Inferring available bandwidth of overlay network paths based on inline network measurement

Cao Le Thanh Man
Graduate School of Information Science and Technology, Osaka University [0]
1-3 Yamadagaoka, Suita, Osaka 560-0871, Japan
mlt-cao@ist.osaka-u.ac.jp

Go Hasegawa and Masayuki Murata
Graduate School of Information Science and Technology, Osaka University
1-3 Yamadagaoka, Suita, Osaka 560-0871, Japan
{hasegawa, murata}@ist.osaka-u.ac.jp

Abstract

We introduce ImSystem, a distributed system that infers real-time information concerning the available bandwidth of all of the paths in the overlay networks, using a technique called inline network measurement. The key concept in ImSystem is that, when the overlay hosts transmit overlay traffic, they deploy the traffic to perform inline network measurements and exchange results with each other. ImSystem performs supplemental active measurements only when overlay traffic is insufficient for inline measurements, and therefore injects very little probe traffic onto the network. In addition, we enhance the system to ImSystemPlus, in which conflicts of the supplemental active measurements are greatly reduced, providing that the IP network topology is known. The simulation results show that the proposed systems have the same accuracy as that when all paths are measured by active measurements, while using only a small amount of probe traffic.

1 Introduction

Overlay networks have been proposed as a way to improve Internet routing, due to quickly detecting and recovering from path outages and periods of degraded performance. Overlay networks are deployed on end-hosts running the overlay protocol software without the cooperation of the core of the network. The end-hosts (overlay nodes) are in charge of routing the overlay traffic. That is, they control the sequence of the overlay nodes that the traffic traverses before reaching its destination. Thus, the network end-hosts should collect network resource information in order to form an overall view of the entire network so as to optimize the path selection. Some metrics of IP network resources are propagation delay, packet loss ratio, capacity, and available bandwidth. When the overlay network obtains sufficient information, the path selection is good, and, in time, the performance of the overlay network can be greatly improved.

We focus on the task of monitoring an important metric of IP network resources: the end-to-end available bandwidth. For routing in the overlay network, the fluctuation of bandwidth should be reported in small time scales. Therefore, the measurement tasks should be performed periodically in short intervals. However, measuring the available bandwidth of N^2 paths of a network, where N is the number of network nodes, requires a great deal of probe traffic. A number of studies [1-3] have focused on reducing the overhead. The methods proposed in these studies utilize the fact that the network paths are overlapping, with the assumption that the topology of the IP network is known. These methods carry out direct measurements on some overlay paths and indirectly estimate the bandwidth on the remained paths, deploying the measurement results of other network paths. However, the advantage of topology information appears to be limited because the amount of required probe traffic is still large, for example, on the order of $Nlog(N)$ [1, 2] or N [3].

[0] The author is at Systems Development Laboratory, Hitachi Ltd. 292 Yoshida, Totsuka, Yokohama 244-0817, Japan since April 2007. Email: lethanhman.cao.eq@hitachi.com

In a previous study [4] we have introduced a new version of TCP, called Inline measurement TCP (ImTCP). Like previous TCP versions, ImTCP can transmit data. However, ImTCP can also measure the available bandwidth of the path followed by TCP packets. When a sender transmits data packets, ImTCP first stores a group of up to several packets in a queue and then subsequently forwards them at a transmission rate determined by the measurement algorithm. Each group of packets corresponds to a probe stream. Then, considering ACK packets as echoed packets, the ImTCP sender estimates available bandwidth according to the algorithm. The simulation results in [4] show that ImTCP can yield measurement results with relative errors smaller than 20% every few RTTs without degrading transmission throughput.

In the present paper, we propose ImSystem, which infers the available bandwidth of all of the overlay network paths in real time. ImSystem utilizes the overlay traffic flows for measurement of the available bandwidth, using inline network measurement. If the transmission of overlay traffic occurs frequently, ImSystem works in a completely silent fashion, sending no probe traffic to the network. The ImSystem injects a small amount of probe traffic to the network only when the overlay traffic is insufficient for obtaining up-to-date information by inline measurement. We also enhance ImSystem to ImSystemPlus. Under the assumption that the topology of the IP network is known, ImSystemPlus predicts the conflicts of the active measurements on the overlapping paths and delays some measurements in order to reduce the number of conflicts.

