SPATIO-TEMPORAL AVAILABLE BANDWIDTH ESTIMATION FOR HIGH-SPEED NETWORKS

Vinay J. Ribeiro, Rudolf H. Riedi and Richard G. Baraniuk

Department of Electrical and Computer Engineering, Rice University
6100 South Main Street, Houston, TX 77005, USA

ABSTRACT

We present a new packet dispersion based active probing scheme for available bandwidth estimation that (1) overcomes system I/O bandwidth limitations on high-speed ($>$1Gbps) networks and (2) determines the location of a path's *tight link*. The scheme is based on the *packet-tailgating* technique earlier used for link capacity estimation and topology identification. Available bandwidth estimation and tight link localization benefits network-aware applications like high-speed grid computing, overlay network routing and server selection. Tight link localization also provides insight into the causes of network congestion and ways to circumvent it.

1. INTRODUCTION

We define the available bandwidth on a network path as the minimum unused capacity of all its links. Numbering links along a path as $i = 0, 1, \dots, N$ and labeling their utilizations u_i and link capacities C_i, the available bandwidth of a path *segment* $h, \dots, k$ $(0 \leq h \leq k \leq N)$ is

$$A[h,k] = \min_{i=h,\dots,k} (1 - u_i)C_i. \qquad (1)$$

The path available bandwidth is $A = A[0,N]$ provides a useful metric for a host of applications including grid computing, overlay network routing, SLA verification, and network monitoring.

Many packet dispersion based available bandwidth tools have been developed in the recent past [1, 2, 3, 4, 5]. All these schemes are based on the principle of self-induced congestion: the probe packets temporarily induce network congestion if and only if the probing bit-rate exceeds the path available bandwidth thus leading to a noticeable increase in queuing delay. The minimum probing bit-rate that causes network congestion hence gives an estimate of the available bandwidth.

High-speed networks with capacities exceeding 1Gbps present several technical challenges for estimation techniques of available bandwidth based on packet dispersion [6]. First, the end-hosts generating and receiving the probing packets could have system I/O bandwidth less than A. Thus probing trains with rate more than the system I/O bandwidth but less than A will experience I/O delays akin to that caused by congestion. As a result the available bandwidth estimates will be conservatively biased. Second, packet arrival interrupts are often coalesced at the NIC cards of end-hosts to increase system efficiency. This adds to the noise in the observed end-to-end packet delay that can affect estimates. We address these challenges in later sections.

We define a path's *tight link* as the one with minimum available bandwidth, that is

$$t = \arg\min_i (1 - u_i)C_i. \qquad (2)$$

Locating the tight link can provide insight into the causes of congestion and ways to circumvent it. Intuition suggests that congestion normally occurs at poorly provisioned peering links or at the very edge of the network [7]. Tight link localization can also enhance certain applications that benefit from knowing whether paths share a common tight link or not [8].

2. PACKET-TAILGATING SCHEMES

Packet-tailgating uses probe trains consisting of large packets interleaved with small tailgating packets. The large packets exit the path midway due to limited TTLs but the small packets travel to the destination while capturing important timing information. Packet-tailgating has been previously used to measure per-hop link capacities [9, 10, 11] and identify network topologies [12].

We propose packet-tailgating for available bandwidth estimation (see Figure 1). Simply replace every probe packet in earlier suggested packet dispersion based available bandwidth schemes with a large packet of size P with TTL l followed closely by a small packet of size p with maximum allowed TTL.

For illustrative purposes consider a CBR probe train of bit rate R and assume that probes encounter CBR fluid cross-traffic at all links. Since large packets exit the path at link l, the effective probing bit rate roughly decreases by a factor of $p/(p + P)$ after link l. Call this bit-rate R_l. We split the analysis into the two following cases.

Case (i): $R < A[0, l]$. In this case the probe packet train will not congest the path on the segment $0, \dots, l$. Under the as-

This work was supported by NSF grant ANI–0099148, DoE SciDAC grant DE-FC02-01ER25462, DARPA/AFRL grant F30602–00–2–0557, and the Texas Instruments Leadership University program.

