A Generalized N-Policy for an $M/M/1$ Queueing System and Its Optimization[1)]

Jongho Bae[2)], Jongwoo Kim[3)], Eui Yong Lee[4)]

Abstract

We consider a generalized N-policy for an $M/M/1$ queueing system. The idle server starts to work with ordinary service rate when a customer arrives. If the number of customers in the system reaches N, the service rate gets faster and continues until the system becomes empty. Otherwise, the server finishes the busy period with ordinary service rate. We obtain the limiting distribution of the number of customers in the system. After assigning various operating costs to the system, we show that there exists a unique fast service rate minimizing the long-run average cost per unit time.

Keywords: $M/M/1$ queue, N-policy, distribution of the number of customers, optimal service rate

1. Introduction

In this paper, we consider a generalized N-policy for an $M/M/1$ queueing system. The server is initially idle and starts to work with service rate $\mu_1 \geq 0$ on a customer's arrival. Customers arrive according to a Poisson process of rate $\nu > 0$. If the number of customers in the system reaches N ever, the server starts to serve customers with faster service rate μ_2 $(\mu_2 \geq \mu_1)$ including the customer being served at the moment, and continues until the system being empty. Otherwise, the server finishes the busy period with the ordinary service rate μ_1. For the stability of the system, we assume that μ_2 is greater than ν. When $\mu_1 = 0$, this policy is reduced to the original N-policy.

Bae, Kim, and Lee(2002) recently introduced a two service rate policy where the service rate is changed depending on the workload rather than the number of customers. They obtained the limiting distribution of the workload process.

---

1) This work was supported by KOSEF through Statistical Research Center for Complex System at Seoul National University.

2) Full-time Instructor, Department of Mathematics, Jeonju University, Jeonju, 560-759, Republic of Korea.

3) Ph.D. Candidate, Department of Mathematics, Pohang University of Science and Technology, Pohang, 790-784, Republic of Korea.

4) Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Republic of Korea.

The N-policy has been studied by many researchers since it was introduced by Yadin and Naor(1963). Heyman(1968)}, Sobel(1969), and Bell(1971) discussed the optimization of the N-policy. Federgruen and Tijms(1980), Yamada and Nishimura(1994), and Nishimura and Jiang(1995) extended the N-policy to the two service rate policies and studied the state probability of the queueing system.

In section 2, we obtain the limiting distribution of the number of customers in the system by using the decomposition technique(Lee and Ahn(1998)), by which we decompose the non-Markovian process of the number of customers into the several Markovian processes for the purpose of analysis.

After assigning operating costs to the system, related to the service rate, heavy traffic, and idle period, in section 3, we calculate the long-run average cost per unit time, and show that there exists a unique fast service rate μ_2 which minimizes the long-run average cost per unit time.

2. Limiting Distribution of the Number of Customers

Let $\{N(t), t\geq 0\}$ be the process of the number of customers in the $M/M/1$ queueing system under the generalized N-policy. Note that the time epoch when the server starts to work after an idle period is an embedded regeneration point of $\{N(t), t\geq 0\}$. Since $\{N(t), t\geq 0\}$ is non-Markovian, we first decompose $\{N(t), t\geq 0\}$ into three processes $\{N_1(t), t\geq 0\}$, $\{N_2(t), t\geq 0\}$, and $\{N_3(t), t\geq 0\}$. Process $\{N_1(t), t\geq 0\}$ is formed by separating the periods of service rate μ_1 from the original process and connecting them together. Process $\{N_2(t), t\geq 0\}$ is similarly formed by separating and connecting the periods of service rate μ_2. Process $\{N_3(t), t\geq 0\}$ is formed by connecting the rest of the original process, that is, $N_3(t)\equiv 0$ for all $t\geq 0$.

