Journal of the Operations Research
Society of Japan
Vol. 23, No. 4, December 1980

ANALYSIS OF THE CONTROL OF QUEUES WITH SHORTEST PROCESSING TIME SERVICE DISCIPLINE

Jeyaveerasingam George Shanthikumar
University of Toronto

(Received November 12, 1979; Revised July 1, 1980)

Abstract We analyze two models of controlled M/G/1 queues with shortest processing time service discipline and removable server. The control policies considered here are the server vacation policy of Levy and Yechiali and the N-control policy of Heyman. The Laplace-Stieltjes transforms of the waiting time distributions, the mean cost rates and the optimal control policies are derived for these two models. Properties of level crossings of regenerative processes and delayed busy cycles are used in our analysis.

1. Introduction

Phipps [9] derived the mean waiting time in a single server queue with exponential arrival and shortest processing time service discipline (M/G/1, SPT queue), treating it as the limiting case of the non-preemptive priority queue analysed by Cobham [3] (and later by Takacs [15]). Recently Shanthikumar and Buzacott [10] using the concept of delayed busy cycle derived the Laplace-Stieltjes transform (LST) of the conditional waiting time distribution function (df) of a customer requiring a processing time p (p-customer) in an M/G/1, SPT queue. Here we derive the LST of the conditional waiting time df, the mean cost rates and the optimal policies for M/G/1, SPT queues with T-control policy of Levy and Yechiali [8] and N-control policy of Heyman [6]. In our analysis we use the concept of delayed busy cycle and the properties of level crossings of alternating regenerative processes (c.f. [10-13]).

In section 2 we give some preliminary results from delayed busy cycle, residual life and level crossing analyses. In section 3 a brief description of the model considered in this paper is given. Two examples, namely M/G/1, SPT queue with T-control and N-control policies are analyzed in section 4.

2. Preliminaries

In the analysis to follow we will be using the notion of delayed busy cycle, residual life and level crossing analysis. Therefore in this section we will present the basic results of these analyses. A delayed busy cycle is normally described by the initial delay and the type of customer arrivals that extends the busy period beyond this initial delay. Let G_0 be the initial delay duration and G_b be the length of the delayed busy period. Then the length of the delayed busy cycle is the total duration $G_0 + G_b \triangleq G_d$. It can be shown (c.f. Conway, Maxwell and Miller [5] page 151) that the LST $\tilde{G}_d(s)$ of the df of the length of the delayed busy cycle caused by an initial delay with df $G_0(\cdot)$ and by a Poisson arrival stream of customers with processing time df $B_0(\cdot)$ and arrival rate λ_0, is given by

$$\tilde{G}_d(s) = \tilde{G}_0(s+\lambda_0-\lambda_0\tilde{G}_c(s)), \text{ where } \tilde{G}_c(s) = \tilde{B}_0(s+\lambda_0-\lambda_0\tilde{G}_c(s)) \tag{1}$$

It should be noted that, throughout this paper, unless otherwise stated, the df of a random variable (rv) X will be denoted by $X(\cdot)$, its LST by $\tilde{X}(\cdot)$ and expected value by $E[X]$. Next we give the residual life of a rv Y. If y_r is the residual life

$$\tilde{Y}_r(s) = (1 - \tilde{Y}(s))/sE[Y] \tag{2}$$

We will next present the basic concepts of level crossing analysis. Level crossing analysis is a relatively new technique developed to solve queueing and related problems. In this analysis the relationship between the number of up- and downcrossings over a level x of the virtual delay process during a regeneration cycle is used as the starting model equation (c.f. Shanthikumar [12, 13] for a detailed treatment). We next describe a special case of regenerative process and present the results for such analysis.

Let X_n, n=1,2,... be a renewal sequence with nonarithmetic df and finite mean. Also define the alternating renewal sequences A_n and I_n such that $X_n = A_n + I_n$. Now define the regenerative process $\{V_t, t \geq 0\}$ with respect to X_n, n=1,2,..., such that if $Z_0 = 0$ and $Z_n = X_1 + X_2 + \ldots + X_n$ (c.f. Cohen [4], Smith [14]), then $V_t^{(n)} \triangleq V_{t+Z_n}$, n=0,1,2,..., have identical probabilistic properties. In the special case of regenerative process considered, the following additional assumptions will be made:

