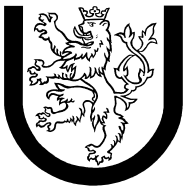


ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

Rychlost konvergence v IP/MPLS sítích

Jméno a příjmení: *Martin Lipinský*
Osobní číslo: ***A05450***
Studijní skupina: *Dálkové studium*
Obor: *INIB-INFB*
E-mail: *martin@lipinsky.cz*
Předmět: *KIV/PRJ5*

Datum odevzdání: 24.12.2007

Obsah

1 Úvod

1.1 Předpokládané znalosti

Dokument předpokládá, že čtenář je obeznámen s principem TCP/IP sítí, procesem směrování paketů a principem fungování dynamických směrovacíchprotokolů (OSPF, IS-IS). Rovněž budu dále v textu předpokládat, že čtenář má povědomí o principu a fungování MPLS sítí a přepínání labelů. (Chápe princip funkce protokolu LDP či RSVP). Je třeba se orientovat v rolích jednotlivých routerů v MPLS topologii (P = Provider, PE = Provider edge, CE = Customer edge). Pro pochopení následujícího textu je užitečná i detailní znalost protokolu BGP (MP-BGP) používaného v MPLS sítích pro přenos směrovacích informací.

1.2 Vymezení pojmů

Konvergencí obecně je myšlen čas, potřebný v síti s redundantní topologií k přesměrování (přepnutí) provozu na záložní či obecně redundantní části sítě. Ve velmi rozsáhlých sítích WAN (Internet) se nezřídka dostáváme na konvergenci v řádech desítek vteřin až minut. Stále častěji je však, zvláště v podnikových sítích, MPLS nasazováno i pro přenos hlasu, kde nahrazuje sítě TDM. Narážíme pak na požadavek dosahovat podobné spolehlivosti (99,999%) a konvergence (typicky desítky milisekund). Spolehlivost s rychlostí konvergence velmi úzce souvisí, neboť v případě požadované spolehlivosti 99,999% (nazývané někdy pět devítek) se bavíme o možnosti výpadku v délce 5,25 minuty za rok. Viz tabulka 1.

Dostupnost	Délka výpadku
90%	36,5 dní
99%	3,36 dní
99,9%	8,76 hodin
99,99%	52,55 minut
99,999%	**5,25 minut**
99,9999%	31,5 sekund

Zařízení, splňující tyto požadavky, musí byt velmi stabilní, schopné softwarových upgradů za běhu systému a bez přerušení forwardování paketů. Druhou klíčovou vlastností se pak stává rychlá konvergence při výpadcích linek, kterým pochopitelně nelze zcela zabránit. I pro datové služby se setkáváme s požadavkem na rychlejší než sekundovou konvergenci (<1s), pro přenos hlasu a hlasovou signalizaci pak méně než 500 milisekund.

2 Jednotlivé aspekty konvergence

2.1 Fyzické médium

Dříve než je možno zabývat se problémem konvergence protokolů, je potřeba vyřešit rychlost detekce výpadku na lince mezi dvěma routery. Jednoduchá situace nastává, pokud je spoj veden lokálním kabelem. U SDH linek je informace o případném přerušení linky také šířena přímo pomocí signálů (LOS, AIS atd). Horší situace nastává u WAN spojů s rozhraním Ethernet, vedených často přes několik L2 zařízení, kdy o výpadku uprostřed se nemusí ze stavu na rozhraní ani jeden ze směrovačů dozvědět (link proti switchi je stále nahoře). Podobný případ

nastane v případě optického rozhraní s duplexem vláken (vysílání – příjem) a přerušením jen jednoho vlákna. Pak jeden z routerů o vzniklém problému na lince vůbec neví. IGP protokol na základě nepřijetí HELO paketů na problém samozřejmě přijde, ale až v čase přesahujícím několik sekund. Cesta implementace subsekundových HELO do IGP situaci zcela neřeší, neboť například u OSPF v implementaci Cisco je DEAD interval stále 1s.

Proto byl pro rozhraní, která nemají vlastní signalizaci problémů na lince, vyvinut a standardizován protokol BFD (Bi-directional forwarding detection)[1]. Ten periodicky posílá HELO pakety v řádu desítek až stovek milisekund oběma směry pro kontrolu průchodnosti linky. V případě nepřijetí tří HELO paketů pak informuje OS směrovače o vyřazení linky z provozu.

