Fundamental Limits of Input Rate Control in High Speed Network

San-qi Li Song Chong
Department of Electrical and Computer Engineering
University of Texas at Austin, TX 78712

Abstract

In this paper[1] *we explore the fundamental limits of input rate control by spectral analysis in frequency domain. Both deterministic and stochastic analyses are developed. Especially, the simple deterministic analysis helps us to gain a great knowledge of performance trade-off for input rate control in high speed network.*

1 Introduction

A central objective of input rate control in high speed network is to prevent nodal congestion, caused by irregular or unexpected bursty input. Many control protocols have been proposed to regulate (or reshape) the input rate process at network entry point, among which the most popular one is the *leaky bucket* control. In principle, an input rate control system is designed to trade packet delay, or selective loss, at source for congestion avoidance within network. Most analyses of input rate control focus on control delay and loss rate, with simplifying assumption of the input process to be Poisson, renewal, or 2-state (3-state) Markov modulated Poisson process [1]-[8]. It is lack of technique, however, to measure the effectiveness of input rate control to avoid network congestions. This is largely because of the complexity involved in the exact modeling of output rate process of the control system [6] [8]. As a result, there is no clear comparison between irregular input and regulated output via control system.

This paper applies a new concept of input spectral characterization, recently developed in queueing theory [9] [10], to measure input rate control effectiveness in frequency domain and therefore to explore the fundamental limits of input rate control to performance improvement. Figure 1 describes a control system with its input rate process generated by source and output rate process injected to network. Our emphasis is placed on the study of interrelationship between input power spectrum $P_I(\omega)$ and output power spectrum $P_O(\omega)$ of the control system. A key observation made in [9] [10] is that the network performance is dominated by input power in low-frequency band. Ideally, an input rate control system is to reduce input power in low-frequency band without causing excess delay/loss of information at the source. Hence, the less the output power in low-frequency band, the less the chance for nodal congestion to occur, and so the less the delay/loss of information within network,

There are two basic approaches to system modeling: the stochastic approach and the deterministic approach. It is generally difficult to characterize the output rate process by stochastic approach since most control systems are nonlinear. Here we first use a deterministic approach to explore the performance tradeoff among input spectrum, queueing delay and output spectrum of the control system. Despite some obvious mathematical looseness of deterministic approach, it has been successfully used to provide many insightful results, which otherwise may not be possible to provide by stochastic analysis [9] [11]. It is also true that many solutions obtained by deterministic analysis can form the basis to stimulate and further enhance the theoretical development by stochastic analysis.

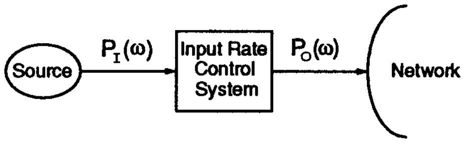

Figure 1: A generic input rate control system

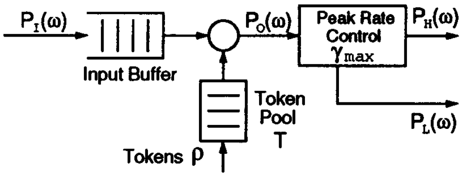

Figure 2: Leaky-bucket input control system

Our deterministic analysis is most simple and clear in concept. Consider a generic stationary random input process. Its first degree property is measured by the average input rate $\bar{\gamma}$; the second degree property is characterized by continuous input power spectrum $P_I(\omega)$. The study in [10] indicates that the queueing performance is much more dependent on $\{\bar{\gamma}, P_I(\omega)\}$ than higher degree input properties. In our modeling, only the first two degree input properties are considered. In order to examine the effect of individual input power spectral components, here we choose an "isolated" sinusoidal input as a test signal to measure the control system response as a function of the input sinusoidal frequency. The analysis can therefore be much simplified.

Our study concentrates on a leaky-bucket control system described in figure 2. The operation is simple. First, each packet in the input buffer, before being moved forward, must

[1] The research reported here was supported by NSF under grant NCR-9015757 and Texas Advanced Research Program under grant TARP-129.

be matched by a token from the pool. Tokens are generated to the pool at constant time interval Δ, unless the pool is full. The token pool acts like a bank credit system, which allows input packets to borrow the service capacity in advance. Before being forward to network, the matched packets can be further divided into two priority streams, measured by $P_H(\omega)$ and $P_L(\omega)$ in frequency domain. The packets in high priority stream is constrained by peak access rate $\gamma_{\max}$, and the rest of packets are in low priority stream. The low priority packets are likely to be dropped within network in case of nodal congestion. There are three control parameters:

- ρ: control utilization factor with $\rho \overset{\text{def}}{=} \overline{\gamma}\Delta$
- T: token pool size
- $\gamma_{\max}$: peak access rate.

In response to the input function $\{\overline{\gamma}, P_I(\omega)\}$, one can therefore measure the control performance by

- $\overline{q}_I$: mean queue length at input-control buffer
- $P_o(\omega)$: output power spectrum which can be further divided into $\{P_H(\omega), P_L(\omega)\}$

For simplicity, we assume infinite buffer size. Hence, the performance tradeoff between $\overline{q}_I$ and $P_o(\omega)$ is to be observed in function of $\{\rho, T, \gamma_{\max}\}$ for each given input $\{\overline{\gamma}, P_I(\omega)\}$. We will capture the effect of each individual control parameter on output spectrum and queueing delay, in response to input spectrum. The analysis is then extended to the next adjacent queue with controlled input, which is used to further evaluate the network response to input rate control.

The following four guiding principles are developed:

1. The overall queueing performance is inherently determined by both input source and network environment; the function of input rate control is simply to trade more input queueing for less network queueing.
2. Due to the large disparity between source generation rate and link transmission rate, a stringent input rate control may unnecessarily increase the user end-to-end delay significantly.
3. While input rate control is ineffective in high frequency band, it will be effective in low frequency band if and only if the user can tolerate excess delay or loss at network entry point.
4. The network performance is insensitive to the adjustment between token pool size T and token generation rate $1/\Delta$ for the leaky bucket control system.

The same control performance, explored by the deterministic analysis, can also be found by stochastic analysis except in very limited cases. In this paper we will use the stochastic approach to obtain the exact solutions of two adjacent queues in response to input power spectrum, where the first queue is a leaky bucket input rate control system and the second queue represents the network.

The paper is organized as follows. Section 2 shows the stochastic modeling of input rate control system. The corresponding deterministic modeling is given in section 3. The main results of this paper are in section 4 for input control performance trade-off based on the deterministic analysis. The stochastic analysis is carried out in section 5 to further explain the solutions obtained in section 4. The paper is then summarized in section 6.

2 Stochastic Modeling

Let us first neglect the peak rate control implemented by $\gamma_{\max}$ in figure 2. Based on fluid flow modeling, which is commonly used for stochastic queueing analysis [6] [12], one can describe the above leaky-bucket queueing system by

$$\tilde{q}(t+\Delta) = \max\{-T, \tilde{q}(t) + \tilde{\gamma}_I(t) - 1\} \qquad (1)$$

where $\tilde{\gamma}_I(t)$ is the input rate random variable at time t, measured in Δ unit. $\tilde{q}(t)$ is a continuous random variable, which is equal to the input buffer content $\tilde{q}_I(t)$ subtracted by the token pool size $\tilde{q}_T(t)$ at time t. Here we add a $\sim$ accent to each of the notations for stochastic analysis. Since both $\tilde{q}_I(t)$ and $\tilde{q}_T(t)$ cannot be simultaneously positive, we have

$$\tilde{q}(t) = \begin{cases} \tilde{q}_I(t) & \text{if } \tilde{q}(t) \geq 0 \\ -\tilde{q}_T(t) & \text{otherwise} \end{cases}$$

That is, $\tilde{q}_I(t) = \max\{0, \tilde{q}(t)\}$. The output rate random variable will then be characterized by

$$\tilde{\gamma}_o(t) = \begin{cases} \frac{1}{\Delta} & \text{if } \tilde{q}(t) \geq 0 \\ \tilde{\gamma}_I(t) & \text{otherwise} \end{cases} \qquad (2)$$

Mathematically, the queueing analysis of such a leaky-bucket queueing system is equivalent to that of a single queue system loaded by ρ (as if taking $T = 0$ here) [2] [5].

