Regular Paper

Analysis of Input Control with Control Delay

AKIRA FUKUDA*, KEN-ICHI KOMAMIZU** and KOUICHI UTSUMIYA***

Many studies have been carried out on input regulation control in queueing models, which is a basic congestion control. Most of the studies assumed that there is no control delay, which means the time from when the control message starts to be sent until the effect of the control appears. However, in real systems, the control delay time cannot be neglected. This paper analyzes a two-level input control queueing model with control delay based on sampling monitoring. The model can deal with cases where the monitoring interval distribution is arbitrary.

Steady state probabilities are analyzed using piecewise Markov process theory and some performance measures are shown. Through several numerical results, the influences of control parameters such as the control delay, threshold values for control, and the monitoring interval distribution on the system performance, are investigated. In particular, we get the remarkable result that there is a mean control delay which gives a minimum loss probability for uncontrolled calls in some cases. Furthermore, from the numerical results, we conjecture that periodically monitoring gives a minimum loss probability for uncontrolled calls under the condition that the system and the control parameters are fixed.

1. Introduction

In computer networks and telecommunication systems such as telephone networks and packet switched networks, if demands to the system exceed the system capacity, congestion will occur. Once the system is in a congested state, system resources are ineffectively used, and throughput degradation occurs. To prevent congestion, it is effective to regulate excess traffic at the entrance of the system [1]. Thus, the input control scheme is a basic congestion control. Therefore, it is important to analyze and investigate input control schemes.

Many studies have been carried out on input regulation control in queueing models [2-8]. Most of the studies assumed that there is no control delay, which means the time from when the control message starts to be sent until the effect of the control appears. However, in real systems, the control delay time cannot be neglected. It consists of some time components. One of the major components is the time to deliver the control message sent at the congested node to the congestion control node. The other is the time elapsed until the influence of the traffic volume already within the network has disappeared, or until the traffic again appears at the network node [7]. There are a few studies of input control with control delay [7, 8]. These studies assumed that the state of the system is continuously monitored. However, in some real system cases input regulation control is usually activated by the information obtained by monitoring factors such as resource utilization at sampled time points. Thus, it is important to analyze and investigate an input control queueing model with control delay based on sampling monitoring.

The control described above should be finally evaluated in networks. When we analyze a network with the control, we should employ a medium/large size of network because:

•In a network, it seems that one of the most essential characteristics of the control is the congestion propagation. However, in a small network such as a two-nodes network, we cannot get the enough observation of the congestion propagation.

Employing a medium/large size of network, we have the following problems:

(1) It is difficult to exactly analyze a medium/large size of queueing network with the control.

(2) Network behavior is complicatedly influenced by many parameters such as network parameters (a network topology, source-destination pairs, traffic patterns, and so on), node parameters (the number of processors and waiting room capacity), and control parameters (threshold values, monitoring interval distributions, mean monitoring interval, and basic control delay times). This indicates that it is very hard work to clarify the control characteristics in the networks.

As the first step to clarify the control characteristics, we exactly analyze a single node (a single queueing

*Department of Computer Science and Communication Engineering, Faculty of Engineering, Kyushu University, 6-10-1 Higashi-ku, Fukuoka-shi, Fukuoka 812, Japan.
**NEC Corporation, Fuchu, Tokyo 183, Japan.
***Department of Computer Science Intelligent Systems, Oita University, Oita 870-11, Japan.

model) with the control, rather than a queueing network, and investigate basic characteristics of the control. We believe that this gives us a prospect of the control characteristics in the networks.

From the viewpoint of the networks, analyzing a single queueing model is based on the following assumption:

•For nodes each of which receives control signals from the particular node, loads are very light, almost negligible.

This paper analyzes exactly a two-level input control queueing model with control delay based on sampling monitoring, and investigates the influences of the control delay times on the system performance. The model can deal with cases where the monitoring interval distribution is arbitrary. By setting the control delay times to zero, the model becomes one proposed in [6].

Steady state probabilities are analyzed using piecewise Markov process theory and some performance measures are shown. Through several numerical results, the effectiveness of the proposed input control method is demonstrated. Furthermore, the influences of control parameters such as the control delay times, threshold values for control, and the monitoring interval distribution on the system performance, are investigated.

The model comes from a out-of-chain routing scheme in telephone networks [9]. The scheme permits calls which overflow from the final in-chain route under the existing routing scheme (far-to-near rotation scheme) to be offered to out-of-chain routes. With the scheme, out-of-chain routed calls are regulated based on the number of vacant facilities in a node. Furthermore, we can adapt the model to a congestion control scheme in telephone networks [10]. With the scheme, calls routed to a particular area are regulated.

2. Model Description

The proposed queueing model is shown in Fig. 1. The system has S processors and K-S waiting room size. Two streams of Poisson calls, which are named Call[1] and Call[2], respectively, are offered to the system. Their arrival rates are λ_1 and λ_2. The service times for both classes of calls are negative exponentially distributed with mean $1/\mu$. Calls are processed by means of a first-in first-out rule. If a call arrives when the waiting room is full, that call can not enter the system and is lost.

