Monitoring overlay path bandwidth using an inline measurement technique

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
{mlt-cao, 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 paths in the overlay networks. The key concept in ImSystem is that, when the overlay hosts transmit overlay traffic, the overlay hosts deploy the traffic to perform inline network measurements. Inline network measurement is a method for measuring available bandwidth using only the packets transmitted in a data flow, instead of injecting probe traffic onto the network. ImSystem performs supplemental active measurement only when overlay traffic is insufficient for inline measurement, and therefore injects very little probe traffic onto the network. We also enhance the system to ImSystemPlus and ImSystemLight, deploying IP network topology information. In ImSystemPlus, the conflicts of the supplemental active measurements are greatly reduced, and, in ImSystemLight, the number of exchange messages sent by the overlay nodes is reduced. The simulation results show that the proposed systems can monitor the bandwidth in real-time while using only a small amount of probe traffic if the overlay traffic is sufficient.

Keywords: End-to-end measurement, available bandwidth, inline measurement, overlay network, active measurement

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].

In a previous study [4] we have introduced a new version of TCP, called Inline measurement TCP (ImTCP). ImTCP can transmit data like previous TCP versions. 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 of the network path between the TCP sender and TCP receiver. We name the technique inline measurement. The simulation results in [4] shows that ImTCP can yield measurement results with relative errors smaller than 20% every few RTTs without degrading transmission throughput. Moreover, studies in [5, 6] validates the measurement results for ImTCP in the real Internet environments.

The present paper is an extended version of our previous work in [7]. In this 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 an inline measurement technique. When the transmission of overlay traffic occurs frequently, ImSystem works in a completely silent fashion, that is, ImSystem sends no probe traffic to the network. The system injects a small amount of probe traffic onto the network only when the overlay traffic is insufficient for obtaining up-to-date information by inline measurement.

We also propose enhanced versions of ImSystem, ImSystemPlus and ImSystemLight. Both two systems work 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. On the other hand, in ImSystemLight, the overlay nodes estimate the bandwidth of a path using information concerning its overlapping paths. They then reduce the number of messages used to exchange measurement results of the paths.

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.

The remainder of the present paper is organized as follows. In Section 2, we discuss some related research. Sections 3, 4 and 5 describe the respective designs of the three proposed systems: ImSystem, ImSystemPlus and ImSystemLight, respectively. In Session 6, we present a number of simulation studies to validate the proposed systems. Finally, Section 7 presents conclusions and a discussion of future research.

2 Related study

Resilient Overlay Networks (RON) proposed in [8] monitors the IP network by active measurements in fixed intervals. Every overlay host sends probe traffic for measurement of propagation delay and packet loss to all other hosts in the network. From the measurement results, the hosts estimate the throughput of data transmission on the overlay paths. The overhead for the information collection is $O(N^2)$, where N is the number of overlay nodes. Therefore, [8] also points out that RON can work only with 50 or fewer nodes. In an effort to reduce the load of probe traffic on the network, [1] introduces a system that infers the available bandwidth of N^2 paths, in which the measurement overhead is reduced to the order of $Nlog(N)$. In this case, the accuracy becomes 90% of that when the measurements are performed on the full mesh. The method requires topology information, which is inferred by network tools such as `traceroute`. In addition, [2] deploys algebraic functions to reduce the measurement overhead to $O(Nlog(N))$. However, this requires a master node for managing all of the data processing. BRoute [3] leverages the fact that most Internet bottlenecks are on path edges as well as the fact that edges are shared by several different paths. BRoute performs bandwidth measurements on a number of paths using a hop-by-hop active measurement tool called Pathneck and infers the bandwidth of the remaining paths using the AS-level topology. In BRoute, the measurement overhead is further reduced to the order of N. This method also requires a master node with which to collect and process data from all hosts. The systems proposed in the present paper do not require a master node with which to manage the entire system while monitoring available bandwidth with a much smaller number of active measurements, compared to the existing methods proposed in [1-3, 8].

3 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. ImSystem programs monitor the available bandwidth information of overlay paths and present this information to the overlay networks. ImSystem is independent of the overlay network; it can work with any overlay routing algorithms. Each ImSystem program collects the available bandwidth of the paths that start from the node where it locates and exchanges the measurement results with each other. The bandwidth information is yielded mainly by inline measurement, the measurement technique that does not inject probe traffic onto the network. Active measurements are also used in case inline measurement results are not available.

