社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

インラインネットワーク計測を用いたオーバーレイネットワークパスの帯域推測手法

Cao Le Thanh Man[†]　　長谷川　剛[††]　　村田　正幸[†]

[†] 大阪大学大学院情報科学研究科〒 560-0871 大阪府吹田市山田丘 1-5
[††] 大阪大学サイバーメディアセンター〒 560-0043 大阪府豊中市待兼山町 1-32
E-mail: †{mlt-cao,murata}@ist.osaka-u.ac.jp, ††hasegawa@cmc.osaka-u.ac.jp

あらまし　オーバーレイネットワークによって、エンドホストがトラヒックの経路選択を行うことができる状況が登場しつつある。最適経路の選択を可能にするためには、オーバーレイノードが経路の帯域情報を知る必要がある。本稿では、各オーバーレイノードで分散的に動作する帯域情報収集システム ImSystem を提案する。ImSystem はデータ転送のために用いられている TCP コネクションを用いた計測、いわゆるインラインネットワーク計測を利用することにより、帯域計測用のトラヒックを減少させる。そのため、帯域情報収集のためにネットワークへ与える負荷が小さいという特長を持つ。さらに、ImSystem をベースに、計測用のトラヒックが互いに重ならないように計測タイミングを調整する ImSystemPlus を提案する。シミュレーション評価結果により、提案システムが常に最新の帯域情報を高い精度で把握することが可能であり、かつ、ネットワークへ与える影響が少ないことを示す。また、オーバレイネットワーク上のトラヒック量が少ない場合においても、提案システムがそれを有効に利用し、計測用トラヒックを大幅に減らすことができることを明らかにする。

キーワード　エンドホスト間計測、利用可能帯域、インライン計測、オーバーレイネットワーク、アクティブ計測

Inferring bandwidth of overlay network paths using inline network measurement

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
[††] Cybermedia Center, Osaka University 1-32, Machikaneyama, Toyonaka, Osaka 560-0043, Japan
E-mail: †{mlt-cao,murata}@ist.osaka-u.ac.jp, ††hasegawa@cmc.osaka-u.ac.jp

Abstract　To optimize the overlay route selections, overlay nodes require end-to-end bandwidth information of the paths between the overlay nodes. In the present paper, we introduce ImSystem, a distributed system installed in the overlay end hosts. ImSystem 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.

Key words　End-to-end measurement, available bandwidth, inline measurement, overlay network, active measurement

1. Introduction

Overlay networks such as RON [1] 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 [2-4] 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 $Nlog(N)$ [2, 3] or N [4].

In a previous study [5] 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. To minimize the transmission delay caused by the packet store-and-forward process, we introduce an algorithm that is similar to the round trip timeout (RTO) calculation in TCP to regulate the packet storage time in the queue. The simulation results in [5] 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. During the time the active measurement in a path is delayed, the system estimates the available bandwidth of the path using the bandwidth information on other 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.

2. Related study

Resilient Overlay Networks (RON) proposed in [1] 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, [1] 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, [2] 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, [3] 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 [4] 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 then 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. ImSystem and ImSystemPlus can collect available bandwidth information with a much smaller number of active measurements, compared to the system described in [1-4].

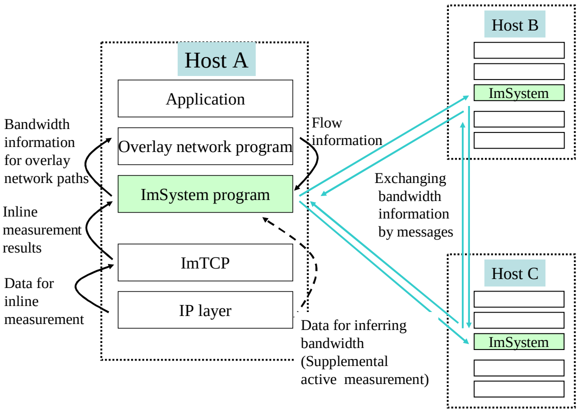

Host A
Application
Overlay network program
ImSystem program
ImTCP
IP layer
Bandwidth information for overlay network paths
Inline measurement results
Data for inline measurement
Flow information
Exchanging bandwidth information by messages
Data for inferring bandwidth (Supplemental active measurement)
Host B
ImSystem
Host C
ImSystem

図 1 Placement of ImSystem

3. ImSystem

3.1 Overview

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 collect the available bandwidth information of all overlay paths and present this information to the overlay networks. ImSystem works independent of the overlay network and can work with any overlay routing algorithms.