Call[2] is regulated to enter the system and is lost by the following two-level on/off control. Let ω_r be the r-th monitoring epoch, and ξ_r be the number of calls in the system at the epoch ω_r. Here, time intervals $(\omega_{r+1}-\omega_r)$ are assumed to be generally and independently distributed in accordance with a general distribution $F(t)$ with mean $1/\sigma$. A control message, that is a switch-on message, M-on, or a switch off message, M-off, is sent at the monitoring epoch ω_r by

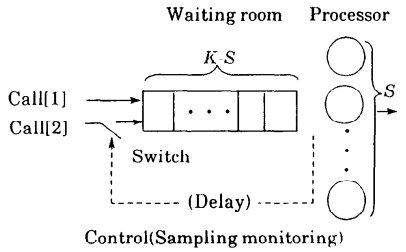


Fig. 1 Queueing model for input control with control delay.

the following control scheme:

$$\begin{cases}\text{If } 0\le\xi_r\le L, \text{ the control message } M\text{-on in sent,}\\ \text{If } L+1\le\xi_r\le H-1, \text{ any control message is not sent,}\\ \text{If } H\le\xi_r\le K, \text{ the control message } M\text{-off is sent,}\end{cases}$$

where L and H are thresholds for control.

When the control message arrives at the switch, the switch is set in the on- or off-position determined by the contents of the arriving message. Namely, if the arriving message is M-on (M-off), the switch is set in the on-position (the off-position).

The control delay times, (which means the times from when the control message starts to be sent until it arrives at the switch), are assumed to be negative exponentially distributed with mean $1/v$. If the previous control message is sending when the control message starts to be sent at the monitoring time point, the previous one is replaced with the current one. By setting the control delay times to zero, this model becomes one proposed in [6].

A two-level control mechanism is necessary to prevent control oscillation. This control type is called hysteresis control.

3. Model Analysis

Let $\xi(t)$ be the number of calls in the system and $C(t)$ be the indicator variable of the state of the switch defined as follows at the instant t:

$$C(t)=\begin{cases}0 \text{ if the switch is in the on-position,}\\ 1 \text{ if the switch is in the off-position.}\end{cases} \tag{1}$$

And, let $\zeta(t)$ be the indicator variable of the state of the control massage currently underway defined by

$$\zeta(t)=\begin{cases}0 \text{ if any control message is not being sent,}\\ 1 \text{ if the control message } M\text{-on is being sent,}\\ 2 \text{ if the control message } M\text{-off is being sent.}\end{cases} \tag{2}$$

The stochastic process $\{\eta(t)\}\equiv\{(\xi(t), C(t), \zeta(t))\}$ forms a piecewise Markov process [11] with the r-th monitoring epoch ω_r as a regeneration point. Therefore, the model can be analyzed in a manner similar to that reported by Kuczura [11].

3.1 State Probability at a Time Point Immediately before Monitoring Time Point

Let $\{\theta_{i,j,k}(l, m, n)\}$ be the regeneration matrix defined by

$$\theta_{i,j,k}(l, m, n) \equiv Prob(\eta(\omega_r+0)=(l, m, n) \mid \eta(\omega_r-0)=(i, j, k)). \quad (3)$$

By considering the control scheme described in Section 2, for $j=0, 1$, $k=0, 1, 2$, we have

$$\left.\begin{aligned} &\theta_{i,j,k}(i, j, 1)=1, && 0 \leqq i \leqq L, \\ &\theta_{i,j,k}(i, j, k)=1, && L+1 \leqq i \leqq H-1, \\ &\theta_{i,j,k}(i, j, 2)=1, && H \leqq i \leqq K, \\ &\theta_{i,j,k}(l, m, n)=0, && \text{otherwise.} \end{aligned}\right\} \quad (4)$$

Let $h_{i,j,k}(l, m, n)$ be the Markovian transition probabilities defined by

$$h_{i,j,k}(l, m, n) = Prob(\eta(\omega_r-0)=(l, m, n) \mid \eta(\omega_{r-1}+0)=(i, j, k)). \quad (5)$$

By considering the control scheme described in Section 2, $h_{i,j,k}(l, m, n)$ are obtained as follows: For $0 \leqq i, l \leqq K$, $j=0, 1$,

$$\left.\begin{aligned} &h_{i,j,k}(l, j, k)=\int_0^\infty P_{i,l}(t, \lambda_j)e^{-\nu t}dF(t), \quad k=1, 2, \\ &h_{i,j,j+1}(l, j, 0)=\int_0^\infty P_{i,l}(t, \lambda_j)(1-e^{-\nu t})dF(t), \\ &h_{i,j,2-j}(l, 1-j, 0)=\int_0^\infty \left\{\sum_{k=0}^{K}\int_0^t P_{i,k}(\tau, \lambda_j)\nu e^{-\nu\tau}) \times P_{k,l}(t-\tau, \lambda_{1-j})d\tau\right\} dF(t), \\ &h_{i,j,0}(l, j, 0)=\int_0^\infty P_{i,l}(t, \lambda_j)dF(t), \\ &h_{i,j,k}(l, m, n)=0, \quad \text{otherwise,} \end{aligned}\right\} \quad (6)$$

where $\lambda_0=\lambda_1+\lambda_2$, and $P_{i,l}(t, \lambda_j)$ are the transition probabilities for the M/M/S/K queueing model with arrival rate λ_j. They are given [12] by

$$P_{i,l}(t, \lambda_j)=p_l^{(j)}+\sum_{k=1}^{K} C_k^{(j)}(i, l)\exp(\gamma_k^{(j)}t). \quad (7)$$