3.1 Inline network measurement

Inline network measurement is the concept of performing active measurement using packets transmitted in a data flow. Inline network measurement can achieve high accuracy while not sending probe traffic into the network. We have developed ImTCP (Inline measurement TCP) [4], a Reno-based TCP that performs inline network measurement. Figure 1 illustrates the working of ImTCP. When ImTCP sender transmits data packets, it periodically stores a group of up to several packets in a queue and subsequently forwards them at determined intervals. Each group of packets corresponds to a probe stream. Then, the ImTCP sender checks the arrival intervals of the corresponding ACK packets. Under the supposition that all packets in a group traverse the same network path, if the arrival intervals are larger than the sending intervals, the available bandwidth of the network path between ImTCP sender and receiver is smaller than the transmission rate of the probe stream. Otherwise, the available bandwidth is larger. Using this fact, the ImTCP sender determines the available bandwidth by changing the transmission rate of probe streams. The measurement algorithm is similar to that of Pathload [9] and PathChirp [10]. However, ImTCP does not search for available bandwidth from 0 bps to the upper limit of the physical bandwidth with every measurement as these algorithms do. Instead, ImTCP limits the bandwidth measurement range using statistical information from previous measurement results so that one measurement is performed in as short as some RTTs. By limiting the measurement range, ImTCP avoids sending probe packets at an extremely high rate and keep the number of probe packets small so that it doesnot affect other data flows. Our experiment results in [4] show that ImTCP can yield measurement results within 20.8% of the actual available bandwidth in intervals as short as some RTTs without degrading transmission throughput.

3.2 Filtering inline measurement results

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 exchange the measurement results with the ImSystem programs in other overlay hosts. The message includes the name of the beginning and end nodes of the path, the result of the measurement performed on that path and the validity term of the result. The validity term will be mentioned in the next Subsection. By exchanging the messages, every ImSystem program can obtain quickly the information of all paths in the overlay networks.

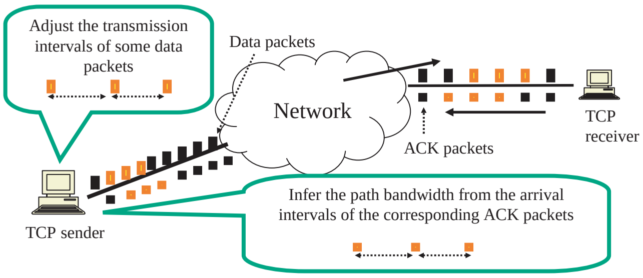

Adjust the transmission intervals of some data packets
Data packets
Network
TCP receiver
ACK packets
TCP sender
Infer the path bandwidth from the arrival intervals of the corresponding ACK packets

Figure 1. Key concept of inline network measurement

Inline measurement yields measurement results in small intervals such as a number of RTTs. Therefore, if the ImSystem programs exchange every result, 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 [11], 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.

3.3 Supplemental active measurement

In the case ImTCP is not available, 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 is deployed in the system, there are cases in which there is no overlay traffic on a certain path for a long time. During this period, ImTCP 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 an inline measurement result. The validity term corresponds to how long the 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 [12]:

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

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:

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

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.

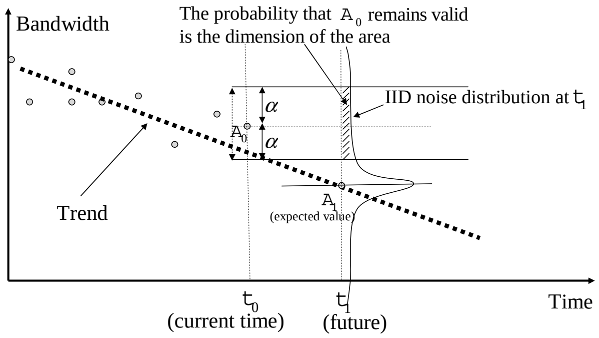

Bandwidth
The probability that A0 remains valid
is the dimension of the area
IID noise distribution at t1
α
α
A0
Trend
A1
(expected value)
t0
(current time)
t1
(future)
Time

Figure 2. The accuracy of the previous result in the future environment

In Figure 2, we assume that at the 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:

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

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.

