A Distributed and Conflict-Aware Measurement Method Based on Local Information Exchange in Overlay Networks

Dinh Tien Hoang*, Go Hasegawa† and Masayuki Murata*
*Graduate School of Information Science and Technology, Osaka University
1-5, Yamadaoka, Suita,Osaka, 565-0871 Japan, Email: {d-hoang,murata}@ist.osaka-u.ac.jp
†Cybermedia Center, Osaka University
1-32, Machikaneyama, Toyonaka, Osaka, 560-0043 Japan, Email: hasegawa@cmc.osaka-u.ac.jp

Abstract—**Network resource information, including available bandwidth, propagation delay and packet loss ratio, should be obtained by measurements for maintaining the effectiveness of overlay network services. However, measurements consume bandwidth that should be used for transferring data. Furthermore, although measurement accuracy can be enhanced by frequent measurements, measuring with high frequency can cause a measurement conflict problem that increases the network load and degrades the measurement accuracy. In this paper, we propose a distributed, conflict-aware measurement method that reduces the measurement conflicts while maintaining high measurement accuracy. The main idea is that overlay nodes exchange the route information and the measurement results with a small number of other overlay nodes while decreasing the measurement frequency. Simulation results show that the relative error in the measurement results can be halved with proposed method, while keeping the total measurement overhead unchanged, compared with the existing method. We also confirm that exchanging measurement results contributes more to the enhancement of measurement accuracy than performing measurements.**

I. Introduction

Overlay networks, which are defined in this paper as an application-level network constructed on the IP network, have been increasingly used to deploy network services due to their ability to produce effective overlay routing. Applications of overlay networks include end-system multicast (e.g., Narada [1]), P2P systems (e.g., BitTorrent [2]), content distribution systems (e.g., Akamai [3]), and resilient routing (e.g., RON [4]).

To maintain and improve the performance of network service, an overlay network should obtain the network resource information of the underlay network, including available bandwidth, propagation delay, and packet loss ratio. In general, these metrics should be measured frequently to obtain high measurement accuracy. RON [4] is one early-stage instance that measures all paths among overlay nodes. The measurement overhead becomes $O(n^2)$, where n is the number of overlay nodes. Therefore, [5] pointed out that the number of overlay nodes that can be applied is up to around fifty. Many solutions have been proposed to reduce measurement overhead [6]–[10]. However, these methods have shortcomings in terms of measurement accuracy [6] or available measurement metrics [8], [9].

Measurement accuracy is affected not only by the way measurements are performed but also by the overlap of underlay paths among overlay nodes. Concurrent measurement tasks of overlapping paths compete on common links for network resources (e.g., processing power at routers and link bandwidth), causing high load on the common links and additional error in the measurement results. [11] addressed this problem and proposed a method that schedules the timing of the measurement tasks of the overlay paths so that measurement conflicts can be avoided completely. However, the measurement frequency in this method is limited because of the heuristic behavior of the proposed scheduling algorithms [12]. Moreover, the methods in [6], [7], [10], [11] require a master node to aggregate the complete topology information of the underlay (IP) network, decide measurement timings, and give instructions to each overlay node. Therefore, the amount of time and network traffic for the aggregation of topology information and instructions are large, and the performance of overlay networks decreases when changes occur in the underlay or overlay networks.

In this paper, we propose a distributed measurement method that can reduce measurement conflicts and obtain high measurement accuracy. In our proposed method, each overlay node exchanges route information with a small number of other overlay nodes to detect the overlapping paths. Overlapping paths with the same source node are measured sequentially to completely avoid measurement conflicts. Overlapping paths with different source nodes are randomly measured to reduce measurement conflicts. The overlay node then exchanges the measurement results with other overlay nodes to statistically improve the measurement accuracy. Our method can also lower the measurement frequencies to reduce the overhead and measurement conflicts. We evaluate our method and compare it with the method in [11] by simulations with both generated and real Internet topologies.

The remainder of this paper is organized as follows. In Section II, we explain our method for detecting the overlapping of overlay paths. Section III describes our technique for reducing measurement conflicts and improving measurement accuracy. We evaluate our proposed method by simulations in Section IV, conclude our paper and discuss future work in Section V.