See Appendix A for calculating $p_l^{(j)}$, $C_k^{(j)}(i, l)$, and $\gamma_k^{(j)}$. Substituting Eq. (7) into Eq. (6), for $0 \leqq i, l \leqq K$, $j=0, 1$,

$$\left.\begin{aligned} &h_{i,j,k}(l, j, k)=p_l^{(j)}\Psi(\nu)+\sum_{r=1}^{K} C_r^{(j)}(i, l)\Psi(\nu-\gamma_r^{(j)}), \quad k=1, 2, \\ &h_{i,j,j+1}(l, j, 0)=p_l^{(j)}\{1-\Psi(\nu)\}+\sum_{r=1}^{K} C_r^{(j)}(i, l)\{\Psi(-\gamma_r^{(j)})-\Psi(\nu-\gamma_r^{(j)})\}, \\ &h_{i,j,2-j}(l, 1-j, 0)=\sum_{k=0}^{K} p_k^{(j)}p_l^{(1-j)}(1-\Psi(\nu)) \\ &\qquad +\sum_{k=0}^{K}\sum_{r=1}^{K} C_r^{(1-j)}(k, l)p_k^{(j)}\nu\frac{\Psi(-\gamma_r^{(1-j)})-\Psi(\nu)}{\gamma_r^{(1-j)}+\nu} \\ &\qquad +\sum_{k=0}^{K}\sum_{r=1}^{K} C_r^{(j)}(i, k)p_l^{(1-j)}\nu\frac{1-\Psi(\nu-\gamma_r^{(j)})}{\nu-\gamma_r^{(j)}} \\ &\qquad +\sum_{k=0}^{K}\sum_{r=1}^{K}\sum_{n=1}^{K} C_r^{(j)}(i, k)C_n^{(1-j)}(k, l)\nu\frac{\Psi(\nu-\gamma_r^{(j)})-\Psi(-\gamma_n^{(1-j)})}{\gamma_r^{(j)}-\gamma_n^{(1-j)}-\nu}, \\ &h_{i,j,0}(l, j, 0)=p_l^{(j)}+\sum_{r=1}^{K} C_r^{(j)}(i, l)\Psi(-\gamma_r^{(j)}), \\ &h_{i,j,k}(l, m, n)=0, \quad \text{otherwise,} \\ &\qquad \gamma_r^{(1-j)}+\nu \neq 0, \quad \nu-\gamma_r^{(j)} \neq 0, \quad \gamma_r^{(j)}-\gamma_n^{(1-j)}-\nu \neq 0, \end{aligned}\right\} \quad (8)$$

where $\Psi(s)$ is the Laplace-Stieltjes transform of $F(t)$:

$$\Psi(s)=\int_0^\infty e^{-st}dF(t). \quad (9)$$

Let $r_{i,j,k}(l, m, n)$ be the one-step transition probabilities of the preregeneration chain defined by

$$r_{i,j,k}(l, m, n) \equiv Prob\{\eta(\omega_r-0)=(l, m, n) \mid \eta(\omega_{r-1}-0)=(i, j, k)\}. \quad (10)$$

From Eqs. (3) and (5), $r_{i,j,k}(l, m, n)$ defined by Eq. (10) is rewritten as follows:

$$r_{i,j,k}(l, m, n)=\sum_{n_3=0}^{2}\sum_{n_2=0}^{1}\sum_{n_1=0}^{L}\theta_{i,j,k}(n_1, n_2, n_3) \times h_{n_1,n_2,n_3}(l, m, n). \quad (11)$$

Substituting Eq. (4) into Eq. (11), for $0 \leqq l \leqq K$, $j, m=0, 1$, and $k, n=0, 1, 2$

$$\left.\begin{aligned} &r_{i,j,k}(l, m, n)=h_{i,j,1}(l, m, n), && 0 \leqq i \leqq L, \\ &r_{i,j,k}(l, m, n)=h_{i,j,k}(l, m, n), && L+1 \leqq i \leqq H-1, \\ &r_{i,j,k}(l, m, n)=h_{i,j,2}(l, m, n), && H \leqq i \leqq K. \end{aligned}\right\} \quad (12)$$

$r_{i,j,k}(l, m, n)$ can be evaluated using Eqs. (8) and (12).

Let $\pi_{i,j,k}$ be the steady state probabilities of the chain imbedded at points immediately before the monitoring time point:

$$\pi_{i,j,k} \equiv \lim_{r\to\infty} Prob(\eta(\omega_r - 0) = (i, j, k)). \tag{13}$$

The distribution $\{\pi_{i,j,k}\}$ can be found using the following equations.

$$\left.\begin{aligned} \pi_{i,j,k} &= \sum_{l=0}^{K}\sum_{m=0}^{1}\sum_{n=0}^{2} \pi_{l,m,n} r_{l,m,n}(i, j, k), \\ \sum_{i=0}^{K}\sum_{j=0}^{1}\sum_{k=0}^{2} \pi_{i,j,k} &= 1. \end{aligned}\right\} \tag{14}$$

The linear system of equations can be solved by means of an ordinary numerical calculation method.

3.2 State Probability at an Arbitrary Time Point

Let $\{q_{i,j,k}\}$ be the stationary distribution at arbitrary time points defined by

$$q_{i,j,k} \equiv \lim_{t\to\infty} Prob(\eta(t) = (i, j, k)). \tag{15}$$

Using the rate conservation principle [11] in the piecewise Markov process, the following state equations can be obtained for $j=0, 1$, $k=0, 1, 2$,

