VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIZERTAČNÍ PRÁCE

Brno, 2017 Ing. Jiří Balej

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

AKTIVNÍ IP GEOLOKACE PRO VERIFIKACI POZIC STANIC V INTERNETU
ACTIVE IP GEOLOCATION FOR VERIFICATION HOST POSITION IN INTERNET

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — **Ing. Jiří Balej**
AUTHOR

ŠKOLITEL — **doc. Ing. Dan Komosný, Ph.D.**
SUPERVISOR

BRNO 2017

ABSTRAKT

Dizertační práce se zabývá způsoby nalezení geografické polohy zařízení v síti Internet při znalosti IP adresy. Tento proces se nazývá IP geolokace a je v současnosti řešen pomocí geolokačních databází nebo za využití výsledků měření síťových parametrů k cílové IP adrese. Nevýhodou dnešních geolokačních databází je, že některé poskytované polohy nejsou správné a mohou vykazovat velkou odchylku od správné polohy. Cílem této práce je vyvinout metodu, která by na základě měření dokázala ověřit správnost pozice z geolokační databáze. Z tohoto důvodu je v práci podrobně rozebrán vliv parciálních částí zpoždění, které ovlivňují výpočet maximální vzdálenosti na základě změřeného zpoždění mezi referenční stanicí a cílovou IP adresou. Ze stejného důvodu je v práci popsáno dlouhodobé měření zpoždění, kde je řešena přesnost IP geolokace za použití kalibračních dat z dřívějších měření. Navržená metoda Cable Length Based Geolocalisation (CLBG) je postavena na vlastnostech dílčích složek zpoždění, které jsou závislé na délce přenosových médií. Metoda ze změřeného obousměrného zpoždění vyloučí vliv zpoždění generovaného mezilehlými prvky a koncovými stanicemi a za použití rychlosti šíření signálu přenosovým médiem určí geografickou vzdálenost. Dále byl experimentálně zjištěn parametr nepřímého vedení kabelů, jež je použit pro určení mezních hranic. Průnik mezních hranic jednotlivých referenčních bodů je následně použit ke stanovení regionu, kde se IP adresa nachází. Výsledky této metody při geolokaci jsou lepší než jednoduché metody (ShortestPing, GeoPing a SOI) a srovnatelné s metodami pokročilejšími (CBG a Octant). Nevýhodou vytvořené metody je velikost regionu, kde se stanice nachází, což je ale dáno jejím účelem. Pro zjištění správnosti informace z geolokační databáze slouží ověření, zda její pozice leží ve zmíněném regionu.

KLÍČOVÁ SLOVA

IP geolokace, lokalizace, geolokační databáze, zpoždění, multilaterace, RTT, CBG, Octant, CLBG.

ABSTRACT

Dissertation thesis deals with methods for finding the location of the device in the Internet, based on knowledge of the IP address. The process is called IP geolocation and is currently solved by geolocation databases or by measurement of network properties to the IP address. The disadvantage of nowadays geolocation databases is an incorrect information about some locations, because they can be in large distance from correct position. The aim of the thesis is to develop a method for verification of a position from geolocation database using delay measurement. Because of it, there is a detail analysis of influence of partial delays on the distance estimation accuracy, calculated using measured delay between the landmark and the target IP address. For the same reason, long-term delay measurement was performed, where the IP geolocation accuracy was compared using calibration data from previous measurements. On this background, Cable Length Based Geolocalisation (CLBG) method is proposed. Principle of this method is built on the properties of partial delays, which depend on the length of transport media. Firstly, the method measures round trip time (rtt), which is subsequently lowered by intermediate devices and end stations delay. The geographical distance is estimated using signal speed in the transport media. Further, the winding media parameter is established, which is used to determine a constraint around the landmark. The intersection of all constraints defines the area, where the target IP is. The IP geolocation using CLBG gives better results than simpler methods (ShortestPing, GeoPing and SOI), in comparison with more advanced methods (CBG and Octant) the accuracy is similar. The disadvantage of the CLBG method is the size of region, where the target lies, but this is due to its purpose. The position found in geolocation database can be checked by evaluation if it lies in the region.

KEYWORDS

IP geolocation, localization, geolocation databases, delay, multilateration, rtt, CBG, Octant, CLBG.

BALEJ, Jiří. *Aktivní IP geolokace pro verifikaci pozic stanic v Internetu*. Brno, 2017, 95 s. Dizertační práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací. Vedoucí práce: doc. Ing. Dan Komosný, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou dizertační práci na téma „Aktivní IP geolokace pro verifikaci pozic stanic v Internetu“ jsem vypracoval(a) samostatně pod vedením školitele dizertační práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor(ka) uvedené dizertační práce dále prohlašuji, že v souvislosti s vytvořením této dizertační práce jsem neporušil(a) autorská práva třetích osob, zejména jsem nezasáhl(a) nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom(a) následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora(-ky)

PODĚKOVÁNÍ

Chtěl bych velmi poděkovat vedoucímu mé dizertační práce doc. Ing. Danu Komosnému, Ph.D. za její odborné vedení, konzultace, cenné rady a podněty pro její zpracování. Také bych rád poděkoval kolegům z Mendelovy univerzity v Brně, konkrétně Ing. Martinu Pokornému, Ph.D., Ing. Petru Zachovi, Ph.D. a Ing. Jiřímu Passingerovi, za předávání zkušeností v oblasti návrhu, správy a výuky počítačových sítí. V neposlední řadě chci poděkovat své rodině a především mé přítelkyni za podporu v průběhu zpracování závěrečné práce.

Brno ....

podpis autora(-ky)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této dizertační práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno ....

podpis autora(-ky)

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

OP Výzkum a vývoj
pro inovace

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

1 ÚVOD

S rozvojem informačních věd dochází také k rozvoji nových služeb. Jednou z nich je automatické nalezení místa, kde se uživatel nachází, což zvyšuje především jeho pohodlí. Geografická lokalizace je způsob nalezení skutečné pozice zařízení a tím pádem uživatele, který toto zařízení využívá. Poloha je určená zpravidla pomocí názvu místa (stát, region, město) nebo geografických (zeměpisných) souřadnic. V tomto případě se jedná o sférické souřadnice skládající se ze dvou údajů – zeměpisné šířky a délky, které jednoznačně definují místo na povrchu zeměkoule. Dříve byly tyto údaje používány pouze k armádním účelům a pro námořní a leteckou navigaci například pomocí GPS. Dnes těchto služeb využívá téměř každý, a to nejen ke zjištění polohy své, ale například polohy svého auta, zařízení či přátel.

Tato práce se zabývá geografickou lokalizací zařízení s IP adresou – tzv. IP geolokací, v praxi se jedná o počítače, servery, mobilní zařízení atd. Tato zařízení nejsou vždy vybavena přijímačem GPS a je tak obtížné určit přesně jejich polohu. V případě, že zařízení disponuje alespoň přijímačem bezdrátového signálu (WiFi, GSM) je možné jej lokalizovat s přesností desítek až stovek metrů [30]. Pokud je geolokace prováděna pouze na základě znalosti IP adresy (IP geolokace) je přesnost mnohonásobně nižší. Při použití geolokačních databází je možné zařízení zařadit do příslušné země v lepším případě určit město, ve kterém se nachází [40, 48]. Při použití aktivní IP geolokace (založené na ICMP odpovědi IP adresy na dotaz) jsou dnes dosahované přesnosti v řádech desítek kilometrů [38, 27].

Uživatelé počítačů, mobilních telefonů a dalších zařízení dnes při návštěvách webových stránek, případně jiných služeb, vědomě a často i nevědomky využívají služeb IP geolokace. Použitím IP geolokace můžeme zajistit například:

- Zvýšení pohodlí uživatele – při prohlížení webových stránek může být na základě zjištěné polohy upraven obsah stránky. Například se jedná o zobrazení lokálního jazyka a měny, místních zpráv, předpovědi počasí či zobrazení výsledků vyhledávání vztahujících se k poloze uživatele [41]. Při velmi přesné lokalizaci (desítky až stovky metrů) mohou být nabídnuty nejbližší body zájmu – autobusová zastávka, bankomat, restaurace, ...
- Zabezpečení důvěrných stránek – při přihlašování k elektronickému bankovnictví a soukromým účtům se navíc kontroluje poloha uživatele a v případě přihlašování z neobvyklé lokality je vyžádáno dodatečné ověření, protože hrozí, že se přihlašuje neoprávněný uživatel.
- Omezení přístupů k lokálnímu obsahu – některá data (filmy, TV pořady, hudba) mohou být omezena pouze na konkrétní zemi, díky IP geolokaci je možné rozlišit oprávněné a neoprávněné uživatele.
- Odhalování pachatelů internetové kriminality – pomocí IP geolokace je možné

určit přibližnou polohu internetových pirátů [8] či serverů s nezákonným obsahem.

- Předcházení zneužití platebních údajů – umožňuje bankám detekovat anomálie v platbách klientů, například pokud byla platba provedena z lokality, kde se klient obvykle nepohybuje.
- Cílená reklama – v dnešní době je IP geolokace často využívána k nabídce místních služeb (např. blízký obchod). K tomu slouží přizpůsobení reklamy lokalitě uživatele, případně typu reklamy (při znalosti demografických dat) pro zvýšení zisku společností.
- Lokalizace tísňových hovorů – v případě volání na tísňovou linku pomocí VoIP (Voice over IP) je možné díky IP geolokaci určit přibližné místo, odkud je voláno.

Mimo výhod plynoucích z geolokace můžeme v souvislosti se znalostí pozice uživatele nalézt i problémy. Obecně vzato, informaci o poloze určité osoby lze považovat za důvěrnou, protože může posloužit k nekalým účelům – sledování osoby, odhalení doby, kdy se daná osoba nevyskytuje doma atd. Z tohoto důvodu je poskytnutí přesných informací, například prostřednictvím WiFi signálu, zpravidla nutné potvrdit uživatelem. Oproti tomu lokalizace pouze na základě IP adresy zařízení je dostupná vždy, neboť tyto informace (v hlavičce paketu) dojdou až příjemci zprávy. Pokud uživatel využívá NAT (Network Adresss Translation) nebo Proxy serveru, není dostupná přímo IP adresa zařízení, ale jiná adresa, často však ze zařízení v blízkosti skutečného uživatele. Možností, jak skrýt svoji identitu a tím ztížit lokalizaci, je použít anonymizační sítě – např. Tor (The Onion Router) [19] nebo využít služeb VPN (Virtual Private Network).

Tato práce se věnuje problému ověření geografické polohy zjištěné z databáze pomocí nově navržené aktivní geolokační metody. Po úvodu do problematiky v kapitole 1, jsou v kapitole 2 podrobně rozebrány veškeré současné přístupy k nalezení polohy stanice při znalosti IP adresy – tzv. IP geolokace. Kromě toho jsou v této kapitole (2) popsány jednotlivé zdroje zpoždění komunikačního řetězce. Následující kapitola 3 uvádí cíl práce, kterým je návrh nové geolokační metody, která s jistotou určí region, ve kterém se stanice nachází a je možné ji použít pro identifikaci chybných záznamů v geolokačních databázích. Kapitola 4 obsahuje analýzu a měření vlastností zpoždění komunikačního řetězce, které přímo ovlivňují výpočet vzdálenosti. Následně je v kapitole 5 uveden popis výsledků dlouhodobého měření zpoždění, za účelem zjistit vliv dříve naměřených dat na přesnost geolokačních metod. Kapitola 6 se pak podrobně věnuje návrhu nové metody i popisu geodetických výpočtů pro určení cílových souřadnic na povrchu Země. Použití této navržené metody v praxi je uvedeno v kapitole 7, kde je ověřeno 5000 záznamů ze tří různých geolokačních databází. Závěru a diskuzi výsledků je věnována kapitola 8.

2 PŘEHLED SOUČASNÉHO STAVU PROBLEMATIKY

V současnosti je pro geolokaci elektronického zařízení používáno několik přístupů. Způsoby geolokace můžeme rozdělit do tří základních skupin, dle jejich principu. První skupina zahrnuje zařízení obsahující rádiový přijímač, který je buď primárně určen ke geolokaci (GPS) nebo je geolokace jeho vedlejší funkcí (GSM, WiFi). Pro další dvě skupiny se obvykle užívá pojem IP geolokace – nalezení geografické pozice stanice za použití IP adresy. Tyto metody se dají rozdělit na pasivní a aktivní. Pasivní metody využívají statických záznamů v databázích, kdežto aktivní metody provádějí měření a na jeho základě je rozhodují o poloze. V praxi se můžeme setkat i s geolokací za využití kombinace těchto dvou přístupů.

Výsledkem geolokace je poloha, kterou je možné určit v několika úrovních – kontinent, stát, region, město, PSČ, přesná adresa a geografické souřadnice (např. ve formátu WGS84). Pasivní IP geolokační služby fungují často na komerční bázi, a tak úroveň přesnosti pozice stanice je obvykle úměrná ceně. Aktivní IP geolokace a geolokace s rádiovým přijímačem většinou vrací polohu ve formátu zeměpisných souřadnic s určitou tolerancí.

Přesnost zjištěné polohy závisí především možnostech zařízení, které k lokalizaci používáme. Pokud je zařízení vybaveno rádiovým přijímačem, přesnost lokalizace je obvykle vyšší (řádově desítky až stovky metrů). Pasivní geolokace dosahuje přesnosti na úrovni města až regionu – dle vyspělosti země a pokročilosti databáze. Nejméně přesnou metodou je v dnešní době aktivní IP geolokace, kde se medián chyby pohybuje v desítkách až stovkách kilometrů, přesnost závisí především na konektivitě měřené stanice a rozmístění referenčních bodů (tzv. landmarků).

Při určování polohy na základě IP adresy může vzniknout chyba, pokud lokalizujeme IP adresu, která ve skutečnosti zařízení nepatří. S těmito případy se setkáváme v denní praxi například, protože mnoho poskytovatelů připojení (ISP) nepřiděluje uživatelům tzv. veřejnou IP adresu, ale pouze privátní (dle RFC 1918 [63]). Ta je následně překládaná na veřejnou IP adresu pomocí techniky NAT (Network Address Translation). Pro účely IP geolokace je poté použita veřejná IP adresa, která nemusí být v lokalitě uživatele. Na druhou stranu k NATu zpravidla dochází ve směrovači blízko uživatele, takže chyba IP geolokace nemusí být v tomto případě velká. Druhým problém je rozšiřující se anonymizace adres, například pomocí Tor, využitím Proxy serveru nebo VPN připojení. V těchto případech je IP geolokace provedena pro IP adresu zařízení, ze kterého provoz z těchto sítí vystupuje. V těchto případech je takovéto zařízení (Tor uzel, Proxy server, VPN koncentrátor) často ve velké vzdálenosti od skutečného uživatele. Posledním problémem, v dnešní době vzrůstajícím

na významu, jsou různé formy loadbalancingu (rozklad zátěže mezi více zařízení) a také směrování typu Anycast (směrování provozu k nejbližšímu z mnoha zařízení). Tímto vznikne problém existence více zařízení se stejnou IP adresou na různých lokalitách. Pro geolokační databáze to znamená evidování záznamu o více polohách jedné IP adresy. Pro aktivní IP geolokační metody je toto dokonce kritickým problémem, protože není možné jednoznačně určit, který stroj je aktuálně měřen. V další práci byla snaha vyhnout se případům popsaným výše, a to především pečlivým výběrem všech použitých stanic.

2.1 Geolokace zařízení s rádiovým přijímačem

Lokalizace zařízení, jež obsahují rádiový přijímač, se provádí především na základě měření síly, směru a zpoždění přijímaného signálu. Mezi nejpoužívanější patří družicové polohové systémy, buňkové systémy pro mobilní komunikaci a bezdrátové lokální (případně metropolitní) sítě.

2.1.1 Družicové systémy pro lokalizaci

GPS a další globální družicové polohové systémy (Glonass, Galileo, Beidou) potřebují k lokalizaci informace o čase a aktuální poloze nejméně čtyř družic na oběžné dráze. Z těchto informací systém dopočítá zpoždění signálu od družice a pomocí multilaterace dojde k velmi přesnému určení polohy. Přesnost GPS a podobných systémů se pohybuje řádech metrů, avšak za použití korekčních informací z pozemních stanic je možné přesnost zvýšit až na jednotky milimetrů [47]. Základním předpokladem pro správné fungování je venkovní anténa.

2.1.2 Systémy mobilní komunikace

Geolokace pomocí GSM systému (případně jeho dalších generací) využívá informací o aktuální síle signálu okolních buněk (BTS). Dle identifikátoru buňky (CID) vysílače GSM signálu je prohledána databáze obsahující její pozici[1]. Po zkombinování údaje o poloze dostupných buněk se sílou jejich signálu je rozhodnuto o geografické pozici zařízení [64]. Tento způsob dosahuje přesnosti v řádech desítek až stovek metrů [75, 3] a je závislý na počtu okolních buněk a údajích o jejich poloze v databázi. Geolokaci v mobilních sítích 5. generace je v současnosti věnováno velká pozornosti např. v [30].

[1] Viz například neoficiální databáze českých BTS na stránce gsmweb.cz

2.1.3 Bezdrátové lokální sítě

WiFi je komerční označení pro standardy IEEE 802.11, které jsou používány v lokálních bezdrátových počítačových sítích. K nalezení pozice takovéto stanice se využívá informací o dostupných přístupových bodech a síle jejich signálu[2]. Dle SSID (Service Set Identifier) a BSSID (tvořené MAC adresou bezdrátové karty) adresy je prohledána databáze přístupových bodů a výsledná pozice je přisouzena poloze přístupového bodu v databázi s nejlepším aktuálním signálem [61]. Přesnost je závislá na dosahu bezdrátového signálu WiFi sítí, proto je obvykle dosahováno přesnosti v řádech desítek metrů.

2.2 Pasivní IP geolokační služby

Pasivní IP geolokace je založena na vyhledávání záznamu o pozici bez použití aktivního měření. Velmi často jedná o databázi IP adres a příslušných geografických údajů, další možností je použít informací ze systému DNS nebo z databáze registrátorů IP adres. Protože jsou všechny tyto systémy založeny na statických záznamech, jsou velice náročné na správu a dochází u nich k systémovým chybám. Často se v databázi nachází chybný záznam – například stejný pro skupinu IP adres, které patří zařízením nacházejícím se na různých místech. Tyto metody mají také problém s konvergencí při přesunu stanice do nové lokace. Oproti těmto nevýhodám je hlavní výhodou rychlost a v případě správného záznamu také přesnost.

2.2.1 IP adresy a způsoby jejich přidělování

Většina geolokačních databází vychází z údajů poskytovaných registrátory IP adres. Zodpovědnost za přidělování veřejných IP adres a čísel autonomních systémů (AS) má organizace IANA (Internet Asigned Numbers Authority), která je rozdělena na pět dílčích koordinačních středisek dle lokality – RIR (Regional Internet Registry)[3]. Záznam v RIR databázi obsahuje rozsah přidělených IP adres a název organizace, jež má tyto IP adresy registrovány. Velmi často se však sídlo registrující organizace neshoduje se skutečnou polohou všech stanic z přiděleného rozsahu IP adres, především z důvodu přesunů IPv4 adres mezi různými lokalitami. Naproti tomu IPv6 má obrovský adresní prostor, kde není nutné přesouvat adresní rozsahy, na druhou stranu však narůstají nároky na velikost databází těchto adres.

Příchod protokolu IPv6 však zavedl také možnost mobility adresy (a to i pro protokol IPv4), což znamená dostupnost stanice pod jednou IP adresou kdekoliv se

[2]Jedna mnoha databází je například wigle.net.

[3]Pro Evropu a část Asie se jedná o síťové koordinační centrum RIPE NCC (Reseaux IP Europeens Network Coordination Centre).

nachází [65]. Naštěstí tento mechanismus není v praxi téměř využíván a tak IP geolokaci neovlivní. Další novinkou v IPv6 je možnost využít směrování typu Anycast, které umožňuje mít více strojů se stejnou IP adresou na různých lokalitách. Přestože je tento mechanismus často využíván pro důležité servery (např. pro kořenové DNS servery) při IP geolokaci jde častěji lokalizaci uživatelských strojů.

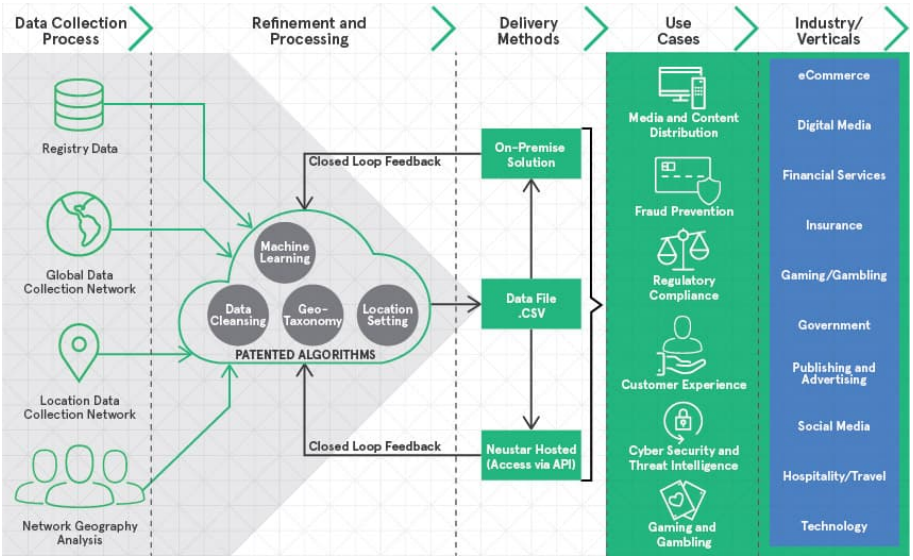
2.2.2 Analýza DNS záznamů

Mezi pasivní metody pro zjištění polohy IP adresy je možné zařadit i analýzu DNS (Domain Name System) záznamů. Použitím zpětného (reverzního) překladu DNS lze pro IP adresu zjistit doménové jméno a z něj je možné vyčíst indicie o poloze. Některé koncové stanice a směrovače mají ve svém doménovém jménu uvedenu zkratku označující město, ve kterém se nachází. Například `10gigabitethernet1-3.core1.prg1.he.net` pravděpodobně označuje směrovač nacházející se v Praze (zkratka PRG) [6]. Tyto záznamy však nejsou nijak zaručeny, aby vznikl chybný záznam, stačí přestěhovat směrovač do nové lokality a nezměnit jeho doménové jméno. V doménových jménech bývají často zakomponovány mezinárodní zkratky států, měst, letišť či meteorologických stanic. Na analýze rekurzivních dotazů na reverzní DNS záznam je postavená metoda GeoTrack [56] a tohoto principu je využito v [72], kde jsou reverzní DNS záznamy pro mezilehlé routery zjištěné pomocí `traceroute`.

Druhou možností, jak využít DNS, je rozšíření definované v RFC 1876 [15], které dovoluje přidat záznam o poloze DNS serveru. Tento záznam se nazývá *Loc* záznam a obsahuje položky: zeměpisná šířka, zeměpisná délka, nadmořská výška, velikost oblasti a přesnost pozice. Bohužel většina DNS serverů nemá tento parametr vyplněn a pokud je vyplněn, tak není nijak ověřený, takže ke geolokaci jej lze použít pouze jako pomocný údaj.

2.2.3 Databáze pro IP geolokaci

V současnosti nejrozšířenější je geolokace pomocí databáze IP adres, a to především díky své jednoduchosti a dobré přesnosti. Existuje celá řada databází, které se liší mezi sebou především kvalitou a počtem záznamů. Některé databáze jsou volně dostupné, jiné vlastní specializované firmy a využití záznamů je placené, případně v omezené míře veřejné. Plnění těchto databází je většinou patentováno (např. [57]) a jedná se o složité mechanizmy data-miningu [26] a analýzy údajů o poloze vyplněných uživateli. Pro zlepšení přesnosti se využívá také detekce anomálií, statistika, demografické údaje oblasti či kontrola pomocí aktivních IP geolokačních metod. Příklad, jak může být pojato plnění geolokační databázi, je na obrázku 2.1, který pochází ze stránek firmy Neustar poskytující svoji databázi IP Intelligence.



Obr. 2.1: Způsob zpracování dat pro IP geolokační databázi IP Intelligence. Převzato z [54]

Vzhledem k různému využití geolokačních databází tyto poskytují různé úrovně informací o IP adrese. V případě, že je nutné například omezit přístup z jiných zemí, postačuje znalost státu, což bývá nejjednodušší verze výstupu. Další verze databází poskytují informace o regionu, městu, PSČ, zeměpisných souřadnicích, typu a rychlosti připojení, poskytovateli připojení (ISP) nebo i demografické údaje jako velikost sídla, hustota zalidnění, bohatství, kriminalita,... Pro potřeby této práce budou využity především zeměpisné souřadnice uváděné u záznamů, přestože se tyto často vztahují k centru města, případně regionu ve kterém se tato IP nachází.

Některé geolokační služby umožňují stažení aktuální verze databáze a její využití ve vlastních programech. Častější je ale zobrazení výstupu na webové stránce nebo prostřednictvím API – např. pomocí standardizovaného rozhraní *W3C Geolocation API* [60]. Podpora IP geolokace je již také zabudovaná v současném standardu HTML verze 5 [29].

V následujících kapitolách jsou uvedeny významné geolokační databáze, výčet však není úplný, neboť nové služby mohou vznikat každý den a pro různé lokality (např. Čína) existují speciální databáze.

GeoIP2

Společnost MaxMind disponuje databází GeoIP2, jejíž obsah je vytvářen analýzou informací o poloze, jež o sobě uživatelé sami vyplnili na různých stránkách a dle [50] obsahuje 99.9999 % veškerých používaných IPv4 a IPv6 adres. GeoIP2 poskytuje tři placené úrovně informací – nejlevnější (stojí \$0,000 1 za záznam) poskytuje pouze

kontinent a zemi ve které se IP adresa nachází. Vyšší úroveň ($0,000 4 za záznam) poskytuje navíc město, PSČ, zeměpisné souřadnice, přesnost a další obdobné informace. Nejdražší verze ($0,002 za záznam) obsahuje navíc informace o typu uživatele, jeho příjmu a další demografické informace. Kromě toho MaxMind poskytuje geolokaci i bezplatně, a to prostřednictvím databáze GeoLite2, která obsahuje informace o zemi a městě, ve kterém se IP adresa nachází. Bezplatná verze je však méně přesná a méně často aktualizovaná.

IP2Location

IP2Location je geolokační databáze pro zjištění různých dat o poloze IPv4 i IPv6 adresy. Celkem je k dispozici 24 různých variant poskytovaných informací počínaje informací o zemi za $49 za rok [34]. Oblíbená verze stojí desetkrát tolik a zahrnuje zemi, region, město, PSČ a zeměpisné souřadnice. Nejvyšší verze stojí $1849 za rok, a kromě již zmíněného poskytuje například informace o časovém pásmu, ISP, rychlosti připojení a kódu nejbližší meteorologické stanice. Databázi je možné si vyzkoušet, zdarma je poskytováno 50 záznamů denně (pro registrované uživatele 200 záznamů denně). Kromě komerční verze a demo přístupů je možné využít i Open source verzi databáze s názvem IP2Location Lite [33], která má však deklarovanou nižší přesnost a stejnou polohu uvádí vždy pro blok 256 IP adres (IPv4 prefix /24).

Komerční verze databáze dle [34] určí správně zemi ve $> 99{,}5$ % případů (město ve > 80 %, oproti tomu Lite verze má přesnost státu > 98 % a města $> 60\%$. Dle webových stránek [34] je každý měsíc aktualizováno 15 % záznamů, což znamená průběžné zpřesňování databáze, případně aktualizace některých záznamů z důvodu přesunu zařízení nebo změny jeho IP adresy.

NetAcuity

Firma Digital Element nabízí již od roku 1999 geolokační data, která v současnosti prodává pod názvem NetAcuity [17]. Stejně jako ostatní databáze poskytuje různé údaje počínaje státem, přes město, zeměpisné souřadnice a typ připojení až k demografickým údajům. Již od roku 2011 databáze obsahuje i IPv6 adresy a v současnosti pokrývá 99,9999 % všech fungujících IP adres.

DB-IP

DB-IP obsahuje cca 19 miliónů IP adres (verzí 4 i 6) především ze Spojených států amerických a dle [16] jsou každý měsíc doplněny nebo opraveny milióny záznamů[4].

[4]Například v červnu 2017 bylo přidány skoro 3 milióny adres, 1 milión adres byl opraven a přibližně 1 milión adres byl odebrán.

Základní informace (stát a město) jsou z databáze poskytovány zdarma, další informace (zeměpisné souřadnice, časové pásmo, ISP a typ připojení) stojí $99 případně $189 za rok.

IPinfo

Geolokační API poskytované stránkou ipinfo.io má vlastní IP geolokační databázi, obsahující jak IPv4, tak IPv6 adresy. Tato databáze byla vytvořena na základu GeoLite2 databáze od MaxMind a stále obsahuje zhruba polovinu záznamů stejných [35]. Úrovně přesnosti jsou obdobné – země, město, PSČ, zeměpisné souřadnice a další. Ceny jsou závislé na počtu dotazů do API za měsíc – 1000 stojí $10 za měsíc, ceny dále vzrůstají až k $400 za měsíc za 320 000 dotazů do API. Bez e-mailové podpory je možné získat i zdarma přístup do databáze, ale pouze do 1000 dotazů na API měsíčně.

IP Intelligence

IP Intelligence je geolokační databáze poskytovaná firmou Neustar. Tato databáze obsahuje informace o 99,99 % ze všech veřejně směrovatelných IPv4 i IPv6 adresách [55]. U každé adresy je evidováno až 30 různých parametrů včetně státu, města i zeměpisných souřadnic. Kromě toho je evidováno i tzv. IP reputation, které indikuje, nakolik může být komunikace s danou adresou ohrožující. Poskytované údaje jsou ve třídách bronze, silver a gold, přičemž všechny obsahují informace o poloze a typu připojení. Silver navíc obsahuje i vlastníka adresy a gold ještě údaje o případné anonymizaci (Tor a Proxy).

IPligence

IP geolokační databáze od IPligence [36] poskytuje data ve třech placených úrovních – Lite, Max a Pro. Tyto se liší cenou (od $39 do $299 za rok updatů) a poskytovanými informacemi. V Lite verzi je pouze kontinent a země, ve verzi Pro jsou mimo jiné město, PSČ a zeměpisné souřadnice. Na stránce je možné navíc využít zdarma službu pro lokalizaci až 30 adres v jediném dotazu.

