Performance Analysis of a Computer Communication System with Time-out Scheme

Kentaro Hoshi[1], Sumito Iijima[1], Yoshitaka Takahashi[2], and Naohisa Komatsu[1]

[1]Faculty of Fundamental Science and Engineering, Waseda University, 3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555, Japan

[2]Faculty of Commerce, Waseda University, 1-6-1, Nishi-waseda, Shinjuku, Tokyo 169-8050, Japan

Motivated by performance evaluation of a computer communication system, we consider a renewal input, general service time, single-server, and infinite-capacity queuing system with generally distributed time-out threshold. We obtain two-moment approximate formulas for the mean system performance measures (including the mean number of customers in the system and the mean response time) by using the diffusion process with the reflecting barrier. Our tele-traffic model includes the standard G/G/1 system without time-out scheme as a special case. From performance comparisons between our diffusion approximation and simulation results, we also provide refining formulas.

***Index Terms*— The G/G/1 queue, time-out scheme, tele-traffic analysis, performance evaluation, diffusion approximation**

I. Introduction

QUEUEING models with time-out schemes are frequently encountered in computer communication systems. By the *time-out scheme* we mean that a customer arriving at a service system has to leave the system when its waiting time reaches a pre-assigned time limitation. This pre-assigned time limitation will be referred to as time-out threshold and denoted by Γ. The customer will receive the service if it's waiting time is less than Γ. The customer will be rejected if its waiting time reaches Γ.

For instance, in a telephone system we can see a situation where a call whose waiting time reaches the time-out threshold Γ will be rejected by a switching node. In a computer network an incoming packet to a processor buffer can be also rejected due to a time out scheme. In a recently-developed web service system, a user's session connection can also be cut off due to a time-out scheme; see Sery & Beale [17] for HTTP server operation.

We use the following queueing notation originally introduced by Kendall: A/B/c-T, where A stands for the arrival process, B the service time distribution, c the number of servers, and the last (-T) stands for the time-out threshold distribution. For example, M/M/1-D signifies a Poisson arrival (exponential inter-arrival time), exponential service time, single-server system with deterministic time-out threshold. E_2/G/1-M signifies a two-stage Erlang arrival, general service time, single-server system with exponential time-out threshold.

There has been much interest in exact approaches for the queueing models with time-out schemes. Barrer [3,4] presented the M/M/1-D system, Finch [6] treated the G/M/1-D system, and Rao [16] analyzed the M/G/1-M system. Stanford [18] formulated the G/G/1-G system. However, these existing results contain complicated numerical calculations including integrals, and it is so hard to calculate a performance measure, e.g., the mean number of customers in the system via these previous results.

The goal of this paper is to present an approximate formula on the mean performance measures (including the mean number of customers in the system and the mean waiting time) for the G/G/1-G system.

The rest of this paper is organized as follows. Section II describes our queueing model in details. We introduce our stochastic assumptions and key notations for the G/G/1-G system. In Section III we approximate the queue-length process by a diffusion process with a reflecting barrier (RB) as in Heyman [8] who treated the standard G//G/1 system without any time-out scheme. We determine the diffusion parameters arising out of the diffusion (Fokker-Planck) equation, which is an essential part of this paper. We then derive a simple approximate formula for the mean system performance measure. The derived formula thru the diffusion approximation is seen to be positive even if the offered traffic is zero (arrival rate λ = 0). Thus we propose a refined approximate formula which is reduced to zero whenever the offered traffic is zero. Section VI verifies the accuracy of our proposed formulas thru the simulation results. In Section V, we summarize our results and mention future works to conclude our remarks.

