Česká společnost uživatelů otevřených systémů EurOpen.CZ
Czech Open System Users' Group
www.europen.cz

35. konference

Sborník příspěvků

Lesní zámeček STAR 4&5
Klínovec
4.–7. října 2009

Programový výbor

Sborník příspěvků z 35. konference EurOpen.CZ, 4.–7. října 2009

Plzeň 2009. První vydání.
Editor: Vladimír Rudolf, Jiří Sitera
Jiří Felbáb
Sazba a grafická úprava: Ing. Miloš Brejcha – Vydavatelský servis, Plzeň
e-mail: servis@vydavatelskyservis.cz

Tisk: Typos, tiskařské závody, s. r. o.
Podnikatelská 1160/14, Plzeň

Příspěvky neprošly redakční ani jazykovou úpravou.

ISBN 978-80-86583-17-4

Obsah

Praktická správa barev v Linuxu

Martin Sivák

E-mail: msivak@redhat.com

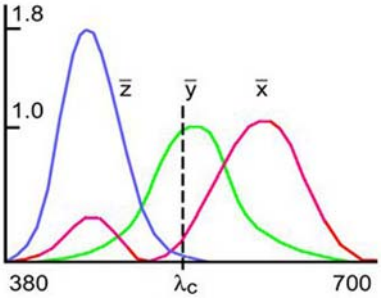
Dnes, kdy se většina podkladů připravuje digitálně je často zapotřebí řešit barevnou shodu předlohy s výstupem. Ať už se jedná o fotografie, reklamní materiály nebo prezentaci, nepřesné barvy mohou negativně ovlivnit pocity, které taková práce vyvolá. Proto je pro prosazení linuxu do grafických odvětví nutné se zabývat i problémem kalibrace barev a podporou příslušných nástrojů.

Trocha teorie

Barva

Co je to vlastně barva a jak ji vnímáme?

Nejdůležitějším principem, který si potřebujeme uvědomit je, že světlo je elektromagnetické vlnění. Základními faktory, které ovlivňují barevnost předmětu jsou potom jeho povrchové vlastnosti a světlo na něj dopadající ze zdroje. Ze světla jsou odfiltrovány složky, které daný materiál pohlcuje a zbytek je odražen a může dopadnout do našeho oka. Lidské oko je neuvěřitelně geniální orgán, který vnímá elektromagnetické vlnění v délkách cca 400 nm až 700 nm. Naše černobílé vidění zařizují tyčinky, které vnímají pouze intenzitu vlnění okolo 500 nm, ale barevné vidění vyžaduje o něco nápaditější „vybavení“. Podobně jako dnes naše počítače, oko vnímá tři základní barvy. Čípky v oku jsou totiž trojího druhu – červené, zelené a modré. Každý čípek vnímá vlnění v určitém okolí své základní barvy a shrneme-li všechny tyto vjemy, získáme naše barevné vidění.



Proč je toto důležité? Na tomto principu fungují i nejpřesnější přístroje pro měření barev – spektrometry. Jsou schopny změřit intenzitu jednotlivých složek světla (s rozlišovací schopností na desítky až jednotky nm) a po aplikaci předpokládané citlivosti čípků v našem oku získají hodnotu zvanou tristimulus (označovanou jako XYZ). A tento barevný systém se používá pro jednoznačnou identifikaci barvy. Důležité je, že barva, kterou nakonec vnímáme, je určena kromě vlastností objektu i světlem, které na něj dopadá. Zajímat by Vás mohla ještě jedna informace, a to že náš mozek podle všeho nevnímá jednotlivé barvy, ale kromě jasu vnímá vztah mezi červenou a zelenou a vztah mezi modrou a žlutou (složenou z červené a zelené). Proto nikdy neuvidíte modro-žlutou nebo červeno-zelenou barvu. Tato maličkost se občas hodí, když potřebujete změnit barevné vyznění dokumentu. Mimo jiné se těmto barevným osám také říká LAB a jelikož mají stejnou schopnost přesně určit barvu jako systém XYZ, tak často se používají místo něj.

Osvětlení a „pamatovací barvy“

Když už tedy víme jak oko vnímá, nemělo by nás překvapit, že dvě různé barvy mohou pod stejným osvětlením vypadat stejně. Představte si například červenou a zelenou desku pod modrou lampou. Našemu oku se budou obě jevit černé. Tento jev se nazývá metamerizmus. Pokud ovšem nebude barva osvětlení tak výrazná a položíte pod ní běžný kancelářský papír, Váš mozek po chvíli usoudí, že vidí bílý papír. Existují totiž některé barvy, na které jsme velmi citliví a jsou pro nás tak důležité, že si mozek podle nich upraví vnímání okolní scény. Jedná se tedy o schopnost našeho mozku provést automatické vyvážení bílé. Mezi tyto barvy patří například již zmíněný bílý papír, barva lidské kůže, modrá barva čisté oblohy nebo zelená barva lesa. Když nebude na fotografii některá z těchto barev sedět ve vztahu k ostatním, budete takovou fotografii považovat za přinejmenším zvláštní. A naopak, pokud si dobrou fotografii zobrazíte na mírně „rozhašeném“ monitoru, ani si nevšimnete, že má barevnou odchylku – Vaše oko a mozek ji automaticky vykompenzovaly. Proto je dobré na oko při kalibracích moc nespoléhat a použít přístroje.

Barevné profily

Pokud chceme reprezentovat barvy v počítači, není ale výše zmíněný XYZ nebo LAB systém výhodný. Zabírá příliš mnoho místa a těžko se s ním pracuje. Proto se používají jiné systémy, přičemž nejvýznamnější z nich budou asi různé varianty založené na RGB a CMYK (a samozřejmě i z historických důvodů). Aby se ale neztratila informace o přesné barvě, kterou jednotlivé kombinace RGB reprezentují, používají se dnes „mapy barev“, kterým říkáme barevné profily. Profil kromě mapy také obsahuje informace o tom, jaká je nejsvětlejší (bílý bod) a nej-

tmavší použitelná barva, což je důležité zvláště pokud mají tyto body nějaký barevný nádech. Profily bychom mohli rozdělit do dvou skupin: První skupinou jsou profily nezávislé na zařízení, používané pro charakterizaci barevných prostorů na reprezentaci dat. Sem patří například nám známé sRGB, AdobeRGB nebo různé formy CMYKu (Fogra, ISO Coated a.j.). Do druhé skupiny patří profily, které popisují chování a vlastnosti nějakého zařízení – monitoru, tiskárny, scanneru nebo třeba projektoru. Každé zařízení totiž interpretuje hodnoty do něj posílané jinak a tak rgb: 100, 100, 100 bude jiná barva na monitoru a jiná barva na domácí tiskárně. Pokud ale máte správně sestavený profil (mapu, chcete-li) pro vstupní data i pro výstupní zařízení, tak dojde před odesláním na výstup k přepočítání hodnot tak, aby odpovídaly stejným hodnotám v systému XYZ nebo LAB a tedy i stejné barvě (nebo alespoň co nejbližší možné). Bohužel výstupní zařízení nejsou často schopná zobrazit všechny barvy, které bychom po nich chtěli, takže se pro hledání té správné výstupní barvy musí použít jistá aproximace. Setkat se můžete se třemi hlavními:

- Absolutní kolorimetrická – Použijí se přesně ty barvy, které umí zařízení použít a u barev, které zařízení neumí se nalezne nejbližší odstín, který už zařízení zvládne.

- Relativní kolorimetrická – Tato varianta se chová stejně jako Absolutní kolorimetrická, ale před hledáním barev ještě všechny posune tak, aby byl bílý bod podle zdrojového profilu zobrazen jako bílý bod podle nového profilu.

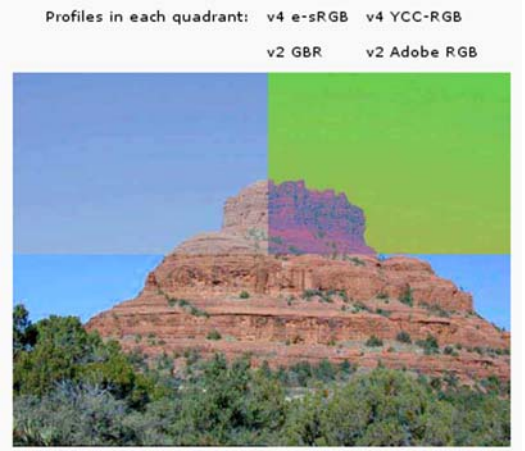
- Perceptuální – Pokusí se změnit barvy tak, aby se zachovaly jejich relativní vztahy a všechny byly stále rozlišeny, obvykle tedy dojde k jistému „sražení“. I když tato metoda způsobí nepřesnost v reprodukci konkrétních jednotlivých barev, jedná se o vhodný způsob například pro fotografie, kde záleží na celkovém vyznění více než na absolutní přesnosti.

Kalibrace vs. profil

I když v textu používáme slovo kalibrace, proces se ve skutečnosti skládá ze dvou částí. První částí je již zmíněná kalibrace tj. nastavení zařízení do stavu, kdy se chová k intenzitě barev pokud možno lineárně a pokaždé stejně. Druhou částí je poté změření tohoto chování a vytvoření charakteristiky neboli profilu (jak jsem již říkal – mapy) zařízení. Profil je poté samozřejmě platný a přesný pouze pokud je nastavení stejné, jaké bylo v době jeho vytvoření. U monitoru to znamená, že máte stejný jas, kontrast a nastavení bílé, ale také že podsvícení má pořád stejnou kvalitu. U tiskárny je těchto faktorů více – tiskárna, její nastavení, inkoust, papír, vlhkost, to všechno má vliv na přesnou reprodukci barev a platnost profilu.

Podpora správy barev v linuxu

Narozdíl od Mac OS X nebo MS Windows není podpora správy barev nedílnou součástí systému, ale je na aplikacích zda budou tuto funkci obsahovat. Potřebné nástroje obsahuje knihovna lcms a převést obrazová data mezi barevnými prostory je otázka asi tří řádků zdrojového kódu. I proto správu barev podporuje čím dál více aplikací – od editorů jako je Gimp, přes správce fotografií typu digiKam nebo bibble, až po jednoduché prohlížeče fotografií typu gqview nebo eye-of-gnome. Ověřit míru podpory konkrétní aplikace si můžete na stránkách ICC konsorcia: http://www.color.org/version4html.xalter.



Nejobvyklejšími nástroji pro tvorbu profilů, kalibraci a měření jsou kompletní balík ArgyllCMS s grafickým rozhraním dispcalGUI a profiler monitorů a scannerů LProf.

Pracovní prostředí a kalibrace monitoru

Kalibrace monitoru je základní předpoklad pro rozumnou práci s barvami. I pokud plánujete jen vystavovat fotografie na internetu nebo pracovat v oblasti webdesignu, je správně seřízený monitor a přesný profil velkou výhodou. Základními parametry, které si při tvorbě profilu pro displej volíme je barevná teplota bílého bodu a požadovaná hodnota pro gamma křivku. Doporučené hodnoty se liší podle prostředí, kde se počítač nachází:

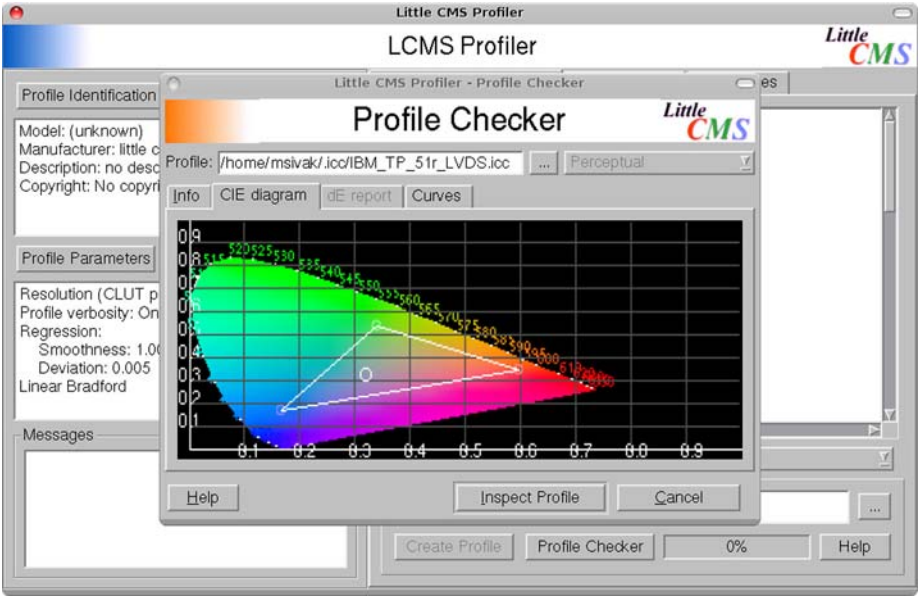
- Pro grafická studia a pracoviště s normovaným zářivkovým osvětlením jsou obvyklé hodnoty 5 000 K a gamma 2.2.
- Pro domácí počítače s běžným rozptýleným osvětlením se doporučuje teplota 6 500 K a gamma 2.2. Pokud si přesto nastavíte 5 000 K nezpůsobíte

tím žádný problém, jen si budete muset zvyknout na výrazně nažloutlé barvy, protože Vaše oko a Váš mozek jsou zvyklé (a za oknem vidí) na mnohem studenější teplotu jasného dne.

- Pro počítače bez možnosti upravit teplotu bílé na monitoru (většina LCD umí jen filtrovat barvu podsvícení) je lepší ponechat teplotu beze změny a gamma 2.2, aby byl zachován co největší rozsah barev. Například můj starý prezentační notebook má hardwarově bílý bod někde v oblasti kolem 6 000 K.

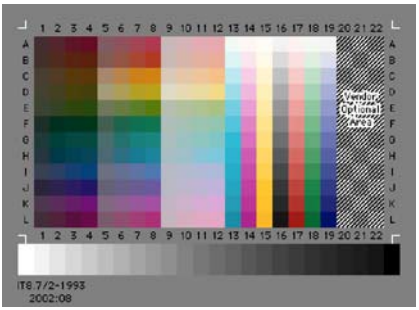
- Jedinou výjimkou pro nastavení gamma, na kterou můžete narazit, jsou některé počítače od Apple, které používaly hodnotu 1.8, protože kdysi nejvíce odpovídala chování k nim dodávané tiskárny a historicky se tato hodnota udržela.

V předchozím seznamu jsem také nakousl problematiku světla v pracovním prostředí. Nejlepší prostředí pro práci s barvami a prohlížení tisků je samozřejmě zatemněné studio s normovaným stropním osvětlením a neutrální barvou na stěnách. Tohoto v domácích podmínkách ovšem dosáhnete jen stěží, ale vhodnou neutrální barvou (bíla nebo světle šedá) na stěnách pracovny a umístěním displeje tak, aby se na něm nezrcadlilo okolí ani dopadající světlo, docílíte alespoň trochu použitelného stavu. Případně roleta a jednoduchá zářivka se známou barevnou teplotou a vysokým CRI (index podání barev; alespoň 90 %) zastanou pro domácí podmínky dostatečnou službu. Zařízeními schopnými změřit barevné vlastnosti monitoru jsou kolorimetry (měří obvykle právě jen tři složky tristimulu, ale jsou výrazně levnější) a spektrometry. Softwarová výbava již byla zmíněna – nejjednodušší je postup při použití programů dispcalGUI nebo LProf. Obvykle se skládá ze tří kroků – Před samotným měřením nechte monitor zahřát, aby veškeré podsvícení dosáhlo provozních podmínek (alespoň 30 minut). Po započetí kalibrace Vám zařízení ukazuje jas a barevnou teplotu a Vy přímo ovládacími prvky displeje upravíte jeho zobrazení tak, abyste dosáhli požadovaných hodnot jasu a barevné teploty. Následně necháte software přeměřit množství barevných čtverců a vypočítat ze získaných informací profil. Hotový profil musíme „nainstalovat“. DispcalGUI toto zařídí za nás, případně využijeme příkazu `dispcal -I profil.icc`, který nakopíruje soubor s profilem na správné místo, zapíše konfigurační soubor .config/color.jcnf a vytvoří spouštěcí dávku v .config/autostart (KDE i Gnome tento soubor respektují a použijí). Ruční zavedení zařídí příkaz `dispcal -L`. Dnešní moderní linuxová prostředí ovšem trpí jedním problémem. Spořiče obrazovky si obvykle pamatují stav před svým spuštěním a kalibrační informace nám z grafické karty vymažou. U některých se to dá řešit zavedením profilu ještě před jejich inicializací. Nejjistější je ovšem spořič úplně vypnout a spoléhat jen na správu napájení tj. na zhasnutí displeje.



Vstupní zařízení a kalibrace scanneru

Vytváření profilů vstupních zařízení a scannerů obzvláště patří mezi ty jednodušší postupy. Vše co potřebujete je kalibrační obrazec se známými hodnotami a vhodný software. Vytištěné obrazce IT8.7 i s daty získáte například u pana Daneše nebo si je můžete objednat z Německa (Wolf Faust). Také pokud vlastníte spektrometrickou sondu, již jste pravděpodobně výtisk dostali, nyní si ho musíte akorát přeměřit podle postupu popsaného u kalibrace tiskáren.



Pokud máte všechno připraveno, použijte některý z programů LProf, Argyll-CMS (zajímá Vás příkaz scanin) nebo VueScan (komerční program podporující mnoho scannerů).

Výstup a tisk

Tvorba profilu pro výstupní zařízení a speciálně pro tiskárny je tím nejsložitějším, čeho se můžete v oblasti správy barev dočkat. Papír totiž nemá žádný vlastní zdroj světla a tak je nutné měřit světlo odražené. Pokud si ještě vzpomínáte na úvodní teorii, použité světlo ovlivňuje hodnoty, které sonda vidí, a proto je nutné na její výrobu použít kvalitní a stálé součástky. To samozřejmě zvedá cenu. Obecně platí, že kolorimetry nejsou schopné tento typ měření provést a je třeba použít dražší spektrometrické sondy. Další možností je i ruční okometrická kalibrace. Tato metoda spočívá v metodě pokus a omyl a úpravách pokročilých vlastností tiskového procesu. Nezískáte takto profil, ale je možné, že se Vám podaří vyladit tisk tak, aby Vám výtisky připadaly věrohodné. Bude Vás to ovšem stát mnoho papíru i inkoustu, a proto nemohu tuto metodu doporučit a dále se budu věnovat pouze kalibraci s použitím sondy. Speciálně zde u kalibrace a profilů tiskárny je na místě zopakovat upozornění na nutnost mít stále stejné parametry tisku (tiskárna, inkoust, papír a nastavení) pro zachování přesnosti a použitelnosti profilu. Nyní již k samotné kalibraci. Z linuxových programů podporuje kalibraci tiskového procesu pouze ArgyllCMS. Tvorba profilu by se dala shrnout do několika málo kroků:

- Podle požadované kvality si necháme spočítat potřebné množství barevných vzorků a sestavíme z nich kalibrační obrazec. (targen a printtarg)
- Vypneme veškerou inteligenci tiskových ovladačů!
- Vypneme správu barev a vytiskneme kalibrační obrazec. Potom ho necháme několik hodin uschnout! (Barvy se prvních pár hodin změní a rozdíl činí až několik dE [dE spočítáme jako vzdálenost dvou XYZ bodů v 3D prostoru. Hodnota 3 a vyšší je již viditelná okem.]).
- Použijeme sondu a software pro změření spektrální odezvy jednotlivých barevných polí. (chartread)
- Software potom podle zvolené barevné teploty v místě pozorování (standardně $D50 = 5\,000\,\mathrm{K}$, ale pokud víme, že fotografie budou viset jinde, můžeme s tím počítat a profil vyrobit už upravený) vypočítá převodní tabulky do profilu. (colprof)

Profil pak nahrajeme na nějaké vhodné umístění a nastavíme ho v aplikacích, které podporují tiskové profily jako GIMP nebo photoprint.

Link profily

Balík ArgyllCMS obsahuje také nástroje potřebné pro automatizaci převodu dokumentů mezi profily. Umožní Vám pomocí řádkového programu collink vytvořit speciální typ profilu, který popisuje převod z jednoho barevného prostoru do druhého, aniž by používal mezivýpočet přes XYZ nebo LAB. Nástrojem cctiff také převedete TIFF soubory podle zadaných parametrů. Spojením těchto dvou možností poté může vzniknout například automatická služba na převod a uložení dokumentů.

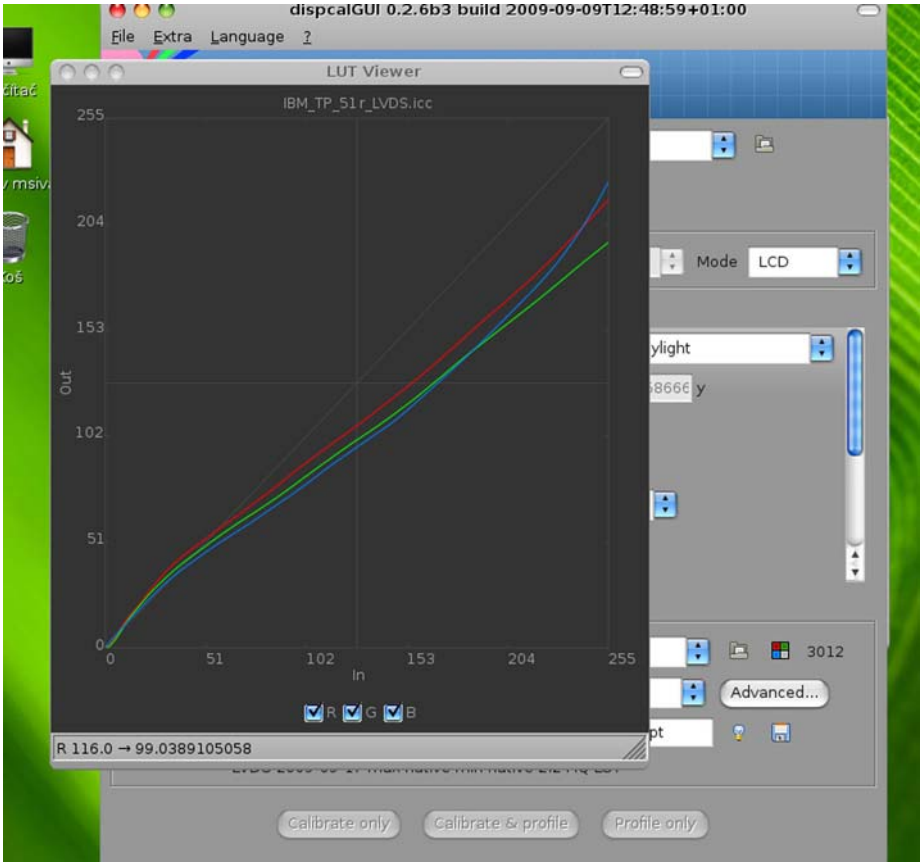
Kontrola profilů a ověření jejich přesnosti

Pro prvotní inspekci a kontrolu profilu můžeme využít dispcalGUI (Extra/Load LUT.. & Show LUT curves) nebo nástroje Profile Checker zabudovaného do již několikrát zmíněného LProf. Jednoduše vybereme profil, který si chceme prohlédnout a LProf nám zobrazí základní informace, korekční křivky a u maticových profilů i CIE diagram se znázorněním zobrazitelných barev.

