

Wide Area Performance Monitoring Using Aggregate Latency Profiles ^{*}

Vladimir Zadorozhny¹, Avigdor Gal², Louiqa Raschid³, and Qiang Ye¹

¹ University of Pittsburgh Pittsburgh, PA
{qye,vladimir}@sis.pitt.edu

² Technion – Israel Institute of Technology Haifa, Israel
avigal@ie.technion.ac.il

³ University of Maryland College Park, MD
louiqa@umiacs.umd.edu

Abstract. A challenge in supporting Wide Area Applications (WAA) is that of scalable performance management. Individual *Latency Profiles (iLPs)* were proposed in the literature to capture latency distributions experienced by clients when connecting to a server; it is a passive measurement made by client applications and is gathered on a continuous basis. In this paper, we propose a scalable technique for managing *iLPs* by aggregating them into aggregate *Latency Profiles (aLPs)*. We use measures such as mutual information and correlation to compare the similarity of pairs of *iLPs*.

1 Introduction

Wide area applications (WAAs) utilize a WAN infrastructure, e.g., the Internet, to connect a federation of hundreds of servers, typically content providers, with tens of thousands of clients. Servers provide services that may range from simple downloads of digital content to complex Web services with multiple interchanges between client and server. It is expected that WAA must scale to millions of client and server pairs. As an example, consider a global name service such as the Handle protocol, an IETF/IRTF standard from CNRI- Corporation for National Research Initiatives [13]. Handle provides a namespace, a name resolution service, and protocols for digital object location and access. The International Digital Object Identifier (DOI) Foundation (www.doi.org) and the community of publishers utilize handles to facilitate the identification and exchange of intellectual property in the digital environment. It is expected that such applications must scale to tens of millions of Handles and thousands of content servers, representing the digital content managed by the publishing community, and large numbers of Handle clients.

A significant challenge in deploying WAA is that of scalable performance management for large numbers of clients. The unpredictable behavior of a dynamic WAN [11, 12] results in a wide variability in access latency (end-to-end delay). There has been extensive research in the networking literature to develop metrics and models to predict latencies, including *Internet distance* and *points of congestion* [1, 4, 9, 11, 12]. There

^{*} This research is supported by NSF Grants IIS0219909 and EIA0130422.

has been research on route aggregation based on IP prefixes exchanged via the Border Gateway Protocol (BGP) and exploiting BGP information to monitor and predict performance [5, 7]. BGP routes expressed as paths via Autonomous Systems (ASes). However, an entire AS may not demonstrate homogeneous behavior, *e.g.*, whenever it spans a large geographic area. Further, the effort to acquire knowledge of the BGP paths between different clients and servers may vary, since some clients and servers do not provide a looking glass service. Finally, while network topology is often a good predictor of latency, it may be the case that there is no available latency data for a closely matching client and server pair. Alternately, a client and server pair with similar BGP routing may not always be a good predictor of latency for the client and server of interest, *e.g.*, if the two servers experience dissimilar workloads, or were associated with dissimilar points of congestion. Latency prediction models based on network characteristics alone would not be appropriate, or would not differentiate the cases described above. This too motivates the complementary need for a management tool and measures that do not rely on extensive (and sometime unavailable) knowledge of the network and its characteristics.

In [10], we proposed *latency profiles* as a conceptual model to characterize the behavior of sources over a WAN. Latency profiles (LPs) are time-dependent latency distributions that capture the changing latencies clients experience when accessing a server; it is measured by client applications or middleware and is gathered passively and on a continuous basis. Latency profiles can be utilized as a WAA monitoring tool, to predict latencies that clients should expect in response to requests, using historical data and recurrent behavior patterns. However, in the presence of hundreds of servers and tens of thousands of clients, managing millions of latency profiles cannot scale. Therefore, we explore in this paper a method for aggregating latency profiles. We propose information theoretic and statistical measures such as mutual information and correlation to compare the similarity of pairs of iLPs. Individual latency profiles (iLPs) will be aggregated into an *aggregate latency profile (aLP)*. A representative latency profile for this aggregate will then be maintained. Whenever a request for service arrives, a prediction will be based on the representative latency profile. Using aLPs allows us to discover aggregate performance patterns that would have been difficult to obtain using network topology and characteristics alone. We empirically show that there is a considerable amount of non-random associations between iLPs. While some of the strong associations can be explained based on physical network topology and characteristics, our experiment also shows that given a group of client and server ASes, with similar (overlap) of BGP routes, there may be a wide variation of the strength of non-random associations between pairs of iLPs.