The simulation results show that the proposed systems can provide up-to-date bandwidth information of overlay network paths while performing few additional active measurements. The proposed systems send almost no probe traffic when the amount of overlay traffic is sufficiently large.

2 ImSystem

ImSystem is formed by software programs (called ImSystem programs) that are installed in overlay nodes. ImSystem is located between the overlay network and the IP network.

2.1 Performing inline network measurement

We assume that ImTCP is deployed in all overlay hosts so that inline network measurement can be performed in every TCP connection used in the transmission of overlay traffic, and ImTCP senders pass all inline measurement results to the ImSystem program. Each ImSystem program sends messages to exchanges the measurement results with the ImSystem programs in other overlay hosts. The message includes a 4-byte field showing the IDs of the paths. Another 4-byte field shows the measurement result of the available bandwidth on that path. By exchanging the 8-byte messages, every ImSystem program can collect the information of all paths in the overlay networks.

Inline measurement yields measurement results in small intervals such as a number of RTTs. Therefore, if the ImSystem programs exchange every results, the number of messages will be extremely large. In order to decrease the number of exchange messages, ImSystem programs send the messages to report the measurement results only when they detect a change in the results. However, the measurement results always fluctuate due to both the measurement errors and actual changes in the available bandwidth. The problem is how to determine which changes in the measurement results were caused by real available bandwidth changes. Here, we introduce Equation (1), as proposed in [5], for abrupt change detection.

$$g_k = (1-\alpha)g_{k-1} + \alpha(y_k - \mu)^2, g_0 = 0. \qquad (1)$$

In Equation (1), y_k is the current inline measurement result, μ is the mean of the K latest results, where K is the number of inline measurement results yielded since the last message was sent. The maximum value of K is set to 15 in the following simulation experiments. In addition, g_k is an indicator of an abrupt change at the current sample, and α is the forgetting parameter, taking a value between 0 and 1. We set α to 0.5 and use a simple threshold rule as follows. If g_k is larger than the threshold (h), then we conclude that an actual change has occurred, otherwise the assumption is that no change occurred. Here, h is set to 120. This value is sufficient to rule out all significant changes in approximately 100-Mbps network paths.

2.2 Supplemental active measurement

In the case ImTCP is not deployed, ImSystem performs active measurements on the paths in every T (s), where T is the maximum length of the time that an active measurement may take. Even when ImTCP deployed, there are cases in which there is no overlay traffic on a certain path for a long time. During this period, ImSystem cannot perform inline measurements and the information concerning the available bandwidth of the path cannot be updated. In such cases, ImSystem waits a short time for new overlay traffic to arrive. The

waiting time depends on how long the current measurement results can maintain their accuracy when the network environment changes with time. We refer to the time as the validity term of the current result. If there is no new overlay traffic during the validity term, ImSystem performs supplemental active measurements on that path in order to update its available bandwidth information.

We now consider the length of the validity term of the current inline measurement result. The validity term corresponds to how long the current measurement result can maintain its accuracy in the future environment. We consider the measurement results delivered in the past as a time series and predict the trend of the changes in the correct value of available bandwidth in the future. By doing this, we can calculate the period in which the current result remains valid.

Here, we apply a model introduced in [6]:

$$X_t = m_t + s_t + Y_t,$$

where X_t is the time series of measurement results. In addition, m_t is the part that shows the trend of the time series and is set to be a linear because the measurement intervals are short. In the case of inline measurements, the intervals are a number of RTTs and the term s_t shows the periodical changes. In a short period, s_t can be considered as a linear change. In addition, Y_t is an independent and identically distributed random variable. Y_t shows the random noise of the measurements. We assume that Y_t has a normal distribution $N(0, \sigma^2)$.

