Les Cahiers du GERAD ISSN: 0711–2440

Congestion Avoidance with Future-Path Information

P. Jacko
B. Sansò

G–2007–70

September 2007

Congestion Avoidance with Future-Path Information

Peter Jacko
Department of Statistics
Universidad Carlos III de Madrid
Av. de Universidad 30
289 11 Leganés (Madrid), Spain
peter.jacko@uc3m.es

Brunilde Sansò
GERAD and Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. Centre-ville
Montréal (Québec) Canada, H3C 3A7
brunilde.sanso@polymtl.ca

September 2007

Les Cahiers du GERAD
G–2007–70

Abstract

In this paper we analyze the trade-off between admission costs and receiver rewards of TCP Tahoe flows competing for buffer space. Since the buffer space is a scarce resource during heavy traffic and congestion epochs, it is important to understand in which circumstances packet dropping may be optimal. We develop a restless bandit model for assessing an economic value of packets at routers they encounter in transmit. We then argue that the economic value of the whole network increases if the packets with lower economic value are the preferred candidates for dropping or marking in congestion avoidance mechanisms. Such changes are arguably expected to lead both to a lower delay and higher network throughput.

Key Words: Congestion Avoidance; Fairness; TCP; Restless Bandits, Marginal Productivity Index; Economic Value.

Résumé

Dans cet article nous analysons le "trade-off" entre les coûts d'admission et les revenus des flots TCP Tahoe qui compétitionnent pour de l'espace tampon. Puisque l'espace tampon est une ressource rare lors des situations de congestion, il est important de comprendre dans quelles circonstances il est optimal de se débarasser des paquets. Nous développons un modèle de "restless bandit" pour évaluer la valeur économique des paquets dans les routeurs. Ensuite, nous utilisons l'argument que la valeur économique du réseau augmente si les paquets avec des valeurs inférieures sont délestés en premier ou marqués dans des mécanismes pour éviter la congestion. On s'attend à ce que ces changements nous amènent à moins de délais et plus de débit.

Acknowledgments: This research has been supported in part by the Spanish Ministry of Education and Science under grant MTM2004-02334 and an associated Postgraduate Research Fellowship, by the Autonomous Community of Madrid-UC3M through grant UC3M-MTM-05-075, and by the European Union's Network of Excellence Euro-NGI. This work was mostly done during the stay of Jacko at GERAD.

1 Introduction

With the growth of traffic volume and traffic heterogeneity in best-effort networks, congestion control has been getting more importance. The initial naïve implementation of the queue tail drop policy showed to be prone to creating various serious problems including bias against bursty traffic and global synchronization, eventually resulting in congestion collapse [1]. Such a *reactive* congestion control has thus a significant negative impact on the efficiency of scarce resources (bandwidth and buffer space) allocation in networks.

Alternative proposals focused on *preventive* congestion control developing *congestion avoidance mechanisms*, such as RED [4], BLUE [3], and a palette of their variants, which try to detect congestion in its early stage and warn the traffic sources expecting that they decrease their transmission rates. Yet, packet losses in the Internet are still high and Quality of Service (QoS) strongly suffers from this fact.

In order to avoid packet losses resulting from congestion avoidance mechanisms, *explicit congestion notification* (ECN) has been proposed. ECN marks a bit in the packet header instead of dropping the packet, notifying the receiver about the congestion experienced during the transit. This information is then echoed back to the sender, which is expected to react.

Nevertheless, the choice of packets to be dropped or marked is done myopically—only considering the current state or recent history of router-based measures (such as queue length, packet loss, link utilization, etc.). In particular, *economic value* of packets is not taken into account in congestion avoidance mechanisms.

Closely tight to congestion avoidance is the issue of *fairness*. It arises in virtually all congestion avoidance mechanisms proposed in the literature, and the approach that seems to be generally accepted is to treat all packets (or flows) fairly according to their size, i.e., each packet Byte is supposed to have the same economic value for the receiver. More importantly, it is implicitly supposed that this value is the same throughout the whole route. There are, however, strong reasons to challenge such an understanding of fairness.

