Wprowadzenie do protokołu BGP (Border Gateway Protocol)

1. Wstęp

Podział protokołów routingu

- Kryterium I (zasada działania)
  - Wektor-odległość: RIP, IGRP
  - Stanu łącza: OSPF, IS-IS
  - Hybrydowe: EIGRP
- Kryterium II (obszar stosowania)
  - W ramach jednego systemu autonomicznego, klasa IGP: RIP, IGRP, EIGRP, OSPF, IS-IS
  - Pomiędzy systemami autonomicznymi, klasa EGP: EGP (Exterior Gateway Protocol), BGP

Komunikacja w protokołach IGP

- RIP, IGRP: regularne przesyłanie informacji o znanych sobie sieciach
- OSPF, EIGRP:
  - Regularnie wysyłane *hello* informujące o aktywności sąsiada
  - Przesyłanie informacji o zmianach tylko po ich wystąpieniu (i w czasie inicjacji)

Komunikacja w protokołach IGP

- RIP v1
  - Adres: 255.255.255.255
  - Enkapsulacja: UDP/IP (port 520)
- RIP v2
  - Adres: 224.0.0.9
  - Enkapsulacja: UDP/IP (port 520)
- OSPF
  - Adres: 224.0.0.5 i 224.0.0.6
  - Enkapsulacja: IP
- IGRP
  - Adres: 255.255.255.255
  - Enkapsulacja: IP

System autonomiczny

- Obszar sieci będący pod wspólną władzą administracyjną i zarządzany w jednolity sposób
- Identyfikowany przez numer: liczbę 16-to bitową
  - 1-64511 – numery publiczne
    - przyznawane przez RIR (Regional Internet Registries)
      - 8323 – CYFRONET-AS2 Metropolitan Area Network
      - 8267 – CYFRONET-AS Metropolitan Area Network
      - 8364 – poznański MAN
    - muszą jednoznacznie identyfikować właściciela
  - 64512-65534 – numery prywatne
    - mogą być wykorzystywane wielokrotnie
- W 2007 wprowadzono numery AS o długości 32 bitów które są kompatybilne ze starszym formatem.

Ważne spostrzeżenie

- *W obszarze zastosowania protokołów EGP (*Exterior Gateway/Routing Protocols*) nie ma jednej władzy administracyjnej*

Rozgłaszanie prefiksów

- System autonomiczny może rozgłaszać określoną liczbę prefiksów (zgrupowanych adresów IP)
- Liczba prefiksów w sieci (tablicy routingu) cały czas rośnie. Jest to związane ze zwiększającą się liczbą wykorzystywanych adresów oraz potrzebą podziału bloków adresów na coraz mniejsze zakresy (podsieci)
  - rok 2011 – 350000 prefiksów
  - rok 2013 – 450000 prefiksów

Prefiksy - ACK Cyfronet

- CYFRONET-AS Metropolitan Area Network
  - 149.156.0.0/16
  - 192.86.14.0/24
  - 192.245.169.0/24
  - 195.150.224.0/19
- CYFRONET-AS2 Metropolitan Area Network
  - 193.193.64.0/19
  - 194.8.45.0/24
  - 194.8.46.0/24
  - 195.150.0.0/16

http://bgp.potaroo.net/cgi-bin/as-report?as=AS8323

Połączenia AS 8267

8501 – PIONIER
6778 – Exatel SA
5550 – TASK, gdański MAN
8508 – SILWEB, Śląsk
12631 – Formus Polska
12990 – Onet.pl
15541 – Ceti

import: from AS8501 action pref=100; accept ANY
import: from AS8323 action pref=100; accept ANY
import: from AS6778 action pref=200; accept AS6778
import: from AS5550 action pref=200; accept AS5550
import: from AS8508 action pref=250; accept AS8508
import: from AS12631 action pref=250; accept AS12631
export: to AS8323 announce ANY
export: to AS8501 announce AS8267 AS8323 AS12990 AS13255 AS15967 AS16290 AS16138 AS15541
export: to AS6778 announce AS8267 AS8323 AS12990 AS13255 AS15967 AS16290 AS16138
export: to AS5550 announce AS8267 AS8323 AS12990 AS13255 AS15967 AS16290 AS16138 AS15541
export: to AS8508 announce AS8267 AS8323 AS12990 AS13255 AS15967 AS16290 AS16138 AS15541
export: to AS12631 announce AS8267 AS8323 AS12990 AS13255 AS15967 AS16290 AS16138 AS15541