II. Detecting overlapping paths

A. Network model and definitions

We consider an overlay network in which the overlay nodes are deployed on the routers. This deployment has been simplified with such techniques as network virtualization [13] and software defined network [14]. Suppose that the network contains m routers, denoted by R_i $(i = 1, ..., m)$. We denote the underlay path between routers R_i and R_j as R_iR_j. If two different paths R_iR_j and R_sR_t share at least one link, R_iR_j (R_sR_t) is an *overlapping path* of R_sR_t (R_iR_j). Suppose that n $(n \leq m)$ overlay nodes are deployed on n routers. Density d

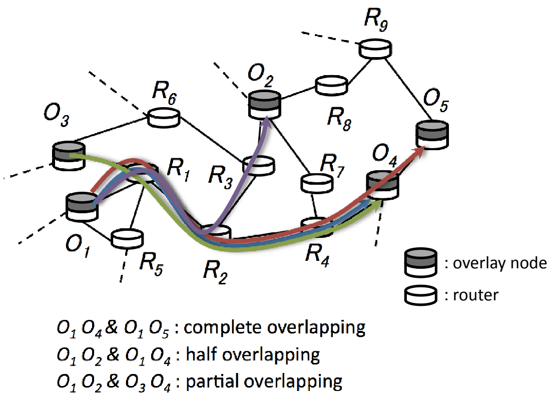

R9
O2
O5
R6
R8
O3
O4
R1
R3
R7
O1
R5
R2
R4
: overlay node
: router
O_1O_4 & O_1O_5 : complete overlapping
O_1O_2 & O_1O_4 : half overlapping
O_1O_2 & O_3O_4 : partial overlapping

Fig. 1. Classification of path overlapping

of the overlay nodes is defined as the ratio of the number of overlay nodes to the number of routers, i.e., $d = n/m$. We denote the overlay nodes as O_i $(i = 1, ..., n)$ and call the path between two overlay nodes an *overlay path*. For overlay path O_iO_j, O_i is the *source node*, and O_j is the *destination node* of the overlay path.

Figure 1 shows our classification of the overlapping state of overlay paths. In this paper, we classify the overlapping states into the following three types:

- *Complete overlapping*: One overlay path completely includes another overlay path.
- *Half overlapping*: Two overlay paths share a route from the source node to a router that is not an overlay node.
- *Partial overlapping*: Two overlay paths share a route that does not include the source node.

For example, in Fig. 1, path O_1O_4 is a complete overlapping path of O_1O_5. Paths O_1O_2 and O_1O_4 have a half overlapping relation. Path O_1O_2 is a partial overlapping path of O_3O_4. Note that the above classification covers all types of the overlapping states.

B. Methods for detecting complete and half overlapping paths

Complete overlapping and half overlapping can be detected by the source node of the overlay path using `traceroute`-like tools, as described in [15]. For example, in Fig. 1, when overlay node O_1 issues `traceroute` to O_4 and O_5, complete overlapping of paths O_1O_4 and O_1O_5 can be detected. Similarly, the shared route from O_1 to router R_2 by paths O_1O_2 and O_1O_4 can be detected when O_1 issues `traceroute` to O_2 and O_4.

C. Method for detecting partial overlapping paths

1) Detecting algorithms: Partial overlapping cannot be precisely detected only by `traceroute`-like tools, because the source nodes of the partial overlapping paths are different. Therefore, in this subsection, we propose the following method for detecting partial overlapping paths.

We demonstrate how overlay node O_i detects the partial overlapping paths. We denote the set of overlay paths whose source nodes are O_i, which contain at least two links and do not completely include other overlay paths as $\mathcal{S}_{O_i}$. We also denote the set of overlay paths whose destination nodes are O_i, which contain at least two links and do not completely include other overlay paths as $\mathcal{D}_{O_i}$. Note that we exclude one-link paths when defining $\mathcal{S}_{O_i}$ and $\mathcal{D}_{O_i}$ since they do not have partial overlapping paths. Also, we do not directly measure the paths that completely include other overlay paths, as described in Subsection III-A1. Our method consists of the following two algorithms that detect the partial overlapping paths of each path in $\mathcal{S}_{O_i}$ and $\mathcal{D}_{O_i}$.

- Algorithm 1:
  O_i detects the partial overlapping paths of each path O_iO_j in $\mathcal{S}_{O_i}$ as follows:
  1) O_i finds the candidates of the partial overlapping paths of O_iO_j by utilizing the information of its half overlapping paths.
     When O_iO_s and O_iO_t are the half overlapping paths of O_iO_j and when the length of the overlapping part of O_iO_j and O_iO_s is smaller than the length of the overlapping part of O_iO_j and O_iO_t, we infer that O_sO_t is a candidate of the partial overlapping path of O_iO_j.
  2) O_i exchanges path information with the source nodes of the candidates to decide the overlapping states between O_iO_j and the candidates.
     O_i exchanges path information with O_s to determine whether O_iO_j and O_sO_t actually have a partial overlapping relation. Furthermore, when receiving path information from other nodes, O_i may find new candidates of the partial overlapping paths. In that case, O_i repeats the information exchange and the decisions of the overlapping states.

