A Proposal of Quasi-Static Approach for Analyzing the Stability of IP Telephony Systems

Masaki Aida*, Chisa Takano†, Masayuki Murata‡, Makoto Imase‡
* Graduate School of System Design, Tokyo Metropolitan University
Hino-shi, 191-0065 JAPAN
Email: maida@sd.tmu.ac.jp
† Graduate School of Information Sciences, Hiroshima City University
Hiroshima-shi, 731-3194, JAPAN
Email: chisa@m.ieice.org
‡ Graduate School of Information Science and Technology, Osaka University
Suita-shi, 565-0871 JAPAN
Email: murata@cmc.osaka-u.ac.jp, imase@ist.osaka-u.ac.jp

***Abstract*—Troubles in commercial IP telephony systems were reported one after another recently in Japan. One of important causes is congestion of control plane. In the current Internet, it has been recognized that controlling congestions caused by overload of control plane becomes important in addition to congestions caused by overload of data plane. In particular, since input traffic including retries tends to cause overload, it is an important issue to avoid congestion from retry traffic. In this paper, we focus on an RSVP-based communication model that is a combination of transmission and processing systems and consider the behavior of retry traffic. In general, users reattempt to set up connections not only when transmission systems are overloaded but also when processing systems in the network are overloaded. The latter is caused by user psychology: an increase in the waiting time for the processing to be completed tends to increase his or her reattempts. Thus, it is important to know interactions between users and the system and to manage both transmission and processing resources properly. Since both traditional Markov approach and simulation technique are difficult to handle these issues, we propose a new approach, quasi-static approach, for analyzing the stability of the RSVP model. This approach is based on the concept of decomposition of timescales like statistical physics. System behaviors are described as a combination of macroscopic behavior described in a human perceptible timescale and microscopic behavior described in a shorter timescale characterized by system's state transitions. Using our approach, we demonstrate the evaluation of IP telephony systems as an example of the RSVP model and show system down probabilities of very small values.**

I. Introduction

In the Internet, congestion due to overload of the control plane has become an important issue in addition to congestion due to overload of the data plane. In particular, retry traffic, such as reattempts to set up a connection or reload a file, tends to cause congestion. Therefore, it is important to develop evaluation models of system stability in order to design long-term stable system operations against retry traffic.

In general, retry traffic is generated by factors on both the data plane and the control plane as described below.

- Retry traffic generated due to a shortage of data-plane resources. A shortage of resources on the data plane causes existing requests for establishing a connection or securing bandwidth to be discarded. This will induce retries.
- Retry traffic generated due to a shortage of control-plane resources. A shortage of processing resources causes the response time of the processing system to increase. This will induce impatient users to retry their requests.

Since retry traffic itself consumes network resources, it is not appropriate to handle congestion on the data plane and that on the control plane separately.

Studies dealing with retry traffic on an M/G/s/s retrial queue model are well known [1], [2], [3]. The expression, M/G/s/s, represents a model in which service requests arise in a Poisson manner, enter the system, receive service from one of s servers, and then leave the system, and in which service requests are discarded if all s servers are busy. An M/G/s/s retrial queue model represents an M/G/s/s model in which discarded service requests are stored in a retrial queue and re-enter the system after a certain elapsed time determined by an exponential distribution (Fig. 1). It is known that the stability of the M/G/s/s retrial queue model can be derived if the length of the retrial queue does not diverge [1], and

$$\frac{\lambda_0}{\mu} < s, \tag{1}$$

where λ_0 is the arrival rate of service requests, excluding retry requests, per unit time, and $1/\mu$ is the average service time.

Although the M/G/s/s retrial queue model incorporates retry traffic that arises due to a resource shortage on the data plane, it does not incorporate retry traffic that arises due to a resource shortage on the control plane.

Reference [4] points out that an increase in the complexity of call control processing due to the addition of new process-

978-0-7695-3106-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICN.2008.20

ing requirements, such as the determination of QoS, causes an increase in the load on the control plane, and discusses the properties of call processing delay. Although [4] addresses both the data and control planes, it does not take retry traffic into consideration.

This paper focuses on RSVP-based communication services including IP telephony, and discusses the properties of retry traffic caused by a resource shortage on the data plane or the control plane. In order to take retry traffic generations into consideration, it is necessary to consider the interaction between users and the system. First, we show quasi-static retry traffic model, which describes behaviors of input traffic including retries under the condition that state transitions of the system occur on a timescale infinitesimally shorter than human perceptible timescale. Next, we introduce quasi-static approach for analysing the stability of input traffic including retries. This approach is based on the concept of decomposition of timescales like statistical physics. System behaviors are described as a combination of macroscopic behavior described in a human perceptible timescale and microscopic behavior described in a shorter timescale characterized by system's state transitions. Finally, using our approach, we demonstrate the evaluation of IP telephony systems as an example of the RSVP model and show system down probabilities.