In our stochastic modeling, $\tilde{\gamma}_I(t)$ represents a stationary random process, to which only the first and second degree input properties, defined by $\{\overline{\gamma}, P_I(\omega)\}$, are assumed to be characterized. This assumption is made for two reasons. First, in practice it is always difficult to measure higher degree properties of random traffic. Second, the queue response is much more dependent on the first two degree input properties than higher ones [10]. Note that the stochastic queueing analysis cannot be carried out unless the time variation of the input rate $\tilde{\gamma}(t)$ is characterized by Markov chain. The technique developed in [10] shows how to construct such an input Markov chain from the given first and second degree input properties. The queue response to input power spectrum can then be evaluated by using the QBD-Folding algorithm developed in [13] [14]. The key problem with the stochastic analysis, however, lies in its difficulty to characterize the output rate process, which essentially measures the effectiveness of the input rate control. The detail stochastic analysis is postponed to section 5. The next two sections focus on the deterministic analysis.

3 Deterministic Modeling

In the deterministic analysis, we use a periodic input function $\gamma_I(t) = \gamma_I(t + t_o)$ for the input rate where t_o is the common period. One can then describe the input queue function by

$$q(t+\Delta) = \max\{-T, q(t) + \gamma_I(t) - 1\}$$

as its counterpart (1). From $\gamma_I(t) = \gamma_I(t + t_o)$ we must also have $q(t) = q(t+t_o)$ in steady state. The average input queue is therefore measured by

$$\overline{q}_I = \frac{1}{t_o} \int_0^{t_o} \max\{0, q(t)\} dt.$$

Similar to (2), we have the output rate function $\gamma_o(t)$ equal to $\frac{1}{\Delta}$ for $q_I(t) \geq 0$, and $\gamma_I(t)$ otherwise, with respect to $\gamma_o(t) =$

$\gamma_o(t+t_0)$. The numerical evaluation of $q(t)$ and $\gamma_o(t)$ therefore becomes most simple and straightforward using deterministic analysis since all functions are periodic. One can also obtain the discrete output spectrum $P_o(k\omega_0)$ by taking Fourier series expansion of $\gamma_o(t)$ at $\omega_o = \frac{2\pi}{t_o}$.

Note that we must keep $\gamma_I(t) \geq 0$, $\forall t$, which means that the DC component always exists in the input spectrum. Without causing information loss by control, the same DC component will appear in the output spectrum. For simplicity, we will neglect such a DC component in the definition of all our power spectral functions. Further, since the power spectral functions are central symmetric, it is only necessary to consider $k > 0$.

To facilitate our performance measurement in frequency domain, let us consider a single sinusoidal input, defined by

$$\gamma_I(t) = \overline{\gamma}(1 + \cos\omega_0 t) \qquad (3)$$

Essentially, the sinusoidal frequency ω_0 can be used to describe the input correlation properties. The lower the ω_0 in frequency domain, the slower the time variation of the input process, and so the higher the input correlation is in time domain. The corresponding power spectrum $P_I(k\omega_0)$ is a single impulse function at $k = 1$. In appendix we derive at $T = 0$ for the zero pool system

$$\overline{q}_I = \frac{1}{\omega_0\Delta}\,\frac{F(\rho)}{1-\rho} \qquad (4)$$

where $F(\rho)$ is an intermediate function only dependent on ρ. Note that ω_0 is defined in radian frequency and Δ is equivalent to the mean service time. It is obvious that the time variation of the input process, which is represented by ω_0 in frequency domain, needs to be measured in unit of mean service time for queueing analysis [10]. This is equivalent to normalize ω_0 by $\frac{1}{\Delta}$. A key property explored in (4) is that $\overline{q}_I \propto \frac{1}{\omega_0\Delta}$. That is, the mean queue response is proportional to the inverse of sinusoidal input frequency. As one will see in section 4, this can be regarded as the most inherent property of queue response to input spectrum. For application of voice and video transmission in high speed network, most of their input powers are in low-frequency band [10], which in our case implies that $\frac{1}{\omega_0\Delta}$ can be substantially large.

To a large extent one may view the single sinusoidal input as a test signal to measure the control system response. As will be seen shortly, many intrinsic properties of input rate control can therefore be revealed by such a simple deterministic analysis, which otherwise can hardly be exposed by stochastic analysis due to the complexity involved.

4 Performance Trade-Off
- *Deterministic Analysis*

In this section we use a single sinusoidal input to explore the performance trade-off in the design of leaky-bucket system for input rate control. Without loss of generality, we assume $\overline{\gamma} = 1$ and so $\rho = \Delta$ to represent the token generation interal. Hence, there is only one source parameter, ω_0, which is used to represent the input power spectral property.

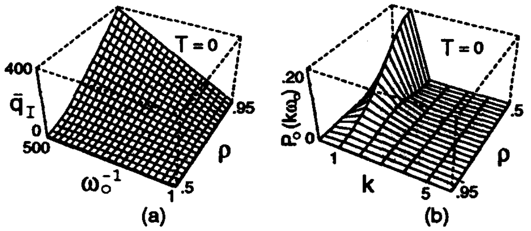

Figure 3: Control performance in function of ρ at $T = 0$: (a) mean queue length (b) output spectral envelope

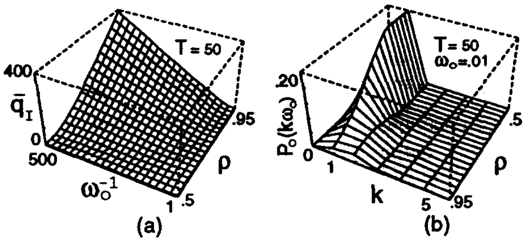

Figure 4: Control performance in function of ρ at $T = 50$: (a) mean queue length (b) output spectral envelope at $1/\omega_0 = 100$

We first consider a zero pool system at $T = 0$. Figure 3a shows the mean queue length performance in function of ρ and $1/\omega_0$. By definition, ρ is also the utilization factor of the control system. This is why the queue response is always improved by the reduction of ρ. On the other hand, as described by (4) at each given ρ, the mean queue response increases linearly with $1/\omega_0$. Plotted in figure 3b is the envelope of the corresponding output power spectrum in function of ρ. From the appendix one can verify that the shape of $P_o(k\omega_0)$ is independent on the frequency unit ω_0 when $T = 0$. Obviously, increasing ρ will reduce the output spectrum in low frequency band. There are two extremes. One is at $\rho = 1$, where $P_o(k\omega_0)$ will contain the DC component only since the buffer is never empty and so $\gamma_o(t)$ becomes constant. One is for ρ to be sufficiently small, such that the buffer is always empty which yields $P_o(k\omega_0) = P_I(k\omega_0)$. For the single sinusoidal input defined in (3), the buffer will always be empty at $\rho = 0.5$. This is why the output spectrum, as $\rho \rightarrow 0.5$ in figure 3, is gradually shifted to a single impulse function at $k = 1$ which is equal to the input spectrum. A clear tradeoff exists between $\overline{q}_I$ and $P_o(\omega)$ through the adjustment of control parameter ρ. Any reduction of $\overline{q}_I$ will cause the increase of $P_o(\omega)$ in low frequency band, and vice versa. Similar observation is made in figure 4 at $T = 50$, except for the shape of $P_o(k\omega_0)$ dependent on ω_0 when $T > 0$.

