Online End-to-End Quality of Service Monitoring for Service Level Agreement Verification

Xiaoyuan Ta and Guoqiang Mao
School of Electrical and Information Engineering
The University of Sydney
NSW 2006 Australia
Email: xiaoyuant@ee.usyd.edu.au, guoqiang@ee.usyd.edu.au

***Abstract*— Service-level-agreement (SLA) monitoring measures network Quality-of-Service (QoS) parameters to evaluate whether the service performance complies with the SLAs. It is becoming increasingly important for both Internet service providers (ISPs) and their customers. However, the rapid expansion of the Internet makes SLA monitoring a challenging task. As an efficient method to reduce both complexity and overheads for QoS measurements, sampling techniques have been used in SLA monitoring systems. In this paper, a novel adaptive stratified sampling strategy is developed based on the stratified sampling with optimum allocation to make the QoS monitoring less intrusive and more efficient. Simulations using real traffic traces are conducted, which show that the proposed algorithm achieves better performance than systematic sampling and Poisson sampling.**

***Index Terms*— SLAs, QoS, packet delay, LMS prediction, adaptive sampling**

I. INTRODUCTION

Internet Service Providers (ISPs) now offer service level agreements (SLAs) routinely to their customers. This has driven the service-providers to seek consistent testing and measurement methods to accurately measures network performance. To develop proper monitoring and performance estimation techniques therefore becomes a key challenge for network management. However, the implementation of measurement becomes increasingly difficult and complex due to the rapid expansion of the Internet. Moreover, the dramatic increase in the speed of wide area backbones presents obstacles to complete statistics collection. The enormous amount of measurement data may significantly increase the cost and resource usage [1].

Sampling-based measurement methods are used to reduce the quantity of control data and resources required for network performance monitoring, and finally to reduce the measurement complexity and cost. The principle of sampling techniques is to investigate the characteristics of a population of elements using a representative subset. In network performance monitoring, the performance metrics (e.g., packet delay, packet loss and jitter) are computed by choosing some particular packets among the entire traffic in the network. Systematic sampling and random sampling are two widely used methods in existing monitoring systems, but both of them have severe limitations. Stratified random sampling with optimum allocation can achieve higher estimation accuracy, but it requires extra statistics from the parent traffic trace, which are not known *a priori* in real applications. To address the challenge, a novel adaptive sampling strategy is proposed, which employs a least-mean-square (LMS) algorithm to predict the required statistics from past observations. The sample size for the next stratum is calculated from the predicted value of the required statistics. The proposed algorithm is applied for packet delay measurements. Simulation results show that the proposed adaptive sampling scheme produces good performance.

This research is supported by Optus through contract BLO-2004-7260 and the Australian research council through Discovery Project DP0559248.

The rest of this paper is organized as follows: Section II introduces three conventional sampling methods, as well as their advantages and disadvantages; Section III describes in detail the proposed adaptive stratified sampling scheme; Section IV introduces the delay traffic trace used for simulation; Section V presents the simulation results using real traffic traces provided by the WAND group; and finally Section VI concludes this paper.

II. SAMPLING TECHNIQUES

Traditional sampling techniques can be classified into three categories: systematic sampling, random sampling and stratified random sampling [1], [2], [3]. Fig. 1 illustrates these three sampling techniques.

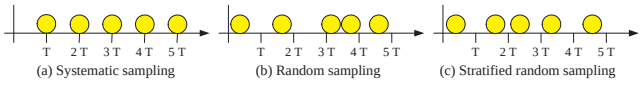

Fig. 1. Sampling techniques

A. Systematic Sampling

Systematic sampling generates sampling traffic according to a deterministic function. Generation of the sampling traffic is triggered by either time (i.e., at fixed intervals) or packet count (i.e., every k-th packet). Fig. 1.(a) shows systematic sampling with a period of T seconds.

The use of systematic sampling always involves the risk of biasing the results. If the systematics (e.g., periodic repetition of an event) in the sampling process resemble the systematics in the observed stochastic process (e.g., occurrence of event of interest in the network), there is a high probability that the estimation will be biased.