Economic value of a network should be assessed through the service it provides—transport of data. Thus, it must be a measure of all *received data* meeting receiver-specified QoS. In this paper we develop a theoretical framework for assessing economic value of packets at routers they encounter on their route to the receiver. We then argue that the economic value of the whole network increases if the packets with lower economic value are the preferred candidates for dropping or marking in congestion avoidance mechanisms.

Dropping a packet on its route implies that all the scarce resources it has consumed so far are wasted. It is then intuitively appealing that when a scarce resource is to be allocated to a packet, the possibility of getting that packet lost in the remainder of its route must be taken into account. In the networks, where precise information about the packet future path may not be available, the "allocation-decision-maker" may infer or estimate required

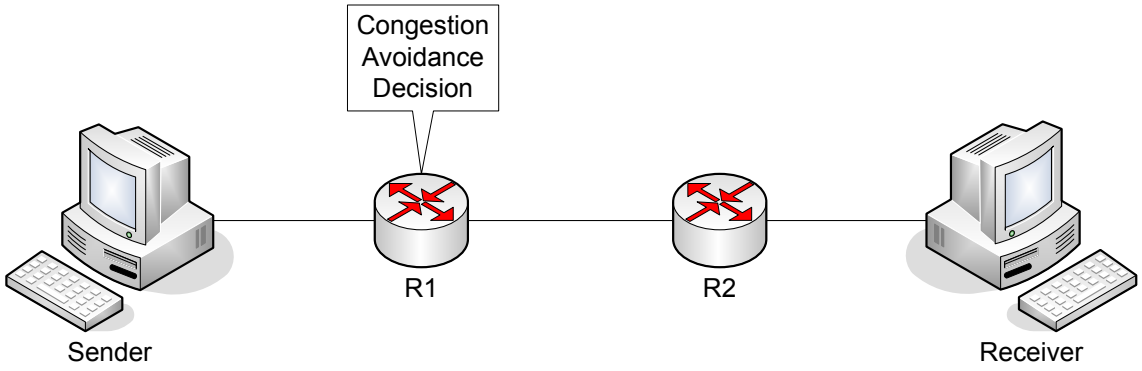


Figure 1: A design of an end-to-end connection.

information. For instance, observation of the ECN bits on the opposite way could be used to estimate it.

To illustrate the idea on an example, consider an end-to-end connection that includes two bottleneck routers, as in Figure 1. If router R2 is busy (yet still have some free space in the buffer) and router R1 is able to anticipate it, then congestion avoidance decisions at R1 should take into account the transmission rate of an incoming flow. If the rate is small, so that R2 would be able to service it, the flow should be admitted at R1. On the other hand, if the transmission rate is too high, so that R2 is likely to drop it, the flow should not be admitted at R1; or, it could be a strong candidate for drop policy implemented in the congestion avoidance mechanism at R1. Thus, defining flow value as the expected number of packets that arrive to the receiver implies that packet value can differ in different points of the path.

The main objective of this paper is to verify that future-path information can improve the resource allocation decisions, help understand how it translates into quantitative terms, and propose improvements of congestion avoidance mechanisms which would take into account an *economic value* of packets in the network. For that end, we analyze *TCP Tahoe* in the framework of *restless bandits*. The restless bandit model is analytically tractable and extremely powerful in assessing the economic value of the flow via the *marginal productivity indices* (see Section 2).

Some of the recent theoretical proposals for future generation networks may benefit from the results we present. For example, the *flow-aware networking* proposal [7] considers every flow separately, and that is exactly how we model the problem in this paper. Further, in such a network, one can introduce certain priority parameters of each flow that would capture their relative importance, e.g. by assigning importance weights to flow receivers. Combining this with the marginal productivity indices may lead to a powerful control mechanism.

In Subsection 1.1 we describe TCP Tahoe in more detail. In Section 2 we introduce briefly the bandit problem, develop a model of TCP Tahoe, and state its optimal control. Practical implementation of our results is outlined in Section 3. In Section 4 we discuss model limitations and an ongoing work.

