Směrování mezi autonomními systémy

Jak funguje Internet?

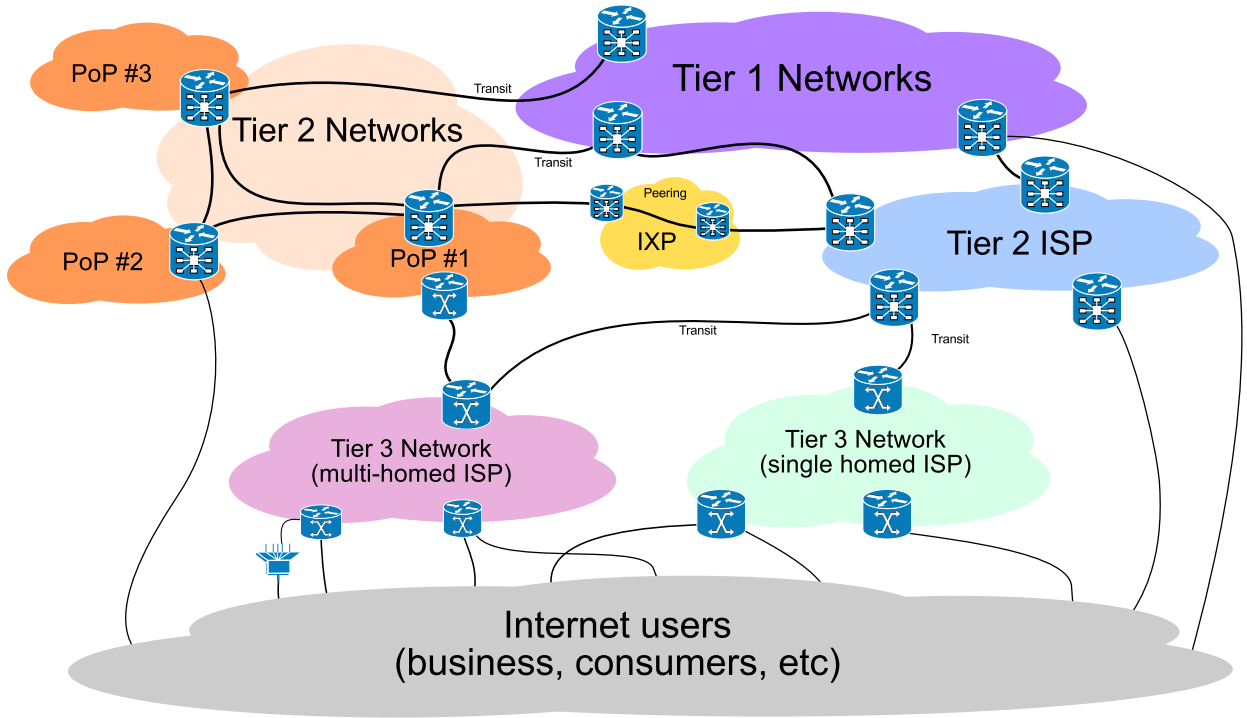
Inter — net
mezi — sítěmi.

ICANN

iana
Internet Assigned Numbers Authority

IETF

RIPE NCC RIPE NETWORK COORDINATION CENTRE



Regulační a řídící orgány

ICANN (Internet Corporation for Assigned Names and Numbers) organizace zastřešující veškeré administrativní činnosti Internetu.

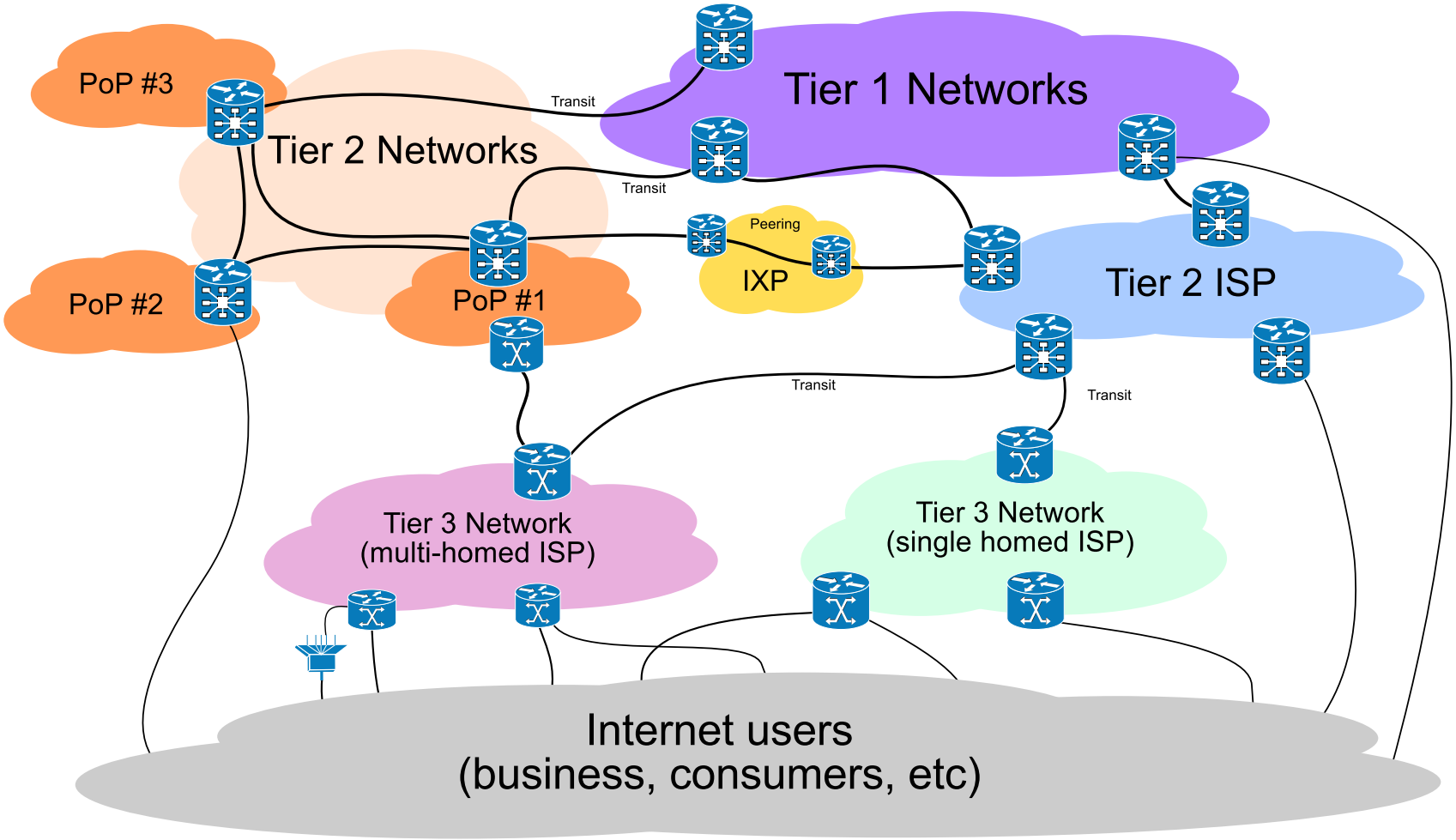
IANA Internet Assigned Numbers Authority, stará se o globální dohled nad přidělováním IP adres, čísel autonomních systémů, správu root DNS

RIR (Regional Internet Registry) – přiděluje adresy LIRům a koncovým zákazníkům, udržuje databázi přidělených IP rozsahů, atd. RIR jsou regionálně rozděleni následovně:

- **ARIN** American Registry for Internet Numbers
- **RIPE NCC** Réseaux IP Européens Network Coordination Centre
- **LACNIC** Latin America and Caribbean Network Information Centre
- **AfriNIC** African Network Information Centre
- **APNIC** Asia-Pacific Network Information Centre

LIR (Local Internet Registry) – organizace přidělující IP adresy koncovým zákazníkům/uživatelům. Obvykle ISP, státní organizace, atd.

Úrovně ISP (neformální členění)



Úrovně ISP (neformální členění)

Tier 1 velcí celosvětoví tranzitní ISP, kteří se mezi sebou propojují bez propojovacích poplatků.

Tier 2 ISP, kteří nakupují konektivitu, aby se dostali do jiných částí Internetu, s některými částmi sítě se propojují po vzájemné dohodě bez poplatků.

Tier 3 takoví ISP, kteří veškerou svoji konektivitu nakupují od větších poskytovatelů.

IXP Internet Exchange Point – obvykle regionální centrum, které poskytuje centralizované propojovací místo mezi místními poskytovateli. Zjednodušuje peering mezi ISP, ISP nemusejí budovat propojovací trasy mezi svými PoP (Point of Presence). V ČR je to např. NIX (`http://www.nix.cz/`), ve světě DE-CIX, AMS-IX, LINX, atd.

IP adresy z pohledu globálního směrování

PA IP (Provider Aggregatable IP addresses) – IP adresy přidělené poskytovatelem, není možné je přenášet od jednoho poskytovatele ke druhému

```
$ whois 217.69.97.0
inetnum:        217.69.97.0 - 217.69.97.15
netname:        IRI-CZ
...
status:        ASSIGNED PA
...
```

PI IP (Provider Independent IP addresses) – IP adresy nezávislé na poskytovateli, je možné je přenášet mezi ISP. Obvykle svázané s vlastním autonomním systémem.

```
$ whois 147.251.0.0
inetnum:        147.251.0.0 - 147.251.255.255
netname:        MUNI-TCZ
...
status:         ASSIGNED PI
...
```

Autonomní systém

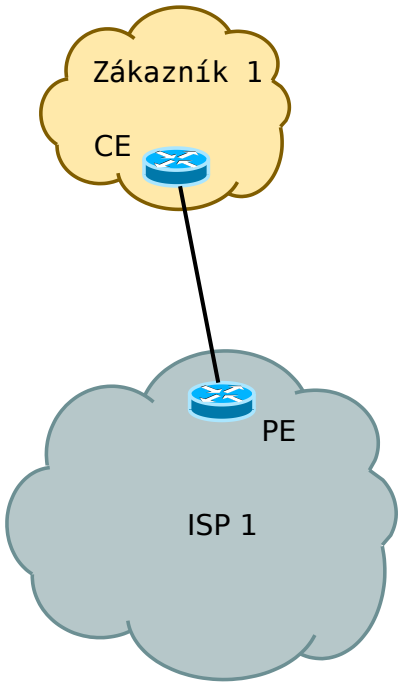
Autonomní systém (AS):

obecné pojetí – je část síťového IP prostoru pod správou jednoho subjektu s vlastní směrovací politikou.

EGP – je část síťového IP prostoru s vlastním číslem AS, podílející se na výměně směrovacích informací pomocí některého EGP (Exterior Gateway Protocol) – v dnešní době BGP (Border Gateway Protocol). Taková část síťového prostoru obvykle používá v rámci AS jeden IGP (Interior Gateway Protocol) (OSPF, IS-IS, EIGRP, atd.) Vydávání čísel AS se řídí pravidly jednotlivých RIR (obvykle je nutné mít vůbec potřebu čísla AS – např. být tranzitním AS nebo být dual-home zákazník). Pro přidělování čísel AS v Evropě v současnosti platí dokument RIPE-525 (srpen 2011).