(i) $V_I^{(n)} \triangleq V_{Z_n+}$ have identical probabilistic properties

with $\Pr\{V_I^{(n)} < x\} = Q(x)$ independent of n,

(ii) $V_T^{(n)} \triangleq V_{Z_n+A_{n+1}-}$ have identical probabilistic properties with

$\Pr\{V_T^{(n)} < x\} = H(x)$ independent of n,

(iii) $V_t = 0$, $t \in (Z_n + A_{n+1}, Z_{n+1})$ and $V_t > 0$, $t \in (Z_n, Z_n + A_{n+1}) \setminus D$ $n=0,1,2,\ldots$, where D is at most a denumerable set and includes the epochs Z_n, $n=0,1,2,\ldots$

(iv) V_t is continuous and strictly decreasing at a rate of one at every $t \in (Z_n, Z_n + A_{n+1}) \setminus D$, $n=0,1,2,\ldots$, and has a non-negative upward jump such that $V_{t+} > V_{t-}$ at every $t \in D \subset [0, \infty)$.

It will be assumed that the number of upward jumps N_n during $(Z_n, Z_n + A_{n+1})$, (excluding those jumps at epochs Z_n, $n=0,1,\ldots$) form a discrete valued renewal sequence with finite mean. Let t_n, $n=1,2,\ldots$, be the epochs of these upward jumps and let $W_n = V_{t_n-}$ for $n=1,2,\ldots$. It will also be assumed that W_n, $n=1,2,\ldots$, form a discrete time regenerative process on the renewal sequence N_n, $n=1,2,\ldots$. Let A be the length of the active phase (AP) $(Z_n, Z_n + A_{n+1})$ and N be the number of upward jumps during (0, A). Now define $D_x \triangleq \#\{t: V_t = x \mid t \in [0, A) \setminus D\}$, the number of downcrossings and $U_x \triangleq \#\{t: V_{t+} > x, V_{t-} < x \mid t \in D \cap [0,A)\}$, the number of upcrossing over level x during an AP [0, A). Then it can be shown that (c.f. Shanthikumar [13])

$$f(x) = (E[U_x] - (1 - H(x)))/E[A] \tag{3}$$

and

$$E[U_x] = (1 - Q(x)) + E\left[\sum_{i=1}^{N} \int_0^x (1 - G_i(x - u))dW_i(u)\right],$$

where $f(x) = dF(x)/dx$, $F(x) = \lim_{t\to\infty} \Pr\{V_t < x \mid t \in AP\}$ is assumed differentiable in $x > 0$, and $G_i(\cdot)$ is the df of the magnitude of the i^{th} $(i=1,2,\ldots,N)$ jump during (0, A). If $W_i = 0$ with probability 1 (wp.1) $\forall$ i and $G(\cdot)$ is the df of the magnitude of an arbitrary jump, then

$$E[U_x] = (1 - Q(x)) + E[N]\ (1 - G(x)) \tag{4}$$

3. The Model

The model class considered in this paper is of a single server queue at

which arrivals form a homogeneous Poisson stream with rate λ and queue for service using a shortest processing time (SPT) service discipline. The server is shut down and started up according to some control policy. In this paper we consider two server control policies, namely multiple server vacations (T-control c.f. Levy and Yechiali [8]) and N-control (c.f. Heyman [6]) policies. In each case it is assumed that the service time df $B(\cdot)$ is continuous and that the induced stationary virtual waiting time process of a p-customer exists and has a unique df which is differentiable.

The customers requiring processing time greater than p will be refered to as ℓ-customers and customers requiring processing time less than p as h-customers. The service time df of an ℓ-customer (w.r.t. a given p) is $B_\ell(x) = [B(x)-B(p)]/[1 - B(p)]$, $x > p$ with the first two moments being $\bar{x}_\ell(p)$ and $\overline{x_\ell^2}(p)$. Similarly for the h-customers the service time df is $B_h(x) = B(x)/B(p)$, $x < p$ with the first two moments being $\bar{x}_h(p)$ and $\overline{x_h^2}(p)$. Since the arrivals are Poisson the arrivals of ℓ- and h-customers are also Poisson processes with rates $\bar{\lambda}(p) \stackrel{\Delta}{=} \lambda(1-B(p))$ and $\lambda(p) \stackrel{\Delta}{=} \lambda B(p)$ respectively. Then define $\bar{\rho}(p) = \bar{\lambda}(p)\bar{x}_\ell(p) = \lambda \int_p^\infty x dB(x)$ and $\rho(p) = \lambda(p)\bar{x}_h(p) = \lambda \int_0^p x dB(x)$. Then $\bar{\rho}(p) + \rho(p) = \lambda\bar{x} \stackrel{\Delta}{=} \rho$, where $\bar{x}$ is the mean service time of an arbitrary customer. We will assume that $\rho < 1$. Similarly let $\bar{W}_0(p) = \bar{\lambda}(p)\overline{x_\ell^2}(p)/2$ and $W_0(p) = \lambda(p)\overline{x_h^2}(p)/2$. Then $W_0(p) + \bar{W}_0(p) = W_0 \stackrel{\Delta}{=} \lambda\overline{x^2}/2$, where $\overline{x^2}$ is the second moment of the service time of an arbitrary customer.