Argyll také obsahuje nástroj iccgamut, který z ICC profilu vygeneruje soubor popisující reprodukovatelné barvy. Tento soubor (i více, pokud chceme porovnávat) je pak pomocí utility viewgam možné zkonvertovat do 3D prostoru, který si zobrazíme pomocí nástroje freewrl. A nakonec, pokud máte změřeny dvě předlohy (například původní a aktuální stav monitoru), tak nástroj verify z balíku ArgyllCMS Vám spočítá a ukáže jak moc jsou měření shodná a ve kterých barvách došlo k největšímu rozdílu.

Závěrem

Správa barev a problematika profilů je velmi rozsáhlé téma. Doufám, že jsem zmínil alespoň ty nejdůležitější problémy a nástroje, se kterými se můžete setkat v linuxu. Teoretické postupy a fyzikální základ platí stejně ve všech běžných systémech, a proto věřím, že jste si odnesli nejenom seznam několika málo užitečných příkazů.



Zdroje a odkazy

[1] Real-World Color Management; Bruce Fraser, Chris Murphy, Fred Bunting; Peachpit Press; 2005.

[2] ArgyllCMS – http://www.argyllcms.com/

[3] dispcalGUI – http://dispcalgui.hoech.net/

[4] LittleCMS – http://www.littlecms.com/

[5] LProf – http://lprof.sourceforge.net/

[6] Photoprint – http://blackfiveimaging.co.uk/index.php?article=02Software%2F01PhotoPrint

[7] terče IT8.7 od pana Daneše – http://www.danes-picta.com/

[8] terče IT8.7 od f. Wolf Faust – http://www.targets.coloraid.de/

[9] VueScan – http://www.hamrick.com/

Fotografické workflow v Linuxu

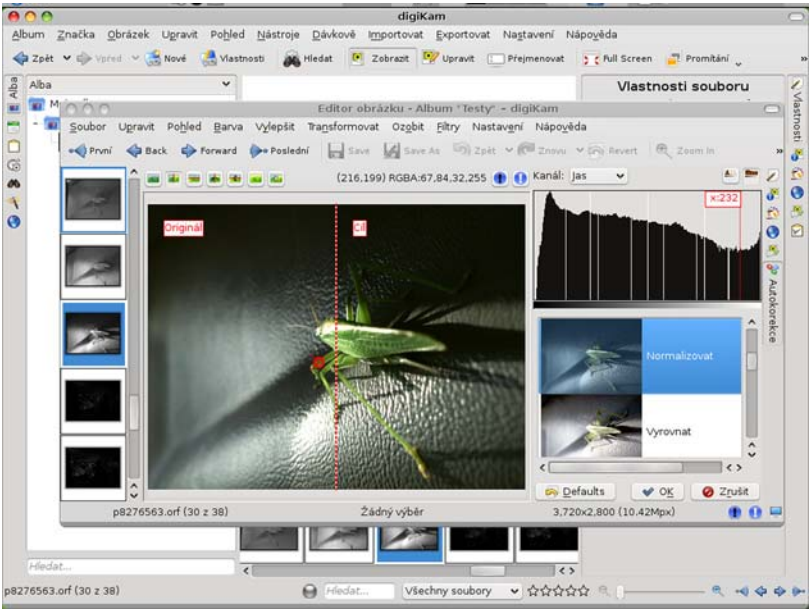
Martin Sivák

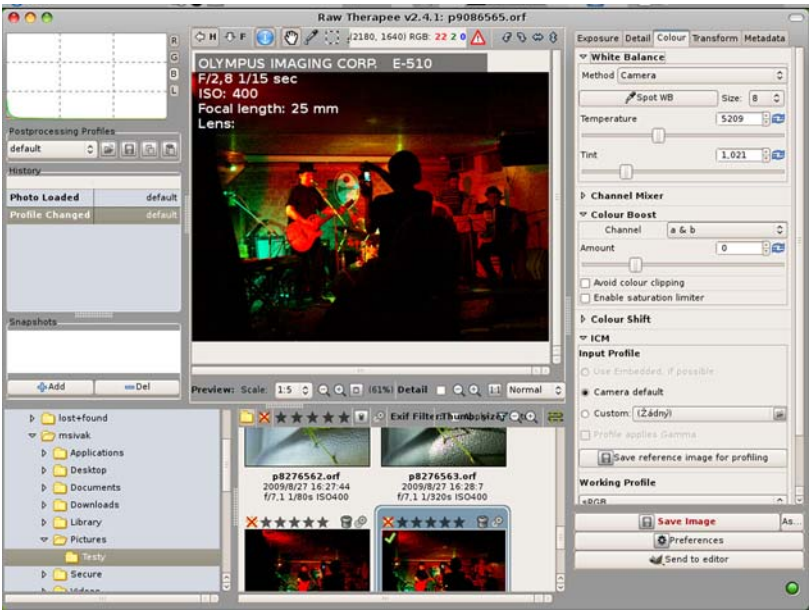
E-mail: msivak@redhat.com

Protože linuxové systémy ještě stále nejsou pro fotografy běžnou volbou, představil bych rád několik možností, jak v tomto systému vykonat činnosti, které s tímto povoláním souvisí. Mezi hlavní časovou náplň fotografa patří hlavně získání fotografií, katalogizace, různé retušovací a adjustační práce a nakonec i příprava pro tisk nebo jiné použití. V následujících několika kapitolách se postupně na všechny tyto úkoly podíváme. Nečekejte ovšem kompletní návod, spíše lehký úvod do mnou vyzkoušených nástrojů, postupů a poznámek k nim.

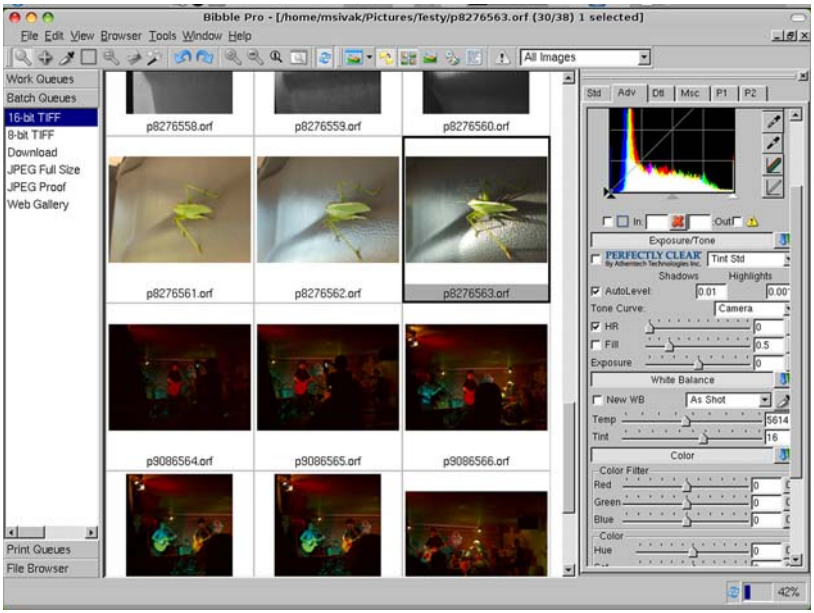
Podpora fotografování

Pro běžnou činnost samozřejmě není podpora operačního systému pro fotografování klíčová, ale existují i oblasti, kde je vhodné se o ní zmínit. Jednou z nich je podpora správy barev a datových formátů používaných v moderních fotoaparátech. O obou těchto oblastech lze již dnes říci, že na linuxových systémech s nimi nemáme problém, i když možná nejsou tak pohodlné, jak by mohly být. Pro správu barev máme knihovnu littlecms a pro práci s RAW formáty projekty libraw a dcraw, který jsou dále využíván v mnoha aplikacích jako digiKam, rawtherapee, rawstudio nebo ufraw. Mezi komerčními produkty se také najdou světlé vlaštovky podporující linux. Například výborný program bibble4 od Bibble labs (bohužel jejich vývojový model je poněkud zvláštní, takže podpora pro všechny nové fotoaparáty je integrována jen do nové vývojové verze, která ještě bohužel postrádá některé užitečné nástroje). Další činností, pro kterou je podpora v OS nezbytná je tzv. tethered shooting. Tj. způsob focení, kdy fotoaparát ovládáte přímo z počítače a na vyfocené snímky se rovnou díváte na velkém displeji. Zde je situace o něco horší, obvykle tento způsob práce podporují pouze oficiální nástroje dodané výrobcem (Olympus studio), které žel často nejsou na linux portované. Pokud tedy tento způsob práce potřebujete, fotoaparáty Canon a Nikon jsou pro vás ty pravé – tethered s nimi podporuje například gphoto2 nebo i již zmíněný bibble4.


digiKam
Album Značka Obrázek Upravit Pohled Nástroje Dávkově Importovat Exportovat Nastavení Nápověda
Vlastnosti souboru
Editor obrázku - Album 'Testy' - digiKam
Soubor Upravit Pohled Barva Vylepšit Transformovat Ozobit Filtry Nastavení Nápověda
Original
Normalizovat
Vyrovnat
Defaults
OK
Zrušit
p8276563.orf (30 z 38)
Žádný výběr
3,720x2,800 (10,42Mpix)


Raw Therapee v2.4.1: p9086565.orf
OLYMPUS IMAGING CORP. E-510
F/2,8 1/15 sec
ISO: 400
Focal length: 25 mm
Lens:
Exposure Detail Colour Transform Metadata
White Balance
Channel Mixer
Colour Boost
Colour Shift
ICM
Input Profile
Working Profile
Save Image
Preferences
Send to editor
p8276562.orf
p8276563.orf


Rawstudio - /home/msivak/Pictures/Testy/p8276563.orf
Soubor Úpravy Fotografie Zobrazení Dávka Nápověda
Priorities
Nástroje Dávka Open
A B C
Expozice
0,00
Sytost
1,00
Odstín
0,00
Kontrast
1,00
Teplota/tónování
0,19
0,32
Sharpen
0,00
Křivka
Připraveno


Bibble Pro - [/home/msivak/Pictures/Testy/p8276563.orf (30/38) 1 selected]
File Edit View Browser Tools Window Help
All Images
Work Queues
Batch Queues
16-bit TIFF
8-bit TIFF
Download
JPEG Full Size
JPEG Proof
Web Gallery
Print Queues
File Browser
p8276558.orf
p8276559.orf
p8276560.orf
p8276561.orf
p8276562.orf
p8276563.orf
p9086564.orf
p9086565.orf
p9086566.orf
Std Adv Dfl Msc P1 P2
Exposure/Tone
PERFECTLY CLEAR
Tint Std
Shadows Highlights
AutoLevel
Tone Curve:
Camera
HR
Fill
Exposure
White Balance
New WB
As Shot
Temp
Tint
Color
Color Filter
Red
Green
Blue
Color
Hue
42%

Získání fotografií z aparátu

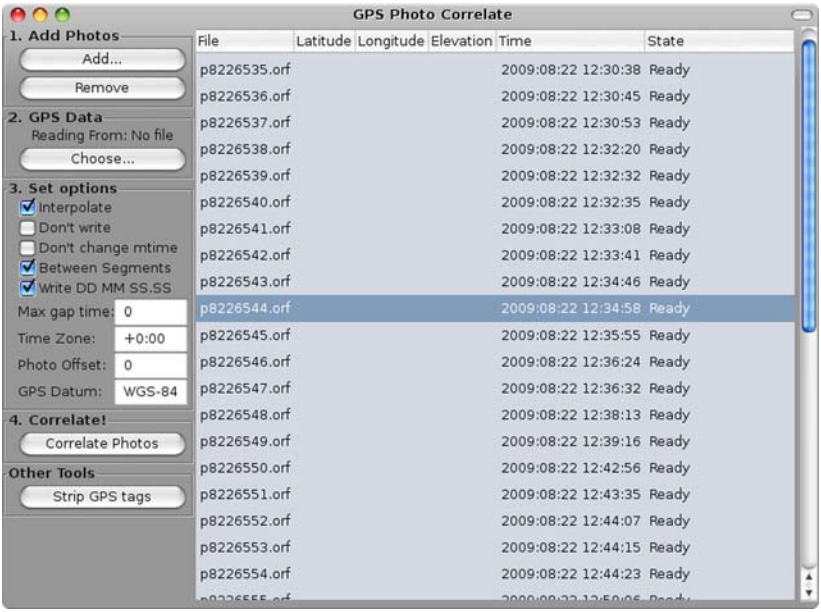
V době nedávno minulé bylo přímé propojení fotoaparátu a počítače zajímavým dobrodružstvím. Naštěstí projekty jako gphoto2 od té doby již stihly dospět a být integrovány do nejběžnějších prostředí, se kterými se dnes můžete setkat. Mnoho výrobců také přešlo na protokol Mass Storage, takže aparát se po připojení objeví jako nový USB disk a není třeba nic řešit. Druhým dnes rozšířeným způsobem je protokol PTP, který býval propagován hlavně značkou Canon. Jednou z věcí, která by nemusela být úplně zřejmá, pokud zpracováváte méně fotografií, je ale rychlost (tedy spíše „pomalost“) přenosu dat přímo z aparátu. Proto, i když vše obvykle funguje i se samotným aparátem, se setkáte s doporučením použít samostatnou čtečku na paměťové karty. Obvykle tím dosáhnete výrazné úspory času při kopírování. Podpora v linuxu pro tak základní věc samozřejmě nechybí.

Anotace dat

Pro nalezení fotografie v archivu je samozřejmě nutné ji správně označit. Nejběžnější formáty typu RAW a formát JPEG podporují dva typy vložených informací – EXIF a IPTC. EXIF obsahuje informace o získání snímku a vytváří ho přímo fotoaparát během focení. IPTC informace jsou určeny pro doprovodné údaje typu autorství, místo a akce, kategorie, nálepky pro snadné vyhledání a podobně. Je vhodné si zažít systém, kdy snímky popíšete hned po zkopírování z aparátu. Později se Vám nebude chtít a v archivu začne vznikat chaos. Proto jako první krok po zkopírování fotografií získaných v terénu provádím „geotagging“. EXIF hlavička totiž umožňuje uložit informaci o přesné GPS pozici, na které byl snímek pořízen. Většina aparátů ale není GPS zařízením vybavena, proto je starost o zapsání tohoto údaje přenesena na fotografa. Já využívám programu GPS Photo Correlation, kterému předám soubor se záznamem trasy ve formátu GPX, adresář s fotografiemi a časový rozdíl mezi systémem GPS a lokálním časem v mém fotoaparátu. Tento program potom na základě časových údajů v trase a ve fotografiích provede spárování fotografií s pozicí, kde se v tu chvíli přijímač nacházel. Další údaje, které zapíši hned na začátku práce s fotografiemi jsou informace o autorství a informace o akci, na které jsem fotografie pořizoval. Programy bibble, digiKam a rawtherapee podporují zápis IPTC informací. Dalším programem, o kterém se podrobněji zmíníme později, s podporou IPTC je jbrout, ale ten pracuje pouze se soubory v JPEG formátu.

Indexace a zálohování dat

Pro správu popsaných fotografií, si můžeme vybrat mezi digiKamem a bibble5 a pokud nepotřebujeme podporu RAW formátů, tak nám poslouží i jBrout (po-



zor na prvotní otázku o přejmenování souborů, silně nedoporučuji odpovídat ano). Ve všech těchto programech můžeme upravovat IPTC data nebo podle nich třídit a vyhledávat. Na síťové zálohování existují i grafické nástroje, ale zatím jsem neobjevil žádný, který by překonal jednoduchost na rsync nástroji založeného skriptu duplicity.

„Vyvolání“ RAW formátu

V předchozích kapitolkách jsem již jemně nakousl téma RAW. Neškodilo by ale pro úplnost uvést, co se pod touto zkratkou skrývá. RAW je obvykle binární, špatně nebo vůbec zdokumentovaný proprietární formát, který (až na jednu výjimku) si každý výrobce vymyslel po svém, aby mohl uložit data získaná z čipu fotoaparátu a zpracovat je až v počítači. Z předchozího odstavce tudíž vyplývá, že za podporou těchto formátů stojí nemálo úsilí a času. Přesto dnes existuje hned několik projektů a nástrojů, které nám práci s nimi umožňují. Za tuto nepříjemnost nám ale RAW formát poskytuje větší množství dat a co je nejdůležitější – možnost nedestruktivní úpravy téměř všech vlastností fotografie. Máme možnost fotografii otočit, zmenšit, upravit barvy, kontrast nebo vyvážení bílé, a to všechno bez ztráty kvality. Jediné co už nelze změnit jsou expoziční hodnoty. Ale i zde je jistý prostor pro korekce, protože RAW obvykle obsahuje více obra-

zové informace, než je možno zobrazit nebo vytisknout. Jako zástupce programů v této kategorii bych možná trošku nečekaně vyzdvihl komerční řešení. Za cenu kolem $ 100 je totiž bibble4 nejrychlejším a nejpohodlnějším programem pro práci s RAW (tedy pokud zatím nepotřebujete novější fotoaparáty), který jsem na linuxu našel. Podporuje totiž dávkové zpracování, kdy pro každou fotografii nastavíte potřebné úpravy a generování výsledného TIFFu nebo JPEGu spustíte až nakonec. Konkurence mu ale roste v podobě nástroje digiKam, který dnes již umí všechno, co od RAW převodníku fotograf potřebuje, možná až na korekci zkreslení konkrétních objektivů. Možností máme samozřejmě více, pořád jsou tu projekty rawtherapee nebo rawstudio. Jejich vývoj pořád pokračuje a pokud nepotřebujete pokročilejší funkce, mohou úlohu převodníku zastat stejně tak dobře.

Retuše a jiné úpravy fotografií

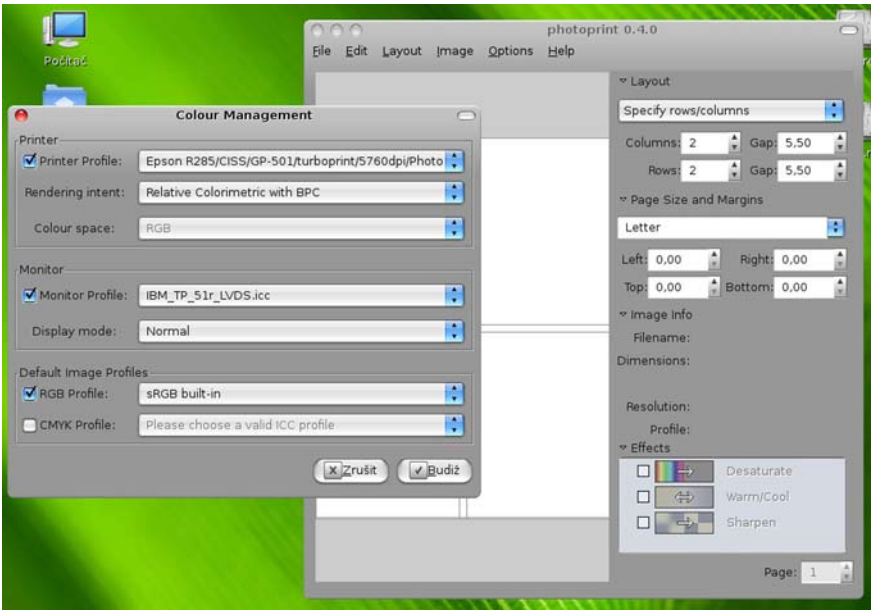
Pro další úpravy je možné samozřejmě použít obligátní GIMP, který pomocí programu ufraw umí RAW soubory i přímo otevírat, bohužel to není moc pohodlné, protože je nutné fotografie otevírat po jedné a je to tudíž poněkud těžkopádné. Ovšem jako doplněk pro retuš fotografie je užitečný, zvláště ve spojení s projektem G'mic (původně Greycstoration), který je výborným „odstraňovačem“ šumu, drátů, plotů a podobných nepříjemností. Pro úpravy ovšem exportujte z RAWu do formátu TIFF. JPEG formát se hodí jen jako výstupní, při jeho úpravách a překódováních příliš trpí kvalita. Dalším programem na práci s bitmapovými daty, tentokrát komerčním, je slovenský Pixel32. Přiznám se, že ho již příliš nesleduji, ale je to zatím jeden z mála editorů, schopných pracovat v 48 bitové barevné hloubce.

Adjustace

Obvyklou úpravou fotografií před zveřejněním je přidání rámečku, vodoznaku nebo podpisu fotografa. Můžete využít nástrojů poskytovaných většími balíky, například Marky a Matty pluginy pro bibble, pluginy v digiKamu nebo přímo úprav v grafických editorech jako třeba Gimp. Také ovšem máte možnost využít „jednoúčelových“ programů a knihoven ImageMagick nebo třeba Python Imaging Library (PIL). Co se efektivity práce týče, nejpohodlnější je samozřejmě využití funkcí programu, ve kterém fotky zpracováváte. Jiná situace již ovšem nastane, pokud fotky někam posíláte automatizovaně a chcete například pouze přidat vodoznak. Pak je rychlejší spustit krátký (obvykle jednořádkový) skript, který celý adresář upraví za Vás.

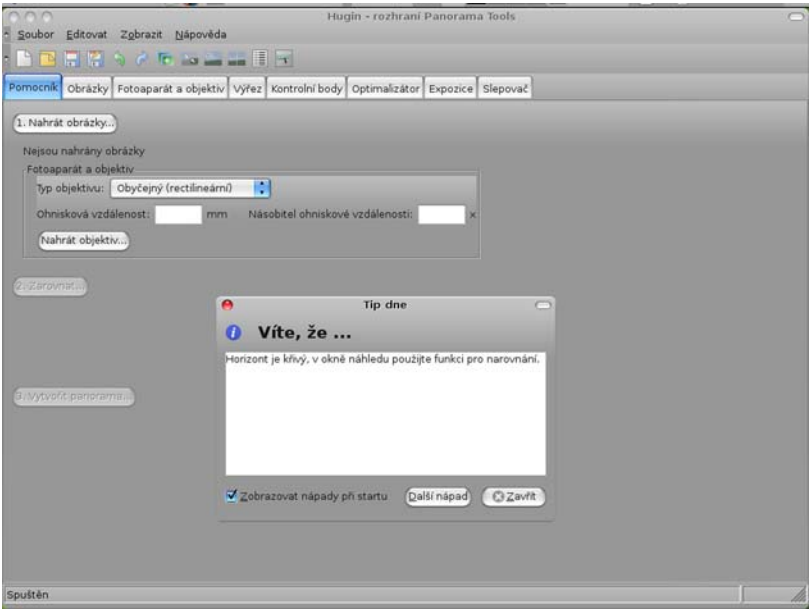
Prezentace fotografií

Ve chvíli, kdy jsou fotografie hotové, potřebujeme obvykle vyřešit způsob jejich prezentace. Nejčastější elektronickou formou bude pravděpodobně webová galerie. S její tvorbou Vám pomůže některý z mnoha časem ověřených projektů. Pro vygenerování statických html stránek pro vložení na CD nebo jednoduchý web je nejsnazším řešením integrovaný generátor alb v programu pro správu fotografií (digiKam, jBrout). Pokud ovšem potřebujete pokročilejší možnosti úpravy vzhledu nebo se Vám jen nelíbí dostupná témata v těchto programech, sáhněte po nástroji JAlbum. Pokud Vám nestačí ani ten, můžete si napsat vlastní. Já už dva roky úspěšně používám svůj (původně) jednoduchý skript, který mi generuje celý můj statický web. Pomocí jazyka python a knihoven PIL a pyexiv2 jsem byl schopný poměrně rychle napsat generátor galerie, včetně získání EXIF a IPTC informací. Druhý směr, kterým se můžete vydat je použití nějaké existující webové služby jako jsou Flickr, Picassa nebo mnoho serverů založených na projektu Gallery2. Pro všechny tyto služby existují buď exportní pluginy, takže fotografie nahrajete z pohodlí Vašeho oblíbeného programu nebo alespoň jednoúčelové nástroje na export. Pro Flickr dokonce vznikl virtuální filesystém (flickrfs), který v linuxu připojíte a potom pomocí běžných příkazů ovládáte svůj účet.