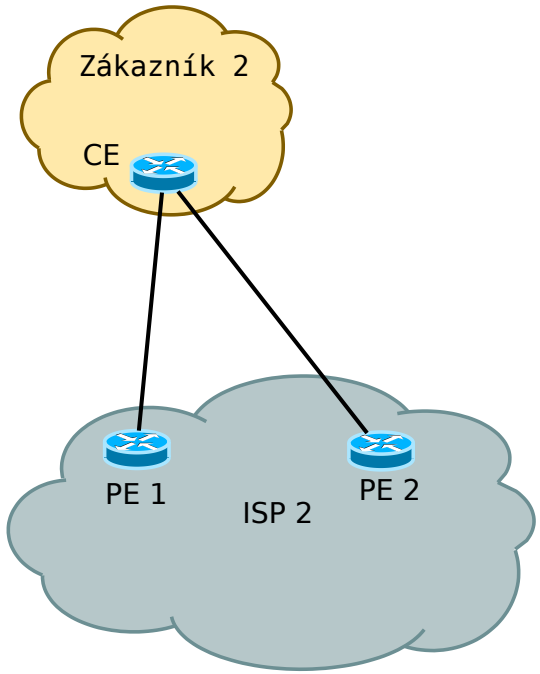
Typy autonomních systémů: single-homed, stub AS

single-homed zákazník, jediná linka k ISP – není nutné běžet žádný směrovací protokol. Vše obstarají statické cesty.



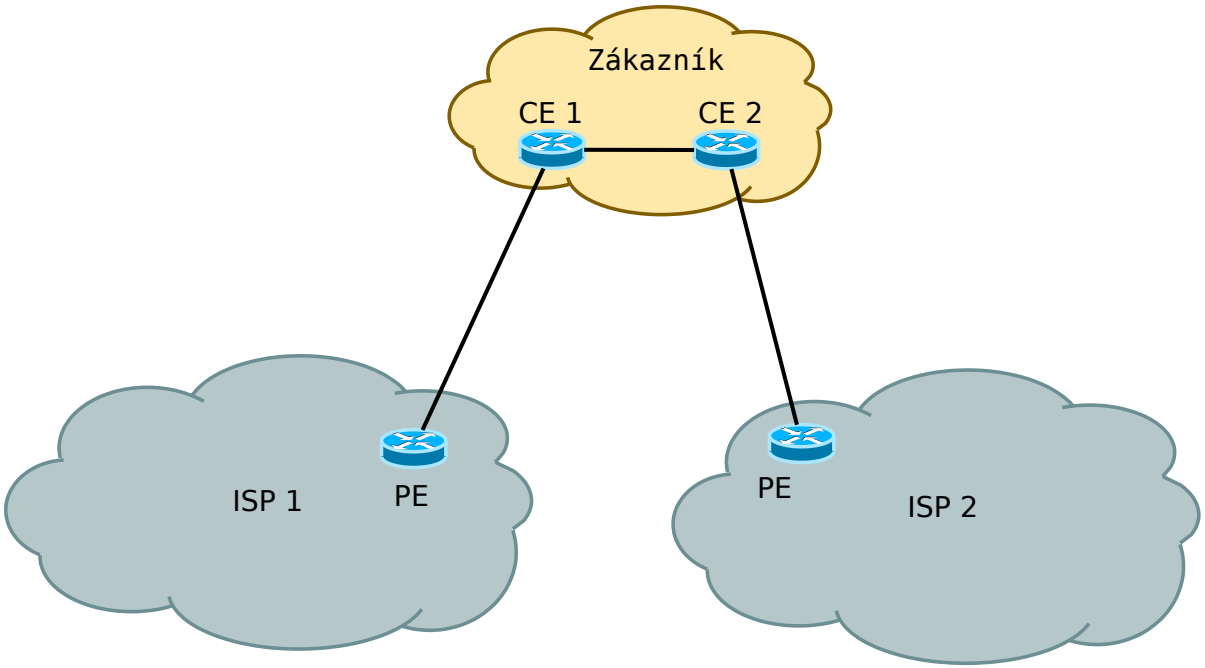
Typy autonomních systémů: single-homed, stub AS, dvě linky k jednomu ISP

single-homed zákazník, dvě linky k ISP – nutné běžet směrovací protokol. V případě BGP není nutné používat veřejně přidělené číslo AS, k dispozici jsou privátně přidělovaná čísla v rozmezí 64512 – 65534 pro 16 bitová čísla AS nebo 4200000000 – 4294967294 pro 32 bitová čísla AS.



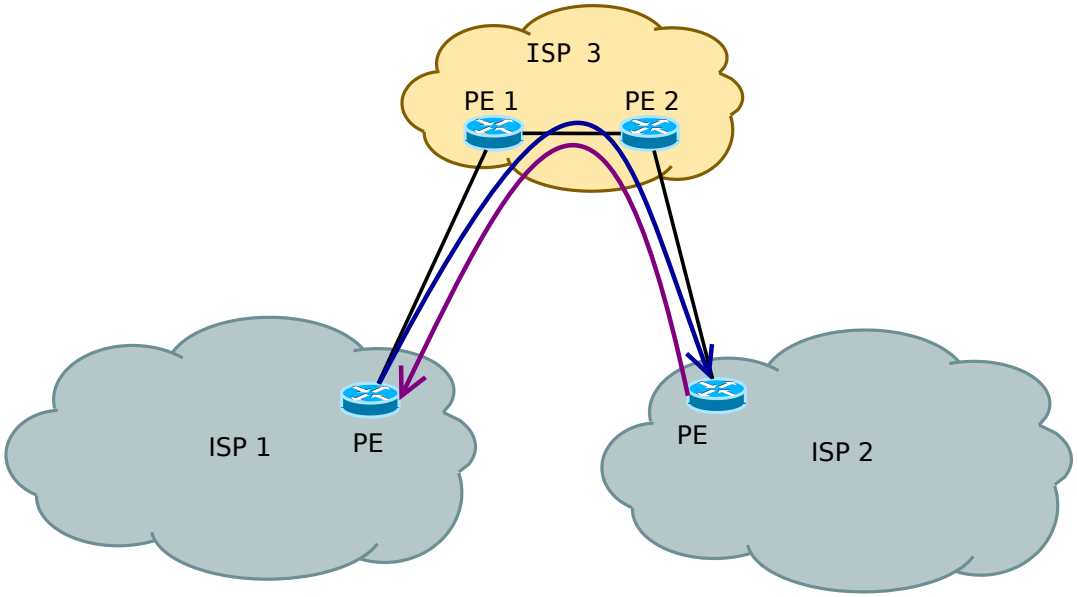
Typy autonomních systémů: multi-homed AS

netranzitní AS připojený ke dvěma ISP – obvykle zákazník nebo místní ISP toužící po vyšší dostupnosti. V tomto případě je již nutné BGP s PI IP adresami a veřejným číslem AS. Zároveň musí být BGP vhodně nakonfigurováno, aby se AS nestal tranzitním (aby nemohli ISP 1 a ISP 2 komunikovat přes linky vedoucí k zákazníkovi). Musí mít PI IP adresy.

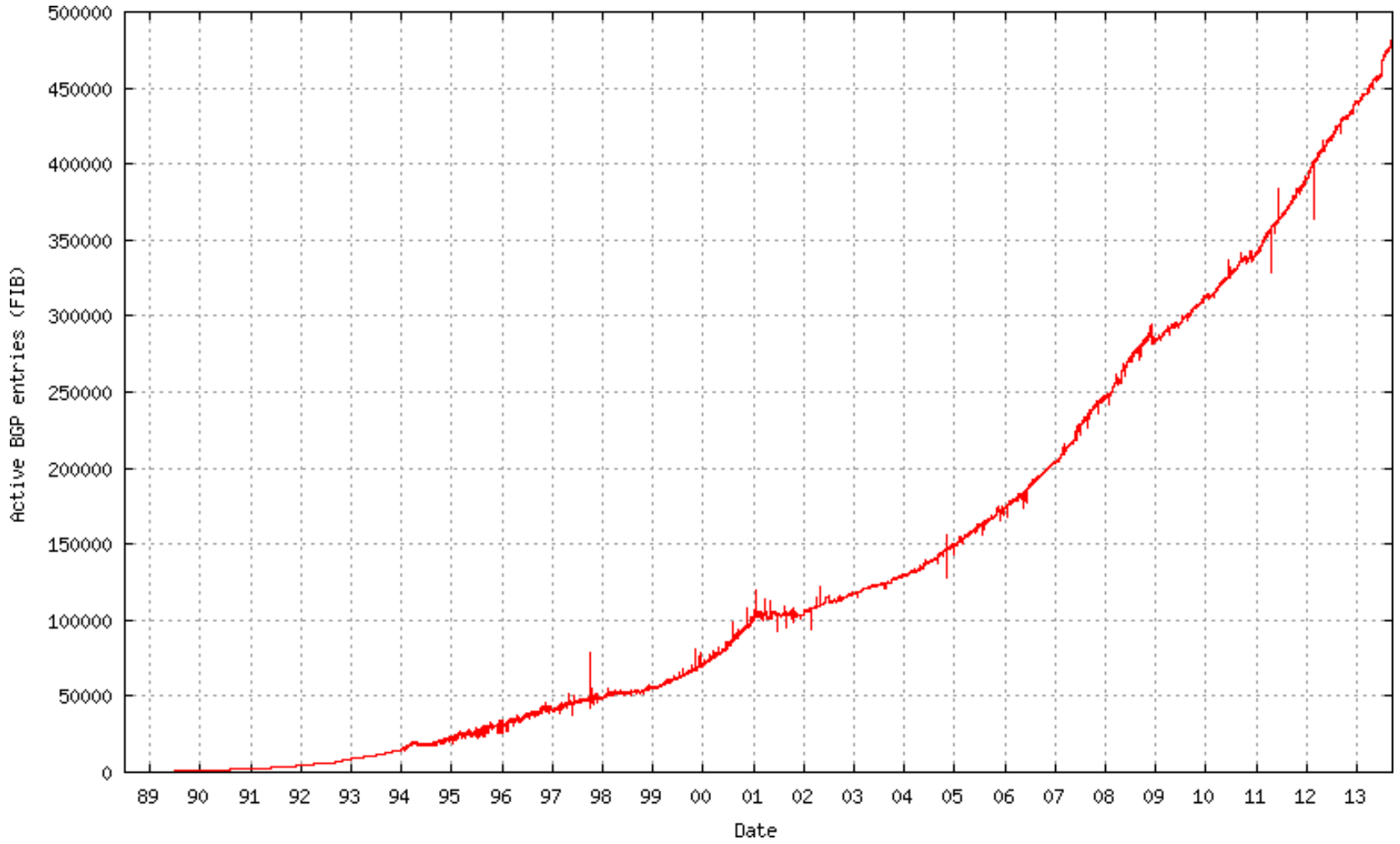


Typy autonomních systémů: tranzitní AS

tranzitní AS – propouštějí provoz mezi AS, se kterými daný AS sousedí. To, jaké cesty a jakým způsobem propouští, záleží na tzv. peerovací (propojovací) politice mezi AS a je řízeno konfigurací BGP.