II. Our Tele-Traffic Model

Let X be an independent, and identically distributed (iid) random variable (rv). The mean, variance, and squared coefficient of variation (cv) for rv X are respectively denoted by E(X), σ_X^2, and C_X^2. We have by definition

$$C_X^2 = \frac{\sigma_X^2}{E(X)^2} \tag{2.1}$$

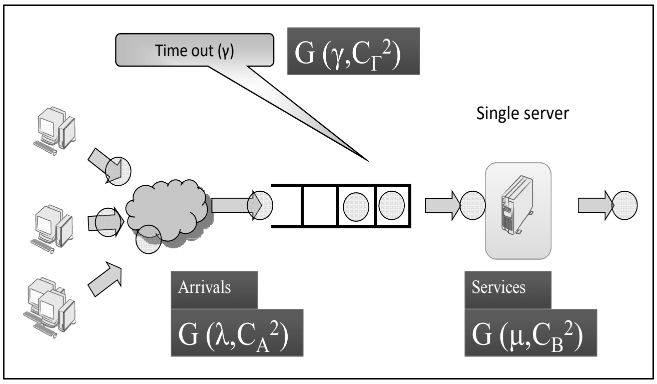


Fig. 1 G/G/1-G queuing system with time-out scheme.

We consider a stochastic service system assuming the followings; see Figure 1.

i) Customers arrive independently each other at a single-server system.

ii) The inter-arrival time of the customers is an iid rv, A. The arrival rate is denoted by λ. We then have

$$\lambda = \frac{1}{E(A)} \tag{2.2}$$

iii) The service time of a customer is an iid rv, B. The service rate is denoted by μ. We also have

$$\mu = \frac{1}{E(B)} \tag{2.3}$$

iv) The capacity of the waiting room is infinite.

v) A customer has to leave the system (or to be rejected) whenever its waiting time reaches a time-out threshold Γ, which is an iid rv. The time-out threshold rate is denoted by γ. Also, we have

$$\gamma = \frac{1}{E(\Gamma)} \tag{2.4}$$

It should be noted that if we let the time-out threshold be infinity ($\Gamma = \infty$) the time-out scheme model is reduced to the standard queueing model without any time-out In other words, the standard model is a very special case of the time-out scheme.

We define the traffic intensity by

$$\rho = \frac{\lambda}{\mu} \tag{2.5}$$

III. Tele-traffic Analysis

1) The Probability Density Function

We approximate the queue-length process (the stochastic process generated by the number of customers in the system at time t) { N(t) | t ≧ 0 } by a diffusion process with RB (Reflecting Barrier) boundary as in Heyman[8]. To be more exact, if we denote by f(x,t) the probability density function (pdf) of N(t):

$$f(x,t)dx \equiv P_r\{x < N(t) < x + dx\}, \tag{3.1}$$

the pdf f(x,t) satisfies the forward Kolmogorov (Fokker-Planck) equation:

$$\frac{\partial}{\partial t} f = -a \frac{\partial}{\partial x} f + \frac{b}{2} \frac{\partial^2}{\partial x^2} f \tag{3.2}$$

subject to the RB boundary condition:

$$\lim_{x \to 0} \left[-af + \frac{b}{2} \frac{\partial}{\partial x} f \right] = 0 \tag{3.3}$$

Here, *a* and *b* be the infinitesimal mean and variance of the process, called as the diffusion parameters in short which will be determined later.

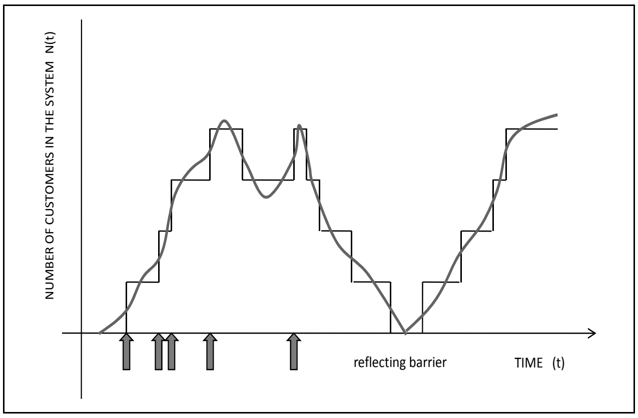


Fig. 2 Modeling via diffusion equation.