Tisk

Monitor nikdy nebude tak estetický jako pěkně zarámovaná velkoformátová fotografie. Proto i tisk patří k procesu zpracování fotografie. Pro tento účel lze využít třeba tiskový dialog programu, který používáte pro zpracování fotografií. Zde je ovšem jedna z linuxových slabin. Tyto dialogy nejsou jednotné, často neposkytují žádnou kontrolu nad barevností a nejsou tedy příliš vhodné. Proto bych pro účely tisku fotografií z linuxu doporučil výborný projekt photoprint, který podporuje správu barev, umožní Vám poskládat několik fotografií na stránku, provést základní úpravy barevnosti a fotografie naposled zkontrolovat. Zde bych si dovolil jednu důležitou poznámku: Mnoho lidí naráží na úskalí přesné reprodukce barev. Pokud nemáte dobře zkalibrovaný monitor (a případně tiskárnu), raději nečekejte přílišnou věrnost. I když se barvy dají „vyladit“ ručně, je to zdlouhavý, někdy i drahý proces a výsledky nikdy nebudou přesné. Pokud tedy plánujete tisknout větší množství fotografií, obětujte nějaké finance a nechte si udělat kalibraci. Uvidíte, že ten rozdíl poznáte.



Další možnosti

Oblíbenou kratochvílí krajinářských fotografů je vytváření panoramatických snímků. I pro tuto úlohu naleznete v linuxu kvalitní nástroj. Jedná se o projekt Hugin, který Vás pomocí jednoduchého průvodce provede vložením foto-

grafií ve formátu JPEG (pokud možno s co nejbližšími parametry expozice), vyhledáním spojovacích bodů a vytvořením výsledného panoramatu. Umožňuje nahrát i více snímků než je třeba a pak postupně vypínat nebo zapínat jejich viditelnost pro dosažení co nejlepšího výsledku. Fotografie jsou dnes samozřejmě pořizovány i pro mnoho jiných účelů než umění nebo krajinky – technickou dokumentací výrobků počínaje a tvorbou plakátů nebo billboardů konče. Proto se sluší alespoň zmínit nějaký program, který je pro tyto účely vhodný. A to je zaprvé a především Scribus. Tento program může směle konkurovat mnoha komerčním DTP programům, takže pokud potřebujete vytvořit vizitku, plakát nebo kalendář, sáhněte rovnou po něm.

Závěr

Jak jste se mohli přesvědčit na předchozích několika řádcích, není situace v linuxových systémech už ani zdaleka tak zoufalá, jak ji někteří prezentují. Nějaké mezery ještě existují v oblasti podpory nových aparátů nebo počítačem řízeného fotografování, ale na všechny běžné činnosti již linux stačí. Nezbývá mi tedy nic jiného než Vám popřát hodně fotografických úspěchů a dobré světlo, bez kterého to nejde především.

Zdroje a odkazy

[1] Bibble Labs – http://bibblelabs.com/

[2] dcraw – http://www.cybercom.net/ dcoffin/dcraw/

[3] digiKam – http://www.digikam.org/

[4] Duplicity – http://duplicity.nongnu.org/

[5] GIMP – http://www.gimp.org/

[6] gphoto2 – http://www.gphoto.org/

[7] GPS Photo Correlation – http://www.freefoote.com/linux_gpscorr.html

[8] G'mic – http://gmic.sourceforge.net/

[9] Hugin – http://hugin.sourceforge.net/

[10] Image Magick – http://www.imagemagick.org/

[11] JAlbum – http://jalbum.net/

[12] jBrout – http://jbrout.manatlan.com/

[13] LittleCMS – http://www.littlecms.com/

[14] Photoprint – http://blackfiveimaging.co.uk/index.php?article=02Software%2F01PhotoPrint

[15] Python Imaging Library – http://www.pythonware.com/products/pil/

[16] Rawstudio – http://rawstudio.org/

[17] RawTherapee – http://www.rawtherapee.com/

[18] Scribus – http://www.scribus.net/

[19] ufraw – http://ufraw.sourceforge.net/

Jak ti vědci počítají

Zdeněk Šustr

E-mail: zdenek.sustr@cesnet.cz

Klíčová slova: grid, grid computing, supercomputing, cloud computing, parallel computing

Abstrakt

Moderním vědeckým přístrojům jako jsou urychlovače částic či velké astronomické dalekohledy se dostává poměrně dost pozornosti dokonce i v populárních médiích. Na jakých strojích se ale zpracovávají nasbíraná data? Na čem se spouští simulace fyzikálních či chemických procesů? Tento článek poskytne základní přehled cest, jimiž se lze ubírat.

1 Superpočítače

Chceme-li rychle počítat, základním impulzem je pořídit si rychlejší počítač. Termín „superpočítač“ není zřetelně vymezen a obvykle se používá k označení počítače, který se řadí (v době svého vzniku) mezi nejvýkonnější dostupná zařízení. Kam tato špička dosahuje v současnosti ilustruje Tab. 1. Z hlediska běžných vědeckých výpočtů mají superpočítače některé nevýhody:

Pořizovací náklady Pořízení superpočítače je jednorázová investice, po níž následuje již jen pozvolné zastarávání. Možosti upgradu jsou velmi omezené. Na druhé straně dovede nabídnout „vymoženosti“ jako např. sdílenou paměť.

Škálovatelnost Existují sice modulární architektury, které do jisté míry škálování umožňují, ale jak hardware superpočítače stárne, přestává být tato možnost zajímavá jak z finačního tak z výkonnostního hlediska.

Sdílené vlastnictví Je sice možné, ale problematické. I v případě, že se výzkumné instituce dokážou dohodnout na společném pořízení superpočítače, narazí na řadu administrativních a účetních překážek.

Tabulka 1: První čtyři místa žebříčku TOP500

č.	Provozovatel	Popis	Výkon [PFLOPS]
1.	DOE/NNSA/LANL	**Roadrunner** – BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3,2 Ghz / Opteron DC 1,8 GHz, Voltaire Infiniband, IBM (129 600 jader)	1,105
2.	Oak Ridge National Laboratory	**Jaguar** – Cray XT5 QC 2,3 GHz, Cray Inc.	1,059
3.	Forschungszentrum Juelich (FZJ)	**JUGENE** – Blue Gene/P Solution, IBM	0,826
4.	NASA/Ames Research Center/NAS	**Pleiades** – SGI Altix ICE 8200EX, Xeon QC 3,0/2,66 GHz, SGI	0,487

Vytížení Aby se pořizovací náklady pokryly, musí být superpočítač dostatečně vytížen. Je sice představitelné, že se jednotlivé vědecké týmy budou vzájemně synchronizovat, aby např. jedna skupina počítala, zatímco další připravuje své vlastní úlohy, ale v praxi je dosažení taková míry souhry velmi administrativně náročné.

Superpočítače jsou pochopitelně nenahraditelné u řešení náročných úloh, které neumožňují masivní paralelizaci. Některé z výše uvedených problémů (např. škálování) překonávají výpočetní clustery složené z většího počtu „standardních“ počítačů. Jiné problémy (např. vytížení) se řeší zapojením vlastních výpočetních prostředků do větších národních či nadnárodních infrastruktur – gridů.

2 Gridy

Narozdíl od superpočítačů se specializovaným hardwarem se gridy staví prakticky z obyčejných PC. Jejich síla je ovšem v množství. Zatímco superpočítač musí stát na jednom místě a případné spoluvlastnictví je obtížné, grid je distribuovaným prostředím, do nějž účastnické organizace vkládají jako příspěvek svůj hardware a samy o něj dále pečují. To s sebou přináší celou řadu výhod:

Distribuované stárnutí Gridový hardware nezastarává najednou. Jednotlivé části se obvykle obměňují s tím, jak se odpojují staré zdroje a zapojují nové. Z hlediska uživatelů to znamená zvýšenou stabilitu prostředí a postupně rostoucí výkon.

Jednoduché škálování Posílení gridu např. vstupem nového partnera, či naopak odpojení některých částí, je rutinní operace.

Snadné zapojení Princip gridu velmi dobře vyhovuje zažité koncepci vlastnictví výpočetních prostředků. Vědecká pracoviště jsou z dlouholeté praxe zvyklá vlastnit výkonný hardware a jeho pořizování se chápe do značné míry jako prestižní záležitost. Zapojením do gridu[1] se může zlepšit míra jeho využití a účastnická organizace zároveň získává další služby, např. vč. možnosti využívat nárazově výkon větší než vložený příspěvek.

2.1 Grid Middleware

Typickým výpočetním prostředím současnosti je rozsáhlý grid osazený gridovým middlewarem, tj. sadou služeb, které zajišťují spolupráci jednotlivých uzlů, uložení a zpřístupnění dat, plánování úloh a dohled nad jejich životním cyklem.[2] Z pohledu uživatele se nejedná o interaktivní počítač, kde by se mohl připojit ke konzoli a pracovat s ním na úrovni jednotlivých příkazů. Naopak, jedná se o distribuované prostředí, do nějž uživatel „vypustí“ nadefinovanou úlohu[3] a pak čeká na její zpracování. Různé komponenty gridu zajistí, že se úloha správně naplánuje a spustí na adekvátním zařízení.

Základní komponenty gridu a jejich podíl na zpracování výpočetní úlohy ilustruje Obr. 1.

Patrně největším gridem pro vědecké výpočty je produkční grid evropského projektu EGEE[4]. Podle údajů z července 2009 má v současnosti ve výpočetních kapacitách 92 tisíce procesorů (144 tisíc jader) a diskovou kapacitu 25 PB. V Česku je do gridu EGEE zapojeno přibližně 1 600 procesorů. Vzhledem k decentralizované povaze – díky níž se obvykle nedají v infrastruktuře spouštět celoplošné výkonnostní testy – lze, bohužel, celkovou výkonnost takového gridu pouze odhadovat. Podle statistik z menších úloh spouštěných řádově na tisících jader si autor troufá zcela subjektivně odhadnout výkon přibližně 1 TFLOPS na 100–150 procesorů, což by pro celý grid EGEE činilo přibližně 0,75 PFLOPS.

Výhradně českým gridovým prostředím je MetaCentrum provozované sdružením CESNET. Jeho uživatelé mají pro své výpočty k dispozici 1 150 procesorů.[5]

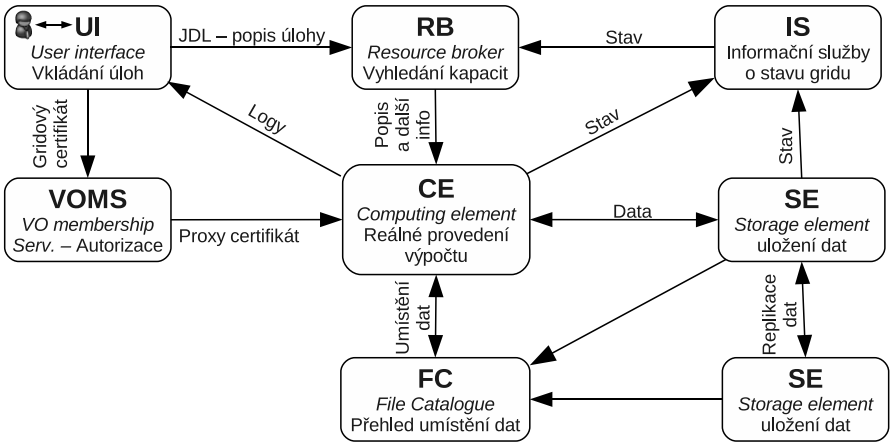
[1] Přičemž toto zapojení by mělo být podle předních vizionářů [2] stejně jednoduché, jako zapojení do elektrické sítě – odtud také samotný výraz „Grid“...

[2] Při tom se skutečně jedná o middleware, který sice ve své distribuci nabízí hotové nástroje pro základní operace, ale hlavně publikuje programové rozhraní, jenž mohou uživatelské skupiny využívat při řešení komplexnějších problémů.

[3] Součástí definice je označení spustitelného příkazu, seznam požadavků na instalované knihovny a jejich verze, odkazy na vstupní data apod.

[4] Enabling Grids for E-sciencE

[5] Údaj ze 17. 9. 2009, uvádí pouze aktuálně využitelné/využité procesory.



Obr. 1: Náznak interakcí gridových komponent při průchodu úlohy

Pro srovnání: 5. generace superpočítače (clusteru) Amálka provozovaného akademií věd má 326 procesorů.

2.2 Cycle Scavengers

Speciálním případem gridových prostředí jsou tzv. *cycle scavengers*, tzn. programy využívající volnou výpočetní kapacitu na osobních počítačích dobrovolníků. Příklady těch nejvýkonnějších uvádí Tab. 2. Tento způsob počítání je téměř výhradně doménou vědeckých výpočtů, obvykle v atraktivních oborech jako genetika či astrofyzika. Méně přitažlivé, kontroverzní či tajné projekty se tímto směrem z pochopitelných důvodů ubírat nemohou. Existují ovšem také komerční programy, které na jedné straně dobrovolníkům za účast v projektu platí a na druhé straně nabízí „vykoupený“ procesorový čas dalším zájemcům na komerční bázi [3].[6]

Zajímavým trendem je posun od běžných osobních počítačů k dalším domácím zařízením, která disponují relativně velkým a při tom často zcela nevyužitým výkonem. Projekt Folding@home, který v současnosti disponuje mezi všemi gridy svého druhu největším výkonem, např. v době vzniku tohoto článku získává více než 25 % výpočetního výkonu z klientů pro herní konzoli PlayStation 3.

[6]Pro ilustraci: Výše odměny se odvozuje od dostupnosti počítače (musí být připojen k Internetu alespoň 20 hodin denně) a propočítaného času. V září 2009 se nabízí odměna 0,10 USD za každý den, kdy byl počítač on-line, plus 0,000 5 USD za propočítanou minutu.

Tabulka 2: Nejvýkonnější infrastruktury skupiny Cycle Scavengers

č.	Projekt	Výkon [PFLOPS]
1.	Folding@home	7,565[7]
2.	BOINC (Seti@home, Einstein@home...)	2,411
3.	GIMPS (Great Internet Mersenne Prime Search)	0,041

3 Cloudy

Cloudy mají s gridy mnoho společného. Jedná se o rozsáhlé výpočetní zdroje, ovšem provozované zpravidla na komerční bázi a zprostředkované ve formě virtuálních strojů, u nichž se zpoplatňuje reálné využití, tzn. propočítaný čas, přenesená data apod. Mimo finanční aspekty je tu ještě několik dalších významných rozdílů:

Homogenita V cloudovém prostředí zpravidla figuruje jeden poskytovatel služeb (např. [4]), což znamená, že prostředí je, co do použitých architektur, značně homogenní. To je pochopitelně výhoda, je-li váš výpočet připraven právě pro takové prostředí. Na druhé straně se může jednat o nevýhodu, pokud nabídnuté prostředí některé požadavky nesplňuje, nebo pokud je potřeba úlohu složitě přizpůsobovat.

Administrativa Z výše uvedeného také plyne, že příjemcem plateb je komerční firma, často zahraniční, zatímco do gridu se zapojují lokální prostředky a velká část nákladů na jejich pořízení a hlavně provoz zůstává doma. Tato skutečnost sice nemá přímý dopad na vhodnost či nevhodnost cloudů pro vědecké výpočty, nicméně pro vlády, které vědecký výzkum financují, se jedná o otázku poměrně zajímavou.

Management úloh Narozdíl od gridů, které se orientují na zpracování jednotlivých úloh, je cloud pouze „holý“ OS, který standardně neobsahuje žádnou podporu pro dávkové spouštění úloh, sledování jejich stavu, dílčí účtování, přenosy vstupních a výstupních dat apod. Pokud má uživatel zájem o takové funkce, musí si je sám doplnit do svého virtuálního systému.

Prominentním poskytovatelem cloudových služeb je americký Amazon. Nabízí služby v různých úrovních a s různým způsobem zpoplatnění (předplacené rezervované zdroje, nebo naopak placení pouze za propočítaný čas). Nejvýkonnější dostupný virtuální stroj – 8 jader s výkonem ekvivalentním 2,5 GHz proce-

[7] Údaj platný v době vzniku článku (září 2009), konvertován do x86 PFLOPS. Špičkový výkon dosažený v květnu 2009 činil 8,5 x86 PFLOPS.

soru Intel Xeon, se 7 GB paměti a s Linuxovou distribucí dle vlastního výběru[8] – vám zpřístupní za paušální cenu 1 820 USD za rok plus 0,32 USD za propočítanou hodinu. Mimo to vám naúčtuje uložení dat (0,11 USD za GB×měsíc) a další drobné příplatky za I/O operace, monitoring, přidělení IP adres apod. Studie [5] např. odhaduje, že vykonání veškeré práce provedené gridem EGEE za rok 2007, by při použití cloudových služeb Amazonu stálo více než 59 miliónů USD.[9]

V srpnu letošního roku zpřístupnil cloudové služby speciálně pro aplikace vyžadující velký výpočetní výkon také výrobce linuxových clusterů Penguin Computing.

4 Virtualizace na objednánku – cloudy v gridech

Potenciální přínos, jenž virtualizace zdrojů nabízí, je zřejmý. Virtualizace umožňuje provoz stabilního prostředí na různých hardwarových prostředcích, jeho migraci dle aktuálních administračních potřeb a přináší celou řadu dalších výhod.

Samostatnou kapitolou je pak možnost zprovoznění vlastních virtuálních strojů v prostředí pro zpracování dávkových úloh [6]. Uživatelům zjednodušuje vstup do „velkého světa“ gridů. Prostředí, na něž jsou zvyklí a jež si ve spolupráci se svými správci vyladili podle svých vlastních potřeb, se jednoduše spustí jako zvláštní typ úlohy ve velkém distribuovaném systému. Pokud zmíněné rozsáhlé výpočetní prostředí poskytne příslušnou podporu, dají se pak z jednotlivých virtuálních strojů stavět clustery se všemi parametry, které bychom očekávali,[10] tzn. např. včetně vytvoření virtuální privátní sítě, do níž se jednotlivé uzly clusteru zapojí bez ohledu na své fyzické umístění. Tím se zjednodušuje i správa takových virtuálních clusterů, zejména co do zabezpečení a správy uživatelských účtů.

Výhody jsou nasnadě. Uživatelé stále pracují v prostředí, na něž jsou zvyklí. Cluster nemusí běžet pořád. Když se nepočítá, virtuální stroje se zastaví a na příslušném hardwaru může v tu dobu počítat někdo jiný. Když se úloha ladí nebo se pouze připravuje prostředí, nemusí běžet všechny uzly, nebo nemusí běžet na

[8] U některých typů zpoplatnění jsou k dispozici také virtuální stroje s MS Windows, ale jsou o 45 % dražší.

[9] Podle údajů dostupných v gridovém monitoringu se v uvedeném roce 2007 v gridu EGEE zpracovaly přibližně 54 milióny úloh. V roce 2008 vzrostl počet zpracovaných úloh na téměř 125 miliónů. S přijatelnou mírou jistoty lze předpokládat, že přímo úměrně tomu by vzrostl i odhad nákladů na pořízení stejné výpočetní kapacity od Amazonu. Výslednou sumu si autor přesto netroufá – z obavy před velkými a potenciálně zavádějícími čísly – uvést konkrétně.

Čísla z prvních osmi měsíců letošního roku navíc indikují meziroční nárůst počtu zpracovaných úloh mezi lety 2008 a 2009 o dalších 90 %.

[10] Vzniká tak prostředí, které je v základních rysech ekvivalentní cloudu.

plný výkon. Naopak v době, kdy výpočet potřebuje maximální výpočetní kapacitu, lze cluster i posílit a poskytnout uživateli dočasně špičkový výkon, který by pro něj nebyl myslitelný, kdyby byl odkázán pouze na vlastní hardwarové prostředky.

Narozdíl od komerčních cloudů, kde je vlastníkem hardwarových zdrojů jeden subjekt, mohou být tyto zdroje v gridu distribuované. Uživatelé mohou do gridu zapojit své vlastní stroje a potom využívat adekvátní výpočetní výkon např. i tehdy, když jejich vlastní hardware neběží z důvodu poruchy nebo údržby. Za zpřístupnění svých vlastních kapacit v době, kdy je sami plně nevyužívají, pak mohou být „odměněni“ navýšením dostupného výkonu v okamžiku, kdy jej potřebují.

Reference

[1] *Top500 List June 2009*, TOP500, 23. 6. 2009, http://www.top500.org

[2] Foster, I., Kesselman, C. *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers, ISBN 1-55860-475-8.

[3] *Gomez Peer*, https://www.gomezpeerzone.com/

[4] *What is AWS?*, Amazon.com, 2009, http://aws.amazon.com/what-is-aws/

[5] Bégin, M. *An EGEE Comparative Study: Grids and Clouds – Evolution or Revolution?*, CERN, 30. 5. 2008.

[6] Sitera, J., Antoš, D. *Virtuální grid*, Brno, CPress Media, a. s., 2009. 2 s. Connect! 7–8/2009, roč. XIV. ISSN 1211-3085.

Jak podpůrná infrastruktura datového centra ovlivňuje jeho možnosti a rozpočet

Vladimír Houška

E-mail: houska.vladimir@completecz.cz

Provozovatelé datových center dnes ve stále větší míře sledují celkové náklady na vlastnictví, ve kterých hrají významnou roli provozní náklady a náklady na změny konfigurace, resp. navyšování výkonu. Jednou z největších položek tvoří systém chlazení – jak z hlediska pořizovací ceny, tak spotřeby energie. Ačkoliv se dnes většina organizací snaží maximálně šetřit, setrvalý růst objemu zpracovávaných dat a stoupající hustota ICT technologií je nutí k navyšování kapacit serveroven a datacenter. Rovněž zajištění nepřetržitého chodu informačních a komunikačních technologií a ochrana systémů před vážnější nehodou vyžaduje adekvátní zázemí a podpůrnou infrastrukturu.

