Journal of Applied Mathematics and Stochastic Analysis
9, Number 4, 1996, 551-562.

ON APPLICATIONS OF EXCESS LEVEL PROCESSES TO (N,D)-POLICY BULK QUEUEING SYSTEMS

JEWGENI H. DSHALALOW
Department of Applied Mathematics
Florida Institute of Technology
Melbourne, FL 32901, U.S.A.

(Received July, 1996; Revised November, 1996)

ABSTRACT

The paper deals with queueing systems in which N- and D-policies are combined into one. This means that an idle or vacationing server will resume his service if the queueing or workload process crosses some specified fixed level N or D, respectively. For the proposed (N,D)-policy we study the queueing processes in models with and without server vacations, with compound Poisson input, and with generally distributed service and vacation periods. The analysis of the models is essentially based on fluctuation techniques for two-dimensional marked counting processes newly developed by the author. The results enable us to arrive at stationary distributions for the embedded and continuous time parameter queueing processes in closed analytic forms, enhancing the well-known Kendall formulas and their modifications.

This article is dedicated to the memory of Roland L. Dobrushin.

Key words: Queueing, Queueing Process, Vacations, N-Policy, D-Policy, (N,D)-Policy, Kendall's Formula, Fluctuation Theory, First Excess Level Theory, Semi-Regenerative Techniques, Embedded Markov Chain.

AMS (MOS) subject classifications: 60K10, 60K15, 60K25.

1. Introduction

D-policy systems form an analytically attractive and practically useful class of queues. However, very few works on this topic appear in the literature, compared to an abundance of its close relatives: *quorum*, *N-policy*, and *hysteresis* (cf. [2,3,7,9,15-28,30-45]). D-policy queues belong to workload dependent systems according to the classification of state-dependency in Dshalalow [13].

The D-policy determines when to end an idle period and begin the following busy period, which starts when the cumulative workload (i.e. a period of time needed to process all available customers) crosses level D. [Note that in this case, a (single) server should take a fixed number of customers in a batch, preferably one, or else the workload would be hard to define.] One of the main reasons for employing D-policy is to minimize server switch-overs, usually followed by start-ups. Of course, this protracts idle or vacation periods, thereby increasing the unfinished work or, equivalently, customers' sojourn times. The latter, is balanced out by the reduction of switch-overs, and the "best" value of D can become a subject of a relevant optimization problem. There have been just a few articles [4-6, 8, 29, 38, 42, 43] on D-policy, perhaps because of the analytical complexity of the subject, and even they target either the waiting time process or some special optimization problems related to the waiting time process rather than the queueing process.

The N- (or *threshold*) policy is another rule specifying the beginning of a busy period. However, the N-policy is tied to the queueing process. Namely, once the server enters an idle period, the policy specifies how many customers (N, the threshold level) should be accumulated in the queue before the server turns on. As in the case of the D-policy, N-policy is designed to minimize server switch-overs. Another common name for N-policy is *removable server*, called so by Yadin and Naor [45] (who were the first to introduce N-policy systems in 1963) and their followers [30, 31, 43, 44].

N-policy is most often combined with vacations. Once the system is exhausted, the server goes on a single vacation and upon his return he checks on the queue, and if the queue has accumulated to at least N customers he begins service; otherwise, he rests and starts servicing as soon as the queue reaches level N. This is referred to as a *single vacation* discipline. In the case of multiple vacations, when the system is exhausted, the server initiates a sequence of vacation segments. This sequence (not any particular segment) is interrupted as soon as the system accumulates to at least N customers. Systems with N-policy and vacations were studied in papers [2, 9, 16, 23, 24, 28, 35, 38, 40, 43].

One of the main difficulties when analyzing queueing processes in D-policy systems is to determine the value of the queue by the beginning of a busy period. A similar difficulty becomes apparent in N-policy systems whenever the input stream is bulk. However, some special results on fluctuation theory obtained in Abolnikov and Dshalalow [1] and further developed in Dshalalow [10-12] enabled Dshalalow and Yellen [14] and Muh [33, 34] to treat bulk N-policy models with and without server vacations.

It is an obvious observation that in spite of their similarities, N- and D-policies carry different advantages and deficiencies with respect to each other. For instance, the system may become alerted by as many as N customers in the system to resume its service or interrupt vacations, while in some situation it would not be necessary, as a large number of customers in the buffer need not be a sign of a long workload for the server, as many of them may require short processing times. However, this is not taken into account, since N-policy does not care about this nuance. On the other hand, D-policy is blind to an abundance of customers which can overstock the buffer while yielding just a reasonable accumulated workload. Therefore, both policies are preferred to each other depending on system configurations and cost factors. So, if the waiting time is too costly but the penalties for occupied seats in the waiting room are reasonable, the N-policy would be a better option. What if all factors need to be taken into consideration?