Geobytes

Geobytes je jeden z nejstarších (od roku 1999) poskytovatelů geolokačních informací na Internetu a obsahuje informace o poloze všech rozsahů IPv4 adres, které se objevují v BGP tabulkách směrovačů v Internetu [21]. Kromě informací o poloze (stát, město, zeměpisné souřadnice) obsahuje Geobytes databáze i demografické informace jako (národnost, měnu, populaci a další). Informace z databáze jsou poskytovány

zdarma až do 16 384 dotazů za hodinu, více je však možné si zaplatit pomocí VIP přístupu ($9.99 za 100 000 dotazů). Dle [21] má databáze přesnost 97 % ve správném určení země a 75 % v určení města (s tolerancí 50 km).

HostIP.Info

HostIP.Info [31] je IP geolokační databáze fungující na principu Open source. Data jsou získávána od dobrovolníků a kdokoliv má možnost nahlásit špatně určenou adresu nebo přidat chybějící záznam. Databáze eviduje informace pouze k blokům 256 IPv4 adres (prefix /24), díky čemuž nedokáže postihnout případy, kdy je tento prefix podsíťován. HostIP.Info poskytuje informace o zemi a městě, ve kterém se IP nachází a dle [31] je denně aktualizovaná.

Software77

Software77 je jedna z prvních (rok vzniku 2004) IP geolokačních databází a je poskytovaná firmou WebNet77 [67]. V dnešní době obsahuje již i adresy IPv6. Databáze je poskytována zdarma pod licencí *Donationware*. Služba obsahuje pouze mapování IP adresy na úroveň státu a je v ní aktualizováno cca 50 záznamů denně [67].

EurekAPI

Služba IP-GeoLoc na stránce eurekapi.com poskytuje pomocí API geolokační data ve třech různých edicích – basic, standard a professional. Edice se liší množstvím poskytovaných informací – počínaje zemí a regionem, konče městem, PSČ, zeměpisnými souřadnicemi a poskytovatelem připojení. Tyto verze se liší také cenou – nejlevnější basic stojí $15 měsíčně, nejdražší professional $30.

2.2.4 Porovnání IP geolokačních databází

Rozdíly mezi výše popsanými databázemi jsou nejen v ceně za jejich použití, ale také v množství záznamů, které obsahují. Většina databází vychází z údajů lokálních registrátorů IP adres (RIR) a tak obsahují veškeré existující veřejné IP adresy, některé z nich však s údaji převzatými od RIR. Ne všechny databáze také obsahují i pomalu se rozšiřující protokol IPv6, ale z větších komerčních databází jej podporují všechny. Zásadní rozdílem mezi databázemi je však přesnost výstupů z nich. Porovnání přesnosti deklarované na webových stránkách některých[5] z nich jsou zobrazeny v tabulce 2.1, která srovnává spolehlivost správného určení země a města (s tolerancí 50 km). Vzhledem k tomu, že tyto údaje poskytují sami provozovatelé databází, může jít o údaje nadsazené případně vypočítané dle přizpůsobené (nezveřejněné) metodiky.

[5]Ne všechny geolokační služby poskytují informace o jejich přesnosti.

Tab. 2.1: Přesnost IP geolokačních databází dle informací z jejich webových stránek.

	země	město (tolerance 50 km)
GeoIP2	99,8 %	81,0 %
IP2Location	99,5 %	80,0 %
NetAcuity	99,9 %	97,0 %
Geobytes	97,0 %	75,0 %

Seriózním porovnáním přesnosti IP geolokačních databází se zabývá několik publikací, neexistuje však žádné důvěryhodné nebo pravidelné srovnání. Prvnímu většímu srovnání se věnoval Poese et. al [59], který srovnal v roce 2011 největší IP geolokační databáze té doby – GeoIP (předchůdce GeoIP2 od MaxMind), InfoDB, IP2Location, Software77 a HostIP.Info. Nejprve je v článku řešeno, jak jsou v databázích organizovány prefixy adres a že pro mnoho z nich odpovídají rozdělení prefixů přidělených RIR. Toto značí, že údaje vychází především z údajů RIR a nezohledňují podsíťování v lokalitách. Ve srovnání přesnosti nejlépe vyšla databáze GeoIP následovaná InfoDB a IP2Location.

Další studie přesnosti databází [66] zahrnuje IP2Location, GeoIP (od MaxMind), GeoBytes, NetAcuity, HostIP.Info, IPligence a také aktivní metodu Spotter [68]. Spolehlivost určení země se u všech případů pohybuje mezi 80–97 %, přičemž nejlepšího výsledku dosáhla databáze NetAcuity. Tato databáze se také nejlépe vypořádala s určením města (79 %) u ostatních databází se přesnost pohybovala okolo 20 %, pouze IPligence měla přesnost určení města necelé jedno procento. Zajímavé je, že aktivní metoda Spotter správně určila město v téměř 28 % případů, což byl třetí nejlepší výsledek.

Center for Applied Internet Data Analysis (CAIDA) vytvořilo v roce 2011 srovnání veřejných a komerčních IP geolokačních databází [32]. Jako veřejné průzkum považuje data od RIR, Software77, HostIP.Info, GeoLite (od MaxMind) a InfoDB (dnešní IP2Location Lite); za komerční pak IPligence, Cyscape, GeoIP a Digital Envoy (dnes NetAcuity). Nejmenší chybu od skutečné lokality dosáhla databáze od Digital Envoy, následovaná IPligence a GeoIP databází. Podrobné výsledky jsou k nalezení v tabulce 2.2, která vychází z dat z [32]. Z veřejných databází nejlépe dopadla databáze GeoLite. Databáze HostIP.Info obsahovala velmi malé množství všech adres (přibližně 16 %) a databáze od Sofware77 se zásadně (pouze 4,5 % rozdílných záznamů) nelišila od veřejných dat RIR.

Druhá studie [76] od CAIDA se věnuje porovnání alokovaných bloků adres z RIR a jejich použití v databázi GeoIP jak pro IPv4, tak i pro IPv6. Databáze GeoIP v té době (2012) nepokrývala pouze 0,4 % alokovaných IPv6 adres. Z pohledu přesnosti

Tab. 2.2: Přesnost IP geolokačních databází dle studie CAIDA [32].

	země	město (tolerance 40 km)
HostIP.Info	94,5 %	67,0 %
IPligence	94,3 %	78,0 %
Cyscape	98,4 %	–
GeoIP	99,1 %	78,0 %
GeoLite	98,9 %	75,0 %
Digital Envoy	96,7 %	93,0 %

určení země byla přesnost databáze GeoIP pro IPv4 vyšší (cca o 5 %) než informace z RIR, avšak pro IPv6 se přesnost téměř nelišila.

Srovnání čínských IP geolokačních databází poskytuje [48] z roku 2015, které srovnává tradiční databázi IP2Location a čtyři lokální databáze Chunzhen, Taobao, Sina a IP138. Přesnost určení země se pro všechny databáze pohybovala okolo 99 %, určení města pak pro databáze Sina a IP138 dosahovalo jistoty u 96,7 % respektive 95 % adres. V určení města byla nejhorší celosvětová databáze IP2Location, která měla 81,4 % správných záznamů.

Poslední srovnání geolokačních databází je [40] z roku 2016, ve kterém jsou porovnány IPv4 a IPv6 databáze DB-IP, IP2Location a GeoIP2. Celkem bylo porovnáno 3206 stanic s IPv4 i IPv6 adresou, přičemž databáze DB-IP a GeoIP2 neobsahovaly 7 % respektive 8 % těchto adres. Z hlediska přesnosti určení místa (s tolerancí do 50 km) dosáhla nejlepších výsledků databáze IP2Location – 61 % správně určených pozic pro IPv4 a 35 % pro IPv6. Z celkových výsledků je zřejmé, že databázové záznamy pro IPv6 adresy stále zaostávají za IPv4 adresami.

Závěr z výše prezentovaných dat je, že dnešní geolokační databáze dokáží s velkou jistotou (nad 95 %) určit zemi, ve které se IP adresa nachází, problematické je stále přesné určení města, kde se přesnost u lepších databází pohybuje okolo 75 %, obvykle však kolem 50 %. Pro zlepšení přesnosti IP geolokačních databází mohou posloužit tzv. aktivní metody, kterým je věnována následující část a také většina dizertační práce.

2.3 Aktivní IP geolokační metody

Aktivní IP geolokace je založena na měření zpoždění a případně dalších síťových parametrů mezi stanicí se známou polohou (referenčním bodem – landmarkem) a lokalizovanou stanicí. Většinou se k lokalizaci jedné stanice používá větší množství referenčních bodů (řádově desítky). Jelikož probíhá několik měření mezi referenčními

body a lokalizovanou stanicí, dochází k navýšení síťového provozu úměrně počtu měření a počtu referenčních bodů.

Princip většiny aktivních metod je založen na korelaci mezi zpožděním a geografickou vzdáleností. To je dáno vlivem zpoždění způsobeného rychlostí šíření signálu v médiu, které je hlavní složkou zpoždění na dlouhých vzdálenostech. Podrobněji se o zdrojích a obvyklé velikosti parciálních zpoždění se píše v kapitole 2.4.

2.3.1 Metoda GeoPing

Geoping [56] je nejstarší geolokační metoda založená na měření zpoždění. Ke své činnost potřebuje velké množství pasivních referenčních bodů (uzlů se známou polohou) a několik aktivních sond (uzlů provádějících měření). Nevýhodou této metody je určení výsledné polohy jako místa, kde leží jeden z referenčních bodů, čímž je omezena přesnost metody. Proto je důležité disponovat množinou s co největším počtem referenčních bodů, které jsou geograficky rovnoměrně rozloženy a jsou připojeny spolehlivým vysokorychlostním spojem. Dále je nutné mít M aktivních sond (doporučeno 7–9 [56]), které dokáží změřit dobu zpoždění k jednotlivým referenčním bodům a cílové stanici. I tyto sondy by měly být geograficky rovnoměrně rozmístěny.

Princip metody je v porovnání vektorů zpoždění příslušejících referenčním bodům ($\mathbf{DV}$) a lokalizované stanici ($\mathbf{DV}'$). Vektor zpoždění obsahuje změřenou dobu přenosu informace mezi referenčním bodem a všemi sondami. Stejný vektor je změřen pro lokalizovanou stanici a následně je srovnán s vektory referenčních bodů k nalezení nejvíce podobného vektoru. Pro určení nejpodobnějšího vektoru je vytvořen M rozměrný prostor (rovný počtu sond), v němž je nalezen vektor s nejmenší eukleidovskou vzdáleností k hledanému vektoru [56]. Výpočet eukleidovské vzdálenosti je proveden pomocí

$$\mathrm{d}(\mathbf{DV}, \mathbf{DV}') = \sqrt{\sum_{i=0}^{M-1} (t_i - t'_i)^2}, \tag{2.1}$$

kde t_i je zpoždění mezi i-tou sondou a referenčním bodem a t'_i je zpoždění mezi i-tou sondou a lokalizovanou stanicí. Výsledná poloha stanice je následně určena jako poloha referenčního bodu s nejnižší eukleidovskou vzdáleností. Dle autorů [56] je medián chyby metody GeoPing 382 km.

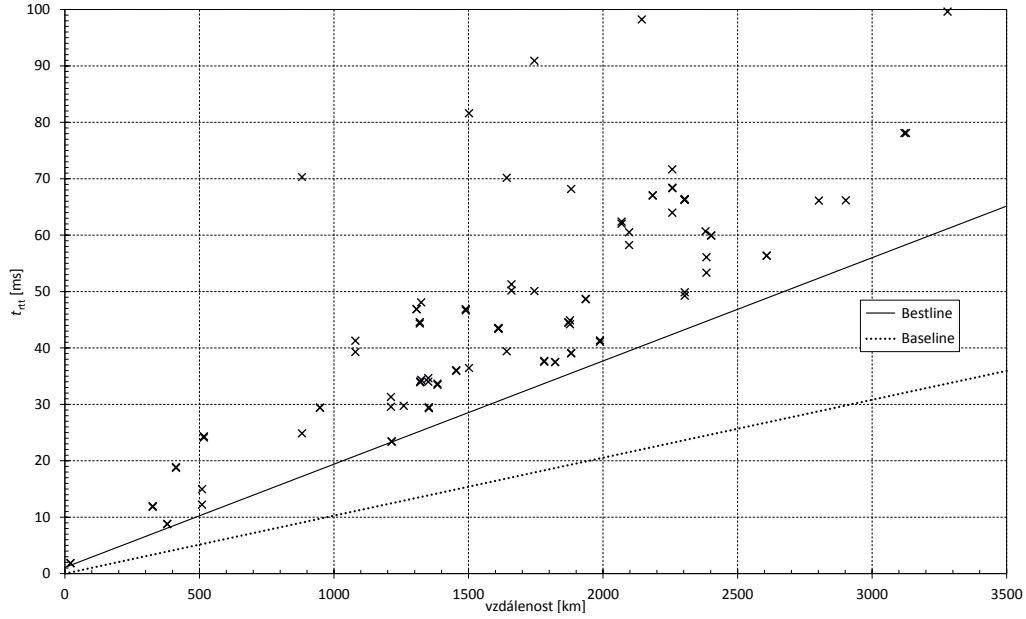
2.3.2 Metoda ShortestPing

Principiálně nejjednodušší IP geolokační metodou založenou na měření zpoždění je ShortestPing [39]. Tato metoda vyžaduje velké množství rovnoměrně rozmístěných referenčních bodů se známou polohou. Metoda zjišťuje zpoždění mezi lokalizovanou stanicí a všemi referenčními body, výsledná pozice je přisouzena poloze referenčního

bodu s nejmenší hodnotou zpoždění. Přestože je tato metoda jednoduchá a výslednou pozici přisuzuje jednomu z referenčních bodů, v některých případech dosahuje tato metoda lepších výsledků než některé složitější metody (např. GeoPing). V publikaci [20] je uveden medián chyby 55 km a průměrná chyba 106 km.

2.3.3 Metoda Constraint Based Geolocation

Constraint Based Geolocation (CBG) [24] ke své činnosti využívá multilaterace známé z rádiového určování polohy. Princip metody tkví ve využití vztahu mezi geografickou vzdáleností a zpožděním k vytvoření tzv. hranice nejvzdálenějšího možného umístění stanice. Tato hranice je určena přepočtem zpoždění na základě tzv. Bestline, což je přímka vytvořená při kalibraci a udává vztah mezi zpožděním a vzdáleností pro příslušný referenční bod. CBG tedy pro činnost potřebuje množinu aktivních referenčních bodů se známou polohou (landmarků).



Obr. 2.2: Graf zpoždění v závislosti na geografické vzdálenosti zjištěný při kalibraci metody CBG.

Před měřením je provedena kalibrace – každý landmark změří zpoždění k ostatním referenčním bodů a k naměřené hodnotě zpoždění $t_{i,j}$ přiřadí geografickou vzdálenost $l_{i,j}$ (viz obrázek 2.2). Pro každý landmark je pak je nalezena přímka (tzv. Bestline) s rovnicí

$$t_{i,j} = m_i l_{i,j} + b_i, \tag{2.2}$$

která leží pod všemi body grafu a zároveň k nim má nejblíže – tím reprezentuje největší poměr zpoždění a vzdálenosti zjištěné kalibrací [24]. Druhá přímka vyznačená

v obrázku 2.2 je tzv. Baseline, která reprezentuje nejzazší fyzicky možnou vzdálenost pro naměřené zpoždění – bere v úvahu jen zpoždění vzniklé rychlostí šíření signálu v optickém vlákně (pomocí konstanty $\frac{2}{3}$c).

K nalezení rovnice přímky 2.2 je třeba využít poznatků z problematiky lineárního programování a nalézt následující minimum

$$\min_{\substack{b_i \geq 0 \\ m_i \geq m}} \left(\sum_{i \neq j} l_{i,j} - m_i t_{i,j} - b_i \right), \tag{2.3}$$

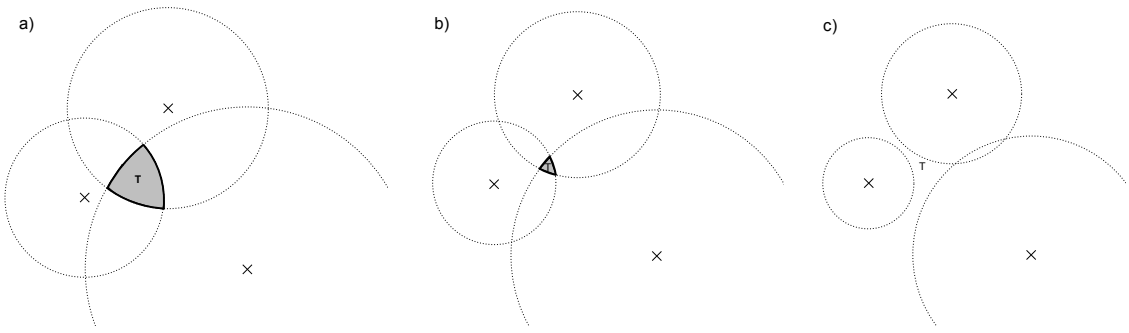
za podmínek nezáporného absolutního členu b_i a lineárního kvocientu m_i většího, než je kvocient Baseline přímky ($m = \frac{4}{3}$c). Při znalostí hodnot b_i a m_i pro i-tý landmark je možné přepočítat naměřené obousměrné zpoždění ($t_{i,T}$) na vzdálenost $l_{i,T}$ mezi landmarkem (i) a targetem (T) pomocí

$$l_{i,T} = \frac{t_{i,T} - b_i}{m_i}. \tag{2.4}$$

Samotná lokalizace pak probíhá tak, že každý referenční bod změří zpoždění k cílové stanici. Toto zpoždění následně referenční bod přepočítá dle rovnice 2.4 pomocí vlastní Bestline přímky na vzdálenost, která se rovná poloměru kruhu, ve kterém se cílová stanice nachází. Cílová pozice stanice je pak určena průnikem kruhů jednotlivých referenčních bodů a nalezením těžiště této oblasti průniku. Velikost průniku určuje také chybovou oblast, ve které se cílová stanice může nacházet. Na obrázku 2.3 b) je průnik kruhů – oblast, kde se nachází cílová stanice. Dle autorů metody [24] CBG je medián chyby pro USA roven 130 km a průměrnou chybu 209 km, pro evropský dataset je to 42 km respektive 106 km.

2.3.4 Metoda Speed of Internet

Metoda Speed of Internet (SOI) [39] je založena na podobném principu jako CBG – vytvoření hranice nejzazší vzdálenosti, kde se cíl může nacházet. Ke své činnosti tedy také potřebuje množinu aktivních referenčních bodů se známou polohou. Rozdíl oproti CBG je v přepočítání zpoždění na vzdálenost, kdy je použita konstanta $\frac{4}{9}$c (Baseline) namísto přímky vypočítané z kalibračních dat [39]. SOI tedy nepotřebuje kalibrační měření, čímž je zmenšena zátěž sítě. Nevýhodou je poté menší přesnost, větší oblast průniku – obr. 2.3 a) a také možnost, že se kruhy neprotnou. To může nastat při podhodnocení vzdáleností, podobně jako na obrázku 2.3 c), kde neexistuje průnik oblastí a není tedy možné určit pozici cíle [24]. Dle autorů [39] je medián chyby metody SOI okolo 180 km.



Obr. 2.3: Zobrazení principu geolokačních metod založených na vytváření hranic (kruhů) okolo referenčních bodů (křížky). Průnik kruhů definuje oblast, ve které se cíl nachází. Na obrázku a) je nadhodnocení velikosti kruhů, zde je cíl bezpečně uvnitř průniku, b) zobrazuje minimalizování velikostí kruhů, stále však bezpečné pro lokalizaci a na c) je chyba při lokalizaci, kdy některé hranice (kruhy) byly podhodnoceny a průnik všech kruhů není možný stejně jako lokalizace cíle (T).

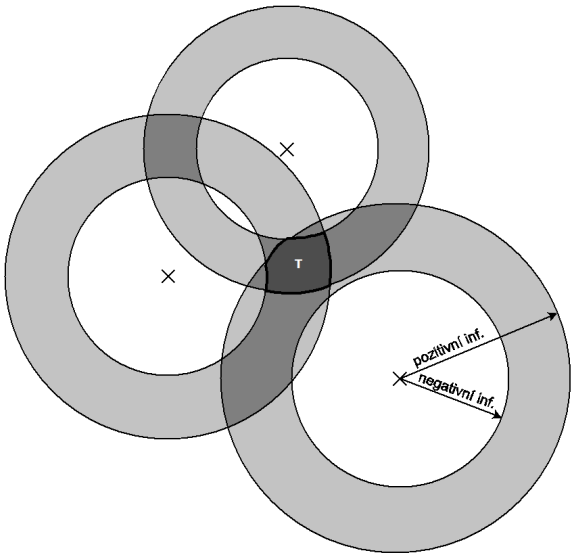
2.3.5 Metoda Topology Based Geolocation

Oproti ostatním metodám bere Topology Based Geolocation (TBG) [39] v úvahu také mezilehlé routery, které je možné zjistit pomocí nástroje `traceroute`. Tyto mezilehlé uzly jsou následně využity pro přesnější určení cíle. Za pomocí lokalizování uzlů po cestě je možné zmenšit region, ve kterém se cílová stanice nachází. Navíc je možné tyto mezilehlé uzly využít v dalších měřeních jako pasivní referenční body, případně pro ještě vyšší přesnost využít analýzy doménových jmen těchto uzlů k získání vyšší jistoty určení polohy routerů. Nevýhodou této metody je nemožnost využít automatizace při použití ruční analýzy doménových jmen. Medián přesnosti metody je bez použití pasivních landmarků 225 km, s jejich použitím 176 km a při prováděné analýze doménových jmen 67 km. Ve stejném pořadí je uvedena i průměrná chyba 253 km, 178 km a 138 km [39].

2.3.6 Metoda Octant

Metoda Octant [73] principiálně vychází z metody CBG, oproti ní však může referenční bod zjistit nejen oblast, kde se cílová stanice nachází, ale navíc i oblast, kde se cílová stanice nemůže nacházet. Tyto oblasti se označují jako pozitivní a negativní vzdálenosti. Negativní vzdálenost udává hranice, za kterými se stanice nemůže nacházet – jedná se o kruhovou oblast blízko referenčnímu bodu. Spojením s pozitivní informací (známou z CBG) vznikne mezikruží, ve kterém se cílová stanice může nacházet. Cílová poloha je pak určena jako průnik těchto mezikruží, čímž

může vzniknout i nekonvexní oblast (obrázek 2.4), která je pak popsána Beziérovými křivkami.



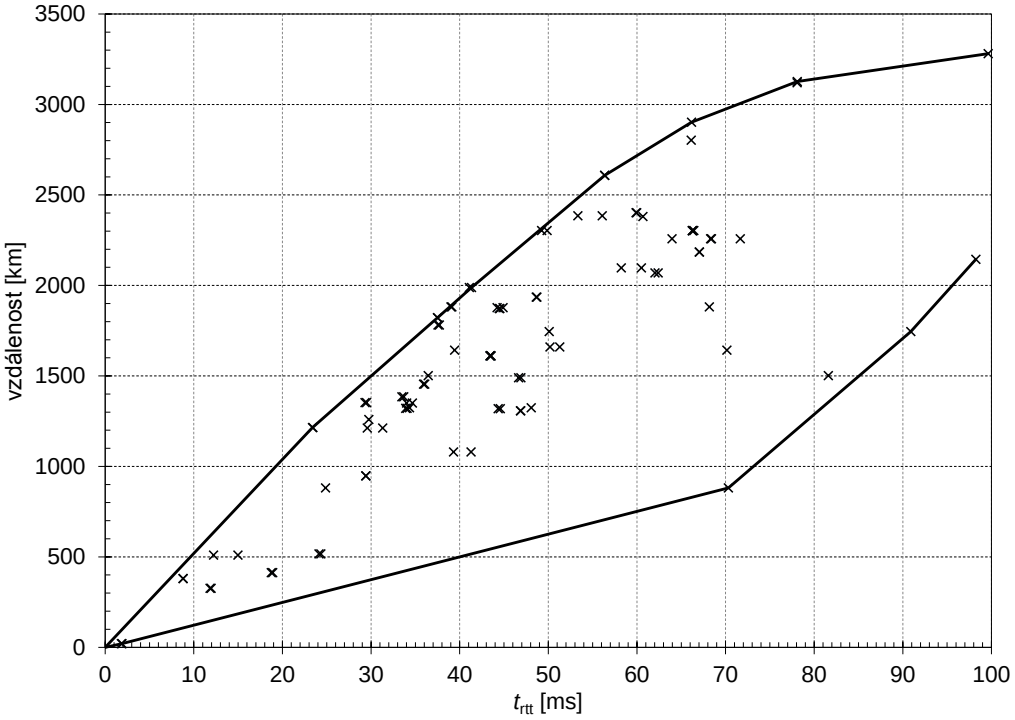
Obr. 2.4: Princip metody Octant založený na pozitivních a negativních informacích, oblast možného výskytu stanice je definována jako mezikruží, cílová poloha je poté určena jako průnik těchto mezikruží.

Pro přepočet zpoždění na vzdálenost se obdobně jako u CBG používá kalibračních dat mezi referenčními body, které jsou pro ilustraci vyneseny do grafu (obrázek 2.5). K přepočtu je však využita konvexní obálka všech změřených dat, pro která platí

$$r_L(t_{\text{rtt}}) \leq \|loc(L) - loc(i)\| \leq R_L(t_{\text{rtt}}) \text{ [73]}, \tag{2.5}$$

kde $\|loc(L) - loc(i)\|$ je vzdálenost mezi pozicí referenčního bodu $loc(L)$ a vzdálené stanice $loc(i)$. Horní část konvexní obálky $R_L(t_{\text{rtt}})$ popisuje přepočet obousměrného zpoždění t_{rtt} na pozitivní mezní hranice (obdobně jako u metody CBG). Dolní část konvexní obálky $r_L(t_{\text{rtt}})$ definuje negativní mezní hranice, kde se cílová stanice nenachází [73].

Octant umožňuje také jako negativní informaci použít obydlenost území a vyřadit tak z výsledku moře a další nepravděpodobné oblasti. Dále Octant zjištěnou polohu zpřesňuje použitím zpětného převodu DNS a hledáním polohy mezilehlých prvků (směrovačů). V publikaci [20] je uveden pro metodu Octant medián chyby 54 km a průměrná chyba má velikost 95 km.

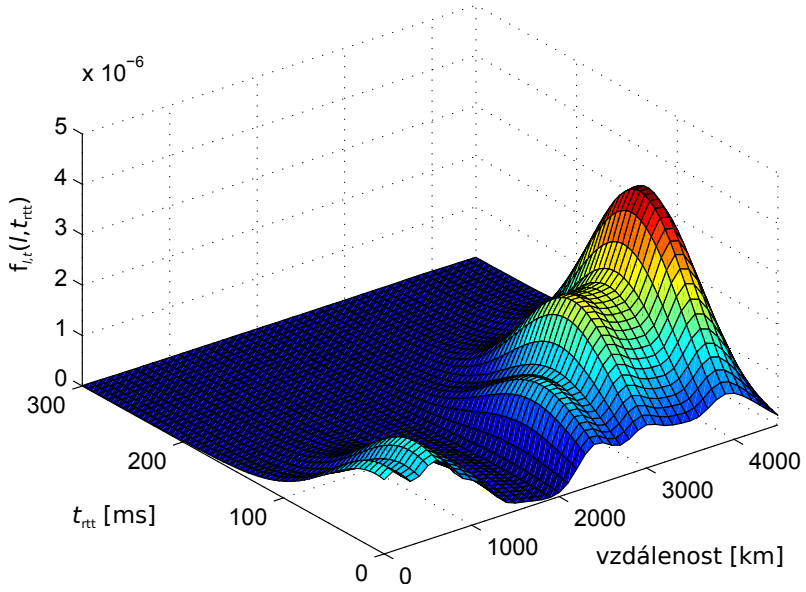


Obr. 2.5: Graf zpoždění v závislosti na geografické vzdálenosti zjištěný při kalibraci metody Octant. Plnou čarou je vyznačena konvexní obálka, která je použita pro výpočet pozitivních a negativních informací.

2.3.7 Metoda Statistical Geolocation

Statistical Geolocation (SG) [74] obdobně jako ostatní metody (CBG, Octant,...) nejprve změří kalibrační data každého referenčního bodu (landmarku). Tato data obsahují vzdálenost a obousměrné zpoždění (RTT) mezi jím samým a ostatními landmarky. Takto vzniknou páry hodnot (vzdálenost, zpoždění), které jsou následně vizualizovány trojrozměrného grafu (viz obrázek 2.6), který vyjadřuje statistickou pravděpodobnost výskytu daného zpoždění a vzdálenosti, jako sdruženou distribuční funkci (joint probability distribution function).

Následně jsou tyto hodnoty aproximovány pomocí jádrového odhadu hustoty (kernel density estimation) kvůli dalšímu využití při určování pravděpodobnosti vzdálenosti k lokalizované stanici. Pro zjištění polohy stanice je využito Force-directed algoritmu, který iterativně zkouší nejvíce pravděpodobné vzdálenosti přepočtené ze zpoždění pomocí jádrového odhadu hustoty jednotlivých landmarků [74]. Autoři metody v [74] uvádějí medián chyby 53 km a průměrnou chybu 92 km.



Obr. 2.6: Statistická pravděpodobnost výskytu daného zpoždění a vzdálenosti, vykreslená jako sdružená distribuční funkce (joint probability distribution function). Převzato z [74].

2.3.8 Metoda GeoWeight

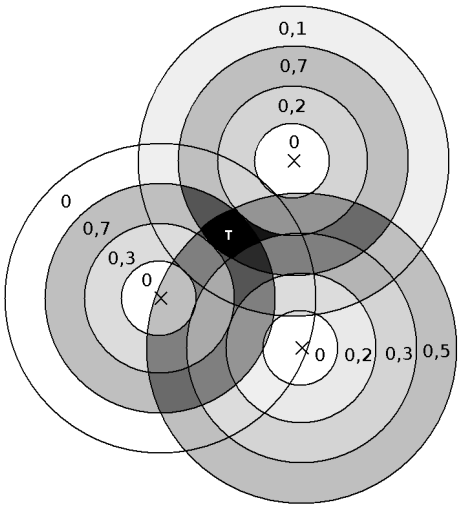
GeoWeight je metoda založená na principech metod CBG a Octant. Vylepšení přichází v rozdělení změřené oblasti možné polohy stanice na několik podoblastí s definovanou pravděpodobností výskytu stanice. Výsledná pozice je opět určena průnikem oblastí, jen se tentokrát jedná o protínající se mezikruží s nejvyšším součtem pravděpodobností – viz obrázek 2.7 [1].