1.1 TCP Tahoe

TCP Tahoe is a simple additive-increase/multiplicative-decrease (AIMD) transmission control protocol (TCP) we model in Section 2.1. The actual packet transmission rate is maintained by the variable *actualWindow*, which ranges between the minimum value of 1 packet and the maximum number of packets given by *advertisedWindow*. It has two phases: *slow start* phase, which is between the minimum transmission rate and *congestionThreshold*, and *congestion avoidance* phase, which is between *congestionThreshold* and *advertisedWindow*.

The dynamics of TCP Tahoe is as follows. After every period with no packets lost, TCP Tahoe doubles the *actualWindow* during the slow start phase, whereas it increments the *actualWindow* by 1 packet during the congestion avoidance phase. If a packet is lost, the generator restarts the transmission rate at the minimum value of 1 packet. Note that it does not incorporate *fast recovery* nor *fast retransmit*.

2 Restless Bandit Model

The (multi-armed) bandit problem [**?**, cf.]]Gittins1979 is a resource allocation model capturing the fundamental trade-off between *exploitation* of the present (reactive control) and *exploration* of the future (preventive control). An appealing feature of bandit models is the *priority-index policy* solution. The classical examples of optimality of an index policy are the $c\mu$-rule for multi-class $M/G/1$ queues [2] and the Gittins index policy for the classic bandit problem [5]. For a more complex *restless bandit* problem [6] introduced so-called *marginal productivity indices* that generalize all the above indices. Well-performing indices often have an economic interpretation, which we exploit in this paper.

To make an analogy with our problem of interest, consider a router with finite buffer (scarce resource), for which several flows (bandits) compete. Flows generate certain reward for receivers, if they arrive to them (i.e., if they are admitted in the buffer). The difficulty is that these flows are dynamically changing their transmission rate, so the rewards may increase or decrease later on. Thus, the problem is whether to exploit the present rewards, or take a myopically-suboptimal action which may yield higher rewards in the future. Based on the bandit problem results, we will decompose this problem, analyze each flow separately, and calculate the marginal productivity indices. In Section 3 we then discuss practical implementation of these indices into congestion avoidance mechanisms in order to improve the economical value of the whole network.

Apart from presenting a novel theoretical framework, [6] proved indexability and derived marginal productivity indices of an admission control problem. He then employed the

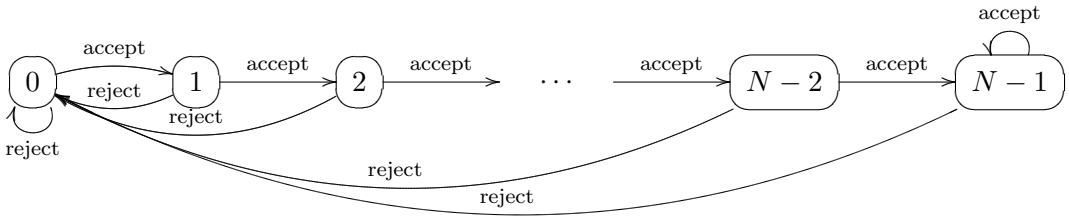


Figure 2: Modeling TCP Tahoe as a Markov decision process.

marginal productivity indices in an index policy heuristic for the problem of routing to parallel queues. The approach we adopt in this paper is somewhat analogous: we first derive the indices for a flow admission control problem and then propose to use these indices in congestion avoidance mechanisms. Yet, the two admission control problems significantly differ. In this paper we face the trade-off between admission costs and receiver rewards of admitted flows, whereas [6] analyzed the trade-off between holding costs of admitted flows and rejection costs of rejected flows. We further allow for a dynamically changing transmission rate (i.e., buffer space requirement), while his model assumed random arrivals.

2.1 Restless Bandit Model of TCP Tahoe

In what follows, *restless bandit* is used as a short name for a binary-action finite-state Markov decision process (MDP) with parametric immediate rewards. In this section we set out to model TCP Tahoe as a restless bandit (see Figure 2).

We set the model in discrete time, defining one time period as one round-trip time (RTT). Let ν be a problem parameter denoting the *admission cost* paid for each unit of buffer required by the TCP. We assume that all packets are of the same size, which is equal one buffer unit.

