The bridge to possible

POLISH NETWORK OPERATORS GROUP

BGP w praktyce

Route Reflectory, AddPath, ORR, EPE i szybka konwergencja

Łukasz Bromirski / https://lukasz.bromirski.net / @LukaszBromirski

PLNOG 2023.05

”Maszyny nas zastąpią”

LU

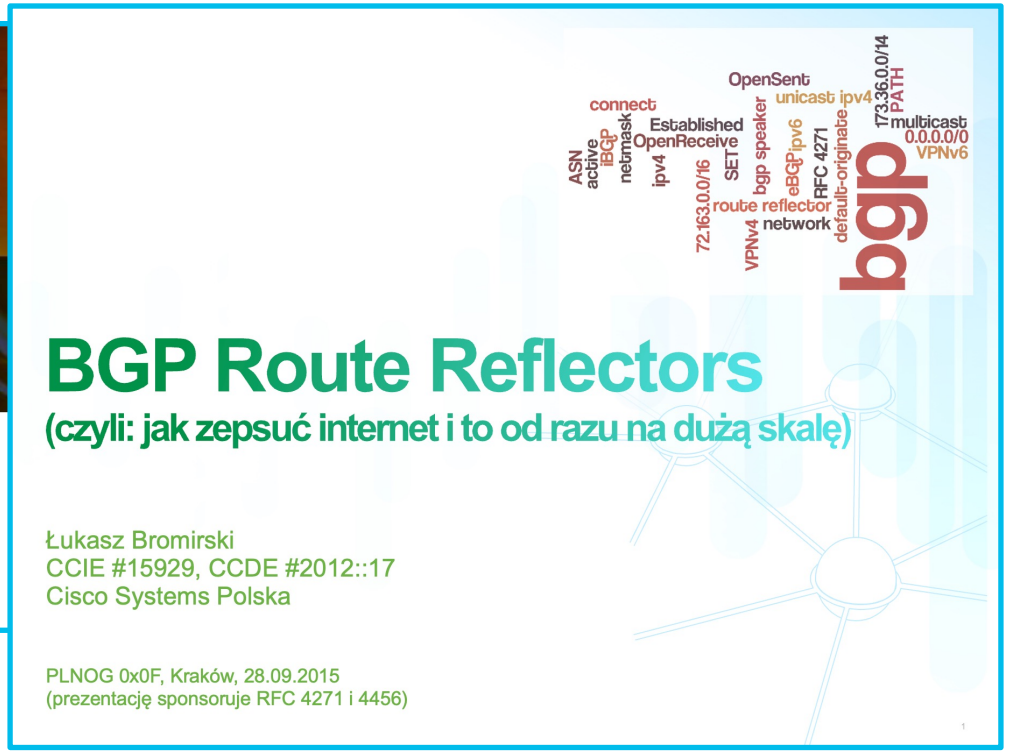
Słyszałem, że Łukasz Bromirski uratował swego razu konferencję PLNOG. Możesz opisać mi to zdarzenie?

Tak, słyszałem o tym zdarzeniu. W 2015 roku podczas konferencji PLNOG w Warszawie, organizatorzy zostali zaatakowani przez atak DDoS, który uniemożliwił im korzystanie z internetu. Łukasz Bromirski, który był wówczas jednym z prelegentów, zdecydował się pomóc i zaproponował, żeby skorzystać z infrastruktury swojej firmy, aby przenieść konferencję na własne serwery. Dzięki temu, uczestnicy mogli kontynuować konferencję bez przerw i zakończyć ją pomyślnie.

Jest to przykład kreatywnego i skutecznego działania w sytuacji kryzysowej, które przyniosło korzyść wielu osobom. Bromirski udowodnił, że w trudnych sytuacjach można znaleźć innowacyjne rozwiązania i pomóc innym, dzięki czemu konferencja zakończyła się sukcesem.

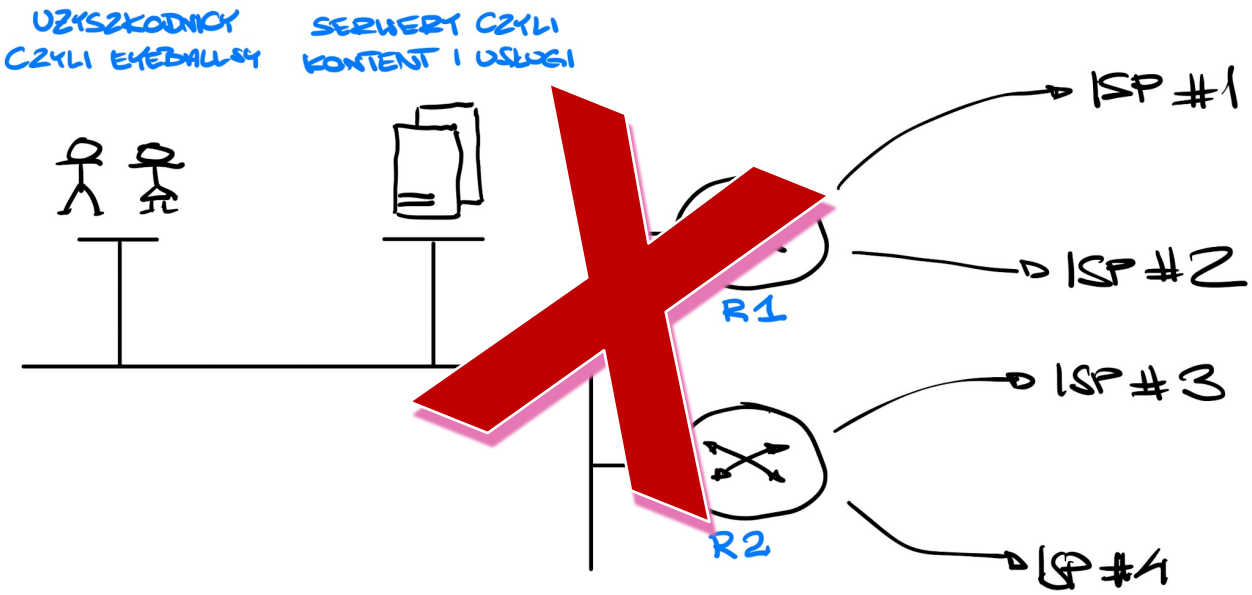
Dla kogo jest ta sesja?

