Nowości w protokole BGP
Optymalizacja routingu

Łukasz Bromirski
lukasz.bromirski@cisco.com
CCIE #159292 R&S/SP, CCDE #2012::17

PLNOG

Rozkład jazdy

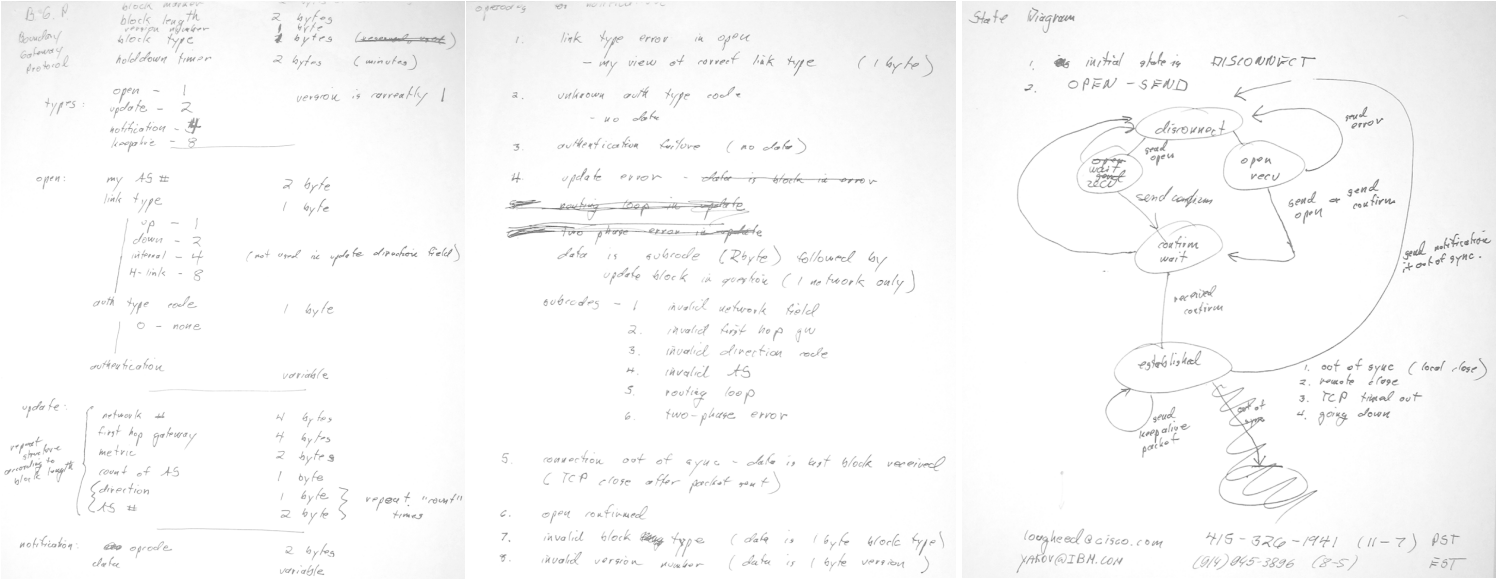
- Protokół BGP
- Skalowanie BGP – platformy i możliwości
- BGP i wysoka dostępność
- BGP i diagnostyka – BMP
- BGP multipathing
- BGP a SDN
- Optymalizacja BGP na brzegu

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Tak stworzono protokół BGP

(na trzech serwetkach, w kawiarni)

- Tyle było potrzeba, aby zbudować prosty, rozszerzalny protokół do obsługi wymiany informacji pomiędzy systemami autonomicznymi w 1989 roku:



12-ste spotkanie IETF: Len Bosack, Kirk Lougheed, Yakov Rekhter

© 2015 Cisco and/or its affiliates. All rights reserved.

Ewolucja warstwy kontrolnej internetu

Usługi coraz częściej i chętniej wykorzystują BGP

Zastosowanie	Kiedyś (~2000)
Internet (Peering)	BGP IPv4
Prywatne IP (L3VPN)	BGP IPv4
Prywatny multicast (Mc VPN)	PIM
Usługi L2 (L2VPN)	LDP VPWS/VPLS
Zapobieganie DDoSom	CLI, ACL, PBR
Monitoring sieciowy	SNMP
Security	Filtry
Wykrywanie bliskości (proximity)	
Skalowanie DC	IGP (ISIS, OSPF) or L2 (Trill, FP, Vlan)
Transport MPLS	LDP
SDN	PBR, OpenFlow (2013), Yang (future)
Transport nakładkowy (overlay)	

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Skoro jesteśmy w 2015 to na czym teraz uruchamia się BGP?

Mamy do dyspozycji systemy 64-bitowe

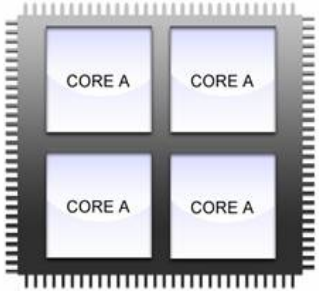
(tak, nawet w sprzęcie sieciowym)

- 2^32 = 4,294,967,296 bajtów
  Realnie – 1.5 do 4 millionów prefiksów BGP
- 2^64 = 18,446,744,073,709,551,616 bajtów
  Realnie – obecnie komercyjnie niespotykana konfiguracja ;P
- IOS-XR
  64 bitowa wersja na NCS 6000
  Oprogramowanie gotowe dla ASR 9000, CRS-3 i XRv (wersja wirtualna)
- IOS – XE
  64 bitowa wersja dla ASR 100x, CSR1000v, ISR 44xx

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Wydajność – CPU mają dzisiaj wiele rdzeni

(tak, nawet w sprzęcie sieciowym)



- Od czasów Motoroli 68030 napędzającej router Cisco 2500 upłynęło wiele wody w Wiśle
- Częstotliwość taktowania wzrosła z 20MHz do 1-2GHz w najpopularniejszych platformach
- Jeden rdzeń zastępują dzisiaj 2, 4, 8, 16
- IOS XR:

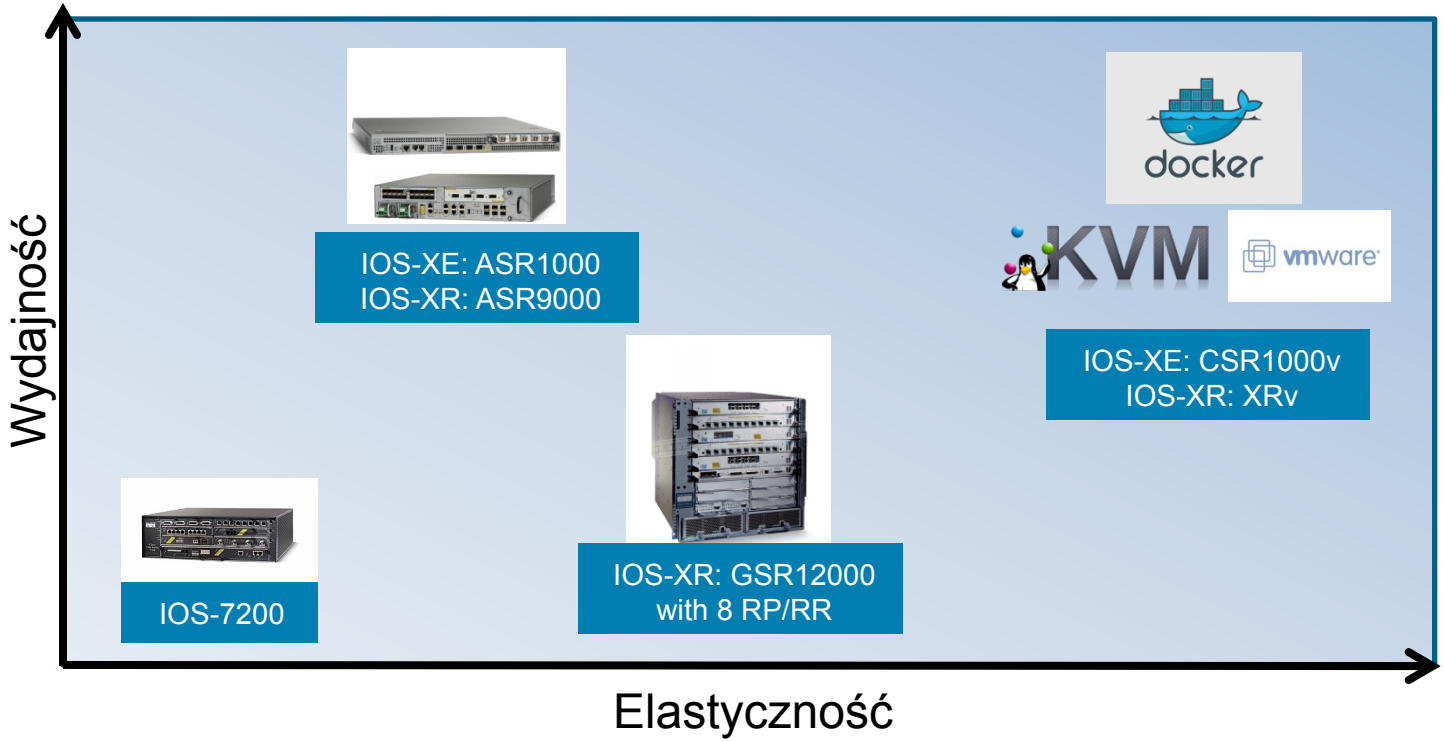
  Osobne wątki dla osobnych zadań i procesów

  Proces BGP rozdzielony na 16 wątków – zopytmalizowany do architektury wieloprocesorowej

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

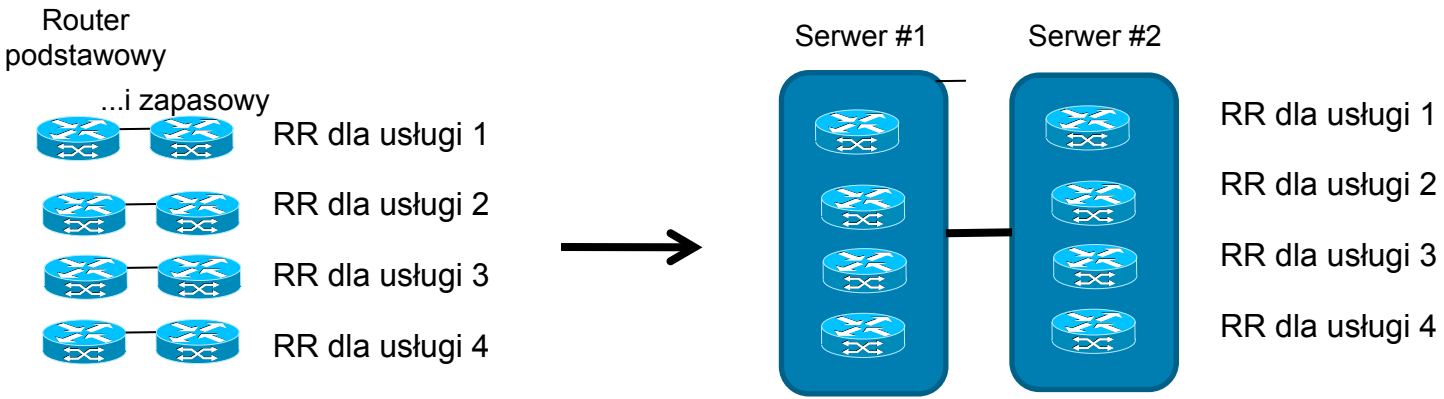
Wirtualne BGP

Ewolucja Route Reflectorów BGP



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Wirtualne Route Reflectory BGP



- Ilość osobnych płaszczyzn kontroli rośnie – wynika z coraz większej ilości usług przenoszonych przez BGP – i napędza to optymalizację ale i wirtualizację funkcji RR BGP
- Oczywiście nie chcemy iść na kompromisy – 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
  - Zarządzanie (Hypervisor)

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP i wysoka dostępność (i szybkie odtwarzanie działania sieci)

