Price-Based Rate Control for Random Access Networks

Clement Yuen and Peter Marbach

Abstract— **We study a price-based rate control mechanism for random access networks. The mechanism uses channel feedback information to control the aggregate packet arrival rate. For our analysis, we use the standard slotted Aloha model with an infinite set of nodes. We show that the resulting Markov chain is positive recurrent. In addition, we characterize the throughput and delay at the operating point of the system and show how the operating point can be set a priori by appropriately choosing the control parameters. We illustrate our results using numerical experiments.**

I. Introduction

We consider rate control for random access networks that support the CSMA family of protocols [1]. These protocols emphasize zero-coordination by allowing nodes to contend for channel access at will ("free-for-all" approach) [1]. However, if two or more nodes transmit a packet at the same time, then the packets interfere ("collide") and are not successfully received by the destination nodes. Such packets become backlogged and must be retransmitted at a later time. An important feature of such random access networks is that the network can become unstable (*i.e.* the number of backlogged packets grows without bound) if the packet generation rate exceeds the channel capacity [1]. As a result, rate control is needed to ensure a stable operation of the network. In this paper, we propose such a rate control mechanism and study its properties.

The work presented in this paper has been motivated by the popularity of wireless local area networks using the IEEE 802.11 standard for channel access. Bandwidth is scarce in wireless local area networks and if the input rates are left uncontrolled, then the network is likely to become congested and buffers saturated. Currently, TCP is used to control the input rates to wireless local area networks [2]; however, it has been shown experimentally that the combination of TCP and the IEEE 802.11 standard leads to a degradation of throughput as the number of active nodes increases [3]. This observation leads us to consider the following question: is this throughput degradation a fundamental property of rate control for random access networks, or does there exist a rate control mechanism which can achieve a sustainable throughput as the number of nodes increases?

To study the above question, we consider a rate control mechanism that is similar to price-based rate control schemes for wireline networks. That is, we assume nodes with elastic traffic and use a control (price) signal to modulate the input rates. We refer to [4]–[9] for a discussion of elastic traffic and how price-based mechanisms can be used to control the transmission rates of nodes with such traffic. As the aggregate input rate and total system backlog cannot be observed directly in random access networks, we use channel feedback information to adapt the control (price) signal. In the event of a collision we increase the price (resulting in a decrease in the total input rate) to avoid an increase in the backlog and additional collisions in the future. When the channel is idle, we decrease the price to stimulate a more efficient use of the network capacity. For this rate control mechanism, we investigate the following questions: (a) what is the maximal sustainable throughput as the number of nodes increases; (b) can a desired throughput be achieved by suitably choosing the parameters of the control mechanism; and (c) how does the rate control scheme interact with the retransmission strategy for backlogged packets, *i.e.* what is the relation between the parameters of the rate control, the retransmission strategy, and the network throughput.

We study the above price-based rate control as applied to random access networks that support the CSMA family of protocols. For our analysis we use standard models for these protocols (as described for example in [1]). These models are simple enough to allow a formal analysis, yet retain all the important characteristics of the random access channel. We start by studying the above rate control for slotted Aloha [1]. Slotted Aloha and its unslotted version (pure Aloha) have been central to the understanding of random access networks. These two protocols have over the years evolved into a rich family of medium access control schemes, most notably CSMA/CD, the Ethernet standard, and CSMA/CA, which forms the basis of the IEEE 802.11 standard. We discuss in Section VIII how the results obtained for the slotted Aloha model can be extended to slotted CSMA and CSMA/CD. Starting with slotted Aloha simplifies the notation and makes the analysis and results more lucid.

We obtain the following results. We show that (1) the proposed rate control mechanism achieves a sustainable throughput as the number of nodes increases; (2) (under appropriate assumptions) there exists a unique operating point which can be used to characterize the sustainable throughput and packet delay; (3) the throughput at the operating point depends solely on the parameters of the rate control mechanism; (4) the delay at the operating point is determined by the retransmission probability for backlogged packets; and (5) a desired operating point can be set a priori by suitably choosing the parameters of the rate control. An interesting implication of our results is that it is not the retransmission strategy, but the rate control mechanism, which dictates the throughput achieved over a random access network. This suggests that a simple (state-

C. Yuen and P. Marbach are with the Department of Computer Science, University of Toronto; email:clhyuen,marbach@cs.toronto.edu.

independent) retransmission strategy for backlogged packets, combined with the right rate control mechanism, might be sufficient to achieve a high throughput in random access networks.

We note that the above rate control mechanism could potentially be combined with price-based rate control for wireline networks to provide end-to-end rate control (see also Section II-C). A detailed discussion of end-to-end rate control is however beyond the scope of this paper.

The rest of the paper is organized as follows. In Section II, we introduce the rate control mechanism that we consider and formulate its dynamics as a Markov chain in Section III. In Section IV, we establish that this Markov chain is positive recurrent; we define the operating point of the chain, and characterize the average system throughput and delay. We illustrate these results using numerical case studies in Section V. In Section VI - VIII, we consider possible extensions. In Section IX we review the related research literature.

II. Problem Formulation

In this section, we describe the channel model and the rate control algorithm that we consider. In Section III, we present a Markov chain formulation based on the standard infinite node model of slotted Aloha [1].

A. The Channel Model

Consider the situation where many nodes communicate over a slotted multiaccess broadcast channel [1]. All packets have the same length and each packet requires one time unit for transmission. All nodes are synchronized so that the transmission of each packet starts at an integer time $t \geq 0$. We refer to the slot beginning at time t as slot t. If two or more nodes transmit a packet in a given time slot, there would be a collision and the receiver would obtain no information about the contents or source of the transmitted packets. Packets that experience a collision become backlogged and are retransmitted at a later time. If all nodes involved in a collision immediately retransmit, another collision would surely occur. Instead, each backlogged node retransmits a packet with a fixed probability q, $0 < q < 1$, in each successive slot until a successful transmission occurs. If only one node sends a packet in a given time slot, the packet is successfully received by the destination node(s). We assume that packets are generated at each node by application(s). Whenever a node receives a new packet from an application, it transmits the packet in the first slot after its arrival, thus risking occasional collisions but achieving a small delay if collisions are rare.

B. The Rate Control Scheme

If the rate at which applications generate packets is left uncontrolled, then the above slotted Aloha channel could get clogged with collisions. To avoid this, we propose a rate control mechanism where we assume that nodes have elastic traffic [4] and use a control (price) signal u to control the input rates to the network. We refer to [5]–[9] for a discussion on how price-based mechanisms can be used to control the transmission rates of nodes with elastic traffic.

Suppose that M nodes with elastic traffic are sharing the channel and each node generates new packets according to an independent Poisson process. Let $\lambda_m(u)$, $m = 1, \ldots, M$, be the rate of the Poisson process under the control u. The rate function $\lambda_m(u)$ is assumed to be decreasing in u so that a higher control u will lead to a reduction in the arrival rate of new packets at node m. Let

$$\lambda(u) = \sum_{m=1}^{M} \lambda_m(u)$$

be the aggregate arrival rate over all nodes.

As new packets are generated according to a Poisson process and backlogged packets are retransmitted with probability q, the total number of transmission attempts in time slot t is given by a random variable Z_t. Let $I[\cdot]$ be the indicator function and let $I[Z_t = 0]$, $I[Z_t = 1]$, and $I[Z_t \geq 2]$ indicate an idle slot, a successful transmission, or a collision in slot t. Let u_t be the control signal that determines the rate of arrivals of new packets in slot t. We then update the control u_t by setting

$$u_{t+1} = \left[u_t - \alpha I[Z_t = 0] + \beta I[Z_t = 1] + \gamma I[Z_t \geq 2]\right]^+, \quad (1)$$

where $[x]^+$ denotes $\max\{x, 0\}$, $x \in \mathbb{R}$, and α, β, γ are parameters such that $\alpha, \gamma > 0$ and $\beta \in \mathbb{R}$. The parameters α, β and γ are control updates corresponding to the observed channel outputs of "idle", "success" and "collision" respectively. Note that updating u only requires that the system to be able to detect a channel event, but does not require that the number of packets involved in a collision to is known. As the aggregate rate is a decreasing function, the above update rule has the following interpretation. The control signal should be increased in the event of a collision to reduce the arrival rate of new packets, and to avoid an increase in the backlog and additional collisions in the future. In the event of an idle slot the control should be lowered to allow nodes to increase their rates to make more efficient use of the channel. In the event of a successful transmission, the control can be left unchanged, or lowered (raised) in the case of a more conservative (aggressive) control scheme.

In the following, we study whether this scheme can stabilize the slotted Aloha channel, and given that α and γ are positive while β can take on any real value, how we should choose the right control parameters to achieve a certain throughput and delay at the operating point of the network. (We will define the notion of an operating point in Section IV-B.)

C. Discussion

Below, we briefly comment on some aspects of the above rate control that are important from a practical point of view. A detailed discussion of these issues is beyond the scope of the paper.

The above rate control assumes a congestion signal u that is common to all nodes in the random access network. One approach to achieve this requirement is by letting the access point (which could be a base-station in a wireless, or a router in a wired setting) update the congestion signal. In order to communicate the control signal to individual nodes, packet marking as proposed for price-based rate control in wireline networks can be used. In [10], we discuss such an approach based on a marking scheme proposed by Athuraliya and Low [8]. Using packet marking has the advantage that it allows one to potentially integrate the above rate control with price-based rate control for wireline networks to provide end-to-end rate control, as discussed also in [10]. We refer to [6]–[9] for a more detailed discussion of practical issues in price-based rate control schemes such as rate adaptation and packet marking mechanisms.

III. Markov Chain Formulation

As we are interested in the situation where a large number of nodes access the network, we use for our analysis the slotted Aloha model with an infinite number of nodes and assume that each node has at most one backlogged packet [1]. It is well-known that this model provides a worst-case analysis for the finite node model, *i.e.* it lower-bounds the throughput, and upper-bounds the delay, obtained in the finite node case. For our analysis, we assume that the aggregate rate function $\lambda(u)$ satisfies the following condition.

Assumption 1: The rate function $\lambda(u)$ is bounded, continuous and strictly decreasing. Furthermore, there exists a positive constant $u_{\max}$ such that $\lambda(u) = 0$ when $u \geq u_{\max}$.

Note that we assume that the arrival rate vanishes when the control signal exceeds the threshold value $u_{\max}$. In Section VI we relax this assumption and only require that $\lambda(u)$ approaches 0 in the limit as u becomes large.

