n\le t}$ and $\{(\tau_n, S_n)\}_{\tau_n\le t}$. Hence, (iii') implies that $\{W(t)\}$, $\{L(t)\}$ and $\{\alpha(t)\}$ are adapted to $\{\mathcal{F}_t\}$, where a stochastic process $\{X(t)\}$ is said to be adapted to the filtration $\{\mathcal{F}_t\}$ if $X(t)$ is measurable with respect to $\mathcal{F}_t$ for each $t \in R$. Thus, γ_n and $\gamma_n + C_n$ may eventually depend on the past histories of $\{W(t)\}$ and $\{L(t)\}$. However, (iii') does not imply the adaptiveness of $\{R_v(t)\}$ with respect to $\{\mathcal{F}_t\}$ because C_n for n satisfying $\gamma_n \le t$ may not be measurable with respect to $\mathcal{F}_t$. We need a stronger condition for this adaptiveness. That is,

(iii") For each $t \in R$ and integer n satisfying $\gamma_n \le t$, C_n is measurable with respect to $\mathcal{F}_t$.

This assumption will be only used in Theorem 3.3.

Define a marked point process $\Psi_v = \{(\gamma_n, C_n)\}$ and a point process N_v by

$$N_v(B) = \#\{n|\gamma_n \in B\} \qquad (B \in \mathcal{B}(R)) \ ,$$

Similarly, define point processes N_b and N_i, respectively, by

$$N_b(B) = \sum_{\tau_n\in B} 1_{\{\alpha(\tau_n-)=1\}} \qquad N_i(B) = \sum_{\tau_n\in B} 1_{\{\alpha(\tau_n-)=0\}} \qquad \text{for } (B \in \mathcal{B}(R)) \ ,$$

N_b and N_i count customers arriving in busy and idle periods, respectively. From assumption (iv), Ψ_v, N_b and N_i are consistent with $\{\theta_t\}$, where a point process N is said to be consistent with $\{\theta_t\}$ if $N(B) \circ \theta_t = N(B + t)$ for all $t \in R$, $B \in \mathcal{B}(R)$ and for $B + t = \{u + t|u \in B\}$. Because N_b and N_i are thinning of N_a, they have finite intensities λ_b and λ_i, respectively, both of which are assumed to be positive. Hence, similarly to N_a, we can define Palm probability measures P_b and P_i of P with respect to N_b and N_i, respectively. We denote the expectations concerning them by E_b and E_i, respectively. In certain cases, we further assume that

(v) The point process N_v has a finite and positive intensity λ_v and $E_v(C_0)$ is finite, where E_v denotes the expectation concerning the Palm probability measure P_v of P with respect to N_v.

We are now in a position to derive general formulas by using the rate conservation law of Miyazawa [18] (Corollary 3.1 of the paper). We first apply the rate conservation law to the stationary process $X(t) = e^{-\theta W(t)}$,

where θ is a nonnegative number. We should be careful that θ_t is an operator on Ω while θ is a number. Since $W'(t) = -\alpha(t)$, we get

$$\begin{aligned}\theta E(e^{-\theta W(0)}; \alpha(0) = 1) &= \lambda E_a\left(e^{-\theta W(0-)}(1 - e^{-\theta S_0})\right) \\ &= \lambda_b E_b\left(e^{-\theta W(0-)}(1 - e^{-\theta S_0})\right) + \lambda_i E_i\left(e^{-\theta W(0-)}(1 - e^{-\theta S_0})\right) \ , \quad (2.3)\end{aligned}$$

where $E(X; D) = E(X1_D)$ for a random variable X and for $D \in \mathcal{F}$. By dividing both sides of (2.3) by θ and letting θ tend to zero, we have

$$P(\alpha(0) = 1) = \rho = \rho_b + \rho_i \ , \quad (2.4)$$

where $\rho_b = \lambda_b E_b(S_0)$ and $\rho_i = \lambda_i E_i(S_0)$.

We next consider the virtual work load process $\{V(t)\}$. Here we assume that the arrivals of customers and of the vacations do not occur simultaneously. Define $X(t) = e^{-\theta V(t)}$, and apply the rate conservation law. Since the right-hand derivative $V'(t) = -1$ for $V(t) > 0$, we have

$$\begin{aligned}&\theta E(e^{-\theta V(0)}; V(0) > 0) \\ &\quad = \lambda E_a\left(e^{-\theta V(0-)}(1 - e^{-\theta S_0})\right) + \lambda_v E_v\left(e^{-\theta V(0-)}(1 - e^{-\theta C_0})\right) \ . \quad (2.5)\end{aligned}$$

Similarly to Eqn. (2.4), we get

$$P(V(0) > 0) = \rho + \lambda_v E_v(C_0) \ . \quad (2.6)$$

We finally consider the queue length process $\{L(t)\}$. Let N_d be a point process generated by the departure instants of customers. Note that N_d has the intensity λ and its Palm probability measure P_d can be defined. In this case we let $X(t) = e^{-\theta R_s(t)} z^{L(t)}$. Since $X'(t) = \theta e^{-\theta R_s(t)} z^{L(t)} 1_{\{\alpha(t)=1\}}$ and the jump instants of $X(t)$ are decomposed into the point processes N_i, N_b and N_d, we get

$$\begin{aligned}&\theta E(e^{-\theta R_s(0)} z^{L(0)}; \alpha(0) = 1) \\ &\quad = \lambda_b E_b(e^{-\theta R_s(0-)} z^{L(0-)}; L(0-) \geq 1)(1 - z) \\ &\quad\ +\lambda_i \left[E_i(e^{-\theta R_s(0-)} z^{L(0-)}; L(0-) \geq 1)(1 - z)\right. \\ &\qquad\qquad \left. +P_i(L(0-) = 0) - z E_i(e^{-\theta S_0}; L(0-) = 0)\right] \\ &\quad\ +\lambda \left[E_d(z^{L(0)}(z - e^{-\theta R_s(0)}); L(0) \geq 1)\right. \\ &\qquad\qquad \left. +P_d(L(0) = 0)z - P_d(L(0) = 0)\right] \ , \quad (2.7)\end{aligned}$$

where E_d denotes the expectation concerning P_d. By letting $\theta = 0$ in (2.7), we get

$$\lambda_b E_b(z^{L(0-)}; L(0-) \geq 1) + \lambda_i E_i(z^{L(0-)}) - \lambda E_d(z^{L(0)}) = 0 \ . \quad (2.8)$$

In particular, by letting $z = 0$ in (2.8), we have

$$\lambda_i P_i(L(0-) = 0) - \lambda P_d(L(0) = 0) = 0 \ . \tag{2.9}$$

Eqn. (2.8) is a version of Finch's formula. The above equations will be used to derive decomposition formulas together with the following PASTA properties.