V konzervativně nastavených směrovačích bývá často z důvodu stability vložen timer, který způsobuje opožděné informování operačního systému o nastalém výpadku. Tato technika zaručuje vysokou stabilitu (odolnost proti flapujícím linkám a jiným periodickým problémům). Znemožňuje ale rychlou detekci problému. Pro rychlou konvergenci je potom potřeba tento timer nastavit na 0 a tím zajistit okamžité informování operačního systému směrovače o nastalé události.

2.2 IGP

Rychlost konvergence sítě může být a často také je přímo závislá na rychlosti konvergence interního směrovacího protokolu (OSPF[2] nebo IS-IS[3]). Oba dané protokoly fungují velmi podobně. Oba jsou zástupci tzv. link-state protokolů, beroucích v úvahu nejen počet skoků ale i další parametry (např. propustnost linky, zastupovanou parametrem metric).

Každá změna v síti způsobí rozeslání LSA paketů s informací o nové topologii všem okolním routerům. Na základě informací z těchto paketů je v každém routeru vybudována databáze s novou topologií. Následuje spuštění Dijkstra[4] (jinak též SPF – shortest path first) algoritmu, který, zjednodušeně řečeno, z grafu linek udělá strom, tedy spočítá nejkratší cesty v síti a následně nainstaluje směrovací tabulku.

Pro rychlost konvergence IGP je tedy klíčové nastavit agresivní timery na rozeslání LSA paketů a následně nakonfigurovat malou prodlevu před spuštěním SPF algoritmu. Při rychle se opakujících periodických problémech je zde ale opět nebezpečí, že router bude jen měnit topologii a tak spotřebuje všechny své prostředky. Proto se používá tzv. dampening, kdy při první události reaguje například za 10ms, pokud velmi brzy poté přijde další změna, počká 100ms, pokud opět nastane změna, počká 1s atd. Tato technologie zajišťuje, že je síť velmi rychlá a neztrácí na stabilitě. Je třeba si uvědomit že nastavení časování IGP je vždy kompromisem mezi rychlostí a stabilitou. Teoreticky by však subsekundová konvergence ani ve velmi rozsáhlých sítích (stovky směrovačů) na dnešních routerech neměla být problém .

