Forty-Sixth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 23-26, 2008

Effective Resource Allocation in a Queue: How Much Control is Necessary?

Krishna Jagannathan, Eytan Modiano and Lizhong Zheng

Abstract— **In this paper, we consider a single-server queue with Poisson inputs and two distinct service rates. The service rate employed at any given instant is decided by a resource allocation policy, based on the queue occupancy. We deal with the question of how often control information needs to be sent to the rate scheduler so as to stay below a certain probability of congestion. We first consider some simple Markovian service rate allocation policies and derive the corresponding control rate vs. congestion probability tradeoffs in closed form. However, since a closed form solution is not possible for more general Markov policies, we resort to large deviation tools to characterize the congestion probabilities of various control policies. We also identify a simple 'two-threshold' policy which achieves the best possible tradeoff between rate of control and the decay exponent of the congestion probability. Finally, we also investigate the impact of control errors on the congestion probability of a resource allocation policy.**

I. Introduction

Resource allocation is an essential part of any practical queueing system. Resource allocation involves assigning service rates to various queues in the system, so as to avoid instability effects that lead to large queueing delays, and to ensure other performance objectives such as fairness. Typically, resource allocation policies try to allocate larger service rates to queues that are congested, and vice-versa; see for example [1], [2]. Most systems use a combination of resource allocation and flow control in order to mitigate congestion in the network, and in order to achieve various performance objectives; see [3], [4]. In this paper, we restrict our attention to the server allocation problem, although many of our observations also hold in a flow control setting.

Since the queue lengths can vary widely over time in a dynamic network, effective resource allocation typically requires the exchange of control information between agents that can observe the various queue lengths in the system, and the service rate schedulers which adapt their rates to the varying queues. This control information can be thought of as being a part of the inevitable protocol and control overheads in a network. Intuitively, if the the rate schedulers have very accurate information about the current queue lengths in the system, resource allocation can be performed very effectively by adapting the service rates appropriately. However, furnishing the schedulers with accurate queue length knowledge might involve communicating a lot of control information. Conversely, if the queue length information is rarely furnished to the scheduler, we can expect large queue variance. It turns out that this tradeoff between the efficacy of congestion control and the rate of control information is a fundamental one in most scenarios. Furthermore, these control signals typically also utilize some part of the communication resources of the system. Therefore, it is of interest to characterize the rate at which control information needs to be exchanged in order to achieve a certain congestion probability. It is also of interest to identify control policies which achieve the lowest congestion probability for a given control rate.

Historically, Gallager was among the first to address the important question of protocol overhead in communication networks. In his seminal paper [5], he derives information theoretic lower bounds on the amount of protocol information needed for network nodes to keep track of source and destination addresses, as well as message starting and stopping times. We are, however, concerned with a very specific type of overhead, namely that of queue length information that needs to be communicated periodically to rate schedulers. In a related paper, Perkins & Srikant consider the problem of maximizing the throughput in a flow controlled single server queue, subject to a maximum congestion probability constraint. When the input rate is constrained to lie within a certain range, they show that a 'bang-bang' type solution is optimal [6]. To the best of our knowledge, the tradeoff between the control information rate and congestion performance has not been explicitly dealt with in the literature.

In the present paper, we consider a single server queue with Poisson arrivals and two distinct service rates. The service rate employed at any instant is decided by a control policy, based on the queue occupancy. We first consider some simple Markovian service rate allocation policies and derive the corresponding control rate vs. congestion probability tradeoffs in closed form. Unfortunately, a closed form solution is not possible for more general Markovian policies. We therefore develop a more tractable large deviation characterization of the congestion probabilities for various control policies. These decay exponent tools are used throughout the rest of the paper. We identify a simple 'two threshold policy' which achieves the optimal congestion probability decay exponent for any given rate of control.

In the first part of the paper, we assume that the control information about the queue occupancy is received instantaneously and perfectly by the controllers. However, in many situations such as in wireless networks, the control

This work was supported by NSF ITR grant CCR-0325401, by NSF grant CNS-0626781, and by ARO Muri grant number W911NF-08-1-0238.

The authors are with the Massachusetts Institute of Technology, Cambridge, MA 02139, USA. krishnaj@mit.edu, modiano@mit.edu, lizhong@mit.edu

978-1-4244-2926-4/08/$25.00 ©2008 IEEE

information about the queue lengths may also be subject to channel errors. We investigate the effect of control errors on the congestion control performance of the two threshold policy. Assuming a simple probabilistic model for errors on the control channel, we show the existence of a critical error probability, beyond which the errors in receiving the control packets lead to an exponential worsening of the congestion probability. However, below the critical probability of error, the probability of congestion is of the same exponential order as an error free system.

The remainder of this paper is organized as follows. We describe our queueing system model in Section II. We introduce Markovian control policies in Section III and make some relevant definitions. Section IV deals with the control rate vs. congestion probability tradeoff for some simple Markovian policies. In Section V, we introduce large deviation exponents, and characterize the decay exponents of the congestion probabilities of various control policies. In Section VI, we introduce errors on the control channel, and investigate the effect of control errors on the congestion probability.