4. Examples

Ex. 1. M/G/1, SPT, T-control

In this example, the server scans the queue T time units after the end of a busy phase, to determine whether there are any customers present in the system. It is assumed that the scan time T is a r.v. with df $T(\cdot)$. If any customers are found in the system at the scan epoch, a busy phase begins and the server is kept in the active state until the system is empty. If no customers are found a busy period of length zero occurs. In either case, the next scan is made T time units after the end of the busy period. The server is said to be on vacation during the scan interval.

If a virtual p-customer enters the system just after the end of the busy phase (that is, at the beginning of the server vacation) the p-customer has to wait in queue for a time period equal to the length of the delayed busy cycle G_T caused by an initial delay of T time units and by an arrival stream of h-customers. Then from (1) we have $\tilde{G}_T(s) = \tilde{T}(s+\lambda(p)-\lambda(p)\tilde{G}_h(s))$, where $\tilde{G}_h(s) =$

$\tilde{B}_h(s+\lambda(p) - \lambda(p)\tilde{G}_h(s))$. This virtual delay will drop at a rate of one until the delayed busy cycle G_T is over. At this point of time, if the virtual p-customer were to arrive he would incur no delay because the system is empty of h-customers. However if there are any jobs at this time, all of them will be ℓ-customers and the virtual delay will jump by an amount equal to the length of a delayed busy cycle caused by an initial delay equal to the minimum processing time of the ℓ-customers available at that time and by the arrival stream of h-customers. If there are no customers available then the busy period will end. Otherwise the virtual delay process will continue in the same fashion as discussed above, a jump occurring at every time it reaches zero by an amount equal to the delayed busy cycle caused by the initial delay equal to the minimum processing time of the available ℓ-customers and by the arrival stream of h-customers, until the system becomes empty.

For every ℓ-customer that goes into service the virtual delay will jump from 0 and the magnitude of an arbitrary jump is equal to the length of a delayed busy cycle G_ℓ caused by an initial delay equal to the service time of an ℓ-customer and arrival stream of h-customers. Then from (1), $\tilde{G}_\ell(s) = \tilde{B}_\ell(s +\lambda(p) - \lambda(p)\tilde{G}_h(s))$. Let C be the length of the busy cycle of the virtual delay process, then the expected number of ℓ-customers served during the busy cycle C is $\bar{\lambda}(p)E[C]$. Comparing this busy cycle to the special case of regenerative process defined earlier we have $H(x) = 1$, $\forall x > 0$, $E[A] = E[C]$, $Q(x) = G_T(x)$, $E[N] = \bar{\lambda}(p)E[C]$, $W_i = 0$ w.p.1 and $G(x) = G_\ell(x)$. Then from (3) and (4) and taking Laplace transforms we get $\tilde{F}(s) = \{1 - \tilde{G}_T(s) + \bar{\lambda}(p)E[C]\ (1 - \tilde{G}_\ell(s))\}/sE[C]$, where $\tilde{F}(s)$ is the LST of the virtual waiting time (also the waiting time because of Poisson arrival) $W_q(p)$ of a p-customer. Using $\lim_{s\to 0} F(s) = 1$, after substituting for $\tilde{G}_T(s)$ and $\tilde{G}_\ell(s)$, we get $E[C] = E[T]/(1 - \rho)$. Substituting this for $\tilde{F}(s)$ we get

$$\tilde{F}(s) = \{(1 - \rho)/sE[T]\}\ \{1 - \tilde{T}(s + \lambda(p) - \lambda(p)\tilde{G}_h(s)\} + (\bar{\lambda}(p)/s)\ \{1 - \tilde{B}_\ell(s + \lambda(p) - \lambda(p)\tilde{G}_h(s)\}, \tag{5}$$

