Implementation and Deployment of a Distributed Network Topology Discovery Algorithm

Benoit Donnet, Bradley Huffaker, Timur Friedman, Kimberly C. Claffy

Implementation and Deployment of a Distributed Network Topology Discovery Algorithm

Benoit Donnet*†, Bradley Huffaker†, Timur Friedman*, kc claffy†
*Université Pierre & Marie Curie – Laboratoire LiP6-CNRS, UMR 7606, Paris, France
email: {benoit.donnet,timur.friedman}@lip6.fr
†CAIDA – San Diego Supercomputer Center, San Diego, USA
email: {benoit,bhuffake,kc}@caida.org

***Abstract*— In the past few years, the network measurement community has been interested in the problem of internet topology discovery using a large number (hundreds or thousands) of measurement monitors. The standard way to obtain information about the internet topology is to use the traceroute tool from a small number of monitors. Recent papers have made the case that increasing the number of monitors will give a more accurate view of the topology. However, scaling up the number of monitors is not a trivial process. Duplication of effort close to the monitors wastes time by reexploring well-known parts of the network, and close to destinations might appear to be a distributed denial-of-service (DDoS) attack as the probes converge from a set of sources towards a given destination. In prior work, authors of this report proposed Doubletree, an algorithm for cooperative topology discovery, that reduces the load on the network, i.e., router IP interfaces and end-hosts, while discovering almost as many nodes and links as standard approaches based on traceroute. This report presents our open-source and freely downloadable implementation of Doubletree in a tool we call traceroute@home. We describe the deployment and validation of traceroute@home on the PlanetLab testbed and we report on the lessons learned from this experience. We discuss how traceroute@home can be developed further and discuss ideas for future improvements.**

I. Introduction

For some time, the problem of internet topology discovery has drawn the attention of the network measurement community. One can see the internet topology at three different levels. The first one, the *IP interface level*, considers internet protocol (IP) interfaces of routers and end systems. Usually, this topology is obtained by using data collected with the probing tool *traceroute* [1]. Traceroute is a networking tool that allows one to discover IP interfaces along the path that data packets take to go from a *source* machine or *monitor* to a *destination* machine. The second level, *the router level*, is an aggregation of the IP level. It can be obtained by using a technique called *alias resolution* [2], [3], [4], [5]. The idea is to summarize all the IP addresses of a router into a single identifier. Finally, the *AS level* provides information about the connectivity of autonomous systems (ASes). An AS is a set of routers that are under the same administrative control. The ASes are interconnected in order to allow IP packets to transit from one network to another, so providing global connectivity. The connectivity information might be inferred from BGP tables [6], BGP being the routing protocol used between ASes. Note that the special problem of determining the topology within a single AS is a separate area of inquiry. It is not, strictly speaking, internet topology discovery, and it is considerably helped by the privileged access available to the administrator of an AS.

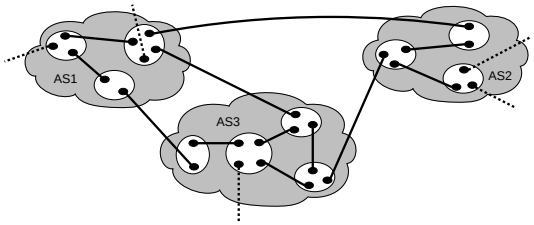


Fig. 1. The different levels of topology

Fig. 1 illustrates the three levels of the internet topology. Black dots represents router interfaces, blank shapes stand for routers and shaded areas for ASes. The plain and dotted lines correspond to links. The IP interface level is illustrated by the links between routers. The router level is obtained when all interfaces of a router are grouped in a single identifier. Finally, the AS level is obtained when we look only at ASes and the links between them.

This report focuses on the IP interface level. More specially, it describes the implementation and deployment story of a topology discovery algorithm, *Doubletree* [7], that reduces load on the network, i.e., router IP interfaces and end-hosts, while discovering nearly the same set of nodes and links as standard approaches based on traceroute. Doubletree was proposed by authors of this report.

Today's most extensive tracing system at the IP interface level, *skitter* [8], uses 24 monitors, each targeting on the order of one million destinations. Authors of this report are responsible for skitter. In the fashion of skitter, *scamper* [9] makes use of several monitors to traceroute IPv6 networks. Other well known systems, such as *RIPE NCC TTM* [10] and *NLANR AMP* [11], each employ a larger set of monitors, on the order of one- to two-hundred, but they avoid probing outside their own network. However, recent work has indicated the need to increase the number of traceroute sources in order to obtain a more complete topology measurement [12], [13]. Indeed, it has been shown that reliance upon a relatively small number of monitors to generate a graph of the internet can

introduce unwanted biases.

One way of rapidly creating a large distributed monitor infrastructure would be to deploy traceroute monitors in an easily downloadable and readily usable piece of software, such as a screensaver. This was first proposed by Jörg Nonnenmacher, as reported by Cheswick et al. [14]. Such a suggestion is in keeping with the spirit of of that have arisen in the past few years. The most famous one is probably SETI@home [15]. SETI@home's screensaver downloads and analyzes radio-telescope data. Others similar projects are Folding@home [16], a computation application that studies protein folding, and, distributed.net [17] a general-purpose distributed computing project. The first publicly downloadable distributed route tracing tool is *DIMES* [18], released as a daemon in September 2004. At the time of writing this report, DIMES counts more than 6,000 agents scattered over five continents.

However, building such a large structure leads to potential scaling issues: the quantity of probes launched might consume undue network resources and the probes sent from many vantage points might appear as a distributed denial-of-service (DDoS) attack to end-hosts. These problems were quantified in our prior work [7]. There are two ways to avoid these problems: the first one is to stay small, as skitter does for instance, but this solution is opposed to the basic idea of scaling up the number of tracing monitors. The second is to trace slowly, as does DIMES. In this case, the problem is that the resulting network snapshot may be blurred bye the routing changes that take place over the course of a probing interval.

The Doubletree algorithm [7] is a first attempt to perform large-scale topology discovery efficiently and in a network friendly manner. Doubletree acts to avoid retracing the same routes in the internet by taking advantage of the tree-like structure of routes fanning out from a source or converging on a destination. The key to Doubletree is that monitors share information regarding the paths that they have explored. If one monitor has already probed a given path to a destination then another monitor should avoid that path. Probing in this manner can significantly reduce load on routers and destinations while maintaining high node and link coverage [7]. By avoiding redundancy, not only is Doubletree able to reduce the load on the network but it also allows one to probe the network more frequently. This makes it possible to better capture network dynamicity (routing changes, load balancing) compared to standard approaches based on traceroute.

This report goes beyond earlier theory and simulation to propose a Doubletree implementation written in Java [19]. We call this prototype *traceroute@home*. traceroute@home is certainly not the first tracerouting tool developed. However, it differs from standard approaches, such as skitter, due to its distributed aspect and its scaling resistance thanks to Doubletree. traceroute@home is easily tunable and extensible for further developments. traceroute@home is completely open-source and freely downloadable [20].

To validate traceroute@home, we deployed it on the PlanetLab [21] testbed. We installed traceroute@home on ten PlanetLab nodes and probed the network, using 200 other PlanetLab nodes as destinations.