II. System Description

Let us first describe a simple model of a queue with service rate control. Fig. 1 depicts a single server queue with Poisson inputs of rate λ. We also assume throughout that the packet sizes are exponentially distributed with mean 1. An observer watches the queue evolution and sends control information to the to the service rate controller, which changes the service rate $S(t)$ based on the control information it receives. The purpose of the observer-controller subsystem is to assign service rates at each instant so as to control congestion in the queue.

For analytical simplicity, we assume that the service rate at any instant is chosen to be one of two distinct values: $S(t) \in \{\mu_1, \mu_2\}$, where $\mu_2 > \mu_1$ and $\mu_2 > \lambda$. The control decisions are sent by the observer in the form of information-less packets. Upon receiving a control packet, the rate controller switches the service rate from one to the other. We focus on Markovian control policies, in which the service rate chosen after an arrival or departure event is only a function of the previous service rate and queue length. Note that due to the memoryless arrival and service time distributions, there is nothing to be gained by using non-Markovian policies.

III. Markovian Control Policies

We begin by defining notions of Markovian control policies, its associated congestion probability, etc.

Let $t > 0$ denote continuous time. Let $Q(t)$ and $S(t)$ respectively denote the queue length and service rate (μ_1 or μ_2) at time t. Define $Y(t) = (Q(t), S(t))$ to be the state of the system at time t. We assign discrete time indices $n \in \{0, 1, 2, \dots\}$ to each arrival and departure event in the queue ("queue event"). Let Q_n and S_n respectively denote the queue length and service rate just after the n^{th} queue event. Define $Y_n = (Q_n, S_n)$. A resource allocation policy assigns service rates S_n after every queue event.

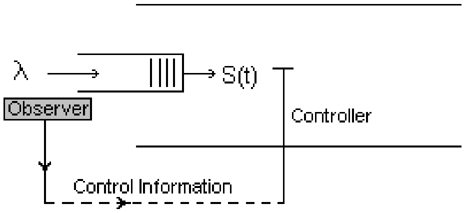

Fig. 1. A single server queue with service rate control.

Definition 1: A control policy is said to be Markovian if it assigns service rates S_n such that

$$\mathbb{P}\{S_{n+1}|Q_{n+1}, Y_n, \dots, Y_0\} = \mathbb{P}\{S_{n+1}|Q_{n+1}, Y_n\}, \quad (1)$$

$\forall n = 0, 1, 2 \dots$.

For a Markovian control policy operating on a queue with memoryless arrival and packet size distributions, it is easy to see that $Y(t)$ is a continuous time Markov process with a countable state space, and that Y_n is the imbedded Markov chain for the process $Y(t)$.

Definition 2: The congestion probability is defined as $\lim_{t\to\infty} \mathbb{P}\{Q(t) \geq M\}$, where M is some congestion limit. We denote it henceforth by $\mathbb{P}\{Q \geq M\}$.

Note that if there is no restriction imposed on using the higher service rate μ_2, it is optimal to use it all the time, since the congestion probability can be minimized without using any control information. However, in a typical queueing system with limited resources, it may not be possible to use higher service rate at all times. There could be a cost per unit time associated with using the faster server, so that it is typically only used when the queue occupancy is high. In this paper, we explicitly impose the constraint that when the queue length is no more than some threshold l, we are forced to use the lower service rate μ_1. If the queue length exceeds l, we are allowed to use the higher rate μ_2 without any additional cost until the queue length falls back to l.

We are now ready to characterize the tradeoff between control rate and congestion performance for some simple Markovian control policies which satisfy the above constraint.

IV. Control rate vs. congestion probability tradeoff

Since we are free to use the higher service rate μ_2 whenever the queue length is larger than l, it is clear that in order to minimize the congestion probability we should switch to the higher service rate whenever the queue length exceeds l, and switch back to the lower rate when it falls back to l. (We will call this the *single-threshold policy*). As an aside, in a system with a cost associated with the use of the faster server, the single threshold policy can be shown to minimize the average cost for any desired congestion probability. However, it suffers from the drawback that it

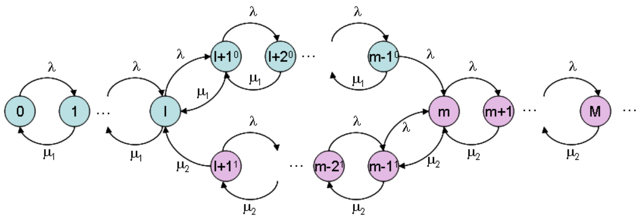

Fig. 2. The Markov process $Y(t)$ corresponding to the two-threshold control policy.

requires frequent exchange of control packets, since the queue length can often toggle between l and $l+1$. It turns out that a simple extension of the single threshold policy gives rise to a family of policies which provide more flexibility with the rate of control.

A. The two threshold policy

As suggested by the name, the service rates in the two threshold policy are switched at two distinct queue length thresholds l and m. Specifically, when the queue length grows past $m = l + k$, the service rate switches to μ_2. Once the higher service rate is employed, it is maintained until the queue length falls back to l, at which time the service rate switches back to μ_1. We will soon see that this 'hysteresis' in the thresholds helps us tradeoff the control rate with the congestion probability.