II. Model That Takes Account of Retry Traffic on Both the Control Plane and the Data Plane

This section describes a retry traffic model of RSVP-based system. This model describes the properties of retry traffic that arises due to resource shortages on the data and control planes, and the method is used to analyze the macroscopic behavior described in a human perceptible timescale.

A. Basic Model

In order to describe retry traffic that arises due to resource shortages on the data and control planes, we have developed a model that incorporates a serial combination of a data-plane model and a control-plane model (Fig. 2).

The data-plane model describes the behavior of data transmission processing, such as securing link bandwidth (eg. call channel), and so we use an M/G/s/s retry queue model for it. If the system fails to secure bandwidth due to a resource shortage on the data plane, it means that all s servers are busy. In this case a service request is kept waiting in the retry queue for a period of time (exponential), and then is re-entered into the system as a new request.

The control-plane model describes the system behavior related to routing and processing of protocols, and so M/M/1 has been adopted for it. However, some of the users who have been kept waiting due to increased processing time may attempt to start a new service request. Therefore, we have modified this model slightly to take account of retry traffic that will arise depending on the length of the queue in the M/M/1 model. Since such retry traffic will be generated without the existing service requests being cancelled, our model assumes that new service requests arise without any service requests waiting in the queue in the M/M/1 model being withdrawn.

The input to the data plane is assumed to follow a Poisson process because M/M/1 is chosen for the control-plane model. This is because service requests that are departed from the control plane and entered in the data plane follow a Poisson process.

B. Quasi-Static Retry Traffic Model

This subsection considers how retry traffic arises from the control-plane system shown in Fig. 2. If retry traffic that is dependent on the length of the queues in the control plane arises, one possible case is that retry traffic volume is proportional to the length of the queue in the control plane. If we assume that service requests waiting in the queue generate retry traffic at a certain rate ϵ, a diagram depicting the speed of state transition with respect to the queue length of the control-plane system is as shown in Fig. 3. In this figure, λ_0 is the arrival rate of service requests, excluding retry traffic, per unit time, while $1/\eta$ is the average service time of the control-plane system. For this system to have a steady-state probability, the infinite sum on the right-hand side of the following equation must exist.

$$p_0 = \left[1 + \sum_{i=1}^{\infty} \prod_{j=0}^{i} \left(\frac{\lambda_0 + j\epsilon}{\eta} \right) \right]^{-1}.$$

Therefore, if $\epsilon > 0$, the system is unstable. Since an increase in retry traffic does not result in the divergence of the waiting time of an actual control plane under normal operating conditions, we can conclude that a model in which retry traffic is generated in proportion to the queue length of the control plane is not realistic.

In general, state transitions of the control-plane system occur on a timescale much shorter than humans can perceive. It is natural to assume that the facts that users grow impatient and reattempt to send service requests are not in response to a queue length at present time but are in response to the average waiting time that occurs on a longer timescale, a timescale long enough for humans to perceive. Let T be the minimum timescale perceptible to humans. We assume that the control-plane system is in a steady-state on a timescale smaller than T, and that retry traffic affects the system on a timescale larger than T. In the following, we assume such "a quasi-static state" for the generation of retry traffic.

- The system can be assumed to be in a steady-state on a timescale smaller than T.
- Changes in the system are observed at discrete times occurring at an interval of T.
- Retry traffic from the system at a certain time, $t = k$, is determined by the steady-state probability of the system at $t = k - 1$. (More specifically, retry traffic from control plane system is proportional to the average queue length at $t = k - 1$, and that from data plane system is proportional to loss tatio at $t = k - 1$.)

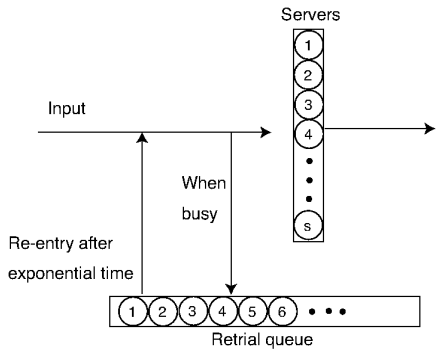

Fig. 1. M/G/s/s with retrial queue model.

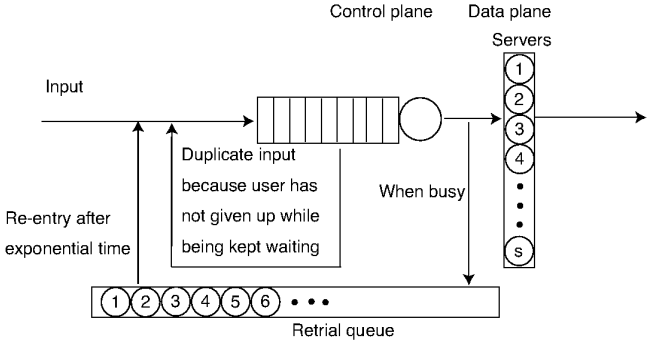

Fig. 2. Model incorporating retry traffic in both the control-plane model (M/M/1) and the data-plane model (M/G/s/s).