Let n_t be the total number of backlogged packets at the beginning of slot t. As each node has at most one backlogged packet, the dynamics of the total number of backlogged packets is given by

$$n_{t+1} = n_t + Y_t - I[Z_t = 1], \tag{2}$$

where Y_t is the random variable indicating the number of newly arrived packets in slot t. Using Eq. (1) and Eq. (2), the rate control can be modelled by a Markov chain (n_t, u_t) with the following transition probabilities,

$$\mathbf{P}\big(n_{t+1} = n', u_{t+1} = u' \,\big|\, n_t = n, u_t = u\big) \tag{3}$$
$$= \begin{cases} e^{-\lambda(u)} nq\,(1-q)^{n-1}, & \\ \qquad\qquad n' = n-1, u' = u + \max\{\beta, -u\} & \\ e^{-\lambda(u)} (1-q)^n, & n' = n, u' = u + \max\{-\alpha, -u\} \\ \lambda(u)\, e^{-\lambda(u)} (1-q)^n, & n' = n, u' = u + \max\{\beta, -u\} \\ e^{-\lambda(u)} \big(1 - (1-q)^n - nq\,(1-q)^{n-1}\big), & \\ \qquad\qquad n' = n, u' = u + \gamma & \\ \lambda(u)\, e^{-\lambda(u)} \big(1 - (1-q)^n\big), & n' = n+1, u' = u + \gamma \\ e^{-\lambda(u)} \dfrac{\big(\lambda(u)\big)^k}{k!}, & n' = n+k, k \geq 2, u' = u + \gamma \\ 0, & \text{otherwise.} \end{cases}$$

Without loss of generality, we can assume that the Markov chain starts at the initial state $(n_0, u_0) = (0, 0)$ and the control signal u_t takes on values only from the countable set $\mathcal{U}$ given by

$$\mathcal{U} = \big\{[-\alpha a + \beta b + \gamma c]^+ \,\big|\, a, b, c \in \mathbb{Z}_+\big\}.$$

Let $\mathcal{S} = \mathbb{Z}_+ \times \mathcal{U}$ be the state space of this Markov chain, where $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$. Then $\mathcal{S}$ is a countable set. Furthermore, one can show that a Markov chain on $\mathcal{S}$ is irreducible and aperiodic [11].

IV. Stability, Operating Point, and Parameters Selection

Without rate control, but assuming a fixed aggregate arrival rate λ, the model given by Eq.(3) is the well-known infinite user model of slotted Aloha which has been shown to be unstable (transient) for any non-zero arrival rate λ and any fixed retransmission probability q [12]. In this section we show that when a rate control scheme is imposed on slotted Aloha the system becomes stable (positive recurrent). Furthermore, we characterize the system throughput and average delay by considering the operating point of the system, and show how we can achieve a desired operating point by appropriately choosing the control parameters α, β, and γ.

A. Stability Analysis

We start out by showing that the Markov chain given by Eq. (3) is positive recurrent. We have the following result.

Proposition 1: Under Assumption 1, the Markov chain given by Eq. (3) is positive recurrent.

As the Markov chain is irreducible, the above proposition implies that every state $(n, u) \in \mathcal{S}$ is positive recurrent. It may appear that Proposition 1 immediately follows from Assumption 1 which states that there are no new packet arrivals when the control signal exceeds the threshold value $u_{\max}$. Unfortunately, this is not the case and a detailed stability analysis is necessary. In fact we prove Proposition 1 in Appendix I by showing that a Lyapunov function $V(n, u)$ which satisfies the Pakes' mean-drift criteria [13] exists.

B. Operating Point Analysis

Having established stability, in the next step we would like to characterize the expected throughput and delay of the system. Note that obtaining the exact values of these quantities is too ambitious a goal as this requires the computation of the steady-state probabilities of the Markov chain. Instead, we approximate the expected throughput and delay by considering the throughput and delay at the operating point of the system.

We use the following notation. Given a state (n, u), consider the functions $d_n(n, u)$ and $d_u(n, u)$ given by

$$d_n(n, u) \triangleq \mathbf{E}\big(n_{t+1} - n_t \,\big|\, n_t = n, u_t = u\big),$$
$$d_u(n, u) \triangleq \mathbf{E}\big(u_{t+1} - u_t \,\big|\, n_t = n, u_t = u\big).$$

Note that d_n and d_u represent the one-step expected change (mean drift) of the backlog and the control signal at state (n, u). To determine the values of d_n and d_u, we make use of the Poisson approximation for small q (as described in [1]), which allows the probability of exactly one packet transmitted in a slot to be approximated by

$$\lambda(u) e^{-\lambda(u)} (1-q)^n + nq e^{-\lambda(u)} (1-q)^{n-1}$$
$$\approx \big(\lambda(u) + nq\big) e^{-\lambda(u) - nq},$$

where n is the backlog and u the control signal at the beginning of the slot. Let

$$G(n,u) = \lambda(u) + nq$$

be the offered load to the channel at state (n,u) [1]. To simplify notation, we will often use G instead of $G(n,u)$ to denote the offered load at state (n,u). The expected changes in the backlog and control signal are then given by

$$\begin{aligned} d_n(n,u) &= \lambda(u) - Ge^{-G}, &\text{(4a)}\\ d_u(n,u) &= \max\{-\alpha,-u\}e^{-G} + \max\{\beta,-u\}Ge^{-G} \\ &\quad + \gamma\big(1-e^{-G}(1+G)\big). &\text{(4b)} \end{aligned}$$

The lower bounds on the control parameters in $d_u(n,u)$ are due to the requirement that the control u is non-negative. We define the operating point as follows.

Definition 1: We call $(n^*,u^*) \in \mathbb{R}_+^2$ an operating point of the system if $d_n(n^*,u^*) = d_u(n^*,u^*) = 0$.
To simplify the notation, we will use G^* to denote the offered load $G(n^*,u^*)$ at an operating point (n^*,u^*).

The above definition implies that at an operating point (n^*,u^*) we have

$$\lambda(u^*) = G^* e^{-G^*}, \tag{5a}$$
$$G^* = \lambda(u^*) + n^* q. \tag{5b}$$

Let

$$S = G^* e^{-G^*} \tag{6}$$

denote the (expected) throughput S at an operating point. Note that the system is in equilibrium at an operating point in the sense that its packet arrival rate $\lambda(u^*)$ is equal to its throughput S. The maximal throughput occurs when $G^* = 1$ and is equal to e^{-1}.

Further, let D denote the (expected) packet delay at an operating point (n^*,u^*). Using Little's result [1] and Eq. (5), this is given by

$$D = \frac{n^*}{\lambda(u^*)} = \frac{G^* - \lambda(u^*)}{q\lambda(u^*)} = \frac{e^{G^*}-1}{q}. \tag{7}$$

Note that the delay increases exponentially with the offered load G^*.

Next, we address the existence and uniqueness of an operating point. We have the following existence result.

Proposition 2: Let Assumption 1 hold. Then there exists an operating point (n^*,u^*) for the system.
We refer to [14] for a proof of the above proposition.

An operating point which exists by virtue of Proposition 2 is not necessarily unique. However, for situations that are of practical interest, uniqueness can be guaranteed. We have the following uniqueness result.

Proposition 3: Let Assumption 1 hold, and furthermore let $\lambda\big(\max\{\alpha,-\beta\}\big) \geq e^{-1}$. Then there exists a unique operating point.
We provide a proof for the above proposition in Appendix II. Proposition 3 essentially states that when the unconstrained arrival rate $\lambda(0)$ is large (so that $\lambda\big(\max\{\alpha,-\beta\}\big)$ is larger than the channel capacity e^{-1}), then there exists a unique operating point. If the unconstrained arrival rate is small, additional operating points with distinct n^*'s and offered load G^*'s might exist. To guarantee uniqueness regardless of the unconstrained arrival rate, one can exclude operating points with large offered load. We have the following uniqueness result.

Proposition 4: Let Assumption 1 hold. Furthermore, let $-\alpha+\beta+(e-2)\gamma \geq 0$. Then there exists a unique operating point.
We refer to [14] for a proof of the above proposition. The condition in Proposition 4 for the control parameters essentially excludes operating points with $G^* > 1$. In general operating points with large offered load exhibit a much larger delay (see Eq. (7)) and ought to be avoided in practice.

In Section V we carry out numerical case studies to investigate how closely S and D approximate the average throughput and delay of the system. Note that S and D can easily be computed as a function of G^*. As we show below, this simplicity allows one to select the control parameters in order to achieve a desired operating point.

C. Selection of the Control Parameters

Suppose we would like to operate the channel at a given offered load $\widehat{G}$ in order to obtain a certain throughput or delay. We are then interested in the following question: Can we choose the parameters α, β, and γ in such a way that the offered load at the resulting operating point is equal to $\widehat{G}$? The following lemma establishes that this is indeed possible.

Lemma 1: Let $\widehat{G}$ be a given offered load. Furthermore let Assumption 1 hold, and let $\alpha > 0$ and $\gamma > 0$ be given control parameters. For

$$\beta = \frac{\gamma}{\widehat{G}}(\widehat{G} + 1 - e^{\widehat{G}}) + \frac{\alpha}{\widehat{G}},$$

every operating point (n^*,u^*) of the system is then such that Eq. (5a) and (5b) hold, and that

$$\begin{aligned} G^* &= \widehat{G} \quad \text{if } \lambda\big(\max\{\alpha,-\beta\}\big) \geq e^{-1},\\ G^* &\leq \widehat{G} \quad \text{if } \lambda\big(\max\{\alpha,-\beta\}\big) < e^{-1}. \end{aligned}$$

We provide a proof for the above lemma in Appendix III.

Lemma 1 implies that when the unconstrained arrival rate $\lambda(0)$ is large (so that $\lambda\big(\max\{\alpha,-\beta\}\big)$ exceeds the channel capacity), it is possible to operate the channel at any desired offered load. When the unconstrained arrival rate is small, one may also achieve smaller offered load; in this case the resulting throughput $G^*e^{-G^*}$ could be higher or lower than the desired throughput $\widehat{G}e^{-\widehat{G}}$.

Note that the value of β in Lemma 1 only depends on α, γ, and $\widehat{G}$, but not on the aggregate rate function $\lambda(u)$ or the retransmission probability q. This result implies that one can choose the control parameters α, β, and γ to achieve a certain throughput S, without any a priori knowledge of the aggregate rate function or the retransmission probability. This property is of importance in practice as it allows one to stabilize the system at a desired throughput S even when traffic dynamics changes (in which case the aggregate rate function $\lambda(u)$ varies), or is not known a priori.

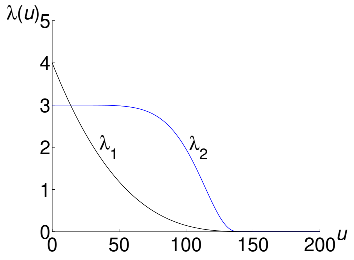
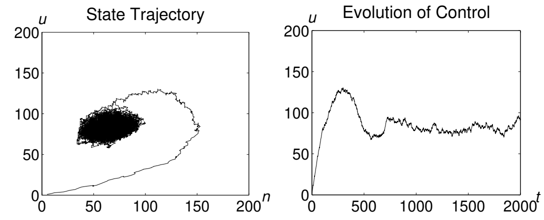
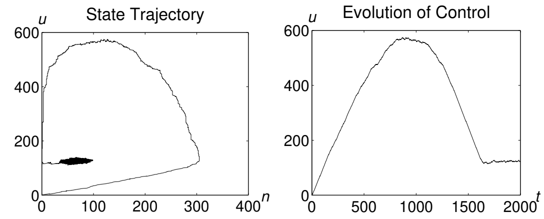


Fig. 1. The two rate functions used in the numerical case studies.



