Achieving scale: Large scale active measurement from PlanetLab (LABS)

Jordan Augé, Marc-Olivier Buob, Timur Friedman (UPMC)

4th PhD School on Traffic Monitoring and Analysis (TMA), 2014

OBJECTIVES

Measuring and monitoring widely distributed infrastructures, such as the Internet, is a challenging task that requires the use of sophisticated tools and methodology.

In this lab, we set as an objective to measure the deployment of the Google Public DNS services, well known under the 8.8.8.8 IP address, and we will also investigate root DNS servers.

STRUCTURE OF THE DOCUMENT

The subject is composed of four parts.

The first part, presents a set of tools and webservices of interest for those willing to better understand the Internet architecture, and will introduce some basic concepts that will help us get a better understanding on how the DNS servers are managed.

In the second part, we will target more specifically the Google Public DNS and try to get some evidences about its architecture, thanks to carefully designed measurements from PlanetLab nodes. This will allow us to design a measurement methodology that we will appy on our measurements.

Before running large scale measurements, the third part will familiarize yourself with the tools we have previously presented, namely:

> TDMI : for running large scale distributed measurements on top of PlanetLab,

> TopHat : for aggregating and combining heterogeneous datasets,

> Manifold: the software that runs those different services.

As an application case, we will characterize the location of nodes in the PlanetLab testbed to evaluate whether it is representative of the whole Internet.

Finally, in the last part, we will proceed to large scale measurements towards both the root DNS servers and Google Public DNS. The first series of measurements will allow us to benchmark our method, since there is plenty of available information for DNS servers. We will then map the deployment of Google Public DNS servers as much as possible.

REQUIREMENTS

This lab refers to some tools and datasets available on a specifically built virtual machine based on VirtualBox. See the TMA PhD school website for more information.

PRELIMINARY

We have already set up a PlanetLab slice and created a SSH keypair for the lab, that will allow you to access the nodes. Please execute the following commands on your virtual machine. They will retrieve the keypair as well as install a set of dependencies that will be useful for the lab.

```
wget http://www.top-hat.info/download/tma-2014/tma2014-scale-setup.sh
chmod +x tma2014-scale-setup.sh
su
./tma2014-scale-setup.sh
```

A. INTERNET ARCHITECTURE

The Internet is made of the interconnection of multiple independent networks, called *Autonomous Systems*, or AS. These ASes exchange traffic according to routing agreements implemented via the BGP protocol.

A.I IP addresses, AS, and their geography

A.I.1. Can you name different types of Autonomous Systems ? Give an example for each type.

A.I.2. Determine your IP address (use the `ifconfig` command).

A.I.3. Determine your public IP address. Can you geolocalize it ? In which AS is it located ? You can copy/paste the URL of the following webservices : ifconfig.me[1], MaxMind[2] and Team Cymru IP-to-ASN mapping service[3]

A.I.4. Can you explain how Team Cymru manages to provide such a mapping ?

A.II Google Public DNS

In this lab, we will study the DNS service offered by Google under the 8.8.8.8 public IP address.

> Google Public DNS is a Domain Name System (DNS) service offered by Google. It functions as a recursive name server providing domain name resolution for any host on the Internet. The service was announced on 3 December 2009, in an effort described as making the web faster and more secure. *source: Wikipedia - Google Public DNS*

A.II.1. Under which AS is 8.8.8.8 hosted ?

A.II.2. Can you geolocate it using MaxMind ? Do you have an idea about the reason ?

A.III Internet paths

We will now study the path taken by your traffic when you contact the Google DNS server. We will start using some PlanetLab nodes. You can connected to the node `ple2.ipv6.lip6.fr` by SSH via the following command:

```
ssh upmc_tma@ple2.ipv6.lip6.fr
```

[1] `http://ifconfig.me`
[2] `https://www.maxmind.com/fr/geoip_demo`
[3] `https://asn.cymru.com/`

Your user is the name of the slice, composed of your origin institution (here UPMC) and an identifier (here: tma). Authentication is made thanks to your private key.

A.III.1. Thanks to the `traceroute` command, print the IP- and AS-level path towards 8.8.8.8. You can use the command `man traceroute` to get some help with its parameters.

A.III.2. How is the mapping IP-to-ASN done by `traceroute` ? You can use Google to search for some information.