B. Random Sampling

Random sampling employs a random distribution function to determine when a sample should be generated. Typically the samples are generated according to a Poisson process or uniform process. As shown in Fig. 1.(b), random sampling may produce a varying number of samples in a given time interval. With random sampling, an unbiased estimate of the QoS metric can be achieved. However, the entirely random nature of the sampling process may also cause the undesirable effect that sampling intervals are not evenly distributed, and therefore the network may not be sampled for a rather long time.

C. Stratified Random Sampling

Stratified random sampling combines the fixed time interval used in systematic sampling with random sampling [4]. Fig. 1.(c) shows stratified random sampling with a period of T and a random sample is generated in each period.

The elements of the parent population are firstly grouped into subsets (i.e., strata), elements of sample are then taken from each subset. Depending on how the sample size (i.e., number of elements) is distributed among strata, stratified sampling can be further classified into proportional allocation and optimum allocation [5]. *Proportional allocation* means that the sample size in each stratum is proportional to the size of parent population in that stratum, while *optimum allocation* means that the sample size in each stratum is proportional to the standard deviation of the variable of interest (e.g., packet delay) in that stratum. In this paper, the procedure of stratified sampling is divided into fixed time intervals (i.e., stratum size) according to the correlation of the elements (e.g., packet delay) to be measured, then sampling packets are selected according to a random process during each interval. The stronger the correlation between packet delays in an interval is, the more accurate the estimation of the mean packet delay will be.

D. Sampling Trigger

The sampling process can be triggered by packet count, timer or packet-content [6]. In count-based sampling methods, the start and the finish of a sampling is triggered by packet count. For example, a count-based systematic sampling deterministically selects every k-th element (e.g., packet) out of the data set. Timer-based sampling methods use a timer instead of a packet count to trigger sampling. When the timer expires, we select the next sampling packet. Packet-content-based sampling methods trigger the sampling process according to the contents of a packet (e.g., TCP SYN packet). Claffy *et al.* [1] show that the performance difference between count-based sampling techniques and timer-based sampling techniques is very small.

E. Performance Comparison between Counted-based Simple Random Sampling and Stratified Sampling

As variance of the sample mean has been widely used as a performance measure [7, pp. 15], [8], the performance of these sampling techniques is compared by comparing the variance of the sample mean of different sampling schemes under the constraint that the sample sizes of different sampling methods are the same. The smaller the variance is, the better performance the sampling technique has. The sampling gain Δ is defined as the difference between the variance of the sample mean of two different sampling techniques [3]. Table I shows the notations used in our analysis:

TABLE I
NOTATIONS USED IN THE ANALYSIS

Property	Parent population	Sample
Number of elements	N	n
Number of elements in the l-th stratum	N_l	n_l
Number of strata	L	L
Mean value	μ	$\bar{y}$
Mean value in the l-th stratum	μ_l	$\bar{y}_l$
Variance of the variable of interest	σ^2	s^2
Standard deviation of the variable of interest	σ	s
Variance of the variable of interest in the l-th stratum	σ_l^2	s_l^2
Standard deviation of the variable of interest in the l-th stratum	σ_l	s_l
Variable of interest (e.g., packet delay)	y	y

Eq. 1 and Eq. 2 present the two assumptions used in the analysis, and they are widely used assumptions in the area [2], [3]:

$$N_l - 1 \approx N_l, \tag{1}$$

$$\frac{n}{N} < 0.05. \tag{2}$$

For stratified sampling, it can be shown that the variance σ^2 is related to the variances in each stratum by:

$$\begin{aligned} & \sigma^2 \\ = \; & \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{1}{N-1}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(y_{li}-\mu)^2 \\ = \; & \frac{1}{N-1}\sum_{l=1}^{L}\sum_{i=1}^{N_l}[(y_{li}-\mu_l)+(\mu_l-\mu)]^2 \\ = \; & \frac{1}{N-1}\sum_{l=1}^{L}(N_l-1)\sigma_l^2 + \frac{1}{N-1}\sum_{l=1}^{L}N_l(\mu_l-\mu)^2 \end{aligned} \tag{3}$$

