Článek byl publikován dne 2.3.2009 na **www.lupa.cz**
http://www.lupa.cz/clanky/proc-a-zda-supronet-shodil-internet

Autoři: Ondřej Surý, Emanuel Petr

Supronet – co se doopravdy stalo?

Od události před čtrnácti dny, kdy došlo k nestabilitě internetu ve velké části světa, byla média a to i ta česká zaplněna zprávami o tom, jak Malý český ISP způsobil světový kolaps. Na tomto příkladu se opět ukázalo, že lynčování je nejjednodušším způsobem soudu a stále také velmi oblíbeným. Jelikož od onoho incidentu uplynulo trochu času a na internetu je mnoho textů a diskuzí, můžeme se s odstupem podívat na to, co se ve skutečnosti stalo, a s jakými chybnými informacemi se můžete setkat.

Již původní článek na Renesysu (http://www.renesys.com/blog/2009/02/the-flap-heard-around-the-worl.shtml) poukazoval na to, že s internetem, který je náchylný ke kolapsu takto velkého rozsahu, je něco špatně. Obecně se dá říci, že na všech stranách došlo k porušení tzv. Postelova zákona (http://en.wikipedia.org/wiki/Postel's_law), které říká: „Buď konzervativní v tom, co děláš, a buď liberální v tom, co přijímáš od ostatních.“ Pojďme si tedy teď podrobně projít reálné příčiny onoho „světového kolapsu“. Tyto příčiny jsou čtyři, seřadím je od nejdůležitějších:

1. Nově objevená chyba ve směrovačích Cisco
2. Chyba ve směrovači MikroTik
3. Nedostatečné nebo žádné filtry u tranzitních operátorů
4. Překlep v konfiguraci směrovače SUPRONETu

Směrovače Cisco – nevalidní aktualizace BGP

Hlavní příčina nestability BGP je schovaná v nově objevené chybě ve směrovačích Cisco. Chyba má číslo CSCsx73770 (http://tools.cisco.com/security/center/viewAlert.x?alertId=17670).

Příčina cyklického rozpadávání BGP spojení a jejich opětovného navazování spočívá nikoli na směrovačích, který zprávu BGP UPDATE přijímá, ale na směrovačích, který tuto zprávu generuje. Pokud délka cesty AS path při prodlužování překročí 255 ASNs, je vygenerována nevalidní BGP UPDATE zpráva, příjemce ohlásí chybu, a následně korektně (dle RFC) spojení ukončí. Směrovače se následně snaží spojení obnovit následkem čehož dochází k opakovaným restartům a generování množství BGP zpráv. Právě tato chyba ohrozila dne 16.února.2009 stabilitu Internetu.