The Markov process $Y(t)$ corresponding to the two-threshold policy is pictorially depicted in Fig. 2. Note that we use the short hand notation $l + i^0$ and $l + i^1$ in the figure, to denote respectively, the states $(Q(t) = l + i, S(t) = \mu_1)$ and $(Q(t) = l + i, S(t) = \mu_2)$. For queue lengths not more than l, we drop the superscripts altogether, since the service rate for such queue lengths is understood to be μ_1. Similarly, the superscripts denoting the service rate is dropped for queue lengths m and larger, since it is understood to be μ_2. Observe that the case $k = 1$ corresponds to a single threshold policy, where the higher rate is employed if and only if the queue length is strictly greater than l.

Intuitively, as the gap between the two thresholds $k = m - l$ increases for a fixed l, the probability of congestion should increase, whereas the rate of control packets sent by the observer should decrease. It turns out that we can fully characterize the rate-congestion tradeoff for the two-threshold policy in closed form.

1) Control rate vs. congestion tradeoff for the two-threshold policy: Define $\rho_2 = \frac{\lambda}{\mu_2} < 1, \rho_1 = \frac{\lambda}{\mu_1}$, and $\eta_1 = 1/\rho_1$. We can solve for the steady state probabilities for each state in the Markov process represented in Fig. 2.

Let us denote the steady state probabilities of the non-superscripted states in Fig. 2 by p_j, where $j \le l$, or $j \ge m$. Next, denote by p^0_{l+i} (p^1_{l+i}) the steady state probability of the state $l + i^0$ ($l + i^1$), for $i = 1, 2, \ldots, k - 1$. By solving for the steady state probabilities of various states in terms of p_l, we obtain:

$$p_i = p_l \eta_1^{l-i}, \quad 0 \le i \le l, \tag{2}$$

$$p^0_{m-1} = \begin{cases} \frac{1-\eta_1}{1-\eta_1^k} p_l, & \eta_1 \ne 1 \\ \frac{p_l}{k}, & \eta_1 = 1 \end{cases},$$

$$p^0_{m-j} = \begin{cases} \frac{1-\eta_1^j}{1-\eta_1} p^0_{m-1}, & \eta_1 \ne 1 \\ j p^0_{m-1}, & \eta_1 = 1 \end{cases} \quad j = 1, 2, \ldots, k-1,$$

$$p^1_{l+j} = \rho_2 \frac{1-\rho_2^j}{1-\rho_2} p^0_{m-1}, j = 1, 2, \ldots, k-1,$$

and

$$p_j = \rho_2^{j-m+1} \frac{1-\rho_2^k}{1-\rho_2} p^0_{m-1}, j \ge m.$$

The value of p_l, which is the only remaining unknown in the system can be determined by normalizing the probabilities to 1:

$$p_l = \begin{cases} \left[\frac{k(1-\rho_2\eta_1)}{(1-\eta_1^k)(1-\rho_2)} - \frac{\eta_1^{l+1}}{1-\eta_1}\right]^{-1}, & \eta_1 \ne 1 \\ \left[l + \frac{k+1}{2} + \frac{\rho_2}{1-\rho_2}\right]^{-1}, & \eta_1 = 1 \end{cases}. \tag{3}$$

Using the steady-state probabilities derived above, we can compute the probability of congestion and the rate of control information in terms of the system parameters.

The average rate of control packets sent can be found by noting that there is one packet transmitted by the observer, every time the state changes from $m-1^0$ to m or from $l+1^1$ to l. The rate (in control packets per second) is therefore given by

$$R = \lambda p^0_{m-1} + \mu_2 p^1_{l+1}.$$

Next, observe that for a positive recurrent chain, $\lambda p^0_{m-1} = \mu_2 p^1_{l+1}$. Thus,

$$R = 2\lambda p^0_{m-1} = \begin{cases} \frac{2\lambda(1-\eta_1)}{1-\eta_1^k} p_l, & \eta_1 \ne 1 \\ \frac{2\lambda p_l}{k}, & \eta_1 = 1 \end{cases}, \tag{4}$$

where p_l was found in terms of the system parameters in (3). The probability of congestion can also be found easily:

$$\mathbb{P}\{Q \ge M\} = \sum_{j \ge M} p_j = \rho_2^{M-m+1} \frac{1-\rho_2^k}{(1-\rho_2)^2} p^0_{m-1}. \tag{5}$$

Let us now investigate the behavior of R (equation (4)) and $\mathbb{P}\{Q \ge M\}$ (equation (5)) as functions of k, for a given l. It is easy to check that as the gap $k = m - l$ increases, the rate of control decreases. In fact, for $\eta_1 < 1$, that is, $\mu_1 < \lambda < \mu_2$, p_l (equation (3)) decreases inversely with k, and R also decreases approximately inversely with k. For $\eta_1 = 1$, i.e., $\lambda = \mu_1 < \mu_2$ we see that p_l decreases inversely with k, and R decreases inversely with k^2. Finally, note that for $\eta_1 > 1$, i.e., $\lambda < \mu_1 < \mu_2$, p_l does not decrease below the constant value $\frac{\eta_1-1}{\eta_1^{l+1}}$ as k increases, while R decreases

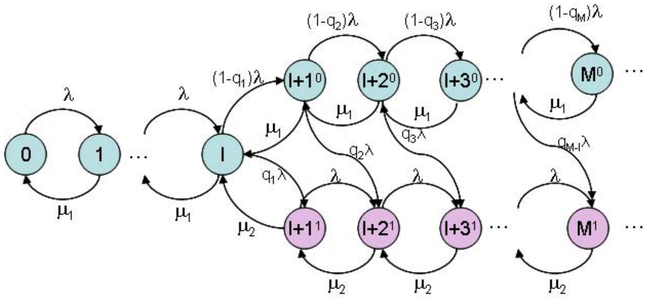

Fig. 3. The Markov process $Y(t)$ for a random control policy.

exponentially with k. On the other hand, the congestion probability increases exponentially with k for all values of η_1. This is an example of the inherent tradeoff that exists between how frequently we control the queue and how much 'risk' of congestion we are willing to tolerate.