Email: {vinay, riedi, richb}@rice.edu, URL: spin.rice.edu.

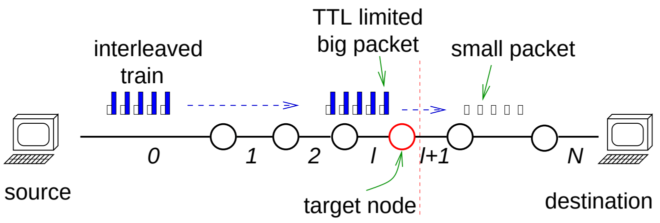

Fig. 1. *Packet train of large packets interleaved with small tailgating packets. Large packets vanish at node l due to TTL expiry.*

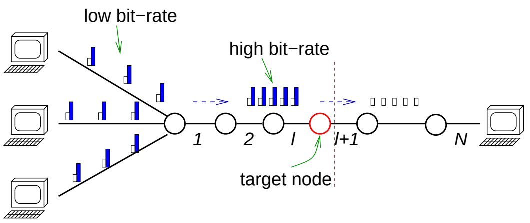

Fig. 2. *Combining multiple sources to increase the probing bit-rate.*

sumption that

$$\frac{A[l+1,N]}{A[0,l]} > \frac{p}{p+P} \tag{3}$$

we have

$$R_l < \frac{p}{p+P}R < \frac{p}{p+P}A[0,l] < A[l+1,N]. \tag{4}$$

Thus we will not observe an increase in queuing delay of probe packets at the destination N.
Case (ii): $R > A[0,l]$. In this case we will observe an increasing delay which allows us to estimate $A[0,l]$. For non-CBR fluid cross-traffic similar arguments can be made by invoking the principle of self-induced congestion. Note that in practice (3) is not an unreasonable assumption since $p/(p+P)$ can be as small as $1/50$.

We next describe how to use packet-tailgating to solve the problems mentioned in Section 1.

3. HIGH-SPEED NETWORK AVAILABLE BANDWIDTH ESTIMATION

This section briefly describes ways to overcome system I/O bandwidth limitations and NIC interrupt coalescence on high-speed networks.

3.1. Limited destination system I/O bandwidth

For now we assume that the source system I/O bandwidth exceeds A but the destination's system I/O bandwidth does not. Regarding the I/O bus at the destination as an extra link numbered $N+1$ on the path, we observe from the discussion in Section 2 that the packet-tailgating technique gives us $A[0,N-1] \approx A$ for $l = N-1$ if the destination system I/O bandwidth is large enough to permit

$$\frac{A[N,N+1]}{A[0,N-1]} > \frac{p}{p+P}. \tag{5}$$

In practice we determine N by finding the smallest TTL that allows the packets to reach the destination.

3.2. Limited source system I/O bandwidth

In case the source system I/O bandwidth is less than A we propose using multiple sources to generate probing trains with bit rates exceeding A (see Figure 2). One source acts as the *master* instructing the other sources when to start transmitting their respective packet trains. Denoting the system I/O bandwidths of the sources as B_i, $i = 1,\ldots,m$, we can in theory obtain a maximum net probing rate of $R = \sum_i B_i$.

Source synchronization and determination of B_i, $i = 1,\ldots,m$ are crucial for the success of this method. Synchronization should not be difficult if sources are on the same LAN and have sub-millisecond RTTs. Transmitting a burst of back-to-back packets between sources j and k will give lower bound estimates of B_j and B_k. Thus by keeping j fixed and varying k one can obtain a tight lower bound on B_j.

3.3. NIC interrupt coalescence

NIC cards typically coalesce packet arrival interrupts to increase system efficiency. For example, a NIC card could delay all packets arriving in a 100 microsecond interval and forward them back-to-back to the kernel. This adds delays to probe packets which can lead to erroneous available bandwidth estimates.