Dostępne przepustowości a wysoka dostępność

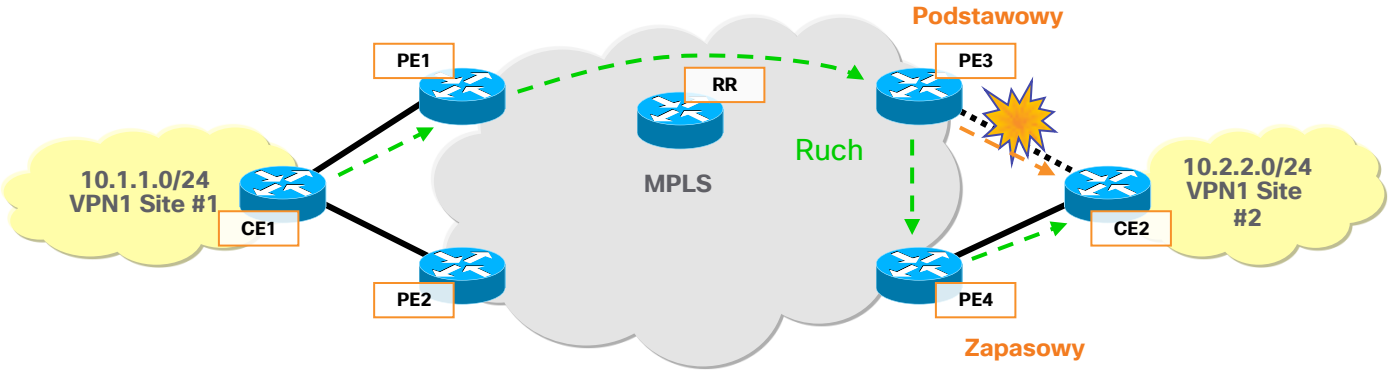
	2000	2015	Przyszłość
Dostępne pasmo	E1/T1 (~2Mbit/s)	100GE	Wiele 100GE, 400Gbits/ 1Tbit/s
Ilość pakietów utraconych na sekundę	400	400.000.000	Powyżej miliarda
Konwergencja BGP	5-10 min	50-200 msec	BGP PIC (*)
Konwergencja OSPF/ISIS	1 min	50 msec	LFA FRR (*)
Multi-path	brak	32	64, 128 (*)
Wysoka dostępność węzła	brak	Pełne HA	NSR(*), ISSU, BGP GSHUT

(*) Innowacja Cisco

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

PIC Edge: ochrona łącza

Rozszerzenie HA dla BGP



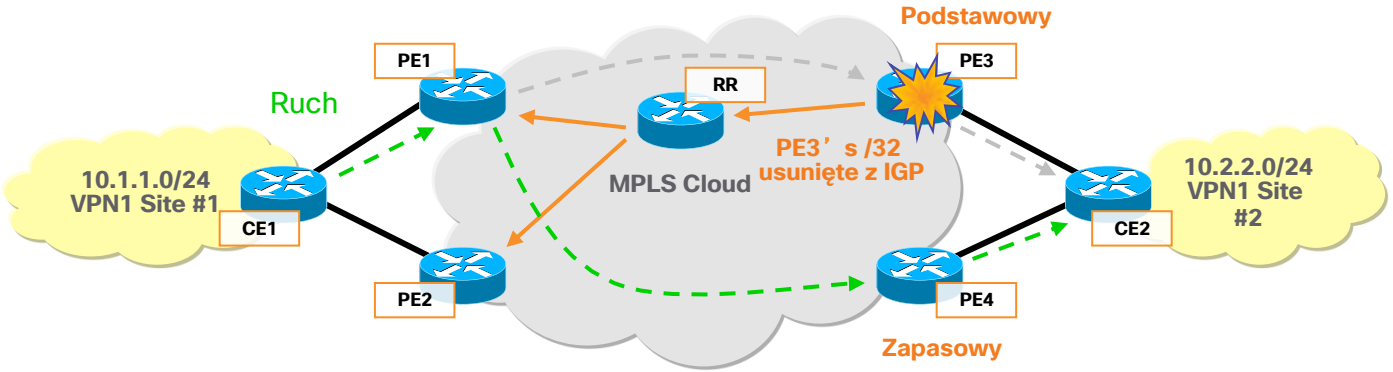
- CEF (dzięki BFD lub informacji od sterownika interfejsu sieciowego) wykrywa awarię PE3-CE2

  CEF podmienia etykiety na ścieżkę zapasową i ruch przechodzi do PE4 oraz na łącze PE4-CE2

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

PIC Edge: ochrona węzła brzegowego

Rozszerzenie HA dla BGP



- PE1 wykrywa utratę adresu /32 routera PE3 w IGP
  CEF podmienia etykietę z PE3 na PE4
- BGP na PE1 wyliczy najlepszą ścieżkę potem – wybierając PE4

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Graceful Shutdown

- Mechanizm pozwala wykonać prace serwisowe na węźle prowadzącym routing BGP bez przerywania ruchu produkcyjnego

- Sąsiedzi przekierowują ruch i po określonym czasie wygaszają sesje BGP
- Modyfikacja/przekierowanie może być wykonane za pomocą Local Preference lub community (GSHUT)

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Lepsza diagnostyka i odporność na błędy