While several works [6, 8, 38] compared N- and D-policies for special cases and concluded that under various conditions one or the other is superior, the above arguments make us believe that a combination of the two policies is superior to either of them. In other words, a policy, referred to as *(N,D)-policy*, can be specified as a rule to terminate every idle period when the queueing or workload processes cross their respective (fixed) values N or D for the first time, whichever comes first. On top of this, the system will include bulk arrival and, optionally, server vacations, which is itself referred to as *T-policy*. To analyze such a system, in particular the queueing process, relevant information regarding the queue length at the beginning of each busy period is crucial. Techniques applied to these and similar situations were recently developed in Dshalalow [11, 12]. They will be discussed in Section 2 and adopted to special queues in Sections 3 and 6. The background model for the investigation is $M^X/G/1$ (with bulk input) modified for (N,D)-policy and with or without server vacations. This is rendered in Sections 4,5 and 6. The queueing process is treated both with continuous time parameter and at departure epochs. All formulas are computationally tractable and presented in closed analytical forms.

2. Background Material

In this section we will present an overview of first excess level techniques applied to a class of delayed renewal processes marked by a vector renewal process.

All processes and random variables will be considered on a probability space $(\Omega, \mathfrak{F}, P)$. Let

$$\tau = \sum\nolimits_{n \geq 0} \varepsilon_{\tau_n}$$

be a delayed renewal process on $\mathbb{R}_+$, where ε_a is a point mass, and let

$$(X,Y) = \{(X_n, Y_n);\ n = 0,1,\ldots\}$$

be a sequence of independent and, for $n \geq 1$, identically distributed random vectors, such that components X_n are discrete-valued in $\{0,1,\ldots\}$ and Y_n are continuous-valued in $\mathbb{R}_+$. (However, for each n, X_n and Y_n need not be independent.) We construct with τ and (X,Y) the marked renewal process

$$Z = \sum\nolimits_{n \geq 0} (X_n, Y_n) \varepsilon_{\tau_n}, \tag{2.1}$$

such that Z is obtained from τ by position independent marking. Consequently,

$$(A,B) = \{(A_n = \sum\nolimits_{k=0}^{n} X_k,\ B_n = \sum\nolimits_{k=0}^{n} Y_k);\ n = 0,1,\ldots\} \tag{2.2}$$

is a two-dimensional delayed renewal process.

Process Z will be described in terms of the following transformations:

$$\gamma_0(z,\vartheta) = \mathbb{E}[z^{X_0} e^{-\vartheta Y_0}],\ \gamma(z,\vartheta) = \mathbb{E}[z^{X_1} e^{-\vartheta Y_1}],\ |z| \leq 1,\ Re(\vartheta) \geq 0, \tag{2.3}$$

$$\alpha_0(z) = \gamma_0(z,0),\ \alpha(z) = \gamma(z,0), \tag{2.4}$$

$$h_0(\theta) = \mathbb{E}[e^{-\theta \tau_0}],\ h(\theta) = \mathbb{E}[e^{-\theta(\tau_1 - \tau_0)}],\ Re(\theta) \geq 0, \tag{2.5}$$

with no restrictions imposed on γ_0, γ, α_0, α, h_0, and h.

For two fixed nonnegative real numbers p_1 and p_2, let $L = (p_1, p_2)$. We say (A,B) *hits threshold* L if, for some $n = 0,1,\ldots$, A_n or B_n exceeds p_1 or p_2, respectively. More formally, we call the integer-valued random variable

$$\nu = \min\{\inf\{j\colon A_j > p_1\},\ \inf\{k\colon B_k > p_2\}\} \tag{2.6}$$

the *termination index*, so that (A,B) hits L at τ_ν, called the *first passage time* of (A,B). The value (A_ν, B_ν) of (A,B) at τ_ν will be called the *first excess of threshold* L or just the *first excess level.*

Now, we introduce the operators

$$D_p^d f(z) = (1-z) \sum\nolimits_{p \geq 0} z^p f(p),\ |z| < 1, \tag{2.7}$$

$$D_p^c f(s) = s \int_{p=0}^{\infty} e^{-sp} f(p) dp,\ Re(s) > 0, \tag{2.8}$$

and their combination

$$D_{p_1,p_2} g(x,y) = D_{p_1}^c \{D_{p_2}^d g(p_1,p_2)(x)\}(y), \tag{2.9}$$

where f and g are relevant integrable Baire functions.

With Taylor-like functionals

$$\mathfrak{D}_x^k(\cdot) = lim_{x \to 0} \frac{1}{k!} \frac{\partial^k}{\partial x^k} \frac{1}{1-x}(\cdot) \tag{2.10}$$

we can restore f subjected to D_p^d:

$$\mathfrak{D}_x^k D_p^d f(x) = f(k) \tag{2.11}$$

and with

$$\mathcal{L}_D(\cdot) = Lapl^{-1}(\tfrac{1}{s}(\cdot))(D), \text{ for some real number } D \geq 0, \tag{2.12}$$

where $Lapl^{-1}$ stands for the inverse of Laplace transform, we can find f transformed by D_p^c. For various special cases throughout the paper, we notice a few elementary properties of the operator $\mathfrak{D}_x^k$:

(a) $\mathfrak{D}_x^k$ is a linear operator with fixed points at every constant function.