4 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 common characteristic of the active measurement algorithms for available bandwidth is that, they require the probe traffic to fill up the unused bandwidth of the target path for some time. Therefore, in the case when the overlapping part of the paths includes a tight link (a link in which the unused bandwidth is smallest in the path), the probe packets of two different measurements may conflict to each others, causing degradation in measurement performance. In addition, the simultaneous transmission of probe traffic on the overlapping parts may cause localized congestion in the networks.

To avoid conflicts of measurements, ImSystemPlus program does not start active measurements right after the validity term of the current measurement result of the path expires. Instead, ImSystemPlus program considers whether or not the active measurements on other paths conflict with its measurement. In case there is high probability of conflict, the program delays its measurement for a certain time.

4.1 Conflict avoidance for measurements

In ImSystemPlus, the messages that the hosts exchange with each other have an additional field, showing the time when the validity term expires. If the term expires without any new overlay traffic transmission appearing on the correspondent path, ImSystemPlus program will perform active measurements to update the result. Therefore, ImSystemPlus program can know when the programs on other hosts schedule the performance of their active measurements. This time is referred to as the measurement time.

Figure 3 shows an example when the new measurement time of Path b is too close (within T (s)) to that of Paths a and c. We assume that Paths a and b, and b and c are overlapping, which means that the paths share one or more links. The active measurements on these paths will then come into conflict with each other.

To avoid probable conflicts on these paths, we propose a strategy that moves the measurement time of Path b to the right side, far away from that of Paths a and c. We consider the probability of moving the measurement time of Path b $k \cdot T$ seconds to the right side, where k = 0, 1, 2, ... To determine the probabilities, we consider the followings:

- Active measurements on Paths a and c may not be performed at the scheduled time due to the arrival of data transmission on these paths. If this probability is high, the probability of moving the measurement time of Path b should be low.
- If the overlapping parts of the Paths a and b, c and d are not so large, then the conflict of the measurement may not cause serious problems. In this case the probability of moving the measurement time of Path b should be low. Next, we explain how to calculate these probabilities.

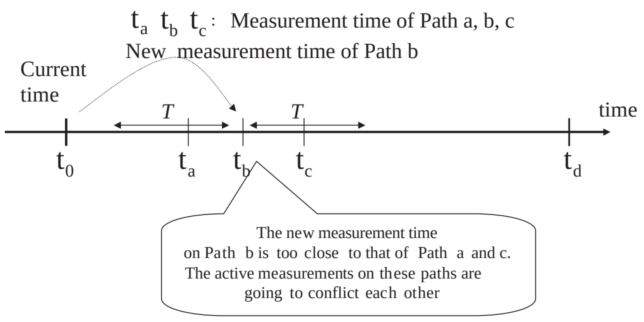

t_a t_b t_c : Measurement time of Path a, b, c
New measurement time of Path b
Current time
T
T
time
t_0
t_a
t_b
t_c
t_d
The new measurement time on Path b is too close to that of Path a and c. The active measurements on these paths are going to conflict each other

Figure 3. Conflict in active measurements on overlapping paths

4.2 Overlapping index

The degree to which two path overlaps each other is related to the effect that the simultaneous measurement on the two paths may have on the network. If the overlapping part is large, the conflict of the measurements will have a worse effect on the network and its performance. ImSystemPlus deploys the concept of path overlapping introduced in [13].

$$Joint(a, b) = \frac{Latency(G)}{min(Latency(a), Latency(b))}.$$

Here, G is the overlapping part of Paths a and b. $Latency()$ shows the transmission delay of the entire network path or part of the network path. $Joint()$ is an index taking a value between 0 and 1, which indicates the degree to which the two paths overlap each other.

4.3 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 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 expires at t_x. An active measurement is scheduled to be performed at this 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 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.

4.4 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

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

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

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

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.

5 ImSystemLight

In ImSystem, in order to keep the bandwidth information up-to-date in each overlay node, each overlay node sends messages to all of the other overlay nodes to report any new measurement results. In this section, we show how to decrease the number of messages while maintaining the highest possible degree of accuracy with respect to the bandwidth information.

