VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE VOIP APLIKACÍ VE STATISTIKÁCH SÍŤOVÉHO PROVOZU

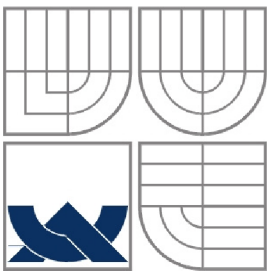
BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

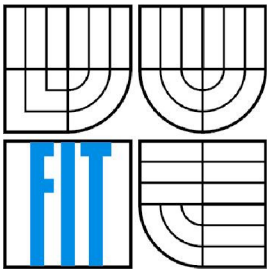
AUTOR PRÁCE ONDŘEJ HAVELKA

AUTHOR

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE VOIP APLIKACÍ VE STATISTIKÁCH SÍŤOVÉHO PROVOZU

DETECTION OF VOIP APPLICATIONS IN NETWORK STATISTICS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — ONDŘEJ HAVELKA
AUTHOR

VEDOUCÍ PRÁCE — ING. MARTIN ŽÁDNÍK
SUPERVISOR

BRNO 2009

Abstrakt

VoIP aplikace umožňují rychlou a levnou komunikaci prostřednictvím internetu. Díky masivnímu rozšíření těchto aplikací v posledních několika letech dochází k nárůstu VoIP provozu, který je třeba umět rozlišit od ostatního provozu. Detekce VoIP aplikací také souvisí s potřebou monitorování dat a klasifikace síťového provozu. Práce se zaměřuje na detekci těchto aplikací: GTalk, ICQ, Skype, WLM a Yahoo. Jsou zde představeny dvě detekční techniky. První je založena na chování VoIP aplikací. Tato technika využívá statistické vlastnosti toků, jako jsou četnost paketů a jejich velikost. Druhá metoda je založená na hledání vzorů v obsahu paketů, kterými jsou přenášena zvuková data pomocí protokolu RTP. Funkčnost metod byla experimentálně ověřena a výsledky jsou v práci prezentovány.

Abstract

VoIP applications allow us fast and cheap communication over the internet. Due to the massive expansion of these applications in last few years, there is growth of VoIP traffic, which is needed to distinguish from other traffic. Detection of VoIP applications is related to data monitoring and network traffic classification. The work focuses on detection of these applications: GTalk, ICQ, Skype, WLM a Yahoo. There are presented two detection techniques. First of them is based on behavior of VoIP applications. This technique uses the statistical characteristics of network flows as packet frequency and packet size. The second method is based on searching for patterns in packet contents. These packets carry voice data using protocol RTP. Both techniques were experimentally verified and the results are presented in the work.

Klíčová slova

Voice over IP, VoIP, Internetová telefonie, detekce hlasového provozu

Keywords

Voice over IP, VoIP, Internet Telephony, voice traffic detection

Citace

Havelka Ondřej: Detekce VoIP aplikací ve statistikách síťového provozu, bakalářská práce, Brno, FIT VUT v Brně, 2009

Detekce VoIP aplikací ve statistikách síťového provozu

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Martina Žádníka. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....

Ondřej Havelka

20. května 2009

Poděkování

Tímto bych chtěl poděkovat vedoucímu své bakalářské práce, Ing. Martinu Žádníkovi, za pomoc a veškerý čas, který mi ochotně a trpělivě věnoval. Také bych chtěl poděkovat Petře Zajíčkové za pomoc při vytváření testovacích datových sad.

Obsah

Úvod

V poslední době se na internetu objevují nové služby, jako jsou telekonference, zvuk na dálku nebo videokonference. Stále častěji lidé přecházejí od tradičních telekomunikačních řešení k novým způsobům komunikace, která probíhá prostřednictvím datových sítí. Hovory, obzvláště dálkové hovory, jsou velmi drahé, pokud jsou přenášeny klasickou telefonní sítí. Trendem dnešní doby je využít síť internet k přenosu hovorů.

Voice over Internet Protocol, dále jen VoIP, je technologie umožňující přenos digitalizovaného hlasu prostřednictvím IP sítí, jako jsou internet, intranet a LAN. VoIP je relativně mladá technologie na rozdíl od klasických pevných telefonních sítí, které také slouží k přenosu hlasu. Technologie VoIP nám umožňuje integrovat hlas do IP sítí a přenášet jej bez potřeby telefonních rozvodů. Neexistuje jednotný standard, který by určoval způsob přenosu hlasu. Proto existuje několik protokolů, uzavřených či otevřených, a každý z nich implementuje přenos hlasu jinak. Protože VoIP využívá k přenosu IP sítě, nelze zaručit spolehlivost a stabilitu. Telefonní sítě, na rozdíl od IP sítí, jsou spolehlivé, stabilní a založené na mezinárodních telekomunikačních standardech.

Existují dvě hlavní příčiny rozmachu VoIP ve firemní sféře - nízká cena hovorů a jedna infrastruktura pro data i hlas. Tyto příčiny mají za následek migraci společností od stávajících hlasových řešení k VoIP. V domácím prostředí nedochází k takovému rozmachu, protože hodně domácích uživatelů má internet spojený s pevnou linkou nebo případně pro ně VoIP hovory nejsou takovým cenovým lákadlem. Svět je v dnešní době propojen pomocí internetu a můžeme tedy předpokládat masivní rozšíření VoIP technologie v budoucnosti.

VoIP telefonie přináší spoustu výhod, mezi něž patří:

Výhody VoIP	
Nízká cena	Komunikace PC-PC je většinou zdarma. Volání na pevné linky po celém světě je ve srovnání s cenou telefonních operátorů levnější (hlavně do vzdálenějších destinací).
Dostupnost	Pro uskutečnění hovoru stačí pouze připojení k internetu. Lidé, kteří jsou na cestách nebo se často stěhují, mohou ocenit přenosnost. Opět stačí pouze připojení k internetu a po přihlášení ke službě, lze telefonovat za stále stejné ceny po celém světě
Přidané služby	Konferenční hovory, přesměrování, ID volajícího jsou služby, které jsou dostupné zdarma, a není třeba za ně platit navíc, jako u telefonních operátorů (pokud tyto služby vůbec poskytují).
Hardware a software	K uskutečnění hovoru nepotřebujeme klasický telefonní přístroj. Vystačíme si pouze s programem nainstalovaným na našem počítači, mikrofonem a reproduktory. Pokud dáme přednost telefonnímu přístroji, je možné jej připojit přes převodník do sítě a uskutečňovat s ním hovory.

Tabulka 0.1 Výhody VoIP technologie

Nevýhody VoIP technologie:	
Nedostupnost v případě výpadku elektrické energie	Klasický telefon je napájen z rozvodů, takže výpadek energie se ho nedotkne a lze s ním volat. V případě VoIP telefonů nelze při výpadku elektřiny telefonovat, a pokud nemáme k dispozici záložní generátory elektrické energie, nefunguje ani internet ani VoIP telefony.
Nelze využívat k tísňovým voláním	Při telefonování na tísňové linky dochází k lokalizaci účastníka a přepojení na nejbližší operační centrum záchranného systému. VoIP technologie umožňuje z principu telefonovat odkudkoliv ze stále stejného čísla a může být těžké zjistit přesnou geografickou polohu účastníka.
Spolehlivost a kvalita	Jako přenosové médium jsou použity IP sítě a s tím je spojena jejich nespolehlivost a ztrátovost paketů během cesty. V případě přetížení sítě jsou VoIP pakety stejně jako ostatní pakety zahazovány, což má za následek špatnou kvalitu hovorů a výpadku zvuku.
Bezpečnost	Většina datových přenosů není šifrována, a to platí i pro VoIP přenosy. Internet je veřejná síť a nelze tedy zaručit, že naše data nikdo neodchytává a neodposlouchává.

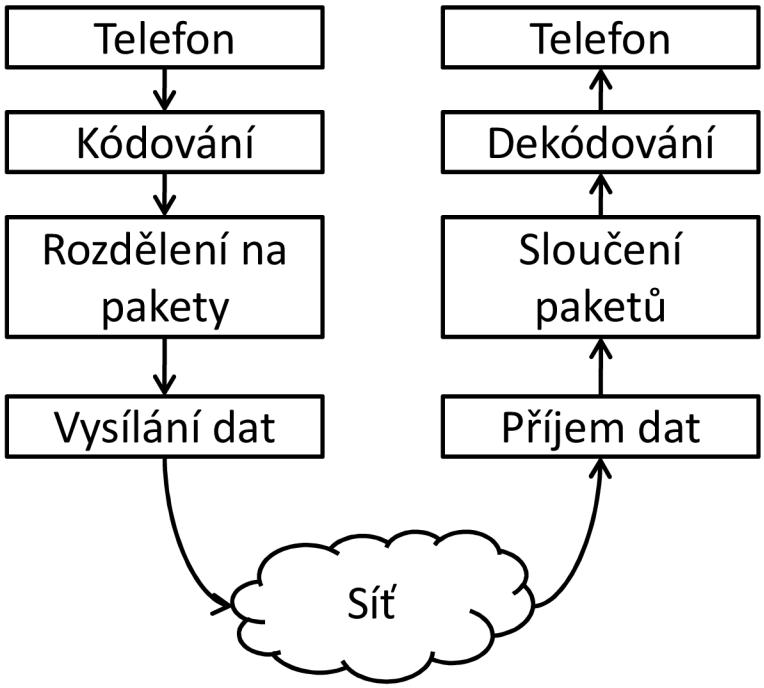
Tabulka 0.2 Nevýhody VoIP technologie

Cílem této práce je snaha o zlepšení možnosti detekce VoIP hovorů v síti, tak aby bylo zachyceno co nejvíce opravdových hovorů, a zároveň aby toky, které nejsou VoIP hovory, nebyly označeny jako VoIP. V práci jsou prezentovány dva algoritmy pro detekci VoIP aplikací, které umožňují rozlišit VoIP provoz od ostatního provozu na síti.

Zbývající část práce je organizována následovně. První kapitola popisuje VoIP technologii a její princip. Kapitola 2 obsahuje informace o charakteristikách VoIP provozu a navrhované algoritmy pro detekci tohoto provozu. Třetí kapitola stručně popisuje program pro detekci a vyhodnocování VoIP provozu. Kapitola 4 obsahuje výsledky testování. Závěr práce obsahuje shrnutí a budoucí možný vývoj práce.

1 Technologie VoIP

Technologie VoIP umožňuje přenos hlasu pomocí IP sítí. Hlasový záznam je třeba rozdělit na části a tyto části odeslat příjemci. Na následujícím obrázku je zobrazen princip VoIP hovoru.



Obrázek 1.1 Princip VoIP přenosu

Zvuk z mikrofonu je navzorkován a převeden do digitální podoby. Poté je digitalizovaný zvuk zakódován a zkomprimován pomocí kodeku. Kódovaná a komprimovaná data jsou rozdělena do paketů, které obsahují informace umožňující rekonstrukci zvukových dat u příjemce. Pakety jsou poslány přes obecnou IP sít. Příjemce přijímá jednotlivé pakety, seřadí je do posloupnosti podle informací v hlavičkách paketů a poté z nich vyjme data. Data jsou na základě informace o typu kodeku dekódována a z nich je rekonstruován původní analogový signál. Tato podkapitola byla převzata z [1].

1.1 Protokoly využívané VoIP relacemi

Síťová vrstva - Protokol IP

Protokol IP (Internet Protocol) je zodpovědný za doručování paketů mezi dvěma počítači. IP pracuje na síťové vrstvě modelu TCP/IP. IP je nespojovaný protokol, to znamená, že není třeba nejdříve vytvořit cestu před samotným přenosem dat. Protokol IP nezaručuje spolehlivost, neumožňuje kontrolu toku ani detekci chyb, proto pakety mohou přijít s chybami, mimo pořadí nebo nemusejí přijít vůbec. Tyto problémy řeší protokoly vyšších vrstev.

VoIP aplikace využívají k přenosu protokol IP. Samotný protokol není pro relace probíhající v reálném čase nejvhodnější, protože je nespolehlivý a nezaručuje včasné doručení dat. O tyto problémy se starají protokoly vyšších vrstev.

Transportní vrstva - protokoly TCP, UDP

VoIP má velmi vysoké požadavky na přenos v reálném čase, a proto jakékoliv kolísání přenosu nebo zpoždění se projeví na kvalitě hovoru, který se může stát nesrozumitelným. TCP protokol garantuje spolehlivé doručení ve správném pořadí, ale za cenu zpoždění jednotlivých paketů, což pro VoIP aplikaci je nepřípustné. TCP je spolehlivá služba a při chybě součtu nebo ztrátě paketu, je daný paket poslán znovu. Opakovaný přenos chybných nebo ztracených paketů je zdrojem zpoždění a kolísání toku. Proto je protokol TCP ve VoIP aplikacích využíván jen pro vytváření spojení a pro kontrolu spojení. Zmíněné vlastnosti TCP jej dělají nevhodným pro přenos hovoru.

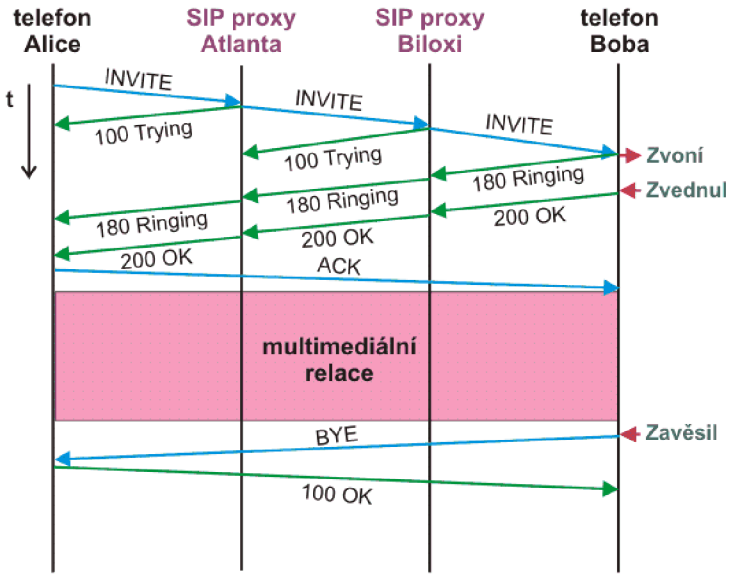
Hlasová komunikace nevyžaduje perfektní spolehlivost, a proto ztráta paketu se projeví pouze jako malý výpadek zvuku. VoIP aplikace využívají pro přenos hovoru protokol UDP, který nezaručuje spolehlivé doručení ani doručení ve správném pořadí. Veškerou kontrolu toku UDP paketů musí provádět samotná aplikace. UDP neumožňuje rozdělit zprávu a na druhém konci ji složit, stejně jako neumožňuje řazení paketů do správné posloupnosti, ale předává data v pořadí, ve kterém přišla. To znamená, že aplikační program využívající UDP musí být schopen zjistit, zda celá zpráva dorazila, jestli je ve správném pořadí a bez výpadků.

