Network Measurement as a Cooperative Enterprise

Sridhar Srinivasan, Ellen Zegura
Networking and Telecommunications Group
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332

Abstract—**Real-time network measurements can be used to improve performance of existing Internet services and support the deployment of new services dependent on performance information (e.g., topologically-aware overlay networks). Internet-wide measurement faces numerous scaling-related challenges, including the problem of deploying enough measurement endpoints for wide-spread coverage. We observe that peer-to-peer networks, made up of "volunteer" hosts around the Internet world, have the potential to provide a level of coverage that greatly exceeds that made possible with the tedious human process of negotiating endpoint locations. We therefore propose a distributed peer-to-peer system that can be queried for network performance information. We sketch the architecture and operation of such a system and briefly relate it to alternative proposals for measurement infrastructures. Finally, we list open problems related to the design and realization of such a system.**

INTRODUCTION

Measurements of network performance are valuable for improving performance, assessing utilization, engineering traffic and validating design choices. We are particularly interested in real-time measurements that can be used to improve the performance of existing Internet services and support the deployment of new services dependent on performance information (e.g., topologically-aware overlay networks).

The challenges involved in constructing an Internet-scale measurement infrastructure are considerable. First, there is the difficulty of coverage, that is, obtaining access to a large number of distributed measurement endpoints. Current measurement systems generally involve human-negotiated access to endpoints either with ISPs and/or friends at diverse locations [3, 6]. Second, there is the difficulty of obtaining accurate measurements, given the time-varying nature of network properties of interest (e.g., loss rate, available bandwidth, latency). Third, there is the issue of overhead. Care must be taken to avoid a measurement process that imposes excessive overhead on the overall system. These challenges have obvious interactions; for example, one can reduce overhead with less accurate measurements or more coarse-grained coverage.

We observe that peer-to-peer networks, made up of "volunteer" hosts around the Internet world, have the potential to provide a level of coverage that greatly exceeds that made possible with the tedious human process of negotiating endpoint locations. *We therefore propose a distributed peer-to-peer system that can be queried for network performance information.* The M-coop (or Measurement cooperative) is a system that answers queries about the path between two arbitrary IP addresses. In addition to performance metric information, the system returns assessments of the metric accuracy and trustworthiness.

Such a system does not, on its own, solve the problem of obtaining accurate measurements. Nor does it solve the problem of measurement overhead. Indeed, because such a system may involve a very large number of end systems, the scaling problem is significant. We will rely on known techniques for dealing with accuracy (e.g., using moving weighted averages); we will introduce mechanisms for reducing the number of end systems that form measurement pairs to help with the scale problem.

Such a system brings with it a number of additional challenges. Well-known are the problems that result from a peer-to-peer system of hosts that may join and leave on a frequent basis [4, 5, 8]. Merely keeping the M-coop system connected can be challenging in this environment. Because the measurement entities are volunteers, and not under any accountable control, we must deal with issues of inaccurate information due to misconfiguration or malicious use. The inclusion of a trustworthiness value recognizes the fact that information quality may vary. We must also consider the question of incentive. What would motivate someone to include their host in an M-coop measurement infrastructure? The limited examples of deployed peer-to-peer systems indicate that people are motivated by self-interest (e.g., Napster, Gnutella) and by a sense of contributing to a larger "good" (e.g., SETI@home). An Internet-scale measurement infrastructure has the potential to tap both sources of motivation.

In the next section, we sketch the design of an M-coop system. In Section-D, we briefly describe related work and then conclude with a section discussing the open problems

in the design and realization of an Internet-wide peer-to-peer measurement system.

AN M-COOP DESIGN

We sketch one possible design of a cooperative measurement system. The system has some features in common with other measurement infrastructures (most notably IDMaps [1] and NIMI [3]). Some similarities and differences are discussed briefly in the Related Work section.

A. The Service

The M-coop system answers queries of the form *(IP1, IP2, measurement type)* where *IP1* and *IP2* are IP addresses. The measurement type may be any network quantity measurable by hosts on a network, e.g., delay, bandwidth, jitter. The system returns the answer to the query along with trust and accuracy parameters if available. As a voluntary peer-to-peer system, the possibility of misinformation is high, so a trust value is reported with the information returned. The trust value is an indication of the past reliability (with respect to quality of information) of the node that responded to the query.

The size of the Internet dictates that any measurement will only be an estimate. To keep the system managable, instead of the measurement being from the requested host, it might be from a "nearby" node on the overlay network. Also, the measurement process may contain some inaccuracies due to changing paths in the Internet, inherent inaccuracies in the measurement process, congestion, etc. The accuracy value tries to quantify the "nearness" of the host to the measurement node as well as the inaccuracies in the measurement process.