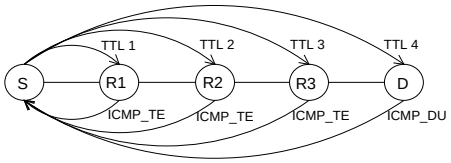


Fig. 2. Traceroute example

In this report, we describe the traceroute@home system. We evaluate the performance of our prototype, point out its weaknesses and the problems encountered in its deployment. We discuss directions for future development of our tool and the opportunity for creating an infrastructure entirely dedicated to network measurement.

The remainder of this report is organized as follows: Sec. II explains how traceroute works; Sec. III describes the Doubletree algorithm and some of its extensions; Sec. IV describes traceroute@home; in Sec. V, we discuss its on PlanetLab nodes; Sec. VI introduces directions for future extensions; finally, Sec. VII summarizes the principal contributions of this report.

II. TRACEROUTE

Traceroute is a networking tool that allows one to discover the path a data packet takes to go from a machine S (the *source* or the *monitor*) to a machine D (the *destination*).

Fig. 2 illustrates how traceroute works. `S` is the source of the traceroute, `D` is the destination and the `Ri`s are the routers along the path. S sends multiple UDP probes, UDP being the *User Datagram Protocol* which is a connectionless transport protocol, into the network with increasing *time-to-live* (TTL) values, the TTL being a field in the IP header indicating how long a packet can circulate in the network. Each time a packet enters a router, the router decrements the TTL. When the TTL value is one, the router determines that the packet has consumed sufficient resources in the network, drops it, and informs the source of the packet by sending back an ICMP *time exceeded* message (`ICMP_TE` in Fig. 2). ICMP, Internet Control Message Protocol, is a protocol for managing errors related to networked machines. By looking at the IP source address of the ICMP message, the monitor can learn the IP address of the router at which the probe packet stopped.

When, finally, a probe reaches the destination, the destination is supposed to reply with an ICMP *destination unreachable* message (`ICMP_DU` in Fig. 2).

Unfortunately, the traceroute behavior explained above is the ideal case. A router along the path might not reply to probes because the ICMP protocol is not activated, or the router is overloaded. In order to avoid waiting an infinite time for the ICMP reply, the traceroute monitor actives a timer when it launches the UDP probe. If the timer expires and no reply was received, then, for that TTL, the machine is considered as *non-responding*.

However, a particular problem occurs when it is the destination that does not reply to probes because, for instance, of a

restrictive firewall. In this case, the destination will be recorded as non responding but it is impossible to know that it was reached. In order to avoid inferring a boundless path, an upper bound on the number of successive non-responding machines is used. For instance, in skitter and in our application, this upper bound is set to five.

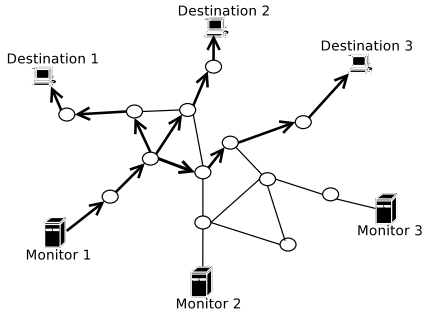
Standard traceroute, as just described, is based on UDP probes. However, two variants exist. The behavior of the traceroute for the intermediate routers is the same as standard traceroute. The difference comes with the destination. The first variant is based on ICMP. Instead of launching UDP probes, the source sends ICMP *Echo Request* messages. The destination is supposed to reply with an ICMP *Echo Reply*. The second sends packets using the *Transport Control Protocol* (TCP) which is a connection-oriented transport protocol. The TCP traceroute aims to bypass most common firewall filters by sending TCP SYN packets. It assumes that firewalls will permit inbound TCP packets to specific ports listening for incoming connections.

III. DOUBLETREE

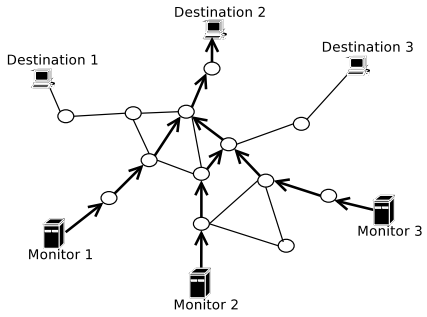
Doubletree [7] is the key component of a coordinated probing system that significantly reduces load on routers and end-hosts while discovering nearly the same set of nodes and links as standard approaches based on traceroute. It takes advantage of the tree-like structures of routes in the context of probing. Routes leading out from a monitor towards multiple destinations form a tree-like structure rooted at the monitor (see Fig. 3(a)). Similarly, routes converging towards a destination from multiple monitors form a tree-like structure, but rooted at the destination (see Fig. 3(b)). A monitor probes hop by hop so long as it encounters previously unknown interfaces. However, once it encounters a known interface, it stops, assuming that it has touched a tree and the rest of the path to the root is also known. Using these trees suggests two different probing schemes: backwards (monitor-rooted tree) and forwards (destination-rooted tree).

For both backwards and forwards probing, Doubletree uses stop sets. The one for backwards probing, called the *local stop set*, consists of all interfaces already seen by that monitor. Forwards probing uses the *global stop set* of $(\text{interface}, \text{destination})$ pairs accumulated from all monitors. A pair enters the stop set if a monitor received a packet from the interface in reply to a probe sent towards the destination address.

A monitor that implements Doubletree starts probing for a destination at some number of hops h from itself. It will probe forwards at $h+1$, $h+2$, etc., adding to the global stop set at each hop, until it encounters either the destination or a member of the global stop set. It will then probe backwards at $h-1$, $h-2$, etc., adding to both the local and global stop sets at each hop, until it either has reached a distance of one hop or it encounters a member of the local stop set. It then proceeds to probe for the next destination. When it has completed probing for all destinations, the global stop set is communicated to the next monitor.



(a) Monitor-rooted



(b) Destination-rooted

Fig. 3. Tree-like routing structures

Doubletree has one tunable parameter. The choice of initial probing distance h is crucial. Too close, and duplication of effort will approach the high levels seen by classic forwards probing techniques [7, Sec. 2]. Too far, and there will be high risk of traffic looking like a DDoS attack for destinations. The choice must be guided primarily by this latter consideration to avoid having probing look like a DDoS attack.

While Doubletree largely limits redundancy on destinations once hop-by-hop probing is underway, its global stop set cannot prevent the initial probe from reaching a destination if h is set too high. Therefore, each monitor sets its own value for h in terms of the probability p that a probe sent h hops towards a randomly selected destination will actually hit that destination. Fig. III shows the cumulative mass function for this probability for skitter monitor `apan-jp`. If one considers as reasonable a 0.2 probability of hitting a responding destination on the first probe, it must chose $h \leq 12$.

Simulation results [7, Sec. 3.2] show for a range of p values that, compared to classic probing, Doubletree is able to reduce measurement load by approximately 70% while maintaining interface and link coverage above 90%.