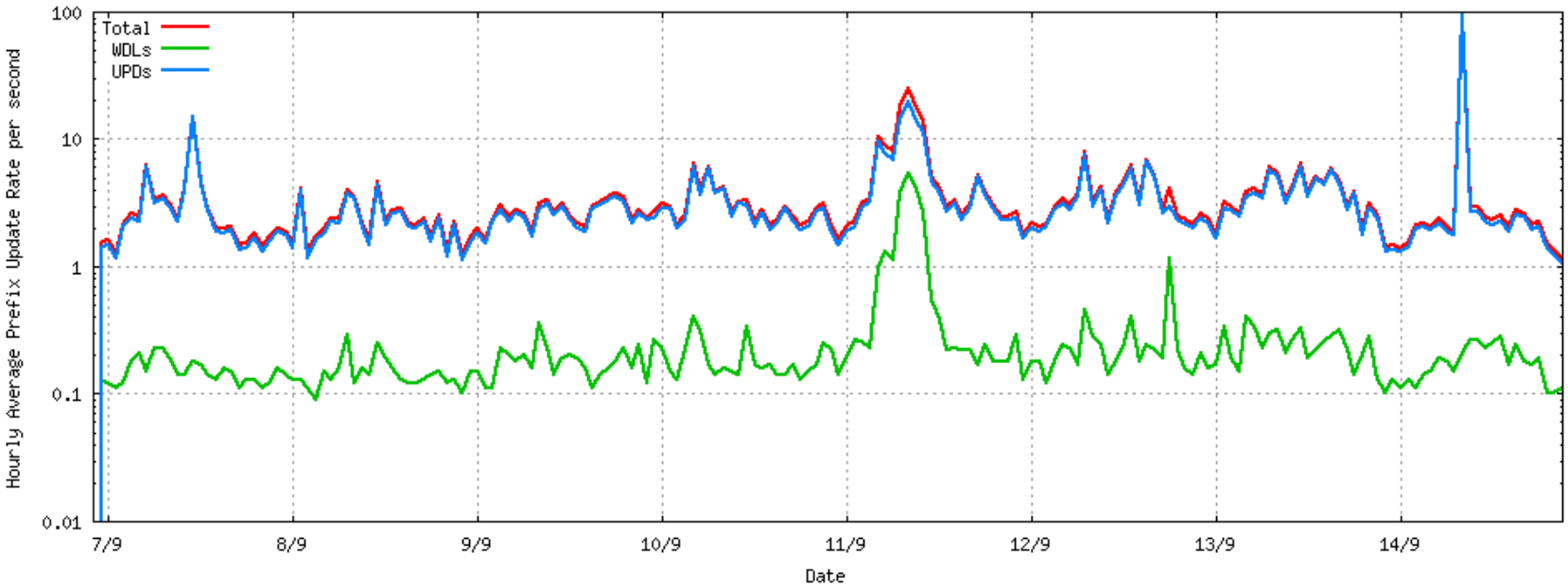


Velikost směrovacích tabulek v Internetu



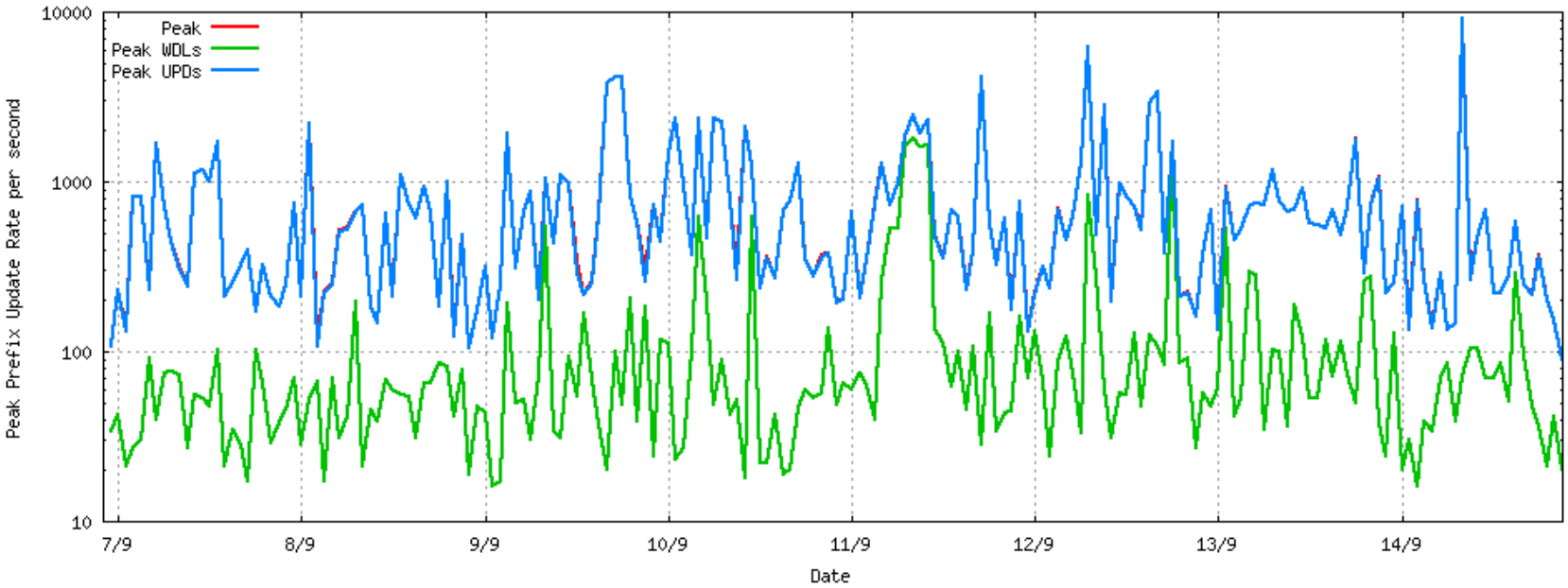
(autor: Geoff Huston `http://bgp.potaroo.net/`)

Četnost změn ve směrovacích tabulkách v Internetu – průměr za vteřinu během hodiny



(zdroj: `http://bgpupdates.potaroo.net/instability/bgpupd.html`)

Četnost změn ve směrovacích tabulkách v Internetu - maxima za vteřinu během hodiny



(zdroj: `http://bgpupdates.potaroo.net/instability/bgpupd.html`)

BGP

BGP (Border Gateway Protocol) – směrovací protokol určený pro směrování ve velkých sítích a mezi autonomními systémy.

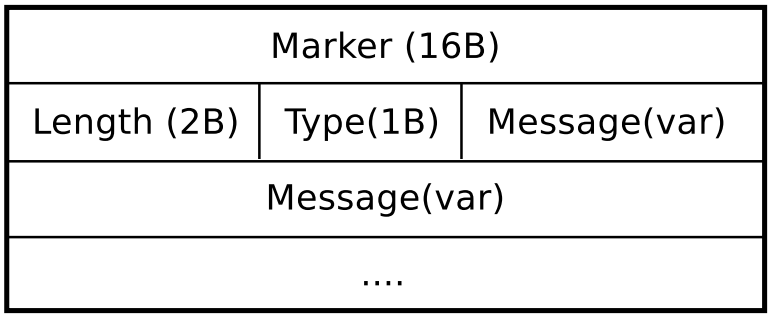
Vlastnosti:

- **rozšiřitelný** – je možné přidávat nové vlastnosti (směrování pro nové protokoly – IPv6, IP multicast, atd.)
- **ovladatelný** – je možné různými způsoby ovládat výběr cesty a šíření směrovacích informací
- **inkrementální** – při změně v síti se nepřenáší celá směrovací tabulka, pouze aktualizace

BGP

- postupně se vyvinul z protokolů EGP, BGP-1 – BGP-3
- RFC 4271 (RFC 1771)
- distance (path) vector protokol
- určen pro velké sítě
- bohatá metrika (atributy cesty)
- běží nad TCP (port 179) (co to znamená?)
  - soused musí být zadaný v konfiguraci
  - spolehlivý přenos
  - je nutné, aby byla cílová adresa dosažitelná (tj. buď přímo připojená nebo zde musí běžet ještě nějaký podkladový IGP směrovací protokol)

Formát předávaných zpráv

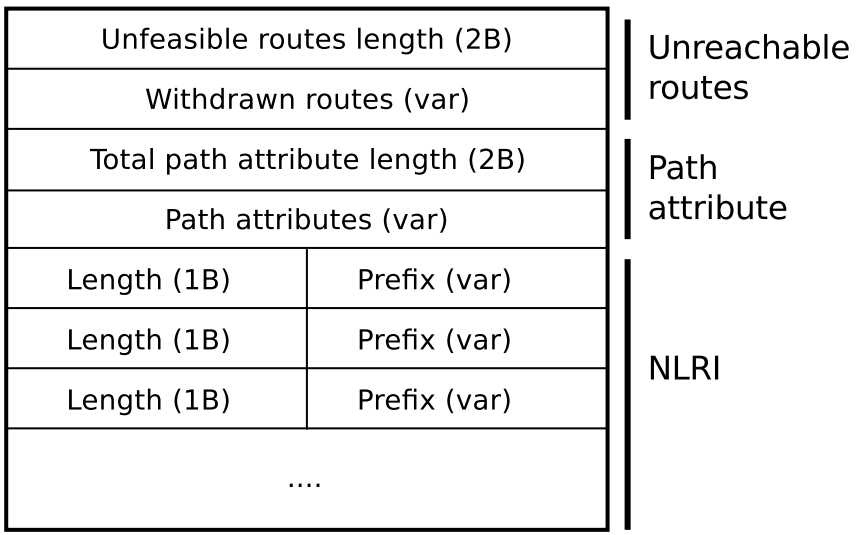


Typy zpráv:

Open	vyjednávání parametrů při ustavování spojení
Update	předávání směrovacích informací
Notification	oznamování chyb
Keepalive	kontrola živosti spojení

Update zprávy

Update zprávy nesou informace o předávaných a rušených cestách včetně atributů pro tyto cesty.



- unreachable routes – seznam NLRI, které jsou vyřazovány ze směrovacích tabulek
- path attributes – atributy cesty pro všechny následující NLRI
- NLRI (Network Layer Reachability Information) – dvojice [délka netmasky, síť]

Atributy cesty

Well-known, mandatory atribut, který musí být součástí každé cesty, každá implementace BGP mu musí rozumět. (ORIGIN, AS_PATH, NEXT_HOP, ...)

Well-known, discretionary atribut, kterému musí rozumět všechny implementace BGP, ale nemusí být součástí každé UPDATE zprávy. (LOCAL_PREF, ATOMIC_AGGREGATE).