(b) For any function ϕ, analytic at zero,

$$\mathfrak{D}_x^k(x^j\phi(x)) = \begin{cases} 0, & k < j \\ \mathfrak{D}_x^{k-j}(\phi(x)), & k \geq j. \end{cases} \tag{2.13}$$

Now, given threshold $L = (p_1, p_2)$, we introduce the transformations

$$\mathfrak{F}(\xi,\theta,z,\vartheta;L) = \mathbb{E}[\xi^{\nu}e^{-\theta\tau_{\nu}}z^{A_{\nu}}e^{-\vartheta B_{\nu}}] \tag{2.14}$$

and

$$\mathfrak{F}^*(\xi,\theta,z,\vartheta;x,s) = (D_{p_1,p_2}\mathfrak{F})(x,s). \tag{2.15}$$

The below result is due to Dshalalow [12].

Theorem 1.

$$\frac{1}{h_0(\theta)}\mathfrak{F}^*(\xi,\theta,z,\vartheta;x,s) = \gamma_0(z,\vartheta) - \gamma_0(xz,\vartheta+s)\frac{1-\xi h(\theta)\gamma(z,\vartheta)}{1-\xi h(\theta)\gamma(xz,\vartheta+s)}. \tag{2.16}$$

With $X_0 = i$ *a.s.*, the Laplace-Stieltjes transform of Y_0 denoted by $\beta_0(\vartheta)$, and under the assumptions $h_0(\theta) = 1$ and $\vartheta = 0$, (2.16) reduces to

$$\mathfrak{F}^*(\xi,\theta,z,0;x,s) = z^i - (xz)^i\beta_0(s)\frac{1-\xi h(\theta)\alpha(z)}{1-\xi h(\theta)\gamma(xz,s)}. \tag{2.17}$$

We will be interested in the transform of the following marginal processes: termination index, first excess level, and first passage time, all derived from (2.17) for appropriate values of the variables ξ, θ, and z, and subsequent use of the inverse transforms. For the upcoming applications, we are going to use the pair $(N-1, D)$ for L and abbreviate $\mathbb{E}[\cdot \mid X_0 = i]$ by $\mathbb{E}^i[\cdot]$.

Termination index

By using properties (a, b) (equation (2.13)) of operator $\mathfrak{D}_x^k$, we obtain:

$$T_L^{(i)}(\xi) = \mathbb{E}^i[\xi^{\nu}] = 1 - (1-\xi)\mathcal{L}_D\mathfrak{D}_x^{N-i-1}\left\{\frac{\beta_0(s)}{1-\xi\gamma(x,s)}\right\}. \tag{2.18}$$

First excess level

$$E_L^{(i)}(z) = \mathbb{E}^i[z^{A_{\nu}}] = z^i - z^i[1-\alpha(z)]\mathcal{L}_D\mathfrak{D}_x^{N-i-1}\left\{\frac{\beta_0(s)}{1-\gamma(xz,s)}\right\}. \tag{2.19}$$

Marginal first passage time

$$P_L^{(i)}(\theta) = \mathbb{E}^i[e^{-\theta\tau_{\nu}}] = 1 - [1-h(\theta)]\mathcal{L}_D\mathfrak{D}_x^{N-i-1}\left\{\frac{\beta_0(s)}{1-h(\theta)\gamma(x,s)}\right\}. \tag{2.20}$$

3. Model with a Dormant Server. Preliminaries

We will start with the following model based on the $M/G/1$ queue. The input to the system is compound Poisson. A single server processes customers one at a time and there is no restriction to their service time distribution. The server becomes idle when the system is exhausted. The idle period lasts until the queue accumulates to N (≥ 1) or more customers or the cumulative service time of all arrived customers will exceed a fixed number D (≥ 0) whichever of the two events comes first. Then, a busy period begins. As it was mentioned in the introduction, while each of the policies is aimed to reduce the number of switch-overs between idle and busy modes (usually followed by start-up periods), its combination, i.e., the (N,D)-policy, limits the "first excess level" from an excessive (accumulated) workload and buffer overstock simultaneously.

We will formalize the system as follows.

Service. Let $T_0 = 0$, T_1, ... be the successive instants of service completions. The service of the nth customer, rendered within the time interval $(T_{n-1}, T_n]$, which is referred to as the nth *service cycle*, lasts σ_n ($\leq T_n - T_{n-1}$) with the PDF $B(t)$, finite first moment b, and Laplace-Stieltjes transform $\beta(\theta) = \mathbb{E}[e^{-\theta\tau_0}]$, common for all $n = 1, 2, \ldots$. If $Q(t)$ is the number of all customers in the system present at time t, including one in service, then let $Q_n = Q(T_n +)$.

Input. The input to the system will be associated with the marginal marked point process $\Pi = \sum_{n \geq 0} X_n \varepsilon_{\tau_n}$ and for convenience specified on the first service cycle. We set $X_0 = Q_0$ and $\tau_0 = 0$ *a.s.* and assign X_1, X_2, ... to the successive batches arriving at the system at τ_1, τ_2, ..., i.e., X_k is the kth increment of the arrival process over the interval $(\tau_{k-1}, \tau_k]$. With the assumption that Π is (delayed) compound Poisson, we have $h(\theta) = \lambda/(\lambda + \theta)$, while the probability generating function (pgf), $\alpha(z)$ of X_1 is arbitrary, with finite mean a. The term "nth customer" mentioned above is not related to a rigid order. Customers are lined up in order of arriving batches, but within every batch the order is arbitrary. $A_k = X_0 + \ldots + X_k$ is the total number of customers arrived at the system by time τ_k.