  We use Fig. 1 to explain how Algorithm 1 works for path O_1O_2. Set $\mathcal{S}_{O_1}$ includes O_1O_2, O_1O_3, and O_1O_4 and does not include O_1O_5 because it completely contains O_1O_4. We infer that path O_3O_4 is a partial overlapping path of O_1O_2, because the length of the overlapping part of O_1O_2 and O_1O_3 is smaller than the length of the overlapping part of O_1O_2 and O_1O_4. O_1 then exchanges path information with O_3 to confirm whether O_1O_2 and O_3O_4 actually have a partial overlapping relation.
- Algorithm 2:
  O_i exchanges the information of the paths in $\mathcal{D}_{O_i}$ with their source nodes to detect their partial overlapping paths as follows.
  1) O_i receives the information of each path in $\mathcal{D}_{O_i}$ from the source node (referred to as O_s) of the path.
  2) O_i detects the partial overlapping paths of each path O_sO_i in $\mathcal{D}_{O_i}$ and sends the information of these paths to O_s.

  We also use Fig. 1 to explain how Algorithm 2 works for path O_2O_4. Set $\mathcal{D}_{O_4}$ includes O_1O_4, O_2O_4, and O_3O_4 and does not include O_5O_4 because it contains only one link. First, O_4 receives the information of paths O_1O_4, O_2O_4, and O_3O_4 from O_1, O_2, and O_3, respectively. O_4 then detects that O_1O_4, O_2O_4, and O_3O_4 are in a partial overlapping relation and sends the information of O_1O_4 and O_3O_4 to O_2.

2) Variation of detecting algorithms: Algorithm 1 includes iterations for information exchange and the decision of the overlapping states. When the number of iterations increases the detection accuracy is enhanced, while the overhead of the information exchange among the overlay nodes also increases. In addition, since Algorithms 1 and 2 can be conducted independently, we set the following four detecting levels to conduct Algorithms 1 and 2 to investigate the trade-off relationships between the detection accuracy and the information exchange overhead.

- detecting level 1: run Algorithm 1 with one iteration.

- detecting level 2: run Algorithm 1 with two iterations.
- detecting level 3: run Algorithm 1 completely.
- detecting level 4: run Algorithms 1 and 2 completely.

We have evaluated our proposed algorithms with four detecting levels by simulation experiments, in terms of detection accuracy and information exchange overhead. For the underlay network topology, we used the AT&T topology obtained from [16]. We also utilized generated topologies based on BA [17] and random models [18]. We generated ten topologies for each model using the BRITE topology generator [19]. All topologies have 523 nodes and 1304 links. We set the density of the overlay nodes to 0.2 and randomly chose them. For averaging the results, the choice of the overlay nodes was taken 100 times for the AT&T topology and ten times for each topology of the BA and random models.

We compared our method with the full-mesh method when evaluating the information exchange overhead. In the full-mesh method, each overlay node sends information of all overlay paths departing from it to all other overlay nodes. We have obtained the following results.

1) Our method needs only 1/6 and 1/3 of the path information exchanges, compared with the full-mesh method, to detect about 60% and 90% of the partial overlapping paths at detecting levels 1 and 4, respectively.
2) The results of detecting levels 2 and 3 are very close, meaning that we only need to run two iterations of the exchange loop of Algorithm 1.

III. Measurement method for overlay paths

In this section, we propose a method that reduces the measurement conflicts based on the status of the path overlapping detected by the method in Section II. First, note that if an overlay path has no overlapping paths, it is unnecessary to consider a method for reducing measurement conflicts. Therefore, we are only concerned with the case of an overlay path that has overlapping paths. We consider the following two cases of overlapping states:

1) When the overlay path completely includes other overlay paths, it is not measured directly.
2) When the overlay path does not include other overlay paths, we adjust the frequency and timing of the measurements to reduce the measurement conflicts.

The detailed mechanisms for the above two cases are described in Subsections III-A1 and III-A2, respectively. In Subsection III-B, we propose a statistical method for improving the accuracy of the measurement results.

A. Reducing measurement conflicts

1) Complete overlapping: In this case, the overlay path that includes the other overlay paths is not measured directly. Instead, the measurement result is estimated based on the measurement results of the overlay paths included in it.