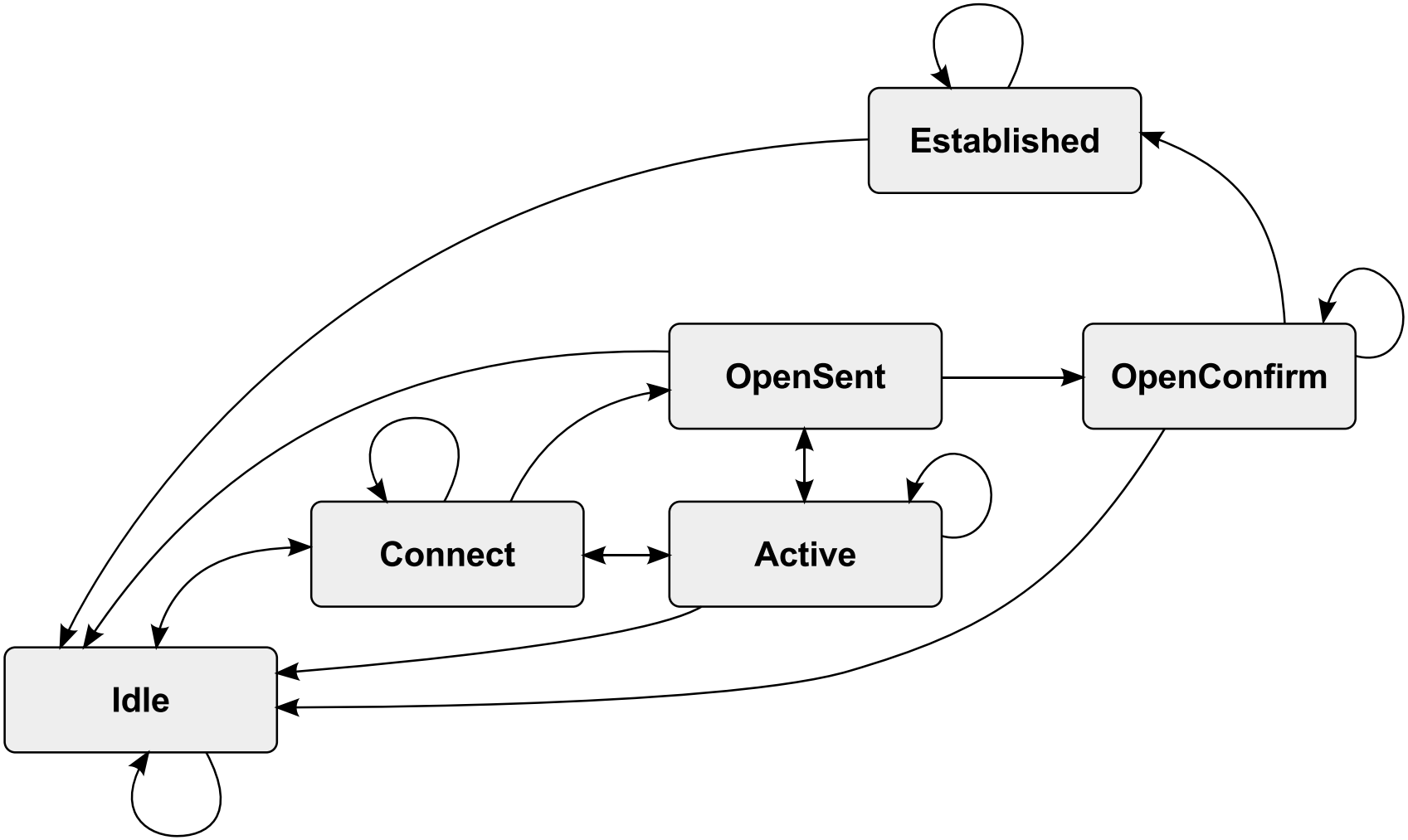
Optional, transitive nemusí znát všechny implementace BGP, nicméně i když je dané implementaci neznámý, musí jej předat svým sousedům. (AGGREGATOR, COMMUNITY)

Optional, non-transitive nemusí znát všechny implementace BGP. Tento atribut se nepředává dalším BGP sousedům. (ORIGINATOR_ID, Cluster List)

Základní informace o cestách na směrovači (Cisco)

```
#show ip bgp
BGP table version is 3417659, local router ID is 217.69.96.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, m multipath, b backup-path, f RT-Filter,
              x best-external, a additional-path, c RIB-compressed,
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI validation codes: V valid, I invalid, N Not found

     Network          Next Hop            Metric LocPrf Weight Path
 *   0.0.0.0          82.119.252.41            0    100      0 29208 i
 *>                   62.168.17.145   4294967295    200      0 2819 i
 *   1.0.0.0/24       82.119.252.41   4294967295             0 29208 15169 i
 *>                   62.168.17.145   4294967295             0 2819 15169 i
 *   1.1.1.0/24       82.119.252.41   4294967295             0 29208 15169 i
 *>                   62.168.17.145   4294967295             0 2819 15169 i
 *   1.2.3.0/24       82.119.252.41   4294967295             0 29208 15169 i
 *>                   62.168.17.145   4294967295             0 2819 15169 i
 *>  1.224.0.0/13     82.119.252.41   4294967295             0 29208 9318 i
 *>  1.232.0.0/13     82.119.252.41   4294967295             0 29208 9318 i
 *>  1.240.0.0/13     82.119.252.41   4294967295             0 29208 9318 i
 *>  1.248.0.0/13     82.119.252.41   4294967295             0 29208 9318 i
 *>  2.16.0.0/23      82.119.252.41   4294967295             0 29208 3257 i
 *>  2.16.4.0/24      82.119.252.41   4294967295             0 29208 3257 i
```

Stavový diagram BGP



Atributy cesty

NEXT_HOP povinný atribut, všechny implementace jej musejí rozpoznat (well-known, mandatory) určuje kam se má paket určený pro tuto síť poslat

AS_PATH povinný atribut, všechny implementace jej musejí rozpoznat. Určuje přes které autonomní systémy je tato síť dostupná. Zabraňuje směrovacím smyčkám (pokud je v AS_PATH číslo vlastního AS, směrovač musí tuto cestu vyhodit)

LOCAL_PREF volitelný atribut, všechny implementace jej musejí rozpoznat (well-known, discretionary). Určuje lokální preferenci dané cesty, šíří se pouze v rámci IBGP, nešíří se mimo hranice AS. Vyhrává největší.

MULTI_EXIT_DISC (MED, Multi Exit Discriminator) určuje metriky více různých cest mezi dvěma AS (tj. nevstupuje do rozhodování o cestách, pokud je stejná cesta přijata ze dvou různých AS). Je to metrika, vyhrává nejmenší.

Update zprávy předávané do směrovacích tabulek a sousedům

- ze všech stejných cest (tj. stejné NET/PREFIX) od všech BGP sousedů se vybere vždy pouze jedna nejlepší cesta
- nejlepší cesta je pak předávána BGP sousedům
- nejlepší cesta vstupuje do procesu instalace do směrovací tabulky směrovače. Zde záleží na administrativní metrice směrovacích protokolů.

Cisco

protokol	adm. metrika
connected	0
static	1
eBGP	20
OSPF	110
ISIS	115
RIP	120
EGP	170
iBGP	200

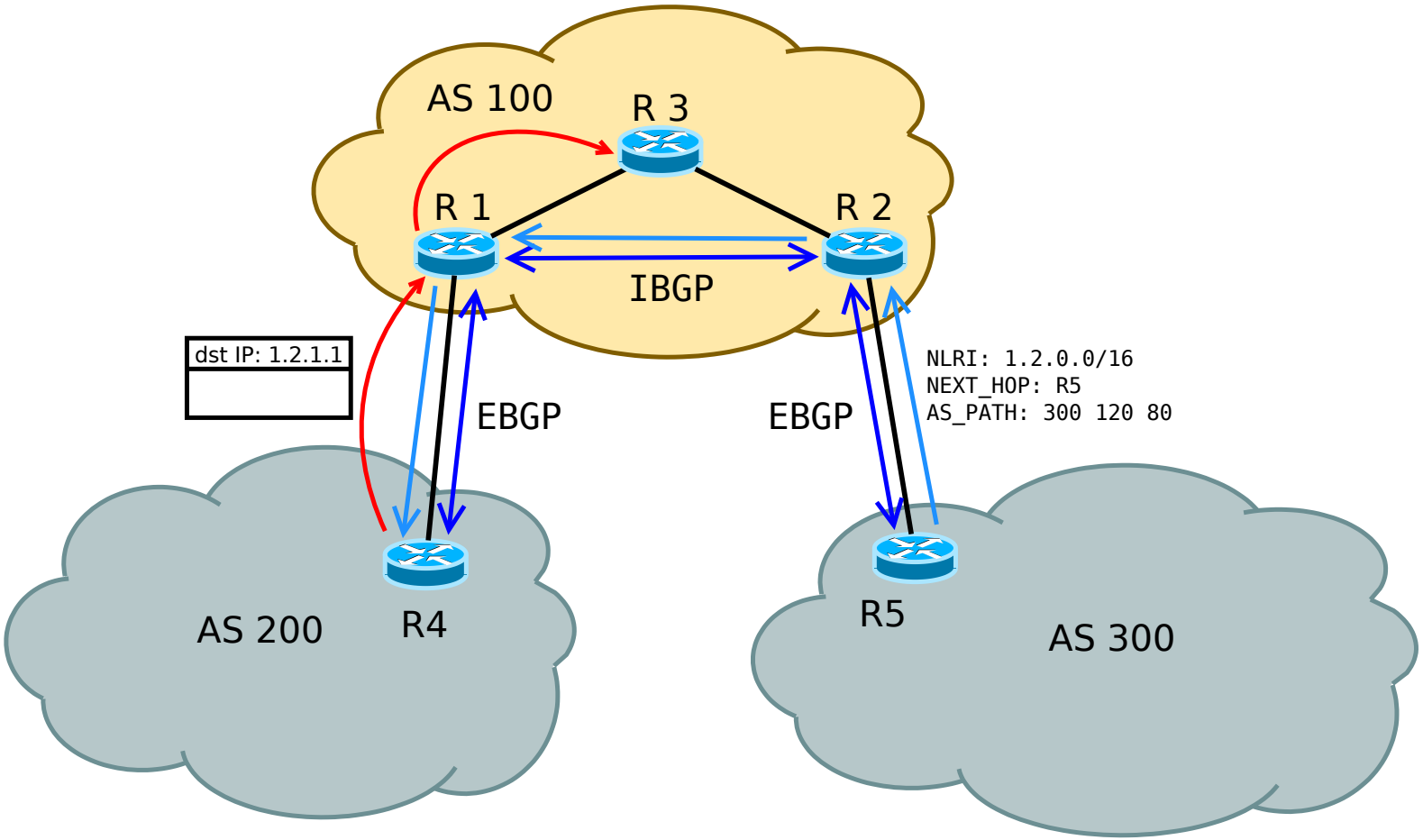
Juniper

protokol	adm. metrika
direct/local	0
static	5
LDP	9
OSPF int.route	10
ISIS L1 int.route	15
ISIS L2 int.route	18
RIP	100
OSPF AS ext.route	150
BGP	170