Busy period discipline. Again, for convenience we specify the processes on the first service cycle. Let Y_k be the cumulative job brought by the kth batch of customers. Hence, the arriving jobs form the marginal marked point process $J = \sum_{n \geq 0} Y_n \varepsilon_{\tau_n}$ with Y_0 being the initial cumulative job due the presence of X_0 customers, and $B_k = Y_0 + \ldots + Y_k$ is the cumulative job accumulated by time τ_k. If $Q_0 = X_0$ is zero, then so is Y_0, and the system turns the server off until the buffer fills up with at least N units or the cumulative job crosses level D, whichever comes first; that is, the busy period begins upon (A,B) hitting threshold $L = (N-1,D)$.

The transformation $\gamma(z,\vartheta) = \mathbb{E}[z^{X_1} e^{-\vartheta Y_1}]$ reduces to

$$\gamma(z,\vartheta) = \alpha(z\beta(\vartheta)) \tag{3.1}$$

with the use of the conditional expectation

$$\gamma(z,\vartheta) = \mathbb{E}[\mathbb{E}[z^{X_1} e^{-\vartheta(\sigma_1 + \ldots + \sigma_{X_1})} \mid X_1]] \tag{3.2}$$

and straightforward probability arguments. For arbitrary i, $\beta_0(s)$ in (2.17) is $[\beta(s)]^i$. Since we consider an exhaustive system, the relevant value of i in formula (2.17) will be zero and because of (3.1) and (2.17) we have the modified expression for the joint transformation of the termination index, first excess level, and first passage time:

$$\mathbb{E}^0[\xi^{\nu} e^{-\theta\tau_{\nu}} z^{A_{\nu}}] = 1 - \left[1 - \xi\frac{\lambda}{\lambda+\theta}\alpha(z)\right]\mathcal{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1 - \xi\frac{\lambda}{\lambda+\theta}\alpha(xz\beta(s))}\right\}. \tag{3.3}$$

Equation (3.3) will therefore yield the following analogs of (2.18-2.20).

Termination index

$$T_L^{(0)}(\xi) = 1 - (1-\xi)\mathcal{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1 - \xi\alpha(x\beta(s))}\right\}. \tag{3.4}$$

$$\overline{T}_L^{(0)} = \mathbb{E}^0[\nu] = \mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\alpha(x\beta(s))}\right\}. \tag{3.5}$$

Marginal first excess level

$$E_L^{(0)}(z) = 1 - [1-\alpha(z)]\mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\alpha(xz\beta(s))}\right\}. \tag{3.6}$$

$$\overline{E}_L^{(0)} = \mathbb{E}^0[A_\nu] = a\overline{T}_L^{(0)}. \tag{3.7}$$

First passage time

$$P_L^{(0)}(\theta) = 1 - \left(1 - \frac{\lambda}{\lambda+\theta}\right)\mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\frac{\lambda}{\lambda+\theta}\alpha(x\beta(s))}\right\}. \tag{3.8}$$

$$\overline{P}_L^{(0)} = \mathbb{E}^0[\tau_\nu] = \tfrac{1}{\lambda}\overline{T}_L^{(0)}. \tag{3.9}$$

4. Embedded Process

The above transformations will be used to derive a formula for the pgf of the queueing process embedded in $Q(t)$ over the sequence $\{T_n\}$ in the steady state.

The queueing process $Q(t)$ is obviously semi-regenerative relative to the point process $\{T_n\}$. Consequently, $\{Q_n, T_n\}$ is a Markov renewal process and $\{Q_n\}$ is an embedded Markov chain with transition probability matrix (tpm) (p_{ij}). This is a Δ_2-matrix which differs from one for the $M^X/G/1$ queue by the first row only. This can be seen from the transition

$$Q_1 = \begin{cases} Q_0 + \mathcal{A}_1 - 1, & Q_0 > 0 \\ A_\nu + \mathcal{A}_1 - 1, & Q_0 = 0, \end{cases} \tag{4.1}$$

where $\mathcal{A}_n$ is the number of customers that arrive during the nth service period. [For $M^X/G/1$ system, A_ν needs to be replaced by X_1, the size of the first arriving batch.] Consequently,

$$\rho < 1, \text{ where } \rho = ab\lambda, \tag{4.2}$$

is the necessary and sufficient condition for the steady state, which we assume to be met throughout the paper.