Note that the condition that the steady-state is achieved on a finite timescale T implies that transitions of the system occur on a timescale infinitesimally shorter than human perceptible timescale. In other words, the system works at infinite high speed.

C. Value of timescale T

Since this paper deals with retry traffic depending on users' average waiting time in timescale perceptible to humans, we treat the behaviors of the control-plane system occurring on human perceptible timescale T and that of shorter than T, separately. To do so, it is necessary to determine the specific value of T. The value of T must satisfy the following requirements:

- T must be a timescale on which humans can actually perceive an increase in the waiting time for their requests to be processed.
- T must be sufficiently longer than the timescale on which state transitions occur in the system so that it is possible to assume that the system is in a steady-state on a timescale smaller than T.

It is well known that the tolerable waiting time (i.e., the maximum length of time for which a user's attention can be kept) for a website to be displayed completely is 8 seconds. To satisfy this so-called 8-second rule, many webpages are so designed that they can be displayed completely within 8 seconds. Measurement of tolerable waiting time in actual experiments verified that a user's interest drifts away after 10 seconds or so [5]. It is possible that a user's tolerable waiting time in our case related to retry traffic is shorter than in the case of webpage browsing because our case can involve urgent service requests, such as attempts to make reservations for popular services. Reference [6] classifies the timescales according to how humans perceive them as follows:

- Delay of 0.1 second is perceived as instantaneous access.
- Delay of 1.0 second is the limit for users' flow of thought to stay uninterrupted.
- Delay of 10 seconds is the limit for keeping users' attention/focus on the dialogue.

The last category coincides fairly well with the 8-second rule for websites in terms of both the timescale and the reason. Since the generation of retry traffic seems to correspond to the second category, one second should be a reasonable value for T.

III. Stability of Input Traffic in Case that System Works at Infinite High Speed

The properties of input traffic and the system's stability are examined here using a quasi-static retry traffic model, a model that takes account of retry traffic caused by resource shortages on both the control plane and the data plane in combination. In addition, this model assumes that the system works at infinite high speed.

Details of the model are as follows. Let $\lambda_0 > 0$ be the traffic input rate without retry traffic. We assume that $\lambda_0 > 0$. Changes in the state of the control-plane system are examined at discrete intervals T. We assume that the retry traffic at time $t = k$ is determined by the steady-state probability of the system at time $t = k-1$. We assume that users whose service requests have not been served by any server of the data-plane system keep sending service requests until they are finally served, and that users kept waiting for their requests to be processed by the control-plane system send service requests at a rate proportional to the average queue length.

Let λ_k be the input load, including retry traffic, at time k. We assume that the input load at time $k+1$ is

$$\lambda_{k+1} = \lambda_0 + \lambda_k B(\rho_k, s) + \epsilon \frac{\rho_k/a}{1-\rho_k/a}, \tag{2}$$

where $\rho_k = \lambda_k/\mu$, $1/\mu$ is the average service time of the data-plane system, a is the ratio of $1/\mu$ to the processing time of the control-plane server (i.e., the ratio between the processing powers of the two servers. $a = \eta/\mu$ where $1/\eta$ is the average service time of the control-plane system), ϵ is a positive constant indicating the intensity of the retry traffic generated due to a control-plane resource shortage. $B(\rho, s)$ is an Erlang B formula, i.e.,

$$B(\rho, s) = \frac{\frac{\rho^s}{s!}}{1+\rho+\frac{\rho^2}{2!}+\cdots+\frac{\rho^s}{s!}}. \tag{3}$$

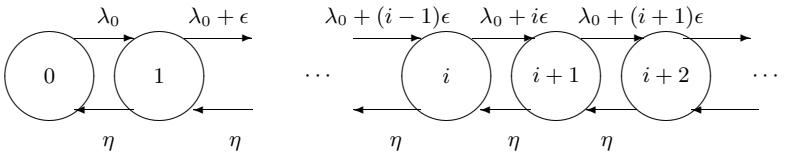

Fig. 3. State transition diagram for the case where retry traffic is generated in proportion to the queue length of the control-plane system

