Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informatika

Ján Garaj

OPTIMALIZÁCIA VYUŽITIA VÝPOČTOVÝCH PROSTRIEDKOV PRI E-MAILOVEJ KOMUNIKÁCII

Bakalársky projekt

Vedúci bakalárskeho projektu:
Ing. Ján Máté
december, 2006

ANOTÁCIA

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informatika

Autor: Ján Garaj
Bakalársky projekt: Optimalizácia využitia výpočtových prostriedkov pri e-mailovej komunikácii
Vedenie bakalárskeho projektu: Ing. Ján Máté
máj, 2006

Úlohou projektu je optimalizovať využitie výpočtových prostriedkov pri emailovej komunikácii. Východiskom práce je analýza súčasnej implementácie typického emailového servera so zameraním na využitie diskovej kapacity. Cieľom je vytvorenie a implementácia mailového softvéru, ktorý by eliminoval nevýhody vlastností dnešných mailových systémov.

ANNOTATION

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Informatics

Author: Ján Garaj
Bachelor Theses: Optimization of the Usage of Computer Systems In E-mail Communication
Supervisor: Ing. Ján Máté
2006, May

The aim of the theses is to optimize the usage of computer systems in e-mail communication. The starting point of it is the analysis of the contemporary implementation of the common e-mail server with the stress laid on the usage of its disc capacity. The aim is to design and implement a e-mail software that would eliminate the negatives of contemporary e-mail system.

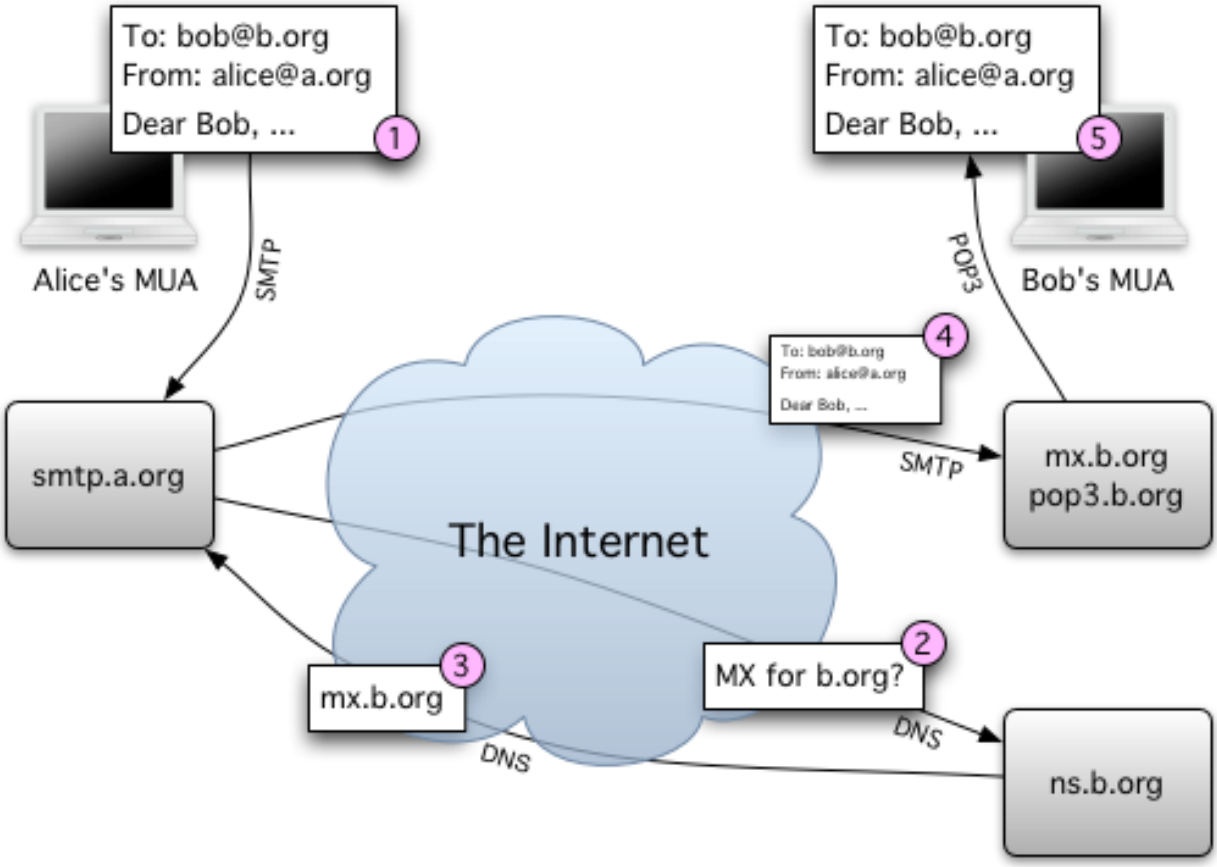
Obsah

1 Úvod

1.1 Elektronická pošta

Elektronická pošta (skrátený názov e-mail alebo email z angl. electronic mail) je metóda na tvorbu, posielanie, ukladanie a prijímanie správ cez elektronický komunikačný systém. Výraz email označuje dve veci. V prvom rade Internetový systém doručovania založený na protokole SMTP (Simple Mail Transfer Protocol) a taktiež intranetový systém umožňujúci používateľom jednej organizácie posielanie správ ostatným členom danej organizácie. Často sa na komunikáciu pomocou správ v rámci organizácie používa Internetový protokol namiesto intranetového systému a tak „email“ časom nadobudol význam obidvoch pojmov.

1.2 Ako pracuje email



Obrázok 1: Schéma udalostí pri posielaní a doručovaní emailu

Typický sled udalostí, emailovej komunikácie (príklad Alica píše email Bobovi) podľa obrázku 1:

1. Alica na vytvorenie emailu používa emailového klienta - Mail User Agent (ďalej ako MUA). Aby každý článok komunikácie vedel kde má smerovať je nutné už pri tvorbe emailu zadať položku „To“, ktorá určuje adresáta emailu. Taktiež sa zadáva aj položka „From“, aby adresát vedel určiť odosielateľa. Samozrejmou súčasťou emailu je aj správa, ktorú používateľ posiela príjemcovi emailu. Po vyplnení položiek emailu a odoslaní emailu používateľom sa emailový klient spojí protokolom SMTP so SMTP serverom, ktorý má uložený vo svojej konfigurácii a odošle email serveru.
2. SMTP server sa u prijatého emailu rozhoduje podľa hlavičky emailu. V hlavičke emailu nájde položku „To“ a podľa jej hodnoty sa rozhodne o ďalšom spracovaní emailu. V prípade, že SMTP server nie je emailovým serverom adresáta, server sa snaží doručiť email ďalej až k serveru adresáta. Aby mohol server túto akciu vykonať potrebuje vedieť kto je emailovým serverom adresáta. V tejto chvíli sa dostáva ku slovu DNS protokol. SMTP server vygeneruje požiadavku na Name Server adresátovej domény so žiadosťou o MX (emailový) záznam pre doménu adresáta.
3. Name Server je server, ktorý má informácie o doménových a subdoménových menách a taktiež aj o emailovom serveri danej domény (domén), ktorú spravuje. V prípade požiadavky poskytuje informácie o menách a IP adresách požadovaných doménových mien, čo sa stane aj v tomto prípade.
4. Po získaní názvu emailového servera adresáta SMTP server pošle SMTP protokolom email na daný server. Typicky sa email doručuje prostredím Internetu. Cestou Internetom môže dôjsť k spracovaniu emailu MTA servermi, ktoré email posielajú k emailovému serveru adresáta, čiže cieľovému serveru. V prípade, že po doručení emailu na cieľový server adresát emailu existuje, server uloží email do odkladacieho priestoru adresáta.
5. Adresát (Bob) používa emailového klienta, ktorý sa v pravidelných intervaloch pripája protokolom POP3 na emailový server. V prípade uloženého mailu na serveri, stiahne daný email na lokálny počítač a sprístupní ho adresátovi.

1.3 Súčasný stav emailovej komunikácie

Analýza súčasnej situácie vychádza z prieskumu Regentskej univerzity, ktorá sa zaoberala Internetom vo svojej práci *How Much Information? 2003*. Uvedená práca cituje spoločnosť International Data Corporation, zaoberajúcu sa marketingovými prieskumami. Táto spoločnosť odhadla objem priemernej dennej celkovej emailovej komunikácie na rok 2006 na 60 biliónov emailov denne. V prieskume sa uvádza aj spoločnosť Forrester Research, ktorá odhadla priemernú veľkosť emailu na 59KB. Vychádzajúc z uvedených čísel vychádza denne viac ako 300TB emailovej celosvetovej komunikácie. Uvedené čísla poukazujú na vysoký stupeň využívania emailov zo strany používateľov Internetu. Na uvedené počínanie používateľov reagovali aj poskytovatelia prístupu na Internet, ktorý pre svojich klientov zriaďovali a ešte zriaďujú emailové servery. Celosvetový počet mailových serverov je dnes vyše 1,5 milióna. Tabuľka 1 [15] ukazuje v akom pomere sú aktuálne (november 2006) zastúpené jednotlivé druhy emailových serverov.