Suppose N_a is a Poisson process with respect to the filtration $\{\mathcal{F}_t\}$. That is, N_a is adapted to $\{\mathcal{F}_t\}$ and $N_a((u, u+h])$ is independent of $\mathcal{F}_t$ for all t, $u > t$ and $h > 0$. Then, as is well known (see e.g. [3]), the stochastic intensity of N_a with respect to $\{\mathcal{F}_t\}$ is the constant λ, and we have, for $t > 0$ and for $D \in \mathcal{F}_{0-}$,

$$E(\int_0^t 1_D \circ \theta_u N_a(du)) = E(\int_0^t 1_D \lambda du) = \lambda t P(D) \ , \tag{2.10}$$

where $\mathcal{F}_{t-} = \cap_{u<t} \mathcal{F}_u$. Eqn. (2.10) is known to characterize a Poisson process. (2.10) with $t = 1$ and the definition of the Palm probability measure (2.1) lead the following result, which is known as *PASTA* (Poisson Arrivals See Time Averages) (*e.g.* see [24]).

- (PASTA property) If N_a is a Poisson process with respect to $\{\mathcal{F}_t\}$, then $P_a(D) = P(D)$ for $D \in \mathcal{F}_{0-}$,

We need another PASTA property.

- (Conditional PASTA property) If the restriction of N_b on busy periods is a Poisson process with respect to $\{\mathcal{F}_t\}$, then $P_b(D) = P(D|\alpha(0) = 1)$ for $D \in \mathcal{F}_{0-}$,

This conditional PASTA property can be proved similarly to the PASTA property because the point process N_b admits the stochastic intensity $\lambda_b \alpha(t-)$ with respect to $\{\mathcal{F}_t\}$ and therefore we have

$$E(\int_0^t 1_D \circ \theta_u N_b(du)) = E(\int_0^t 1_D \lambda_b \alpha(u-) du) = \lambda_b t P(D; \alpha(0) = 1) \ . \tag{2.11}$$

Note that it is also a direct consequence of the conditional version of PASTA due to König and Schmidt [15].

Remark 2.3 We can not apply those PASTA properties to processes unless they are adapted to $\{\mathcal{F}_t\}$. Hence (iii) (equivalently (iii')) and, in a certain case, (iii") are essential in our applications.

3. STANDARD DECOMPOSITION FORMULA I

In this section we consider typical decomposition formulas found in the literature and their generalizations. We are concerned with $M/GI/1$ type and $M/GI/1$ queues with general server vacations. Here, we call the vacation model of Section 2 $M/GI/1$ type if it satisfies assumptions (ii), (iii) and (iv) and that the input process in busy periods is a compound Poisson process. Note that we do not make such assumptions for idle periods. This kind of models have been discussed in [23] and [7]. The terminology $M/GI/1$ type is due to Shanthikumar [23]. Let us introduce notations for the $M/GI/1$ type queue. Let λ_b^* be the intensity of its Poisson process restricted on busy periods, and G_b and G_i be the distribution functions of the service times of customers arriving in busy and idle periods, respectively. Note that G_b and G_i represent the distributions of S_n with respect to P_b and P_i, respectively. By the assumptions of the $M/GI/1$ type queue, S_n of customers arriving in busy periods are $i.i.d.$ with respect to P_b. Because the fraction of the time when the server is busy is ρ, it is easy to see that $\lambda_b = \rho\lambda_b^*$. We define $\rho_b^* = \lambda_b^* E_b(S_0)$. Similarly we denote the intensity of the point process N_i restricted to idle periods by λ_i^* and define $\rho_i^* = \lambda_i^* E_i(S_0)$. Since $\lambda_i = (1-\rho)\lambda_i^*$, (2.4) implies

$$\rho = \rho\rho_b^* + (1-\rho)\rho_i^* \ . \tag{3.1}$$

We first consider the workload distribution for the above $M/GI/1$ type queues with general server vacations. In this and the following sections, we drop parameters n and t of random variables X_n and $X(t)$, respectively, if their distributions do not depend on the parameters, and denote its Laplace-Stieltjes transform (LST for short) with respect to P (P_y) by $\hat{X}(\theta)$ and ($\hat{X}_y(\theta)$), and its distribution function by $\tilde{X}(x)$ ($\tilde{X}_y(x)$) for $y = a, v, b, i$. Similarly, the LST of distribution function F are denoted by $\hat{F}(\theta)$. Then (2.3) becomes

$$\begin{aligned}&\theta E(e^{-\theta W(0)}; \alpha(0) = 1)\\ &\quad = \rho\lambda_b^* \hat{W}_b^-(\theta)(1 - \hat{G}_b(\theta)) + (1-\rho)\lambda_i^* E_i(e^{-\theta W(0-)}(1 - e^{-\theta S_0}))\ ,\end{aligned} \tag{3.2}$$

where $\hat{W}_b^-(\theta)$ is LST of $W(0-)$ with respect to P_b. Hence, (3.1), (3.2) and the conditional PASTA property with (2.4) yield the following result.

Theorem 3.1 For the $M/GI/1$ type queue with general server vacations,

we have

$$\hat{W}_b^-(\theta) = \hat{W}_{M/GI/1(b)}(\theta)\frac{E_i(e^{-\theta W(0-)}(1-e^{-\theta S_0}))}{E_i(S_0)\theta} , \tag{3.3}$$

and, in particular, if $E_b(W(0-)) < \infty$,

$$E_b(W(0-)) = E(W_{M/GI/1(b)}) + E_i(W(0-)) + \frac{E_i(S_0^2)}{2E_i(S_0)} , \tag{3.4}$$

where

$$\hat{W}_{M/GI/1(b)}(\theta) = \frac{(1-\rho_b^*)\theta}{\theta - \lambda_b^*(1-\hat{G}_b(\theta))} ,$$

i.e. $W_{M/GI/1(b)}$ is the stationary workload of the no-vacation $M/GI/1$ queue with the arrival rate λ_b^* and the service time distribution function G_b.

Eqn. (3.3) is equivalent to (3.10) of [7]. To see this fact, we apply Neveu's cycle formula (*e.g.* see [4]) concerning P_v to the second fraction of (3.3). Then, we get

$$\begin{aligned}(1-\rho)\lambda_i^* E_i((e^{-\theta W(0-)}(1-e^{-\theta S_0})) &= \lambda_v E_v\left(\sum_{n=1}^{n_v} e^{-\theta W(\tau_n-)}(1-e^{-\theta S_n})\right) \\ &= \lambda_v E_v\left(e^{-\theta W(0-)} - e^{-\theta W(\tau_{n_v})}\right) ,\end{aligned}$$

where n_v is the number of customers arriving the idle period starting time 0, and the last equality follows from the fact that the server does not work during idle periods. Substituting this formula to (3.3), we obtain Doshi's [7] result. While he considered the workload processes concentrated on busy periods with additional work, we here consider the process on the whole time axis, which gives more detailed expression (3.3).