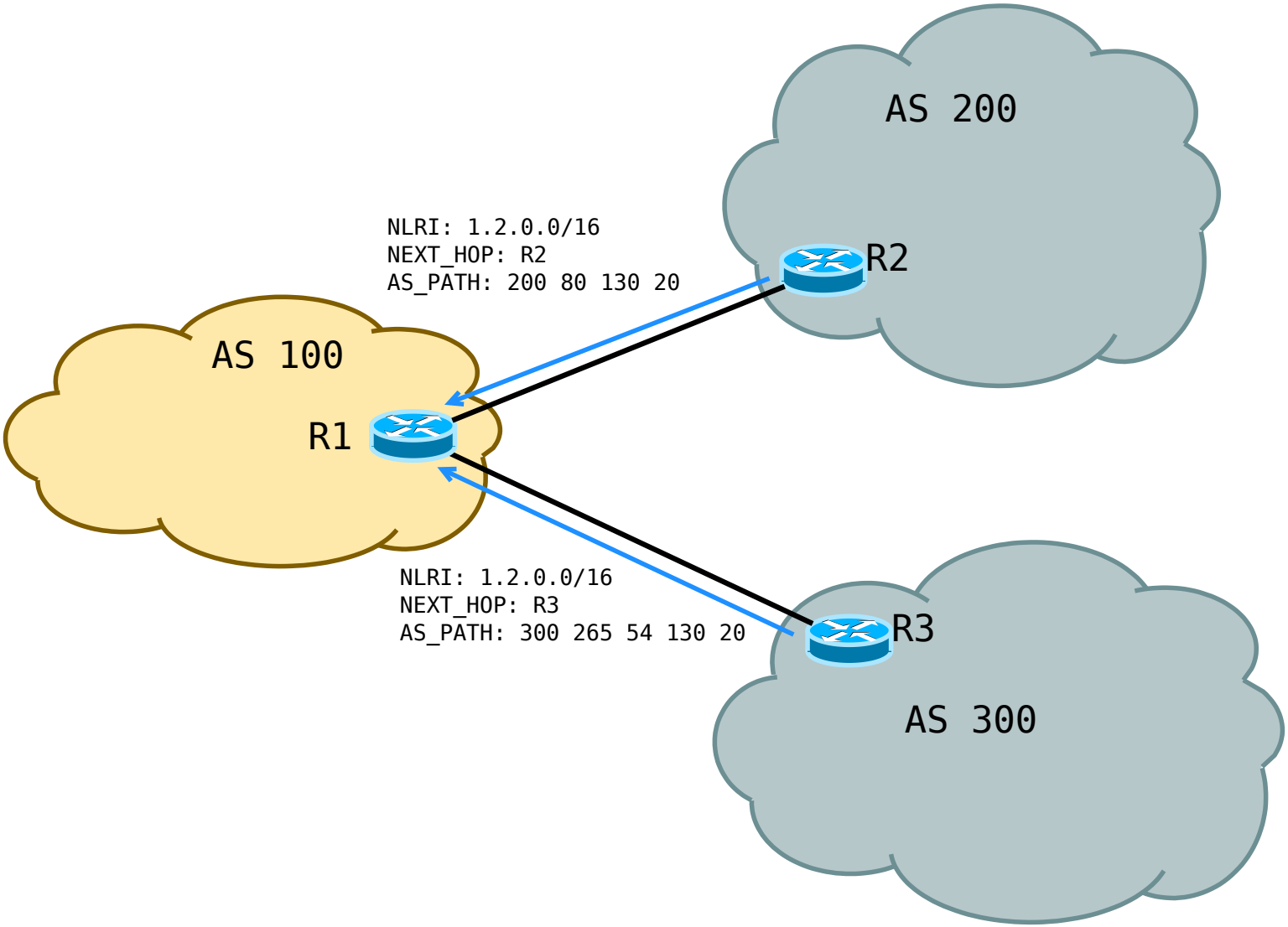
Rozhodovací algoritmus pro výběr nejlepší cesty

1. vyřaď cesty, které nejsou synchronizované s IGP a cesty s nedostupnou NEXT_HOP adresou
2. (preferuj největší weight (lokální na směrovači) – pouze Cisco)
3. preferuj nejvyšší local preference LOCAL_PREF (globální uvnitř AS)
4. preferuj cestu pocházející z tohoto routeru
5. preferuj nejkratší AS_PATH
6. preferuj nejnižší origin code ($IGP < EGP < INCOMPLETE$)
7. preferuj nejnižší MED
8. preferuj EBGP cesty proti IBGP cestám
9. pro IBGP cesty, preferuj cestu s nejbližším IGP NEXT_HOPem
10. pro EBGP cesty, preferuj nejstarší cestu
11. preferuj cesty ze směrovače s nejnižším BGP router ID
12. preferuj cesty s kratším Cluster-list atributem
13. preferuj cesty od BGP souseda s nejnižší IP adresou

Proč je (skoro) nutná synchronizace?



příklad rozhodování při výběru nejlepší cesty



Příklad využití komunit k řízení toku

```
29208:4100 - do not advertise to NIX
29208:4110 - 1x prepend to NIX
29208:4200 - do not advertise to SIX
29208:4210 - 1x prepend to SIX
29208:4300 - do not advertise to DECIX
29208:4310 - 1x prepend to DECIX
29208:4999 - do not advertise to ANY-transit
29208:5000 - Do not adverise to next AS(Tiscali)
29208:5001 - give routes localpref below normal customer route
29208:5002 - give routes localpref below normal peer route
29208:5003 - give routes localpref below transit route
29208:6000 - Do not export to next AS(Telia)
29208:6001 - Set local pref 50 within AS1299 (lowest possible)
29208:6002 - Set local pref 150 within AS1299 (equal to peer, backup)
```

Agregace

Agregace vytváří sumární cesty (tzv. agregáty) ze sítí obsažených v BGP tabulce.

- jednotlivé cesty mohou být zároveň agregovány nebo potlačeny
- agregovaná cesta bude propagována pokud existuje v BGP tabulce alespoň jedna podsíť této agregace
- jednotlivé sítě jsou stále propagovány v odchozích BGP updatech (pokud není řečeno jinak)
- atributy:

  Atomic aggregate (indikuje ztrátu AS_PATH informace), nesmí být odstraněn pokud už byl nastaven

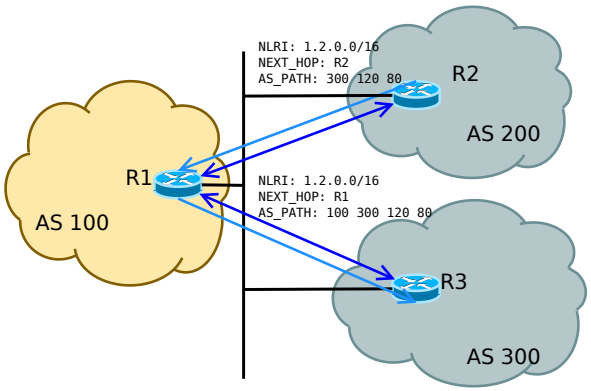
  Aggregator číslo AS a IP adresa routeru generující agregovanou adresu (vhodné pro ladění).

Manipulace s atributy

U každé přijaté cesty můžeme manipulovat s atributy. K tomu potřebujeme:

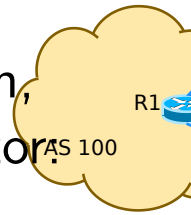
- metodu, jak najít cestu odpovídající našim zvoleným kritériím (match)
- možnost měnit atributy cesty (change/set)
- možnost jak celou cestu vyřadit (filter)

Manipulace s atributy - NEXT_HOP

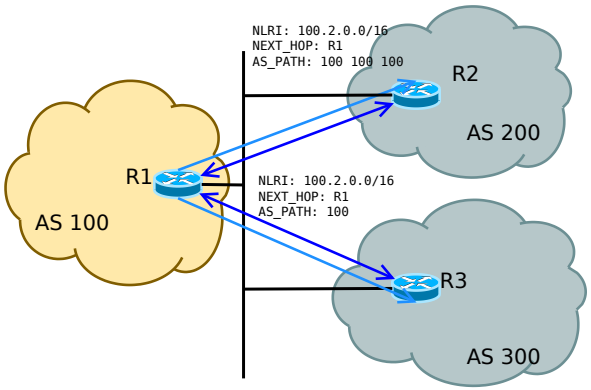


Manipulace s atributem NEXT_HOP umožňuje směrovači říct, že má pakety k cíli cesty směrovat jinam, než odkud mu přišel update (pozor: takový NEXT_HOP musí být ze směrovače dosažitelný)

Příklad: sdílené médium a neexistence přímé BGP relace mezi AS 200 a AS 300. (Směrování řídí route server v AS 100).

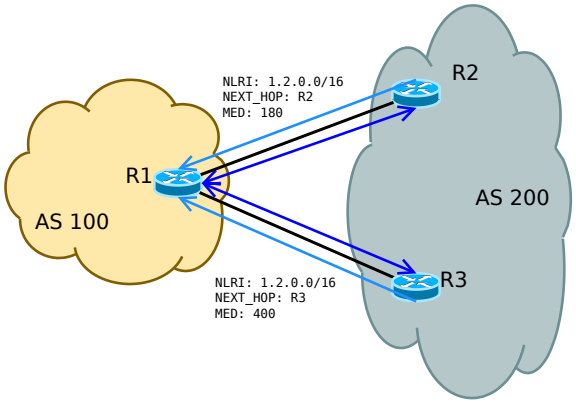


Manipulace s atributy - AS_PATH



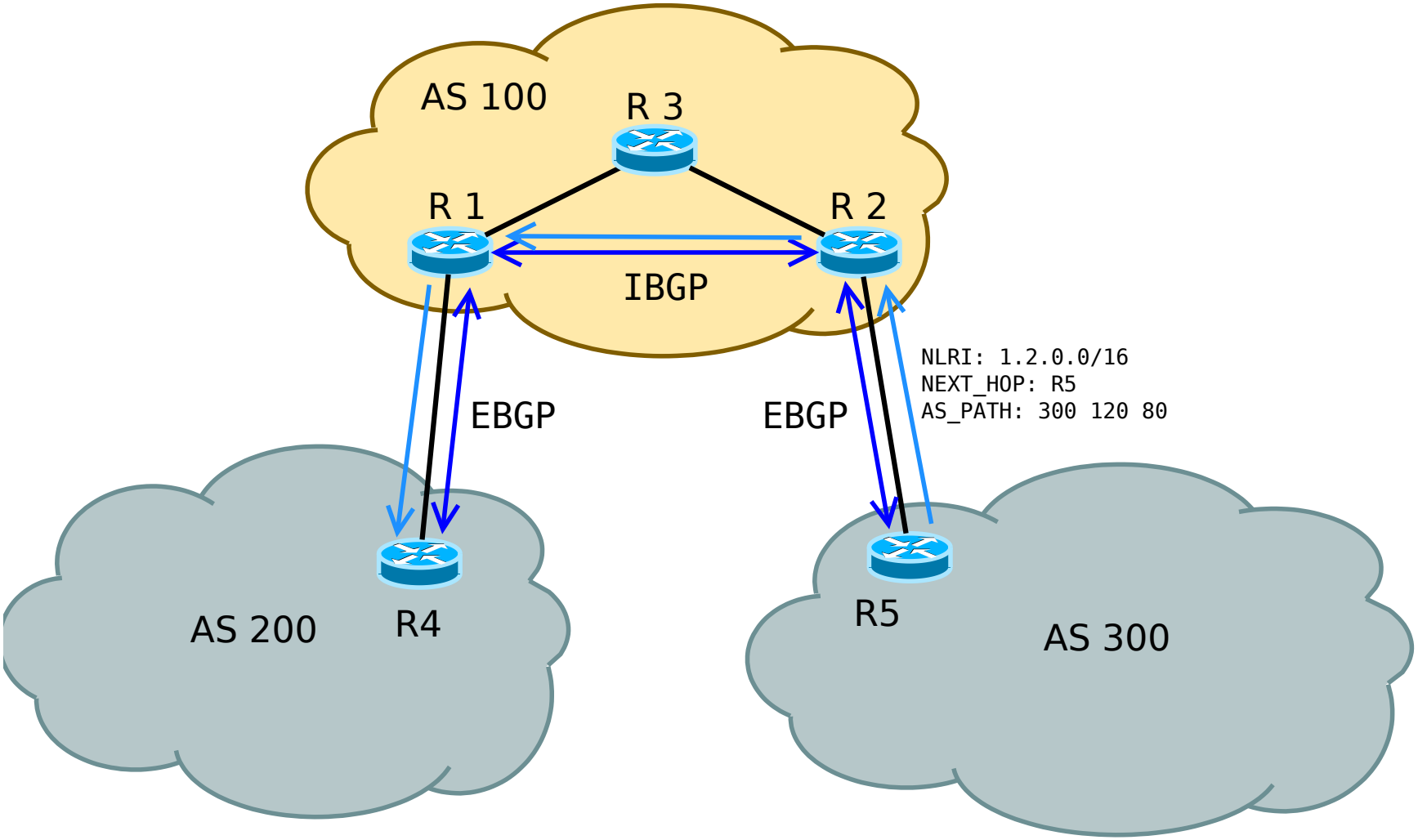
Ovlivnit příchozí provoz z jiných AS je možné pomocí ovlivňování délky AS_PATH. Z AS_PATH není možné prvky rušit, je možné tam pouze přidávat další čísla AS (obvykle několikrát za sebou číslo svého vlastního AS). Čím delší AS_PATH, tím méně preferovaná cesta (viz pravidlo č.5 při výběru cesty)