A.III.3. This node was located in Paris, France. Do the same measurement from another PlanetLab node in Denver, in the US: `planetlab2.cs.du.edu`.

A.III.4. What types of the AS traversed ? You can use Google to find some information.

A.III.5. If this was representative of PlanetLab, what would be your conclusions about the types of networks involved in the testbed ? What consequence might this have on measurements ?

A.III.6. Again, supposing this was representative, what does this suggest about the peering agreements of Google for 8.8.8.8 with other ASes ?

B. ANYCAST

In this section, we will continue using PlanetLab nodes to understand the routing behaviour of 8.8.8.8. Open two terminals connected to the nodes we used previously, and which are located in different locations in the world (EU and US):

> ple2.ipv6.lip6.fr

> planetlab2.cs.du.edu

B..1. From one node, launch a few `ping` probes towards destination 8.8.8.8 (use the -c option). Is the performance good ?

B..2. If you were runnning the same ping on other nodes, again the performance would be roughly the same. Can you conclude something ?

B..3. Several factors originating from the PlanetLab node itself, or from the network can affect your measurements. Can you cite a few ?

B..4. What can we do to get an rough estimation of the propagation delay ?

B..5. Get an estimation of the propagation delay towards 8.8.8.8 from both nodes.

The propagation delay are $d_{EU} = 8.371$ms, and $d_{US} = 11.021$ms.

> According to Google, as of 2013, Google Public DNS is the largest public DNS service in the world, handling more than 130 billion requests per day. *source: Wikipedia - Google Public DNS*

Maybe you already get an idea about the deployment of Google Public DNS servers ? Let's investigate a bit more...

B.I Localizing PlanetLab nodes

As a preliminary step, we want to get the localization of PlanetLab nodes. This data is not available in the web interface, but fortunately, PlanetLab provides a XMLRPC API that can be called to retrieve latitude and longitude information from the nodes.

The following Python script would allow us to retrieve latitude and longitude from the nodes, by calling two API functions: `GetNodes` and `GetSites`[4]. We will see later how to retrieve this information more conveniently.

```
vim node-lat-lon.py
```

```
#!/usr/bin/env python
#! -*- coding: utf-8 -*-

import xmlrpclib, pprint

API_URL    = "https://www.planet-lab.eu:443/PLCAPI/"
AUTH = {
    'AuthMethod': 'password',
    'Username'  : 'my-planetlab-username',
    'AuthString': 'my-planetlab-password',
}

srv = xmlrpclib.ServerProxy(API_URL, allow_none=True)

nodes = srv.GetNodes(AUTH, {}, ['hostname', 'site_id'])
sites = srv.GetSites(AUTH, {}, ['site_id', 'latitude', 'longitude'])

map_sites  = dict()
for site in sites:
    map_sites[site['site_id']] = site

for node in nodes:
    site = map_sites[node['site_id']]
    latitude = site.get('latitude')
    longitude = site.get('longitude')
    if not latitude: latitude = 0
    if not longitude: longitude = 0

    print "%s\t%f\t%f" % (node['hostname'], latitude, longitude)
```

```
./node-lat-lon.py > node-lat-lon.dat
grep ple2.ipv6.lip6.fr node-lat-lon.dat
grep planetlab2.cs.du.edu node-lat-lon.dat
```

```
ple2.ipv6.lip6.fr       48.852500       2.278490
planetlab2.cs.du.edu    39.679800       -104.963000
```

[4] https://www.planet-lab.eu/db/doc/PLCAPI.php

B.II Some speed of light considerations...

Let's compute the geodesic distance (the "as the crow flies" distance) between the two nodes. This is the minimal distance between the two point at the surface of the earth. Various models and approximation exist to compute this distance.

We advise you to use Python; you can use the python-geopy module [5], which should already be installed in your VM.

You can also use the language of your choice but we might not be able to offer answers or support. Search for packages allowing you to compute the geodesic distance between two nodes (eg. using the Vincenty formula, but others could be appropriate also).

Note that this script will be reused in the latest part of the lab, for the treatment of large scale datasets.