$$\left.\begin{aligned} &\{\lambda_j + \mu_i + (\delta_{k,1}+\delta_{k,2})\nu\} q_{i,j,k} + (\delta_{k,0}+\delta_{k,2})\sigma\pi_{i,j,k} = \lambda_j q_{i-1,j,k} + \mu_{i+1} q_{i+1,j,k} + \delta_{j,0}\delta_{k,0}\nu(q_{i,0,1}+q_{i,1,1}) \\ &\qquad + \delta_{j,1}\delta_{k,0}\nu(q_{i,0,2}+q_{i,1,2}) + \delta_{k,1}\sigma(\pi_{i,j,0}+\pi_{i,j,2}), \quad 0 \leqq i \leqq L, \\ &\{\lambda_j + \mu_i + (\delta_{k,1}+\delta_{k,2})\nu\} q_{i,j,k} = \lambda_j q_{i-1,j,k} + \mu_{i+1} q_{i+1,j,k} + \delta_{j,0}\delta_{k,0}\nu(q_{i,0,1}+q_{i,1,1}) \\ &\qquad + \delta_{j,1}\delta_{k,0}\nu(q_{i,0,2}+q_{i,1,2}), \quad L+1 \leqq i \leqq H-1, \\ &\{(1-\delta_{i,K})\lambda_j + \mu_i + (\delta_{k,1}+\delta_{k,2})\nu\} q_{i,j,k} + (\delta_{k,0}+\delta_{k,1})\sigma\pi_{i,j,k} = \lambda_j q_{i-1,j,k} + \mu_{i+1} q_{i+1,j,k} + \delta_{j,0}\delta_{k,0}\nu(q_{i,0,1}+q_{i,1,1}) \\ &\qquad + \delta_{j,1}\delta_{k,0}\nu(q_{i,0,2}+q_{i,1,2}) + \delta_{k,2}\sigma(\pi_{i,j,0}+\pi_{i,j,1}), \quad H \leqq i \leqq K, \end{aligned}\right\} \tag{16}$$

where

$$\mu_i = \begin{cases} min(i, S)\mu & 0 \leqq i \leqq K, \\ 0 & i = K+1, \end{cases} \tag{17}$$

$$\delta_{i,j} = \begin{cases} 0 & i \neq j, \\ 1 & i = j. \end{cases} \tag{18}$$

From the state diagram, which is not shown in this paper for lack of space, $q_{i,j,k}$ are divided into two groups. One group is for $0 \leqq i \leqq K$ and $(j, k) = (0, 0), (0, 1), (1, 1)$. The other is for $0 \leqq i \leqq K$ and $(j, k) = (0, 2), (1, 0), (1, 2)$. The number of independent equations in Eq. (16) are variables $q_{i,j,k}$ for each group is $3(K+1)-1$ and $3(K+1)$, respectively. Therefore, $q_{i,j,k}$ can not be obtained from only Eq. (16). Another equation for each group is needed. Since the state which consists of $0 \leqq i \leqq K$ and $(j, k) = (0, 0), (0, 1), (1, 1)$, does not change for consecutive monitoring time points, we have

$$\sum_{i=0}^{K} q_{i,0,0} + q_{i,0,1} + q_{i,1,1} = \sum_{i=0}^{K} \pi_{i,0,0} + \pi_{i,0,1} + \pi_{i,1,1}. \tag{19}$$

Similarly

$$\sum_{i=0}^{K} q_{i,0,2} + q_{i,1,0} + q_{i,1,2} = \sum_{i=0}^{K} \pi_{i,0,2} + \pi_{i,1,0} + \pi_{i,1,2}. \tag{20}$$

$q_{i,j,k}$ can be evaluated using Eqs. (16), (19) and (20).

3.3 Performance Measures

(1) Loss probability for Call[2] at the switch, $B_{2,SW}$:

$$B_{2,SW} = \sum_{i=0}^{K}\sum_{j=0}^{2} q_{i,1,j}. \tag{21}$$

(2) Loss probability for Call[2] at the system, $B_{2,SYS}$:

$$B_{2,SYS} = \sum_{i=0}^{2} q_{K,0,i}. \tag{22}$$

(3) Loss probability for Call[1], B_1, for Call[2], B_2:

$$B_1 = \sum_{i=0}^{1}\sum_{j=0}^{2} q_{K,i,j}, \tag{23}$$

$$B_2 = B_{2,SW} + B_{2,SYS}. \tag{24}$$

(4) Mean waiting time for Call[1], W_1, for Call[2], W_2:

$$W_1 = \frac{1}{S\mu}\sum_{i=S}^{K-1}(i-S+1)\left(\sum_{j=0}^{1}\sum_{k=0}^{2} q_{i,j,k}\right) \Big/ (1-B_1), \tag{25}$$

$$W_2 = \frac{1}{S\mu}\sum_{i=S}^{K-1}(i-S+1)\left(\sum_{j=0}^{2} q_{i,0,j}\right) \Big/ (1-B_2). \tag{26}$$

(5) Regulation control probability, R:

R is the probability that the system is in the regulation state, that is the switch is in the off-position. R is given as

$$R = B_{2,SW}. \tag{27}$$

(6) Regulation control frequency, F:

F is the average number of transitions from off-position to on-position or from on-position to off-position of the switch per unit time. F is given as

$$F = \nu\left(\sum_{i=0}^{K} q_{i,1,1} + \sum_{i=0}^{K} q_{i,0,2}\right). \tag{28}$$

Under the steady state, the first term is equal to the second term in the right-hand side of Eq. (28).

4. Numerical Results and Discussions

Through all numerical results, the parameters used are $K=10$, $S=5$, $\mu=1.0$, and $\lambda_1=1.0$. Through Figs. 2–6, the monitoring interval distribution is the unit distribution (D).

(1) Effectiveness of the input control

Figure 2 shows the loss probabilities as a function of λ_2 for Calls[1], B_1, and for Calls[2], B_2, both at the switch, $B_{2,SW}$, and at the entrance of the system, $B_{2,SYS}$. The parameters used in Fig. 2 are $(L, H)=(3, 7)$, $\sigma=1.0$, and $v=2.0$. For comparison, the loss probability without the control is shown, where $B_1=B_2$. The loss probability for Calls[1], B_1, with the control is smaller than that without the control. A high percentage of the lost Calls[2] are rejected at the switch. Therefore, the control is effective.