Eqn. (3.3) serves a decomposition formula (see Theorem 3.1 of [7]). We here derive more detailed decomposition formulas by assuming a further condition on the service times of customers arrived in idle periods.

Corollary 3.1 Under the conditions of Theorem 3.1, if the service times of customers arrived in idle periods are *i.i.d.* with a distribution function G_i and independent of the arrival process, then we have

$$\tilde{W}_b^-(x) = \tilde{W}_{M/GI/1(b)} * G_i^e * \tilde{Y}_i(x) , \tag{3.5}$$

$$\tilde{W}(x) = \rho\tilde{W}_{M/GI/1(b)} * G_i^e * \tilde{Y}_i(x) + (1-\rho)\tilde{Y}(x) , \tag{3.6}$$

where "$*$" denotes the convolution of distribution functions, $G_i^e(x) = \frac{1}{E_i(S_0)} \int_0^x (1 - G_i(u))du$ $(x \geq 0)$, and Y_i and Y are the workloads observed by an arbitrary customer arriving in idle periods and of the one at an arbitrary time in idle period, respectively.

Proof (3.5) is a direct consequence of (3.3). From (2.4) and (3.2), we have

$$\rho \hat{W}(\theta) - (1-\rho)\hat{Y}(\theta) = \rho \rho_b^* \hat{W}_b^-(\theta)\hat{G}_b^e(\theta) + (1-\rho)\rho_i^* \hat{Y}_i(\theta)\hat{G}_i^e(\theta) \ ,$$

where $\hat{G}_b^e(\theta) = \frac{1-\hat{G}_b(\theta)}{\theta E_b(S_0)}$. Substituting (3.3) into this formula, we get (3.6) after some manipulations using (3.1).

Remark 3.1 Eqn. (3.6) is a generalization of Theorem 9.1 of [7], in which the arrival process in idle periods are assumed to be Poisson. In particular, if the input process Ψ_a is compound Poisson with the service time distribution function G, then (3.6) reduces to

$$\tilde{W}(x) = \tilde{W}_{M/GI/1} * \tilde{Y}(x) \ , \tag{3.7}$$

which is the well-known decomposition formula (*e.g.* see [2] and [7]), where $W_{M/GI/1}$ is the stationary workload of the no-vacation $M/GI/1$ queue with the arrival rate λ and the service time distribution G.

We next consider the queue length distribution for the $M/GI/1$ type queue with general server vacations. We here assume two additional conditions.

(a) A customers in service is not preempted by other customers, and, if he is interrupted by a server vacation, his service is resumed when the server returns,

(b) The service times $\{S_n\}$ of all customers are *i.i.d.* with a distribution function G with respect to P_a.

Concerning assumption (a), we remark that the interruption of service by server vacations is not allowed in [7] and [23]. Assumption (b) implies that $\{S_n\}$ are also *i.i.d* with the same distribution function G concerning P, P_b and P_i. Since $R_s(0) = S_0$ on $\{L(0) \geq 1\}$ *a.s.* P_d by assumption (a), Eqn.

(2.7) implies

$$\begin{aligned}
&\theta E(e^{-\theta R_s(0)} z^{L(0)}; \alpha(0) = 1) \\
&\quad = \rho\lambda_b^* E_b(e^{-\theta R_s(0-)} z^{L(0-)}; L(0-) \geq 1)(1-z) \\
&\quad\quad +(1-\rho)\lambda_i^* \left[E_i(e^{-\theta R_s(0-)} z^{L(0-)}; L(0-) \geq 1)(1-z) \right. \\
&\quad\quad\quad\quad \left. +P_i(L(0-)=0) - P_i(L(0-)=0) z\hat{G}(\theta) \right] \\
&\quad\quad +\lambda \Big[E_d(z^{L(0)}; L(0) \geq 1)(z - \hat{G}(\theta)) \\
&\quad\quad\quad\quad +P_d(L(0)=0)z - P_d(L(0)=0)] \ . \qquad (3.8)
\end{aligned}$$

Let $\theta = \lambda_b^*(1-z)$ in Eqn. (3.8). Then, by using (2.4), (2.8) and the conditional PASTA property, we have

$$\begin{aligned}
&(1-\rho)\lambda_i^* \left[E_i(e^{-\theta R_s(0-)} z^{L(0-)}; L(0-) \geq 1)(1-z) - P_i(L(0-)=0) z\hat{G}(\theta) \right] \\
&\quad +\lambda \left[E_d(z^{L(0)}; L(0) \geq 1)(z - \hat{G}(\lambda_b^*(1-z))) + P_d(L(0)=0)z \right] = 0 \ .
\end{aligned}$$

Hence, by using (3.1) and (2.9), we have the following result.

Theorem 3.2 For the $M/GI/1$ type queue with general server vacations satisfying assumptions (a) and (b), we have

$$\begin{aligned}
E_d(z^{L(0)}) \ &= \ \frac{(1-\rho_b^*)(1-z)}{\hat{G}(\lambda_b^*(1-z)) - z} \\
&\quad \times \left[E_i(e^{-\lambda_b(1-z)R_s(0-)} z^{L(0-)}; L(0-) \geq 1) \right. \\
&\quad\quad \left. + \hat{G}(\lambda_b^*(1-z)) P_i(L(0-)=0) \right] \ . \qquad (3.9)
\end{aligned}$$

Remark 3.2 It is not surprising that $R_s(0-)$ appears in Eqn. (3.9). This is because we allow the interruption of service by a server vacation.

We can interpret Eqn. (3.9) as follows. The queue length distribution just after the departure instant equals that of the conventional $M/GI/1$ queue in which the number of customers arrived during the service of the customer who starts busy period equals in distribution the number of customers arrived during the remaining service time of the interrupted customer of the vacation model plus the number of customers observed by an arbitrary chosen customer arriving in idle period. In particular, if $R_s(0)$ and $L(0-)$ are

independent under P_i, then Eqn. (3.9) implies the following decomposition.

$$E_d(z^{L(0)}) = \frac{(1-\rho_b^*)(1-z)}{\hat{G}(\lambda_b^*(1-z))-z} \times \Big[\hat{R}_{s,i}(\lambda_b(1-z))E_i(z^{L(0-)}; L(0-) \geq 1) + \hat{G}(\lambda_b^*(1-z))P_i(L(0-)=0)\Big] \ . \tag{3.10}$$

We now assume that

(c) A customer in service is not preempted by server vacation.