However, one possible obstacle to successful deployment of Doubletree concerns the communication overhead from sharing the global stop set among monitors. Tracing from 24 monitors, i.e., the same quantity of probing monitors as

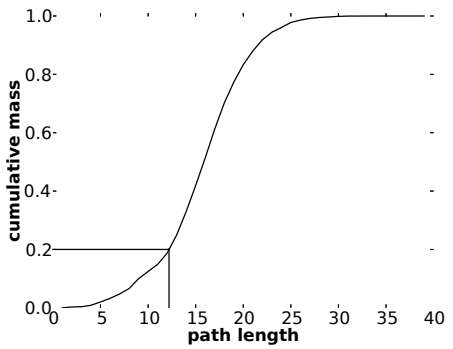


Fig. 4. Cumulative mass plot of path lengths from skitter monitor `apan-jp`

skitter, to a relatively small set of just 50,000 destinations with $p = 0.05$ produces a set of 2.7 million $(\text{interface}, \text{destination})$ pairs. As pairs of IPv4 addresses are 64 bits long, an uncompressed stop set based on these parameters requires 20.6 MB.

A way to reduce this communication overhead is to use *Bloom filters* [22] to implement the global stop set. A Bloom filter [23] summarizes information concerning a set in a bit vector that can then be tested for set membership. An empty Bloom filter is a vector of all zeroes. A key is registered in the filter by hashing it to a position in the vector and setting the bit at that position to one. Multiple hash functions may be used, setting several bits to one. Membership of a key in the filter is tested by checking if all hash positions are set to one. A Bloom filter will never falsely return a negative result for set membership. It might, however, return a false positive. For a given number of keys, the larger the Bloom filter, the less likely is a false positive. The number of hash functions also plays a role.

In prior work [22], we have shown that, when $p = 0.05$, using a bit vector of size 10^7 and five hash functions allows nearly the same coverage level as a list implementation of the global stop set while slightly reducing the redundancy on both destinations and internal interfaces and yielding a compression factor of 17.3.

Donnet and Friedman, co-authors of this report, also proposed an enhancement to the forwards stopping rule based on Classless Inter-Domain Routing (CIDR) address prefixes [24]. The idea is to aggregate the destinations set by recording the CIDR address prefixes of destinations rather than the full IP address. This allows one to reduce the amount of communication required by Doubletree. Instead of sharing a set of $(\text{interface}, \text{destination})$ pairs, monitors will share a set of $(\text{interface}, \text{prefix\_destination})$ pairs. The shortest the prefix, the more destinations can be represented by a single entry, but also the more likely the entry will generate false positives. With this simple mechanism, load on destinations can be further reduced while maintaining the coverage accuracy around 90%. When combined with a Bloom filter, one further reduces the global stop set size, providing a compression factor of 57.1.

IV. TRACEROUTE@HOME IMPLEMENTATION

This section describes the traceroute@home implementation. Sec. IV-A presents our design choices. Sec. IV-B introduces the macroscopic functioning of the cooperative system that makes use of the Doubletree algorithm. Sec. IV-C takes a microscopic look at traceroute@home by explaining the behavior of a monitor and each module composing it. Sec. IV-D discusses the general messages framework.

A. Design Choices

We implemented traceroute@home in Java [19]. We choose Java as the development language because of two reasons: the large quantity of available packages and the possibility of abstracting ourselves from technical details. As a consequence, the development time was strongly reduced. Unfortunately, Sun does not provide any package for accessing packet headers and handling raw sockets, which is necessary to implement traceroute. Instead of developing our own raw sockets library, we used the open-source *JSocket Wrench* library [25]. We modified the JSocket Wrench library in order to support multithreading. Our modifications are freely available [20].

We aimed for the design of traceroute@home to be easily extended in the future by ourselves but also by the networking community. For instance, concerning the messages exchanged by monitors, we define a general framework for messages, making creation and handling of new messages easier. In addition to that, traceroute@home is readily tunable due to a configuration file that is loaded by the application at its starting. Our implementation is freely available [20].

We designed our application by considering two levels: the *microscopic* level and the *macroscopic* level.

From a macroscopic point of view, i.e., all the monitors together, the monitors are organized in a ring, adopting a round robin process. At a given time, each monitor focuses on its own part of the destination list. When it finishes probing its part, it sends information to the next monitor and waits for data from the previous one, if it was not yet received. Sec. IV-B explains this macroscopic aspect of traceroute@home.

From a microscopic point of view of our implementation, i.e., a single monitor, a monitor is composed of several modules that interact with each other. Our implementation is thread-safe, as a monitor is able to send several probes at the same time. Further, topological information collected by a monitor is regularly saved to XML files. Sec. IV-C explains this microscopic level of traceroute@home.

B. System Overview

The simulations conducted in prior work [7] were based on a simple probing system: each monitor in turn covers the destination list, adds to the global stop set the $(\text{interface}, \text{destination})$ pairs that it encounters, and passes the set to the subsequent monitor.

This simple scenario is not suitable in practice: it is too slow, as an iterative approach allows only one monitor to probe the network at a given time. We want all the monitors probing in

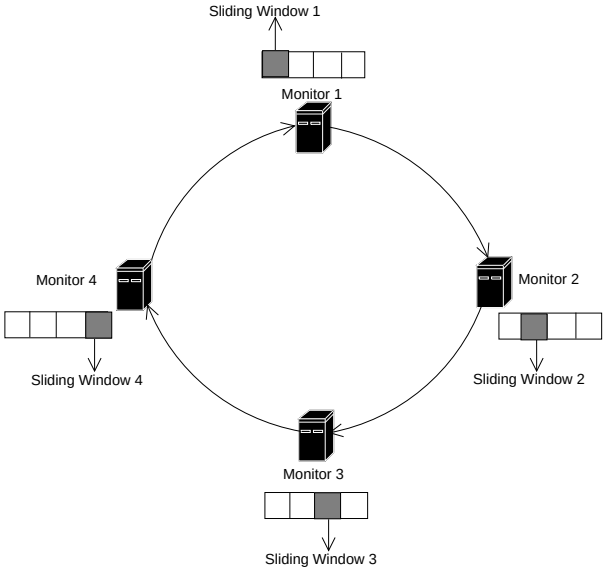


Fig. 5. Doubletree with sliding windows

parallel. However, how would one manage the global stop set if it were being updated by all the monitors at the same time?

An easy way to parallelize is to deploy several *sliding windows* that slide along the different portions of the destination list. At a given time, a given monitor focuses on its own window, as shown in Fig. 5. There is no collision between monitors, in the sense that each one is filling in its own part of the global stop set. The entire system counts m different sliding windows, where m is the number of Doubletree monitors. If there are n destinations, each window is of size $w = n/m$. This is an upper-bound on the window size as the concept still applies if they are smaller.

A sliding window mechanism requires us to decide on a step size by which to advance the window. We could use a step size of a single destination. After probing that destination, a Doubletree monitor sends a small set of pairs corresponding to that destination to the next monitor, as its contribution to the global stop set. It advances its window past this destination, and proceeds to the next destination. Clearly, though, a step size of one will be costly in terms of communication. Packet headers (see Sec. IV-D for details about packet format) will not be amortized over a large payload, and the payload itself, consisting of a small set, will not be as susceptible to compression as a larger set would be.