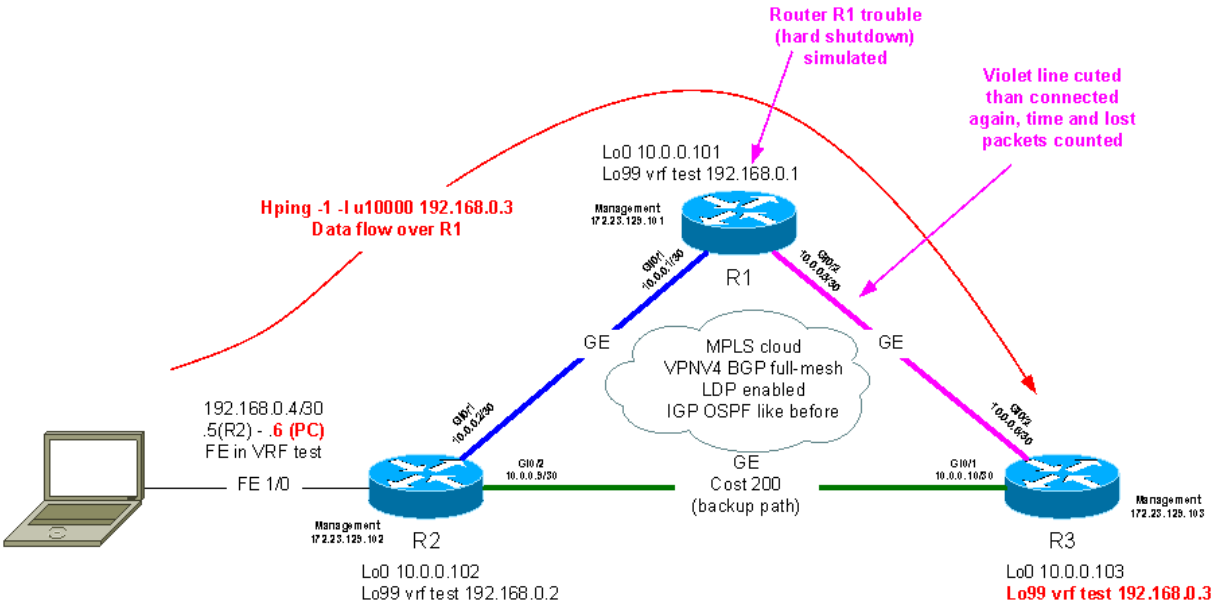
2.3 MPLS / LDP

V rozsáhlejších MPLS sítích se nejčastěji setkáváme se signalizačním protokolem LDP, řídícím automatické sestavení cest MPLS sítí. Pokud nesestavujeme statické tunely či MPLS Fast Reroute záložní cesty, ale spolehneme se pouze na spuštěné LDP, měla by konvergence teoreticky odpovídat zhruba rychlosti, jakou dosahujeme se samotným protokolem IGP (ať už OSPF nebo IS-IS). Zde ale narážíme na problém náběhu nové linky při výpadku. Správně nastavené IGP velmi rychle zjistí že linka je opět funkční a zahrne ji do směrovacích tabulek. Bohužel LDP protokolu trvá výměna labelů nějaký čas, po který směrovač pakety zahazuje. Výsledkem je velmi rychlá konvergence při výpadku, konvergence při opětovném náběhu linky ale trvá několik sekund. Řešení je několik. Každý z výrobců používá jiný, i když v principu velmi podobný systém, viz tabulka 2.

Výrobce	Řešení LDP blackholingu
Cisco	Synchronizace LDP a IGP. IGP počká, dokud se na lince neustaví LDP neighbourship.
Juniper	IGP zareaguje okamžitě, ale na linku nastaví metriku nekonečno až do doby, než se ustaví LDP neighbourship.
Alcatel / Timetra	IGP počká s náběhem linky předem definovaný čas, který se považuje za dostatečný, aby mezitím LDP stačil navázat neighbourship.

3 Testovací síť a konfigurace

Pro úvodní měření byly použity routery 7206VXR firmy Cisco Systems (http://www.cisco.com/en/US/products/hw/routers/ps341/index.html), vybavené procesorovou kartou NPE-1G . Kompletní konfigurace routerů viz přílohy. Směrovače byly nakonfigurovány jako MPLS PE routery. Uvnitř MPLS sítě byl použit směrovací protokol OSPF a signalizační protokol LDP. Pro přenos VPN informací mezi routery bylo ustaveno spojení protokolem MP-BGP.



Ve směrovačích byl po celou dobu testu použit IOS c7200-adventerprisek9-mz.124-15.T1.bin, který jako jeden z mála podporuje všechny klíčové technologie a protokoly pro rychlou konvergenci. Bohužel se jedná o vývojovou větev operačního systému IOS. Dosud neexistuje produkční verze, která by všechny potřebné technologie obsahovala.

Na síti byla nakonfigurována L3VPN (v Cisco terminologii VRF) s názvem „test“. Uvnitř této VPN byla ze stanice 192.168.0.6 testována dostupnost loopbacku na routeru R3 192.168.0.3. Při různých nastaveních jsme testovali dobu, po kterou byla přerušena síťová komunikace v případě rozpojení fialové linky nebo tvrdého vypnutí routeru R1, přes který data tekla (viz administrativně nastaveny „cost“ na zelené lince). Poté jsme měřili totéž při obnovení původní topologie sítě (spojená linka / nastartovaný router R1).

3.1 Standardní konfigurace

Pro referenci bylo provedeno i měření na „standardní konfiguraci“. To znamená, že všechny protokoly (OSPF, BGP a LDP) byly ve svých standardních konfiguracích. Zároveň to znamená že nebyly nakonfigurovány žádné techniky pro rychlou konvergenci (BFD, LDP-IGP synchronizace ani nulová prodleva na rozhraních). Mělo se ukázat, o kolik je možno konvergenci zrychlit použitím níže uvedených konfigurací. Kompletní nastavení je zobrazeno v **příloze A**.

3.2 Konfigurace rychlé IGP a vypnutí prodlevy na rozhraních

OSPF bylo nastaveno pro rychlou konvergenci a na rozhraních byly použity příkazy, zajišťující okamžité informování procesů při detekci chyby hardwarem rozhraní. Relevantní část konfiguraci najdete v tabulce níže, kompletní nastavení směrovačů pak v **příloze B**.

```
interface FastEthernet1/1
 ip address ...
 carrier-delay msec 0
 ip ospf  network point-to-point
 dampening
!
router ospf 1
 timers throttle spf 50 50 5000
 timers throttle lsa all 0 20 5000
 timers lsa arrival 15
 timers pacing flood 15
!
```