2 Wide Area Performance Monitoring

Figure presents a WAA performance monitoring architecture. There are three types of nodes, namely clients, content servers, and performance monitors (PMs). Clients continuously download data from content servers and passively construct individual iLPs. PMs manage large collections of iLPs; this is done by aggregating iLPs into a smaller number of aLPs; PMs then manage some number of aLPs and the associated

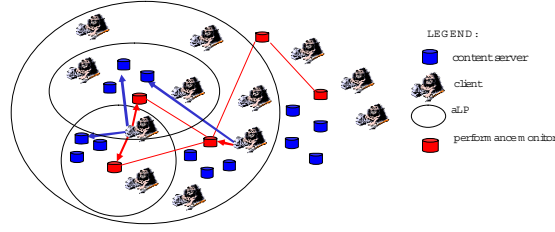


Fig. 1. WAA monitoring architecture based on performance profiles

iLPs. Clients consult PMs to obtain a prediction. The scope of an aLP is depicted in Figure 1 by ellipses, where each ellipse contains clients and servers for each an iLP can be constructed.

Suppose a latency prediction is requested for a pair (c, s) representing client c and server s . Suppose also the PM does not have an associated iLP , from the same client AS of c to the same server AS of s , that can be directly used to predict latency. Alternatively, the system does not have sufficient resources to continuously maintain all profiles. Assume further that there exist an iLP_1 associated with a client/server pair (c_1, s) for a different client AS than that of c , but to the same server AS as of s . Similarly, there is an iLP_2 for client/server pair (c, s_1) (same client AS as c and different server AS) and iLP_3 for client/server pair (c_1, s_1) (different client AS as c and different server AS as s). Now the PM can choose either iLP_1 or iLP_2 to make a prediction for the client/server pair (c, s) . It is also possible that there exists strong non-random associations between iLP_1 , iLP_2 and iLP_3 . In this case, the best estimate of access latency for (c, s) is possibly obtained by aggregating iLP_1 , iLP_2 and iLP_3 into an aggregate latency profile aLP , and choosing a representative profile.

2.1 Individual and Aggregate Latency Profiles

Given a client c , a server s , an object of size b , and a temporal domain T , an *individual latency profile* is a function $iLP_{c,s} : T \times \mathcal{N} \rightarrow \mathbb{R}^+ \cup \{TO\}$. $iLP_{c,s}(t, b)$ represents the end-to-end delay for a request from server s at time t , given as either a real number or using TO to represent a timeout. $iLP_{c,s}$ comes in two flavors, similar to [3]. One flavor measures time-to-first, which depends on factors such as workload at the server and size of the requested object. The other flavor measures time-to-last, which has a greater dependency on network bounds. Due to the stochastic nature of the network, $iLP_{c,s}$ is clearly a random variable.

An *aggregate latency profile* aLP_{iLP} combines a set of n individual latency profiles $iLP = \{iLP_{c_i, s_i}\}_{i=1}^n$. We construct an aLP by grouping iLP s with similar characteristics that are non-randomly associated with each other; this will ensure that the grouping will benefit the prediction ability of the aLP. For this grouping, we rely on information theoretic and statistical measures computed for the pair-wise association of iLPs. In particular, we use mutual information [2], and correlation [8]. A higher mutual information between two iLP s means that those iLP s are non-randomly associated. Conversely, a mutual information of zero means that the join distribution of iLP s holds no more information than their individual distributions. A higher correlation between two iLP s can also indicate that those iLP s are non-randomly associated. In general, there is no straightforward relationship between correlation and MI [6]. While correlation captures linear dependence, mutual information is a general dependence measure.

After constructing an *aLP* from a set of non-random associated *iLPs*, we can improve the prediction of an *iLP* by using observations of other *iLPs* in the *aLP*. Recall that using an observation of a random variable Y which is related to a random variable X in some way, *e.g.*, Y is non-randomly correlated with X , an optimal mean-square-error estimator of X given Y is the conditional expectation of X given Y , $E(X|Y)$ [8]. We use conditional expectation to utilize the meaningful relationships within an *aLP* in order to improve latency prediction.

3 Experiments

In this section we report on part of our experiences with constructing *aLPs*. The experimental data was collected over the CNRI Handle testbed [13], – an emerging IETF/IRTF standard that provides a global name service for use over WANs. We gathered data from November to December 2002. The data is typically PDF files that are reachable via Handle resolution. We report on the performance of 22 clients (2 each on 11 client ASes) accessing 10 servers, yielding 220 *iLPs*. We explored two approaches for grouping *iLPs* in *aLPs*, namely using mutual information and using correlation.

We group strongly related *iLPs* in *aLP*. We applied conditional expectation to estimate individual latencies using observations of latencies from a representative *iLPs* within one *aLP*. All our *aLPs* in this experiment consisted of two *iLPs*. For each *aLP* $\{iLP_i, iLP_j\}$ we estimated latencies of *iLP*_{*i*} using observations of *iLP*_{*j*}, *i.e.*, we choose *iLP*_{*j*} as a representative profile.