(a) Simulation run for λ_1



(b) Simulation run for λ_2

Fig. 2. Trajectory of the Markov chain (n_t, u_t) (left) and evolution of the control signal (right) for λ_1 and λ_2.

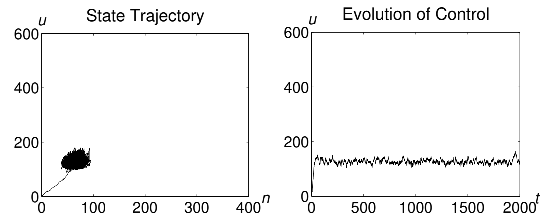


Fig. 3. Trajectory of the Markov chain (n_t, u_t) (left) and evolution of the control signal (right) for λ_2 with $\gamma = 6$.

V. Numerical Results

To illustrate the behavior of the rate control mechanism, we carry out several numerical case studies for the infinite node model of Section III. We first investigate how well the operating point analysis of Section IV-B predicts the actual system performance. Next, we discuss how the convergence behavior of the algorithm can be improved.

A. System Performance

Suppose we always want to operate the system at the maximal throughput e^{-1}. Suppose also that the retransmission probability q is equal to 0.01, and that we consider two aggregate rate functions given by

$$\lambda_1(u) = \left[4\left(1 - \tfrac{u}{150}\right)^3\right]^+ \text{ and } \lambda_2(u) = \left[3\left(1 - \left(\tfrac{u}{140}\right)^6\right)^3\right]^+,$$

as shown in Figure 1. The given rate functions can be regarded as characterizations of the traffic dynamics of different applications. Using the results from the previous section, we choose the control parameters $\alpha = \gamma = 1$ arbitrarily, while $\beta = 0.2817$ is chosen as in Lemma 1 to obtain the achievable desired offered load $G^* = 1$. The operating point[1] for this setup is equal to $(n_1^*, u_1^*) = (63.21, 82.29)$ for the rate function λ_1 and $(n_2^*, u_2^*) = (63.21, 124.86)$ for λ_2. The throughput and delay at the operating point are equal to $S = e^{-1} = 0.368$ and $D = 171.82$ respectively for both cases.

Starting at state $(0, 0)$, we simulated the Markov chain given by Eq. (3) for 100,000 slots. Figure 2 shows the simulation runs for both cases as the chain (n_t, u_t) evolves in their state spaces, as well as the evolution of the control signals for the first 2000 slots. After a transient, the system "hovers" around the operating points (n_1^*, u_1^*) and (n_2^*, u_2^*). For λ_1, the measured average throughput is equal to 0.369 and the measured average delay is equal to 172.57; whereas for λ_2, the measured average throughput and the measured average delay are 0.363 and 179.55 respectively. Note that these quantities are remarkably close to the predicted values of 0.368 and 171.82. This simulation confirms that the approximate analysis of Subsection IV-B satisfactorily predicts the actual system performance, and that the throughput and delay achieved are independent of the rate function concerned.

It is important to note that the fact that the system hovers around the operation point (n^*, u^*) is not undesirable, but indeed necessary for the rate control algorithm to ensure stability. To show this, we give the following argument. Suppose that the algorithm would actually converge to the operating point. The system would then behave like a slotted Aloha channel with a fixed arrival rate $\lambda(u^*)$ and retransmission probability q. However, as mentioned in Section IV, this channel has been shown to be unstable. This implies that to ensure stability, the arrival rate needs to be adapted on a fast time-scale to better react to backlog fluctuations; such adaptation leads to the "hovering" behavior observed.

B. Convergence Behavior

An important feature of a rate control mechanism is the time it takes to converge to the vicinity of the operating point. Clearly, long convergence time is unfavorable for handling traffic dynamics due to, for instance, changes in applications. In the above simulation, it took roughly 700 slots for λ_1 to "converge" to a local area of the operating point, whereas this convergence time is close to 2000 slots in the case of λ_2. We note that it is possible to reduce the convergence time in both cases by boosting the control parameter γ. Intuitively, a large γ reduces the arrival rate more intensely when backlog accumulates and collisions prevail. This lessens the rate of accumulation and prevents a large buildup of backlog, thereby shortening the time required to bring the backlog down to the operating level n^*.

[1] The operating point can be computed numerically using Eq. (5a)-(5b).

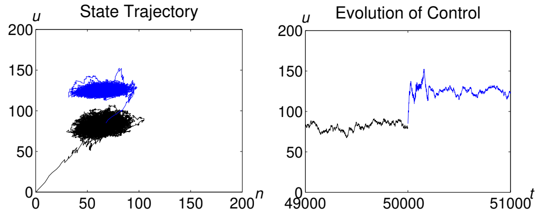


Fig. 4. Trajectory of the Markov chain (n_t, u_t) (left) and evolution of the control signal (right) with λ_1, λ_2 and adaptation of γ.

To illustrate this intuition, we redid the simulation using the rate function λ_2 with $\gamma = 6$ (and $\beta = -3.310$ to maintain $G^* = 1$). The results are shown in Figure 3. The measured throughput and delay are 0.352 and 182.56 respectively. In this case it only takes around 25 slots to get to the neighborhood of the operating point, which is a dramatic improvement in the convergence time. Note, however, that the larger γ introduces more fluctuations to the control signal, as evident from Figure 3, as well as a performance loss with respect to the average throughput and delay.

The above simulations for the rate function λ_2 suggest that perhaps the best strategy for choosing γ is to assume a large value for speedy convergence to the neighborhood of the operating point, but a small value to achieve minimal variance and high system performance after the convergence. We demonstrate this approach by the following simulation, which also illustrates the performance of the rate control mechanism in the face of traffic dynamics. The simulation utilizes λ_1 as the rate function for the first 50000 slots, and λ_2 for the subsequent 50000 slots, to model a change in traffic dynamics. During the first 200 slots for each rate function, γ is set to 4 to decrease the convergence time. After that, γ reverts to 1 to attain minimal variance of the control signal. The results are shown in Figure 4, with the evolution of the control signal shown for the 2000 slots during which the rate function changes. For the first 50000 slots, the average backlog is equal to 63.72 while the average control is equal to 82.61. The measured throughput is 0.368 and the measured delay is 172.50. For the second 50000 slots, the average backlog is 64.36 while the average control is 125.36. The measured throughput and the measured delay are 0.364 and 177.13 respectively.

The discussion of this subsection illustrates that the convergence behavior of the system can be improved by suitably choosing the control parameters α, β, and γ. However, there remain several design issues from a practical standpoint: (i) is there an optimized γ to achieve maximum convergence or minimum variance in the above approach? (ii) how does one detect a drastic change in traffic dynamics to invoke the use of a larger γ? (iii) how does one determine when the system has "converged" and the smaller γ should be used? A detailed discussion of these issues is beyond the scope of the paper.

VI. Extensions

In this section, we consider possible extensions of the basic model given in Section II. We first consider the situation where nodes have different delay requirements. Next, we allow nodes to retransmit backlogged packets more aggressively when the system backlog is small and collisions are less likely to occur.

A. Delay Differentiation

In the previous sections, we considered the case where each node uses the same retransmission probability q to schedule packet retransmissions. However, some nodes (applications) might have stringent delay requirements and we would like them to be able to retransmit packets more aggressively. In this subsection, we discuss such an approach.

Assume that nodes can be classified into C classes where all nodes in the same class have (roughly) the same delay requirements. To account for different delay requirements we associate with class c the retransmission probability q_c where classes with strict delay requirements are assigned a higher value. We make the following assumption.

Assumption 2: There exist positive constants $q_{\min}$ and $q_{\max}$ such that $0 < q_{\min} \le q_c \le q_{\max} < 1$, for $c = 1, \ldots, C$.

Note that the state of the Markov chain is now given by $(\underline{n}, u)$, where $\underline{n} = (n_1, \ldots, n_C)$ is the vector indicating the total number of backlogged packets in the different classes. Using the set $\mathcal{U}$ as defined in Subsection III, let $\mathcal{S} = \mathbb{Z}_+^C \times \mathcal{U}$ be the countable state space of the Markov chain. For a given state $(\underline{n}, u) \in \mathcal{S}$, let $\lambda_c(u)$ be the total arrival rate of packets of class c. Then the offered load due to class c is given by

$$G_c(\underline{n}, u) = \lambda_c(u) + n_c q_c.$$

Further, let

$$n = \sum_{c=1}^{C} n_c$$

be the total number of backlogged packets, and let

$$q(\underline{n}) = \frac{1}{n} \sum_{c=1}^{C} n_c q_c$$

be the average retransmission probability. Then the total offered load $G(\underline{n}, u)$ at state $(\underline{n}, u)$ is given by

$$G(\underline{n}, u) = \sum_{c=1}^{C} G_c(\underline{n}, u) = \lambda(u) + nq(\underline{n}),$$

where $\lambda(u) = \sum_{c=1}^{C} \lambda_c(u)$ is the aggregate arrival rate of the system. We define an operating point of the system as follows.

Definition 2: We call $(\underline{n}^*, u^*) \in \mathbb{R}_+^{C+1}$ an operating point of the system if $d_u(\underline{n}^*, u^*) = d_{c,n}(\underline{n}^*, u^*) = 0$, where

$$\begin{aligned} d_u(\underline{n}^*, u^*) &= \max\{-\alpha, -u^*\}e^{-G^*} + \max\{\beta, -u^*\}G^* e^{-G^*} \\ &\quad + \gamma\big(1 - e^{-G^*}(1 + G^*)\big), \\ d_{c,n}(\underline{n}^*, u^*) &= \lambda_c(u^*) - G_c^* e^{-G^*}, \qquad c = 1, \ldots, C, \end{aligned}$$

with $G_c^* = G_c(\underline{n}^*, u^*)$ and $G^* = G(\underline{n}^*, u^*)$ as before to simplify the notation.

Note that the above definition implies that each class c is in equilibrium at the operating point in the sense that its throughput $S_c = G_c^* e^{-G^*}$ is equal to its arrival rate $\lambda_c(u^*)$.