On the other hand, a step size equal to the size of the window itself poses other risks. Suppose a monitor has completed probing each destination in its window, and has sent the resulting subset of the global stop set on to the following monitor. It then might be in a situation where it must wait for the prior monitor to terminate its window before it can do any further useful work.

A compromise must be reached, between lowering communications costs and continuously supplying each monitor with useful work. This implies a step size somewhere between 1 and w. For our implementation of Doubletree, we let the user decide the step size. This is a part of the XML configuration file that each Doubletree monitor loads at its start-up (see Sec. IV-C). Future work might reveal information about how to tune the step size of a monitor.

C. Inside a traceroute@home Monitor

Fig. 6 shows the different modules composing a traceroute@home monitor and the way they interact with each other (the arrows) and with their environment, i.e., hard disk or network (circles).

First, a traceroute@home monitor loads an XML file (*config.xml*) that contains configuration information, such as the number of sliding windows and the name (or IP address) of the next monitor in the round robin process. This XML file must strictly follow a DTD (Document Type Definition).[1] In the current version of the application, the XML file must be present on the machine running the software. In the following versions (See Sec. VI), we can imagine that a traceroute@home monitor will upload this file from a remote server. Note also that the destination list and sliding window information must also be present on the machine. We can envisage that these will also be remotely available in the future.

This configuration file is processed by the *Agent*. The Agent is the heart of a traceroute@home monitor as it first creates all other modules, manages them and, finally, allows the various components to interact with each other.

The Agent creates the *StopSet* module that implements the stop set data structure [7]. Our stop set implementation is multithread safe. Two types of implementations are proposed: list and Bloom filter. The list is the basic implementation and can be used for the local stop set as well as the global stop set. The Bloom filter implementation [22] may only be used for the global stop set. The hash functions needed by the Bloom filter are emulated with the SHA-1 algorithm [26]. Depending on the XML configuration file, the global stop set may be compressed or not before being sent to the next monitor. Note that, for consistency reasons, each monitor in the system must use the same implementation for the global stop set.

Doubletree is a cooperative algorithm. The different monitors have to share their global stop set. They thus exchange messages. The purpose of the *Message* component is therefore to build, parse and store messages (before their handling by the Agent) sent and received by a monitor. We explain in detail the Message component by describing our general message framework in Sec. IV-D.

A traceroute@home monitor, through the Message module, is able to handle messages. To send and receive them, it makes use of the *Communication* component that allows a monitor to interact with other monitors. It sends messages to a given monitor when the Agent orders it and listens to potential connections for incoming messages. Incoming messages are parsed and stored by the Message module before their handling by the Agent. In order to make this module as efficient as

[1]For interested readers, the DTD is available online [20].

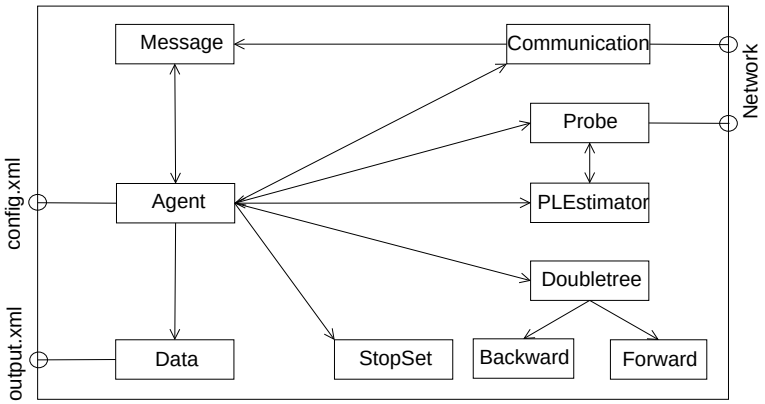


Fig. 6. A traceroute@home monitor's modules

possible, it was implemented using a *selector*. A selector provides the ability to do readiness selection, which enables multiplexed I/O operations. A selector makes it possible for a single thread to manage many I/O channels simultaneously. This corresponds to the `select()` operation in C.

The *Probe* module builds probes (UDP or ICMP), assigns them a TTL value, sends them into the network and waits for eventual ICMP replies. The Probe module is created by a client that wants to probe the network. An example of such a client is the Agent. The client may require multithreading in its network exploration. Therefore, the Probe module allows one to send multiple probes at the same time. Another unique thread listens to incoming ICMP replies messages and dispatches them to the client. The client is in charge of the matching between a probe sent and the eventual ICMP reply. This matching is possible because the ICMP reply (`destination unreachable` or `time exceeded`) contains the IP header and the 8 first bytes of the original datagram (refer to Sec. II for details about how traceroute works). Placing a unique source port number in the UDP header of each outgoing probe allows the returning ICMP replies to be identified. Moors discusses the reasons for varying the source port of UDP datagram instead of the destination port [27, Sec. III].

The *PLEstimator* (for "Path Length Estimator") module is in charge of discovering path lengths for the current sliding window. It sends to destinations UDP probes with a TTL of 64 and a high destination port. If a destination replies with a `Destination Unreachable` ICMP message, the PLEstimator is able to know the distance by looking at the TTL field contained in the IP header copied in the ICMP message payload. A simple subtraction allows one to know the path length. With all the path length information received, the PLEstimator builds the path length CDF, as shown in Fig. III. As suggested in prior work [7], the PLEstimator will use, by default, a p value of 0.05 in order to determine the h value that must be used for the current sliding window. The PLEstimator is a client of the Probe module.

The *Doubletree* module defines an interface describing the general behavior of a probing scheme. This interface is implemented in two ways, such that it defines the two probing scheme behaviors, i.e., backwards and forwards, proposed by prior work [7]. However, the Doubletree module is not a Probe module client. In this case, multithreading is managed by the Agent. The Doubletree module is only used to decide, according to a probe reply (which might be empty if the router did not reply to probes), if a stopping condition is reached. Four stopping conditions are defined: *i)* normal (the destination, forwards probing, or the first hop, backwards probing, is reached), *ii)* stop set, *iii)* loop and *iv)* gap. A gap occurs when the monitor encounters five successive non-responding interfaces. The Doubletree module is also in charge of deciding the next TTL value.

When a traceroute@home monitor has finished probing a part of the sliding window corresponding to the step size, it sends the corresponding stop set to the subsequent monitor but it also saves the topological information that it has gathered. Maintaining information about the topology discovered and saving this information on the hard disk is the *Data* module's job. The topological information is transformed into an XML file according to a DTD.[2] The XML format was chosen in order to facilitate eventual later data migration into a more complex data type. An XML format also makes the data handling easier. For each destination probed, we currently record the following information: the source (i.e., the IP address of the monitor), the destination, a timestamp, the backwards probing stopping reason as described above, the backwards probing stopping distance, the forwards probing stopping reason, the forwards probing stopping distance, and the path discovered. The path is composed of several hops, a hop being composed of the TTL value and three IP addresses or "*", corresponding to a non-responding interface. If the IP address corresponds to a responding router, then the round-trip time (RTT) for the probe and the response is added.

D. Message Exchange

Monitors in the system have to share information about what was previously discovered. This is achieved by regularly

[2]Interested readers may find it online [20].