Applying the approximation in Eq. 1, and multiplying both sides by a common factor $\frac{1}{n}(1-\frac{n}{N})$, where $(1-\frac{n}{N})$ is the finite population correction (fpc) factor, it can be obtained that:

$$\begin{aligned} & \frac{1}{n}(1-\frac{n}{N})\sigma^2 \\ = \; & \frac{1}{n}(1-\frac{n}{N})\sum_{l=1}^{L}\frac{N_l}{N}\sigma_l^2 + \frac{1}{nN}(1-\frac{n}{N})\sum_{l=1}^{L}N_l(\mu_l-\mu)^2. \end{aligned} \tag{4}$$

The variance of the sample mean with simple random sampling is [5, pp. 15]:

$$Var_{ran}(\bar{y}) = (1-\frac{n}{N})\frac{\sigma^2}{n}. \tag{5}$$

For stratified random sampling, the variance of the sample mean is given by [7, pp. 91]:

$$Var_{st}(\bar{y}) = \sum_{l=1}^{L}(\frac{N_l}{N})^2(\frac{N_l - n_l}{N_l})\frac{\sigma_l^2}{n_l}. \quad (6)$$

If *proportional allocation* is used, then n_l is given by [7, pp. 91]):

$$n_l = n\frac{N_l}{N}. \quad (7)$$

The variance of the sample mean for proportional allocation becomes:

$$Var_{prop}(\bar{y}) = \frac{1}{n}(1 - \frac{n}{N})\sum_{l=1}^{L}\frac{N_l}{N}\sigma_l^2. \quad (8)$$

Comparing Eq. 4 and Eq. 5 with Eq. 8, it can be shown that when the total sample size n is the same:

$$Var_{ran}(\bar{y}) = Var_{prop}(\bar{y}) + \frac{1}{n}(1 - \frac{n}{N})\frac{1}{N}\sum_{l=1}^{L}N_l(\mu_l - \mu)^2. \quad (9)$$

Hence, the sampling gain of the stratified sampling with proportional allocation is:

$$\Delta_{prop} = Var_{ran}(\bar{y}) - Var_{prop}(\bar{y}), \quad (10)$$

$$= \frac{1}{nN}(1 - \frac{n}{N})\sum_{l=1}^{L}N_l(\mu_l - \mu)^2 \geq 0. \quad (11)$$

The sampling gain is positive, which indicates performance improvement can be achieved in moving from simple random sampling to stratified sampling with proportional allocation.

If *optimum allocation* is used, then n_l is given by [7, pp. 97]:

$$n_l = \frac{nN_l\sigma_l}{\sum_{k=1}^{L}N_k\sigma_k}. \quad (12)$$

The variance of the sample mean for optimum allocation can be obtained from Eq. 6 and Eq. 12:

$$Var_{opt}(\bar{y}) = \frac{1}{n}(\sum_{l=1}^{L}\frac{N_l}{N}\sigma_l)^2 - \frac{1}{N}\sum_{l=1}^{L}\frac{N_l}{N}\sigma_l^2. \quad (13)$$

From Eq. 8 and Eq. 13, we can derive the difference between $Var_{prop}(\bar{y})$ and $Var_{opt}(\bar{y})$:

$$Var_{prop}(\bar{y}) - Var_{opt}(\bar{y}) = \frac{1}{nN}\sum_{l=1}^{L}N_l(\sigma_l - \bar{\sigma_l})^2 \geq 0, \quad (14)$$

where $\bar{\sigma_l}$ is:

$$\bar{\sigma_l} = \sum_{l=1}^{L}\frac{N_l}{N}\sigma_l. \quad (15)$$

Therefore with the same sample size n, ignoring the fpc factor, the sampling gain of the stratified sampling with optimum allocation in comparison with simple random sampling is:

$$\Delta_{opt} = Var_{ran}(\bar{y}) - Var_{opt}(\bar{y}) \quad (16)$$

$$= \frac{1}{nN}[\sum_{l=1}^{L}N_l(\mu_l - \mu)^2 + \sum_{l=1}^{L}N_l(\sigma_l - \bar{\sigma_l})^2] \quad (17)$$

$$\geq 0. \quad (18)$$

Based on Eq. 11 and Eq. 14, we can conclude that stratified sampling with proportional allocation performs better than the simple random sampling, and stratified sampling with optimum allocation performs better than stratified sampling with proportional allocation.