Notice that processes $\{N_1(t), t\geq 0\}$ and $\{N_2(t), t\geq 0\}$ are now Markovian regenerative processes. In both processes, we will call each separated segment a cycle. Each idle period of the original process becomes a cycle in $\{N_3(t), t\geq 0\}$. Observe that $\{N_1(t), t\geq 0\}$ finishes a cycle either at state 1 or $N-1$, and the probability of finishing at $N-1$ can be obtained from the gambler's ruin problem (Ross(1996), pp.186-188) as follows:

$$p=\begin{cases}\dfrac{1-(\mu_1/\nu)}{1-(\mu_1/\nu)^N} & \text{if } \mu_1\neq\nu,\\ \dfrac{1}{N} & \text{if } \mu_1=\nu.\end{cases} \tag{1}$$

Let $P_i(n)$ be the limiting distribution of $\{N_i(t), t\geq 0\}$ and T_i be the length of a cycle in $\{N_i(t), t\geq 0\}$, for $i=1,2,3$. Let $P(n)$ be the limiting distribution of $\{N(t), t\geq 0\}$ and T be the length of a cycle, the period between two successive regeneration points, in $\{N(t), t\geq 0\}$. Suppose we earn a reward at a rate of one per unit time when process $\{N(t), t\geq 0\}$ is equal to n. Then, by applying the renewal reward theorem(Ross(1996), p.133), we can see that, for $n=0,1,2,\ldots$,

$$P(n) = \frac{E[\text{reward during } T]}{E[T]} = P_1(n)\frac{E[T_1]}{E[T]} + pP_2(n)\frac{E[T_2]}{E[T]} + \frac{1}{\nu E[T]} I_{\{n=0\}}, \tag{2}$$

where $E[T] = E[T_1] + pE[T_2] + 1/\nu$ and $I_{\{A\}}$ is the indicator of event A. We, now, evaluate $P_1(n)$, $E[T_1]$, $P_2(n)$, and $E[T_2]$ in the following two subsections.

2.1. Limiting Distribution of $N_1(t)$

The corresponding balance equations are

$$\begin{aligned}(\nu+\mu_1)P_1(1) &= \nu P_1(N-1) + \mu_1 P_1(1) + \mu_1 P_1(2),\\ (\nu+\mu_1)P_1(n) &= \nu P_1(n-1) + \mu_1 P_1(n+1), \qquad \text{for } 2 \le n \le N-2,\\ (\nu+\mu_1)P_1(N-1) &= \nu P_1(N-2),\end{aligned}$$

which, with $\sum_{n=1}^{N-1} P_1(n) = 1$, have a unique solution given by

$$P_1(n) = \begin{cases} \dfrac{1-(\mu_1/\nu)\,1-(\mu_1/\nu)^{N-n}}{N1-(\mu_1/\nu)-1+(\mu_1/\nu)^N} & \text{if } \mu_1 \neq \nu, \\ \dfrac{2(N-n)}{N(N-1)} & \text{if } \mu_1 = \nu, \end{cases} \tag{3}$$

for $1 \le n \le N-1$.

To obtain $E[T_1]$, we observe that, after a busy period begins in the original process $\{N(t), t \ge 0\}$, process $M(t) = N(t) - (\nu - \mu_1)t$ is a martingale with $E[M(0)] = E[M(t)] = 1$ until $N(t)$ reaches either state 0 or N. We also observe that T_1 is equal in distribution to $T_1^* = \min\{t \ge 0 | N(t) = 0 \text{ or } N\}$. Applying the optional sampling theorem(Karlin and Taylor(1975), p.259) to $M(t)$ with Markov time T_1^* gives

$$E[M(0)] = E[M(T_1^*)] = N \cdot \Pr\{N(T_1^*) = N\},$$

where $\Pr\{N(T_1^*) = N\}$ is equal to p given in equation (1). Hence, $E[T_1] = E[T_1^*]$ is obtained as follows:

$$E[T_1] = \begin{cases} \dfrac{N/\nu}{1-(\mu_1/\nu)^N} - \dfrac{1/\nu}{1-(\mu_1/\nu)} & \text{if } \mu_1 \neq \nu, \\ \dfrac{N-1}{2\nu} & \text{if } \mu_1 = \nu. \end{cases} \tag{4}$$

2.2. Limiting Distribution of $N_2(t)$

The corresponding balance equations are

$$\begin{aligned}(\nu+\mu_2)P_2(1) &= \mu_2 P_2(2),\\ (\nu+\mu_2)P_2(n) &= \nu P_2(n-1) + \mu_2 P_2(n+1), \qquad \text{for } n \ge 2, n \neq N\\ (\nu+\mu_2)P_2(N) &= \mu_2 P_2(1) + \nu P_2(N-1) + \mu_2 P_2(N+1),\end{aligned}$$

which, with $\sum_{n=1}^{\infty} P_2(n)=1$, have a unique solution given by

$$P_2(n)=\begin{cases}\frac{1}{N}(1-(\nu/\mu_2)^n) & \text{for } 1\le n\le N-1,\\ \frac{1}{N}(\nu/\mu_2)^{n-N}(1-(\nu/\mu_2)^N) & \text{for } n\ge N.\end{cases} \tag{5}$$