3.3 Spuštění protokolu BFD

Protože je zřejmé, že zvláště na WAN linkách s LAN protokolem Ethernet není dobře vyřešena detekce poruch, je na všech třech linkách v našem testovacím scénáři spuštěn protokol BFD. Opět uvádím příslušnou část konfigurace, kompletní výpis je možno najít v **příloze C**.

```
interface FastEthernet1/1
 bfd interval 50 min_rx 50 multiplier 3
 bfd neighbor 10.0.0.1
!
router ospf 1
  bfd all-interfaces
!
```

3.4 Synchronizace LDP a IGP

Aby nedocházelo k blackholingu provozu v MPLS síti, byla v další části nakonfigurována synchronizace OSPF a protokolu LDP. Zapnutí je velmi jednoduché, viz tabulka. Zároveň bylo zapnuto i pomalé nabíhání OSPF při startu routeru pro zabránění blackholingu z důvodu ustavení OSPF přes bootující router v době, kdy ještě není schopen směrovatpakety. Kompletní výpis viz **příloha D**.

```
router ospf 1
  mpls ldp sync
  max-metric router-sla on-startup 90
```

4 Měření

Cílem měření bylo odladění konfigurace routerů pro rychlou konvergenci a ověření teorií uvedených výše. Z poznatků z praxe i teoretických úvah se zdálo, že by neměl být problém dosáhnout rychlosti konvergence menší než jednu sekundu. Tento předpoklad měl být potvrzen nebo vyvrácen.

Všechna následující měření byla provedena za stejné situace pětkrát za sebou, a do tabulek byl zaznamenán nejnižší počet ztracených paketů, nejvyšší počet ztracených paketů a průměrná hodnota, a to jak při vzniku problému (porucha) tak i při odstranění problému (náběh).

Měření bylo z nedostatku lepších měřicích přístrojů prováděno notebookem s OS Linux a programem hping[5], u kterého je možnost nastavit konstantní odstup vysílání paketů v mikrosekundách. Při nastavení dle tabulky bylo změřeno že program generuje cca 70 paketů za vteřinu. Potom není problém podle počtu ztracených paketů určit čas výpadku obousměrné komunikace s postačující přesností. Kompletní tabulka všech měření je k dispozici v **příloze E**.

```
# cca 70 paketů za sekundu
hping2 -1 -i u10000 192.168.0.3
```

4.1 Výpadek aktivní linky (obousměrně)

Při obousměrném výpadku linky (čisté přerušení simulované příkazem shutdown na rozhraní Gi0/2 směrovače R1) vědí o výpadku oba směrovače zároveň a v závislosti na použitém čipsetu nejpozději za cca 20ms (údaje získané přímo od techniků firmy Cisco Systems). Rozdíly jsou v řádech jednotek milisekund. Neměli bychom být tedy závislí na BFD a nebo IGP detekci problému. Problém byl simulován přerušením metalického spoje mezi routery R1 a R3 (fialová linka).

Konfigurace	Akce	min lost packets	average lost packets	max lost packets
Standardní konfigurace	porucha	2444	**2534**	2636
	náběh	0	**120**	337
Rychlé IGP	porucha	10	**31**	67
	náběh	242	**310**	375
Navíc na linkách BFD	porucha	25	**46**	68
	náběh	0	**121**	304
Navíc LDP-IGP sync.	porucha	8	**32**	69
	náběh	0	**0**	0

Problém s rychlostí konvergence při pádu linky v tomto scénáři vyřeší již rychlé časování IGP. Problém blackholingu při náběhu linky řeší až „IGP LDP synchronizace“. Nejhorší naměřená rychlost konvergence v poslední konfiguraci byla cca **0,97** s.

4.2 Výpadek aktivní linky (jeden směr)

Při přerušení jen jednoho vlákna zůstává jeden ze směrovačů na základě detekce signálu v nevědomosti o výpadku na lince a detekce je provedena až na základě BFD či IGP. Předpokládá se tedy o něco pomalejší konvergence než v prvním případě. Oproti předchozímu případu je patrný značný rozdíl mezi jen zkonfigurovaným rychlým IGP a nasazením BFD protokolu.