Assume the steady state from now on. We denote the steady-state pdf by

$$f(x) = \lim_{t \to \infty} f(x,t) \tag{3.4}$$

The diffusion equation and RB boundary condition becomes the following equations:

$$0 = -a \frac{d}{dx} f + \frac{b}{2} \frac{d^2}{dx^2} f \tag{3.5}$$

$$\lim_{x \to 0} \left[-af + \frac{b}{2} \frac{d}{dx} f \right] = 0 \tag{3.6}$$

Solving the equation (3.5) under the RB condition (3.6) for f(x), we straightforwardly have:

$$f(x) = -\frac{2a}{b} e^{\frac{2a}{b}x} \tag{3.7}$$

2) *The Mean Performance Measures*

The mean number of customers in the steady state, E(N), can be derived from the pdf f(x) as:

$$E(N) = \int_0^\infty x f(x) dx \\ = -\frac{b}{2a} \tag{3.8}$$

The mean response time, E(R) is now obtained thru Little's formula [1,2] which links the mean number of customers in the system (time-average) and the mean response time (customer-average):

$$E(R) = \frac{E(N)}{\lambda} = -\frac{b}{2a\lambda} \tag{3.9}$$

3) *Diffusion Parameters*

It remains to decide the diffusion parameters *a* and *b*. For our G/G/1-G time-out model, we denote by *N*(*t*) the number of customers in the system at time *t*. Similarly, we denote by $N_A(t)$, $N_\Gamma(t)$ and $N_B(t)$ the cumulative number of arrivals, the cumulative number of reaching time-out thresholds, and the cumulative number of departures during the time period

$$(0,\ t] = \{x \in R;\ c < x \le t\} \tag{3.10}$$

We then have

$$N(t) = N(0) + N_A(t) - N_\Gamma(t) - N_B(t) \tag{3.11}$$

For any renewal process {M(t), t≧0} with mean m and variance V of the inter-event time, we have

$$M(t) \sim N\left(\frac{t}{m},\ \frac{t}{m^3}V\right) \tag{3.12}$$

where $N(\alpha,\beta^2)$ signifies the normal distribution with mean α, variance β^2, and ～ signifies the asymptotically equality in distribution as time goes to infinity (t → ∞). See Hoshi et.al.[11] for the proof of (3.11).

Thus, we have

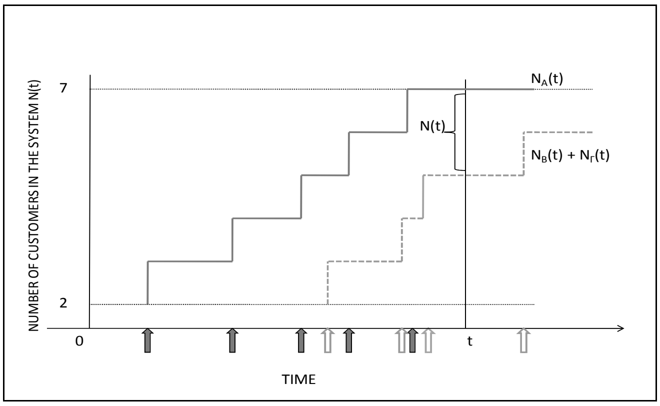
$$\lim_{t\to\infty} \frac{E(M(t))}{t} = \frac{1}{m},\quad \lim_{t\to\infty} \frac{\sigma^2_{(M(t))}}{t} = \frac{V}{m^3} \tag{3.13}$$

Since {N$_A$(t), t≧0} is a renewal process, we have from (3.12)

$$\lim_{t\to\infty} \frac{E(N_A(t))}{t} = \lambda,\quad \lim_{t\to\infty} \frac{\sigma^2_{(N_A(t))}}{t} = \lambda^3 \sigma_A^2 \tag{3.14}$$

$$\left(\because m \Leftrightarrow \frac{1}{\lambda}\quad V \Leftrightarrow \sigma_A^2\right)$$

The process {N$_\Gamma$(t) , t≧0} is not always renewal, but, in heavy traffic (ρ is close to 1), the server is expected to be occupied most of the time, so we can regard $N_\Gamma(t)$ as a renewal process. Thus, following the argument above, we have



z Fig.3 A sample path of the number of customers in the system.

$$\lim_{t\to\infty} \frac{E(N_\Gamma(t))}{t} = \gamma,\quad \lim_{t\to\infty} \frac{\sigma^2_{(N_\Gamma(t))}}{t} = \gamma^3 \sigma_\Gamma^2 \tag{3.15}$$