Před první lokalizací je nutné provést kalibraci metody změřením zpoždění mezi referenčními body. Následně jsou vytvořena rovnoměrná pásma vzdáleností, kterým jsou přiřazena odpovídající naměřená zpoždění. Dle počtu přisouzených zpoždění je každé vzdálenosti přidělena odpovídající pravděpodobnost. Při lokalizování stanice jsou pak naměřenému zpoždění přiřazena pásma vzdáleností a jejich pravděpodobnosti.

Metoda GeoWeight má dle autorů [1] medián chyby 44 km a průměrnou chybu 170 km.

2.3.9 Metoda Posit

Posit [20] je metoda založená na rozložení pravděpodobnosti, které vytvořeno pomocí trénovaní na datasetu cílů se známou polohou. Pro lokalizovanou stanici je



Obr. 2.7: Metoda GeoWeight definuje pro různá rozmezí vzdáleností od referenčních bodů pravděpodobnosti výskytu stanice. Cílová oblast je určena průnikem prstenců s nejvyšším součtem pravděpodobností.

z referenčních stanic změřeno zpoždění, které je následně na základě rozložení pravděpodobnosti převedeno na vzdálenost. Průnik těchto vzdáleností od referenčních stanic vytvoří region, ve kterém se cíl nachází. Následně se v tomto regionu vypočítá logaritmická věrohodnostní funkce (log-likelihood) pro všechny v ní umístěné monitorovací stanice a landmarky. Monitorovací stanice nebo landmark s nejvyšší pravděpodobností je poté určen jako pozice hledané stanice. Autoři metody Posit [20] uvádějí medián chyby 32,9 km a průměrnou chybu 74,3 km.

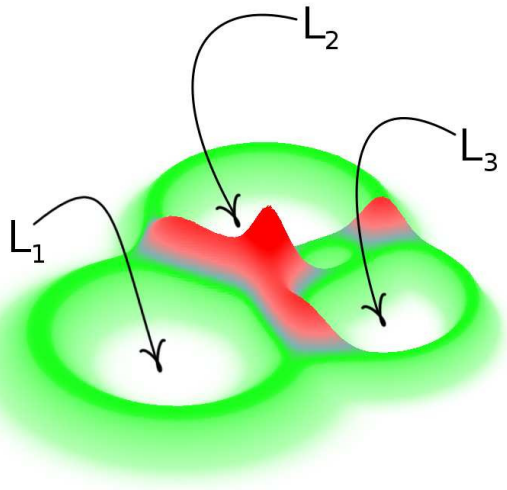
2.3.10 Metoda Spotter

Metoda Spotter [46] také vychází z metody CBG a využívá kalibrace mezi jednotlivými referenčními body. Kalibrační data jsou poté podrobena statistické analýze a na tomto základě je vytvořeno normální (Gaussovo) rozdělení pravděpodobnosti vzdáleností l, které má hustotu pravděpodobnosti $f_{t_{\mathrm{rtt}}}(l)$ definovanou dle

$$f_{t_{\mathrm{rtt}}}(l) \approx \frac{1}{\sqrt{2\pi}\sigma(t_{\mathrm{rtt}})} \cdot \mathrm{e}^{\left(-\frac{(l-\mu(t_{\mathrm{rtt}}))^2}{2\sigma^2(t_{\mathrm{rtt}})}\right)} \text{ [46]}, \tag{2.6}$$

kde $\mu(t_{\mathrm{rtt}})$ značí střední hodnotu rozdělení a $\sigma^2(t_{\mathrm{rtt}})$ rozptyl pro příslušnou hodnotu změřeného zpoždění t_{rtt}. Následně je změřeno zpoždění od referenčního bodu k lokalizované stanici a dle toho je vytvořena kružnice (se středem v referenčním bodě). Na jejím blízkém okolí je definována hustota pravděpodobnosti dle zkalibrovaného Gaussova rozdělení (rovnice 2.6). Průnik hustot pravděpodobnosti všech referenčních bodů vytvoří v místě průniku region s vysokou pravděpodobností (součet všech

pravděpodobností) výskytu cílové stanice (viz obrázek 2.8). Autoři metody umožňují využít jejich lokalizační metodu online spotter.etomic.org a v publikaci [46] uvádějí medián chyby metody 30 km.



Obr. 2.8: Způsob lokalizace stanice pomocí nalezení místa s nejvyšší pravděpodobností výskytu (červená barva). Pravděpodobnosti jsou vykresleny okolo landmarků L_1, L_2 a L_3. Převzato z [46].

2.3.11 Metoda Spring Based Geolocation

Metoda Spring Based Geolocation [25] obdobně jako jiné metody nejprve provede kalibraci měřením zpoždění mezi ostatními landmarky a z nich pomocí metody nejmenších čtverců spočítá převodní funkci mezi zpožděním a vzdáleností. K lokalizaci stanic je však využito principu k nalezení rovnovážného stavu pružin. Tento systém byl dříve využit autory systému Vivaldi [13] pro predikci zpoždění. U SBG je však rovnovážný stav využit k určení polohy hledané stanice. Autoři metody v [25] uvádějí její medián chyby 60 km a průměrnou chybu 73,4 km.

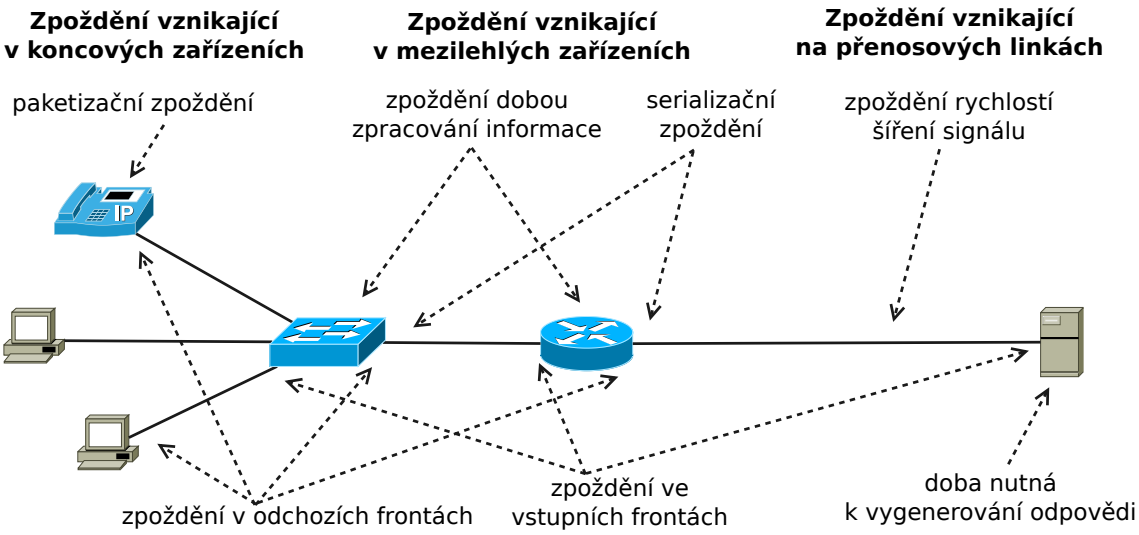
2.3.12 Srovnání aktivních geolokačních metod

Aktivní geolokační metody uvedené výše patří mezi ty v současnosti nejvýznamnější vzhledem k tomu, že jsou často referovány v aktuálních publikacích. Kromě těchto metod, existují řada dalších – například Dragoon [27, 28], Aliade [10], geolokace založená na neuronových sítích [37]. Taktéž v průběhu času vzniklo velké množství článků věnujících se různým vylepšením metod například [18, 14, 23] .

U všech aktivních metod popsaných výše v samostatných podkapitolách je kromě principu funkce metody, uvedena i přesnost, kterou ve většině případů uvádějí samotní autoři metody. Tyto výsledky jsou však těžko porovnatelné mezi sebou, neboť každý autor používá jiný počet a rozmístění měřících uzlů, testuje na vlastních datasetech cílů a v neposlední řadě výsledky ovlivňuje lokalita, ve které byla měření provedena – většina publikací se zaměřuje na evropský kontinent nebo USA. Srovnání jednotlivých metod je možné najít u autorů metody novější, jejíž výsledky jsou porovnávány s metodami staršími – např. [73, 1, 20], případně existují samostatná srovnání jednotlivých metod [5, 78]

2.4 Parametry komunikačního řetězce ovlivňující IP geolokaci

Za zpoždění je v telekomunikacích považován čas, který informace stráví na cestě od zdroje k příjemci. Zpoždění vznikající na jednotlivých částech komunikačního řetězce (obrázek 2.9) má různý charakter, velikost i vliv na celkové zpoždění [80].



Obr. 2.9: Zdroje zpoždění a místo jejich vzniku.

Pokud zkoumáme zpoždění podrobněji, zjistíme, že při opakovaných měřeních stejné přenosové trasy má zpoždění podobnou velikost. V publikaci [6] autoři rozdělují celkové zpoždění na jeho *deterministickou* část a *stochastickou* část. Deterministické zpoždění má konstantní velikost, kterou lze vypočítat (jedná se o minimální čas potřebný pro přenos zprávy). Hodnota celkového zpoždění nemůže být nikdy menší než velikost deterministické části zpoždění. Oproti tomu stochastické zpoždění má náhodný charakter a je ovlivněné aktuálním stavem sítě (dobou zpracování

v odchozích frontách, zatížením síťových prvků, rušením na přenosovém médiu – nutnost retransmise). Toto zpoždění může mít i nulovou hodnotu a nemusí celkové zpoždění ovlivňovat, naopak v případě velkého zatížení sítě může velikost tohoto zpoždění převyšovat deterministickou část. Stochastické zpoždění má zpravidla charakter Gamma rozdělení s výrazným dozvukem (heavy tail) [6].

2.4.1 Zpoždění vznikající v koncových zařízeních

Koncovými zařízeními jsou zdroj a cíl přenášené informace, které předpřipraví zprávu pro přenos, označí ji adresou svojí i příjemce a zapouzdří ji do paketů. Na druhé vrstvě referenčního modelu ISO/OSI se paket zapouzdří do rámce, či buňky (dle použité technologie) a následně se vyšle bit po bitu na přenosovou linku. Přijímací stanice k získání přenášené informace musí provést opačný postup. Na celkové zpoždění koncového zařízení t_{KZ} má vliv několik dílčích zpoždění, která jsou rozebrána v následujících podkapitolách. Zpoždění v odchozích frontách, serializační a deserializační zpoždění se také týká koncových zařízení při odesílání, respektive příjmu zprávy, ale primárně se vztahuje k mezilehlým zařízením, proto jsou podrobnosti k těmto typům zpoždění uvedeny v kapitolách 2.4.3 a 2.4.3.

Paketizační zpoždění

Paketizační zpoždění (t_{pak}) je doba, potřebná pro vytvoření paketu – přenosové jednotky v IP sítích. Závisí na rychlosti zpracování informace počítačem. Pokud se však jedná o interaktivní komunikaci, závisí toto zpoždění též na velikosti paketu, objemu informace a vzorkovací frekvenci. Při tomto typu komunikace se vždy čeká, než je naplněna velikost paketu hodnotami několika po sobě jdoucích vzorků (např. úryvků řeči). To způsobuje zpoždění prvního vzorku v paketu. Při depaketizaci je toto zpoždění vyrovnáno pozdržením ostatních vzorků paketu tak, aby byly stejně zpožděny jako první vzorek. Např. u kodeku hlasu G.723.1 je t_{pak}=30 ms [53].

Doba nutná k vygenerování odpovědi

Při analýze obousměrného zpoždění t_{rtt} (RTT – Round Trip Time) patří mezi složky zpoždění i doba potřebná k vygenerování odpovědi cílové stanice (t_{odp}). Velikost tohoto zpoždění je závislá na koncové stanici, operačním systému a jejím vytížení. Tím pádem ji není možné snadno definovat a k jejímu zjištění je potřebné provést měření na skutečných zařízeních.

Laki et al. [45] provedl měření doby nutné k vygenerování odpovědi stanic v síti GEANT2 pomocí nástrojů `ping` a `traceroute`. Princip měření spočíval ve změření

RTT přes N L3 přeskoků a následně změřil zpoždění mezi každými sousedními routery. Poté od změřeného t_{rtt} na celém komunikačním řetězci odečetl součet t_{rtt} měření mezi jednotlivými sousedy. Tímto bylo získán $N-1$ násobek doby nutné k vygenerování odpovědi. Doba nutná k vygenerování odpovědi byla ve většině případů mezi 300 μs a 1000 μs, přičemž nejvíce zastoupena byla hodnota okolo 500 μs. V této hodnotě jsou zahrnuty i paketizační, serializační a deserializační zpoždění obou koncových zařízení t_{KZ2}, proto je tuto hodnotu možné považovat jako $t_{\text{KZ,zdroj}} + t_{\text{KZ,cil}}$.

2.4.2 Zpoždění vznikající na přenosových linkách

Přenosové linky jsou část transportní soustavy, která je přímo závislá na fyzické pozici stanic. Linky ke stanicím nejsou vedeny nejkratší cestou, ale kabely jsou pokládány na vhodných místech (například podél silnic, železnic a spolu s dálkovým vedením elektrické energie) [6]. Další prodloužení cesty paketu může způsobit směrování. Přestože existuje k cíli fyzicky kratší linka, směrovače mohou vybrat linku jinou, například z důvodu cenové politiky, nebo rychlosti linek. Tímto narůstá délka trasy a roste i zpoždění způsobené dobou šíření signálu. Kromě fyzické vzdálenosti stanic je důležitá také technologie přenosu signálu, která ovlivňuje především přenosovou rychlost, a tedy i serializační zpoždění. Doba zpoždění přenosově linky (t_{PL}), zahrnuje zpoždění způsobené dobou šíření signálu. Serializační zpoždění ovlivněné rychlostí linky je uvedeno v rámci zpoždění mezilehlých zařízení.

Zpoždění vzniklé rychlostí šíření signálu

Zpoždění vzniklé rychlostí šíření signálu v přenosovém médiu t_{rs} (v jiné literatuře nazývané také propagační zpoždění) je přímo závislé na délce přenosové linky l a rychlosti šíření signálu v_{rs} v ní, dle rovnice

$$t_{rs} = \frac{l}{v_{\text{rs}}} \text{ . [45]} \tag{2.7}$$

Elektromagnetický signál se šíří médiem rychlostí, která je odvozená od rychlosti světla ve vakuu (c = 299 792 458 m/s), respektive c je maximální možná rychlost šíření. Protože je nejrozšířenějším médiem pro transportní sítě optické vlákno, je vhodné na dlouhých trasách počítat s rychlostí šíření signálu přibližně v_{rs} = 194 895 km/s, tj. 0,65c. U metalických kabelů je rychlost šíření signálu médiem 0,75c (v_{rs} = 224 844 km/s) a při bezdrátovém přenosu vzduchem je přibližně rovna c (v_{rs} = 299 792 km/s) [62]. Pro dlouhé vzdálenosti (nad 1000 km) má toto zpoždění hlavní podíl na celkovém zpoždění. Například při mezikontinentální komunikaci je hodnota tohoto zpoždění v řádu desítek až stovek milisekund [81].

2.4.3 Zpoždění vznikající v mezilehlých zařízeních

Mezilehlé zařízení je jakýkoliv aktivní prvek na cestě mezi koncovými zařízeními. Tato zařízení mají mnoho rozličných úkolů, svým rozsahem zahrnující různé vrstvy referenčního modelu ISO/OSI. Mezi nejdůležitější úkoly patří směrování a přepínání, tedy přeposílání datových jednotek nejkratší (nejvýhodnější) cestou k cíli. Zpoždění v mezilehlých zařízení t_{MZ} se skládá ze serializačního a deserializačního zpoždění, doby potřebné pro přesunutí datové jednotky ze vstupu na výstup, a doby čekání ve výstupních frontách.

Vzhledem k složitosti všech složek zpoždění vznikajících v mezilehlých zařízeních (více viz následující podkapitoly) autoři publikace [45] využívají, jako aproximaci zpoždění jednoho L3 prvku, vlastním měřením zjištěnou hodnotu $t_{\mathrm{MZ}} = 100$ μs. Tato hodnota je také průměrem měření zjištěných v publikacích [11, 51].

Starší měření zpoždění mezilehlého L3 zařízení bylo provedeno pomocí speciálního nástroje (Test box) vložením 0–3 směrovačů mezi dva Test boxy. Průměrná naměřená hodnota zpoždění pro směrovač Cisco 3630 byla 224 μs (pro 100 B paket) [6].

Serializační a deserializační zpoždění

Serializační zpoždění (t_{ser}) je doba od vyslání prvního bitu do odeslání posledního bitu rámce na linku. Hodnotu tohoto zpoždění ovlivňuje velikost rámce a přenosová rychlost linky. Serializační zpoždění vzniká za každým aktivním prvkem, ve kterém dojde k uložení rámce do paměti. Serializační zpoždění má deterministický charakter, při znalosti velikosti zprávy a rychlosti přenosové linky, jej můžeme přesně určit. Dobu serializace bitů na linku je možné spočítat pomocí vzorce

$$t_{ser} = R \cdot F, \ [45] \tag{2.8}$$

kde je vynásobena přenosová rychlost linky R a velikost rámce[6] F. V případě použití vysokorychlostních linek – 10 Gbit/s a rychlejších je i pro největší možný rámec toto zpoždění menší než 2 μs.

Deserializační zpoždění (zpoždění ve vstupních frontách) t_{des} je doba, kterou rámec stráví ve vstupní vyrovnávací paměti (buffer), než jsou načteny všechny potřebné bity (hlavička, či celý paket). V případě změny přenosové rychlosti mezi vstupním a výstupním portem je potřeba načíst celý paket, v ostatních případech záleží na metodě přeposílání rámců. Proto je velikost deserializačního zpoždění rovna serializačnímu zpoždění t_{ser} příchozí linky vztažené na počet bitů ukládaných do vyrovnávací paměti.

[6] Protože je na 2. vrstvě RM ISO/OSI velmi často nasazena technologie Ethernet, bývá největší velikost rámce 1518 B. Pro jiné technologie lze tuto hodnotu dohledat v literatuře [62].

V případě L2 zařízení (např. přepínač) se v dnešní době nejčastěji používaná metoda načtení celé zprávy do zařízení (*store and forward*), kdy se uplatní stejné serializační i deserializační zpoždění. Můžeme se však setkat i s metodami *cut-through* nebo *fragment free*, u nichž se načte pouze prvních 14 B, respektive 64 B a poté je rámec dále odeslán odchozí linkou.

Doba zpracovávání informace

Je to čas t_z, který zařízení potřebuje pro přesunutí paketu ze vstupu do výstupní fronty. Minimální hodnota závisí na výkonu zařízení a prováděných operacích (např. směrování, překlad adres, značkování pro QoS, filtrování portů, ...). Celková doba zpracování se zvyšuje se zvyšujícím se zatížením prvku. Každé zařízení má definovanou propustnost, která se ne vždy rovná zatížení při maximálním využití všech portů. Moderní síťové prvky mají většinu funkcionalit podporovaných hardwarovým čipem, a tak při standardním provozu bude tato doba minimální. Každopádně je nemožné stanovit deterministicky tuto dobu pro různé typy zařízení a jejich funkcionality.

Zpoždění v odchozích frontách

Hodnotu zpoždění v odchozích frontách (t_f) není možné deterministicky určit, protože je závislá na aktuálním zatížení prvku. Zpoždění v odchozích frontách se uplatňuje nejen směrovače a přepínače, ale i pro koncové stanice. Velikost tohoto zpoždění je proměnná a dá se aktuálně zjistit podle celkové velikosti paketů ve frontě a rychlosti odchozí linky podle vztahu

$$t_f = \frac{1}{\mu - \lambda_p} = \frac{1}{\frac{R}{F} - \lambda_p} \text{ [43].} \tag{2.9}$$

Při rychlosti odchozí linky 10 Mbit/s (R) a průměrné velikosti rámce 1250 B (F) je odchozí fronta schopna obsloužit až 1000 paketů za sekundu (μ). Pokud by přicházelo 500 paketů za sekundu (λ_p), pak by průměrné zpoždění bylo 2 ms (t_f) a fronta byla zatížená (ρ) na 50 %. Zatížení fronty je možné vypočítat dle vztahu

$$\rho = \frac{\lambda}{\mu} \text{ [43].} \tag{2.10}$$

Tyto rovnice jsou platné pouze, pokud je použito řazení do front typu FIFO. V případě prioritních front, či jiných metod obsluhy front, využívaných pro zajištění QoS, není možné určit průměrnou dobu zpoždění.

2.4.4 Zpoždění celého komunikačního řetězce

Jak je zobrazeno na obrázku 2.9, zdrojem zpoždění v síti je každá část sítě. Celkové zpoždění je dáno součtem zpoždění za jednotlivé části sítě. V závislosti na počtu N

mezilehlých zařízeních (počet linek $N+1$) je možné vyjádřit jednosměrné zpoždění t jako

$$t = t_{\mathrm{KZ,zdroj}} + \sum_{i=0}^{N} t_{\mathrm{PL},i} + \sum_{i=0}^{N-1} t_{\mathrm{MZ},i} + t_{\mathrm{KZ,cil}} \ . \tag{2.11}$$

V případě podrobnějšího pohledu na síť, kdy budou rozlišeny i jednotlivé druhy zpoždění se předchozí rovnice rozšíří následovně

$$t = t_{\mathrm{pak,zdroj}} + t_{\mathrm{f,zdroj}} + t_{\mathrm{ser,zdroj}} + \sum_{i=0}^{N} t_{\mathrm{rs},i} + \sum_{i=0}^{N-1} (t_{\mathrm{des},i} + t_{\mathrm{z},i} + t_{\mathrm{f},i} + t_{\mathrm{ser},i}) + t_{\mathrm{des,cil}} \ . \tag{2.12}$$

Pro obousměrné zpoždění t_{rtt} (RTT) je možné vyjádřit podrobně zpoždění pro celý komunikační řetězec při symetrickém směrování zprávy takto

$$\begin{aligned} t_{\mathrm{rtt}} = {} & t_{\mathrm{pak,zdroj}} + t_{\mathrm{f,zdroj}} + t_{\mathrm{ser,zdroj}} + t_{\mathrm{des,zdroj}} + 2 \cdot \sum_{i=0}^{N} t_{\mathrm{rs},i} + \\ & + 2 \cdot \sum_{i=0}^{N-1} (t_{\mathrm{des},i} + t_{\mathrm{z},i} + t_{\mathrm{f},i} + t_{\mathrm{ser},i}) + t_{\mathrm{des,cil}} + t_{\mathrm{odp,cil}} + t_{\mathrm{pak,cil}} + t_{\mathrm{ser,cil}} + t_{\mathrm{f,cil}} \ . \end{aligned} \tag{2.13}$$

Méně podrobně, ale na podobném principu je popsán popis obousměrného zpoždění komunikačního řetězce v [45] jako

$$t_{\mathrm{rtt}} = t_{\mathrm{odp,cil}} + \sum_{i=0}^{N-1} (t_{\mathrm{z},i} + t_{\mathrm{f},i} + t_{\mathrm{ser},i} + t_{\mathrm{rs}}(\mathbf{n}_{i-1}, \mathbf{n}_i)) \ , \tag{2.14}$$

kde $t_{\mathrm{rs}}(\mathbf{n}_{i-1}, \mathbf{n}_i)$ znamená zpoždění způsobené rychlostí šíření signálu mezi dvěmi zařízeními, přičemž do celkového počtu zařízení N jsou započítány zdroj i cíl.

Popis složek deterministické části zpoždění $t_{\mathbf{d}}$ můžeme nalézt v [6], kde je uvedena následující rovnice

$$t_{\mathrm{d}} = t_{\mathrm{zd}} + t_{\mathrm{ser}} + t_{\mathrm{rs}} \ , \tag{2.15}$$

kde t_{zd} reprezentuje deterministickou část zpoždění nutnou k přesunu datové jednotky ze vstupu na výstup zařízení. Tato část doby zpracování informace je minimální možná – tedy daná fyzickou charakteristikou prvků. Část zpoždění způsobeného dobou zpracování, která je závislá na aktuálním zatížení (popř. konfigurací prvku), je definována jako t_{zs} a je využita při popisu stochastické části zpoždění t_{s} [6] pomocí

$$t_{\mathrm{s}} = t_{\mathrm{zs}} + t_{\mathrm{f}} \ , \tag{2.16}$$

kde je navíc doba strávená ve frontách t_{f}, která je závislá na množství dat čekajících na dostupnost odchozí linky. Součtem deterministické a stochastické části zpoždění je definováno celkové zpoždění komunikačního řetězce.

3 CÍLE DIZERTAČNÍ PRÁCE

V dnešní době je využití IP geolokace na denním pořádku, vzhledem k velkému počtu webových stránek, které ji používají k přizpůsobení stránky návštěvníkovi nebo naopak k omezení služeb z některých zemí. Pro správce sítě a experty z oblasti kyberkriminality je důležité také zjistit, odkud síťový provoz pochází (na základě zdrojové IP adresy). K tomu všemu jsou v převážné míře využívány IP geolokační databáze. Jejich přesnost je velmi dobrá – s jistotou určí zemi a při určení místa (města, popř. regionu) jsou přesné v cca 75 % případů [40, 48]. Problematické však jsou v těchto případech záznamy o IP adresách, které nejsou pravdivé. Tyto vznikají především z důvodu změny IP adresy zařízení, při přesunu zařízení či při vložení chybného záznamu do databáze [59]. Z tohoto ohledu je důležité najít způsob, jak tento chybný záznam s jistotou rozpoznat a případně předat informace, které pomůžou vlastníkům databází záznam opravit.

Dizertační práce se zabývá aktivními IP geolokačními metodami a jejich využitím pro ověření záznamů z IP geolokačních databází. Aktivní metody jsou založeny na měření síťových parametrů z referenčních bodů zvaných landmarky. Při využití fyzikálních zákonů (např. rychlost šíření signálu v médiu) a statistických údajů komunikace v Internetu je možné stanovit vzdálenost od referenčního uzlu (landmarku), za kterou již lokalizovaná stanice nemůže ležet. Pokud vytvoříme průnik takto vzniklých oblastí okolo jednotlivých landmarků získáme oblast, ve které se s jistotou cílová stanice nachází. Když pro tuto stanici vezmeme zeměpisné souřadnice z geolokační databáze, můžeme snadno ověřit, zda tyto souřadnice leží uvnitř vymezené oblasti a rozhodnout, jestli je informace o poloze z geolokační databáze důvěryhodná.

Hlavním cílem dizertační práce je navrhnout geolokační metodu, která bude s jistotou definovat oblast, ve které se hledaná stanice nachází a následně tuto metodu aplikovat na záznamy z vybraných geolokačních databází a nalézt tak záznamy, které nejsou důvěryhodné. Tato metoda bude založena na fyzických vlastnostech síťových prvků – jak aktivních (router, switch, ...), tak pasivních (přenosová média). Funkčnost metody bude ověřena na vybraném vzorku testovacích adres a její výsledky budou porovnány s obdobnými metodami. Následně bude vytvořen dataset IP adres, pro které budou získány záznamy z geolokačních databází k ověření vytvořenou metodou. Výstupem bude především statistika počtu adres mimo oblast zaručené polohy, druhotně pak přesnost záznamů vzhledem k výsledkům aktivního měření.

Dílčí cíle vedoucí ke splnění hlavního cíle dizertační práce:

- Prvotním cílem je prozkoumat vlastnosti zpoždění, určit parciální složky zpoždění a definovat, které složky jsou ovlivněny geografickou vzdáleností. Pro ne-

deterministicky určitelné složky zpoždění bude nutné stanovit statisticky jejich hraniční hodnoty. K tomu bude využito veřejných charakteristik a dokumentů popisujících síť CESNET2, ve které bude následně provedeno měření za účelem porovnání teoretických předpokladů s reálnými vlastnostmi datové sítě.

- Následně bude dlouhodobým měřením zpoždění v experimentální síti PlanetLab [58] zjištěna variace zpoždění a její vliv na kalibraci geolokačních metod. Porovnávána bude především proměnlivost velikosti zpoždění a délky komunikační cesty v průběhu denní doby, fáze pracovního týdne a delšího časového období. Na tomto základě pak bude možné stanovit kalibrační proces u nově navržené metody.
- Na základě deterministických složek zpoždění a hraničních hodnot stochastických činitelů bude vyvinuta nová geolokační metoda. Tato metoda bude oproti jiným s jistotou definovat oblast, ve které se musí hledaná stanice nacházet.
- Vyvinutá metoda bude následně otestována na datasetu IP adres se známou polohou, aby se ověřila její věrohodnost. Následně bude provedeno měření na obsáhlém datasetu IP adres, pro které budou zjištěny zeměpisné souřadnice z vybraných geolokačních databází. Výsledkem pak bude určení procenta adres, které se dle měření nenachází v oblasti se zaručenou polohou a také odchylka, od polohy ohlášené jinými aktivními geolokačními metodami.

4 ANALÝZA PARAMETRŮ OVLIVŇUJÍCÍCH ODHAD VZDÁLENOSTI

Tato kapitola se věnuje analýze síťových parametrů mezi stanicemi se známou polohou za účelem zjištění typických hodnot parametrů, které ovlivňují přepočet zpoždění na vzdálenost pro účely IP geolokace. Většina aktivních geolokačních metod uvedených v kapitole 2.3 potřebuje pro svoji činnost odvodit vzdálenost mezi landmarkem (stanicí se známou polohou) a lokalizovanou stanicí. K tomu je využito měření zpoždění při přenosu informace. Vzhledem k nemožnosti využít k měření lokalizovanou stanici je možné pouze změřit čas od odeslání zprávy do přijetí její odpovědi (obousměrné zpoždění – RTT). Pokud zanedbáme malé procento případů, kdy je provoz v Internetu směrován asymetricky (podrobněji v [12]), můžeme předpokládat, že po odečtení času nezbytného pro vygenerování odpovědi je polovina obousměrného zpoždění rovna zpoždění jednosměrnému.

K odvození velikosti jednosměrného zpoždění je nutné od obousměrného zpoždění odpočítat čas nutný k vygenerování odpovědi cílovou stanicí a také čas zpracování zdrojovou stanicí (odeslání a příjem), které je podrobněji popsané v kapitole 2.4.1. Pro výpočty v této kapitole vyjdeme ze závěrů autorů publikace [45], kteří využívají nejnižší jimi zjištěnou dobu nutnou k vygenerování odpovědi ($t_{\text{odp}} = 300$ μs). Díky tomu je možné naměřené hodnoty obousměrného zpoždění dále analyzovat jako jednotlivé složky jednosměrného zpoždění tak, jak je uvedeno v kapitole 2.4.