We do not address the question of who is allowed to make queries. In the spirit of cooperatives, one might imagine that only participants are allowed access to the community information. This sort of access control (or any other) is orthogonal to the base system design.

B. The Architecture

Architecturally, the system is an overlay network of Internet hosts running M-coop software. Nodes connected by edges in the overlay form measurement peers, hence an important issue is the construction of the overlay graph to support accurate measurements without undue overhead. Measurements are taken by the endpoints of the edge in two ways, actively, by sending probe packets to each other, and passively, by monitoring the system traffic that traverses this edge of the graph.

For scalability, each node on the network is assigned an "area of responsibility" or AOR, defining a set of addresses for which it can answer queries. The AOR is assigned when the node joins the network. It changes as other nodes join and leave the network.

A query to the system, *(IP1, IP2, measurement type)* is first routed to the the node which has *IP1* in its AOR. We denote this node *R(IP1)* to indicate it is responsible for *IP1*. If the measurement information is available, node *R(IP1)* will reply, along with the available accuracy and trust information. If the data is not available, it may trigger a measurement or a new query. This new query, called a composition query, will traverse a path on the overlay from *R(IP1)* to *R(IP2)* collecting metric data about the links traversed. This data is then returned as the reply to the composition query and finally as a reply to the original query.

A node on the system thus consists of three modules:

1. **Routing.** This module is responsible for maintaining the overlay, communicating with the peer nodes and routing queries and responses through the overlay.
2. **Measurement.** This module performs measurements between itself and its measurement peers, verifies the measurements obtained by other nodes and responds to queries about the node's AOR.
3. **Trust.** This module maintains the trust database, performs trust metric calculations and responds to trust queries.

C. Architecture Details and Operation

C.1 Routing

There are many ongoing research efforts trying to develop better methods of locating data in a distributed system. These efforts are directed towards scalability, reliability, graceful degradation under dynamic conditions, and efficient search [4, 5, 8]. Based on this, we assume that a method of locating data on the overlay network is available to us and we will use the generic term *routing* to imply that a packet is using one of the above methods to reach its destination node.

The routing module is responsible for routing data queries and their responses through the overlay based on the AOR of the nodes. It also responsible for maintaining the measurement overlay.

C.2 Measurement

Measurements are taken from the node to its measurement peers. The set of measurement peers is identified when the node joins the system and is updated as needed when peer nodes join or depart. The set of measurement peers is selected to map onto the underlying network topology in the following fashion. If a node in the overlay network is the only one in an autonomous system (AS), then it has one edge for each AS level neighbor in the underly-

ing network. If there are multiple nodes in the same AS, a clustering protocol is run when new nodes join or depart to ensure that there is only one node which has edges to the AS level neighbors. The remaining nodes are organized to provide redundancy and intra-AS measurement. The AS level edges join this node to the overlay network nodes responsible for the IP addresses in the AS level neighbors. The intent is that the measurements obtained by the M-coop system will then better approximate the values seen by a packet on the underlying network. The measurement data obtained is stored with meta-information such as time of measurement, whether it is composed from other measurement data, whether it is cached, the accuracy, trust value, etc. Measurements may be taken periodically, triggered by options in a data query, and/or determined passively by examining packets in the system.

We performed some simple experiments to verify that such an AS-level composition of paths gives reasonable estimates of the original query between two IP addresses. We used data from the UW-3 and UW-4b datasets used for the Detour study[6]. These are end-to-end `traceroute` measurements collected from public `traceroute` servers. Details of the data collection method can be found in the referenced study. For each dataset, we calculate the average latency between two nodes *a* and *b* and also compute the AS-level path between them. The composition path is computed by finding a node *c* which lies on the AS-level path from a to b, and calculating *(a,c)+(c,b)* for all possible such *c*s. This simulates our scheme in which the composition packets are forwarded at the AS-level from the source AOR to the destination AOR. The results are shown in Fig. 1 as a cdf of the ratio of *((a,c)+(c,b))/(a,b)*. It can be observed that most of the composition values are within a factor of two of the actual path value. These results are similar to those obtained by IDMaps [1] and show that forwarding at the relatively coarse level of ASes can give reasonable estimates.

Measurements are of two types, data and verification. Data measurements are performed between a node and its measurement peers and these are reported in response to queries. Verification measurements are performed to verify the responses of the peer nodes. These measurements are part of the process by which the trust module calculates the trust value to be assigned to the peer nodes.