Similarly, we have

$$\lim_{t\to\infty} \frac{E(N_B(t))}{t} = \mu,\quad \lim_{t\to\infty} \frac{\sigma^2_{(N_B(t))}}{t} = \mu^3 \sigma_\Gamma^2 \tag{3.16}$$

The diffusion coefficients (a, b) have the following relationship as in Newell [15]:

$$a = \lim_{t\to\infty} \frac{E(N(t))}{t},\quad b = \lim_{t\to\infty} \frac{\sigma^2_{(N(t))}}{t} \tag{3.17}$$

It follows from (3.11),(3.13)-(3.17) that

$$a = \lim_{t\to\infty} \frac{E(N_A(t)) - E(N_\Gamma(t)) - E(N_B(t))}{t} \\ = \lambda - \mu - \gamma \quad (a < 0) \tag{3.18}$$

Similarly, we have

$$b = \lim_{t\to\infty} \frac{\sigma^2_{(N_A(t))} + \sigma^2_{(N_\Gamma(t))} + \sigma^2_{(N_B(t))}}{t} \\ = \lambda^3\sigma_A^2 + \mu^3\sigma_B^2 + \gamma^3\sigma_\Gamma^2 \\ = \lambda C_A^2 + \mu C_B^2 + \gamma C_\Gamma^2 \tag{3.19}$$

In the M/M/1-M system, we can derive these diffusion parameters (a, b) via a direct approach, see Appendix.

4) *Approximate Formulae via Diffusion Process*

Substituting our obtained diffusion parameters (3.18), (3.19) into (3.8) yields an approximate formula on the mean number of customers in the system:

$$E(N_{Diff}) = \frac{\lambda C_A^2 + \mu C_B^2 + \gamma C_\Gamma^2}{2(\mu + \gamma - \lambda)} \qquad (3.20)$$

Similarly, substituting the diffusion parameters (3.18), (3.19) into (3.9), we have an approximate formula on the mean response time:

$$E(R_{Diff}) = \frac{\rho C_A^2 + C_B^2 + \dfrac{\gamma}{\mu} C_\Gamma^2}{2(\lambda + \rho\gamma - \rho\lambda)} \qquad (3.21)$$

5) Refining Formulae

Whitt[23] noted that refining the diffusion approximations is necessary, since the diffusion approximations do not result in the well-known explicit formulae for special cases. However, there are very few explicit formulas for our time-out scheme models. Therefore, as for our refining, we adopt a trivial situation where the mean number of customers [E(N)] should be zero (the system should be idle) when the offered traffic is zero (ρ = 0). Observing this crucial but trivial point we have the following refinement equation:

$$E(N_{\mathrm{Re}f}) = E(N_{Diff}) - (1-\rho)\cdot E(N_{Diff})\big|_{\rho=0} \qquad (3.22)$$

Our refinement $E(N_{\mathrm{Prpo}})$ defined by the right-hand side of (3.20) is zero when the offered traffic is zero (ρ = 0), satisfying the trivial point as mentioned above. The refinement $E(N_{Prpo})$ converges to our diffusion approximation $E(N_{Diff})$ as the traffic intensity tends to unity (ρ → 1).

$$E(N_{\mathrm{Re}f}) = \frac{\lambda C_A^2 + \mu C_B^2 + \gamma C_\Gamma^2}{2(\mu+\gamma-\lambda)} - \frac{(1-\rho)\cdot\left(\mu C_B^2 + \gamma C_\Gamma^2\right)}{2(\mu+\gamma)} \qquad (3.23)$$

$$E(R_{\mathrm{Re}f}) = \frac{C_A^2 + \dfrac{C_B^2}{\rho} + \dfrac{\gamma}{\lambda} C_\Gamma^2}{2(\dfrac{1}{\rho} + \dfrac{\gamma}{\lambda} - 1)} - \frac{(\dfrac{1}{\lambda} - \dfrac{1}{\mu})\cdot\left(\mu C_B^2 + \gamma C_\Gamma^2\right)}{\dfrac{2}{\lambda}(\mu+\gamma)} \qquad (3.24)$$

IV. Numerical Examples

We compare our approximation (diffusion approximation and refined approximation) results with the simulated results. We present the mean number of customers E(N) in the system as a function of the traffic intensity *ρ*. Here, we assume the time is normalized by the mean service time, i.e., μ =1.0.