Pro následující analýzu je nezbytné stanovit přesnost, s jakou je nezbytné měření provádět. K tomu vyjdeme z podkapitoly 2.4.2 věnující popisu propagačního zpoždění, které ovlivňuje dobu potřebnou k přenesení informace přes definovanou vzdálenost. Jako transportní médium uvažujeme optické vlákno, metalický kabel nebo bezdrátový přenos vzduchem. Pro vzdálenost 1 km přímou cestou je propagační zpoždění na těchto médiích 5,13 μs respektive 4,45 μs a 3,34 μs (podrobněji viz kapitola 2.4.2). Vzhledem k tomu, že většina měření v práci je prováděna pomocí obousměrného měření, můžeme po odečtení všech ostatních vlivů na zpoždění předpokládat, že přibližně 10 μs[1] znamená vzdálenost 1 km.

Většina měření v práci probíhá na operačním systému Linux[2] pomocí zabudovaného nástroje `ping`, který ve výstupu zobrazuje naměřené zpoždění s přesností na jednotky μs. Avšak dle měření provedených v článku [2] dosahuje `ping` přesnosti $\pm$0,1 ms, což bude i přesnost s kterou budeme pracovat dále v práci. Vzhledem k informacím z předchozího odstavce, můžeme předpokládat přesnost výpočtů na cca 10 km.

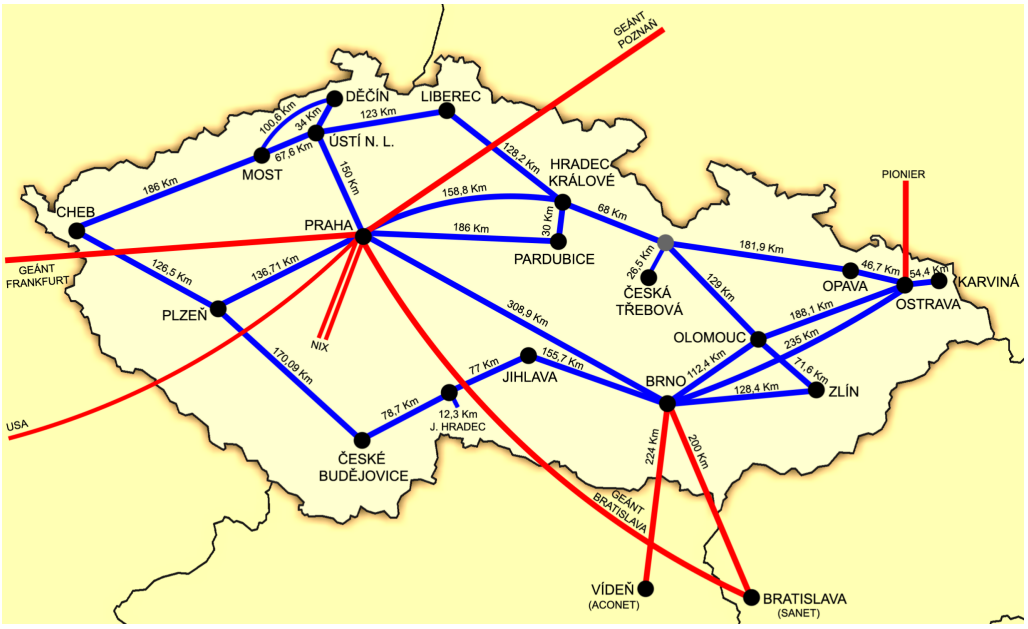
[1] Přesně vyjádřeno pro optiku, metaliku a vzduch je to 10,26 μs respektive 8,90 μs a 6,68 μs.

[2] Jednalo se převážně o distribuce založené na Red Hat Linuxu.

4.1 Měření v rozlehlé síti se známou topologií

Cílem měření v síti CESNET2 bylo srovnat naměřené zpoždění s předpokládanou hodnotou přenosového zpoždění a odvodit minimální, průměrnou a střední hodnotu zpoždění pro jedno zařízení pracující na síťové vrstvě (L3) referenčního modelu ISO/OSI.

Cesnet je sdružení vysokých škol a akademických pracovišť, jehož cílem je rozvoj páteřní akademické sítě CESNET2 (Czech Scientific and Education NETwork 2). Tato síť propojuje univerzitní města vysokorychlostními datovými okruhy. Topologie sítě je složena z několika kruhů procházejících omezeným počtem měst. Páteř sítě tvoří optický přenosový systém, na němž je nasazeno DWDM (Dense Wavelength Division Multiplexing) a na třetí vrstvě je IP/MPLS (Multiprotocol Label Switching) [9]. Celá síť CESNET2 je bohatě dokumentovaná a je možné zjistit mnoho informací o topologii sítě i parametrech jednotlivých linek (na obr. 4.1 je zobrazena topologie sítě). Pro další použití byla v dokumentaci nalezena skutečná délka optických kabelů mezi jednotlivými městy [9], která je v obrázku 4.1 také zaznačena.



Obr. 4.1: Délka optických kabelů mezi jednotlivými městy v síti CESNET2 [81], údaje z roku 2011.

K měření byly využity dva uzly sítě PlanetLab, umístěné v Praze a Brně, které měřily zpoždění pomocí ICMP dotazů k serverům na různých místech sítě CESNET2. Měření bylo provedeno opakovaně (pětkrát v různých pracovních dnech a hodinách)

a pokaždé bylo odesláno celkem 121 zpráv ICMP typu Echo Request během 120 s [3]. Ze všech hodnot bylo vybráno minimum, za účelem eliminace aktuálního zatížení přenosových prvků – stochastické složky zpoždění. Jednosměrné zpoždění t bylo získáno odečtením hodnoty doby nutné k vygenerování odpovědi t_{odp} (dle kapitoly 2.4.1 300 μs) od obousměrného zpoždění t_{rtt} a jejím rozpůlením dle

$$t = \frac{t_{\mathrm{rtt}} - t_{\mathrm{odp}}}{2}. \tag{4.1}$$

Toto můžeme provést, neboť předpokládáme pro oba směry stejnou cestu, což bylo experimentálně ověřeno mezi uzlem v Praze a Brně. K podrobnému prozkoumání přenosové trasy bylo využito nástroje `traceroute`, který zjistí počet mezilehlých zařízení, zpoždění k nim, jejich IP adresu a doménové jméno.

4.1.1 Zpoždění na jedno mezilehlé zařízení

Pro výpočet zpoždění je použito známé délky optického kabelu (přenosové cesty), počtu mezilehlých zařízení a změřeného obousměrného zpoždění. Počet mezilehlých zařízení zahrnuje pouze zařízení pracující na třetí vrstvě referenčního modelu OSI, nezahrnuje tedy přepínače, rozbočovače, zesilovače signálu či jiné pasivní nebo aktivní prvky pracující na nižší než třetí vrstvě ISO/OSI modelu [4]. Počet zařízení pracujících na třetí vrstvě byl zjištěn z velikosti TTL[5] příchozího paketu. Vzhledem k tomu, že obvyklé hodnoty TTL (255, 128, 64 a 32) používané operačními systémy jsou známé [52], dopočet není složitý. Zjištěný počet mezilehlých zařízení byl také ověřen programem `traceroute`.

Změřené hodnoty obousměrného zpoždění t_{rtt} a počtu mezilehlých zařízení N jsou pro pět vybraných organizací uvedeny v tabulce 4.1. Velikost jednosměrného zpoždění t byla odvozena od změřeného obousměrného zpoždění dle 4.1. Skutečná délka cesty mezi městy (l_{skut}) byla zjištěna z dokumentace sítě CESNET2, přičemž zanedbáváme délku vedení v lokalitě stanice. Výpočet zpoždění pro jedno mezilehlé zařízení (t_{mz}) byl proveden pomocí rovnice

$$t_{\mathrm{mz}} = \frac{t - \frac{l_{\mathrm{skut}}}{v_{\mathrm{rs}}}}{N}, \tag{4.2}$$

kde v_{rs} rychlost šíření signálu v médiu. V našem případě je přenosovým médiem optika – tedy $v_{\mathrm{rs}} = 0,65\mathrm{c}$, kde c je rychlost světla ve vakuu.

[3] Dle doporučení – RFC 2544 [7].

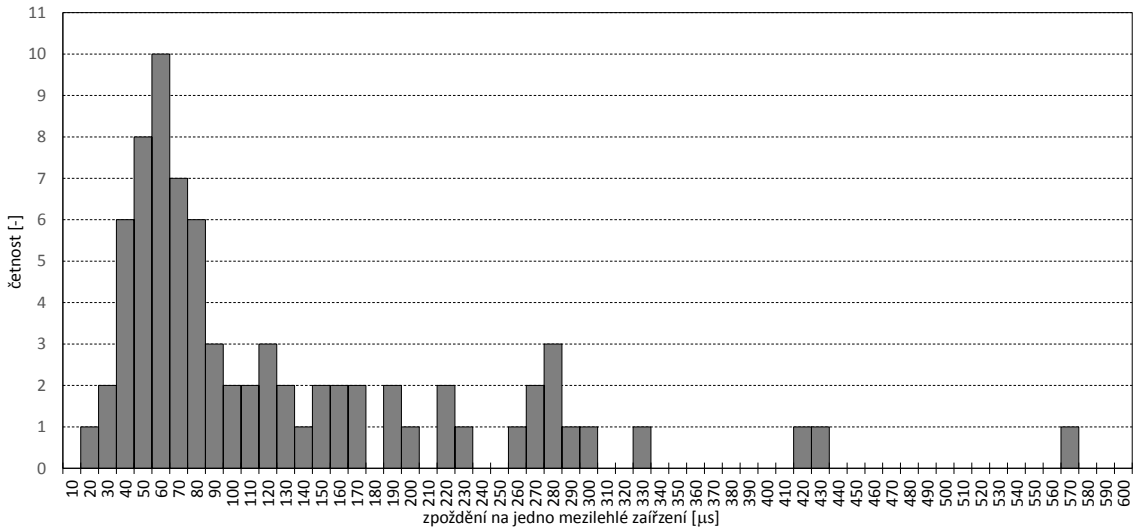
[4] Dle dokumentace zařízení a provedených měření mají tyto prvky zanedbatelné zpoždění v porovnání s ostatními zdroji zpoždění [44].

[5] Time To Live – parametr záhlaví IP paketu. Každé zařízení pracující na třetí vrstvě referenčního modelu OSI jej musí při příjmu snížit o jedna.

Tab. 4.1: Změřené a vypočítané hodnoty pro měření k vybraným cílům [81].

město	**doménové jméno**	t_{rtt} [ms]	N [-]	t [ms]	l_{skut} [km]	t_{mz} [ms]
Zlín	www.utb.cz	2,009	4	1,005	128,4	0,086
Ostrava	vpn.vsb.cz	3,125	4	1,563	235,0	0,089
Praha	www.cvut.cz	4,148	6	2,074	308,9	0,082
Plzeň	www.lfp.cuni.cz	5,892	6	2,946	445,6	0,110
Cheb	fennel.fek.zcu.cz	7,036	7	3,518	572,1	0,083

Histogram na obrázku 4.2 zobrazuje statistické rozložení spočítaných hodnot zpoždění na jeden L3 prvek (zařízení pracující na třetí vrstvě referenčního modelu ISO/OSI). Dle histogramu můžeme prohlásit, že pro 75 % hodnot je zpoždění nižší než 160 μs, což se rovná přenosovému zpoždění na vzdálenost 31 km. Minimální vypočítané zpoždění bylo 15 μs a hodnota dolního kvartilu byla 53 μs. Průměrná hodnota vypočítaných zpoždění na jeden prvek je 124 μs, medián je roven 77 μs [79]. Tyto hodnoty jsou srovnatelné s výsledky z článku [45], kde autoři uvádějí průměrnou hodnotu zpoždění směrovače 101 μs.



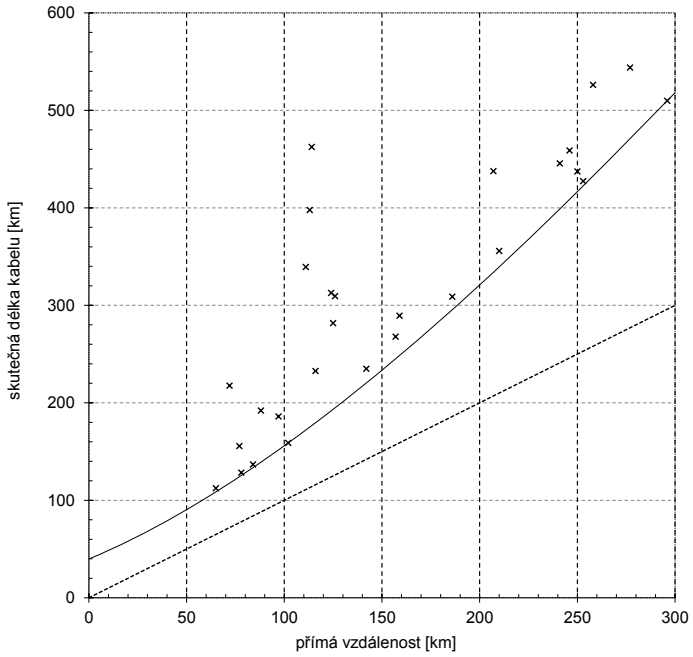
Obr. 4.2: Histogram zpoždění pro jedno mezilehlé zařízení zjištěné měřením a výpočtem, každý sloupec je široký 0,01 ms [79].

4.1.2 Srovnání délky kabelu s přímou vzdáleností

Aktivní geolokační techniky určí polohu cílové stanice jako průnik oblastí zjištěných jednotlivými referenčními body. Poloměr oblasti je zjištěn přepočtem zpoždění

na geografickou vzdálenost, která je ovlivňována nejen přenosovým zpožděním, ale i skutečnou délkou, která se může od přímé vzdálenosti výrazně lišit. Proto byl vyjádřen poměr skutečné délky kabelů ku přímé vzdálenosti ($\frac{l_{\text{skut}}}{l_{\text{prima}}}$). Na obrázku 4.3 je graf tohoto poměru vzdáleností mezi různými místy v síti CESNET2, kde je přerušovanou čarou zvýrazněn poměr 1:1. Plnou čarou je v obrázku 4.3 vyznačena křivka, která leží pod všemi zjištěnými páry, ale je jim nejblíže.

Pokud výsledky vyjádříme statisticky, je poměr v průměru 1,99, medián 1,92, minimum 1,38 a maximum 4,64. Srovnatelné výsledky uvádí Laki et al. [45], kde je uveden průměrný poměr 2,17 mezi přímou vzdáleností a skutečnou délkou kabelů.

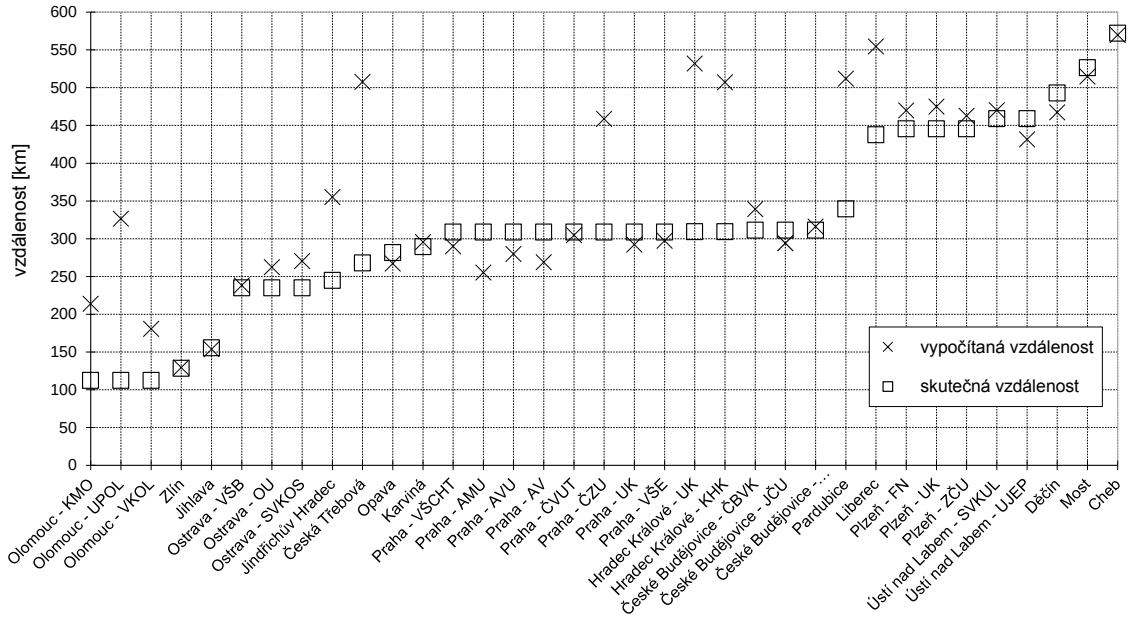


Obr. 4.3: Srovnání přímé vzdálenosti a skutečné délky kabelů v síti CESNET2 [77]. Přerušovanou čarou je vyznačen poměr přímé vzdálenosti ku skutečné 1:1, plná čára aproximuje nejnižší zjištěné poměry.

4.1.3 Porovnání vypočtené a skutečné délky trasy

Na základě zjištěných informací o síti CESNET2 byl proveden zkušební výpočet délky trasy, aby mohl být porovnán se skutečnou vzdáleností a byla prokázána korelace mezi vypočtenou a skutečnou délkou. V grafu na obrázku 4.4 je zobrazeno srovnání délky tras mezi Brnem a ostatními univerzitními městy. Některá města jsou v grafu několikrát, protože byla změřena každá vysoká škola ve městě. Výsledky ukazují, že většina vypočtených vzdáleností se blíží skutečnosti. Avšak v grafu se vyskytuje přibližně čtvrtina hodnot, jejichž relativní chyba je větší než 20 %, což je

způsobeno velmi zatíženými prvky zvyšujícími celkové zpoždění nebo směrováním datové jednotky po delší cestě.



Obr. 4.4: Graf vypočítané a skutečné délky cesty, pro organizace ze sdružení Cesnet [80].

5 VARIABILITA ZPOŽDĚNÍ V PRŮBĚHU ČASU

Pro návrh metody sloužící k verifikaci údajů z IP geolokačních databází je nutné nejprve prostudovat nakolik je zpoždění proměnné v čase. V této kapitole bude provedeno měření, za účelem dlouhodobého ověření stability charakteristických údajů zpoždění, které jsou popsány v kapitole 4 a jsou využity v kapitole 6 věnující se návrhu nové metody. Dle toho bude možné stanovit nakolik je pro novou metodu potřebná kalibrace.

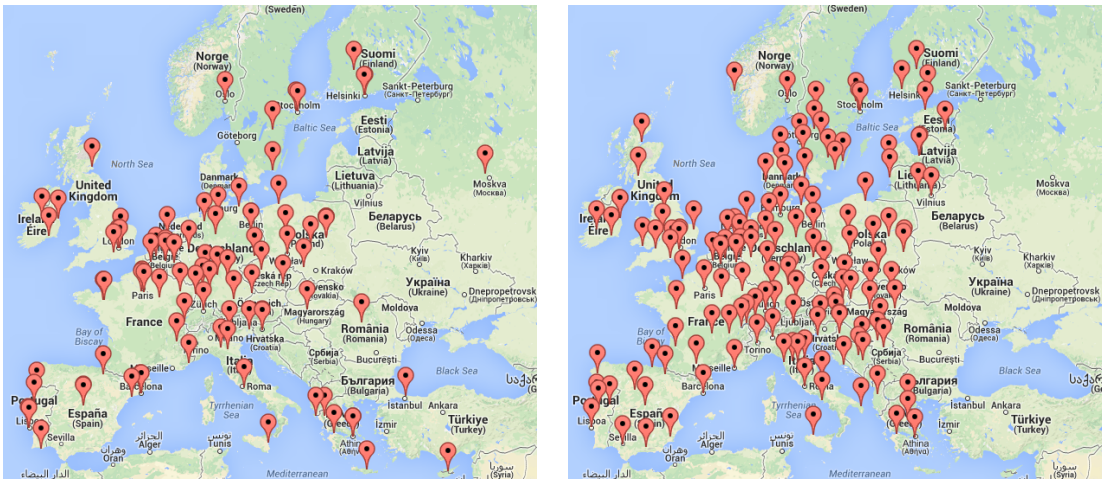
IP geolokační metody popsané v publikacích [1, 24, 25, 46, 73] předpokládají provedení kalibrace před zahájením skutečného měření. Kalibrace slouží k vytvoření převodní funkce mezi změřeným zpožděním a geografickou vzdáleností, důvodem k jejímu provedení před samotným měřením je zajištění co nejvyšší přesnosti. Naopak velkou nevýhodou provedení kalibrace před měřením je prodloužení doby nutné k realizaci samotného měření (předchází mu kalibrace). Další nevýhodou je zvýšení síťového provozu.

Tato kapitola se zabývá možností využití kalibračních dat z předešlé kalibrace za předpokladu zachování stejné přesnosti geolokace a také tím, nakolik neprovedení kalibrace degraduje výsledky. Pro toto ověření je v této kapitole provedeno měření pomocí základních geolokačních metod (CBG [24], Octant [73] a pro výpočet vzdálenosti také Spotter [68]) za použití různě starých kalibračních dat. Díky tomu je možné nejen definovat proměnlivost zpoždění, ale i vliv této proměnlivosti na výsledky IP geolokačních metod.

5.1 Metodologie

Měření v této kapitole probíhalo v experimentální síti PlanetLab [58], jejíž stanice jsou rozmístěny po celém světě, přičemž vyšší hustotu rozmístění mají stanice ve více obydlených a rozvinutých částech světa. Všechna měření probíhala na serverech PlanetLabu umístěných v Evropě, protože je zde největší hustota uzlů. Pro následná měření byla snaha vybrat z každé lokality, kde jsou umístěny servery PlanetLab, právě jeden uzel. Bohužel, některé uzly byly v době měření mimo provoz, případně měly jiné problémy (např. nedostupné SSH). Celkově bylo vybráno 93 uzlů rozmístěných po celé Evropě, jejichž polohy jsou zobrazeny na obrázku 5.1 a). Tyto uzly jsou dále využity jako aktivní landmarky, které dokáží provést měření zpoždění k jakémukoliv zařízení s veřejnou IP adresou.

Část b) na obrázku 5.1 zobrazuje polohu cílů – pasivních serverů se správnou (známou) polohou. Dataset těchto cílů se skládá z IP adres webových a jiných veřej-



a) b)

Obr. 5.1: Polohy serverů využitých pro ověření kalibrace IP geolokačních metod. a) polohy landmarků (serverů ze sítě PlanetLab), b) polohy ostatních uzlů využitých jako pasivní cíle.

ných serverů univerzit a významných firem. Dataset byl tvořen s cílem, aby se v něm nacházela alespoň jedna, ale maximálně tři IP adresy, z každého významného Evropského města. Poloha IP adres z datasetu byla zjištěna z adresy sídla organizace, neboť u velkých organizací předpokládáme, že většinou provozují vlastní server ve své lokaci. Takto zjištěná adresa byla následně ověřena pomocí geolokační databáze IP2Location [34] a pokud se tyto dvě adresy lišily, byla tato adresa z datasetu odebrána. Celkově jsme takto získali 151 IP adres s ověřenou polohou, méně zastoupená v datasetu je především východní Evropa.

Měření bylo opakovaně spouštěno po dobu necelých 4 měsíců (15 týdnů) tak, aby byla posbírána veškerá data ke každé celé hodině. Konkrétně měření probíhalo přibližně 30 minut a bylo spouštěno každou hodinu ve tři čtvrtě – aby mohlo být vztaženo k celé hodině. Při měření mezi každou dvojicí stanic bylo odesláno 30 dotazů v 0,5 sekundových intervalech, z nichž byla vybrána minimální hodnota. Zároveň dle velikosti TTL u odpovědi byl určen počet mezilehlých zařízení. Takto sebraná data posloužila k další analýze, která je podrobněji představena dále.

5.2 Sledování změny zpoždění v Internetu v průběhu času

Při analýze sesbíraných dat bylo v prvé řadě přistoupeno ke sledování změny zpoždění mezi stejnými uzly v Internetu v průběhu časového úseku. Převážná většina aktivních geolokačních metod provádí přepočet RTT (obousměrného zpoždění) na

vzdálenost za pomoci kalibrační funkce sestavené z párů hodnot RTT a geografické vzdálenosti. Vzhledem k tomu, že geografická vzdálenost mezi místy je konstantní, je potřebné zjistit, nakolik změna zpoždění vyvolá změnu kalibrační funkce.

Pro nově navrhovanou metodu bude důležité stanovit, jakým způsobem bude nezbytné řešit její kalibraci. Existuje předpoklad, že metoda bude založena převážně na deterministických částech zpoždění, které by měly být konstantní a závislé na fyzických parametrech spojení, díky čemuž by měl být vliv změny zpoždění v průběhu času zanedbatelný.

5.2.1 Změna zpoždění v průběhu času

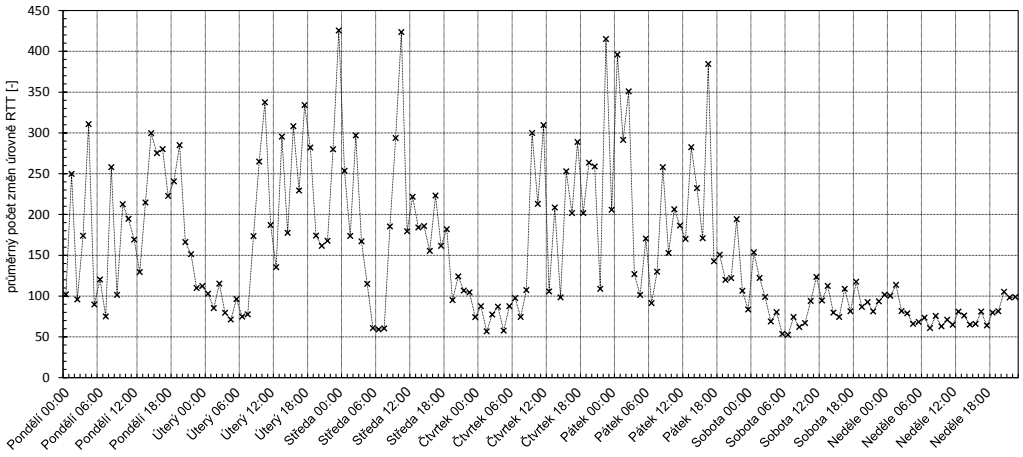
Celkově bylo provedeno měření mezi devadesáti stanicemi se zaručenou polohou, což znamená údaje o přibližně devíti tisíci párech stanic. Při měření každou hodinu po dobu šesti týdnů bylo nasbíráno okolo 6,5 miliónu jednotlivých měření, každé sestávající se z třiceti opakování ICMP dotazu a odpovědi. Vzhledem k přesnosti ±0,1 ms nástroje `ping` zmiňované v [2], byla považována změna velikosti zpoždění o 0,2 ms za nepřesnost měření a v dalších úvahách je jako změna zpoždění považovaná změna větší než 0,2 ms. Takovéto změny byly identifikovány u 195 tisíc po sobě jdoucích vzorků. Celkem 125 tisíc měření (2 % ze všech) vykazovalo změnu zpoždění, která trvala alespoň dvě hodiny. U zbytku změn (1 % ze všech měřených případů) se jednalo pouze o krátkodobé změny vyskytující se právě v jednom vzorku.

Po dobu šesti týdnů došlo k trvalejší změně (stejná hodnota po alespoň 2 hodiny) u každého páru v průměru 15krát (medián 9). Dočasné změny, kdy hodnota RTT měla při dalším měření hodnotu jinou, proběhly v průměru 7,5krát u každého páru (medián byl 2).

Zároveň bylo možné pozorovat, že zpoždění u příslušného páru mění hodnoty mezi několika málo zpožděními (s tolerancí ±0,1 ms), což může znamenat například střídání různých cest. V průměru se u jednoho páru vystřídalo naměřené RTT 5,2 různých hodnot (medián 5).

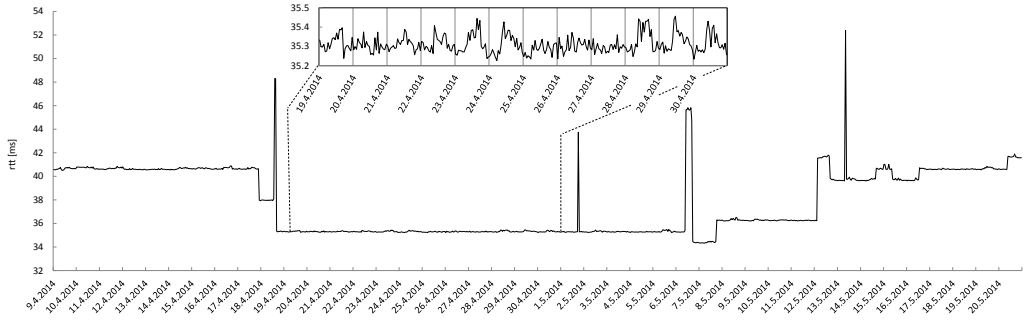
Při dalším pozorování změn úrovně zpoždění bylo zjištěno, že k nim dochází častěji v pracovních dnech než o víkendu. Podrobněji toto ukazuje graf na obrázku 5.2, kde je zobrazen průměrný počet změn v jednotlivých hodinách týdne.

K ilustraci typického průběhu zpoždění mezi místy bylo vybráno měření mezi městy Modena v Itálii (`planetlab-2.ing.unimo.it`) a Madrid ve Španělsku (`utet.ii.uam.es`), které se dají považovat za „průměrný“ příklad průběhu zpoždění. Průběh je zobrazen na obrázku 5.3, kde je zobrazen vývoj velikosti obousměrného zpoždění po dobu šesti týdnů (měřeno každou hodinu), přičemž graf začíná v 0:00 středoevropského času. V části grafu přibližující téměř konstantní průběh lze



Obr. 5.2: Průměrný počet změn úrovně RTT v rámci jednoho týdne.

pozorovat výchylku pohybující se maximálně ±0,1 ms a tato může být způsobena chybou nástroje, popř. zvýšeným zatížením v průběhu pracovního dne.