C.3 Trust

The trust value of a node in the context of a particular link is a measure of its reliability in the past. We assume that past behaviour is an indication of future actions, i.e., if a node has been providing reasonable accurate responses about a link in the past, it is likely to provide a reasonably

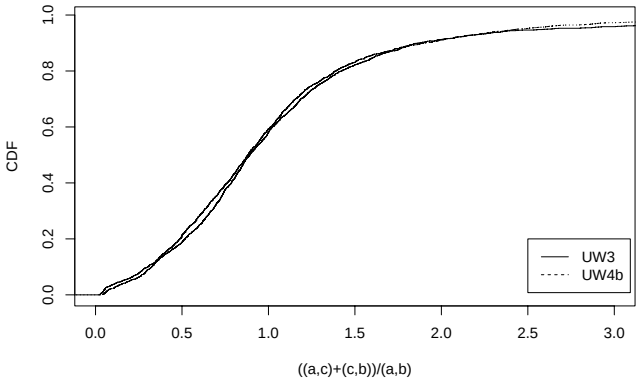

Fig. 1. CDF of composition of AS-level paths

accurate response when queried now. Since trust is used in a specific context, we will use the term “trust of a node” to implicitly mean the trust of a node with respect to a specific link.

The node checks the operation of its peers using a *verification process*, which is run regularly. The results of this process are used to calculate the trust value of a node. This trust factors in the time since the last verification process was run as well as reports from other nodes on the trust of the node in concern. In the system, a node reports on the trust of its immediate neighbours. It also gathers information about nodes two and three hops away which is then reported only if a query about that node’s trust passes through.

We now explain the verification process in more detail. In Fig. 2, node *a* responds to trust queries about the nodes *b*, *c* and *d*. Periodically, *a* runs the verification process on the nodes *b*, *c* and *d*. The verification process is in two parts: *a* queries *b* about the path from *b* to *e*, which is *b*’s neighbour; *a* also performs a measurement from itself to *e* directly. Since *a* knows the value of the measurement of the *a*-*b* link, it can estimate the *b*-*e* link and compare it with the value reported by *b*. The two values thus obtained are then used to update the trust of node *b*.

It is important to note that this measurement has higher chances of being inaccurate and so a single value which doesn’t tally with the reported value may not be enough to affect the trust value of the node *b*.

This verification mechanism requires that the point-to-point measurements made by the node be independent of the trustworthiness of the other end point, and hence reliable. This can be partly achieved if at least some of the measurements can be performed without the cooperation of the other node (perhaps at the operating system

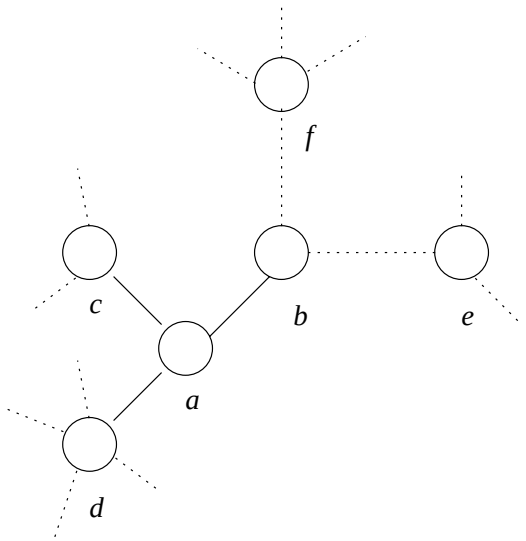

Fig. 2. Verification

level). For example, a ping query to measure latency does not reach the application layer of the other end-point and hence is harder to affect. Given this assumption, *a* can verify the functioning of *b* by making a direct measurement to *e* to estimate the *b*-*e* link. Since *b* is in *a*'s neighbouring AS and similarly, *e* in *b*'s, it is reasonable to suppose that the direct measurement by *a* will produce a good estimate. Another solution could be to have *a* measure directly to other nodes in *b*'s AS, if they exist. These solutions partly address the problem of verification, but do not provide a completeley reliable method of veryfing a neighboring node's measurements.

D. AOR Assignment

The AOR assignment for a node takes place when the node joins the overlay network. The startup procedure for nodes joining the network assumes two things: a node is capable of finding its AS number[1]; and knows the IP address of an existing node on the overlay. (The case of the first node in the overlay is discussed seperately.) A further assumption that is useful, but not required is that a node has access to the list of ASes connected to its AS.

On startup, the node *n* contacts the node *e* which is already a member of the overlay. In its initial message, *n* advertises the entire AS as its AOR. Since the overlay already exists, some node *s* already has the AS in its AOR and so *e* returns the address of *s*, to *n*. The node *n* then contacts *s* with the same advertisement. If *s* is not in the AS, *s* splits its AOR, and relinquishes claim to the AS and *n* now has the AS in its AOR. If *s* is in the same AS as *n*, a clustering protocol is run to enable *n* to locate and peer with intra-AS nodes.

When *n* is the first node in the overlay, it assigns all AS numbers to its AOR and waits for further nodes to join the network.