We show later that the two threshold policy has the optimal exponential decay rate for the congestion probability. Next, we consider another simple control policy which is not optimal in the exponential sense. However, the analysis of the uniform random control policy will be useful later when we consider error prone control signals, and the corresponding decay exponents.

B. Uniform random control

In this policy, if an arrival occurs to a queue with l or more packets, and if the current service rate is μ_1, the observer sends a control packet with probability q. The decision to send a control packet is independent of such decisions in the past, and of the queue arrival and file size distributions. If the control packet was sent, the service rate switches to μ_2, and stays there until the queue length falls back to l. On the other hand, if the packet was not sent, the service rate stays at μ_1, and the above actions are repeated the next time an arrival occurs to a queue with l or more packets. Note that for $q = 1$, the policy corresponds to the single threshold policy, which also coincides with the two-threshold policy with $k = 1$.

More generally, we can also consider a non-uniform random control policy, where the observer sends a control packet with probability $q_i, i = 1, 2, \ldots$ whenever an arrival occurs to a queue with $l+i-1$ packets. The Markov process $Y(t)$ corresponding to the random control policy is shown in Fig. 3. The uniform random control policy corresponds to having each q_i to be equal to q in Fig. 3. Further, note that the two threshold policy is also a special case of Fig. 3, with $q_k = 1$, and $q_i = 0, i = 1, 2, \ldots, k-1$. In fact, the Markov process in Fig. 3 is general enough to encompass all the Markovian control policies that we will be interested in. However, the chain in Fig. 3 is not analytically tractable for general values of $q_i, i = 1, 2, \ldots$. We will only analyze the uniform random control policy in this subsection.

An important feature of the uniform random control policy is that, unlike in the two threshold policy, there is no fixed queue length threshold beyond which the higher rate is employed. In other words, there is a non-zero probability of not having switched to the higher service rate even if the queue length is very large.

1) Control rate vs. congestion tradeoff for the uniform random control policy: Intuitively, we can see that a smaller value of q leads to a lower rate of control, and a higher congestion probability. We now analytically characterize this tradeoff by solving for the steady state probabilities as before. Evidently, the first $l+1$ states of this system satisfy (2). We obtain the steady state probabilities of each state in terms of p_l, the steady state probability of the state l :

$$p^0_{l+i} = s^i p_l, i \geq 1,$$

and

$$p^1_{l+i} = \begin{cases} \frac{p_l q \rho_2}{1-s}\left(\frac{s^i - \rho_2^i}{s-\rho_2}\right), & s \neq \rho_2, \\ \\ \frac{p_l q \rho_2}{1-\rho_2} i \rho_2^{i-1}, & s = \rho_2. \end{cases},$$

where s is defined by

$$s = \frac{1+\rho_1 - \sqrt{(\rho_1 - 1)^2 + 4q\rho_1}}{2}. \quad (6)$$

Finally, we normalize all probabilities to obtain p_l in terms of the system parameters:

$$p_l = \left[\frac{1-\eta_1^{l+1}}{1-\eta_1} + \frac{s}{1-s} + \frac{q\rho_2}{(1-s)^2(1-\rho_2)}\right]^{-1}. \quad (7)$$

If $\eta_1 = 1$, the first term in (7), which is not defined, should be replaced by $l+1$.

Now that we know the steady state probabilities under this policy, we can easily obtain the rate of control packets and congestion probability, as we did for the two threshold policy. We find that the rate of control packets is given by

$$R = \frac{2\lambda q p_l}{1-s}, \quad (8)$$

and the congestion probability by (9).

We plot the probability of congestion (9) as a function of R, as we vary $q \in (0, 1]$. We compare the control rate vs. congestion tradeoff for the uniform random control policy with the curve obtained for the two-threshold policy, in Fig. 4. In the figure, the solid curve is the tradeoff curve for the two threshold policy, while the dotted curve corresponds to the uniform random control policy. Both curves correspond to the same set of system parameters, $l = 5, M = 25$, $\lambda = 1\,\text{sec}^{-1}$, $\mu_1 = 0.8\,\text{sec}^{-1}$, and $\mu_2 = 1.2\,\text{sec}^{-1}$. We see that the two-threshold policy has a strictly smaller congestion probability for a given rate of control, except at the right most point, where the curves coincide. This right extreme point corresponds to $k = 1$ and $q = 1$, respectively, where both the schemes coincide with the single threshold policy as mentioned earlier. The two-threshold policy was found to dominate the uniform random policy for various other combinations of system parameters as well.

$$\mathbb{P}\{Q \geq M\} = \begin{cases} \frac{s^{M-l}}{1-s} p_l + \frac{p_l q \rho_2}{(1-s)(s-\rho_2)} \left[\frac{s^{M-l}}{1-s} - \frac{\rho_2^{M-l}}{1-\rho_2} \right], & s \neq \rho_2, \\ \frac{\rho_2^{M-l}}{1-\rho_2} p_l + \frac{p_l q}{(1-\rho_2)^3} \rho_2^{M-l}(\rho_2 + (M-l)(1-\rho_2)), & s = \rho_2. \end{cases} \tag{9}$$