Remark 1: Via input buffering, more powers in low frequency band are reduced and shifted to high frequency band as the token generation interval ρ increases.

To study the impact of T, we fix ρ at 0.8. Figure 5 shows the mean queue response to sinusoidal input in function of T and $1/\omega_0$. It is clear that the token pool size T has to be sufficiently large in order to reduce the mean queue length when $1/\omega_0$ is high (i.e., when more input powers are in low frequency band). The corresponding output spectrum is also displayed in figures 5b and 5c with respect to $1/\omega_0 = 100$ and 500.

Remark 2: Via token pooling, more powers remain in low frequency band as the pool size T increases.

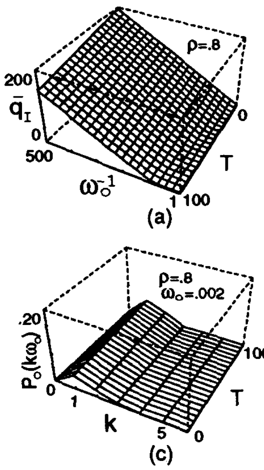

Figure 5: Control performance in function of T at $\rho = 0.8$: (a) mean queue length (b) output spectral envelope at $1/\omega_0 = 100$ (c) output spectral envelope at $1/\omega_0 = 500$

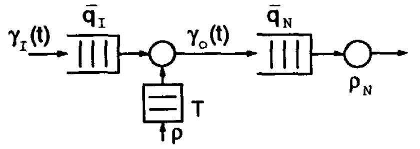

Figure 6: Input control with network queueing

As one can see, by proper selection of ρ and T one can always trade more input queueing delay for less output powers in low frequency band, or vice versa. Nevertheless, the inherent nature of queue response to input frequency, i.e., $\bar{q}_I \propto 1/\omega_0$, is basically unchanged at each given ρ and T. In other words, the selection of ρ and T for the design of input rate control must be strongly dependent on input spectrum.

Once we understand the performance trade-off by individual selection of ρ and T, one may raise a question on their joint adjustment. Consider that a practical design of input rate control is always subject to some delay constraint. Choosing a fixed mean queue length $\bar{q}_I$ to be the subjective condition, our objective in the joint adjustment of ρ and T is to minimize the delay within network. For simplicity we use a single queue to represent the network. Further, to isolate the input control from rest traffic in network, all the arrivals are assumed to be the departures from the input control system, represented by $\gamma_o(t)$ in figure 6. Denote the mean queue length of the network by $\bar{q}_N$ and its utilization factor by ρ_N, respectively. Since $\gamma_o(t)$ is periodic, the analysis of $\bar{q}_N$ based on $\gamma_o(t)$ is just like the analysis of $\bar{q}_I$ based on $\gamma_I(t)$.

Figure 7a shows the solution of ρ in function of T and $1/\omega_0$, under a subjective condition $\bar{q}_I = 20$. Note that in order to keep $\bar{q}_I = 20$ we must have a sufficiently small ρ, or a significantly large T, especially when more input powers are in low frequency band (i.e., when $1/\omega_0$ is large). Choose $\rho_N = 0.8$. Also plotted in figure 7b is the overall performance

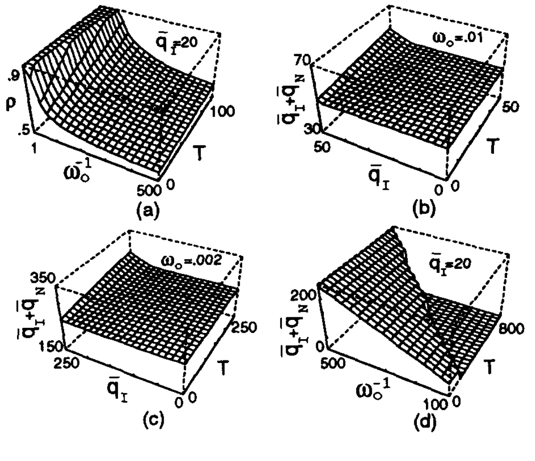

Figure 7: Control performance in function of T subject to a fixed $\bar{q}_I$ at $\rho_N = 0.8$: (a) solution of ρ at $\bar{q}_I = 20$ (b) overall queue $\bar{q}_I + \bar{q}_N$ at $1/\omega_0 = 100$ (c) overall queue $\bar{q}_I + \bar{q}_N$ at $1/\omega_0 = 500$ (d) overall queue $\bar{q}_I + \bar{q}_N$ in function of T and $1/\omega_0$ for $\bar{q}_I = 20$

$\bar{q}_I + \bar{q}_N$ in function of T subject to a fixed $\bar{q}_I$ when $1/\omega_0 = 100$. Note that T is jointly adjusted with ρ under each subjective condition $\bar{q}_I \in [0, 50]$. Clearly, the overall performance $\bar{q}_I + \bar{q}_N$ is basically independent on the joint adjustment of ρ and T. Similar observation is made in figure 7c at $1/\omega_0 = 500$. That is, the overall performance is essentially captured by the source parameter $1/\omega_0$ and the network parameter ρ_N, and it can hardly be changed by input rate control. This acts like a conservation law for the design of input rate control system. Hence, there is a fundamental limit to input rate control.

Remark 3: The overall queueing performance is inherently determined by both input source and network environment; the function of input rate control is simply to trade more input queueing for less network queueing. Further, the network performance is insensitive to the joint adjustment of ρ and T; for simplicity one may design a single parameter control system by having $T = 0$.

Also depicted in figure 7d is the sensitivity of $\bar{q}_N + \bar{q}_I$ to both T and $1/\omega_0$ for a fixed $\bar{q}_I$ (= 20). Note that T cannot be too large for small $1/\omega_0$ in order to keep $\bar{q}_I = 20$. The overall queueing performance is much dependent on the input powers in low frequency band. Again, we have seen the general trend $\bar{q}_I + \bar{q}_N \propto 1/\omega_0$. That is, the more the input powers in low frequency band, the longer the queue will be. Essentially, the only way to reduce $\bar{q}_N$ by input rate control is to proportionally lift $\bar{q}_I$. In other words, unless the source has a large input buffer and can tolerate excess delay, the input rate control can hardly improve the network performance. This is true especially when we consider that the low-frequency powers are the main cause to drive the network to congestion [9],[10]. Hence, such input rate control cannot be effectively applied to real-time traffic like voice and video due to the stringent service time constraint [15], but it can be used effectively to regulate nonreal-time traffic to avoid network congestion.