<msg> = <header><payload>$^{?}$
<header> = <length><type>
<length> = <integer$_{16}$>
<type> = <stopset>
<stopset> = “0”
<integer$_{16}$> = “4” | ...| “65535”

Fig. 7. ABNF for a message

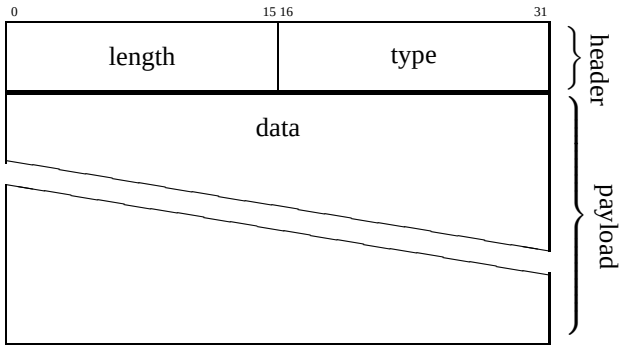


Fig. 8. General packet format

exchanging their global stop set.

A trivial implementation could have a monitor simply send its global stop set as a byte stream to another monitor, without any additional information. However, such a simple mechanism would make further extensions difficult, specially in the case of a peer-to-peer or overlay system used to manage the whole system (see Sec. VI).

For that reason, we provide a general framework for messages exchanged between monitors. This framework might be easily extended in order to manage any type of messages.

In this section, we describe the message framework we propose and a specific message that is actually the only one currently implemented: STOPSET. Both will be first described in Augmented Backus-Naur Form (ABNF) [28] and second as a byte stream.

A message in a traceroute@home system is composed of the two parts: the *header* and the *payload*. The header is mandatory and gives information about the message length and its type. As opposed to the header, the payload is optional. Some messages, for instance a heartbeat-like message, do not need any payload. The ABNF for a message is given in Fig. 7.

Fig. 8 provides a byte-stream vision of a message. Length and type are expressed as 16 bit integers. Note that the length includes both the header and the payload. The minimum length of a message is thus four bytes, i.e., the header length. The type and length encoding is big-endian.

In the current traceroute@home implementation, a monitor sends and receives only one type of message: the STOPSET. This message contains information about the topology discovered by a monitor. A STOPSET message is sent when, for a given sliding window, a step size is reached. Two pieces of information (number of the sliding window and the step

<payload> = <window><slice><impl>
<stopset>
<window> = <integer$_{8}$>
<slice> = <integer$_{8}$>
<impl> = <stype><ip><compress>
<stype> = <bit>
<ip> = <bit>
<compress> = <bit>
<stopset> = <byte>$^{+}$
<integer$_{8}$> = “1” | ...| “32767”
<byte> = <bit>*8
<bit> = “0” | “1”

Fig. 9. ABNF for the STOPSET message

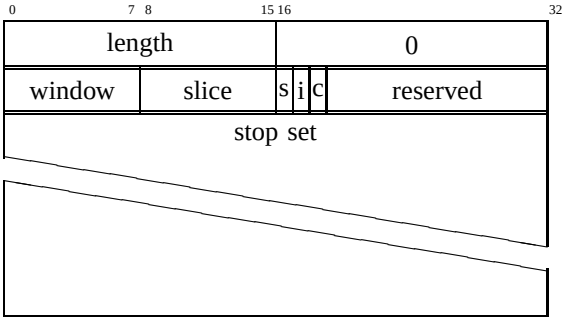


Fig. 10. STOPSET packet format

size considered) must be present in the STOPSET message so that the receiver might identify the message and link it to a portion of a sliding window. In the ABNF, the step size is called a *slice*. A STOPSET message must provide information on implementation details: the type of stop set (the *stype*, list or Bloom filter), the type of network probed (the *ip*, IPv4 or IPv6)[3] and the eventual compression (the *compress*) of the stop set. Of course, it also contains the relevant portion of the global stop set. The STOPSET payload in an ABNF format is given by Fig. 9.

Fig. 10 shows the STOPSET message as a byte stream. We see that 13 bits are reserved for future extensions. Padding is used to fill in this unused part of the packet. The stop set is encoded as a byte array. If the stop set is implemented as a list, a group of 8 bytes refers to a global stop set key. The first four bytes represent the interface address and the last four bytes represent the destination address. The principle is identical modulo group size in case of IPv6 addresses.

V. DEPLOYMENT STORY

This section talks about the deployment and validation of traceroute@home on PlanetLab nodes, the different difficulties we encountered (not necessarily related to PlanetLab) and the way we solved them (Sec. V-A). Sec. V-B presents our deployment results.

[3]Currently, only IPv4 is implemented.

Usability	Number
offline	120
unreachable	69
broken	0
ok	448
total	637

TABLE I
PLANETLAB NODES AVAILABILITY, DECEMBER 2005

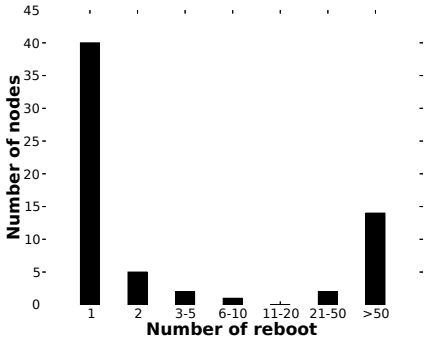


Fig. 11. PlanetLab nodes reboot, one week (December 2005)

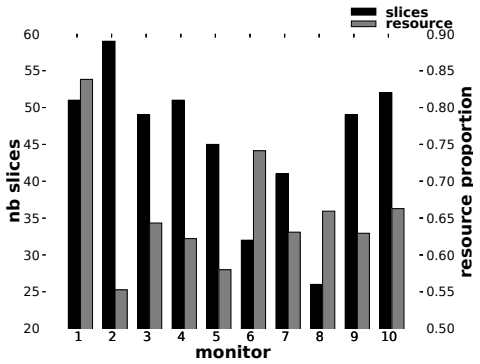


Fig. 12. traceroute@home monitors evaluation (December 20th 2005)

A. Difficulties Encountered

Some routers along the path may be poorly configured. It seems that they, when building the ICMP message, can modify the original datagram. Several ICMP messages were returned with the source and destination ports changed in the original datagram. This is a critical issue as the source port of the originating UDP datagram is different for each datagram, as explained in Sec. IV-C, in order to identify the thread that sends the datagram. This problem can be avoided by also checking the destination address in the original IP header.

By definition, a PlanetLab node is minimalist in the sense that it provides a nearly empty file system. The only environments provided consist of Perl (version 5.8.3) and Python (version 2.3.3). By default, there are no compilation possibilities (no make, no gcc, no g++) and no Java environment. We had to install a Java runtime environment, on each nodes supposed to run traceroute@home.

Table I describes the availability of PlanetLab nodes in December, 2005. *Offline* nodes are those that are either being installed or having long term issues. When gathering these statistics, 18.9% of the PlanetLab nodes were offline. *unreachable* nodes are in production, i.e., the PlanetLab system is running, but not reachable via SSH. 10.9% of the nodes were unreachable. *broken* nodes are those that have failed tests but that can be logged into via SSH as root. No PlanetLab nodes were broken. *ok* nodes are those that can be used normally. 70.2% of the PlanetLab nodes were up but it was important for us to check availability before running any experiment.