Figure 4 considers the M/M/1-D system. We assume γ = 0.5, 0.1, 0.01. Our proposed refined approximation is seen to be accurate very well. The diffusion approximation accuracy is not bad except for light and moderate traffic.

Figure 5 considers the M/M/1-M system indicating the 95% confidence interval via Student-*t* distribution. We assume γ=0.01. We see the almost same accuracy of our approximations as in Figure 4.

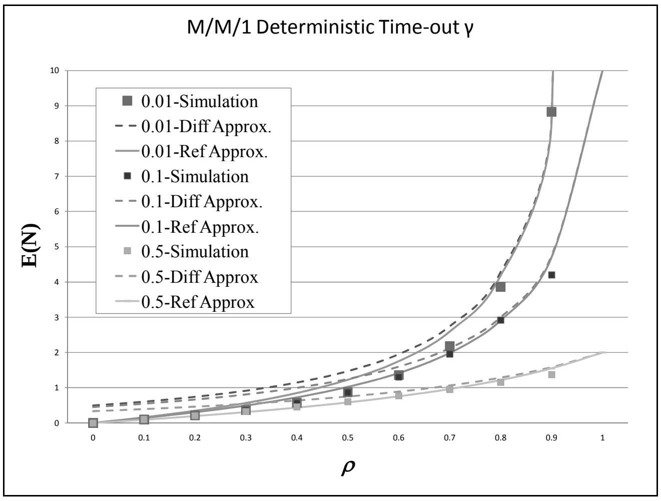


Fig. 4 M/M/1-D system performance.

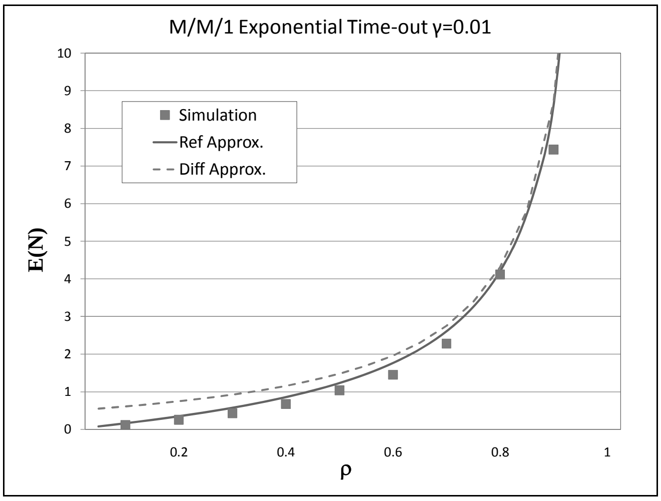


Fig. 5 M/M/1-M system performance.

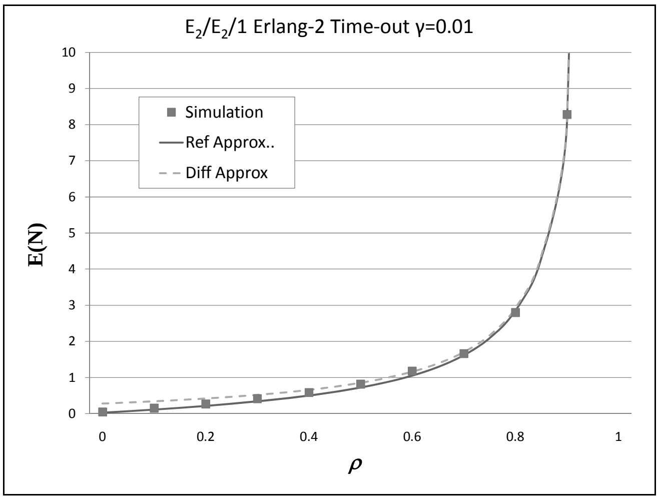


Fig. 6 E_2/ E_2/1-E_2 system performance.

Figure 6 considers the $E_2/E_2/1$-E_2 system. We also assume γ=0.01. We see the almost same accuracy of our approximations as in Figures 5 and 6.