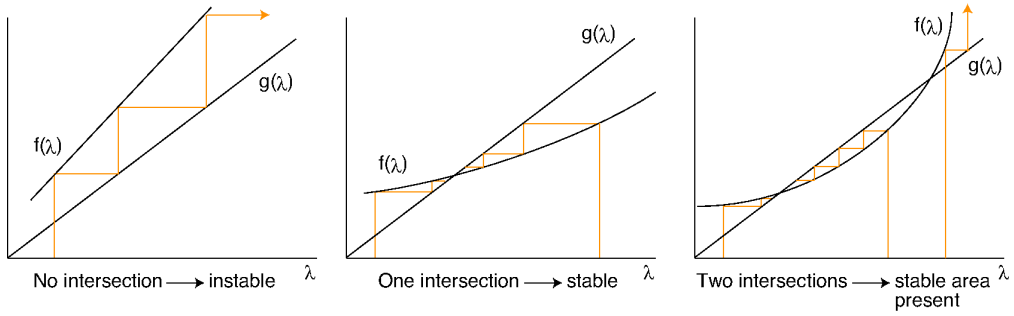

Fig. 4. System stability depends on the presence of intersections between $f(\lambda)$ and $g(\lambda)$.

Next, we will consider the stability of the system. We assume that the requirement for the system being stable is that traffic, including retry traffic, does not diverge after a sufficient elapsed time. Namely,

$$\lim_{k\to\infty} \lambda_k < \infty. \tag{4}$$

This can be verified by defining the functions of λ, $f(\lambda)$ and $g(\lambda)$ as follows, and examining whether they intersect.

$$f(\lambda) = \lambda_0 + \lambda\, B(\rho, s) + \epsilon\, \frac{\rho/a}{1-\rho/a}, \tag{5}$$

$$g(\lambda) = \lambda. \tag{6}$$

Since $\lambda_0 > 0$, $f(0) > g(0)$. The relationship between $f(\lambda)$ and $g(\lambda)$ in some typical cases is as shown in Fig. 4. If the input traffic at a certain time is λ, the input traffic at the next time becomes $\{g^{-1} \circ f\}(\lambda)$, followed by $\{g^{-1} \circ f\}^2(\lambda)$, etc. In general, the input traffic after an elapse of n unit times becomes $\{g^{-1} \circ f\}^n(\lambda)$. As shown in the left-most chart in Fig. 4, if $f(\lambda)$ and $g(\lambda)$ do not intersect, $\lim_{n\to\infty}\{g^{-1} \circ f\}^n(\lambda) = \infty$. If, on the other hand, the two functions intersect as shown in the middle chart, and if $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) < 1$, then the system is stable. If the two intersects in the manner shown in the right-hand side of Fig. 4, and if $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) > 1$, then the system is stable for all λ to the left of the right-most intersection.

If the retry traffic from the control-plane system is negligible ($\epsilon = 0$ or $a = \infty$), $f(\lambda) = \lambda_0 + \lambda\, B(\rho, s)$. Considering that $\lambda_0 > 0$ and $B(\rho, s) < 1$, $f(\lambda)$ and $g(\lambda)$ intersect at one point if (1) is satisfied, and do not intersect at all if (1) is not satisfied. If, on the other hand, the retry traffic from the control-plane system is not negligible ($\epsilon > 0$ and $a < \infty$), the third term on the right-hand side of (5) always renders $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) > 1$. Therefore, $f(\lambda)$ and $g(\lambda)$ may intersect at two points or at one point (a point of contact), or do not intersect at all.

Next, let us look at system stability using examples with specific values. We will consider the case where $\lambda_0 = 9,000$, $\mu = 1$, $a = 50,000$, and $s = 10,000$. The behavior of $f(\lambda)$ and $g(\lambda)$ for $\epsilon = 0.1$, 1, and 10 is shown in Fig. 5. In order to consider the influence of retry traffic on the control plane and the data plane separately, Fig. 5 shows the first and the second terms (original traffic and contribution of the data plane), and the third term (contribution of the control plane) on the right-hand side of (5) separately, in addition to the value of $f(\lambda)$. This figure implies that the behavior of $f(\lambda)$ is hardly influenced from the value of ϵ. This result is convenient for us. This is because the value of ϵ is hard to measure from real systems and hard to determine. This insensitivity enable us to roughly determine the value of ϵ.

IV. Quasi-Static Approach

In the previous sections, we assume the condition that the steady-state is achieved on a finite timescale T. This implies that the system works at infinite high speed. However, since system speed is finite in real systems, the steady-state is never achieved in finite time interval. To avoid this problem, we propose "quasi-static approach". In this approach, system behaviors are described as a combination of macroscopic behavior described in a human perceptible timescale and microscopic behavior described in a shorter timescale characterized by system's state transitions. In other words, the quasi-static approach gives way to decompose timescale into a human perceptible timescale and a shorter timescale.