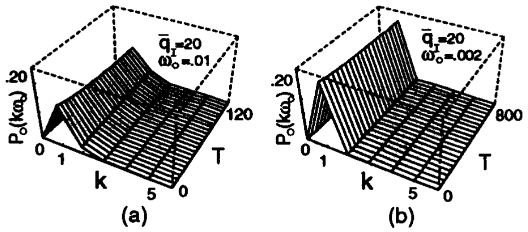

Figure 8: Output spectral envelope in function of T subject to $\bar{q}_I = 20$ (a) at $1/\omega_0 = 100$ (b) at $\omega_0 = 500$

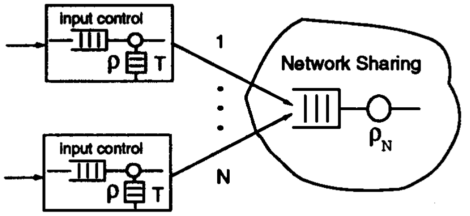

Figure 9: Input control in multiplexing of N homogeneous sources

Figure 8 shows the output spectrum $P_o(k\omega_0)$ at $1/\omega_0 =$ 100 and 500, with respect to the control system in figure 7 under $\bar{q}_I = 20$. As one can see, the output spectrum $P_o(k\omega_0)$ is always dominated by the single input sinusoidal frequency ω_0 at $k = 1$. The value of $P_o(k\omega_0)$ at $k = 1$ slightly reduces as T increases at $1/\omega_0 = 100$, while it basically remains unchanged by T at $1/\omega_0 = 500$. We then come across a question: why the network queue $\bar{q}_N$ is unaffected by the reduction of $P_o(k\omega_0)$ in low frequency band (i.e., at $k = 1$) at $1/\omega_0 = 100$. This is due to the effect of phase interference amongst the harmonic frequencies [9]. It is known that the power spectrum is phase blind. In stochastic modeling, the phase spectrum represents the third degree property. In our case, the output rate function $\gamma_o(t)$ cannot be fully recovered from $P_o(k\omega_0)$ without phase spectrum. As found in [9] [10], however, the impact of input phase spectrum on queue is much less significant than that of input power spectrum. In practice, the statistic measurement of phase spectrum is also much more difficult than that of power spectrum. Therefore, without knowledge of higher degree output properties, the power spectrum can generally be used to measure the control effectiveness as found in figures 3-5.

So far we have separated the input control system from rest traffic in network. From user's point of view, the performance should be measured by end-to-end delay, including the ***input delay*** for control and the ***network delay*** for multiplexing. Similarly, we use a single queue to represent network, which is now shared by N homogeneous sources as described in figure 9. N can be very large in practice due to the large disparity between input source generation rate and network link transmission rate. For simplicity, the same input control parameters, T and ρ, are assumed to apply at each source. Note that the function of input rate control is to reduce the network delay (or congestion) at the expense of increasing its own input delay. On the other hand, simply based on the large number law, one will expect that the network queueing is more efficient than the input queueing especially when N is large. Let us now examine the control trade-off between the average input delay and network delay, denoted by d_I and d_N, Assume the network queue is always loaded at $\rho_N = 0.8$. There are four design parameters: ω_0, T, ρ and N. Figure 10 shows the results of d_I and d_N in function of each individual parameter. Except for the parameter to be tuned, the rest three parameters are fixed accordingly by $1/\omega_o = 100$, $\rho = 0.8$, $T = 20$ and $N = 20$.

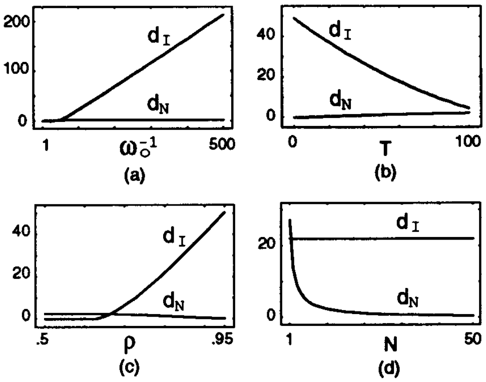

Figure 10: Performance of input delay and network delay: (a) in function of $1/\omega_0$, (b) in function of T, (c) in function of ρ, (d) in function of N, where the rest parameters are fixed accordingly by $1/\omega_0 = 100$, $T = 50$, $\rho = 0.8$, $N = 20$ and $\rho_N = 0.8$

Figure 10a shows that, once the input control ρ and T are fixed, the input delay can be much increased as more input powers are in low frequency band, while the network delay is basically unaffected. In other words, the input rate control does have the effect of blocking low frequency powers at network entry point, which otherwise may cause nodal congestions in network. But, this is done at the high expense of input delay. The results in figures 10b, 10c further indicate that the input delay can also be much increased by tightening the input rate control (as to reduce T or increase ρ), while the improvement on network delay is negligible. It is obvious that the optimal solution for the minimum of $d_I + d_N$ is to entirely remove the input rate control. Again, this is caused by the disparity between source generation rate and link transmission rate, measured by $N > 1$. The input delay, as compared with the network delay, is much more sensitive to ω_0, ρ and T for large N. Figure 10d shows the network delay improvement as N increases, with respect to the given values of ω_0, ρ and T.

Remark 4: Due to the large disparity between source generation rate and link transmission rate, a stringent input rate control may unnecessarily increase the user end-to-end delay by significant amount.

The above argument of having no input rate control is made purely from delay performance's point of view. Of course, many other important factors need to be considered in practical system design. For example, excessive bursty input generated by a source can be highly unpredictable. It is then

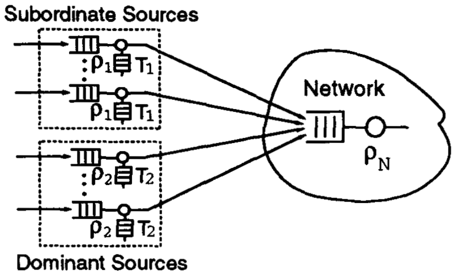

Figure 11: Input control in multiplexing of dominant and subordinate sources

essential to provide a direct and tight feedback loop from congestion point to input source for control purposes. Consider that the time for relaying congestion information from network internode to input source can be too long. One effective way is to simply block highly unpredicted excessive traffic at network entry points via input rate control. The performance trade-off explored here provides us a guiding principle for the design of effective input rate control.

For integration of heterogeneous sources, we consider the multiplexing of two diverse traffic types as shown in figure 11. Each type consists of five homogeneous sources. The input rate control for source type 1 is defined by (ρ_1, T_1), and for source type 2 by (ρ_2, T_2). The average source input queue is denoted by $\overline{q}_{I1}$ and $\overline{q}_{I2}$, respectively. The network queue $\overline{q}_N$ is used to measure the congestion. That is, the smaller the $\overline{q}_N$, the less the probability for network to congest. Again, the input function at each source is assumed to be a single sinusoidal with identical amplitude. The diversity of the two traffic types is characterized by the significant difference of the two sinusoidal frequencies ω_1 and ω_2. Choose $1/\omega_2 >> 1/\omega_1$ so that the type 2 traffic is dominant and the type 1 traffic is subordinate. In the example we have $1/\omega_1 = 100$ and $1/\omega_2 = 1000$. Unless otherwise specified, the control parameters are assigned by $(\rho_1, T_1) = (0.8, 50)$ and $(\rho_2, T_2) = (0.6, 50)$. The network is loaded at $\rho_N = 0.8$. Figures 12a,b show the effect of type 1 traffic control on both input and network queues, in function of ρ_1 and T_1. It is obvious that the network performance can hardly be improved by the control of the subordinate traffic type 1. On the contrary, as shown in figure 12c,d, the input control on the dominant traffic type 2 can significantly reduce the network queue (but at the high expense of input queue).

Remark 5: While input rate control is ineffective in high frequency band, it will be effective in low frequency band if and only if the user has a large storage capacity to tolerate excess delay at network entry point.