Earlier analysis is performed on count-based sampling techniques. Since the difference between count-based sampling and timer-based sampling is very small [1], the same conclusion may also extend to the timer-based sampling techniques.

III. ADAPTIVE STRATIFIED SAMPLING ALGORITHM

In the last section, we have shown that stratified sampling with optimum allocation has the best performance. However Eq. 12 implies that stratified sampling with optimum allocation requires the knowledge of the variance of the parent population in the l-th stratum, i.e., σ_l, in order to allocate the sample size in the l-th stratum. This requirement is unrealistic for online monitoring. In this section, we develop an adaptive stratified sampling algorithm, which uses the least-mean-square algorithm to predict the value of σ_l for sample size allocation. The proposed algorithm is then applied to packet delay sampling.

A. Least-mean-square Algorithm

The LMS algorithm is one of the most widely used adaptive linear algorithm. The computational procedure for the LMS algorithm is listed in the following [9, pp. 655]:

- Compute required output

$$\hat{x}_k = \sum_{i=0}^{m-1}w_k(i)x_{k-1-i} = \mathbf{W}_k^T\mathbf{X}(k), \quad (19)$$

  where m is the order of the predictor, $\mathbf{X}(k)$ is the input vector and $\mathbf{W}_k$ is the prediction coefficient vector.

$$\mathbf{X}(k) = [x_{k-1}, x_{k-2}, ..., x_{k-m}]^T, \quad (20)$$

$$\mathbf{W}_k = [w_k(0), w_k(1), ..., w_k(m-1)]^T. \quad (21)$$

  Initially, each weight $w_k(i)$ is set to an arbitrary fixed value.
- Compute the prediction error

$$e_k = x_k - \hat{x}_k. \quad (22)$$

- Update the coefficient vector

$$\mathbf{W}_{k+1} = \mathbf{W}_k + 2\upsilon e_k\mathbf{X}(k), \quad (23)$$

  where υ is the step size.

B. Prediction of the Sample Size within a Stratum

It has been shown in Section II-C that for stratified sampling with optimum allocation, the sample size within a stratum is:

$$n_l = \frac{nN_l\sigma_l}{\sum_{k=1}^{L}N_k\sigma_k}. \quad (24)$$

To simplify the estimation of n_l, an assumption is made that the parent population size N_l is approximately the same in each stratum, i.e.,

$$\frac{N_l}{N_k} \approx 1, \quad l \neq k. \tag{25}$$

This assumption is valid when the parent population size N_l is very large and the stratum size is a constant in time. This assumption has been validated using the real traffic trace. Fig. 2 shows the ratio N_k/N_1 of the real traffic trace with stratum size $= 50, 100, 130$ and 200 seconds respectively, where $N_k,\ k = 1, 2, ..., L$ is the total number of packets within the k-th stratum of the real traffic trace and N_1 is the total number of packets within the 1-st stratum of the real traffic trace. We can see that the ratio N_k/N_1 is bounded in the interval $[0.8, 1.2]$.

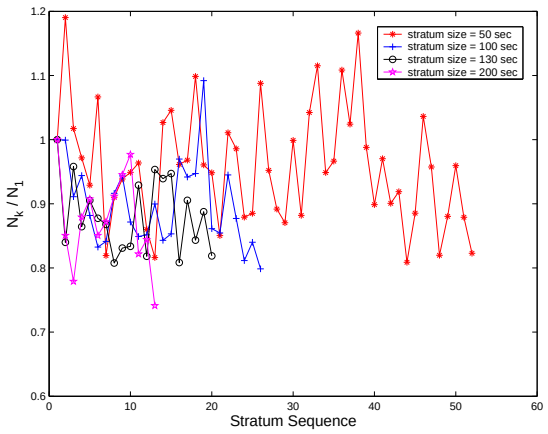

Fig. 2. Ratio of packet number between different strata

Using the assumption, Eq. 24 can be simplified as:

$$n_l \approx \frac{n\sigma_l}{\sum_{k=1}^{L}\sigma_k} = \frac{n\sigma_l}{L\bar{\sigma}_s} = \varphi\sigma_l, \tag{26}$$

$$\varphi = \frac{n}{\sum_{k=1}^{L}\sigma_k} = \frac{n}{L\bar{\sigma}_s}. \tag{27}$$

In real applications, φ can be simply treated as a proportionality constant which determines the sampling rate. φ can be chosen empirically and a larger φ will produce a higher sampling rate.