PlanetLab nodes reboot periodically for scheduled maintenance and upgrades. Fig. 11 shows the number of reboot during one week in December 2005. 64 PlanetLab nodes were involved in rebooting, for a total of 18,577 reboots. Among these 64 nodes, 40 rebooted only once. However, 14 nodes rebooted more than 50 times. Among these 14 nodes, 10 nodes rebooted more than 1,000 times during the one week period and one node rebooted 3,144 times. As it is not clear how to know which PlanetLab node will reboot and when, a long-term experiment must allow for the possibility that a portion of the nodes will reboot.

Fig. 12 evaluates the performances of the ten PlanetLab nodes we used as traceroute@home monitors (see Sec. V-B for details about the traceroute@home monitors). The performance statistics were gathered on December 20th, 2005. The horizontal axis shows the PlanetLab nodes we used as traceroute@home monitors. The left-side vertical axis gives the number of active slices on each monitor (black bar). A slice is the PlanetLab term for an account. The right-side vertical axis gives the proportion of network resources used by the most consuming slice on each monitor (grey bar). By network resources, we understand the quantity of information sent and received.

Regarding first the quantity of slices per monitor, we see that the maximum, 59, is reached with the monitor labeled 2. The minimum is 32 for the monitor 6. On average, a PlanetLab node chosen for being a traceroute@home monitor hosted 45 slices. These statistics give us an idea of a PlanetLab node load, as each PlanetLab node is supposed to affect resources to a slice in a best effort way.

The right-hand of the vertical axis informs us of the proportion of network resources used, on each monitor, by the most active slice. It oscillates between 0.5526 (monitor 2) and 0.8383 (monitor 1). On average, 0.65601 of the network resources are used by a single slice.

B. Deployment Success

As described in prior work [7], security concerns are paramount in large-scale active probing. It is important to not trigger alarms inside the network with Doubletree probes. It is also important to avoid burdening the network and the destination hosts. It follows from this that the deployment of a cooperative active probing tool must be done carefully, proceeding step by step, from an initial small size, up to larger-

scales. Note that this behavior is strongly recommended by PlanetLab [29, Pg. 5].

Our application was deployed to only ten PlanetLab nodes. We selected ten nodes based on their relatively high stability (i.e., remaining up and connected), and their relatively low load. These traceroute@home monitors are scattered around the world: North America (USA, Canada), Europe (France, Spain, Switzerland, Spain), and Asia (Japan, Korea). In the future, we will wish to scale up the number of monitors to, at least, the skitter scale (i.e., 24 monitors).

The destination list consists of $n = 200$ PlanetLab nodes randomly chosen amongst the approximately 300 institutions that currently host PlanetLab nodes. Restricting ourselves to PlanetLab nodes destinations was motivated by security concerns. By avoiding tracing outside the PlanetLab network, we avoid disturbing end-systems that do not welcome probe traffic. None of the ten PlanetLab monitors (or other nodes located at the same place) belongs to this destination list. The sliding window size of $w = n/m$ consists of twenty destinations. We consider two step sizes (i.e., slices) by window, so each slice counts ten destinations.

Finally, each traceroute@home monitor was configured as follows: the probability p was set to 0.05, the global stop set implementation was the list (i.e., the standard implementation) and no compression was required before sending the STOPSET messages.

The experiment was run on the PlanetLab nodes on Dec. 20th 2005. All the traceroute@home monitors were started at the same time. The experiment was finished when each monitor had probed the entire destination list.

A total of 2,703 links and 2,232 nodes were discovered. We also encountered 2,434 non-responding interfaces (routers and destinations). We recorded 36 invalid addresses. Invalid addresses are, for example, private addresses [7, Sec. 2.1].

Table II shows the different reasons for stopping backwards and forwards probing for each traceroute@home monitor. It further indicates the h value computed by each monitor. The last row of the table indicates the mean for each column.

Looking first at the backwards stopping reasons, we see that the stop set rule strongly dominates (98.6% on average). On average, normal stopping (i.e., reaching the first hop) occurs only 0.65% of the time.

Fig. 13 shows the stopping distance (in terms of hops), for a given monitor, `Uoregon`, when probing backwards and forwards. The vertical line indicates the h value computed by `Uoregon`. Results presented in Fig. 13 are typical for the other traceroute@home monitors.

We see that more than 90% of the backwards stopping occurs at a distance of 5, that is to say the distance corresponding to $h - 1$. In 2.5% of the cases, the probing stops between hop 1 (normal stopping) and hop 4. Except for hop 1, the other stops are caused by the stop set, probably due to very short paths. They illustrate the cases in which the first probe sent with a TTL of h directly hits a destination.

Looking now at the forwards stopping reasons in Table II, we see that the gap rule (five successive non responding interfaces) plays a greater rule. We believe that these gaps occur when a destination does not respond to probes because of a restrictive firewall or because the PlanetLab node is down.

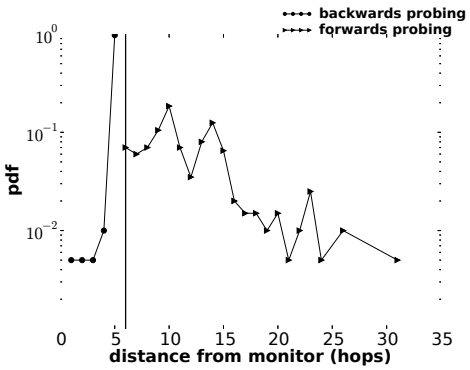


Fig. 13. Stopping distance for the `Uoregon` monitor

monitor	size
`Blast`	12.31
`Cornell`	8.77
`Ethz`	7.41
`Inria`	7.85
`Kaist`	11.41
`Nbgisp`	12.84
`Paris`	10.96
`UCSD`	10.44
`Uoregon`	11.62
`Upc`	10.29
mean	10.39

TABLE III
TOTAL STOPSET MESSAGE SIZE (IN KB) PER MONITOR

On average, in 58% of the cases, the stop set rule applies, and in 28.2% of the cases, the normal rule applies. The normal rule proportion might be seen as high but we have to keep in mind that a Doubletree monitor starts with an empty stop set. Therefore, during the first sliding window, the only thing that can stop a monitor, aside from the gap rule, is an encounter with the destination.

Looking at the stopping distance in Fig. 13, we see that the distances are more scattered for forwards probing than for backwards probing. Regarding the forwards probing, a peak is reached at a distance of 10 (18.5% of the cases). In 7% of the cases, the monitor stops probing at a distance of 6, that is equal to the value h. It could correspond to the stop set rule application or the normal rule, by definition of p. Recall that p defines the probability of hitting a destination with the probe sent with a TTL equals to h. For our experiment, we set $p = 0.05$, meaning that in 5% of the cases the first probe sent by a monitor will hit a destination.

Table III shows the total size of STOPSET messages (in KB) sent by each monitor. The size takes into account the header of the message (4 bytes) and the payload.