Let us now study the impact of peak rate control on the two priority output streams $P_H(\omega)$ and $P_L(\omega)$ in figure 2. For sinusoidal input in (3) at $\overline{\gamma} = 1$ we have $\gamma_{\max} \in (1/\rho, 2)$ for effective peak rate control. Again, the input control ρ and T are designed under the constraint $\overline{q}_I = 20$. Here we fix T at 50 while ρ is adjusted in function of $1/\omega_0$ to satisfy $\overline{q}_I = 20$. As in figure 6, the network is symbolicly represented by a single queue which is loaded by the control output only. Let the queue be loaded at $\rho_N = 0.8$ when both priority output streams are accepted by the network. Figure 13a shows the

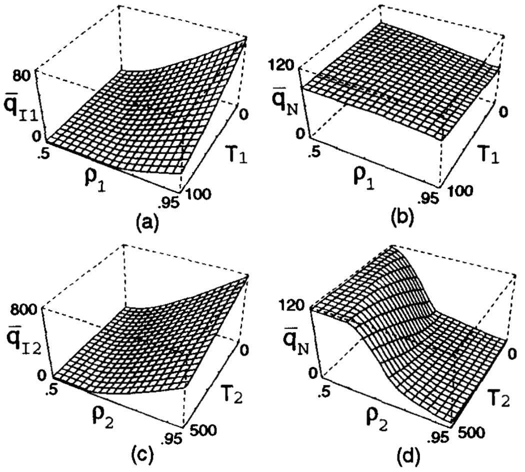

Figure 12: Performance of input delay and network delay: (a,b) control on subordinate sources at $1/\omega_0 = 100$ in function of T_1 and ρ_1, (c,d) control on dominant sources at $1/\omega_0 = 1000$ in function of T_2 and ρ_2, where the rest parameters are fixed by $\rho_1 = 0.8, T_1 = 50, \rho_2 = 0.6, T_2 = 50$ and $\rho_N = 0.8$ accordingly.

overall queueing performance $\overline{q}_I + \overline{q}_N$ in function of $1/\omega_0$ and $\gamma_{\max}$. Note that we must have $\gamma_{\max} \geq 1/\rho$ where ρ is adjusted by $1/\omega_0$ to keep $\overline{q}_I = 20$. Also displayed in figure 13b is the corresponding average loss rate L, caused by the network blocking of the entire low priority stream. Since no traffic is in low priority at $\gamma_{\max} = 2$, the loss rate becomes zero while the queue reaches its maximum for each given ω_0. As $\gamma_{\max}$ decreases, the loss rate arises while the queue falls. Nevertheless, at each given $\gamma_{\max}$ we find the same basic trend $\overline{q}_I + \overline{q}_N \propto 1/\omega_0$ in figure 13a. The same behavior can be found as we change the subjective conditions $\overline{q}_I$ and T. It means that, when more input powers are in low frequency band, the only way to reduce $\overline{q}_N$ is to increase the loss rate L for a fixed input queue constraint $\overline{q}_I$. Both priority power spectra, $P_H(k\omega_0)$ and $P_L(k\omega_0)$, are shown in figures 13c,d in function of $\gamma_{\max}$ at $1/\omega_0 = 200$, in association with the original output spectrum $P_o(k\omega_0)$ in figure 13e. Note that both $P_H(k\omega_0)$ and $P_L(k\omega_0)$ are dependent each other. As is found, via peak access rate control one can shift some of the low frequency powers from the high priority spectrum to the low priority one.

5 Performance Trade-Off
- *Stochastic Analysis*

The inherent properties of input rate control, explored in the above section by the deterministic analysis, can also be obtained by the stochastic analysis but with much more complexities. As is recalled in section 2, for stochastic queueing analysis we need the input rate process $\tilde{\gamma}_I(t)$ to be modulated by Markov chain, described by $\{Q, \vec{\gamma}\}$. Q is an $N \times N$ state transition rate matrix, which is assumed to be diagonalizable. $\vec{\gamma}$ is a vector for the input rate associated with each individual state, given by $\vec{\gamma} = [\gamma_0, \gamma_1, ..., \gamma_{N-1}]$. By spectral

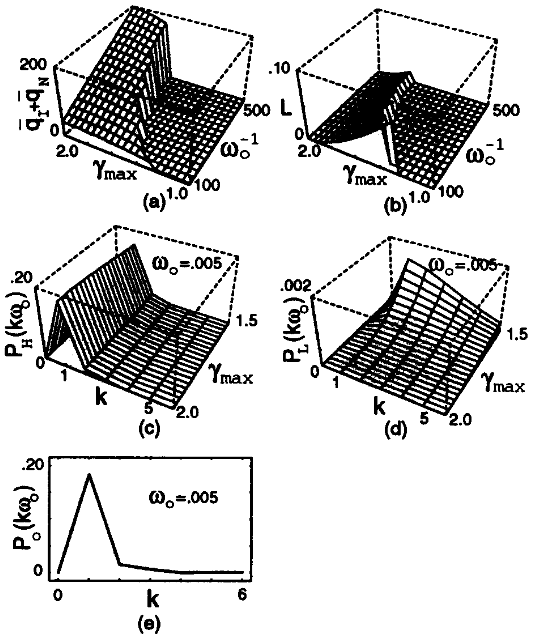

Figure 13: Performance in function of peak access rate γ_{max} (a) overall queue (b) loss rate (c) high priority output spectrum (d) low priority output spectrum (e) total output spectrum (at $1/\omega_0 = 200$)

decomposition,

$$Q = \sum_{l=0}^{N-1} \lambda_l \vec{g}_l \vec{h}_l \tag{5}$$

where both $\vec{g}_l$ and $\vec{h}_l$ are the right and left eigenvectors for the eigenvalue λ_l. For stationary Markov chain, one of the eigenvalues, denoted by λ_0, must be zero, while the rest eigenvalues satisfy $\text{Re}\{\lambda_l\} < 0$. Its input power spectrum is then readily expressed by [10]

$$P_I(\omega) = 2\pi\overline{\gamma}^2\delta(\omega) + \sum_{l=1}^{N-1} \psi_l b_l(\omega) \tag{6}$$

with

$$\psi_l = \sum_k \sum_n \gamma_k \gamma_n g_{kl} h_{nl} \pi_k, \qquad b_l(\omega) = \frac{-2\lambda_l}{\lambda_l^2 + \omega^2},$$

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} b_l(\omega) d\omega = 1 \tag{7}$$

where g_{kl} is the k-th element of $\vec{g}_l$ and h_{nl} is the n-th element of $\vec{h}_l$. $\overline{\gamma}$ is the average input rate. $2\pi\overline{\gamma}^2\delta(\omega)$ is the DC term for non-negative input rate. Each eigenvalue component, $b_l(\omega)$ in (7), represents a bell-shape curve located at the central frequency $\omega_l = \text{Im}\{\lambda_l\}$ and weighted by ψ_l. The shape of each bell, before being weighted, is measured by its half-power bandwidth $BW_l = -2\text{Re}\{\lambda_l\}$. Both ω_l and BW_l are defined in radian frequencies.

For simplicity we consider a two-state Markov chain, alternating between ON and OFF periods, which is commonly used as a building block for construction of voice and video sources. Define

$$Q = \begin{bmatrix} -\alpha_{off} & \alpha_{off} \\ \alpha_{on} & -\alpha_{on} \end{bmatrix}, \quad \vec{\gamma} = \begin{bmatrix} 0 \\ \gamma_{on} \end{bmatrix}^T,$$

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -\alpha_{off} - \alpha_{on} \end{bmatrix} \tag{8}$$

where γ_{on} is the input rate while in ON-state. For two-state Markov chain, λ_1 must be real and so we have $\omega_1 = 0$ and $BW_1 = -2\lambda_1$. Hence, the single bell on its power spectrum is always centered at zero-frequency, expressed by

$$P_I(\omega) = 2\pi\overline{\gamma}_1{}^2\delta(\omega) + \frac{\overline{\gamma}_1{}^2 C_{\gamma_1}^2 BW_1}{(BW_1/2)^2 + \omega^2} \tag{9}$$

with $\overline{\gamma}_1 = \epsilon\gamma_{on}$ where ϵ is the source activity factor given by $\alpha_{on}^{-1}/(\alpha_{on}^{-1} + \alpha_{off}^{-1})$. $C_{\gamma_1}^2$ is the squared coefficient of input rate variation, equal to $(1-\epsilon)/\epsilon$.