Aplikační vrstva

SIP

Session Initiation Protocol (SIP) je signalizační protokol využívaný pro sestavování a ukončování multimediálních relací, jako jsou hlasové hovory, videohovory nebo multimediální konference přes internet. Protokol je možné využít pro vytvoření, změnu a ukončení relace, skládající se z jednoho nebo několika multimediálních toků.

SIP protokol je protokol aplikační vrstvy TCP/IP modelu. SIP je nezávislý na typu transportní vrstvy a může běžet nad TCP nebo UDP na portu 5060. Protokol SIP je textově založený, podobně jako protokol HTTP.



Obrázek 1.2 Ukázka SIP transakce, (dostupný z WWW: http://upload.wikimedia.org/wikipedia/cs/4/4c/SIP.png)

RTP

Real-time Transport Protocol (RTP) [2] je protokol pro přenos audiovizuálních dat prostřednictvím sítě internet. RTP je nejvíce využívaným protokolem použitým v IP telefonii. RTP negarantuje spolehlivé doručení, ale číslování paketů umožňuje příjemci detekovat výpadky paketů. Kontrolu přenosu dat provádí RTP Control Protocol (RTCP). Jeho hlavním účelem je poskytnout informaci o kvalitě hovoru periodickým posíláním statistických informací účastníkům multimediální relace. RTP protokol se nestará o sestavování hovoru, ale pouze přenáší multimediální obsah.

1.2 Důvody detekce VoIP

Existuje mnoho důvodů proč detekovat VoIP provoz. Jedním z nich je potřeba internetového poskytovatele vědět, co přenáší po svých linkách, aby mohl podle toho přizpůsobit svou síť. Dalším argumentem pro detekci VoIP hovorů může být snaha o jejich zkvalitnění a jejich prioritizaci před ostatními daty, což pro poskytovatele může být další zdroj příjmů. Pokud mají zákazníci platit za VoIP služby, tak chtějí, aby tyto služby byly spolehlivé, kvalitní a dostupné.

Ve firemní sféře je důvodem detekce VoIP aplikací snaha zaměstnavatelů, aby jejich zaměstnanci netelefonovali v pracovní době. Hovory pomocí klasických analogových telefonů směřující přes firemní ústřednu ven, je možné zakázat, případně kontrolovat a tím omezit nepracovní hovory. V případě VoIP hovorů nelze spolehlivě rozlišit mezi VoIP hovory a ostatními daty, a proto nelze zakázat určitému PC veškerý datový provoz mimo firemní síť. Pokud budeme schopni rozlišit VoIP hovory od ostatních dat, bude možné filtrovat a případně blokovat jen VoIP provoz.

Dalším důvodem může být blokace VoIP přenosu. S rozmachem VoIP hovorů přicházejí telekomunikační operátoři o své zisky, protože firmy přecházející na VoIP telefonii jim již neplatí za zprostředkování hovorů. Pokud dokážeme detekovat tyto VoIP hovory, můžeme je blokovat a zájemcům je za poplatek povolit.

Strach ze zneužití VoIP hovorů ke kriminalitě, případně terorismu je další příčinou, proč tyto hovory detekovat. Klasické hovory, jak přes pevnou, tak i mobilní síť, mohou být monitorovány a může být včas odhalen pokus o spáchání trestného činu nebo teroristického útoku. V případně VoIP hovorů, pokud nevíme, že daný tok dat je hovor, tak ho nemůžeme monitorovat a případně sledovat. Také u klasických hovorů víme přesně, kde se obě komunikující osoby nacházejí, což nemusí platit v případě VoIP.

1.3 Související práce

Potřeba detekce VoIP v síťovém provozu se objevila s masovým rozšířením VoIP v posledních čtyřech letech. Z toho vyplývá, že techniky pro detekci tohoto provozu jsou ve svém počátku a trpí nedostatečnou spolehlivostí, přestože je jim věnována pozornost v mnoha vědeckých pracích.

Klasická metoda založená na analýze portů je nedostatečná, protože vychází z předpokladu, že každá služba využívá svůj definovaný port. Pokud se na daném portu objeví nějaký provoz,

tak se předpokládá, že je to protokol využívaný danou službou. Bohužel některé VoIP protokoly nepoužívají určité definované porty, ale spíše dynamicky zvolené, aby předešly detekci. Navíc některé aplikace, jako např. Skype, mohou využívat pro komunikaci port 80, pokud se nemohou dostat přes firewall. Metoda analýzy portů může být v některých situacích stále dostačující, ale v případě dále prezentovaných aplikací se stává nepoužitelnou, protože tyto aplikace nepoužívají pevně stanovené porty a použití této metody dává velké procento falešně-pozitivních výsledků. Jak zmiňuje J.P.Curtis [3]: „Na rozdíl od ostatních typů provozu, VoIP nemůže být jednoduše identifikován pomocí IP adres nebo použitím portů. Protože měření byla prováděna na skutečných sítích, tak bezpečnostní nařízení umožňovaly přístup pouze k hlavičkám paketů, ne k obsahu paketů. Proto není identifikace VoIP provozu triviální úkol.“

Curtis, Cleary, McGregor a Pearson [3] se zaměřili na porovnání výkonu UDP a TCP protokolu. Dochází k nárůstu VoIP provozu, který využívá UDP protokol, protože má lepší odezvu v real-time aplikacích a menší režii. Aplikace založené na UDP neimplementují kontrolu zahlcení. Autoři si pokládají otázku, jestli růst VoIP, resp. UDP provozu negativně neovlivní aplikace využívající protokol TCP. K výzkumu využívají simulaci prostředí. TCP provoz je http a UDP je protokol H.323. Úvodní hypotéza se potvrdila. UDP provoz má nečekaný vliv na TCP data v situacích, kdy dojde k zahlcení sítě. V případě kde je přítomno větší množství UDP provozu, tak UDP agresivně vytlačí TCP provoz. Pokud ovšem je přítomen jen TCP provoz, tak se pásmo rozdělí rovnoměrně.

Idrees a Khan [4] se zabývají obecnou charakteristikou hlasového provozu a jeho detekce. Prezentují algoritmus, který bude schopen efektivně identifikovat přítomnost hlasové komunikace na síti, bez ohledu na použitý protokol. Tento algoritmus je založený na charakteristikách toku, jako jsou četnost paketů a jejich velikost.

Bonfiglio, Mellia, Meo, Rossi a Tofanelli [5] se zaměřují na identifikaci hovorů pomocí aplikace Skype. Představili dvě techniky pro detekci aplikace Skype v síťovém provozu. Chi-Square Classifier(CSC) – technika založená na Pearsonově Chi-Square testu detekuje otisk Skype ze struktury paketů. CSC odhaduje, zda je zpráva šifrovaná, případně která část. Cílem je otestovat, jestli zpráva splňuje jeden z formátu zpráv aplikace Skype. Další technika, Naive Bayesian Classifiers (NBC), je založená na stochastických charakteristikách aplikace Skype, jako jsou četnost příchodu paketů a jejich velikost. K ověření výsledků vyvinuli autoři Payload Based Classifier (PBC), který je založený na tradiční technice inspekce paketů zkombinované s analýzou chování koncových stanic, což umožňuje spolehlivě detekovat Skype a jím tvořený síťový provoz. Výše zmíněné dvě techniky byly testovány na dvou datových sadách: Campus – 95hodinový provoz zachycený na univerzitní lince. ISP – jednodenní provoz italského poskytovatele internetu. Výsledky jejich práce jsou v následujících třech tabulkách.

		N	OK	FP	FP%	FN	FN%
PBC	E2E	1014	—	—	—	—	—
	E2O	163					
NBC	E2E	1236	726	510	0.68	288	28.40
	E2O	441	153	288	0.38	10	41426
CSC	E2E	2781	984	1797	2.40	30	2.96
	E2O	161	157	4	0.01	6	3.68
NBC ∧	E2E	716	710	6	0.01	304	29.98
CSC	E2O	147	147	0	0.00	16	9.82
TOT	≥ 100	76025	—	—	—	—	—
		487729					

Tabulka 1.1 Výsledky UDP toků, Data Campus

		N	OK	FP	FP%	FN	FN%
PBC	E2E	65	—	—	—	—	—
	E2O	125					
NBC	E2E	27437	50	27387	73.73	15	23.08
	E2O	295	124	171	0.46	1	0.80
CSC	E2E	191	57	134	0.36	8	12.31
	E2O	190	123	67	0.18	2	1.6
NBC ∧	E2E	51	49	2	0.01	16	24.62
CSC	E2O	163	122	41	0.11	3	2.40
TOT	≥ 100	37212	—	—	—	—	—
		258634					

Tabulka 1.2 Výsledky UDP toků, Data ISP

		CAMPUS	ISP
NBC	E2E	20910	60
	E2O	2034	646
CSC	E2E	403996	46876
	E2O		
NBC ∧	E2E	621	12
CSC	E2O	313	0
TOT	≥ 100	1646424	108831
		23856424	1614553

Tabulka 1.3 Výsledky UDP toků, obě množiny dat

Z výsledků vyplynuly tyto závěry: NBC je efektivní v identifikaci hlasového provozu nezávisle na aplikaci. CSC je účinný v identifikaci Skype provozu využívajícího protokol UDP. Použití technik NBC a CSC dohromady je velmi efektivní pro detekci hlasové komunikaci pomocí aplikace Skype, protože poskytují téměř nulové falešně-pozitivní výsledky.

Ve studii [6] autoři uvádí: „Pro poskytnutí spolehlivých VoIP služeb je nezbytné aplikovat řízení provozu a filtrování povoleného VoIP provozu od zakázaného.“ Jejich technika je založená na analyzování vzorů v paketech, chování toku a interakci. Přijaté pakety jsou analyzovány na základě vzorů. Pokud nesouhlasí s referenčními vzory, jsou tyto pakety označeny jako neznámé a jejich vzor je přidán do referenčních. Chování toku je závislé na použitém hlasovém kodeku a parametrech paketů (přenosová rychlost a četnost paketů). Jako interakce je označena výměna signalizačních zpráv pro sestavení spojení a výměna dat, využívajících sestavené spojení.

2 Návrh řešení

2.1 VoIP aplikace

Tato bakalářské práce se zaměřuje na detekci pěti VoIP aplikací:

- Google Talk (dále jen Gtalk) [7]
- ICQ [8]
- Skype [9]
- Windows Live Messenger (dále jen WLM) [10]
- Yahoo Messenger (dále jen Yahoo) [11]

Specifikace jednotlivých protokolů.

Všechny zkoumané aplikace podporují výměnu zpráv, posílání souborů, hlasovou komunikaci a některé umožňují i přenos videa. Následující specifikace jsou zaměřeny na sestavení a průběh hlasové komunikace v jednotlivých aplikacích.

Aplikace	Popis
Gtalk	Aplikace využívá k posílání zpráv otevřený protokol XMPP. VoIP komunikace je v Gtalku založena na protokolu Jingle, což je rozšíření Jabber/XMPP protokolu. Protokol Jingle umožňuje peer-to-peer signalizaci pro hlasovou a video komunikaci. Samotný multimediální obsah je přenášen pomocí RTP protokolu.
ICQ	Aplikace využívá pro signalizaci hovorů protokol SIP, který je přenášen v šifrovaném SSL kanálu. Hovor je přenášen pomocí protokolu RTP.
Skype	Protokol aplikace Skype je uzavřený, nejsou známy jeho specifikace a nevíme, jaké techniky používá pro ustavení hovoru, pro jeho kontrolu a samotný přenos hlasových dat.
WLM	WLM využívá ke komunikaci mezi klienty protokol MSNP (Microsoft Notification Protocol).WLM provádí signalizaci pomocí protokolu SIP, který je zakódován kódováním base64. Zakódovaný protokol SIP je pak přenášen jako parametr zprávy protokolu MSNP. Pro přenos hovoru WLM využívá protokol RTP
Yahoo	Yahoo využívá pro sestavování hovorů protokol SIP, který je přenášen šifrovaným SSL kanálem. Hovory jsou přenášeny protokolem RTP.

Tabulka 2.1 Specifikace zkoumaných VoIP aplikací

Tok

Základní jednotkou pro detekci VoIP hovorů je tok. Tok je skupina paketů, které mají společné tyto vlastnosti:

1. Zdrojová IP adresa
2. Cílová IP adresa
3. Zdrojový port
4. Cílový port

5. IP protokol
6. Rozhraní
7. IP Type of Service

Aplikace může při hovoru komunikovat s druhou stranou na několika portech různými protokoly, ale samotná zvuková data jsou přenášena jen jedním tokem. Pokud v průběhu hovoru dojde k výpadku a aplikace pokračuje v komunikaci na jiném portu, jedná se o nový tok, který nemá, z pohledu definice toku, nic společného s předchozím průběhem hovoru.