Our restless bandit model of TCP Tahoe can be defined as follows:

- *State space* is $\mathcal{N} = \{0, 1, \ldots, N-1\}$, with N possible states; state n is defined by a pair (w_n^1, r_n^1), where $w_n^1 > 0$ is interpreted as the buffer utilization required by the TCP (in packets/RTT) and r_n^1 is a reward.
- *Actions admitting* and *rejecting* the flow are available in each state.
- *Dynamics if admitted*: If the TCP is in state n and the flow is admitted at a given period, then during that period it generates reward r_n^1, a cost νw_n^1 must be paid, and the generator moves to state $n+1$ for the next period (or remains in $N-1$, if it is already there).

- *Dynamics if rejected*: If the TCP is in state n and the flow is rejected at a given period, then there is no cost for buffer utilization nor any reward, and it moves to state 0 for the next period.

We will interchangeably call the action of admitting the active action; rejecting will also be called the passive action. Further, we suppose that states are ordered increasingly, so that $w_n^1 < w_m^1$ for any $n, m \in \mathcal{N}$ with $n < m$. The reward r_n^1 can be interpreted as measuring a one-period economic value (utility) of the admitted flow for the receiver; we use an *expected goodput* measure in Section 3. Note that passive action means rejecting the *entire* flow.

For TCP Tahoe, state n is an abstract concept denoting that its current transmission rate is w_n^1 packets/RTT (i.e., w_n^1 is the value of the *actualWindow* variable), yielding those packets a reward r_n^1. If *advertisedWindow* is assumed to be constant over time, then $w_{N-1}^1 = advertisedWindow$. Further, we have $w_0^1 = 1$.

2.2 Optimization Problem

We consider the problem of finding an optimal *index policy* control of TCP Tahoe over an infinite time horizon, for both *discounted criterion* with discount factor $0 < \beta < 1$ and *long-run average criterion*. In the remainder of this section we follow the analysis introduced in [6]. Our analytical focus will be on the former, whose marginal productivity indices can be directly extended to the latter by taking limit $\beta \to 1$. These two criteria are the most appropriate for applications such as computer networks.

From the MDP theory it follows that there exists an optimal stationary policy independent of the initial state, therefore we narrow our focus only to those policies and represent them via *active sets* $\mathcal{S} \subseteq \mathcal{N}$. In other words, a policy $\mathcal{S}$ prescribes to be active in states in $\mathcal{S}$ and passive in states in $\mathcal{S}^{\mathrm{C}} := \mathcal{N} \setminus \mathcal{S}$. This view is crucial in this approach, as it admits a combinatorial optimization formulation of the optimal control problem, which we develop next.

Let us denote $f_n^{\mathcal{S}}$ the total expected discounted reward under policy $\mathcal{S}$ starting from state n, defined by

$$f_n^{\mathcal{S}} = \mathbb{E}_n^{\mathcal{S}} \left[\sum_{t=0}^{\infty} \beta^t r(t) \right], \tag{1}$$

where $r(t)$ is the reward at time t identified by the actual state and the action applied. The symbol $\mathbb{E}_n^{\mathcal{S}}$ denotes the conditional expectation given that the initial state is n and the policy applied is $\mathcal{S}$. Similarly, $g_n^{\mathcal{S}}$ is the total expected discounted work under policy $\mathcal{S}$ starting from state n

$$g_n^{\mathcal{S}} = \mathbb{E}_n^{\mathcal{S}} \left[\sum_{t=0}^{\infty} \beta^t w(t) \right], \tag{2}$$

where $w(t)$ is the buffer utilization (or simply *work*) at time t identified by the actual state and the action applied. Then, formulated for initial state n, the optimization problem is

$$\max_{\mathcal{S}\subseteq\mathcal{N}} f_n^{\mathcal{S}} - \nu g_n^{\mathcal{S}}. \tag{3}$$

2.3 Marginal Productivity Indices

Throughout this subsection we suppose that the immediate reward r_n^1 is concave in w_n^1 and present an optimal index policy for such case. Because of space restrictions, we omit detailed analysis and proofs.