The power spectrum is directly additive for superposition of independent input processes, except for the DC term. Denote the power spectrum of the k-th independent input process by $P_{Ik}(\omega)$ and its DC term by $2\pi\overline{\gamma}_k{}^2\delta(\omega)$, where $\overline{\gamma}_k$ represents the average input rate. Excluding the DC term, one can generally write

$$P_I(\omega) = \sum_{k=1}^{M} P_{Ik}(\omega)$$

Since the DC term in $P_I(\omega)$ is physically determined by the superimposed average input rate, it is given by $2\pi\overline{\gamma}^2\delta(\omega)$ with $\overline{\gamma} = \sum_{k=1}^{M}\overline{\gamma}_k$. For the k-th process defined by $\{Q_k, \vec{\gamma}_k\}$, the overall input process will be described by $\{Q, \vec{\gamma}\}$ with

$$Q = Q_1 \oplus Q_2 \oplus \ldots \oplus Q_M, \qquad \vec{\gamma} = \vec{\gamma}_1 \oplus \vec{\gamma}_2 \oplus \ldots \oplus \vec{\gamma}_M. \tag{10}$$

Hence, for the superposition of M homogeneous 2-state Markov chains we get

$$P_I(\omega) = 2\pi M^2\overline{\gamma}_1{}^2\delta(\omega) + \frac{M\overline{\gamma}_1{}^2 C_{\gamma_1}^2 BW_1}{(BW_1/2)^2 + \omega^2} \tag{11}$$

Assume that the leaky-bucket system has finite buffer size of K_I in packets. As was shown in [2] [5], the mathematical modeling of leaky-bucket system, defined by $\{\rho, T, K_I\}$, is equivalent to that of single-server queue system with buffer size $K_I + T$ and loaded by ρ. The corresponding service rate is fixed at $\mu = 1/\Delta$. With a $\{Q, \vec{\gamma}\}$ Markovian input process, one can thus characterize the leaky-bucket system by a finite QBD process. The state of such a QBD process is defined by levels and phases. Each level represents a buffer occupancy for $-T \leq \tilde{q}(t) \leq K_I$; each phase on a level corresponds to a state of input Markov chain. The overall QBD transition matrix will then be expressed in the following block tri-diagonal form:

$$G = \begin{bmatrix} Q - \underline{\alpha} & \underline{\alpha} & & & \\ \underline{\beta} & Q - \underline{\beta} - \underline{\alpha} & \underline{\alpha} & & \\ & \ddots & \ddots & \ddots & \\ & & \underline{\beta} & Q - \underline{\beta} - \underline{\alpha} & \underline{\alpha} \\ & & & \underline{\beta} & Q - \underline{\beta} \end{bmatrix} \tag{12}$$

where G is finite and irreducible. The steady state distribution vector $\vec{\pi}$ will be the unique solution to the equations:

$$\vec{\pi}G = 0, \quad \vec{\pi}\vec{e} = 1$$

where $\vec{e} = [1, 1, ..., 1]^T$. $\underline{\alpha}$ represents the queue expansion rate matrix while in overload, given by $diag[(\tilde{\gamma}_I - \mu)^+]$ where $\tilde{\gamma}_I = \gamma_k$ when the input Markov chain is in phase k with $0 \leq k \leq N$. Similarly, $\underline{\beta}$ is the queue reduction rate matrix while in underload, equal to $diag[(\mu - \tilde{\gamma}_I)^+]$. The symbol $(\cdot)^+$ denotes the larger of 0 or its argument. Notice that here $\tilde{q}(t)$ assumes a discrete state space, unlike the continuous fluid flow model proposed in [12]. This discrete model was originally introduced and well examined in [13],[14] and referred to as Markov-modulated fluid-flow model. The queueing performance is measured by the mean queue size $\overline{q}_I$ and the average packet loss rate L_I for each given ρ and T, in response to the input power spectrum $P_I(\omega)$. The results are obtained by using the QBD-Folding algorithm developed in [13][14].

Denote a QBD state in level j at phase k by (j, k). The transition probability from (j, k) to (n, m) in time interval t is then defined by $\left[e^{Gt}\right]_{(j,k),(n,m)}$. The output rate process $\tilde{\gamma}_o(t)$ is also a stationary random process, which is defined on the QBD state space by

$$\tilde{\gamma}_o(j, k) = \begin{cases} \mu & \text{if } j > 0 \\ \min\{\gamma_k, \mu\} & \text{if } j = 0 \\ \gamma_k & \text{else} \end{cases} \tag{13}$$

where μ is the service rate and γ_k is the input rate in phase k. The output rate autocorrelation function will thus be expressed by

$$R_o(t) = \sum_{j,k}\sum_{n,m} \tilde{\gamma}_o(j, k)\tilde{\gamma}_o(n, m) \left[e^{G|t|}\right]_{(j,k),(n,m)} \pi_{j,k}$$

where $\pi_{j,k}$ is the steady state probability in state (j, k), i.e., $\pi_{j,k} \in \vec{\pi}$. Numerically we always find that G is diagonalizable. By spectral decomposition as done in (5),

$$G = \sum_{i,l} \lambda_{(i,l)}\vec{g}_{(i,l)}\vec{h}_{(i,l)} \tag{14}$$

we get

$$R_o(t) = \sum_{i,l} \psi_{(i,l)} e^{\lambda_{(i,l)}|t|} \tag{15}$$

with

$$\psi_{(i,l)} = \sum_{j,k}\sum_{n,m} \tilde{\gamma}_o(j, k)\tilde{\gamma}_o(n, m) g_{(j,k)(i,l)} h_{(n,m)(i,l)} \pi_{j,k} \tag{16}$$

where $g_{(j,k)(i,l)}$ is the (j, k)-th element of $\vec{g}_{(i,l)}$ and $h_{(n,m)(i,l)}$ is the (n, m)-th element of $\vec{h}_{(i,l)}$. Taking Fourier transform of $R_o(t)$, we obtain the output rate power spectrum in frequency domain,

$$P_o(\omega) = \sum_{i,l} \frac{-2\psi_{(i,l)}\lambda_{(i,l)}}{\lambda^2_{(i,l)} + \omega^2} \tag{17}$$

In numerical study we consider four homogeneous 2-state Markov chains with its power spectrum in (9), defined by

$$(BW_1, \overline{\gamma}_1, C_{\gamma_1}) = (0.01, 0.25, 1.22)$$

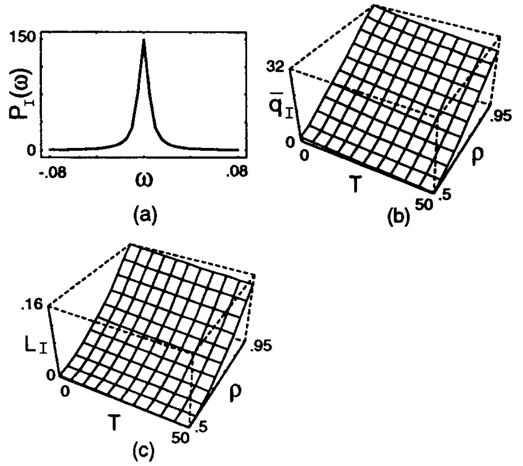

Figure 14: Stochastic control performance in function of ρ and T: (a) input power spectrum (b) mean queue length (c) average loss rate