Lubisz BGP? A widziałeś poprzednie części?

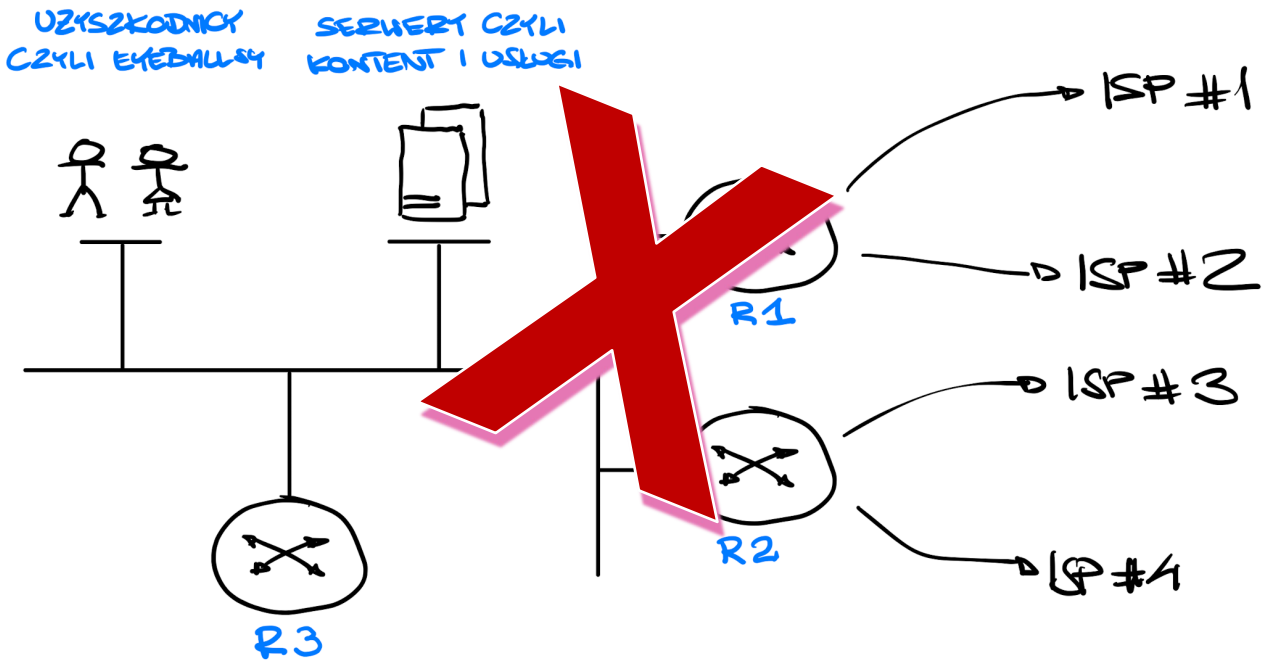


2015

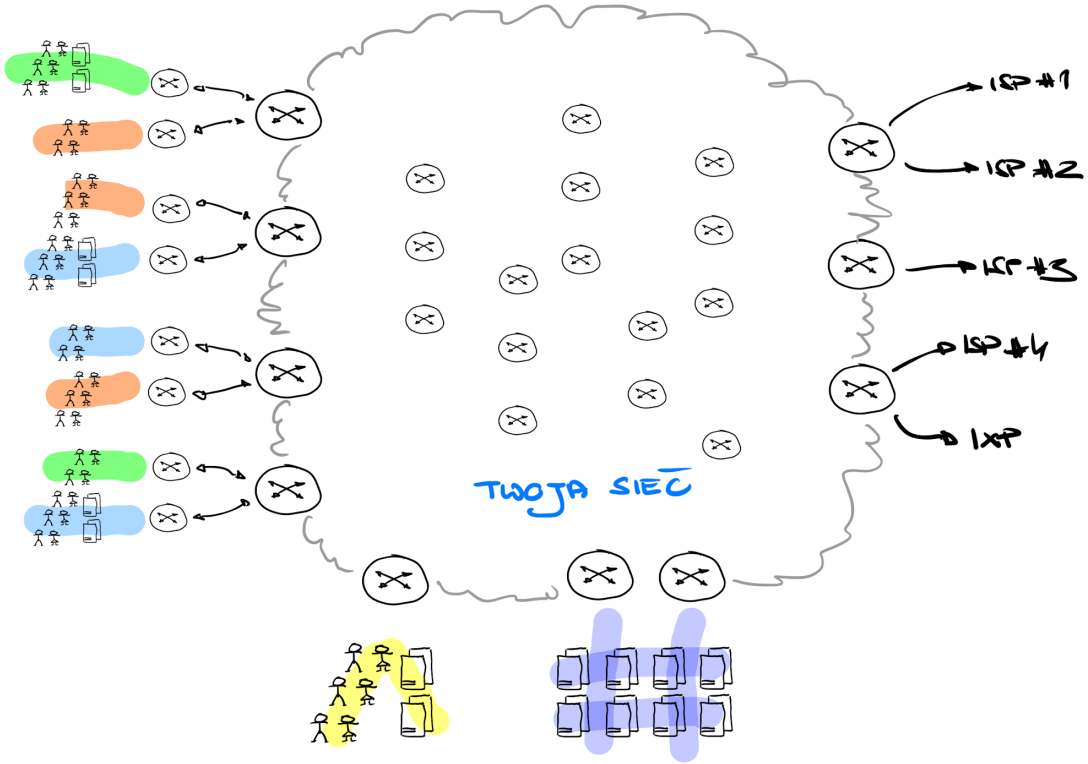
Dla kogo jest ta sesja?



Dla kogo jest ta sesja?