A STOPSET message is sent by a monitor when it reaches a step size (i.e. a slice) in the current sliding window. As we define for our experiment two step sizes per sliding window

	Backwards				Forwards				
monitor	loop	gap	stop set	normal	loop	gap	stop set	normal	**h**
Blast	0	0	99.5	0.5	2	17	50	31	7
Cornell	0	0	99	1	0	13.5	69.5	17	7
Ethz	1	0	98.5	0.5	2	10.5	52	35.5	11
Inria	1.5	0	97.5	1	1	4	67	28	15
Kaist	0	0	99	1	0.5	10.5	64.5	24.5	9
Nbgisp	0.5	4	95	0.5	3.5	30.5	22	44	7
LiP6	0	0	99.5	0.5	1	9.5	62.5	27	11
UCSD	0	0	99.5	0.5	0	10.5	60.5	29	7
Uoregon	0	0	99.5	0.5	0	7	74.5	18.5	6
Upc	0.5	0	99	0.5	1	14	57.5	27.5	15
mean	0.35	0.4	98.6	0.65	0.11	12.7	58	28.2	9

TABLE II

STOPPING REASONS (IN %) AND h VALUE PER MONITOR

	time		
monitor	**total**	**probing**	**waiting**
Blast	24	23	1
Cornell	28	15	13
Ethz	20	12.5	7.5
Inria	32	12.5	19.5
Kaist	23	13	10
Nbgisp	26	10	16
LiP6	21	9	12
UCSD	22	13.5	8.5
Uoregon	31	31	0
Upc	27	18.5	8.5

TABLE IV

RUNNING TIME (IN MINUTES)

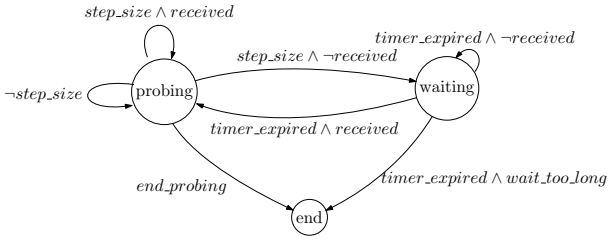


Fig. 14. Probing/waiting state interactions

and as we deploy our prototype on ten PlanetLab nodes, each monitor sent 20 STOPSET messages. We tune each Doubletree monitor in order to use the list implementation of the stopset.

The monitors do not exchange their entire stop set. They only send an update that contains the $(\text{interface}, \text{destination})$ pairs discovered during the current step size probing.

In Table III, we can see that a monitor sends a total of between 7.41 KB (`Ethz`) and 12.84 KB (`Nbgisp`) to the subsequent monitor. On average, a monitor sends 10.39 KB of stop set information into the network.

During our experimentation, the traceroute@home application did not flood the network with STOPSET messages. However, our prior work [22] has shown, on a larger destination list, that it can grow to excessive sizes. In this case, we recommend configuring a traceroute@home monitor to first use the Bloom filter implementation of the stop set and second compress it before sending it in the network.

Table IV shows, for each traceroute@home monitor, the running time (in minutes) in terms of probing and waiting. The waiting period occurs when a monitor has finished its sliding window or a slice in a given sliding window and is waiting for the global stop set that should be sent by the previous monitor in the round-robin topology. We see that nearly all monitors have to wait. A waiting period, in our implementation, lasts 30 seconds. When the timer expires, the monitor checks if it received a new message. If so, the waiting period ends and a new probing period begins. Otherwise, it sleeps during 30 seconds. To avoid infinite waiting, if after 40 sleeping periods (i.e., 20 minutes), nothing was received, the monitor quits with an error message. Fig. 14 illustrates the interactions between the probing state and the waiting state.

We believe that these long waiting periods are due to a characteristic of the PlanetLab IP stack. It seems that when ICMP replies are received by the stack, the `recvfrom`() function does not read them immediately. As the timer set on the listening socket never expires in this case, we think that the `recvfrom`() function is waiting for the permission to access the IP stack. It looks like the resource is owned (or locked) by another process on the PlanetLab node. Note that this behavior was also noticed by other Planet-Lab users [30].

VI. FURTHER WORK

This section discusses possible extensions and ways to improve traceroute@home (Sec. VI-A). It also discusses some key points to enhance measurement infrastructures in general (Sec. VI-B).

A. traceroute@home

One of the main aspects we would like to address in the near future is the stability of the whole system. Currently, it is like dominos: when a monitor fails, the whole system fails.

From a long term point of view, we aim to develop a peer-to-peer (p2p) or overlay application to manage the whole system. However, it is not yet obvious how this p2p/overlay should work. We need a transitional solution. Currently, our main concern is monitor failure recovery.

A simple solution would be to build a centralized server. We keep the basic round robin functioning of the system (see Sec. IV-B) but in the middle, we place a server. In this case, the system has a star topology. The server's job will be to maintain the coherency of the round robin structure.

The basic idea is the following: the server knows the entire topology of the system and, for a given monitor, to which monitor it is supposed to send its global stop set and from which it must receive it. Regularly, the server checks each monitor's state by sending messages of type HEARTBEAT. When a monitor receives a HEARTBEAT message, it is supposed to reply immediately with a HEARTBEAT_ACK message.

Non receiving consecutively, e.g., three HEARTBEAT_ACK messages from a given monitor leads to its removal from the topology. The central server then begins a maintenance (or reorganization) phase of the topology. We estimate that the system can still work while there are at least two working monitors.

The second aspect we would like to tackle in the near future is load balancing between monitors. Each traceroute@home monitor focuses on its own part of the destination list, as described in Sec. IV-B. However, a problem arises on the step size by which to advance the window. A manually tunable step size does not eliminate blocking situations in which a monitor is waiting for the prior monitor to terminate its window before it can do any further useful work. Some monitors may potentially wait a long time before receiving the needed information (see Table IV). This might happen because some monitors are slower (as they are more heavily loaded) than others.

We plan to develop, in our future version, a way to balance the load between monitors. This will imply that the sliding window size will differ from one monitor to another. Some monitors will work harder while others will maintain a low probing rate.

Currently, a person who controls a traceroute@home system might use it in a malicious way in order perform DDoS attacks. Nothing is done to prevent this misuse of our tool. However, the centralized solution also has the opportunity to improve the security in traceroute@home.

The new version of traceroute@home should also allow communicating entities (monitors and servers) to mutually authenticate themselves through cryptographic level *authentication*. The next version must also prevent third parties from eavesdropping network communications, i.e., guarantee their *confidentiality*. It is unfortunately impossible to fully prevent eavesdropping from administrators of machines running our programs. In addition to that, the next version of traceroute@home must be able to guarantee the *integrity* of the results. We should be able to identify tampering if and when it happens. While it is not reasonable to expect that we will be able to detect subtle modifications, we shall reject absurd results and stop accepting input from those trying to submit them.

We are likely to extensively use cryptographic means. They have all the features we need and good tools are already available. We should use a *public key infrastructure* (PKI) to create and manage certificates for both clients and servers. Communicating entities will thus have the ability to easily verify their peer's identity before proceeding.

Furthermore, in addition to the Doubletree prototype robustness increase, this centralized infrastructure opens interesting perspectives, in particular for the development of a general network monitoring tool, along the lines suggested by COMNI Workshop [31] in which we were active. We can imagine an extension to our tracerouting tool in order to provide additional measurement services that can be used for network monitoring. This could differ from Scriptroute [32] as the monitors have the opportunity to cooperate.