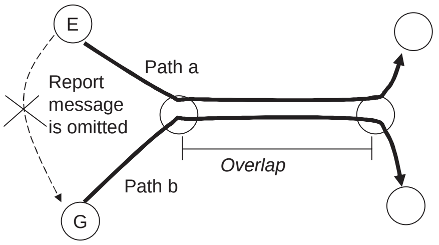


Figure 4. Reducing report messages in ImSystemLight

Like ImSystemPlus, ImSystemLight also deploys information about the topology of the underlying IP network to reduce the traffic caused by the communication between overlay nodes. Figure 4 is used to explain how ImSystemLight reduces the report messages. A node (node E) omits report messages relating to the inline measurement results on path a that should be sent to node G if there is more than one path starting from G that overlaps path a. On the other hand, node G, which always updates its database when a report concerning the bandwidth of path a arrives, will instead estimate the bandwidth of path a automatically, in case the report is omitted by node E and the previous information becomes invalid. Among the paths starting from node G, let path b be the path that has the longest overlap with path a. Under the assumption that path a and path b share the same tight link (the link with smallest available bandwidth), node G uses the available bandwidth of path b as the estimated bandwidth of path a.

The probability that the tight link of path a appears on the overlapping part of path a and path b, is:

$$g(a, b) = \frac{Latency(Overlap(a, b))}{Latency(a)}.$$

Here, $Overlap(a, b)$ is the overlapping part of path a and path b. In ImSystemLight, the probability that node E reports the measurement result of path a to node G $(p(E, a, G))$ is set as follows:

$$p(E, a, G) = 1 - g(a, b).$$

We use this setting because, if g is low, the probability of incorrect estimation is high. In this case, node E should report the measurement results to avoid the

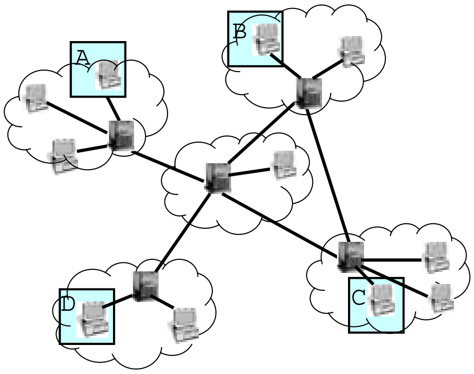


Figure 5. Network topology for examine the work of ImSystem

degradation in information accuracy. On the other hand, if g is high, the estimation is reliable. In this case, node E can omit the message without causing a significant degradation in bandwidth information accuracy. Note that the messages will not be omitted if they report measurement results for active measurements. These results always have high accuracy, so that they are important for maintaining the accuracy of the information collected by the systems.

6 Simulation experiments

In this chapter, we evaluate the performance of the proposed systems in some different network topologies.

6.1 Collecting bandwidth information in ImSystem

We first examine the work of ImSystem in a simple topology shown in Figure 5. 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 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].

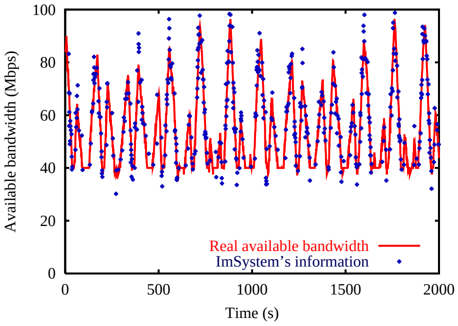


Figure 6. Information about path D-B

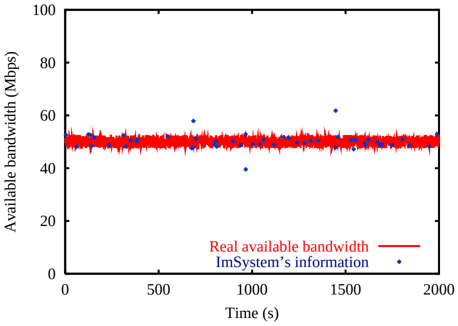


Figure 7. Information about path B-C

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. The relative error is calculated from the ImTCP simulation results in [4].