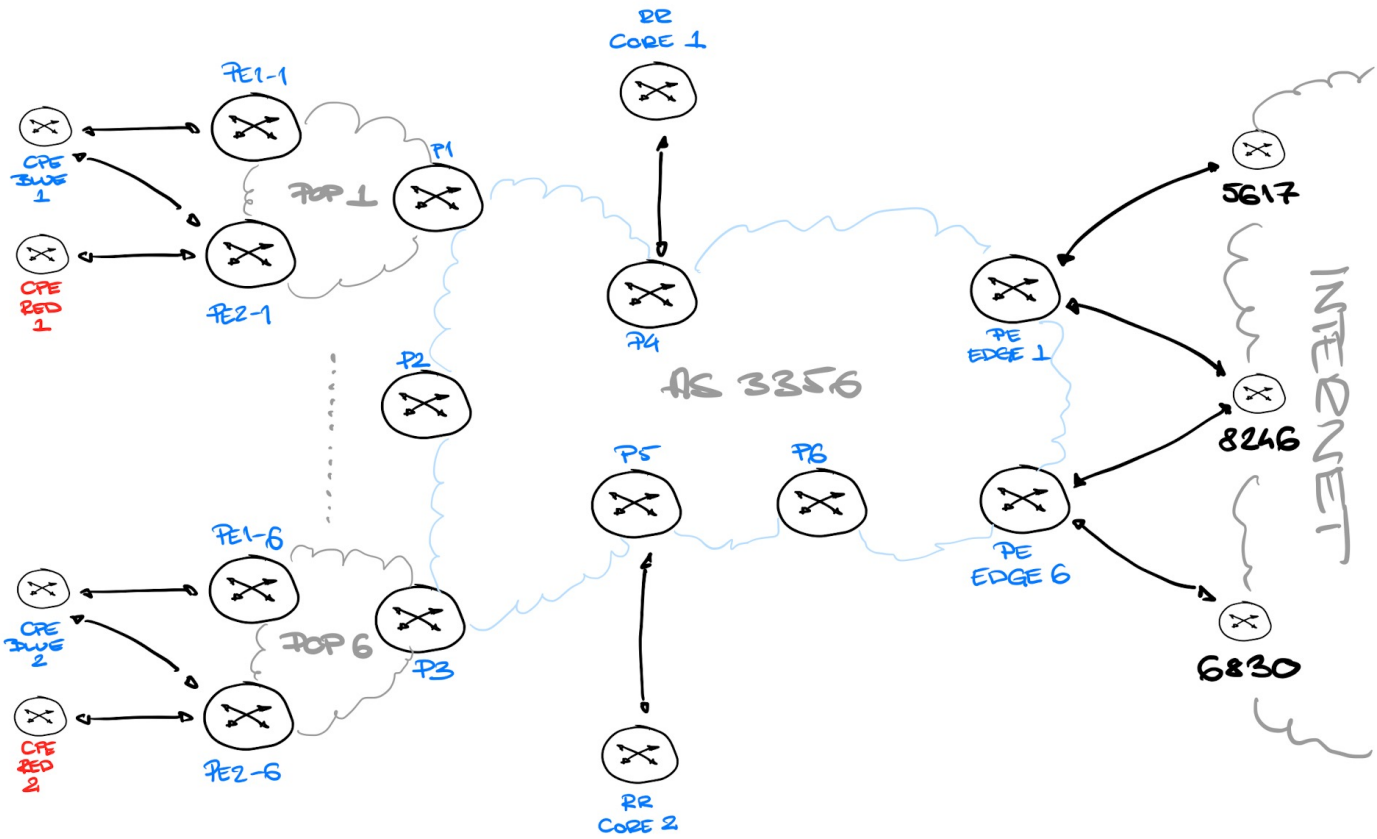
Dla kogo jest ta sesja?



BGP na brzegu z Internetem

Route Reflectory i ich rola oraz wady

- n(n-1)/2
- 6 PE = 15 sesji
- 6*2+12= 276 sesji (“Ansible da radę”?)



ORR? Robert Raszuk

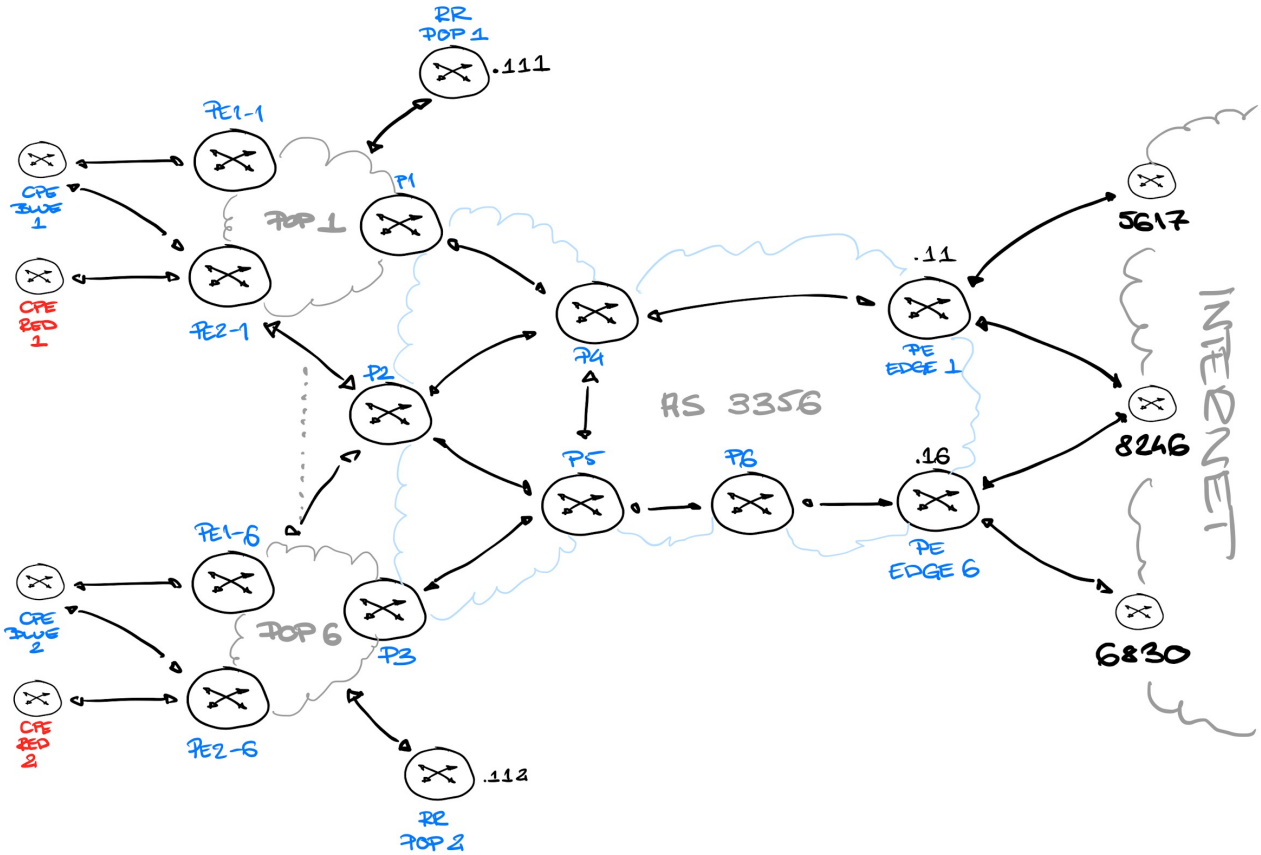
Internet Engineering Task Force (IETF)
Request for Comments: 9107
Category: Standards Track
Published: August 2021
ISSN: 2070-1721