Another interesting future undertaking would be to make our traceroute@home prototype IPv6 networks aware, allowing thus the use of Doubletree in IPv6 networks. Currently, the prototype can only probe IPv4 networks. In the near future, we would like to increase its capabilities to IPv6 networks. We believe that the current version can be easily extended in order to support IPv6. The main work should be done in the JSocket Wrench library, to handle IPv6 messages and sockets. Note that standard IPv6 traceroute, such as scamper [9], or more complex tools, such as *atlas* [33], already exist.

B. Measurement Infrastructure

To test our prototype, we choose the PlanetLab infrastructure because it offers an easy access to a relatively large quantity of nodes. However, despite this apparent simplicity, we encountered several difficulties, as mentioned in Sec. V-A.

In a certain sense, these problems were expected as PlanetLab is a *testbed* oriented towards overlays and peer-to-peer networks. It is not an infrastructure entirely dedicated to network measurement or the deployment of measurement tools. Such tools have to share resources (CPU, memory, network access) with all the users, to strongly limit their use of disk space, cannot control node management.

Consequently and inspired by the COMNI workshop [31], we believe it is time to think about an infrastructure entirely dedicated to network measurements and network monitoring. This infrastructure should allow us to go beyond the experimental environment of PlanetLab.

In the fashion of PlanetLab, the nodes composing this infrastructure should be numerous and geographically diverse. Further, this infrastructure should be carefully engineered in order to avoid attacks from the outside world and to avoid abuse from users.

VII. CONCLUSION

In this report, we described our Java implementation of an efficient and cooperative topology discovery algorithm, Doubletree. We implemented the algorithm in a tool we call traceroute@home. traceroute@home is freely available and easy to extend.

We first discussed the global functioning of the system and, next, we introduced the internal architecture of a traceroute@home monitor. We also explained the message frame-

work proposed for our prototype. This message framework is easy to extend for further improvements.

In order to test our implementation, we deployed our prototype on a few PlanetLab nodes and evaluated its performance.

We finally identified some weaknesses in our prototype and proposed several ideas for further development. We also introduce a discussion about the opportunity of developing a networking measurement infrastructure.

REFERENCES

[1] V. Jacobsen et al., "traceroute," man page, UNIX, 1989, See source code: ftp://ftp.ee.lbl.gov/traceroute.tar.gz.

[2] R. Govindan and H. Tangmunarunkit, "Heuristics for internet map discovery," in *Proc. IEEE INFOCOM*, 2000.

[3] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with Rocketfuel," in *Proc. ACM SIGCOMM*, 2002, http://www.cs.washington.edu/research/networking/rocketfuel/.

[4] J. J. Pansiot and D. Grad, "On routes and multicast trees in the internet," *ACM SIGCOMM Computer Communication Review*, vol. 28, no. 1, 1998.

[5] R. Teixeira, K. Marzullo, S. Savage, and G. Voelker, "In search of path diversity in ISP networks," in *Proc. Internet Measurement Conference (IMC)*, 2003.

[6] "Route views, University of Oregon Route Views project," http://www.antc.uoregon.edu/route-views/.

[7] B. Donnet, P. Raoult, T. Friedman, and M. Crovella, "Efficient algorithms for large-scale topology discovery," in *Proc. ACM SIGMETRICS*, 2005.

[8] B. Huffaker, D. Plummer, D Moore, and k claffy, "Topology discovery by active probing," in *Proc. Symposium on Applications and the Internet*, 2002.

[9] "IPv6 scamper," WAND Network Research Group. Web site: http://www.wand.net.nz/~mjl12/ipv6-scamper/.

[10] F. Georgatos, F. Gruber, D. Karrenberg, M. Santcroos, A. Susanj, H. Uijterwaal, and R. Wilhelm, "Providing active measurements as a regular service for ISPs," in *Proc. Passive and Active Measurement (PAM) Workshop*, 2001.

[11] A. McGregor, H.-W. Braun, and J. Brown, "The NLANR network analysis infrastructure," *IEEE Communications Magazine*, vol. 38, no. 5, 2000.

[12] A. Lakhina, J. Byers, M. Crovella, and P. Xie, "Sampling biases in IP topology measurements," in *Proc. IEEE INFOCOM*, 2003.

[13] A. Clauset and C. Moore, "Traceroute sampling makes random graphs appear to have power law degree distributions," cond-mat 0312674, arXiv, 2004.

[14] B. Cheswick, H. Burch, and S. Branigan, "Mapping and visualizing the internet," in *Proc. USENIX Annual Technical Conference*, 2000.

[15] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, "SETI@home: An experiment in public-resource computing," *Communications of the ACM*, vol. 45, no. 11, 2002.

[16] S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande, "FOLDING@home and GENOME@home: Using distributed computing to tackle previously intractable problems in computational biology," in *Computational Genomics*, 2002, http://folding.stanford.edu/.

[17] "distributed.net," http://www.distributed.net.

[18] Y. Shavitt and E. Shir, "DIMES: Let the internet measure itself," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 5, 2005, http://www.netdimes.org.

[19] Java Sun Microsystems, "JDK 1.4.2," http://java.sun.com.

[20] B. Donnet, "traceroute@home 1.0," https://www-rp.lip6.fr/site_npa/site_rp/traceroutеathome-1.0.tar.gz.

[21] "PlanetLab project," Web site: http://www.planet-lab.org.

[22] B. Donnet, T. Friedman, and M. Crovella, "Improved algorithms for network topology discovery," in *Proc. Passive and Active Measurement Workshop (PAM)*, 2005.

[23] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, 1970.

[24] B. Donnet and T. Friedman, "Topology discovery using an address prefix based stopping rule," in *Proc. Eunice Workshop*, 2005.

[25] "Jsocket wrench R04," 2004, http://jswrench.sourceforge.net/.

[26] D. Eastlake and P. Jones, "US secure hash algorithm 1 (SHA-1)," RFC 3174, Internet Engineering Task Force, 2001.

[27] T. Moors, "Streamlining traceroute by estimating path lengths," in *Proc. IEEE International Workshop on IP Operations and Management (IPOM)*, 2004.

[28] D. Crocker and P. Overell, "Augmented BNF for syntax specifications: ABNF," RFC 2234, Internet Engineering Task Force, 1997.

[29] L. Peterson, V. Pai, N. Spring, and A. Bavier, "Using PlanetLab for network research: Myths, realities, and best practices," Design Note PDN–05–028, PlanetLab Consortium, 2005.

[30] Planet-Lab users mailing list, ," Mar. 2006, http://lists.planet-lab.org/pipermail/users/2006-March/0018

[31] kc claffy, M. Crovella, T. Friedman, Shannon C., and N. Spring, "Community-oriented network measurement infrastructure (COMNI)," 2005, Workshop Report.

[32] N. Spring, D. Wetherall, and T. Anderson, "Scriptroute: A public internet measurement facility," in *Proc. 4th USENIX Symposium on Internet Technologies and Systems*, 2002, http://www.cs.washington.edu/research/networking/scriptrout

[33] D. C. Waddington, F. Chang, R. Viswanathan, and B. Yao, "Topology discovery for public IPv6 network," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 3, 2003.