It turned out that the proposal approximation is excellent accuracy by these examples of the numerical value. It is shown that it is enough practicably accuracy at a heavy traffic about a diffusion approximation formulae.

V. Conclusion

A computer communication system performance leads to our general (G/G/1-G) *time-out model*. Our queueing model includes the previously treated queueing models [3,4,6,14,16] as special cases, and it is identical to the one by Stanford [18]. However, Stanford's approach does not enable us to obtain the mean system performance measures (the mean number of customers in the system, and the mean response time).

As our analytical approach we have taken the diffusion approximation by Heyman [8] who treated the standard G/G/1 model without any time-out scheme. We have continued Heyman's effort, and determined the *diffusion parameters* for our time-out model. Based on the diffusion parameters, we have derived the mean system performance measures. We have further refined to propose approximate formulae on the mean performance measures for the time-out model.

It is our main contribution to present two-moment very simple approximate formulae on the mean performance measures for the (G/G/1-G) time-out model. It is left as a future work to seek for more accurate formulae which are consistent to the exact result for a special (i.e. M/M/1-M) model. It also remains to apply our results for obtaining the optimal time-out threshold.

Appendix

A) An Alternative Derivation of the Diffusion Parameters for the M/M/1-M System

We consider an M/M/1-M system. Recall that customers arrive according to a Poisson process with rate λ, and the service-time is exponentially distributed with mean μ^{-1}. The time-out threshold is also exponentially distributed with mean γ^{-1}.

Let *N(t)* denote the number of customers in the system at time *t*, and we introduce the transition probability as

$$\pi(t,n;n_0) = \Pr\{N(t)=n \,|, N(0)=n_0\} \qquad \text{(A.1)}$$

By using the Markov process theory [1,2], we have the following equation:

$$\begin{aligned}&\frac{\partial}{\partial t}\pi(t,n;n_0) = \lambda\pi(t,n-1) + (\mu+\gamma)\pi(t,n+1)\\ &-(\lambda+\mu+\gamma)\pi(t,n)\end{aligned} \qquad \text{(A.2)}$$

for any time t > 0 and any nonnegative integer n=1,2,$\cdots$. The initial condition is given by

$$\pi(0,n;n_0) = \begin{cases} 1 & if \quad n = n_0 \\ 0 & if \quad n \neq n_0 \end{cases} \qquad \text{(A.3)}$$

and the boundary condition is given by

$$\pi(t,n;n_0) = 0 \quad n < 0 \quad t \geq 0 \qquad \text{(A.4)}$$

For any bivariate function G, we have Taylor's expansion as

$$\begin{aligned}&G(x+h,y+k)\\ &= G(x,y) + \frac{\partial}{\partial x}Gh + \frac{\partial}{\partial y}Gk + \frac{1}{2!}\frac{\partial}{\partial x^2}G(x,y)h^2 \\ &+2\frac{\partial^2}{\partial x\partial y}G(x,y)hk + \frac{\partial^2}{\partial y^2}G(x,y)k^2\end{aligned} \qquad \text{(A.5)}$$

Putting *x=t*, *y=n*, *h=0* in (A.5), we have

$$G(t,x+k) = G(t,x) + \frac{\partial}{\partial x}Gk + \frac{\partial^2}{\partial x^2}Gk^2 \qquad \text{(A.6)}$$

$$\begin{aligned}G(t,x-1) &= G(t,x) - \frac{\partial}{\partial x}G + \frac{1}{2}\frac{\partial^2}{\partial x^2}G\\ G(t,x+1) &= G(t,x) + \frac{\partial}{\partial x}G + \frac{1}{2}\frac{\partial^2}{\partial x^2}G\end{aligned} \qquad \text{(A.7)}$$

Replacing G(t, x) by the transition probability $\pi(t, n; n_0)$, and comparing term by term (A.2) subject to (A.3)

$$\frac{\partial}{\partial t}\pi = -(\lambda-\mu-\gamma)\frac{\partial}{\partial x}\pi + \frac{(\lambda+\mu+\gamma)}{2}\frac{\partial^2}{\partial x^2}\pi \qquad \text{(A.8)}$$