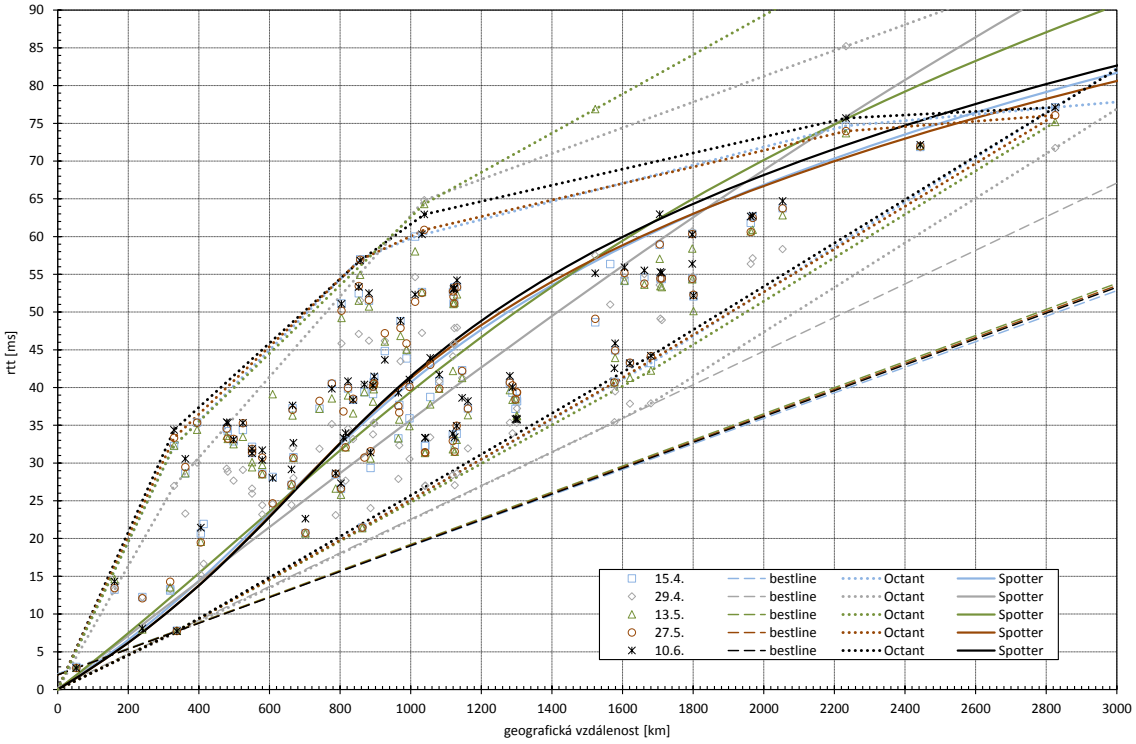


Obr. 5.3: Minimální změřené RTT mezi uzly `planetlab-2.ing.unimo.it` (Modena, Itálie) a `utet.ii.uam.es` (Madrid, Španělsko) v průběhu času.

V průběhu šesti týdnů došlo u této dvojice uzlů ke změně celkem čtrnáctkrát a pro 11 případů z toho změna této hodnoty trvala déle než hodinu (jedno měření). Vzhledem ke stabilitě nové hodnoty obousměrného zpoždění je pravděpodobné, že byla způsobena změnou cesty – např. směrováním přes jiný autonomní systém nebo použití jiného připojení cílového nebo zdrojového AS. Je proto pravděpodobné, že nová cesta znamenala využití cesty s rozdílnou délkou přenosového média (narostlo propagační zpoždění). Z údajů o TTL víme, že se změnil i počet mezilehlých zařízení (zpoždění dobou zpracování informace a v odchozích frontách). Důvodem vzniku krátkodobých změn, vyskytujících se pouze v jednom měření, může být dočasná změna směrování (podobně jako v případě dlouhodobých změn), ale také krátkodobé zatížení mezilehlých nebo koncových zařízení.

5.2.2 Vliv změn obousměrného zpoždění na kalibrační funkci

V předchozí kapitole je ukázáno, jak se obousměrné zpoždění mění v čase, kdy v průměru se vyskytnou 2–3 změny za týden u každého měřeného páru. Za změnu úrovně zpoždění považujeme změnu větší než 0,2 ms, z důvodu možné chyby měření. Vezmeme-li v úvahu propagační zpoždění, tak změna o 0,2 ms znamená změnu vzdálenosti o přibližně 20 km (více viz 2.4.2).



Obr. 5.4: Graf kalibrační funkce pro referenční bod v Modeně, Itálie (`planetlab-2.ing.unimo.it`). Zobrazuje vývoj kalibračních dat v průběhu 6 týdnů – zobrazena jsou data sesbíraná v úterý ve 21:00 každých 14 dní.

Kalibrační funkce IP geolokačních funkcí založených na hranicích (např. CBG a Octant) berou v úvahu především maximální, případně minimální poměr mezi obousměrným zpožděním a vzdáleností. Z tohoto důvodu ani větší změna zpoždění nemusí znamenat změnu kalibrační funkce. K tomu dochází pouze, pokud se změní hraniční hodnoty v kalibračním grafu (viz obrázek 5.4). Naproti tomu metody založené na použití statistického proložení kalibračních dat (např. Spotter, SG a SBG) jsou na jakoukoliv změnu náchylné a více změn může vyvolat i úplnou změnu kalibrační funkce.

Na obrázku 5.4 je zobrazený kalibrační graf pro referenční uzel `planetlab-2.ing.unimo.it` umístěný v Modeně (Itálie). V grafu je zobrazeno 5 různých kalibračních dat zachycených každé druhé úterý v 21:00 po dobu necelých dvou měsíců.

Většina hodnot RTT kolísá okolo stejné úrovně a tím pádem vliv na kalibrační funkce IP geolokačních metod je malý, což dokládá i změna přímky Bestline popř. konvexní obálky u metody Octant. Výjimkou jsou data z 29. 4., kdy došlo k větší změně RTT (pro mnoho párů je nižší) a tím i převodní funkce vypočítané z tohoto měření jsou více odlišné. Vliv těchto odlišností na přesnost geolokace je podrobně ukázán v dalších kapitolách.

5.3 Výpočet vzdálenosti při použití kalibračních dat z předchozích měření

Pro určení nakolik změny zpoždění ovlivňují přesnost aktivní IP geolokace bylo provedeno měření testující použití kalibračních dat z různých časových úseků pomocí různých geolokačních metod. Vybrány byly metody, které si vytvářejí před začátkem geolokace převodní funkci mezi obousměrným zpožděním a vzdáleností, tedy například CBG, Octant a Spotter. Tato část se věnuje výpočtu geografické vzdálenosti za použití kalibračních dat z předchozích měření. Výpočet vzdálenosti mezi stanicemi na základě znalosti obousměrného zpoždění je klíčovým prvkem při IP geolokaci aktivními metodami. V této části je řešeno, nakolik výpočet s kalibračními daty z předchozích měření ovlivní přesnost výpočtu.

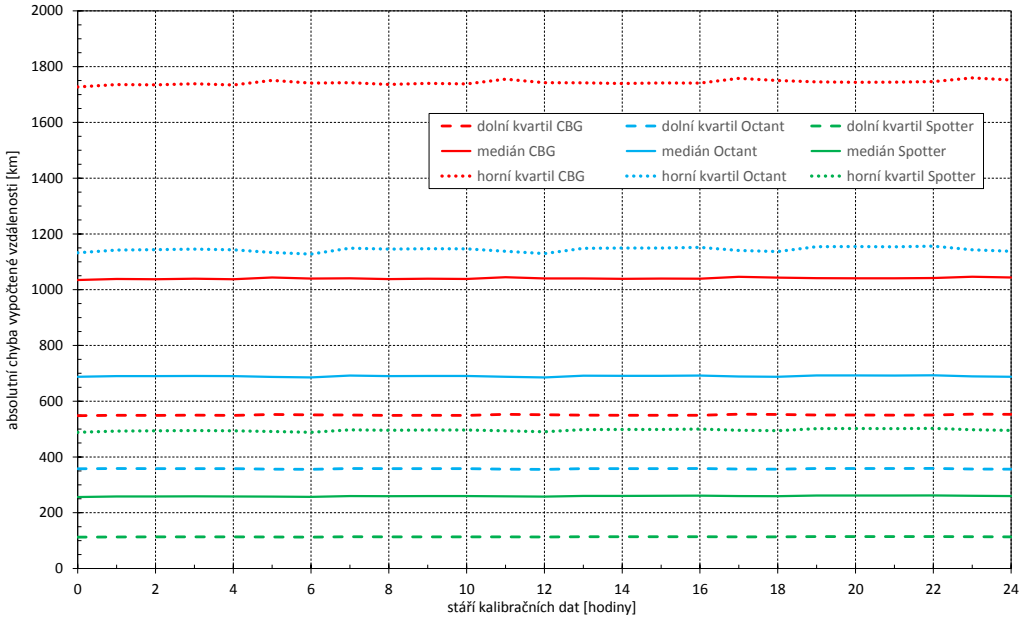
Výpočet geografické vzdálenosti na základě znalosti obousměrného zpoždění probíhá pomocí funkce vypočítané z kalibračních dat. Pro metodu CBG je tento výpočet popsán v kapitole 2.3.3 a k nalezení Bestline přímky (rovnice 2.2) je použito lineárního programování za podmínek dle rovnice 2.3. Metoda Octant využívá definice konvexní obálky (viz rovnice 2.5), pomocí níž je možné změřené zpoždění převést na pozitivní a negativní vzdálenosti (viz kapitola 2.3.6), podrobnější popis metody Octant je možné nalézt v [73]. Z metod aproximujících všechny hodnoty zpoždění byla vybrána metoda Spotter, jež využívá normální rozdělení pravděpodobnosti (viz rovnice 2.6) pro mapování zpoždění na pravděpodobnost vzdálenosti. Popis funkce metody Spotter je v kapitole 2.3.10 a podrobněji také v původní publikaci [46].

Celkově je v této kapitole použito 19 tisíc vzdáleností mezi měřící stanicí (landmarkem) a cílem. Pro získání výsledků uvedených dále, byl použit výpočet vzdálenosti za použití kalibračních dat z různých časových úseků. Následující podkapitoly popisují chybu výpočtu vzdálenosti v závislosti na různém stáří kalibračních dat.

5.3.1 Vliv stáří kalibračních dat řádech hodin na výpočet vzdálenosti

Protože je síťový provoz variabilní v průběhu dne a ve špičce je několikanásobně větší než v době minima, je tato podkapitola věnována zkoumání vlivu stáří kalibračních dat v rámci jednoho dne. Otázkou je, nakolik ovlivní vyšší (případně nižší) zatížení síťových prvků zpoždění, které je použito jak pro výpočet převodní funkce, tak pro samotný výpočet vzdálenosti. Na druhou stranu byla z dat získaných měřením použita minimální hodnota (tak jako je tomu u většiny aktivních IP geolokačních metod), právě kvůli eliminaci krátkodobého zatížení.

Celkově byly vyhodnoceny všechny kombinace kalibračních dat z různých časových úseků, získaných kontinuálním měřením po dobu šesti týdnů. To činilo přibližně 1000 různých měření pro každou dvojici bodů (celkově okolo sedmi miliónů výpočtů vzdálenosti). Z těchto dat byly zjištěny kvartily absolutní chyby výpočtu vzdálenosti, které jsou zobrazeny v grafu na obrázku 5.5. Počátek (hodnota 0 na ose x) znamená použití aktuálních kalibračních dat, další hodnoty na ose x znamenají použití x hodin starých kalibračních dat.



Obr. 5.5: Kvartily chyby výpočtu vzdálenosti pomocí metod CBG, Octant a Spotter při využití starších kalibračních dat – od jedné do dvaceti čtyř hodin.

Chyba výpočtu vzdálenosti je pro dolní kvartil a medián po celou dobu téměř konstantní, pro metody CBG a Octant dochází ke zvýšení chyby maximálně o 1 %, u metody Spotter maximálně o 2 %. Zvýšení chyby u horního kvartilu všech metod se pohybuje mezi 2 % a 3 % a i v grafu na obrázku 5.5 je znatelné její kolísání. Chyby

vypočtené vzdálenosti jsou velké – medián 1045 km, 688 km a 260 km pro metody CBG, Octant a Spotter při použití kalibračních dat jeden den starých. Vzhledem k tomu, že výsledná pozice geolokace je zjištěná (u metod CBG a Octant) průnikem hranic tvořených těmito vzdálenostmi, nemusí se chyby vypočtené vzdálenosti výrazně projevit ve výsledcích geolokace[1].

5.3.2 Vliv stáří kalibračních dat v rámci dnů na výpočet vzdálenosti

Obdobně jako na síťový provoz působí průběh dne, jsou v rámci datových toků znatelné také rozdíly mezi jednotlivými dny v týdnu (nejmarkantnější rozdíl je mezi pracovními dny a svátky). Výpočet vzdálenosti může ovlivnit použití dat z víkendového dne (síťový provoz je nižší) v den pracovní. Tato část se věnuje vlivu použití kalibračních dat ze stejného času, ale z jiného dne.

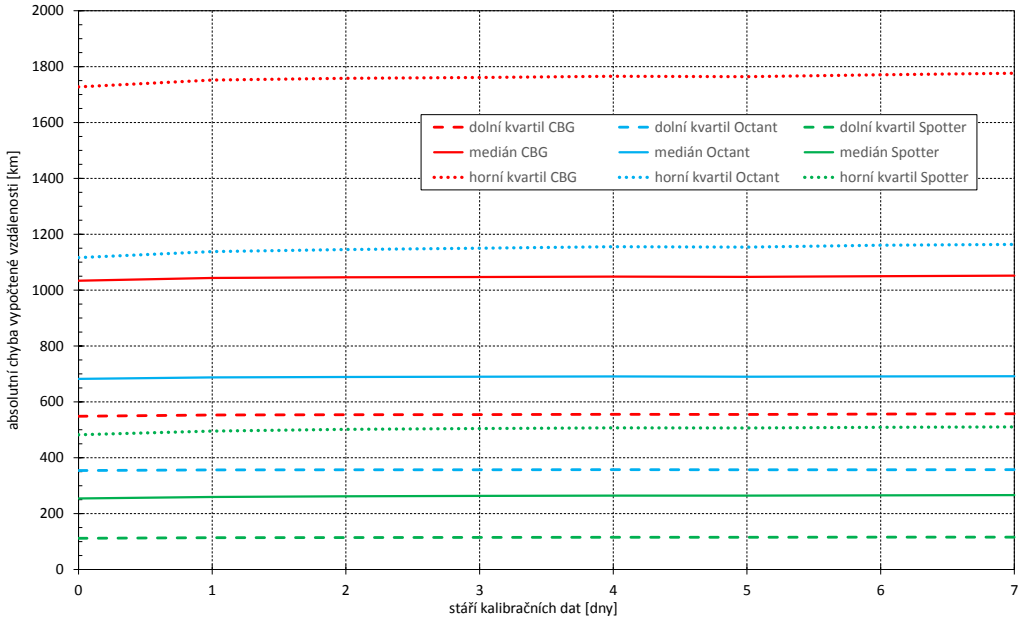
Data pro tuto část byla sbírána po dobu čtvrt roku, kdy probíhalo měření čtyřikrát každý den (po šesti hodinách). V průměru bylo získáno okolo 300 změřených dat pro každou vzdálenost (celkově okolo 4 miliónů). Na obrázku 5.6 je vykreslen vývoj změny kvartilů absolutní chyby při použití různě starých kalibračních dat (od jednoho do sedmi dnů), hodnoty v bodě nula jsou při použití aktuálních kalibračních dat.

Kvartily pro absolutní chybu vypočtené vzdálenosti za použití kalibračních dat den až sedm dní starých vykazují konstantní ráz, pouze chyba u horního kvartilu viditelně mírně stoupá (viz obrázek 5.6). U metody CBG dojde je relativní zvětšení chyby o 1,6–2,8 % (dolní kvartil až horní kvartil). Metoda Octant má relativní zvětšení chyby pro dolní kvartil o 0,9 %, pro medián o 1,4 % a pro horní kvartil o 4,2 %. K největšímu zvýšení chyby došlo u vzdálenosti vypočítané metodou Spotter – dolní kvartil o 3,7 %, medián o 4,7 % a horní kvartil o 5,8 %. Z těchto informací vyplývá, že při neprovedení kalibrace po dobu jednoho týdne dojde ke zvětšení chyby vypočtené vzdálenosti přibližně o 5 %.

5.3.3 Vliv stáří kalibračních dat za čtvrt roku na výpočet vzdálenosti

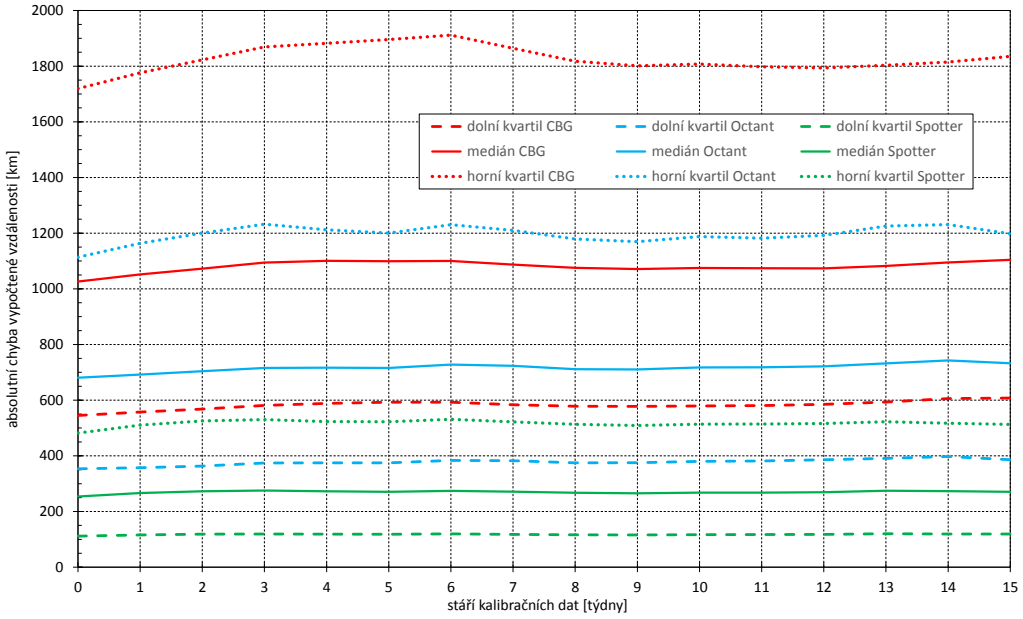
Poslední srovnání výpočtu vzdálenosti využívá data ze stejného času i dne v týdnu tak, aby jejich případný vliv byl eliminován. Obdobně jako v předchozím případě

[1]Důvodem je to, že ne všechny hranice určují výsledný průnik. Některé hranice jsou natolik široké, že zahrnují celou oblast průniku a často i celé mezní hranice jiného landmarku.



Obr. 5.6: Kvartily chyby výpočtu vzdálenosti pomocí metod CBG, Octant a Spotter, při využití starších kalibračních dat (jeden den až týden).

bylo využito dat posbíraných během čtvrt roku. Toto pro každou vzdálenost znamená v průměru 250 různých kombinací změřených dat. Pro delší rozmezí (15 týdnů) mezi změřenými a kalibračními daty to bylo v průměru pouze 20 různých kombinací.



Obr. 5.7: Kvartily chyby výpočtu vzdálenosti pomocí metod CBG, Octant a Spotter, při využití starších kalibračních dat (jeden až patnáct týdnů).

Na obrázku 5.7 jsou zobrazeny kvartily chyby vypočtené vzdálenosti jednotlivými metodami (CBG, Octant a Spotter). Přestože jejich chyba nevzrůstá přímo (vyjma horních kvartilů), je zřejmé, že po patnácti týdnech jsou vyšší než na počátku. U všech metod je znatelný nárůst chyby pro kalibrační data z doby 3–6 týdnů před měřením, která následně opět klesne. U metody CBG se chyba výpočtu vzdálenosti zvýší maximálně o 11,5 % (dolní kvartil), o 7,6 % (medián) a o 11,2 %. Pro metodu Octant je zvýšení chyby podobné – o 12,5 %, respektive o 9,1 % a o 10,7 %. Metoda Spotter měla změnu chyby vypočtené vzdálenosti nižší než metody CBG a Spotter, konkrétně se chyba zvýšila maximálně o 7,8 % pro dolní kvartil, o 8,5 % pro medián a o 10,2 % pro horní kvartil.

5.4 Přesnost IP geolokace při použití kalibračních dat z předchozích měření

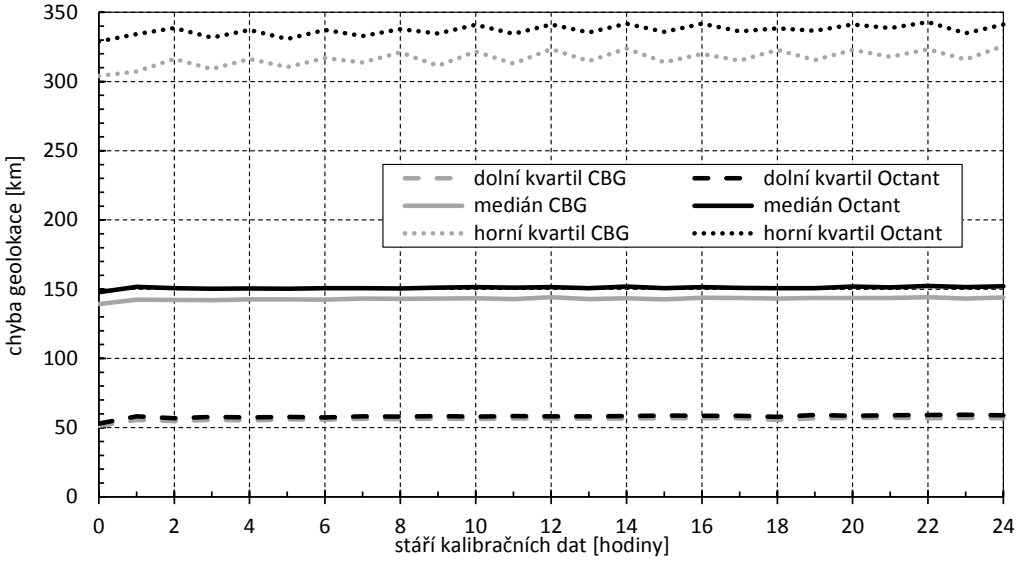
Předcházející kapitoly ukazovaly vliv použití neaktuálních kalibračních dat na výpočet vzdálenosti při znalosti obousměrného zpoždění. Z těchto výsledků bylo zřejmé, že použití neaktuálních dat způsobí snížení přesnosti výpočtu vzdálenosti. Na druhou stranu IP geolokační metody založené na vytváření hranic (CBG, Octant a další), které určují nejvzdálenější bod, kde se zařízení může nacházet, nemusí být touto chybou natolik ovlivněny. To je z důvodu, že tyto metody využívají průniku všech hranic a některé vypočtené hranice nemusí přímo ovlivňovat region průniku.

Obdobně jako v předchozí kapitole, jsou výsledky rozděleny do tří skupin dle stáří kalibračních dat (1 den, 1 týden a čtvrt roku). Kontrolní dataset cílů (zařízení, které byly lokalizovány) obsahoval 150 IP adres. Ne při všech měřeních bylo možné všechny cíle lokalizovat, a to především z důvodu podhodnocení geografické vzdálenosti, díky čemuž nebylo možné udělat průnik hranic. Tento problém narůstal především při použití starších kalibračních dat.

5.4.1 Přesnost IP geolokace při použití den starých kalibračních dat

Tato část se zaměřuje na vliv stáří kalibračních dat v řádech hodin na výslednou polohu určenou IP geolokací. Výsledky v kapitole 5.3.1 ukazují, že se stářím kalibračních dat mírně roste chyba vypočítané vzdálenosti (maximálně o 4,6 %). Jaký má tato chyba vliv na výsledek geolokace je možné vidět v grafu na obrázku 5.8.

V grafu jsou zobrazeny kvartily rozložení chyby pro metody CBG a Octant. Nejnižší chyba je při použití aktuálních kalibračních dat (hodnota 0 na ose x) a následně mírně narůstá (obdobně jako chyba výpočtu vzdálenosti). Pro dolní kvartil je tento



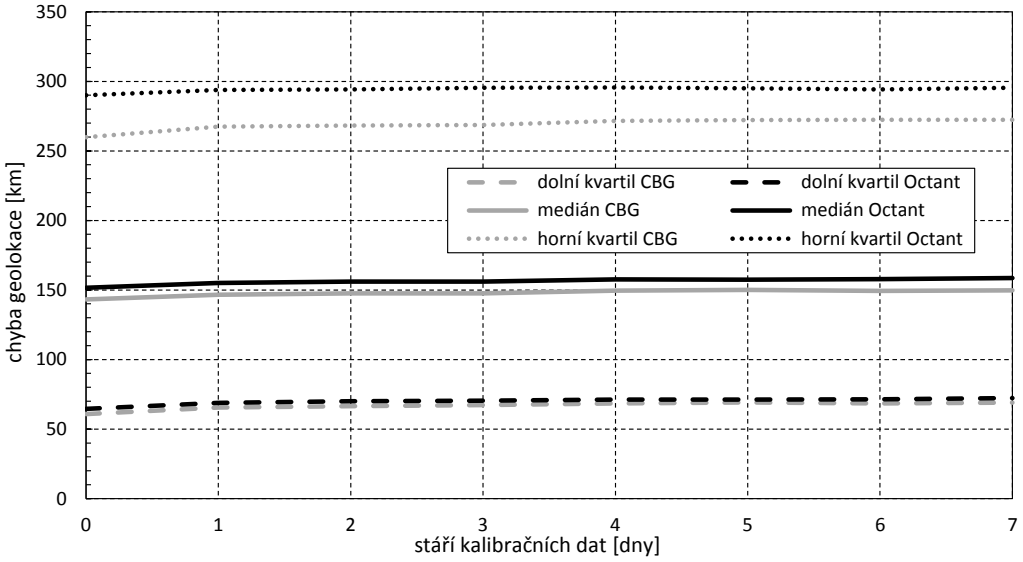
Obr. 5.8: Graf chyby geolokace za použití kalibračních dat získaných v časovém rozmezí hodina až 24 hodin. Zobrazení vývoje kvartilů chyby pro metody CBG a Octant.

nárůst o 11,4 % pro CBG a o 12,1 % pro Octant, medián narostl o 3,6 % respektive o 3,1 % a horní kvartil o 7,1 % respektive o 4,3 %. Z tohoto můžeme usoudit, že přesnost IP geolokace klesá nejvíce u cílů nalezených s vysokou přesností, u ostatních je rozdíl v použití den starých geolokačních dat relativně malý.

5.4.2 Přesnost IP geolokace při použití týden starých kalibračních dat

Kapitola navazuje na část 5.3.2, kde byl sledován vliv až týden starých kalibračních dat na výpočet vzdálenosti. Chyba u odhadu vzdálenosti narostla maximálně o necelých 5 % při použití sedm dnů starých kalibračních dat (pro metody CBG a Octant).

Obrázek 5.9 ukazuje obdobný graf jako v předchozí kapitole – průběh kvartilů chyby geolokace pomocí metod CBG a Octant. Obdobně jako u chyby geolokace při použití den starých dat (předchozí kapitola) je i zde znatelný mírný nárůst chyby IP geolokace. Konkrétně relativní nárůst chyby byl pro dolní kvartil o 13,7 % (CBG) a o 12 % (Octant), pro medián o 4,8 % respektive o 4,6 % a pro horní kvartil o 4,8 % a o 1,9 %. Celkově opět sledujeme mírný, ale ne kritický nárůst chyby, který je nejmarkantnější u čtvrtiny nejpřesněji zjištěných pozic.



Obr. 5.9: Graf chyby geolokace za použití kalibračních dat den až týden starých. Zobrazení vývoje kvartilů chyby pro metody CBG a Octant.

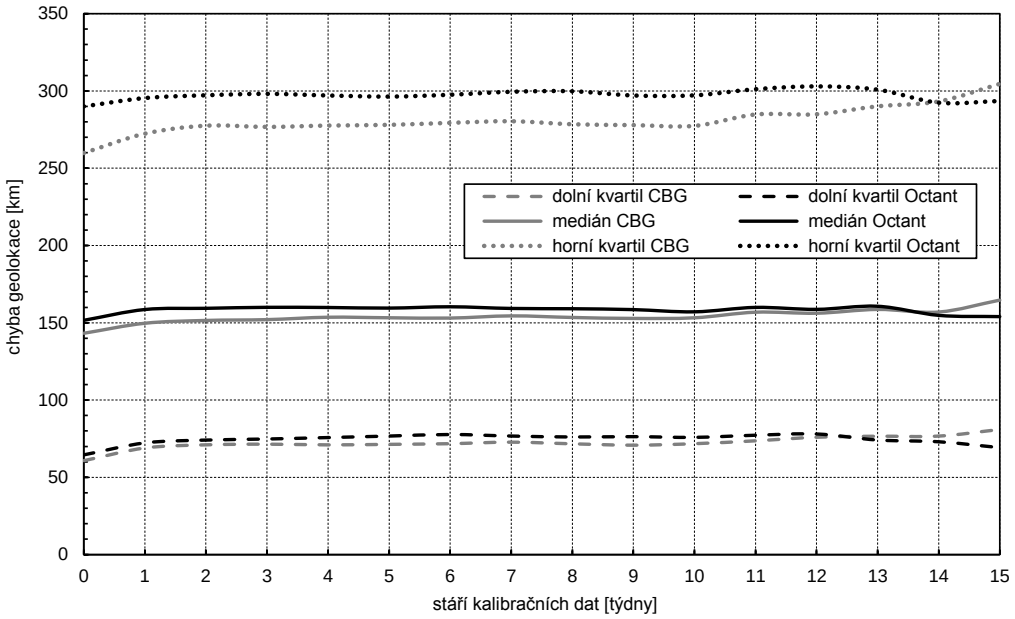
5.4.3 Přesnost IP geolokace při použití čtvrt roku starých kalibračních dat

Vzhledem k tomu nakolik v kapitole 5.3.3 kolísá přesnost výpočtu vzdálenosti, je důležité ukázat, jaký vliv to má na přesnost samotné geolokace. Při srovnání grafů na obrázcích 5.7 a 5.10 je zřejmé, že vzdálenost zjištěná při použití starých kalibračních dat kolísá obdobným způsobem i u chyby IP geolokace.

Dolní kvartil lokalizovaných zařízení má chybu vyšší o 33,2 % (u CBG) a o 20,9 % (pro Octant), u mediánu vzrostla chyba o 15 % respektive o 6 % a horní kvartil má chybu vyšší o 17,2 % u metody CBG a o 4,5 % u metody Octant. Zde se výrazně projevuje vliv přesnosti na geolokaci zařízení s největší přesností, u nichž přesnost nejvíce poklesla.