We rewrite X_t as follows:

$$X_t = a_0 + b_0 t + Y_t,$$

where a_0 and b_0 are fixed values that can be calculated using the integrated moving average method. Variance σ is also calculated from the disparity in the trend and the measurement results.

In Figure 1, we assume that at the current time t_0, ImSystem sends messages to report the measurement result of A_0. Based on the measurement results just before t_0, we determine the trend of the changes in the available bandwidth of the path, as shown by the line in the figure.

We next consider the timing t_1 in the future. We examine the probability that the real available bandwidth remains at approximately A_0. This is the probability that the real available bandwidth appears in $[A_0 - \alpha, A_0 + \alpha]$, where α is $0.2A_0$, since study in [4] shows that the relative errors of ImTCP measurement results are within 20%. At this timing, the expected value of the measurement result, A_1, is:

$$A_1 = a_0 \cdot t_1 + b_0.$$

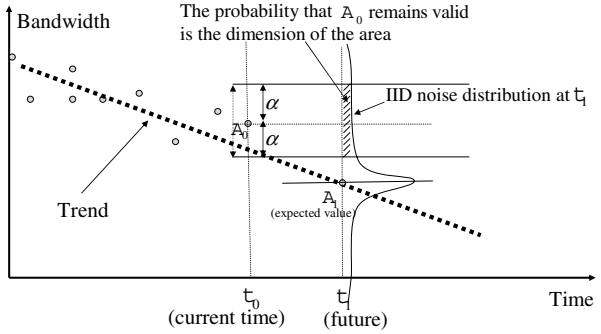

Figure 1. The accuracy of the previous result in the future environment

We assume that the measurement results at the time t_1 has the distribution $N(A_1, \sigma^2)$. Thus, the probability that the measurement result falls in $[A_0 - \alpha, A_0 + \alpha]$ is

$$q_{t_1} = \int_{A_0-\alpha}^{A_0+\alpha} \frac{1}{\sqrt{2\pi}\sigma} exp - \frac{(x - (a_0 \cdot t_1 + b_0))^2}{2\sigma^2} dx.$$

We assume that the measurement result A_0 becomes invalid at the time t_1 if the probability q_{t_1} falls below 1%. The validity term is then calculated as $t_0 - t_1$ where t_1 is the smallest solution of the following inequality:

$$\int_{A_0-\alpha}^{A_0+\alpha} \frac{1}{\sqrt{2\pi}\sigma} exp - \frac{(x - (a_0 \cdot t_1 + b_0))^2}{2\sigma^2} dx \leq 0.01.$$

Thus, the validity term is long if the available bandwidth does not change significantly. That is, a_0 is approximately zero. Then, ImSystem can save active measurements. On the other hand, if the available bandwidth changes dramatically, ImSystem will perform active measurements just after inline measurement to quickly update the bandwidth information.

3 ImSystemPlus

In ImSystem and other previously proposed systems [1-3], there are the cases in which two or more overlapping paths are probed by active measurements at the same time. The simultaneous measurements on the overlapping parts may cause degradation in measurement performance and localized congestion in the networks. In this section, we improve ImSystem to ImSystemPlus, which can solve the problem, providing that the IP network topology is known.

Unlike ImSystem, the ImSystemPlus program does not start active measurements just after the validity term of the current measurement result of the path

expires. Instead, the ImSystemPlus program considers the timing of active measurements on other paths, whether or not the they conflict with its measurement. In case there is high probability of conflicts, the ImSystemPlus program delays its measurement for a certain time.

3.1 Probability that a scheduled active measurement will be performed

We model the arrivals of data transmission on each overlay path as a Poisson process. The intervals between two arrivals then have an exponential distribution, $E_x(\lambda)$. The intervals between two arrivals on Path x (x is a, b, c ...) has the distribution of $E_x(\lambda_x)$, where λ_x is calculated based on the transmission history of Path x.