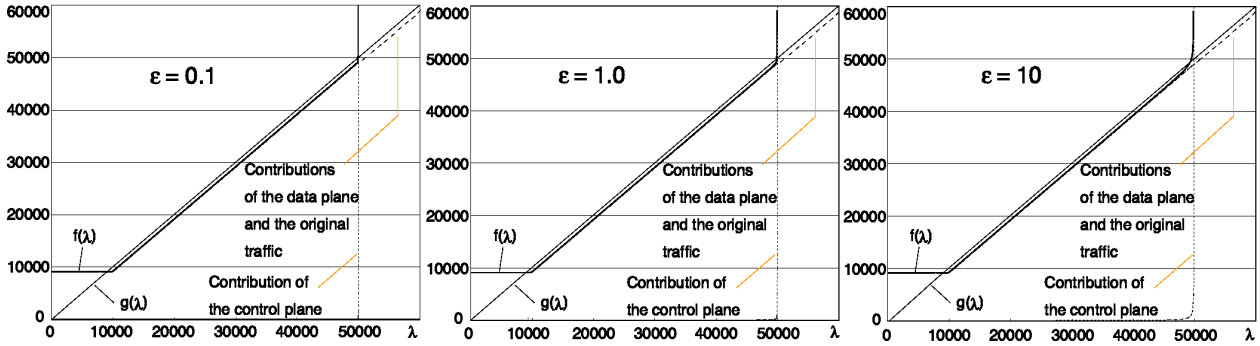

Fig. 5. Relationship between $f(\lambda)$ and $g(\lambda)$ when ϵ is varied.

A. Concept

Let us consider how to examine the stability of input traffic under the condition that the system works at finite speed. First, we clarify problems in traditional approaches to examine the stability. Next, we clarify the aims of the proposed quasi-static approach by showing that the characteristics of our approach are suitable to avoid the problems.

Let n be the number of service requests arisen in an interval of T. If retry traffic from control plane system is proportional to the average queue length of M/M/1 at time k (in an interval of T), the third term on the right-hand side of (2) is replaced with

$$\epsilon \frac{1}{n} \sum_{i=1}^{n} Q_i^k, \tag{7}$$

where Q_i^k denotes the number of requests in M/M/1 system at immediately before the i-th request arrival. If system works at infinite speed, that is, $k =$ constant and $n \to \infty$, we have

$$\lim_{n\to\infty} \epsilon \frac{1}{n} \sum_{i=1}^{n} Q_i^k = \epsilon \frac{\rho_k/a}{1-\rho_k/a} \quad \text{a.s.} \tag{8}$$

Let us consider the system works at very low speed, for example $n = 1$. This case corresponds that human can detect the present system state, ability of human growth very high or system speed is very low. In this case, the next input traffic depens only on the latest value of Q_1^k. So, the system can be described as a Markov model. In general, if $n > 1$, we can use a Markov model with n-dimensional state space, which is constructed by the past n states $\{Q_1^k, Q_2^k, \ldots, Q_n^k\}$. However, if the system works at high speed, $n \gg 1$, the state space explodes and becomes intractable.

Next, we consider applicability of simulation technique. If we require the system down probability, for example, to be less than 10^{-6} in long-term operations of IP telephony system, more than 10^6 (usually 10^8 to 10^9) runs of the corresponding long-term simulation are required. Such a large scale simulation is not realistic.

Aims of quasi-static approach are as follows. In order to analyze the stability of the high speed (but finite) system,

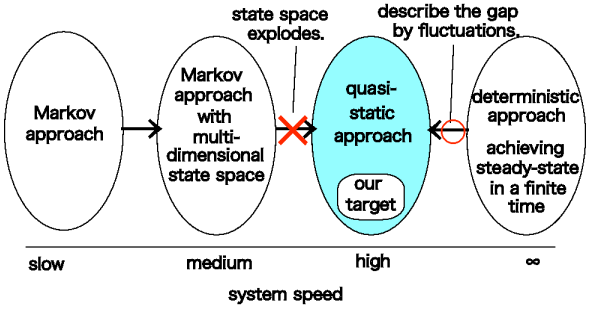

Fig. 6. Concept of quasi-static approach.

our quasi-static approach first examines the system behavior at infinite speed, and considers the gap between the system behavior at finite speed as "fluctuations" (see Fig. 6). In other words, the change factor in the state of the system is decomposed into two, one is deterministic change factor and is described by system behavior at infinite speed of the system and the other factors are described as stochastic fluctuations. This approach is effective to the model of the scale that cannot be realistically handled by the Markov model and the simulation technique.

B. Modeling

Expediently, we replace the discrete time k with a continuous time t, and denote the amount of the input traffic (including retry traffic) at time t by $X(t)$. At this time, temporal evolution of input traffic $X(t)$ can be written in the form of the Langevin equation as

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t) = F(X(t)) + R(X(t), t), \tag{9}$$

where $F(X)$ denotes deterministic transitions governed by the behaviour of input traffic at infinite system speed, and $R(X,t)$ denotes stochastic fluctuations characterized by the gap between the system behaviors at infinite and finite speeds.