5.5 Poznatky užitečné pro návrh nové geolokační metody

Z výsledků prezentovaných v této kapitole je zřejmé, že stáří kalibračních dat má vliv na výpočet vzdálenosti i na přesnost geolokace samotné. Pro výpočet vzdálenosti tato chyba vzrostla maximálně o 12,5 % při použití 14 týdnů starých kalibračních dat. Při samotné IP geolokaci došlo ke zvýšení chyby až o 33 % při použití 15 týdnů starých geolokačních dat. Z tohoto důvodu můžeme poznamenat, že je důležité provádět kalibraci geolokačních metod (CBG, Octant a podobných) a to alespoň



Obr. 5.10: Graf chyby geolokace za použití čtvrt roku starých kalibračních dat. Zobrazení vývoje kvartilů chyby pro metody CBG a Octant.

jedenkrát týdně. V tomto případě můžeme čekat nárůst chyby maximálně o 10 %, která bude s rostoucím časem od poslední kalibrace narůstat.

Pro návrh nové geolokační metody je důležité také uvést, že charakter zpoždění je po většinu času trvalý. Ke změnám došlo pouze ve 3 % po sobě jdoucích vzorcích změřených dat. Dvě třetiny těchto změn jsou dlouhodobého rázu – trvají dvě hodiny a déle. U těchto změn byla zaznamenána i změna počtu mezilehlých zařízení a pravděpodobně došlo ke změně směrovací cesty. Z tohoto můžeme potvrdit předpoklad, že změna změřeného zpoždění, je ovlivněna počtem mezilehlých zařízení a délkou přenosových médií. Těchto vlastností bude využito při návrhu metody, která se bude opírat o rychlost šíření signálu přenosovým médiem a dle analýzy zpoždění odpočítávat dobu strávenou v mezilehlých zařízeních. Nově navrhovaná metoda bude postavena na vlastnostech zpoždění přenosového řetězce, kde jeho vybrané parametry budou odvozeny měřením. Z těchto důvodů není předpoklad zásadního vlivu kalibrace na novou metodu jako je tomu u metod CBG, Octant a Spotter, u kterých je přepočet zpoždění na vzdálenost přímo ovlivněn zjištěnými daty z kalibrace. Přesto je doporučeno provádět opakovaná měření popsaná v kapitole 4.1 nebo v [45] za účelem přizpůsobení parametrů vytvořené metody aktuálnímu stavu.

6 GEOLOKACE ZALOŽENÁ NA ANALÝZE PŘENOSOVÉHO ZPOŽDĚNÍ

Cílem této kapitoly je popsat návrh nové geolokační metody pro potřeby ověření důvěryhodnosti pozic zařízení (IP adresy) poskytovaných geolokačními databázemi. Metoda bude pracovat na principu vytvoření hranic v nejzazší vzdálenosti od referenčního bodu (landmarku), kde se může lokalizovaná IP adresa nacházet. Průnikem těchto oblastí (obdobně jako u metod CBG, SOI a dalších) vznikne region, ve kterém se na základě fyzických charakteristik komunikačního řetězce stanice musí nacházet. V případě, že se informace o poloze z databáze v regionu nenachází, je v databázi tento záznam označen za chybný a může být nahrazen jiným dostupným záznamem, například z jiného zdroje (způsoby plnění databází jsou rozebrány v kapitole 2.2.3).

6.1 Předpoklady pro návrh geolokační techniky

Při návrhu metody budeme na jedné straně vycházet z komplexního popisu zpoždění, který je podrobně uveden v kapitole 2.4. Na straně druhé stojí vlastnosti komunikačního řetězce, které je možné získat. Základním předpokladem pro návrh metody je nemožnost využít měřenou stanici (cíl) k provedení aktivního měření. Z tohoto důvodu jsme při návrhu metody omezeni na obecně podporované mechanismy jako je překlad IP adresy na doménové jméno a odpověď stanice na ICMP dotazy. Oboje může být v některých případech nefunkční, a to v případech, kdy neexistuje reverzní DNS záznam nebo má stanice příjem ICMP zpráv filtrován či zakázáno odesílání odpovědí na tyto zprávy. Dále je možné zjistit počet mezilehlých L3 zařízení pomocí nástroje `traceroute` nebo výpočtem z hodnoty TTL uvedené ve zprávě odpovídajícího cíle. Vzhledem k tomu, že výstup `traceroute` není vždy úplný[1], je vhodnější se spolehnout na výpočet počtu L3 zařízení na základě TTL návratové zprávy (podrobněji popsáno v kapitole 4.1.1).

Omezením lokalizace na základě IP adresy jsou systémy, které maskují IP adresu za jinou. Nejčastěji je skutečná IP adresa stanice (obvykle tzv. privátní adresa dle RFC 1918 [63]) nahrazena jinou adresou pomocí mechanismu NAT[2]. V těchto případech dochází k IP geolokaci pro adresu vystupující do sítě Internet – vnější rozhraní zařízení provádějícího NAT. U IP geolokace je předpokládáno, že zařízení provádějící NAT je v blízkosti skutečného zařízení (domácí router, lokální síťový uzel poskytovatele atd.).

[1] Pravděpodobný důvod je neodesílání nebo filtrování zpráv ICMP typu 11 některými síťovými zařízeními.

[2] Mimo NAT může změnu IP adresy cílového zařízení vyvolat také použití Proxy serveru, anonymizační sítě Tor a podobné.

6.2 Parciální složky zpoždění pro výpočet geografické vzdálenosti

Dle analýzy složek zpoždění provedené v kapitole 2.4 můžeme obousměrné zpoždění t_{rtt} definovat rovnicí 2.13. Pokud budeme uvažovat maximální možnou vzdálenost zdroje a cíle, kdy přenosové linky jsou vedeny nejkratší možnou cestou (přímo), můžeme z rovnice 2.13 vytknout zpoždění zapříčiněné přenosovými linkami t_{rs}, jak je ukázáno v následující rovnici

$$2\sum_{i=0}^{N} t_{\mathrm{rs},i} = t_{\mathrm{rtt}} - t_{\mathrm{pak,zdroj}} - t_{\mathrm{ser,zdroj}} - t_{\mathrm{des,zdroj}} - 2\sum_{i=0}^{N-1}(t_{\mathrm{des},i} + t_{\mathrm{z},i} + t_{\mathrm{f},i} + t_{\mathrm{ser},i}) + \\ + t_{\mathrm{des,cil}} - t_{\mathrm{odp,cil}} - t_{\mathrm{pak,cil}} - t_{\mathrm{ser,cil}} \,. \tag{6.1}$$

Zpoždění způsobené rychlostí šíření signálu dle rovnice 2.7 závisí pouze rychlosti šíření signálu médiem v_{rs} a délce přenosového média l. Pokud budeme na dálkových trasách předpokládat majoritní zastoupení optických vláken[3], můžeme pro rychlost šíření v médiu použít hodnotu $v_{\mathrm{rs}} = 194\ 895$ km/s, přesněji 0,65c. Výpočet maximální vzdálenosti l_{celk}, skládající se ze součtu všech délek přenosových médií, je poté možné zapsat po úpravě předcházející rovnice takto

$$l_{\mathrm{celk}} = 2(N+1)v_{\mathrm{rs}}\left(t_{\mathrm{rtt}} - t_{\mathrm{KZ,zdroj}} - 2\sum_{i=0}^{N-1}(t_{\mathrm{des},i} + t_{\mathrm{z},i} + t_{\mathrm{f},i} + t_{\mathrm{ser},i}) - t_{\mathrm{KZ,cil}}\right) . \tag{6.2}$$

Vzhledem k tomu, že všechna mezilehlá zařízení identifikovaná v rovnicích 6.1 a 6.2 pracují na třetí vrstvě RM ISO/OSI (L3), je u nich nutné načíst nejprve celou zprávu do paměti – tj. uplatní se stejné t_{des} a t_{ser}. Za předpokladu nejnižší dnes používané rychlosti v páteřních linkách (1 Gbit/s) je serializační zpoždění pro 98 B zprávu (velikost měřící ICMP zprávy i odpovědi) rovno 0,784 μs. Dále můžeme uvažovat zaokrouhlenou hodnotu 2 μs na jedno L3 zařízení (součet serializačního a deserializačního zpoždění).

Zpoždění vzniklé dobou strávenou v odchozích frontách t_{f} je stochastického charakteru a závisí na zatížení výstupního rozhraní prvku. Pokud budeme uvažovat minimální možnou hodnotu, je tato rovna nule. S touto hodnotou budeme počítat i nadále, neboť z každého měření je použita minimální zjištěná hodnota, přesto ale je možné, že v některých případech toto zpoždění nulové nebude.

[3]Metalická vedení pro připojení koncových stanic musí být z podstaty technologie Ethernet dlouhá maximálně 100 m, což je při odhadu vzdáleností stovek kilometrů zanedbatelná vzdálenost. Obdobně můžeme přistoupit i k použití metalického vedení případně bezdrátových spoje v přístupových sítích, délka je obvykle do 10 km. Vypočítaná vzdálenost při použití rychlosti šíření pro optiku by v tomto případě byla 8,7 km.

Dalším zpožděním, které ovlivňuje výpočet vzdálenosti je doba nutná k předání zprávy ze vstupu zařízení jeho na výstup t_z. Tuto dobu lze u některý produktů najít v katalogových listech, ale vzhledem nemožnosti rozpoznat výrobce mezilehlých uzlů, vyjdeme v dalších úvahách z výsledků uvedených v kapitole 4.1.1. Konkrétně pro prvotní návrh metody budeme pracovat se změřeným prvním kvartilem, tj. $t_{MZ} = 53\ \mu s$. Tato hodnota vzhledem ke způsobu zjištění zahrnuje nejen dobu nutnou k předání zprávy mezi vstupem a výstupem, ale serializační zpoždění t_z i zpoždění v odchozích frontách t_f.

Koncová zařízení, kromě parametrů stejných jako pro mezilehlá zařízení (t_f, t_z, t_{des} a t_{ser}) ovlivňují výpočet také dobou nutnou k vygenerování zprávy t_{pak} a odpovědi na ní t_{odp}. V dalších úvahách budeme předpokládat všechny tyto hodnoty jako jedno zpoždění vzniklé v koncových zařízeních $t_{KZ,zdroj} + t_{KZ,cil} = t_{KZ2}$, která dle [45] má minimální hodnotu $300\,\mu s$.

6.3 Vliv nepřímého vedení kabelů na výpočet vzdálenosti

Přenosové linky mezi zdrojovým a cílovým zařízením nekopírují přímou cestu mezi nimi, ale naopak jsou často několikanásobně delší. To je dáno především tím, že přenosová média jsou často pokládána podél významných silnic, železnic, elektrických vedení a produktovodů. Další vliv hraje směrovací politika mezi autonomními systémy. Vlivem nepřímého vedení kabelů a směrování pro účely geolokace se zabývaly například tyto publikace [6, 45, 5].

V kapitole 4.1.2 je v síti se známou topologií linek ukázán vliv jejich nepřímého vedení. Konkrétně využijeme minimální zjištěnou hodnotu $\frac{l_{skut}}{l_{prima}} = 1,38$, která pro vypočtené hodnoty představuje minimální násobek prodloužení trasy. Tuto hodnotu dále použijeme jako koeficient k_{nv} pro určení přímé vzdálenosti l_{prima} z vypočítané celkové vzdálenosti l_{celk}.

Autoři [45] využívají místo koeficientu vlastní empiricky zjištěnou hodnotu r_{nv}, která však zahrnuje i rychlost přenosových linek v_{rs}. Minimální jimi zjištěná hodnota pro výpočet vzdálenosti, ve které cíl s jistotou bude je $r_{nv} = 0,47c$. Když vezmeme v úvahu charakter námi zvoleného koeficientu k_{nv} můžeme převod mezi těmito parametry stanovit následující rovnicí

$$k_{nv} = \frac{v_{rs}}{r_{nv}} = \frac{0,65c}{0,47c}\ , \tag{6.3}$$

kde po dosazení výše uvedených hodnot vyjde stejná hodnota ($k_{nv} = 1,38$), jako byla zjištěna v kapitole 4.1.2.

Výpočet přímé geografické vzdálenosti l_{prima}, při znalosti obousměrného zpoždění t_{rtt} a počtu mezilehlých L3 zařízení N, poté můžeme provést pomocí této rovnice

$$l_{\text{prima}} = \frac{2(N+1)v_{\text{rs}}\left(t_{\text{rtt}} - 2Nt_{\text{MZ}} - t_{\text{KZ2}}\right)}{k_{\text{nv}}} \,. \tag{6.4}$$

6.4 Geodetický aparát pro výpočet geografické polohy

Vypočítaná vzdálenost se promítá na povrch Země, jejíž tvar lze v globále aproximovat modely, kterými můžou být koule, elipsoid nebo geoid [69]. S rozvojem GPS se rozšířil popis polohy pomocí zeměpisných souřadnic na referenčním elipsoidu WGS84[4]. K výpočtu vzdáleností na povrchu tohoto rotačního elipsoidu je možné využít tzv. Vincentyho rovnic [71].

Pro účely IP geolokace je nezbytné nalézt hranici mezní vzdálenosti (constraint) od referenčního (měřícího) bodu. Pomocí průniku takto nalezených hranic se vytvoří region, ve kterém se cílová stanice nachází. Zeměpisné souřadnice polohy cíle jsou následně určeny jako těžiště nalezeného regionu.

Souřadnice bodu v definované vzdálenosti od referenčního bodu

Prvním krokem bude nalezení hranice mezní vzdálenosti, kterou pro další použití popíšeme jako polygon spojující body na této hranici. Pro nalezení bodu ve vzdálenosti l po zemském povrchu pod úhlem α_1 od počátečního bodu se souřadnicemi ϕ_1 (zeměpisná šířka) a L_1 (zeměpisná délka) je možné použít přímé Vincentovy formule (Direct Vincenty Formula) [71].

Pro použití rovnic je nutné definovat U (snížená zeměpisná šířka) podle $U_1 = \arctan\left((1-f)\tan\phi_1\right)$. Ta je využita pro výpočet úhlové vzdálenosti z rovníku k počátečnímu bodu σ_1 pomocí první Vincentovy rovnice [71]

$$\sigma_1 = \arctan\left(\frac{\tan U_1}{\cos\alpha_1}\right) \,. \tag{6.5}$$

Ke zjištění úhlu spojnice mezi body na rovníku α je nutné vypočítat druhou Vincentovou rovnicí [71]

$$\sin\alpha = \cos U_1 \sin\alpha_1 \,. \tag{6.6}$$

Pro další výpočty je nutné definovat velikost $u^2 = \cos^2\alpha\left(\frac{a^2-b^2}{b^2}\right)$, jež je využita pro výpočet velikosti proměnných A a B. S těmi je dále pracováno při iterativním

[4]Velikosti poloos WGS84 elipsidu jsou $a = 6\ 378\ 137\,\text{m}$, $b = 6\ 356\ 752,3142\,\text{m}$, reciprocká hodnota zploštění $f = 298,257\ 223\ 6$ [69].

výpočtu úhlové vzdálenosti mezi body a lze je definovat pomocí rovnic

$$A \ = 1 + \frac{u^2}{16384}\left\{4096 + u^2\left[-768 + u^2(320 - 175u^2)\right]\right\} \ , \tag{6.7}$$

$$B = \frac{u^2}{1024}\left\{256 + u^2\left[-128 + u^2(74 - 47u^2)\right]\right\} \ [71]. \tag{6.8}$$

V dalších výpočtech je využita hodnota úhlové vzdálenosti mezi rovníkem a středem spojnice bodů σ_m. Následující tři rovnice jsou iterativně opakovány, dokud nedojde k změně úhlové vzdálenosti mezi body (σ) o menší než stanovenou hodnotu.

$$2\sigma_m = 2\sigma_1 + \sigma \ , \tag{6.9}$$

$$\begin{aligned}\Delta\sigma = B\sin\sigma\Big\{\cos(2\sigma_m) + \tfrac{1}{4}B\Big[\cos\sigma\Big(-1 + 2\cos^2(2\sigma_m)\Big) - \\ - \tfrac{B}{6}\cos(2\sigma_m)(-3 + 4\sin^2\sigma)\Big(-3 + 4\cos^2(2\sigma_m)\Big)\Big]\Big\} \ ,\end{aligned} \tag{6.10}$$

$$\sigma = \frac{l}{bA} + \Delta\sigma \ . \tag{6.11}$$

Jakmile je dosaženo požadované přesnosti σ, může být výsledná zeměpisná šířka ϕ_2 vypočtena pomocí rovnice

$$\phi_2 = \arctan\left(\frac{\sin U_1\cos\sigma + \cos U_1\sin\sigma\cos\alpha_1}{(1-f)\sqrt{\sin^2\alpha + (\sin U_1\sin\sigma - \cos U_1\cos\sigma\cos\alpha_1)^2}}\right) \ [71]. \tag{6.12}$$

Zeměpisná délka hledaného bodu L_2 je následně určena jako

$$L_2 = L + L_1 \ , \tag{6.13}$$

přičemž rozdíl v zeměpisné délce L je zjištěn pomocí

$$L = \lambda - (1 - C)f\sin\alpha\left\{\sigma + C\sin\sigma\left[\cos(2\sigma_m) + C\cos\sigma(-1 + 2\cos^2(2\sigma_m))\right]\right\} \ [71], \tag{6.14}$$

při znalosti C a λ (rozdíl v zeměpisné délce na pomocné kouli), které je možné vypočítat pomocí následujících dvou Vincentyho rovnic [71]

$$C = \frac{f}{16}\cos^2\alpha\Big[4 + f(4 - 3\cos^2\alpha)\Big] \ , \tag{6.15}$$

$$\lambda = \arctan\left(\frac{\sin\sigma\sin\alpha_1}{\cos U_1\cos\sigma - \sin U_1\sin\sigma\cos\alpha_1}\right) \ . \tag{6.16}$$

6.4.1 Hranice okolo referenčního bodu

Jestliže aplikujeme rovnice z předchozí kapitoly, za použití různých hodnot azimutů z výchozího bodu, získáme souřadnice bodů na mezní hranici od referenčního bodu. Hustotu těchto bodů je možné volně nastavit pomocí k_α, kterým definujeme počet bodů na hranici a jednotlivé úhly α_i můžeme získat dle

$$\alpha_i = i\frac{2\pi}{k_\alpha} \ . \tag{6.17}$$

Získané zeměpisné souřadnice jednotlivých bodů (ϕ_i, L_i) pak tvoří polygon popisující mezní vzdálenost okolo referenčního bodu.

6.4.2 Region průniku všech mezních vzdáleností

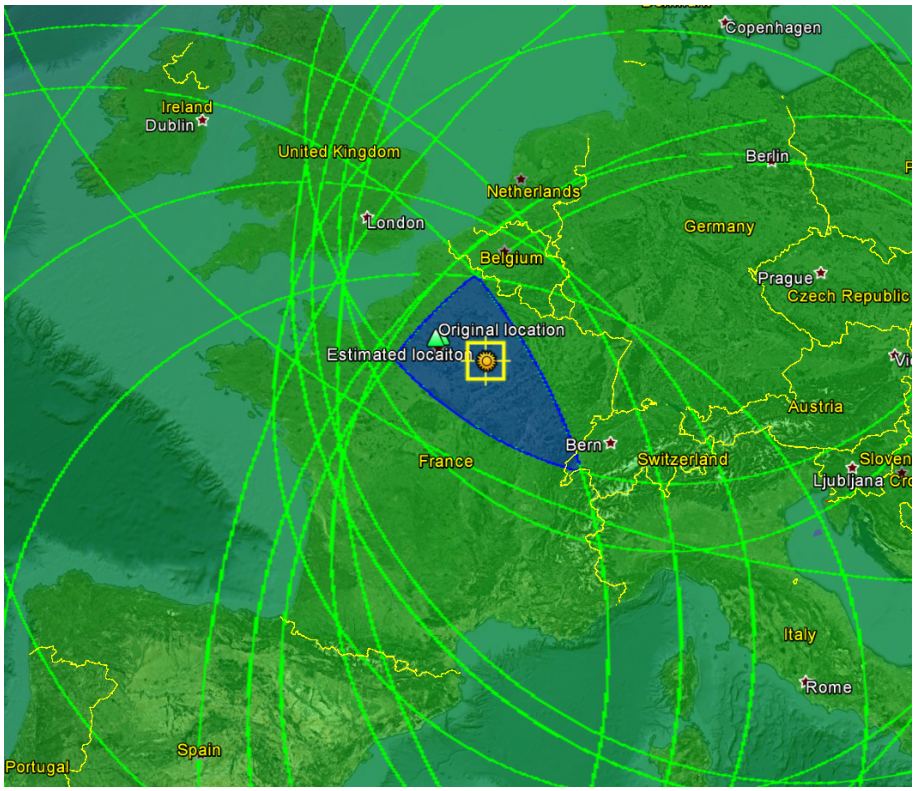
Vzhledem k dostupnému výpočetnímu výkonu dnešní doby, bylo přistoupeno k nalezení regionu průniku všech hranic referenčních bodů M pomocí ověření, zda jednotlivé body z polygonů popisujících hranice mezní vzdálenosti (ϕ_i, L_i) leží ve vzdálenosti l_j od jednotlivých referenčních uzlů (LM_j). Pro každou dvojici tvořenou referenčním bodem LM_j a hraničním bodem jiného referenčního bodu $\mathrm{LM}_{j+1}(\phi_i, L_i)$ byla pomocí inverzní Vincentyho formule [71] zjištěna vzdálenost $l\left(\mathrm{LM}_j(\phi, L), \mathrm{LM}_{j+1}(\phi_i, L_i)\right)$. Pokud tato vzdálenost byla menší nebo rovna než vypočtená vzdálenost l pro všechny referenční body (LM), došlo k zahrnutí tohoto bodu do množiny $\boldsymbol{R}$, která tvoří region průniku.

Body obsažené v množině $\boldsymbol{R}$, byly následně uspořádány tak, aby tvořily konvexní polygon souřadnic $\boldsymbol{R} = \{(\phi_i, L_i)\}$. Vizualizace polygonu vypočítaného pomocí výše popsané metody je na obrázku 6.1. Zde jsou na podkladu map z `Google Earth` vykresleny zeleně hranice mezních vzdáleností od referenčních bodů. Vypočítaný polygon je následně zobrazen modrou barvou, v obrázku je také ukázka IP geolokace pomocí navržené metody – *Estimated location* je vypočítaná poloha (těžiště regionu), *Original location* je skutečná poloha zařízení.

6.4.3 Těžiště a obsah regionu průniku mezních vzdáleností

Výstupem většiny aktivních IP geolokačních metod je vypočtená poloha měřené stanice a pro definici přesnosti plošná výměra vypočteného regionu. Zjištění souřadnic měřeného zařízení probíhá pomocí nalezení těžiště (centroidu). Pokud se výsledný region rozléhá na omezeném území, které lze aproximovat rovinou, je možné pro výpočet obsahu S tohoto regionu použít vztah

$$S \ = \frac{1}{2}\sum_{i=1}^{N} \phi_i \left(L_{i+1} - L_{i-1}\right) \ , \tag{6.18}$$



Obr. 6.1: Vizualizace výpočtu pozice stanice navrženou IP geolokační metodou. Zelené kruhy znázorňují hranice okolo jednotlivých referenčních bodů. Průnik všech hranic je vyznačen modrou barvou. *Estimated location* je vypočítaná poloha (těžiště regionu), *Original location* je skutečná poloha zařízení.

kde pořadí souřadnici všech (N) vrcholů polygonu (ϕ_i, L_i) je orientováno po směru hodinových ručiček [4]. Zjištěného obsahu regionu S je následně využito při výpočtu souřadnic těžiště regionu $(\phi_\mathrm{t}, L_\mathrm{t})$ pomocí následujících dvou rovnic

$$\phi_\mathrm{t} = \frac{1}{6S} \sum_{i=1}^{N} (\phi_i + \phi_{i+1}) (\phi_i L_{i+1} - \phi_{i+1} L_i) \ , \tag{6.19}$$

$$L_\mathrm{t} = \frac{1}{6S} \sum_{i=1}^{N} (L_i + L_{i+1}) (\phi_i L_{i+1} - \phi_{i+1} L_i) \ [4]. \tag{6.20}$$

Pro zajištění vyšší přesnosti, především v případě rozlehlého polygonu je možné využít metod pro výpočet centroidu a obsahu polygonu nad geoprostorovými daty. V tomto případě je ale nutné počítat s vyšší výpočetní náročností. Tyto funkce jsou připravené například v knihovně `org.geotools` pro programovací jazyk *Java* [22]. Tato knihovna obsahuje též funkci, která ověří, zda pozice leží uvnitř vypočítaného regionu. Toho je možné využít při ověřování platnosti geografických souřadnic, které poskytují IP geolokační databáze.

6.5 Kalibrační proces u metody CLBG

Jak bylo ukázáno v kapitole 5 pro zajištění přesnosti geolokačních metod CBG a Octant je nutná kalibrace. Tyto metody jsou však založeny na převodu zpoždění na vzdálenost pomocí aktuálně změřených dat. Metoda CLBG je oproti tomu postavená na fyzických charakteristikách zpoždění (převážně jeho deterministické části) a vliv na výsledný výpočet mají pouze parametry velikosti zpoždění na jedno mezilehlé zařízení t_{MZ}, zpoždění koncových zařízení t_{KZ2} a koeficient nepřímého vedení kabelů k_{nv}. Protože pro zjištění těchto hodnot bylo použito minimálních zjištěných hodnot, není nutné provádět kalibraci těchto hodnot před každým měřením. Zároveň předpokládáme, že vlastnosti, na kterých jsou závislé (vnitřní architektura síťových zařízení, poloha dálkových vedení), se nemění dramaticky v čase. Z těchto důvodů je vhodné provádět kalibraci těchto hodnot pouze výjimečně – například jedenkrát ročně pro ověření jejich stálé platnosti a případné aktualizaci.

Kalibraci velikosti zpoždění na jedno mezilehlé zařízení t_{MZ} je možné provést podle postupu uvedeného v kapitole 4.1.1 nebo podle experimentálního měření provedeného v [45]. Pro zjištění hodnoty zpoždění způsobené koncovými zařízeními t_{KZ2} je potřeba postupovat podle [45]. Ke zjištění koeficientu nepřímého vedení kabelů k_{nv} je možné využít postupu uvedeného v kapitole 4.1.2, nebo vyjít z měření uvedeného v [45] a výsledný parametr r_{nv} následně přepočítat dle rovnice 6.3. U všech hodnot je důležité použít minimální zjištěné hodnoty, případně nízkého percentilu (např. 1 %) zjištěných hodnot.

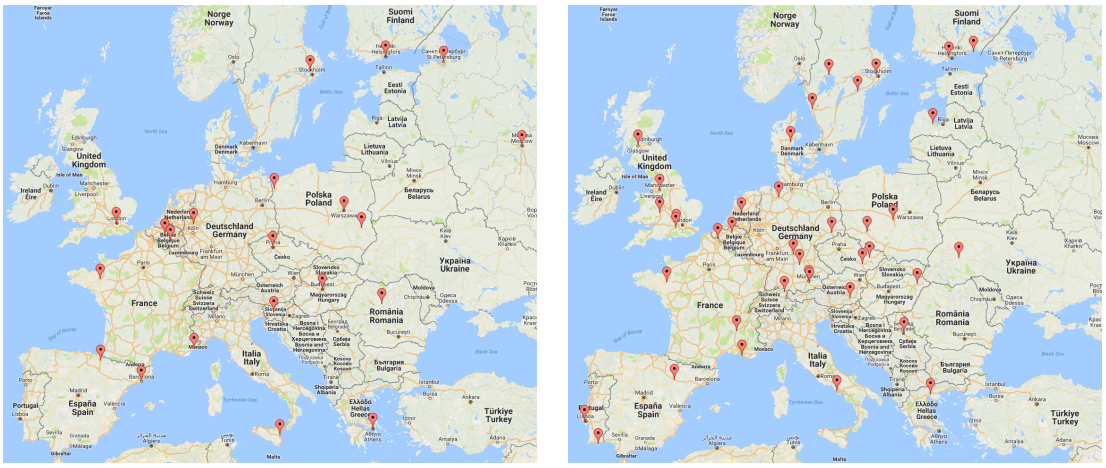
6.6 Srovnání výsledků nové metody

V kapitolách 6.2 a 6.3 je popsán postup přepočtu zpoždění na geografickou vzdálenost a je k tomu využito rovnice 6.4. Tato metoda je založena přepočtu zpoždění na vzdálenost dle velikosti propagačního zpoždění [83]. Protože její výsledky jsou závislé na délce přenosových médií (kabelů), byla nazvána Cable Length Based Geolocalization (CLBG) [84]. Následující podkapitoly poskytují srovnání této metody s jinými aktivními geolokačními metodami.

6.6.1 Porovnání výsledků CLBG s metodami GeoPing, ShortestPing a SOI

Výsledky metody CLBG (navržené v rámci doktorského studia) jsou srovnávány s geolokačními metodami GeoPing, ShortestPing a SOI. Měření bylo provedeno v síti PlanetLab pomocí 21 referenčních bodů (landmarků), které byly rovnoměrně rozmístěny po celé Evropě – obrázek 6.2 a). Jako cílů bylo využito 39 serverů se známou

polohou – servery významných evropských organizací, případně nevyužité servery ze sítě PlanetLab – obrázek 6.2 b). Z každého referenčního bodu bylo opakovaně změřeno zpoždění ke všem serverům a z opakovaných měření byla vybrána nejmenší hodnota – pro potlačení stochastické složky zpoždění způsobené např. zatížením mezilehlých prvků sítě.