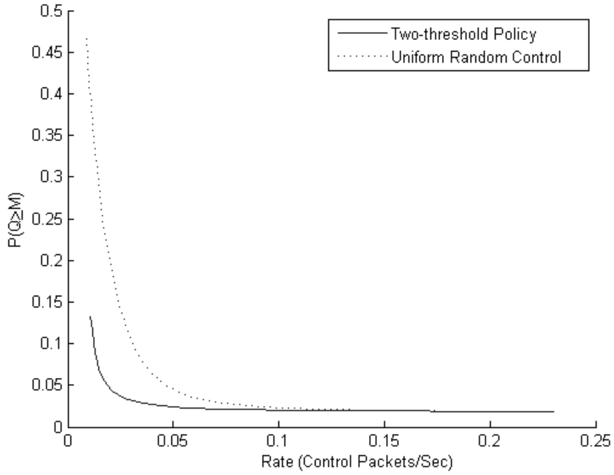

Fig. 4. Comparison between the two-threshold policy and uniform random control policy.

In spite of the above numerical evidence that the two-threshold policy dominates the uniform random policy, proving it analytically is very cumbersome, as evinced by the complicated expressions encountered above. Even if we did manage to show the above result analytically, the rate-congestion tradeoff for more general control policies shown in Fig. 3 would still remain out of reach. In order to circumvent this situation, we resort to large deviation techniques, which characterize the *exponential rate of decay* of the congestion probability in M, as a function of the control rate. We show that the two-threshold policy has the best possible exponential rate of decay of the congestion probability for any given rate of control.

V. Large Deviation Exponent Characterization of Congestion Probability

In many queueing systems, the congestion probability decays exponentially in the buffer size M. (See for example, the congestion probabilities in (5) and (9)). Furthermore, when the buffer size gets large, the exponential term dominates all other sub-exponential terms in determining the decay probability. It is therefore useful to focus only on the exponential rate of decay, while ignoring all other sub-exponential dependencies of the congestion probability on the buffer size M. Such a characterization is obtained by using the so called large deviation exponent (LDE).

Suppose that $f(M)$ is a function of M that tends to zero as $M \to \infty$. Define the LDE of $f(M)$ to be

$$E = \lim_{M \to \infty} -\frac{1}{M} \ln f(M)$$

if the limit exists. If $E = 0$, then $f(M)$ decays to zero sub-exponentially in M. If E is positive but finite, $f(M)$ decays exponentially with M. If E is infinite, then $f(M)$ decays to zero faster than a simple exponential (e.g. $f(M) = e^{-M^2}$). Note that the LDE is insensitive to all sub-exponential dependencies on M. For example, $f_1(M) = M^{1000}e^{-2M}$, and $f_2(M) = \frac{e^{-2M}}{M}$ have the same LDE, namely 2, even though f_1 is much larger than f_2 for large M. In spite of its gross insensitivity, large deviation exponents are widely used in several areas, including queueing theory and information theory, to study highly unlikely events. We now use decay exponents to characterize the congestion probability of various control schemes.

For a given control policy, define the LDE corresponding to the decay rate of the congestion probability as

$$E = \lim_{M-l \to \infty} -\frac{1}{M-l} \ln \mathbb{P}\{Q \geq M\}.$$

Here, we define the LDE with respect to $M - l$ getting large, since there is no control applied when $Q \leq l$. We now compute and compare the LDE for some simple control policies described earlier.

A. LDE for the two threshold policy

It is clear from (5) that the only term that is exponential in M is $\rho_2^{M-m} = \rho_2^{M-l-k}$. The LDE is therefore given by $E = \lim_{M-l\to\infty} \frac{M-l-k}{M-l} \ln \frac{1}{\rho_2}$. Now assume that k scales with M sub-linearly, so that $\lim_{M-l\to\infty} \frac{k(M)}{M-l} = 0$, and the LDE of the two threshold policy becomes

$$E = \ln \frac{1}{\rho_2}. \tag{10}$$

On the other hand, it is clear from (4) that the control rate can be made arbitrarily small, if $k(M)$ tends to infinity. We thus have the following result.

Proposition 1: If k grows to infinity sub-linearly in M, the two threshold policy achieves a constant LDE for any rate of control. In particular, as the control rate approaches zero, the congestion probability increases sub-exponentially in M.

B. LDE for the uniform random control policy

The uniform random control policy was introduced earlier because it exhibits a more interesting and non-constant LDE characteristic. Further, the analysis and the LDE characteristics of this policy throw light on what happens when the control channel is error prone, as discussed in the next section. The congestion probability expression in (9) consists of *two* distinct terms that decay exponentially in M, namely s^{M-l} and ρ_2^{M-l}. We need to determine which of them dominates the rate of decay, and hence decides the LDE.