(2) Influence of mean control delays on system performance

Figures 3(a), 3(b), and 3(c) show loss probability for Calls[1], regulation control probability, and mean waiting time for Calls[1], respectively, in the case where $(L, H)=(4, 5)$.

Figure 3(a) implies that the optimum mean control delay, which gives a minimum B_1, exists when the mean monitoring interval is not too short and the traffic is heavy. This phenomenon is confirmed by additional examples, which are not shown for lack of space. We show a possibility that the phenomenon occurs as follows.

Let us consider what happens in the cases where the traffic of Call[2] is heavy and the mean control delay, $1/\nu$, is (a) short, (b) medium, or (c) long by typical examples in Figs. 4(a)–(c), where $H=L+1$ for simplicity.

In case (a), shown in Fig. 4(a), the control message M-off is sent to the switch at the monitoring time point τ_1, because the number of calls in the system at τ_1 exceeds the threshold L. The message M-off arrives at the switch after the control delay $1/\nu$, and the switch turns into the off-position at the time t_1. The message M-on is sent to the switch at τ_2, because the control delay is short and the time interval τ_2-t_1 is long enough for the number of calls in the system to become lower than L at τ_2. Next, the switch turns into the on-position at t_2. These phenomena are repeated. That is, the switch turns into the off/on/off/on/off-positions at t_1, t_2, t_3, t_4, t_5, respectively. In this case, each cycle of control consists of two states of the switch, one on and one off. Therefore, the regulation control probability $R=1/(1+1)=1/2$ in Fig. 4(a).

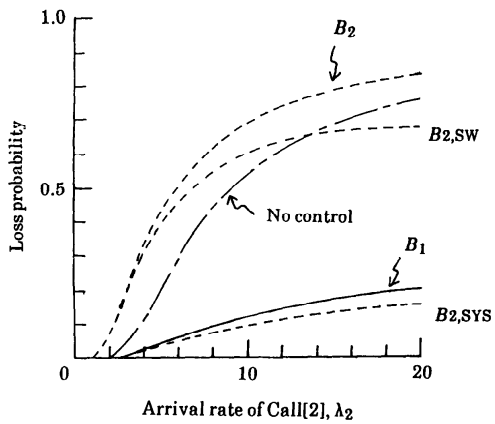


Fig. 2 Loss probability vs. arrival rate of Call[2].

Let us consider case (b), shown in Fig. 4(b), where the control delay is medium. The message M-off is sent to the switch at τ_2, because the time interval τ_2-t_1 is not long enough for the number of calls in the system to become lower than L at τ_2. In this case, the switch is in the off/off/on/off/off-positions during the time periods (t_1, t_2), (t_2, t_3), (t_3, t_4), (t_4, t_5), (t_5, t_6), respectively. There are one on-state and two off-states of the switch in each cycle. Therefore, $R=2/(1+2)=2/3$ in Fig. 4(b).

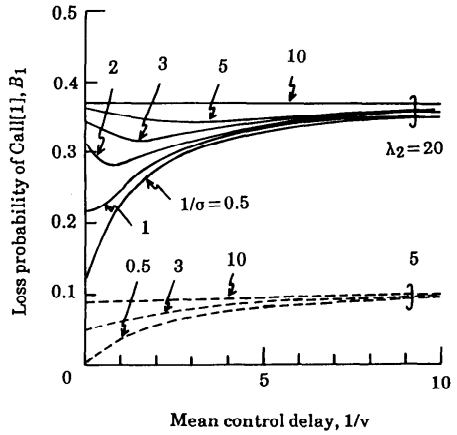
In case (c), shown in Fig. 4(c), the switch has the same on-off position pattern as case (b) during a time period from t_1 to t_4. However, the message M-on is sent to the switch at τ_4, because the time interval τ_4-t_3 is not long enough for the number of calls in the system to exceed L at τ_4. In this case, the switch is in the off/off/on/on/off-positions during the time periods (t_1, t_2), (t_2, t_3), (t_3, t_4), (t_4, t_5), (t_5, t_6), respectively. There are two on-states and two off-states in each cycle. Therefore, $R=2/(2+2)=1/2$ in Fig. 4(c).

As shown above, the value of R is greatest for case (b), which is confirmed by Fig. 3(b). The example in Fig. (4) is one in case where the system behavior is not stochastic. However, the similar discussion described above seems to be applied in a case where it is stochastic. It is likely that a large number of Calls[1] can enter the system at the neighborhood of point where the number of Calls[2] rejected by the control is greatest, namely where the value of R is greatest. Above discussion means that the optimum mean control delay, which gives a minimum B_1, exists when the mean monitoring interval is not too short and traffic is heavy.

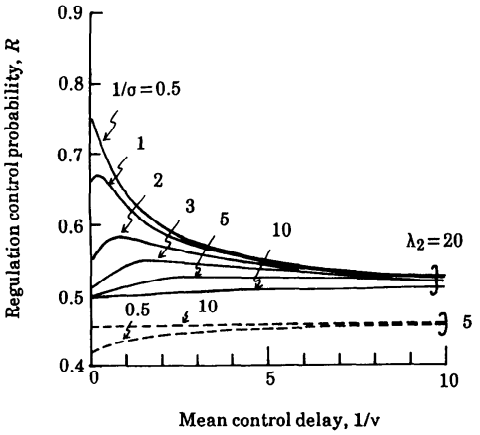
The discussion described by using Fig. 4 suggests that R has a local maximum value in the case where the mean control delay time is greater than the mean monitoring interval. However, Figure 3(b) shows that it is not, which is confirmed by additional examples.