Assumption 1 (Concave Rewards) *There is a real-valued function r with $r(0) \geq 0$, which is concave on the domain $\{0, w_0^1, \ldots, w_{N-1}^1\}$, such that $r_n^1 = r(w_n^1)$.*

Theorem 1 *Under concave rewards the problem* (3) *is indexable, having the marginal productivity index of state n for the discounted criterion be given by*

$$\nu_n = \frac{r_n^1 + \sum_{m=0}^{n-1} \beta^{m+1}(r_n^1 - r_m^1)}{w_n^1 + \sum_{m=0}^{n-1} \beta^{m+1}(w_n^1 - w_m^1)} \tag{4}$$

and for the long-run average criterion by

$$\nu_n = \frac{(n+1)r_n^1 - \sum_{m=0}^{n-1} r_m^1}{(n+1)w_n^1 - \sum_{m=0}^{n-1} w_m^1}. \tag{5}$$

Proposition 1 *Under concave rewards we have $\nu_n \leq r_n^1/w_n^1$.*

3 Practical Considerations

We now narrow our focus to problem (3) under long-run average criterion, which is more appropriate for real-time situations. The discounted criterion could be useful in networks with relatively large round-trip times or short lifetime (not discussed here). Practitioners would also agree that flows that are, before arriving at the receiver, expected to encounter subsequent gateways with buffers (see Figure 1 for illustration) will satisfy the assumption of concave rewards.

Table 1: The maximum integer values of *advertisedWindow* assuring concavity of $r(w)$.

$l \setminus p$	0.0001	0.001
1	19998	1999
5	3999	399

3.1 Indices for TCP Tahoe

For the sake of simplicity, we suppose that the TCP Tahoe parameters can be expressed as powers of 2. In particular, $advertisedWindow = 2^{W-1}$ and $congestionThreshold = 2^{T-1}$ for some positive integers $T \leq W$. For TCP Tahoe we have $w_n^1 = 2^n$ for all $n \leq T-1$, and $w_n^1 = 2^{T-1} + n - (T-1)$ for all $T \leq n \leq 2^{W-1} - 2^{T-1} + T - 1$, having $T + 2^{W-1} - 2^{T-1}$ states. Recall that the *work* w_n^1 gives the number of packets sent by the flow generator.

The reward function is the *expected goodput*, i.e. the expected number of useful Bytes received by the receiver: $r(w) = (1-p_1)^w(1-p_2)^w \dots (1-p_l)^w ws$, where $0 \leq p_k < 1$ is the probability of losing a packet on the k-th link of the connection *after* it passes the router, at which we implement the congestion avoidance mechanism. Further, $s > 0$ is the useful size in Bytes of each packet and we assume $l \geq 1$.

Thus, the receiver gets a reward of ws if the *entire* flow arrives, and nothing if at least one of the packets is lost in the network. See the actually considered connection in Figure 1, but with l links (i.e., $l-1$ routers) along the connection after the router R1. In the following, we simplify the expression assuming a constant dropping probability $p_k = p$ for all $k = 1, 2, \dots, l$, so that $r(w) = (1-p)^{lw} ws$. The function $r(w)$ is concave if and only if $w \leq \dfrac{-2}{l \ln(1-p)}$, which is approximately $2/(lp)$ (see Table 1).

Now we are ready to see the marginal productivity indices of TCP Tahoe in quantitative terms. Figure 3 displays the marginal productivity indices with different values of *congestionThreshold*, for $advertisedWindow = 512$ packets and packet size $s = 1$. State 0 with flow equal to 1 packet has the highest priority index, and the indices are smaller for larger *actualWindow*. That is, the economic value of each packet becomes smaller as the transmission rate increases.

Apart from the TCP Tahoe with no congestion avoidance phase (discrete points), Figure 3 exhibits the marginal productivity indices for TCP Tahoe with *congestionThreshold* equal to a half and a quarter of *advertisedWindow*. The inclusion of the congestion avoidance phase slightly diminishes the indices in the congestion avoidance interval, while those that are below *congestionThreshold* are not affected. Recall the formula (5): the indices only depend on the states with smaller flows. In particular, they are independent of *advertisedWindow*.

The marginal productivity indices are nonincreasing and they are below the curve $r(w_n^1)/w_n^1$ (Proposition 1), included in Figure 3 for comparison. Further, especially for values within the congestion avoidance phase and close to the maximum concavity-assuring *actualWindow* value, the index is negative. That is, if the admission cost parameter ν is positive, the optimal action is to reject that flow. Even if we had $\nu = 0$, which could be the case of infinite buffer, it is optimal not to admit the flow.