2.2 Statistické vlastnosti VoIP provozu

Všech pět zkoumaných aplikací bylo podrobeno analýze. S každou aplikací byl uskutečněn jeden pětiminutový hovor, který byl zachycen pomocí nástroje Wireshark [12]. Po prozkoumání nasbíraných dat, byly získány tyto statistické vlastnosti:

- Čas mezi prvním a posledním paketem
- Počet paketů
- Počet bajtů
- Průměrný počet paketů za sekundu
- Průměrná velikost paketů
- Minimální velikost paketů
- Maximální velikost paketů
- Průměrný počet bajtů za sekundu
- Průměrný počet megabitů za sekundu.

Následující tabulka zobrazuje výše zmíněné statistické vlastnosti jednotlivých VoIP aplikací.

	VOIP GTalk	**VOIP ICQ**	**VOIP WLM**	**VOIP Skype**	**VOIP yahoo**
Čas	293,254	308,834	308,402	302,765	305,767
Paketů	16398	19503	12988	18462	13104
Bajtů	2106419	2778879	1611240	2688504	1764643
Paketů/sek.	55.917	63.150	42.898	59.863	42.856
Prům.vel.paketů	128.456	142.485	124.056	145.624	134.664
Min.vel.paketu	67	62	45	55	54
Max.vel.paketu	276	304	516	251	313
Bajtů/sek.	7182.929	8997.962	5321.748	8717.533	5771.197
Mbit/sek.	0.057	0.072	0.043	0.070	0.046

Tabulka 2.2 Souhrn statistik VoIP aplikací

Pro porovnání VoIP provozu s ostatním internetovým provozem jsou v tabulce 2.3 zobrazeny ostatní internetové aplikace (web, stažení souboru, torrent, video z youtube, TFTP přenos a online hra).

	Online hra	TFTP	Torrent	Video	Web	Stažení souboru
Čas	276.5	276.5	302,473	135,297	304,770	305,551
Paketů	15184	39697	47529	1146	9024	29601
Bajtů	1894695	12265622	30169183	12463602	5550447	32206024
Paketů/sek.	54.915	130.055	157,135	84,599	29,609	96,878
Prům.vel.paketů	124.782	308.981	634,753	1088,905	615,076	1088,008
Min.vel.paketu	58	46	54	54	54	54
Max.vel.paketu	1442	558	1514	1514	1514	1514
Bajtů/sek.	6852.419	40184.523	99741,82	92120,22	18211,89	105403,25
Mbit/sek.	0.055	0.321	0,798	0,737	0,146	0,843

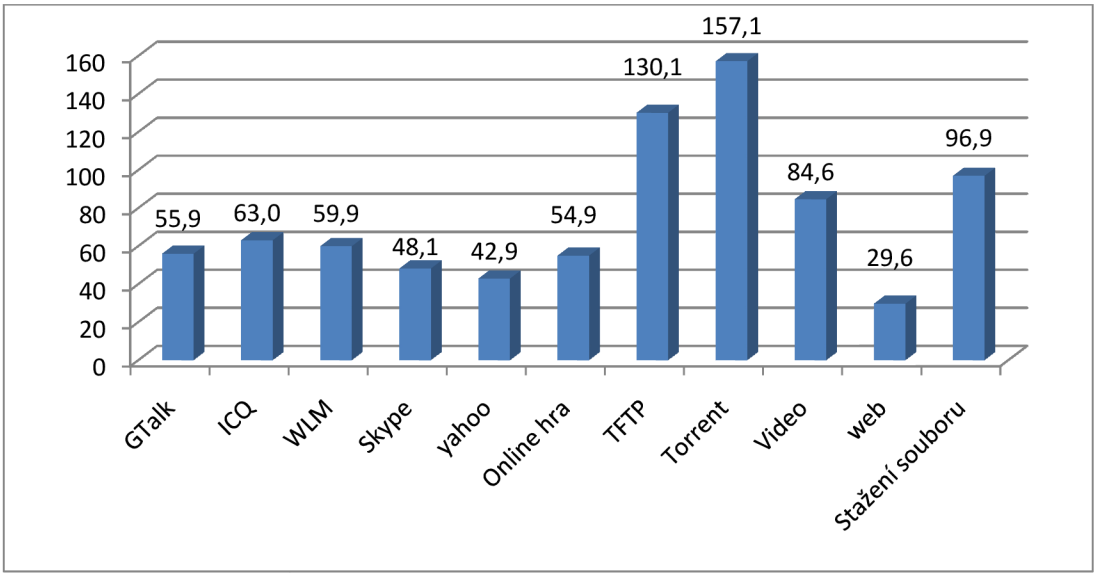
Tabulka 2.3 Souhrn statistik ostatních aplikací

Po srovnání uvedených charakteristik jednotlivých aplikací je vidět, že každá VoIP aplikace má jiné charakteristiky. Pokud tyto aplikace srovnáme s jinými internetovými aplikacemi lze vidět, že VoIP aplikace mají podobné charakteristiky. To bylo následně ověřeno na větší datové sadě, jejíž popis a výsledky jsou uvedeny v kapitole 4.

Velký vliv na výše uvedené charakteristiky má rychlost internetového připojení. U stahování souboru, torrentu nebo TFTP přenosu záleží na rychlosti internetového připojení. Pokud je přípojka pomalá, tak parametry, jako počet paketů za sekundu a bajtů za sekundu, se posunou směrem dolů a budou nižší, než na rychlejších přípojkách. Toto ovšem neplatí u VoIP aplikací, resp. není tak znatelný rozdíl, jako např. u stahování souboru, kde stahuji data umístěná na serveru, a doba přenosu souboru záleží jen na rychlosti připojení. Hlasové aplikace potřebují vždy přenést data, která se aktuálně vytvářejí kódováním hlasu z mikrofonu. Nelze tedy přenášet to, co ještě nebylo nahráno, a proto je charakteristika VoIP aplikací stejná jak na 20Mbit, tak i na 2Mbit přípojce.

Na základě pozorování lze odvodit dvě vlastnosti VoIP provozu.

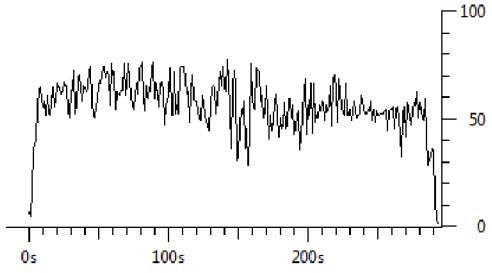
1. **Průměrný počet paketů za sekundu je u VoIP aplikací přibližně stejný**



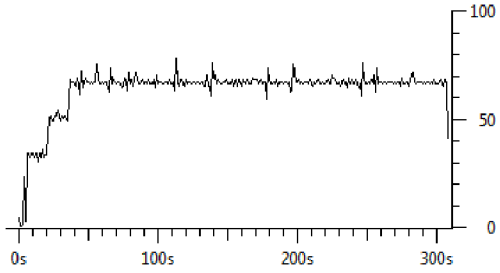
Obrázek 2.1 Průměrný počet paketů za sekundu v jednotlivých aplikacích

Při pohledu na graf na obrázku 2.1 vidíme, že počet paketů za sekundu je u VoIP aplikací přibližně stejný a pohybuje se v hodnotách od 40 do 60 paketů za vteřinu. Ostatní aplikace mají tyto hodnoty vyšší, případně nižší v závislosti na tom, jak jsou datově náročné. Stahování dat pomocí torrentu nebo TFTP přenos jsou datově náročné služby, a proto mají vysokou četnost paketů za sekundu.

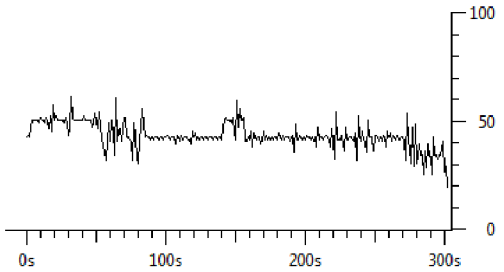
Následujících pět grafů (Obrázek 2.2 - Obrázek 2.6) zobrazují počet paketů v každé vteřině pěti minutového hovoru. Osa x jsou vteřiny hovoru a osa y je počet paketů v každé vteřině.



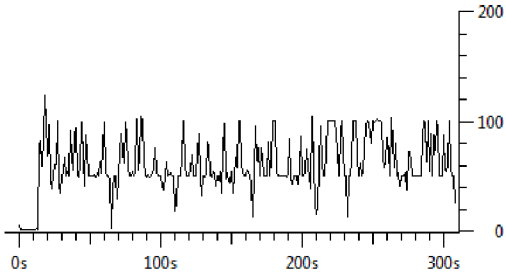
Obrázek 2.2 Gtalk – počet paketů za sekundu



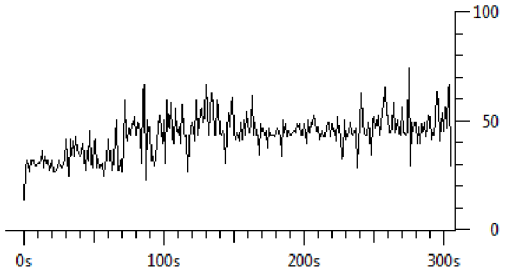
Obrázek 2.3 ICQ – počet paketů za sekundu



Obrázek 2.4 Skype – počet paketů za sekundu

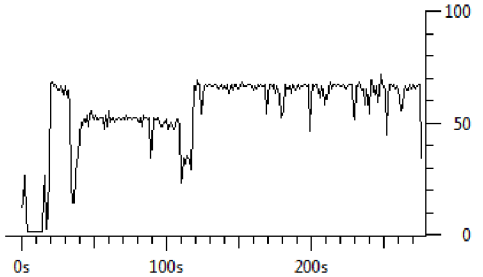


Obrázek 2.5 WLM – počet paketů za sekundu

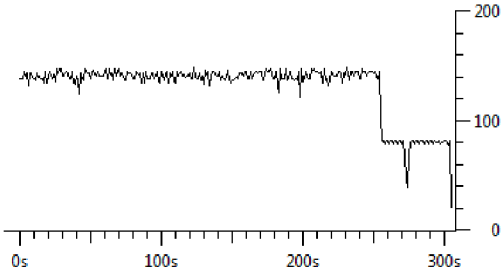


Obrázek 2.6 Yahoo – počet paketů za sekundu

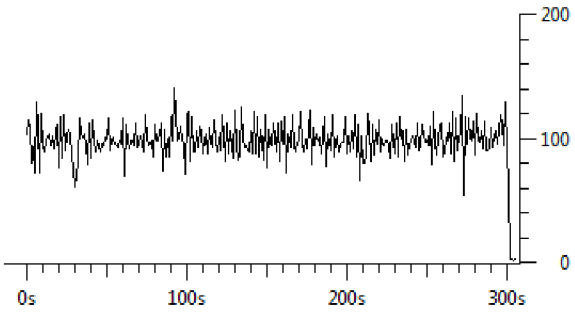
Pro srovnání zobrazuje následujících šest grafů datový průběh ostatních aplikací, jako jsou online hra, TFTP přenos, stahování souboru, web, torrent, video z youtube. Na grafech je zachycen počet paketů v každé vteřině v průběhu pěti minut



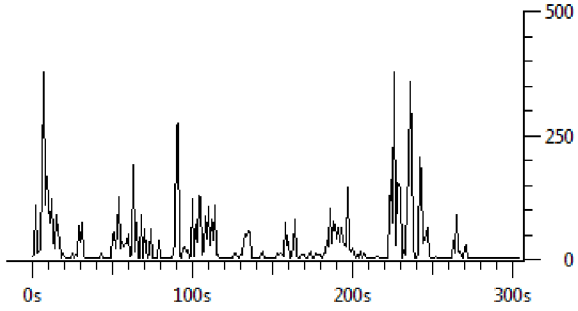
Obrázek 2.7 Online hra – počet paketů za sekundu



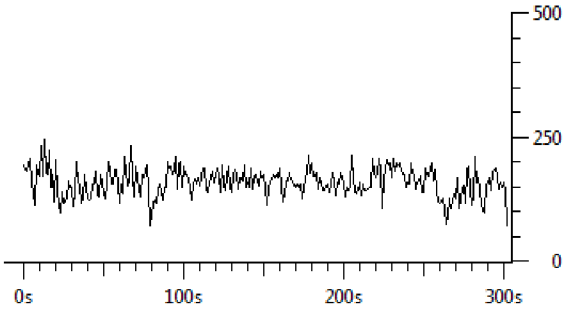
Obrázek 2.8 TFTP – počet paketů za sekundu



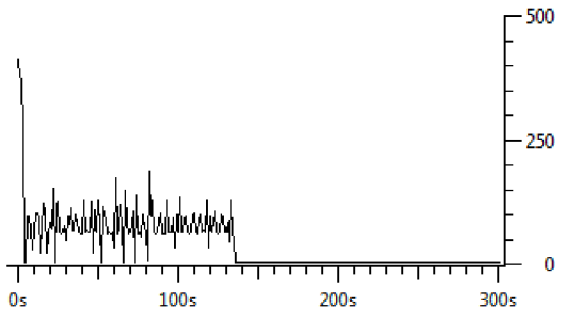
Obrázek 2.9 Stažení souboru – počet paketů za sekundu



Obrázek 2.10 Web – počet paketů za sekundu



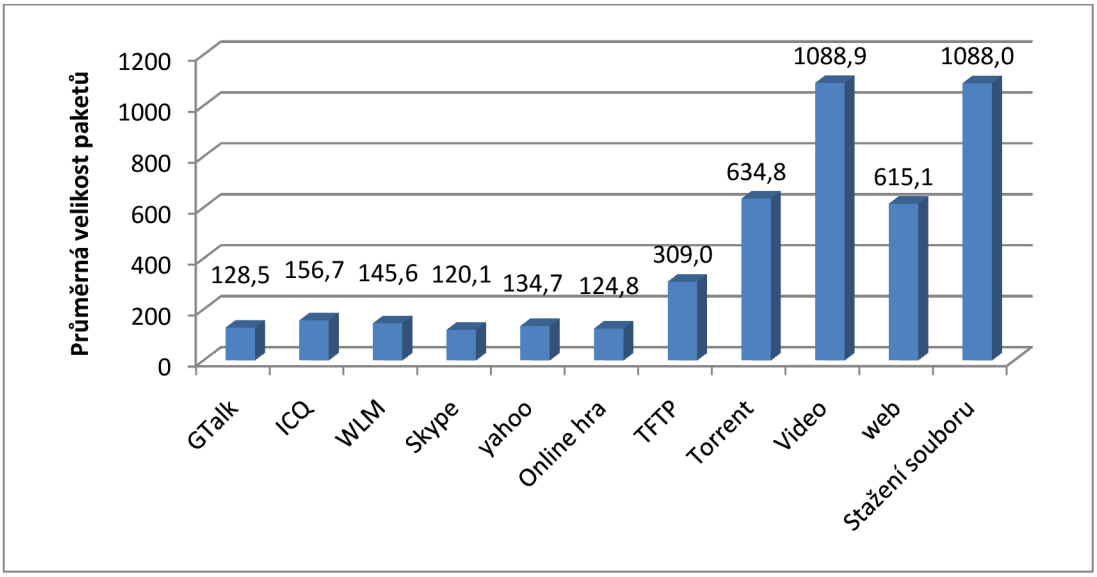
Obrázek 2.11 Torrent – počet paketů za sekundu



Obrázek 2.12 Video z youtube – počet paketů za sekundu

Na předchozích grafech lze vidět, že každá aplikace má jiný průběh počtu paketů za sekundu. Přesto si lze všimnout, že všechny VoIP aplikace mají tento počet přibližně stejný.

