A Packet Burst-based Inline Network Measurement Mechanism

Cao Le Thanh Man, Go Hasegawa and Masayuki Murata

Graduate School of Information Science and Technology, Osaka University
1-3 Yamadagaoka, Suita, Osaka 560-0871, Japan
E-mail: {mlt-cao, hasegawa, murata}@ist.osaka-u.ac.jp

Foreword

We introduce a new measurement method that can cope with interrupt coalescence techniques common in high-speed network adapters and reduce the impact of CPU overhead and other difficulties with system clocks. The proposed method adjusts the number of packets that are transmitted in a burst of an active TCP connection and estimates the available bandwidth by observing the inter-intervals of the bursts. The measurement results show that the method perform well in 1-Gbps networks.

Extended Abstract

We focus on a new challenge regarding active bandwidth measurement from end hosts. Measuring network from end hosts has the advantage of not requiring the cooperation of the routers along the path. We investigate the bandwidth measurement of 1-Gbps or faster network paths, which are becoming increasingly popular. In such high-speed networks, active measurement tools based on packet spacing must overcome the following problems.

- First, measurement in fast networks requires short transmission intervals of the probe packets. However, regulating such short intervals causes a heavy load on the CPU.
- Second, network cards for high-speed networks usually employ Interrupt Coalescence (IC) [1], which rearranges the arrival intervals of packets and causing bursty transmission, so that the algorithms utilizing the packet arrival intervals do not work properly.

In the present study, we introduce ICIM (Interrupt Coalescence -aware inline measurement), a new end-to-end bandwidth measurement approach that overcomes the above-mentioned two problems. Inline measurement is the idea of "plugging" the active measurement mechanism into an active TCP connection. This method has the advantage of requiring no extra traffic to be sent on the network, and provides fast and accurate measurement [2]. ICIM is be employed to the sender TCP.

IC has been shown to be detrimental to TCP self-clocking. The absolute timer, the default setting for IC of Intel Gigabit Ethernet Controllers [1], works as follows. It starts to count down upon receipt of the first packet. Subsequent packets do not alter the countdown. Once the timer reaches zero, the controllers generate an interrupt to pass all of the packets to the OS in a bursty manner. The length of the timer is decided by the parameter *RxAbsIntDelay*. Thus, IC causes the ACK packets to arrive at the sender in bursts, and this bursty arrival in turn causes bursty transmission of data packets and, subsequently, bursty transmission of ACK packets from the TCP receiver. According to one study [3], with IC, 65% of ACKs arrive at the sender TCP with intervals of less than 1 μs, because they are delivered to the kernel with a single interrupt. Meanwhile, without IC, almost no ACK packets arrive with small intervals.

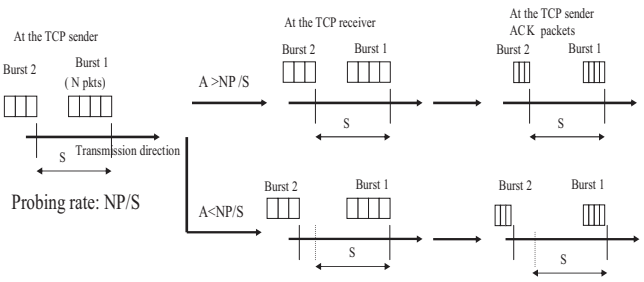

Fig. 1 Packet burst-based available-bandwidth measurement principle

The main idea of ICIM is to exploit the burst of data packets in TCP under the effects of IC to measure the available bandwidth. The TCP sender adjusts the number of packets involved in a burst and checks the intervals of the corresponding ACK packet bursts to investigate the available bandwidth.

The measurement principle is shown in Figure 1. Suppose that two bursts of packets are sent at the interval S. The number of packets in Burst 1 is N. C is the capacity of the bottleneck link. C_{Cross} is the average transmission rate of the cross traffic over the bottleneck link, and P is the packet size. Then, the amount of traffic that enters the bottleneck link during the period from the point at which the first packet of Burst 1 reaches the link until the point at which the first packet of Burst 2 reaches the link will be: $C_{Cross} \cdot S + N \cdot P$. If the amount is larger than the transfer ability of the link during this period, considered to be $C \cdot S$, then Burst 2 will go to the buffer of the link. This results

in a tendency for the interval between the two bursts to increase after leaving the bottleneck link. We can write that the burst interval will be increased if

$$C_{Cross} \cdot S + N \cdot P > C \cdot S \quad (1)$$

or,

$$\frac{N \cdot P}{S} > C - C_{Cross}$$

Note that $C - C_{Cross}$ is the available bandwidth (A) of the bottleneck link. Therefore, Eq. (1) becomes

$$\frac{N \cdot P}{S} > A$$

Thus, by sending numerous bursts with various values of NP/S (by changing N), we can search for the value of the available bandwidth A.