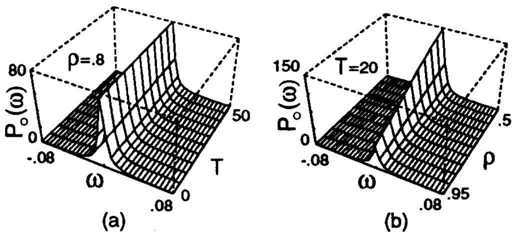

Figure 15: Output spectrum by stochastic control analysis (a) in function of T at $\rho = 0.8$ (b) in function of ρ at $T = 20$

as is also plotted in figure 14a. To avoid the numerical difficulty we choose the buffer size $K_I = 80$ and limit the pool size in $0 \leq T \leq 50$. Figures 14b and 14c show the performance of $\overline{q}_I$ and L_I in function of T and ρ for the leaky-bucket system. Exactly as we have observed in figures 3-5 by the deterministic analysis, both queue and loss performances are improved by the increase of T or the reduce of ρ. T has to be sufficiently large to improve the queueing performance since in our case most input powers are in low-frequency band, which is consistent with what we have found in figure 5a. Also displayed in figure 15 are the output spectrum in function of T at $\rho = 0.8$, as well as in function of ρ at $T = 20$. Just as what has been stated in remarks 1 and 2, more input powers are reduced in low frequency band at the output as the control is to be strengthened. By comparison, the effect of ρ on output power spectrum in figure 15b is very much like the one found in figures 3b,4b by the deterministic analysis. Note that the output spectrum at $\rho = 0.5$ in figure 15b is about identical to the input spectrum in figure 14a. Similar comparison is drawn between figure 15a and figure 5c for the effect of T when input powers are in low-frequency band.

For the input control with network queueing in figure 6, we introduce an output rate vector $\vec{\gamma}_o = [\tilde{\gamma}_o(j, k)]$ as defined

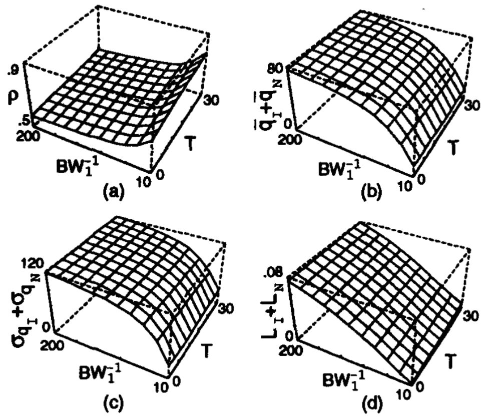

Figure 16: Stochastic control performance in function of T and BW_1^{-1} subject to $\overline{q}_I = 10$ (a) solution of ρ (b) overall mean queue length (c) overall queue standard deviation (d) overall average loss rate, with respect to $\rho_N = 0.8$, $K_I = 80$ and $K_N = 255$

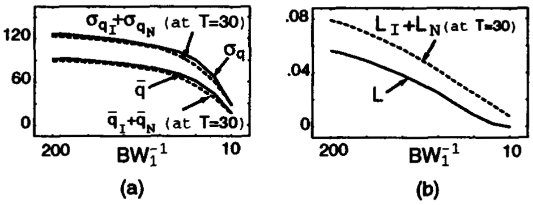

Figure 17: Overall performance comparison between controlled and non-controlled systems in function of BW_1^{-1} at $\rho_N = 0.8$, $K_I = 80$ and $K_N = 255$

by (13). Just like the input rate process $\{Q, \tilde{\gamma}\}$, the output rate process can be exactly characterized by a Markov modulated process, described by $\{G, \tilde{\gamma}_o\}$. Note that the state space of $\tilde{\gamma}_o(t)$ is $(K_I + T + 1) \times (M + 1)$, identical to that of the above entire QBD process. One can then formulate the network queue by another finite QBD process, which has the same structure as in (12) for the input queue process, except with a much larger phase space. Denote the network buffer size by K_N and the constant service rate by μ_N. We have $\rho_N = \overline{\gamma}/\mu_N$, which is the network utilization factor if no packet loss occurs at both input and network queues. The network performance will be measured by the mean queue length, queue standard deviation and average loss rate, denoted by $\overline{q}_N$, σ_{qN} and L_N respectively. As we had in figure 7, one can tune the input control parameters (ρ, T) under a subjective condition $\overline{q}_I$, and then evaluate $(\overline{q}_N, \sigma_{qN}, L_N)$ by the QBD-Folding algorithm. In numerical study we follow the same example used in figures 14,15 for the construction of $\{G, \tilde{\gamma}_o\}$, except to keep $0 \leq T \leq 30$ for simplicity. The size of the phase space for the network QBD process will then be in the range of 405 to 555 for $0 \leq T \leq 30$ at $K_I = 80$. Further, for the four i.i.d. 2-state Markov chains of the original input process we still fix $(\overline{\gamma}_1, C_{\gamma_1})$ at (0.25, 1.22), but change the central-bandwidth BW_1 to reflect the variation of the original input power spectrum $P_I(\omega)$ in figure 14a. BW_1 is like the sinusoidal frequency ω_0 in section 4. The smaller the BW_1, the more the input powers are shifted from high-frequency band to low-frequency band, while the average input power, given by $M\overline{\gamma}_1{}^2 C_{\gamma_1}^2$ in (11), remains unchanged.

Similar to figure 7a, we show in figure 16a the solution of ρ in function of T and BW_1^{-1} subject to a fixed input mean queue length $\overline{q}_I = 10$. The reason we choose $\overline{q}_I = 10$ is to keep the average input loss rate L_I relatively small. Assume $K_N = 255$ and $\rho_N = 0.8$ for the network environment. Displayed in figure 16b are the corresponding overall mean queue length $\overline{q}_I + \overline{q}_N$, which is analogous to the results in figure 7d. Also in figures 16c,d we plot the solution of $\sigma_{qI} + \sigma_{qN}$ and $L_I + L_N$. It is obvious that all these results are virtually independent on T. As already indicated by remark 3, the overall performance is insensitive to the joint adjustment of ρ and T. On the other hand, both queue and loss performances are very sensitive to the input power spectrum. The higher the BW_1^{-1} is, the more the input powers are in low frequency band, and so the larger the queue and the higher the loss rate. The effect of BW_1 is somewhat like the effect of ω_0 on queueing performance (although it is not linear as found in figure 7d).

Now, let us suppose that one can remove the input control in figure 6, and so the two separate queues are merged into a single one. We would then have a single queue system with capacity $K = K_I + K_N$ and service rate μ_N to support the original input process $\{Q, \tilde{\gamma}\}$. Such a non-controlled system is also measured by the mean queue length, queue standard deviation and average loss rate, denoted by $(\overline{q}, \sigma_q, L)$. In figure 17a we compare the results of both $(\overline{q}_I + \overline{q}_N, \sigma_{qI} + \sigma_{qN})$ and $(\overline{q}, \sigma_q)$ in function of BW_1^{-1} for the controlled and non-controlled systems. The controlled system is designed by $T = 30$ under the subjective condition $\overline{q}_I = 10$. The total buffer capacity of the two systems are identical. Since the controlled system has two separate finite-buffer queues, its average loss rate, as found in figure 17b, is always greater than that of the non-controlled system. This also explains why in figure 17a the queueing performance of the controlled system is slightly better than that of the non-controlled system. Nevertheless, the queueing difference between the two systems is always negligible. This study clearly indicates that the function of input rate control is simply to trade more input queueing for less network queueing. Once again, this is consistent to the remark 3 made by the deterministic analysis. As compared to the deterministic analysis, not only the stochastic analysis is much more difficult but also one can only solve much limited systems.