Let $e_c \in \mathbb{R}_+^C$ denote the c-th unit vector. Using the Poisson approximation, the state transition probabilities are given by

$$\mathbf{P}\left(\underline{n}_{t+1} = \underline{n}', u_{t+1} = u' \,\middle|\, \underline{n}_t = \underline{n}, u_t = u\right) \tag{8}$$
$$= \begin{cases} n_c q_c e^{-G(\underline{n},u)}, & \underline{n}' = \underline{n} - e_c, u' = u + \max\{\beta, -u\} \\ e^{-G(\underline{n},u)}, & \underline{n}' = \underline{n}, u' = u + \max\{-\alpha, -u\} \\ \lambda(u) e^{-G(\underline{n},u)}, & \underline{n}' = \underline{n}, u' = u + \max\{\beta, -u\} \\ e^{-\lambda(u)} - \left[1 + \sum_{c=1}^C n_c q_c\right] e^{-G(\underline{n},u)}, & \underline{n}' = \underline{n}, u' = u + \gamma \\ \lambda_c(u) e^{-\lambda(u)} - \lambda_c(u) e^{-G(\underline{n},u)}, & \underline{n}' = \underline{n} + e_c, u' = u + \gamma \\ e^{-\lambda(u)} \displaystyle\prod_{c=1}^C \frac{\left(\lambda_c(u)\right)^{k_c}}{k_c!}, & n_c' = n_c + k_c, \sum_{c=1}^C k_c \geq 2, \\ & u' = u + \gamma \\ 0, & \text{otherwise.} \end{cases}$$

We then have the following stability result (see Appendix I for a proof).

Proposition 5: Let Assumption 1 and 2 hold, then the Markov chain given by Eq. (8) is positive recurrent.

Using similar arguments as presented in Subsection IV-B, we can show that there always exists an operating point, and that the operating point is unique under the conditions given in Proposition 3 and 4. The system throughput S at an operating point $(\underline{n}^*, u^*)$ is given by $S = G^* e^{-G^*}$, and the throughput S_c for class c is given by

$$S_c = G_c^* e^{-G^*}.$$

At the operating point, the average delay D over all classes is given by

$$D = \frac{e^{G^*} - 1}{q(\underline{n}^*)}.$$

The average delay D_c for nodes in class c is given by

$$D_c = \frac{e^{G^*} - 1}{q_c}.$$

The above result shows that delay differentiation among nodes of different classes according to their respective requirements is achieved at the operating point. Furthermore, a desired offered load G^* can again be achieved by choosing the control parameters α, β, and γ as given by Lemma 1, and the delay D_c of class c at the operating point can be set by choosing the retransmission probability q_c.

B. Dynamic Retransmission Probabilities

So far, we have assumed that backlogged nodes always retransmit with a fixed probability. However, when the backlog is small collisions are less likely to occur, and nodes could be allowed to retransmit more aggressively. In this subsection, we consider an approach that adaptively changes the retransmission probability to achieve a better utilization of the channel. As individual nodes do not know the exact number of backlogged packets, we use the control signal u as an indicator of the size of the backlog and adapt the retransmission probabilities accordingly. Let $q(u)$ be the resulting retransmission probability function. We make the following assumption.

Assumption 3: The retransmission probability function $q(u)$ is non-increasing in u and we have that $\lim_{u \to +\infty} q(u) = 0$. Furthermore, there exists a non-negative and non-increasing function $\mathcal{G}(u)$, as well as positive constants a_l and a_h such that

$$a_l \leq \left| \frac{\mathcal{G}(u+p) - \mathcal{G}(u)}{q(u)} \right| \leq a_h,$$

where $u \geq 0$ and $p \in \{-\alpha, \beta, \gamma\}$.

The above assumption guarantees that the function $q(u)$ neither decreases too fast, nor too slowly. Examples of functions that satisfy the above assumption are $q(u) = e^{-bu}$ and $q(u) = (1 + bu)^{-r}$, where b and r are positive constants and $r > 1$. This setting can be seen to be very similar to that described by Kelly in [15]; in fact with $b = r = 1$, the second example above is asymptotically the function used in [15].

For this extension, we can relax the assumption on the rate function $\lambda(u)$ as follows.

Assumption 4: The rate function $\lambda(u)$ is bounded, continuous and strictly decreasing, and we have $\lim_{u \to +\infty} \lambda(u) = 0$.

Under the above assumption, we can show that the system is positive recurrent (see [14] for a detailed derivation). Furthermore, it can be shown that there exists an operating point as given by Definition 1. The throughput and average delay at the operating point are given by Eq. (6) and (7) where we use the retransmission probability $q(u^*)$ at the operating point to compute the delay in Eq. (7). Using Lemma 1, control parameters can be chosen to achieve a desired offered load $\widehat{G}$ without prior knowledge of the rate function $\lambda(u)$. However, in order to predict the delay D at the operating point, we need to know the rate function $\lambda(u)$ in order to determine the control u^*, and the retransmission probability $q(u^*)$, at the operating point. This suggests that it might be advantageous to use fixed, rather than dynamic, retransmission probabilities as this allows one to more conveniently predict the delay at the operating point.

By similar arguments as in the previous subsection, delay differentiation can be achieved. We refer to [14] for a detailed derivation.

VII. Numerical Results: Dynamic Retransmission Probabilities and Delay Differentiation

Below, we illustrate the results of the previous section, where we focus on the performance at the operating point. Similar to Section V the convergence time of the mechanism can be improved by suitably choosing the control parameter γ. We omit here a detailed discussion of this issue.

A. Dynamic Retransmission Probabilities

Suppose that the network provider wants to operate the system at the maximal throughput e^{-1}, and chooses the control parameters $\alpha = \gamma = 1$, and $\beta = 0.2817$ to obtain the operating point $G^* = 1$. The rate function is chosen to be

$$\lambda(u) = \frac{40}{(1+u)^{1.5}}, \qquad u \geq 0,$$

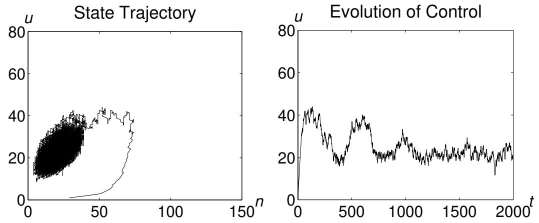


Fig. 5. Trajectory of the Markov chain (n_t, u_t) (left) and evolution of the control signal (right) using the retransmission function $q(u) = \frac{1}{(1+u)^{1.1}}$.

and the function determining the retransmission probabilities is given by

$$q(u) = \frac{1}{(1+u)^{1.1}}, \qquad u \geq 0.$$

Note that the above rate function is different from the rate function in Section V. In particular, the above rate function satisfies the relaxed Assumption 4, but not Assumption 1. The function $q(u)$ for the dynamic retransmission probabilities satisfies Assumption 3. The operating point for this setup is equal to $(n^*, u^*) = (19.68, 21.78)$ and the throughput and delay at the operating point are equal to $S = e^{-1} = 0.368$ and $D = 53.51$ respectively.

Fig. 5 shows the system trajectory for a simulation run of $100,000$ steps starting with state $(0, 0)$ and the evolution of the control signal for the first 2000 slots. After a transient, the system "converges" around the measured operating point $(20.65, 22.79)$. The measured average system throughput is equal to 0.369 and the measured average delay is equal to 55.92, again well approximated by the predicted values of $S = 0.368$ and $D = 53.51$.

B. Delay Differentiation: Fixed Retransmission Probabilities

Suppose that we can distinguish between two classes of nodes which have very different delay requirements. In particular, class 2 nodes are not delay-sensitive and can tolerate a delay roughly 10 times larger than class 1 nodes.

We first consider the case where nodes retransmit backlogged packets with a fixed probability and associate the retransmission probabilities

$$q_1 = 10 \cdot q_2 = 0.01,$$

with class 1 and 2 nodes. Furthermore, the aggregate arrival rates for nodes of class 1 and 2 respectively are chosen to be

$$\lambda_1(u) = \lambda_2(u) = \left[2\big(1 - \tfrac{u}{150}\big)^3\right]^+.$$

These rate functions satisfy Assumption 1 as required. The control parameters are as given above. For this set up the operating point is $(\underline{n}^*, u^*) = \big((31.61, 316.1), 82.29\big)$ with throughput $S = 0.368$ and delay $D_1 = 171.82$ and $D_2 = 1718.2$ for class 1 and 2 nodes respectively.

We simulated the chain for 100,000 steps starting with state $\big((0, 0), 0\big)$ (see Fig. 6). The measured operating point is equal to $\big((31.52, 314.11), 82.39\big)$. The measured average system throughput is equal to 0.369 and the measured average delay is equal to 169.49 and 1685.4 for class 1 and 2 nodes respectively, providing a delay differentiation by roughly a factor of 10 as intended. Again, the throughput and delay at the operating point approximates well the measured system performance.

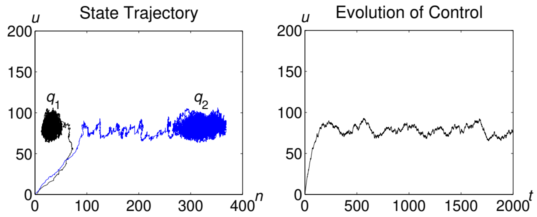


Fig. 6. Trajectory of the Markov chain $(\underline{n}_t, u_t)$ (left) and evolution of the control signal (right) for retransmission probabilities $q_1 = 10 \cdot q_2 = 0.01$.

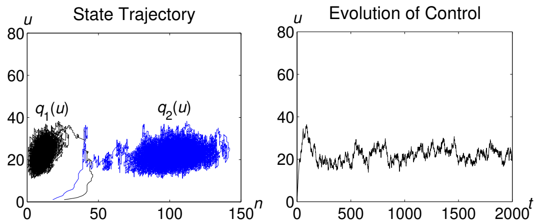


Fig. 7. Trajectory of the Markov chain $(\underline{n}_t, u_t)$ (left) and evolution of the control signal (right) for retransmission probabilities $q_1(u) = 10 \cdot q_2(u) = \frac{1}{(1+u)^{1.1}}$.

C. Delay Differentiation: Dynamic Retransmission Probabilities

Next, we consider delay differentiation for dynamic retransmission probabilities and associate the retransmission functions

$$q_1(u) = 10 \cdot q_2(u) = \frac{1}{(1+u)^{1.1}}$$

and rate functions

$$\lambda_1(u) = \lambda_2(u) = \frac{20}{(1+u)^{1.5}}$$

with class 1 and 2 nodes. The control parameters are as given above. For this set up the operating point is $(\underline{n}^*, u^*) = \big((9.84, 98.42), 21.78\big)$ with throughput $S = 0.368$ and delay $D_1 = 53.51$ and $D_2 = 535.08$ for class 1 and 2 nodes respectively.

We simulated the chain for 100,000 steps starting with state $\big((0, 0), 0\big)$ (see Fig. 7). The measured operating point is equal to $\big((10.03, 101.48), 22.47\big)$. The measured average system throughput is equal to 0.367 and the measured average delay is equal to 54.56 and 551.34 for class 1 and 2 nodes respectively, providing a delay differentiation by roughly a factor of 10 as intended.