Note that equation (A.8) corresponds to the diffusion (Fokker-Planck) equation (3.2). Comparing the coefficients of (A.8) and (3.2), we have the diffusion parameters (a, b) as the followings:

$$\begin{aligned}a &= \lambda - \mu - \gamma\\ b &= \lambda + \mu + \gamma\end{aligned} \qquad \text{(A.9)}$$

References

[1] H. Akimaru, and K. Kawashima, *Teletraffic*, Springer -Verlag, London, 1993

[2] A.O. Allen, *Probability, Statistics, and Queuing Theory with Computer Science Applications*, Academic Press, Boston, 1990.

[3] D.Y. Barrer, "Queueing with impatient customers and indifferent clerks," *Oper. Res*, vol.5, no.5, pp.644-649, October 1957.

[4] D.Y. Barrer, "Queueing with impatient customers and ordered service," *Oper. Res*, vol.5, no.5, pp.650-656, October 1957.

[5] D.R. Cox, *Renewal Theory*, Chapman & Hall, London: Methuen, 1962.

[6] P.D. Finch, "Deterministic customer impatience in the queueing system GI/M/1," *Biometrika,* vol.47 pp.45-52, 1960.

[7] E. Gelenbe, and I. Mitrani, *Analysis and Synthesis of Computer Systems*, Academic Press, New York, 1980.

[8] D.P. Heyman, "A diffusion model approximation for the GI/G/1 queue in heavy traffic," *Bell Syst. Tech.* J, vol.54, pp.1637-1646, 1975.

[9] T. Honma, "*Introduction to statistical mathematics*," Morikita, Tokyo, 1984. (in Japanese)

[10] K. Hoshi, S. Iijima, Y. Takahashi, and N. Komatsu, "A diffusion process approximation for the GI/GI/1 queuing system with time-out scheme," *Tech. Rept.*, vol.107, no.378, IN2007-103, pp.25-30, 2007. (in Japanese)

[11] K. Hoshi, Y. Takahashi, and N. Komatsu, “A stochastic approach for deriving the parameters arising out of a diffusion model,” *Tech. Rept*, vol.108, no.17, CQ2008-11, pp.59-64, 2008.(in Japanese)

[12] S. Kasahara, “Internet traffic modeling : towards queueing theory for the internet design,” *Tech. Rept*, Vol.101, No.647, pp.25-30, 2002. (in Japanese)

[13] G. Kimura, and Y. Takahashi, “Traffic analysis for a token ring system with exhaustive or gated service -renewal batch inputs model-,” *IEICE Trans.*, vol.J71-B, pp.129-137, 1988.

[14] I.N. Kovalenko, “Some queuing problems with restrictions,” *Theory of Prob. and its Appl.* vol.6, no.2, pp.204-208, 1965.

[15] G.F. Newell, *Application of Queuing Theory*, Chapman and Hall, London 1971.

[16] S.S. Rao, “Queueing with balking and reneging in M/G/1 systems,” *Metrika*, vol.12, no.1, pp.174-188, 1968.

[17] P.G. Sery, and J. Beale, *Red Hat Linux Internet Server*, Willy, Indiana, 2003.

[18] R.E. Stanford, “Reneging phenomena in single server queues,” *Math. Oper. Res*, vol.4, pp.162-178, 1979.

[19] A. Takahashi, Y. Takahashi, S. Kaneda, and N. Shinagawa, “Diffusion approximations for the GI/G/c/K queue,” *Proc. 16th ICCCN 2007*, pp.681-686, 2007.

[20] Y. Takahashi, “Diffusion approximation for the single server system with batch arrivals of multi-class calls,” *IEICE Trans.*, vol.J 69-A, pp.317-324, 1986.

[21] T. Takine, M. Murata, “Queue in communication network,” *Oper. Res*, vol. 43, no. 5, pp.264− 271, 1998. (in Japanese)

[22] W. Whitt, *Heavy Traffic Limit Theory*, Chapman and Hall, London, 1971.

[23] W. Whitt, “Refining diffusion approximations for queues” *Oper. Res. Letters*, vol.1, no.5, pp. 165-169, 1982.