Manipulace s atributy - MED



Multi Exit Discriminator (MED) se používá výhradně mezi dvěma AS k výběru lepší cesty v případě dvou a více linek mezi AS. Menší vyhrává.

iBGP - interní BGP



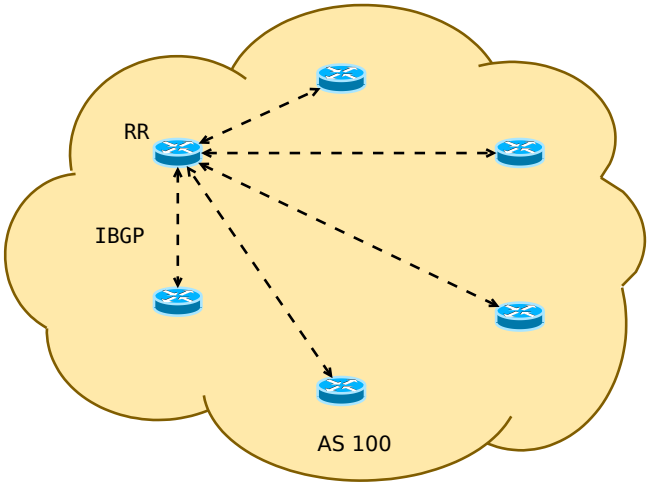
iBGP

- číslo AS se do atributu AS_PATH přidává pouze při předávání do jiného AS. Tj:
- nefunguje detekce smyček
- ochrana: iBGP směrovač nepředává jinému iBGP směrovači cesty, které dostal od jiného iBGP souseda
- problém: nedostane všechny cesty
- řešení: iBGP musí být full-mesh
- spojení mezi iBGP sousedy se obvykle navazuje na loopbacková rozhraní. Z toho plyne:
- je obvykle nutné mít spuštěný nějaký IGP protokol, aby mělo BGP jak navázat spojení.

Pravidla Pro iBGP

- cesty do iBGP souseda se nepředávají jinému iBGP sousedovi
- nemění se atribut cesty AS_PATH
- nemění se atribut cesty NEXT_HOP

Route Reflektory (RR)



Full-mesh iBGP je konfiguračně náročné - každé přidání nového iBGP směrovače znamená zásah do všech ostatních iBGP směrovačů.
Route reflektory tedy porušují základní pravidlo iBGP a předávají iBGP sousedům i cesty, které dostaly od jiného iBGP souseda.
Aby se předešlo nebezpečí single point of failure, je možné route reflektory sdružovat do skupin. Tyto směrovače pak sdílejí společný atribut `Cluster_ID`

Konfederace

Konfederace umožnují rozdělit jeden AS tak, aby vnitřně vypadal jako více autonomních systémů. Navenek vypadá jako jeden AS. Takové rozdělení umožňuje provádět filtrování a další manipulace cestami stejně jako by to byly nezávislé AS.

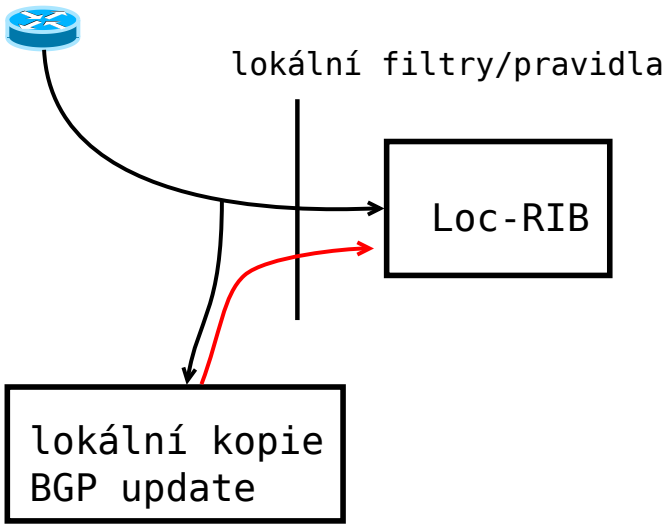
Route refresh

Route refresh je rozšíření protokolu BGP (RFC 2918). Umožňuje BGP směrovači říct BGP sousedovi, že má znovu poslat celou směrovací tabulku, jako by se jednalo o zahájení relace. Znovupřenesení všech cest je nutné např. při změně směrovací politiky.

Výhody:

- nedojde k resetu BGP spojení a nutnosti ho znovu navazovat
- nedojde ke zrušení cesty ze směrovacích tabulek (zrušení může být v BGP penalizováno)

Soft rekonfigurace



V případě, že soused neumí Route refresh, může BGP směrovač použít techniku **soft rekonfigurace**. V takovém případě si BGP směrovač uloží všechny cesty, které od BGP souseda dostal do samostatné tabulky.
V případě, kdy chce změnit směrovací politiku nevyžádá si route refresh od souseda, ale využije tuto tabulku.

\+ rychlost a chování

\- paměťové nároky

Outbound route filtering

V BGP relacích je běžné, že přijímající BGP směrovač některé cesty během zpracování prostřednictvím filtrů zahodí. Taková zahozená cesta pak nemusí být vůbec přenášena. Aby bylo možné tohoto dosáhnout, bylo do BGP integrováno rozšíření RFC 5291 **Outbound Filtering Capability**. Toto rozšíření umožňuje přijímajícímu BGP směrovači poslat vysílajícímu BGP směrovači sadu filtrů, kterými jsou některé cesty zahozeny ještě před vysíláním.

```
#sh bgp ipv4 unicast neighbors 10.0.0.1
BGP neighbor is 10.0.0.1,  remote AS 65080, internal link
 Inherits from template CORE-SESSION for session parameters
 BGP version 4, remote router ID 10.0.0.1
 Neighbor capabilities:
    Route refresh: advertised and received(new)
    Four-octets ASN Capability: advertised and received
    Address family IPv4 Unicast: advertised and received
    ipv4 MPLS Label capability: advertised and received
```

Rozšíření o IPv6

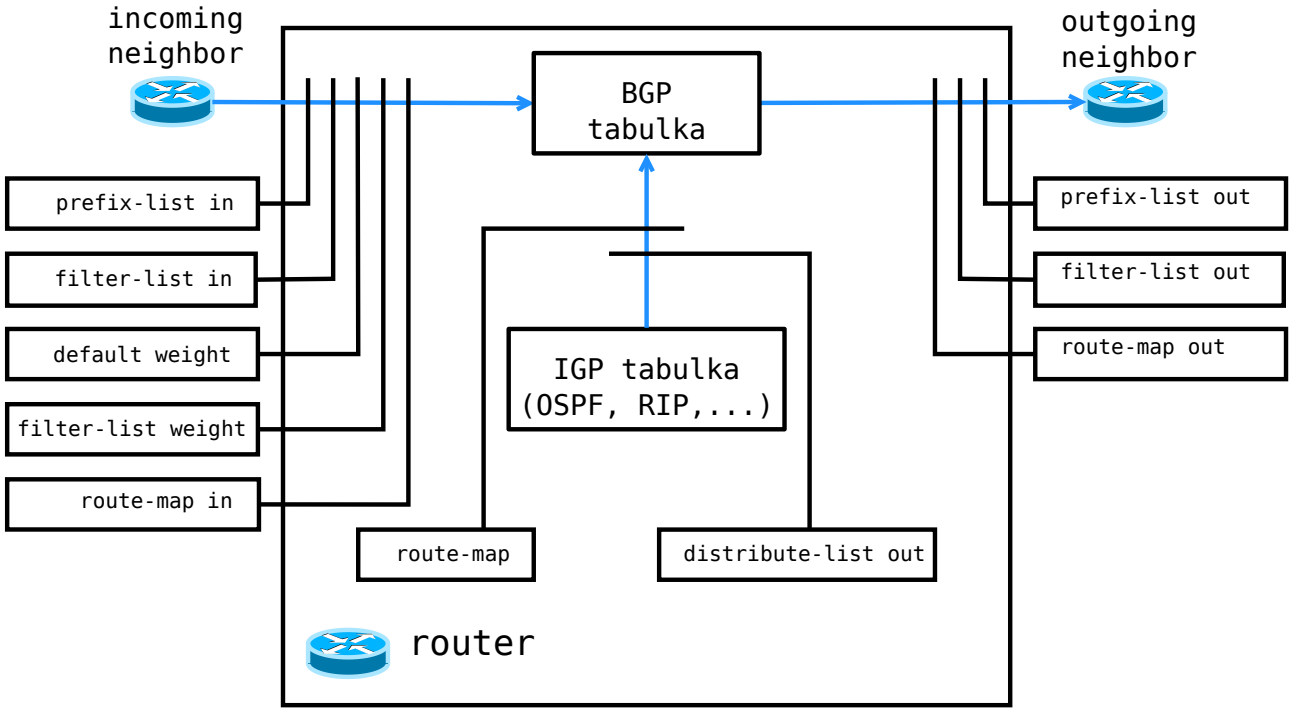
BGP umožňuje díky své rozšiřitelnosti přenášet směrovací informace i jiných protokolů. V dnešní době hlavně IPv6 (RFC 2545)

- link-local adresy používané v jiných IPv6 směrovacích protokolech nejsou pro BGP příliš vhodné (NEXT_HOP se v IPv4 při předávání v rámci IBGP nemění)
- změna NLRI informace - atribut NEXT_HOP se přesunul do NLRI, může se skládat ze dvou IPv6 adres - jedné globální a jedné link-local adresy (pouze v případě, kdy entita identifikovaná globální IPv6 adresou sdílí společnou L3 síť s BGP speakerem)
- to obvykle znamená, že spojovací sítě mezi BGP sousedy musejí být adresovány globálními IPv6 adresami (to např. u OSPFv3 pro IPv6 neplatí a není vhodné)
- BGP je nezávislé na přenosovém protokolu, může být přenášeno po IPv4, to komplikuje určení NEXT_HOPu
- doporučuje se mít pro IPv4 a IPv6 dvě různé BGP relace - jednu pro IPv4 komunikující po IPv4 a jednu pro IPv6 komunikující po IPv6 (navíc lepší pro jednoduchou detekci fungování IPv6 - neustaví se BGP relace)

