DETERMINISTIC PACKET MARKING FOR MAXIMUM LINK PRICE ESTIMATION

R.W. Thommes and *M.J. Coates*
Department of Electrical and Computer Engineering, McGill University
3480 University St, Montreal, QC, Canada H3A 2A7
Tel : 514-398-7137 Fax : 514-398-4470
Email: rthomm@tsp.ece.mcgill.ca, coates@ece.mcgill.ca

ABSTRACT

A recently proposed congestion control algorithm, MaxNet, achieves MaxMin fairness for a variety of utilization functions. MaxNet requires every source to have information about the price – a measure of congestion – of the most congested link in the path to the sink. In this paper we describe a deterministic packet-marking algorithm which conveys the maximum link price to the source. Our algorithm may be incorporated into MaxNet or any other congestion control scheme based on maximum link prices. The approach we describe achieves high efficiency while making as few changes as possible to the TCP/IP protocol suite. The algorithm makes use of the 2-bit ECN field in the IP header to allow routers to encode price information, and the 1-bit ECE field in the TCP header to allow the sink to communicate its maximum link price estimate back to the source.

1. INTRODUCTION

In a highly utilized Internet-type network, congestion control is necessary to achieve acceptable performance. Congestion control algorithms signal each source with a measure of the amount of congestion, referred to as a price, along the end-to-end path from source to sink. A source may then adjust its transmission rate accordingly. Algorithms currently deployed in the Internet, such as TCP Reno, make use of the total path price (defined as the sum of individual link prices making up the path). In a simple, commonly used implementation, routers along the path drop packets to convey a measure of total path price to the source. There are several proposed techniques which utilize packet-marking rather than packet-dropping to indicate the total path price. These include two probabilistic marking techniques: Random Exponential Marking (REM) [1], and Random Additive Marking (RAM) [2], as well as a deterministic marking scheme we have previously proposed [3].

A newly proposed congestion control algorithm, MaxNet [4], only requires sources to have information about the most congested link in a path. MaxNet is an attractive alternative, because, as the authors show, it achieves *MaxMin* fairness for a wide range of *utility functions*. A utility function is defined by the relationship between source transmission rate and the congestion price. MaxMin fairness indicates that every source in a network is transmitting at the maximum rate possible without lowering the rate of another source transmitting at an equal or lower rate. In order to convey the maximum link price (MLP), the authors indicate that a packet "must include bits to communicate the complete congestion price", but the details lie outside the scope of their paper. This serves as the primary motivation for our paper. We will specify a deterministic marking algorithm which estimates the MLP along an end-to-end path from source to sink in a TCP/IP network, and conveys the information back to the source. The algorithm is a modified, more efficient version of our marking scheme for total path price estimation [3]. In order to make the potential deployment of our algorithm more attainable, we have attempted to make as few changes as possible to the TCP/ IP protocol suite. Given the restriction the TCP protocol places on one's ability to embed or feed back information, the approach we describe is a very efficient mechanism for conveying MLP information to the source.

2. BACKGROUND INFORMATION

Our MLP estimation algorithm does not require any changes to the packet header fields, but does use the 2-bit ECN (Explicit Congestion Notification) field in the IP packet header and the 1-bit ECN-Echo (ECE) field in the TCP header in a different manner than what is described in RFC 3168 [5]. Since the document is a proposal which has not been standardized, there is still flexibility in how routers choose to mark packets. Our algorithm does retain 00 an indication that ECN is not supported, as suggested by the RFC. However, contrary to the RFC specification, we require that all packets have their ECN field initialized to 01, and that routers may set the ECN field to either 10 or 11 according to our specification below.

The algorithm also makes use of the `IPid` field, but does not modify its value. As described in RFC 791 [6], the 16-bit `IPid` field provides a means of distinguishing fragments in order to facilitate reassembly. In order to achieve a low probability of two packets belonging to the same flow having the same `IPid` value while both are in flight, most Internet hosts either assign consecutive `IPid`s to successive packets in a flow or assign random `IPid`s.