To establish the measurement peers, *n* queries *s* for the nodes responsible for the neighbouring ASes. *n* then contacts these nodes with a peer request to make them peers on the overlay. This process is simple if a list of connected ASes is available[2]. To maintain the integrity of the overlay, *n* may also peer with other nodes according to the indexing protocol used for the overlay.

RELATED WORK

There have been several prior projects that concern measurement infrastructures (e.g., IDMaps [1], NIMI [3], SPAND [7] and Remos [2]). Our work is most closely to the IDMaps project, so we limit our related work discussion to that project.

IDMaps [1] is a proposal for a global infrastructure for gathering and distributing Internet host distance information. The goal of the IDMaps project is to provide distance metrics between two hosts on the Internet in an accurate and timely manner. The IDMaps architecture consists of a network of Tracers, which gather Internet distance information, and Clients, which use this information to estimate distances between hosts. The distance estimate between any two IP addresses is calculated from the Address Prefixes (APs) that contain the IP addresses, serving a similar function to our Areas of Responsibility. The calculation is performed by finding the APs to which the IP addresses belong, locating the systems or "boxes" to which the APs are closest and then running a spanning-tree algorithm to to find the shortest distance between the two boxes. This calculation requires that a substantial portion of the box connection topology must be maintained.

The actual box-box topology can be achieved in two ways, the Hop-by-Hop (HbH) and the End-to-End (E2E) models. In the HbH model, every transit backbone router is modeled as a box and the calculation is the sum of inter-AS and intra-AS paths from one AP to the other. The distances on these paths are calculated by the Tracers probing the routers at random intervals. In the E2E model, the Tracers are the boxes and the distances are calculated as the sum of the AP to box distances and the distance between the two boxes.

Our goals for the M-coop system are to provide a generalized metric collection and distribution infrastructure that is simple and rapid to deploy on a large scale. The informa-

[1] A repository of AS information is available at www.arin.net/whois.

[2] NLANR maintains such a list at http://moat.nlanr.net/AS/

tion returned by the system also contains some indication of how reliable (in terms of accuracy and trustworthiness) the information is. Our approach to the problem is similar to the HbH model proposed in the IDMaps architecture but our method of distance estimation and information dissemination is fundamentally different. We intend to have a little more complexity at the nodes gathering the distance information to avoid the problem of maintaining a global view of the box topology. We also try to address the deployment of the system in the Internet by means of our peer-to-peer design.

Open Questions

- **Participation.** Will such a scheme generate enough participation to achieve critical mass, i.e., a level where the query results are a good approximation of the actual values? Related to this issue is the broader question of what will motivate people to participate in peer-to-peer systems. Will people eventually subscribe to peer-to-peer systems, like they subscribe to magazines? Or will they contribute their host to peer-to-peer systems, a la charitable donations? What are the best analogies to the peer-to-peer experience?
- **Generality.** Can a single system be used to satisfy the different measurement requirements of the diverse applications which might want to take advantage of this service? Can such a system be used as a common measurement service for peer-to-peer systems to use for optimizing their operation?
- **Usefulness of Parameters.** Can trust and accuracy be made useful to applications?
- **Composition.** Can composition of measurements from intermediate hops give meaningful values for the actual measurement between two IP addresses?
- **Collusion.** Collusion is a problem in trust systems. Can the amount of collusion required to subvert the system be made large enough to deter attacks?

Acknowledgments

The authors would like to acknowledge the helpful suggestions of the anonymous reviewers. We would also like to thank Andy Collins and Stefan Savage for providing the datasets for the experiments.

References

[1] P. Francis, S. Jamin, V. Paxson, L. Zhang, D. Gryniewicz, and Y. Jin. An architecture for a global internet host distance estimation service. In *IEEE/ACM Trans. on Networking*, October 2001.

[2] N. Miller and P. Steenkiste. Collecting network status information for network-aware applications. In *Proceedings of Infocom'00*, Tel Aviv, March 2000.

[3] V. Paxson, J. Mahdavi, A. Adams, and M. Mathis. An architecture for large-scale internet measurement. In *IEEE Communications*, volume 36, pages 48–54, August 1998.

[4] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content-addressable network. In *Proceedings of the ACM SIGCOMM '01*, San Diego, CA, September 2001.

[5] A. Rowstron and P. Druschel. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In *Middleware*, 2001.

[6] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson. The end-to-end effects of internet path selection. In *Proceedings of the ACM SIGCOMM'99*, Boston, MA, September 1999.

[7] S. Seshan, M. Stemm, and R. H. Katz. A network measurement architecture for adaptive applications. In *Proceedings of Infocom '00*, Tel Aviv, March 2000.

[8] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A peer-to-peer lookup service for internet applications. In *Proceedings of the ACM SIGCOMM '01*, San Diego, CA, September 2001.