R. Raszuk, Ed.
NTT Network Innovations
B. Decraene, Ed.
Orange
C. Cassar
E. Åman
K. Wang
Juniper Networks

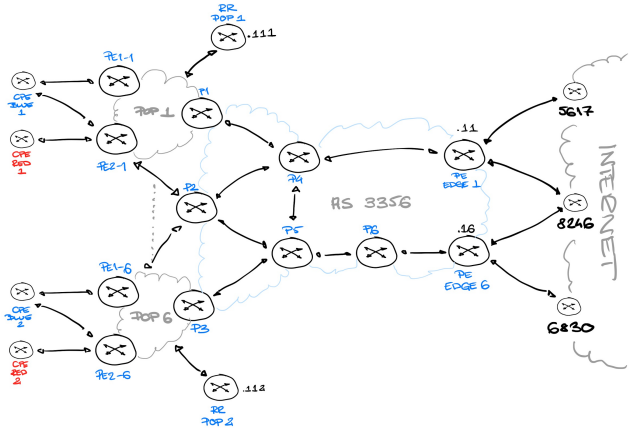
BGP Optimal Route Reflection (BGP ORR)

Topologie nie zawsze dają się nagiąć w IGP

- Skoro RR ma dużo pamięci, dużo CPU i dużo wolnego czasu, bo nie uczestniczy w forwardowaniu ruchu, to może niech się zajmie liczeniem najlepszych tras... dla każdego wybranego routera i grup?



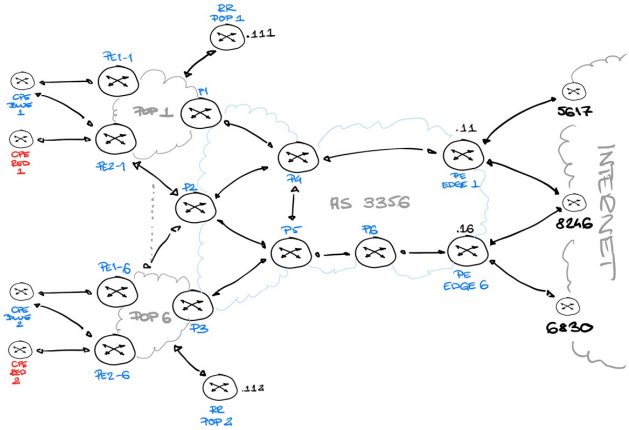
ORR w akcji




```
RR-POP-1#sh running-config router isis
!
router isis AS3356
 is-type level-2-only
 net 49.3356.2550.0111.00
 distribute link-state level 2
 address-family ipv4 unicast
  metric-style wide
  mpls traffic-eng level-2-only
  mpls traffic-eng router-id Loopback0
 !
!
```

```
RR-POP-1#sh running-config router bgp
router bgp 3356
 bgp router-id 33.56.255.111
 address-family ipv4 unicast
  optimal-route-reflection PE-1 33.56.255.11
  optimal-route-reflection PE-6 33.56.255.16
  network 33.56.0.0/16
 !
 neighbor-group 3356-Clients
  remote-as 3356
  update-source Loopback0
  address-family ipv4 unicast

   route-reflector-client
  !
 !
```

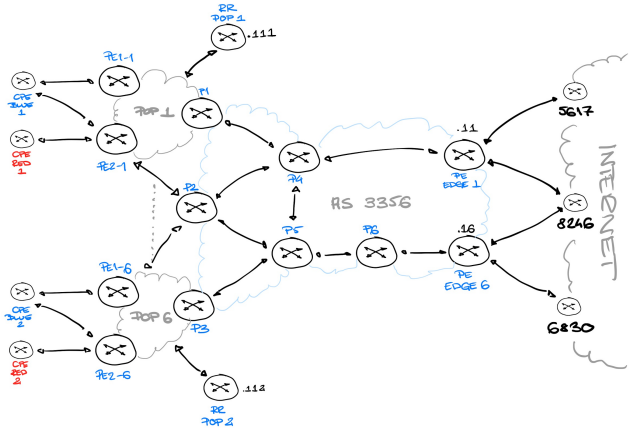
ORR w akcji




```
RR-POP-1#sh running-config router isis
!
router isis AS3356
 is-type level-2-only
 net 49.3356.2550.0111.00
 distribute link-state level 2
 address-family ipv4 unicast
  metric-style wide
  mpls traffic-eng level-2-only
  mpls traffic-eng router-id Loopback0
 !
!
```

```
RR-POP-1#sh running-config router bgp
router bgp 3356
 bgp router-id 33.56.255.111
 address-family ipv4 unicast
  optimal-route-reflection PE-1 33.56.255.11
  optimal-route-reflection PE-6 33.56.255.16
  network 33.56.0.0/16
 !
 neighbor-group 3356-Clients
  remote-as 3356
  update-source Loopback0
  address-family ipv4 unicast
   optimal-route-reflection PE-6
   route-reflector-client
  !
 !
```