At high rates of control, (i.e., $q \approx 1$), the value of s is close to zero, as can be easily seen from (6). Thus, the value of ρ_2^{M-l} dominates the rate of decay of the congestion probability, so that $E = \ln \frac{1}{\rho_2}$ for 'high' rates of control, where $\rho_2 > s$. As we decrease q and hence decrease the rate of control, s increases gradually, and for a certain value of q, say $q = q^*$, we will have $s = \rho_2$. At this point, the rates of decay of the two exponential terms are equal, and $E = \ln \frac{1}{\rho_2}$ holds for $q \geq q^*$. Note that the LDE for this range is the same as that of the two-threshold policy.

Next, as the rate decreases further, we have $q < q^*$, and $s > \rho_2$, so that s^{M-l} determines the rate of decay of the congestion probability. This means that for $q < q^*$, the LDE is given by $E = \ln \frac{1}{s}$. Since $s > \rho_2$ for this range, the LDE is clearly smaller than $-\ln \rho_2$. Furthermore, note that the LDE for this range is no longer a constant function of the rate of control, since s itself is a function of q.

We therefore conclude that the uniform random control policy has two distinct 'regimes' of operation, when it comes to the decay rate of the congestion probability. For the 'high rates of control' regime characterized by $q \geq q^*$, the decay exponent is a constant function of rate, and it is equal to the decay exponent of the two threshold policy. However, for 'low rates of control' ($q < q^*$,) the LDE decreases as the rate of control decreases. It is an easy exercise to determine the 'knee point' q^* by solving for q that satisfies $s = \rho_2$. This yields

$$q^* = (1 - \rho_2\eta_1)(1 - \rho_2). \tag{11}$$

We finally investigate the behavior of the LDE for the uniform random policy, as the rate of control approaches zero. Here, we distinguish two cases, namely $\eta_1 \leq 1$, and $\eta_1 > 1$. In the former case, as the rate of control approaches zero ($q \downarrow 0$), we can see from (6) that s approaches unity, so that the LDE given by $E = \ln \frac{1}{s}$ approaches zero. On the other hand, when $\eta_1 > 1$, it follows from (6) that $\lim_{q\downarrow 0} s = \rho_1$, so that the LDE approaches $\ln \frac{1}{\rho_1}$ as the rate of control approaches zero. To summarize: the LDE is constant, equal to $\ln \frac{1}{\rho_2}$, for $R > R(q^*)$. As the control rate decreases from $R(q^*)$ to zero, the LDE decreases monotonically from $\ln \frac{1}{\rho_2}$ to zero ($\frac{1}{\rho_1}$) if $\eta_1 \leq 1$ ($\eta_1 > 1$) . We state the conclusions of this subsection in the following proposition.

Proposition 2: For the uniform random control policy, the rate of control packets and the decay exponent of the congestion probability are given parametrically in terms of q by the following expressions:

$$R(q) = \frac{2\lambda q p_l}{1-s},$$

$$E(q) = \begin{cases} \ln \frac{1}{\rho_2}, & q \geq q^*, \\ \ln \frac{2}{1+\rho_1-\sqrt{(\rho_1-1)^2+4q\rho_1}}, & q < q^*, \end{cases} \tag{12}$$

where s, p_l and q^* are given by (6),(7), and (11) respectively.

To conclude this subsection, we present a plot of the LDE as a function of the control rate for the two threshold policy as well as the uniform random policy, in Fig. 5. The system parameters are: $l = 5, \lambda = 1\,\text{sec}^{-1}, \mu_1 = 0.5\,\text{sec}^{-1}, \mu_2 = 2\,\text{sec}^{-1}$.

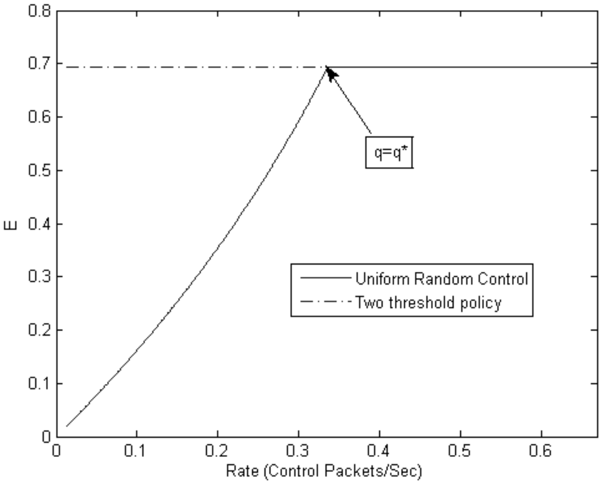

Fig. 5. LDE as a function of control rate.

Remark 1: Large deviation theory has been widely applied to study congestion and overflow behaviors in queueing systems. Tools such as the Kingman bound [7] can be used to characterize the LDE of any G/G/1 queue. Large deviation framework also exists for more complicated queuing systems, with correlated inputs, several sources, finite buffers etc., see for instance [8]. However, for *controlled* queues, where the input or service rates can vary based on queue length history, simple large deviation formulae do not exist. It is remarkable that for a single server queue with Markovian control, we are able to obtain rather intricate LDE characterizations such as in Fig. 5, just by applying 'brute force' steady state probability computations.

C. An upper bound on the LDE for generic Markovian policies

In this subsection, we show that the LDE of the two-threshold policy cannot be beaten by any control policy.

Theorem 1: The two threshold policy has the best possible LDE for any given rate of control, among all control policies.

The theorem is a simple consequence of the fact that no policy can have a faster rate of decay for the congestion probability than the M/M/1 queue with service rate μ_2.