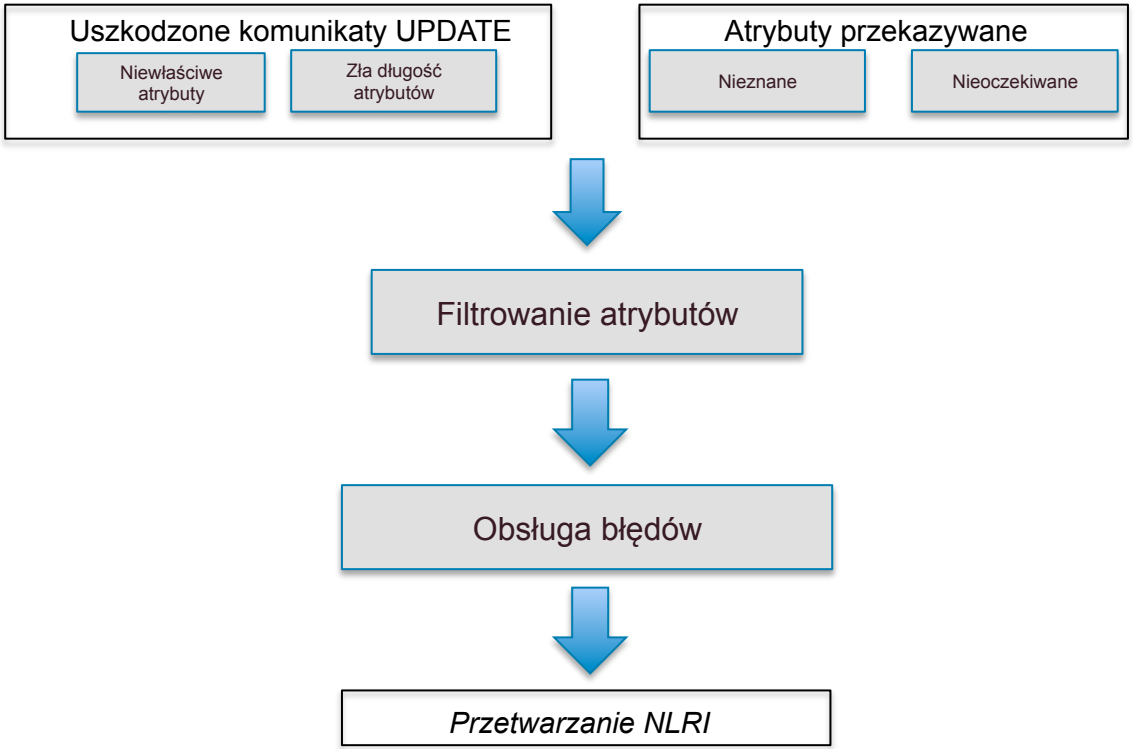
Rozszerzona obsługa błędów w BGP

- Obsługa sytuacji problemowych w BGP w dużej mierze opiera się o zerwanie sesji
- Problemy z błędnymi implementacjami BGP i nowymi implementacjami/ skryptami BGP (SDN, jednorożce ☺) powodują propagowanie się różnego rodzaju sytuacji niepożądanych
- Atrybuty “niespotykane”, takie jak ATTR_SET, CONFED w AS4_PATH potrafią zabić całą sieć przez zabicie routera
- Powoli kończy się standaryzacja obsługi błędów i sytuacji krytycznych w BGP dzięki nowemu draftowi*

* https://tools.ietf.org/html/draft-ietf-idr-error-handling-18

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Rozszerzona obsługa błędów w BGP - logika



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Rozszerzona obsługa błędów w BGP

- Zdefiniowane sytuacje wyjątkowe*:
  - Otrzymanie komunikatu NOTIFICATION i zerwanie sesji – błędy krytyczne i nie obsługiwane poniżej
  - Wyłączenie kombinacji AFI/SAFI – RFC4760 określa procedurę wyłączenia konkretnej rodziny protokołów bez zrywania sesji
  - Obsługa tak jak wycofanie (WITHDRAW) – obsługa komunikatu sprowadza się do logicznej zamiany wszystkich prefiksów z rozgłoszenia tak, jak gdyby zostały rozgłoszone jako wycofywane – co oznacza między innymi wycofanie z Adj-RIB-In
  - Pominięcie atrybutu – tylko tam, gdzie atrybut nie wpływa na wybór ścieżki lub jej instalację

* https://tools.ietf.org/html/draft-ietf-idr-error-handling-18

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Rozszerzona obsługa błędów w BGP

- Obsługa konkretnych atrybutów jako wycofanie rozgłoszenia (‘withdraw’)

```
router bgp 65002
 neighbor 2001:db8:1::254 path-attribute treat-as-withdraw 100 in
```

- Odrzucanie konkretnych atrybutów z komunikatu UPDATE

```
router bgp 65002
 neighbor 2001:db8:1::254 path-attribute discard 14 in
 neighbor 2001:db8:1::254 path-attribute discard range 17 254 in
```

- Co zostało odrzucone/rozpoznane jako nieznane:

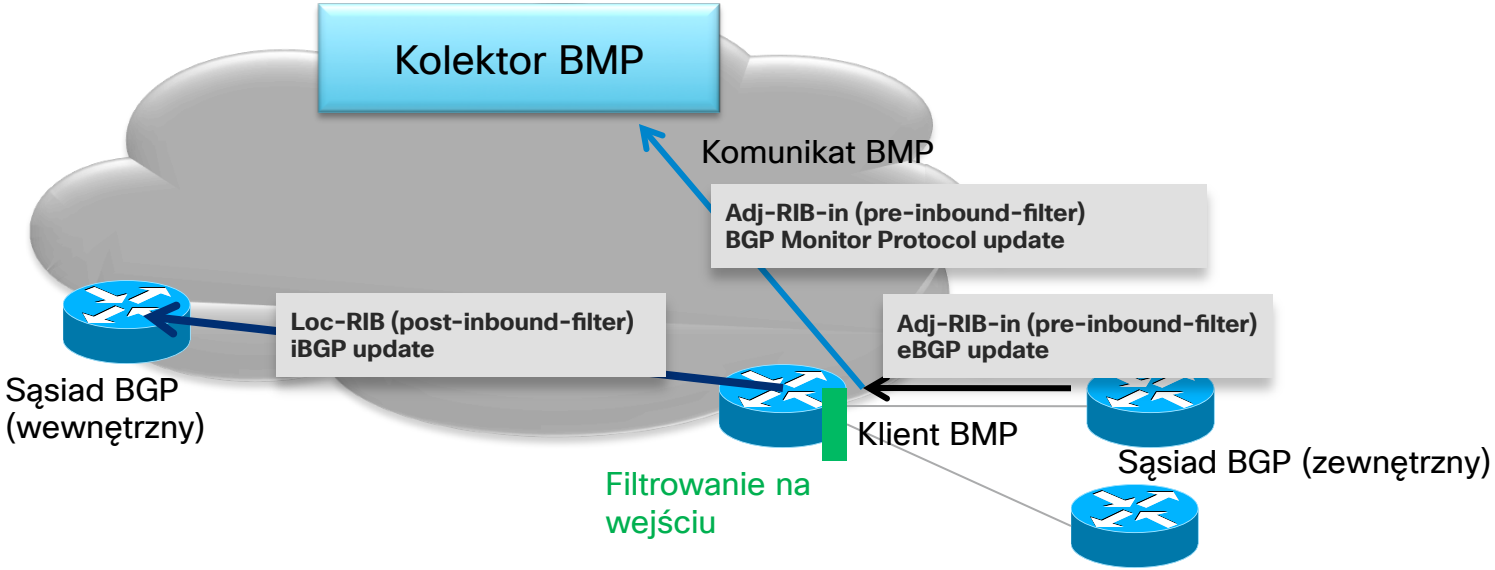
```
show ip bgp path-attribute [unknown|discard]
```