From the theory of a standard $M/M/1$ queueing system, we observe that

$$T_2 = N\cdot(\text{ busy period of standard } M/M/1 \text{ queueing system})$$

in distribution, since in a cycle of $\{N_2(t), t\ge 0\}$, the process starts with N customers. Hence,

$$\begin{aligned}E[T_2] &= N\cdot E[\text{busy period of standard } M/M/1 \text{ queueing system}]\\ &= \frac{N/\mu_2}{1-(\nu/\mu_2)}\end{aligned} \tag{6}$$

Thus, from equations (4) and (6), we finally have

$$E[T]=\begin{cases}\frac{N/\nu}{1-(\mu_1/\nu)^N}\frac{\mu_2-\mu_1}{\mu_2-\nu}+\frac{\mu_1}{\nu(\mu_1-\nu)} & \text{if } \mu_1\neq\nu,\\ \frac{N+1}{2\nu}+\frac{1}{\mu_2-\nu} & \text{if } \mu_1=\nu.\end{cases} \tag{7}$$

3. Optimization

In this section, after assigning four operating costs to the system, we show that there exists a unique fast service rate minimizing the long-run average cost per unit time. Let $h(\mu)$ be the running cost per unit time with $h(0)=0$ while the server is working with service rate μ, and $f(x)$ be the setting cost to increase the service rate to $\mu_2=x$. We assume that $x\ge\mu_1>\nu$. We also assign the cost to the heavy traffic. Let $g(n)$ be the cost per unit time while $N(t)=N+n$, for $n=1,2,3,\ldots$. Finally, let $c\ge 0$ be the cost per unit time while the server is idle. This can be considered as a maintenance cost or the penalty for under-utilization of the server.

We assume that $h(\mu)$, $f(x)$, and $g(n)$ are all nonnegative increasing functions. We also assume that h and f are secondly differentiable convex functions which of course include the linearly increasing functions. We also assume that $\sum_{n=1}^{\infty} g(n)r^n<\infty$ for all r, $0\le r\le\nu/\mu_1<1$. For instance, if $g(n)$ is a nonnegative increasing polynomial of n, for $n\ge 1$, then $g(n)$ satisfies the assumptions.

We first calculate the long-run average cost per unit time when $\mu_2=x$. By the renewal reward theorem(Ross(1996), p.133), the long-run average cost per unit time is given by

$$\frac{E[\text{cost during } T]}{E[T]}. \tag{8}$$

From the results in section 2, we have

$$E[T]=a+p\frac{N}{x-\nu},$$

where

$$a=E[T_1]+E[T_3]=\frac{1}{\mu_1-\nu}-\frac{N/\nu}{(\mu_1/\nu)^N-1}+\frac{1}{\nu},$$

and p is given in equation (1). The expected costs during a cycle but the cost related to the heavy traffic are easily obtained as follows:

$$\begin{aligned}E[\text{running cost during } T] &= h(\mu_1)E[T_1]+h(x)pE[T_2]\\ &= h(\mu_1)\left(a-\frac{1}{\nu}t\right)+h(x)\frac{pN}{x-\nu},\end{aligned}$$

$$E[\text{setting cost during } T]=pf(x),$$

and

$$E[\text{cost for idle period during } T]=cE[T_3]=\frac{c}{\nu}.$$

The expected cost related to the heavy traffic can be calculated from equation (8) by noting that

$$\begin{aligned}&E[\text{cost for heavy traffic during } T]\\ &=E[\text{cost for heavy traffic per unit time}]\cdot E[T]\\ &=\sum_{n=0}^{\infty}g(n)P(N+n)\cdot E[T]\\ &=\frac{p}{x-\nu}\left\{1-\left(\frac{\nu}{x}\right)^N\right\}\sum_{n=1}^{\infty}g(n)\left(\frac{\nu}{x}\right)^n.\end{aligned}$$