Although the minimum control delay time may be fixed after building a network, the control delay time to be implemented into a real system can be modified in the case where the optimum control delay is grater than the minimum one. In this case, the control delay is set to be the optimum one.

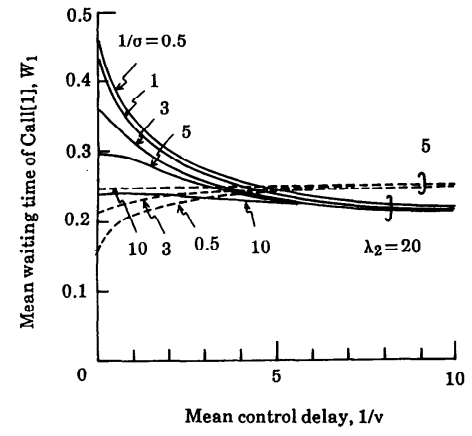
It is difficult to derive a formula of the optimum control delay, because the system behavior is stochastic and a large number of parameters such as the arrival rates of Calls[1] and Calls[2], the thresholds, the mean



(a) Loss probability of Call[1]

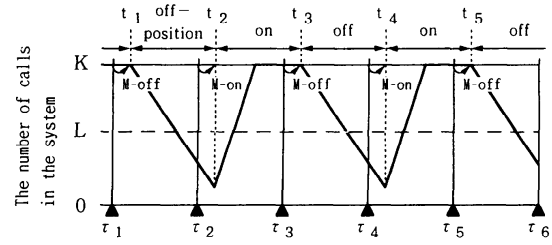


(b) Regulation control probability

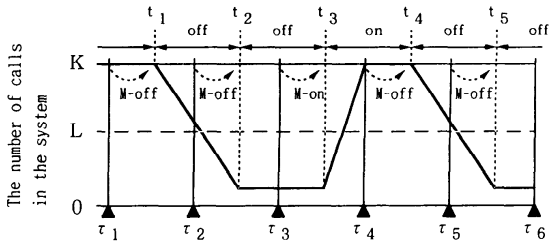


(c) Mean waiting time of Call[1]

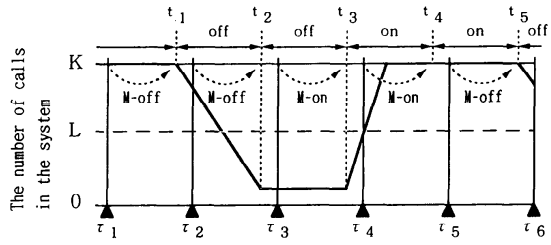
Fig. 3 System performance vs. mean control delay.



(a) The control delay is short.



(b) The control delay is medium.



(c) The control delay is long.

▲ τ_i : monitoring time point

t_i : control message arriving time point at the switch

Fig. 4 Examples in the cases where the control delay is (a) short, (b) medium, or (c) long.

monitoring interval, etc., can affect the system performance. It is one of further studies to derive an approximate formula for the optimum control delay.

In Fig. 3(c), the slopes of W_1 are in opposite directions for light and heavy traffic. The reason is explained by Fig. 5. Figure 5 shows $W_{1,on}$, $W_{1,off}$, and a probability Q in the case where $1/\sigma=1$ in Fig. 3. In the case of $\lambda_2=5$, Fig. 3(c) does not show the slope of W_1 under $1/\sigma=1$ for convenience of illustration. The values of W_1 under $1/\sigma=1$ are between that under $1/\sigma=0.5$ and that under $1/\sigma=3$. The values of $W_{1,on}$ and $W_{1,off}$ mean the mean waiting time for Calls[1] entering the system when the switch is in the on-position and in the off-position, respectively. The value of Q means a ratio of the number of Calls[1] entering the system when the switch is in the off-position to the number of Calls[1] entering the system. $W_{1,on}$, $W_{1,off}$, and Q are given in Appendix B. A relationship between W_1, $W_{1,on}$, $W_{1,off}$, and Q is given by Eq. (B·4) in Appendix B.

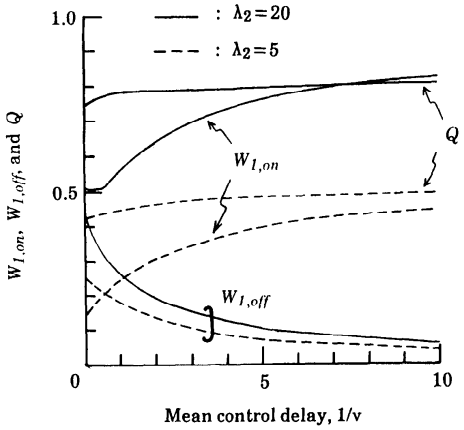


Fig. 5 Two conditional mean waiting times for Call[1], $W_{1,on}$, $W_{1,off}$, and a ratio, Q vs. mean control delay.

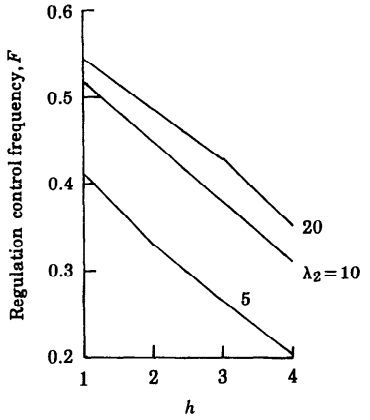


Fig. 6 Regulation control frequency vs. differences between two thresholds.