Assume that the last measurement result of Path x is expired at t_x. An active measurement is scheduled to be performed at that time. However, during the period from the current time t_0 to t_x, a data transmission may arrive. In this case, the active measurement scheduled at t_x will not be performed. Due to the loss of the memory property of an exponential distribution, the probability that there is no data transmission during the period from the current time t_0 to t_x is: $P_x = e^{-\lambda_x \cdot (t_x - t_0)}$. This is also the probability that active measurement is performed at t_x.

3.2 Probability for moving measurement time

When the new measurement time t_y of the measurement result on Path y is decided, we examine other measurement times that are approximately t_y in order to determine if there is any probable conflict measurements. We calculate the sum (Q) of the probability of the probable conflict measurements at approximately time t_y:

$$Q(t_y) = \begin{cases} S(t_y) & S(t_y) < 1 \\ 1 & S(t_y) \geq 1 \end{cases}$$

where

$$S(t_y) = \sum_{x; t_y - T < t_x < t_y + T} P_x \cdot joint(x, y). \qquad (2)$$

Here, $joint(x, y)$ is an index taking a value between 0 and 1, which indicates the degree to which the two paths x and y overlap each other [7].

The probability that we do not move the measurement time t_y to the right side is:

$$H^0 = 1 - Q(t_y)$$

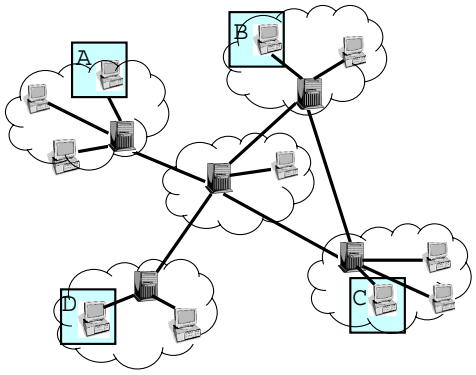

Figure 2. Network topology for examine the work of ImSystem

Similarly, the probability that we set the measurement time of Path y to $t_y + k \cdot T$ is:

$$H^{k \cdot T} = \prod_{h=0..k-1} Q(t_y + h \cdot T) \cdot (1 - Q(t_y + k \cdot T))$$

Here, $k = 1, 2...$ Note that when k is sufficiently large, the part P_x of $S(t_y + k \cdot T)$ calculated in Equation (2) approaches 0 (because when t_x is sufficiently large, the probability that there is no data transmission in the period $[t_0, t_x]$ approaches 0). Then, $Q(t_y + kT) = 0$ and H^{hT} with $h > k$ will be 0. This means that the measurement time cannot be delayed for a long time.

4 Simulation experiments

4.1 Bandwidth information of ImSystem

We first examine the work of ImSystem in a simple topology shown in Figure 2. There is a four-node overlay network built upon an IP network. The capacity of the links in IP network is 100 Mbps. In addition to overlay flows, non-overlay traffic also exists on the IP link, referred to as cross-traffic. The rate of cross traffic at one link is uniformly distributed in $[M - 0.05M, M + 0.05M]$, where M is the average rate, independent of the rate changes at other links. M changes as follows. After every second, M is increased by b Mbps. When M reaches 60 Mbps, it is decreased by b Mbps every second, until reaching 0 Mbps. M is then increased by b Mbps every second, and so on. b is randomly determined in the range [1, 50] Mbps. For the links on the path between B and C, the average rate of cross traffic M is kept constant at 50 Mbps. Overlay flows at the overlay paths are generated according to a Poisson process with an average arrival rate of F. All