Jak se měří provozní efektivita

Jedním z velmi diskutovaných témat světa informačních technologií je Green IT. Pro nás tkví jeho hlavní význam ve sledování efektivity systému chlazení, tedy především poměru spotřeby energie výrobníků studené vody a CRAC jednotek a spotřeby energie chlazených ICT technologií. Vzhledem k současným cenám energií mohou účty za elektřinu během průměrné 10leté životnosti datacentra snadno převýšit účty za pořízení veškeré IT techniky.

Nejrozšířenějším způsobem je poměřování spotřeby celého objektu datacentra (a potažmo jeho hlavních částí, např. chlazení) ke spotřebě vlastní IT techniky (o kterou zde jde na místě prvním). Výsledný parametr se označuje PUE a občas se používá i jeho převrácená hodnota DCiE (v tomto případě je spotřeba ICT techniky uvedena jako procentuální část celkové spotřeby datacentra). U počítače volně stojícího v prostoru (nebo datacentra na severním pólu chlazeného větrem, bez osvětlení a zálohování) se PUE rovná 1, u současných datacenter někdy překonává 3. Přitom za přijatelné se dnes považují hodnoty pod 2. To znamená, že pokud má ICT příkon 50 kW, nesmí odběr celého centra přesáhnout 100 kW. (více viz např. http://www.TheGreenGrid.org nebo česky http://www.CompleteCZ.cz/GreenIT.html)

Největší prostor ke zvýšení efektivity je zvyšování efektivity systému chlazení, tedy především poměru spotřeby energie výrobníků studené vody a CRAC jednotek a spotřeby energie chlazených ICT technologií. Avšak ještě než se vrhnete na plánování maximálně sofistikovaného řešení, ověřte si rozpočet. I když provozovatelé Green IT přístupy požadují, málokterý investor chápe, že zpočátku budou stát více peněz. Pokud například zákazník požaduje roční návratnost investice do „zelených" technologií, nemůže si koupit to nejlepší řešení na trhu.

Trochu jiná, pro energeticky úsporné technologie příznivější situace, se otevírá při modernizaci či rozšiřování starších, často hodně neefektivních datových center. Většina nových jednotek bude sama o sobě efektivnějších, a když je doplníme např. o free-cooling a optimalizujeme rozložení techniky a distribuci chladu uvnitř datacentra, poměrně snadno snížíme dosavadní provozní náklady nebo s podobnými náklady uchladíme výkonnější technologie.

Další souvislosti a vazby

Největší balíky peněz si při stavbě datového centra, vedle vlastní IT techniky, vyžádá systém chlazení a energetický systém (distribuce napájení a záložní zdroje) – zároveň mají v koncepci datacentra velké vzájemné vazby. Pro systém napájení a zálohování jsou např. významné rozběhové proudy výrobníků chlazené vody – jednak spoluurčují požadovaný příkon technologického celku a zároveň i potřebnou kapacitu, resp. výkon záložních zdrojů energie. U každého většího datacentra centra je požadován nepřerušený chod, takže záložní systémy musí zajistit energii pro IT technologie i veškerou nezbytnou infrastrukturu, tj. i pro chillery.

Proto může být v některých situacích výhodné použít jednotky s kompresorem v provedení Turbocor, který má startovací proud 2 A – na rozdíl od běžných 200 až 500 A klasických scroll kompresorů. Turbokompresory mají celkově výrazně vyšší účinnost – při částečné zátěži i třikrát až čtyřikrát – což znamená, že při provozu šetří elektrickou energii. Toto řešení sice stojí přibližně o třicet procent víc, ale kromě provozních nákladů se ušetří i na zmíněné energo části – na menších dimenzích rozvodů i na méně výkonných UPS a diesel agregátech. Řešení se bude hodit i v situaci s pevně omezeným příkonem daného objektu.

Dalším důležitým hlediskem je zmíněné budoucí rozšiřování datacentra. Dodavatel například zákazníka, trochu bez ohledu na jeho konkrétní potřeby, přesvědčí na řešení pomocí přímého chlazení stojanů. Když pak uživatel potřebuje doplnit několik nových prvků (nová síťová zařízení, další servery či disková pole), musí zakoupit další, v dané chvíli zbytečně naddimenzovaný modul nebo si musí pořizovat jiný, odlišný systém ke chlazení ostatních technologií – v obou případech dojde ke zbytečnému nárůst provozních nákladů.

A opačně, když je stávající systém CRAC jednotek na hranici své kapacity, může být někdy spíše než jeho rozšiřování výhodné doplnit do systému jeden či dva stojany s přímým chlazením (např. APC, RITTAL či SAIFOR), přemístit do nich zařízení s nejvyšší hustotou (nejvyšším ztrátovým tepelným výkonem), a „odlehčit“ tak zbývající části sálu.

A do třetice, pokud například stávající chlazení přes zdvojenou podlahu nezvládá uchladit horní části několika stojanů s vyšší hustotou, není většinou potřeba zavádět nový systém nebo zvyšovat celkový chladicí výkon. Často postačí doplnit aktivní podlahové moduly (např. UNIFLAIR AFM), případně v kombinaci s inteligentním řízením tlaku v podlaze AFPS či dalšími technologiemi a infrastruktura datacentra může sloužit ještě nějaký čas bez větších rekonstrukcí.

Pár příkladů

Pokud by například provozovatel většího datového centra zůstal pouze u přímého chlazení (kdy chladicí jednotky mezi racky chladí pouze jeden či dva sousední stojany), musí počítat s nároky na údržbu mnoha malých jednotek, při rozšiřování musí dokupovat další jednotky a jim odpovídající stojany jedné značky, musí samostatně řešit chlazení UPS a dalších technologií mimo stojany. Za čas může skončit v podobné situaci, k jaké došlo u serverů – kdy náklady na údržbu a správu mnoha malých, různě opotřebovaných zařízení uživatele vedou k investicím do jejich konsolidace, tj. převedení provozu do menšího počtu výkonnějších strojů. Od určité velikosti se prostě vyplatí chlazení zdvojenou podlahou. Jednou ze zářných referencí z poslední doby je instalace přesné klimatizace UNIFLAIR v evropském datacentru EasyNet, kde se vhodnou kombinací technologií daří vyrobit 12 656 MWh chladu se spotřebou pouhých 743 MWh elektřiny.

Na druhou stranu je nerozumné doporučovat do serverovny, kde v nejbližších letech budou jen dva nebo tři stojany, zdvojenou podlahu a CRAC jednotku – mnohem jednodušší budou právě racky s přímým chlazením a chiller s odpovídající rezervou. Rovněž v případě velmi stísněných prostor bývá přímé chlazení stojanů vhodnější variantou, kdy nedochází ke snížení výšky místnosti a nejsou potřeba složitější stavební úpravy.

Při volbě řešení se naopak není třeba obávat stojanů s vysokou hustotou a větší produkcí odpadního tepla, třeba až ke 30 kW/rack – zkušenosti ukazují, že uchladit je lze jak s pomocí mezirackového chlazení, tak i zdvojené podlahy. Podobně je to s mírou účinnosti toho či onoho řešení – důležitá je především celková koncepce systému a dotažení zdánlivě drobných detailů reálného provedení. (zajímavé je např. studie Intelu: Air-Cooled High-Performance Data Centers: Case Studies and Best Methods na adrese http://www.intel.com/it/pdf/air-cooled-data-centers.pdf)

A co zdůraznit na závěr? Problematiku chlazení datového centra je potřeba chápat jako součást infrastruktury, kde navazuje na další prvky a kde jeden ovlivňuje druhý. Zároveň je dobré mít na paměti, že ke skutečně efektivnímu řešení často vede kombinace různých technologií a „vychytávek“. Nakupujte proto u odborníků :)

Architektura sběrnic PCI, PCI-X a PCI-Express

Tomáš Martínek

E-mail: martinek@liberouter.org

Klíčová slova: PCI, PCI-X, PCI Express, systémová sběrnice

Abstrakt

Příspěvek je věnován popisu vývoje sběrnic rodiny PCI, konkrétně sběrnic PCI, PCI-X a PCI-Express. Stručně jsou shrnuty principy společné pro všechny typy sběrnic jako např. struktura paměťového prostoru nebo způsob alokace prostředků v době inicializace systému. Zdůrazněny jsou rozdíly v architektuře jednotlivých sběrnic, komunikačním protokolu, způsobu generování přerušení, zabezpečení přenosu dat nebo obsluhy a zotavení z chyb.

1 Úvod

Sběrnice typu PCI tvoří v současné době standard pro připojení přídavných adaptérů a periferních zařízení v rámci architektury personálních počítačů. Jejich vývoj je koordinován organizací PCI-SIG (Peripheral Component Interconnect Special Interest Group), která spravuje a vydává specifikace sběrnicí od fyzické úrovně (definice elektrických signálů, tvar konektoru, apod.) až po podrobný popis komunikačního protokolu.

Historie vývoje sběrnicí typu PCI sahá až do roku 1993, kdy byla uvolněna jedna z prvních specifikací sběrnice typu PCI pracující na frekvenci 33MHz. O dva roky později byla tato základní verze doplněna o variantu pracující na dvojnásobné frekvenci 66 MHz. Obě tyto verze však obsahovaly velké množství nedostatků jako jsou např. nízká efektivita, nedostatečná správa chyb a komplikace při dalším zvyšování pracovní frekvence potažmo přenosové kapacity sběrnice. V roce 1999 byla proto uvolněna nová specifikace sběrnice s označením PCI-X. Oproti předchozímu typu již dokázala mnohem efektivněji pracovat

s přenosovou kapacitou a odstranila téměř všechny nedostatky svého předchůdce. Stále vyšší požadavky na přenosovou kapacitu sběrnicí, snížení jejich ceny a nástup nových technologií však vedly v roce 2002 na opuštění tohoto konceptu paralelních sběrnic řízených hodinovým signálem. Nová generace s označením PCI-Express byla postavena na vysokorychlostních, plně duplexních sériových linkách a paketově orientovaným komunikačním protokolem.

Tabulka 1: Propustnost sběrnicí PCI a PCI-X

Typ	Frekvence	Max. propustnost	Počet slotů
PCI 32-bit	33 MHz	133 MB/s	4–5
PCI 32-bit	66 MHz	266 MB/s	1–2
PCI-X 32-bit	66 MHz	266 MB/s	4
PCI-X 32-bit	133 MHz	533 MB/s	1–2
PCI-X 32-bit	266 MHz	1 066 MB/s	1
PCI-X 32-bit	533 MHz	2 131 MB/s	1

Propustnost jednotlivých typů sběrnicí PCI a PCI-X je znázorněna v následující tabulce. Hodnotu v MB/s lze jednoduše vypočítat jako součin frekvence a šířky sběrnice. Teoretické hodnoty propustnosti se tedy pohybují od 133 MB/s až po 2 131 MB/s. Kromě typů uvedených v tabulce existují také varianty sběrnicí se šířkou 64 bitů, jejich propustnost pak dosahuje dvojnásobných hodnot. Velmi zajímavou informací je údaj o počtu slotů, které norma definuje pro jednotlivé typy. Z tabulky je možné si všimnout, že se zvyšující se frekvencí se počet slotů na sběrnici snižuje. Tato vlastnost je ovlivněna především technologií výroby, neboť pro vyšší frekvence je návrh desky spojů komplikovanější a splnění časových kritérií pro správnou funkci sběrnice je složitější. Pro vysoké frekvence norma povoluje pouze jeden slot na sběrnici. Pokud je potřeba více slotů, potom musí být od sebe odděleny skrze PCI bridge, který zajistí převzorkování a konkrétní zpracování protokolu.

Tabulka 2: Propustnost sběrnicí PCI Express

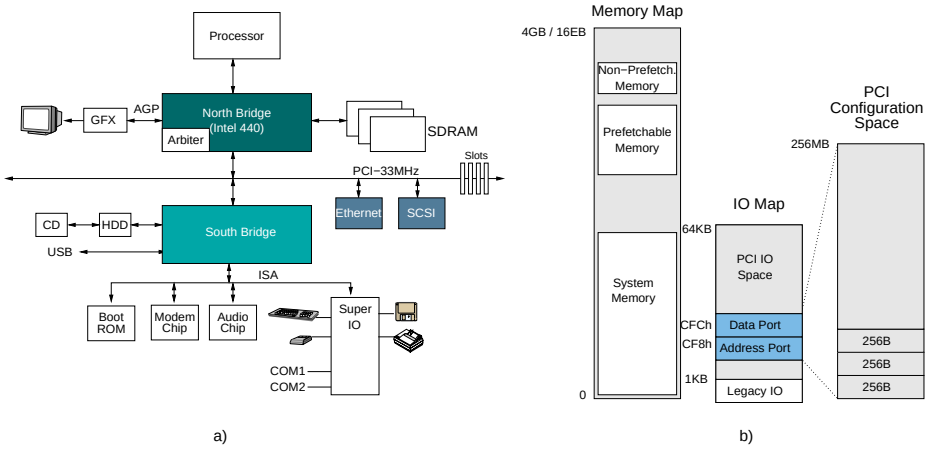
Šířka linky	x1	x2	x4	x8	x12	x16	x32
Propustnost [GB/s]	0.5	1	2	4	6	8	16

U sběrnice typu PCI Express je propustnost dána rychlostí a počtem sériových linek. Ty v základní verzi dosahují přenosové kapacity 2.5 Gb/s pro každý směr (příjem i vysílání). Linky jsou skládány k sobě paralelně v rozsahu od 1 do

32 a hodnota propustnosti se tak pohybuje v rozsahu od 0.5 do 16 GB/s. Poznámka: do těchto hodnot je již započítána režie kódování 8/10 (každých 8 bitů dat je zakódováno do 10 bitů). V současné době je k dispozici již druhá generace této sběrnice s propustností 5 Gb/s na linku a připravuje se třetí generace, která propustnost opět zdvojnásobí (konec roku 2010).

1.1 Model architektury systému se sběrnicí PCI

Model architektury pro systémy se sběrnicí PCI je zobrazen na obrázku 1a. Architektura je rozdělena do několika částí na základě toho, jakou propustnost jednotlivé části pro svoji funkci potřebují. Zařízení vyžadující vysokou propustnost dat jsou propojeny skrze tzv. *North bridge* (severní most). Zde patří především procesor, systémová paměť a grafický adaptér. *North bridge* spojuje všechny tyto části a realizuje efektivní přenosy mezi příslušnými rozhraními (AGP, DRAM, CPU). North bridge mimo jiné tvoří také přípojný bod celé PCI sběrnice, kde se předpokládá připojení adaptérů s nižšími požadavky na propustnost (např. SCSI řadič, síťový adaptér, apod.). Nakonec, pomalá zařízení, která potřebují pro svoji funkci jen velmi omezenou přenosovou kapacitu (myš, klávesnice, modem, IDE řadič apod.), jsou připojena skrze tzv. *South bridge* (jižní most). V celkovém modelu pak *South bridge* může vystupovat jako jedno z PCI zařízení. Na tomto místě je důležité poznamenat, že se jedné pouze o model. Reálné architektury se od tohoto modelu mohou lišit. Zejména implementace *North* a *South bridge* je obvykle provedena v rámci jedné čipové sady, kde skutečná realizace a rozvrstvení závisí na konkrétním výrobci.



Obr. 1: a) Model architektury počítače se sběrnicí PCI, b) Struktura paměťového prostoru PCI

Důležitá je také struktura paměťového prostoru systému založeného na PCI. Rozlišují se zde tři paměťové prostory:

- *Paměťový prostor* – dopovídá fyzickému paměťovému prostoru systému. V závislosti, zda je použita 32bitová resp. 64bitová architektura je velikost tohoto prostoru 4 GB resp. 16 EB. Spodní část tohoto prostoru je rezervována pro paměť RAM fyzicky dostupnou v konkrétním počítači (např. 512 MB nebo 1 GB). Nad touto pamětí se nachází potencionálně volný prostor, který se využívá pro mapování paměťových prostorů nejrůznějších periferních zařízení připojených na sběrnici PCI, AGP, apod.

  Pro ilustraci, pokud je v systému připojen grafický adaptér nebo zvuková karta, každé z těchto periferních zařízení obsahuje svou interní paměť. A právě tato paměť se zpravidla mapuje do prostoru nad hranicí fyzické RAM. Tento způsob přináší své výhody. Programy, které využívají periferní zařízení nemusí přistupovat k jejich paměti složitým způsobem skrze IO řídící registry, ale mohou pracovat přímo s jejich obsahem tak, jako by byl součástí paměti RAM. Tyto programy dokonce nemusí ani tušit, že pracují s pamětí umístěnou mimo oblast fyzické RAM.
- *IO prostor* – odpovídá paměťovému prostoru určenému pro IO operace. Programy resp. operační systém obvykle využívaly tento prostor pro přístup k řídícím registrům periferních zařízení (např. sériový port, paralelní port apod.). V moderních systémech se již tento prostor příliš nepoužívá a veškerá paměť periferních zařízení je mapována do globálního paměťového prostoru systému (viz předchozí bod).
- *Konfigurační paměťový prostor* – tento prostor je určen pro uložení konfiguračních informací o jednotlivých PCI zařízení. Každé PCI zařízením připojené v systému má v tomto prostoru vymezenu svou část, kde jsou uloženy například informace o identifikaci zařízení, jeho typu (grafický adaptér, síťový adaptér apod.) a také informace o namapování prostoru adaptéru do globálního paměťového prostoru systému (viz bod 1).

V rámci tohoto modelu a adresového prostoru umožňuje PCI sběrnice komunikovat mezi libovolnými body připojenými na PCI. Díky této vlastnosti je možné velmi efektivně implementovat například:

- *DMA přenosy* – Přenos bloku dat mezi PCI zařízením a hlavní paměti RAM. PCI zařízení inicializuje příslušnou transakci na sběrnici, *North bridge* na tuto transakci odpoví a provede čtení/zápis z/do paměti RAM, která je k němu připojena.
- *Peer-to-peer přenosy* – Přenos bloku dat mezi PCI zařízeními navzájem. Pokud PCI zařízení zná identifikaci/adresu cíle, potom nepotřebuje k přenosu dat CPU a může jej provést samo.

Tato vlastnost nebyla do té doby zcela běžná. Preferoval se model, kdy většinu operací inicializovalo samo CPU a pokud bylo potřeba vykonat DMA operaci, používal se k tomuto účelu specializovaný DMA řadič. S příchodem PCI pozbyl tento centralizovaný DMA řadič svého významu.

2 PCI

2.1 Způsob přístupu ke sběrnici

Na sběrnici PCI může být připojeno více zařízení současně. Aby nedošlo ke kolizi při přístupu ke sběrnici je použit centralizovaný způsob arbitrace. Arbitr je umístěn v *North bridge* a ke každému zařízení jsou přivedeny dva signály s označením REQ, GNT (Request, Grant). Zařízení, které chce přistoupit ke sběrnici musí nejprve nastavit signál REQ. V dalším kroku centrální arbitr rozhoduje o přidělení sběrnice a pokud je volná, povolí zařízení přístup aktivací signálu GNT. Pokud o sběrnici žádá více zařízení současně, je úkolem arbitru rozhodnou komu oprávnění předá. Oddělené signály REQ, GNT umožňují, aby arbitrace mohla probíhat na pozadí právě probíhající transakce. Díky této skryté arbitraci, může po ukončení transakce přistoupit v co nejkratším čase ke sběrnici další zařízení.

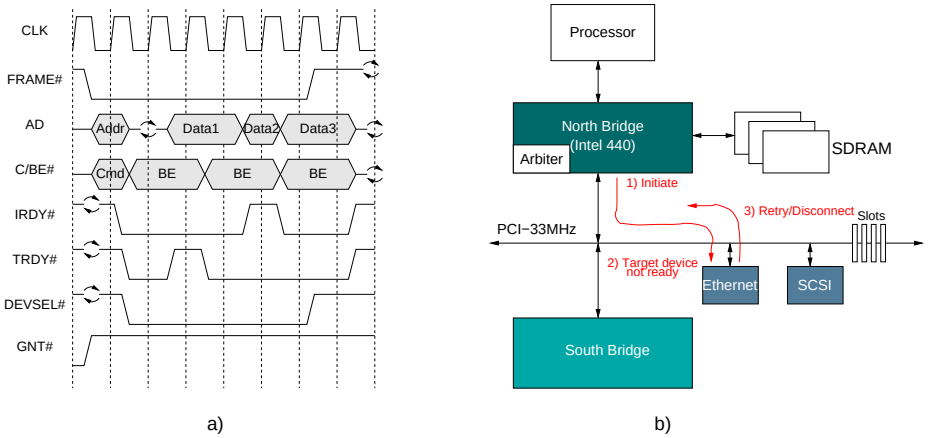
2.2 Komunikační protokol

Pokud chtějí libovolné dvě zařízení na sběrnici PCI komunikovat, potom musí dodržet definovaný protokol. Zařízení, které komunikaci inicializuje je označováno jako *Initiator*. Zařízení, se kterým je komunikace navázána je označováno jako *Target* (cíl). Jakmile *Initiator* získá přístup ke sběrnici, vystaví adresu zařízení, se kterým chce komunikovat. Pokud *Target* rozpozná svou adresu, potvrdí příjem transakce a začíná datový přenos.

Mezi hlavní výhody použitého komunikačního protokolu patří: (1) Adresa je multiplexována s daty. Tento způsob vede na úsporu adresových vodičů, které by jinak musely vést paralelně s datovými. Úspora vodičů má vliv na levnější a rychlejší realizaci PCI sběrnice na desce plošných spojů. (2) Jedná se obecně o blokový přenos dat. Čím více dat se přenáší v rámci jedné transakce, tím menší je režie komunikace a lepé je využita přenosová kapacita sběrnice. (3) *Target* i *Initiator* mají kontrolu nad stavem transakce, oba ji mohou pozastavovat a oba ví, zda byla přerušena. Naopak nevýhodou je skutečnost, že *Target* nikdy neví předem, kolik dat se bude v průběhu transakce přenášet. Tato vlastnost může komplikovat návrh vstupně/výstupních bufferů na straně zařízení.

2.3 Protokol Retry/Disconnect

Protokol *PCI Retry* je použit v situaci, kdy není cílové PCI zařízení (*Target*) schopno přijmout čtecí transakci a zpřístupnit svá lokální data. Tato situace může být způsobena například tím, že požadované data jsou uložena v určité



Obr. 2: a) Komunikační protokol sběrnice PCI, b) Protokol PCI Retry/Disconnect

části adaptéru, která není ihned přístupná. Protokol *PCI Retry* řeší tuto situaci následovně: (1) *Initiator* inicializuje čtecí operaci směrem k *Target* zařízení. (2) *Target* rozpozná svou adresu, potvrdí příjem operace, ale zjistí že nemá připravena požadovaná data. Poznámka: pokud *Target* ví, že bude schopen data připravit do 16 hodinových taktů, potom může protokol pozdržet vložením čekacích stavů, v opačném případě *Target* odloží operací příkazem *Retry* a začne si připravovat požadovaná data. (3) *Initiator* na základě příkazu *Retry* transakci zruší a pokusí se o její založení později.