Niektóre polskie AS

- AS8938 ENERGIS-IP Energis Polska IP Network
- AS12827 WIRTUALNAPOLSKA Wirtualna Polska S.A.
- AS20553 HPPOLAND-AS Hewlett Packard Polska (HPO)
- AS20778 PKOBP Powszechna Kasa Oszczędności Bank Polski S.A.
- AS25439 PWPW-AS Polska Wytwórnia Papierów Wartościowych S.A
- AS25506 TVP-AS Telewizja Polska S.A.
- AS30759 ALCATEL-POLSKA-AS Alcatel Polska S.A
- AS33900 TPSA-MPLSNET-AS Telekomunikacja Polska MPLSNET
- AS1887 NASK-ACADEMIC NASK Research and Academic Network in Poland (tylko uczelnie)

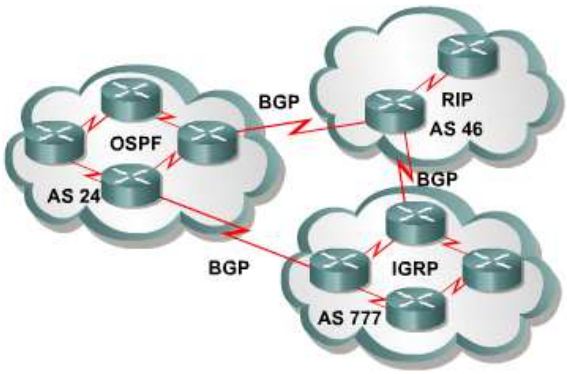
2. Informacje ogólne o BGP

Informacje ogólne

- Protokół routingu zewnętrznego: routing pomiędzy systemami autonomicznymi
- Budowany w oparciu o doświadczenie zdobyte na protokole EGP
- *de facto* standard w komunikacji pomiędzy routerami w Internecie
- tzw. policy routing; wymaga dużej ingerencji (i wiedzy) administratora
- Nie ma tradycyjnej metryki – wykorzystuje atrybuty i algorytm wyboru

Obszar zastosowania BGP

- Przede wszystkim pomiędzy systemami autonomicznymi (eBGP)
- Czasem też wewnątrz systemu autonomicznego (iBGP)
  - najczęściej w sieci ISP



Grafika: Cisco Systems

Rodzina protokołów BGP

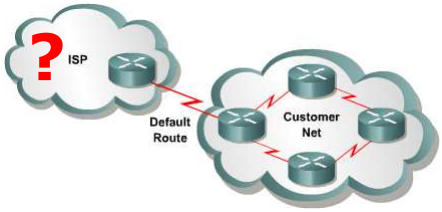
- BGP-1, RFC 1105, 1989
  - zastąpienie EGP
- BGP-2, RFC 1163, 1990
- BGP-3, RFC 1267, 1991
- BGP-4, RFC 1771, 1995
  - Routing bezklasowy
- Zmiany w BGP-4: RFC 4271, 2006

Działanie routera IP

- Informacje o dostępności sieci zgromadzone w tablicy routingu
- Sekwencja wpisów postaci:
  - Adres IP
  - Maska adresu
  - Adres routera następnego skoku/nazwa interfesju
- Algorytm wyboru: dopasowanie najdłuższej maski; im bardziej specyficzna trasa, tym lepiej
- Wpis domyślny
  - Zmniejszenie rozmiaru tablicy
  - Delegowanie odpowiedzialności za przekazanie pakietu dalej