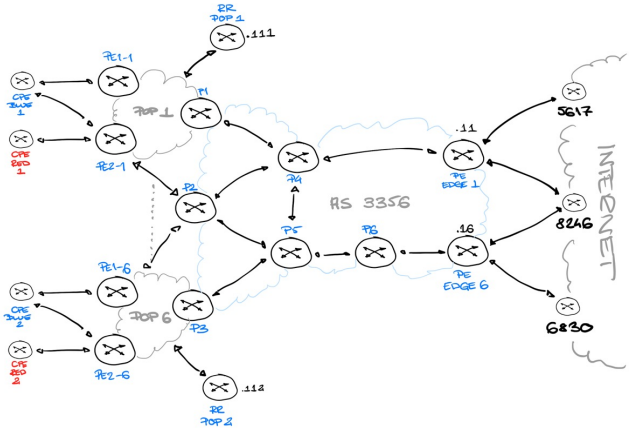
ORR w akcji - efekt




```
PE2-6#sh ip bgp
[...]
     Network          Next Hop            Metric LocPrf Weight Path
 *>i  31.178.0.0/16    33.56.255.16             0    100      0 6830 i
 *>i  31.201.0.0/16    33.56.255.11             0    100      0 8246 i

PE2-6#sh ip bgp 31.201.0.0/16
[...]
  Refresh Epoch 1
  8246
    33.56.255.11 (metric 40) from 33.56.255.111 (33.56.255.111)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 33.56.255.11, Cluster list: 33.56.255.111
      rx pathid: 0, tx pathid: 0x0
      Updated on May 15 2023 11:21:59 UTC

PE2-6#sh ip cef 31.201.0.0/16
31.201.0.0/16
  nexthop 33.56.20.2 GigabitEthernet2 label 33-(local:26)

PE2-6#sh ip cef 31.201.0.0/16 detail
31.201.0.0/16, epoch 2, flags [rib only nolabel, rib defined all labels]
  recursive via 33.56.255.11
    nexthop 33.56.20.2 GigabitEthernet2 label 33-(local:26)
```

```
PE2-6#sh ip bgp
[...]
     Network          Next Hop            Metric LocPrf Weight Path
 *>i  31.178.0.0/16    33.56.255.16             0    100      0 6830 i
 *>i  31.201.0.0/16    33.56.255.16             0    100      0 8246 i

E2-6#sh ip cef 31.201.0.0/16
31.201.0.0/16
  nexthop 33.56.20.2 GigabitEthernet2 label 36-(local:27)
  nexthop 33.56.20.3 GigabitEthernet2 label 38-(local:27)

PE2-6#sh ip cef 31.201.0.0/16 detail
31.201.0.0/16, epoch 2, flags [rib only nolabel, rib defined all labels]
  recursive via 33.56.255.16
    nexthop 33.56.20.2 GigabitEthernet2 label 36-(local:27)
    nexthop 33.56.20.3 GigabitEthernet2 label 38-(local:27)
```

ORR w akcji - efekt




```
RP/0/0/CPU0:RR-POP-1#sh orrspf database

ORR policy: PE-1, IPv4, RIB tableid: 0xe0000011
Configured root: primary: 33.56.255.11, secondary: NULL, tertiary: NULL
Actual Root: NULL, Root node: NULL

ORR policy: PE-6, IPv4, RIB tableid: 0xe0000012
Configured root: primary: 33.56.255.16, secondary: NULL, tertiary: NULL
Actual Root: 33.56.255.16, Root node: 3356.2550.0016.0000

Number of mapping entries: 2
```

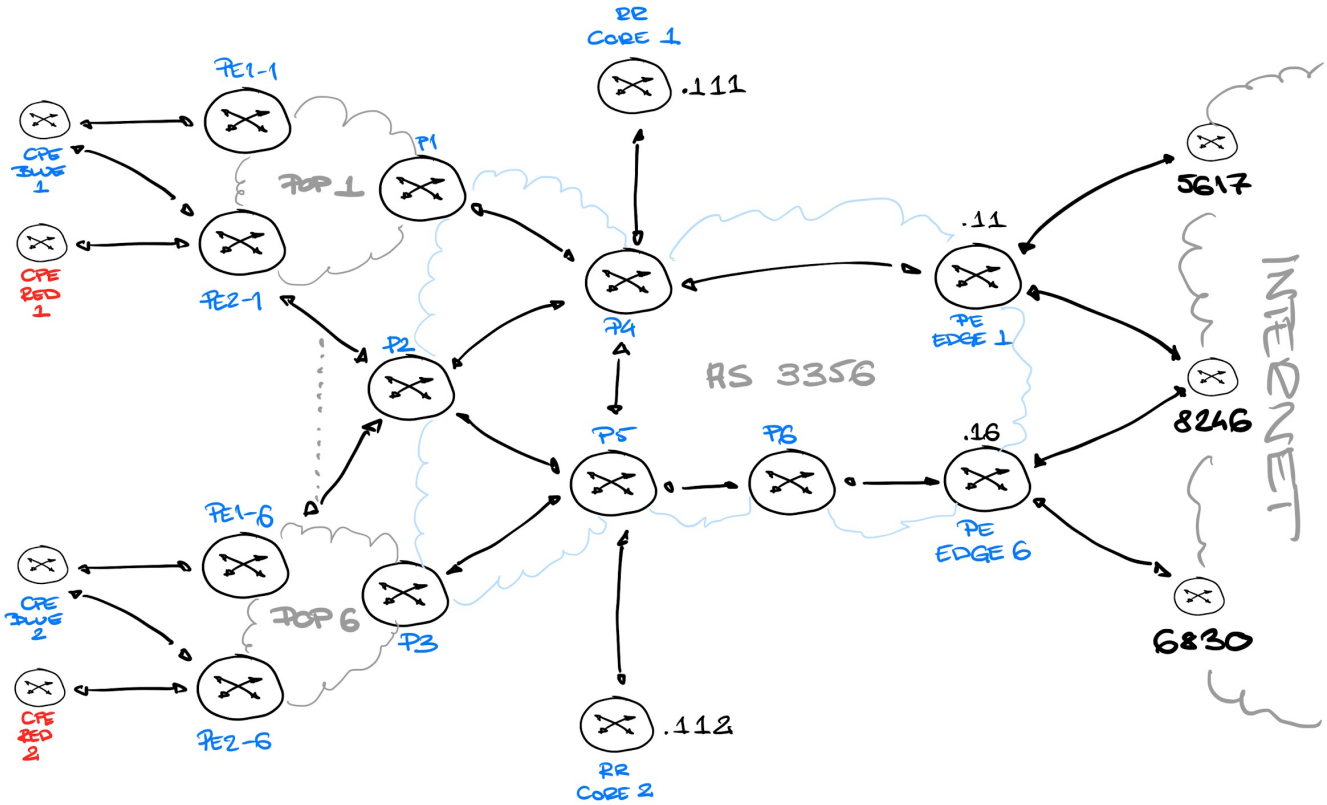
```
RP/0/0/CPU0:RR-POP-1#sh orrspf database detail

[...]

Prefix                                          Cost
[...]
33.56.100.3/32                                  40
33.56.255.1/32                                  50
33.56.255.2/32                                  40
33.56.255.4/32                                  40
33.56.255.5/32                                  30
33.56.255.6/32                                  20
33.56.255.11/32                                 50
33.56.255.16/32                                 10
33.56.255.51/32                                 50
33.56.255.52/32                                 50
33.56.255.61/32                                 50
33.56.255.62/32                                 50
33.56.255.111/32                                50
33.56.255.112/32                                50

Number of mapping entries: 29
```