Figure 1 shows the placement of ImSystem, as well as the its relationships to the overlay network and IP networks. Each ImSystem program collects the available bandwidth measurement results delivered by ImTCP senders located at the same host. ImSystem programs also observe the transmissions of the overlay host to decide the time to perform active measurements, for the case in which there are no transmissions for a long time. These programs exchange the measurement results with each other so that every ImSystem program has full information concerning all paths in the overlay networks. Next, we explain in detail the working of ImSystem programs.

3.2 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. The 4-byte ID field is sufficient to distinguish all of the paths in a network with as many as 1,000 nodes. 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

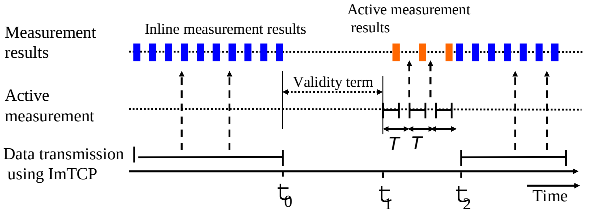

Measurement results
Inline measurement results
Active measurement results
Active measurement
Validity term
T T
Data transmission using ImTCP
t0
t1
t2
Time

図 2 Relationship among validity term, inline measurements and active measurements

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 [6], for abrupt change detection.

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

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. When an ImSystem program detects a change, it reports the average of K results to other ImSystem programs via the messages.

3.3 Suplemental active measurement

The start and stop of an overlay traffic flow is beyond the control of ImSystem. Therefore, 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. The active measurements are performed periodically in every T (s) intervals, where T is the maximum length of the time that an active measurement may take. Considering all of the existing measurement tools, 15 s is thought to be a good value for T. Active measurement tools can be Pathload [7] or any other end-to-end available bandwidth measurement tool.

Figure 2 shows the relationship between the validity term, inline measurements, and active measurements. Before t_0, ImSystem receives inline measurement results from the ImTCP connection. Here, t_0 is the time that the transmission of the overlay traffic stops, and $t_1 - t_0$ is the validity term of the last inline measurement results yielded at t_0. During this period, no transmission of overlay traffic occurs. At t_1, ImSystem starts active measurements. At t_2, a new overlay traffic transmission occurs. The active measurement then stops, and ImSystem receives the bandwidth information from ImTCP connection again.

We now consider the length of the validity term of the current inline measurement result, which is yielded at t_0 in Figure 2. 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.

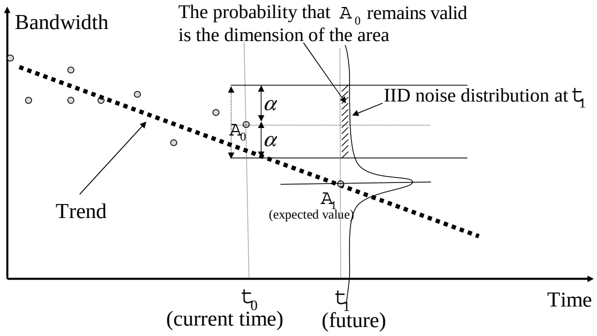

Bandwidth
The probability that A0 remains valid is the dimension of the area
IID noise distribution at t1
α
α
A0
Trend
A1
(expected value)
t0
(current time)
t1
(future)
Time

図 3 The accuracy of the previous result in the future environment

Here, we apply a model introduced in [8]:

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

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 a be 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 3, 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 robability that the real available bandwidth appears in $[A_0 - \alpha, A_0 + \alpha]$, where α is $0.2A_0$, since study in [5] 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