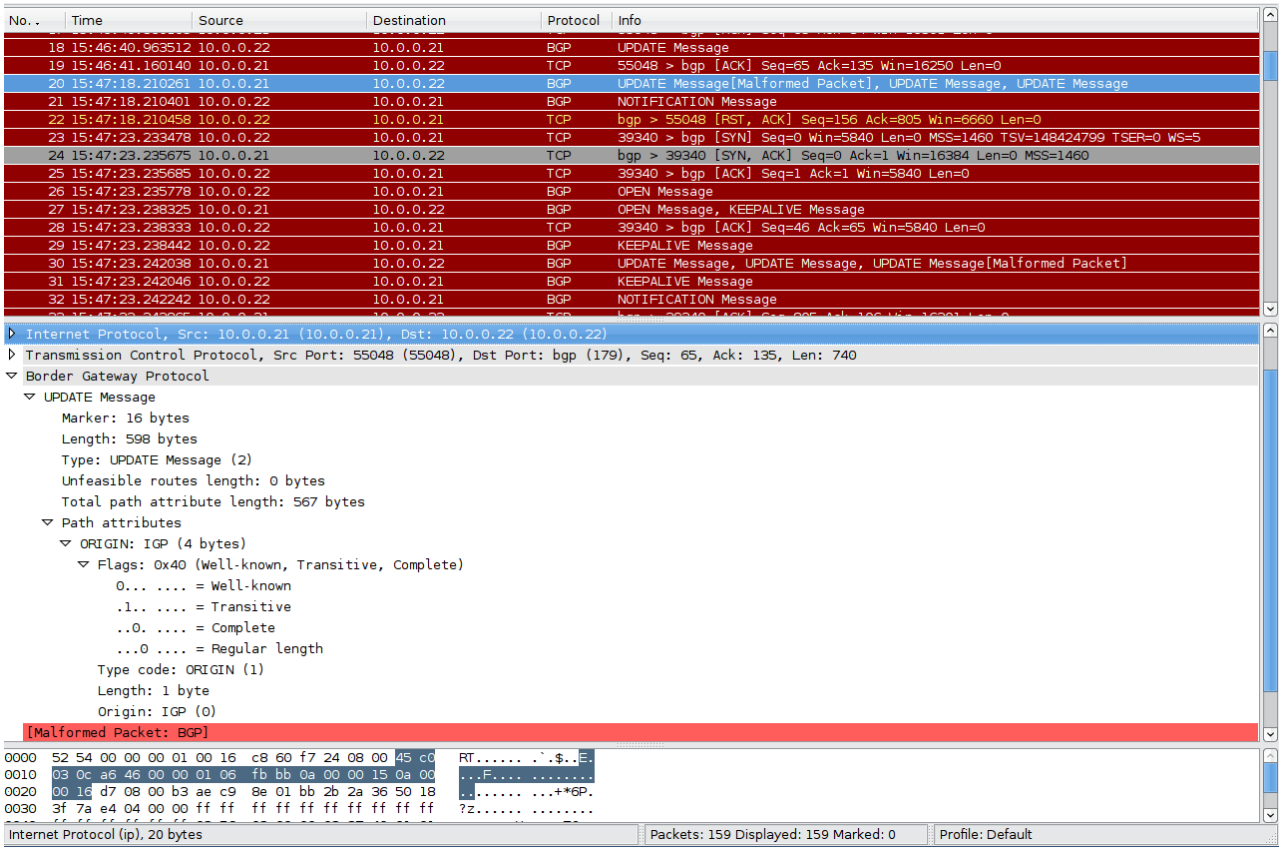
Konkrétní ukázka ze směrovače Juniper po přijetí vadné BGP update zprávy:

```
Feb 28 01:53:34.710355 bgp_read_v4_update:8010: NOTIFICATION sent to 10.0.0.21
(External AS 64501): code 3 (Update Message Error) subcode 11 (AS path
attribute problem)
```

Cisco záznam:

```
*Feb 27 22:42:42.077: %BGP-3-NOTIFICATION: received from neighbor 10.0.0.20
3/11 (invalid or corrupt AS path) 0 bytes
*Feb 27 22:42:42.077: %BGP-5-ADJCHANGE: neighbor 10.0.0.20 Down BGP
Notification received
```

Výpis komunikace a ukázka špatné zprávy BGP UPDATE:



Jak lze této chybě zabránit? Možnosti jsou dvě. Buď nepoužívat příkaz „set as-path prepend“ nebo v lepším případě implementovat omezení délky AS path u importovaných prefixů příkazem „bgp maxas-limit“. Druhý způsob je zodpovědnější a obecně zvyšuje bezpečnost dalších BGP peerů.

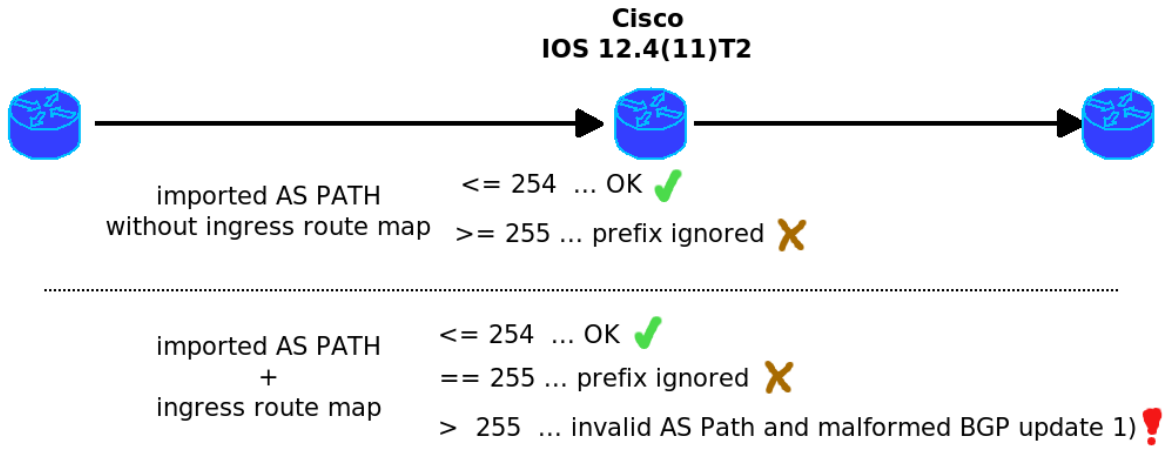
Na kolik AS path omezit? Pokud vezmeme v potaz, že starší IOS měl problém už s AS path > 128 (Cisco bug ID CSCdr54230 - opraveno v IOS 12.2), je vhodná hodnota někde z rozsahu 70-100. Osobně bychom doporučili používat hodnotu 70.

```
#bgp maxas-limit ?
  <1-2000> Number of ASes in the AS-PATH attribute
```