Since the standard deviation of packet delay σ_l is the true value of the parent delay trace, which cannot be obtained in real applications, it is approximated by the corresponding standard deviation of sampling packet delay s_l. Then the LMS algorithm is employed to predict s_l from its past values. Hence, the estimator $\hat{n}_l$ of sample size for the l-th stratum is computed by:

$$\hat{n}_l = \varphi\hat{s}_l, \tag{28}$$

$$\hat{s}_l = \sum_{i=0}^{m-1} w_l(i)s_{l-1-i}, \tag{29}$$

$$e_l = s_l - \hat{s}_l, \tag{30}$$

$$w_{l+1}(i) = w_l(i) + 2ve_l s_{l-1-i},\ i = 0, 1, ..., m-1 \tag{31}$$

The predictor order can be obtained using the AICC Criterion for order selection [10, pp. 171].

C. Estimation Error

The estimation error in $\hat{n}_l$ may increase the variance of the sample mean, i.e., decrease the measurement accuracy of the adaptive sampling method. From Eq. 6, the actual variance of the sample mean using the predicted stratum sample size $\hat{n}_l$ is:

$$Var_{act}(\bar{y}) = \sum_{l=1}^{L}(\frac{N_l}{N})^2\frac{\sigma_l^2}{\hat{n}_l} - \sum_{l=1}^{L}(\frac{N_l}{N})^2\frac{\sigma_l^2}{N_l}. \tag{32}$$

From Eq. 13 and Eq. 32, we can derive the relative error between them:

$$\frac{Var_{act}(\bar{y}) - Var_{opt}(\bar{y})}{Var_{opt}(\bar{y})} = \frac{1}{n}\sum_{l=1}^{L}\frac{(\hat{n}_l - n_l)^2}{\hat{n}_l} \tag{33}$$

$$= \frac{1}{n}\sum_{l=1}^{L} n_l\frac{(\phi_l - 1)^2}{\phi_l}, \tag{34}$$

where $\phi_l = \hat{n}_l/n_l$. When $\phi = 0.9$, the relative error between $Var_{act}(\bar{y})$ and $Var_{opt}(\bar{y})$ is $0.0111 = 1.11\%$; when $\phi = 1.2$, the relative error between $Var_{act}(\bar{y})$ and $Var_{opt}(\bar{y})$ is $0.0333 = 3.33\%$. We can see that the impact of the estimation error on the measurement accuracy is marginal when $\hat{n}_l$ is reasonably accurate.

IV. Parent Traffic Trace

In order to compare the performance of different sampling techniques, experiments are necessary. In this paper, all experiments are performed using a one-way delay trace as the parent traffic trace. This delay trace is generated by importing a real traffic trace into Opnet Modeler. This real traffic trace ("20010613-060000-e1.gz") was collected by the WAND research group at the University of Waikato Computer Science Department. It was captured between 6.00 a.m. and 8.54 a.m. on June 13th, 2001 on a 100Mbps Ethernet link. IP headers in the traffic trace are GPS synchronised and have a time accuracy of 1 μs. More information on the traffic trace and the measurement infrastructure can be found on the research group's website [11].

The network topology used in the Opnet Modeler is shown in Fig. 3. The selection of network nodes (e.g., switch, router, link) and background traffic utilisations of the links are shown in Table II. The first 2600-second part of the entire trace is

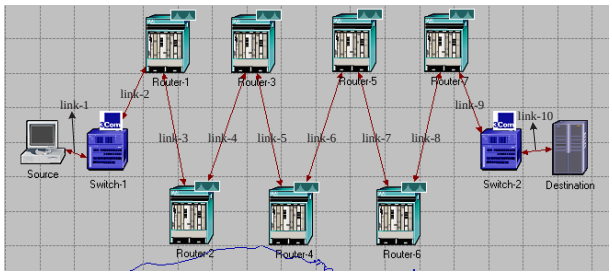

Fig. 3. Network topology used in Opnet Modeler.