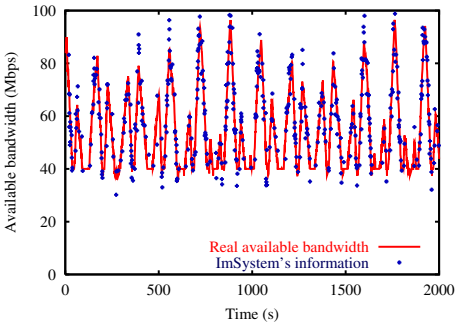

Figure 3. Information about path D-B

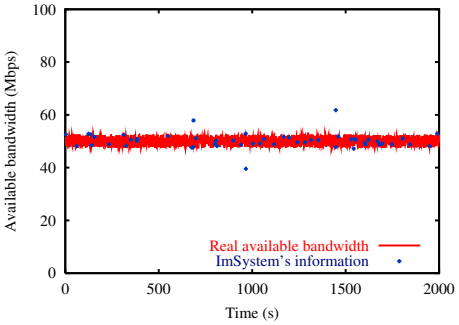

Figure 4. Information about path B-C

overlay paths have the same value of F. Overlay flow duration has exponential distribution with an average of 20 s. Overlay flow rate is uniformly distributed in the range [100 Kbps, 1 Mbps].

The active measurement is assumed to be Pathload. The time required by one active measurement is 10 s. Active measurement results are uniformly distributed in $[A - 0.1A, A + 0.1A]$, where A is the real available bandwidth value. The active measurement rate is 250 Kbps. The time required for one inline measurement is set to 1 s. In fact, ImTCP can yield results in smaller intervals. We assume that ImSystem takes an average of the measurement results every second. Inline measurement results are uniformly distributed in $[A + 0.2A, A - 0.2A]$, where A is the real available bandwidth value.

Figure 3 shows the changes of the real available bandwidth in the overlay path from host D to host B. In this case F is set to 0.2. The figure also shows how the ImSystem program on the third host (host A) observes the bandwidth on this path. In this case, since the bandwidth changes dramatically over time, ImSystem updates the information frequently. Similarly, Figure 4 shows how the ImSystem program on host A observers the available bandwidth of the path from B to C. The real value of the available bandwidth

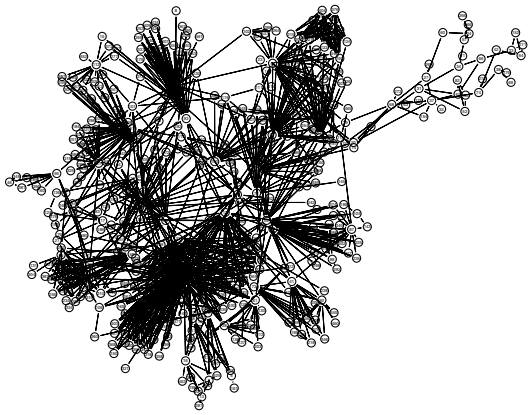

Figure 5. Sprint network topology

is also shown. Since the non-overlay traffic on the path is set at a constant 50 Mbps, the available bandwidth of the path does not fluctuate significantly. Therefore, in this case, we can see that ImSystem does not update the value very often.

4.2 Accuracy of bandwidth information and the amount of probe traffic

We next evaluate ImSystem and ImSystemPlus in a larger network topology. We use the topology of Sprint backbone network shown in Figure 5, which is inferred by Rocketfuel. The topology includes 467 nodes and 1280 links. The suppositions on cross traffic, overlay traffic as well as measurement tools are the same as the simulation mentioned in the previous subsection. However, in this simulation, the overlay network has 10 nodes, which are randomly distributed in the IP network.

We perform 10 simulations with different distributions of overlay nodes. Figure 6 shows the average relative errors between the collected bandwidth information the real bandwidth values. The maximum and minimum values in 10 simulations are also shown. Figure 7 shows the average as well as the maximum, minimum value of the amount of probe traffic for active measurements performed by these systems. The horizontal axes of these figures show the average arrival rates of the overlay traffic at the overlay nodes (F). For comparison, we also perform the simulations where the active measurement results are periodically deployed in all overlay paths at fixed T and $2T$ intervals, where T is the maximum time for an active measurement to be performed. T is set to 15 (s). For avoiding conflicts in the measurements, the nodes start their measurements in random times.