At this point, we conjecture that the two threshold policy achieves the lowest congestion probability for any given rate of control, among all policies which use the lower service rate for queue lengths not exceeding l. To prove this claim, one has to only show optimality among all Markovian policies, as mentioned earlier.

VI. The effect of control errors on queue congestion

In this section, we relax the assumption that the control packets sent by the observer are received without errors by the controller. We investigate the effect of control errors on

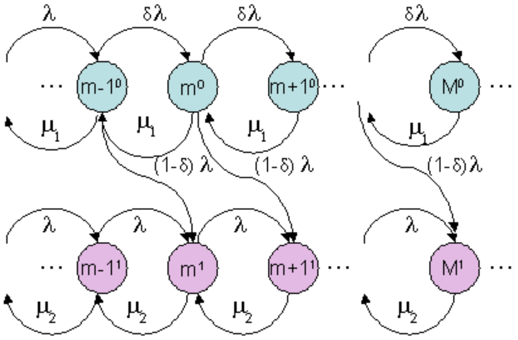

Fig. 6. The Markov process $Y(t)$ corresponding to the error prone two-threshold policy. Only a part of the state space is shown.

the probability of congestion in the queue. We use a simple probabilistic model for the errors on the control channel. In particular, we assume that any control packet sent by the observer can be lost with some probability δ, independently of other packets. Using the decay exponent tools developed earlier, we show the existence of a critical value of the error probability, say δ^*, beyond which the errors in receiving the control packets lead to an exponential worsening of the congestion probability.

A. The two-threshold policy over an error-prone control channel

In this subsection, we analyze the behavior of the two-threshold policy operating on a control channel with probability of error δ. We consider the two-threshold policy owing to its simplicity, and the fact that it has the best possible LDE among all Markovian policies when the control channel is perfect.

As described earlier, in the two threshold policy, the observer sends a control packet when the queue length reaches $m = l + k$. This packet may be received by the flow controller with probability $1 - \delta$, in which case the service rate switches to μ_2. The packet may be lost with probability δ, in which case the service continues at the lower rate μ_1. We assume that if a control packet is lost, the observer immediately knows about it[1], and sends another control packet the next time an arrival occurs to a system with at least $m - 1$ packets.

The Markov chain corresponding to the error prone two threshold policy is shown in Fig. 6. Notice that the Markov chain is quite similar to that of the uniform random policy, if we replace δ with $1 - q$. As a matter of fact, the uniform random control policy can be viewed as a single threshold policy, with errors on the control channel occurring with probability $1 - q$. In the light of the above observation, we might expect that the error prone two threshold scheme might exhibit an LDE behavior similar to that of the uniform random control policy. This is indeed the case. To show this directly, we can write down the balance equation for the Markov chain corresponding to the error-prone two-threshold policy, and explicitly solve for the congestion probability.

Analogously to (6), define

$$s(\delta) = \frac{1 + \rho_1 - \sqrt{(1 + \rho_1)^2 - 4\rho_1\delta}}{2}. \tag{13}$$

It can be shown that the congestion probability has two terms that decay exponentially in M, namely ρ_2^{M-l-k} and $s(\delta)^{M-l-k}$. When the probability of a control error δ is close to zero, $s(\delta)$ is also small, so that ρ_2^{M-l-k} dominates the decay rate of the congestion probability. As δ increases, $s(\delta)$ also increases, and for a certain value $\delta = \delta^*$, we will have $s(\delta^*) = \rho_2$, and the two rates of decay will be equal. Finally, for $\delta > \delta^*$, we have $s(\delta) > \rho_2$, so that the rate of decay is dominated by $s(\delta)^{M-l-k}$. As $\delta \to 1$, note that $s(\delta)$ approaches ρ_1 if $\rho_1 < 1$, or unity if $\rho_1 > 1$.

We conclude from the above discussion that for a two threshold policy with an error prone control channel, there are two distinct regimes of operation. In particular, for 'small enough' error probability ($\delta < \delta^*$), the exponential rate of decay of the congestion probability is the same as in an error free system. However, for $\delta > \delta^*$, the decay exponent begins to take a hit, and therefore, the congestion probability suffers a drastic increase. This 'critical' error probability δ^* can be obtained easily by solving for $s(\delta) = \rho_2$. Indeed, we find that

$$\delta^* = \frac{\rho_2}{\rho_1}(1 + \rho_1 - \rho_2). \tag{14}$$

We summarize our findings above in the following theorem.

Theorem 2: Consider a two threshold policy in which k grows sub-linearly in M. Assume that the rate increasing control packets sent by the observer can be lost with probability δ, independently of other control packets. Then the LDE corresponding to the congestion probability is given as a function of δ as

$$E(\delta) = \begin{cases} \ln \frac{1}{\rho_2}, & \delta \leq \delta^*, \\ \ln \frac{2}{1+\rho_1 - \sqrt{(\rho_1+1)^2 - 4\delta\rho_1}}, & \delta > \delta^*. \end{cases}, \tag{15}$$

where δ^* is given in (14).

Fig. 7 shows a plot of the decay exponent as a function of the error probability δ, for $\rho_1 > 1$ as well as $\rho_1 < 1$. The 'knee points' in both the plots correspond to δ^* for the stated values of ρ_1 and ρ_2.