Then, Eqn. (3.10) reduces to

$$E_d(z^{L(0)}) = \frac{(1-\rho_b^*)(1-z)\hat{G}(\lambda_b^*(1-z))}{\hat{G}(\lambda_b^*(1-z))-z} E_i(z^{L(0-)}) \ , \tag{3.11}$$

which is equivalent to (7.12) of [7], where our formula is again more detailed similarly to the case of the workload. Eqn. (3.11) reads

$$\tilde{L}_d(x) = \tilde{L}_{M/GI/1(b)} * \tilde{L}_i(x) \ , \tag{3.12}$$

where $\tilde{L}_{M/GI/1(b)}$ is the stationary queue length distribution of the no-vacation $M/GI/1$ queue with the arrival rate λ_b and the service time distribution function G.

We finally discuss the virtual workload process $\{V(t)\}$. We consider it for the $M/GI/1$ queue with general server vacations under the additional assumptions (iii") and (v). We denote the service time distribution function by G. Since (iii") together with (iii') implies that $\{V(t)\}$ is adapted to $\{\mathcal{F}_t\}$, (2.5), (2.6) and the PASTA property yield the following result due to Doshi [6] (see Section 6.4 of his paper).

Theorem 3.3 For the $M/GI/1$ queue with general server vacations satisfying the additional assumptions (iii") and (v), we have

$$\hat{V}(\theta) = \hat{W}_{M/GI/1}(\theta)\left(\frac{1-\rho-\rho_v}{1-\rho} + \frac{\rho_v E_v(e^{-\theta V(0-)}(1-e^{-\theta C_0}))}{(1-\rho)\theta E_v(C_0)}\right) \ , \tag{3.13}$$

and, in particular, if $E(V(0)) < \infty$,

$$E(V(0)) = E(W_{M/GI/1}) + \frac{\rho_v}{1-\rho}\left(E_v(V(0-)) + \frac{E_v(C_0^2)}{2E_v(C_0)}\right) \ , \tag{3.14}$$

where $\rho_v = \lambda_v E_v(C_0)$.

Remark 3.3 As discussed in [6], (3.13) holds in the more general situation where another vacations may arrive as the secondary work during the periods of vacations and they are accumulated. Kella and Whitt ([13]), [14] and Bardhan and Sigman [1] studied similar decomposition for a Levy process reflected at the origin and with an additional jump process. Their results correspond to Eqn. (3.13) with $\rho + \rho_v = 1$. Theorem 3.3 also includes the recent results of Leung [16] (Theorem 1 and Corollary 1 of his paper) in which the exhaustive vacation policy is assumed.

4. STANDARD DECOMPOSITION FORMULAS II

It is well-known that, for the $GI/GI/1$ queue with the exhaustive and multiple vacation policy, if C_n are $i.i.d.$ and independent of the input process, then Eqn. (3.13) with $\rho + \rho_v = 1$ holds. Lucantoni et al. [17] generalized it for a Markovian arrival process (see Section 3). Doshi [8] further generalized it by assuming the existence of limiting distributions and the independence of the input process and the sequence of the vacation lengths. We here give an alternative proof to Doshi's [8] results under assumptions (i) and (ii). In this section we assume for simplicity that the arrivals of customers and of the vacations do not occur simultaneously.

To make an argument transparent, we first consider the case of $GI/GI/1$ queue. Assume that $\rho < 1$, which is equivalent to assumption (ii) in this case. Then, as we mentioned in Remark 2.1, there exists a unique stationary process for the virtual workload in the no-vacation model. Consider the no-vacation model. Let $R_a^{(nv)}(t)$ be the remaining arrival time at time t. Here (nv) indicates no vacation model. This convention will be used in what follows. Then, $\{(V^{(nv)}(t), R_a^{(nv)}(t))\}$ is a stationary Markov process under P. Denote the distributions of $(V^{(nv)}(0), R_a^{(nv)}(0))$ with respect to P and P_a by $\nu^{(nv)}$ and $\nu_a^{(nv)}$, respectively. Note that $\nu_a^{(nv)}$ is reduced to one dimensional Palm distribution of $V^{(nv)}(t)$ because $R_a^{(nv)}(t-) = 0$ when a customer arrives at t. We denote LSTs of $\nu^{(nv)}$ and $\nu_a^{(nv)}$ by $\phi^{(nv)}(\theta, \eta)$ and $\phi_a^{(nv)}(\theta)$, respectively. Then, applying the rate conservation law to $X(t) = e^{-\theta V^{(nv)}(t) - \eta R_a^{(nv)}(t)}$, we get, similarly to (2.4),

$$(\theta + \eta)\phi^{(nv)}(\theta, \eta) = \theta\phi_0^{(nv)}(\eta) + \lambda\phi_a^{(nv)}(\theta)(1 - \hat{G}(\theta)\hat{F}(\eta)) \ , \qquad (4.1)$$

where $\phi_0^{(nv)}(\eta) = \int_0^\infty e^{-\eta u}\nu^{(nv)}(0, du)$ and F is the distribution of interarrival time T_n.

We next consider the corresponding vacation model with the exhaustive and multiple vacation policy. We assume that C_n are *i.i.d.* with a finite means and independent of the input process. Define the following LSTs.

$$\begin{aligned} \phi(\theta,\eta) &= \phi^{(nv)}(\theta,\eta)\hat{C}^e(\theta)\ , \qquad \phi_a(\theta) = \phi_a^{(nv)}(\theta)\hat{C}^e(\theta)\ , \\ \phi_v(\eta) &= \frac{\phi_0^{(nv)}(\eta)}{1-\rho}\ , \end{aligned} \tag{4.2}$$

where $\hat{C}^e(\theta) = \frac{1-E(e^{-\theta C_0})}{\theta E(C_0)}$. Then, from (4.1), we get

$$(\theta+\eta)\phi(\theta,\eta) = \lambda\phi_a(\theta)(1-\hat{G}(\theta)\hat{F}(\eta)) + \frac{(1-\rho)}{E(C_0)}\phi_v(\eta)(1-\hat{C}(\theta))\ . \tag{4.3}$$