One solution to the problem of interrupt coalscence is to use long probing trains with duration several times that of the interrupt coalescence interval (see Figure 3). While the fine grained timing delay of probes will be corrupted by coalescence noise, the global trends in delays will indicate whether network congestion has occurred or not. For example if the NIC coalesces packets in 100 microsecond intervals then an increase in average delay of probes over consecutive 500 microsecond intervals will still give a good indication of congestion. One can further refine the estimates by considering only the last of the back-to-back packets arriving at the kernel for each coalescence interval since their observed delay will be more accurate than that of other packets. In practice the coalescence intervals must be determined from the packet time stamps at the destination.

4. TIGHT LINK LOCALIZATION

Determining the location of the tight link is straightforward using packet-tailgating probing trains. Intuitively the available bandwidth $A[0,l]$ will remain constant for $l \geq t$ where t is the tight link. Mathematically we estimate t as

$$\widehat{t} = \min\{i : A[0,j] \approx A[0,N]\ \forall j \geq i\}. \tag{6}$$

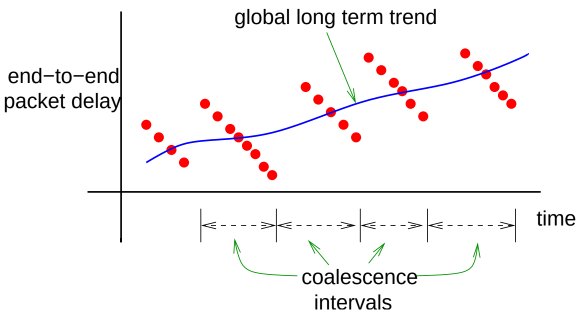

Fig. 3. *Available bandwidth estimation via global delay trends is robust to interrupt coalescence.*

In case the router at the tight link does not decrement the TTL, this method will only provide an approximate location of the tight link.

5. CURRENT WORK

The packet-tailgating technique is currently being tested out using packet *chirps* as well as packet *trains*. Performance results will be available in November 2003.

6. REFERENCES

[1] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," *Proc. ACM SIGCOMM*, 2002.

[2] B. Melander, M. Björkman, and P. Gunningberg, "A new end-to-end probing and analysis method for estimating bandwidth bottlenecks," *Global Internet Symposium*, 2000.

[3] N. Hu and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," *IEEE JSAC Special Issue in Internet and WWW Measurement, Mapping, and Modeling*, vol. 21, no. 6, 2003.

[4] G. Jin and B. Tierney, "Netest: A tool to measure the maximum burst size, available bandwidth, and achievable throughput," *ITRE*, August 2003.

[5] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathchirp: Efficient available bandwidth estimation for network paths," *Passive and Active Measurement Workshop*, 2003. San Diego, CA.

[6] G. Jin and B. Tierney, "System capability effects on algorithms for network bandwidth estimation," *Internet Measurement Conference*, 2003.

[7] A. Akella, S. Seshan, and A. Shaikh, "An empirical evaluation of wide-area Internet bottlenecks," *Internet Measurement Conference*, 2003.

[8] D. Rubenstein, J. Kurose, and D. Towsley, "Detecting shared congestion of flows via end-to-end measurement," *Proc. ACM SIGMETRICS*, 2000.

[9] K. Lai and M. Baker, "Measuring link bandwidth using a deterministic model for packet delay," *In Proc. ACM SIGCOMM*, 2000.

[10] A. Pasztor and D. Veitch, "Active probing using packet quartets," *Internet Measurement Workshop*, 2002.

[11] K. Harfoush, A. Bestavros, and J. W. Byers, "Measuring bottleneck bandwidth of targeted path segments," *IEEE INFOCOM*, 2003.

[12] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang, "Maximum likelihood network topology identification from edge-based unicast measurements," *ACM SIGMETRICS*, 2002.