Podobný princip je použit i v případě, že při čtecí transakci je přečten určitý počet dat, ale v průběhu přenosu *Target*, zjistí, že další data není schopen dodat bez větší časové prodlevy. V tomto případě *Target* rozpojí transakci příkazem *Disconnect* a *Initiator* se pokusí transakci založit po čase znovu avšak od pozice, kde naposledy skončil. Z pohledu přenosové kapacity sběrnice, se protokol *PCI Retry/Disconnect* chová neefektivně, protože při opětovném pokusu o přečtení dat se plýtvá přenosovou kapacitou média, aniž by došlo k přenosu datového slova.

2.4 Princip přerušení

Podobně jako u ostatních typů sběrnicí, také sběrnice PCI podporuje signalizaci přerušení. Tento mechanismus je velmi užitečný signalizaci asynchronních událostí jako jsou například příchod nových dat do adaptéru, dokončení jejich zpracování nebo ukončení jejich blokového přenosu do paměti RAM. Pro signalizaci přerušení používá sběrnice PCI čtyři signály: *INTA*, *INTB*, *INTC* a *INTD*.

Tyto signály jsou přivedeny ke všem zařízením a ty pak přerušení signalizují aktivací příslušného signálu. Řadič přerušení je umístěn s *South bridge*. Jeho úkolem je vyhodnotit prioritu přerušení a zaslat procesoru signál *INTR*. Procesor na základě tohoto signálu přeruší svoji činnost a provede obsluhu přerušení.

Prosím všimněme si, že PCI používá pouze čtyři přerušovací signály pro všechna zařízení. Pokud je potřeba více než čtyři přerušení dochází ke sdílení přerušovacích signálů ze strany několika zařízení. Při obsluze přerušení, pak musí operační systém projít všechna PCI zařízení, která sdílí daný přerušovací vodič a identifikovat, které přerušení skutečně vyvolalo. Pro identifikaci přerušení je tedy generováno několik čtecích transakcí na sběrnici a dochází tak ke plýtvání její přenosové kapacity.

2.5 Obsluha chyb

Pro zajištění ochrany přenosu dat po sběrnici podporuje PCI jednoduchý systém detekce parity. V průběhu každé transakce jednotlivé PCI zařízení kontrolují paritu vystavované adresy a dat. V případě, že je detekována chybná parita, zařízení nastaví signály *PERR* (Parity Error) a *SERR* (System Error). V *South bridge* je umístěna speciální logika, která vyhodnocuje tyto stavy a generuje nemaskovatelné přerušení do procesoru signálem *NMI*. Takto vzniklá chyba je v architektuře na bázi PC identifikována jako velmi kritická a dochází k pádu systému z něhož se nelze zotavit. Na některých typech systému proto bývá tato kontrola vypnuta. Tento způsob detekce chyb je z obecného hlediska nedostačující i sáohledem na to, že parita dokáže odhalit pouze lichý počet chyb.

2.6 Shrnutí nevýhod PCI

Na závěr této kapitoly proveďme krátké shrnutí hlavních nevýhod a nedostatků sběrnice PCI.

- Malá efektivita, využití sběrnice je cca 50 %! Tento problém si lze zjednodušeně představit tak, že z každých 100 hodinových cyklů na sběrnici PCI, se pouze 50 použije pro skutečný přenos dat, zbytek přenosové kapacity je plýtváno na režii sběrnice. Tato režie zahrnuje: režie komunikačního protokolu, vícenásobné opakování přístupu *PCI RetryDisconnect*, čekací stavy, apod.

- Nedostačující výkon: fyzická realizace povoluje max. frekvenci 66 MHz, není možné proto připojit periferie jako jsou 10 Gb Ethernet, výkonné RAID pole, apod. Velký počet vodičů (pro 64bitovou variantu) navíc komplikuje návrh desky a zvyšuje její cenu.

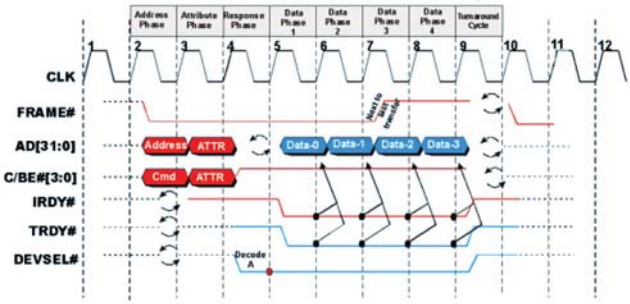
- Při přenosu není definována velikost dat: tato skutečnost komplikuje návrh PCI zařízení a jejich správu bufferu při přenosu dat.
- Zpracování přerušení: sdílené signály pro signalizaci přerušení *INTA..INTD*, driver operačního systému musí zjistit, kdo přerušení vyvolal.
- Správa chyb: parita je nedostačující, při identifikaci chyby se systém zhroutí a není definován způsob zotavení.
- Není podporována technologie *Hot Plug*: připojení/odpojení nového zařízení za chodu. Potřeba například pro serverové systémy, kdy při výpadku PCI zařízení je potřeba jej vyměnit bez odstavení systému z provozu.
- Není podporováno řízení spotřeby (*Power Management*): velmi důležitá vlastnost pro moderní a vestavěné systémy, kde je velký požadavek na nízkou spotřebu systému.

Tyto hlavní nedostatky se v dalších letech snaží odstranit nastupující technologie typu PCI-X a PCI Express.

3 PCI-X

Sběrnice s technologií PCI-X je nástupcem původní technologie PCI. Hlavním cílem tvůrců tohoto typu sběrnice bylo odstranit klíčové nedostatky svého předchůdce a současně co nejvíce zachovat kompatibilitu již s existujícími zařízeními. Přitom máme na mysli kompatibilita, jak na úrovni hardware, tak i na úrovni software. Základní vlastnosti a rozdíly oproti technologii PCI lze shrnout do několika následujících bodů:

- *Změny v komunikačním protokolu* – Komunikační protokol zaznamenal několik klíčových změn:
  - Do protokolu přibyla hned za vystavením adresy fáze *Attrib*, která je použita pro specifikaci délky transakce. Díky této vlastnosti již všechna *Target* zařízení znají velikost přenášených resp. požadovaných dat a jsou schopny lépe spravovat své vnitřní vyrovnávací buffery.
  - Sběrnice PCI-X již nepodporuje čekací stavy v okamžiku, kdy se již začaly přenášet první data. Čekací stavy jsou povoleny pouze mezi fázi rozpoznávání adresy a začátkem přenosu dat. Navíc minimální velikost přenášených dat je pro PCI-X 128 bajtů.
  - Protokoly *PCI Retry/Disconnect* byly zrušeny a místo nich je definován nový efektivnější způsob tzv. *rozdělených transakcí*. Hlavní



Obr. 3: Komunikační protokol sběrnice PCI-X

myšlenkou tohoto principu je, že *Initiator* se již opakovaně nedotazuje na dostupnost dat, ale pošle *Target* čtecí požadavek pouze jednou na začátku. *Target* si data připraví a sám je pošle zdrojovému uzlu. Součástí odpovědi je také příznak (*Tag*), podle kterého zdroj rozpozná a jaký požadavek se jednalo. Tímto způsobem se již neplýtvá přenosovou kapacitou sběrnice.

Efektivita přenosu dat: Díky výše uvedením opatřením stoupla průměrná efektivita využití přenosové kapacity sběrnice až na 85 %.

- *Obsluha přerušení* – pro signalizaci přerušení musí zařízení typu PCI-X podporovat tzv. *MSI protokol*. Tento protokol již není založen na sdílených signálech *INTA-INTD*, ale místo toho, každé zařízení signalizuje přerušení zápisem speciální zprávy na adresu řadiče přerušení umístěného v *South bridge*. Každé zařízení si může alokovat několik přerušovacích vektorů a operační systém již nemusí prohledávat jednotlivá zařízení, aby zjistil které z nich přerušení vyvolalo.

- *Obsluha chyb* – zařízení PCI-X již podporují systém ECC. Vznik jednobitových chyb je automaticky opravován. Ze vzniku vícenásobných chyb je pak dále definován způsob zotavení bez pádu celého systému.

- *Fyzická úroveň přenosu dat* – na fyzické úrovni bylo provedeno také několik zásadních změn. Nejprve všechny signály sběrnice jsou registrovány tzn. že mezi samotný signál a pinem je vložen registr. Vložení tohoto registru způsobí, že délka vodiče se fyzicky zkrátí a sběrnice může pracovat na výrazně vyšší frekvenci. Kromě frekvence má toto registrování vliv i na počet slotů, který v základní verzi PCI-X 1.0 vzrostl na 4 sloty na sběrnici.

Poslední výrazné zrychlení přinesla specifikace PCI-X 2.0, která povolila použití technologie *DDR* (Double Data Rate) resp. *QDR* (Quad Data Rate). Tyto

technologie umožňují přenos až dvou resp. čtyř datových slov v rámci jednoho taktu hodinového taktu Zvýšení rychlosti má však negativní vliv na počet připojených slotů v rámci jedné sběrnice. Specifikace PCI-X 2.0 povoluje maximálně 1 konektor na jedné sběrnici. Pokud je potřeba připojit více slotů, musí být od sebe odděleny skrze přídavný *PCI bridge*.

4 PCI Express

PCI Express je vásoučasné době nejmodernějším sběrnicí zárodiny PCI. Oproti předchozím typů PCI a PCI-X, již není založena na širokých paralelních sběrnicích, ale na vysokorychlostních plně-duplexních sériových linkách. Vájeden okamžik je tedy možné komunikovat oběma směry současně. Rychlost každé linky je v základní verzi 2.5 Gb/s, pro oba směry celkem 5 Gb/s. Tyto plně duplexní linky lze navíc paralelně řadit vedle sebe a vytvářet postupně sběrnici sávyšší přenosovou kapacitou. Současná norma podporuje zapojení až 32 linek vedle sebe.

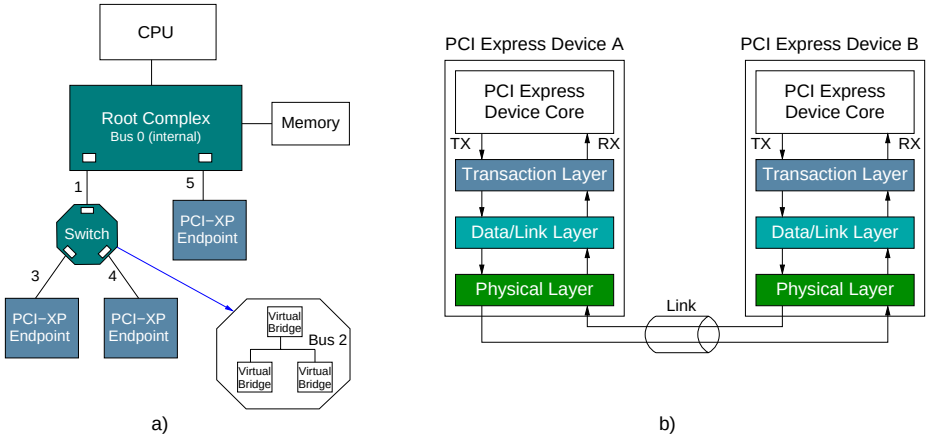
Komunikace na sběrnici PCI Express probíhá skrze paketovou komunikaci. Pokud je spoj mezi dvěma zařízeními tvořen více než jednou linkou, jsou data paketu rovnoměrně zasílána do jednotlivých linek. Paket je rozprostřen do čtyř linek tak, že první bajt je zaslán linkou 0, druhý linkou 1, třetí linkou 3 atd. až do ukončení přenosu. V okamžiku, kdy se PCI Express zařízení připojí do systému, musí se sádruhou stranou dohodnout na způsobu komunikace, tzn. počtu linek, rychlosti apod. Zátohoto důvodu musí každé PCI Express zařízení povinně podporovat alespoň komunikaci po jedné lince.

Základní topologie sběrnice PCI Express je znázorněna na obrázku 4. Topologie je organizována do stromu, kde kořenový uzel tvoří prvek sánázvem *Root Complex*. Tento prvek zastává funkci *North* a *South bridge* a dále propojuje prvky sápožadavkem na vysokou propustnost dat (CPU a RAM). Prvky válistových uzlech stromu jsou označovány jako *Endpointy* a reprezentují jednotlivá koncová zařízení. Poslední částí topologie jsou přepínače tzv. *Switch* komponenty, které tvoří hierarchii stromu a směrují provoz do jednotlivých větví.

Každá transakce na sběrnici reprezentuje tok dat složený sájednoho nebo více paketů o maximální velikost 4 kB. Podobně jako u síťových technologií, je však potřeba, aby se v rámci topologie všechny prvky dohodly, jakou maximální velikost paketu (MTU) budou používat. Tato velikost má přímý vliv na velikosti bufferů podél celé infrastruktury a tudíž i cenu. V současných systémech je velikost MTU obvykle nastavována na hodnoty 128 nebo 256 bajtů.

I ostatní základní funkce byly v této generaci sběrnice PCI dále vylepšeny. Výpis těch nejvýznamnějších je následující:

- *Obsluha přerušení* – Způsob signalizace přerušení probíhá pomocí protokolu MSI podobně jako u sběrnice PCI-X. Přerušení se jednoduše signali-



Obr. 4: a) Topologie sběrnice PCIe, b) Vrstvový model sběrnice PCIe

zuje zasláním paketu obsahující informaci, které zařízení přerušení vyvolalo včetně vektoru přerušení.

- *Řízení spotřeby (Power Management)* – Oproti předchozím technologiím, PCI Express umožňuje regulovat spotřebu jak pro linky, tak pro samotné zařízení. Řízení spotřeby lze ovládat buď skrze software (zasláním správy) nebo plně automaticky. Např. váokamžiku, kdy neprobíhá na daném spoji komunikace po delší dobu, přepne se linka do režimu sánižší spotřebou. Podobně lze uspávat i jednotlivá zařízení. Pro řízení spotřeby linky se používají čtyři stavy: L0, L1, L2 a L3. Pro zařízení pak režimy: D0, D1, D2, D3-Hot a D3-Cold.

- *Správa chyb* – Každý paket je opatřen CRC kódem sávelmi dobrou zabezpečovací schopností. Vzniklé chyby se zapisují do logovacích registrů a systém se záchyb zotavuje na několika úrovních bez zásahu systémového software.

- *Hot Plug* – V průběhu činnosti systému lze připojovat/odpojovat nové PCI zařízení. Pro tyto účely je na desce ke každému slotu speciální tlačítko a dvojice LED diod pro signalizaci stavu napájení. Tato vlastnost je důležitá např. pro servery nebo systémy, které jsou neustále v provozu (je potřeba vyměnit porouchané zařízení za běhu).

Všechny tyto funkce byly logicky rozděleny do několika vrstev, podobně, jak je tomu např. u síťových zařízeních (podle modelu ISO/OSI), kde každá závrstev je zodpovědná za určitou funkci a mezi vrstvami je přesně definované

komunikační rozhraní. Vrstvový model sběrnice PCI Express je tvořen třemi vrstvami:

1. *Fyzická vrstva* – realizuje přenos dat po lince na nejnižší úrovni, zahrnuje digitální i analogovou část. Mezi hlavní funkce patří kódování 8/10, vyrovnání zpoždění dat mezi jednotlivými linkami a inicializační a trénovací proces linek.

2. *Linková vrstva* – stará se o spolehlivý přenos paketu mezi dvěma sousedními uzly. Pro tyto účely je každý paket rozšířen o CRC kód a sekvenční číslo. V případě detekce chyby je paket automaticky přeposlán znovu, aniž by do toho musel zasahovat systémový software.

3. *Transakční vrstva* – řídí přenos paketu mezi libovolnými uzly na nejvyšší úrovni. Zajišťuje směrování paketů, dodržení kvality služeb (QOS), Flow control a mnoho dalších vlastností.

5 Shrnutí

Pokud shrneme vlastnosti první generace sběrnice typu PCI, je potřeba poznamenat, že byla poměrně úspěšnou sběrnicí, zejména s ohledem na dříve požívané standardy (např. ISA). PCI se vyznačovala efektivnějším využitím přenosové kapacity vodičů díky multiplexování adresy a dat, což navíc zjednodušilo a zlevnilo výrobu desky plošného spoje. DMA přenosy jsou realizovatelné přímo v základním protokolu bez nutnosti existence specifického DMA řadiče. Arbitrace o sběrnici probíhá skrytě a neukusuje z přenosové kapacity, apod. I přes všechny tyto přínosné vlastnosti je PCI považována za sběrnici s nízkou efektivitou přenosu dat (cca 50 %), která je způsobována vkládáním čekacích stavů, neefektivním protokolem *PCI Retry/Disconnect*, obsluhou sdílených přerušení apod. Seznam nevýhod pak uzavírá velmi nízká spolehlivost sběrnice zabezpečené pouze paritou, bez možnosti zotavení z chyb.

Tvůrci druhé generace sběrnice s označením PCI-X se snažili odstranit téměř všechny nevýhody předchozího typu. Zdokonalen byl komunikační protokol, byly odstraněny čekací stavy, *PCI Retry/Disconnect* nahradil mnohem efektivnější model rozdělených transakcí, byla odstraněna sdílená přerušení a sběrnice se stala odolnější proti chybám s alespoň základní možností zotavení. Nakonec se ale ukázalo, že se jednalo spíše o slepou větev, která záplatovala nevýhody svého předchůdce. Její vývoj byl ukončen především proto, že díky použité technologii, paralelní sběrnice řízené hodinovým signálem, nebylo možné dále zvyšovat přenosovou kapacitu sběrnice a reagovat tak na požadavky nových zařízení.

Nová generace s označením PCI Express byla založena na zcela nové technologii přenosu dat využívajících vysokorychlostního přenosu po sériových linkách.

Přenos dat již nebyl synchronizován vzhledem k centrálnímu hodinovému signálu, ale tento hodinový signál byl zakódován přímo do přenášených dat. Tento princip způsobil výrazné zvýšení přenosové kapacity linek sběrnice. Zavedením paketově orientovaného přenosu se navíc snížil počet potřebných vodičů a většina důležitých funkcí sběrnice (přerušení, Hot Plug, power management, obsluha chyb, apod.) bylo možné jednoduše rozlišit pouze na základě typu použitého paketu. Jedinou nevýhodou sériového přenosu může pro některé typy aplikací být zvýšená latence přenášených dat. Tuto latencí však lze, do jisté míry, skrýt například využitím hierarchie pamětí cache.

Poděkování

Autor by rád poděkoval za podporu výzkumnému záměru s označením MSM, 6383917201 – Optická síť národního výzkumu a její nové aplikace.

Reference

[1] Ravi Budruk, Don Anderson, and Ed Solari. *PCI Express System Architecture*. Pearson Education, 2003.

[2] Tom Shanley. *PCI-X System Architecture with CD*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2000.

[3] Tom Shanley and Don Anderson. *PCI System Architecture*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

Ethernet 40/100 Gb

Jiří Novotný

E-mail: novotny@ics.muni.cz

Klíčová slova: Ethernet, IEEE 802.3, FPGA

Abstrakt

Článek se bude zabývat evolucí vývoje ethernetu od rychlosti 10 Mb/s až po 100 Gb/s. Budou vysvětleny základní principy, srovnány jednotlivé vývojové stupně z pohledu návrhu architektury a podrobně rozebrán nejnovější návrh 40/100 Gb/s. Součástí bude i diskuse možnosti realizace a návrhu adaptoru ethernetového adaptoru pro rychlost 40/100 Gb/s na bázi obvodů FPGA a specializovaných obvodů rozhraní.

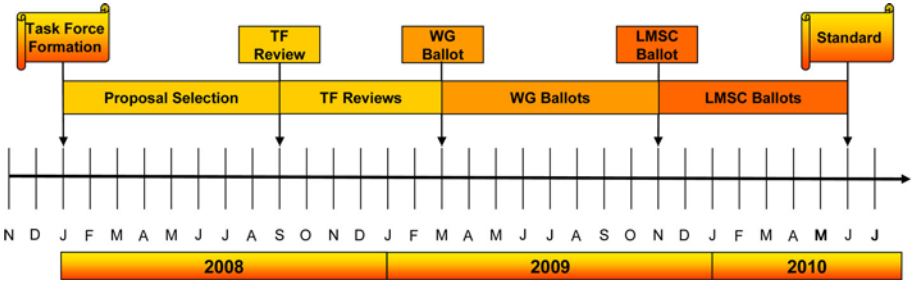
1 Úvod

Ethernet [4] je v současné době nejrozšířenější používanou síťovou technologií. Původní nasazení bylo určeno do lokálních sítí (LAN) později se začal používat v městských sítích (MAN). V dnešní době je používán i v rozlehlých sítích (WAN), kde stále více nahrazuje původně dominující technologii SDH/SONET. V návrhu 1 GbE a vyšších verzích je uvažováno i použití pro komunikaci jednotlivých bloků uvnitř výpočetních a komunikačních systémů.

První verze ethernetu vznikla v letech 1973–1975 ve firmě XEROX. V roce 1980 vyšel návrh standardu ethernetu a konečný návrh byl schválen jako IEEE 802.3 v roce 1982. Původní návrh ethernetu využíval koaxiální kabel a metodu detekce kolizí CSMA/CD na rychlosti 10 Mb/s. V roce 1983 uvedla firma 3COM na trh kartu 10 MbE.

V polovině 80. let byl uveden standard ethernetu na kroucené dvojlince (4 páry) na konektoru RJ45. Koncové stanice v tomto případě nejsou připojeny na jeden kabel, ale propojeny bod-bod do HUBu.

V roce 1995 vydalo IEEE standard 802.3u 100 MbE s autonegociací, kde obě propojené strany navzájem vyjednávají nejlepší možný způsob komunikace



Obr. 1: Časový harmonogram standardizace 40/100 GbE

(10 Mb/s oproti 100 Mb/s a halfduplex oproti fullduplexu). 1 GbE verze ethernetu (včetně 1000BASE-TX) byla standardizována v roce 1999 – IEEE 802.3ab a 10 GbE v roce 2002 jako standard IEEE 802.3ae.

V prosinci roku 2007 byl zahájen proces standardizace 40/100 GbE – IEEE 802.3ba, který má být dokončen v polovině roku 2010. Poprvé v historii obsahuje návrh dvě rychlosti – 40 Gb/s a 100 Gb/s. Při přípravě standardu probíhaly diskuse, která rychlost bude pro novou generaci výhodnější, v konečné verzi jsou v návrhu standardu obě rychlosti.

Robert Metcalf, jeden z duchovních otců ethernetu předpokládá, že terabitový ethernet bude dostupný okolo roku 2015 [8].