“Gdzie znajdę numery atrybutów i ich znaczenie?”
http://www.iana.nl/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BMP – BGP Management Protocol

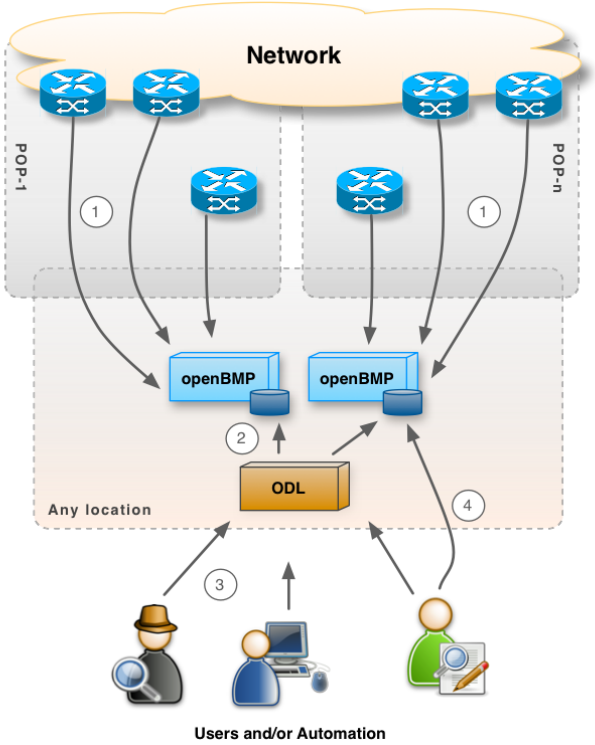
Kolektor/klient BMP



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Po co BMP?

(prawie) pełen obrazek



1. Urządzenia obsługujące sygnalizację BMP wysyła komunikaty do kolektora BMP

2. Komunikaty zbierane są przez plugin SQL z kontrolerów takich jak Open Daylight (ODL) czy NCS (TAIL-f).

3. Administratorzy, inżynierowie sieciowi, skrypty mogą współpracować z ‘prostszym’ interfejsem ODL/NCS na pełnych danych diagnostycznych vs na żywej infrastrukturze routingowej (czego nikt nie lubi)

Przykłady konfiguracji z OpenBMP:
http://openbmp.org/#!docs/ROUTER_CONFIG.md

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Po co BMP?

Jest parę innych rozwiązań...

- Parę dobrych powodów:
  - Looking Glasses - IPv4, IPv6, and VPN4
  - Analityka tras – czasy konwergencji, historia prefiksów, monitoring zmian polityk BGP
  - Analityka inżynierii ruchowej – zmiany dynamicznie dostosowywane na podstawie szerokiego spojrzenia na sieć
  - Symulacje – na podstawie widoku utrzymywanego przez BMP symulacja nakładania polityk BGP
- ...oczywiście możliwości są – jak zwykle – “nieograniczone”*

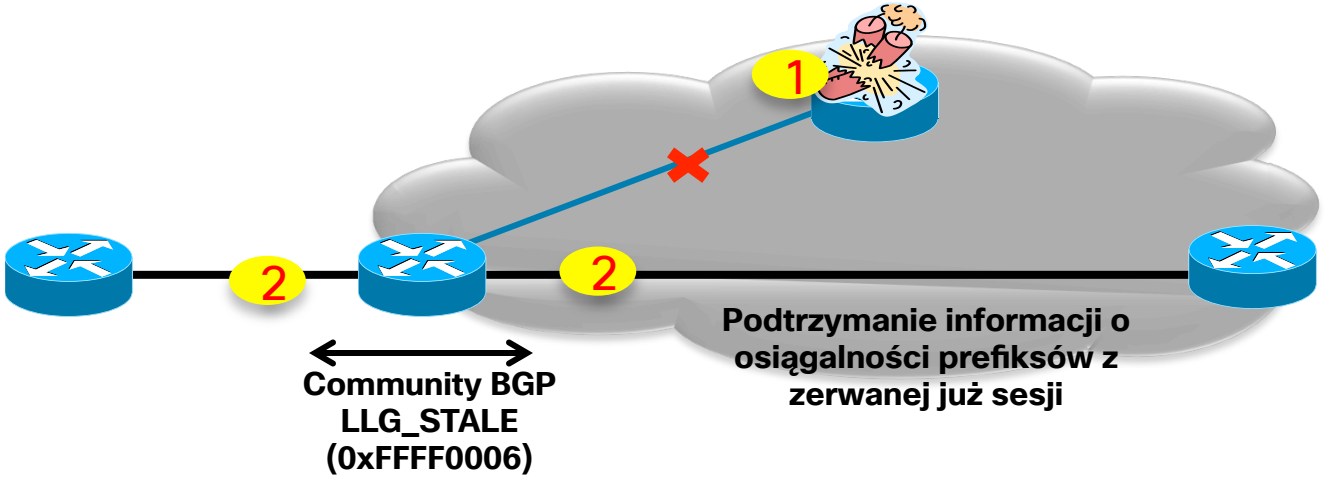
* Jeśli przez ‘nieograniczone’ rozumiemy załadowanie wszystkiego do SQLa i zabawę danymi ☺

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP persistence

(albo Long-Lived Graceful Restart – powyżej 4095 sekund)

- Podtrzymanie informacji z zamkniętej sesji BGP dla konkretnych kombinacji AFI/SAFI
  http://tools.ietf.org/html/draft-uttaro-idr-bgp-persistence-03



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Sygnalizacja wielu ścieżek w BGP