The invariant probability measure $p = (p_0, p_1, \ldots)$ of $\{Q_n\}$ can be sought in the form of its pgf $P(z)$, which satisfies the equation

$$P(z) = \textstyle\sum_{i=0}^{\infty} p_i P_i(z), \ |z| \leq 1, \tag{4.3}$$

along with the boundary condition

$$P(1) = 1, \tag{4.4}$$

where $P_i(z) = \mathbb{E}^i[z^{Q_1}]$ is the pgf of the ith row of the tpm (p_{ij}). With

$$\mathbb{E}^i[z^{\mathcal{A}_1}] = \beta(\lambda(1-\alpha(z))), \text{ in notation, } \mathrm{b}(z), \tag{4.5}$$

and (4.1) we have

$$P_i(z) = \begin{cases} z^{i-1}\mathrm{b}(z), & i > 0 \\ z^{-1}E_L^{(0)}(z)\mathrm{b}(z), & i = 0. \end{cases} \tag{4.6}$$

By (4.3) and (4.6) we easily arrive at the Kendall-like formula

$$P(z) = p_0 \mathrm{b}(z)\frac{1 - E_L^{(0)}(z)}{\mathrm{b}(z) - z}. \tag{4.7}$$

p_0 is determined from boundary condition (4.4) applied to (4.7) and equation (3.7), yielding

$$p_0 = \frac{1-\rho}{a\overline{T}_L^{(0)}}. \tag{4.8}$$

The mean stationary service cycle is defined as

$$c = PC, \text{ where } C = (C_i = \mathbb{E}^i[T_1];\ i = 0, 1, \ldots)^{\mathrm{T}}, \tag{4.9}$$

satisfying

$$C_i = \begin{cases} \mathbb{E}^0[\tau_\nu] + b = \frac{1}{\lambda}\overline{T}_L^{(0)} + b, & i = 0 \\ b, & i > 0. \end{cases} \tag{4.10}$$

(See (3.9).) Now, (4.8-4.10) yield

$$c = \frac{1}{a\lambda}, \tag{4.11}$$

which is the same value as that for the $M^X/G/1$ system, a quite surprising result.

The Kendall-like formula for the conventional $M^X/G/1$ model without (N,D)-policy follows from (4.7) and (4.8) by taking $N = 1$ in $L = (N-1,D)$, thereby reducing $E_L^{(0)}(z)$ to $\alpha(z)$ and $\overline{T}_L^{(0)}$ to 1.

5. Continuous Time Parameter Queueing Process

The analysis of the continuous time parameter process is based on semi-regenerative techniques, which give much quicker results than the more popular method of supplementary variables provided the invariant probability measure of the embedded process is known. It starts with the evaluation of transition probabilities

$$K_{ik}(t) = \mathbb{P}^i\{Q(t) = k,\ T_1 > t\} \tag{5.1}$$

on the first service cycle, where $\mathbb{P}^i\{\cdot\} = \mathbb{P}\{\cdot \mid Q_0 = i\}$, and the formation of matrix $K(t) = (K_{ik}(t))$, which is called the *semi-regenerative kernel.*

The stationary probability measure $\pi = (\pi_0, \pi_1, \ldots)$ exists given the ergodicity condition $\rho < 1$, regardless of the initial state of the system, and is conveniently sought in the form of its pgf $\pi(z)$:

$$\pi(z) = \tfrac{1}{c}ph(z), \tag{5.2}$$

where c is the stationary mean value of the service cycle and $ph(z)$ is the scalar product of the invariant probability measure p of the embedded process (derived in the previous section) and vector $h(z) = (h_i(z); i = 0, 1, \ldots)^{\mathrm{T}}$ of the pgf's of the respective rows of the integrated (over $\mathbb{R}_+$) semi-regenerative kernel $K(t)$. (See cf. Dshalalow [13] for a pertinent reference.)

Let

$$\Phi_j(t) = \mathbb{P}\{A_\nu = j, \tau_\nu \le t\} \tag{5.3}$$

be the joint PDF of the first excess level and first passage time of process (A,B) relative to the threshold L, and let

$$\delta_j(t) = \mathbb{P}^0\Big\{ \textstyle\sum_{n \geq 0} X_n \varepsilon_{\tau_n}[0,t] = j\Big\}, \tag{5.4}$$

i.e., the probability that j customers arrive in interval $[0,t]$. Then, by simple probability arguments,

$$K_{0k}(t) = \textstyle\sum_{j=1}^{k} (\delta_{k-j}(\cdot)[1 - B(\cdot)]) * \Phi_j(t), \; k = 1, 2, \ldots, \tag{5.5}$$

where $*$ is the convolution operator,

$$K_{00}(t) = e^{-\lambda t}, \tag{5.6}$$

and for $i > 0$,

$$K_{ik}(t) = \begin{cases} \delta_{k-i}(t)[1 - B(t)], & k \geq i \\ 0, & k < i. \end{cases} \tag{5.7}$$

Now,

$$h_i(z) = \textstyle\sum_{k=0}^{\infty} z^k \int_0^{\infty} K_{ik}(t)dt \tag{5.8}$$

yields

$$h_0(z) = \tfrac{1}{\lambda} + E_L^{(0)}(z)\Delta(z), \tag{5.9}$$

$$h_i(z) = z^i \Delta(z), \tag{5.10}$$

where

$$\Delta(z) = \frac{1 - \mathrm{b}(z)}{\lambda[1 - a(z)]}. \tag{5.11}$$

Finally, by (5.2) and with (5.9-5.11), we have

$$\pi(z) = a p_0 + \frac{az}{\mathrm{b}(z)} \Delta(z) P(z), \tag{5.12}$$

where $P(z)$ and p_0 satisfy formulas (4.7) and (4.8).