We use Fig. 2 to explain this method. As shown in this figure, path O_iO_w completely includes path O_iO_j. When O_i issues `traceroute` to O_w, the `traceroute` packet goes through O_j, which learns that it is on path O_iO_w. O_j then measures path O_jO_w and transmits the result to O_i, which also learns that O_j is on path O_iO_w, based on the `traceroute` result. Then O_i does not directly measure path O_iO_w; it only measures path O_iO_j. O_i estimates the measurement result of path O_iO_w from the measurement result of path O_iO_j and that of path O_jO_w received from O_j. See [15] for details. Note that this method dramatically reduces the number of measurement paths, especially when the density of the overlay nodes is large [15]. Furthermore, the reasonable measurement accuracy of such a spatial composition method has been confirmed [20].

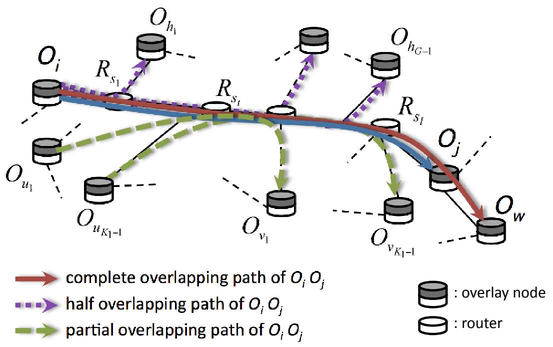


Fig. 2. Example for explaining our proposed measurement method

2) Half and partial overlapping: We explain our proposed method by describing the detailed behavior for overlay path O_iO_j shown in Fig. 2. We assume that O_iO_j has $(G_{i,j} - 1)$ half overlapping paths $(G_{i,j} \geq 1)$. For simplicity, we rewrite $G_{i,j}$ as G. We denote path O_iO_j as path 1, and each of its half overlapping paths as path p $(2 \leq p \leq G)$. Furthermore, we assume that, with the method described in Section II to detect partial overlapping paths, path p $(1 \leq p \leq G)$ has $(K_p - 1)$ partial overlapping paths $(K_p \geq 1)$.

Overlay node O_i can avoid the measurement conflicts between half overlapping paths 1, 2, ... and G simply by measuring them sequentially. On the other hand, because the source nodes of the partial overlapping paths of path p are different, measurement conflicts between them cannot be avoided completely. Therefore, we propose a technique that combines a sequential measurement for half overlapping paths and a random measurement for partial overlapping paths. We define the *measurement frequency* as follows. We assume that the time required for each measurement task is identical for all overlay paths and denote it as τ. We also assume that the measurement results of path p are aggregated in the time duration of T_p $(T_p \geq \tau)$. We call T_p an *aggregation period.* When a path is measured q $(q \leq T_p/\tau)$ times at an aggregation period, its measurement frequency at that aggregation period is defined as $f_p = q\tau/T_p$.

We introduce β_p as a value that reflects the dispersion of the measurement results of path p at an aggregation period. Note that the method to determine β_p is beyond the scope of this paper. β_p can be calculated based on the statistics of the measurement results or using the method in [9]. We set measurement frequency f_p proportional to β_p for all paths, i.e., $f_1/\beta_1 = f_2/\beta_2 = ... = f_G/\beta_G$. To avoid measurement conflicts between half overlapping paths, the sum of their measurement frequencies should be equal to or less than one, i.e., $\sum_{p=1}^{G} f_p \leq 1$. So we have $f_p \leq \beta_p/(\sum_{s=1}^{G} \beta_s)$.

To reduce the probability of measurement conflicts between path p and its $(K_p - 1)$ partial overlapping paths, we set the measurement frequency of path p to a value equal to or less than $1/K_p$, i.e., $f_p \leq 1/K_p$. In addition, we keep the measurement frequencies as large as possible to obtain as many measurement results as possible. Therefor, the measurement frequency of path p is decided based on the following equation:

$$f_p \quad = \quad \min\{\beta_p/(\sum_{s=1}^{G}\beta_s), 1/K_p\}. \tag{1}$$

Next, we explain our method for randomly deciding the measurement timings of path p so that the probability that the measurement of path p is carried out becomes f_p. We define a *measurement cycle* for the measurements of paths 1, 2, ... and G. We also divide the measurement cycle into multiple *measurement time slots*, each of which is assigned to the measurement of each path. We consider a scheme for allocating the measurement timings of paths p to these measurement time slots as follows.