3. ALGORITHM SPECIFICATION

Our algorithm requires every link price s_i to be bounded and normalized: $0 \leq s_i \leq 1$. Each router calculates a b-bit quantization of the congestion price of outgoing links. The key idea of the algorithm is that there is a finite number of possible quantized prices a link may take on, and each packet ascertains whether any links currently take on 1 of 2 particular prices. The specific link prices a given packet is concerned with depends on its *probe type*. A packet's probe type is determined by its `IPid` field. There are $m = 2^{b-1}$ unique probe types, one for each of the possible values taken on by the $b-1$ most-significant bits (MSBs) of a link price. For example, if $b = 4$, probe type 0 is concerned with the two prices whose MSBs are 000. It ascertains if there is a link price (or link prices) on the path in the range of [0000,0001], and, if so, what the maximum price is in this range. Similarly, probe type 7 is concerned with prices in the range [1110,1111]

When a packet arrives at a router, the router calculates the packet's probe type using a modulo operation: $\text{ProbeType} = \texttt{IPid} \bmod m$. The routers accesses a static look-up table mapping probe types to the link prices with which they are concerned. If a router determines a match between the link price and the probe type, it modifies the packet's ECN field based on that field's current value and the least significant bit (LSB) of the link price. If the ECN field is 01, and the LSB is 0, it sets the ECN field to 10. If the LSB is 1, it sets the ECN field to 11.

Once the sink has received at least one instance of each probe type, it can inspect the ECN fields to determine the maximum quantized price of any link. It scans through the probe types, in order of highest to lowest associated price ranges. Proceeding in this manner, the first probe the sink finds to have an ECN field not equal to 01 determines the MLP.

In order to notify the source of the current MLP, the sink makes use of the ECE bit in packets sent back to the source. Each such packet encodes one of the b bits of the currently estimated MLP. Packets are mapped to a different set of probe types based on the `IPid` field, in this case using a modulo b operation. Probe type 0 carries the LSB of the price, while probe type $b-1$ carries the MSB. Once the source has receive one instance of each sink probe type, it can discern the MLP.

4. PERFORMANCE ANALYSIS

As the MLP estimation algorithm proceeds, it will always be in one of three phases. The first phase, *Outdated Estimate* (OE) begins immediately after a change in the MLP. During this interval, neither the sink nor source estimates are correct. We note that a correct estimate e_c is defined as differing from the true MLP p_m by no more than the maximum possible quantization error: $p_m - \frac{1}{2^{(b+1)}} \leq e_c \leq p_m + \frac{1}{2^{(b+1)}}$. Upon receiving instances of the downstream probe types reflecting the new MLP, the sink corrects its estimate. At this point the *Correct Receiver Estimate* (CRE) phase begins. It lasts until the source receives the upstream probe types necessary to mirror the correct estimate of the sink. Thus begins the *Correct Estimate* (CE) phase, which lasts until the next MLP change. Two price changes in quick succession may result in either the CE or both the CE and CRE phase being of length zero.

Next we will examine the characteristics these phases, including the distribution of their lengths and the estimation errors at the source. Lengths are defined according to the number of downstream packets received, and in the case of fixed-rate packet arrivals may be converted to time lengths by multiplying by the rate parameter. The length of an OE interval depends on whether the associated MLP change was an increase or decrease. In the case of an increase, the sink must only receive a probe type carry information about the interval in which the new MLP lies. With randomly generated `IPid`s, the number of packets that the sink sees before receiving the one necessary probe type is geometrically distributed with parameter $1/m$. With sequentially increasing `IPid`s, the distribution is uniformly distributed in the interval $[1, m]$. If the MLP decreases, the sink must receive a probe type carrying information on the interval of the previous MLP so that the sink can deduce that no links lie in that price range anymore. It may also require a probe type with information on the new MLP interval. However, in some cases it will already have this information and can form a correct estimate after receiving a probe type "clearing" the previous MLP (recall that the sink stores the value of the most previously received instance of each probe). The proportion of instances in which a price decreases requires the sink receiving only one probe type before forming a correct estimate is dependent on the number of links making up the path, the frequency at which individual link prices change, and the possible magnitudes of price changes. The distribution of the estimation error during the OE and CRE phases is also dependent on these factors.

The length of the CRE depends on the number of bits that have changed between the sink's current correct MLP estimate and its previous estimate. If $1 \leq q \leq b$ bits have changed, the source has to receive q unique upstream probe types to correct its estimate. In the case of random `IPid`s, the corresponding number of packets the source has to receive is distributed according to a sum of geometric distributions with parameters $\frac{b}{q}, \frac{b-1}{q}, ..., \frac{1}{q}$. With sequential `IPid`s, the number of required packets is distributed in the interval $[q, b]$.

The distribution of the CE phase length again depends on the frequency at which link prices change, and the distribution of the other two phases. The only source of error during this interval in the MLP estimate is due to quantization. If the MLP is uniformly distributed, the error will be uniformly distributed in the interval $[\frac{-1}{2^{b+1}}, \frac{1}{2^{b+1}}]$, and the expected mean-squared error is $\frac{1}{12 \cdot 2^{2b}}$.