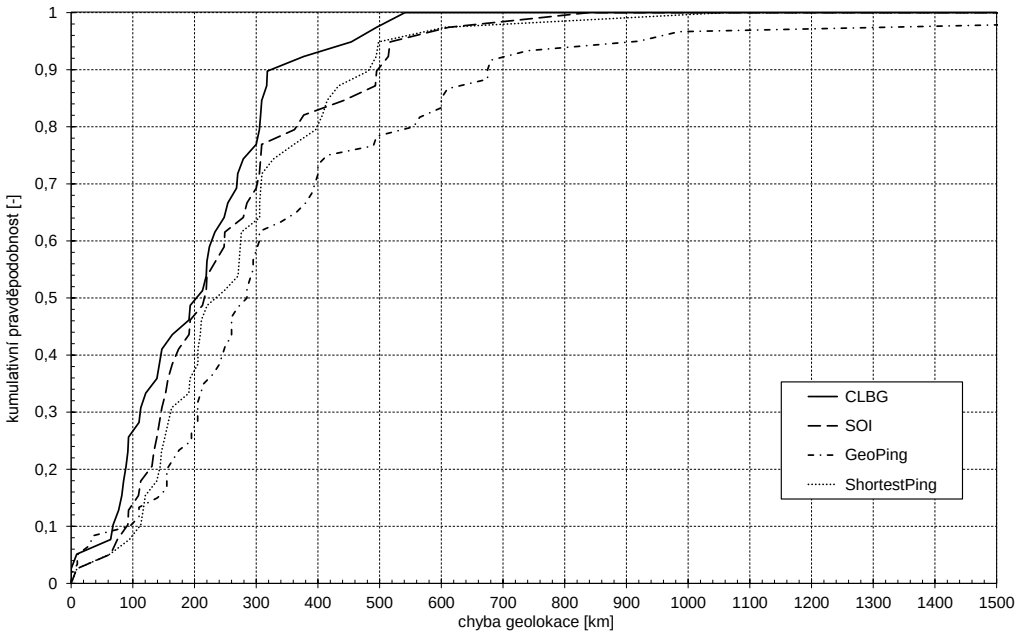
a) b)

Obr. 6.2: Rozmístění serverů využitých pro geolokaci pomocí metody CLBG. a) zobrazuje polohy referenčních bodů (serverů ze sítě PlanetLab), b) polohy ostatních uzlů využitých jako pasivní cíle.

Srovnání chyb IP geolokace pomocí metod CLBG, GeoPing, ShortestPing a SOI je v grafu kumulativní pravděpodobnosti na obrázku 6.3. Nejlépe při srovnání přesnosti vychází metoda CLBG, která má chybu pro dolní kvartil 102 km, medián chyby je 213 km, horní kvartil má chybu 290 km a průměrná chyba má hodnotu 208 km. Při porovnání křivek v grafu můžeme určit pořadí dalších metod SOI, ShortestPing a GeoPing (ve stejném pořadí medián chyby – 219 km, 248 km a 285 km). Zajímavé je, že jednoduchá metoda ShortestPing překonává složitější metodu GeoPing. Na druhou stranu metoda GeoPing je v pro cca 9 % cílů nejpřesnější, pravděpodobně je v těchto případech pasivní uzel blízko hledanému cíli.

6.6.2 Porovnání výsledků CLBG s metodami Octant, CBG a SOI

Druhé porovnání přesnosti metody CLBG vychází z měření popsaného v následující kapitole 7. Popis metodologie měření je možné najít v podkapitole 7.1. Pro toto měření bylo použito 21 měřících serverů a 5000 cílových adres, jejich rozložení je zobrazeno na obrázku 7.1. Porovnávána je přesnost pro pozice z datasetu ze zdroje [70].

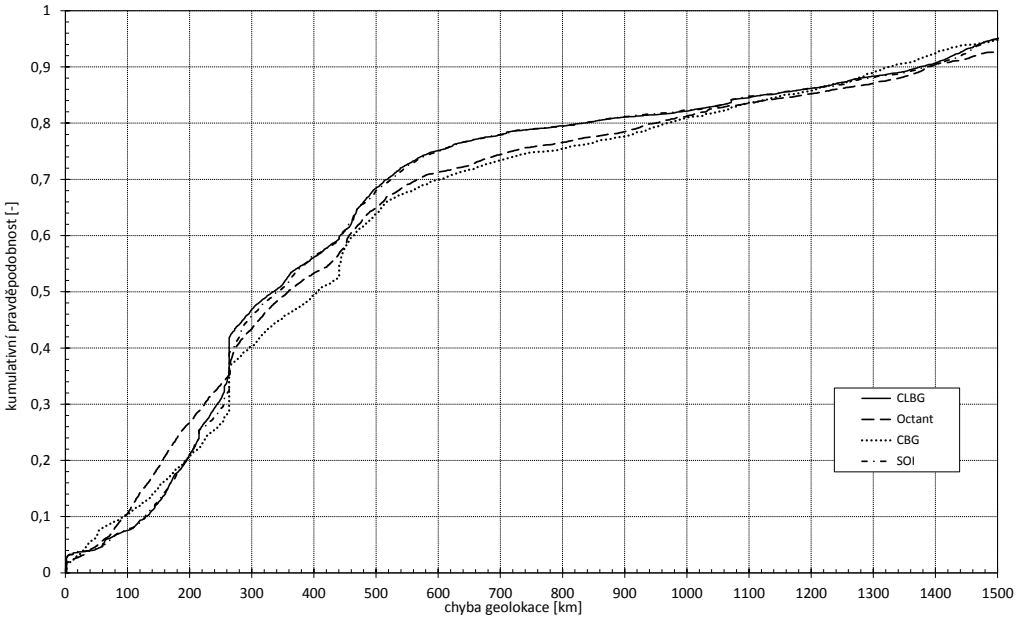


Obr. 6.3: Kumulativní distribuční funkce pravděpodobnosti pro chybu geolokace metod CLBG, SOI, ShortestPing, GeoPing [84].

Graf kumulativní pravděpodobnosti na obrázku 6.4 ukazuje dosahované přesnosti jednotlivých metod. Z výsledků není možné jednoznačně určit metodu s nejnižší chybou, protože pro různé vzorky hodnot se pořadí různí. Pro dolní kvartil hodnot dosahuje nejpřesnějších výsledků metoda Octant (chyba 186 km), stejnou chybu měla metoda CLBG a SOI (215 km), nejvyšší chybu (235 km) měla CBG. Medián chyby byl nejnižší pro metodu CLBG (334 km), následována metodou SOI (343 km), Octant (358 km) a CBG (406 km). V horním kvartilu byla nejpřesnější metoda CLBG (596 km), těsně za ní byla metoda SOI (601 km), dále Octant (719 km) a CBG (775 km).

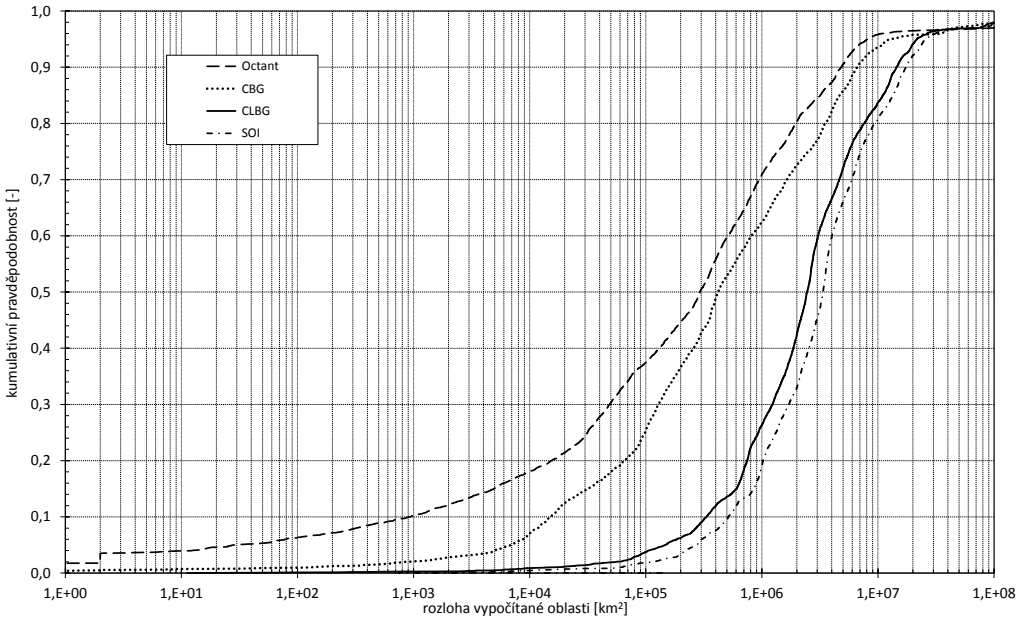
Přesnost metody CLBG nebyla hlavním kritériem při jejím návrhu. Především bylo důležité navrhnout metodu, která s jistotou určí region, ve kterém se stanice nachází. Zároveň je podmínkou spolehlivé funkce metody to, aby bylo možné vypočítat průnik mezních hranic – minimalizovat možnost podhodnocení hranic. Metoda CLBG a také metoda SOI měly pouze 0,48 % cílových adres, které nebylo možné lokalizovat a to z důvodu dočasné nedostupnosti IP adres při měření. Oproti tomu metoda Octant nebyla schopna lokalizovat 9 % cílových adres, ve většině případů z důvodu podhodnocení hranic. Metoda CBG takovýchto adres měla dokonce 24 %.

Nevýhodou metody CLBG a potažmo i metody SOI je velikost regionu, který zahrnuje možné pozice cíle. Důvodem je, že metoda CLBG byla navržena tak, aby generovala region, ve kterém se na základě analýzy zpoždění cílová stanice vždy nachází. Ve většině případů je tento region velký a často zahrnuje území celého státu,



Obr. 6.4: Kumulativní funkce pravděpodobnosti chyby geolokace metody CLBG ve srovnání s metodami Octant, CBG a SOI [82].

někdy i několika. Porovnání plochy regionu vypočítaných jednotlivými metodami je na obrázku 6.5. Pro vyjádření plochy regionu bylo použito logaritmické měřítko osy x. Vertikální osa opět zobrazuje kumulativní pravděpodobnost. Z grafu je zřejmé, že nejnižší obsah regionu má metoda Octant, která využívá tzv. negativních hranic, aby vyloučila málo pravděpodobné oblasti. Metoda CLBG má druhý největší vypočítaný region (větší má metoda SOI), ale toto bylo předpokládáno vzhledem k povaze metody.



Obr. 6.5: Semilogaritmický graf kumulativní pravděpodobnosti plochy regionu zjištěného pomocí metod CLBG, Octant, CBG a SOI [82].

7 OVĚŘENÍ DŮVĚRYHODNOSTI ZÁZNAMŮ GEOLOKAČNÍCH DATABÁZÍ

Cílem kapitoly je pomocí vytvořené metody CLBG provést ověření důvěryhodnosti záznamů z geolokačních databází. Jak je blíže popsáno v kapitole 2.2.3, jedná se o statické záznamy o poloze IP adresy, které v některých případech obsahují chybné záznamy s výraznou odchylkou od skutečné polohy. V této kapitole bude ověřeno pomocí metody CLBG, zda poloha IP adresy, oznamovaná různými geolokačními databázemi, je reálná. Tedy zda v oblasti, kterou ohlašuje CLBG leží dané souřadnice a pokud neleží je možné databázi navrhnout opravu záznamu, například použitím jiného zdroje informace. Pro srovnání výsledků metody CLBG s jinými, jsou v této kapitole referovány i metody CBG a Octant.

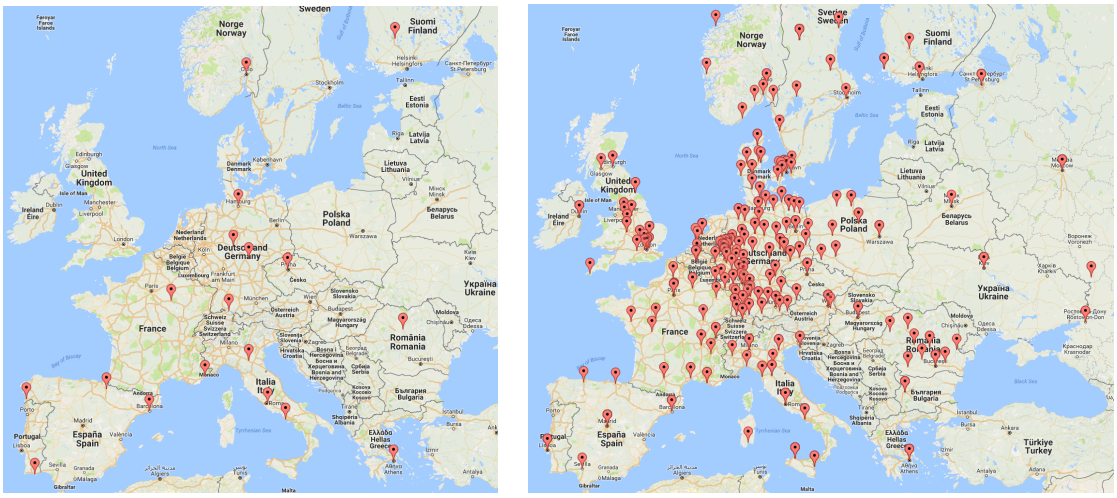
7.1 Metodika měření

7.1.1 Měřící stanice (landmarky)

Všechny použité aktivní IP geolokační metody (CBG, Octant i CLBG) využívají měřících stanic se známou polohou (landmarků). Pro tuto část práce byla použita experimentální síť PlanetLab [58], která obsahuje servery rozmístěné po celém světě a umožňuje přístup k nim za účelem provádění výzkumů. Oblast prováděného experimentu byla omezena na evropský kontinent, a tak ze sítě PlanetLab bylo vybráno celkem 21 aktivních serverů. Rozmístění serverů sloužících jako landmarky je na obrázku 7.1 a). Vzhledem k výsledkům uvedeným v kapitole 5 byla před každým měřením provedena kalibrace metod CBG a Octant k dosažení co nejpřesnějších výsledků.

7.1.2 Dataset cílových uzlů

Dataset cílových uzlů byl získán ze stránek projektu iPlane [70] provozovaného Univerzitou ve Washingtonu. V tomto datasetu byly ponechány pouze IP adresy umístěné na evropském kontinentě, aby se z nich následně náhodnou funkcí vybralo 5000, které budou otestovány. Jejich rozmístění je na obrázku 7.1 b). Pro tyto adresy byly zjištěny zeměpisné souřadnice jejich polohy pomocí tří geolokačních databází. Konkrétně se jednalo o veřejné databáze *Geobytes* [21] a *GeoLite2* [49] a demo verzi placené databáze *IP2Location* [34]. Jako údaje z další „databáze“ byly použity zeměpisné souřadnice uvedené přímo v datasetu projektu iPlane – dále označované jako *poloha z datasetu*.



a) b)

Obr. 7.1: Rozmístění měřících stanic a cílů pro ověření přesnosti geolokačních databází. a) zobrazuje polohy měřících stanic (serverů ze sítě PlanetLab), b) polohy ostatních uzlů využitých jako pasivní cíle.

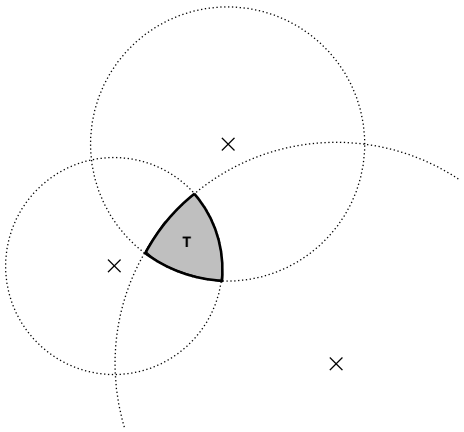
7.1.3 Postup měření

K měření bylo využito již zmiňovaných serverů ze sítě PlanetLab z nichž bylo postupně provedeno měření obousměrného zpoždění ke všem 5000 cílům. Toto bylo provedeno ve dvou várkách (vždy po 2500 cílů) kvůli opětovnému provedení kalibrace, neboť měření ke 2500 IP adresám trvalo přibližně deset hodin. Ke každému cíli bylo obousměrné zpoždění změřeno celkem třicetkrát a z těchto výsledků byla vybrána minimální hodnota kvůli eliminaci krátkodobého zatížení přenosových linek a mezilehlých zařízení (zpoždění dobou zpracování informace a v odchozích frontách). Počet mezilehlých zařízení byl spočítán na základě hodnoty TTL uvedené v odpovědi vzdálené stanice. Po provedení všech měření byla data stažena z měřících stanic a vyhodnocení bylo provedeno offline.

7.2 Ověření důvěryhodnosti údajů z geolokačních databází

Hlavním cílem této kapitoly bylo ověřit důvěryhodnost dat poskytnutých geolokačními databázemi. Na základě změřených dat (obousměrné zpoždění a počet mezilehlých uzlů) mezi referenčními uzly (landmarky) a adresami z datasetu, byla pomocí výše popsaných metod vypočítána předpokládaná poloha cíle a především region, kde se cíl nachází. Následně bylo vyhodnoceno, zda se souřadnice získané z jed-

notlivých geolokačních databází nacházejí ve vytyčené oblasti (ukázka oblasti je na obrázku 7.2).



Obr. 7.2: Vytvoření oblasti, ve které se nachází měřená stanice (cíl).

Výpočet oblasti, ve které se cíl nachází, v některých případech selhal kvůli podhodnocení vzdáleností – ty neutvořily průnik. Pro metodu CLBG k tomuto došlu v 0,5 % případů (cíl neodpověděl na ICMP zprávu), u metody CBG u 24 % a u metody Octant u 9 % [82].

V tabulce 7.1 jsou uvedeny procentuální počty záznamů z geolokačních databází, jejichž poloha byla uvnitř vypočítané oblasti. Při srovnání výsledků jednotlivých metod můžeme pozorovat, že metoda CLBG má nejvyšší procento záznamů z databází ve vypočítaném regionu. Toto je dáno tím, že tato metoda vytváří větší region, který však udává hranice, kde se dle fyzikálních zákonů stanice může nacházet. Metody CBG a Octant jsou svou podstatou vytvořeny pro co nejpřesnější geolokaci a tak je jejich region menší. Z tohoto důvodu je i více záznamů z datasetů, které v jejich regionech neleží. Nevyplývá z toho ale jistota, že poloha získaná z databáze je špatná. Proto v dalších srovnáních budeme upřednostňovat výsledky získané metodou CLBG.

Při porovnání výsledků jednotlivých databází v tabulce 7.1 můžeme dojít k závěru, že údaje z databáze *Geobytes* jsou nejméně důvěryhodné, neboť až 32 % záznamů leží mimo vypočtenou oblast. Databáze *GeoLite2*, která je neplacenou verzí komerční databáze *GeoIP2* je důvěryhodná pro více než 83 % záznamů. A nejlépe z uvedeného porovnání vychází demo verze placené databáze *IP2Location*, poloha jejíchž uzlů je v téměř 92 % případů ve vypočteném regionu. Za zmínku stojí i kontrola polohy uvedené v originálním datasetu (poloha z datasetu), která je dle metody CLBG správná v podobném množství případů jako poloho z databáze *IP2Location*. To, že není ve všech případech „poloha z datasetu“ správná, může být zapříčiněno stářím údajů v datasetu (3,5 roku), případně chybami v samotném datasetu.

Tab. 7.1: Procentuální vyjádření množství IP adres ležících v regionu, vypočítaném pomocí metod CBG, Octant a CLBG [82].

	CLBG	CBG	Octant
poloha z datasetu	91.9 %	46.5 %	44.0 %
Geobytes	68.0 %	39.4 %	36.5 %
GeoLite2	83.1 %	48.7 %	45.0 %
IP2Location	91.7 %	45.2 %	45.1 %

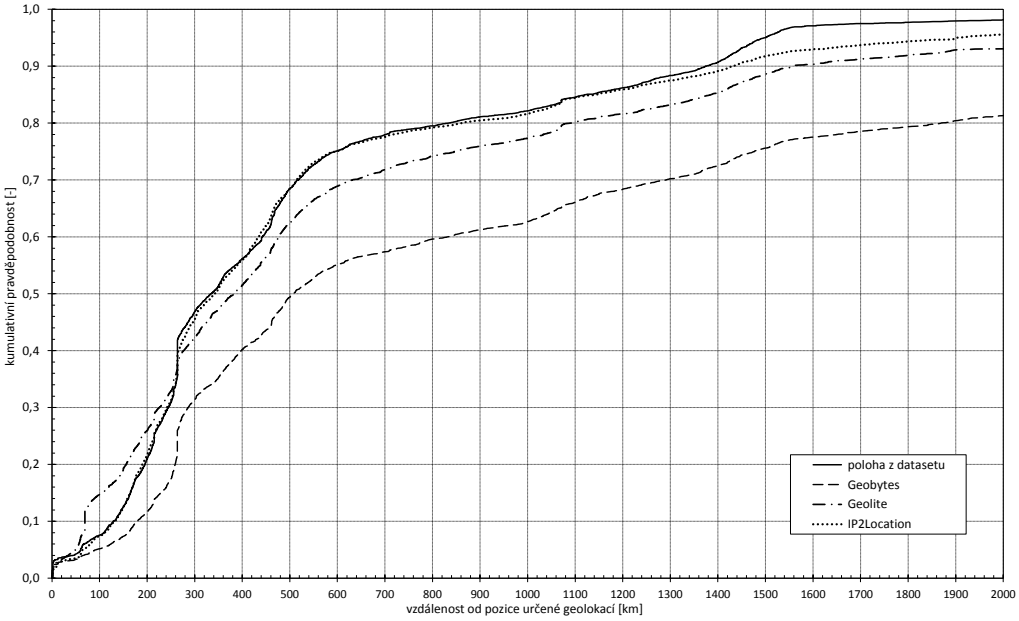
Srovnání databází, provedené v této kapitole, nebylo vytvořeno za účelem diskreditovat některé databáze, ale především pro ověření funkčnosti navržené metody. Celkově je však možné na základě těchto výsledků uvést, že obecně mají geolokační databáze při ohlašování přesné polohy chybovost – v našem případě minimálně 8 %. Cílem práce bylo nabídnout nástroj pro vylepšení těchto databází odstraněním chybných záznamů a jejich nahrazením informacemi z jiných zdrojů.

7.3 Srovnání přesnosti geolokačních databází

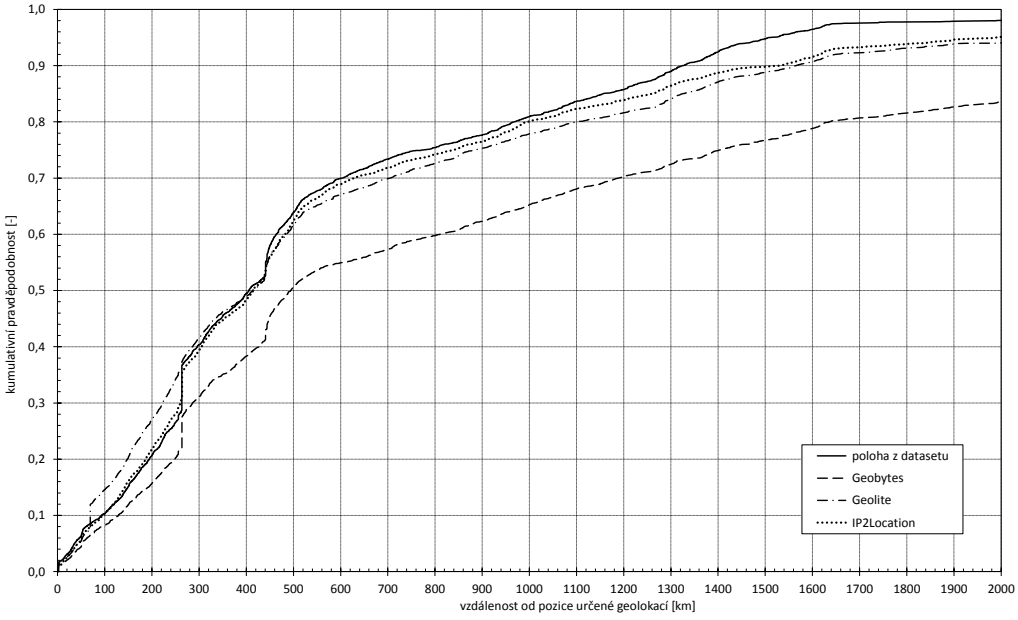
Přestože v předchozí kapitole je uvedeno, že cílem práce není hodnocení jednotlivých geolokačních databází, je možné díky změřeným datům porovnat i přesnost jednotlivých databází. Aktivní geolokační metody mají výrazně nižší přesnost než údaje z geolokačních databází [42], na druhou stranu jsou založeny na měření k cílovému zařízení. Z tohoto důvodu jsou v určení oblasti, kde se stanice nachází, důvěryhodnější.

Graf kumulativní pravděpodobnosti odchylky pozice vypočítané aktivní metodou CLBG od skutečné polohy je na obrázku 7.3. Z grafu a výsledků je patrné, že pouze 3 % vypočítaných pozic metodou CLBG se shoduje s pozicí udávanou databázemi (tolerance 5 km). Pro metody CBG a Octant (obrázky 7.4 a 7.5) jsou to pouze necelá 2 %.

Obecně je možné z grafů všech tří metod vyvodit, že databáze *Geobytes* dosahuje nejmenší přesnosti – při porovnání polohy s aktivními metodami. U databáze GeoLite2 je možné uvést, že je pro dolní kvartil záznamů nejpřesnější, dále však její přesnost klesá. Toto může být dáno tím, že se jedná o neplacenou verzi databáze, která může obsahovat některé záznamy více přesné a jiné záměrně méně (donucení ke koupi přístupu do databáze). Velmi podobnou přesnost vykazuje demo verze databáze IP2Location a poloha uvedená v originálním datasetu. Porovnání odchylky vzdálenosti od vypočtené polohy pomocí jednotlivých metod je uvedeno v tabulce

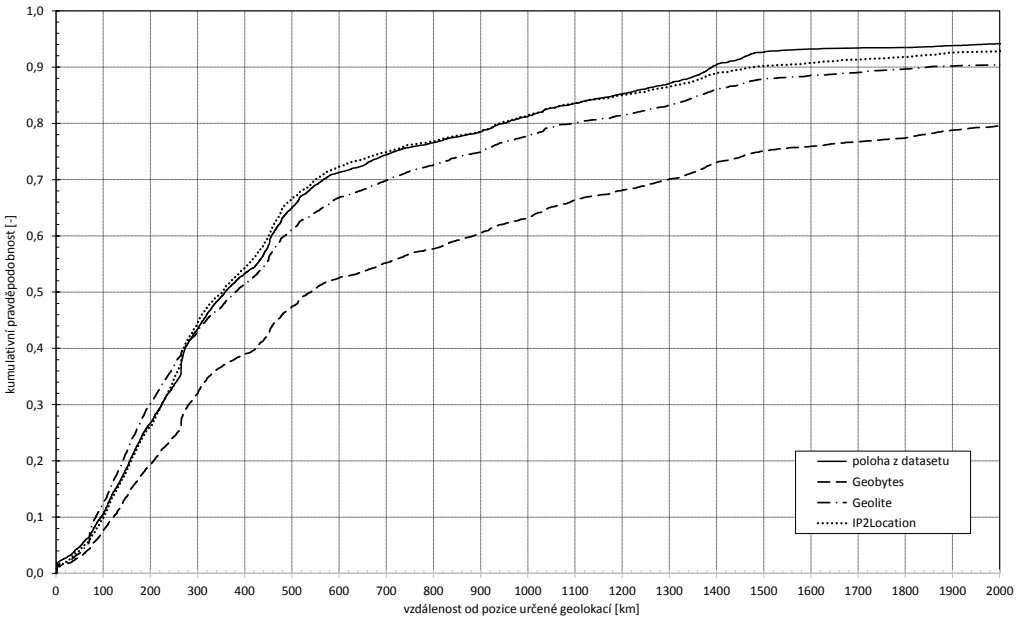


Obr. 7.3: Kumulativní distribuční funkce pravděpodobnosti chyby polohy vypočítané metodou CLBG od polohy získané z geolokačních databází [82].



Obr. 7.4: Kumulativní distribuční funkce pravděpodobnosti chyby polohy vypočítané metodou CBG od polohy získané z geolokačních databází [82].

7.2, kde jsou vypsány kvartily odchylek pro jednotlivé databáze.



Obr. 7.5: Kumulativní distribuční funkce pravděpodobnosti chyby polohy vypočítané metodou Octant od polohy získané z geolokačních databází [82].

Tab. 7.2: Odchylky polohy vypočítané metodami CBG, Octant a CLBG od pozice udávané databázemi.

	metoda CLBG			metoda CBG			metoda Octant		
	dolní kvartil	medián	horní kvartil	dolní kvartil	medián	horní kvartil	dolní kvartil	medián	horní kvartil
poloha z datasetu	215 km	334 km	596 km	235 km	406 km	775 km	186 km	359 km	719 km
Geobytes	263 km	507 km	1476 km	264 km	494 km	1406 km	256 km	540 km	1493 km
GeoLite2	190 km	383 km	596 km	183 km	414 km	882 km	168 km	380 km	704 km
IP2Location	215 km	342 km	846 km	226 km	414 km	839 km	190 km	353 km	903 km

8 ZÁVĚR

Hlavním cílem práce bylo vytvořit nástroj pro nalezení chybných záznamů v geolokačních databázích. Chybné záznamy jsou pak na základě zjištění nástroje odstraněny nebo nahrazeny informacemi z jiných zdrojů. Zmiňovaným nástrojem je nová geolokační metoda, která je schopna určit region, ve kterém se stanice musí nacházet. Metoda je založena na vlastnostech zpoždění v rámci celého komunikačního řetězce. Dílčí zpoždění identifikované v komunikačním řetězci byly předtím určeny na základě měření. Pro ověření funkčnosti metody bylo provedeno její porovnání se srovnatelnými přístupy a také její použití pro ověření polohy informací z geolokačních databází.

Prvním předpokladem pro návrh metody byla znalost jednotlivých složek zpoždění (kapitola 2.4) a především měření, za účelem nalezení jejich obvyklých hodnot. V kapitole 4 bylo popsáno provedené měření v síti se známou topologií. Výsledkem tohoto měření byla hodnota zpoždění na jeden mezilehlý prvek a parametr nepřímého vedení kabelů. Následně bylo ukázáno využití těchto hodnot při výpočtu vzdálenosti.

V dalším kroku bylo provedeno dlouhodobé sledování variace zpoždění a vliv této variace na IP geolokaci. Na základě měření v průběhu 4 měsíců bylo zjištěno, že zpoždění má po delší časový úsek (dny až týdny) konstantní úroveň. Změny se vyskytují u přibližně 3 % po sobě jdoucích hodnot, kdy 2 % znamenají skok zpoždění na jinou „stabilní“ úroveň, zbytek jsou jednorázové změny. Vliv všech změn na IP geolokaci byl prověřen použitím metod CBG, Octant a Spotter, u kterých byl testován přepočet zpoždění na vzdálenost s kalibračními daty z jiných časových období. Dále byla provedena IP geolokace (u metod CBG a Octant) při použití kalibračních dat z dřívějších měření. Ukázáno bylo, že se stářím kalibračních dat narůstá chyba výpočtu vzdálenosti, která byla nejvyšší při použití 6 týdnů starých dat (chyba až o 12,5 % vyšší). Na samotnou IP geolokaci mělo stáří kalibračních dat největší vliv u přesných hodnot (dolní kvartil) chyba byla vyšší až o 33 %.