As shown in Fig. (5), $W_{1,on}(W_{1,off})$ is an increasing (decreasing) function of the mean control delay, because the system in the on-position (off-position) of the switch is congested (is not congested) as the control delay increases. The value of Q under $\lambda_2=20$ is greater than that under $\lambda_2=5$, because the probability that the switch is in the off-position under $\lambda_2=20$, is greater than that under $\lambda_2=5$. From the above discussion and Eq. (B·4) in Appendix B, the line of W_1 slopes upwards for light traffic and downwards for heavy traffic.

(3) Effectiveness of the hysteresis control

Figure 6 shows the regulation control frequency, F, as a function of h. The values of a lower threshold, L, and a higher threshold, H, are set to satisfy the equations $L=S-h$, $H=S+h$, respectively. The parameters used in Fig. 6 are $\sigma=1.0$ and $v=2.0$. Obviously, F decreases as h increases. Thus, the hysteresis control is effective to reduce the control switching frequnecy. However, it does not mean that a control under which F decreases is good. It should be considered from various viewpoints of system performance.

(4) Influence of the monitoring interval distributions on system performance

Figure 7 shows B_1 in the case where the monitoring interval distributions are the unit distribution(D), exponential distribution(M), and 2-stage hyperexponential distribution(H_2). The parameters used in Fig. 7 are $(L, H)=(3, 7)$ and $\sigma=1.0$. To determine the distribution H_2, a symmetric condition [14] is used, and the value of the coefficient of variation is set at 2.0. The value of the coefficient of variation becomes smaller in the order of the distributions H_2, M, D [15], that is, it is smallest in D. B_1 decreases as the value of the coefficient of variation becomes smaller. The same phenomenon, confirmed by additional examples, is shown through numerical results with no control delay in [6]. Therefore, we conjecture that periodically monitoring(D) gives a minimum loss probability for Calls[1] under the condition that the system and the control parameters (mean monitoring interval, mean control delay, and threshold values for control) are fixed.

5. Conclusions

A two-level input control model with control delay based on sampling monitoring has been analyzed and discussed. Performance measurements have been calculated from the steady state probabilities. This model can deal with cases where the monitoring interval distribution is arbitrary. Through several numerical results, the effectiveness of the proposed input control method is demonstrated. Furtheremore, the influence of the control parameters such as the control delay, the threshold values for control, and the monitoring interval distribution on the system performance, are investigated numerically. In particular, we get the remarkable result that there is a mean control delay which gives a minimum loss probability for uncontrolled calls, Calls[1], under heavy traffic conditions. Furthermore from the numerical results, we conjecture that periodical monitoring(D) gives a minimum loss probability for Calls[1] under the condition that the system and the control parameters are fixed.

This paper mainly focuses on the characteristics for uncontrolled calls. In case where the control is implemented into real systems, the values of the control parameters should be decided after general consideration of total characteristics of the system.

Further studies in progress are:

i) to analyze a model where the control delay times are deterministically distributed,

ii) to prove the conjecture that periodical monitoring gives a minimum loss probability for uncontrolled calls under the condition that the system and the control parameters are fixed,

iii) to derive an approximate formula for the optimum

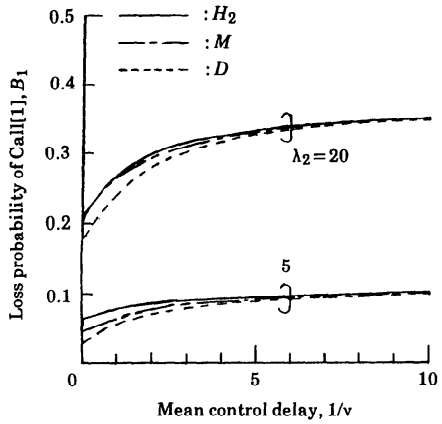


Fig. 7 Loss probability of Call[1] vs. mean control delay.

control delay giving a minimum loss probability for uncontrolled calls under heavy traffic conditions, and
iv) to extend the model to a network model.

References

1. GERLA, M. and KLEINROCK, L. Flow Control: A Comparative Survey, *IEEE Trans. Commun.*, **COM-28**, 4 (1980), 553-574.
2. MATSUMOTO, J. and MORI, H. Flow Control in Packet-switched Networks by Gradual Restrictions of Virtual Calls, *IEEE Trans. Commun.*, **COM-29**, 4 (1981), 466-473.
3. HSIA, W. S. and SCOTT, M. A Finite Queue with Some Control on Service and Arrival Rates, Cahiers du C.E.R.O., **25**, 1-2 (1983), 129-141.
4. FUKUDA, A. Analysis of Input Control Based on Discrete Monitoring in Mixed Input Queueing Model, *Trans. IECE Japan*, **J67-B**, 12 (1984), 1339-1346.
5. FUKUDA, A. Input Control Based on Sampling Monitoring Using Call Gapping Control in Mixed Input Queueing Model, *Trans. IECE Japan*, **J68-B**, 9 (1985), 953-961.
6. KAWASHIMA, K. Queueing Analysis for Input Regulation Method Employing Periodic Monitoring and Control, *Eur. J. Oper. Res.*, **23** (1986), 100-107.
7. VAN AS H. R. Transient Analysis of Markovian Queueing Systems and Its Application to Congestion-control Modeling, *IEEE J. Select. Areas Commun.*, **SAC-4**, 6 (1986), 891-904.
8. MIRCHANDANEY, R., TOWSLEY, D. and STANKOVIC, J. Analysis of the Effects of Delays on Load Sharing, *IEEE Trans. Comput.*, **38** (1989), 1513-1525.
9. FUKUDA, A. Control Algorithms for Out-of-Chain Routing in a Telephone Network, *IECE Trans.*, **J67-B** (1984) (in Japanese), 113-120.
10. NAKAJIMA, S. A Design Method of Control Parameters for a Code Blocking Control System on a Telephone Network, *IECE Trans.*, **J66-B**, 7 (1983) (in Japanese), 837-844.
11. KUCZURA, A. Piecewise Markov Processes, *SIAM J. Appl. Math.*, **24**, 2 (1973), 169-181.
12. RIORDAN, J. Stochastic Service Systems, John Wiley Inc., New York (1962).
13. MACHIHARA, F. Transition Probabilities of Markovian Service System and Their Application, *Rev. of E.C.L.*, **29**, 3-4 (1981), 155-169.
14. MORSE, P. M. Queues Inventories and Maintenance, John Wiley Inc., New York (1958).
15. KLEINROCK, L. Queueing Systems Vol. 1, John Wiley, New York (1975).