For your information, in R, we can use the `distVincentyEllipsoid` command from the `geosphere`package [6] [7].

```
> library(geosphere)
> lip6 <- c(2.27849, 48.8525);
> ud <- c(-104.963, 39.6798)
> distVincentyEllipsoid(lip6, ud) / 1000
```

```
[1] 7881,050
```

We will assume those nodes are connected by optical fibre, whose index of refraction can be considered to be $r_f \approx 1.52$. The index of refraction of a material is calculated by dividing the speed of light in a vacuum $c_0 \approx 300000$km/s by the speed of light in the medium.

B.II.1. Compute the minimum bound on the amount of time the light would take to travel from one node to another.

B.II.2. Compare this result to the delay measurements you got previously ? What can you conclude ? Making a figure will help.

B.III Understanding anycast

B.III.1. From your understanding on how routing works, can you explain how works this deployment, which is denoted as *BGP anycasting*, or simply *anycast* ?

B.III.2. How can administrators manage the different nodes remotely from a central location ?

B.III.3. Beyond DNS, do you know or can you guess what anycast is used for in the Internet ?

B.III.4. In your opinion, what are the pros and cons of using anycast for a service ?

B.III.5. Do you see how the same functionality could be implemented (providing the same resources at different points on the network) without relying on the routing layer ?

We have a method to distinguish between several anycast instances, based on delay measurements from known landmarks. This method is usually referred as the *geo-inconsistency* method.

Our next steps will be to evaluate the efficiency and accuracy of this method, and to measurement the deployment of the Google Public DNS service from PlanetLab.

To make our life easier, we will use the Manifold tool we have introduced before to run the measurements from PlanetLab towards 8.8.8.8.

[5] `http://geopy.readthedocs.org/en/latest/`

[6] `http://www.inside-r.org/packages/cran/geosphere/docs/distVincentyEllipsoid`

[7] You can load R packages using the following command: `library(PACKAGE);`

C. Introduction to Manifold

This section is an introduction to Manifold in order to get familiar with the tool and discover its functionalities. This tool will make our life easier in order to perform the large number of distributed measurements that will be needed for understanding the DNS architecture.

We will use the TopHat service, running at UPMC, which is a deployment of Manifold where a set of data sources of interest have been preconfigured. It includes the services from geolocalization and IP-to-ASN mapping we have used previously, as well as TDMI (TopHat Dedicated Measurement Platform) which will allow us to launch active measurements such as `ping` and `traceroute` from PlanetLab nodes.

Due to the way Internet works, it is natural that our results will depend on the number and location of our vantage points. Let's try to get some insight in this.

C.I Using Manifold shell

Manifold should have been installed by the script you ran at the beginning of this lab. It provides a shell that will allow us to issue queries to the service interactively.

```
manifold-shell -x -U https://clitos.ipv6.lip6.fr:7080 -L ERROR -z anonymous
```

The `-x` option inform the shell to contact the remote service by XML/RPC. The following URL is an alternative deployment of TopHat we have set up for this lab, since the service is not publicly open yet. We then ask the tool to display only error messages. Finally, we use anonymous authentication for the purpose of this lab.

C.II Discovering metadata

Platforms exposing data through Manifold use a representation similar to the one used in relational databases: tables. It is not surprising that the format for queries is then similar to SQL.

In a first step, we will issue a query towards a special table, `local:object`, which sits in a `local` namespace, to retrieve the set of available tables. This special queries allows us to retreive metadata, or the schema of the information available in Manifold.

Run the following queries to get familiar with the system:

```
>>> SELECT table FROM local:object
```

Let's look at the `ip` table more in details, which gives us some information related to IP addresses. The following query lists the fields available in the table `ip` and their description. These are the fields you can use in a query to get some informations.

```
>>> SELECT columns.name, columns.description FROM local:object WHERE table == 'ip'
```

Metadata describe the full range of available information, in this case related to an IP address, even if they do not belong to a single platform. In our setup, this information is drawn from interconnected platforms named `maxmind`, `tc` (Team Cymru) and `dns`. This can be verified by giving the name of the platform as the namespace for the query.

```
>>> SELECT columns.name, columns.description FROM maxmind:object
        WHERE table == 'ip'
>>> SELECT columns.name, columns.description FROM tc:object
        WHERE table == 'ip'
>>> SELECT columns.name, columns.description FROM dns:object
        WHERE table == 'ip'
```

C.III Query databases and webservices

Let's use Manifold to obtain the same date as in Section A..