2. Průměrná velikost paketů je menší než v ostatním provozu



Obrázek 2.13 Průměrná velikost paketů

VoIP aplikace využívají menší pakety, aby nedocházelo k fragmentaci po cestě od zdroje k cíli a tím nedocházelo ke zpožďování paketů. Dalším důvodem použití malých paketů je minimalizace výpadků hovoru při ztrátě jednoho či více paketů. Pokud zakódovaný hlas rozdělíme na velké pakety, které pošleme k cíli, a několik po sobě jdoucích paketů nedorazí do

cíle, tak dojde k výpadku větší části hovoru. Jestliže jsou použity menší pakety, tak při výpadku několika, po sobě jdoucích paketů nemusí být výpadek slyšitelný. Ostatní aplikace, například stahovaní souboru, využívají větší pakety, protože se snaží poslat v paketu co největší množství informací a tím zmenšit počet paketů vyslaných do sítě.

Tabulka 2.4 obsahuje porovnání statistických vlastností VoIP aplikace (ICQ) se stahováním souboru a online hrou. Na první pohled je patrné, že se VoIP aplikace a stahování souboru výrazně liší v charakteristice provozu. Při stahování souboru jsou odchozí pakety malé a jejich četnost je menší než u příchozích paketů, které jsou podstatně větší a četnější. To je dáno povahou provozu při stahování souboru. Příchozí pakety obsahují data, kdežto odchozí pakety jsou převážně potvrzovací a obsahující pouze TCP hlavičku. V případě ICQ je vidět, že četnost paketů i jejich průměrná velikost jsou přibližně stejné. I toto je dáno povahou provozu, kdy jde o hlasovou komunikaci, kde obě strany posílají zakódovaná zvuková data.

Při porovnání VoIP a online hry je vidět, že mají velmi podobné charakteristiky. Celková průměrná velikost paketů a celkový průměrný počet paketů za sekundu jsou přibližně stejné. Online hra je tedy velmi podobná VoIP aplikacím. Jediná charakteristika, ve které se online hra výrazně odlišuje od VoIP aplikací, je maximální velikost paketu.

		ICQ	**Stažení souboru**	**Online hra**
odchozí	Paketů	9879	8547	8171
	Bajtů	1366858	464076	1291642
	Paketů/sek.	31.988	27.972	29.552
	Průměrná velik. paketů	138.360	54.297	158.076
	Minimální velik. paketů	62	54	58
	Maximální velik. paketů	304	732	1442
	Bajtů/sek.	4425.863	1518.819	4671.397
	Mbit/sek.	0.035	0.012	0.037
Příchozí	Paketů	9624	21054	7013
	Bajtů	1412021	31741948	603053
	Paketů/sek.	31.162	68.905	25.363
	Průměrná velik. paketů	146.719	1507.645	85.991
	Minimální velik. paketů	86	54	58
	Maximální velik. paketů	299	1514	839
	Bajtů/sek.	4572.100	103884.438	2181.022
	Mbit/sek.	0.037	0.831	0.017
Celkově	Čas	302.011 sec	305.551 sec	276.5
	Paketů	19503	29601	15184
	Bajtů	2778879	32206024	1894695
	Paketů/sek.	63.150	96.878	54.915
	Průměrná velik. paketů	142.485	1088.005	124.782
	Minimální velik. paketů	62	54	58
	Maximální velik. paketů	304	1514	1442
	Bajtů/sek.	8997.962	105403.250	6852.419
	Mbit/sek.	0.072	0.843	0.055

Tabulka 2.4 Porovnání VoIP aplikace (ICQ), stahování souboru a online hry

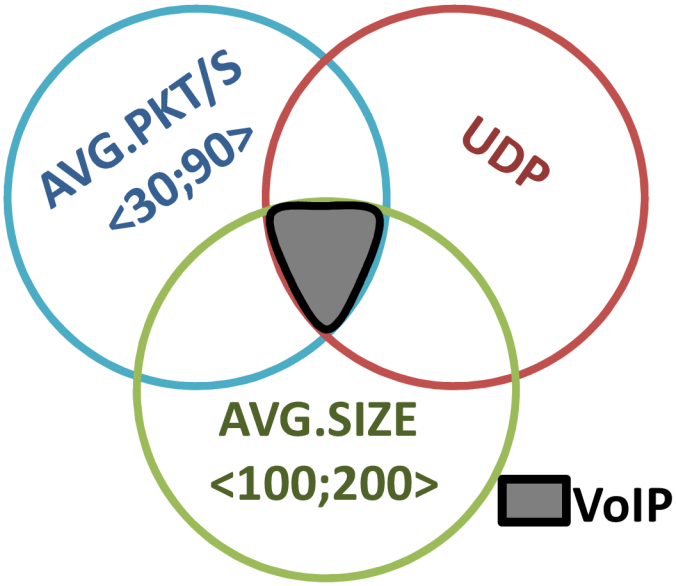
2.2.1 Detekční algoritmus

Základní algoritmus je založený na detekci chování aplikace ve statistikách toků. Algoritmus se skládá ze tří podmínek, jejichž splněním je daný tok vyhodnocen jako VoIP hovor. Pokud není jedna nebo více podmínek splněna, tak daný tok není vyhodnocen jako VoIP hovor. Algoritmus byl publikován v [4].

Základní algoritmus pro detekci VoIP aplikace:

1. Daný tok obsahuje pouze UDP pakety
2. Průměrná velikost paketů je od 100 do 200 bajtů
3. Četnost paketů za sekundu je 30-90

= > daný tok je VoIP hovor



Obrázek 2.14 Grafické znázornění základního algoritmu založeného na chování aplikace

2.2.2 Rozšířený detekční algoritmus

Základní detekční algoritmus nyní rozšíříme o další dvě charakteristiky. Z tabulky 2.2 je vidět, že VoIP aplikace mají omezenou velikost paketů. Největší velikost je 516 bajtů u aplikace Skype. Proto přidáme do základního algoritmu další zpřesňující podmínku:

Všechny pakety daného zkoumaného toku musí být menší než 550 bajtů.

Další charakteristika, kterou lze odvodit z tabulky 2.2 je poměr příchozích a odchozích bajtů (*ratio_io*). Tento poměr lze spočítat takto:

$$ratio\_io = \frac{bytes_{in}}{bytes_{out}}$$

Kde $bytes_{in/out}$ je poměr příchozích/odchozích bajtů daného toku. Vždy dělíme menší hodnotu větší hodnotou, tak aby výsledný poměr byl v intervalu <0;1>. Nezáleží na tom, zda v daném toku je více příchozích nebo odchozích dat, jde nám pouze o jejich poměr.

Tento poměr je u VoIP aplikace ICQ $1366858 / 1412021 = 0{,}968015348$ a v případě stahování souboru $464076 / 31741948 = 0{,}014620275$. K algoritmu, který má nyní čtyři podmínky, přidáme další podmínku:

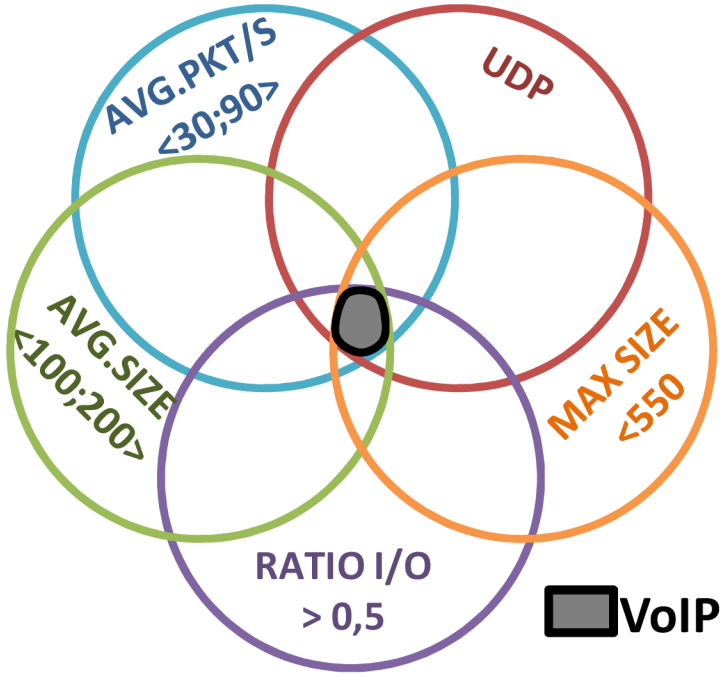
Poměr příchozích a odchozích dat musí být větší než 0,5.

Konečnou podobu rozšířeného algoritmu tvoří 5 podmínek v konjunkci. Aby daný tok byl algoritmem označen jako VoIP hovor, musí platit všechny současně, což je znázorněno na následujícím obrázku.

Rozšířený algoritmus pro detekci VoIP aplikace:

1. Daný tok obsahuje pouze UDP pakety
2. Průměrná velikost paketů je od 100 do 200 bajtů
3. Četnost paketů za sekundu je 30-90
4. Žádný paket není větší než 550 bajtů
5. Poměr příchozích a odchozích bajtů je větší než 0,5

= > daný tok je VoIP hovor



Obrázek 2.15 Grafické znázornění rozšířeného algoritmu založeného na chování aplikace

2.3 Datová charakteristika

Čtyři z pěti aplikací využívají RTP protokol pro přenos hlasu. RTP protokol umožňuje přenos dat ze zdroje k cíli v reálném čase. RTP je nezávislý na typu sítě a přenosovém protokolu, i když je často užíván nad UDP. Každý paket protokolu RTP obsahuje RTP hlavičku, která je zobrazena v tabulce 2.5, a dále jsou již samotná zakódovaná zvuková data.

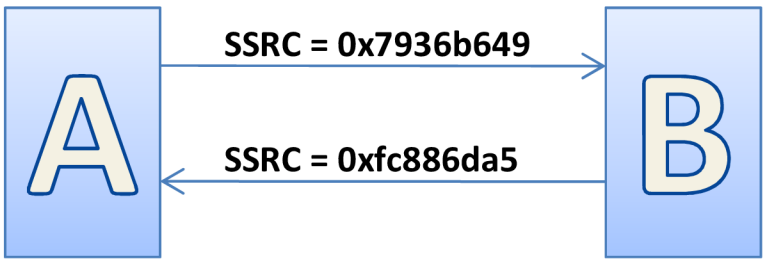
bit offset	0-1	2	3	4-7	8	9-15	16-31
0	Ver.	P	X	CC	M	PT	Sekvenční číslo
32	Časová značka						
64	SSRC identifikátor						
96	CSRC identifikátor (volitelně) ...						

Tabulka 2.5 RTP hlavička

Z použití protokolu RTP vyplývají následující dvě věci:

1. **Pakety obsahují identifikátor toku**

Identifikátor toku (SSRC identifikátor) jsou bity 64-95(9-12 bajt) RTP hlavičky. Tato hodnota je vybrána náhodně aplikací tak, aby dva synchronizační identifikátory uvnitř jedné RTP relace nebyly stejné. Obrázek 2.16 ukazuje příklad RTP proudu a jeho SSRC identifikátorů. Je vidět, že příchozí i odchozí proud má každý svůj vlastní SSRC identifikátor. Z toho vyplývá, že všechny pakety v rámci jedné relace v jednom směru musí mít stejné SSRC identifikátory. Tato vlastnost umožňuje detekovat RTP relaci.



Obrázek 2.16 Příklad SSRC identifikátoru v RTP proudu

2. **První dva bajty hlavičky mají určitou hodnotu**

Další věc, kterou lze odvodit z použití RTP protokolu, je hodnota prvních bajtů. Hlavička RTP paketu začíná označením verze RTP protokolu. V dnešní době se využívá RTP verze 2, proto hlavička RTP paketu začíná číslem 10_b, resp. bity 0-1 jsou nastaveny na 2 (10_b). Z pozorování vyplynulo, že první bajt obsahuje hodnotu 0x80. Druhý bajt je podle specifikace RTP hlavičky typ kódování hlasu (Payload type - PT), použitého v RTP proudu. Typy 96(0x60) – 127 (0x7F) nejsou přesně specifikovány a záleží na aplikaci, jaký kodek použije. V pozorováních obsahovaly první dva bajty jednotlivých aplikací tyto hodnoty:

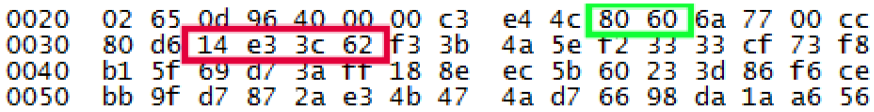
Aplikace	Hodnota prvních dvou bajtů
Gtalk	80 67
ICQ	80 60
WLM	80 72
Yahoo	80 67

Tabulka 2.6 Hodnoty prvních dvou bajtů (hexadecimálně)

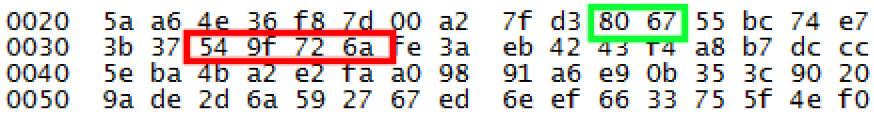
Následující dva obrázky zobrazují obě zmíněné charakteristiky RTP proudu.