When a path is measured at one measurement time slot of the measurement cycle, the probability that the measurement of the path is carried out becomes $1/G$. Therefore, we compare f_p with $1/G$ when considering the measurement timings of path p. We assume that $f_1 \geq f_2 \geq ... \geq f_G$ without loss of generality. For convenience, we define dummy value $f_0 = 1$. Since $\sum_{s=1}^{G} f_s \leq 1$, $0 \leq l < G$ exists, such that $f_0 \geq ... \geq f_l \geq 1/G \geq f_{l+1} \geq ... \geq f_G$.

If $l = 0$, meaning $f_p \leq 1/G, \forall 1 \leq p \leq G$, one measurement time slot in the measurement cycle is enough to allocate measurement timings for each path p.

On the other hand, $l > 0$ means that for path s where $s > l$, one measurement time slot is enough to allocate its measurement timings. For path t where $t \leq l$, one measurement time slot is not enough for allocating its measurement timings to satisfy its measurement frequency. In this case, the measurement time slot allocated to path s where $s > l$ is also used to measure path t where $t \leq l$ when path s is not measured.

In detail, we propose the following scheme for allocating the measurement timings of all paths.

1) Randomly decide the measurement order of path p ($1 \leq p \leq G$) at one measurement circle, and allocate the measurement time slot for each path.
2)
   - If $l = 0$,
     We measure path p with the probability of Gf_p at the measurement time slot allocated to it.
   - If $l \geq 1$,
     - For path t where $t \leq l$, we measure it at the measurement time slot allocated to it.
     - For path s where $s > l$, we measure it with the probability of Gf_s at the measurement time slot allocated to it.
       If path s ($s > l$) is not measured, the measurement time slot is used to measure path t ($t \leq l$) with the probability of $(f_t - 1/G)/\delta$, where $\delta = \sum_{s=l+1}^{G}(1/G - f_s)$.

B. Statistical method for improving the accuracy for measurement results

In the proposed measurement methods in Subsection III-A, because it is impossible to completely avoid measurement conflicts with partial overlapping paths, the accuracy of the measurement results decreases due to measurement conflicts. Therefore, in our proposed method, overlay nodes exchange measurement results and use statistical processing to improve measurement accuracy. We assume the measuring metric is delay.

We use Fig. 2 to explain the method for path O_iO_j. We assume that the overlapping parts of O_iO_j and its half and partial overlapping paths are divided by routers R_{s_1}, R_{s_2}, ..., R_{s_l}. In the proposed method, the delay measurements are individually conducted for overlapping parts $R_{s_1}R_{s_2}$, $R_{s_2}R_{s_3}$, ..., $R_{s_{l-1}}R_{s_l}$ as well as for end-to-end path O_iO_j. In detail, O_i measures the delays to routers R_{s_1}, R_{s_2}, ..., R_{s_l} and calculates the delay of $O_iR_{s_1}$, $R_{s_1}R_{s_2}$, ..., $R_{s_{l-1}}R_{s_l}$ and $R_{s_l}O_j$ as follows, where the delays of $O_iR_{s_1}$, $O_iR_{s_2}$, ..., $O_iR_{s_l}$, and O_iO_j are denoted as $t_{O_iR_{s_1}}$, $t_{O_iR_{s_2}}$,..., $t_{O_iR_{s_l}}$, $t_{O_iO_j}$, respectively.

$$\begin{aligned} t_{R_{s_k}R_{s_{k+1}}} &= t_{O_iR_{s_{k+1}}} - t_{O_iR_{s_k}} \quad ,k = 1, ..., l-1 \\ t_{R_{s_l}O_j} &= t_{O_iO_j} - t_{O_iR_{s_l}} \end{aligned}$$

When part $O_iR_{s_1}$ or $R_{s_k}R_{s_{k+1}}$ is the overlapping part of O_iO_j and its half overlapping path $O_iO_{h_a}$ ($1 \leq a \leq G - 1$), $t_{O_iR_{s_1}}$ or $t_{R_{s_k}R_{s_{k+1}}}$ is used to calculate the measurement results of both paths O_iO_j and $O_iO_{h_a}$. When part $R_{s_k}R_{s_{k+1}}$ or $R_{s_l}O_j$ is the overlapping part of O_iO_j and its partial overlapping path $O_{u_b}O_{v_b}$ ($1 \leq b \leq K_1 - 1$), O_i sends $t_{R_{s_k}R_{s_{k+1}}}$ or $t_{R_{s_l}O_j}$ and its measurement timing to O_{u_b}, so that O_{u_b} can use $t_{R_{s_k}R_{s_{k+1}}}$ or $t_{R_{s_l}O_j}$ to calculate the measurement result of path $O_{u_b}O_{v_b}$.