C.III.1. First, contact the individual platforms (using the namespace) to request respectively country and AS information.

C.III.2. Remove the namespace and issue a single query to the integrated view.

C.IV Some statistics about PlanetLab

Information about PlanetLab nodes and sites that we have previously retrieved from the XMLRPC API via a Python script is also available transparently in Manifold, via the `node` table. Notice that there is no need to make two separate calls anymore (to the equivalent of `GetNodes()` and `GetSites()` functions), this is done automatically by Manifold.

The shell offers you the possibility to store the results of your query inside a variable and dump it into a CSV file.

```
>>> \$variable_name = SELECT \dots FROM \dots
>>> SHOW
>>> SHOW \$variable_name
>>> DUMP \$variable_name INTO "/path/to/file.csv"
```

C.IV.1. Run a Manifold query to get a list of nodes, their country and their ASN.

C.V Analysis of PlanetLab diversity

Run the R software, it will allow you to do some simple statistics about the data.

```
R
```

You can run the follwing commands to load the file you just created and make some statistics.

```
nodes <- read.csv("/home/tma/nodes.csv")
summary(nodes)
summary(nodes$country_name)
summary(nodes$asn)
\dots
```

C.V.1. What can you conclude about PlanetLab diversity ?

C.VI Adding local data sources

Depending on the remaining time, you might want to skip this section, and proceed directly to the last part : "Measuring the DNS architecture". In this case, just read through this section to get an idea of the functionalities of Manifold.

One remark that is often done for measurement studies based on PlanetLab is the bias towards academic networks, since a huge proportion of the nodes is hosted by universities, and far less on commercial networks. You might already have noticed this previously.

One way to verify this would be to cross the data about PlanetLab nodes with a characterization of the ASes they belong to. An old but interesting dataset is provided by GeorgiaTech [**?**].

We provide a copy of this dataset (in CSV format) in the archive you downloaded at the beginning of this lab. This dataset is available online[8].

[8] `http://www.ece.gatech.edu/research/labs/MANIACS/as_taxonomy/data/as2attr.tgz`

In order to join this local data with the information offered by TopHat, we need to configure and run a local Manifold router. We first add the TopHat platform we used to query directly via the shell. You can copy/paste these commands.

```
manifold-add-platform tophat TopHat manifold none
        '{"url": "https://clitos.ipv6.lip6.fr:7080/"}'
```

As for the local CSV file, Manifold already provides an adapter allowing to query this data source. It can be configured simply by configuring a new platform with the corresponding file path. Usually, column names are guessed from the CSV file, but they can be given to Manifold via the platform configuration (which is needed in our case). You can copy/paste these commands.

```
manifold-add-platform georgiatech
        "Georgia Tech Autonomous System Taxonomy Repository"
        csv none '{"as":{"filename": "/tmp/as2attr.txt", "fields":
        [["asn", "int"], ["as_description", "string"],
        ["num_providers", "int"], ["num_peers", "int"],
        ["num_customers", "int"], ["num_prefixes_24", "int"],
        ["num_prefixes", "int"], ["asn_class", "string"]],
        "key": "asn"}}' 0
manifold-enable-platform georgiatech
```

Finally, we then run the shell with an integrated router, with the set of platforms we have configured.

```
manifold-shell -z router
```

C.VI.1. Looking at metadata, verify that the `ip` table now includes those additional fields (mainly `asn_class`.

C.VII Verify PlanetLab academic bias

C.VII.1. Request for each planetlab node its hostname, country, AS number and AS class, and store this inside a csv file (eg. `/home/tma/nodes.csv`).

C.VII.2. Like previously done, use R to see the diversity in terms of AS types. What can you conclude ?

C.VII.3. What consequence this observation might have on the discovery of anycast instances ?

D. Measuring the DNS architecture

D.I In search of ground truth

We have devised a methodology to measure the deployment of anycast instances (we will see later how to extend the reasoning to more than 2 nodes). Though, we do not know neither its efficiency, nor its accuracy.

A significant issue often encountered in measurements is that usually we have no ground truth to which to compare our results. Most of the time we are searching for evidences like we have done here, and it is hard to know about the issues of the model, the measurement artefacts, or the bias introduced by our measure. A significant example is the topology of the Internet, which remains largely unknown. At best we can estimate some of its properties.