TABLE II

SELECTION OF NETWORK NODES AND BACKGROUND TRAFFIC UTILISATIONS OF LINKS

Nodes	Description	Background traffic utilisation
Switch-1,2	$3\ Com's\ Switch\ 3800$	N/A
Router-1,2,..,7	$CISCO\ 12008$	N/A
Link-1,10	$100Mbps\ \ Link$	0%
Link-2,3,8,9	$100Mbps\ \ Link$	50%
Link-4,7	$100Mbps\ \ Link$	70%
Link-5,6	$100Mbps\ \ Link$	55%

then imported into the Opnet Modeler. After the simulation, we obtain a one-way delay traffic trace of a duration of 2600 seconds with 577718 packets. For the purpose of our study, we treat the 2600-second delay traffic trace as the parent traffic trace. Table III shows the summary statistics for the packet delay, packet size and inter-arrival time of the parent traffic trace.

TABLE III

SUMMARY STATISTICS FOR PACKET DELAY, PACKET SIZE AND INTER-ARRIVAL TIME OF THE PARENT DELAY TRACE

Property	Min.	Max.	Mean	Var.
Packet delay (ms)	41.092	141.305	86.024	8529
Packet size (bytes)	64	1518	440.5	302080
Inter-arrival Time (ms)	0.006	203.3280	4.5181	74.4127

V. SIMULATION RESULTS

In this section, we perform simulations with different sampling methods (i.e., timer-based systematic sampling, timer-based Poisson sampling, stratified sampling with optimum allocation and the proposed adaptive stratified sampling) and compare their performance. The parent delay trace used for simulation is the one-way delay trace presented in Section IV.

For stratified sampling with optimum allocation, the parameters required to calculate the sample size n_l are all true values from the parent delay trace. It is used as a benchmark, which represents the best sampling performance that can be achieved. The sample delay traces are selected directly from the parent delay trace. The sampling goal is to estimate the mean packet delay μ and the variance of packet delay σ^2 of the parent delay trace.

Several C programs were developed for sampling the sample delay traces and calculating the estimated mean packet delay $\hat{\mu}$ and the estimated variance of packet delay $\hat{\sigma}^2 = s^2$ from the sample delay traces, where $\hat{\mu}$ is the mean packet delay of the sample delay trace and s^2 is the variance of packet delay of the sample delay trace. For simulation, each kind of sampling (e.g., systematic sampling, systematic sampling) is repeated a number of times, and the random seed in the C programs is updated in each repetition. Let M denote the number of repetitions (i.e., sampling rounds). So after M sampling rounds, we obtain M different sample delay traces. The estimated mean delay $\hat{\mu}$ and estimated variance of delay s^2 are calculated for each sample delay trace in the M sampling rounds. Then we can obtain M estimated mean delay, i.e., $\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_M$ and M estimated variance of packet delay, i.e., $s_1^2, s_2^2, ..., s_M^2$. The absolute error of the estimated mean, i.e., $|\hat{\mu}_i - \mu|$, and the absolute error of the estimated variance, i.e., $|s_i^2 - \sigma^2|$ are also calculated for the M sampling rounds, where the true values μ and σ^2 are obtained in Section IV and shown in Table III.

To compare the performance of different sampling methods, several metrics are used, which are:

- Average value of the sample mean ($AMean$): the average value of the sample mean of the M sample delay traces. $AMean = \frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_i$, where M is the sampling rounds, $\hat{\mu}_i$ is the mean value of the i-th sample delay trace in the M sample delay traces. The smaller the difference between $AMean$ and μ is, the better the performance is.
- Average sample variance ($AVar$): the average value of the sample variance of the M sample delay traces. $AVar = \frac{1}{M}\sum_{i=1}^{M} s_i^2$, where s_i^2 is the variance of the i-th sample delay trace. The smaller the difference between $AVar$ and σ^2 is, the better the performance is.
- Mean square error (MSE) of the sample mean $\hat{\mu}_i$: $MSE = \frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_i - \mu)^2$. The smaller the MSE is, the higher the accuracy is.
- Absolute error of estimated mean ($AEMean$): $|\hat{\mu}_i - \mu|$, the smaller $|\hat{\mu}_i - \mu|$ is, the lower the variance of the sample mean $Var(\hat{\mu})$ is.
- Absolute error of estimated variance ($AEVar$): $|s_i^2 - \sigma^2|$, the smaller $|s_i^2 - \sigma^2|$ is, the better can s^2 estimate the true variance σ^2.