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

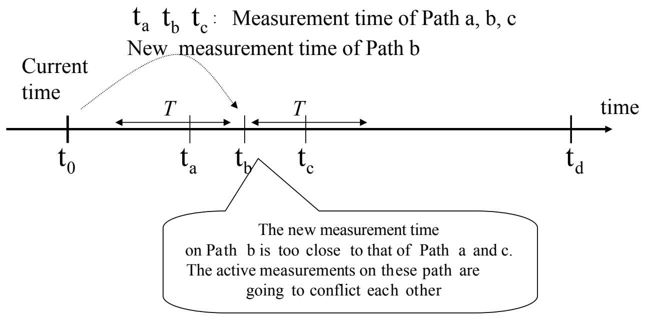
We assume that the measurement result A_0 becomes invalid at the time t_1 if the probability q_{t_1} falls below 1%. The validity term is then calculated as $t_0 - t_1$ where t_1 is the smallest solution of the following inequality:

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

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

ImSystem does not require the information about the topology of the under-laying IP network. However inferring the information by network tools such as `traceroute` is not a complicated job. In the related studies [2-4], the information is assumed to be available. Therefore, in this section, we examine how to improve ImSystem when the IP network topology is known.



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 path are going to conflict each other

図 4 Conflict in active measurements on overlapping paths

4.1 Overview

In ImSystem and other previously proposed systems [2-4], 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.

The information about the topology of the IP network enables ImSystem to overcome the problem mentioned above. That is, the timing of supplemental active measurements is adjusted to avoid the conflicts between the measurements on overlapping paths. ImSystem with the new function is referred to as ImSystemPlus.

Unlike ImSystem, the ImSystemPlus program does not start active measurements just after the alidity 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. During the delay time, the ImSystemPlus program estimates the available bandwidth of the path using the methods proposed in [2-4], instead of using the expired measurement results. The estimation is based on the measurement results on the overlapping paths, which are reported from ImSystemPlus programs locating on other nodes.

4.2 Conflict avoidance for measurements

In ImSystemPlus, the messages that the hosts exchange with each other have an additional 4-byte 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, the 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 4 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.

4.2.1 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 [9].

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

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.

4.2.2 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.

4.2.3 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, using the $joint()$ index as the parameter.

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

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. Simulation experiments

5.1 Simulation setting

We perform flow-level simulations to evaluate the performance of ImSystem and ImSystemPlus. The topology used for the simulations is shown in Figure 5. There is a four-node overlay network built upon an IP network. The parameters used in the simulation are as follows.

- The capacity of the links in IP network: 100 Mbps.
- IP link propagation delay: 0.1 s.
- IP routing: Shortest paths.