The measurement of Google public DNS deployment is no exception to the rule.

Fortunately for us, some root DNS server deployments use anycast, and their deployment is public[9]. (although it might sometimes be incomplete or not up to date). In addition, they often embed debug functionalities that can help us identify which anycast instance we are talking to, even though they all

[9] http://root-servers.org/

answer with the same address. This is described in RFC4892[10] : "Requirements for a Mechanism Identifying a Name Server Instance":

cmttFor some time, the commonly deployed Berkeley Internet Name Domain (BIND) implementation of the DNS protocol suite from the Internet Systems Consortium [BIND] has supported a way of identifying a particular server via the use of a standards-compliant, if somewhat unusual, DNS query. Specifically, a query to a recent BIND server for a TXT resource record in class 3 (CHAOS) for the domain name "HOSTNAME.BIND." will return a string that can be configured by the name server administrator to provide a unique identifier for the responding server. (The value defaults to the result of a gethostname() call). This mechanism, which is an extension of the BIND convention of using CHAOS class TXT RR queries to sub-domains of the "BIND." domain for version information, has been copied by several name server vendors.

A refinement to the BIND-based mechanism, which dropped the implementation-specific label, replaces "BIND." with "SERVER.". Thus the query label to learn the unique name of a server may appear as "ID.SERVER.".

For example, both the F and K root servers use anycast and answer those CHAOS queries. We will use those deployments to benchmark our methodology, before applying it to 8.8.8.8.

For example, the `dig` command, identify the instance of the DNS server `f.root-servers.org` and `f.root-servers.org` we are talking to:

D.I.1. Using the second PlanetLab node (in the US), run the previous `dig` command to identify the instance you are talking to. Can you guess something from the naming ?

D.I.2. Repeat the measurement a couple of times. If you notice differences, what do you notice ? How do you explain these differences ?

D.II Obtaining data and ground truth

Obtaining a ground truth requires us to run simultaneous measurements of delay (using `ping`), and identification of the anycast instance (using `dig`) from a large number of nodes.

The integrated dataset of measurements targetting the F-ROOT and K-ROOT root DNS servers, as well as the Google Public DNS has been prepared in advance:

You can download the datasets at this address:
`http://www.top-hat.info/download/tma-2014/datasets/`
The file is composed of the following columns, separated by a space character:

> hostname: the hostname of the PlanetLab node;

> latitude: its latitude;

> longitude: its longitude;

> delay: the propagation delay towards the anycast instance (as measured by `ping`);

> instance (or *unknown* when it is not available): the name of the anycast instance (as measured by `dig`).

D.III Extending the method to multiple nodes

Previously, we have introduced the *geo-inconsistency* method to distinguish between two anycast instances. We will now extend this method to multiple instances.

Suppose we are looking at the measurement from node i, at position p_i. The delay to the anycast instance is d_i. Denote A_i the area that can be reach by the light during the time d_i.

D.III.1. Using set notation, with nodes 1 and 2, what is the necessary condition on A_1 and A_2 so that we can distinguish the two anycast instances with the *geo-inconsistency* method ?

D.III.2. What is the necessary condition for 3 nodes ? for N nodes ?

[10] `http://www.ietf.org/rfc/rfc4892.txt`

D.III.3. In practice, this condition will not be verified by all our measurements. What should be done in order to be in a position to apply the method ?

D.III.4. How can we discover as many instances as possible ? Clearly formulate the problem.

D.IV Implementing the method

D.IV.5. Make a program that will try to determine as much instances as possible, based on the previous datasets. You can make an approximated algorithm here.

D.IV.6. Run your program against `froot.dat` and `kroot.dat`.

D.IV.7. Looking at the root server website[11], what do you think of the completeness of the method ?

D.IV.8. What would be necessary for a better coverage ?

D.IV.9. What about the method accuracy ? does it present inconsistencies (two different instances that are in fact the same) ?

D.IV.10. What might be the reasons for these inaccuracies ?

D.IV.11. Run your program against `google.dat`. What are your results ?

D.IV.12. If you still have time, plot your results on a map.

D.IV.13. Given your understanding of Internet routing, and the results you have observed for paths and delays, what might be a good strategy for deploying anycasted instances ?

D.IV.14. Do you have ideas on how to further analyze anycasted services ?

[11] `http://root-servers.org/`