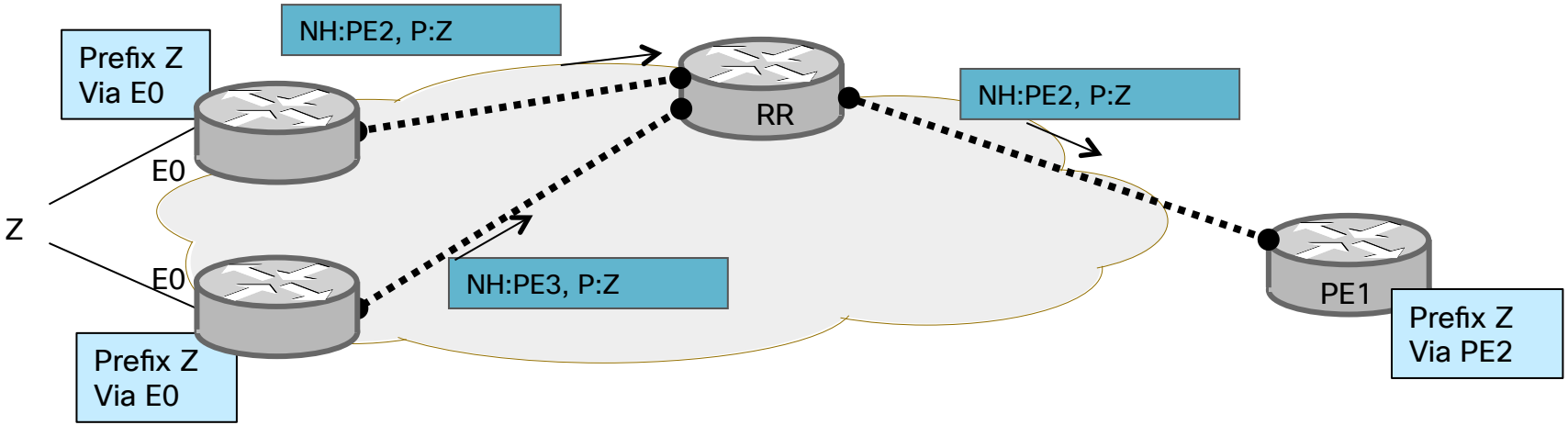
Po co nam w ogóle wiele ścieżek?

- Konwergencja (i jej czas)
  - BGP Fast Convergence (2 i więcej ścieżek w bazie BGP)
  - BGP PIC Edge (2 i więcej ścieżek od razu zaprogramowane w FIBie)
- Multipath load balancing
  - ECMP LB (głównie w Data Center, lub przy wielu łączach 10GE przed migracją do 100GE)
- Zapobieganie oscylacjom routingu
- Obsługa scenariusza routingu z ‘gorącym ziemniakiem’

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Wybór najlepszej ścieżki przez BGP

Route Reflector vs pełna siatka iBGP



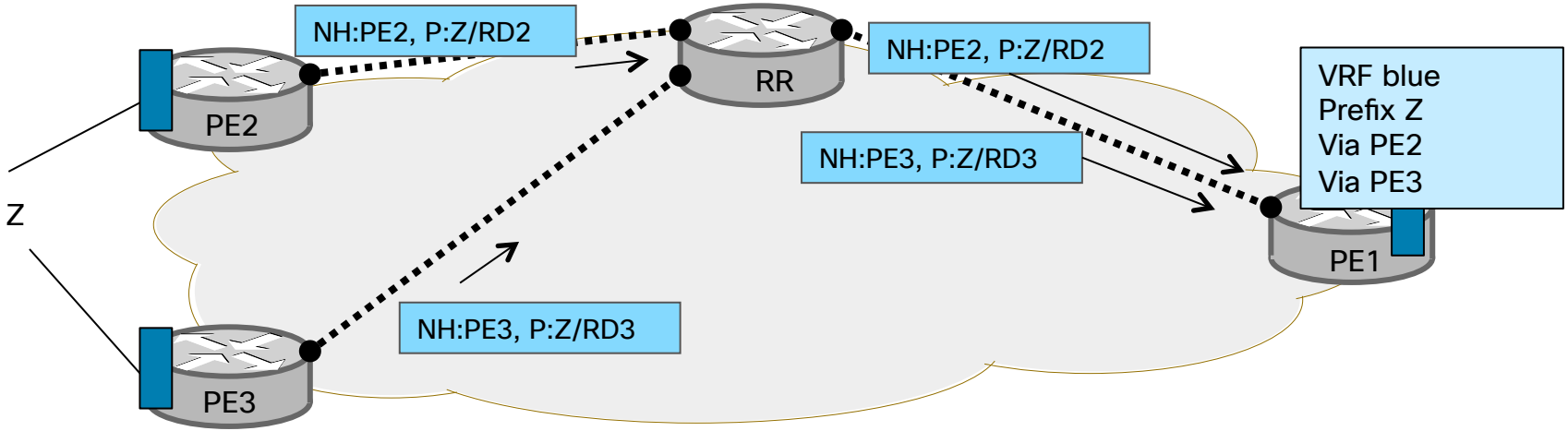
© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Rozwiązania

- Unikalny RD dla MPLS VPN
- BGP Best External
- BGP Add-Path

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

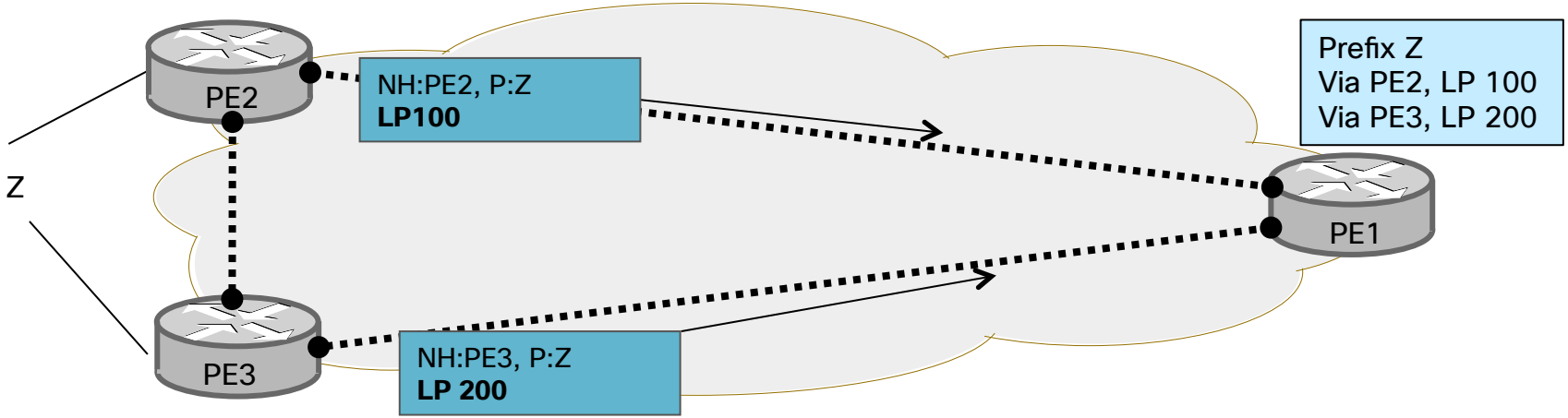
Unikalny RD w MPLS VPN



- Unikalny RD dla VRFów → unikalne NLRI VPNv4/v6
- RR wykonuje algorytm najlepszej ścieżki dla dwóch różnych adresów
- Rekomendowane rozwiązanie dla MPLS-VPN