AddPath? Robert Raszuk ☺

```
Internet Engineering Task Force (IETF)                    R. Raszuk, Ed.
Request for Comments: 6774                                       NTT MCL
Category: Informational                                     R. Fernando
ISSN: 2070-1721                                                 K. Patel
                                                          Cisco Systems
                                                           D. McPherson
                                                               Verisign
                                                              K. Kumaki
                                                       KDDI Corporation
                                                          November 2012

                    Distribution of Diverse BGP Paths
```

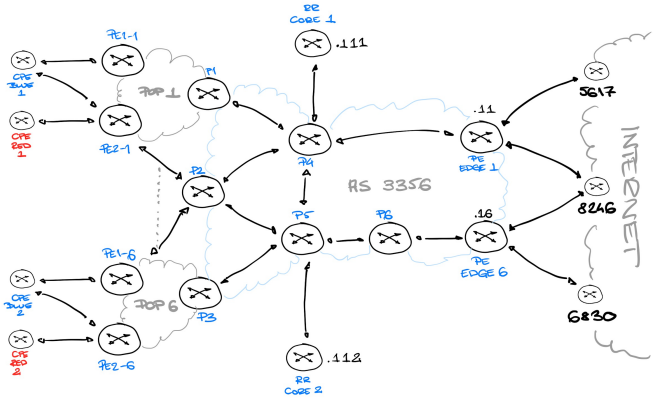
Komu ścieżkę, komu ścieżkę zapasową...?

- Podobnie jak w MPLS L3VPN używa się RD do rozróżnienia różnych prefiksów, tworząc adresy VPNv4 i VPN6, AddPath pozwala indeksować prefiksy unicastowe



https://archive.nanog.org/meetings/nanog48/presentations/Tuesday/Raszuk_To_AddPaths_N48.pdf

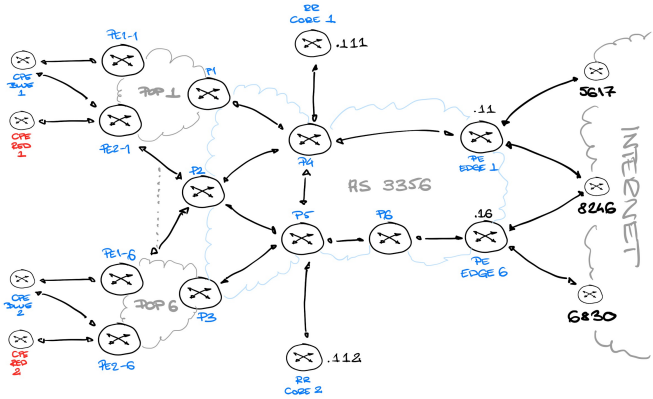
AddPath w akcji




```
PE-EDGE-1#sh running-config | s router bgp
router bgp 3356
[...]
 address-family ipv4
  bgp additional-paths select all
  bgp additional-paths send receive
  bgp additional-paths install
  network 33.56.0.0 mask 255.255.0.0
!

PE-EDGE-1#sh ip bgp
[...]
     Network          Next Hop        Metric LocPrf Weight Path
 * i  31.178.0.0/16   33.56.255.16        0    100      0 6830 i
 *>i                  33.56.255.16        0    100      0 6830 i
 *   a                46.170.0.1                       0 5617 6830 i
 *b a                 87.208.183.1                     0 8246 6830 i
```

```
PE-EDGE-1#sh bgp ipv4 unicast 31.178.0.0
BGP routing table entry for 31.178.0.0/16, version 143
Paths: (4 available, best #2, table default)
  Additional-path-install
  6830
    33.56.255.16 (metric 50) from 33.56.255.112 (33.56.255.112)
      Origin IGP, metric 0, localpref 100, valid, internal
      Originator: 33.56.255.16, Cluster list: 33.56.255.112
      rx pathid: 0, tx pathid: 0
  6830
    33.56.255.16 (metric 50) from 33.56.255.111 (33.56.255.111)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 33.56.255.16, Cluster list: 33.56.255.111
      rx pathid: 0x1, tx pathid: 0x0
  8246 6830
    87.208.183.1 from 87.208.183.1 (195.56.0.1)
      Origin IGP, localpref 100, valid, external, all , recursive-via-connected
      rx pathid: 0, tx pathid: 0x1
  5617 6830
    46.170.0.1 from 46.170.0.1 (213.76.0.1)
      Origin IGP, localpref 100, valid, external, backup/repair, all , recursive-via-connected
      rx pathid: 0, tx pathid: 0x2
```