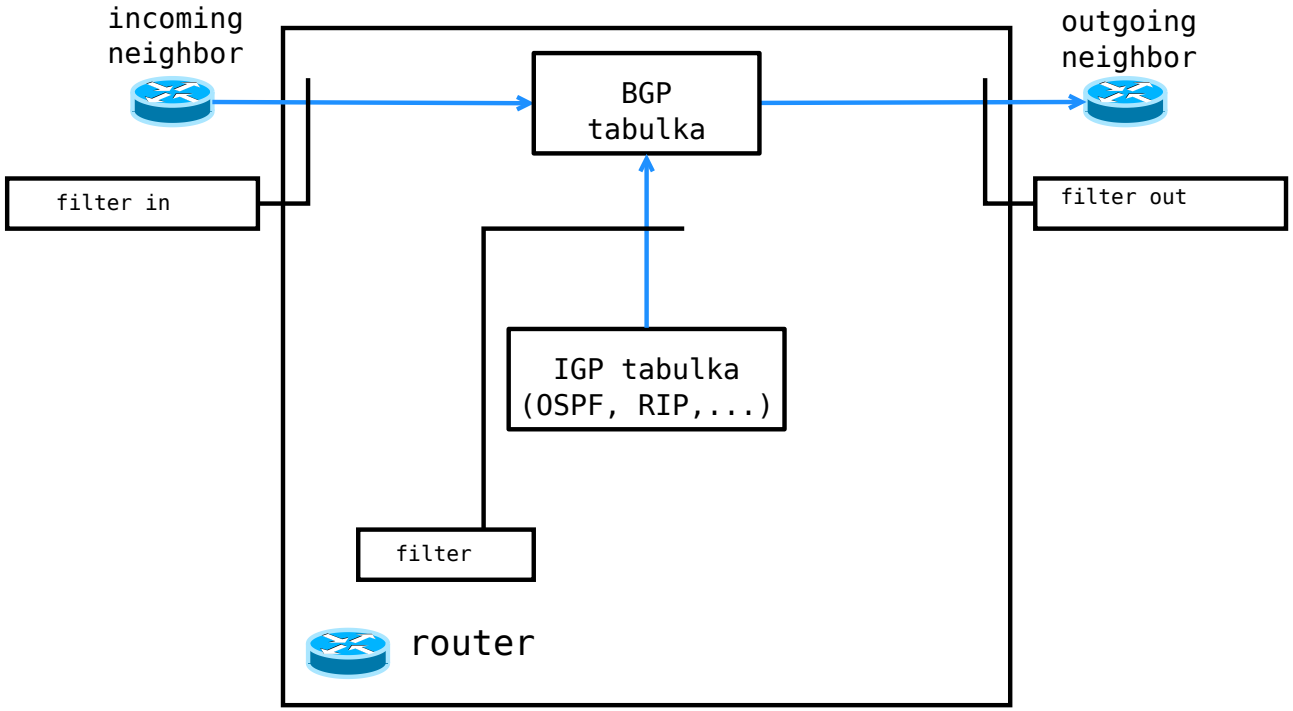
Bezpečnost

- MD5 autentizace
- nikdy nepřijímat od externích BGP sousedů cesty na svoje sítě (EBGP cesty mají lepší administrativní metriku než IGP)
- omezení maximálního počtu přijatých cest
- možnost omezit navázání EBGP spojení pouze pro přímého souseda kontrolou a nastavením IP TTL

BGP - Cisco


incoming
neighbor
outgoing
neighbor
BGP
tabulka
prefix-list in
filter-list in
default weight
filter-list weight
route-map in
IGP tabulka
(OSPF, RIP,...)
route-map
distribute-list out
prefix-list out
filter-list out
route-map out
router

BGP - Mikrotik

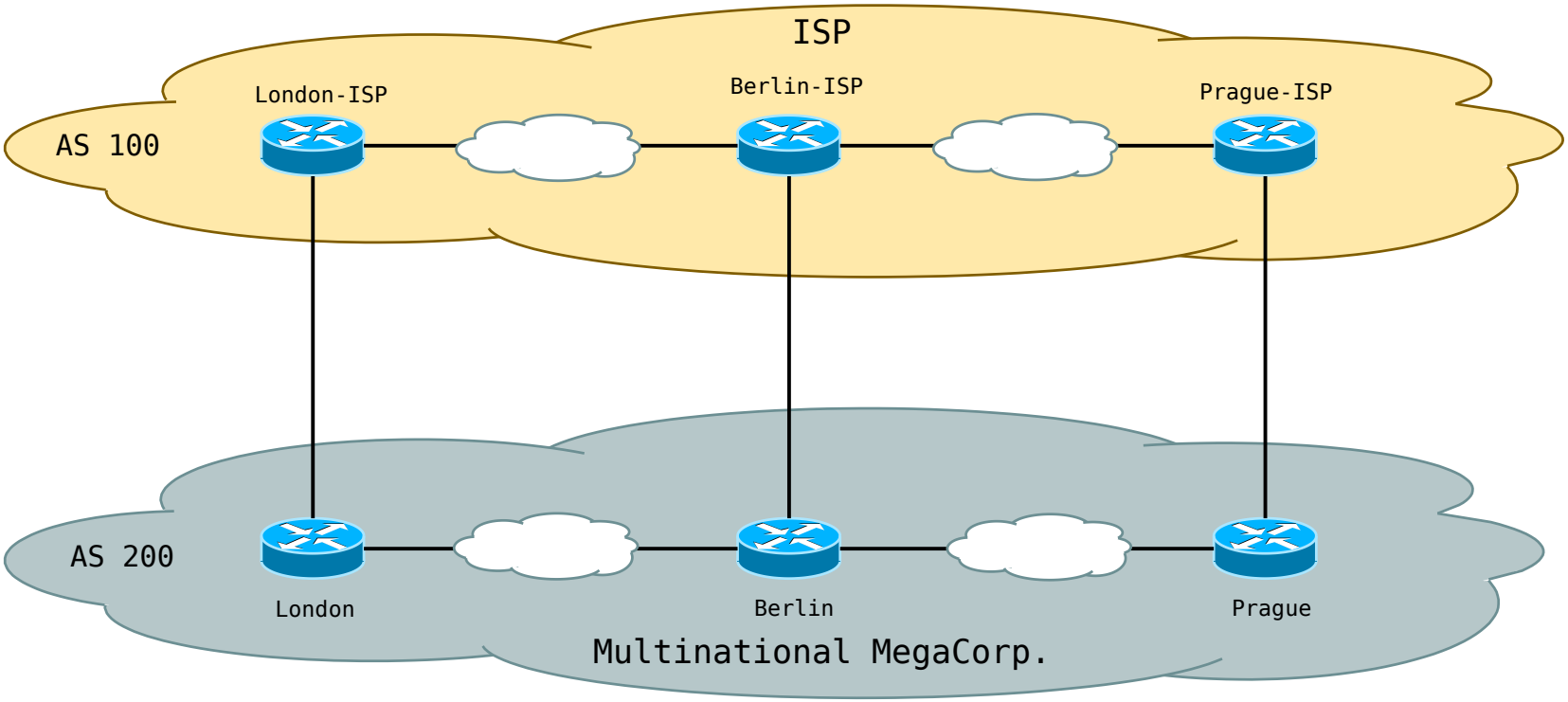


Mikrotik - filtry směrovacích informací

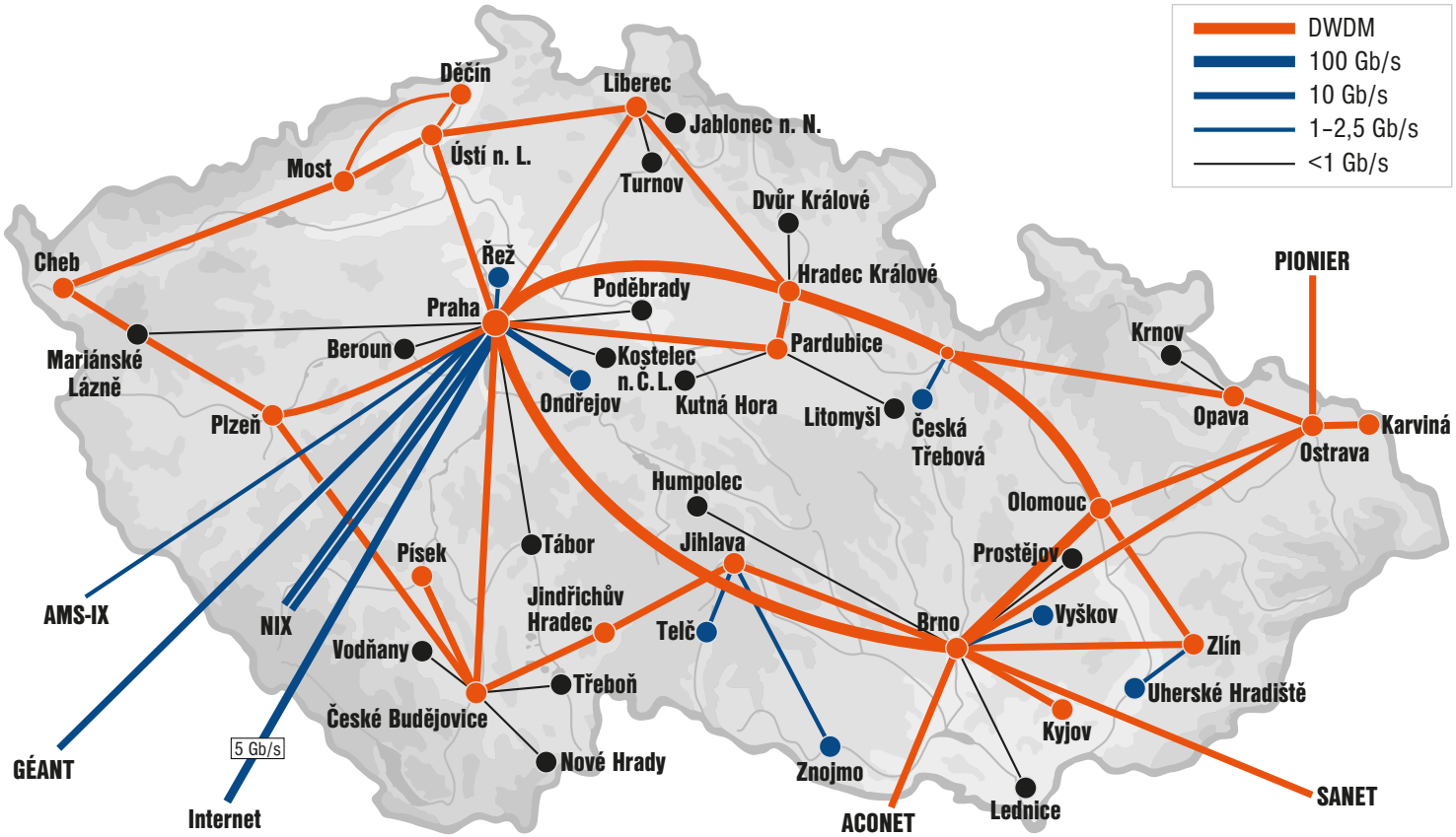
```
/routing bgp instance
set default as=30

/routing bgp peer
add name=toISP1 remote-address=192.168.1.1 remote-as=10
add name=toISP2 remote-address=192.168.2.1 remote-as=20

/routing bgp peer
set isp1 in-filter=isp1-in out-filter=isp1-out
set isp2 in-filter=isp2-in out-filter=isp2-out

/routing filter
add chain=isp2-out prefix=10.1.1.0/24 action=accept set-bgp-prepend=3
add chain=isp2-out prefix=10.1.2.0/24 action=accept set-bgp-prepend=3
add chain=isp2-out action=discard
```