D. Discussion

The above results illustrate that the operating point analysis is indeed useful in predicting the system performance in terms of expected throughput and delay. As discussed in Section VI, the system delay is a function of the retransmission probabilities at individual nodes; therefore, by suitably choosing the retransmission probabilities q_c, or the functions $q_c(u)$, a desired average delay can be achieved. However, when dynamic retransmission probabilities are used, the demand function $\lambda(u)$ has to be known in order to compute u^* and the retransmission probabilities $q_c(u^*)$ at the operating point. In this sense it is advantageous to choose fixed rather than dynamic retransmission probabilities as it allows one to predict the average delay at the operating point without prior knowledge of the rate functions.

VIII. Rate Control for Slotted CSMA/CD

The analysis of Section IV can be extended to slotted CSMA and CSMA/CD. Below, we outline the extension to the standard slotted CSMA/CD model with an infinite number of nodes (as described for example in [1]); the extension to slotted CSMA can be done in a similar way. The extension of the analysis and results of slotted Aloha to slotted CSMA/CD only requires notational changes and we state the results in this section without proofs.

A. Slotted CSMA/CD

The basic operation of slotted CSMA/CD is the same as for slotted Aloha except that (a) in idle and collision slots a duration τ is needed to detect an idle channel after a transmission ends, and (b) if a new packet arrives at a node while a transmission is in progress, then the packet becomes a backlogged packet and is retransmitted with probability q, $0 < q < 1$, after each subsequent idle slot. New packets arriving during an idle slot are transmitted in the next slot (as before). For this model, it is well-known (see for example [1]) that the offered load at state (n, u) is given by $g(n, u) = \tau\lambda(u) + nq$. To simplify the notation, we will again use g to denote the offered load at state $g(n, u)$. The throughput under the offered load g is given by (see for example [1])

$$\frac{ge^{-g}}{\tau + ge^{-g} + \tau\big[1 - (1+g)e^{-g}\big]},$$

and the maximum throughput is $1/(1 + 3.31\tau)$ achieved at $g = 0.77$.

B. Formulation as a Markov Chain

Consider again the rate control scheme of Section II where the control signal u_{t+1} at the beginning of time slot $t+1$ is given by

$$u_{t+1} = \big[u_t - \alpha I[Z_t = 0] + \beta I[Z_t = 1] + \gamma I[Z_t \geq 2]\big]^+.$$

The above rate control scheme as applied to slotted CSMA/CD can again be modelled by a Markov chain (n_t, u_t), $t \geq 0$, where n_t is the total number of backlogged packets and u_t the control signal at the beginning of time slot t. The transition probabilities of this Markov chain are given by

$$P(n_{t+1} = n', u_t = u' \mid n_t = n, u_{t-1} = u) \tag{9}$$
$$= \begin{cases} e^{-\tau\lambda(u)} nq(1-q)^{n-1} \frac{(\lambda(u))^{Y_t}}{Y_t!} e^{-\lambda(u)}, \\ \qquad n' = n - 1 + Y_t, u' = u + \max\{\beta, -u\} \\ e^{-\tau\lambda(u)}(1-q)^n, \\ \qquad n' = n, u' = u + \max\{-\alpha, -u\} \\ \tau\lambda(u)e^{-\tau\lambda(u)}(1-q)^n \frac{(\lambda(u))^{Y_t}}{Y_t!} e^{-\lambda(u)}, \\ \qquad n' = n + Y_t, u' = u + \max\{\beta, -u\} \\ e^{-2\tau\lambda(u)} \frac{(\tau\lambda(u))^{Y_t}}{Y_t!}(1 - (1-q)^n - nq(1-q)^{n-1}), \\ \qquad n' = n + Y_t, u' = u + \gamma \\ \tau\lambda(u)e^{-2\tau\lambda(u)}(1 - (1-q)^n)\frac{(\tau\lambda(u))^{Y_t}}{Y_t!}, \\ \qquad n' = n + 1 + Y_t, u' = u + \gamma \\ \frac{(\tau\lambda(u))^k}{k!} e^{-2\tau\lambda(u)} \frac{(\tau\lambda(u))^{Y_t}}{Y_t!}, \\ \qquad n' = n + k + Y_t, u' = u + \gamma, k \geq 2, \\ 0, \qquad \text{otherwise.} \end{cases}$$

where Y_t is the random variable indicating the number of newly arrived packets in slot t. We then have the following stability result.

Corollary 1: Under Assumption 1, the Markov chain given by Eq. (9) is positive recurrent.

C. Operating Point Analysis

As in Section IV-B, we consider the operating point of the Markov chain given by Eq. (9). Recall that for a given state (n, u) the one-step expected changes in the backlog, and the control, respectively, are given by

$$d_n(n, u) \triangleq \mathbf{E}\big(n_{t+1} - n_t \,\big|\, n_t = n, u_t = u\big),$$
$$d_u(n, u) \triangleq \mathbf{E}\big(u_{t+1} - u_t \,\big|\, n_t = n, u_t = u\big).$$

Using the Poisson approximation (see [1] for a derivation), we obtain that

$$d_n(n, u) = \lambda(u)\Big(\tau + ge^{-g} + \tau\big[1 - (1+g)e^{-g}\big]\Big) - ge^{-g},$$
$$d_u(n, u) = \max\{-\alpha, -u\}e^{-g} + \max\{\beta, -u\}ge^{-g} + \gamma\big(1 - e^{-g}(1+g)\big).$$

We call $(n^*, u^*) \in \mathbb{R}_+^2$ an operating point of the system if $d_n(n^*, u^*) = d_u(n^*, u^*) = 0$. To simplify the notation, we will again use g^* to denote the offered load $g(n^*, u^*)$ at an operating point (n^*, u^*).

The above definition implies that at an operating point (n^*, u^*) we have

$$\lambda(u^*) = \frac{g^* e^{-g^*}}{\tau + g^* e^{-g^*} + \tau\big[1 - (1+g^*)e^{-g^*}\big]}, \tag{11a}$$
$$g^* = \tau\lambda(u^*) + n^* q. \tag{11b}$$

We have the following results.

Corollary 2: Under Assumption 1, there exists an operating point for CSMA/CD.

Corollary 3: Let Assumption 1 hold and furthermore let $\lambda\big(\max\{\alpha, -\beta\}\big) > 1/(1 + 3.31\tau)$. Then there exists a unique operating point.

Corollary 4: Let Assumption 1 hold and furthermore let $-\alpha + 0.77\beta + \gamma(e^{0.77} - 1.77) \geq 0$. Then there exists a unique operating point.

D. Selection of Control Parameters

The offered load at an operating point can be set again by the procedure as given by Lemma 1.

Corollary 5: Let $\hat{g}$ be a given offered load. Furthermore let Assumption 1 hold, and let $\alpha > 0$ and $\gamma > 0$ be given control parameters. For

$$\beta = \frac{\gamma}{\hat{g}}(\hat{g} + 1 - e^{\hat{g}}) + \frac{\alpha}{\hat{g}},$$

every operating point (n^*, u^*) of the system is then such that Eq. (11a) and (11b) hold, and that

$$g^* = \hat{g} \quad \text{if } \lambda\big(\max\{\alpha, -\beta\}\big) \geq \frac{1}{1 + 3.31\tau},$$
$$g^* \leq \hat{g} \quad \text{if } \lambda\big(\max\{\alpha, -\beta\}\big) < \frac{1}{1 + 3.31\tau}.$$

In order to maximize the throughput, the offered load g^* at the operation point should be equal to 0.77, instead of 1 as was the case for slotted Aloha.

IX. Related Work

Random access networks, and in particular slotted Aloha, have been extensively studied throughout the past decades [16], [17]. An extensive treatment on random access communication protocols can be found for example in [1], [18], [19]. Most of this work has focused on retransmission strategies for backlogged packets. Below, we highlight the work on rate control.

Kleinrock and Lam [20] and Mittal and Venetsanopoulos [21] have studied an ON-OFF rate (input) control for slotted Aloha where the transmission rate switches between two values, depending on the backlog with respect to a given threshold. However, their rate control mechanism requires that the total system backlog to be globally known, which is an unrealistic assumption. Note that the rate control considered in this paper does not make this assumption as it is based on (locally available) channel feedback information. The approach presented in [20] and [21] also differs from the rate control mechanism considered in this paper in the sense that we use a continuous rate function, rather than an ON-OFF control. This enables a “smoother” adaptation of the input rates in the spirit of price-based rate control schemes for point-to-point networks [6]–[8].

Price-based control schemes for random access networks have been proposed in [22], [23]. In [23], Jin and Kesidis consider a price-based scheme to determine the retransmission probabilities at individual nodes. The work by Jin and Kesidis assumes that all nodes are saturated and always have a packet to send, and does not address rate control for stabilizing the system. In [22], Battiti *et al* propose and analyze a mechanism where pricing is used to control the number of hosts contending for channel access. The control scheme proposed by Battiti *et al* requires a priori knowledge of the rate function $\lambda(u)$ in order to achieve a desired operating point [22]. The control scheme presented in this paper relaxes this assumption as the parameters α, β, and γ for achieving a desired operating point can be determined without knowing the rate function $\lambda(u)$.

The infinite node model [1] that we use in our analysis has played a central role in the theoretical understanding of random access networks; it is well-known that this model allows a qualitative measure of the goodness of control algorithms. Specifically, it has been widely used in the stability analysis of retransmission strategies of slotted Aloha [15], [24]–[26]. While best suited for the study of fundamental properties of control mechanisms, it is known that such a model does not accommodate detailed aspects of specific CSMA protocols such as the IEEE 802.11 standard. A more accurate mathematical model for the IEEE 802.11 protocol has been presented by Cali *et al* [27] to study its capacity and control mechanisms. In particular, this model has been used to study TCP throughput [28] and price-based congestion control [22] in IEEE 802.11 networks. The model of Cali *et al* assumes a finite number of nodes operating in asymptotic conditions, *i.e.* all nodes are saturated and always have a packet ready for transmission.

X. Conclusions

We presented a rate control mechanism for random access networks which we analyzed using the slotted Aloha model with an infinite set of nodes. The mechanism relies on a control signal which is updated based on the channel output. The resulting Markov chain model was shown to be stable. In addition, we characterized the throughput and delay at the operating point to approximate the average system throughput and delay. The rate control scheme was also shown to have the important property that one can choose control parameters a priori to stabilize the system at a desired operating point and without knowledge of the aggregate rate function. We illustrated our results through numerical case studies. These numerical results revealed a trade-off with respect to convergence time and variance; the larger we choose the control parameters the faster the system converges to the vicinity of the operating point, but the larger is the variance of the system trajectory around the operating point. An in-depth study of this issue is future research.

