IMPROVEMENT OF GEOLOCATION METHODS

Jiří Balej
Doctoral Degree Programme (1), FEEC BUT
E-mail: jiri.balej@phd.feec.vutbr.cz

Supervised by: Dan Komosný
E-mail: komosny@feec.vutbr.cz

Abstract: The paper deals with geolocation of device with an IP address. The passive and active methods and their major representatives are described. Next part contains the measurement results and their use for determining delay per hop and real cable length to straight distance ratio. These values are used in the proposed technique which aim is to improve accuracy of constrint-based IP geolocation methods. New method tries to cut off delay caused by intermediate devices and it considers the cable winding factor.

Keywords: Geolocation, CBG, SOI, distance, delay.

1 INTRODUCTION

Geolocation estimates geographic position of person, car, device,. . . This paper deals with the problem of finding position of computer or device with an IP address. The real-world position is valuable for many purposes like locating nearby points of interest, set up local environment (language, currency,. . .), identify place of origin of emergency call and aiming the advertisement. Each purpose of geolocation has own location accuracy requirements, once it is just country other time it demands accuracy in kilometers. The geolocation techniques can be classified as passive and active. Passive methods use databases and manually administered records, active methods measure delay and other network properties like number of intermediate devices (hops).

1.1 PASSIVE GEOLOCATION METHODS

The passive methods are simpler than active, but their disadvantage is the requirement to manually update data about location, which is time expensive and includes a lot of faulty records. On the other hand these methods are the most used nowadays for their simplicity. The two basic principles are described below, but their combinations also exist.

IP address database investigation is the first and the most used way for geolocation. To find a node location it is needed to check out one of many databases which contain IP addresses, country, organization and their estimated location. The location is estimated primarily from the Regional Internet Registries (RIRs), for Europe it is RIPE NCC[1]. Supplementary sources for better accuracy are information from local Internet Service Providers (ISP) and the results of data mining techniques, which are based on data filled by users into web forms. These databases contain a lot of private data, so public access to this data is not advisable due to privacy and security issues. Another disadvantage is the reliability of the records. The location of the host may be not the same as that of the registered office of owner [1]. Most of commercial geolocation products use IP databases, which are updated manually by many administrators and use several information sources to ensure accuracy and to remove faulty records.

[1] www.ripe.net

DNS records give two pieces of information about host location. The first is the TLD (Top Level Domain), which usually belongs to a specific country. The second may be a reference to the city, where the host is situated (like abbreviation PRG means Prague). Some DNS servers support the extension of DNS referred in RFC 1876[2], which allows information about the location of DNS server to be added to databases records. This method can contain errors, when domain country code mismatch with country where host is located. Other errors can be caused by the relocation of servers, without properly changing the names [1].

1.2 ACTIVE GEOLOCATION METHODS

Geolocation techniques based on latency measurement have been developed recently. These methods are based on measuring latency to hosts with a known location. These nodes are called landmarks and they can be located all over the world. The ping utility is used to identify latency between the host and landmarks. Some methods measure latency from landmarks to host, some from host to landmarks. Certain methods use additional network topology information, which can be discovered using the traceroute tool. This tool detects intermediate devices working on the third OSI (Open Systems Interconnection) layer.

The fundamental of active methods is in relation between delay and geographic distance. Main impact on delay has the length of communication trace, but also the type of transmission medium and the time needed for switching and routing (finding a path) to destination. The overall delay consists of the particular latencies originated at the elements of communication chain. The delay can be divided into two parts – deterministic and stochastic [3]. The stochastic part contains mostly the queuing and processing latency of devices on load. This part can be suppressed by choosing the minimal measured value of delay for calculations. Deterministic part contains the minimal queuing and processing latency of devices, transmission and propagation delay of transport lines. Most of geolocation methods count only propagation delay and neglect the impact of other deterministic factors. Another drawback of nowadays geolocation methods is the neglecting of cable winding, because cables are mostly buried along highways, railways or long-distance power lines.