6 Summary

In this paper we have measured the effectiveness of input rate control in frequency domain. Based on spectral analysis we are able to explore the fundamental limits of input rate control. Four guiding principles have been developed. First, the overall queueing performance is inherently determined by both input source and network environment; the function of input rate control is simply to trade more input queueing for less network queueing. Second, due to the large disparity between source generation rate and link transmission rate, a stringent input rate control may unnecessarily increase the user end-to-end delay by significant amount. Third, while

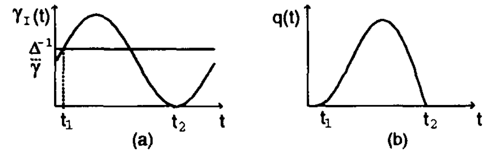

Figure A-1: Queue response to single sinusoidal input (a) input function (b) queue function

input rate control is ineffective in high frequency band, it is effective in low frequency band if and only if the user can tolerate excess delay or loss at network entry point. Fourth, the network performance is insensitive to the adjustment between token pool size and token generation rate for the leaky bucket control system. As one can see, the simple deterministic analysis of input rate control developed here has helped us to gain a great knowledge of performance trade-off for input rate control in high speed network.

Appendix

The continuous fluid flow model at $T = 0$ can be expressed by $\frac{dq}{dt} = \gamma_I(t) - \Delta^{-1}$ for $q(t) \geq 0$ where Δ^{-1} is the service rate. For a single sinusoidal input $\gamma_I(t) = \overline{\gamma}[1 + \cos(\omega_0 t)]$, as described in figure A-1, one can introduce two time indexes, t_1 and t_2, such that $q(t) > 0$ for $t_1 < t < t_2$ and $q(t) = 0$ elsewhere within a sinusoidal period. We then get $q(t) = \int_{t_1}^{t}[\gamma_I(t) - \Delta^{-1}]dt$, which yields

$$q(t) = \frac{\overline{\gamma}}{\omega_0}(\sin\omega_0 t - \sin\omega_0 t_1) - (\Delta^{-1} - \overline{\gamma})(t - t_1) \qquad (A-1)$$

For the total queue accumulation

$$\int_{t_1}^{t_2} q(t)dt = \frac{\overline{\gamma}}{\omega_0^2}(\cos\omega_0 t_1 - \cos\omega_0 t_2) - \frac{\overline{\gamma}}{\omega_0}(t_2 - t_1)\sin\omega_0 t_1$$

$$-\frac{\Delta^{-1} - \overline{\gamma}}{2}(t_2 - t_1)^2$$

The average queue is then given by $\overline{q}_I = \frac{\omega_0}{2\pi}\int_{t_1}^{t_2} q(t)dt$. From $q(t_2) = 0$ in (A-1) we also get

$$t_2 - t_1 = \frac{\rho}{\omega_0(1-\rho)}(\sin\omega_0 t_2 - \sin\omega_0 t_1) \qquad (A-2)$$

Taking the above equation into the queue accumulation, we finally obtain

$$\overline{q}_I = \frac{1}{2\pi\omega_0\Delta}\;\frac{\rho}{1-\rho}[(1-\rho)(\cos\omega_0 t_1 - \cos\omega_0 t_2)$$

$$-\rho\sin\omega_0 t_1(\sin\omega_0 t_2 - \sin\omega_0 t_1) - \frac{\rho}{2}(\sin\omega_0 t_2 - \sin\omega_0 t_1)^2]$$

From $\gamma_I(t_1) = \Delta^{-1}$ one can write $\cos\omega_0 t_1 = \rho^{-1} - 1$. Also, re-express $q(t_2) = 0$ in (A-2) by

$$\frac{\sin\omega_0 t_2 - \sin\omega_0 t_1}{\omega_0 t_2 - \omega_0 t_1} = \frac{1}{\rho} - 1$$

It is obvious that both $\omega_0 t_1$ and $\omega_0 t_2$ are uniquely determined by ρ. One can then generally write $\overline{q}_I = \frac{1}{\omega_0\Delta}\frac{F(\rho)}{1-\rho}$.

References

[1] M. Sidi, W.Z. Liu, I. Cidon and I. Gopal, "Congestion Control Through Input Rate Regulation" Proc. *Globecom'89*, pp. 1764-1768.

[2] A. Berger, "Performance Analysis of a Rate-Control Throttle where Tokens and Jobs Queue," *IEEE J. Sel. Areas Commu.* SAC-9, Feb. 1991, pp. 165-170.

[3] E.P. Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks," *IEEE J. Sel. Areas Commu.* SAC-9, April 1991, pp. 325-334.

[4] M. Butto, E. Cavallero and A. Tonietti, "Effectiveness of the 'Leaky Bucket' Policing Mechanism in ATM Networks," *IEEE J. Sel. Areas Commu.* SAC-9, April 1991, pp. 335-342.

[5] K. Sohraby and M. Sidi, "On the Performance of Bursty and Correlated Sources Subject to Leaky Bucket Rate-Based Access Control Schemes," Proc. *IEEE Infocom'91*, pp. 426-434.

[6] A.I. Alwalid and D. Mitra, "Analysis and Design of Rate-Based Congestion Control of High Speed Networks, I: Stochastic Fluid Models, Access Regulation" *Queueing Systems*, Vol. 9, 1991, pp. 29-64.

[7] K.K. Leung, B. Sengupta and R.W. Yeung, "Queueing Analysis of a Credit Manager for Flow Control of High Speed Networks," Proc. *IEEE Infocom'92*, pp. 2368-2377.

[8] M. Murata, Y. Ohba and H. Miyahara, "Analysis of Flow Enforcement Algorithm for Bursty Traffic in ATM Networks," Proc. of *IEEE Infocom'92*, pp. 2453-2462.

[9] S.Q. Li and C.L. Hwang, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis," Proc. of *IEEE Infocom'92*, pp. 382-394 (which received the Conference Paper Award of Infocom'92).

[10] S.Q. Li and C.L. Hwang, "Queue Response to Input Correlation Functions: Continuous Spectral Analysis," presented at the 7th IEEE Computer Communication Workshop in Hilton Head Island, SC, Oct. 1992 (also submitted to IEEE/ACM Trans. on Networking, July 1992).

[11] J.-C. Bolot and A. Shankar, "Analysis of a Fluid Approximation to Flow Control Dynamics," Proc. *IEEE Infocom'92*, pp. 2398-2407.

[12] D. Anick, D. Mitra and M.M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources," *Bell Sys. Tech. J.*, 61, 1982, pp. 1871-1894.

[13] J. Ye and S.Q. Li, "Analysis of multi-media traffic queues with finite buffer size and overload control - Part I: algorithm," Proc. of *Infocom'91*, April 1991, pp 1464-1474.

[14] J. Ye and S.Q. Li, "Analysis of multi-media traffic queues with finite buffer size and overload control - Part II: Application," Proc. of *Infocom'92*, May 1992, pp. 848-859.

[15] S. Q. Li and H.D. Sheng "Discrete Queueing Analysis of Multi-Media Traffic with Diversity of Correlation and Burstiness Properties", Proc. of *IEEE Infocom'91*, April 1991, pp. 368-381 (also accepted by *IEEE Trans. Commu.*).