Zelený obdélník – první dva bajty hlavičky

Červený obdélník – SSRC identifikátor



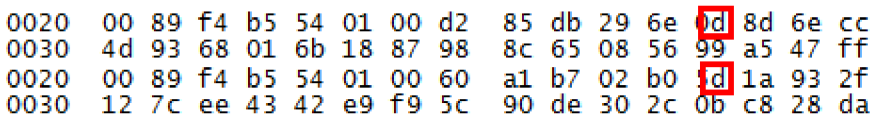
Obrázek 2.17 Data paketu - ICQ



Obrázek 2.18 Data paketu - Gtalk

Skype

Předchozí dvě charakteristiky se týkají všech zkoumaných aplikací, kromě Skype, protože jeho data jsou šifrovaná a nelze v nich nalézt ani jednu ze dvou uvedených vlastností. Přestože je provoz šifrovaný, lze nalézt jeden společný vzor v datových paketech hovoru. Ve spodních čtyřech bitech třetího bajtu se vyskytuje hexadecimální hodnota 0xD. Tato hodnota se vyskytuje vždy. Tato vlastnost je zobrazena na obrázku 2.19, který obsahuje dva po sobě následující datové pakety hovoru jednoho směru. V červených obdélnících je vyznačena spodní část třetího bajtu dat UDP paketu ve dvou po sobě jdoucích paketech toku.

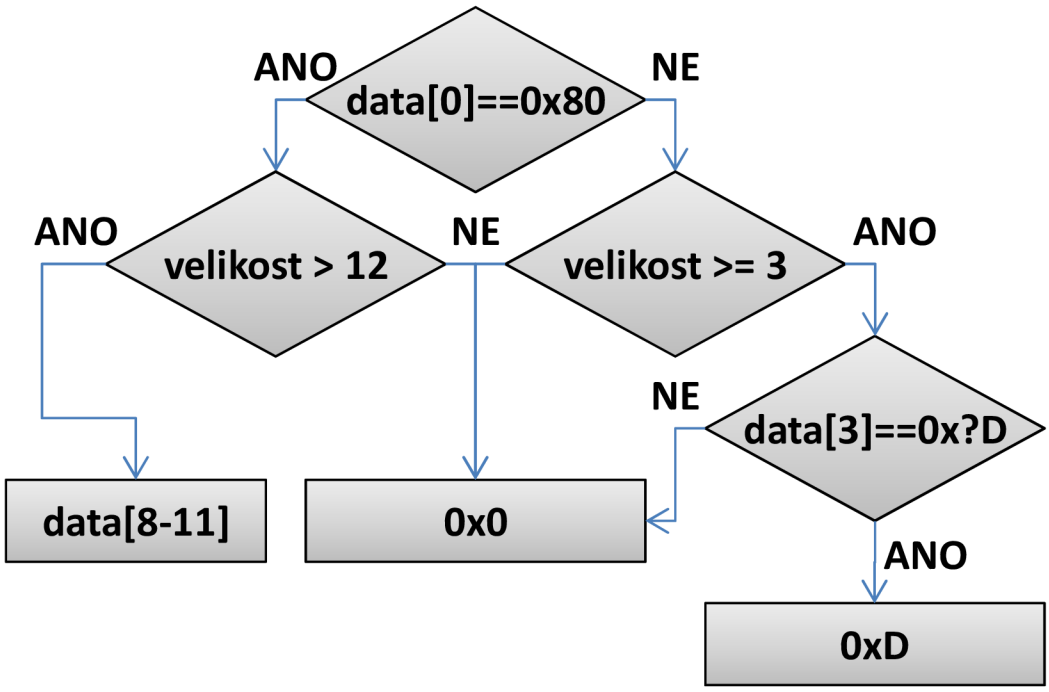


Obrázek 2.19 Data dvou po sobě jdoucích paketů (Skype)

Stejně jako u RTP protokolu se vyskytoval SSRC identifikátor, tak u Skype protokolu je spodní část třetího bajtu konstantní.

2.3.1 Detekční algoritmus

Z předchozích uvedených charakteristik jsem navrhl následující detekční algoritmus založený na hledání vzorů v obsahu paketů VoIP hovorů. V průběhu zachytávání paketů si program ukládá určitou hodnotu z každého paketu. Způsob výběru této hodnoty z datové části paketu je zobrazen v diagramu na obrázku 2.20. Pokud příchozí paket obsahuje na prvních 2 bajtech 0x80, jedná se pravděpodobně o RTP protokol, a proto si zapamatuji 9-12 byte (identifikátor toku). Pokud paket neobsahuje 0x80, ale na spodních čtyřech bitech třetího bajtu je hodnota 0xD, jedná se nejspíše o data aplikace Skype, a proto si zapamatuji hodnotu 0xD (Skype). V ostatních případech si zapamatuji nulu. Tato hodnota je uložena v toku podle toho, jestli je daný paket příchozí nebo odchozí.



Obrázek 2.20 Diagram výběru hodnoty

Při exportu toku se vypočítá poměr počtu výskytů nejčetnějšího vzoru v daném toku k počtu paketů pro každý směr zvlášť. Pokud je nejčetnějším vzorem 0, tak je poměr nastaven také na hodnotu 0. Poměry pro příchozí a odchozí směr jsou vypočítány takto:

$$ratio_{in} = \frac{\max\_count_{in}}{count_{in}}$$

$$ratio_{out} = \frac{\max\_count_{out}}{count_{out}}$$

Celkový poměr (Ratio) je vypočten jako vážený průměr příchozích a odchozích poměrů takto:

$$ratio = \frac{\left((ratio_{in} * count_{in}) + (ratio_{out} * count_{out})\right)}{(count_{in} + count_{out})}$$

Popis pojmů užitých v předchozích rovnicích je v následující tabulce.

Termín	Popis
max_count_{xx}	počet výskytů nejčetnějšího vzoru v příchozích/odchozích datech
$count_{xx}$	počet příchozích/odchozích paketů
$ratio_{xx}$	poměr nejčetnější příchozí/odchozí hodnoty

Tabulka 2.7 Popis termínů použitých v rovnicích

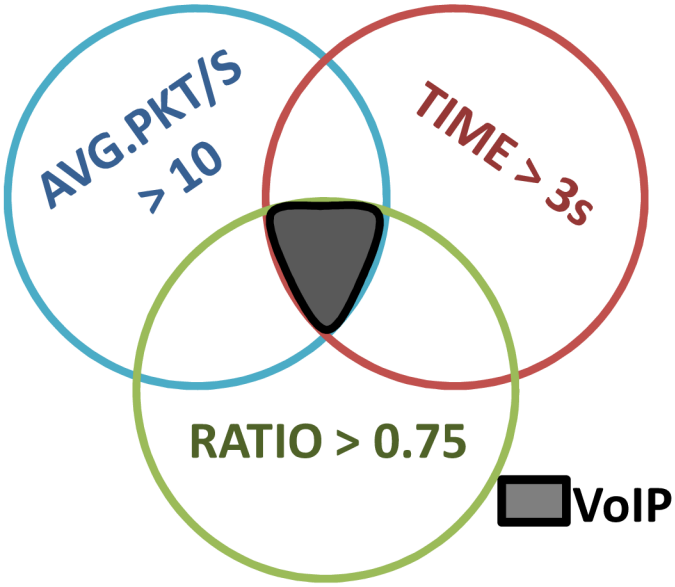
Pokud je daný tok VoIP hovor a používá RTP protokol, tak v příchozím i odchozím směru bude jedna hodnota (SSRC identifikátor) zastoupena ve velmi vysokém poměru (90% – 100%) k ostatním hodnotám, v případě aplikace Skype to bude hodnota 0xD. Ostatní dvě podmínky

(čas a průměrný počet paketů za sekundu) jsou pro omezení vlivu krátkých toků, které by mohly mít vysoký poměr jednoho vzoru.

Rozšířený algoritmus pro detekci VoIP aplikace:

1. Čas je větší než 3 sekundy
2. Četnost paketů za sekundu je větší než 10 paketů za sekundu
3. Poměr jedné hodnoty k ostatním hodnotám paketů je větší než 0,75

= > daný tok je VoIP hovor

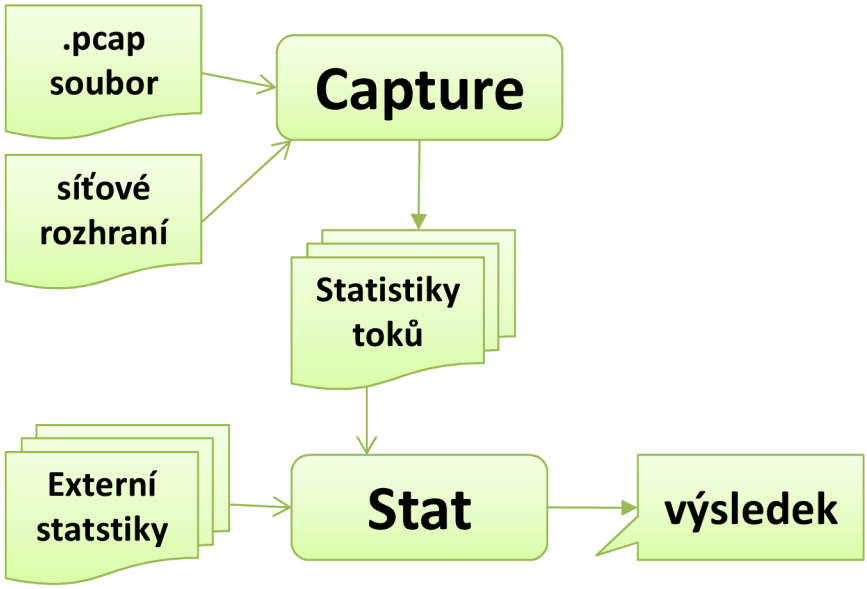


Obrázek 2.21 Grafické znázornění algoritmu založeného na detekci vzorů

Uvedený detekční algoritmus funguje pro všech 5 zkoumaných aplikací. Jakmile je provoz šifrován, tunelován nebo není možné získat data z paketů, tak je uvedený algoritmus nefunkční.

3 Implementace

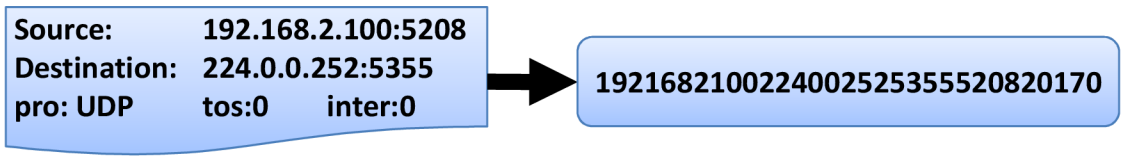
Aplikace pro detekci VoIP hovorů je napsaná v jazyce C/C++. Je možné ji provozovat na platformě Windows nebo Linux. Aplikace využívá volně dostupnou knihovnu libpcap [13] pro sběr paketů ze síťové karty nebo souboru. Program je složen ze dvou samostatných modulů. Schéma aplikace je znázorněno na následujícím obrázku.



Obrázek 3.1 Schéma aplikace

3.1 Modul Capture

Modul může běžet v jednom ze dvou režimů. Buď zachytává pakety přímo ze síťové karty, nebo je vstupem soubor v pcap formátu, který obsahuje dříve zachycené pakety. Z informací získaných z každého příchozího paketu je vytvořen hash, který umožní tento paket zařadit do toku.



Obrázek 3.2 Schéma vytvoření hodnoty hash

Hodnota hash slouží k vyhledávání příslušnosti k toku. Všechny toky jsou uloženy v paměti a toky se vyhledávají podle hodnoty hash. Pokud již je vytvořený hash v paměti, tak jsou přidány do toku informace o paketu, jako jeho velikost, časové razítko a hodnota. Pokud ještě tok v paměti není, vytvoří se nová třída, která v sobě uchovává statistické informace o toku. Toky jsou průběžně exportovány a vždy je jim nastaven nějaký důvod, pro který byl daný tok exportován.

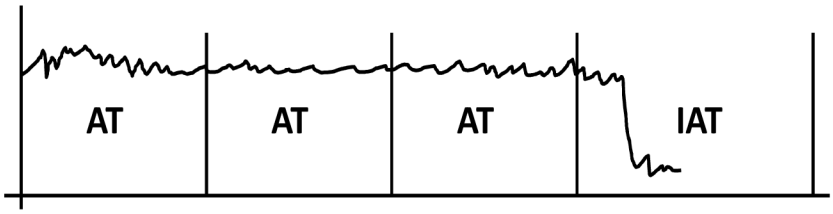
Při přidávání paketu do toku se nejdříve zkontroluje, jestli tok nepřekročil příchodem paketu jeden z limitů AT nebo IAT. Pokud ano, je tok exportován a důvod je nastaven na AT nebo IAT, podle toho, který z limitů tok překročil. Při exportu toku jsou veškeré statistické

vlastnosti vymazány a příchozí paket je prvním paketem toku. Pro odstranění krátkých toků, které již nemají další příchozí pakety je v aplikaci implementován ještě jeden mechanismus. Aplikace jednou za 30 sekund projde všechny toky a exportuje ty, které mají interval od jejich posledního paketu do aktuálního času větší než je hodnota IAT. Poté je tok kompletně vymazán z paměti i s instancí třídy, která jej tvořila. Důvod exportu je IAT-GLOBAL. Pokud aplikace v režimu čtení ze souboru dojde na konec souboru, tak jsou všechny zbývající toky exportovány a vymazány z paměti. V tomto případě je nastaven důvod EOF a aplikace je ukončena. Pracuje-li aplikace v režimu živého sběru dat a dojde-li k zachycení signálu SIGINT, jsou všechny toky exportovány. Důvod je nastaven na EOC a aplikace je ukončena.