Přestože rychlost ethernetu vzrostla od počátku o pěr řádů a jsou používána rozdílná fyzická média, formát ethernetového rámce zůstává v podstatě stále stejný. V současné době jsou nejrozšířenější varianty ethernetu s konektorem RJ45 (na kroucené dvojlince) v lokálních sítích a na optických vláknech pro propojení na větší vzdálenosti.

2 Architektura ethernetu

V OSI modelu pokrývá ethernet první (fyzickou) a druhou (datovou) vrstvu. V článku se zaměříme na podrobný popis fyzické vrstvy.

2.1 Datová vrstva

Data z datové vrstvy přicházejí do MAC bloku, který formátuje data do ethernetového rámce a provádí řízení fyzické vrsvy. Formát nejpoužívanějšího typu rámce je následující:

Datové rámce jsou navzájem odděleny mezipaketovou mezerou.

Tabulka 1: Formát ethernetového rámce

počet byte	význam
7	preambule
1	start symbol
6	cílová adresa
6	zdrojová adresa
4	případný VLAN tag
2	délka/typ
42–1 500	data
4	kontrolní součet

Tabulka 2: Struktura fyzické vrstvy 10 MbE až 100 GbE

10 Mb/s	**100 Mb/s**	**1 Gb/s**	**10 Gb/S**	**40 Gb/s**	**100 Gb/s**
MAC	MAC	MAC	MAC	MAC	MAC
REC	REC	REC	REC	REC	REC
MII	*MII*	*GMII*	*XGMII*	*XLGMII*	*CGMII*
PLS	PCS	PCS	PCS	PCS	PCS
AUI	PMA	PMA	PMA	PMA	PMA
PMA	PMD	PMD	PMD	PMD	PMD
MDI	*MDI*	*MDI*	*MDI*	*MDI*	*MDI*
MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM

kde:

MAC	Media Acces Control – datová vrstva
REC	Reconciliation – mapování signálů datové na fyzickou vrstvou
MII	Media Indipendent Interface – rozhraní nezávislé na typu média
GMII	Gigabit MII
XGMII	10 Gigabit MII
XLGMII	40 Gigabit MII
CGMII	100 Gigabit MII
AUI	Attachment Unit Interface – rozhraní pro připojení média
PLS	Physical Layer Signaling – kódování u 10 MbE
PCS	Physical Coding Sublayer – kódování znaků
PMA	Physical Medium Attachment sublayer – kódování pro dané médium
PMD	Physical Medium Dependent sublayer – připojení na médium
MDI	Medium Dependent Interface – rozhraní zpro připojení média

2.2 Fyzická vrstva

V následující části se zaměříme na nejdůležitější aspekty jednotlivých verzí ethernetu. Vlastnosti 10 MbE, 100 MbE a 1 GbE budou uvedeny pouze stručně. Podrobněji bude popsána verze 10 GbE, jejíž vlastnosti jsou důležité pro popis verze 40/100 GbE.

Základní rozdělení fyzické vrstvy pro rychlosti od 10 Mb/s do 100 Gb/s je v tabulce 2.

2.2.1 10 MbE

První verze ethernetu byla navržena pro rychlost 10 Mb/s. Struktura fyzické vrstvy je jednodušší než v dalších verzích. Použité kódování je typu Manchester s využitím frekvenčního pásma na 50 %. Používaná fyzická média jsou typu koaxiální kabel (několik typů), 10BASE-T na konektoru RJ45 a multimódové optické vlákno 10BASE-F. Konektory pro tato média jsou buď vyvedeny z karty rozhraní (v pozdějších verzích) nebo je použit AUI konektor, na který je připojen převodník s konektorem pro dané rozhraní.

2.2.2 100 MbE

Ve verzi 100 MbE je model fyzické vrstvy poněkud složitější a zůstáva s výjimkami, které budou popsány u dalších verzí, v podstatě stejný. Kódování znaků zabezpečuje subvrstva PCS, převod pro dané médium subvrstva PMA a vlastní připojení k médiu subvrstva PMD. Kódování u nejpoužívanější verze 100BASE-X je 4B/5B s využitím přenosového pásma 80 %. V praxi se nejvíce prosadila verze s konektorem RJ45 – 100BASE-TX a optické vlákno (monomód i multimód) 100BASE-FX.

2.2.3 1 GbE

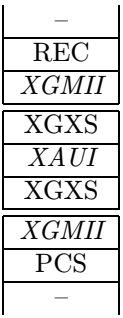
U 1 GbE je použito pro přenos na optických vláknech složité kódování 8B/10B s využitím pásma 80 %. Toto kódování zabezpečuje signál bohatý na hrany, což je důležité pro obnovu hodin signálu a stejný počet jedniček a nul, což je nutné pro zachování střední úrovně signálu. Nejpoužívanější média jsou 1000BASE-SX (optika multimód) 1000BASE-LX (optika monomód). Pro přenos na kroucené dvojlince 1000BASE-T je použito kódování 4D-5PAM, data jsou přenášena po všech 4 párech obousměrně s použitím potlačení ozvěny (echo cancelation). Základem tohoto systému je odečtení známé úrovně vysílaného signálu od skutečně přijatého signálu. Moderní karty umožňují použít na jednom konektoru RJ45 rychlosti 10 Mb/s, 100 Mb/s a 1 Gb/s. Autonegociace je u 10 MbE a 100 MbE doporučená, u 1 GbE je povinná. Vlastní transceiver je buď součástí desky rozhraní nebo je použita klec (GBIC, SFP), do které je vložen. Tato klec je připo-

jena na rozhraní 1000BASE-X a je možno do ní vložit buď optický transceiver nebo transceiver s konektorem RJ45. V případě RJ45 je v tranceiveru relativně komplikovaný obvod pro převod 1000BASE-X na 1000BASE-T.

2.3 10 GbE

U 10 GbE verze dochází ke změně při mapování dat z datové do fyzické vrstvy. V předchozích verzích je v rozhraní MII, respektive GMII, použito 8 datových bitů, hodinový signál a řídící signály. Ve verzi 1 GbE jsou data taktována na rychlosti 125 Mhz, což je frekvence, kterou dnešní obvody zvládají bez problému. Při zachování stejného rozhraní by v případě 10 GbE rychlost taktování musela stoupnout na 1.25 Ghz, což je pro dnes používané obvody příliš vysoká frekvence. Proto byla definována šířka dat na 32 bitů se čtyřmi kontrolními signály (pro každý byte jeden, kdy „0“ znamená data a „1“ kontrolní znak). Hodiny mají takt 156.25 Mhz DDR (Double Data Rate – data jsou přenášena na nástupné i sestupné hraně hodin). Vzhledem k potřebnému počtu vodičů pro propojení integrovaných obvodů rozhraním XGMII je v 802.3 standardu definována volitelná subvrstva XGMII extender (viz tabulka 3).

Tabulka 3: Rozšíření fyzické vrstvy o XAUI rozhraní



kde:

XGCS	XGMII Extender Sublayer – připojení rozšiřující subvrstvy
XAUI	rozšiřujici rozhraní

Rozšiřující rozhraní XAUI (tabulka 4) je tvořeno 4 páry diferenciálních vodičů (pro každý směr). Pro přenos dat je použit sériový přenos s rychlostí 3.125 Gb/s s kódováním 8B/10B (je použit stejný kódovací mechanismus jako u 1 GbE). Použití XAUI přináší úsporu vodičů (16 oproti 74 u XGMII), značné

zjednodušení plošného spoje a snížení přeslechů. Na druhé straně je potřeba zabezpečit zarovnání (alignment) všech čtyř kanálů a nutnost použít specializovaný vysokorychlostní obvod (ten může být i uvnitř většího obvodu).

Mimo nejběžnější variantu 10GBASE-R, která bude popsána dále, jsou definovány 10GBASE-LX4 (4 vlnové délky, kódování 8B/10B) a 10GBASE-W (mezi PCS a a PMA je vložena WAN subvrstva).

10GBASE-R používá kódování 64B/66B (podstatně jednodušší, než 8B/10B) s využitím přenosového pásma přes 96 % a následný scrambling polynomem:

$$G(x) = 1 + x^{39} + x^{58}$$

Tabulka 4: Srovnání XGMII a XAUI rozhraní

XGMII	D7..D0/C0	D15..D8/C1	D23..D16/C2	D31..D24/C3	CLK
XAUI	Lane 0	Lane 1	Lane 2	Lane 3	–

V případě použití XAUI jsou jednotlivé skupiny (data + řídící signál) mapovány do příslušného sériového kanálu.

Začátek každého rámce musí být zarovnán na Lane 0, obsahuje Start symbol (0xFB/1) a preambuli. Ta je tvořena 6ti znaky 0xAA/0 a znakem 0xAB/0. Úvodní část rámce má 8 byte přenesených v jednom taktu hodin (je použit přenos DDR). Data začínají opět na Lane 0.

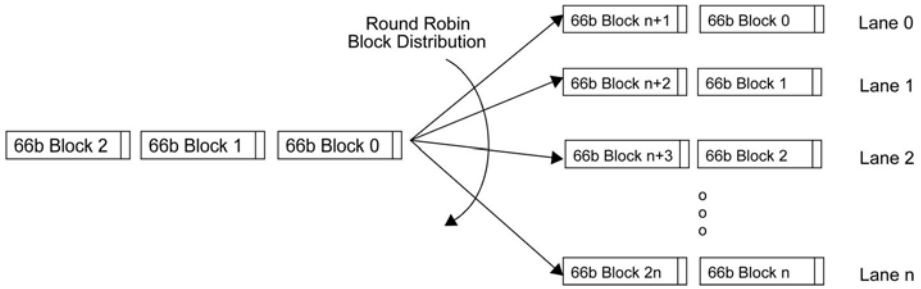
Tabulka 5: Struktura preambule

Lane 0	Lane 1	Lane 2	Lane 3
1111 1011 /1	1010 1010 /0	1010 1010 /0	1010 1010 /0
1010 1010 /0	1010 1010 /0	1010 1010 /0	1010 1011 /0

Rámec je ukončen znakem Konec (0xFD/1) a následuje mezipaketová mezera. Stejně jako u 1 GbE mohou být transceivery buď součástí desky rozhraní nebo jsou v klecích. Na rozdíl od 1 GbE je variant transceiverů mnohem více. Mezi nejpoužívanější patří XENPACK, XFP a SFP+.

2.4 40/100 GbE

Koncem roku 2007 začal proces standardizace zatím poslední verze ethernetu, který má být dokončen v polovině roku 2010. Na rozdíl od předchozích verzí



Obr. 2: Rozdělení do kanálů [5]

se jedná o dvě rychlosti 40 Gb/s a 100 Gb/s. Zatímco konstrukce 40 GbE je v současných technologických možnostech (návrh řešení bude popsán v další kapitole), konstrukce 100GbE bude mnohem obtížnější.

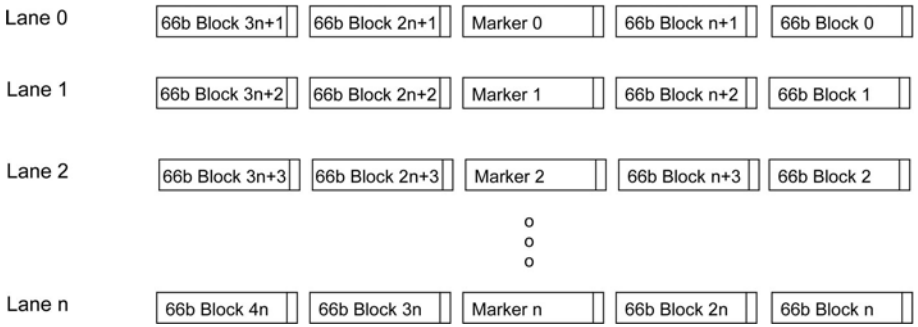
Na úrovni mapování datové vrstvy do fyzické došlo ke změně proti předchozí 10 GbE verzi. Rozhraní má nyní šířku 8 byte s příslušnými kontrolními signály a hodinami. Pro verzi 40 GbE (rozhraní XLGMII) je potřeba takt hodin 312.5 Mhz DDR, případně lze použít širší datovou sběrnici (128 b s taktem 156.25 Mhz), což je realizovatelné i v dnešních FPGA obvodech. Pro verzi 100 GbE (rozhraní CGMII) je při šířce 64 bitů taktovací frekvence 781.25 Mhz DDR. Je také možno použít širší datovou sběrnici, ale obě řešení již naráží na limity současných obvodů.

Tabulka 6: Rozhraní XLGMII a CGMII

D7..D0/C0	D15..D8/C1	...	D63..D55/C7	CLK

Kódování znaků (64B/66B) a scrambling ve vrstvě PCS jsou převzaty z 10 GbE, poté následuje rozdělení do bloků. V případě 40 GbE do 4, u 100 GbE do 20 (obr. 2). Každý blok zpracová nezávisle 64 bitů z XLGMII/CGMII. Pro synchronizaci jednotlivých bloků se používají datové značky (obr. 3).

Po zarovnání do bloků dochází ve vrstvě PMA k přemapování linek z PCS vrstvy do linek výstupního média. Pro jednodušší propojení obvodů je definován Attachment Unit Interface (XLAUI pro 40 GbE a CAUI pro 100 GbE). Fyzická realizace vlastního rozhraní 40 GbE je v návrhu standardu definována pomocí 4 fyzických linek, u 100 GbE je doporučeno ještě řešení s deseti fyzickými linkami. Mimo tato mapování je možno použít i některá další (například mapování na jeden fyzický kanál), jejichž realizace zatím může být mimo realné technické možnosti. Na kratší vzdálenosti je možno použít více optických vláken, na delší se počítá s DWDM – použití několika vlnových délek na jednom optickém vlákně.



Obr. 3: Datové značky [5]

Tabulka 7: Rozdělení do bloků

XLGMII	D63..D0	D127..D64	D191..D128	D255..D192
Kódování a scrambling	B65..B0	B131..B66	B197..B132	B263.. BD198
Blok	Block 0	Block1	Block2	Block 3
CGMII	D63..D0	D127..D64	...	D1279..D1216
Kódování a scrambling	B65..B0	B131..B66	...	B1319..D1254
Blok	Block 1	Block2	...	Block 19

Tabulka 8: Podrobný popis fyzické vrstvy 40/100GbE

Reconciliation	Reconciliation
XLGMII	*CGMII*
40GBASE-R PCS	100GBASE-R PCS
PMA(4:4)	PMA(20:10)
XLAUI	*CAUI*
PMA(4:4/1)	PMA(10:10/4)
PMD	PMD
MDI	*MDI*
40GBASE-R	100GBASE-R

kde:

PMA(4:4)	PMA přemapování 40GbE
PMA(20:10)	PMA přemapování 100GbE
XLAUI	40 Gigabit AUI – 4 linky
CGMII	100 Gigabit AUI – 10 linek
PMA(4:4/1)	volitelné přemapování do 4 nebo jedné fyzické linky
PMA(20:10/4)	volitelné přemapování do 10 nebo 4 linek

Tabulka 9: Fyzická média pro 40/100 GbE

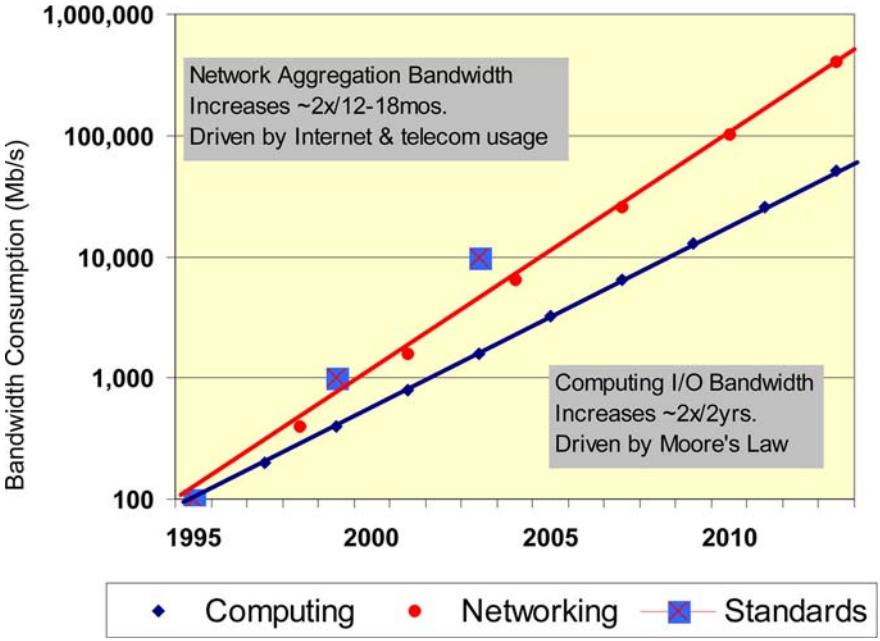
Nejméně	40 Gigabit Ethernet	100 Gigabit Ethernet
1m backplane	40GBASE-KR4	
10 m kabel	40GBASE-CR4	100GBASE-CR10
100 m optika MM	40GBASE-SR4	100GBASE-SR10
10 km optika SM	40GBASE-LR4	100GBASE-LR4
40 km optika SM		100GBASE-ER4

3 Návrh adaptoru 40 GbE

Návrh hardwarově akcelerovaných síťových karet má na CESNETu ve spolupráci s Masarykovou univerzitě a VUT v Brně dlouhou tradici [3]. První karta COMBO6 [1] byla vyvinuta v roce 2002, karta interface pro 10 GbE – COMBO-2XFP byla vyvinuta v rámci projektu SCAMPI [7] na jaře roku 2004. V roce 2008 byla navržena nová rodina karet COMBOv2, která umožňuje zpracovávat rychlosti v oblasti do 100 Gb/s. COMBOv2 [2] je vývojová platforma, sestávající se z výkonné základní karty a karty rozhraní. V současné době je k dispozici základní karta COMBO-LXT a karty rozhraní COMBOI-1G4 (4×1 GbE) a COMBOI-10G2 (2×10 GbE). Ve vývoji je karta rozhraní COMBO-10G4TXT (4×10 GbE), která je osazena obvodem VIRTEX XC5VTX150T a obvody rozhraní 10 GbE AEL 2005.Tuto kartu lze použít jako „běžnou“ hardwarově akcelerovanou kartu se 4 10 GbE rozhraními nebo ji lze použít k prvotním experimentům se 40 GbE. Z architektury 40 GbE vyplývá, že lze použít 4 nezávislé 10 GbE kanály, s datovými synchronizačními značkami. Použité obvody 10 GbE rozhraní tyto značky sice nepodporují, ale přesto bude možné provést alespoň některé základní testy na krátkou vzdálenost. Plnohodnotné řešení pro 40GbE závisí na dostupnosti vhodných obvodů rozhraní a také na konečné verzi standardu IEEE 802.3ba. Předpokládaná realizace této verze je v roce 2010.

4 Závěr

V historii vývoje ethernetu je vidět zajímavý zlom. Fyzická vrstva ethernetu je do rychlosti 1 Gb/s čistě sériová (s výjimkami u rozhraní na kroucené dvojlince) se šířkou rozhraní do datové vrstvy 1 byte. U vyšších verzí dochází ke zvýšení šířky rozhraní fyzické vrstvy na 4 (10 GbE) případně 8 (40/100 GbE) byte. Současně je také standardizováno použití většího množství přenosových optických kanálů u 10 GbE (10GBASE-LX4), i když se v praxi neprosadilo. V případě 40/100 GbE jsou zatím v návrhu standardu doporučovány varianty



Obr. 4: Srovnání propustnosti sítí s rychlostí výpočetních prvků [6]

se 4, případně 10 optickými kanály. Přenos po jednom optickém kanálu je sice ve standardu uvažován ale jeho fyzická realizace bude velice obtížná.

Je vidět, že se stále rostoucí rychlostí začíná přechod na vyšší šířku datových sběrnic a zejména na sériově-paralelní přenos i po optických vláknech. Tento vývoj koreluje i s rychlostí taktování procesorů, která v několika posledních letech zůstává stejná. Obecně je nyní zvyšování výkonu výpočetních systémů závislé na paralelním zpracování, což je poněkud problematické s ohledem na zaběhnuté uvažování vývojářů. Vzhledem k tomu, že zvyšování rychlosti výpočetních prvků je blízko fyzikálního limitu, zatímco zvyšování počtu tranzistorů v integrovaných obvodech stále stoupá, je potřeba aby vývojáři začali více využívat paralelního zpracování, což může vést ke značným změnám v používaných algoritmech.

Situace v oblasti přenosu dat je ještě komplikovanější protože probíhá rozevírání nůžek mezi rostoucími požadavky na přenosové kapacity (Gilderův zákon) a výpočetní kapacitou (Moorův zákon) viz obrázek 4.

Vývoj ethernetu jistě přinese v příštích letech nové otázky a jejich řešení, zejména v oblasti optických systémů a nasazení uvnitř výpočetních a komunikačních systémů. Bude také zajímavé sledovat, zda se naplní předpověď Roeberta Metcalfa o příchodu 1 TbE v roce 2015.

References

[1] CESNET, z.s.p.o. *Description of COMBO Cards*. http://www.liberouter.org/hardware.php.

[2] CESNET, z.s.p.o. *Description of COMBOv2 Cards*. http://www.liberouter.org/hardware.php?flag=2.

[3] CESNET, z.s.p.o. *Official Web Pages of Liberouter Project*, 2008. http://www.liberouter.org.

[4] IEEE. *IEEE 802.3-2008*, 2008. http://standards.ieee.org/getieee802/802.3.html.

[5] John D'Ambrosia, David Law, Mark Nowell. *40 Gigabit Ethernet and 100 Gigabit EthernetTechnology Overview*, 2008. http://www.ethernetalliance.org/files/static page files/ 83AB2F43-C299-B906-8E773A01DD8E3A04/40G 100G Tech overview(2).pdf.

[6] Mark Nowell, Vijay Vusirikala, Robert Hays. *Overview of Requirements and Applications for 40 Gigabit and 100 Gigabit Ethernet*, 2007. http://www.ethernetalliance.org/files/static page files/ D13DCE87-1D09-3519-AD13E838D3CB0181/126 OVERVIEW AND APPLICATIONS2.pdf.

[7] SCAMPI project. *Official Web Pages of SCAMPI Project – IST-2001-32404*, 2005. http://www.ist-scampi.org.

[8] wikipedia. http://en.wikipedia.org/wiki/Ethernet.

NetFlow, monitorování IP toků a bezpečnost sítě

Jan Vykopal

E-mail: vykopal@ics.muni.cz

Klíčová slova: NetFlow, bezpečnost, incident, anomálie, IDS, CSIRT

Abstrakt

NetFlow je dnes synonymem pro monitorování síťových toků. Tato původně proprietární technologie firmy Cisco je nyní uvolněna a rozvíjena světovou komunitou, popsána v několika RFC a zažívá velký rozmach.