GeoPing method locates host by mapping it to the most representative landmark. A host's real-world position is determined by similarity the host delay vector to landmark delay vector. Each active landmark measures delay to other landmarks and creates own delay vector. The host to get own geographic position have to measure delay to some landmarks and compare it with delay vector of active landmarks. The geographic position of active landmark with most similar delay vector is announced as host position [2].

Constraint Based Geolocation (CBG) method computes the geographic distance from delay and use the multilateration method to determine host location. The delay is measured to landmarks with known location. Fig. 1 a) shows the intersection of circles announced as region of possible location. Circles have center in landmark and their radius is geographic distance constraint from landmark to host. Constraint is distance computed from delay using relation between distance and delay, which is gathered from landmark calibration [2].

Speed of Internet (SOI) method is similar to CBG but much simpler, because it does not require initial measurement between landmarks. SOI uses constant for geographic distance computation. The constant suppresses delay from intermediate devices and the influence of cable winding. Delay to distance conversion factor was defined as $\frac{4}{9} \cdot c$ and it is used instead of constant for speed of propagation c_{ew} [2].

The goal of improvement is to reduce circles intersection area like in Fig.1 b). But if the constraints are too low, the circles do not have to intersect and there is no region of possible location. This is

[2] www.rfc-editor.org/rfc/rfc1876.txt

called underestimation of circle boundaries and it is shown in Fig. 1 c). Considering that delay does not depend only on the distance between hosts, it is possible to enhance the accuracy of geolocation by analysis of communication delay and its sources.

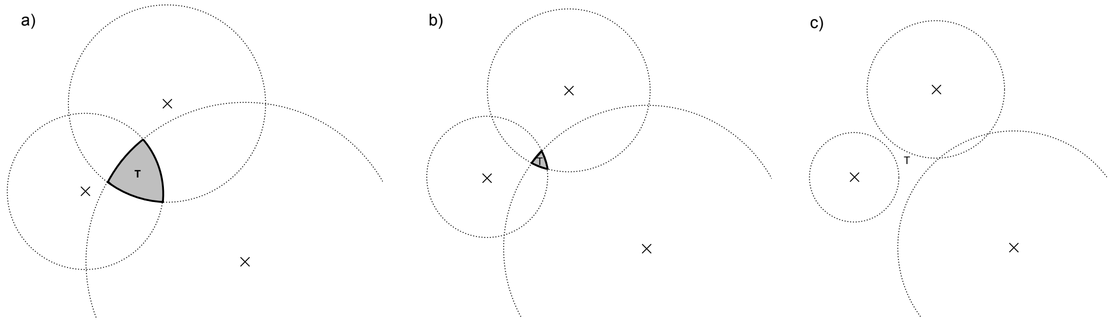

Figure 1: Circles represent the furthermost target position and the intersection of all circles bounds the possible target location. In a) is the region of possible position large, in b) is the intersection area lower – the target position is determined more precisely, in c) the intersection does not exist.

2 LATENCY MEASUREMENT IN CESNET2 NETWORK

The CESNET2 network was chosen for very rich documentation and the availability of nodes. The information from Cesnet documentation used was the length of cables between cities. The measurement was done on two Planetlab[3] stations located in Prague and Brno. The Round Trip Time (RTT) was measured using the ping tool. The measurement was repeated for five times (in different days and hours), the request message was sent every second for 120 seconds to each address in set. The lowest RTT value was chosen to eliminate the load of intermediate devices – the stochastic part of delay. One-way delay was computed as half of RTT. We assume the routing symetry, because the path between PlanetLab nodes in Brno and Prague from traceroute printout was the same for both directions. The number of hops was gained from returned IP packets TTL (Time to Live) value, which is reduced by every layer 3 device on trace. The initial TTL value can be one of the few default values (255, 128, 64, 60 and 32)[4]. The number of hops represents devices working on layer 3 (devices on first nad second layer are hidden).