VII. Concluding Remarks and Future Work

In this paper, we considered some simple Markovian service rate allocation policies, and derived the tradeoff between the rate of control and the associated congestion probability. We also introduced and used large deviation exponents to analyze more general Markovian control policies. We identified a simple two threshold policy that achieves the best possible tradeoff between rate of control and the decay exponent of the congestion probability. Finally, we analyzed the impact of control channel errors on the congestion probability of

[1] This is an idealized assumption; in practice, delayed feedback can be obtained using ACKS.

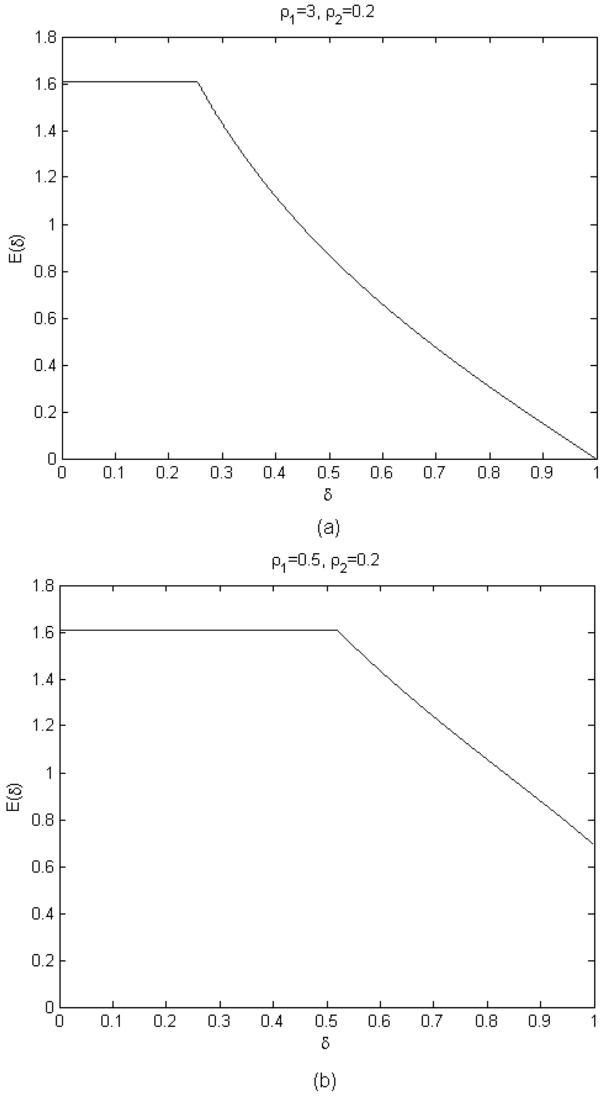

Fig. 7. LDE as a function of δ for (a)$\rho_1 > 1$ and (b)$\rho_1 < 1$.

the two threshold policy. Using a simple probabilistic model for control errors, we showed the existence of a critical error probability, beyond which the congestion probability undergoes an exponential worsening.

Suppose we are given a control channel with a probability of control error greater than the critical probability of error δ^*. If we apply the two threshold policy over this error prone control channel, the LDE of the congestion probability will take a hit, as shown in Section VI. However, addition of some form of redundancy to the control packets may help mitigate control errors, which in turn may improve the LDE performance. However, addition of redundancy to the control information also increases the amount of system resources used for control, and hence reduces the amount of resources available to service the data. Thus, we encounter another basic tradeoff: adding redundancy to control signals ensures their correct reception, which is *desirable* from the point of view of congestion probability, but on the one hand, it reduces the amount of service bandwidth available, which it turn has an *adverse* effect on the congestion level in the sysytem. Understanding this basic tradeoff and characterizing just how robust the control signals need to be in a given system, are topics under investigation, and will be the subject of a future publication.

Finally, we remark that even though we only considered the resource allocation problem in this paper, our results also hold for a single-server queue with flow control, with minor modifications.

REFERENCES

[1] L. Tassiulas, A. Ephremides, *Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks.* IEEE Trans. Aut. Contr. **37**, 1936–1948, 1992.

[2] M.J. Neely, E. Modiano, C.E. Rohrs, *Dynamic power allocation and routing for time-varying wireless networks.* INFOCOM Proceedings, 2003.

[3] M. J. Neely, E. Modiano, and C. Li, *Fairness and Optimal Stochastic Control for Heterogeneous Networks,* IEEE INFOCOM Proceedings, March 2005.

[4] X. Lin, N. Shroff and R. Srikant, *On the Connection-Level Stability of Congestion-Controlled Communication Networks* , To appear in the IEEE Transactions on Information Theory.

[5] R. G. Gallager, *Basic Limits on Protocol Information in Data Communication Networks*, IEEE Transactions on Information Theory, Vol. IT-22, No. 4, July 1976, pp. 385-398.

[6] J R. Perkins, R. Srikant, *The Role of Queue Length Information In Congestion Control and Resource Pricing*, Proceedings of the 38th Conference on Decision & Control, Phoenix, Arizona, USA, December 1999.

[7] R. G. Gallager, *Discrete Stocastic Processes*, Kluwer Academic Publishers, 1996, pp. 234.

[8] A. Ganesh, N. O'Connel, D. Wischik, *Big Queues*, Springer-Verlag, 2004.