The preceding discussion has alluded to the two sources of error in our MLP estimation algorithm: quantization noise, and outdated estimates due to delay between changes in the MLP and arrivals of probe types conveying the updated information. There is an inherent trade off between these two errors. A larger value of b reduces the expected quantization noise but increases the estimation delay, i.e. the length of OE and CRE. Given a set of network parameters – the rate and magnitudes of link prices change, the upstream and downstream packet rate, and the `IPid` behaviour – an obvious question is how to choose b to minimize the root mean-squared error (RMSE) of the estimate. Solving this problem analytically is not feasible since not all error and interval length distributions are known. Instead, we will approach this problem using simulation.

5. Simulation Results

Our simulation models a TCP connection over a path comprised of 20 individual links. The source sends data packets to the sink, and the sink sends only pure ACK packets. Since many TCP implementations send one ACK for every two data packets received [7], we fix the upstream packet rate at one half the downstream rate. The price of each link is initially uniformly distributed in the interval $[0, 1]$. Subsequently, each link price changes independently after a fixed number of downstream packets have been sent. The magnitude of each price change is normally distributed, and there are reflective boundaries at 0 and 1. The `IPid`s are randomly generated. Figure 1 provides an example of how a 4-bit version of our algorithm estimates the changing MLP under these conditions.

We are interested in examining the effect that b has on the length of OE and CRE. The empirical CDFs of the interval lengths for b set to 3, 4, and 5, are presented in Figure 2. In all cases, the standard deviation of link price changes is set to 0.1, and prices change after every 100 downstream packets. The mean OE lengths, in order of increasing b are 4.88, 9.78, and 22.23 packets. For the CRE lengths, the means are 6.41, 8.53, and 10.60 As expected, the average lengths of both intervals tend to increase with b.

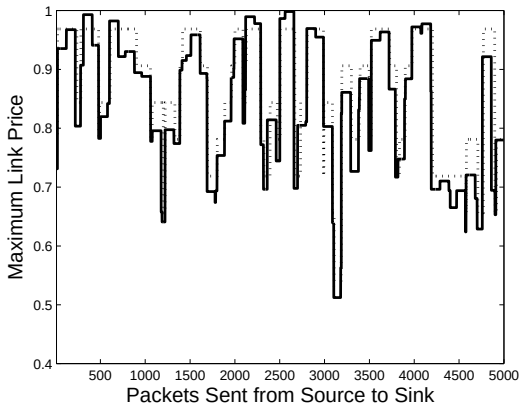
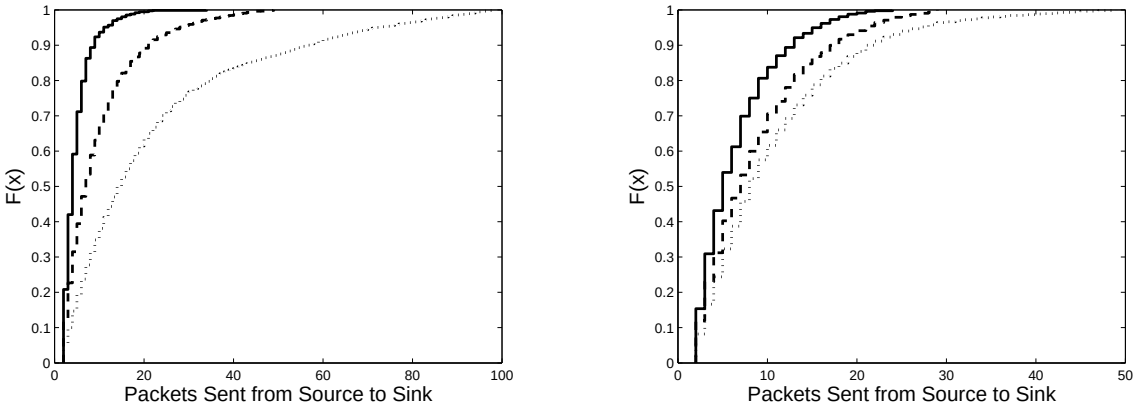


Fig. 1. Maximum link price estimation example. The solid line indicates the actual MLP, and the dotted line the estimate at the receiver. The price of each link changes after every 100 packets, and the change is normally distributed with mean 0 and standard deviation 0.2.

Next, we examine the RMSE of the source estimate during each phase with the same simulation parameters as above. The results are compiled in Table I. From the perspective of the source, the distinction between the OE and CRE phase is largely irrelevant and therefore these two intervals were combined in the simulation. The RMSE during the CE is decreasing in b and significantly smaller than during the OE/CRE phase in all cases. The OE/CRE RMSE is essentially independent of b. It is also worth noting that the CE RMSE values all lie within 16% of the theoretical RMSE for uniform quantization.