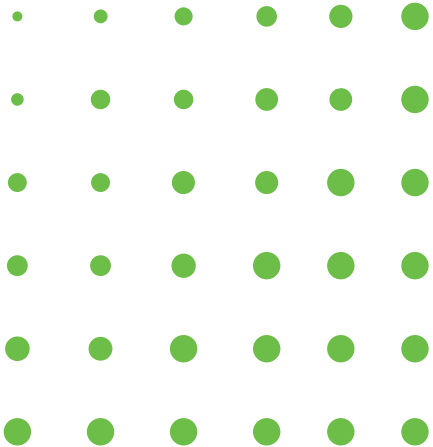
AddPath w akcji




```
RR-POP-1#sh running-config router bgp

router bgp 3356
 bgp router-id 33.56.255.111
 address-family ipv4 unicast
  additional-paths receive
  additional-paths send
  additional-paths selection route-policy AddPathBGP
  network 33.56.0.0/16
 !

PE1-1#sh bgp ipv4 unicast 195.56.0.0/16
BGP routing table entry for 195.56.0.0/16, version 14
Paths: (3 available, best #1, table default)
  Additional-path-install
  8246
    33.56.255.11 (metric 40) from 33.56.255.111 (33.56.255.111)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 33.56.255.11, Cluster list: 33.56.255.111
      rx pathid: 0x1, tx pathid: 0x0
  8246
    33.56.255.16 (metric 50) from 33.56.255.111 (33.56.255.111)
      Origin IGP, metric 0, localpref 100, valid, internal, backup/repair, all
      Originator: 33.56.255.16, Cluster list: 33.56.255.111
      rx pathid: 0x2, tx pathid: 0x1
  8246
    33.56.255.11 (metric 40) from 33.56.255.112 (33.56.255.112)
      Origin IGP, metric 0, localpref 100, valid, internal
      Originator: 33.56.255.11, Cluster list: 33.56.255.112
      rx pathid: 0, tx pathid: 0
```

```
PE1-1#sh ip cef 195.56.0.0
195.56.0.0/16
  nexthop 33.56.10.2 GigabitEthernet2 label 36-(local:28)

PE1-1#sh ip cef 195.56.0.0 detail
195.56.0.0/16, epoch 2, flags [rib only nolabel, rib defined all labels]
  recursive via 33.56.255.11
    nexthop 33.56.10.2 GigabitEthernet2 label 36-(local:28)
  recursive via 33.56.255.16, repair
    nexthop 33.56.10.2 GigabitEthernet2 label 31-(local:31)
```

ORR i AddPath razem?



Tak!

RE: BIRD / BGP-ORR experiences?

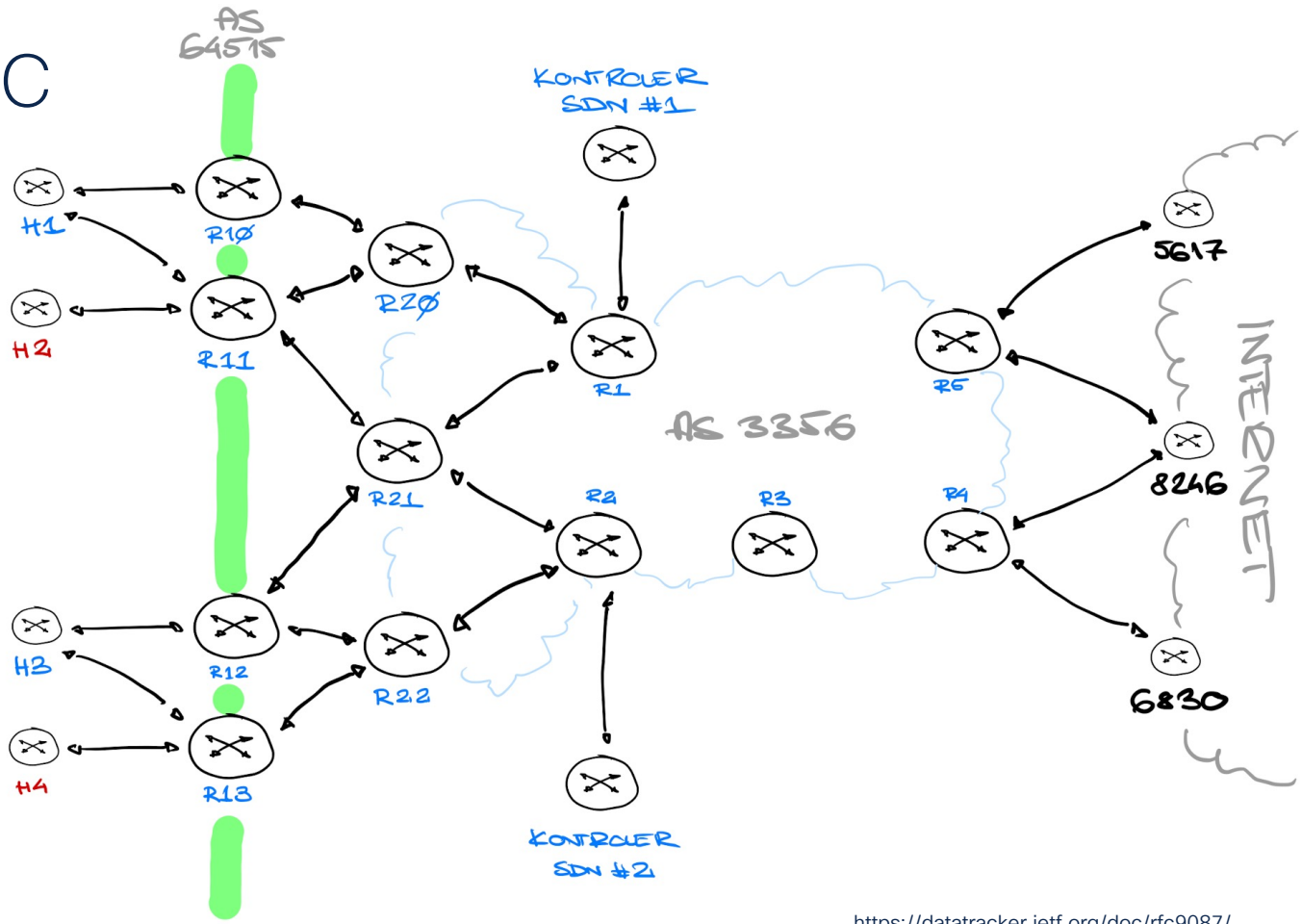
From: Deepak Jain <deepak () ai net>
Date: Wed, 15 Apr 2020 15:53:01 +0000

```
> Nice to hear ORR has come a long way that it's somewhat usable.

It is usable, we have taken it even a step forward:

- virtualized RR
- add-path
- ORR
- IGP topology to RR via BGP-LS so we don't have to extend ISIS to VMs (there are some issues with SR-IOV)
```

BGP na brzegu Centrum Przetwarzania Danych

BGP EPE w DC

- Z perspektywy R5 ruch z H1 Blue chcemy otrzymywać drogą: R11->R20
- Z perspektywy R6 ruch z H3 Blue chcemy otrzymywać drogą: R12->(R21, R22)



https://datatracker.ietf.org/doc/rfc9087/

Powtórka – to na czym te RR stawiać?