Příspěvek se zaměřuje na využití trvalého monitorování síťových toků v práci bezpečnostního týmu organizace. Na příkladu nasazení na Masarykově univerzitě budou předvedeny výhody a přínosy této technologie. Taktéž bude představena potřebná hardwarová a softwarová infrastruktura včetně volně dostupných nástrojů.

1 Úvod

Úvod do monitorování toků a technologie NetFlow včetně problematiky sběru a uchovávání provozních záznamů popsal Tomáš Košňar v [1]. Autor shrnuje své zkušenosti s nasazením v páteřní síti CESNET2.

Cílem tohoto příspěvku je seznámit čtenáře s možnostmi dalšího využití nasbíraných dat zejména pro detekci anomálií a průniků. I když je síť Masarykovy univerzity připojena právě do akademické sítě CESNET2, sběr a ukládání dat je realizováno jinak než v případě sítě CESNET2. Navíc je přítomna nová vrstva celého řetězce: automatizované zpracování uložených dat.

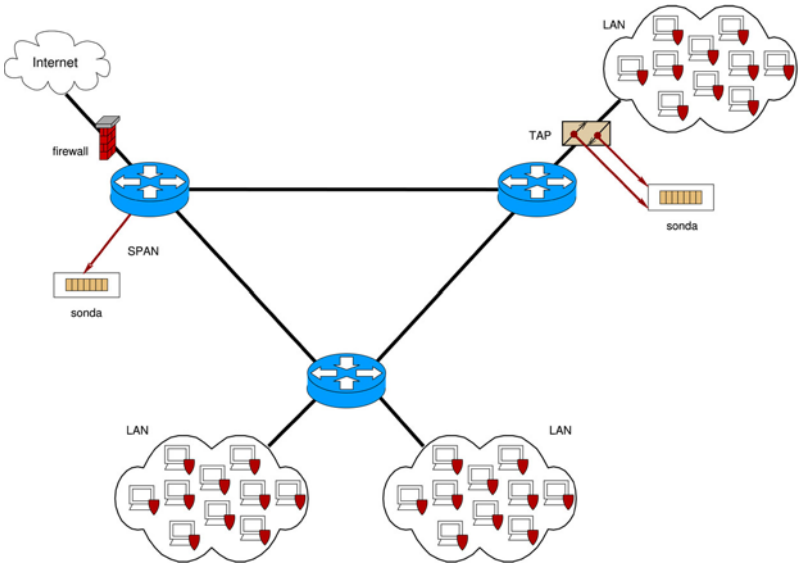
2 Monitorování IP toků a bezpečnost sítě

Každý bezpečnostní tým musí zajišťovat základní službu zvanou *incident handling*. Ta zahrnuje reakce na nahlášené bezpečnostní incidenty od třetích stran (např. od národního CSIRT nebo týmu jiné organizace) nebo přímo od zaměstnanců či jiné jednotky organizace. Reakce spočívá ve vyšetření incidentu, v zmírnění dopadu, oznámení výsledku ohlašovateli a poučení se z incidentu např. v podobě úpravy bezpečnostní politiky.

Vyšetřování nahlášeného bezpečnostního incidentu může být náročné nejen časově, ale i organizačně. Proto se jeví jako pohodlnější cesta incidentům předcházet. Navíc nelze předvídat, kdy incident nastane, zda se vyskytne v daném čase pouze jeden a jaký bude mít dopad. Co se týče síťové bezpečnosti, z těchto důvodů je výhodné nasadit systém pro detekci či prevenci průniků (Intrusion Detection/Prevention System, IDS/IPS). Ten průběžně hlásí zjištěné incidenty či průniky (IDS), nebo je rovnou potlačuje (IPS). Dostupné systémy založené na hledání řetězců a vzorů v obsahu celých paketů mají tyto společné slabiny: příliš mnoho falešných poplachů, které zahlcují operátora či správce a nízká propustnost (typicky maximálně stovky Mb/s). Pokročilejší systémy vyžadují definovat stávající bezpečnostní politiku. V mnoha organizacích však tato politika buď formálně neexistuje, nebo je to znalost distribuovaná mezi více lidí, případně je jen velmi jednoduchá (v případě některých univerzitních sítí i záměrně) a v nejhorším případě neexistuje vůbec.

Soustavné monitorování toků může poskytnout data pro detekční skripty a nástroje, které mohou nahradit či vhodně doplnit stávající systémy operující převážně na vyšších síťových vrstvách. Pro bezpečnostní aplikace je základním požadavkem bezztrátové vzorkování v poměru 1 : 1. Ztráta jediného paketu může zanést nežádoucí nepřesnost. Při použití vhodných softwarových nástrojů a relativně malém objemu zpracovávaných dat (do 1 Gb/s) lze tomuto požadavku dostát i za velmi nízkých nákladů. V případě vysokorychlostních sítí (10 Gb/s a více) je nutno použít pro sběr dat hardwarové sondy. Neméně důležité je vhodné rozmístění zdrojů dat a jejich zapojení do stávající infrastruktury. Sonda na přípojném místě do internetu či export z hraničního směrovače bude monitorovat všechen příchozí a odchozí provoz z naší sítě, avšak komunikace mezi jednotlivými podsítěmi zůstane nadále skryta. Typické rozmístění zdrojů provozních dat v koncové síti je na Obrázku 1. Jde o tzv. *izolované monitorování* (pojem definován v [1]). Kromě sondy, která pokrývá přípojku do internetu, je nasazena další na přípojku vybraného segmentu sítě.

Komunikace koncové stanice v takovém segmentu se strojem mimo celou síť je zachycena dvakrát a jednou taktéž komunikace této stanice v rámci sítě. To může být výhodné při návrhu některých detekčních skriptů. Na druhé straně nasazení a správa více sond (příp. exportu ze směrovačů) je ekonomicky náročnější: větší



Obr. 1: Typické izolované rozmístění sond (zdrojů provozních dat) v koncové síti

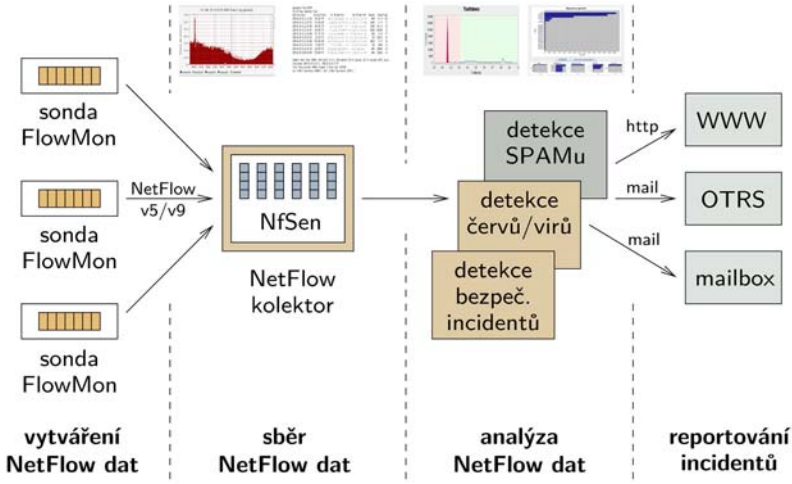
pořizovací náklady (více sond a tapů) či investice do výkonnějšího hardware směrovačů a v neposlední řadě také více lidských zdrojů.

3 Nasazení na Masarykově univerzitě

3.1 Sondy, primární zdroj dat

S výzkumem a vývojem v oblasti síťových prvků má sdružení CESNET ve spolupráci s Masarykovou univerzitou (MU) a Vysokým učením technickým v Brně dlouholeté zkušenosti. Výsledkem řešení evropských projektů a výzkumného záměru sdružení CESNET *Optická síť národního výzkumu a její nové aplikace* je hardwarově akcelerovaná sonda FlowMon [9]. Již první prototypy byly nasazeny na univerzitní síti. Sonda je schopna zpracovávat provoz na vysokorychlostních sítí (10 Gb/s a více). Současně byla testována i další dostupná řešení, zejména softwarové NetFlow sondy *fprobe*[1] [11] a *nprobe* [12]. Na rozdíl od hardwarově akcelerované sondy je softwarová sonda tvořena běžným serverem se síťovou kartou, operačním systémem Linux, samotnou aplikací, která přijímá pakety na síťovém rozhraní, extrahuje klíčové položky toku, vytváří a odesílá NetFlow záznamy na úložiště. Typicky se využívá knihovny *libpcap* a některé akcelerované

[1] Dostupná jako balíček v repozitářích Debian/Ubuntu a dalších.



Obr. 2: Architektura monitorovacího řetězce na Masarykově univerzitě

metody příjmu paketů (nejznámější je PF_RING). Na MU sbíráme nevzorkovaná data ze sond FlowMon [10] firmy Invea-Tech (spin-off MU), doplňkově pak ze směrovačů od firmy Cisco. Sondy jsou připojeny přes tapy (Test Access Port), pouze nouzově přes SPAN (Switched Port Analyzer) porty aktivních prvků. Použití SPAN portu je levnější a jednodušší (není nutno kupovat a instalovat tapy), avšak na zrcadlené lince dochází v lepším případě ke změně pořadí paketů, časovým prodlevám v porovnání s originální linkou a v horším případě až ke ztrátě paketů. [2] To je pro použití v bezpečnostních aplikacích nepřípustné.

3.2 Datové úložiště

Úložiště provozních záznamů je dalším článkem v celém řetězci. Rozhodli jsme se pro nasazení open source NetFlow kolektoru NfSen/nfdump. Za vývojem těchto nástrojů stojí švýcarská síť národního výzkumu a vzdělávání SWITCH. Soubor nástrojů NfSen/nfdump vznikl na základě praktické potřeby a je testován v produkčních podmínkách. Naše zkušenosti ukazují, že jde o jeden z nejlepších volně dostupných NetFlow kolektorů (šířen pod licencí BSD). Jako *nfdump* [8] je označována sada nástrojů pro příjem a zpracování NetFlow dat. Ovládá se z příkazové řádky. Dotazovací jazyk je podobný jako u nástroje tcpdump. *NfSen* [7] je grafická webová nadstavba, která mj. umožňuje nasazení vlastních zásuvných modulů, které periodicky zpracovávají přijatá NetFlow data. Nfsen/nfdump a NetFlow sondy vyvinuté sdružením CESNET jsou součástí tzv. „security toolsetu“ vyvinutého a testovaného v rámci projektu GN2.

Kolektor může být na stejném stroji jako sonda, ale častěji bývá provozován na vyhrazeném stroji s velkou diskovou kapacitou. Data získaná ze sond jsou pak zasílána po síti přes protokol UDP. Sondy provozované na MU mají nastaveno jak lokální, tak i vzdálené ukládání záznamů. Lokální ukládání je nastaveno z důvodu zálohy dat. Produkčně se pracuje s daty na tzv. centrálním kolektoru. Co se děje po příchodu paketu s NetFlow záznamy na kolektor? Data jsou agregována v předem zvoleném časovém okně a ukládána na disk jako binární soubory v adresářové struktuře podle data a profilu. Pro každý zdroj dat je vyhrazen jeden kanál, profil může obsahovat jeden či více kanálů. Pokud je vyčerpáno místo na disku, dojde k rotaci, k přepisu nejstarších dat. Na MU používáme 5minutové časové okno. Velikost uložených dat se tak pohybuje od několika MB až po asi 30 MB agregovaných dat za pět minut ve špičkách. Při velikosti profilu 3 TB máme v současnosti k dispozici až 5 měsíců staré záznamy o veškeré komunikace mezi naší sítí a zbytkem světa.

3.3 Další analýza a zpracování

Dosud popsaný řetězec neprovádí žádná další zpracování nasbíraných a uložených dat. Pro usnadnění práce bezpečnostního týmu CSIRT-MU vznikla sada jednoúčelových skriptů ve skriptovacím jazyce BASH, které vyhodnocují NetFlow data z pohledu síťových anomálií. V plánu je integrace těchto skriptů jako rozšiřujících modulů do kolektoru NfSen. Celý monitorovací řetězec je na Obrázku 2.

Jeden ze skriptů detekuje vertikální i horizontální skenování TCP portů. Pokud se zaměříme pouze na stroje ze spravované sítě, které skenují ostatní počítače v síti či vně sítě, je tato aktivita velmi podezřelá. Často totiž ukazuje na pokusy o rozšíření červů na další stroje. Provozní záznamy obsahují pro TCP toky i informace o příznacích (flags), takže je možno vhodným dotazem dostat všechny IP adresy strojů, které v daném časovém okně prováděly skenování. Tento seznam je i s příslušným portem, časovým razítkem prvního pokusu a celkovým počtem pokusů uložen na webový server v adresářové struktuře podle data. Analytik má tak přístup ke kompletní historii výstupů skriptu přes HTTPS. Současně tyto výstupy zpracovává mailový robot, který na základě seznamu podsítí a kontaktních adres automaticky rozešle upozornění na skenování příslušnému správci.

Podobně funguje i monitorování provozu na segmentu vyhrazeném pro síťové pasti (honeypoty). NetFlow sonda monitoruje přímo síťové rozhraní, přes které prochází veškerý provoz honeypotu. Skript spouštěný periodicky opět prochází nasbíraná data a hledá IP adresy z vnitřku sítě, které se pokusily kontaktovat past. Podle definice pasti by mělo jít o útočníky, legitimní uživatelé by se neměli do pasti vůbec dostat.

V síti MU lze odesílat poštu jen přes vybrané SMTP servery. Další skript sleduje objemové anomálie v provozu a využívá znalosti rozmístění sond. Porov-

nává provoz, který projde přes firewall a přes sondu do internetu s odchozím provozem ze síťových segmentů. Pokud je „rozdíl provozu“ neprázdný, může to ukazovat na neúspěšné spammery či chyby v konfiguraci.

Další skript kontroluje, zda mají komunikující stroje v sítí MU nastaveny reverzní DNS záznamy. Stroj (IP adresa) bez reverzního záznamu může ukazovat na zapomenutý stroj a tedy potenciální slabé místo. Každý den jsou na základě NetFlow dat získány IP adresy všech komunikujících strojů. Ty jsou pak přeloženy nástrojem *host*. V případě neexistujícího záznamu je IP adresa uložena do výstupního seznamu. Současně jsou přepočítány měsíční statistiky. Vycházíme z předpokladu, že nejdříve je potřebné zjednat nápravu u strojů, které jsou bez reverzního záznamu trvale.

Kromě těchto a dalších jednoúčelových skriptů, provozujeme i online verzi systému MyNetScope [6], který pracuje se složitějšími detekčními vzory a mechanismy. V současnosti se jedná se o detekci slovníkových útoků na SSH servery [4] a budování profilů koncových zařízení [5]. Taktéž testujeme prototyp síťového IDS zvaný CAMNEP [3], který je založen na více detekčních agentech, kteří společně vydávají rozhodnutí o důvěryhodnosti jednotlivých toků. Oba tyto systémy stojí v celém monitorovacím řetězci hned za zdroji dat. Obsahují totiž vlastní kolektory dat a vlastní detekční aplikace.

Sondy, kolektory a hlavně nadstavby pracující s NetFlow daty jsou dále rozvíjeny naším týmem v projektu *Bezpečnost informačních a komunikačních systémů Armády ČR*. Cílem projektu je analyzovat jednotlivé vzory chování (např. již zmíněné slovníkové útoky) a specifikovat postupy, jak tyto vzory detekovat a automaticky odvrátit z nich plynoucí hrozby. V současné době pracujeme na detekci infiltrace cizího zařízení do sítě, konkrétně na detekci NAT (Network Address Translation). K tomuto účelu používáme sondy, které exportují záznamy rozšířené o některé další položky z IP datagramu (tzv. „rozšířené NetFlow“).

4 Zkušenosti z provozu

V září roku 2009 bylo v produkčním provozu celkem 8 sond, z toho jedna desetigigabitová monitorující připojení MU do sítě CESNET2 (a tedy i do internetu). Doplňkově jsou také ukládána data exportovaná ze 6 směrovačů. Celkový monitorovaný provoz dosahuje ve špičkách průměrně až 4 Gb/s, 8 000 toků za sekundu a 500 000 paketů za sekundu. Vývoj a testování pokročilých aplikací zpracovávající NetFlow data (např. detekční vzor SSH útoku) potvrdilo, že pro připojení sond není možné použít SPAN port aktivního prvku. Tyto aplikace jsou citlivé na správné pořadí paketů, časové značky atp. Pokud toto není zaručeno, výsledek detekce je nepřesný. Dalším důležitým předpokladem, zejména pro korelaci s dalšími datovými zdroji, je synchronizace času na sondách i kolektorech s časovým serverem.

Jedno výkonnější a větší centrální úložiště[2] je v produkčním provozu, na vývoj skriptů a testování je vyhrazeno další, méně výkonné[3]. Na serverech běží 64bitový operační systém CentOS a nástroje NfSen 1.3 a nfdump 1.5.7. Prázdninové měsíce roku 2009 přinesly menší zátěž pro disky. Za měsíce červen až srpen jde o desítky GB (agregovaných!) záznamů měsíčně, měsíc květen zaujímá asi 120 GB diskového prostoru.

Při vyšetřování bezpečnostního incidentu často potřebujeme znát odpověď na otázku *Komunikoval někdy stroj s IP adresou A.B.C.D s nějakým naším strojem?*. To znamená, že je nutno prohledat všechna uložená data. Vyhodnocení jednoduchého dotazu trvá minuty či desítky minut. Není tedy možno klást interaktivní dotazy. Pro usnadnění práce jsme se rozhodli vyvinout jednoduchou databázi komunikujících párů strojů založenou na PostreSQL. Příchozí NetFlow data jsou průběžně zpracovávány a pokud se v daném dni ještě neobjevil daný komunikační pár, je uložen do databáze včetně jeho časové značky. Pokud už v databázi záznam o páru existuje, nic se neprovede. Vyhledávání v této databázi je pak rychlejší a to ušetří analytikovi čas.

Detekce skenujících strojů výrazně přispěla ke snížení šíření červu Conficker v síti MU v dubnu až červnu 2009. Za tuto dobu jsme rozeslali asi stovku hlášení na fakulty a uživatelům připojeným přes bezdrátové sítě MU. Díky obecnému pojetí detekce byly kromě toho odhaleny i dosud neznámé viry, které nefigurovali v databázích předních výrobců antivirů.

Dobré zkušenosti i máme s profilováním strojů. Např. jsme odhalili jinak nedetekovaný distribuovaný DOS útok na DNS server, nezabezpečené rozhraní síťové tiskárny či web server s volně přístupným nelegálním obsahem.

5 Závěr

Výše uvedené zkušenosti z provozu dokladují, že přínos využití provozních záznamů o síťové komunikaci v oblasti bezpečnosti je poměrně velký. Samotný sběr dat a jejich ukládání nestačí, přidanou hodnotu je právě další automatizované zpracování. Tradiční přístupy detekce anomálií, které pracují s celými pakety, lze zčásti nahradit detekčními algoritmy, které pracují s toky. Výhodou je větší propustnost. Popsaný přístup však má i své nedostatky, sám o sobě tedy nemůže nahradit ostatní techniky. Cílem projektu Fondu rozvoje CESNET *Inteligentní logovací server* je vývoj systému, který bude vhodně korelovat data ze stávajících heterogenních zdrojů (síťových sond, honeypotů, serverů, černých listin...). Tímto by mělo být docíleno synergického efektu zapojených nástrojů.

[2] 2× Intel Xeon E5430 2.66 GHz, 16 GB RAM, 16× HDD, RAID 5, 10 TB
[3] 2× Intel Xeon E5130 2 GHz, 8 GB RAM, 2× 1 TB HDD

Reference

[1] Košňar, T. Benefity a úskalí plošného souvislého sledování IP provozu na bázi toků při řešení bezpečnostních hlášení. In *Sborník příspěvků z 34. konference EurOpen.CZ, 17.–20. května 2009*, 2009. s. 23–37.

[2] Zhang, J. a Moore, A. W. Traffic trace artifacts due to monitoring via port mirroring. In *Proceedings of the Fifth IEEE/IFIP E2EMON*, 2007, s. 1–8.

[3] CAMNEP Webová stránka projektu.
Dostupné online: http://agents.felk.cvut.cz/projects/camnep/. 2009.

[4] Vykopal, J., Plesník, T. a Minařík, P. Network-based Dictionary Attack Detection. In *Proceedings of International Conference on Future Networks (ICFN 2009)*. Bangkok. IEEE Computer Society, 2009, s. 23–27. ISBN 978-0-7695-3567-8

[5] Minařík, P., Vykopal, J. a Krmíček, V. Improving Host Profiling with Bidirectional Flows. In *Proceedings of International Symposium on Secure Computing (SecureCom09)*. Vancouver, Canada. IEEE Computer Society, 2009, s. 231–237. ISBN 978-0-7695-3823-5

[6] MyNetScope Webová stránka aplikace.
Dostupné online: http://www.advaict.cz/mynetscope. 2009.

[7] NfSen Webová stránka projektu.
Dostupné online: http://nfsen.sourceforge.net/. 2009.

[8] nfdump Webová stránka projektu.
Dostupné online: http://nfdump.sourceforge.net/. 2009.

[9] FlowMon (Liberouter) Webová stránka projektu.
Dostupné online: http://www.liberouter.org/flowmon/. 2009.

[10] FlowMon (Invea) Webová stránka projektu. Dostupné online:
http://www.invea-tech.com/cs/products/flowmon-probes. 2009.

[11] fprobe Webová stránka projektu.
Dostupné online: http://fprobe.sourceforge.net/. 2009.

[12] nProbe Webová stránka projektu.
Dostupné online: http://www.ntop.org/nProbe.html. 2009.

FlowMon – inovativní přístup v oblasti monitorování a bezpečnosti počítačových sítí

Jiří Tobola

E-mail: tobola@invea.cz

Abstrakt

Spolehlivá a dobře zabezpečená počítačová síť je klíčem pro úspěšné fungování každé organizace. Již krátkodobý výpadek sítě znamená narušení infrastruktury společnosti a může způsobit škody v řádech miliónů korun, poškození dobrého jména společnosti a nespokojenost nebo dokonce ztrátu zákazníků. Podobné problémy přináší i „pouhé" ne zcela správné fungování počítačové sítě projevující se například sníženou dostupností a pomalou odezvou kritických aplikací (podnikové systémy, VoIP). Následující článek popisuje, jak takovýmto situacím úspěšně předcházet a čelit za pomoci moderních monitorovacích technik na bázi NetFlow.

Dlouhá léta byl synonymem pro monitorování a dohled nad počítačovou sítí protokol SNMP. Současná doba, kdy na dostupnosti a správné funkcionalitě počítačové sítě závisí fungování většiny organizací, si však žádá modernější a efektivnější prostředky. Ty musí v reálném čase poskytovat detailní statistiky o síťovém provozu, které jsou klíčové pro efektivní správu a účinné zabezpečení počítačových sítí. Právě takové statistiky nabízí technologie NetFlow.