Typ emailového servera	Počet serverov	Percentuálny podiel
Sendmail	297 343	32.90%
Microsoft Exchange	187 955	20.79%
Exim	158 381	17.52%
Postfix	115 219	12.75%
IMail	46 791	5.18%
MDaemon	21 071	2.33%
MailEnable	18 169	2.01%
ostatné	58 929	6.50%

Tabuľka 1: Rozdelenie používaných mailových serverov vo svete podľa typu

1.4 Porovnanie emailových serverov

V nasledujúcom porovnaní sa budem zaoberať iba prvými štyrmi emailovými servermi uvedených v tabuľke 1, nakoľko majú spoločne takmer 85% majoritný podiel na celkovom trhu emailových serverov.

Každý z porovnávaných serverov má svoje výhody, vie spracovať veľké množstvo emailov, spolupracovať s rôznymi databázami, má svoju dokumentáciu a možno aj vlastnosti, ktoré konkurenčný softvér nemá.

Celkovo bude porovnanie zamerané na nasledujúce kritéria:

- jednoduchosť administrácie
- bezpečnosť
- výkon
- ostatné

Všetky porovnania sa odkazujú na výsledky uvedené na obrázku 2 [17] a v tabuľke 2 [18]. Obrázok 2 ukazuje výsledky vybraných serverov v základnej konfigurácii (t.j. bez integrovaného antivírusového systému) a taktiež v konfigurácii s antivírusovým systémov na rôznych súborových systémoch. Tabuľka 2 obsahuje výsledky konkrétneho aplikačného testu.

Sendmail

- jednoduchosť administrácie – výhodou je jeden konfiguračný súbor, ktorý je však na druhej strane veľmi zložitý. Na konfiguráciu sa používa pomerne zložitý makrojazyk M4.
- bezpečnosť – od počiatku existencie Sendmailu, roku 1981, dodnes bolo v Sendmaily objavených niekoľko desiatok závažných chýb, ktoré sú archivované v rôznych databázach chýb po celom svete ako sú National Vulnerability Database alebo Open Source Vulnerability Database.
- výkon – v základnej konfigurácii (bez antivírusu) je spomedzi testovaných OpenSource serverov má relatívne najnižšiu priepustnosť emailov
- ostatné – z historického hľadiska bol Sendmail prednastaveným emailovým serverom vo viacerých Linuxových distribúciách a aj vďaka tomuto faktu je a ešte je veľmi nasadzovaný. K výhodám Sendmailu patrí efektívna implementácia tzv. „*milter*“ filtrov, voľne prekladaných ako „mailových filtrov“, ktoré sa využívajú ako rozhranie pre softvér tretích strán a taktiež port na Windows. Nevýhodou je monolitickosť servera.

Microsoft Exchange

- jednoduchosť administrácie – používateľsky prívetivé grafické rozhranie, od verzie 2007 dostupné aj skriptovacie rozhranie Windows PowerShell
- bezpečnosť – databázy chýb softvéru archivujú niekoľko desiatok bezpečnostných chýb. Microsoft však pravidelne vydáva updaty na tento svoj produkt.

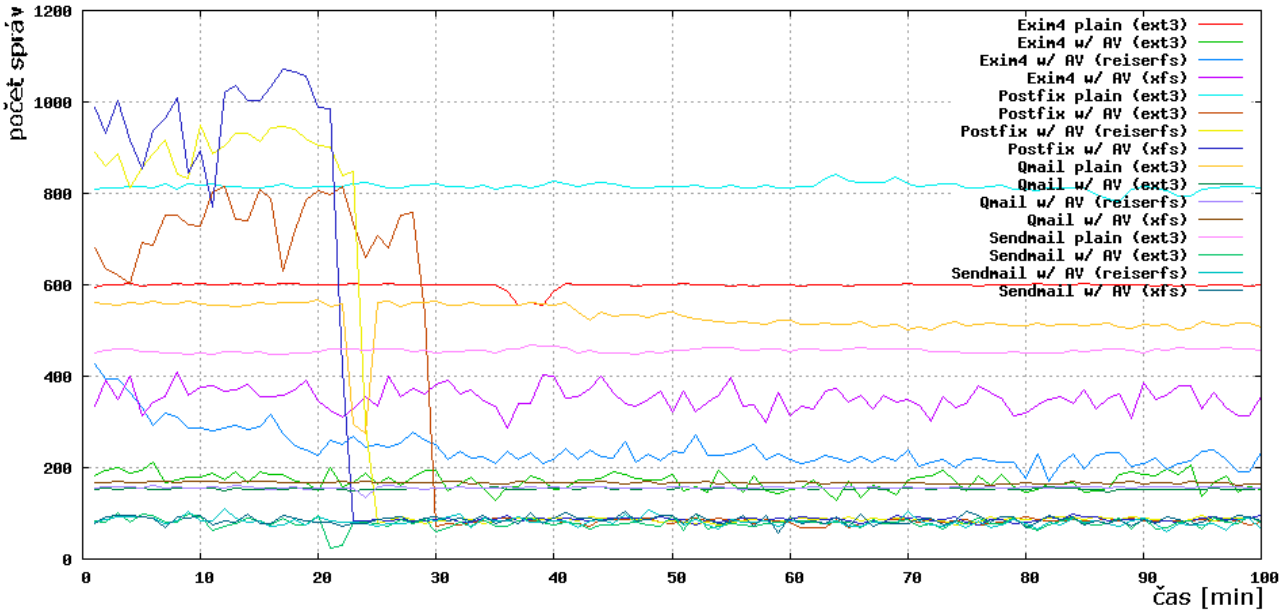
- výkon – podľa dostupných oficiálnych zdrojov [19] je jeho emailová priepustnosť porovnateľná s ostatnými emailovými servermi
- ostatné – Exchange je možné nasadiť ako groupwarové riešenie pre firmy. V spolupráci s ďalšími produktmi Microsoftu vie „tlačiť“ emaily aj na bezdrôtové prístroje siete GSM, tzv. „BlackBerry“. Nevýhodou je uzavretý zdrojový kód a proprietárna licencia k produktu. Formát ukladania emailov je iba do databázy, avšak má implementovanú funkciu „*single instance storage*“ [16], ktorý ukladá do databázy jednu správu určenú viacerým adresátom iba raz.

Exim

- jednoduchosť administrácie – jednoduchá, obsahuje iba jeden konfiguračný súbor
- bezpečnosť – necelá desiatka chýb hodnotených ako vysoko nebezpečné svedčia o celkovo dobrej bezpečnosti
- výkon – z testovaných serverov sa umiestnil zhruba v strede, vyniká svojou časovou stabilitou pri spolupráci s antivírusovým systémom.
- ostatné – Exim má otvorený zdrojový kód je možné ho spustiť aj pod operačným systémom Windows použitím Cygwin emulátoru. Pri správnom nastavení je možné volať perl interpreter priamo z konfiguračného súboru. Exim má aj vlastný filtrovací jazyk. Nevýhoda je monolitickosť tohto emailového servera.

Postfix

- jednoduchosť administrácie – priemerná, obsahuje niekoľko konfiguračných súborov, z ktorých sa však edituje zvyčajne iba jeden
- bezpečnosť – počet bezpečnostných od vzniku tohto servera sa pohybuje niečo mierne nad 10, takže je všeobecne považovaný za bezpečný
- výkon – vo väčšine kategórií mu patri prvenstvo a taktiež aj v celkovom hodnotení je najvýkonnejší, t.j. má najväčšiu priepustnosť emailov
- ostatné – používatelia oceňujú najmä jednoduchosť a bezpečnosť, ktorú autor matematik Wietse Venema, dosiahol modulárnosťou servera na rozdiel od zaužívaných monolitických riešení. Nevýhodou je neexistencia verzie na OS Windows.