We now apply Corollary 3.2 of [20], which simultaneously characterizes stationary distributions at an arbitrary point of time and just before jump instants. For this purpose, we first note that $\{(V(t), R_a(t))\}$ is a self-clocking jump process, SCJP for short, of Miyazawa [20] because the process is Markov and has jumps at the instants when either $V(t)$ or $R_a(t)$ attains 0. Here, there are two macrostates, one represents $V(t) > 0$ the other $V(t) = 0$. We next note that Eqn. (4.3) corresponds to (7) of [20]. Hence, by assumption (ii), Eqn. (4.3) shows that $\phi(\theta,\eta)$, $\phi_a(\theta)$ and $\phi_v(\eta)$ are stationary distributions of $(V(t), R_a(t))$ at an arbitrary point of time, just before the arrival instants of customers and just before the arrival instants of vacations, respectively. Thus, from (4.2), we get the well-known decomposition formulas (see e.g. [6]):

$$\hat{V}(\theta) = \hat{V}^{(nv)}(\theta)\hat{C}^e(\theta)\ , \qquad \hat{V}_a(\theta) = \hat{V}_a^{(nv)-}(\theta)\hat{C}^e(\theta)\ . \tag{4.4}$$

By using the above argument, let us prove the decomposition under the assumptions that $\Psi_a = \{(\tau_n, S_n)\}$ is a stationary marked point process satisfying assumptions (i) and (ii) and that $\{C_n\}$ is stationary sequence of random variables and independent of Ψ_a. We supplement the state $(V(t), R_a(t))$ by the histories of the input and the vacation durations up to just before time t, which are denoted by $\boldsymbol{H}_i(t) \equiv \{(S_n, T_n)\}_{n\le n(t)}$ and $\boldsymbol{H}_v(t) \equiv \{C_n\}_{n\le m(t)}$, respectively, where $n(t) = \sup\{n|\tau_n < t\}$ and $m(t) = \sup\{m|\gamma_m < t\}$. Then $\{(V(t), R_a(t), \boldsymbol{H}_i(t), \boldsymbol{H}_v(t))\}$ is a Markov process, and, in particular, a SCJP, too. This SCJP has infinitely many clocks but only two of them are active. Hence we can also apply Corollary 3.2 of [20] to the SCJP (see page 551 of [20]). Note that the rate conservation law of [18] yields, for the

no-vacation model,

$$(\theta+\eta)E(e^{-\theta V^{(nv)}(0)-\eta R_a^{(nv)}(0)};D_i)=\theta E(e^{-\eta R_a^{(nv)}(0)};D_i,V^{(nv)}(0)=0)$$
$$+\lambda\left[E_a(e^{-\theta V^{(nv)}(0-)};D_i^-)-E_a(e^{-\theta(V^{(nv)}(0-)+S_0)-\eta T_1};D_i)\right]\ , \quad (4.5)$$

and, for the vacation model under assumptions (iv),

$$(\theta+\eta)E(e^{-\theta V(0)-\eta R_a(0)};D_i,D_v)$$
$$=\lambda_v\left[E_v(e^{-\eta R_a(0)};D_i,D_v^-)-E_v(e^{-\eta R_a(0)-\theta C_0};D_i,D_v)\right]$$
$$+\lambda\left[E_a(e^{-\theta V(0-)};D_i^-,D_v)-E_a(e^{-\theta(V(0-)+S_0)-\eta T_1};D_i,D_v)\right]\ , \quad (4.6)$$

where $D_i=\{(S_n,T_n)\in B_n^{(2)}(\forall n\le 0)\}$, $D_v=\{C_n\in B_n(\forall n\le 0)\}$, $D_i^-=\{(S_n,T_n)\in B_{n+1}^{(2)}(\forall n<0)\}$ and $D_v^-=\{C_n\in B_{n+1}(\forall n<0)\}$ for arbitrary given sequences of $B_n^{(2)}\in\mathcal{B}(R^2)$ and $B_n\in\mathcal{B}(R)$. Then, we can construct stationary distributions of $(V(t),R_a(t),\boldsymbol{H}_i(t),\boldsymbol{H}_v(t))$ from those of the no-vacation model in a similar way as we did for the case of the $GI/GI/1$ queue. That is, since D_i and D_v (D_i^- and D_v^-) are independent under P and P (P_a), we have

$$E(e^{-\theta V(0)-\eta R_a(0)};D_i,D_v) = E(e^{-\theta V^{(nv)}(0)-\eta R_a^{(nv)}(0)};D_i)\frac{E(1-e^{-\theta C_0};D_v)}{\theta E(C_0)}\ ,$$
$$E_a(e^{-\theta V(0-)};D_i^-,D_v) = E_a(e^{-\theta V^{(nv)}(0-)};D_i^-)\frac{E(1-e^{-\theta C_0};D_v)}{\theta E(C_0)}\ ,$$
$$E_a(e^{-\theta(V(0-)+S_0)-\eta T_1};D_i,D_v)$$
$$=E_a(e^{-\theta(V^{(nv)}(0-)+S_0)-\eta T_1};D_i)\frac{E(1-e^{-\theta C_0};D_v)}{\theta E(C_0)}\ ,$$
$$\lambda_v E_v(e^{-\eta R_a(0)};D_i,D_v^-) = E(e^{-\eta R_a^{(nv)}(0)};D_i,V^{(nv)}(0)=0)P(D_v^-)\ ,$$
$$\lambda_v E_v(e^{-\eta R_a(0)-\theta C_0};D_i,D_v)$$
$$=E(e^{-\eta R_a^{(nv)}(0)};D_i,V^{(nv)}(0)=0)E(e^{-\theta C_0};D_v)\ . \quad (4.7)$$

Here, note that $P(D_v^-)=P(D_v)$. Thus we get the following results due to Doshi [8] (see Theorem 5.1 of his paper).

Theorem 4.1 For a single server vacation model with a stationary input satisfying (i) and (ii), if the vacation policy is exhaustive and multiple and if $\{C_n\}$ are stationary sequence with a finite expectation and independent of the input, then there exists a stationary process $\{V(t)\}$ for the virtual workload and we get

$$\hat{V}(\theta)=\hat{V}^{(nv)}(\theta)\hat{C}^e(\theta)\ , \qquad \hat{V}_a^-(\theta)=\hat{V}_a^{(nv)-}(\theta)\hat{C}^e(\theta)\ , \quad (4.8)$$

in particular, if $E(C_0^2)$ is finite,

$$\begin{aligned} E(V(0)) &= E(V^{(nv)}(0)) + \frac{E(C_0^2)}{2E(C_0)}\,, \\ E_a(V(0-)) &= E_a(V^{(nv)}(0-)) + \frac{E(C_0^2)}{2E(C_0)}\,. \end{aligned} \tag{4.9}$$

Remark 4.1 For the exhaustive vacation policy, if the service discipline is FCFS, then the virtual workload observed by arriving customers is identical with the waiting time. Hence, the second equations of (4.8) and (4.9) are decompositions of the waiting time distribution and its mean, respectively, for the FCFS queue.