Finally, we use statistical processing for the data obtained by information exchange to calculate the measurement result of path O_iO_j. First, using the gathered values with the above method, we obtain the average value of the measurement results of $O_iR_{s_1}$, $R_{s_1}R_{s_2}$, ..., $R_{s_{l-1}}R_{s_l}$, and $R_{s_l}O_j$, which are denoted as $\bar{t}_{O_iR_{s_1}}$, $\bar{t}_{R_{s_1}R_{s_2}}$, ..., $\bar{t}_{R_{s_{l-1}}R_{s_l}}$, and $\bar{t}_{R_{s_l}O_j}$, respectively. The measurement result of path O_iO_j is then calculated as follows.

$$\bar{t}_{O_iO_j} \quad = \quad \bar{t}_{O_iR_{s_1}} + \sum_{k=1}^{l-1}\bar{t}_{R_{s_k}R_{s_{k+1}}} + \bar{t}_{R_{s_l}O_j} \tag{2}$$

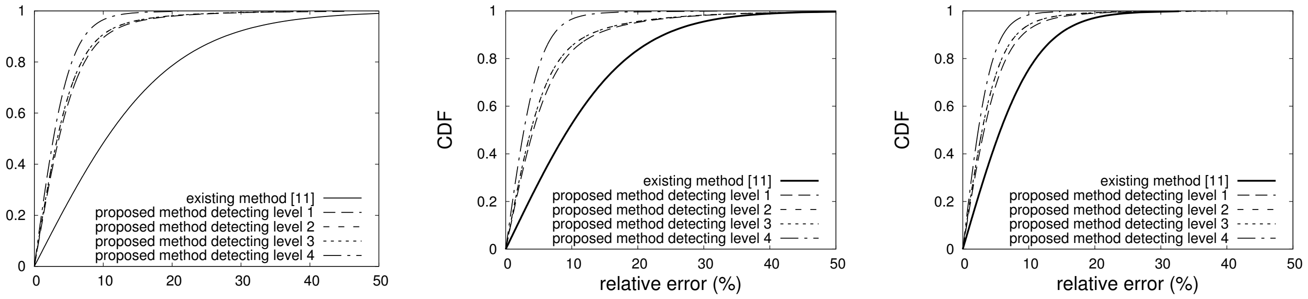
IV. Performance evaluation

A. Evaluation method

To evaluate the proposed method, we compare it with the method in [11], which also improves measurement accuracy by avoiding measurement conflicts. The authors of [11] proposed some heuristic algorithms from graph theory to divide measurement tasks into some groups, so that each group contains only measurement tasks of non-overlapping paths. The measurement tasks in the same group are simultaneously performed, while the measurement tasks in the different groups are sequentially performed. Therefore, measurement conflicts between overlapping paths are avoided completely.

We assume that the measuring metric is delay. We compare the proposed method and the method in [11] with the following metrics:

- Measurement accuracy
  We use the relative error of the measurement results as a metric to evaluate the measurement accuracy of the methods.
- System overhead
  We consider the following three kinds of overheads in conducting the measurements.
  - Path information accessing overhead
    This is caused when each overlay node uses `traceroute`-like tools to access the information of the overlay paths.
  - Measurement overhead
    This is caused when performing measurements on the overlay paths.
  - Information exchange overhead
    This is caused when overlay nodes exchange information of the overlay paths and measurement results with other overlay nodes.



(a) AT&T topology (b) BA topology (c) Random topology

Fig. 3. Relative error of measurement results

TABLE I
Average number of measurements, measurement results and concurrent measurements of one link during an aggregation period

Method	number of measurements			number of measurement results			number of concurrent measurements		
	AT&T	BA	Random	AT&T	BA	Random	AT&T	BA	Random
Existing method [11]	10.626	16.034	37.753	10.626	16.034	37.753	1.000	1.000	1.000
Proposed method detecting level 1	7.918	10.531	20.424	130.323	141.577	187.547	1.031	1.030	1.040
Proposed method detecting level 2	8.213	10.593	20.538	136.889	148.918	203.068	1.029	1.027	1.036
Proposed method detecting level 3	8.211	10.602	20.538	136.852	149.077	203.199	1.029	1.027	1.036
Proposed method detecting level 4	6.798	9.286	19.162	168.294	210.704	277.161	1.022	1.018	1.022

The relative error of the measurement result is calculated by:

$$\epsilon = \frac{|\bar{t} - t^*|}{t^*} \qquad (3)$$

where t^* and $\bar{t}$ are the real delay and average values of the measurement results, respectively.