Pozn.: Některé zdroje odkazují na implicitní hodnotu 75. Ta byla však platná do IOS verze 12.3, od IOS 12.4 je hodnota zvýšena na 255. V určitých podverzích IOS 12.4 umožňuje příkaz „bgp maxas-limit“ zadat více než 255 ASNs. Limit však zůstane stále 255 a to i v případě, neuvedete-li příkaz vůbec.

Testováno s IOS 12.4(11)T2.

Pro import platí:



1) if an egress route map is applied, then the prefix is ignored and the BGP update will be correct

Pozn.: Zajímavá je možnost, na kterou jsme přišli v rámci laboratorního testování. Chybný prefix v odchozím BGP UPDATE lze eliminovat aplikací route-map (i prázdné) na odchozí prefixy. Tento chybný prefix je z BGP UPDATE zprávy odstraněn.

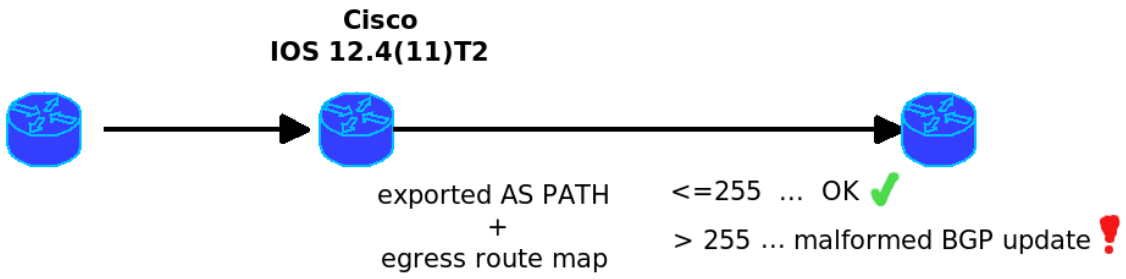
Normálně vypadající AS path pro importovaný prefix 10.111.0.0/16.

```
cisco# show ip bgp
*  10.111.0.0/16    10.0.0.111                             0 64503 64503
64503 ... 64503 i
```

Znetvořené AS path při importu, kde došlo k prodloužení AS path nad 255.

```
cisco# show ip bgp
*> 10.111.0.0/16    10.0.0.111                             0  i
```

Pro export platí:



Směrovače MikroTik - špatné zpracování parametrů

Dokumentace ke směrovači MikroTik (http://www.mikrotik.com/testdocs/ros/2.9/routing/filter.php) říká:

```
set-prepend (integer: 0..16) - specifies how many times the router should
prepend its AS number to the AS_PATH BGP attribute value for this route
```

Příkaz k prodloužení AS path by měl akceptovat pouze hodnoty 0-16. Administrátor ze SUPRONET-u zadal místo násobku, číslo svého AS 47868. Implementace MikroTiku neprovedla žádnou kontrolu, jak bychom předpokládali, a místo nahlášení chyby použila dolních 8-bitů z čísla autonomního systému. Zbytek po dělení 256 z čísla autonomního systému 47868 je 252, což je přesně počet, který byl připojen do AS path. Nezáleží tedy na absolutní hodnotě čísla autonomního systému, ale na jeho konkrétní hodnotě – pokud by číslo autonomního systému bylo jen o čtyři větší nestalo by se pravděpodobně vůbec nic, protože zbytek po dělení by byl roven 0.

Proč administrátor SUPRONETu zadal konfiguraci chybně můžeme jen spekulovat, ale pokud srovnáme například Cisco a Juniper, tak oba dva směrovače mají pro prodloužení AS path dva příkazy: U prvního je parametrem výčet čísel autonomních systémů a druhý má jako parametr násobek. Nejspíše tedy došlo k záměně a díky chybějící kontrole konfigurace společně s chybou ve směrovačích Cisco tato lidská chyba mohla napáchat tolik škody.