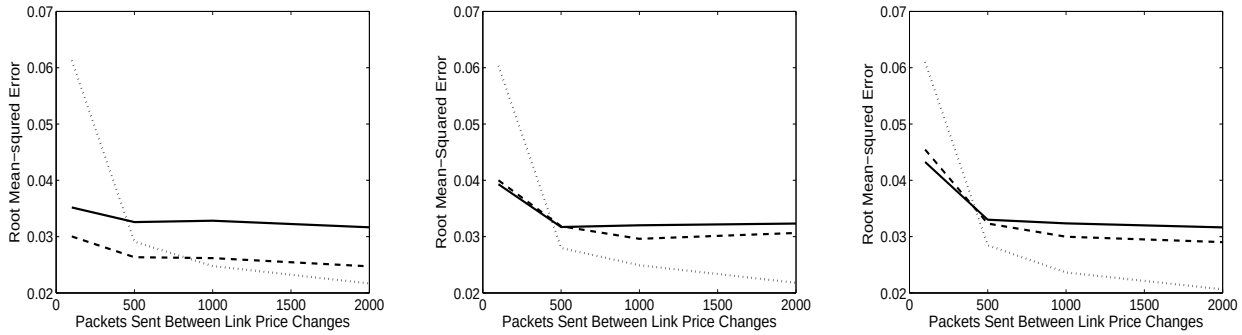
In order to explore the problem of minimizing the estimation RMSE, we consider simulations over a range of link price change-rates (LPCR) and magnitude distributions. The results are illustrated in Figure 3. In all cases, 5-bit quantization results in the largest RMSE for the higher LPCRs, and eventually achieves the best performance as the rate decreases. This can be explained by the following observations: As the LPCR decreases, the expected lengths of the OE and CRE intervals are essentially unaffected, but the length of the CE phase tends to grow. Furthermore, the length of the OE and CRE interval grows with b . Since the RMSE is much higher in the OE and CRE intervals than during the CE phase, it is advantageous to limit the lengths of these intervals by choosing a smaller b when the LPCR is high. However, since increasing b results in a lower RMSE



(a) Empirical CDF of OE Interval Length

(b) Empirical CDF of CRE Interval Length

Fig. 2. Empirical Cumulative Distribution Function of the lengths of the Outdated Estimate and Correct Receiver Estimate intervals. The solid, dashed, and dotted lines represent, respectively, 3,4, and 5 bit price quantization



(a) Price changes normally distributed with $\mu = 0$, $\sigma = 0.1$

(b) Price changes normally distributed with $\mu = 0$, $\sigma = 0.2$

(c) Price changes normally distributed with $\mu = 0$, $\sigma = 0.3$

Fig. 3. Root mean-squared error of MLP estimate as a function of the delay between link price changes. The solid, dashed, and dotted lines represent, respectively, 3,4, and 5 bit price quantization

Phase:	*3-bit quantization*	*4-bit quantization*	*5-bit quantization*
OE and CRE	0.1030	0.0924	0.0995
CE	0.0317	0.0158	0.0076

TABLE I. Root mean-squared error of estimates during OE/CRE and CE phases

during the CE phase, as the LPCR is lowered, the CE phase eventually becomes long enough to warrant the choice of a larger b.

6. CONCLUSION

We have presented a novel algorithm allowing a source in a TCP/IP network to determine the maximum level of congestion of any link along the path to the sink, and examined the problem of choosing an optimal number of quantization bits to minimize the mean-squared estimation error of the algorithm.

REFERENCES

[1] S. Athuraliya, V.H. Li, S.H. Low, and Q. Yin, "REM: Active queue management," *IEEE Network*, vol. 15, pp. 48–53, May 2001.

[2] M. Adler, J-Y Cai, J.K. Shapiro, and D. Towsley, "Estimation of congestion price using probabilitic packet marking," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003.

[3] R.W. Thommes and M.J. Coates, "Determinstic packet marking for congestion price estimation," in *IEEE Infocom 2004*, March 2004.

[4] Bartek Wydrowski and Moshe Zukerman, "Maxnet: A congestion control architecture for maxmin fairness," *IEEE Communications Letters*, vol. 6, no. 11, pp. 512–514, November 2002.

[5] K. Ramakrishnan, S. Floyd, and D. Black, "The addition of explicit congestion notification (ECN) to IP," Sept. 2001, IETF RFC 3168.

[6] J. Postel, "Internet protocol," Sept. 1981, IETF RFC 791.

[7] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, "Modeling tcp throughput: a simple model and its empirical validation," in *ACM SIGCOMM*, Sept. 1998.