Obrázok 2: Priebeh priepustnosti vybraných OpenSource emailových serverov na rôznych súborových systémoch a s rôznymi konfiguráciami

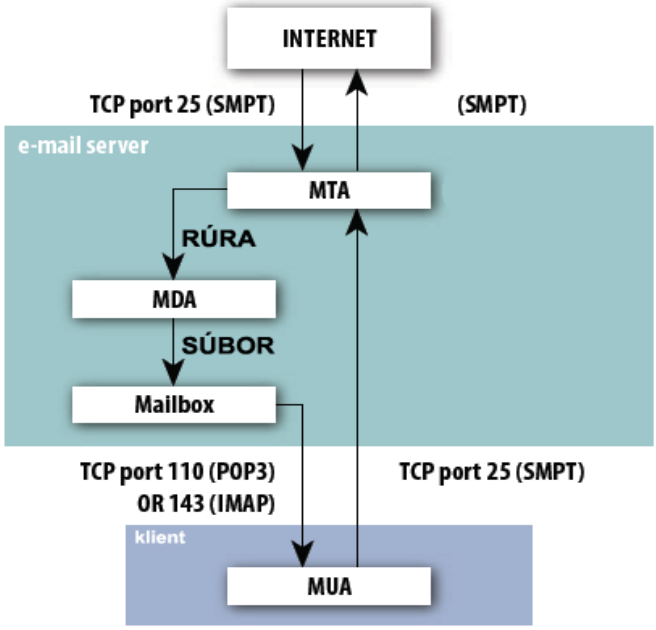
server	trvanie spojenia [s]	trvanie doručenia [s]	sync zápis	async zápis	priepustnosť [mail/s]	poznámka
1000 emailov pre 1 adresáta s 20 vláknami, 50 emailov na spojenie						
exim	56	262	6,621	6,785	3.8	manual queue run unsafe
postfix	49	62	1,898	4,110	16.1	-
sendmail	136	136	8,172	7,993	7.3	no softupdates FreeBSD 4.4-RC

Tabuľka 2: Testovanie vybraných OpenSource emailových serverov

2 Analýza problémovej oblasti

2.1 Emailový server

Vzhľadom k výkonu a vzhľadom k všetkým spomenutým vlastnostiam emailového servera Postfix, vrátane modularity, otvoreného zdrojového kódu a jednoduchej konfigurácii sa v ďalších častiach budem podrobne zaoberať emailovým serverom Postfix. Vnútornú štruktúru Postfixu, ktorá platí všeobecne pre väčšinu emailových serverov na operačnom systéme UNIX, znázorňuje obrázok 3. Vždy je prítomný Mail Transfer Agent (ďalej označovaný skratkou MTA), ktorý zabezpečuje SMTP komunikáciu smerom do a z Internetu a k emailovým klientom. MTA počúva na TCP porte číslo 25, ktorý je určený na SMTP komunikáciu.

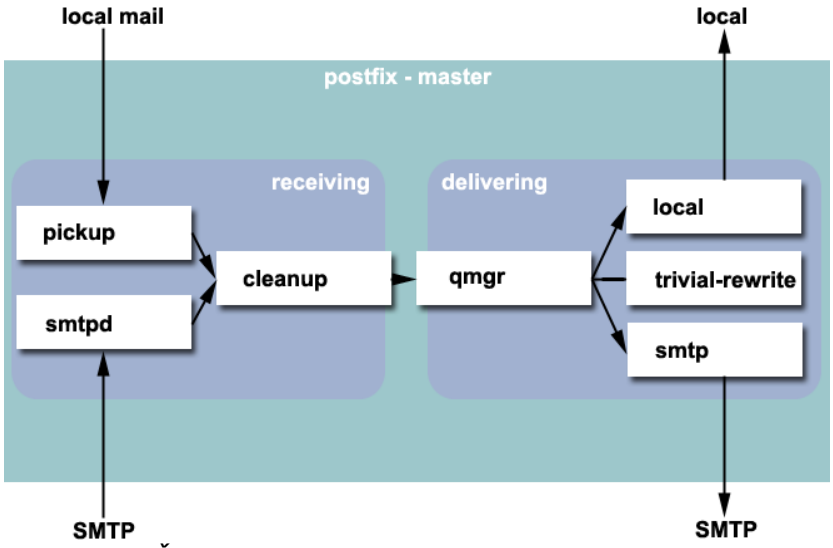


Obrázok 3: Schéma emailového systému na OS UNIX

Ďalšou súčasťou Postfixu je Mail Delivery Agent (ďalej označovaný skratkou MDA), ktorého úlohou je doručovanie emailov do lokálnych emailových adresárov, prípadne do mailboxov používateľov. Z MTA zvyčajne získava emaily cez dátovod (rúru) a jeho výstupom je už konkrétny súbor v emailovom adresári. Postfix má v konfiguračnom súbore direktívu, ktorá umožnuje použitie použitie externého MDA, čo bude pravdepodobne užitočná vlastnosť pre moju ďalšiu prácu.

2.2 Emailový server Postfix

Postfix je implementovaný ako jeden hlavný server (proces "*master*"), ktorý spúšťa obslužné démony vykonávajúce špecifické operácie podľa potreby. Jednotlivé služby sa spúšťajú ako samostatné procesy a zabezpečujú napr. posielanie a prijímanie mailov z Internetu, doručovanie lokálnych mailov atď. Počet procesov, ktoré daný typ požiadavky obsluhujú, je určený konfiguráciou.



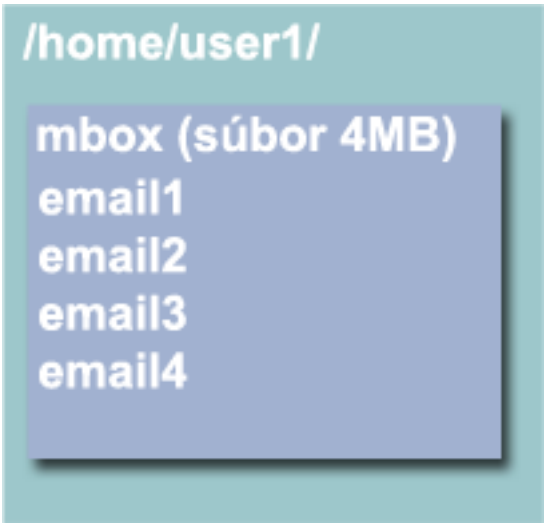
Obrázok 4: Členenie a spolupráca modulov v serveri Postfix

Stručný popis funkcie démonov servera Postfix podľa obrázku 4:

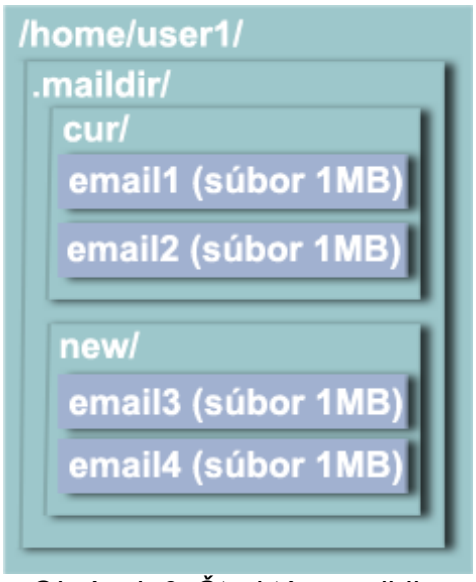
- *smtpd*: počúva na porte a prijíma SMTP požiadavky. Všetky prijaté správy sú presmerované na démona "*cleanup*"
- *pickup*: čaká na lokálne napísané emaily a presmeruje ich obsah na démona "*cleanup*"
- *cleanup*: spracúva prijatý mail (pridáva chýbajúce hlavičky a pod.), vkladá ho do fronty prijatých emailov a informuje démona "qmgr" o jeho príchode
- *qmgr:* čaká na prijaté emaily a zabezpečuje ich doručenie. Spôsob doručenia určí démon "*trivial-rewrite*"
- *trivial-rewrite*: prepisuje adresu do štandardizovanej formy. Démon pripája meno domény k lokálnym emailom bez jej uvedenia a pod. Okrem toho určuje, čo sa stane s emailom, ako a kam sa bude doručovať na základe adresy
- *local*: doručuje mail do lokálnych schránok na serveri
- *smtp*: smtp klient Postfixu. Doručuje maily z mailovej fronty, ktoré sú určené pre iné mailové servery.

2.3 Spôsoby ukladania emailov

Ako už bolo spomenuté všetky emaily sa používateľom ukladajú na súborový systém, buď do mailboxu alebo do maildiru. Rozdielna štruktúra mailboxu a maildiru je načrtnutá na obrázkoch 5 a 6. Ďalšia možnosť je využitie databázy ako úložisko mailov. Výhoda tohto riešenia je v jednoduchej manažovateľnosti a flexibilnosti, avšak celý emailový systém je plne závislý od vlastností a spoľahlivosti použitej databázy.