For our analysis, we adopted the infinite user model of slotted Aloha. However, the analysis and results obtained here can be extended to the finite user case [14]. The study of the finite user case poses additional challenges as nodes can now have multiple backlogged packets. The results for the infinite user model can provide insights in the finite user case in the sense that they reveal the worst-case performance of the rate control mechanism when the number of nodes becomes large. In addition, we currently investigate a variant of the rate control mechanism where nodes individually compute and maintain the control signal. Such a scheme is of interest for ad-hoc networks as it does not require a dedicated node (such as a base-station) to maintain and communicate the control signal.

A. Acknowledgments

We thank the reviewers for their suggestions which have significantly improved the paper. We thank Yiping Gong for deriving the extension to slotted CSMA/CD in Section VIII.

REFERENCES

[1] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Prentice-Hall, Inc., 1992.

[2] V. Jacobson, "Congestion Avoidance and Control," in *Proceedings of Symposium on Communications Architectures and Protocols (SIGCOMM)*. Stanford, California, United States: ACM Press, August 1988, pp. 314–329.

[3] H. Wu, Y. Peng, K. Long, C. S, and J. Ma, "Performance of Reliable Transport Protocol over IEEE 802.11 Wireless LAN: Analysis and Enhancement," in *Proceedings of IEEE Infocom*, June 2002.

[4] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal on Selected Areas in Communication*, vol. 13, no. 7, September 1995.

[5] F. Kelly, "Charging and Rate Control for Elastic Traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[6] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

[7] S. H. Low and D. E. Lapsley, "Optimization Flow Control, I: Basic Algorithm and Convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, December 1999.

[8] S. Athuraliya and S. H. Low, "Optimization Flow Control, II: Implementation," May 2000, submitted for publication.

[9] S. Kunniyur and R. Srikant, "End-to-end Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks," in *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, Tel Aviv, Israel, March 2000, pp. 1323–1332, vol. 3.

[10] C. Yuen and P. Marbach, "End-to-End Rate Control for Networks with Random Access Links," in *Proceedings of the IEEE CCECE*, Niagra Falls, ON, April 2004.

[11] C. Yuen, "Rate Control with Pricing in Contention-Based Access Networks," Master Thesis, University of Toronto, 2002.

[12] G. Fayolle, E. Gelenbe, and J. Labetoulle, "Stability and Optimal Control of the Packet Switching Broadcast Channel," *Journal of the Association for Computing Machinery*, vol. 24, no. 3, pp. 375–386, July 1977.

[13] A. G. Pakes, "Some conditions for ergodicity and recurrence of Markov chains," *Operations Research*, vol. 17, pp. 1058–1061, 1969.

[14] C. Yuen and P. Marbach, "Service Differentiation in Random Access Networks," University of Toronto, Technical Report CSRG-472, July 2003.

[15] F. Kelly, "Stochastic Models of Computer Communication Systems," *Journal of the Royal Statistical Society*, vol. B 47, no. 3, pp. 379–395, 1985.

[16] J. L. Massey, Ed., *Special Issue on Random Access Communications, IEEE Transactions on Information Theory*, vol. IT-31, March 1985.

[17] A. Ephremides and B. Hajek, "Information Theory and Communication Networks: An Unconsummated Union," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2416–2434, October 1998.

[18] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis*, ser. Telecommunication Networks and Computer Systems Series. Springer-Verlag New York Inc., 1990.

[19] B. S. Chlebus, *Randomized Communication in Radio Networks*, ser. Handbook of Randomized Computing. Kluwer Academic Publishers, July 2001, vol. I, ch. 11, pp. 401–456.

[20] L. Kleinrock and S. Lam, "Packet switching in a multiaccess broadcast channel: Dynamic control procedures," *IEEE Transactions on Communications*, vol. COM-23, no. 9, pp. 891–904, September 1975.

[21] K. Mittal and A. N. Venetsanopoulos, "A Note on Optimal Input Control Policy for an ALOHA Access Scheme," *IEEE Transactions on Communications*, vol. 39, no. 2, pp. 197–199, February 1991.

[22] R. Battiti, M. Conti, E. Gregori, and M. Sabel, "Price-based Congestion-Control in Wi-Fi Hot Spots," in *Proceedings of First IEEE/ACM WiOpt Workshop*, March 2003.

[23] Y. Jin and G. Kesidis, "A Pricing Strategy for an ALOHA Network of Heterogeneous Users with Inelastic Bandwidth Requirements," in *Proceedings of the Conference on Information Sciences and Systems*, Princeton University, United States, March 2002.

[24] B. Hajek and T. van Loon, "Decentralized Dynamic Control of a Multiaccess Broadcast Channel," *IEEE Transactions on Automatic Control*, vol. AC-27, no. 3, pp. 559–569, June 1982.

[25] L. P. Clare, "Control procedures for slotted Aloha systems that achieve stability," in *Proceedings of the ACM SIGCOMM on communications architecture & protocols*, Stowe, VT, August 1986, pp. 302–309.

[26] R. L. Rivest, "Network Control by Bayesian Broadcast," *IEEE Transactions on Information Theory*, vol. IT-33, no. 3, pp. 323–328, May 1987.

[27] F. Cali, M. Conti, and E. Gregori, "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," *ACM/IEEE Transactions on Networking*, vol. 8, no. 6, pp. 785 – 799, December 2000.

[28] R. Bruno, M. Conti, and E. Gregori, "Throughput Evaluation and Enhancement of TCP Clients in Wi-Fi Hot Spots," in *Proceedings of Wireless On-Demand Network Systems 2004*, 2004, pp. 73–86.

[29] F. G. Foster, "On the stochastic matrices associated with certain queuing processes," *Annals of Mathematical Statistics*, vol. 24, no. 3, pp. 355–360, September 1953.

[30] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, ser. Communication and Control Engineering Series. Springer-Verlag London Ltd., 1993.

APPENDIX I
PROOF OF PROPOSITION 1 AND 5

In this appendix, we provide a proof for Proposition 5. Note that Proposition 1 is a special case of Proposition 5.

Let $(\underline{n}, u) \in \mathcal{S}$, and consider the *mean drift* of the Lyapunov function $V(\underline{n}, u)$ at a given state $(\underline{n}, u)$,

$$\mathbf{E}\big(\Delta V \mid \underline{n}, u\big) \triangleq \mathbf{E}\big(V_{t+1} - V_t \;\big|\; (\underline{n}_t, u_t) = (\underline{n}, u)\big),$$

where $V_t = V(\underline{n}_t, u_t)$ denotes the value of the Lyapunov function at time t for $t \geq 1$. We have the following well-known *mean drift criteria* [13], [29], [30].

Proposition 6: The Markov chain given by Eq. (8) is positive recurrent if and only if there exist a non-negative function $V(\underline{n}, u), (\underline{n}, u) \in \mathcal{S}$, positive constants δ and b, and a finite set $\mathcal{F} \subset \mathcal{S}$ such that

$$\text{(i)} \;\; \mathbf{E}\big(\Delta V \mid \underline{n}, u\big) \leq -\delta, \qquad (\underline{n}, u) \notin \mathcal{F},$$
$$\text{(ii)} \;\; \mathbf{E}\big(\Delta V \mid \underline{n}, u\big) < b, \qquad (\underline{n}, u) \in \mathcal{F}.$$

We use Proposition 6 to prove system stability and consider a Lyapunov function of the form

$$V(\underline{n}, u) = \bar{q}u + \Gamma e^{\Lambda(u) + n\bar{q}}, \tag{12}$$

where $\bar{q}$ and Γ are positive constants such that

$$q_{\max} < \bar{q} < 1,$$

$q_{\max}$ as given in Assumption 2, and

$$\Gamma = 2 \cdot \max\{\alpha + \beta, \gamma - \beta\}.$$

For the function $\Lambda(u)$ we have the following assumption.

Assumption 5: The function $\Lambda : \mathbb{R} \to \mathbb{R}_+$ is continuous, non-increasing, and bounded. Furthermore, it has the following properties: there exist constants $\hat{u} > u_{\max}$ and $\epsilon, \xi > 0$ such that

(a) $\Lambda(u) = \Lambda(0)$ for $u < 0$

(b) $\Lambda(u) = 0$ for $u \geq \hat{u}$

(c) $\Lambda(u+\gamma) - \Lambda(u) < \begin{cases} (1 - e^{\bar{q}})\lambda(u) - \epsilon & 0 \leq u \leq \hat{u} - \gamma, \\ (1 - e^{\bar{q}})\lambda(u) & \hat{u} - \gamma < u < \hat{u} \end{cases}$

(d) $\Lambda(\hat{u} - \gamma + \beta) \leq \bar{q} - \xi$.

Note that we have

$$\lim_{u \to +\infty} V(\underline{n}, u) = +\infty$$

and

$$\lim_{n_c \to +\infty} V(\underline{n}, u) = +\infty, \qquad c = 1, \ldots, C$$

and that V is a non-negative function. At the end of this appendix, we show that a Lyapunov function of the above form indeed exists. For now, we assume that this is the case and show in the following proof that this Lyapunov function can be used to satisfy the conditions in Proposition 6.

Proof: Let V be a Lyapunov function of the above defined form. Then we have

$$\begin{aligned} &\mathbf{E}\Big(V_{t+1} - V_t \mid (\underline{n}_t, u_t) = (\underline{n}, u)\Big) \\ &= \mathbf{E}\big(\Delta V \mid \underline{n}, u\big) \\ &= \mathbf{E}\Big(\bar{q}\Delta u + \Gamma e^{\Lambda(u+\Delta u)+(n+\Delta n)\bar{q}} - \Gamma e^{\Lambda(u)+n\bar{q}} \mid \underline{n}, u\Big), \end{aligned} \tag{13}$$

where $\Delta n = n_{t+1} - n_t$ is the one-step difference in the system backlog given that $(\underline{n}_t, u_t) = (\underline{n}, u)$, and $\Delta u = u_{t+1} - u_t$ is the one-step difference of the control signal. To simplify notation, we use the following convention for our derivation of the mean drift at a given state $(\underline{n}, u) \in \mathcal{S}$,

$$\begin{aligned} q &\triangleq q(\underline{n}), & G &\triangleq G(\underline{n}, u), \\ \lambda &\triangleq \lambda(u), & \Lambda &\triangleq \Lambda(u), \\ \Lambda_\alpha &\triangleq \Lambda(u-\alpha), & \Lambda_p &\triangleq \Lambda(u+p) \quad p \in \{\beta, \gamma\}. \end{aligned}$$

In addition, we consider the relaxed update rule for the congestion signal u given by

$$u_{t+1} = u_t - \alpha I[Z_t = 0] + \beta I[Z_t = 1] + \gamma I[Z_t \geq 2].$$

The inclusion of the constraint, as given in Eq. (1), that the congestion signal cannot be negative requires only notational changes in our proof and is therefore omitted to simplify the derivation.