where $\tilde{G}_h(s) = \tilde{B}_h(s + \lambda(p) - \lambda(p)\tilde{G}_h(s))$. Then

$$E[W_q(p)] = \lim_{s\to 0} - \frac{\partial F(s)}{\partial s} = \{W_0 + [(1-\rho)E[T^2]/2E[T]]\}/(1-\rho(p))^2$$

is obtained from (5) and using the identities given in section 3. Since the expected waiting time of an arbitrary customer is $E[W_q(SPT, T)] = \int_0^\infty E[W_q(p)]$ $dB(p)$ we get after nothing that $E[W_q(SPT)] = \int_0^\infty \{W_0 dB(p)/(1-\rho(p))^2\}$ is the mean waiting time of an arbitrary customer in an M/G/1, SPT queue and $E[W_q(FCFS)] =$

$W_0/(1-\rho)$ is the mean waiting time in an M/G/1 queue with first come first served (FCFS) discipline,

$$E[W_q(SPT,T)] = \{\frac{W_0}{1-\rho} + \frac{E[T^2]}{2E[T]}\} \; (E[W_q(SPT)]/E[W_q(FCFS)])$$

It can be easily verified that the quantity within the double brackets is the mean waiting time $E[W_q(FCFS, T)]$ in an M/G/1 queue with FCFS service discipline and T-control policy (c.f. Levy and Yechiali [8]). This leads to an interesting conservation identity

$$E[W_q(SPT, T)]/E[W_q(SPT)] = E[W_q(FCFS, T)]/E[W_q(FCFS)].$$

More will be said about this later in the conclusions. Now let R be the shut down and start up cost and h be the cost of waiting per unit time for a customer. Then the cost rate c(SPT, T) is equal to $R/E[C] + \lambda h E[W_q(SPT, T)]$.

$$c(SPT, T) = \frac{R(1-\rho)}{E[T]} + \lambda h\{\frac{W_0}{1-\rho} + \frac{E[T^2]}{2E[T]}\} \; (\frac{E[W_q(SPT)]}{E[W_q(FCFS)]})$$

For deterministic scan time T it is easily shown that the optimal scan time T*(SPT, T) is given by

$$T^*(SPT, T) = T^*(FCFS, T) \sqrt{\frac{E[W_q(FCFS)]}{E[W_q(SPT)]}}$$

where

$$T^*(FCFS, T) = \sqrt{\frac{2R(1-\rho)}{\lambda h}}$$

is the optimal deterministic scan time for the FCFS case.

Ex. 2. M/G/1, SPT, N-control

We will now consider an M/G/1 queue with SPT service discipline and N-control policy. Under an N-control policy, the server is deactivated when there is no customer in the system and activated when there are N customers in the system. The arrival of the first N-1 customers after server deactivation forms the blocking phase (BLP) where the server is inactive and an arrival will not activate the server. The inter arrival time between the (N-1)th and Nth customer forms the waiting phase (WP) where a customer arrival will activate the server and the time between server activation and the next deactivation forms the busy phase (BP) where the server is active (c.f. Shanthikumar [12,

13]). Now we will analyse these three phases separately. Let $F(\cdot|i)$ be the df of the virtual waiting time $W_q(p)$ of a p-customer in phase $i \in \{BLP, WP, BP\}$.

(i) Blocking phase (BLP)

If a virtual p-customer were to arrive just after the end of BP (that is, at the beginning of BLP) he will have to wait for the arrival of N-1 other customers for the server to be active again. However the virtual p-customer could obtain service only when the system is empty of any h-customers. That is the virtual p-customer will incur a virtual delay composed of: (i) the time $A^{(N-1)}$ taken for N-1 arrivals and (ii) the length of the delayed busy cycle G_{N-1} caused by an initial delay I_{N-1} equal to the total processing time of the h-customers among the N-1 customers discussed above, and by an arrival stream of h-customers. As time goes by the virtual delay will drop at a rate of one until these N-1 customers arrive. Just before the (N-1)th arrival the virtual delay will be equal to G_{N-1}. Therefore the stationary df of the virtual delay is the convolution of the df $G_{N-1}(\cdot)$ of the delayed busy period and the df $A_r^{(N-1)}(\cdot)$ of the residual life of (N-1) inter arrival time $A^{(N-1)}$.