We use the M/M/1 queueing model for each link in the network to calculate t^* and $\bar{t}$. We assume that each measurement on a link causes the increase in the link utilization, that results in the increase of the delay and delay jitter at the link. When the number of concurrent measurements on a link increases, the link utilization also greatly increases, causing additional error in the delay measurements.

The system overhead, denoted by A, is calculated by:

$$A = \frac{s_a + s_m + s_e}{d} \qquad (4)$$

where d is the duration during which the measurements were performed, and s_a, s_m and s_e are the sizes of the data packets used for accessing the path information, measuring, and exchanging the path information and the measurement results, respectively. We use second as the unit of d and bit as the unit of s_a, s_m and s_e. Therefore, the unit of A is bit per second (bps).

B. Simulation settings

In obtaining the following simulation results, our assumptions on the network topologies, the number and the distribution of the overlay nodes are the same as those mentioned in Subsection II-C2. We use the shortest path algorithm for underlay routing.

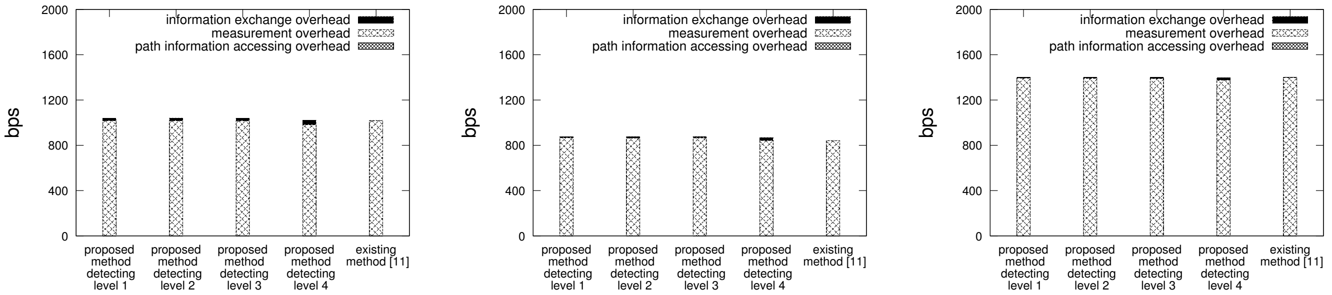
Value β_p, which is used for calculating the measurement frequencies by Eq. (1), is determined based on the coefficient of variance of the measurement results. Furthermore, we adjust the measurement frequencies in our method so that the system overheads of the proposed method and the method in [11] are the same.

We assume that we utilize `traceroute` to access information of overlay paths, and use `ping` to measure their delays. The size of each `traceroute` packet and `ping` packet is 28 and 475 bytes, respectively. We set the time of each measurement task $\tau = 1$ (second). The aggregation period is set to one hour, and the interval between two times of path information accessing is set to ten hours. We adjust the link capacity and the arrival rate of traffic so that the utilization of each link in the network becomes 0.5. We also assume that each measurement task increases the link utilization by 0.005.

C. Evaluation results and discussions

1) Measurement accuracy: Figure 3 shows the distribution of the relative error in the measurement results. The relative errors in our method are about half of those in the method in [11]. In our method, the relative errors decrease from detecting levels one to four, and the measurement accuracy of detecting level four greatly surpasses the other detecting levels.

To explain these results, we use the evaluation results of the parameters related to measurement accuracy. Table I shows the average number of measurements of an overlay path, the average number of the measurement results of a link (in our method) or a path (in the method in [11]) gathered during an aggregation period, and the average number of concurrent measurements performed at a link. In the method in [11], because measurement results are not exchanged among overlay nodes, the number of aggregated measurement results of an overlay path equals its measurement times. Furthermore, because the measurement conflicts are avoided completely, the average number of concurrent measurements remains one for all links. On the other hand, in our method, as explained in Subsection III-B, the aggregated measurement results of each



(a) AT&T topology

(b) BA topology

(c) Random topology

Fig. 4. Average system overhead of one link

link of an overlay path include the results obtained from the measurements performed by its source node and the results received from other overlay nodes. Furthermore, the average number of concurrent measurements of a link is very close to one, because we reduce the measurement conflicts by adjusting the measurement frequencies based on the status of the path overlapping.

As shown in this table, in our method, although the number of measuring times is smaller than that in the method in [11], the number of aggregated measurement results is much larger, and the number of measurement conflicts is small. Therefore, the measurement accuracy of our method surpasses the method in [11].