Routing pomiędzy ISP

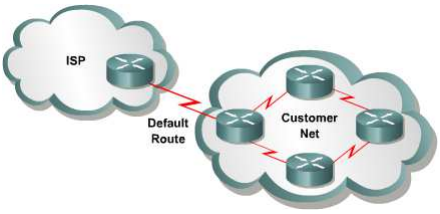
- Prawie zawsze jest wymagana komunikacja obustronna: sieć musi wiedzieć jak przesyłać do nas dane
- Ważna jest symetria
- Ważna jest swoboda wyboru operatora



Grafika: Cisco Systems

Rodzaje systemów autonomicznych (1)

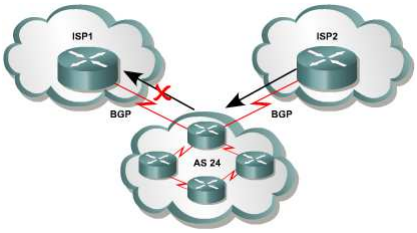
- *Single-homed, stub network*
  - Jedna trasa domyślna na zewnątrz
  - Możliwa konfiguracja
    - Trasa domyślna
    - Routing dynamiczny IGP
    - Routing dynamiczny BGP – po co?



Grafika: Cisco Systems

Rodzaje systemów autonomicznych (2)

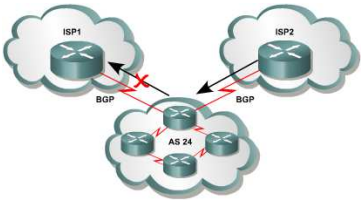
- *Multihomed nontransit*
  - Więcej niż jedno wyjście na świat
  - Dołączone do jednego lub więcej dostawców
  - Nie zezwala się na ruch tranzytowy
    - Nie ogłasza się sieci otrzymanych na drugim wyjściu
  - BGP nie jest wymagany, choć jest polecany



Grafika: Cisco Systems

Rodzaje systemów autonomicznych (2)

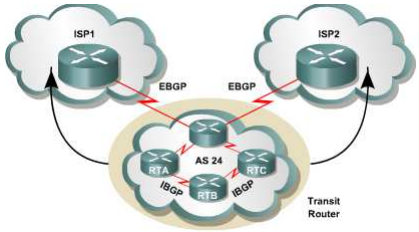
- *Multihomed nontransit*
  - Dołączone do jednego dostawcy
    - Ważna symetria
    - Równoważenie obciążenia lub łącze zapasowe
    - Duże wsparcie BGP
      - Kontrola ruchu wychodzącego
      - Kontrola ruchu przychodzącego
  - Dołączone do wielu dostawców
    - Dywersyfikacja ruchu ☺
    - Wymagania jak wyżej
  - Problem przestrzeni adresowej



Grafika: Cisco Systems

Rodzaje systemów autonomicznych (3)

- *Multihomed transit*
  - Więcej niż jedno wyjście na świat
  - Dołączone do jednego lub więcej dostawców
  - Zezwala się na ruch tranzytowy
  - BGP jest uruchomiony
    - na styku sieci (EBGP)
    - wewnątrz sieci (IBGP)



Grafika: Cisco Systems

Podstawowa terminologia BGP

- Sąsiad, *peer*
  - W tym samym AS:
    - *internal peer*
    - BGP: *internal* BGP, iBGP
  - W innym AS:
    - *external peer*
    - BGP: *external* BGP, eBGP
  - Niekoniecznie bezpośrednio przyległy router (!)
- Prefiks
  - Adres z maską, np. 149.156.97.0/24

Komunikat BGP

- Marker – pole w którym mogą być zawarte dane potrzebne do autentykacji. W przypadku braku mechanizmu autentykacji lub komunikatu typu OPEN są tam umieszczone same jedynki (0xFFFF)
- Length – całkowita długość komunikatu BGP razem z nagłówkiem (19- 4096 bajtów)
- Type – określa rodzaj komunikatu
- Data – dane BGP

Marker	Length	Type	DATA
16 bajtów	*2 bajty*	*1 bajt*	*zmienna*

Komunikaty BGP

- open: używany przy nawiązywaniu połączenia BGP
- update: wysyłany w razie wystąpienia zmian
- notification: informuje o błędach
- keepalive: podtrzymuje komunikację między sąsiadami
- route refresh: służy do przesłania żądania wysłania grupy adresów z bazy BGP bez konieczności zamykania i otwierania sesji (np. po zmianie konfiguracji na routerze)