Given that k out of N-1 customers are h-customers we have $I_{N-1}(x) = B_h^{(k)}(x)$ equal to k fold convolution of $B_h(\cdot)$ with itself. Noting that P_k = Pr {k h-customers out of N-1} = $\binom{N-1}{k}(B(p))^k(1 - B(p))^{N-1-k}$ and $I_{N-1}(x) = \sum_{k=1}^{N-1}(B_h^{(k)}(x)p_k)$, it can be easily shown that $\tilde{I}_{N-1}(s) = (1 - B(p)[1 - \tilde{B}_h(s)])^{N-1}$. Then from (1) we have $\tilde{G}_{N-1}(s) = (1 - B(p)[1 - \tilde{G}_h(s)])^{N-1}$. Also from (2) we have $\tilde{A}_r^{(N-1)}(s) = \lambda(1 - \tilde{A}^{N-1}(s))/[s(N-1)]$, where $\tilde{A}(s) = \lambda/[\lambda+s]$ is the LST of an exponential inter arrival time. Then

$$\tilde{F}(s|BLP) = \{\lambda(1-\tilde{A}^{N-1}(s))(1-B(p)[1-\tilde{G}_h(s)])^{N-1}\}/[s(N-1)] \tag{6}$$

(ii) Waiting phase (WP)

Just after the (N-1)th arrival the waiting phase (WP) begins. If the virtual p-customer were to arrive during WP the server will be activated immediately. Therefore the virtual waiting time remains constant at G_{N-1} during WP. The WP will end at the arrival of the Nth customer. Therefore

$$\tilde{F}(s|WP) = \tilde{G}_{N-1}(s) = (1 - B(p)[1 - \tilde{G}_h(s)])^{N-1} \tag{7}$$

(iii) Busy phase (BP)

Arrival of the N-th customer after server deactivation commences the busy phase. If the virtual p-customer comes along with the N-th arrival, then the

virtual delay will be equal to the delayed busy cycle G_N with initial delay equal to the total processing time of all h-customers among these N-customers and arrival stream of h-customers. Similar to the derivation of $\tilde{G}_{N-1}(s)$ we have $\tilde{G}_N(s) = (1 - B(p)\ [1 - \tilde{G}_h(s)])^N$. However if the virtual p-customer were to arrive just after the N-th arrival, the virtual waiting time will be the same as above except when all of the N customers are ℓ-customers. In this case the virtual delay will be equal to the delayed busy cycle with an initial delay equal to the minimum processing time of these N ℓ-customers and arrival stream of h-customers (in G_N this delay would be counted as zero w.p.1). After this the virtual delay will drop with time at a rate of one until it reaches zero. If a virtual p-customer were to arrive when the virtual delay reaches zero he will immediately go into service because there is no h-customer in the system. However, if the virtual p-customer does not enter the queue, an ℓ-customer with the shortest processing time will be taken up for service. Therefore the virtual delay will jump by an amount equal to the delayed busy cycle caused by a delay equal to the minimum processing time among those jobs available at that time and arrival of h-customers. However if there is no customer available the busy phase of the virtual delay process will end. Otherwise the virtual delay process will continue in the same fashion as discussed above, a jump occurring at every point it reaches zero by an amount equal to the delayed busy cycle caused by the initial delay equal to the minimum processing time of the available ℓ-customers and by the arrival stream of h-customers, until the system becomes empty. Similar to example 1, the df of the magnitude of an arbitrary jump is $G_\ell(\cdot)$. If C is the length of BP, the expected number of such jumps (including a possible jump at the beginning of BP) is $\bar{\lambda}(p)E[C] + N(1 - B(p))$. Then comparing BP to the special case of regenerative process discussed in section 2, we have $H(x) = 1$, $\forall\ x > 0$,
$E[A] = E[C]$, $W_i = 0$ w.p.1. Even though $Q(x) \neq G_N(x)$,
$E[N] \neq \bar{\lambda}(p)E[C] + N(1 - B(p))$ and $G(x) \neq G_\ell(x)$, since
$G_N = 0$ when all N customers are ℓ-customers and the jump at the BP corresponding to this case is counted in $N(1 - B(p)) + \bar{\lambda}(p)E[C]$, we have $E[U_x] = (1-G_N(x)) + [\bar{\lambda}(p)E[C] + N(1-B(p))]\ (1-G_\ell(x))$. Then from (2) and taking Laplace transforms we get

$$\tilde{F}(s|BP) = \{(1-\tilde{G}_N(s)) + [\bar{\lambda}(p)E[C] + N(1-B(p))](1-\tilde{G}_\ell(s))\}/sE[C].$$