We also observe that when the detecting level of the proposed method is four, the number of measurement results is the largest, but the number of concurrent measurements is the smallest. Therefore, the measurement accuracy at detecting level four outperforms those at other detecting levels.

2) System overhead: Figure 4 shows the average values of the system overhead of the method in [11] and our proposed method with four detecting levels. The system overheads of these methods are almost equal. Furthermore, the measurement overhead occupies the most part of the system overhead, and the information exchange overhead is very small while the path information accessing overhead is negligible. This is because the size of the measurement traffic is much larger than the size of the traffic of information exchange and path information accessing. In our method, the information exchange overhead of detecting level four is slightly larger while the measurement overhead is smaller than those of the other detecting levels. This means that by shifting some amount of overhead from measurement to information exchange, we can significantly improve the measurement accuracy.

We finally conclude that from the results in Figs. 3 and 4, in our method, the detecting level four is the most effective for improving measurement accuracy.

V. Conclusion

In this paper, we proposed a distributed overlay network measurement method that reduces measurement conflicts by detecting path overlappings and adjusting the measurement frequencies and the measurement timings of overlay paths. We also proposed a method to improve measurement accuracy by exchanging measurement results among neighboring overlay nodes. Simulation results show that the relative error in the measurement results of our method can be decreased by half compared with the existing method when the total overheads of both methods are equal. We also confirmed that exchanging measurement results contributes more to the enhancement of measurement accuracy than performing measurements. In the future, we plan to construct a measurement system that applies our proposed method and investigate its effectiveness in real environments.

Acknowledgment

This work was supported in part by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan.

References

[1] Y. Chu, S. Rao, S. Seshan, and H. Zhang, "A case for end system multicast," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1456 – 1471, Oct. 2002.

[2] BitTorrent Home Page, available at http://www.bittorrent.com.

[3] Akamai Home Page, available at http://www.akamai.com.

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

[5] A. Nakao, L. Peterson, and A. Bavier, "Scalable routing overlay networks," *ACM SIGOPS Operating Systems Review*, vol. 40, pp. 49–61, Jan. 2006.

[6] C. Tang and P. McKinley, "On the cost-quality tradeoff in topology-aware overlay path probing," in *Proc. ICNP 2003*, Nov. 2003.

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

[8] N. Hu and P. Steenkiste, "Exploiting internet route sharing for large scale available bandwidth estimation," in *Proc. IMC 2005*, Oct. 2005.

[9] C. L. T. Man, G. Hasegawa, and M. Murata, "Monitoring overlay path bandwidth using an inline measurement technique," *IARIA International Journal on Advances in Systems and Measurements*, vol. 1, no. 1, pp. 50–60, 2008.

[10] Y. Gu, G. Jiang, V. Singh, and Y. Zhang, "Optimal probing for unicast network delay tomography," in *Proc. IEEE INFOCOM 2010*, Mar. 2010.

[11] M. Fraiwan and G. Manimaran, "Scheduling algorithms for conducting conflict-free measurements in overlay networks," *Computer Networks*, vol. 52, pp. 2819–2830, 2008.

[12] D. T. Hoang, G. Hasegawa, and M. Murata, "A distributed measurement method for reducing measurement conflict frequency in overlay networks," in *Proc. IEEE CQR 2011*, May 2011, pp. 1–6.

[13] N. M. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Computer Networks*, vol. 54, no. 5, pp. 862 – 876, 2010.

[14] J. Rubio-Loyola, A. Galis, A. Astorga, J. Serrat, L. Lefevre, A. Fischer, A. Paler, and H. Meer, "Scalable service deployment on software-defined networks," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 84 –93, Dec. 2011.

[15] G. Hasegawa and M. Murata, "Scalable and density-aware measurement strategies for overlay networks," in *Proc. ICIMP 2009*, May 2009, pp. 21–26.

[16] N. Spring, R. Mahajan, and C. Wetherall, "Measuring isp topologies with rocketfuel," in *Proc. ACM SIGCOMM 2002*, Jan. 2002.

[17] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[18] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 1617–1622, Dec. 1988.

[19] BRITE: Boston university Representative Internet Topology gEnerator, available at http://www.cs.bu.edu/brite/index.html.

[20] G. Hasegawa and M. Murata, "Accuracy evaluation of spatial composition of measurement results in overlay networks (in Japanese)," *IEICE technical report*, vol. 110, no. 39, pp. 1–6, May 2010.