Obrázok 5: Štruktúra mailboxu



Obrázok 6: Štruktúra maildiru

V mailboxe sa emaily ukladajú sekvenčne v jednom súbore. Nevýhoda tohto spôsobu archivácie emailov sa objaví pri veľkom počte uložených emailov, keď klientský program má vyhľadať jeden konkrétny email. Čas trvania operácie bude priamo úmerný veľkosti celého mailboxu. Túto spomenutú nevýhodu eliminuje maildir tým, že pri ukladaní využíva hierarchické vlastnosti súborového systému. Konkrétne pri maildir-e sa jednotlivé emaily ukladajú do samostatných súborov, ktoré sa členia do adresárovej štruktúry. Výhodou tohto riešenia je, že pri vyhľadávaní a ostatných operáciách sa neprehľadáva obsah jedného súboru, ale prehľadáva sa adresárová štruktúra. K výhodám maildirov a taktiež mailboxov je možné ešte priradiť skutočnosť, že používateľ si môže vytvárať vlastné adresáre v rámci už existujúcej adresárovej štruktúry podľa svojho vlastného uváženia a podľa svojich potrieb. Táto vlastnosť sa v plnej miere využíva pri prístupe k pošte protokolom IMAP.

2.4 Efektívnosť využívania diskovej kapacity serverom Postfix

Na reálne zistenie ako efektívne využíva mailový server Postfix diskovú kapacitu som uskutočnil jednoduchý test. Na počítači typu PC s nainštalovaným OS Linux som vytvoril 10 nových testovacích používateľov posttest1 až posttest10. OS zabezpečoval iba služby nevyhnutné pre beh mailového servera a ostatné služby neboli spustené. V tomto stave som poslal všetkým testovacím používateľom mail obsahujúci 10MB textových ASCII znakov. V rámci dostupných nástrojov súborového systému som porovnal využitie kapacity súborového systému, kde sa ukladali maily pred testom a po teste. Rozdiel o veľkosti 100MB zodpovedal, došlým 10 emailom, každý o veľkosti málo väčšej než 10MB. Pri podrobnejšej analýze som určil skutočnosť, že všetky emaily sú identické až na pár riadkov v hlavičke mailu. Na obrázkoch 7 a 8 sú hlavičky emailov 2 používateľov, rozdiely sú podčiarknuté.

```
Return-Path: <root@vlk.ynet.sk>
X-Original-To: posttest1@vlk.ynet.sk
Delivered-To: posttest1@vlk.ynet.sk
Received: by vlk.ynet.sk (Postfix, from userid 0)
 id C936D8CD9D; Mon, 11 Nov 2006 15:16:23 +0100 (CET)
Received: from localhost (localhost [127.0.0.1])
 by vlk.ynet.sk (Postfix) with ESMTP id C7BF62C2A;
 Mon, 11 Dec 2006 15:16:23 +0100 (CET)
Date: Mon, 11 Nov 2006 15:16:23 +0100 (CET)
From: root <root@vlk.ynet.sk>
To: undisclosed-recipients: ;
...
```

Obrázok 7: Začiatok hlavičky emailu doručeného používateľovi posttest1

```
Return-Path: <root@vlk.ynet.sk>
X-Original-To: posttest2@vlk.ynet.sk
Delivered-To: posttest2@vlk.ynet.sk
Received: by vlk.ynet.sk (Postfix, from userid 0)
 id C936D8CD9D; Mon, 11 Nov 2006 15:16:23 +0100 (CET)
Received: from localhost (localhost [127.0.0.1])
 by vlk.ynet.sk (Postfix) with ESMTP id C7BF62C2A;
 Mon, 11 Dec 2006 15:16:23 +0100 (CET)
Date: Mon, 11 Nov 2006 15:16:23 +0100 (CET)
From: root <root@vlk.ynet.sk>
To: undisclosed-recipients: ;
...
```

Obrázok 8: Začiatok hlavičky emailu doručeného používateľovi posttest2

Pri teste bol použitý jeden identický obsah emailu, ktorý sa doručil fyzicky do mailového adresára každému používateľovi. Z pohľadu využitia diskovej kapacity je to zbytočné plytvanie miestom, nakoľko sa dva prípadne viac súborov s rovnakým obsahom ukladá viacnásobne na súborový systém.

2.5 Špecifikácia požiadaviek

Požiadavky na operačný systém:

- kompatibilný s optimalizovaným emailovým serverom Postfix

Požiadavky na navrhovaný optimalizačný modul:

- spoľahlivosť – každý email, ktorý dostane modul na spracovanie, úspešne spracovaný, v prípade chybového stavu bude tento stav oznámený predchádzajúcemu modulu, respektívne bude v prípade fatálnej chyby vygenerovaný email odosielateľovi o príčine nedoručenia emailu
- bezpečnosť – modul neumožní žiadny neoprávnený prístup k emailovej komunikácii používateľom, ktorí nemajú na túto činnosť oprávnenie
- efektívnosť využitia diskovej kapacity – hlavný dôraz je kladený na túto vlastnosť modulu
- flexibilnosť – každý server pracuje v iných podmienkach, ktorým je potrebné sa prispôsobiť, čo zabezpečia konfiguračné direktívy
- otvorený zdrojový kód – zabezpečí ďalší možný rozvoj modulu pre potreby možno aj ďalších emailových serverov

3 Návrh riešenia

3.1 Možné spôsoby riešenia

Vyššie popísaný test poukázal na problém neefektívneho obsadzovania diskovej kapacity emailovým serverom Postfix. Na riešenie uvedeného problému je vhodné využiť vlastnosť používaných súborových systémov, ktoré umožňujú pristupovať k jednému fyzicky uloženému súboru aj z viacerých miest v súborovom systéme. Konkrétnou a užitočnou vlastnosťou sú linky. Hlavná myšlienka riešenia: pri lokálnom doručovaní zistiť, či už daný mail nebol v určitom časovom rámci doručený a ak bol, aktuálnemu adresátovi vytvoriť v mailovom adresári iba link na už raz uložený mail.

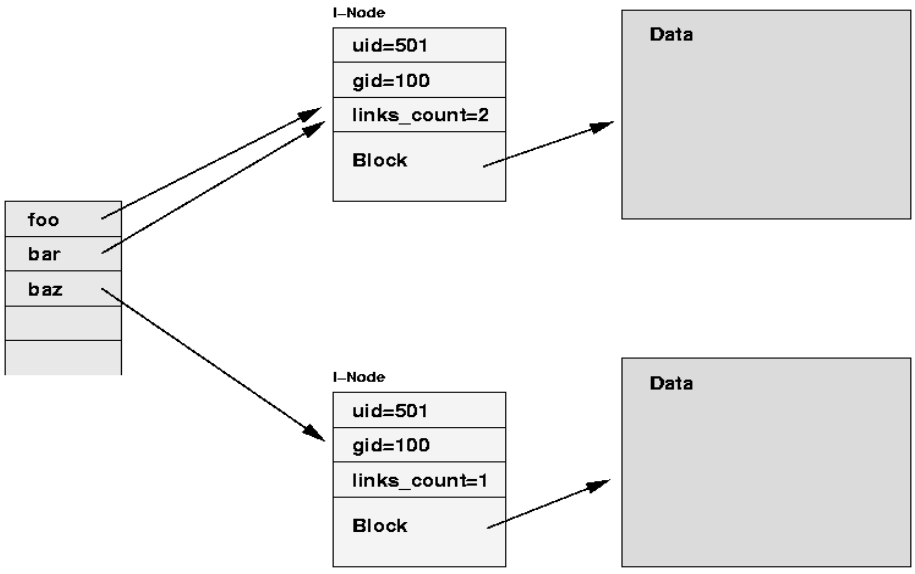
3.2 Linky - hardlink a symlink, iné riešenia

Rozoznávame dva druhy linkov. Hardlinky a symlinky. Každý druh ma svoje výhody aj nevýhody.