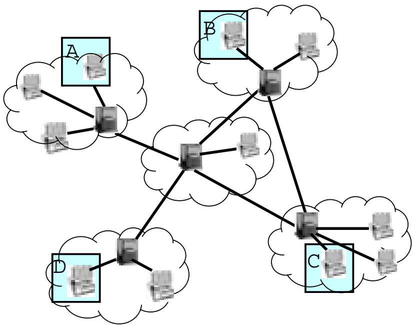


図 5　Network topology

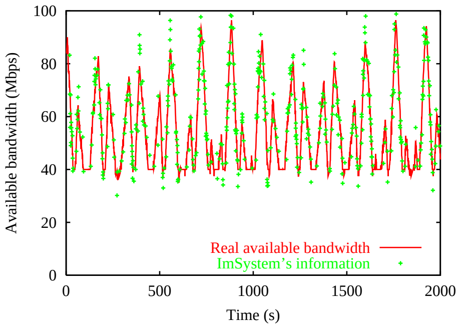


図 6　Measurement result of path D-B

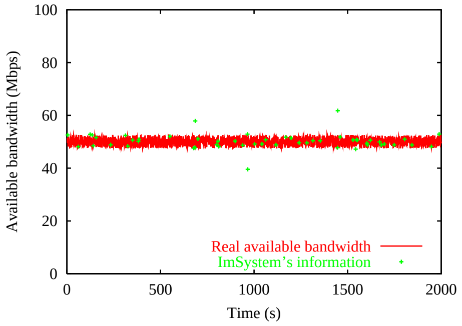


図 7　Measurement result of path B-C

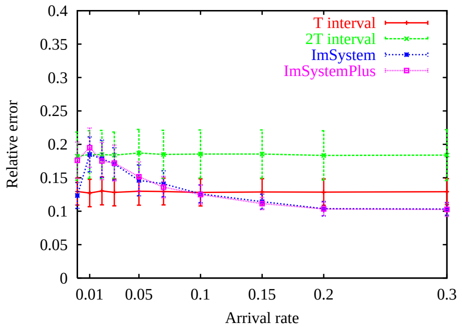


図 8　Relative errors of the collected information at node A

- Cross traffic on IP link: 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 linearly with times 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. Only for the links on the path between B and C, M is kept constant at 50 Mbps.
- Overlay traffic: 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: exponential distribution with an average of 20 s.
- Overlay flow rate: uniformly distributed in the range [100 Kbps, 1 Mbps].
- Active measurement: The active measurement is assumed to be Pathload.
  - The time required for 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 setting is based on the study results in [10] which show that Pathload consumes at least 2.5 Mbps for one measurement.
- 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 [5].
- Simulation time: 2000 s.

5.2 Bandwidth information of ImSystem

Figure 6 shows the changes of the real available bandwidth in the overlay path from host D to host B during 2000 s of the simulation. 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 7 shows how the ImSystem program on host A observer 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 does not update the value very often.

5.3 Accuracy of bandwidth information and the amount of probe traffic

For comparison, we also perform the simulations when the active measurement results are periodically deployed in all overlay paths at fixed intervals. The intervals are set to T and $2T$, 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. The average arrival rates of the overlay traffic at all nodes (F) are set to 0, 0.01, 0.05... as shown in the horizontal axis of Figure 8.

Figure 8 shows the average relative errors between the bandwidth information that host A collects and the real bandwidth values. The points in the line “T” shows the average relative error when the active measurements are performed every T (s). The length of the error bar at each point shows the variance of the relative error of the information on the paths. Similarly, the line “2T” shows the same information about results when we set the measurement intervals to $2T$ (s). The lines “ImSystem” and “ImSystemPlus” show the relative errors when the proposed systems are deployed. The figure shows that, when there is no overlay traffic, ImSystem has the same error with that when active measurements perfomed in every T (s). That is because in the case there is no overlay traffic, ImSystem also performs active measurements on the paths in every T (s). On the other hand, ImSystemPlus avoids the conflicting 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.

When the arrival rate of overlay flows is about 0.01 or larger, the two proposed systems use the inline measurement results and decrease a large amount of probe traffic, as can be seen in Figure 9. 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. That is because the proposed systems can perform inline measurement often in every overlay paths. When the arrival rate of overlay flow comes to 0.3, the proposed systems almost do not use the active measurements.