Now substituting for $\tilde{G}_N(s)$ and $\tilde{G}_\ell(s)$ and using $\lim_{s\to 0} \tilde{F}(s|BP) = 1$, we get $E[C] = N\bar{x}/(1 - \rho)$. Feeding this for $\tilde{F}(s|BP)$ and noting that $\bar{\lambda}(p) = \lambda(1 - B(p))$ we get

$$\tilde{F}(s|BP) = \left(\frac{1-\rho}{sN\bar{x}}\right)\{(1-(1-B(p)[1-\tilde{G}_h(s)])^N)\} + \frac{1}{s\bar{x}} \cdot$$

$$\{(1-B(p))[1-\tilde{G}_\ell(s)]\} \tag{8}$$

It can be shown that the LST $\tilde{F}(s)$ of the virtual waiting time (also the waiting time because of Poisson arrivals [11]) of a p-customer is equal to $\Sigma_i \tilde{F}(s|i)Pr\{i\}$, $i \in \{BLP, WP, BP\}$, where $Pr\{i\}$ is the steady probability that the system is in state i (c.f. Shanthikumar [11]). Since these probabilities $Pr\{i\}$ are independent of the scheduling disciplines from [12], $Pr\{BLP\} = (N-1)(1-\rho)/N$, $Pr\{WP\} = (1-\rho)/N$ and $Pr\{BP\} = \rho$. Then from (6), (7) and (8) we get,

$$\tilde{F}(s) = \{\lambda(1-\rho)(1-\tilde{A}^{N-1}(s))\tilde{G}_1^{N-1}(s) + s(1-\rho)\tilde{G}_1^{N-1}(s) + \lambda(1-\rho) \cdot$$

$$(1-\tilde{G}_1^N(s)) + \lambda N(1-B(p))(1-\tilde{G}_\ell(s))\}/(sN) ,$$

where

$$\tilde{G}_1(s) = (1 - B(p)[1 - \tilde{G}_h(s)]) \tag{9}$$

and

$$\tilde{G}_h(s) = \tilde{B}_h(s + \lambda(p) - \lambda(p)\tilde{G}_h(s))$$

Then from $E[W_q(p)] = \lim_{s\to 0} -\frac{\partial \tilde{F}(s)}{\partial s}$ and using the identities given in section 3, we get after some algebra $E[W_q(p)] = \{W_C + (N-1)(1-\rho)/2\lambda\}/(1-\rho(p))^2$. Then similar to example 1, with obvious notations we get

$$E[W_q(SPT, N)] = \left\{\frac{W_0}{1-\rho} + \frac{N-1}{2\lambda}\right\}(E[W_q(SPT)]/E[W_q(FCFS)]) .$$

It can be easily verified that the quantity within the double brackets is the mean waiting time $E[W_q(FCFS, N)]$ in an M/G/1, FCFS queue with N-control policy (c.f. Heyman [6]). This leads to the conservation identity

$$(E[W_q(SPT, N)]/E[W_q(SPT)]) = (E[W_q(FCFS, N)]/E[W_q(FCFS)]) .$$

Similar to example 1 the cost rate c(SPT, N) is given by

$$c(SPT, N) = \frac{R\lambda(1-\rho)}{N} + \lambda h \left\{\frac{W_0}{1-\rho} + \frac{N-1}{2\lambda}\right\} \left(\frac{E[W_q(SPT)]}{E[W_q(FCFS)]}\right)$$

and the optimal policy for N is the nearest integer to N*(SPT, N) which minimizes c(SPT, N) where

$$N^*(SPT, N) = N^*(FCFS, N) \sqrt{\frac{E[W_q(FCFS)]}{E[W_q(SPT)]}} ,$$

and

$$N^*(FCFS, N) = \sqrt{\frac{2R\lambda(1 - \rho)}{h}} .$$

The optimal policy for the FCFS case is the nearest integer to $N^*(FCFS, N)$ that minimizes $c(FCFS, N) = R\lambda(1 - \rho)/N + \lambda h\{W_0/(1 - \rho) + (N - 1)/2\lambda\}$. Note that