ICIM utilizes *search ranges* to perform faster and more accurate measurement. This is the idea of limiting the bandwidth measurement range using statistical information from previous measurement results rather than searching from 0 bps to the upper limit of the physical bandwidth for every measurement. The measurement algorithm of ICIM is as follows:

1. Set the initial search range to $(T, 2 \cdot T)$ where T is the throughput of TCP.
2. Search for the available bandwidth in the decided search range. ICIM then set the probing rate of k packet burst pairs to k points B_i,

$$B_i = B_l + \frac{B_u - B_l}{k-1}(i-1) \ \ (i = 1, ..., k)$$

   where (B_l, B_u) is the search range. If from burst number j, $j = 1..k$, the arrival interval of the bursts becomes larger, then B_j is considered to be the value of the available bandwidth in that measurement. k is set to 4 in the following simulation.
3. Calculate the next search range. The new measurement result is first added to the database. Then we use the 95% confidential interval of the data stored in the database as the width of the next search range, and the new measurement result is used as the center of the search range.
4. Wait for Q then return to Step 2 and start the next measurement. We set $Q = 2RTTs$ in the following simulation. During the waiting time Q, TCP transmits packets in the normal manner.

We examine the measurement results for ICIM through ns-2 simulations. We implement ICIM via Reno TCP, and use the topology shown in Figure 2 for the simulation. The sender and receiver of TCP are connected through 1-Gbps access links and a bottleneck link. The NICs of both the sender and receiver

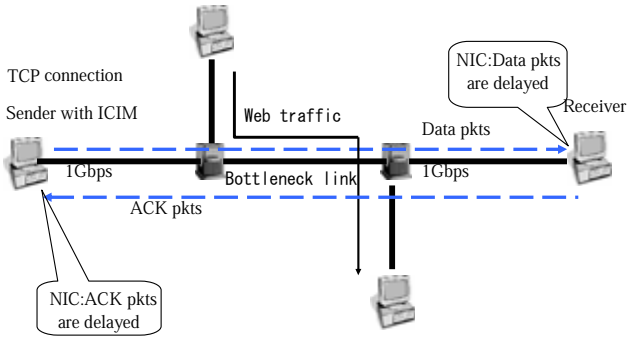

Fig. 2 Simulation topology

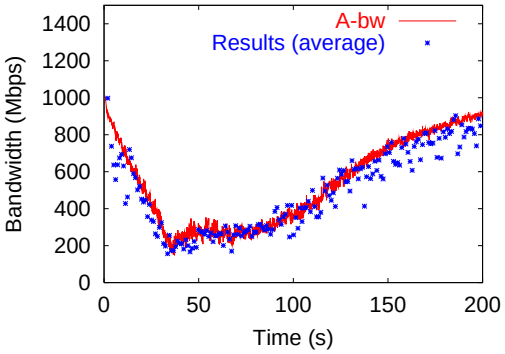

Fig. 3 Measurement results for ICIM

host employ IC with an absolute timer. The cross traffic on the bottleneck link is made up of Web traffic involving a large number of active Web document accesses. We use a Pareto distribution for the Web object size distribution. We use 1.2 as the Pareto shape parameter with 12 KBytes as the average object size. The number of objects in a Web page is 20. The capacity of the bottleneck link is set to 1-Gbps.

Figure 3 shows the average measurement results for ICIM for each second. The available bandwidth is calculated as the capacity of the bottleneck link minutes the total amount of Web traffic passing the link, and shown by the curved line "A-bw". We can see that ICIM can quickly detect the A-bw, even in such a high-speed network.

At present, we are evaluating the performance of ICIM in a real network environment and investigating the measurement mechanism for the capacity of high-speed networks.

References

[1] Intel, "Interrupt Moderation Using Intel Gigabit Ethernet Controllers," available at `http://www.intel.com/design/network/applnots/ap450.pdf(2003)`.

[2] Cao Le Thanh Man, Go Hasegawa and Masayuki Murata, "Available bandwidth measurement via TCP connection," in *Proceedings of the 2nd E2EMON Workshop 2004*, Oct. 2004.

[3] R. Prasad, M. Jain and C. Dovrolis, "Effects of interrupt coalescence on network measurements," in *Proceedings of the 5th PAM Workshop 2004*, Apr. 2004.