2.1 THE CALCULATION OF LATENCY PER HOP

The amount of delay representing one device on the route is hard to determine, because it depends on many different and varying parameters. This experiment shows possibility of finding delay per hop from the measured latency, length of cable and number of hops. The result of equation

$$d_{\mathrm{id}} = \frac{d - \frac{l_{\mathrm{true}}}{c_{\mathrm{ew}}}}{n} \tag{1}$$

is delay for one hop on route (d_{id} – one intermediate device). For calculation, it is needed to know one-way delay d, the length of cables between cities l_{true} and number of hops n. The c_{ew} is constant for speed of signal propagation in optic $\frac{2}{3} \cdot c$ where c is the speed of light [2].

The histogram in Fig. 2 shows the distribution of calculated delays per hop. It can be summarized that delay per hop is higher than 40 μs for 87 % samples. Value per hop can be subtracted from measured delay for each hop on trace. If this result is used for geographic distance calculation, then

[3]www.planet-lab.org

[4]Differs by software.

the computed distance will be lower. For example the geolographic distance can be reduced by more than 100 km (if the number of hops in both directions is bigger than 6).

The average value for all calculated delays per hop is 122 μs (median 76.5 μs). The results presented in paper [4] are similar – authors measured mean value of router processing delay as 101 μs.

2.2 REAL CABLE LENGTH TO STRAIGHT DISTANCE RATIO

For geolocation techniques, the derivation of real distance from computed length of cable can be useful. This is the reason why the ratio of real cable length to straight distance ($\frac{l_{\text{true}}}{l_{\text{straight}}}$) is expressed. The cable length from Cesnet documentation for the shortest trace was used to show how it differs from straight distance. Fig. 3 shows line that represents the equality of straight distance and cable length. The dashed line shows the minimal ratio between straight distance and cable length. The ratio of real length of cable to straight distance is in average 1.99 and the median is 1.92. Both values are similar and express that the real length of cables is approximately twice longer than straight distance. For the purpose of geolocation, we use minimal cable length to straight distance ratio 1.56, which suppress the cable winding factor and also the target will be inside the intersection of circles.

Authors in [5] empirically established the winding of cable as the modified constant for speed of signal propagation – the minimum about 0.09, the maximum 0.47 and the average around 0.3. In this paper the constant for speed of signal propagation has the minimum 0.14, the maximum 0.42 and the average 0.33.

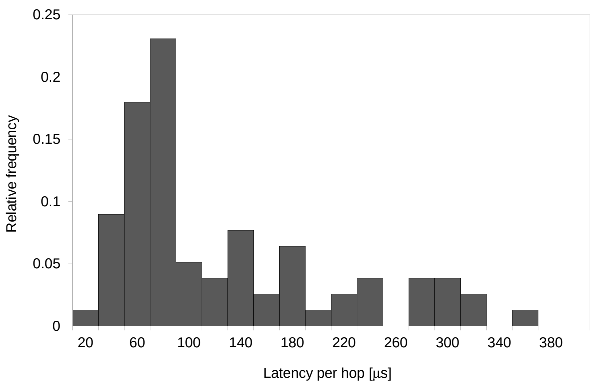

Figure 2: Distribution of delay per hop based on measurement, each bar is 0.02 ms wide.

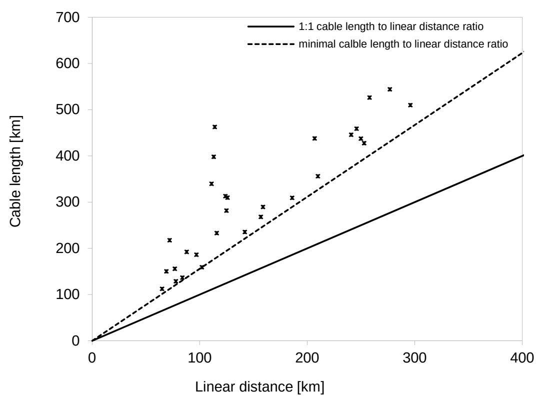

Figure 3: Relation between linear distance and length of cable.