First, we consider details of $F(X)$. We use approximation of Erlang B formura as

$$B(\rho, s) = \begin{cases} 0, & (\rho \leq s), \\ 1 - s/\rho, & (\rho > s). \end{cases} \tag{10}$$

This approximation means the portion of input traffic that exceed the server capacity $s\mu$ is lost, and it is loss-less if input traffic is less than or equal to $s\mu$. That is, $\lambda\, B(\rho, s) = \lambda - s\mu$ for $\lambda > s\mu$. Figure 5 implies this is appropriate approximation. From $F(X) = f(X) - g(X)$, we have an approximation of $F(X)$ as

$$F(X) = \begin{cases} \lambda_0 - X + \dfrac{\epsilon\, X/(a\mu)}{1 - X/(a\mu)}, & (X \leq s\mu), \\ \lambda_0 - s\mu + \dfrac{\epsilon\, X/(a\mu)}{1 - X/(a\mu)}, & (s\mu < X). \end{cases}$$

Next, we consider details of $R(X, t)$. $R(X, t)$ represents the fluctuations of input traffic, which consists of three types of traffic: the original traffic, retries caused by server losses, and retries caused by waiting at M/M/1. From Fig. 5, we can recognize that the retries from M/M/1 is very small in ordinary system operations. Therefore, we only take the original traffic and loss retries into consideration. In order to determine the magnitude of the fluctuations, we should investigate the variance of input traffic. For original traffic, the variance of input traffic is equal to ite input rate. So, we should know the variance of loss retry traffic.

We evaluate the number of losses in M/M/s/s system with $s = 10,000$. Figure 7 shows the result. The horizontal axis denotes input traffic X, and the vertical axis denotes the variance of the number of losses occurred during an interval T under the condition that the actual input traffic is X. Note that the number of input traffic during the interval T is not always λT, even if the rate of input traffic is λ. The results shown in Fig. 7 are obtained under the condition of $X = \lambda T$. This is because we want to decompose fluctuations of loss-retries into that caused by fluctuation of input traffic itself and that caused by fluctuations of the number of losses. From this figure, we can recognize that the variance of number of losses is like the step function. So, we adopt the approximation of the variance as: it is zero when $X \leq s\mu$, and it is a constant c when $s\mu < X$.

Consequently, fluctuations can be represented as

$$R(X, t) = \begin{cases} \sqrt{X}\, \xi(t), & (X \leq s\mu), \\ \sqrt{X + c}\, \xi(t), & (s\mu < X), \end{cases}$$

where $\xi(t)$ is fluctuation noise and is independent of X. Here, the behavior of $\xi(t)$ is caused by effects from a lot of number of independent users, and we assume that $\xi(t)$ obeys the standard normal distribution by the central limit theorem and is white noise

$$\mathrm{E}[\xi(t)] = 0, \quad \mathrm{E}[\xi(t)\, \xi(t')] = \delta(t - t').$$

In order to erase X dependence from the second term on the right-hand side in (9), we introduce a new variable $Y(t)$

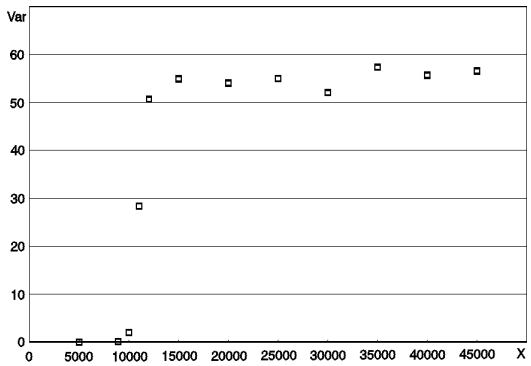

Fig. 7. Variance of the number of losses in M/M/s/s system with $s = 10000$.

as,

$$Y(t) = 2\sqrt{X(t) + c(X)}, \tag{11}$$

where

$$c(X) = \begin{cases} 0, & (X \leq s\mu), \\ c, & (s\mu < X). \end{cases} \tag{12}$$

Then, since the time derivatrive of $Y(t)$ satisfies

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = \frac{1}{\sqrt{X(t) + c(X)}} \frac{\mathrm{d}X(t)}{\mathrm{d}t}, \tag{13}$$

the Langevin equation (9) is rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t} Y(t) = G(Y) + \xi(t), \tag{14}$$

where

$$G(y) = \frac{1}{\sqrt{x + c(x)}}\, F(x), \tag{15}$$

and

$$y = 2\sqrt{x + c(x)}. \tag{16}$$

In order to investigate the deterministic transitions of (14), we consider the potential function $U(Y)$ as

$$U(y) = -\int G(y)\mathrm{d}y = -\int G(y)\frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}x.$$