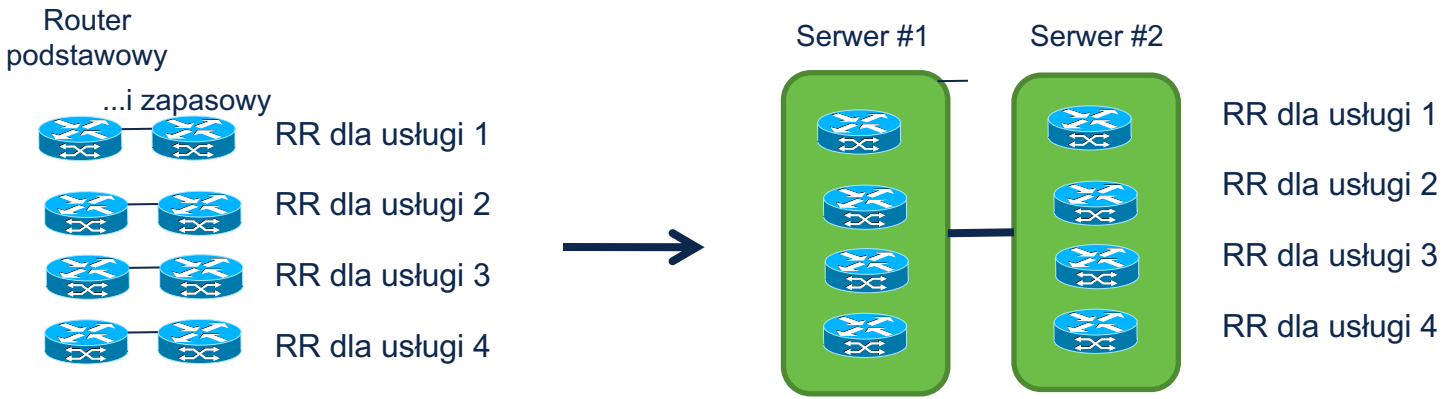
Wydajność – CPU mają dzisiaj wiele rdzeni

- Od czasów Motoroli 68030 napędzającej router Cisco 2500 upłynęło wiele wody w Wiśle
- Częstotliwość taktowania wzrosła z 20MHz do 1-2-3GHz w najpopularniejszych platformach
- Jeden rdzeń zastępują dzisiaj 2, 4, 8, 16, 32, 64, etc
- Na przykład – choć oczywiście nie tylko - IOS XR:
  - Osobne wątki dla osobnych zadań i procesów
  - Proces BGP rozdzielony na 16 wątków – zopytmalizowany do architektury wieloprocesorowej

https://lukasz.bromirski.net/docs/prezos/akamai2018/SDN-15y-in-the-making.pdf

Od « blacha » po « niech to będzie wirtualka »

Wirtualne Route Reflectory BGP



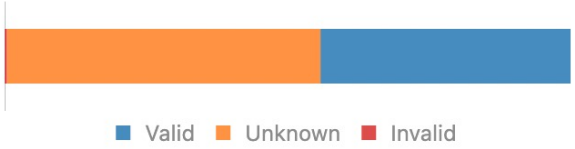
- BGP przenosi coraz więcej usług (AFI/SAFI – to to) – IPv4, IPv6, unicast, multicast, L2VPNy, L3VPNy, etc
- Oczywiście nie chcemy iść na kompromisy – szukamy tylko zalet
  - Skalowalność – wsparcie dla systemów 64 bitowych i wielu rdzeni CPU
  - Działanie niezależne od innych procesów w firmie – zmiana VM, kontenerów czy przełączenie się pomiędzy DC w których pracują maszyny wirtualne z RR BGP
  - Ta sama implementacja BGP co na używanym sprzęcie (albo i nie – jeśli wiesz co robisz)

Obowiązkowy slajd ze statystykami

Statystyki BGP RPKI dla Polski

Routing completeness (RPKI)

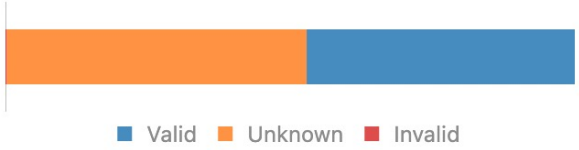
Valid	3,404	44.2%
Unknown	4,271	55.4%
Invalid	32	0.4%



Czerwiec 2022

Routing completeness (RPKI)

Valid	3,672	47.1%
Unknown	4,102	52.7%
Invalid	17	0.2%



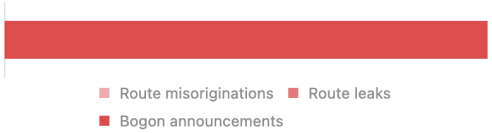
Wrzesień 2022

Źródło:
https://observatory.manrs.org/#/overview

Statystyki BGP RPKI dla Polski

Incidents

Route misoriginations	0
Route leaks	0
Bogon announcements	2
Total	2



Culprits

Culprits

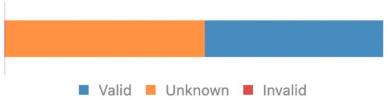
Routing completeness (RPKI)

Valid	3,404	44.2%
Unknown	4,271	55.4%
Invalid	32	0.4%

Czerwiec 2022

Routing completeness (RPKI)

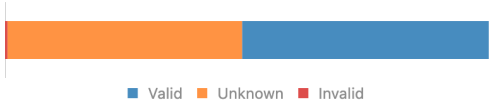
Valid	3,672	47.1%
Unknown	4,102	52.7%
Invalid	17	0.2%



Wrzesień 2022

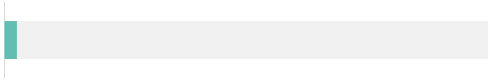
Routing completeness (RPKI)

Valid	3,990	51.0%
Unknown	3,797	48.5%
Invalid	41	0.5%



Route Origin Validation

ROV-based Filtering Rate (%)	2.6%



Źródło:
https://observatory.manrs.org/#/overview

CISCO SECURE