Hlavní pozornost v dizertační práci je věnovaná návrhu nové geolokační metody, která byla dle své závislosti na propagačním zpoždění přenosových linek nazvána Cable Length Based Geolocalisation (CLBG) [84]. Princip metody tkví ve víceparametrovém přepočtu obousměrného zpoždění na přímou geografickou vzdálenost. Přepočet bere v úvahu počet zařízení na třetí vrstvě OSI a jejich minimální zpoždění. Dále také zpoždění vzniklé v koncových zařízeních a parametr nepřímého vedení kabelů ku přímé vzdálenosti. Po výpočtu maximální vzdálenosti od měřícího bodu (ve které se cílová stanice může nacházet) je nutné zjistit region průniku všech takto získaných hranic. K tomu je využito geodetických výpočtů popsaných v kapitole 6.4. Výsledná geografická poloha cílové stanice je určena v těžišti zjištěného regionu.

Metoda CLBG byla následně porovnána s ostatními používanými metodami, přičemž jednodušší metody (ShortestPing, GeoPing a SOI) překonávala v přesnosti. Při srovnání se složitějšími metodami (CBG a Octant) vykazovala měla srovnatelnou přesnost.

Návrh metody CLBG probíhal s ohledem na to, aby byla schopna určit oblast, kde se na základě přenosových charakteristik musí zařízení nacházet. Z těchto důvodů je oblast vytyčená touto metodou mnohem větší, než je tomu u obdobných metod. Uplatnění metoda najde v případech, kde je třeba ověřit polohu IP adresy, získané například z geolokační databáze. Toto bylo ukázáno v poslední kapitole, kde na vzorku 5000 cílových adres byla ověřena jejich poloha zjištěná ze tří geolokačních databází. U nejlepší databáze z testu (*IP2Location*) bylo ukázáno, že pro 8 % IP adres se její pozice nacházela mimo region vypočítaný metodou CLBG. Tímto bylo demonstrováno použití metody a také ukázán postup, jak je možné ověřit jakoukoliv polohu měřením v síti Internet.

LITERATURA

[1] Arif, M. J. – Karunasekera, S. – Kulkarni, S.: GeoWeight: internet host geolocation based on a probability model for latency measurements. In *Proceedings of the Thirty-Third Australasian Conferenc on Computer Science* – Volume 102, ACSC '10, Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2010, ISBN 978-1-920682-83-5, s. 89–98. URL `http://dl.acm.org/citation.cfm?id=1862199.1862209`

[2] Arlos, P. – Fiedler, M.: Accuracy Evaluation of Ping and J-OWAMP. In *Swedish National Computer Networking Workshop*, Luleå, 2006.

[3] Balakrishnan, M. – Mohomed, I. – Ramasubramanian, V.: Where is That Phone?: Geolocating IP Addresses on 3G Networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, New York, NY, USA: ACM, 2009, ISBN 978-1-60558-771-4, s. 294–300, doi:10.1145/1644893.1644928. URL `http://doi.acm.org/10.1145/1644893.1644928`

[4] Bayer, T.: *Algoritmy v digitální kartografii*. Univerzita Karlova v Praze, Nakladatelství Karolinum, první vydání, 2008, ISBN 978-80-246-1499-1, 252 s.

[5] Bendale, J. – Kumar, J. R.: Review of different IP geolocation methods and concepts. *International Journal of Computer Science and Information Technologies*, ročník 5, č. 1, 2014: s. 436–440.

[6] Bovy, C. J. – Mertodimedjo, H. T. – Hooghiemstra, G., et al.: Analysis of End-to-end Delay Measurements in Internet. In *Passive and Active Measurement Workshop-PAM*, Vol. 2002, s. 8, URL `http://www.pamconf.net/2002/Analysis_of_End_to_end_Delay_Measurements_in_Internet.pdf`

[7] Bradner, S. – McQuaid, J. – *RFC 2544: Benchmarking Methodology for Network Interconnect Devices*, [online], 1999. URL `https://www.ietf.org/rfc/rfc2544.txt`

[8] Butkovic, A. – Orucevic, F. – Tanovic, A.: Using Whois Based Geolocation and Google Maps API for support cybercrime investigations. In *Applied electromagnetics, wireless and optical communication*s, 2013: s. 194-200.

[9] Cesnet: *CESNET, zájmové sdružení právnických osob*, [online], c2017. URL `http://www.cesnet.cz`. [cit. 25. 8. 2017]

[10] Chandrasekaran, B. – Bai, M.; Schoenfield, M. et al.: Alidade: IP Geolocation without Active Probing. *Department of Computer Science, Duke University*, Technical Report, CS-TR-2015-001, 2015.

[11] Constantinescu, D. – Popescu, A.: Modeling of One-Way Transit Time in IP Routers. In *Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services (AICT-ICIW'06)*, Feb 2006, s. 16–16, doi:10.1109/AICT-ICIW.2006.132.

[12] Crotti, M. – Gringoli, F. – Salgarelli, L.: Impact of Asymmetric Routing on Statistical Traffic Classification. In *Proceedings of the 28th IEEE Conference on Global Telecommunications*, GLOBECOM'09, Piscataway, NJ, USA: IEEE Press, 2009, ISBN 978-1-4244-4147-1, s. 655–662. URL `http://dl.acm.org/citation.cfm?id=1811380.1811488`

[13] Dabek, F. – Cox, R. – Kaashoek, F. et al.: Vivaldi: A Decentralized Network Coordinate System. In *SIGCOMM Comput. Commun. Rev.*, ročník 34, č. 4, Srpen 2004: s. 15–26, ISSN 0146-4833, doi:10.1145/1030194.1015471. URL `http://doi.acm.org/10.1145/1030194.1015471`

[14] Dahnert, A.: HawkEyes: An advanced IP Geolocation approach: IP Geolocation using semantic and measurement based techniques. In *2011 Second Worldwide Cybersecurity Summit (WCS)*, June 2011, s. 1–3.

[15] Davis, C. – Vixie, P. – Goodwin, T. et al.: *RFC 1876: A Means for Expressing Location Information in the Domain Name System*. Leden 1996. URL `http://tools.ietf.org/html/rfc1876`

[16] db-ip.com: *IP Geolocation and Network Intelligence*, [online], c2017. URL `http://db-ip.com/` [cit. 25. 5. 2017]

[17] Digital Envoy: *Trusted Geolocation - Digital Element*, [online], c2017. URL `http://www.digitalelement.com/geolocation/` [cit. 23. 5. 2017]

[18] Ding, S. – Luo, X. – Yin, M. et al.: An IP geolocation method based on rich-connected sub-networks. In *2015 17th International Conference on Advanced Communication Technology (ICACT)*, Červenec 2015, ISSN 1738-9445, s. 176–181, doi:10.1109/ICACT.2015.7224779.

[19] Dingledine, R. – Mathewson, N. – Syverson, P.: Tor: The Second-Generation Onion Router. In *In Proceedings of the 13th USENIX Security Symposium*, 2004, s. 303–320.

[20] Eriksson, B. – Barford, P. – Maggs, B. et al.: *Posit: An Adaptive Framework for Lightweight IP Geolocation*. Technická zpráva, BUCS-TR-2011-018, Computer Science Department, Boston University, Červenec 2011. URL `http://www.cs.bu.edu/techreports/pdf/2011-018-posit.pdf`

[21] Geobytes: *Simple Solutions*. [online]. URL `http://geobytes.com` [cit. 3. 5. 2017]

[22] GeoTools.org: GeoTools The Open Source Java GIS Toolkit. [online], c2016. URL `http://geotools.org/` [cit. 3. 5. 2017]

[23] Gueye, B. – Uhlig, S. – Ziviani, A. et al.: Leveraging Buffering Delay Estimation for Geolocation of Internet Hosts. In *Proceedings of the 5th International IFIP-TC6 Conference on Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, NETWORKING'06, Berlin, Heidelberg: Springer-Verlag, 2006, ISBN 3-540-34192-7, 978-3-540-34192-5, s. 319–330, doi: 10.1007/11753810_27. URL `http://dx.doi.org/10.1007/11753810_27`

[24] Gueye, B. – Ziviani, A. – Crovella, M. et al.: Constraint-based geolocation of internet hosts. In *IEEE/ACM Trans. Netw.*, ročník 14, č. 6, Prosinec 2006: s. 1219–1232, ISSN 1063-6692. URL `http://dx.doi.org/10.1109/TNET.2006.886332`

[25] Gunther, S. – Schlamp, J. – Carle, G.: Spring-based geolocation. In *Network Operations and Management Symposium (NOMS), 2012 IEEE*, Duben 2012, ISSN 1542-1201, s. 546–549, doi:10.1109/NOMS.2012.6211952.

[26] Guo, C. – Liu, Y. – Shen, W. et al.: Mining the Web and the Internet for Accurate IP Address Geolocations. In *IEEE INFOCOM 2009*, Duben 2009, ISSN 0743-166X, s. 2841–2845, doi:10.1109/INFCOM.2009.5062243.

[27] Hillmann, P. – Stiemert, L. – Dreo, G. et al.: On the Path to High Precise IP Geolocation: A Self-Optimizing Model. *International Journal of Intelligent Computing Research (IJICR)*, ročník 7, číslo 1, březen 2016, ISSN: 0976-9013.

[28] Hillmann, P. – Stiemert, L. – Rodosek, G. D. et al.: Dragoon: Advanced modelling of IP geolocation by use of latency measurements. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, Prosinec 2015, s. 438–445, doi:10.1109/ICITST.2015.7412138.

[29] Holdener, A. T.: *HTML5 Geolocation*. O'Reilly Media, Inc., první vydání, 2011, ISBN: 978-1-449-30472-0.

[30] Holland, O. – Dohler, M.: Geolocation-Based Architecture for Heterogeneous Spectrum Usage in 5G. In *2015 IEEE Globecom Workshops (GC Wkshps)*, Prosinec 2015, s. 1–6, doi:10.1109/GLOCOMW.2015.7414189.

[31] HostIP.info: *IP Address Lookup Hostip.info*, [online]. URL `http://www.hostip.info` [cit. 24. 5. 2017]

[32] Huffaker, B. – Fomenkov, M. – Claffy, k.: *Geocompare: a comparison of public and commercial geolocation databases*. Technická zpráva, Centre for Advanced Internet Architectures (CAIDA), Květen 2011.

[33] IP2Location: *Free IP Geolocation Database: IP2Location LITE*, [online], c2017. URL `http://lite.ip2location.com/` [cit. 24. 5. 2017]

[34] IP2Location: *IP Address to Identify Geolocation Information*, [online], c2017. URL `http://www.ip2location.com` [cit. 24. 5. 2017]

[35] ipinfo.io: *IP Address Details - ipinfo.io*, [online].
URL `http://ipinfo.io/` [cit. 3. 6. 2017]

[36] IPligence: *IP Address Location of Web Visitors & Geolocation services*, [online], c2013. URL `http://www.ipligence.com/` [cit. 3. 6. 2017]

[37] Jiang, H. – Liu, Y. – Matthews, J. N.: IP geolocation estimation using neural networks with stable landmarks. In *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Duben 2016, s. 170–175, doi:10.1109/INFCOMW.2016.7562066.

[38] Jinxia, W. – Xiaoyan, X. – Min, Y. et al.: IP Geolocation Technology Research Based on Network Measurement. In *2016 Sixth International Conference on Instrumentation Measurement, Computer, Communication and Control (IMCCC)*, Červenec 2016, s. 892–897, doi:10.1109/IMCCC.2016.97.

[39] Katz-Bassett, E. – John, J. P. – Krishnamurthy, A. et al.: Towards IP geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, IMC '06, New York, NY, USA: ACM, 2006, ISBN: 1-59593-561-4, s. 71–84. URL `http://doi.acm.org/10.1145/1177080.1177090`

[40] Kester, J.-J.: Comparing the Accuracy of IPv4 and IPv6 Geolocation Databases. In *24th Twente Student Conference on IT*, ročník 24, číslo 1, 2016.

[41] Kliman-Silver, C. – Hannak, A. – Lazer, D. et al.: Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, New York, NY, USA: ACM, 2015, ISBN 978-1-4503-3848-6, s. 121–127, doi:10.1145/2815675.2815714. URL `http://doi.acm.org/10.1145/2815675.2815714`

[42] Komosný, D. – Vozňák, M. – Bezzateev, S. et al.: The Use of European Internet Communication Properties for IP Geolocation. *Information technology and control*, ročník 45, č. 01, 2016: s. 77–85, ISSN 1392-124X, doi:http://dx.doi.org/10.5755/j01.itc.45.1.11062.

[43] Koton, J.: *Moderní síťové technologie*. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, Brno, 2014, ISBN 978-80-214-5026-4.

[44] Krumnikl, M.: *Měření latence síťových prvků*, [online], 2005, str. 15. URL `http://www.cs.vsb.cz/grygarek/SPS/projekty0405/LatenceSW_kru106.pdf`

[45] Laki, S. – Mátray, P. – Hága, P. et al.: A model based approach for improving router geolocation. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, ročník 54, č. 9, Červen 2010: s. 1490–1501, ISSN 1389-1286.URL `http://dx.doi.org/10.1016/j.comnet.2009.12.004`

[46] Laki, S. – Mátray, P. – Hága, P. et al.: Spotter: A model based active geolocation service. In *Proceedings IEEE INFOCOM, Shanghai, 2011*, s. 3173–3181. URL `http://dx.doi.org/10.1109/INFCOM.2011.5935165`

[47] Lemmon, T. R. – Gerdan, G. P.: The Influence of the Number of Satellites on the Accuracy of RTK GPS Positions. *The Australian Surveyor*, ročník 44, číslo 1, 1999: str. 7.

[48] Li, H. – He, Y. – Xi, R. et al.: A Complete Evaluation of the Chinese IP Geolocation Databases. In *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Červen 2015, s. 13–17, doi:10.1109/ICICTA.2015.13.

[49] MaxMind: *GeoLite2 Free Downloadable Databases: MaxMind Developer Site*, [online], c2017. URL `https://dev.maxmind.com/geoip/geoip2/geolite2/` [cit. 8. 6. 2017]

[50] MaxMind: *IP Geolocation and Online Fraud Prevention*, [online], c2017. URL `https://www.maxmind.com` [cit. 8. 6. 2017]

[51] Mochalski, K. – Micheel, J. – Donnelly, S.: Packet Delay and Loss at the Auckland Internet Access Path. In *Passive and Active Network Measurement: 5th International Workshop*, 2002, s. 205–214.

[52] Mukaddam, A. – Elhajj, I. H.: Hop count variability. In *2011 International Conference for Internet Technology and Secured Transactions*, Prosinec 2011, s. 240–244.

[53] Nagireddi, S.: *VoIP Voice and Fax Signal Processing*. Wiley, 2008, s. 592, ISBN: 978-0-470-37786-4. URL http://201-shi.yolasite.com/resources/VOIP%20Voice%20and%20Fax%20Signal%20Processing.pdf

[54] Neustar: *Custom GeoPoint – Accurate IP Geolocation*, [online], c2017. URL https://www.neustar.biz/risk/compliance-solutions/ip-intelligence/custom-geopoint [cit. 8. 6. 2017]

[55] Neustar: *IP Intelligence, Geolocation, and Compliance Solutions*, [online], c2017. URL https://www.neustar.biz/risk/compliance-solutions/ip-intelligence [cit. 8. 6. 2017]

[56] Padmanabhan, V. N. – Subramanian, L.: An investigation of geographic mapping techniques for internet hosts. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, San Diego, California, USA, ročník 31, č. 4, Srpen 2001: s. 173–185, ISSN 0146-4833. URL http://doi.acm.org/10.1145/964723.383073

[57] Parekh, S. – Friedman, R. – Tibrewala, N. et al.: *Systems and methods for determining collecting and using geographic locations of internet users*. Digital Envoy, inc., Červen 2004, US Patent 6,757,740. URL https://www.google.com/patents/US6757740

[58] PlanetLab: *An open platform for developing, deploying, and accessing planetary-scale services*, [online], c2017. URL http://www.planet-lab.org [cit. 18. 4. 2017]

[59] Poese, I. – Uhlig, S. – Kaafar, M. A. et al.: IP geolocation databases: unreliable?. In *ACM SIGCOMM Computer Communication Review*, ročník 41, č. 2, Duben 2011: s. 53–56, ISSN 0146-4833, doi:10.1145/1971162.1971171. URL http://doi.acm.org/10.1145/1971162.1971171

[60] Popescu, A.: *Geolocation API Specification*, [online] Červenec 2014. URL https://dev.w3.org/geo/api/spec-source.html [cit. 5. 2. 2017]

[61] Post, C. C. – Woodrow, S.: Location is Everything: Balancing Innovation, Convenience, and Privacy in Location-based Technologies. In *Ethics and Law on the Electronic Frontier*, číslo 6.805/STS. 487, Prosinec 2008: s. 61.

[62] Pužmanová, R.: *Moderní komunikační sítě od A do Z*. 2. vydání, Brno, Computer Press, 2006, s. 432 ISBN 978-8-025-11278-6.

[63] Rekhter, Y. – Moskowitz, B. – Karrenberg, D. et al.: *RFC 1918: Address Allocation for Private Internets*. Leden 1999. URL `https://tools.ietf.org/html/rfc1918`

[64] Sarikaya, B.: Geographic Location in the Internet. *The Springer International Series in Engineering and Computer Science*, ročník Vol. 691. Kluwer Academic Publishers, Červen 2002, s. 214, ISBN 978-1-4020-7097-6.

[65] Satrapa, P.: *Internetový protokol verze 6*. 3. vydání, CZ.NIC, 2011, s. 409, ISBN 978-80-904248-4-5.

[66] Shavitt, Y. – Zilberman, N.: A geolocation databases study. In *IEEE Journal on Selected Areas in Communications*, ročník 29, č. 10, 2011: s. 2044–2056, ISSN: 0733-8716. URL `https://www.cl.cam.ac.uk/~nz247/publications/JSAC2011-Geolocation.pdf`

[67] software77.net: *IP to Country Database (IPV4 and IPV6)*, [online], c2017. URL `http://software77.net/geo-ip/` [cit. 16. 7. 2017]

[68] Spotter: *Geolocation service*, [online],
URL `http://spotter.etomic.org/` [cit. 28. 1. 2017]

[69] Talhofer, V.: *Základy matematické kartografie*. Univerzita Obrany, Brno, 2007, s. 167, ISBN 978-80-7231-297-9.

[70] University of Washington: *iPlane: Datasets*, University of Washington. URL `http://iplane.cs.washington.edu/data/data.html` [cit. 11. 8. 2013]

[71] Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. In *Survey review*, ročník 23, č. 176, 1975: s. 88–93.

[72] Wei, L. – Ren, G. – Shi, L. et al.: How does the recursive undns algorithm affect the accuracy of an IP geolocation system? In *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Červenec 2013, s. 1060–1064, ISBN 978-1-4673-5253-6, doi:10.1109/FSKD.2013.6816353.

[73] Wong, B. – Stoyanov, I. – Sirer, E. G.: Octant: a comprehensive framework for the geolocalization of internet hosts. In *Proceedings of the 4th USENIX conference on Networked systems design & implementation*, NSDI'07, Berkeley, CA, USA: USENIX Association,

2007, s. 23–23. URL https://www.usenix.org/conference/nsdi-07/octant-comprehensive-framework-geolocalization-internet-hosts

[74] Youn, I. – Mark, B. – Richards, D.: Statistical Geolocation of Internet Hosts. In *Computer Communications and Networks, 2009. ICCCN 2009. Proceedings of 18th Internatonal Conference on*, Srpen 2009, ISSN 1095-2055, s. 1–6, doi: 10.1109/ICCCN.2009.5235373.

[75] Zandbergen, P. A.: Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. In *Transactions in GIS*, ročník 13, číslo s1, Červenec 2009, s. 5–25, doi:10.1111/j.1467-9671.2009.01152.x.

[76] Zander, S.: *On the Accuracy of IP Geolocation Based on IP Allocation Data – Technical Report*. Technická zpráva, Centre for Advanced Internet Architectures (CAIDA), 2012.

PUBLIKACE AUTORA

[77] Balej, J.: Improvement of Geolocation Methods. In *Proceedings of the 17th conference Student EEICT 2011*, 2011, s. 16–20. URL http://www.feec.vutbr.cz/EEICT/2011/sbornik/03-Doktorske%20projekty/01-Elektronika%20a%20komunikace/01-xbalej02.pdf

[78] Balej, J.: Srovnaní přesnosti aktivních geolokačních technik, [online], *Access server*, 2012, ISSN 1214-9675. URL http://access.feld.cvut.cz/view.php?cisloclanku=2012070001

[79] Balej, J. – Komínek, O. – Rajnoha, M.: Geographic Distance Estimation for IP Geolocation. In *Proceedings in Electronic International, Interdisciplinary Conference EIIC 2013*, ročník 2, číslo 1, 2013, s. 584-588, ISBN 978-80-554-0762-3.

[80] Balej, J. – Komosný, D.: Zdroje zpoždění při komunikaci v Internetu. *Elektrorevue - Internetový časopis (http://www.elektrorevue.cz)*, ročník 2010/42, s. 1–7, ISSN: 1213-1539. URL http://elektrorevue.cz/cz/clanky/komunikacni-technologie/0/zdroje-zpozdeni-pri-komunikaci-v-internetu/

[81] Balej, J. – Komosný, D. – Kathiravelu, G.: Mapping round-trip time into length suitable for geolocation. In *ICT2011 proceeding*, Únor 2011, s. 63–68, ISBN: 978-80-214-4231-3.

[82] Balej, J. – Komosný, D. – Zach, P.: How can measurement of delay be helpful for geolocation databases? In *Enterprise and Competitive Environment: Conference Proceedings*, 2017, s. 24–32, ISBN 978-80-7509-499-5. URL https://ece.pefka.mendelu.cz/sites/default/files/imce/ECE2017_fin.pdf

[83] Komosný, D. – Balej, J. – Sathu, H. et al.: Impact of Intermediate Device Latency on IP Geolocalisation. In *Proceedings of the 2011 International Conference on Telecommunication Systems, Modeling and Analysis*, Praha, 2011, s. 39–48, ISBN 978-0-9820958-4-3.

[84] Komosný, D. – Balej, J. – Sathu, H. – aj.: Cable length based geolocalisation. *Przeglad Elektrotechniczny*, ročník 88, č. 7 A, Červenec 2012: s. 26–32. (IF=0,244)

SEZNAM ZKRATEK

API	Application programming interface
AS	Autonomous System
BGP	Border Gateway Protocol
BSSID	Basic Service Set Identifier
BTS	Base Transceiver Station
CAIDA	Center for Applied Internet Data Analysis
CBG	Constraint Based Geolocation
CESNET2	Czech Scientific and Education NETwork 2
CID	Cell ID
CLBG	Cable Length Based Geolocalisation
DNS	Domain Name System
DWDM	Dense Wavelength Division Multiplexing
FIFO	First In First Out
GPS	Global Positioning System
GSM	Global System for Mobile communications
HTML	HyperText Markup Language
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protokol
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protokol
IPv4	Internet Protokol verze 4
IPv6	Internet Protokol verze 6
ISP	Internet Service Provider
L3	Layer 3 RM ISO/OSI
MAC	Media Access Control
MPLS	Multiprotocol Label Switching
NAT	Network Address Translation
OSI	Open System for Interconnection
PING	Packet InterNet Groper
PSČ	Poštovní směrovací číslo
QoS	Quality of Service
RFC	Request for Comments
RIPE NCC	Reseaux IP Europeens Network Coordination Centre
RIR	Regional Internet Registry
RM ISO/OSI	Reference Model International Organization for Standardization/Open Systems Interconnection
RTT	Round Time Trip

SBG	Spring Based Geolocation
SG	Statistical Geolocation
SOI	Speed of Internet
SSH	Secure Shell
SSID	Service Set Identifier
TBG	Topology Based Geolocation
Tor	The Onion Routing
TTL	Time To Live
VIP	Very Important Person
VoIP	Voice over Internet Protocol
VPN	Virtual Private Network
WGS 84	World Geodetic System 1984
WiFi	Wireless Fidelity

SEZNAM SYMBOLŮ A VELIČIN

A	pomocná proměnná pro výpočet Vincentyho rovnic
a	délka hlavní poloosy elipsoidu (pro WGS84 $a = 6\,378\,137$ m)
α	úhel spojnice mezi dvěma body svíraný s rovníkem
α_i	úhel spojnice z bodu i ke druhému bodu na zemském povrchu
B	pomocná proměnná pro výpočet Vincentyho rovnic
b	délka vedlejší poloosy elipsoidu (pro WGS84 $b = 6\,356\,752,3142$ m)
b_i	absolutní člen přímky Bestline pro landmark i (metoda CBG)
C	pomocná proměnná pro výpočet Vincentyho rovnic
c	rychlost světla c = 299 792 458 m/s
DV	distance vektor zpoždění referenčního bodu (metoda GeoPing)
DV′	distance vektor zpoždění lokalizované stanice (metoda GeoPing)
F	velikost rámce (datové jednotky)
$f_{t_{\text{rtt}}}(l)$	funkce hustoty pravděpodobnosti normálního rozdělení (metoda Spotter)
ϕ_i	zeměpisná šířka bodu i
(ϕ_i, L_i)	zeměpisné souřadnice bodu i
k_α	hustota bodů na hranici pro určení mezních hranic okolo landmarku
k_{nv}	parametr vlivu nepřímého vedení přenosových médií
L	rozdíl v zeměpisné délce dvou bodů (výpočet pomocí Vincentyho rovnic)
L_1	zeměpisná délka bodu i
l	délka přenosového média
l_{celk}	vypočtená celková délka přenosových médií
$l_{i,j}$	vzdálenost mezi body i a j
l_{prima}	přímá geografická vzdálenost mezi stanicemi
l_{skut}	skutečná délka využitých přenosových médií mezi stanicemi
λ	rozdíl v zeměpisné délce na pomocné kouli (Vincentyho rovnice)
λ_{p}	počet přicházejících paketů za sekundu
$loc(i)$	popis zeměpisné polohy bodu i
M	počet aktivních měřících bodů
m_i	kvocient přímky Bestline pro landmark i (metoda CBG)
μ	maximální množství obsloužených paketů frontou za sekundu
$\mu(t_{\text{rtt}})$	střední hodnota normálního rozdělení pro obousměrné zpoždění
N	počet mezilehlých L3 zařízení
R	přenosová rychlost
$\boldsymbol{R}$	množina bodů popisujících region průniku mezních hranic
$R_L(t_{\text{rtt}}$	popis horní části konvexní obálky metody Spotter
r	reciprocká hodnota zploštění elipsoidu (pro WGS84=$298,257\,223\,6$)
$r_L(t_{\text{rtt}}$	popis dolní části konvexní obálky metody Spotter

r_{nv}	parametr rychlosti přenosových linek a nepřímého vedení kabelů
ρ	procentuální zatížení fronty
S	obsah regionu vymezujícího polohu hledané stanice
σ	úhlová vzdálenost mezi body
σ_m	úhlová vzdálenost mezi rovníkem a středem spojnice bodů
$\sigma^2(t_{\mathrm{rtt}})$	rozptyl normálního rozdělení pravděpodobnosti pro obousměrné zpoždění
t_{KZ}	celkové zpoždění způsobené koncovým zařízením
t_{KZ2}	zpoždění obou koncových zařízení dohromady
t_{MZ}	zpoždění způsobené mezilehlým zařízením
t_{PL}	zpoždění vzniklé na přenosových linkách
t	jednosměrné zpoždění komunikačního řetězce
t_{d}	deterministické zpoždění
t_{des}	deserializační zpoždění (zpoždění ve vstupní frontě zařízení)
t_{f}	doba strávená v odchozích frontách zařízení
t_i	zpoždění mezi i-tou sondou a referenčním bodem (metoda GeoPing)
t'_i	zpoždění mezi i-tou sondou a lokalizovanou stanicí (metoda GeoPing)
$t_{i,j}$	obousměrné zpoždění mezi body i a j
t_{odp}	doba potřebná k vygenerování odpovědi cílovou stanicí
t_{pak}	paketizační zpoždění – doba potřebná k vytvoření paketu
t_{rs}	zpoždění způsobné dobou šíření signálu médiem
t_{rtt}	velikost obousměrného zpoždění
t_{s}	stochastické zpoždění
t_{ser}	serializační zpoždění
t_{z}	zpoždění způsobené dobou zpracování informace mezilehlým zařízením
t_{zd}	deterministická část doby zpracování informace u mezilehlého zařízení
t_{zs}	stochastická část doby zpracování informace u mezilehlého zařízení
U	snížená zeměpisná šířka prvního bodu (dle Vincentyho rovnic)
v_{rs}	rychlost šíření signálu přenosovým médiem

Curriculum Vitæ

Jiří Balej

Osobní informace

Datum narození:	28. července, 1986
Místo narození:	Brno
Adresa:	Vlnitá 4, 627 00 Brno
E-mail:	jirka.balej@gmail.com

Vzdělání

2010–2017	Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno, obor: Teleinformatika, titul: Ph.D. (předpokládané ukončení: 2017)
2008–2010	Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno, obor: Telekomunikační a informační technika, titul: Ing.
2005–2008	Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Technická 3058/10, 616 00 Brno, obor: Teleinformatika, titul: Bc.
2001–2005	Střední průmyslová škola elektrotechnická v Brně, Kounicova 684/16, 602 00 Brno, obor: Slaboproudá elektrotechnika.

Zahraniční stáže

2017 (únor–červen)	University of Patras (Řecko), Department of Electrical & Computer Engineering.

Odborné certifikace

2016	CCNA Security – Cisco Certified Network Associate Security
2015	MTCWE – MikroTik Certified Wireless Engineer
2014	JNCIA-Junos – Juniper Network Certified Associate JunosOS
2013	MTCTCE – MikroTik Certified Traffic Control Engineer
2013	MTCNA – MikroTik Certified Network Associate
2013	CCNA R&S – Cisco Certified Network Associate Routing & Switching

Zaměstnání

2014–	*Projektant a analytik informačních systémů*, Mendelova univerzita v Brně, Ústav informačních technologií, Zemědělská 1655/1, 613 00 Brno.
2013–2014	*Technicko hospodářský pracovník*, Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, Technická 3082/12, 616 00 Brno.
2012–	*Akademický pracovník – asistent*, Mendelova univerzita v Brně, Provozně ekonomická fakulta, Ústav informatiky, Zemědělská 1655/1, 613 00 Brno.

Publikace

- Publikace v časopisech s impaktním faktorem: 1 + 1 (*odesláno k publikování*)
- Publikace v časopisech indexovaných databází Scopus: 4
- Publikace v jiném odborném periodiku: 2
- Publikace v konferenčních sbornících: 7
- Konferenční příspěvky indexované databází WoS: 2 + 1 (*odeslán k indexaci*)