We computed the average relative estimation errors for all *iLP* pairs (*aLPs*) considered in our experiment. Relative estimation error is defined as $abs(x - x_{est})/x$, where x and x_{est} are actual and estimated latencies correspondingly. For each *aLP* $\{iLP_i, iLP_j\}$ we average the relative errors of estimation of all individual latencies from *iLP*_{*i*}. Figure 2 plots the distribution of the average relative estimation error. We observe that variability of the relative error is considerable. Figure 2 shows that major part of estimation errors (about 9000 estimations) is in a good range of $[0, 1]$. However, more than 1000 estimation errors are large (above 3), and as we see from Figure 2, they can be as much as 75. Meanwhile, from our experiments we found that *practically all of the large estimation errors spread over areas of low MI (< 0.4) and low correlation (< 0.2)*.

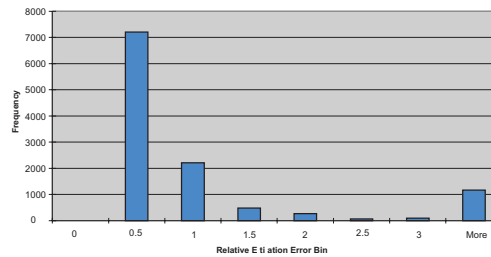


Fig. 2. Distribution of Average Relative Estimation Error

We observed that using MI and correlation to construct *aLPs* does not always guaranty the best latency estimation, but it helps to maintain good estimation quality. Moreover, avoiding non-related representative *iLPs* efficiently eliminates large estimation errors. We conclude that aggregating non-randomly associated latency profiles can practically assist in wide area performance monitoring.

4 Conclusion

We have presented the concept of an aggregate latency profiles as a scalable methodology for utilizing latency profiles. Mutual information and correlations are compared in their ability to explore useful aggregate latency profiles. Our experiments show that in general correlation serves better is generating aggregate latency profiles and in predicting latencies. We plan on implementing our methods in a prototype, allowing the generation of aggregate latency profiles and testing them out in retrieving documents based on handle information. We are going to use more advanced prediction techniques such as Neural Networks and Web Prediction Tool [14], to fully utilize prediction power of aggregate latency profiles.

References

1. P. Francis, S. Jamin, V. Paxson, L. Zhang, D. Grynowicz, and Y. Jin. An architecture for a global internet host distance estimation service. 1999.
2. F.Reza. *An Introduction to Information Theory*. McGraw-Hill, 1961.
3. J.-R. Gruser, L. Raschid, V. Zadorozhny, and T. Zhan. Learning response time for websources using query feedback and application in query optimization. *VLDB Journal*, 9(1):18–37, 2000.
4. S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavin, and L. Zhang. On the placement of internet instrumentation. In *Proceedings of IEEE InfoComm*, 2000.
5. B. Krishnamurthy and J. Wang. On network-aware clustering of web clients. In *Proc. SIGComm*, pages 97–110, 2000.
6. W. Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, (60), 1990.
7. Z. Mao, C. Cranor, F. Douglis, M. Rabinovich, O. Spatscheck, and J. Wang. A precise and efficient evaluation of the proximity between web clients and their local dns servers, 2002.
8. W. Mendenhall and T. Sincich. *Statistics for Engineering and the Sciences*. Macmillan Publishing, 1985.
9. V.N. Padmanabhan and L. Subramanian. An investigation of geographic mapping techniques for internet hosts. *Proceedings of the SIGCOMM*, 2001.
10. L. Raschid, H.-F. Wen, A. Gal, and V. Zadorozhny. Latency profiles: Performance monitoring for wide area applications. In *Proceedings of the Third IEEE Workshop on Internet Applications (WIAPP '03)*, San Jose, CA, June 2003.
11. D. Rubenstein, J. Kurose, and D. Towsley. Detecting shared congestion of flows via end-to-end measurement. *Proceedings of the ACM SIGMETRICS Conference*, 2000.
12. M. Stemm, S. Seshan, and R. Katz. A network measurement architecture for adaptive applications. In *Proceedings of IEEE InfoComm*, 2000.
13. S. Sun and L. Lannom. Handle system overview. *IRDM/IRTF Draft*, 2001, http://www.idrm.org/idrm_drafts.htm, 2001.
14. V. Zadorozhny, L. Raschid, T. Zhan, and L. Bright. Validating an access cost model for wide area applications. *Proceedings of the International Conference on Cooperative Information Systems (CoopIS 2001)*, pages 371–385, 2001.