The above results can be summarized as the following theorem.

Theorem 2. *The stationary probability distribution π of the queueing process $Q(t)$ exists given the condition $\rho < 1$ and satisfies formulas* (5.12), (4.7), *and* (4.8).

For $N = 1$, (5.12) reduces to the well-known result for the $M^X/G/1$ system:

$$\pi(z) = \frac{a(1 - z)}{1 - \alpha(z)} P(z). \tag{5.13}$$

6. Model with Multiple Vacations

In this section we consider a variant of the preceding model, in which the exhausted system lets the server go on vacation. We will adopt the vacation rule known in the literature as *multiple vacations.* This is specified as follows. Once the last customer leaves the system, the server leaves the system too. [He may be assigned for a maintenance or other duties.] The whole vacation trip, called the *vacation period*, consists of multiple segments. At the end of each vacation segment, the server returns to the system and checks on whether the queue is filled up to the desired level; if it does not, the server leaves again and keeps on going until the condition specified for the busy period policy is met. Then the service is resumed. Note that none of the individual vacation segments is interrupted even if the queue hits a specified level.

We assume that all vacation segments are independent of each other and stochastically equivalent with a common distribution $V(t)$, different from the service time distribution, the Laplace-Stieltjes transform $\varphi(\theta)$, and finite mean v. We will show how the first excess level theory can be adopted to this type of model by means of a few minor modifications. We will preserve all other assumptions made regarding the input and service.

First, idle periods are replaced by vacation periods. A service cycle starts with a vacation period if the preceding departure exhausts the system. (Again, for convenience, the situation will be applied to the first service cycle.) Then the server leaves the system and returns to the system periodically in accordance with the renewal process $r_1, r_2, \ldots$. The nth interrenewal period corresponds to the nth vacation segment, during which time the input delivers X_n^* customers with

$$\mathbb{E}[z^{X_n^*}] = \varphi[\lambda(1-\alpha(z)]. \tag{6.1}$$

The value Y_n^* will represent the nth increment of the workload brought by X_n^* customers during the nth vacation segment. Analogously to formula (3.2), we have that

$$\gamma(z,\vartheta) = \mathbb{E}[z^{X_1^*} e^{-\vartheta Y_1^*}] = \mathbb{E}[(z\beta(\vartheta))^{X_1^*}] = \varphi\{\lambda[1-\alpha(z\beta(\vartheta))]\}. \tag{6.2}$$

The two-dimensional benchmark process, to which the busy period policy is applied, is now

$$(A,B) = \{(A_n = \textstyle\sum_{k=0}^{n} X_k^*,\ B_n = \sum_{k=0}^{n} Y_k^*);\ n = 0,1,\ldots\}. \tag{6.3}$$

(A,B) should, as before, hit level $L = (N-1,D)$ to have the server stop vacationing and resume his service. The number of vacation segments made prior to this will equal the termination index ν. The first passage time will obviously be r_ν (replacing τ_ν). $h(\theta)$ will be replaced by $\varphi(\theta)$, now standing for $\mathbb{E}[e^{-\theta r_1}]$, and the marginal transformation $\gamma(z,0)$ satisfies formula (6.1).

The analog of the joint transformation in (3.3) for our case is

$$\begin{aligned} &\mathbb{E}^0[\xi^\nu e^{-\theta r_\nu} z^{A_\nu}] \\ &= 1 - [1 - \xi\varphi(\theta)\varphi[\lambda(1-\alpha(z))]]\mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\xi\varphi(\theta)\varphi[\lambda(1-\alpha(xz\beta(s))]}\right\}, \end{aligned} \tag{6.4}$$

which yields corresponding expressions for termination index, first excess level, and first passage time. We will introduce only a few later on.

We analyze the process $\{Q_n\}$ embedded in $Q(t)$ over the service completions, T_0, T_1, …, by using the same arguments as for the previous model with very little impact on the pgf $P(z)$:

$$P(z) = p_0 \mathrm{b}(z)\frac{1-E_L^{(0)}(z)}{\mathrm{b}(z)-z}, \tag{6.5}$$

where $E_L^{(0)}(z)$ is the marginal functional of (6.4):

$$E_L^{(0)}(z) = \mathbb{E}^0[z^{A_\nu}] = 1 - [1 - \varphi[\lambda(1-\alpha(z))]]\mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\varphi[\lambda(1-\alpha(xz\beta(s))]}\right\}, \tag{6.6}$$

and

$$p_0 = \frac{1-\rho}{a\lambda v \overline{T}_L^{(0)}}, \tag{6.7}$$

where the mean value of the termination index is

$$\overline{T}_L^{(0)} = \mathbb{E}^0[\nu] = \mathfrak{L}_D \mathfrak{D}_x^{N-1}\left\{\frac{1}{1-\varphi[\lambda(1-\alpha(x\beta(s))]}\right\}. \tag{6.8}$$

The ergodicity condition $\rho = ab\lambda < 1$ and the value $c = \frac{1}{a\lambda}$ of the mean stationary service cycle are also the same as for the previous model.