Komunikat OPEN

- Zestawiane jest połączenie warstwy transportowej
  - Połączenie punkt-punkt (!)
  - TCP/179
  - Użycie niezawodnego protokołu eliminuje
    - konieczność implementacji retransmisji
    - konieczność numerowania pakietów
    - konieczność generowania potwierdzeń, itp.
- Przesyłane są m.in. numer AS, numer wersji BGP
- Ustalany jest czas pomiędzy komunikatami KEEPALIVE

Komunikat OPEN

0xFFFF	Length	Type=1	DATA

Version	AS	Hold Time	BGP Identifier	Parameter Length	Optional Parameters
4					TLV
1 bajt	*2 bajty*	*2 bajty*	*4 bajty*	*1 bajt*	*zmienna*

- Hold Time – 0 lub minimum 3 sekundy (keepalive nie częściej niż co 1 s). Zero oznacza, że nie wykorzystywane jest keepalive
- BGP Identifier – id routera wysyłającego
- Parameter Length – całkowita długość pola *Optional Parameters*
- Optional Parameters – np. parametry dotyczące autentykacji sesji, obsługa Route Refresh itp.

Komunikat UPDATE

- Pozwala na przesłanie informacji o trasach pomiędzy sąsiadami BGP
- Wysyłany po zestawieniu połączenia i w razie zaistnienia zmian
- Przesyła informacje
  - Identyfikującą opisywaną trasę: prefiks
  - Identyfikujące usuwane (*withdrawn*) trasy
  - Opisujące ścieżkę: jej atrybuty

Komunikat UPDATE

Unfeasible Routes Length	Withdrawn Routes	Total Path Attribute Length	Path Attributes	Network Layer Reachability Information
2 bajty	*zmienna*	*2 bajty*	*zmienna*	*zmienna*

Komunikat UPDATE

Path Attributes							
Attribute Type						Attribute Length	Attribute Value
Attrbute Flags					Attribute Code		
O	T	P	E	0000			
1 bajt					*1 bajt*	*1-2 bajtów*	*zmienna*

- O - bit Optional, wskazuje czy atrybut jest typu optional (1)
- T - bit Transitive, wskazuje czy atrybut jest typu transitive (1)
- P - bit Partial/Complete wskazuje czy wszystkie routery na ścieżce obsługują parametr (1/0)
- E - bit Extended Length informuje czy Attribute Length zajmuje 1 czy 2 bajty (0 oznacza 1 bajt)
- …

Komunikat NOTIFICATION

- Informuje o błędach
- Wysyłany w razie wystąpienia dowolnego błędu (też po przekroczeniu czasu *hold time*)
- Powoduje zamknięcie połączenia BGP

Marker	Length	Type	Error Code	Error Subcode	DATA
		3			
16 bajtów	*2 bajty*	*1 bajt*	*1 bajt*	*1 bajt*	*zmienna*

Komunikat KEEPALIVE

- Podtrzymuje aktywność na łączu – monitorowanie istnienia komunikacji
- W stabilnie działającej sieci to jedyne komunikaty przesyłane pomiędzy sąsiadami
- Sugerowane ustawienia
  - *Keepalive interval*: 30 sek.
  - *Hold time interval*: 90 sek.
- Wielkości domyślne na routerach Cisco
  - *Keepalive interval*: 60 sek.
  - *Hold time interval*: 180 sek.