Mimochodem na http://bgpmon.net/maxASpath.php se lze přesvědčit o tom, že Uherský Brod není jediný na světě a ani první, kdo poslal takto dlouhou cestu. Týden předtím poslala indonéská společnost Core Mediatech (D-NET), PT prefix, který byl měl cestu jen o jedničku kratší.

Důležité je také říci, že MikroTik i přes svou chybu nemohl vygenerovat AS path, které by bylo v rozporu s RFC4271. Ve specifikaci BGP je délka atributů omezena pouze maximální velikostí samotné BGP UPDATE zprávy (4kB). Bohužel v praxi se stává, že pokud nejsou pevně stanoveny limity, výsledné implementace to řeší svými limity, případně nejsou nastaveny limity žádné, a dochází k fatálním chybám.

Nedostatečné nebo žádné filtry u tranzitních operátorů

Na směrovačích Cisco je od verze IOS 12.4 implicitní limit na maximální počet čísel autonomních systémů v cestě nastavený na 255. Pokud bude tato hodnota vyšší bude taková cesta zahozena. Ostatní směrovače (Juniper, Quagga, Bird), které jsme zkoušeli (http://blog.nic.cz/2009/02/25/as-path-prepending/), byly v pořádku schopny naimportovat i cesty delší (konkrétně 560). Dle dalšího testování se také ukazuje, že i Quagga (verze < 0.99.10) má problémy při přeposílání cest delších než 255.

Jak se ovšem ukázalo před 14ti dny, je více než žádoucí nespoléhat se na implicitní konfigurace, a omezit BGP relace od svých zákazníků na rozumné limity. Pokud by poskytovatel záložního tranzitního připojení odfiltroval takto dlouhé cesty, nemuselo vůbec dojít k propagaci do páteřní sítě a tím způsobit nestabilitu internetu na celém světě.

Pár poznámek na závěr aneb sovy nejsou tím, čím se zdají být

Po vychladnutí hlav a důkladnější analýze se zdá, že onen nešťastník ze SUPRONETu, kterému byla přisuzována hlavní vina za nestabilitu internetu před čtrnácti dny, je spíše oběť náhody a kombinací několika chyb dohromady. Bohužel byl výpadek natolik velký, že se začalo ukazovat prstem dříve než se provedla důkladnější analýza, aby se zjistila pravá příčina onoho masivního výpadku připojení v některých částech světa.

Někteří komentátoři (především v mainstreamových médiích) se podivovali nad tím, že se České republiky výpadek nedotkl. Jelikož nemám v ruce tvrdá čísla, dovolím si zde lehkou spekulaci. Jelikož chyba vznikla v České republice, tak většina sítí v České republice je relativně blízko (nikoli geograficky, ale počtem BGP prependů), nestihl počet čísel autonomních systémů v AS path dosáhnout čísla 255. Další možnosti toho, proč některé země nebyly postiženy a některé ano, můžou být v nepoužívání Cisco

zařízení, dále např. malý počet cest, kterými je celá země připojena do páteřní sítě a dobře nastavené filtry, případně příliš velká „vzdálenost“ - BGP relace se rozpadla dříve než prefix s dlouhou cestou stihl dorazit. Přesnost měření společnosti Renesys se také opírá o počet sond, které v konkrétní části sítě mají. Počty vzorků se můžou v různých místech výrazně lišit a způsobovat tak zkreslení výsledků.

A ještě jedna poznámka na závěr, kterou bychom chtěli zdůraznit – rozpojování BGP relací v případně chybné zprávy BGP UPDATE je korektní chování a není omezeno jen na maximální počet prefixů. Stejné chování se projevuje například, pokud směrovač, který umí zpracovávat 4-bajtové čísla autonomních systémů, dostane skrz operátora, který tyto rozšířené BGP atributy pouze přeposílá dálě, nevalidní data v rozšiřujících atributech pro 4-bajtové ASN (Zdroj: NANOG http://mailman.nanog.org/pipermail/nanog/2009-January/006816.html).

Další zdroje:

1. http://blog.ioshints.info/2009/02/oversized-as-paths-cisco-ios-bug.html
2. http://blog.ioshints.info/2009/02/root-cause-analysis-oversized-as-paths.html
3. http://bgpmon.net/blog/?p=125