Remark 4.2 Doshi [8] proved (4.8) under the assumptions that the limiting distributions of $V(t)$ and $V(\tau_n-)$ exist as t and n go to ∞ when the system starts at time 0 and the sequence $\{\sum_{i=1}^{n} C_i\}_{n=1}^{\infty}$ generates a stationary point process on $[0,\infty)$. His proof seems not correct because he concluded that, if $\{D_n\}_{n=1}^{\infty}$ is a stationary sequence and if $\{N(k)\}_{k=1}^{\infty}$ is a sequence of nonnegative integer-valued random variables satisfying $0 \le N(k+1) - N(k) \le 1$ and depending on $\{D_n\}$, then $D_{N(k)}$ is identically distributed for all $k \ge 1$ (see the proof of Theorem 5.1 of [8]). Furthermore, no sufficient conditions have not been obtained for the existence of those limiting distributions under general assumptions like (i) and (ii). On the other hand, in Theorem 4.1, we need assumptions (i) and (ii) but the existence of the stationary process, which can start with the specific initial distribution given by (4.7) at time 0, is verified. Remark that (4.7) themselves provide the more detailed decomposition formulas.

5. THE WORKLOAD DISTRIBUTION OF THE GI/GI/1 VACATION MODEL

We next consider the vacation model with a renewal arrival process and *i.i.d.* service times under assumptions (ii)-(iv). We denote the distribution functions of the interarrival T_n and of the service time S_n by F and G, respectively. As obtained in Theorem 4.1, if the vacation policy is exhaustive, then we can get the decomposition of the virtual work load distributions. But, for a general vacation policy or for other queueing characteristics, there

seems no decomposition results for the case of a general input. We try this problem for the work load distribution. Of course, we can not expect clean decomposition results in this case. We here get stochastic inequalities by restricting a class of the interarrival time distributions.

For the $GI/GI/1$ vacation model, we have, from (2.3),

$$\theta E(e^{-\theta W(0)}; \alpha(0) = 1) = \lambda \hat{W}_a^-(\theta)(1 - \hat{G}(\theta)) \ , \tag{5.1}$$

where $W^- = W(0-)$. After some manipulations, (5.1) yields

$$\hat{W}(\theta) = \frac{(1-\rho)\theta E(e^{-\theta W(0)}|\alpha(0) = 0) + \rho\theta(\hat{W}_a^-(\theta) - \hat{W}(\theta))\hat{G}^e(\theta)}{\theta - \lambda(1 - \hat{G}(\theta))} \ , \tag{5.2}$$

where $\hat{G}^e(\theta) = \frac{(1-\hat{G}(\theta))}{\theta E(S_0)}$. Let $Y(t)$ be the conditional work load at time t when the server is idle. Since $\hat{Y}(\theta) = E(e^{-\theta W(0)}|\alpha(0) = 0)$, (5.2) is equivalent to

$$\tilde{W}(x) = \tilde{W}_{M/GI/1} * \tilde{Y}(x) + \frac{\rho}{1-\rho}\tilde{W}_{M/GI/1} * G^e * (\tilde{W}_a^- - \tilde{W})(x) \ . \tag{5.3}$$

If the arrival process N_a is Poisson, then (5.3) leads the well-known decomposition formula (3.7) because $\tilde{W} = \tilde{W}_a^-$ by the PASTA property. For the non-Poisson case, we consider stochastic inequalities. For distribution functions F_1 and F_2, F_1 is called stochastically larger than F_2, which is denoted by $F_1 \geq_{st} F_2$, if $1 - F_1(x) \geq 1 - F_2(x)$ for all x. On the other hand, for a distribution function F of a nonnegative random variable with a finite mean m_F is said to be $NBUE$ ($NWUE$) if, for all $t \geq 0$,

$$\int_t^\infty (1 - F(u))du \leq (\geq) m_F(1 - F(t)) \ .$$

Here NBUE (NWUE) is the abbreviation of New Better (Worse) than Used in Expectation (see e.g. [24]). We state the following lemma, which is well-known for the conventional $GI/GI/1$ queue (*e.g.* see 4.6.12 of [10]).

Lemma 5.1 If the interarrival time distribution function F is NBUE (NWUE), then $\tilde{W} \geq_{st} (\leq_{st}) \tilde{W}_a^-$.

Proof Let an arbitrary $x \geq 0$ be fixed and suppose that F is NBUE. By the inversion formula of point processes (*e.g.* see [18]), we have

$$P(W(0) > x) = \lambda E_a \left(\int_0^{T_1} 1_{\{W(s) > x\}} ds \right) \ ,$$

where 1_D is the indicator function of a set D. Define

$$A(t) = \int_0^t \alpha(s)ds \ .$$

Then, $W(t) = W(0) - A(t)$ on the event $\{T_1 > t\}$ because the service discipline is work conserving. Hence, by using the independence assumption of $\{T_n\}$ together with (iii) and applying partial integration, we have

$$\begin{aligned} P(W(0) > x) &= \lambda \int_0^\infty P_a(W(0) - A(s) > x) P_a(T_1 > s) ds \\ &= P_a(W(0) > x) + \lambda \int_0^\infty \int_s^\infty P_a(T_1 > u) du d_s P_a(W(0) - A(s) > x) \ . \end{aligned}$$

Since $P_a(W(0) - A(s) > x)$ is nonincreasing in s, the definition of NBUE:

$$\int_s^\infty P_a(T_1 > u)du \le \frac{1}{\lambda} P_a(T_1 > s) \qquad (\forall s \ge 0) \ ,$$

implies

$$\begin{aligned} P(W(0) > x) &\ge P_a(W(0) > x) + \int_0^\infty P_a(T_1 > s) d_s P_a(W(0) - A(s) > x) \\ &= \int_0^\infty P_a(W(0) - A(s) > x) dP_a(T_1 \le s) = P_a(W(0-) > x) \ . \end{aligned}$$

The next theorem is an immediate consequence of (5.3) and Lemma 5.1.

Theorem 5.1 For the $GI/GI/1$ queue with general server vacations satisfying assumptions (ii), (iii) and (iv), if the interarrival time distribution function F is NBUE (NWUE), then we have

$$\tilde{W} \le_{st} (\ge_{st}) \ \tilde{W}_{M/GI/1} * \tilde{Y} \ . \tag{5.4}$$

Remark 5.1 Unfortunately, similar arguments can not be applied to the virtual work load and to the queue length.