Komunikat KEEPALIVE

- Komunikat KEEPALIVE zawiera tylko nagłówek BGP

Marker	Length	Type
		4
16 bajtów	*2 bajty*	*1 bajt*

Komunikat ROUTE REFRESH

Marker	Length	Type	AFI	Reserved	SAFI
		5			
16 bajtów	*2 bajty*	*1 bajt*	*2 bajty*	*1 bajt*	*1 bajt*

- AFI (*Address Family Identifier)* - IPv4 lub IPv6 (np. wyślij wszystkie adresy IPv4)
- SAFI (*Subsequent Address Family Identifier)* - unicast lub multicast

3. Atrybuty ścieżki BGP
4. Proces decyzyjny BGP

Atrybuty ścieżki

- W protokołach IGP: metryka i jej składowe
- BGP: atrybuty
  - Charakteryzują ogłaszaną ścieżkę
  - Sekwencja wiadomości postaci <typ, długość, wartość> (TLV)
  - Cztery rozłączne klasy
- Atrybuty BGP są przenoszone w komunikatach aktualizacyjnych (UPDATE) protokołu BGP.

Klasy atrybutów ścieżki

- *well-known (standardowe)*: obsługa jest wymagana przez każdą implementację BGP
  - *mandatory (obowiązkowe)*: musi być zawarty w każdej wiadomości *Update*
  - *discretionary (uznaniowe)*: nie musi być zawarty w każdej wiadomości *Update*
- *optional (opcjonalne)*: obsługa nie jest wymagana przez implementację BGP
  - *transitive (przenośne)*: komunikat *Update* powinien być przekazany dalej, nawet gdy parametr nie jest zrozumiany
  - *nontransitive (nieprzenośne)*: komunikat *Update* nie jest przekazywany dalej gdy parametr nie jest zrozumiany

Atrybut ORIGIN (pochodzenie)

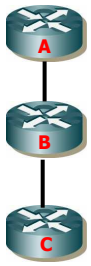
- Klasa *well-known mandatory*
- Opisuje pochodzenie wiadomości
  - IGP: z protokołu wewnętrznego
  - EGP: z protokołu zewnętrznego EGP (RFC 904)
  - *Incomplete*, nieokreślone lub nieznane, np. z redystrybucji
- Kolejność uwzględniania w metryce: IGP, EGP, *incomplete*
- Atrybut ustawiany przez AS, który wygenerował informację o sieci, nie modyfikowany przez inne ASy

Atrybut AS_PATH (ścieżka)

- Klasa *well-known mandatory*
- Zawiera sekwencję numerów AS, przez które wiedzie trasa
- Proces dodawania kolejnych numerów (na początek listy): *AS prepending*
- Sesje iBGP nie zwiększają tej listy
- Możliwe kilkukrotne dodanie tego samego numeru AS
- Pozwala na wykrycie pętli

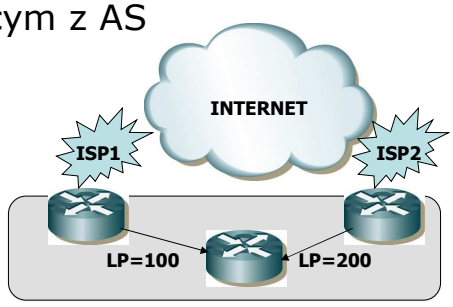
Atrybut NEXT_HOP (następny skok)

- Klasa *well-known mandatory*
- Adres routera następnego skoku
- Nie zawsze adres sąsiedniego routera
  - Gdy routery otrzymujący i wysyłający są w różnych AS, adres ogłaszającego routera
  - Gdy routery otrzymujący i wysyłający są w tym samym AS:
    - prefiks jest z tego samego AS: adres routera ogłaszającego
    - prefiks jest spoza AS: adres routera z poprzedniego AS
  - Niekiedy wymagany jest tzw. *recursive lookup*



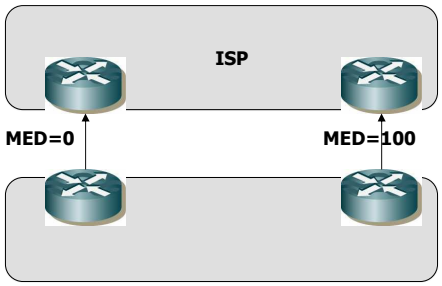
Atrybut LOCAL_PREF (lokalna preferencja)

- Klasa *well-known discretionary*
- Ma sens jedynie w komunikacji pomiędzy routerami iBGP, nie jest przekazywany do innych AS
- Używany do preferowania któregoś z routerów przy ruchu wychodzącym z AS
- Wyższe wartości są preferowane
- 100 jest wartością domyślną