http://wiki.mikrotik.com/wiki/Manual:Routing/Routing_filters

Příklad - nadnárodní společnost s nadnárodním ISP

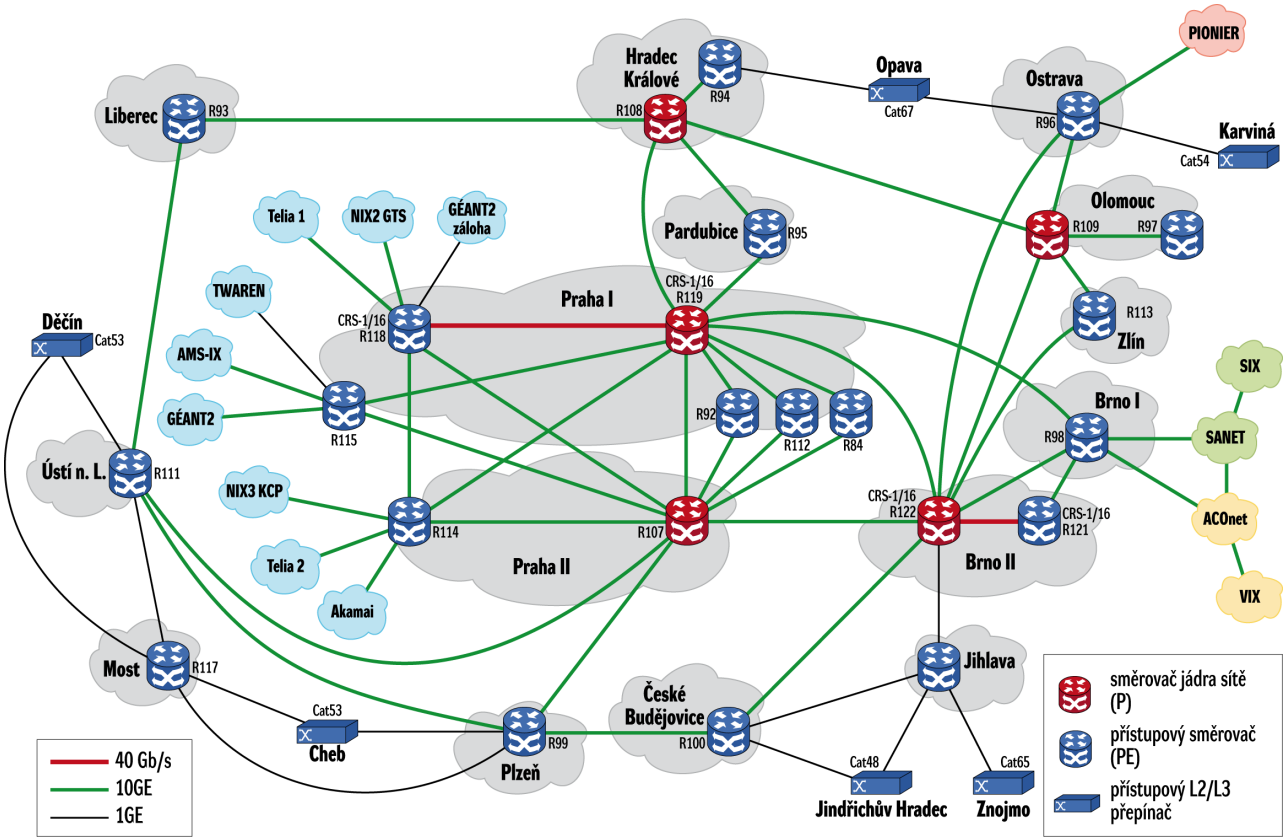


Příklad - CESNET 2, fyzická topologie



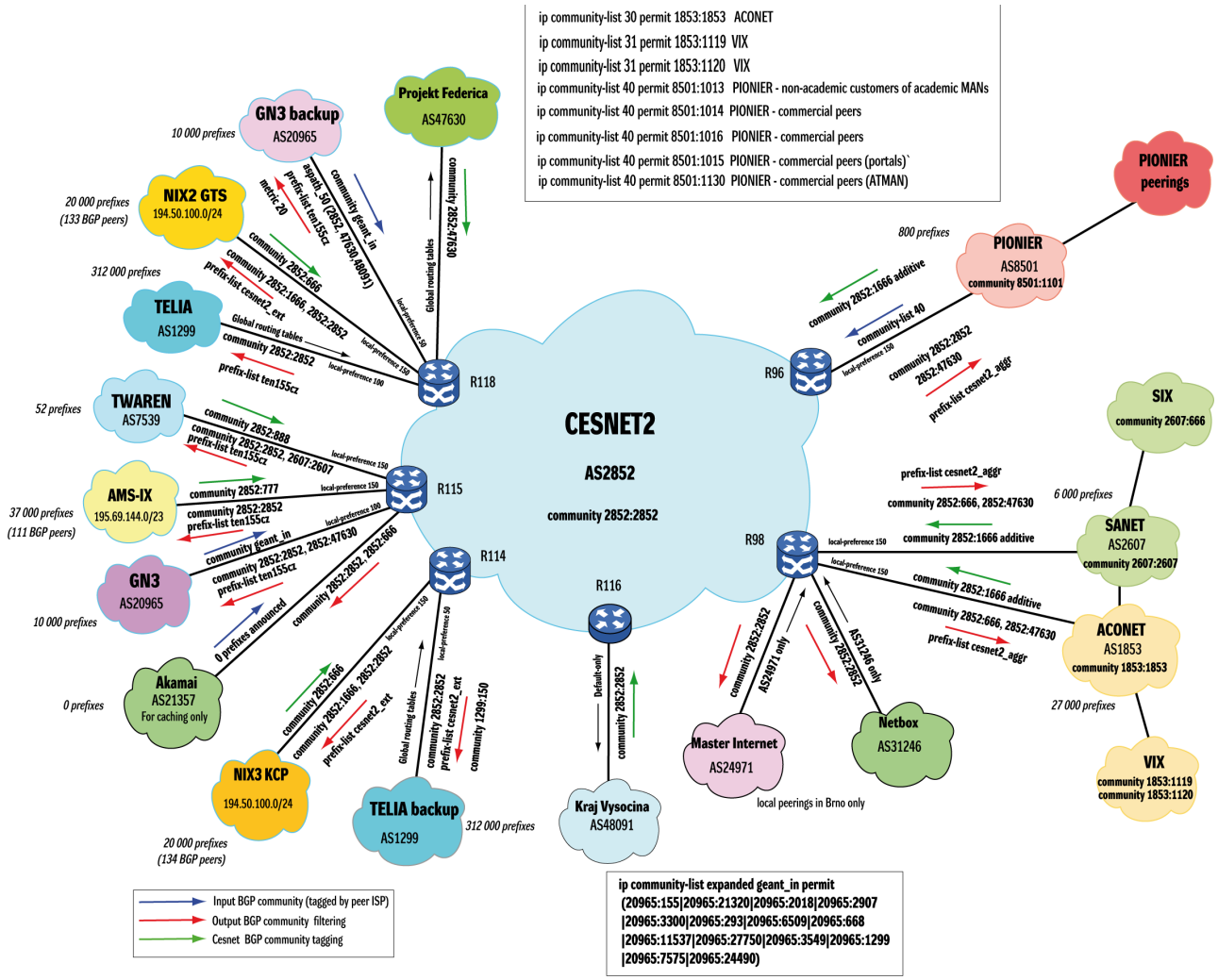
http://www.cesnet.cz/sluzby/pripojeni/topologie/

Příklad - CESNET 2, zapojení



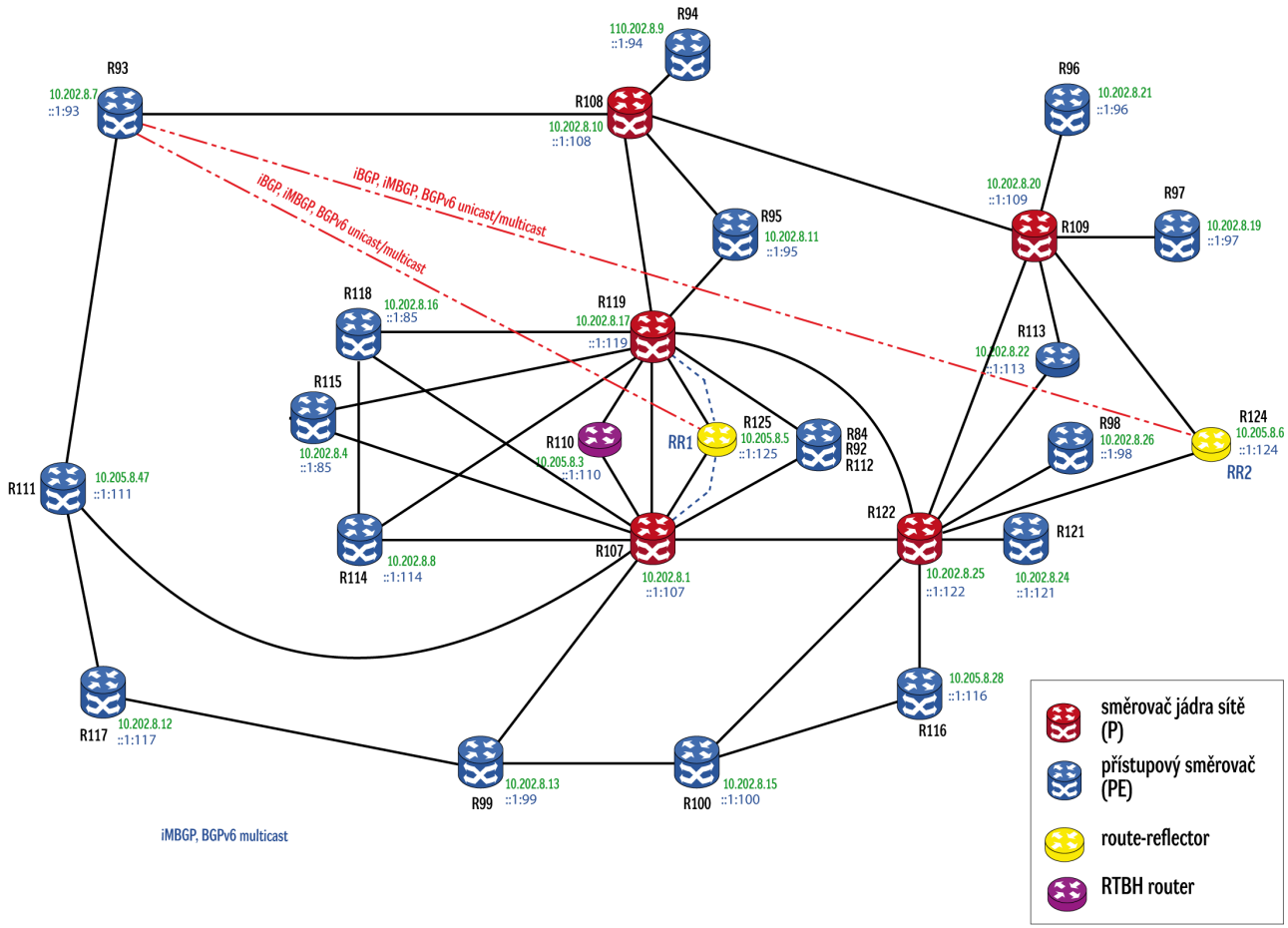
http://archiv.cesnet.cz/doc/techzpravy/2010/bgp-design-optimization/

Příklad - CESNET 2, BGP peering



http://archiv.cesnet.cz/doc/techzpravy/2010/bgp-design-optimization/

Příklad - CESNET 2, vnitřní struktura



Odkazy

- Looking glass `http://www.bgp4.as/looking-glasses`
- RIPE stat `https://stat.ripe.net/`

Literatura

- Sam Halabi, Danny McPherson: Internet Routing Architectures
- Doyle, Carroll: Routing TCP/IP, Volume II
- Zhang, Bartell: BGP Design and Implementation
- RFC 4271, 2545, 2918, ...