6. THE CASE OF MARKOVIAN ARRIVAL PROCESS (MAP)

Because decomposition formulas deeply depend on Poisson structure of the arrival process for general vacation policies as we have seen in Section 3, it might be the next step to consider the arrival processes generated by continuous time Markov chain. For the exhaustive and multiple vacation policy, this kind of vacation models have been studied by Lucantoni et al.

[17]. They got vector versions of decomposition formulas for the queue length and for the virtual waiting time by using matrix geometric analysis. We here consider a vacation model with the same arrival process but for a general vacation policy. Of course, it is difficult to give detailed solutions as given in [17], but we can also find a kind of vector decomposition formulas.

We first introduce the Markovian arrival process. Let $\{J(t)\}$ be a Markov process with a finite state space $K = \{1, 2, \ldots, m\}$, with a transition probability matrix $\boldsymbol{P} = \{p_{i,j}\}$ at jump instants and with exponentially distributed sojourn times which have mean $\frac{1}{\mu_j}$ for $J(t) = j$. Here, the sojourn times only depend on the present states. We assume that $\boldsymbol{P}$ is irreducible, which implies the existence of a stationary distribution for $J(t)$. So far, $\{J(t)\}$ is assumed to be stationary under P. We split $p_{i,j}$ into two components $q_{i,j}$ and $r_{i,j}$ so that $p_{i,j} = q_{i,j} + r_{i,j}$, and a customer arrives with probability $\frac{r_{i,j}}{p_{i,j}}$ when the transition from i to j takes place, where $q_{i,i} = 0$ for all i. This arrival process is exactly same as MAP of Lucantoni et al. [17] although the descriptions are a little different, for example, we have not used an absorbing state. Let τ_n be the n-th arrival time of a customer and S_n his service time. We assume that S_n are $i.i.d.$ with distribution function G and independent of the arrival process. Thus the input process is given by $\Psi_a = \{(\tau_n, S_n)\}$. We call this input process a Markovian input process. Let τ_n^0 be the n-th jump epoch of $J(t)$, and $S_n^0 = S_m$ if $\tau_n^0 = \tau_m$ for some m and $S_n^0 = 0$ otherwise. We call $\Psi_0 \equiv \{(\tau_n^0, S_n^0)\}$ a potential input process. In the following, we are concerned with this potential input process instead of Ψ_a. Thus, we define the vacation model introduced in Section 2 for the marked point process Ψ_0. We here assume (iii) and that $\{(\alpha(t), W(t), J(t))\}$ and Ψ_0 are jointly stationary. Furthermore, we assume that the process $\{(W^{(nv)}(t), J^{(nv)}(t))\}$ has a stationary distribution for the corresponding queueing model without server vacations.

Similarly as in Section 2, let N_0 be a point process generated by $\{\tau_n^0\}$, and denote its intensity by λ_0 and a Palm distribution of P concerning N_0 by P_0. Since the inter-jump times of $J(t)$ are exponentially distributed, the inversion formula of stationary point processes implies

$$\begin{aligned} E(e^{-\theta W(0)}; J(0) = j) &= \lambda_0 E_0(\int_0^{\tau_1^0} e^{-\theta W(s)} 1_{\{J(s)=j\}} ds) \\ &= \frac{\lambda_0}{\mu_j} E_0(e^{-\theta W(0-)}; J(0-) = j) \ . \end{aligned} \tag{6.1}$$

Thus, similarly to (2.3), the rate conservation law of Miyazawa [18] yields

$$\theta E(e^{-\theta W(0)}; \alpha(0) = 1, J(0) = j) = \mu_j E(e^{-\theta W(0)}; J(0) = j) - \sum_{i \in K} \mu_i E(e^{-\theta W(0)}; J(0) = i)(q_{i,j} + r_{i.j}\hat{G}(\theta)) \ . \quad (6.2)$$

Define $\hat{W}_j(\theta) = E(e^{-\theta W(0)}; J(0) = j)$ and $\hat{Y}_j(\theta) = E(e^{-\theta W(0)}; \alpha(0) = 0, J(0) = j)$ for $j \in K$. Then (6.2) is equivalent to

$$\theta \hat{W}_j(\theta) = \theta \hat{Y}_j(\theta) + \mu_j \hat{W}_j(\theta) - \sum_{i \in K} \mu_i \hat{W}_i(\theta)(q_{i,j} + r_{i.j}\hat{G}(\theta)) \ . \quad (6.3)$$

Define row vectors $\hat{\boldsymbol{Y}}(\theta) = (\hat{Y}_1(\theta), \cdots, \hat{Y}_m(\theta))$ and $\hat{\boldsymbol{W}}(\theta) = (\hat{W}_1(\theta), \cdots, \hat{W}_m(\theta))$, and matrixes Q and R by letting their components $Q_{i,j} = \mu_i q_{i,j}$ for $i \neq j$, $Q_{i,i} = -\mu_i$ and $R_{i,j} = \mu_i r_{i,j}$ for all (i, j). Then, from (6.3), we get

$$\hat{\boldsymbol{W}}(\theta) \left(\theta I + Q + \hat{G}(\theta)R\right) = \theta \hat{\boldsymbol{Y}}(\theta) \ ,$$

and hence

$$\hat{\boldsymbol{W}}(\theta) = \theta \hat{\boldsymbol{Y}}(\theta) \left(\theta I + Q + \hat{G}(\theta)R\right)^{-1} \ , \quad (6.4)$$

where $\left(\theta I + Q + \hat{G}(\theta)R\right)^{-1}$ exists for $\Re(\theta) > 0$ except for finitely many singular points because the left-hand side of (6.4) is analytic for $\Re(\theta) > 0$ (see [11]).

Remark 6.1 Let $\pi_j = P(J(0) = j)$. By letting $\theta = 0$ in (6.2), we can see that $\{\pi_i\}$ is a stationary distribution of the continuous time Markov chain with the infinitesimal generator $Q + R$. On the other hand, $\lambda_0 = \sum_{i=1}^m \mu_i \pi_i$ by (6.1), and the arrival rate $\lambda = \sum_{i=1}^m \sum_{j=1} m \mu_i \pi_i r_{i,j}$. Thus, we have the traffic intensity $\rho = E_0(S_0) \sum_{i=1}^m \sum_{j=1} m \mu_i \pi_i r_{i,j}$.