Atrybut MULTI_EXIT_DISC

- Klasa *optional nontransitive*
- Używany do poinformowania sąsiedniego AS o preferowanym punkcie wejścia do naszej sieci
- Ma sens jedynie w komunikacji do jednego AS, nie jest przekazywany do innych AS
- Niższe wartości są preferowane



Proces decyzyjny BGP (skrót)

1. Najwyższa wartość LOCAL_PREF
2. Preferuj trasy uzyskane lokalnie (na tym routerze)
3. Trasy o najmniejszej długości listy AS_PATH
4. Trasy o najmniejszej wartości *origin code* IGP<EGP<*incomplete*
5. Najmniejsza wartość MULTI_EXIT_DISC
6. Najmniejsza wartość metryki do routera NEXT_HOP

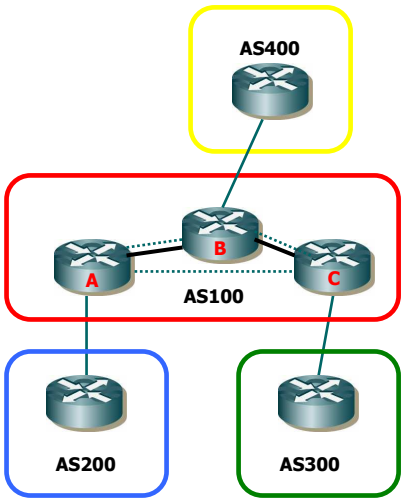
Proces decyzyjny BGP

1. Usuń trasy z niedostępnym adresem NEXT_HOP
2. Najwyższa waga WEIGHT (lokalny parametr BGP, 1-65535, określony tylko dla urządzeń Cisco, nie przesyłany do innych routerów)
3. Najwyższa wartość LOCAL_PREF
4. Preferuj trasy uzyskane lokalnie (na tym routerze)
5. Trasy o mniejszej długości listy AS_PATH
6. Trasy o mniejszej wartości *origin code* IGP<EGP<*incomplete*
7. Najmniejsza wartość MULTI_EXIT_DISC
8. Preferuj trasy uzyskane z eBGP nad iBGP
9. Najmniejsza wartość metryki do routera NEXT_HOP

Proces decyzyjny BGP

9. W przypadku włączonego BGP multipath zapisz trasę w tablicy routingu
10. Preferuj starsze trasy (otrzymane wcześniej)
11. Preferuj trasy, które mają niższy router-id
12. Preferuj trasę, która została wysłana od sąsiada (neighbor) z niższym adresem IP

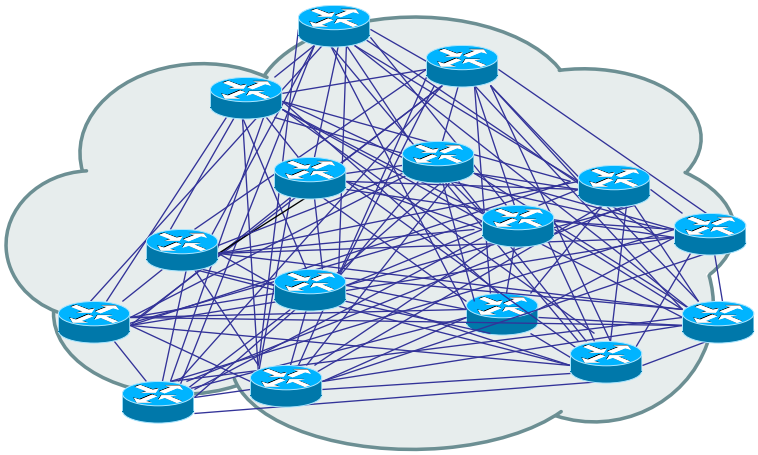
5. Protokół iBGP



Protokół iBGP

- iBGP używany praktycznie tylko w sieciach *multihomed*: właściwe rozgłaszanie informacji w przypadku wielu dostawców
- Nie jest zmieniana wartość AS_PATH (nie ma ochrony przez pętlami)
- iBGP nie ogłasza tras otrzymanych od innych routerów iBGP (pętle!)
- Połączenie routerów musi być typu każdy z każdym (*fully meshed*)
  - Wada, nie używać jako jedynego protokołu IGP