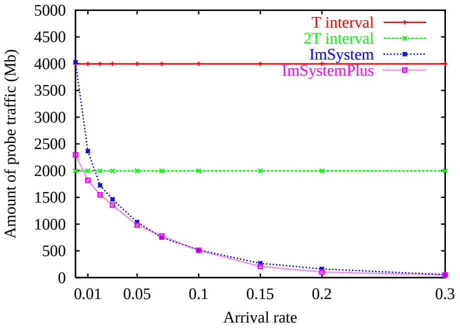


図 9　Amount of probe traffic

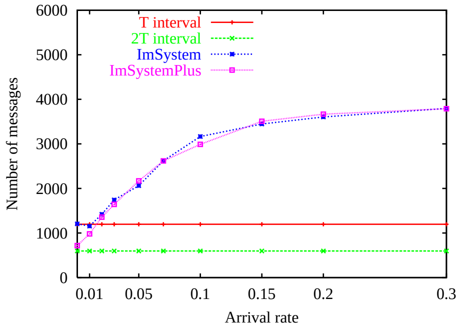


図 11　Number of messages sent for all paths

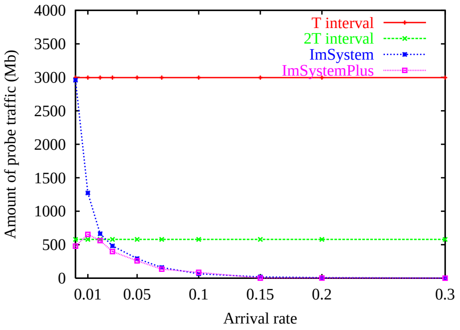


図 10　Amount of probe traffic that conflicts each others

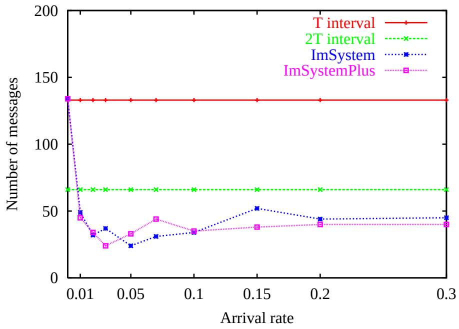


図 12　Number of messages sent for path B-C

5.4 Conflicting probe traffic

We next examine the probe traffic that conflicts with the other 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 10. ImSystemPlus successfully reduces the amount of conflicting probe traffic, compared to ImSystem. This is due to the function of detecting and avoiding conflicts in measurement of ImSystemPlus. Note that for the case in which the active measurements are periodically deployed in T or $2T$ intervals, the measurements are performed at random times. However, the conflict measurements in these case are still numerous.

5.5 Number of exchange messages

The number of messages that the end-hosts use to exchange the measurement results is shown in Figure 11. The proposed systems require a number of messages that is four times greater than that when active measurement is performed in T intervals. The reason for this is that, the available bandwidths of most of the overlay paths in this simulation fluctuate greatly, as shown in Figure 6, so that the ImSystem and ImSystem programs send several messages to update the bandwidth information. This results in small relative errors as shown in Figure 8. The size of a message is only 8 bytes for ImSystem and 12 bytes for ImSystemPlus. Therefore, the load by the traffic caused by the messages is far lower than the probe traffic of active measurements performed in “T” and “2T” intervals. Therefore, we can conclude that the proposed systems cause very little load on network.

Finally we examine the number of messages that Host B sent to update the bandwidth information of overlay path from Host B to Host C. We show the results in Figure 12. In this path, the non-overlay traffic fluctuates slightly around 50 Mbps. The changes in the available bandwidth of this path, when $F = 0.2$, are shown in Figure 7. In this case, ImSystem and ImSystemPlus programs send fewer messages than in the case when active measurements are performed in T and $2T$ intervals. Thus, when the available bandwidth does not change significantly, the proposed systems adaptively decrease the number of messages.

6. Conclusions

In the present paper, we proposed ImSystem and ImSystemPlus, systems that collect available bandwidth information of all end-to-end paths in an overlay network. The proposed systems work based on inline network measurements that work inside the active overlay data flow. Therefore, these systems inject little probe traffic onto the network while inferring the available bandwidth in a real-time fashion.

In the future, we intend to examine the implementation of the proposed systems and evaluate their performance in real network environments.

Acknowledgements

This work is partly supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) from the Ministry of Internal Affairs and Communications of Japan.

文　献

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

[2] C. Tang and P. McKinley, “On the cost-quality tradeoff in topology-aware overlay path probing,” in *Proceedings of the 11th IEEE Conference on Network Protocols (ICNP)*, Nov. 2003.

[3] 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.

[4] N. Hu and P. Steenkiste, “Exploiting internet route sharing for large scale available bandwidth estimation,” in *Proceedings of IMC’05*, Oct. 2005.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] J. Strauss, D. Katabi, and F. Kaashoek, “A measurement study of available bandwidth estimation tools,” in *Proceedings of IMC 2003*, Oct. 2003.