Důvod exportu	Popis	Doba exportu
AT	Active timeout; doba od prvního paketu toku do aktuálně příchozího paketu	Při příchodu paketu
IAT	Inactive timeout; doba od posledního paketu toku do aktuálně příchozího paketu	Při příchodu paketu
IAT-GLOBAL	Inactive timeout; doba od posledního paketu toku do aktuálního času	Globálně v aplikaci
EOF	End of file; v režimu načítání souboru bylo dosaženo konce souboru,	Globálně v aplikaci
EOC	End of capture; v režimu živého sběru paketů, byl ukončen sběr dat	Globálně v aplikaci

Tabulka 3.1 Popis důvodů exportu toků v aplikaci Capture

Tok, který probíhá delší dobu, než jsou limity AT a IAT, je rozdělen na více samostatných toků, které mají společný pouze hash. Grafické znázornění takto rozděleného toku je zobrazeno na obrázku 3.3. Vyhodnocování, zda se jedná o VoIP hovor, probíhá pro každou část toku zvlášť. Pokud je jedna z částí vyhodnocena jako VoIP, potom je celý tok označen jako VoIP.



Obrázek 3.3 Ukázka rozdělení toku pomocí limitů AT a IAT

Příklad výstupu modulu Capture.

```
+====================== Net Flow ======================
Source            : 89.103.90.166:63601
Destination       : 192.168.0.137:16384
Protocol          : 17(UDP)
TOS               : 0
Interface         : 0
Hash              : 192168013789103901661638463601017O
Start time        : 23:16:21.402475
End time          : 23:16:51.986223
Total time        : 30.584
Release           : 1(AT),
```

```
----------------------  Statistics  ----------------------
----------------------   Incoming   ----------------------
Packets          : 815
Bytes            : 128661
Avg. packet/sec  : 26.648
Avg. bytes/s     : 4206.842
Avg. packet size : 157.866
Packet min size  : 62
Packet max size  : 303
Std.dev          : 44.019
Avg. Mbit/s      : 0.034
----------------------   Outgoing   ----------------------
Packets          : 880
Bytes            : 123949
Avg. packet/sec  : 28.773
Avg. bytes/s     : 4052.773
Avg. packet size : 140.851
Packet min size  : 62
Packet max size  : 308
Std.dev          : 38.335
Avg. Mbit/s      : 0.032
----------------------     Total    ----------------------
Packets          : 1695
Bytes            : 252610
Avg. packet/sec  : 55.422
Avg. bytes/s     : 8259.615
Avg. packet size : 149.032
Packet min size  : 62
Packet max size  : 308
Std.dev          : 42.035
Avg. Mbit/s      : 0.066
----------------------   Patterns   ----------------------
Inbound          : 21B67732=811x(0.995)
Outbound         : 3F1672DC=876x(0.995)
+============================================================
```

3.2 Modul Stat

Modul slouží k vyhodnocování toků. Vstupem jsou statistiky toků exportované modulem Capture, případně externí statistiky, které musí být ve stejném formátu jako statistiky vytvořené modulem Capture. Modul Stat nejdříve načte informace o daném toku z textového souboru se statistikami toků. Pomocí interních funkcí program určí, zda se jedná o VoIP tok nebo ne, a tuto informaci společně se stručným popisem toku vytiskne na standardní výstup. Detekční schopnost zabudovaných algoritmů, lze ovlivňovat nastavením limitních hodnot pomocí argumentů příkazové řádky při spuštění programu.

Příklad výstupu modulu Stat:

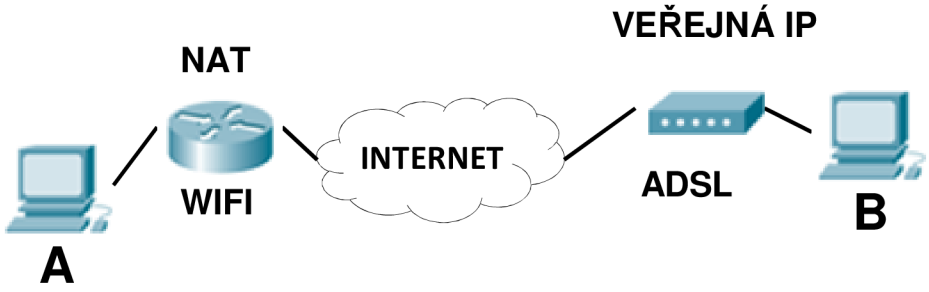
```
+======================= Net Flow ========================
89.103.90.166:63601 -> 192.168.0.137:16384, protocol: 17(UDP)
hash    :192168013789103901661638463601017O
stat    :1,JE VOIP (pro:UDP,pckt/s:55.42,size:149.03)
stat_ext:1,JE VOIP (pro:UDP,pckt/s:55.42,size:149.03,
                    ratio_io:0.96,max:308)
pattern :1,JE VOIP (pckt/s:55.42,time:30.58,ratio:1.00)
+========================================================
```

4 Testování a výsledky

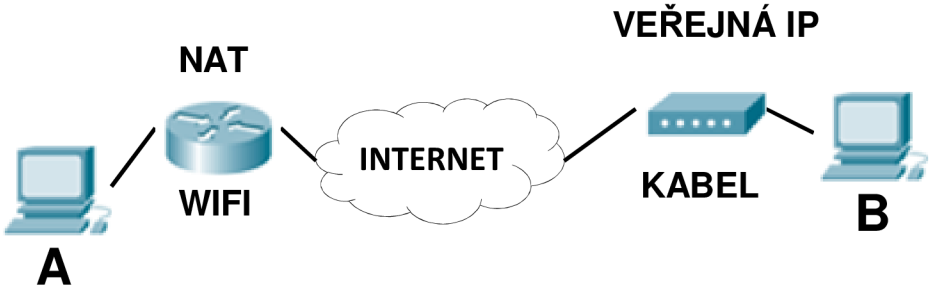
4.1 Popis datové sady

4.1.1 Sada A - Hovory jednotlivými aplikacemi

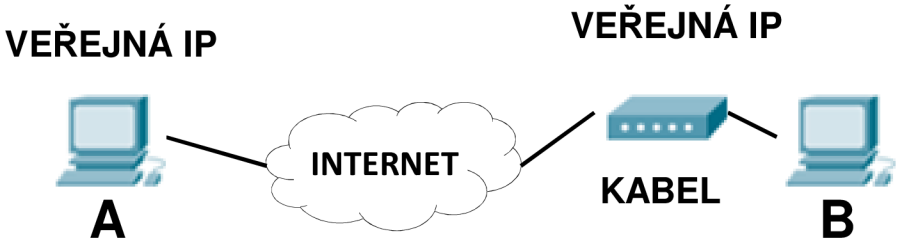
V průběhu jednoho měsíce bylo uskutečněno 5x45 hovorů (45 hovorů každou z pěti zkoumaných VoIP aplikací). Tyto hovory byly provedeny na třech různých sítích. Schémata zapojení sítí jsou znázorněna na obrázcích 4.1 – 4.3. V daném zapojení sítě bylo provedeno 15 hovorů každou aplikací. Hovory byly obousměrné, příchozí i odchozí, a délky hovorů byly 30 a 120 sekund. Rozdělení souborů a jejich parametry jsou v tabulce 4.1.



Obrázek 4.1 Schéma zapojení sítě 1



Obrázek 4.2 Schéma zapojení sítě 2



Obrázek 4.3 Schéma zapojení sítě 3

Číslo souboru	Směr	Délka	Stanice A	Stanice B
00 - 04	A → B	30s	ISP: WIFI NAT	ISP: ADSL VEŘ. IP
05 - 09	B → A	30s		
10 - 14	A → B	120s		
20 - 24	A → B	30s	ISP: WIFI NAT	ISP: KABEL VEŘ. IP
25 - 29	B → A	30s		
30 - 34	A → B	120s		
40 - 44	A → B	30s	ISP: CESNET VEŘ. IP	ISP: KABEL VEŘ. IP
45 - 49	B → A	30s		
50 - 54	A → B	120s		

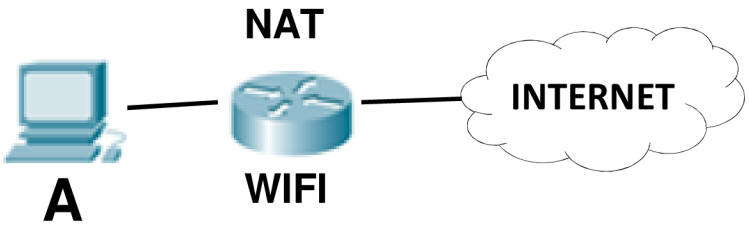
Tabulka 4.1 Popis souborů datové sady A

4.1.2 Sada B - Data z projektové akademické sítě

Sada B obsahuje 12 GB dat z projektové akademické sítě. V této sadě se nenachází žádný VoIP hovor. Sada je použita pro zhodnocení detekčních algoritmů z hlediska falešně pozitivních nálezů. Čím méně falešně pozitivních nálezů, tím je detekční schopnost algoritmu lepší.

4.1.3 Sada C - Jednodenní provoz PC

Sada C obsahuje 18 hodinový provoz PC připojeného k internetu podle schématu na obrázku 4.4. Soubor má 1,8 GB a obsahuje 20 hovorů. Dále obsahuje ostatní provoz jako je surfování na internetu, stahování souborů, online hra, TFTP přenos, poslech internetového rádia, sledování internetové televize. Po celou dobu získávání dat, bylo všech pět VoIP aplikací spuštěno a byly průběžně využívány ke komunikaci a zmíněným hovorům.



Obrázek 4.4 Schéma připojení do internetu

4.2 Vliv parametrů na úspěšnost detekce

Testování bylo provedeno na datové sadě A, která obsahuje 225 hovorů v 225 souborech. V každém souboru je pouze jeden hovor a ostatní provoz je omezen jen na systémové služby, proto výskyt jiných datových toků je minimální. Tento předpoklad je důležitý pro použitou metodu testování. Pokud algoritmus nevyhodnotí jako VoIP hovor žádný tok, tak výsledkem je jeden falešně negativní záznam. Pokud je pouze jeden tok vyhodnocen jako VoIP hovor, tak je daný soubor vyhodnocen jako jeden pozitivní záznam. Je-li v souboru označeno více VoIP hovorů, tak jejich počet, zmenšený o jeden opravdový hovor, jsou falešně pozitivní záznamy.

Níže uvedené grafy zobrazují procentuální počet úspěšně detekovaných hovorů v závislosti na nastavení parametrů prvního detekčního algoritmu, tedy algoritmu založeného na statistických vlastnostech VoIP provozu. Na ose x jsou konkrétní hodnoty limitů pro algoritmus a na ose y je procentuální úspěšnost těchto hodnot. Stoprocentní úspěšnost znamená, že všech 225 VoIP hovorů bylo označeno jako VoIP. Hodnota 0% znamená, že žádný z 225 toků nebyl označen jako VoIP. Na uvedené hodnoty nemají vliv falešně pozitivně označené toky.

Postupně byly pro každé testování přidávány omezující podmínky. První test má pouze jednu omezující podmínku (UDP tok) a jsou testovány limity četnosti paketů za vteřinu. Další test má již dvě omezující podmínky – UDP tok a rozmezí četnosti paketů. Omezující podmínky musí být přítomny, protože nastavování např. maximální velikosti paketu bez jiných omezujících podmínek přináší absolutně nepoužitelné výsledky.

Zde je seznam termínů vyskytujících v další části textu a jejich vysvětlení:

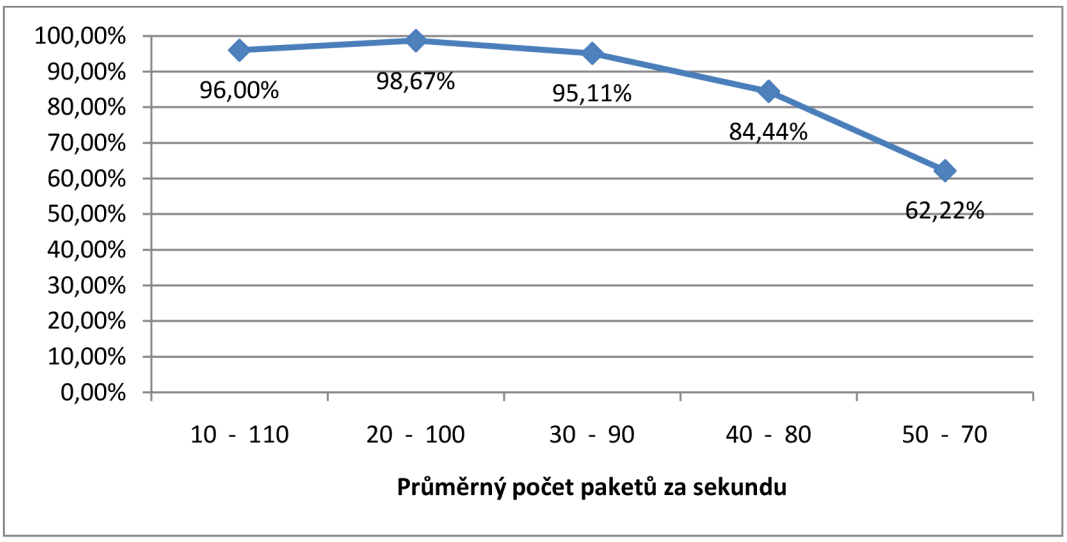
Zkratka	Popis
pckt/s	průměrný počet paketů za sekundu
avg.size	průměrná velikost paketů
ratio_io	poměr příchozích a odchozích bajtů
max	maximální velikost paketu
time	čas od prvního do posledního paketu toku

4.2.1 Algoritmus založený na chování aplikace

Grafy v této podkapitole zobrazují vliv nastavení parametrů algoritmu na procentuální úspěšnost detekce pomocí detekčního algoritmu, který je založený na statistických vlastnostech VoIP toků.