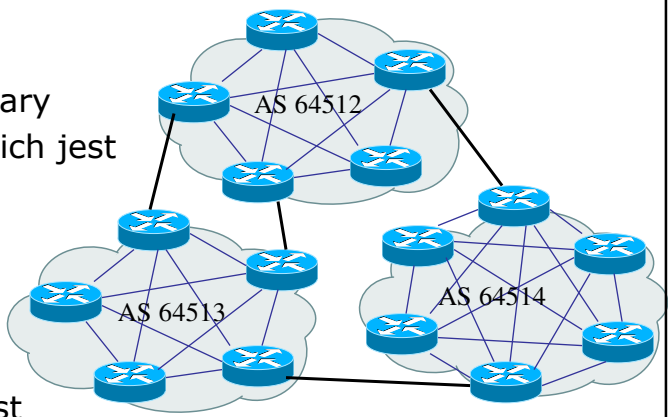
iBGP w dużym AS



Grafika: Intro to BGP: All Day Tutorial, Avi Freedsman

Konfederacje

- Podział AS na mniejsze obszary
- W każdym z nich jest połączenie punkt-punkt

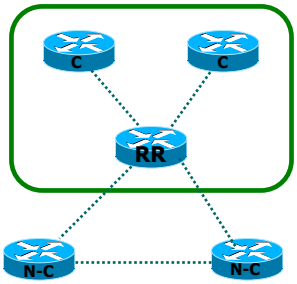


- Mechanizm jest przeźroczysty dla sieci zewnętrznych

Grafika: Intro to BGP: All Day Tutorial, Avi Freedsman

Route reflector – propagator tras

- Podział routerów iBGP w ramach AS na klientów i nie-klientów
- Nie-klient rozgłasza informację do wszystkich swoich klientów
- Router RR przekazuje informacje do wszystkich innych routerów i nie ma potrzeby tworzenia pełnej sieci (*full mesh)*
  -> topologia gwiazdy



połączenie iBGP

6. Stabilność tras

Stabilność tras

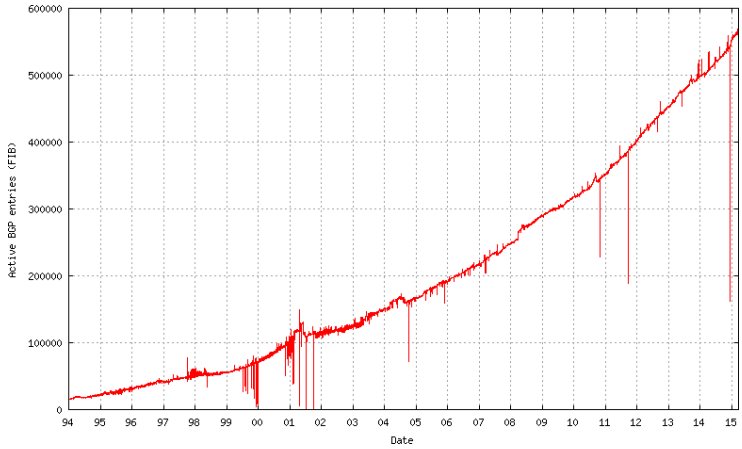
- Gdy sieć się pojawia: informacja jest propagowana
- Gdy sieć znika: informacja jest propagowana
- Często migotające sieci (*flapping*) powodują nadmierny ruch
- Mechanizm *route dampening* pozwala identyfikować takie sieci

Route dampening

- (Niestabilna) sieć otrzymuje karę (*penalty = 1000*) za każdym razem gdy pojawia się i znika
- Jeśli kary przekroczą założony limit (2000) – informacja o sieci nie jest rozgłaszana
- Wielkość kary jest automatycznie obniżana o połowę co ustalony czas
- Po obniżeniu się poniżej ustalonego progu (750) informacja o sieci znów może być ogłaszana
- Trzykrotne pojawienie się i zniknięcie (*flap*) oznacza w praktyce wyłączenie trasy na 30 minut