$$\lambda T^*(SPT, T) = N^*(SPT, N)$$

and

$$\lambda T^*(FCFS, T) = N^*(FCFS, N) .$$

5. Conclusions

In this paper we have derived the LST of the waiting time df, the mean cost rates and the optimal control policies using the notion of delayed busy cycle, residual life and level crossing analysis, for the M/G/1, SPT queue with T-control and N-control policies. Heyman [6] derived the mean waiting time for M/G/1 non-preemptive priority queue with T-control and single vacation policies. Bell [1,2] analyzed the optimal control policies for such priority queues and Tijms [16] derived the mean waiting time in an M/G/1 priority queue with two classes and N-control policy. The method used in this paper can be easily extended for these models. These and other extensions will be reported shortly.

In the examples considered here we noted a form of conservation identity and it is $(E[W_q(SD, CP)]/E[W_q(SD)]) = C(CP)$, where, SD is the service discipline FCFS or SPT, CP is the control policy and C(CP) is a constant for each control policy CP, which is T- or N-control. This conservation identity holds for more general service disciplines (eg. non-preemptive priority, scheduling within generations, truncated shortest processing time, etc., c.f. Shanthikumar [12]) and control policies (eg. single vacation policy of Levy and Yechiali [8]). Use of this identify to analyze other control policies will be discussed elsewhere.

Acknowledgement

The author would like to thank the Canadian Commonwealth Scholarship and Fellowship Committee for providing the financial support for this research. The helpful comments of the referees are gratefully acknowledged.

References

[1] Bell, C.E.: Optimal Average Cost Operating Policy for an M/G/1 Queueing System with Removable Server and Several Priority Classes. *Tech. Report* 144, Dept. of OR, Stanford University, 1971.

[2] Bell, C.E.: Optimal Operation of an M/G/1 Priority Queue with Removable Server. *Operations Research*, 21 (1973), 1281-1290.

[3] Cobham, A.: Priority Assignment in Waiting Line Problems. *Operations Research* 2 (1954), 70-76.

[4] Cohen, J.W.: On Regenerative Processes in Queueing Theory. *Lecture Notes, Econ. & Math.* 121, Springer, Berlin, 1976.

[5] Conway, R.W., Maxwell, W.L. and Miller, L.W.: *Theory of Scheduling*. Addison-Wesley, Reading, Mass., 1967.

[6] Heyman, D.P.: Optimal Operating Policies for M/G/1 Queueing Systems. *Operations Research*, 16 (1968), 362-382.

[7] Heyman, D.P.: A Priority Queueing System with Server Interference. *SIAM J. Applied Math.* 17 (1969), 74-82.

[8] Levy, Y. and Yechiali, U.: Utilization of Idle Time in an M/G/1 Queueing System. *Management Sci.* 22 (1975), 202-211.

[9] Phipps, T.E.: Machine Repair as a Waiting Line Problem. *Operations Research* 4 (1956), 76-86.

[10] Shanthikumar, J.G. and Buzacott, J.A.: The Conditional Waiting Time in an M/G/1 Queue with Shortest Processing Time Discipline. *Working Paper* #79-004, Department of Industrial Engineering, University of Toronto, Toronto, 1979.

[11] Shanthikumar, J.G.: Some Properties of the Alternating Regenerative Processes. *Working Paper* #79-018, Department of Industrial Engineering, University of Toronto, Toronto, 1979.

[12] Shanthikumar, J.G.: Approximate Queueing Models of Dynamic Job Shops. *Ph. D. Thesis*, Department of Industrial Engineering, University of Toronto, Toronto, 1979.

[13] Shanthikumar, J.G.: Some Analyses on the Control of Queues Using Level Crossings of Regenerative Processes. *J. Applied Prob.* (1980), (to appear).

[14] Smith, W.L.: Regenerative Stochastic Processes. *Proc. Royal Soc. of London*, A232 (1955), 6-31.

[15] Takacs, L.: Priority Queues. *Operations Research* 12 (1964), 63-74.

[16] Tijms, H.: A Control Policy for a Priority Queue with Removable Server. *Operations Research* 22 (1974), 833-837.

Mr. J. G. Shanthikumar
Assistant Professor
Dept. of IE & OR
441 Link Hall
Syracuse University
Syracuse, New York 13210
U. S. A.