Monitorování na bázi NetFlow

NetFlow je v současnosti nejrozšířenější průmyslový standard pro měření a monitorování počítačových sítí na základě IP toků. Tok je v terminologii NetFlow definován jako sekvence paketů se shodnou pěticí údajů: cílová/zdrojová IP adresa, cílový/zdrojový port a číslo protokolu. Pro každý tok je zaznamenávána doba jeho vzniku, délka jeho trvání, počet přenesených paketů a bajtů a další údaje. V tradiční NetFlow architektuře jsou statistiky vytvářeny pomocí směrovačů a odesílány na kolektory (datová úložiště) k dalšímu zpracování, vizualizaci a analýzám.

Zatímco SNMP statistiky poskytují jenom souhrnné informace o provozu a neumožňují vidět, co se v síti doopravdy děje (jaké je rozložení provozu, kdo síť nejvíce zatěžuje), NetFlow statistiky poskytují detailní informace o tom kdo komunikoval s kým, kdy, jak dlouho, jak často, nad kterým protokolem a kolik bylo přeneseno dat. Tyto statistiky umožňují sledování vytížení sítě v reálném čase, monitorování aktivit uživatelů i služeb, optimalizaci síťové infrastruktury, sledování reálného využití Internetu, dodržování vyhlášky o elektronické komunikaci, či odhalování a prokazování bezpečnostních incidentů. Tím šetří finance vynaložené na správu počítačových sítí, usnadňují práci síťových administrátorů a zvyšují spokojenost koncových uživatelů a zákazníků.

V minulosti bránila rozšíření monitorovacích systémů na bázi NetFlow jejich vysoká pořizovací cena a dostupnost pouze pro omezený počet směrovačů. Tyto nevýhody odstranila česká společnost INVEA-TECH, a.s., (www.invea.cz), která uvedla na trh kompletní, cenově dostupné řešení FlowMon pro monitorování sítí s využitím technologie NetFlow. Portfolio produktů FlowMon zahrnuje výkonné autonomní sondy pro všechny typy sítí až do rychlosti 10 Gb/s, kolektory pro uložení, zobrazení a analýzy síťových statistik a další rozšiřující moduly (detekce anomálií, dohled nad sítí, inteligentní reporting).

Sondy FlowMon

Sondy FlowMon reprezentují výkonné monitorovací zařízení určená pro ethernetové sítě na rychlostech od 10 Mb/s do 10 Gb/s. Sondy sledují komunikaci na síti, vytvářejí statistiky plně kompatibilní s NetFlow standardem a odesílají je na vestavěný či externí kolektor. Sondy se typicky umisťují na vstupní a výstupní body sítě, kritická místa či linky s největšími přenosy dat. Vlastní připojení sondy do sítě se realizuje pomocí mirror portu směrovače či přepínače, nebo přímým vložením do linky s využitím optického nebo metalického rozbočovače (TAPu).

Řada produktů FlowMon sond zahrnuje standardní modely pro běžné sítě a hardwarově akcelerované modely pro kritické a vysoce vytížené linky. Sondy jsou dostupné pro metalická i optická rozhraní a nabízejí až 6 monitorovacích portů na zařízení. Výhodou proti konkurenčním produktům je garantované zpracování všech dat bez ztráty paketu a integrace NetFlow kolektoru přímo na sondách.

FlowMon monitorovací centrum

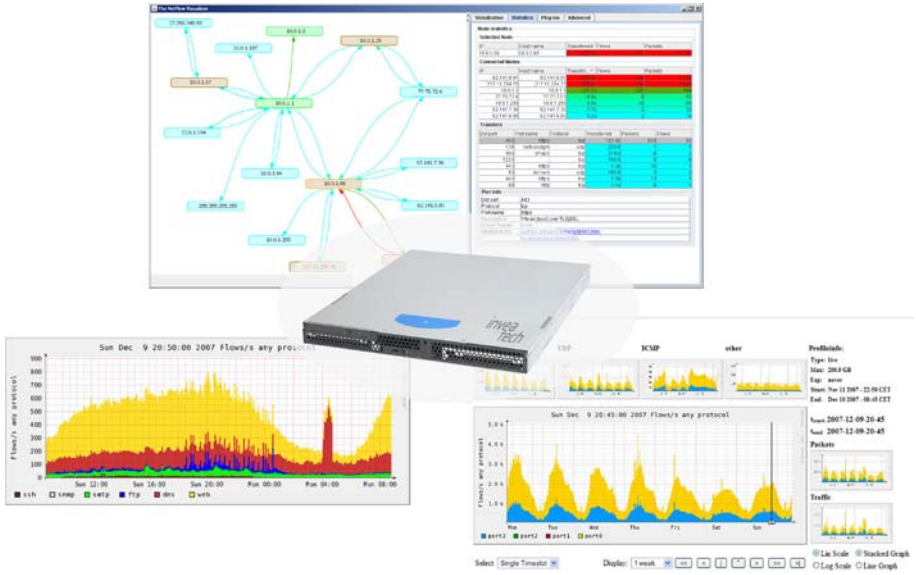
NetFlow data generovaná sondou jsou zasílána na integrovaný, nebo na externí kolektor. Jako kolektor lze využít libovolnou aplikaci třetích stran nebo předkon-

figurovanou aplikaci – FlowMon monitorovací centrum. Vestavěná verze tohoto nástroje integrovaná přímo na sondě je určená pro rychlé seznámení s technologií NetFlow a nabízí kompletní řešení pro menší a střední sítě. Samostatná verze realizovaná na vyhrazeném serveru (FlowMon kolektor) naopak nabízí maximální výkon a profesionální řešení pro sběr dat z více sond ve větších sítích.

Přístup k uloženým NetFlow datům je realizován přes zabezpečené webové rozhraní. Intuitivní ovládání FlowMon monitorovacího centra umožňuje zobrazovat síťové statistiky v podobě grafů a tabulek s různým časovým rozlišením, generovat takzvané top N statistiky, filtrovat data dle požadovaných kritérií, tvořit uživatelské profily, provádět bezpečnostní analýzy, nebo nastavovat generování automatických upozornění na požadované události, např. porušení bezpečnostní politiky. Pomocí rozšiřujících modulů je možné funkcionalitu dále rozšířit například o SNMP dohled nad sítí nebo automatickou detekci anomálií.

Závěr

Nasazení pokročilého monitorovacího řešení FlowMon založeného na sledování toků umožňuje organizacím předcházet ztrátám v důsledku nedostupnosti sítě, snižovat náklady na provoz a zabezpečení sítě, ochránit investice do síťové infrastruktury, zvýšit spolehlivost a dostupnost sítě a maximalizovat spokojenost jejich uživatelů a zákazníků.



Monitoring sítí pomocí NetFlow dat – od paketů ke strategiím

Martin Rehák, Karel Bartoš, Martin Grill, Jan Stiborek, Michal Svoboda

E-mail: jmeno.prijmeni@agents.felk.cvut.cz

Klíčová slova: monitoring sítí, NetFlow, detekce anomálií, bezpečnost

Abstrakt

Cílem prezentace bude seznámit uživatele s celou škálou pohledů na síť modelovanou pomocí NetFlow dat, od čistých NetFlow dat přes jednoduché statistické modely, jejich kombinace, modely důvěryhodnosti, až po obecnější přístupy zahrnující model možného protivníka, vlastních zdrojů a model útoku. Prezentace bude doprovázena praktickými ukázkami založenými na systému CAMNEP, a bude zahrnovat i krátkou diskuzi vhodnosti uzavřeného-otevřeného přístupu pro různé části řešení.

1 Úvod

Cílem našeho příspěvku je představení systému CAMNEP [8] a souvisejících technologií, ať už otevřených či uzavřených, které umožňují snadné a rekonfigurovatelné bezpečnostní monitorování širokého spektra sítí. Systém CAMNEP je určen k detekci neautorizovaných toků (intruzí) v sítích pomocí aplikace metod bahaviorální analýzy [9] (Network Behavior Analysis (NBA)) a detekce anomálií na statistiky síťového provozu ve formátu NetFlow. Na rozdíl od klasických technik detekce intruzí umožňuje tento přístup i odhalení nových útoků doposud nepopsaných pomocí pravidel či vzorů používaných klasickými IDS (IPS) systémy [7]. Tato schopnost je u klasických systémů behaviorální analýzy vykoupena významným zvýšením počtu falešných poplachů, který tyto techniky činí buď zcela nenasaditelnými, nebo nasaditelnými jen s nastavenou nízkou citlivostí, a tedy i s nižší šancí na odhalení skutečných útoků. Systém CAMNEP řeší tento problém pomocí vícestupňové integrace metod detekce anomálií v algoritmu založeném na modelování důvěryhodnosti a pomocí technik autonomic computing, které podstatně snižují kvalifikační a časové nároky na nastavení a správu systému.

2 Historie projektu, výzkum a související publikace

Systém CAMNEP vznikl jako vedlejší produkt výzkumných projektů financovaných CERDEC, US Army prostřednictvím ERO/ITC-A London. Projekty probíhaly v letech 2007 (vývoj systému) a červen 2008-prosinec 2009 (samokonfigurace, distribuce a reakce na inteligentního protivníka). Projekt je realizován Centrem agentních technologií katedry kybernetiky FEL ČVUT. Na projektu se rovněž podílí Ústav výpočetní techniky Masarykovy univerzity, který řeší úlohy sběru a předzpracování dat a podílí se na operační testování systému (a v minulosti i na úlohách vizualizace). Výzkumné výsledky projektu byly a jsou v současnosti publikovány na mezinárodních konferencích a v odborných recenzovaných časopisech.

3 Zachycení a zpracování dat ve formátu NetFlow

Systém využívá jako vstupu dat ve formátu NetFlow, který je realizován týmem Jiřího Novotného, Pavla Čeledy a Vojtěcha Krmíčka z UVT MU Brno. Sondy FlowMon [11, 2, 1], která data sbírají na síti jsou založeny na programovatelných hradlových polích (FPGA), které jsou schopny přesně monitorovat provoz až do rychlosti 10 Gbps. Z tohoto provozu poskytují data ve formátu Netflow, který obsahuje informaci o síťových tocích v okamžiku pozorování. Jeden síťový tok [3] je definován jako množina paketů, které sdílí společnou zdrojovou a cílovou IP adresu, protokol a zdrojový a cílový port. K těmto základním informacím je dále přiřazen počet paketů a jejich úhrnná velikost pozorovaná během období, kdy byl tok agregován. Sondy FlowMon, které v našem systému sběr dat provádí, jsou vyvíjeny v rámci projektu Liberouter, a jsou integrovány s kolektorem nfdump/nfsen, který příjmá a ukládá a částečně i zpracovává data. Ta jsou poté pomocí specializované komponenty předzpracována, jsou z nich dopočítány vlastnosti nezbytné pro afektivní detekci anomálií a tato data jsou poté předána do detekční vrstvy systému CAMNEP.

4 Kolaborativní detekce anomálií

Detekční vrstva systému CAMNEP má za úkol vybrat z množiny všech zaznamenaných spojení/toků pouze ta, která predstavují neobvyklou či nežádoucí aktivitu v síti, a agregovat je do jednotlivých incidentů. Tyto incidenty jsou poté reportovány bezpečnostním administrátorům, kteří mohou následně provést odpovídající reakce v rámci bezpečnostních opatření na síti. Detekční vrstva

v systému CAMNEP je založena na multiagentním systému AGLOBE, ve kterém spolupracuje několik agentů (samostatných entit), vykonávajících vlastní detekci. Výsledky jednotlivých agentů jsou v průběhu detekčního procesu agregovány k zajištení komplexního pohledu na daný síťový provoz. Každý detekční agent je založen na jedné z námi predem vybraných metod detekce anomálií. Výběr těchto metod se řídil podmínkou vzájemné odlišnosti v pohledu na charakteristiky síťového provozu, pomocí kterých lze dosáhnout rozpoznávání incidentu na síti. V současné době využíváme pět základních přístupů, které jsme více či méně modifikovali pro vzájemnou kompatibilitu a vhodnost nasazení v našich podmínkách.

- Metoda MINDS [4] modeluje počty příchozích a odchozích spojení jednotlivých strojů (v kombinaci s číslem portu) v průběhu času a detekuje vzájemné odchylky v těchto časových řadách ů velikost odchylky určuje míru anomálie.

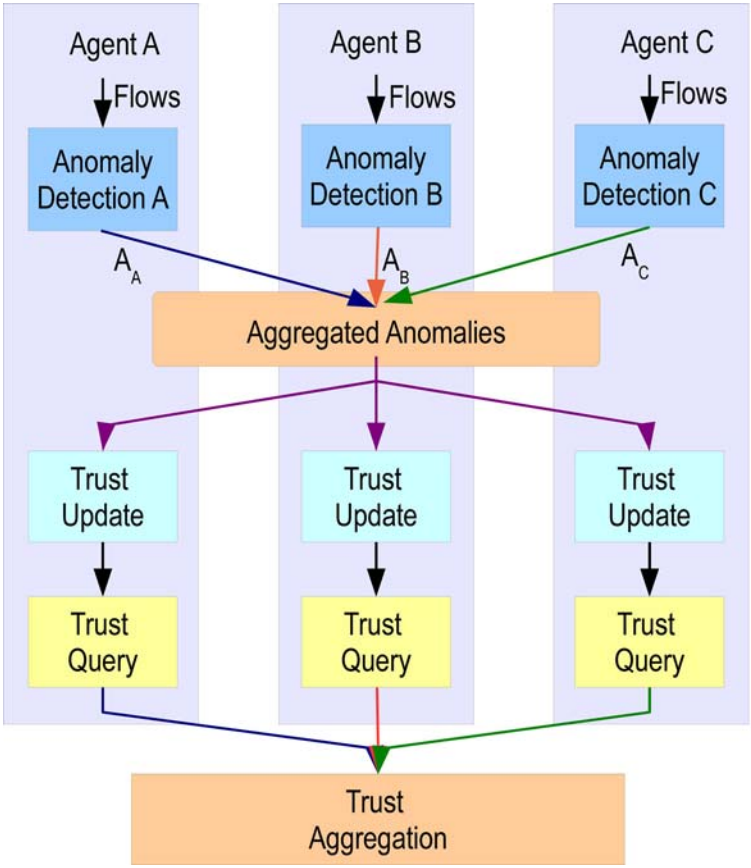
- Metoda autorů Xu et al. [12, 13], založená na použití statických klasifikačních pravidel, umísťuje jednotlivá spojení do vícedimenzionálního prostoru entropií IP adres a portů. Toto umístění poté rozhoduje o nomálnosti daného spojení.

- Další metoda Lakhina et al, [5] využívá k modelování objemu provozu jednotlivých zdrojů provozu (počty spojení, paketů a bytů) statistickou metodu PCA (Principal Component Analysis), pomocí které a na základě sady předchozích a současného pozorování rozděluje provoz na normální a reziduální část. Velikost reziduální části se poté využívá ke stanovení míry anomálie každého zdroje provozu.

- Další metoda stejných autorů [6] je analogická s předchozí metodou, pouze s tím rozdílem, že PCA používá k separaci reziduální části statistické rozdělení parametrů provozu každého zdroje, zejména entropií IP adres a portů.

- Poslední metoda – TAPS [10] – je speciálně vyvinuta k detekci ruzných typu skenování portů a sítí a využívá k tomu velikost poměru počtu odchozích IP adres a počtu různých odchozích portu v kombinaci s metodou sekvenčního testování hypotéz.

Vlastní detekce je rozdělena do fází. V první fázi každý agent na základě své metody detekce anomálií přiřazuje jednotlivým spojením jejich míru anomálie. Pro každé spojení takto získáme od každého agenta jednu hodnotu. Tyto hodnoty se v závěru první fáze agregují v celkovou míru anomálie pro každé spojení.

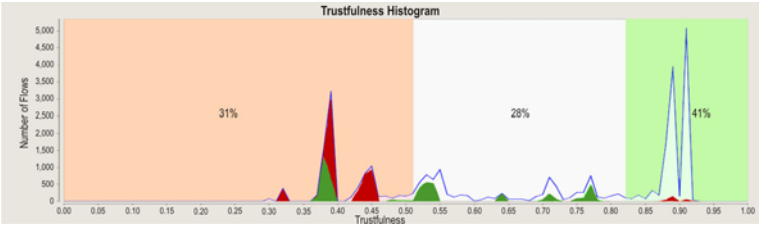


Obr. 1: Proces detekce distribuovaný mezi agenty

Metody detekce anomálií mají obecně velkou výhodu v tom, že jsou schopné detekovat i dosud neznámé síťové události. Tato schopnost je však vykoupena obecně horší klasifikací, zejména kvůli vysokému počtu false positives (legitimního provozu klasifikovaného jako nežádoucího). Proto si každý agent společně s metodou detekce anomálií udržuje rovněž svůj tzv. důvěryhodnostní model, který používá ke stanovení důvěryhodnosti daného spojení na základě dlouhodobého sledování provozu. Důvěryhodnostní model využívá ke stanovení důvěryhodnosti práci se shluky ve vícedimenzionálním metrickém prostoru, do kterého umistuje dané spojení na základě hodnot jeho pozorovaných vlastností a charakteristik, jež byly v první fázi detekce využívány daným agentem k určení míry anomálie. Důvěryhodnost daného spojení je poté určena na základě jeho přiřazení do shluků a důvěryhodnosti těchto shluků.

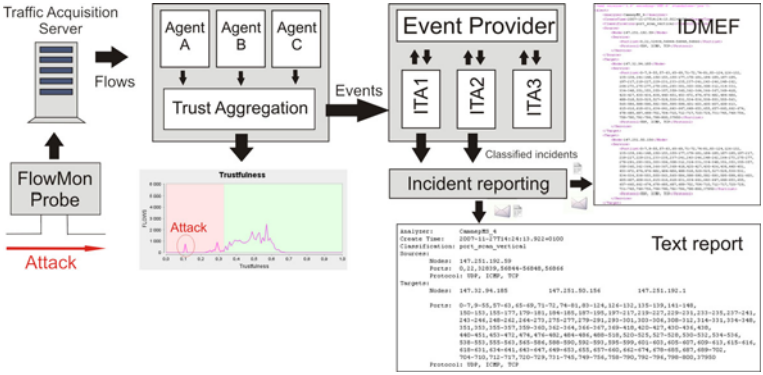
5 Uživatelské rozhraní a použití systému

Systém je koncipován jako dohledový. Vstupy, které v pravidelných intervalech přijímá od sond monitorujících síťový provoz zprqcuje pomocí výše uvedeného postupu a zobrazí je ve formě histogramu, ve kterém rozděluje provoz podle míru důvěryhodnosti na ose od 0 do 1. Toky, které jsou považovány za důvěryhodné jsou umístěny v pravé části spektra, zatímco nedůvěryhodné toky jsou umístěny v levé části spektra.



Obr. 2: Příklad výstupu systému. Plné barevné plochy vyznačují aposteriorní klasifikaci provozu, pozadí dělí provoz na důvěryhodný a nedůvěryhodný

Na příkladu si lze povšimnout rozdělení osy X do 3 regionů. Toky, které jsou v levé části, a tedy považovány za nedůvěryhodné, jsou seskupeny do událostí a tyto události jsou systémem roztříděny podle svého charakteru, tak aby bylo možné zpracovat každý útok podle příslušných pravidel. Toto zpracování je zachceno na posledním schématu.



Obr. 3: Integrace systému do širšího rámce dohledového systému (Pavel Čeleda)

Toto schéma rovněž ilustruje široké spektrum integračních možností pro náš systém. Jako datové zdroje je možno použít širokou řadu sond, od open source řešení (fprobe či nprobe), proprietárních řešení (invea-tech), specializovaný HW (liberouter) a i řadu routerů vyšších řad od renomovaných výrobců. Jejich snadná interoperabilita a záměnnost je zajišťěna respektováním standardu IPFIX (NetFlow), který je implementován i na straně široké škály software určeného pro sbírání, ukládání a další zpracování NetFlow dat. Na tento software se pak přirozeně napojují řešení pro analýzu dat, extrakci událostí a jejich řízené zpracování v produkčním prostředí. Systém CAMNEP zapadá do kategorie nástrojů analytických, kde se, jak již bylo zmíněno, profiluje zejména kvalitou detekce a vysokou citlivostí. Významnou vhodou kvalitní a respektované de-facto standardizace je možnost sestavit řešení na míru konkrétnímu problému, které může kombinovat jak otevřené, tak i proprietární komponenty. Pestrá nabídka komponent umožňuje selektivně investovat (ať už zakoupením propretární komponenty či podporou open-source projektu) do těch částí, které jsou pro daný problém klíčové a snížit náklady a komplexitu řešení tam, kde to je možné.

Reference

[1] CESNET, z. s. p. o., collective of authors. *Family of COMBO Cards.* http://www.liberouter.org/hardware.php, 2007.

[2] CESNET, z. s. p. o., collective of authors. *Liberouter Project.* http://www.liberouter.org/projects.php, 2007.

[3] Cisco Systems. *Cisco IOS NetFlow.* http://www.cisco.com/go/netflow, 2007.

[4] Levent Ertoz, Eric Eilertson, Aleksandar Lazarevic, Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava, and Paul Dokas. Minds – minnesota intrusion detection system. In *Next Generation Data Mining.* MIT Press, 2004.

[5] Anukool Lakhina, Mark Crovella, and Christophe Diot. Diagnosis Network-Wide Traffic Anomalies. In *ACM SIGCOMM '04*, p. 219–230, New York, NY, USA, 2004. ACM Press.

[6] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining Anomalies using Traffic Feature Distributions. In *ACM SIGCOMM, Philadelphia, PA, August 2005*, p. 217–228, New York, NY, USA, 2005. ACM Press.

[7] Stephen Northcutt and Judy Novak. *Network Intrusion Detection: An Analyst's Handbook.* New Riders Publishing, Thousand Oaks, CA, USA, 2002.

[8] Martin Rehak, Michal Pechoucek, Karel Bartos, Martin Grill, Pavel Celeda, and Vojtech Krmicek. Camnep: An intrusion detection system for high-speed networks. *Progress in Informatics*, (5):65–74, March 2008.

[9] Karen Scarfone and Peter Mell. *Guide to intrusion detection and prevention systems (idps).* Technical Report 800–94, NIST, US Dept. of Commerce, 2007.

[10] Avinash Sridharan, Tao Ye, and Supratik Bhattacharyya. *Connectionless port scan detection on the backbone.* Phoenix, AZ, USA, 2006.

[11] Pavel Čeleda, Milan Kováčik, Tomáš Koníř, Vojtěch Krmíček, Petr Špringl, and Martin Žádník. *FlowMon Probe.* Technical Report 31/2006, CESNET, z. s. p. o., 2006. http://www.cesnet.cz/doc/techzpravy/2006/flowmon-probe/.

[12] Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharrya. Reducing Unwanted Traffic in a Backbone Network. In *USENIX Workshop on Steps to Reduce Unwanted Traffic in the Internet (SRUTI)*, Boston, MA, July 2005.

[13] Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya. Profiling internet backbone traffic: behavior models and applications. In *SIGCOMM '05*, p. 169–180, New York, NY, USA, 2005. ACM Press.