Then simulations for these four sampling methods are performed respectively. Each simulation is repeated for 222 times (i.e., $M = 222$). For timer-based systematic sampling, the sampling interval is specified as 1 second. For timer-based Poisson sampling, the mean sampling interval is 1 second. For the proposed adaptive sampling, the prediction parameters and stratum size are adjusted to the appropriate values. The predictor order is 4; the initial weights are: $w_l(0) = 0.257$, $w_l(1) = 0.210$, $w_l(2) = 0.209$ and $w_l(3) = 0.260$; the step size is: $\upsilon = 0.02$; the stratum size is: 50 seconds. These final values of the predictor order, initial weights and stratum size are obtained by using a different traffic trace ("20010612-060000-e1.gz"), which was captured between 6.00 a.m. and 8.54 a.m. on June 12th, 2001 on a 100Mbps Ethernet link [11], from that used in Section IV. As shown in Fig. 4, the error in predicting the standard deviation of sampling packet delay in each stratum, i.e., e_l in Eq. 30, is approximately independent, which indicates a good performance of the prediction algorithm. The total sample size n is specified as 2600 in order to make sure it has the same sample size as the timer-based systematic sampling and the timer-based Poisson sampling ($n = samping\ duration/sampling\ rate = 2600/1 = 2600$). For stratified sampling with optimum allocation, the stratum size is also specified as 50 seconds and the total sample size is 2600.

Table IV shows the simulation results. It can be seen that the stratified sampling with optimum allocation achieves the best performance. The proposed adaptive sampling scheme produces approximately the same performance as the stratified

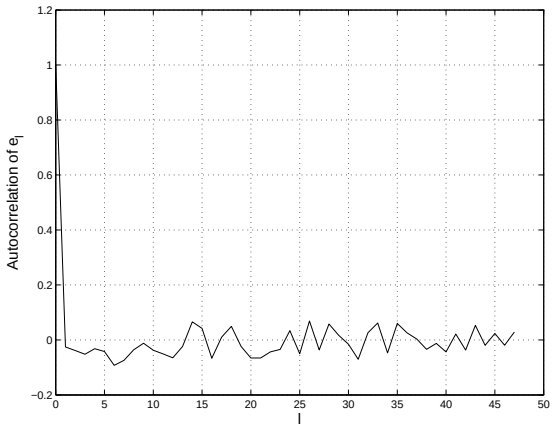

Fig. 4. Autocorrelation of prediction error e_l in Eq. 30.

TABLE IV

SIMULATION RESULTS OF SAMPLING TESTS WITH DIFFERENT SAMPLING METHODS. (TRUE VALUES ARE: $\mu = 86.824\ ms$, $\sigma^2 = 8529$)

Sampling method	M	$AMean$	$AVar$	MSE
Systematic	222	74.803 ms	5996	145
Poisson	222	64.998 ms	4710	478
Stratified with optimum allocation	222	88.023 ms	8959	5
Adaptive stratified	222	84.895 ms	8081	10

sampling with optimum allocation; it performs better than the timer-based systematic sampling and the timer-based Poisson sampling. Fig. 5 shows the absolute error of the estimated

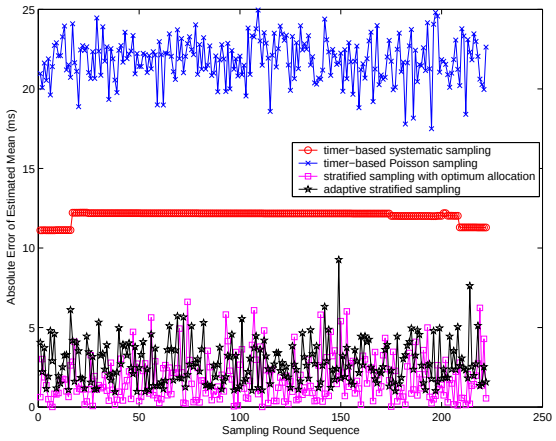

Fig. 5. Comparison of Absolute Error of Estimated Mean for different sampling methods. Stratum size: 50 seconds, sampling rounds: 222.

mean and Fig. 6 shows the absolute error of the estimated variance for the 222 sampling rounds. These also indicate that the proposed adaptive stratified sampling gives higher accuracy of estimate than the timer-based systematic sampling and the timer-based Poisson sampling.