Since the function in the last integral is

$$G(y)\frac{\mathrm{d}y}{\mathrm{d}x} = \begin{cases} \dfrac{\lambda_0}{x} + \dfrac{\epsilon}{a - x} - 1, & (x \leq s\mu), \\ \dfrac{\lambda_0 - s\mu}{x + c} + \dfrac{1}{x + c} \cdot \dfrac{\epsilon}{a - x}, & (s\mu < x), \end{cases}$$

$U(y)$ without a constant of integration is wriiten as

$$U(y) = \begin{cases} -\lambda_0 \log x + \epsilon \log(a - x) + x, & (x \leq s\mu), \\ -(\lambda_0 - s\mu)\log(x + c) \\ \qquad -\epsilon \dfrac{c}{c - a} \log(x + c) \\ \qquad\quad -\epsilon \dfrac{a}{c - a} \log(a - x), & (s\mu < x). \end{cases}$$

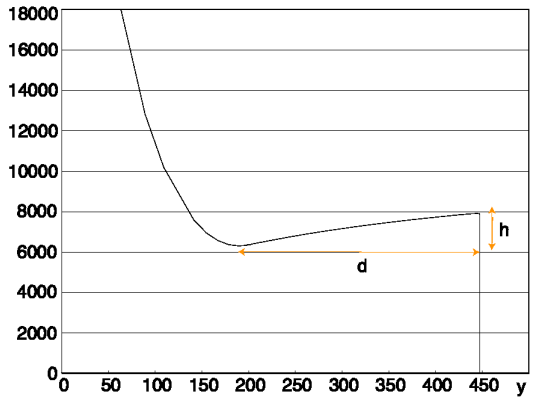

Fig. 8. Example of potential function $U(y)$.

Figure 8 shows the form of the potential function $U(y)$ with parameters of $s = 10,000$, $\epsilon = 0.1$, $\lambda_0 = 9,000$, $\mu = 1$, $a = 50,000$, and $c = 56$. The horizontal axis denotes y and the vertical axis denotes $U(y)$. The minimal point corresponds to left intersection, and the wall corresponds to the right intersection in Fig. 5. If there are no fluctuations, that is, $\xi(t) = 0$ in (14), Y becomes stable at the minimal point. If there are fluctuations, there is some probability to fall off from the potential wall. This situation corresponds to the instabilization of the system and the volume of input traffic diverges.

To prevent the instabilization of the system, there are two strategies. One is increase h and the other is increase d, depicted in Fig. 8. By increasing the number of communication channels s, h increases. By increasing the service rates of M/M/1 (call-control system) η (or a), d increases.

The form of the Langevin equation (14) is not easy to handle to evaluate the actual probability of system down. So, we use the temporal evolution equation of the probability density function of Y. Let $p(y,t)\,\mathrm{d}y$ be the probability that Y satisfies $y < Y \leq y + \mathrm{d}y$ at time t. It is well known that the Langevin equation (14) is equivalent to the Fokker-Planck equation as

$$\frac{\partial}{\partial t}\,p(y,t) = \left\{ -\frac{\partial}{\partial y}\,G(y) + \frac{1}{2}\,\frac{\partial^2}{\partial y^2} \right\} p(y,t), \qquad (17)$$

where the first and second terms on the right-hand side of this equation represent deterministic and stochastic state transitions, respectively. This equation describe a diffusion process in the potential function $U(y)$.

V. Numerical Examples

This section demonstrates the quasi-static approach by evaluating system down probabilities of an IP telephony system. The details of our model is as follows. The number of communication channels is $s = 10,000$, the mean holding time of a channel is $1/\mu = 3$ min, the rate of the original input traffic (excluding retries) $\lambda_0 = 9,000$ calls per 3 min, and the ratio of the service rates of M/M/1 (call-control system) to μ is $a = \eta/\mu = 50,000$. The last one means that the mean service time of the call-control system is 3.6 ms and it has five times capability of the total service rate of 1,000 call channels. In addition, we set $\epsilon = 0.1$ and $c = 56$.

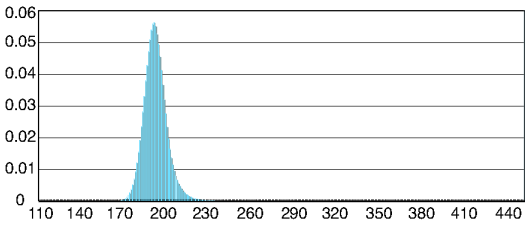

Fig. 9. The probability density function $p(y,t)$ at the elapsed time $t = 3$ haurs.