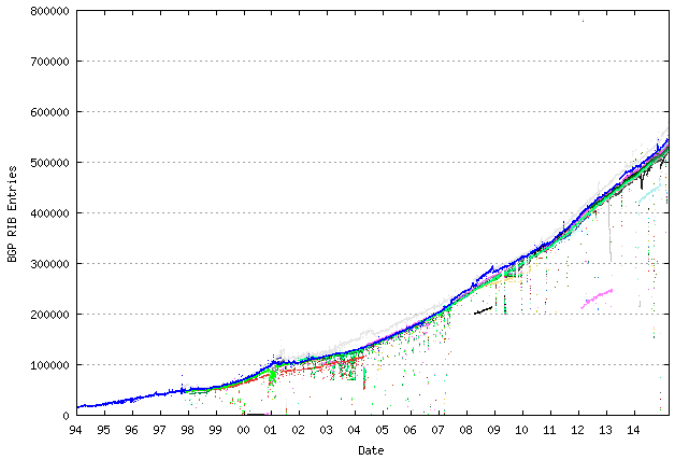
7. Podsumowanie

University of Oregon AS6447: ilość wpisów BGP



Grafika: http://bgp.potaroo.net

Wzrost wielkości tablic BGP



Grafika: http://bgp.potaroo.net

Konfiguracja BGP

router bgp *as-number*
neighbor *ip-address* **remote-as** *as-number*

network *major-network-number //rozgłasza sieci główne*

network *ip-prefix-address* **mask** *subnet-mask //*
włącza rozgłaszanie bezklasowe ale prefix (sieć, podsieć) musi pasować do wpisu z tablicy routingu

neighbor *ip-address* **update-source** *loopback 0*

Przykładowe komendy BGP

```
router bgp 60000
  network 1.1.1.0 mask 255.255.255.0
  neighbor 111.111.111.111 remote-as 60001
  neighbor 111.111.111.111 update-source loopback 0
  neighbor 111.111.111.111 next-hop-self
  neighbor 111.111.111.111 ebgp-multihop 2
  neighbor 1.1.1.1 filter-list 1 in
  neighbor 1.1.1.1 distribute-list 1 out
  neighbor 1.1.1.1 weight 777
  neighbor 1.1.1.1 route-map USTAWWAGE in
  neighbor 1.1.1.1 route-reflector-client
  bgp default local-preference 777
```

Przykład tablicy BGP i tablicy routingu

Tablica BGP

Address	Prefix	AS-Path	Next-hop	Communities	Other attr.
10.0.0.0	/8	42 13	1.2.3.4	37:12	
....					

Tablica routingu

Protocol	Address	Prefix	Next-hop	Outgoing interface
BGP	**10.0.0.0**	/8	**1.2.3.4**	---
OSPF	1.2.3.0	/24	1.5.4.1	ethernet 0
conn.	1.5.4.0	/24	---	ethernet 0

Podsumowanie

- Protokół dystans-wektor (z pewnymi uzupełnieniami)
- Wykorzystuje protokół TCP oraz port o numerze 179 do aktualizacji danych
- Cała tablica routingu jest przesyłana jedynie podczas początkowej sesji BGP
- Sesje BGP wykorzystują mechanizm "keepalive„ (utrzymanie połączenia)
- Każda zmiana w sieci powoduje przesłanie informacji o aktualizacji
- BGP posiada własną tablicę i wykorzystuje różnorodne atrybuty (pochodzenie, następny skok, …)
- Obsługuje VLSM i CIDR

Literatura

- RFC 4271
  - A Border Gateway Protocol 4 (BGP-4)
- RFC 3065
  - Autonomous System Confederations for BGP
- RFC 4456
  - BGP Route Reflection - An Alternative to Full Mesh IBGP
- RFC 2439
  - BGP Route Flap Damping

KONIEC