Recall the transition probabilities as given by Eq. (8). Using the above conventions, we obtain that

$$\begin{aligned} &\mathbf{E}\Big(e^{\Lambda(u+\Delta u)+(n+\Delta n)\bar{q}} \mid \underline{n}, u\Big) \\ &= e^{-\lambda+n(\bar{q}-q)} e^{\Lambda_\alpha} \\ &\quad + e^{-\lambda+n(\bar{q}-q)} e^{\Lambda_\beta} \left(\lambda + nqe^{-\bar{q}}\right) \\ &\quad - e^{-\lambda+n(\bar{q}-q)} e^{\Lambda_\gamma} \left(1 + \lambda e^{\bar{q}} + nq\right) \\ &\quad + e^{-\lambda+n\bar{q}} e^{\Lambda_\gamma} \sum_{k=0}^{\infty} \frac{\left(\lambda e^{\bar{q}}\right)^k}{k!} \\ &= e^{-\lambda+n(\bar{q}-q)} \left(e^{\Lambda_\alpha} - e^{\Lambda_\gamma}\right) \\ &\quad + e^{-\lambda+n(\bar{q}-q)} \left(e^{\Lambda_\beta} - e^{\Lambda_\gamma+\bar{q}}\right)\left(\lambda + nqe^{-\bar{q}}\right) \\ &\quad + e^{\Lambda_\gamma + n\bar{q} + \lambda(e^{\bar{q}}-1)}, \end{aligned} \tag{14}$$

where we have used the fact that $\sum_{k=0}^{\infty} \frac{(\lambda e^{\bar{q}})^k}{k!} = e^{\lambda e^{\bar{q}}}$. Using this result in Eq. (13), we arrive at

$$\begin{aligned} &\mathbf{E}(\Delta V \mid \underline{n}, u) \\ &= \bar{q}(\beta-\gamma)Ge^{-G} - \bar{q}(\gamma+\alpha)\, e^{-G} + \bar{q}\gamma \\ &\quad + \Gamma e^{n(\bar{q}-q)-\lambda}\left(e^{\Lambda_\beta} - e^{\Lambda_\gamma+\bar{q}}\right)\left(\lambda + nqe^{-\bar{q}}\right) \\ &\quad + \Gamma e^{n(\bar{q}-q)-\lambda}\left(e^{\Lambda_\alpha} - e^{\Lambda_\gamma}\right) \\ &\quad + \Gamma e^{n\bar{q}}\left(e^{\Lambda_\gamma+\lambda(e^{\bar{q}}-1)} - e^{\Lambda}\right). \end{aligned} \tag{15}$$

Also note that

$$\mathbf{E}(\Delta u \mid \underline{n}, u) \leq A, \tag{16}$$

where $A = \max\{\beta, \gamma\} > 0$. Furthermore, by assumption the functions λ and Λ are bounded and we have that at each state $(\underline{n}, u) \in \mathcal{S}$ the mean-drift $\mathbf{E}(\Delta V \mid \underline{n}, u)$ is finite (see Eq. (15)).

Using the above results, we are now in a position to show that the Lyapunov function $V(\underline{n}, u)$ satisfies Condition (i) and (ii) of Proposition 6. To construct a set $\mathcal{F} \subset \mathcal{S}$ to be used in the conditions we analyze three regions $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$, of the state space $\mathcal{S}$, such that $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$.

(a) $\mathcal{S}_1 = \big\{(\underline{n}, u) \in \mathcal{S} \mid u \geq \hat{u} - \min\{-\alpha, \beta\}\big\}$

At a given state $(\underline{n}, u) \in \mathcal{S}_1$, we have by assumption that

$$\Lambda = \Lambda_\alpha = \Lambda_\beta = \Lambda_\gamma = 0$$

and $\lambda = 0$. Using these observations in Eq. (15), and the fact that $e^{n(\bar{q}-q)} \geq 1$ and $e^{-\bar{q}} - 1 < -\bar{q}/2$ for $q < \bar{q} < 1$, we obtain that

$$\begin{aligned} &\mathbf{E}(\Delta V \mid \underline{n}, u) \\ &= \bar{q}(\beta-\gamma)Ge^{-G} - \bar{q}(\gamma+\alpha)e^{-G} + \bar{q}\gamma \\ &\qquad + \Gamma e^{n(\bar{q}-q)}\big(e^{-\bar{q}} - 1\big)G \\ &< \bar{q}(\beta-\gamma)Ge^{-G} - \bar{q}(\gamma+\alpha)e^{-G} + \bar{q}\gamma - \tfrac{1}{2}\bar{q}\Gamma G \\ &= \bar{q}\Big[(\beta-\gamma)Ge^{-G} - (\gamma+\alpha)e^{-G} + \gamma - \tfrac{1}{2}\Gamma G\Big]. \end{aligned}$$

Consider the function $f(G)$ given by

$$f(G) = (\beta-\gamma)Ge^{-G} - (\gamma+\alpha)e^{-G} + \gamma - \tfrac{1}{2}\Gamma G.$$

Note that for $G \geq 0$ we have $e^G \geq 1 + G$, and we obtain for the derivative of $f(G)$ on $[0, \infty)$ that

$$\begin{aligned} f'(G) &= e^{-G}\big((\alpha+\beta) + (\gamma-\beta)G - \tfrac{1}{2}\Gamma e^G\big) \\ &\leq \tfrac{1}{2}e^{-G}\big(\Gamma + \Gamma G - \Gamma e^G\big) \leq 0. \end{aligned}$$

Hence f is decreasing on $[0, \infty)$. Combining the above results, we have for all states $(\underline{n}, u) \in \mathcal{S}_1$ that

$$\mathbf{E}(\Delta V \mid \underline{n}, u) < \bar{q}f(G) \leq \bar{q}f(0) = -\bar{q}\alpha.$$

(b) $\mathcal{S}_2 = \big\{(\underline{n}, u) \in \mathcal{S} \mid 0 \leq u < \hat{u} - \gamma\big\}$

At a given state $(\underline{n}, u) \in \mathcal{S}_2$, we have by assumption that

$$\Lambda_\gamma + \lambda\big(e^{\bar{q}} - 1\big) - \Lambda < -\epsilon$$

for some positive constant ϵ. This implies that

$$e^{\Lambda_\gamma + \lambda(e^{\bar{q}}-1)} - e^{\Lambda} < e^{\Lambda}\big(e^{-\epsilon} - 1\big).$$

Also by assumption, the function Λ is non-negative and non-increasing, hence we have

$$e^{\Lambda_\gamma + \lambda(e^{\bar{q}}-1)} - e^{\Lambda} < -D_\epsilon$$

for $D_\epsilon = 1 - e^{-\epsilon}$. Furthermore, by assumption the functions λ and Λ are bounded. Using these observations in Eq. (15), it follows that there exist positive constants B, C, and D, such that for all states $(\underline{n}, u) \in \mathcal{S}_2$ we have

$$\begin{aligned} &\mathbf{E}(\Delta V \mid \underline{n}, u) \\ &< \bar{q}A + e^{n(\bar{q}-q)}(BG + C) - e^{n\bar{q}}D \\ &= \bar{q}A + e^{n\bar{q}}\Big[e^{-nq}\big(B(\lambda + nq) + C\big) - D\Big], \end{aligned}$$

where

$$\lim_{n \to +\infty} e^{-nq}\big(B(\lambda + nq) + C\big) = 0.$$

It follows that for all states $(\underline{n}, u) \in \mathcal{S}_2$ we have

$$\lim_{n\to+\infty} \mathbf{E}(\Delta V \mid \underline{n}, u) = -\infty,$$

and there exists a positive integer N_1 such that for all states $(\underline{n}, u) \in \mathcal{S}_2$ with $n \geq N_1$, we have

$$\mathbf{E}(\Delta V \mid \underline{n}, u) \leq -\bar{q}\alpha.$$

(c) $\mathcal{S}_3 = \{(\underline{n}, u) \in \mathcal{S} \mid \hat{u} - \gamma \leq u < \hat{u} - \min\{-\alpha, \beta\}\}$
At a given state $(\underline{n}, u) \in \mathcal{S}_3$, we have by assumption that $\Lambda_\gamma = 0$. Also by assumption the function Λ is non-increasing. Hence we obtain that

$$\Lambda_\beta - (\Lambda_\gamma + q) = \Lambda_\beta - q \leq \Lambda(\hat{u} - \gamma + \beta) - q \leq -\xi$$

for some positive constant ξ. This implies that there exists a positive constant B_ξ such that

$$e^{\Lambda_\beta} - e^{\Lambda_\gamma + q} < -B_\xi.$$

Furthermore, by assumption λ and Λ are bounded and $e^{\Lambda_\gamma + \lambda(e^{\bar{q}}-1)} \leq e^{\Lambda}$. Using these observations in Eq. (15), it follows that there exist positive constants B and C such that for all states $(\underline{n}, u) \in \mathcal{S}_3$ we have

$$\mathbf{E}(\Delta V | \, \underline{n}, u) < \bar{q}A + e^{n(\bar{q}-q)}(-BG + C).$$

It follows that for all states $(\underline{n}, u) \in \mathcal{S}_3$ we have

$$\lim_{n\to+\infty} \mathbf{E}(\Delta V \mid \underline{n}, u) = -\infty,$$

and there exists a positive integer N_2 such that for all states $(\underline{n}, u) \in \mathcal{S}_3$ with $n \geq N_2$, we have

$$\mathbf{E}(\Delta V \mid \underline{n}, u) \leq -\bar{q}\alpha.$$

Using the observations made for the three regions of the state space $\mathcal{S}$, we construct the set $\mathcal{F}$ used in Condition (i) and (ii) of Proposition 6 as follows. Let

$$N = \max\{N_1, N_2\},$$

where the integers N_1 and N_2 are as obtained above for the sets $\mathcal{S}_2$ and $\mathcal{S}_3$ respectively. Furthermore, let

$$\mathcal{F} = \{(\underline{n}, u) \in \mathcal{S} \mid u < \hat{u} - \min\{-\alpha, \beta\},\ n < N\}.$$

Note that the set $\mathcal{F}$ is finite. Finally, let

$$\delta = \bar{q}\alpha.$$

Using the above derived results for the regions $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$, we obtain

$$\mathbf{E}(\Delta V \mid \underline{n}, u) \leq -\bar{q}\alpha = -\delta \qquad (\underline{n}, u) \notin \mathcal{F},$$

and Property (i) of Proposition 6 follows.

Property (ii) follows immediately from the fact that the set $\mathcal{F}$ is finite and for every state $(\underline{n}, u) \in \mathcal{S}$ (and therefore for every state $(\underline{n}, u) \in \mathcal{F}$) the mean-drift $\mathbf{E}(\Delta V \mid \underline{n}, u)$ is finite (see Eq. (15) and (16)). ■

We have proved Proposition 5 under the assumption that a Lyapunov function of the defined form exists. Note that proving the existence of such a Lyapunov function amounts to showing the existence of the function Λ with the required properties. Below, we construct such a function.