TABLE I
The probabilities of system in unstable and the elapsed times.

elapsed time (hours.)	probability of system in unstable
1.0	3.0×10^{-11}
1.5	3.1×10^{-10}
2.0	8.9×10^{-10}
2.5	1.6×10^{-9}
3.0	2.4×10^{-9}

As the initial condition, the probability density function $p(y,0)$ is δ-function at the minimal point of the potential function, that is, the initial state is deterministically given at the stable point. The potential wall is an absorption state. When probability reaches this point once, it cannot return. It falls off from the potential wall and the volume of input traffic diverges.

Using above setting, we conducted numerical experiments and evaluate the probabilities of absorption during certain time intervals. That is the probabilities that system goes down at least once during certain time intervals.

Figure 9 shows the probability density function $p(y,t)$ at $t = 3$ hours. The horizontal axis denotes y and the vertical axis denotes $p(y,t)$. Seemingly, tail of the density function on the right is slightly heavy though this seems to be near normal distribution. This is recognized through the form of the potential function shown in Fig. 8. The potensial wall exixts at $y \simeq 445$ and the probability is being absorbed little by little in this point.

Table I shows the probabilities of absorption (system goes down at least once) for some elapsed times. Although these probabilities are very small from technological point of view, they are increasing with elapsed time. The probabilities of the stabilized operation of IP telephony system are the following. When assuming that busy hour(s) continues for one hour every day, the probability that the system goes down at least once for five years is about 5.4×10^{-8}, that for ten years is about 1.1×10^{-7}. When assuming that busy hour(s) continues for three houre every day, the probability that the system goes down at least once for five years is about 4.3×10^{-6}, that for ten years is about 8.7×10^{-6}.

VI. Conclusions

In this paper, we have proposed the performance evaluation technique to achieve the stabilized operation of RSVP-based communication systems including IP telephony system, and have shown numerical examples as a demonstration of our technique.

In order to describe congestions caused by overloads of control plane and data plane, we have considered a combined model of both M/M/1 and M/G/s/s models and introduce retry traffic from users. In particular, retries caused by overload of control plane is related to user psychology. Thus, it is important to know interactions between users and the system and to manage both transmission and processing resources properly. So, we have introduced the quasi-static traffic model in which the steady-state is achieved on a finite timescale. That is the situation that the system works at infinite high speed.

Next, we have proposed the quasi-static approach. This approach is based on the concept of decomposition of timescales like statistical physics. System behaviors are described as a combination of macroscopic behavior described in a human perceptible timescale and microscopic behavior described in a shorter timescale characterized by system's state transitions. The former is given by the quasi-static traffic model under the condition that the system works at infinite high speed. The latter is given by stochastic fluctuations describing the gap between behaviors of infinite and finite speed of systems. This approach is effective to attack our problem although both traditional Markov approach and simulation technique are difficult to handle this problem.

Our approach enables us to clarify intuitive understanding of the mechanism of the system instabilization. We have demonstrated the evaluation of IP telephony systems as an example of the RSVP model and show system down probabilities of very small values.

Our numerical example shows the system down probability is very small if the service rates of M/M/1 (call-control system) has five times capability of total capacity of communication links. On the other hand, we have other results that the system down probability becomes extremely high if the capacity of call-control system is not enough. Thus, the designs of control and data plane resources are mutually related and should be considered together.

Hereafter, we further investigate the stability of IP telephony system, and consider extensions of the quasi-static approach for applying to other communication models.

Acknowledgements

A part of this research was made possible by the Grant-in-Aid for Scientific Research (A) No. 16200003 (2004–2007) from the Japan Society for the Promotion of Science.

References

[1] G.I. Falin and J.G.C. Templeton, Retrial Queues, Chapman & Hall, London, 1997.

[2] Hamada Alshaer and Eric Horlait, "The joint distribution of server state and queue length of M/M/1/1 retrial queue with abandonment and feedback," 8th International Symposium on DSP and Communication System, Australia, Oct. 2005.

[3] Weixin Shang, Liming Liu and Quan-Lin Li, "Tail asymptotics for the queue length in an M/G/1 retrial queue," Queueing Systems, vol. 52, pp. 193–198, 2006

[4] Ren-Hung Hwang, Chia-Yi, James F. Kurose and Don Towsley, "On-call processing delay in high speed networks," IEEE/ACM Transactions on Networking, vol.3, no.6, pp.628–639, Dec. 1995.

[5] Fiona Fui-Hoon Nah, "A study on tolerable waiting time: how long are Web users willing to wait?," Behaviour & Information Technology, vol.23, no.3, pp.153–163, May 2004.

[6] Jakob Nielsen, "Response times: the three important limits.," Excerpt from Chapter 5 of Usability Engineering by J. Nielsen, Academic Press, 1993. Available at `http://www.useit.com/papers/responsetime.html`.