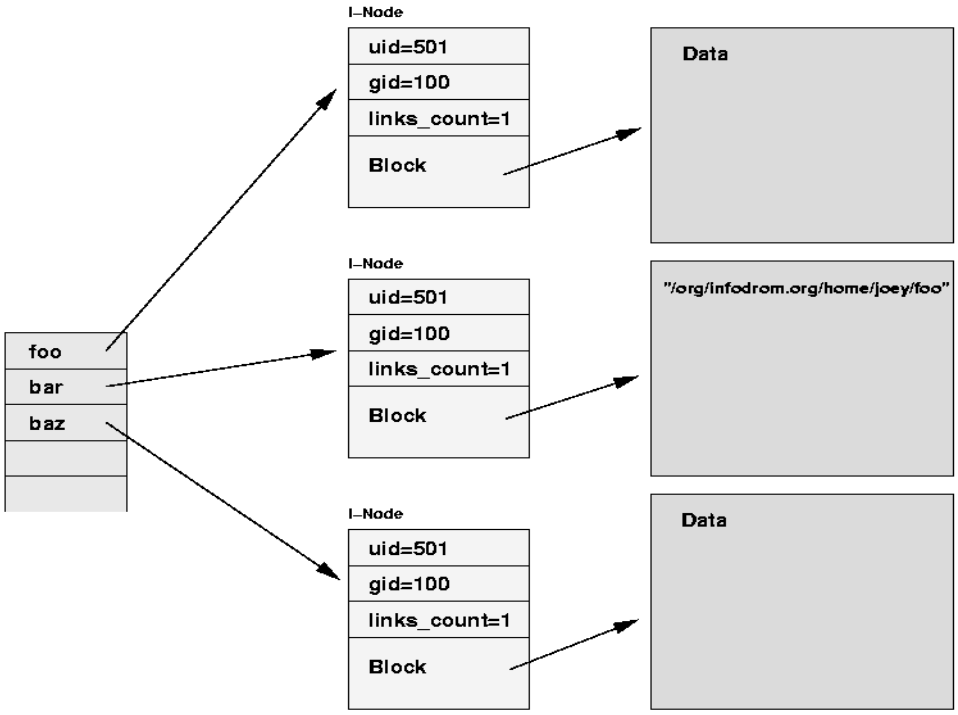
Hardlink je označenie situácie kedy jedno číslo inodu je odkazované viacerými menami súboru. Tieto mená sa môžu nachádzať v tom istom adresári, alebo v rôznych adresároch, ale musia sa nachádzať na tom istom súborovom systéme. Pri takomto usporiadaní sa zmena obsahu súboru alebo jeho atribútov (prístupové práva, čas, veľkosť, vlastník, …) prejaví na všetkých miestach, kde je dané číslo inodu odkazované. Pristupovanie k súboru cez ľubovoľné meno je rovnocenné. Výhodou hardlinkov je skutočnosť, že priamo v inode na odkazovaný súbor existuje počítadlo (reference counter), ktorý zaznamenáva počet existujúcich hardlinkov na daný súbor.

Symlink (tiež nazývaný softlink) je situácia, kedy adresár miesto toho, aby názvu súboru priradil číslo inodu, priradzuje mu meno iného súboru. Výhodou symlinku (v porovnaní s hardlinkom) je to, že môže prekračovať hranice súborového systému. Nevýhodou je to, že odkazovaný súbor nenesie žiadnu informáciu o tom, že naň niekto odkazuje. Ak je cieľový súbor zmazaný, dostávame neplatný symlink. Prístupové práva sa overujú vzhľadom na cieľový súbor.

Praktický prípad linkov aj s hodnotami *reference counter* (prípadne sa nazývajú *links_count*) znázorňujú obrázky 9 a 10.

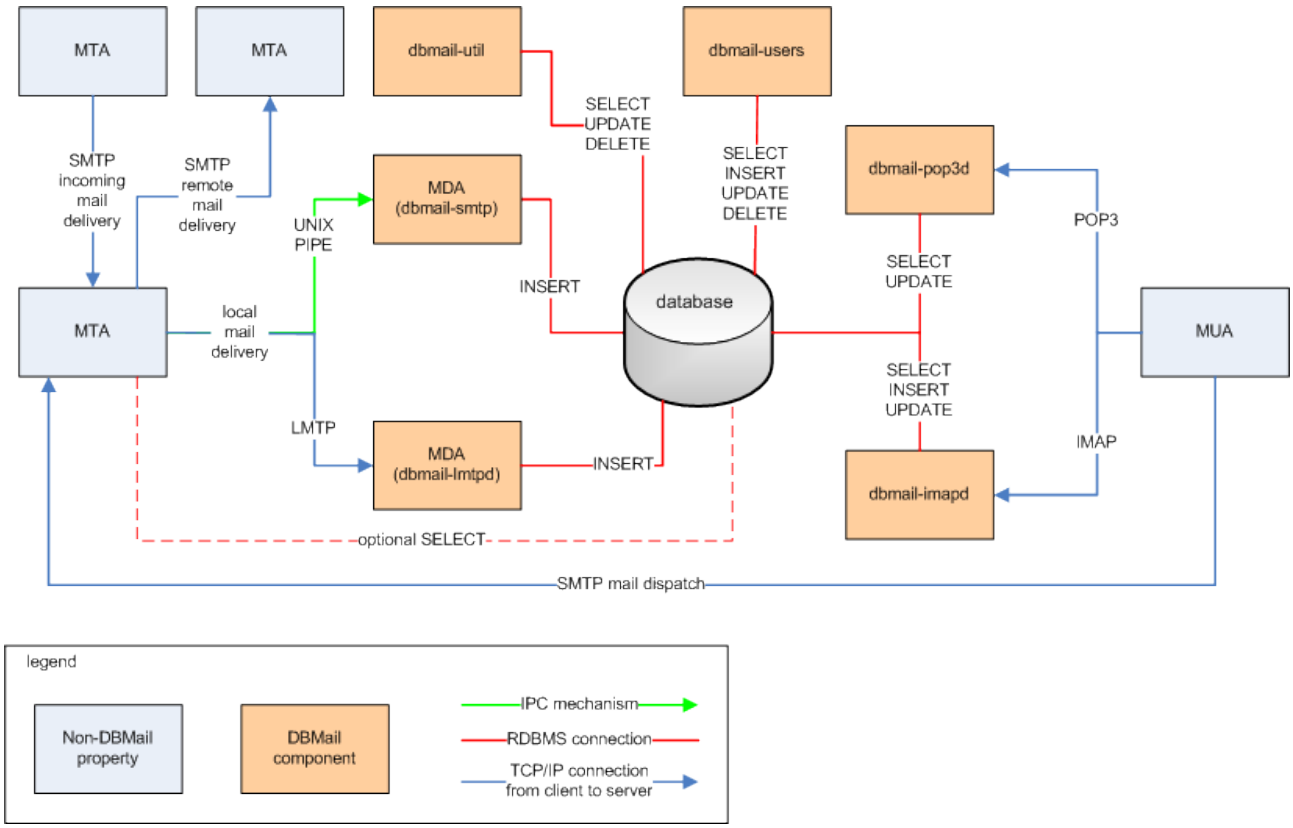


Obrázok 9: Princíp hardlinkov



Obrázok 10: Princíp symlinkov

Ďalší možný spôsob riešenia efektívneho využívania diskového priestoru je využitie databázy ako úložiska emailov. Výhodou je škálovateľnosť, manažovateľnosť, rýchlosť, flexibilita. Nevýhoda je v samotnej databáze, ktorá je ďalším elementom v celom procese a v porovnaní s ukladaním do súboru sa pri práci s databázou spotrebuje väčšie množstvo výpočtového výkonu. Tento výpočtový výkon je potrebný na databázový manažment. Použitím databázy sa celý emailový systém stane zložitejším a aj náchylnejším na možné chyby. Napriek spomenutým nevýhodám sa databáza využíva aj v praxi. Samotný emailový server *Microsoft Exchange* využíva databázu na odkladanie emailov. Z nekomerčných verzií je najznámejším balík programov pod menom *DBMail*, ktorého štruktúra je zobrazená na obrázku 11. Nespornou nevýhodou koncepcie použitia databázy je nutnosť použitia osobitných démonov služieb (IMAP, POP3), ktorí vedia spolupracovať s emailami v databáze. Bežne používaní démoni nemajú zvyčajne implementované databázové funkcie a tak balík *DBMail* obsahuje vlastné implementácie imap a pop3d démonov.