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

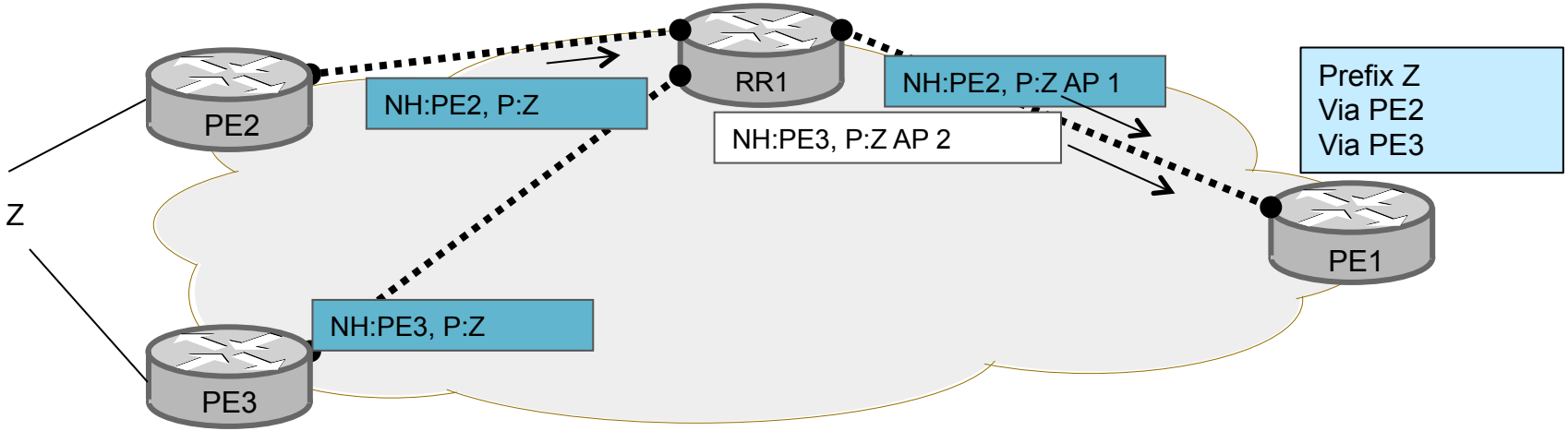
BGP Best External



- Przy wykorzystaniu tej funkcji PE2 nadal rozgłosi do PE1 (lub RR) swoją trasę mimo gorszych atrybutów (których potencjalnie nauczy się od PE3)
- PE1 i PE3 mają po dwie ścieżki

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Add-Path



- Add-Path rozgłosi od 2 do X dodatkowych ścieżek
- Oczywiście router odbierający ścieżki (RR, PE) musi obsługiwać to rozszerzenie

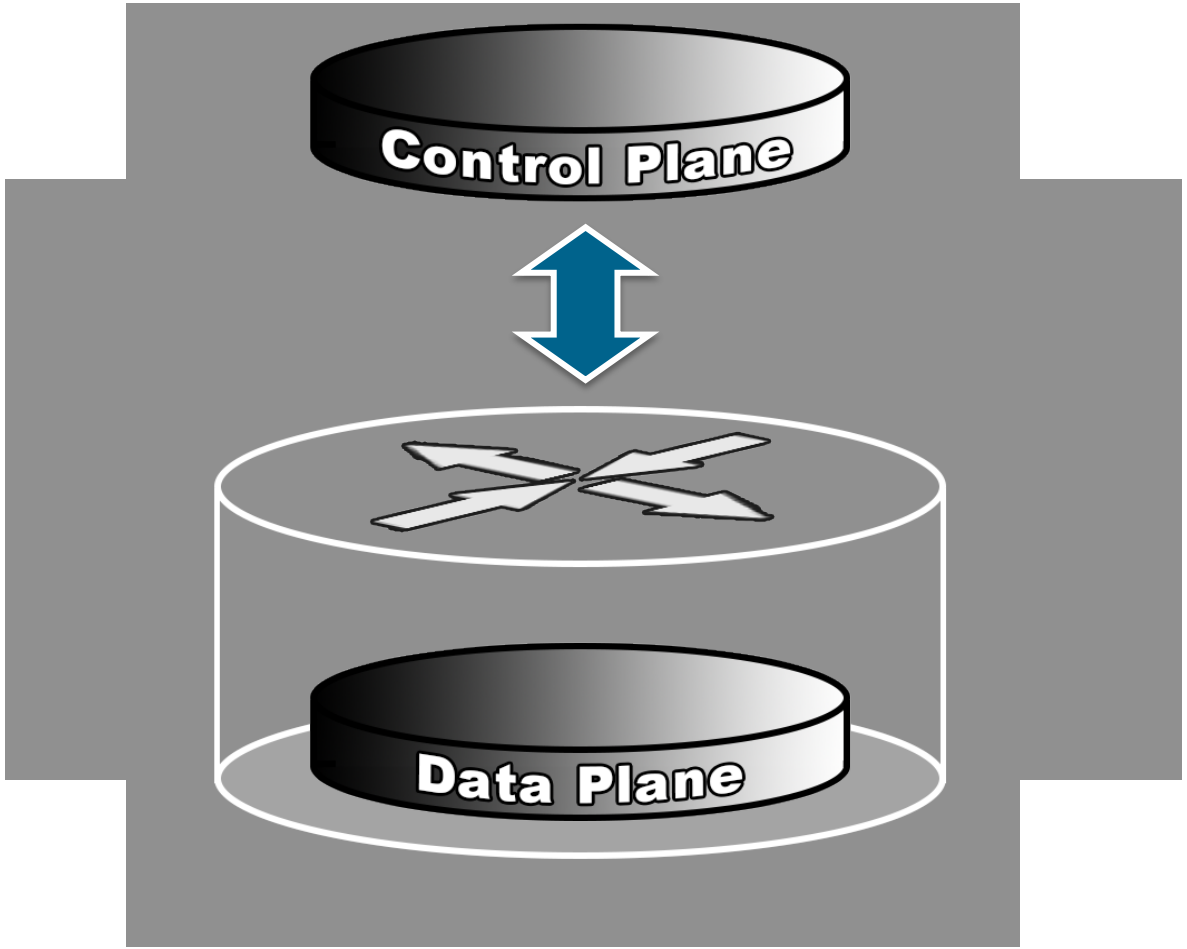
© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Add-Path - rodzaje