Nastavení četnosti paketů za sekundu

Z grafu je patrné, že algoritmus nejlépe detekuje VoIP toky při nastavení limitů pro průměrný počet paketů za sekundu na hodnoty od 20 do 100 paketů za sekundu.

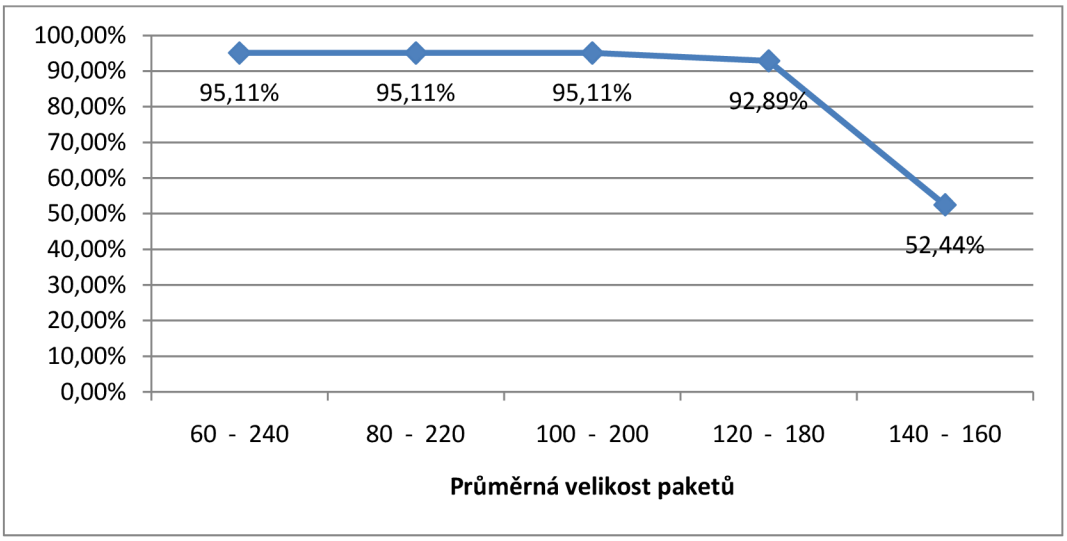


Obrázek 4.5 Průběh úspěšnosti detekce při různých hodnotách četnosti paketů za sekundu

Omezující podmínky: UDP

Nastavení průměrné velikosti paketů

Rozmezí průměrné velikosti paketů 100-200 bajtů se ukázalo jako nejlepší. Další zvětšování rozmezí již nemá na procento správně detekovaných VoIP toků vliv. Zmenšování rozmezí se negativně projevilo na úspěšnosti detekce a procento správně detekovaných VoIP toků klesalo. Pokud více nastavení dává stejné procento úspěšně detektovaných VoIP hovorů, vybereme to nejvíce omezující, abychom zamezili výskytu falešně-pozitivních toků. V tomto případě vybereme nastavení 100-200 bajtů.

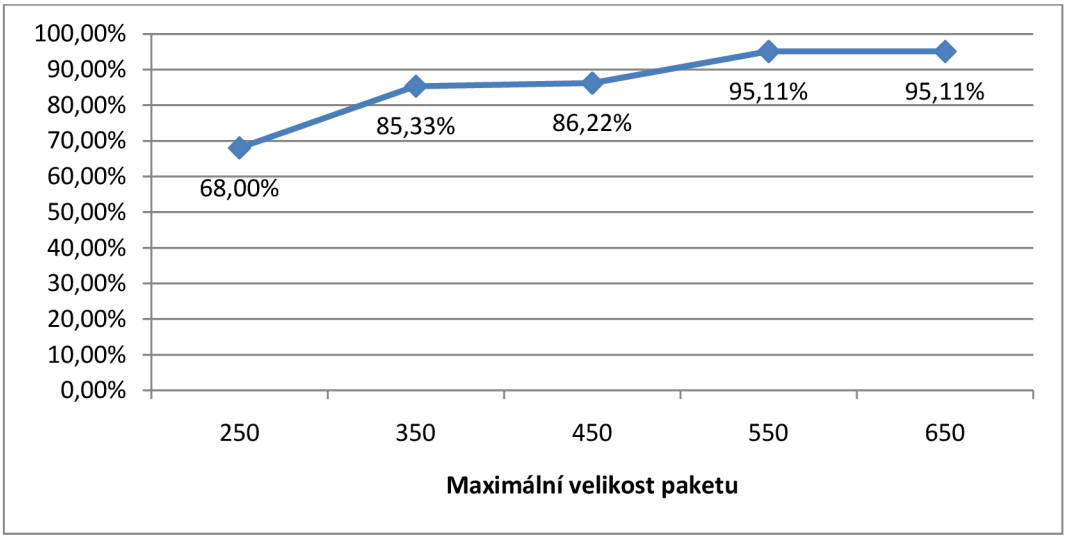


Obrázek 4.6 Průběh úspěšnosti detekce při různých hodnotách průměrné velikosti paketů

Omezující podmínky: UDP ∧ pckt/s ∈ <30;90>

Nastavení maximální velikosti paketu

Omezení maximální velikosti paketu v daném toku umožňuje eliminovat toky, které mají podobné charakteristiky jako VoIP toky, ale ke komunikaci využívají velké pakety. Příliš nízká hodnota má za následek nízkou úspěšnost detekce. Z grafu je patrné, že detekční schopnost algoritmu se nezmění od hodnoty 550 výše.

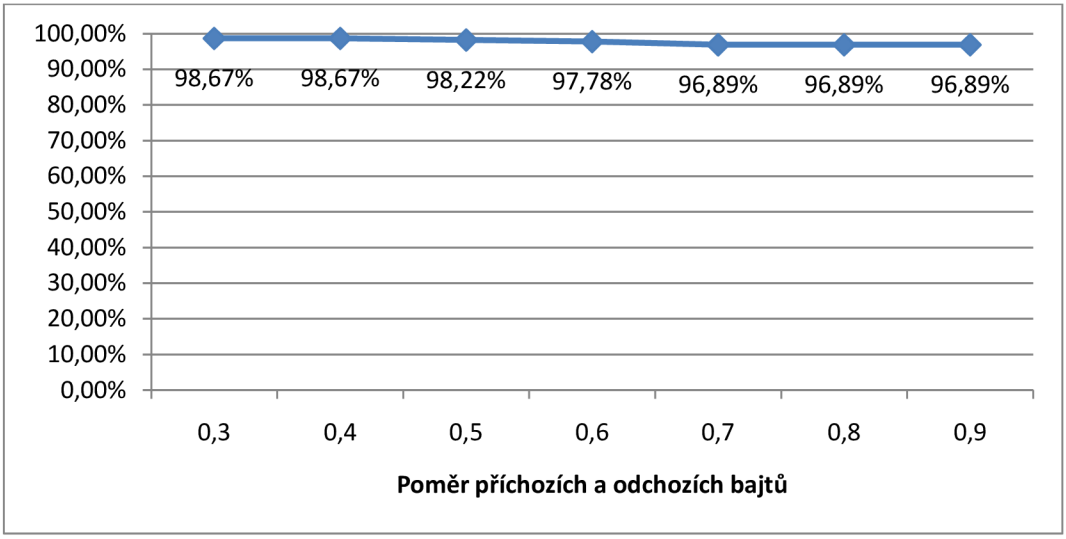


Obrázek 4.7 Průběh úspěšnosti detekce při různých hodnotách maximální velikosti paketů

Omezující podmínky: UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>

Nastavení poměru příchozích a odchozích bajtů

Poměr příchozích a odchozích bajtů nemá velký vliv na úspěšnost detekce jako výše uvedená tři kritéria, ale umožňuje nepatrně zpřesnit detekci. Algoritmus má nejlepší výsledky při nastavení limitu na hodnotu 0,4.



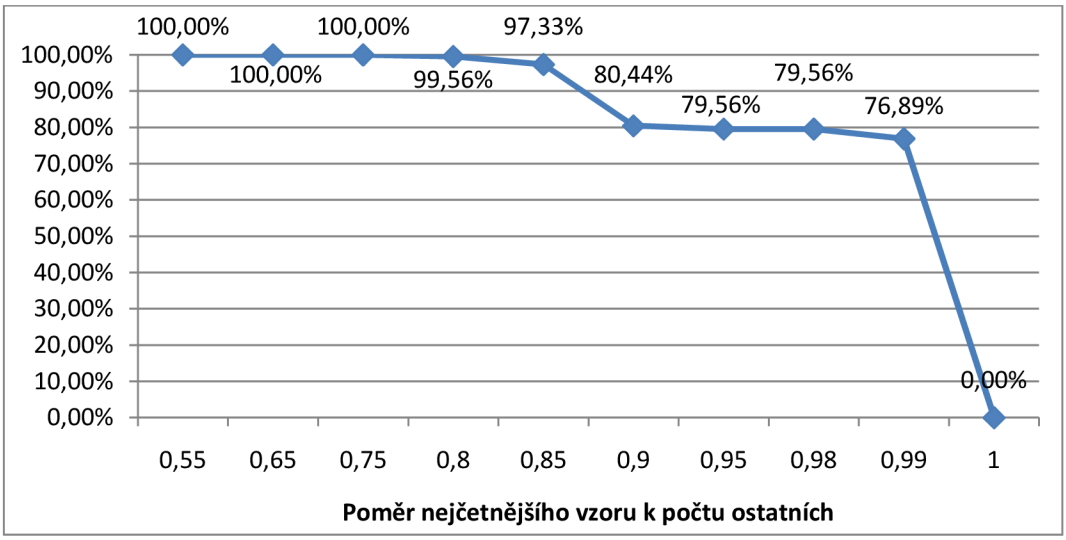
Obrázek 4.8 Průběh úspěšnosti detekce při různých hodnotách poměru příchozích a odchozích dat

Omezující podmínky: UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>

4.2.2 Algoritmus založený na hledání vzorů

Nastavení poměru nejčetnějšího vzoru

V grafu na obrázku 4.9 je zobrazen průběh úspěšnosti detekce při zvyšování limitu pro poměr počtu nejčetnějšího vzoru k počtu ostatních. Z grafu plyne hodnota 0,75 jako limitní hodnota, od níž výše se detekční schopnost snižuje. Způsob výpočtu poměru je vysvětlen v kapitole 2.3.1.



Obrázek 4.9 Průběh úspěšnosti detekce při různých hodnotách poměru výskytu nejčetnějšího vzoru k počtu ostatních

Omezující podmínky: time > 3 ∧ pckt/s > 10

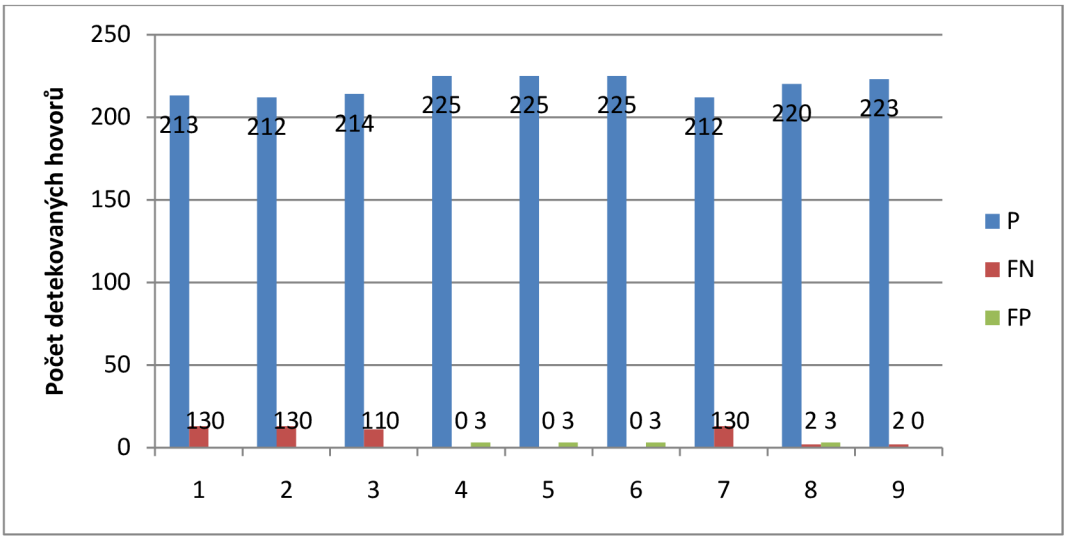
4.3 Test úspěšnosti detekce

Níže uvedené grafy zobrazují počty pozitivních, falešně pozitivních a falešně negativních toků v jednotlivých testovacích datových sadách pro oba detekční algoritmy. V grafech jsou použity následující pojmy:

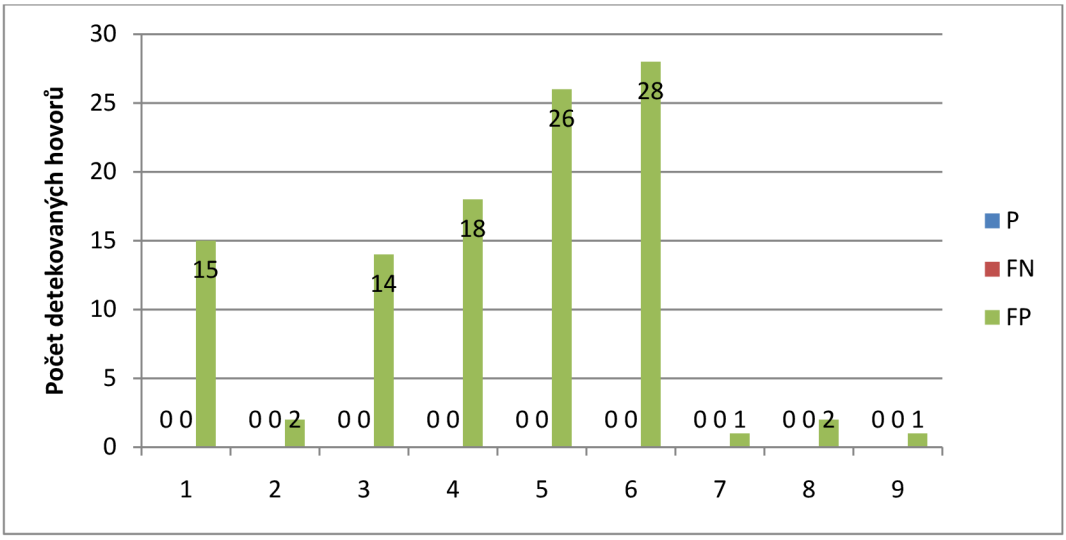
	P	Pozitivní	VoIP toky označené jako VoIP
	FN	Falešně negativní	VoIP toky neoznačené jako VoIP
	FP	Falešně pozitivní	Ostatní toky označené jako VoIP

4.3.1 Algoritmus založený na chování aplikace

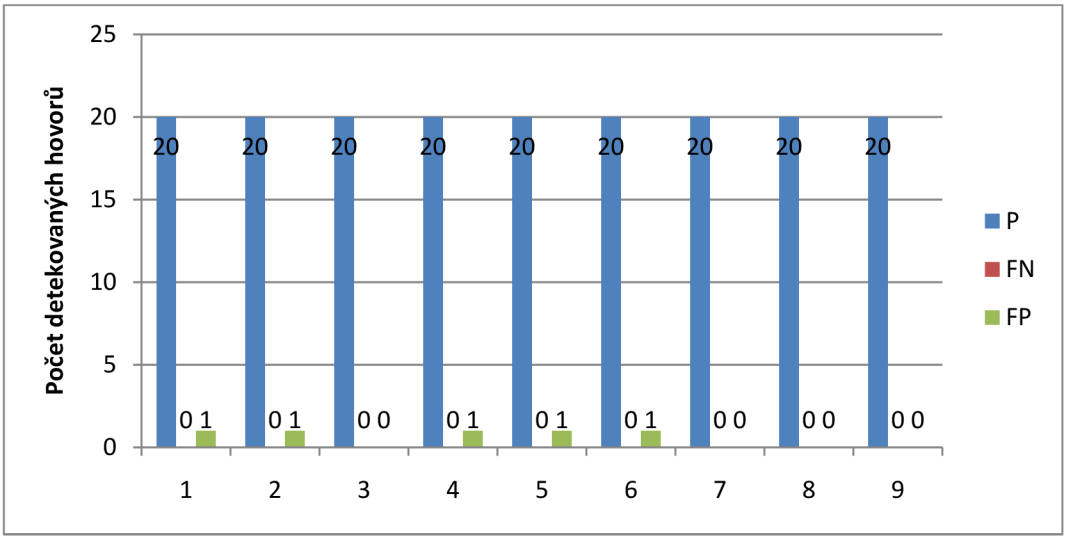
Čísla 1-9 na ose x jsou různá nastavení parametrů detekčního algoritmu. Jejich přesné hodnoty jsou zapsány v tabulce 4.2. Osa y zobrazuje počet detekovaných toků pro dané nastavení algoritmu. V sadě A je 225 hovorů, v sadě B je 0 hovorů a v sadě C je 20 hovorů.



Obrázek 4.10 Sada A – výsledky detekce VoIP hovorů pomocí statistického algoritmu



Obrázek 4.11 Sada B - výsledky detekce VoIP hovorů pomocí statistického algoritmu



Obrázek 4.12 Sada C – výsledky detekce VoIP hovorů pomocí statistického algoritmu

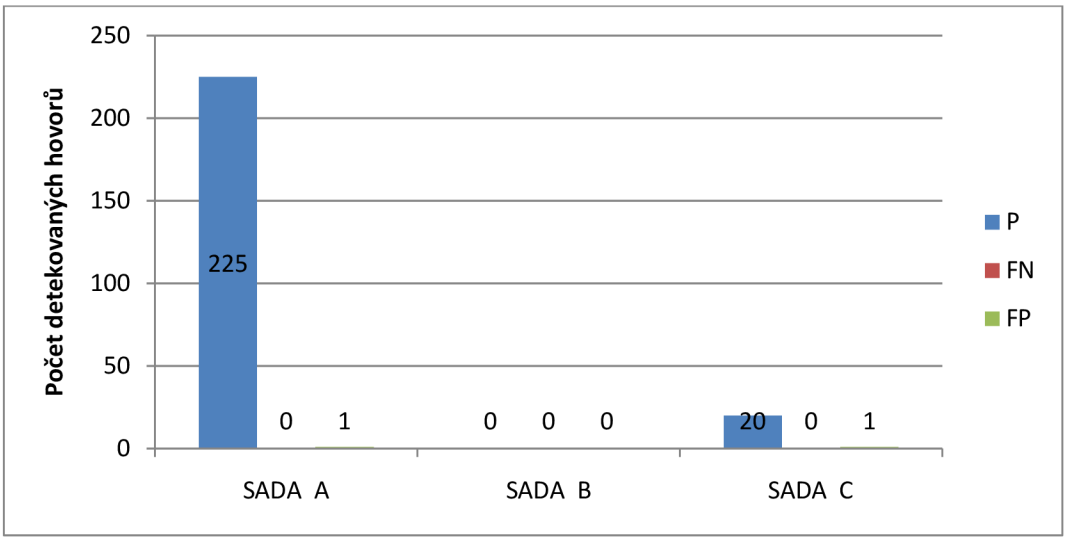
Číslo	Podmínky
1	UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>
2	UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200> ∧ ratio_io ∈ <0,4; 1)
3	UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200> ∧ max ∈ <0; 550)
4.	UDP ∧ ((pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>) ∨ (pckt/s ∈ <20;100> ∧ ratio_io ∈ <0,4; 1)))
5.	UDP ∧ ((pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>) ∨ (pckt/s ∈ <20;100> ∧ max ∈ <0; 550)))
6.	UDP ∧ ((pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200>) ∨ (pckt/s ∈ <20;100> ∧ max ∈ <0; 550)) ∨ (pckt/s ∈ <20;100> ∧ ratio_io ∈ <0,4; 1)))
7.	UDP ∧ pckt/s ∈ <30;90> ∧ avg.size ∈ <100;200> ∧ ratio_io ∈ <0,4; 1) ∧ max ∈ <0; 550)
8.	UDP ∧ pckt/s ∈ <20;100> ∧ avg.size ∈ <100;200> ∧ ratio_io ∈ <0,4; 1) ∧ max ∈ <0; 550)
9.	UDP ∧ pckt/s ∈ <25;100> ∧ avg.size ∈ <100;200> ∧ ratio_io ∈ <0,4; 1) ∧ max ∈ <0; 550)

Tabulka 4.2 Nastavení parametrů statistického detekčního algoritmu

Z grafů lze vidět, že algoritmus dává při různých nastavení parametrů různé výsledky. Hodnoty jednotlivých parametrů byly nastaveny tak, aby algoritmus dosahoval, co nejlepších výsledků. Nastavení 1 je základní detekční algoritmus uvedený v kapitole 2.2.1. Nastavení 2 a 3 jsou základní algoritmus s přidanými omezujícími podmínkami. V grafu na obrázku 4.10 lze vidět, že velmi dobrých výsledků dosáhly nastavení číslo 4-6. Ovšem na dalším grafu (Obrázek 4.11) jsou vidět velmi vysoké hodnoty falešně pozitivních toků pro nastavení 4-6. Položky 7-9 jsou rozšířený detekční algoritmus s různými hodnotami parametrů pro vylepšení detekce. Nejlepších výsledků dosáhlo nastavení číslo 9 - téměř 100% detekce v případě sady A, jen jeden falešně pozitivní hovor v sadě B a nakonec v sadě C dosáhl algoritmus úplné, ideální úspěšnosti.

4.3.2 Algoritmus založený na hledání vzorů

Algoritmus založený na detekci vzorů v obsahu paketů je velmi úspěšný v detekci hovorů pomocí zkoumaných VoIP aplikací. Ve všech třech sadách má 100% úspěšnost detekce skutečných VoIP toků. Jako falešně pozitivní toky označil algoritmus toky obsahující RTP pakety, které ale nepatřily k hlavnímu toku, v němž byl hovor přenášen.



Obrázek 4.13 Výsledky detekce VoIP hovorů pomocí algoritmu hledajícího vzory

Omezující podmínky: time > 3∧ pckt/s > 10 ∧ ratio > 0,75

Závěr

Práce se zabývala detekcí VoIP aplikací na základě jejich charakteristických vlastností. Byly představeny dvě metody detekce. První z nich, založená na chování VoIP aplikací, je převzatá metoda, která byla vylepšena tak, aby dosahovala přesnějších výsledků. Druhá metoda, založená na hledání vzorů v obsahu paketů, je vlastní metoda, která umožňuje spolehlivě detekovat aplikace využívající RTP protokol pro přenos hovorů. Metody byly experimentálně vyladěny na testovací sadě tak, aby dosahovaly přesnějších výsledků. Žádná z metod nedosáhla ideálu a neměla 100% úspěšnost detekce, ale tomuto ideálu se přiblížily. Ideální algoritmus má stoprocentní úspěšnost, označí všechny hovory správně a nemá žádný falešně pozitivní ani falešně negativní výsledek.

Při vytváření této bakalářské práce jsem se podrobněji seznámil s problematikou detekce VoIP aplikací a získal jsem nové znalosti o principu fungování VoIP aplikací. Tyto znalosti mi pomohly při analyzování VoIP provozu a díky nim jsem mohl navrhnout algoritmus, který umožňuje jejich spolehlivější detekci. Práce splnila své zadání, ale algoritmy bylo možné ověřit jen v omezené míře, protože nebyla k dispozici datová sada z reálného provozu.

V případě budoucího vývoje je možné se zabývat těmito věcmi:

- Studium charakteristik dalších VoIP aplikací a rozšíření detekčních algoritmů tak, aby byly schopny detekovat více aplikací.
- Rozšíření programu o funkci umožňující rozlišení mezi jednotlivými aplikacemi na základě určitých znaků, jimiž se daná aplikace vyznačuje.
- Je možné provést experimenty na různorodější datové sadě, zjistit důvody selhání detekce, upravit parametry detekčních algoritmů a tím vylepšit úspěšnost detekce.

Literatura

[1] HE, Qinxia. Analysing the Characteristics of VoIP Traffic. [s.l.], 2007. 95 s. Diplomová práce. Dostupný z WWW: <http://library2.usask.ca/theses/available/etd-07132007-120004/unrestricted/thesis.pdf>.

[2] SCHULZRINNE, H., et al. RTP : A Transport Protocol for Real-Time Applications, RFC 1889 [online]. Internet Engineering Task Force, 1996 [cit. 2009-03-20]. Dostupný z WWW: <http://www.ietf.org/rfc/rfc1889.txt>.

[3] CURTIS, J.P. , et al. Measurement of Voice Over IP Traffic. Proceedings PAM2000 [online]. 2000 [cit. 2009-03-20]. Dostupný z WWW: <http://moat.nlanr.net/Papers/VoIPpaper.pdf>.

[4] IDREES, Fauzia, KHAN, Uzma Aslam. A Generic Technique for Voice over Internet Protocol (VoIP) Traffic Detection. IJCSNS [online]. 2008, vol. 8, no. 2 [cit. 2009-03-20], s. 52-59. Dostupný z WWW: < http://paper.ijcsns.org/07_book/200802/20080207.pdf>.

[5] BONFIGLIO, Dario, et al. Revealing Skype Traffic : When Randomness Plays with You. SIGCOMM [online]. 2007, vol. 37, no. 4 [cit. 2009-03-20], s. 37-48. Dostupný z WWW: <http://ccr.sigcomm.org/online/files/fp118-bonfiglio.pdf>.

[6] KITAMURA, Tsutomu, et al. Traffic Identification for Dependable VoIP. NEC Technical Journal [online]. 2006, vol. 1, no. 3 [cit. 2009-03-20]. Dostupný z WWW: <http://www.nec.co.jp/techrep/en/journal/g06/n03/t060304.pdf>.

[7] Google Talk : Chat online and make free internet calls [online]. c2009 [cit. 2009-04-15]. Dostupný z WWW: <http://www.google.com/talk/>.

[8] ICQ.com : community, people search and messaging service! [online]. c1998-2009 [cit. 2009-04-15]. Dostupný z WWW: <http://www.icq.com/>.

[9] Skype : volání přes Internet [online]. c2009 [cit. 2009-04-15]. Dostupný z WWW: <http://www.skype.com/intl/cs/>.

[10] Windows Live Messenger [online]. c2009 [cit. 2009-04-15]. Dostupný z WWW: <http://download.live.com/?sku=messenger>.

[11] Yahoo! Messenger : Chat, Instant message, SMS, PC Calls and More [online]. c2009 [cit. 2009-04-15]. Dostupný z WWW: <http://messenger.yahoo.com/>.

[12] Wireshark : Go deep [online]. [2009] [cit. 2009-04-15]. Dostupný z WWW: <http://www.wireshark.org/>.

[13] TCPDUMP / LIBPCAP [online]. 2009 [cit. 2009-04-15]. Dostupný z WWW: <http://www.tcpdump.org/>.

[14] IP Protocol suite [online]. c1998-2009 [cit. 2009-03-20]. Dostupný z WWW: <http://www.networksorcery.com/enp/topic/ipsuite.htm>.

Seznam příloh