3 THE IMPROVEMENT OF CONSTRAINT-BASED GEOLOCATION METHODS

The measured results can be used to increase accuracy of CBG or SOI geolocation method. The principle of proposed improvement is firstly to subtract delay per hop d_{id} (computed in previous section) for each device on L3 OSI from one way delay. Then multiply result by constant for winding of cable c_{w} (also computed in previous section). The equation for distance from landmark is then

$$l = \frac{1}{c_{\text{w}}} \cdot c_{\text{ew}} \cdot (d - n \cdot d_{\text{id}}). \tag{2}$$

The first experiment was done for host in Jindrichuv Hradec (Czech Republic). The measurement was done from four landmarks located in Stuttgart, Prague, Wroclaw and Brno. As constant for winding of cable was used minimal calculated value from previous section $c_{\text{w}} = 1.56$. For the latency per hop we used minimal delay for at least 87 % samples from previous section ($d_{\text{id}} = 0.04$ μs). For comparison the same measurement was done using Constraint Based Geolocation and Speed of the Internet method. The measurement results are shown in Fig. 4. The first part a) shows the region of possible locations calculated by SOI – it has area 166 737 km^2. Second figure b) contains result from CBG measurement with area about 90 thousands square kilometers and the last part c) is for proposed method with area 57 645 km^2. The region of last measurement is the smallest of all measurements. It has area 1.58 times smaller than CBG method and 2.89 times smaller than SOI method.

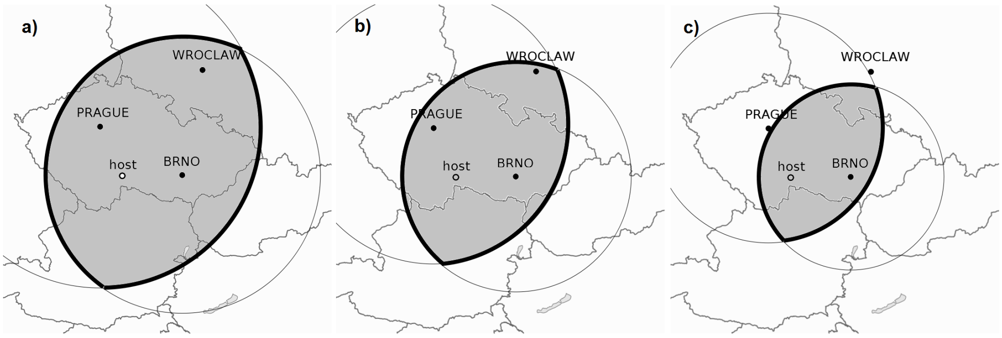

Figure 4: Circles show the boundaries of possible position of host, the intersection is the result. In a) Speed of Internet method, b) CBG method and c) new method.

4 CONCLUSION

This paper describes geolocation methods for finding a possition of IP nodes. There are two groups of these methods – passive and active. Passive methods are nowadays most used methods, but the biggest drawback is their manual maintaining. In contrast the active methods need only a set of landmarks with known position. Three active geolocation methods were described in the beginning of paper and their pros and cons were mentioned. This paper was aimed at accuracy improvement of these methods.

The paper contribution is definition of the real cable length to linear distance ratio. The delay per one hop was calculated for each measured sample and it was found that more than 87 % samples has the per hop latency higher than 40 μs.

The identified real cable length to linear distance ratio was used in first test of proposed improvement of geolocation. Result of single measurement was the noticeable improvement. In the future work, the proposed method will be globally tested and evaluated.

REFERENCES

[1] Sarikaya, Behcet (Ed.), *Geographic Location in the Internet*, 2002, 240p, ISBN: 978-1-4020-7097-6.

[2] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Tom Anderson, and Yatin Chawathe., *Toward IP Geolocation Using Delay and Topology Measurements* In *IMC '06*.

[3] BOVY, C. J., et al., *Analysis of End-to-end Delay Measurement in Internet* [online]. In *Passive & Active Measurement: PAM 2002*, 2002. s. 8.

[4] CONSTANTINESCU, D., POPESCU, A., *Modeling of One-Way Transit Time in IP Routers* . In *AICT/ICIW 2006*, 2006, s. 10.

[5] Laki, S., Mátray, P., Hága, P., Csabai, I., Vattay, G., *A model based approach for improving router geolocation* In *Computer Networks*, 2010, pp 1490-1501.