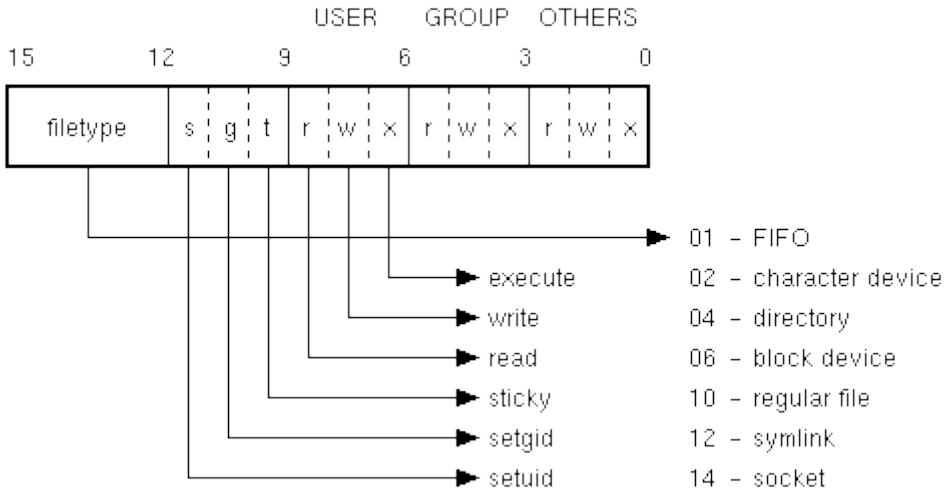
Obrázok 11: Schéma a koncepcia DBMail

3.3 Prístupové práva

Jeden mail, môže byť doručený niekoľkým používateľom. Napríklad jedna pracovná skupina zamestnancov dostane mailom pracovnú úlohu. Navrhovaný mailový systém ošetrí vzniknutú situáciu uložením jednej kópie mailu prvému adresátovi a u ostatných adresátoch detektne existenciu dotyčného mailu a v mailových adresároch ostatných adresátov vytvorí iba linku na už uložený email. Problém vyvstane s akými právami uložiť email resp. linku. Problém linky je vyriešený nakoľko ona samotná ukazuje na mail, čiže práva sa nevzťahujú na linku, ale priamo na linkovaný súbor - email.

Triviálnym riešením je využitie štandardných bitových práv. (vlastník, skupina, ostatní - čítanie/zápis/vykonávanie). Význam jednotlivých bitov a členenie práv je znázornený na obrázku 12. Nevýhodou tohto riešenia, je situácia, keď všetci adresáti emailu sú používatelia z jednej skupiny avšak existujú aj ďalší členovia tejto skupiny, ktorý tento súbor (email) nedostali a tak im prístupové práva neprináležia. Správne nastavenie bitových práv by v tomto prípade mohli byť vyriešené vytvorením 2 nových skupín

(skupina, ktorá dostala email a skupina, ktorá nedostala email) a následným správnym priradením jednotlivých členov pracovnej skupiny do novovzniknutých skupín. Na túto akciu sú však potrebné práva administrátora systému a je zrejmé z tohto prístupu neustále vznikanie nových skupín. Počet skupín v systéme je stanovený OS a bežne sa pre 32 bitové OS pohybuje na úrovni 65 536, čo je pre veľké emailové servery nepostačujúce.



Obrázok 12: Bitové práva a význam jednotlivých bitov

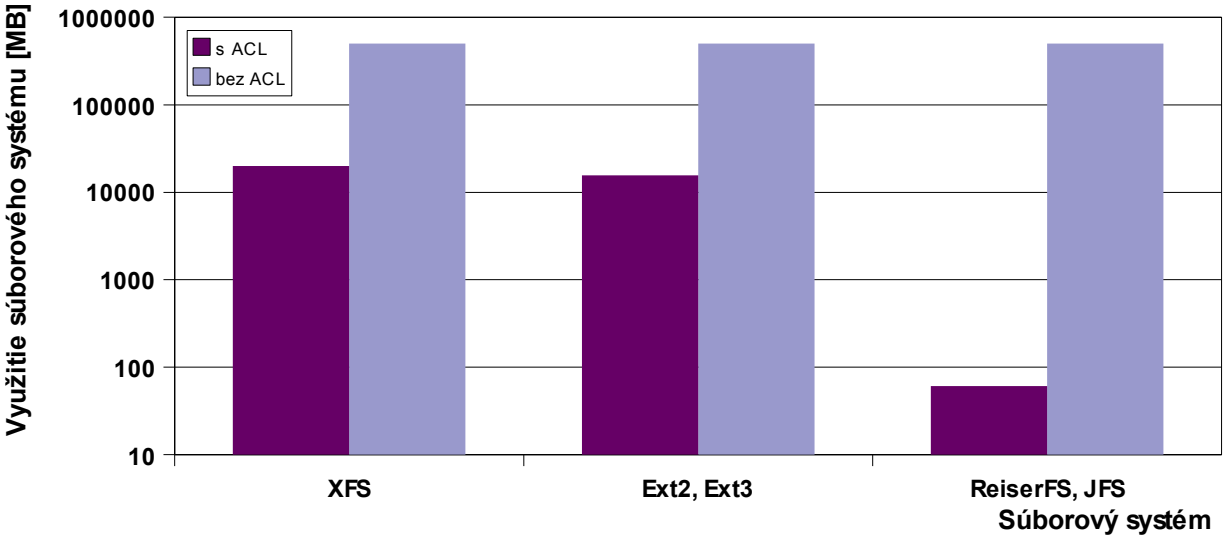
Aby sa zbytočne nevytváral veľký počet skupín, len na zabezpečenie prístupových práv k emailom, je namieste použitie POSIX ACL (Acces Control List). ACL umožňujú vymenovať prístupové práva pre jednotlivých používateľov ku konkrétnemu súboru alebo adresáru. Obmedzenia na počet ACL záznamov sú závislé na použitom type súborového systému ako ukazuje tabuľka 3.

Súborový systém	Max. počet ACL záznamov
XFS	25
Ext2, Ext3	32 (môže byť aj viac – určené veľkosťou bloku)
ReiserFS, JFS	8191

Tabuľka 3: Maximálny počet ACL záznamov v závislosti od súborového systému

Všeobecné obmedzenie na počet ACL záznamov akýmkoľvek súborov systéme s podporou ACL je možné jednoducho obísť uložením každého x-tého. emailu, kde x je rovné limit súborového systému na počet ACL+1. Následné došlé identické emaily sa budú linkovať na naposledy uložený email. Aj pri súborovom systéme s najmenším počtom

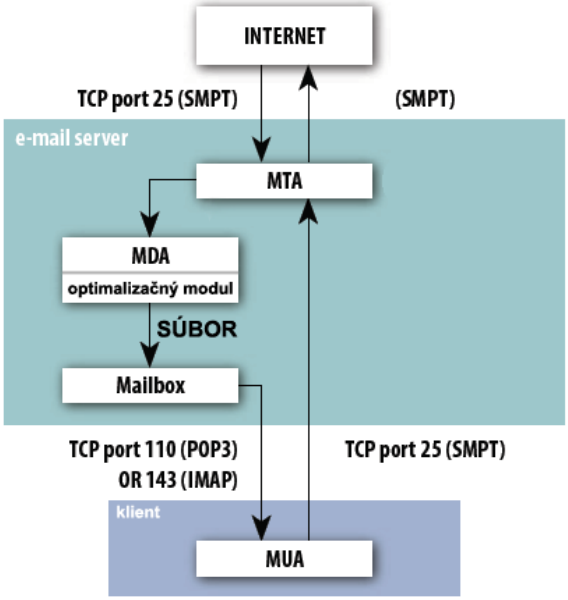
možných uložených ACL záznamov (XFS) sa ušetrí diskový priestor o veľkosti 24 krát veľkosť dotyčného emailu, čo predstavuje značné ušetrenie priestoru. Grafickú reprezentácia ušetrenej diskovej kapacity pri jednotlivých súborových systémoch za predpokladu 100 000 identických 5MB emailov, pričom veľkosť linkov vzhľadom na veľkosť emailu zanedbávame, je znázornené na obrázku 13.



Obrázok 13: Využitie súborového systému pri optimalizovanom ukladaní emailov s použitím a bez použitia ACL

3.4 Integrácia s emailovým serverom Postfix

Postfix zabezpečujúci doručenie emailu do používateľského mailového adresára nie je navrhnutý na optimalizáciu emailovej komunikácie pomocou linkov a tak je nutné optimalizačný modul umiestniť na koniec spracovávania, tesne za MDA, ako je znázornené na obrázku 14. Pri integrácii pred MDA by nastal problém ako sa má správať MDA keď dostane duplikát už existujúceho emailu a taktiež ako mu túto skutočnosť oznámiť.