Let $\beta^- = \min\{\beta, 0\}$ and let

$$\hat{u} = u_{\max} + \gamma - \beta^-.$$

Furthermore, let the positive constants ϵ and ξ be given by

$$\epsilon = \frac{\bar{q}}{4}\left(\frac{\gamma}{\gamma - \beta^-}\right)$$

and

$$\xi = \frac{\bar{q}}{2},$$

where $\bar{q}$, $q_{\max} < \bar{q} < 1$, is the positive constant in the Lyapunov function. We then define Λ as follows.

$$\Lambda(u) = \begin{cases} m\lambda_{\max} + \frac{\bar{q}}{2}\left(1 + \frac{u_{\max}}{\gamma - \beta^-}\right), & u < 0 \\ m\lambda(u) + \frac{\bar{q}}{2}\left(1 + \frac{u_{\max} - u}{\gamma - \beta^-}\right), & 0 \leq u < u_{\max} + \gamma - \beta^- \\ 0, & u \geq u_{\max} + \gamma - \beta^- \end{cases}$$

where m is chosen such that

$$m > \max\{e^{\bar{q}} - 1, \hat{m}\} > 0,$$

and

$$\hat{m} = \max_{0 \leq u < u_{\max} - \gamma} \frac{(1 - e^{\bar{q}})\lambda(u)}{\lambda(u+\gamma) - \lambda(u)}.$$

Note that the function Λ as defined above is continuous, bounded and non-increasing. Next, we verify the four properties (a)-(d) that we stated in Assumption 5 for the function Λ.

By definition, we have that $\Lambda(u) = \Lambda(0)$ for $u < 0$, and $\Lambda(u) = 0$ for $u \geq \hat{u}$; Property (a) and (b) follow immediately for this choice of Λ.

To verify Property (c), we consider two different cases. First, when $0 \leq u \leq u_{\max} - \beta^-$, we have

$$\begin{aligned} &\Lambda(u+\gamma) - \Lambda(u) = m\big(\lambda(u+\gamma) - \lambda(u)\big) - \frac{\bar{q}}{2}\left(\frac{\gamma}{\gamma - \beta^-}\right) \\ &\leq \begin{cases} \hat{m}\big(\lambda(u+\gamma) - \lambda(u)\big) - 2\epsilon, & u < u_{\max} - \gamma \\ -\big(e^{\bar{q}} - 1\big)\lambda(u) - 2\epsilon, & u_{\max} - \gamma \leq u < u_{\max} \\ -2\epsilon, & u_{\max} \leq u \leq u_{\max} - \beta^- \end{cases} \\ &< \big(1 - e^{\bar{q}}\big)\lambda(u) - \epsilon. \end{aligned}$$

Next, when $u_{\max} - \beta^- < u < u_{\max} + \gamma - \beta^-$, $\Lambda(u)$ is decreasing so that $\Lambda(u+\gamma) - \Lambda(u) = -\Lambda(u) < 0 = \big(1 - e^{\bar{q}}\big)\lambda(u)$. Combining the above results establishes Property (c).

Finally, note that

$$\Lambda(u_{\max} + \gamma - \beta^- - \gamma + \beta) \leq \Lambda(u_{\max}) = \frac{\bar{q}}{2} = \bar{q} - \xi,$$

and Property (d) is satisfied.

Appendix II
Proof of Proposition 3

Suppose Assumption 1 holds, and furthermore, $\lambda\big(\max\{\alpha, -\beta\}\big) \geq e^{-1}$. We first establish the following equivalence: (n, u) solves $d_n(n, u) = d_u(n, u) = 0$ (Eq. (4)) if and only if it solves

$$d_n(n, u) = d_u(G) = 0, \tag{17}$$

where $G = \lambda(u) + nq$, and

$$d_u(G) = -\alpha e^{-G} + \beta G e^{-G} + \gamma\big(1 - e^{-G}(1+G)\big) \quad (18)$$

is the relaxed drift for the control signal u. This is such that proving Proposition 3 amounts to showing the existence of a unique operating point (n^*, u^*) for Eq. (17).

To show the above equivalence, suppose (n, u) solves Eq. (4). Now $\lambda\big(\max\{\alpha, -\beta\}\big) \geq e^{-1} \geq Ge^{-G} = \lambda(u)$, and therefore $u \geq \max\{\alpha, -\beta\}$. This implies that (n, u) also solves Eq. (17). By similar arguments, we obtain its converse.

Hence it suffices to prove the uniqueness of an operating point for Eq. (17). We argue in two steps as follows: 1) We show below that there exists a unique $G^* = \lambda(u^*) + n^* q$ such that $d_u(G^*) = 0$. Then 2) observe that for a given $G^* \geq 0$, we have $\lambda(0) > \lambda\big(\max\{\alpha, -\beta\}\big) \geq e^{-1} \geq G^* e^{-G^*}$, and hence there exists a unique $u^* > 0$ such that $d_n(G^*, u^*) = \lambda(u^*) - G^* e^{-G^*} = 0$. From 1) and 2), it follows that there exists a unique operating point (n^*, u^*).

To show 1), we consider three different cases of β for the given α and γ. The derivative of $d_u(G)$ is given by

$$d_u'(G) = e^{-G}\big((\alpha+\beta) + (\gamma-\beta)G\big), \quad G \geq 0.$$

Case (1): $\beta > \gamma > 0$: Note that $d_u\Big(\frac{-\alpha-\beta}{\gamma-\beta}\Big) = e^{-\frac{-\alpha-\beta}{\gamma-\beta}}(\beta - \gamma) + \gamma > \gamma > 0$ and that $\lim_{G\to+\infty} d_u(G) = \gamma > 0$. Furthermore, observe that $d_u'(G) \leq 0$ for $G \in \Big[\frac{-\alpha-\beta}{\gamma-\beta}, \infty\Big)$. As $d_u(G)$ is continuous in G, it follows that there exists no $G^* \in \Big[\frac{-\alpha-\beta}{\gamma-\beta}, \infty\Big)$ such that $d_u(G^*) = 0$. Next, we note that $d_u(0) = -\alpha < 0$ and that the derivative $d_u'(G)$ is strictly positive for $G \in \Big[0, \frac{-\alpha-\beta}{\gamma-\beta}\Big)$. Combining this with the above results, we obtain that there exists exactly one $G^* \in \Big[0, \frac{-\alpha-\beta}{\gamma-\beta}\Big)$ such that $d_u(G^*) = 0$.

Case (2): $-\alpha \leq \beta \leq \gamma$: We have that $\lim_{G\to+\infty} d_u(G) = \gamma > 0$ and that $d_u(0) = -\alpha < 0$. Also the derivative $d_u'(G)$ is strictly positive for all $G \geq 0$. The two results combined show that there exists exactly one $G^* \in [0, \infty)$ such that $d_u(G^*) = 0$.

Case (3): $\beta < -\alpha < 0$: First we note that

$$\begin{aligned} d_u\Big(\frac{-\alpha-\beta}{\gamma-\beta}\Big) &= e^{-\frac{-\alpha-\beta}{\gamma-\beta}}(\beta-\gamma) + \gamma \\ &< \Big(1 - \frac{-\alpha-\beta}{\gamma-\beta}\Big)(\beta-\gamma) + \gamma = -\alpha < 0 \end{aligned}$$

and that $d_u(0) = -\alpha < 0$. Further, the derivative $d_u'(G)$ is strictly negative when $G \in \Big[0, \frac{-\alpha-\beta}{\gamma-\beta}\Big)$. Since $d_u(G)$ is continuous in G, it follows that there does not exist $G^* \in \Big[0, \frac{-\alpha-\beta}{\gamma-\beta}\Big)$ such that $d_u(G^*) = 0$. Similarly, we have that $\lim_{G\to+\infty} d_u(G) = \gamma > 0$, and that the derivative $d_u'(G)$ is strictly positive for all $G \in \Big[\frac{-\alpha-\beta}{\gamma-\beta}, \infty\Big)$. From the above results in this case, there exists exactly one $G^* \in \Big[\frac{-\alpha-\beta}{\gamma-\beta}, \infty\Big)$ such that $d_u(G^*) = 0$.

Figure 8 illustrates the three cases of β. The above result shows that a unique $G^* > 0$ exists such that $d_u(G^*) = 0$ for the given α, γ, and any β. Using (5a), the unique control signal u^* at the operating point is thus given by $u^* = \lambda^{-1}\big(G^* e^{-G^*}\big)$,

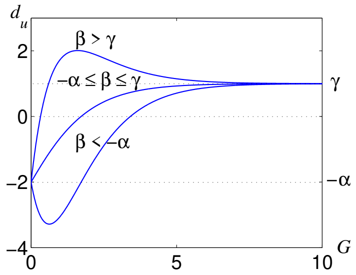


Fig. 8. Drift $d_u(G)$ for different choices of β.

where λ^{-1} denotes the inverse of λ. And the backlog n^* is given by $n^* = \frac{G^* - \lambda(u^*)}{q} = \frac{G^*}{q}\big(1 - e^{-G^*}\big)$ from (5b). (n^*, u^*) is therefore the unique operating point required for Proposition 3. ■

Appendix III
Proof of Lemma 1

Suppose Assumption 1 holds for the given α and γ. Let $\widehat{G}$ be the given offered load and β be as given in the lemma. Note that, by rearranging terms, the choice of β is such that $\widehat{G}$ is a solution to the relaxed drift equation Eq. (18),

$$-\alpha e^{-\widehat{G}} + \beta \widehat{G} e^{-\widehat{G}} + \gamma\Big(1 - e^{-\widehat{G}}(1 + \widehat{G})\Big) = 0.$$

Suppose $\lambda\big(\max\{\alpha, -\beta\}\big) \geq e^{-1}$. Then by Proposition 3, there exists a unique operating point (n^*, u^*) such that Eq. (5) holds. Also by definition, the drift Eq. (4b) at the operating point vanishes, such that

$$\begin{aligned} &\max\{-\alpha, -u^*\} e^{-G^*} + \max\{\beta, -u^*\} G^* e^{-G^*} \\ &\quad + \gamma\big(1 - e^{-G^*}(1 + G^*)\big) = 0. \end{aligned}$$

Following the argument in Appendix II for the equivalence of the drift Eq. (4b) and the relaxed drift Eq. (18) in this case, G^* also uniquely solves

$$-\alpha e^{-G^*} + \beta G^* e^{-G^*} + \gamma\big(1 - e^{-G^*}(1 + G^*)\big) = 0.$$

Since G^* is unique, we have that $G^* = \widehat{G}$.

Suppose $\lambda\big(\max\{\alpha, -\beta\}\big) < e^{-1}$. It can be shown [14] that in this case multiple operating points may exist, depending on the choice of α, γ and $\widehat{G}$, for which Eq. (5) holds. But for any such operating point G^*, we have that $G^* \leq \widehat{G}$. Due to space constraints, however, we omit the details here. ■