The main formula for the pgf $\pi(z)$ for continuous time parameter process also does not change with respect to its relationship with $P(z)$:

$$\pi(z) = ap_0 + \frac{az}{\mathrm{b}(z)}\Delta(z)P(z), \tag{6.9}$$

where

$$\Delta(z) = \frac{1-\mathrm{b}(z)}{\lambda[1-\alpha(z)]}. \tag{6.10}$$

The above will be convenient to summarize in the following statement.

Theorem 3. *For the model under (N,D)-policy with multiple vacations, the queueing process $Q(t)$ has a unique stationary distribution $\pi = (\pi_0, \pi_1, \ldots)$, given the ergodicity condition $\rho < 1$ (which is sufficient and necessary for the embedded process $\{Q_n\}$), and it is expressed in the form of its pgf $\pi(z)$ satisfying formulas* (6.9) *and* (6.10), *in which the expressions for the pgf $P(z)$ are given by formulas* (6.5-6.8).

7. Summary and Open Problems

The (N,D)-policy seems to be a rather advanced way of controlling queueing and workload processes in bulk systems with and without server vacations. With the use of the first excess level techniques it enabled us to analyze the queueing process (which previously was difficult even for a regular D-policy systems) and derive its stationary distributions in closed analytic forms. The studied models can be extended by employing state dependent input and service, which would enhance the versatility of the systems and further motivate the rationale of (N,D)-policy.

A practical generalization of the above systems would be the one with batch service, along with its perspective use of the hysteretic control. The latter means that the system abandons a busy period whenever the queue drops below some specified level (say, r) and ends the following idle or vacation period when the queue accumulates to a level N ($\geq r$). The real obstacle here is its difficulty in determining the workload when making a decision on resuming service based on whether or not it hits level D, as it is hard to predict how the server will form servicing batches and keep track on the rests of customers. A clever solution to this problem could lead to a significant enhancement over the proposed models.

References

[1] Abolnikov, L. and Dshalalow, J.H., A first passage problem and its applications to the analysis of a class of stochastic models, *J. Appl. Math. Stoch. Analysis*, **5**:1 (1992), 83-98.

[2] Babitsky, A.V., $M/G/1$ vacation model with limited service discipline and hybrid switching-on policy, *Math. Probl. Engin.* (to appear).

[3] Baker, K., A note on operating policies for the queue $M/M/1$ with exponential startups, *INFOR*, **11**:1 (1973), 71-72.

[4] Balachandran, K.R., Queue length dependent priority queues, *Manag. Sci.*, **17**:7 (1971), 463-471.

[5] Balachandran, K.R., Control policies for a single server system, *Manag. Sci.*, **19**:9 (1973), 1013-1018.

[6] Balachandran, K.R. and Tijms, H., On the D-policy for the $M/G/1$ queue, *Manag. Sci.*,**21**:9 (1975), 1073-1076.

[7] Borthakur, A., Medhi, J., and Gohain, R., Poisson input queueing system with startup time and under control-operating policy, *Compt. Oper. Res.*, **14**:1 (1987), 33-40.

[8] Boxma, O.J., Note on a control problem of Balachandran and Tijms, *Manag. Sci.*, **22**:8 (1976), 916-917.

[9] Chae, K.C. and Lee, H.W., $M^X/G/1$ vacation models with N-policy: heuristic interpretation of the mean waiting time, *J. Oprnl. Res. Soc.*, **46** (1995), 258-264.

[10] Dshalalow, J.H., On termination time processes, in *Studies in Applied Probability*, Edited

by J. Galambos and J. Gani, Essays in honor of Lajos Takács, *J. Appl. Prob.*, Special Volume **31A** (1994), 325-336.

[11] Dshalalow, J.H., First excess levels of vector processes, *J. Appl. Math. Stoch. Anal.*, **7**:3 (1994), 457-464.

[12] Dshalalow, J.H., Excess level processes in queueing, in: *Advances in Queueing* (ed. by J.H. Dshalalow), 243-262, CRC Press, Boca Raton, FL 1995.

[13] Dshalalow, J.H., Queueuing systems with state dependent parameters, in: *Frontiers in Queueing* (ed. by J.D. Dshalalow), 61-116, CRC Press, Boca Raton, FL 1997.

[14] Dshalalow, J.H. and Yellen, J., Bulk input queues with quorum and multiple vacations, *Math. Probl. Engin.*, **2**:2 (1996), 95-106.

[15] Federguen, A. and So, K.C., Optimality of threshold policy in single-server queueing systems with server vacations, *Adv. Appl. Prob.*, **23** (1991), 388-405.

[16] Heyman, D.P., Optimal operating policies for $M/G/1$ queueing systems, *Oper. Res.* **16**:2 (1968), 362-382.

[17] Hong, J.W. and Lie, C.H., Mean waiting time analysis of cyclic server system under N-policy, *J. Kor. Oper. Res. Soc.*, **18**:3 (1993), 51-63.