Figure 6 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. Simi-

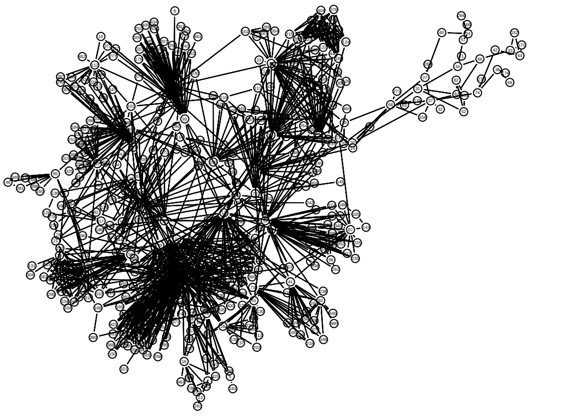


Figure 8. Sprint network topology

larly, Figure 7 shows how the ImSystem program on host A observes the available bandwidth of the path from B to C. The real value of the available bandwidth 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 updates the bandwidth information in larger time intervals in order to decrease the number of messages sent into the network.

6.2 Simulation setting

We next evaluate ImSystem, ImSystemPlus, and ImSystemLight in larger network topologies. We deploy the following three topologies for IP networks in the simulation experiments.

- The topology of the Sprint network. We use the topology of the Sprint backbone network which is inferred by Rocketfuel [14]. The topology includes 467 nodes and 1280 links. Figure 8 illustrates the topology.
- Random topology. The network begins with an initial topology of three nodes. We then add new nodes to the initial topology and create links from the new nodes to the existing nodes. The probability that a new node has a link to node i is

$$p_i = 0.01 + (1 - 0.01)^n \frac{1}{n}$$

  where n is the number of existing nodes. This probability ensures that the new node is a connected node, that is, the new node has a link to at least one of the other node. The final topology has the same node number as the Sprint network and has 1282 links.
- BA model [15] topology. The network begins with an initial topology of three nodes. We then add new nodes to the initial topology until the number of nodes becomes the same as that of the Sprint network (467 nodes). The probability p_i that the new node has a link to node i is

$$p_i = \frac{\sum_j k_i}{k_j}$$

  where k_i is the degree of node i. As the number of nodes reaches 467, the topology has 1388 links.

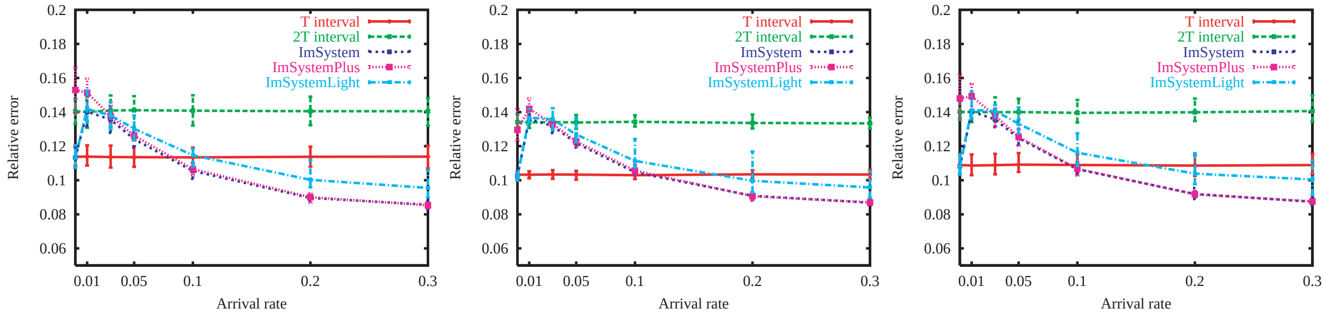
In the following simulations, the assumptions on cross traffic, overlay traffic and measurement tools are the same as those for the simulation mentioned in the previous subsection. The overlay network has 10 nodes, which are randomly distributed in the IP network. We perform 10 simulations with different distributions of overlay nodes. The time for each simulation is 2000 s.

For comparison, we also perform the simulations in which the active measurement results are periodically deployed in all overlay paths (full mesh) 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)). In order to reduce the number of conflicts in the measurements, the nodes begin measurements at random times.