Roughly speaking, the marginal productivity index decreases sublinearly from 1 to 0 as the *actualWindow* increases from 1 packet to the maximum concavity-assuring value. Note that Figure 3(d) shows valid indices only for *actualWindow* smaller than 400 packets (see Table 1). For larger values the concavity assumption is not satisfied and the curve is only illustrative.

Finally, comparing Figure 3(a) to (d) we can observe the effect of future-path packet dropping probability. In Figure 3(a), with the smallest dropping probability and the smallest number of connection links, the economic value of each packet diminishes slightly, maintaining around 90% of its value even at the transmission rate of 512 packets. Figure 3(c) presents roughly ten-times more deteriorated connection with respect to Figure 3(a) and suggests that each packet transmitted at the highest rate only has a half of the single-transmitted packet value. In even less reliable connection in Figure 3(d) the packet value decreases dramatically, and becomes negative even at moderate transmission rates.

3.2 Implementation of Priority Indices in Congestion Avoidance Mechanisms

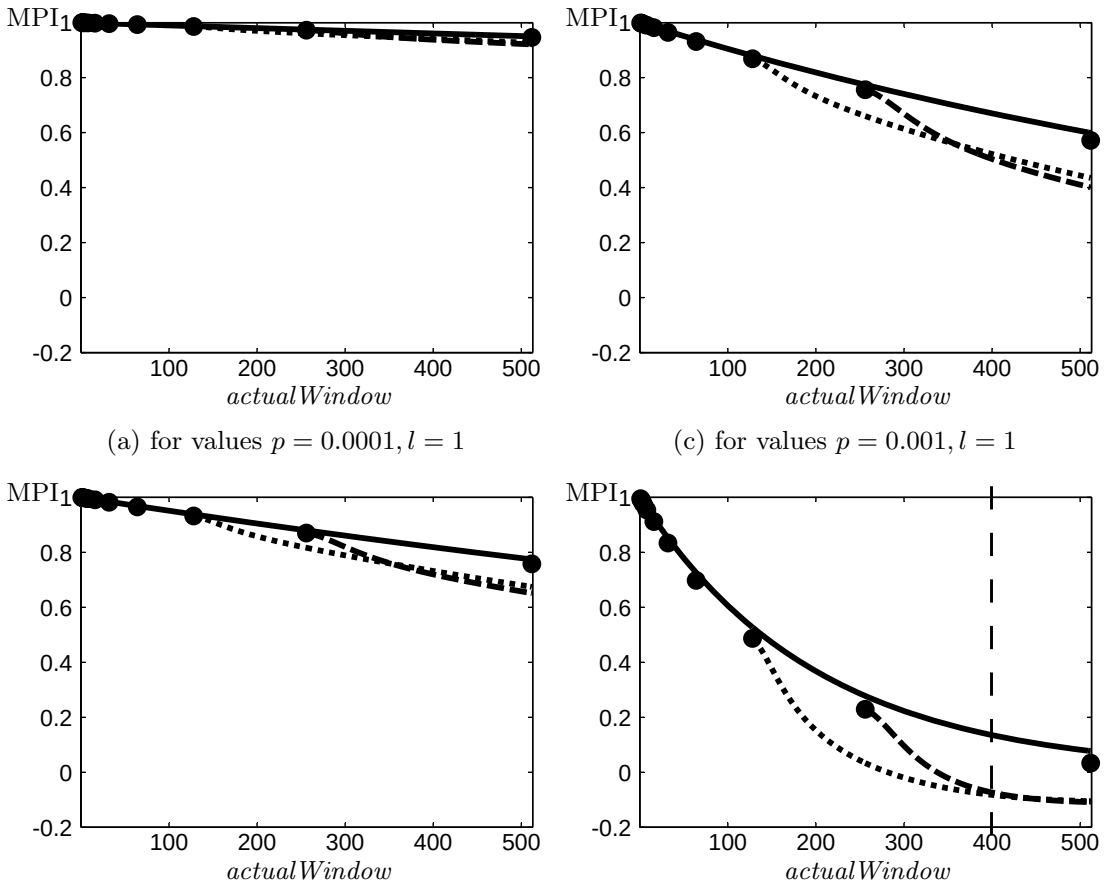
Consider any congestion avoidance mechanism (e.g., RED) with the following property: on an arrival of a packet, it calculates the probability of dropping it, generates the random event, and eventually drops the packet. We will now discuss how this mechanism may be modified so that it takes into account the economic value of the packet during the dropping decision stage.