[18] Jaiswal, N.K. and Simha, P.S., Optimal operating policies for the finite-source queueing process, *Oper. Res.*, **20**:3 (1972), 698-707.

[19] Kramer, M., Stationary distributions in a queueing system with vacation times and limited service, *Queueing Sys.*, **4** (1989), 57-68.

[20] Kubat, P. and Servi, L.D., Cyclic service queues with very short service times, *European J. Opnl. Res.*, **53** (1991), 172-188.

[21] LaMaire, R.O., $M/G/1/N$ vacation model with varying E-limited service discipline, *Queueing Sys.*, **11** (1992), 357-375.

[22] Lee, D-S., A two-queue model with exhaustive and limited service discipline, *Stoch. Mod.*, **12**:2 (1996), 285-305.

[23] Lee, H.-S. and Srinivasan, M.M., Control policies for the $M^X/G/1$ queueing system, *Mgt. Sci.*, **35**:6 (1989), 708-721.

[24] Lee, H.-S., Optimal control of the $M^X/G/1/K$ queue with multiple server vacations, *Comput. Oper. Res.*, **22**:5 (1995), 543-552.

[25] Lee, H.W., Lee, S.S., and Chae, K.C., Operating characteristics of $M^X/G/1$ queue with N-policy, *Queueing Sys.*, **15** (1994), 387-399.

[26] Lee, H.W., Lee, S.S., Park, J.O., and Chae, K.C., Analysis of $M^X/G/1$ queue with N-policy and multiple vacations, *J. Appl. Prob.*, **31** (1994), 467-496.

[27] Lee, H.W., Lee, S.S., and Chae, K.C., A fixed-size batch service queue with vacations, *J. Appl. Math. Stoch. Anal.*, **9**:2 (1996), 205-219.

[28] Lee, S.S., Lee, H.W., Yoon, S.H., and Chae, K.C., Batch arrival queue with N-policy and single vacation, *Compt. Oper. Res.*, **22**:2 (1995), 173-189.

[29] Li, J. and Niu, S-C., The waiting time distribution for the $GI/G/1$ queue under D-policy, *Prob. Engineer. Inform. Sci.*, **6** (1992), 287-308.

[30] Loris-Teghem, J., Hysteretic control of an $M/G/1$ queueing system with two service time distributions and removable server, in: *Point Processes and Queueing Problems*, Colloquia Mathematica Societatis János Bolyai, Hungary, **24** (1978), 291-305.

[31] Loris-Teghem, J., Imbedded and non-imbedded stationary distributions in a finite capacity queueing system with removable server, *Cah. Centr. d'Etud. Rech. Opér.*, **26**:1-2 (1984), 87-94.

[32] Medhi, J. and Templeton, J.G.C., A Poisson input queue under N-policy and with general start up time, *Compt. Oper. Res.*, **19**:1 (1992), 35-41.

[33] Muh, D.C.-R., A bulk queueing system under N-policy with bilevel service delay discipline and start-up time, *J. Appl. Math. Stoch. Anal.*, **6**:4 (1993), 359-384.

[34] Muh, D.C.-R., On a Class of N-Policy Multilevel Control Queueing Systems, Doctoral Thesis, Florida Tech, Melbourne, FL 1994.

[35] Neuts, M.F., Generalizations of the Pollaczek-Khinchine integral equations in the theory of queues, *Adv. Appl. Prob.*, **18** (1986), 952-990.

[36] Park, J.O. and Lee, H.W., Optimal strategy in N-policy system with early set-up, *J. Opnl. Res. Soc.* (to appear).

[37] Ramaswami, R. and Servi, L., The busy period of the $M/G/1$ vacation model with a Bernoulli schedule, *Stoch. Mod.*, **4**:3 (1988), 507-521.

[38] Rubin, I. and Zhang, Z., Switch-on policies for communications and queueing systems, in: *Proceedings of the Third International Conference on Data Communication*, 329-339, Elsevier, North-Holland, Amsterdam, 1988.

[39] Shanthikumar, J.G., Optimal control of an $M/G/1$ priority queue via N-control, *Amer. Journ. Math. Manag. Sci.*, **1** (1981), 191-212.

[40] Takagi, H., Time-dependent process of $M/G/1$ vacation models with exhaustive service, *J. Appl. Prob.*, **29** (1992), 418-429.

[41] Takine, T. and Hagesawa, T., A note on $M/G/1$ vacation systems with waiting time limits, *Adv. Appl. Prob.*, **22** (1990), 513-518.

[42] Talman, A.J.J., A simple proof of the optimality of the best N-policy in the $M/G/1$ queueing control problem with removable server, *Statistica Neerl.*, **32** (1979), 143-150.

[43] Teghem, J., Jr., Optimal control of queues: removable servers [Tutorial paper XIX], *Belgian J. Oper. Res. Stat. Comp. Sci.*, **25**:2-3 (1985), 99-128.

[44] Teghem J., Jr., Control of the service process in a queueing system, *Europ. J. Oper. Res.*, **23** (1986), 141-158.

[45] Yadin, M. and Naor, P., Queueing systems with removable service station, *Oper. Res. Quart.*, **14** (1963), 393-405.