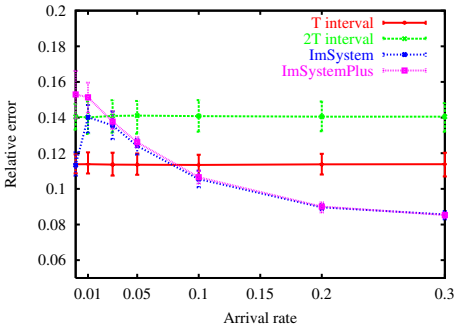

Figure 6. Relative errors of collected bandwidth information.

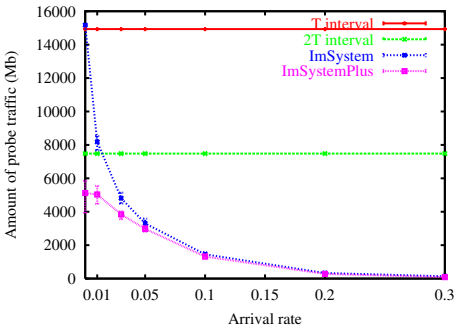

Figure 7. Probe traffic

Figure 6 shows that in case there is no overlay traffic, ImSystem has the same error as when active measurements are performed in every T (s). ImSystemPlus avoids the conflict of measurements so it sends to the network less probe traffic, and therefore the error of the bandwidth information is a little larger than that of ImSystem. The two proposed systems show their advantages when the arrival rate of overlay flow becomes higher than 0.1. They both introduce error smaller than when the paths are actively measured in T intervals, while their probe traffic is only 1/8 or smaller.

4.3 Effectiveness of ImSystemPlus

We finally examine the probe traffic that conflicts with other traffic in the present simulations and calculate the amount of probe traffic that shares one or more links with one or more other probe traffic. The results are shown in Figure 8. ImSystemPlus can decrease at most 80% of the conflict probe traffic that exists in ImSystem. The proportion is highest when where there is no overlay traffic. This is due to the function of detecting and avoiding conflicts in measurement of ImSystemPlus.

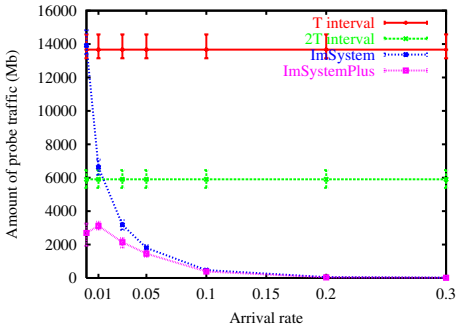

Figure 8. Conflict probe traffic

5 Conclusions

In the present paper, we proposed ImSystem and ImSystemPlus, inline network measurements based systems that collect available bandwidth information of all end-to-end paths in an overlay network.

In future works, we will examine the impact of ImSystem and ImSystemPlus on overlay network's performance. We will also implement the proposed systems and evaluate their performance in real network environments.

References

[1] C. Tang and P. McKinley, "On the cost-quality tradeoff in topology-aware overlay path probing," in *Proceedings of the 11th ICNP*, Nov. 2003.

[2] Y. Chen, D. Bindel, H. Song, and R. Katz, "An algebraic approach to practical and scalable overlay network monitoring," in *Proceedings of ACM SIGCOMM 2004*, Aug. 2004.

[3] N. Hu and P. Steenkiste, "Exploiting internet route sharing for large scale available bandwidth estimation," in *Proceedings of IMC'05*, Oct. 2005.

[4] C. Man, G. Hasegawa, and M. Murata, "ImTCP: TCP with an inline measurement mechanism for available bandwidth," *Computer Communications*, vol. 29, no. 10, pp. 1614–2479, 2006.

[5] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., 1993.

[6] P. J. Borockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer-Verlag NewYork, Inc., 1996.

[7] M. Zhang and J. Lai, "A transport layer approach for improving end-to-end performance and robustness using redundant paths," in *Proceedings of the USENIX 2004 Annual Technical Conference*, June 2004.