Konfigurace	Akce	min lost packets	average lost packets	max lost packets
Standardní konfigurace	porucha	2304	**3295**	4634
	náběh	0	**26**	130
Rychlé IGP	porucha	605	**1780**	2697
	náběh	0	**203**	372
Navíc na linkách BFD	porucha	14	**26**	66
	náběh	0	**120**	318
Navíc LDP-IGP sync.	porucha	13	**21**	37
	náběh	0	**0**	0

Ve výsledcích je vidět, že problém s rychlostí konvergence při pádu vyřeší až BFD, neboť v tomto případě ani rychlé IGP není schopno včas detekovat jednostranný problém na lince. Až BFD problém řeší. Problém blackholingu při náběhu linky řeší až „IGP LDP synchronizace“. Nejhorší naměřená rychlost konvergence v poslední konfiguraci byla cca **0,53** s.

4.3 Výpadek P směrovače

Výpadek P směrovače jsme simulovali jeho tvrdým vypnutím během provozu. Doba výpadku by se neměla lišit od druhého případu, dokonce by měla být bližší prvnímu případu. Zajímavější je situace při opětovném náběhu směrovače do provozu.

Konfigurace	Akce	min lost packets	average lost packets	max lost packets
Standardní konfigurace	porucha	199	**288**	344
	náběh	1970	**2173**	2397
Rychlé IGP	porucha	0	**40**	85
	náběh	3	**2423**	4141
Navíc na linkách BFD	porucha	22	**50**	83
	náběh	0	**2753**	3689
Navíc LDP-IGP sync.	porucha	8	**37**	69
	náběh	0	**0**	0

V tabulce je jednoznačně vidět, že zatímco pád routeru a jeho detekce probíhá podle stejného scénáře jako v případě pádů linek, velkým problémem je náběh routeru. Zde rychlé IGP problém blackoligu prohlubuje, neboť se zvětšuje odstup mezi dobou, kdy je na lince již ustavené IGP a dobou, než proběhne výměna MPLS labelů protokolem IGP. Problém náběhu routeru zde odstraní až „IGP LDP synchronizace“. Nejhorší naměřená rychlost konvergence v poslední konfiguraci byla cca **0,99** s.

5 Závěr

Na základě měření byla ověřena praktická dosažitelnost méně než sekundové konvergence na současné generaci směrovačů v prostředí IP/MPLS sítě, a to i bez použití pokročilých technik pro rychlou konvergenci (například MPLS Fast Reroute). Při standardním nastavení směrovačů byla průměrná rychlost konvergence při všech třech druzích výpadku **29,1** s a nejhorší pak celých **66,2** s tedy více než minutu! V poslední, odladěné variantě konfigurace pak byl průměrný dosažený čas **0,43** s a nejhorší naměřený čas z celkem patnácti provedených měření byl **0,99** s.

Potvrdilo se, že rychlost je závislá na dvou faktorech. Na detekci samotného problému a na následné reakci na něj. Zatímco rychlost konvergence IGP je v Cisco operačním systému k dispozici již dlouhou dobu, detekce výpadků na WAN linkách na bázi Ethernetu protokolem BFD je relativní novinkou, dosud nedosažitelnou na nižších řadách směrovačů. Poslední z použitých technologií, tedy synchronizace IGP a LDP pouze odstraňuje obecnou nenavázanost těchto dvou protokolů na sebe navzájem a tak vznikající blackholing. Bohužel i tato vlastnost je dostupná pouze v experimentálních a vývojových verzích Cisco IOSu.

Rád bych upozornil, že všechny naměřené výsledky pocházejí z laboratorních podmínek. Mohou sice sloužit jako vodítko pro nastavení Vaší sítě, nemohou však postihnout všechny situace, ke kterým právě na Vaší síti může docházet. Obecně doporučuji opatrnost. Při ladění rychlejší konvergence na již běžící síti bych začal konzervativnějším nastavením timerů a následným dlouhodobým testováním. Předejdete tak problémům s případnou nestabilitou při

špatném chování některého směrovače a/nebo linky. Vždy je důležité zvážit, jak rychlou konvergenci ve vaší síti vlastně potřebujete. Pro datové služby jistě i konvergence kolem 2 s bude považována za velmi rychlou. Honba za milisekundami pak postrádá smysl a je lepší dát přednost stabilině.

Seznam použité literatury

1: Ward David; Haas Jeffrey, BFD Charters, 2007, http://www.ietf.org/html.charters/bfd-charter.html
2: Wikipedia, http://en.wikipedia.org/wiki/Open_Shortest_Path_First, ,
3: Wikipedia, http://en.wikipedia.org/wiki/IS-IS, ,
4: Wikipedia, http://en.wikipedia.org/wiki/Dijkstra's_algorithm, ,
5: Salvatore Sanfilippo, http://www.hping.org/, ,