6.3 Accuracy of bandwidth information and the amount of probe traffic

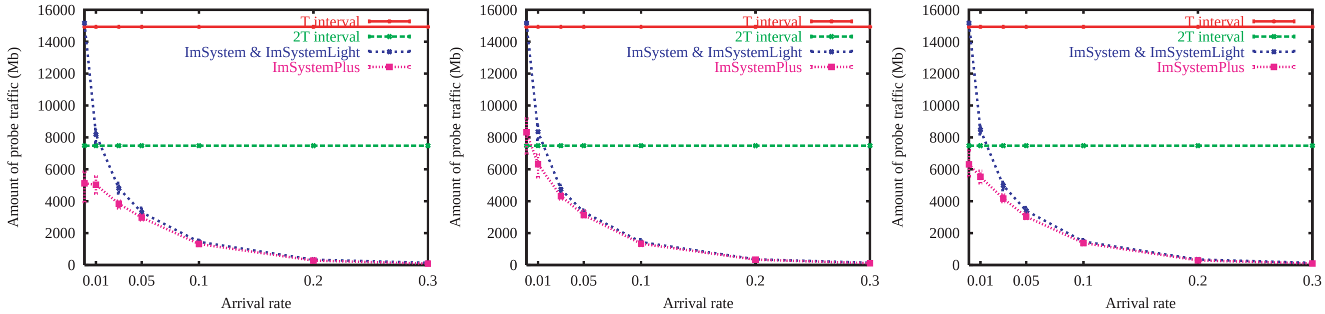
Figure 9 shows the average as well as the maximum, minimum value of the amount of probe traffic for active measurements performed by ImSystem, ImSystemPlus and ImSystemLight through 10 simulations in three different topologies. Also shown are the relative errors when the active measurement are performed in all overlay paths in every T intervals, in the curve "T interval", and $2T$ intervals, in the curve "$2T$ interval". The horizontal axes of these figures show values of F, the average arrival rates of the overlay traffic at the overlay nodes.

From Figure 9 we can see that, in all the topologies, the relative errors of the systems have the same trend. That is, in case there is no overlay traffic, ImSystem, ImSystemPlus as well as ImSystemLight has the same error as when active measurements are performed in every T (s). The three proposed systems show their advantages when the arrival rate of overlay flow becomes higher than 0.1; they introduce error smaller than when the paths are actively measured in T intervals. ImSystemPlus avoids the conflict of measurements so it sends to the network less probe traffic,



(a) Sprint network topology (b) Random topology (c) BA model topology

Figure 9. Relative errors of collected bandwidth information



(a) Sprint network topology (b) Random topology (c) BA model topology

Figure 10. Probe traffic

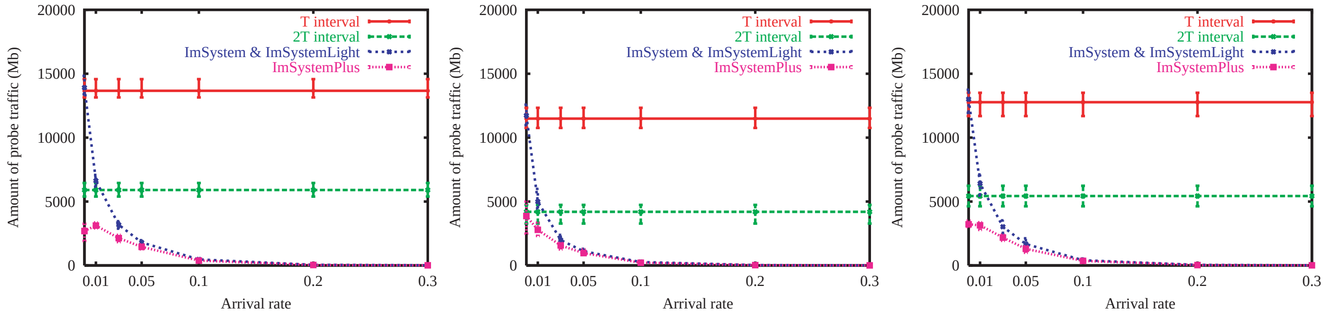
and therefore the error of the bandwidth information is a little larger than that of ImSystem. ImSystemLight also has a little larger error than the others because, as we will show later, it reduces the messages exchanged between overlay nodes.

Comparing Figure 9(b) with Figure 9(a) and Figure 9(c), we can see that in the Sprint network and BA model topologies, the error of bandwidth information collected by active measurement in T or $2T$ intervals is higher than that in a random topology. That is because in the Sprint network and BA model topologies, many overlay network paths are sharing the same IP links, this leads to the fluctuation of available bandwidth due to the overlay traffic.

Figures 10(a), 10(b) and 10(c) show the amount of active probe traffic sent during a simulation. We can see that proposed systems always use much smaller active measurement than when active measurements are performed in T or $2T$ intervals. When the arrival rate of overlay flow comes to 0.3, the proposed systems completely do not use the active measurements. ImSystemPlus uses less probe traffic than ImSystem and ImSystemLight in 2000 s of the simulation because it tends to delay the conflicting measurements.