(Received October 24, 1990; revised July 1, 1991)

Appendix A

Transition probabilities for the M/M/S/K queueing model [13].

Let λ and μ be the arrival rate and the service rate in the M/M/S/K queueing model, respectively. In particular, for convenience, we use λ and $P_{i,l}(t)$ rather than λ_j and $P_{i,l}(t, \lambda_j)$, respectively, which are used in section 3. Let $P_{i,l}(t)$ be the probability that the system is in state l at time t, given that it was in state i at time zero, where the state is defined as the number of calls in the system.

$P_{i,l}(t)$ are given by

$$P_{i,l}(t)=p_l+\sum_{k=1}^{K} C_k(i,l)\exp(\gamma_k t), \quad 0\leqq i, l\leqq K, \quad (A\cdot 1)$$

where p_l are steady state probabilities for the M/M/S/K queueing model;

$$p_l=\begin{cases}(a^l/l!)p_0, & 1\leqq l\leqq S,\\ (a^S/S!)\rho^{l-S}p_0, & S+1\leqq l\leqq K,\end{cases} \quad (A\cdot 2)$$

$$p_0=\left(\sum_{n=0}^{S}(a^n/n!)+(a^S/S!)\sum_{n=1}^{K-S}\rho^n\right)^{-1}, \quad (A\cdot 3)$$

where $a=\lambda/\mu$ and $\rho=a/S$.

$C_k(i,l)$ is given by

$$C_k(i,l)=(D_i(\gamma_k)D_l(\gamma_k)/D_K(\gamma_k)D'_{K+1}(\gamma_k))\lambda^{K-i}\prod_{n=l+1}^{K}\mu_n, \quad (A\cdot 4)$$

where $\mu_i=\min(i,S)\mu$ for $1\leqq i\leqq K$ and γ_k are eigenvalues of the following infinitesimal generator A;

$$A=\begin{bmatrix}-\alpha_0 & \mu_1 & & & 0\\ \lambda & -\alpha_1 & \mu_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \\ 0 & & & \lambda & -\alpha_{K-1} & \mu_K\\ & & & & \lambda & -\alpha_K\end{bmatrix} \quad (A\cdot 5)$$

where $\alpha_0=\lambda$, $\alpha_i=\lambda+\mu_i$ for $1\leqq i\leqq K-1$, and $\alpha_K=\mu_K$.

The eigenvalues γ_k are calculated with ordinary numerical methods. A effective numerical calculation method of the eigenvalues is shown in [13].

The $D_n(s)$ in Eq. (A·4) is an eigenpolynominal for the matrix A, given by

$$\left.\begin{aligned}&D_0(s)=1,\\ &D_1(s)=s+\alpha_0,\\ &D_{n+1}(s)=(s+\alpha_n)D_n(s)-\lambda\mu_n D_{n-1}(s), \quad 1\leqq n\leqq K,\end{aligned}\right\} \quad (A\cdot 6)$$

and $D'_n(s)$ is its derivative;

$$\left.\begin{aligned}&D'_0(s)=0,\\ &D'_1(s)=1,\\ &D'_{n+1}(s)=D_n(s)+(s+\alpha_n)D'_n(s)-\lambda\mu_n D'_{n-1}(s),\\ &1\leqq n\leqq K.\end{aligned}\right\} \quad (A\cdot 7)$$

Appendix B

Expression of $W_{1,on}$, $W_{1,off}$, and Q.

$W_{1,on}$, $W_{1,off}$, and Q are given by

$$W_{1,on}=\frac{1}{S_\mu}\sum_{i=S}^{K-1}(i-S+1)\left(\sum_{k=0}^{2}q_{i,0,k}\right)\Bigg/\sum_{i=0}^{K-1}\sum_{k=0}^{2}q_{i,0,k}, \quad (B\cdot 1)$$

$$W_{1,off}=\frac{1}{S_\mu}\sum_{i=S}^{K-1}(i-S+1)\left(\sum_{k=0}^{2}q_{i,1,k}\right)\Bigg/\sum_{i=0}^{K-1}\sum_{k=0}^{2}q_{i,1,k}, \quad (B\cdot 2)$$

$$Q=\sum_{i=0}^{K-1}\sum_{k=0}^{2}q_{i,1,k}\Bigg/\sum_{i=0}^{K-1}\sum_{j=0}^{1}\sum_{k=0}^{2}q_{i,j,k}. \quad (B\cdot 3)$$

A relationship between W_1, $W_{1,on}$, $W_{1,off}$, and Q is given by

$$W_1=W_{1,on}(1-Q)+W_{1,off}Q. \quad (B\cdot 4)$$