VI. CONCLUSION

In this paper, we proposed a novel adaptive stratified sampling scheme for online end-to-end SLA monitoring. This proposed sampling scheme is based on the stratified sampling

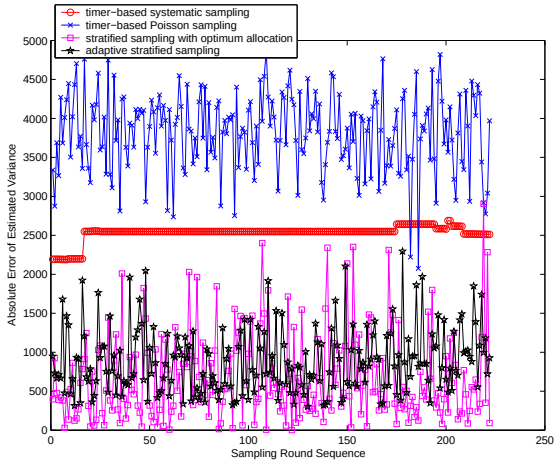

Fig. 6. Comparison of Absolute Error of Estimated Variance for different sampling methods. Stratum size: 50 seconds, sampling rounds: 222.

with optimum allocation. It employs an LMS algorithm to predict the extra statistics required to compute the sample size within a stratum in stratified sampling with optimum sampling, which are unknown *a priori* in real applications. We then performed simulations using real traffic network traffic, and compared the performance of the proposed adaptive sampling strategy with other sampling methods. The simulation results showed that the proposed adaptive sampling scheme performed better than the timer-based systematic sampling and the timer-based Poisson sampling. We have also investigated the impact of the estimation error (i.e., estimate n_l by $\hat{n}_l$) on the expected accuracy of estimation. The theoretical analysis and the simulation results both demonstrated that the impact was marginal.

REFERENCES

[1] K. Claffy, G. Polyzos, and H.-W. Braun, "Application of sampling methodologies to network traffic characterization," *ACM SIGCOMM Computer Communication Review*, vol. 23, no. 4, pp. 194–203, 1993.

[2] T. Zseby, "Deployment of sampling methods for sla validation with non-intrusive measurements," in *Proceedings of Passive and Active Measurement Workshop*, 2002.

[3] ——, "Stratification Strategies for Sampling-based Non-intrusive Measurements of One-way Delay," in *The Passive and Active Measurement Workshop*, 2003.

[4] E. Hernandez, M. Chidester, and A. George, "Adaptive Sampling for Network Management," *Journal of Network and Systems Management*, vol. 9, no. 4, 2001.

[5] S. K. Thompson, *Sampling*, 2nd ed. New York: John Wiley and Sons, 2002.

[6] P. D. Amer and L. N. Cassel, "Management of sampled real-time network measurements," in *Local Computer Networks, 1989, Proceedings 14th Conference on*. Mineapolis, MN, USA: IEEE, October, 1989, pp. 62–68.

[7] W. G. Cochran, *Sampling Techniques*, 2nd ed. New York: John Wiley and Sons, 1964.

[8] I. Cozzani and S. Giordano, "Traffic sampling methods for end-to-end qos evaluation in large heterogeneous networks," *Computer Networks and ISDN Systems*, vol. 30, no. 16-18, pp. 1697–1706, 1998.

[9] E. C.Ifeachor and B. W.Jervis, *Digital Signal Processing-A Practical Approach*, 2nd ed. London: Pearson Education Limited, 2002.

[10] P. Brockwell and R. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. New York: Springer, 2002.

[11] Online, "Auckland-VI trace data, http://pma.nlanr.net/Traces/Traces /long/auck/6/."