6.4 Number of conflicting active measurements

We next 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 other probe traffic. The results are shown in Figure 11. In all three topologies, ImSystemPlus can eliminate most of the conflicting probe traffic that exists in ImSystem. The proportion is highest when there is no overlay traffic. This is due to the function of detecting and avoiding conflicts in the measurement of ImSystemPlus. The characteristics of the topologies have a slight effect on ImSystemPlus. In a random network, the overlapping index of the overlay paths is



(a) Sprint network topology

(b) Random topology

(c) BA model topology

Figure 11. Conflicting probe traffic

Table 1. Number of the exchange messages

Topology	F	ImSystem	ImSys.Light	Ratio
Sprint network	0	5435	5435	1.00
	0.01	4577	4206	0.92
	0.03	6184	5233	0.85
	0.05	7862	6462	0.82
	0.1	10982	8787	0.80
	0.2	13805	10963	0.79
	0.3	14740	11718	0.79
Random	0	5435	5435	1.00
	0.01	4557	4321	0.95
	0.03	6214	5535	0.89
	0.05	7928	6960	0.88
	0.1	10853	9382	0.86
	0.2	13504	11625	0.86
	0.3	14417	12409	0.86
BA model	0	5435	5435	1.00
	0.01	4733	4326	0.91
	0.03	6599	5501	0.83
	0.05	8400	6820	0.81
	0.1	11646	9182	0.79
	0.2	14463	11325	0.78
	0.3	15435	12106	0.78

small so that ImSystemPlus cannot reduce as many probe conflicts as in the Sprint network and BA model topologies.

6.5 Number of exchange messages

Table 1 shows the average number of messages that a node in ImSystem and ImSystemLight sends during the simulation. The last column shows the ratio between the ImSystem and ImSystemLight. ImSystem deploys inline measurement, the measurement accuracy of which is not as high as that of stand-alone measurement tools. Therefore, the nodes exchange many messages in order to increases the accuracy of the collected information. ImSystemLight, as expected, uses fewer messages in comparison with ImSystem while maintaining the highest possible accuracy. The topology also has little affect on the work of ImSystemLight. In the Sprint network and BA model topologies, the overlay paths overlap significantly so that ImSystemLight can reduce the number of exchange messages to a greater degree than in a random network.

7 Conclusion

In the present paper, we proposed ImSystem, which monitors the available bandwidth of all end-to-end paths in an overlay network in real time. The proposed system is based primarily on inline network measurement, that is, ImSystem deploys active overlay data flow for measurement. Therefore, the system injects little probe traffic onto the network while inferring the available bandwidth in a real-time fashion. We also proposed ImSystemPlus and ImSystemLight. In these systems, conflicts in measurement traffic and the data exchanged between overlay nodes are reduced.

In future works, we will examine how scalable are the proposed systems. We will also implement 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. L. T. 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] T. Tsugawa, G. Hasegawa, and M. Murata, "Background TCP data transfer with inline network measurement," *IEICE Transactions on Communications*, vol. E89-B, pp. 2152–2160, Aug. 2006.

[6] T. Tsugawa, C. L. T. Man, G. Hasegawa, and M. Murata, "Inline bandwidth measurements: Implementation difficulties and their solutions," in *Proceedings of E2EMON 2007*, May 2007.

[7] C. L. T. Man, G. Hasegawa, and M. Murata, "Inferring available bandwidth of overlay network paths based on inline network measurement," in *Proceedings of ICIMP 2007*, July 2007.

[8] D. Andersen, H. Balakrishnan, M. Kaashoek, and R. Morris, "Resilient overlay networks," in *Proceedings of SOSP 2001*, Oct. 2001.

[9] M. Jain and C. Dovrolis, "End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput," in *Proceedings of ACM SIGCOMM 2002*, Aug. 2002.

[10] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "PathChirp: Efficient available bandwidth estimation for network paths," in *Proceedings of PAM 2003*, Apr. 2003.

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

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

[13] 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.

[14] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with Rocketfuel," in *Proceedings of SIGCOMM 2002*, Aug. 2002.

[15] A. Barabasi and R. Albert, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.