We next consider the case of no server vacations but having the same input process. In this case, we distinguish notations from the case of the vacation model by putting a superscript (nv) such as $W_j^{(nv)}$. Then, it is known that

$$\hat{\boldsymbol{W}}^{(nv)}(\theta) = \theta \boldsymbol{w}^{(nv)} \left(\theta I + Q + \hat{G}(\theta)R\right)^{-1} \ , \quad (6.5)$$

where $\boldsymbol{w}^{(nv)} = (w_1^{(nv)}, \cdots, w_m^{(nv)})$ for $w_j^{(nv)} = P(W^{(nv)}(0) = 0, J(0) = j)$ (see Eqn. (37) of [17]). We note that Eqn. (6.5) is also obtained by a similar argument as used above to get (6.4). That is, similarly to (6.3), we have

$$\theta \hat{W}_j^{(nv)}(\theta) = \theta w_j^{(nv)} + \mu_j \hat{W}_j^{(nv)}(\theta) - \sum_{i \in K} \mu_i \hat{W}_i^{(nv)}(\theta)(q_{i,j} + r_{i.j}\hat{G}(\theta)) \ , \quad (6.6)$$

which implies (6.5).

Comparing (6.4) with (6.5), we see that Eqn. (6.4) is a kind of decomposition expression, which generalizes the Poisson arrival case and Theorem 11 of [17] for the exhaustive and multiple vacation policy.

Remark 6.2 From Theorems 2 and 10 of [17], we can see $\hat{\boldsymbol{Y}}(0) \neq \boldsymbol{w}^{(nv)}$ for the exhaustive and multiple vacation policy. In general, $\hat{\boldsymbol{Y}}(0)$ may depend on the vacation policy. For example, the server vacations may start only when $J(t) = j$ for a fixed j and the vacation lengths are $i.i.d.$. Thus, we need specific conditions for the vacation policy to determine $\hat{\boldsymbol{Y}}(0)$.

We finally note a decomposition for the waiting time at arrival instants, which means the virtual workload observed by arriving customers. Denote its LST by $\hat{W}_{0,j}(\theta)$, which equals $E_0(e^{-\theta W(0-)}; J(0-) = j)$. Let $\hat{\boldsymbol{W}}_0(\theta) = (\hat{W}_{0,1}(\theta), \cdots, \hat{W}_{0,m}(\theta))$. Then, Eqn. (6.1) implies

$$\hat{\boldsymbol{W}}_0(\theta) = \frac{\theta}{\lambda_0} \hat{\boldsymbol{Y}}(\theta) \left(\theta I + Q + \hat{G}(\theta) R\right)^{-1} R \,, \tag{6.7}$$

where $\hat{\boldsymbol{Y}}_0(\theta)$ is a row vector with the i-th component $E_0(e^{-\theta W(0-)}; \alpha(0) = 0, J(0) = i)$. Similarly, we have

$$\hat{\boldsymbol{W}}_0^{(nv)}(\theta) = \frac{\theta}{\lambda_0} \boldsymbol{w}_0^{(nv)} \left(\theta I + Q + \hat{G}(\theta) R\right)^{-1} R \,, \tag{6.8}$$

where $\boldsymbol{w}^{(nv)}$ is a row vector with the i-th component $P_0(W^{(nv)}(0-) = 0, J(0-) = j)$. Thus we get a decomposition expression for the waiting time at arrival instants, too.

Acknowledgements

I am grateful to Professor Ronald W. Wolff for his valuable comments and to the anonymous referees for their helpful comments.

References

[1] Bardhan, I. and Sigman, K., Rate conservation law for stationary semi-Martingales, Probability in Informational Engineering and Sciences (1993).

[2] Boxma, O. J. and Groenendijk, W. P., Pseudo-conservation laws in cyclic-service systems, J. Appl. Prob. 24 (1987), 949-964.

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

[4] Bremaud, P., An elementary proof of Sengupta's invariance relation and a remark on Miyazawa's conservation principle. J. Appl. Prob. 28 (1991), 950-954.

[5] Doshi, B. T., Note on stochastic decomposition in a $GI/G/1$ queue with vacations or set-up times, J. Appl. Prob. 22 (1985), 419-428.

[6] Doshi, B. T., Queueing systems with vacations - A survey, Queueing systems 1 (1986) 29-66.

[7] Doshi, B. T., Conditional and unconditional distributions for $M/G/1$ type queues with server vacations. Queueing Systems 7 (1990), 229-252.

[8] Doshi, B. T., Generalization of the stochastic decomposition for single server queues with vacations, Stochastic Models 6 (1990), 307-333.

[9] Doshi, B. T., Level-crossing analysis of queues, in *Queueing and Related Models* edited by U. N. Bhat and I. V. Basawa, Oxford University Press (1992), 3-33.

[10] Franken, P., K嗜ig, D., Arndt, U. and Schmidt, V., *Queues and Point Processes*, Wiley, Chichester (1982).

[11] Heffes, H. and Lucantoni, D. M., A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. IEEE J. Sel. Areas Comm., Special Issues on Network Performance Evaluation 4 (1986), 856-868.

[12] Keilson, J. and Servi, L. Oscillating random walk models for $GI/G/1$ vacation systems with Bernoulli schedules, J. Appl. Prob. 23 (1986), 790-802.

[13] Kella, O. and Whitt, W., Queues with server vacations and Levy processes with secondary jump input. Ann. of Appl. Prob. 1 (1991), 104-117.

[14] Kella, O. and Whitt, W., Useful martingales for stochastic storage processes with Levy input. To appear in J. Appl. Prob. (1992).

[15] König, D. and Schmidt, V., Extended and conditional versions of the PASTA property, Adv. Appl. Prob. 22 (1990), 510-512.

[16] Leung, K. K., On the additional delay in an $M/G/1$ queue with generalized vacations sand exhaustive service, Opns. Res. 40 (1992), s272-s283.

[17] Lucantoni, D. M., Meier-Hellstern, K. S. and Neuts, M. F., A single server queue with server vacations and a class of non-renewal arrival processes, Adv. Appl. Prob. 22 (1990), 676-705.

[18] Miyazawa, M., The derivation of invariance relations in complex queueing systems with stationary inputs, Adv. Appl. Prob. 15 (1983), 874-885.

[19] Miyazawa, M., The intensity conservation law for queues with randomly changed service rate, J. Appl. Prob. 22 (1985), 408-418.

[20] Miyazawa, M., The characterization of the stationary distributions of the supplemented Self-clocking Jump Process, Math. of OR 16 (1991), 547-565.

[21] Miyazawa, M., Palm distribution and time-dependent RCL on stationary increment scheme and their applications, Res. Rep. of Science University of Tokyo (1992).

[22] Miyazawa, M., Rate conservation laws: a survey, to appear in Queueing Systems (1993).

[23] Shanthikumar, J. G., On stochastic decomposition in $M/G/1$ type queues with generalized server vacations, Oper. Res. 36 (1988), 566-569.

[24] Wolff, R. W., *Stochastic Modeling and The Theory of Queues.* Prentice Hall, New Jersey (1989).