The economic value of a packet can be evaluated via the marginal productivity index multiplied by its goodput size in Bytes, say $\nu_n s_n$. Let the dropping probability calculated by the congestion avoidance mechanism for this packet at a given time be p_n. We next discuss what should be the dropping probability p_m if a packet with economic value $\nu_m s_m$ arrived instead.

For fair admission control we may want to impose that the *expected economic loss* for dropping packets should be equal. Hence,

$$p_m = \frac{\nu_n s_n}{\nu_m s_m} p_n. \tag{6}$$

As a consequence, either all or none of the incoming packets have a zero dropping probability.



(a) for values $p = 0.0001, l = 1$

(c) for values $p = 0.001, l = 1$

(b) for values $p = 0.0001, l = 5$

(d) for values $p = 0.001, l = 5$

Figure 3: Marginal productivity indices (MPI) as a function of *actualWindow* between 1 and 512 packets. The indices refer to TCP Tahoe with *congestionThreshold* = 512 packets (discrete points), 256 packets (dotted line), and 128 packets (dashed line), respectively. The solid curve depicts the function $r(w)/w$.

When the buffer is heavily congested and $p_n = 1$, then all other packets should experience the same dropping probability. Yet in that case, according to (6), the dropping probability p_m will be larger than 1 for less valued packets, and smaller than 1 for more valued packets.

An alternative formula that satisfies $p_n = 1 \Longleftrightarrow p_m = 1$, is the following:

$$p_m = 1 - (1 - p_n)^{\frac{\nu_n s_n}{\nu_m s_m}}. \tag{7}$$

Note that the two formulae are roughly equivalent for small values of dropping probability p_n. Any of them can hence be implemented in the original RED, in which the dropping probability is maintained at very low levels (except for the aggressive stage when all incoming packets are being dropped). For congestion avoidance mechanisms, in which the dropping probability smoothly increases to 1, the latter should be the preferred formula.

4 Conclusions and Ongoing Work

In this paper we have developed a model of TCP Tahoe and derived its optimal index policy. We have further proposed an implementation of those indices, measuring a marginal productivity of admitting a packet, in congestion avoidance mechanisms, arguing that such a modification can lead to more efficient utilization of network scarce resources.

The main limitation for generalization of our results is that TCP Tahoe does not have an implementation of fast recovery nor fast retransmit. An analogous analysis in the restless bandit framework for other TCP variants is under development. Nevertheless, the main drawback of any restless bandit model is the allowance for only one type of reaction to congestion.

From the practical point of view, however, we believe that the outcome of our model is roughly preserved also in more complicated mechanisms. The reason being that the economic value depends on the actual transmission rate much more strongly than on other aspects of the dynamics of the mechanism. These considerations are part of on-going work.

References

[1] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang. RFC 2309: Recommendations on queue management and congestion avoidance in the Internet. Available as RFC 2309, April 1998. ftp://ftp.isi.edu/in-notes/rfc2309.txt.

[2] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Advances in Applied Probability*, 17(1):237–238, 1985.

[3] W. Feng, K. G. Shin, D. D. Kandlur, and D. Saha. The BLUE active queue management algorithms. *IEEE/ACM Transactions on Networking*, 10(4):513–528, 2002.

[4] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, 1993.

[5] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177, 1979.

[6] J. Niño-Mora. Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Mathematical Programming, Series A*, 93(3):361–413, 2002.

[7] S. Oueslati and J. Roberts. A new direction for quality of service: Flow-aware networking. In *Proceedings of the Next Generation Internet Networks*, pages 226–232, Roma, Italy, 18-20 April 2005.