Obrázok 14: Integrácia optimalizačného modulu do Postfixu

Po analýze problému navrhujem riešenie s nasledovnými požiadavkami:

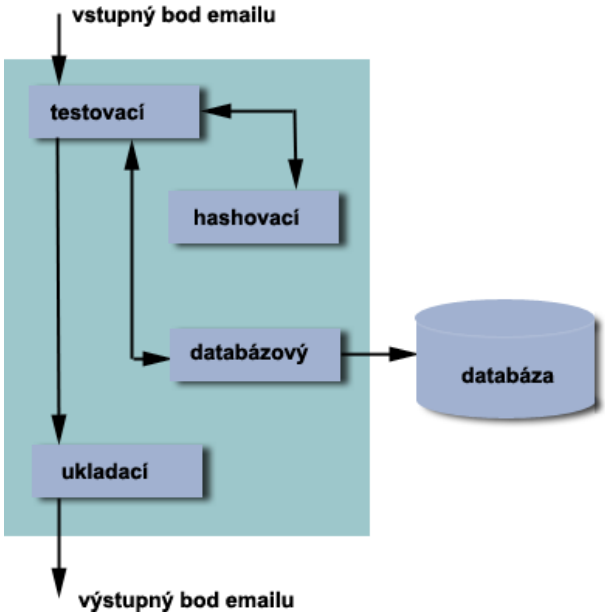
- operačný systém kompatibilný s Postfix
- úložisko mailov maildir
- súborový systém s implementovaným ACL - ext2/3, ReiserFS, XFS, JFS

Operačný systém je vybraný v závislosti od požiadaviek Postfixu, na ktorý budem vytvárať optimalizačný modul. Maildir je požadovaný vzhľadom k tomu, že na samostatný email v mailbox-e nie je možné aplikovať osobitné ACL práva rozdielne od ostatných uložených emailov. Z analýzy vyplýva aj nutnosť ACL pre správne zabezpečenie prístupových práv jednotlivým používateľom, nakoľko bitové práva majú nevhodné obmedzenia, spomenuté vyššie pri analýze problémovej oblasti.

3.5 Vnútorné členenie optimalizačného modulu

Nasledujú obrázok 15 zachytáva navrhované vnútorné členenie optimalizačného modulu. Testovací modul prijme zo vstupu email a na jeho dátovú časť zavolá hashovaciu funkciu, ktorú si používateľ zvolí v konfiguračnom súbore optimalizačného modulu. Informácie potrebné na zistenie duplicitnosti emailu (hash, veľkosť tela správy, časový údaj) sa budú uchovávať v databáze. Navrhnutá štruktúra tabuľky je popísaná v tabuľke 4. Databáza má oproti ukladaniu údajov do súboru výhodu predovšetkým v rýchlom prehľadávaní údajov, čo bude mať priamy vplyv na rýchlosť rozhodovania a celkového

času spracovania emailu celým modulom. V súčasnosti dokáže spolupracovať Postfix s viacerými typmi databáz. V záujme zachovania maximálnej flexibility bude obsahovať aj navrhovaný modul databázové rozhranie pre bežne používané databázy ako MySql, PostqreSQL a iné. Podľa získaných hodnôt sa testovací modul rozhodne o uložení emailu do štandardného súborového tvaru v maildire alebo uložení emailu iba ako linku. Následne posunie informáciu čo a ako uložiť ukladaciemu modulu na uloženie.



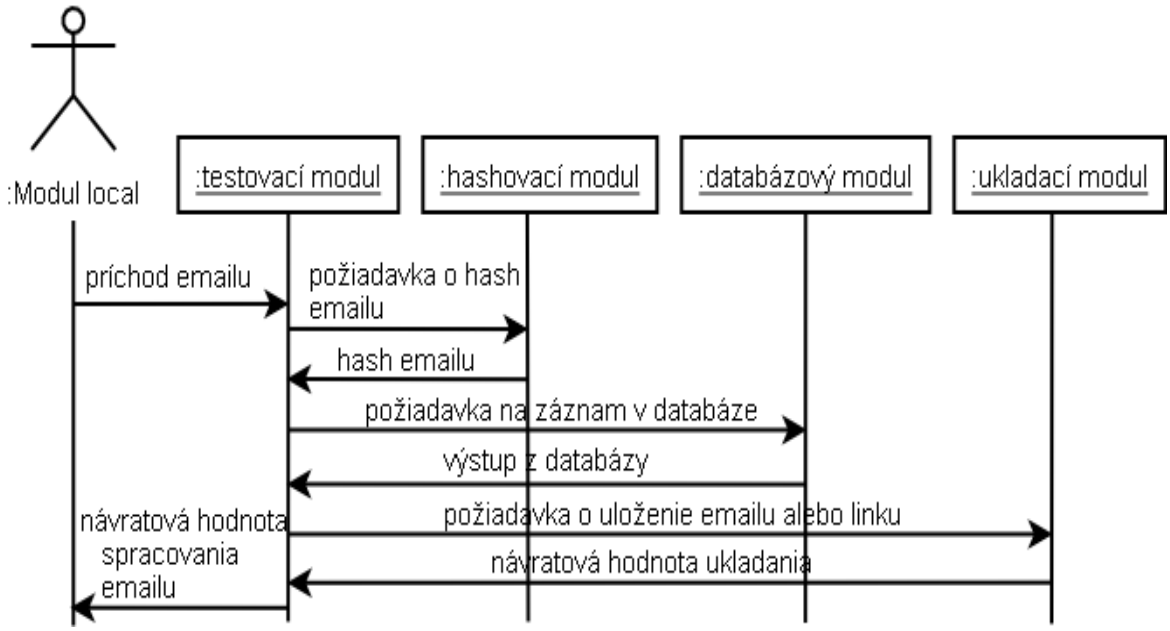
Obrázok 15: Vnútorné členenie optimalizačného modulu

Pri navrhovanom optimalizovanom linkovanom ukladaní emailov vzniká problém s informáciami, ktoré používateľ môže mať prístupné, aj keď by k nim nemal mať žiaden prístup. Z obrázkov 7 a 8 je vidieť, že v prípade linkovania emailu používateľa posttest2 na email používateľa posttest1, tento vidí v hlavičke *X-Original-To a Delivered-To* s údajmi, ktoré neprináležia jeho osobe. Z pohľadu používateľa sú tieto informácie nevýznamné a využívajú sa iba pri doručovaní emailu. Z tohto dôvodu je ich možné vymazať, o čo sa bude starať ukladací modul.

Položka tabuľky	Poznámka
email_id	**primárny kľúč tabuľky**
email_hash	údaj o hash hodnote emailu
email_hidden_flag	indikátor hovoriaci o tom, či má email skrytú hlavičku
email_timestamp	časová pečiatka príchodu emailu

Tabuľka 4: Štruktúra tabuľky pre záznam údajov pre optimalizačný modul

Obrázok 16 prestavuje sekvenciu postupností pri typickom spracovaní emailu navrhovaným optimalizačným modulom.



Obrázok 16: Sekvenčný diagram navrhovaného optimalizačného modulu

4 Technická dokumentácia

4.1 Implementácia optimalizačného modulu

Optimalizačný modul je implementovaný v jazyku C, nakoľko väčšina sotvéru pod OS Linux je napísaná v tomto jazyku.

Základná štruktúra, s ktorou pracujú jednotlivé funkcie implementácie je email. Podľa charakteru pracujú jednotlivé funkcie s prílušnými položkami nasledovnej štruktúry:

```
typedef struct {
        char *head;         //hlavička emailu
        char *body;         //telo emailu
        char *hash;         //hash tela emailu
        char *to;           //meno lokálneho užívateľa
        char *homedir;      //domovský adresár
        char *filepath;     //cesta spolu s názvom email súboru
        int to_uid;         //uid lokálneho užívateľa
        int done;           //flag,  spracovany
        long id;            //id záznamu v databáze
        long inode;         //číslo inodu
} email;
```

Ďalšie použité štruktúry:

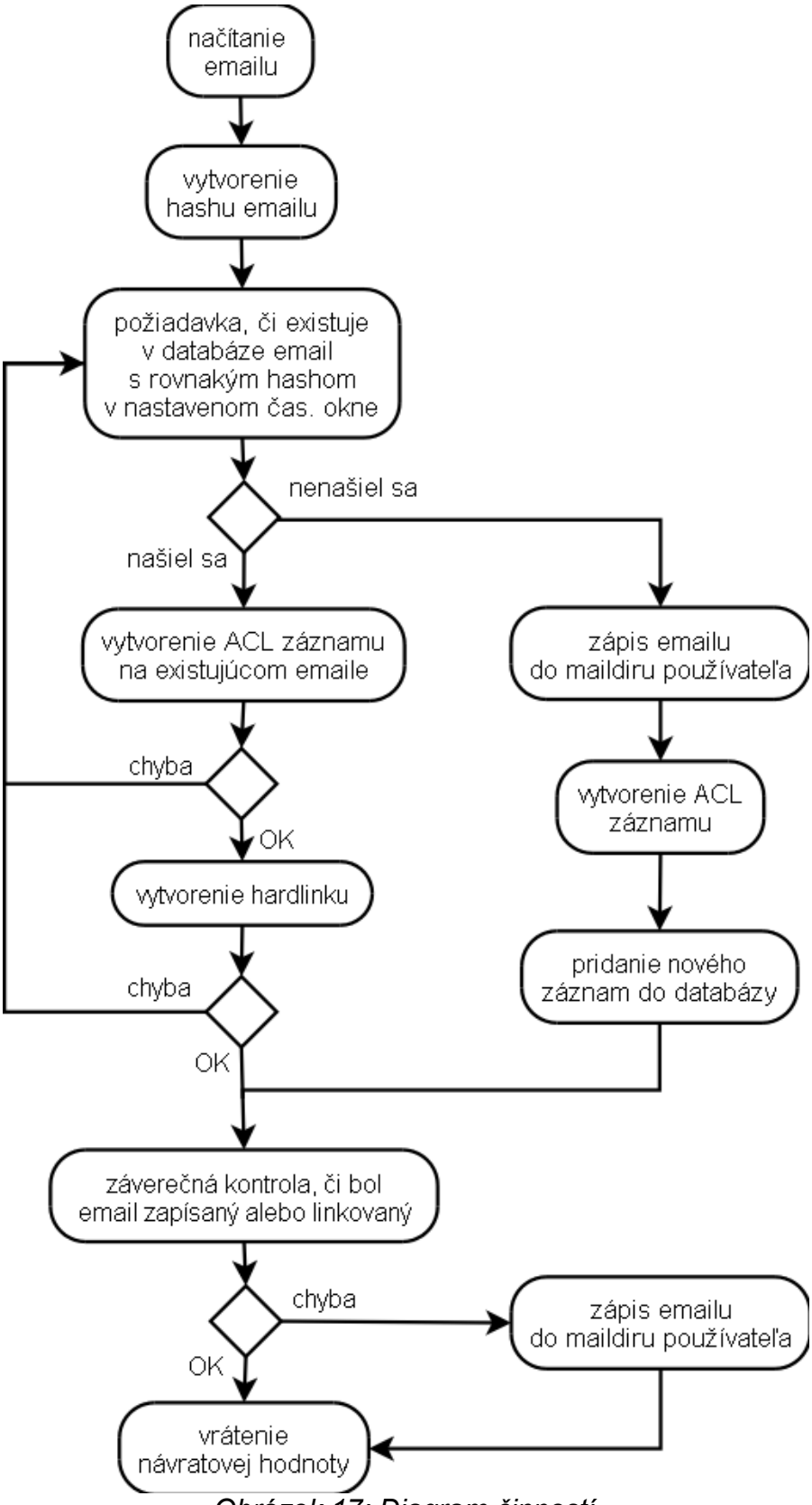
- config – konfiguračné údaje
- SHA256_CTX – štruktúra využívaná v hashovacej funkcii

Ako bolo spomenuté, na identifikáciu emailu sa využíva hash tela emailu. S veľkou váhou na bezpečnostné hľadisko bol vybraný hashovací algoritmus SHA256, nakoľko v súčasnosti neexistujú algoritmi, ktorými by bolo možné vytvoriť pre daný algoritmus kolízie.

Databázová vrstva je vytvorená knižnicou libdbi, ktorá umožnuje relatívne jednoduchou zmenou v konfiguačnom súbore zmeniť typ databázy.

Samotný démon je navrhnutý ako služba, ktorá počúva na predefinovanom porte a po príchode nového emailu sa najprv vytvorí nový proces, ktorý email spracuje.

4.2 Diagram činností



Obrázok 17: Diagram činností

5 Zhodnotenie

Projekt v aktuálnom stave analyzuje problémovú oblasť a navrhuje riešenia. Popri riešeniach mapuje taktiež problémy navrhnutých riešení. Ďalšia práca projektu zahrňuje štúdium možností modulu do samotného Postfixu, prípadne štúdia samotnej aplikácie, ktorej by sa odovzdávali emaily na konečné doručenie. Optimalizácia by v tom prípade bola externá aplikácia funkčnými vlastnosťami podobná procmailu.

6 Slovník pojmov problémovej oblasti

SMTP – Simple Mail Transfer Protocol
Protokol doručujúci a prenášajúci emaily. Špecifikuje ho RFC0821.

POP3 – Post Office Protocol Version 3
Protokol na výber emailov z mailboxu. Špecifikuje ho RFC1939.

IMAP – Internet Message Access Protocol
Protokol na sprístupňujúci emaily v mailboxe. Špecifikuje ho RFC2060.

DNS – Domain Name System
Systém umožňujúci preklad doménových adries na IP adresy.

ISP – Internet Service Provider
Poskytovateľ internetového pripojenia.

MUA – Mail User Agent
Program spracujúci emaily (MS Outlook, Mozilla Mail, Eudora, atď.) , ktorý používateľovi zjednodušuje prístup a spracovanie emailov (napr. čítanie, vytváranie, tlač, triedenie atď.).

MTA – Mail Transfer Agent

Program (inštancia Sendmail, Postfix, Qmail, Exim, atď.) prijímajúci emailové správy z lokálnej a/alebo vzdialených domén, ktoré potom preposiela ďalej. Všeobecne sa hovorí, že MTA je agent prenášajúci maily k ďalším MTA a/alebo MUA.

MDA – Mail Delivery Agent

Program (maildrop, procmail, deliver, local.mail, atď.) doručujúci email do používateľského mailboxu.

MAILBOX

Súbor alebo súborová štruktúra na disku, slúžiaca na ukladanie emailov.

Poznámka: Existuje niekoľko formátov mailboxov v Linuxových OS.

ACL – Acces Control List

súčasť rozšírených atribútov súborového systému umožňuje definovať prístupové práva pre jednotlivých používateľov systému

inod

dátová štruktúra používaná v súborovom systéme, ktorá popisuje jeden súbor. Nesie informácie o tom, aké sú prístupové práva k danému súboru, kedy došlo k poslednej zmene, k prístupu k súboru, kto je jeho vlastníkom, vlastniacou skupinou, počet hardliniek a predovšetkým informácie o tom, kde sa v súborovom systéme nachádza obsah súboru.

7 Zoznam použitej literatúry

[1] Kyle D. Dent: *Postfix: The Definitive Guide*, O'Reilly ISBN: 0-596-00212-2

[2] Craig Hunt: *sendmail Cookbook*, O'Reilly ISBN: 0-596-00471-0

[3] Eset, spol. s r. o. 2006. *NOD32 for Linux Server* http://u4.eset.com/manuals/nod_ug_linux_en.pdf

[4] Postel, Jonathan B. 1982. *SIMPLE MAIL TRANSFER PROTOCOL (RFC 821)* http://www.ietf.org/rfc/rfc0821.txt

[5] Myers, J., Rose, M. 1996. *Post Office Protocol - Version 3 (RFC 1939)* http://www.ietf.org/rfc/rfc1939.txt

[6] Crispin, M. 1996. *INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1 (RFC 2060)* http://www.ietf.org/rfc/rfc2060.txt

[7] Emery, Van. 2005. *Using ACLs with Fedora Core 2 (Linux Kernel 2.6.5)* (nov. 2006) http://www.vanemery.com/Linux/ACL/linux-acl.html

[8] Grünbacher, Andreas. 2003. *POSIX Access Control Lists on Linux* (nov. 2006) http://www.suse.de/~agruen/acl/linux-acls/online/

[9] Tuchbreiter, Jochen. 2005. mailing list acl-devel, „*Maximum number of acls on directory on ext3: 28? (bs does not matter, 1024/2048/4096 tested)*“ (nov. 2006) http://acl.bestbits.at/pipermail/acl-devel/2005-January/001829.html

[10] domovská stránka mailového servera *Postfix* (nov. 2006) http://www.postfix.org

[11] domovská stránka *Procmail* (nov. 2006) http://www.procmail.org

[12] domovská stránka *Spamassassin* (nov. 2006) http://spamassassin.apache.org

[13] Regents of the University of California. 2003. *How Much Information? 2003* (nov. 2006) http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm

[14] Shearer, Dan . *MTA Comparison* (nov. 2006) http://shearer.org/MTA_Comparison

[15] *Internet Research Reports 2006 Mail (MX) Server Survey* (nov. 2006) http://www.securityspace.com/s_survey/data/man.200610/mxsurvey.html

[16] Knotek, M. prepis online diskusie *Zabezpečení operačních systémů* (dec. 2006) http://www.microsoft.com/cze/technet/chat/chats/chat20061004.mspx

[17] Rako, Korunić, Dobrenić: *Testiranje performansi OpenSource SMTP poslužitelja* (dec. 2006) http://dkorunic.net/pdf2/MTA-testiranje.pdf

[18] Matthias Andree: MTA Benchmark (dec. 2006) http://www-dt.e-technik.uni-dortmund.de/~ma/postfix/bench2.html

[19] HP *ProLiant DL360 G4 server benchmark results for Microsoft Exchange Server 2003* (dec. 2006) http://h71019.www7.hp.com/ActiveAnswers/downloads/BenchmarkResults_Microsoft_Exchange2003onProLiantDL360G4_34proc.pdf