Hence, we obtain $C(x)$, the long-run average cost per unit time when $\mu_2=x$, as follows:

$$C(x)=p\frac{(x-\nu)\left(f(x)+h(\mu_1)\frac{a-(1/\nu)}{p}+\frac{c}{p\nu}\right)+Nh(x)+B(x)}{a(x-\nu)+pN},\qquad(9)$$

for $x\geq\mu_1>\nu$, where

$$B(x)=\left\{1-\left(\frac{\nu}{x}\right)^N\right\}\sum_{n=0}^{\infty}g(n)\left(\frac{\nu}{x}\right)^n,\quad x\geq\mu_1>\nu.$$

We now show the uniqueness of the fast service rate minimizing $C(x)$. We need the following lemma:

Lemma

(a) $\lim_{x\to\infty}B(x)=0.$

(b) $\lim_{x\to\infty}B'(x)=0.$

(c) $\lim_{x\to\infty}(x-\nu)B''(x)=0.$

(d) $B''(x)\geq 0.$

We are ready to show the uniqueness of x minimizing $C(x)$. From equation (9), we have

$$C'(x)=p\frac{G_1(x)-G_2(x)}{\{a(x-\nu)+pN\}^2},$$

where

$$G_1(x) = (x-\nu)f(x)\{a(x-\nu)+pN\} + pNf(x) + \frac{cN}{\nu} + Nh'(x)\{a(x-\nu)+pN\} - aNh(x) + Nh(\mu_1)(a-1/\nu)$$

and $G_2(x) = aB(x) - B'(x)\{a(x-\nu)+pN\}$. Notice that $G_1(x)$ is increasing and $G_2(x)$ is decreasing. Hence, $G_1(x)-G_2(x)$ is increasing. It follows from Lemma that $\lim_{x\to\infty} G_2(x)=0$, and it can be shown that $\lim_{x\to\infty} G_1(x)\geq 0$. Therefore we have the following conclusions:

(i) When $G_1(\mu_1)\geq G_2(\mu_1)$, $C(x)$ is an increasing function for $x\geq\mu_1$ and hence minimized at $x=\mu_1$.

(ii) When $G_1(\mu_1)<G_2(\mu_1)$, $C(x)$ is minimized at $x=x^*>\mu_1$ and x^* is the unique solution of $G_1(x)=G_2(x)$.

References

Bae, J., Kim, S., and Lee, E.Y. (2002), A P_λ^M-Policy for an $M/G/1$ Queueing System, *Applied Mathematical Modelling*, to appear.

Bell, C.E. (1971), Characterization and Computation of Optimal Policies for Operating an $M/G/1$ Queuing System with Removable Server, *Operations Research*, 19, 208-218.

Federgruen, A. and Tijms, H.C. (1980), Computation of the Stationary Distribution of the Queue Size in an $M/G/1$ Queueing System with Variable Serve Rate, *Journal of Applied Probability*, 17, 515-522.

Heyman, D.P. (1968), Optimal Operating Policies for $M/G/1$ Queuing Systems, *Operations Research*, 16, 362-382.

Karlin, S. and Taylor, H.M. (1975), *A First Course in Stochastic Processes*, 2nd edition, Academic Press, New York.

Lee, E.Y. and Ahn, S.K. (1998), P_λ^M-Policy for a Dam with Input Formed by a Compound Poisson Process, *Journal of Applied Probability*, 35, 482-488.

Nishimura, S. and Jiang, Y. (1995), An $M/G/1$ Vacation Model with Two Service Modes, *Probability in the Engineering and Information Sciences*, 9, 355-374.

Ross, S.M. (1996), *Stochastic Processes*, 2nd edition, John Wiley & Sons, New York.

Sobel, M.J. (1969), Optimal Average-Cost Policy for a Queue with Start-up and Shut-down Costs, *Operations Research*, 17, 145-162.

Yadin, M. and Naor, P. (1963), Queueing Systems with a Removable Service Station, *Operational Research Quarterly*, 14, 393-405.

Yamada, K. and Nishimura, S. (1994), A Queueing System with a Setup Time for Switching of the Service Distribution, *Journal of the Operations Research Society of Japan*, 37, 271-286.