- IETF draft definiuje parę różnych rodzajów Add-x-Path:
  - **Add-n-path**: RR wykona obliczenia dla wszystkich tras i wyśle N najlepszych do PE
    - Przykładowe zastosowanie: trasa główna + N tras zapasowych
  - **Add-all-path**: RR wykona obliczenie tylko dla pierwszej ścieżki, ale i tak wyśle komplet do PE
    - Przykładowe zastosowanie: routing ‘hot potato’ mimo obecności RR, ECMP load-balancing w DC
  - **Add-all-multipath+backup**: RR wykona obliczenie dla najlepszej ścieżki, wyśle wszystkie trasy o tym samym koszcie i dodatkowo jedną zapasową do PE
    - Przykadowe zastosowanie: ECMP load-balancing w DC

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

SDN, BGP i kolorowe jednorożce



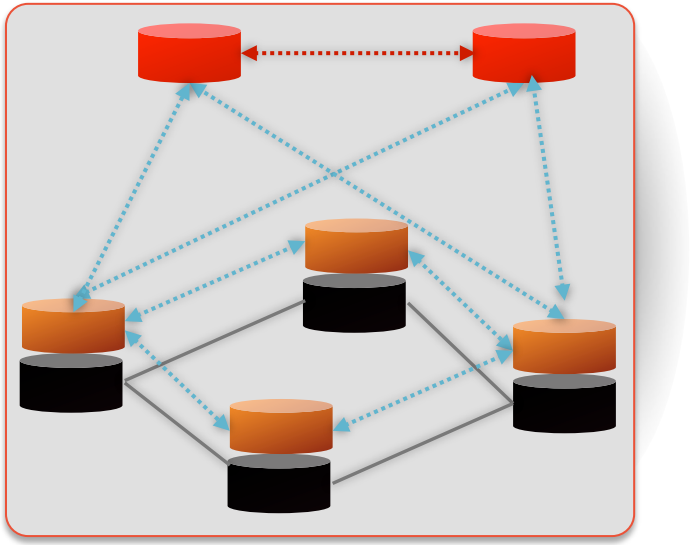
“SDN to separacja control od data plane”

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

SDN – kompromis ekstremów

(czyli podejście hybrydowe)

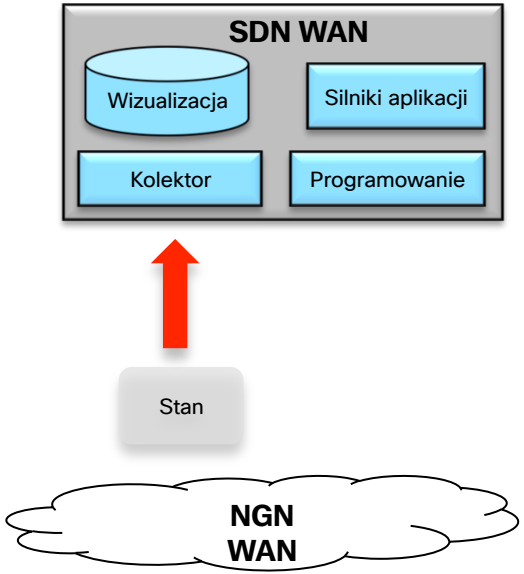
- 20 lat inwestycji w rozproszoną warstwę kontroli – CapEx, umiejętności, doświadczenie, ekspertyza – producentów i klientów
- “Wygrzane” oprogramowanie – sprawdzone w najdziwniejszych sytuacjach i awariach
- Jak pokazuje praktyka, SDNowa warstwa kontroli lepiej sprawdza się w obsłudze wyjątków niż całościowo ruchu (choć to kwestia scenariusza i podejścia)



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Przykład – optymalizacja sieci WAN

- Kontroler SDN chcący zoptymalizować ruch w sieci WAN (lub np. sieci szkieletowej) wymaga “widoczności” ruchu w sieci, w szczególności:
  - Topologii
    - IGP (OSPF, ISIS)
    - BGP (Link State)
  - Pojemność/obciążenie
    - SNMP, NetFlow, NetConf/Yang
    - IS-IS, OSPF, rozszerzone atrybuty TE dla BGP LS (w przyszłości)

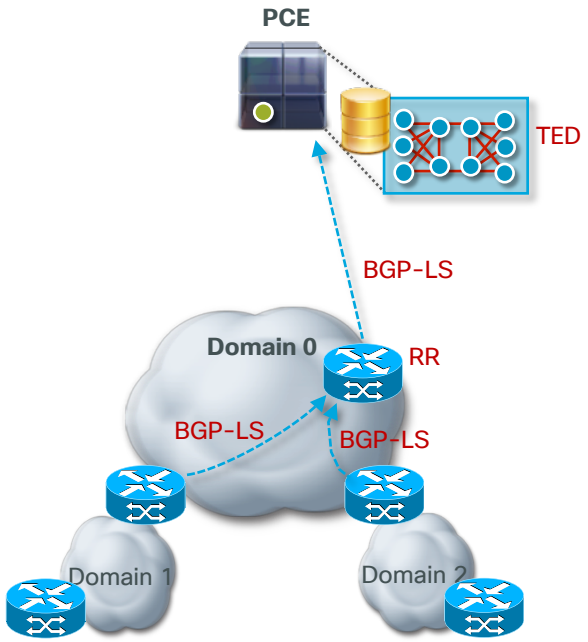


© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Link State

Perspektywa z 10,000 metrów

- BGP może zostać użyty do rozgłaszania stanu łącz oraz TE sieci (BGP LS)
- Nowa rodzina protokołów w konfiguracji BGP – ale brak innych zmian dla całości sieci
- BGP-LS rozgłasza informacje przechowywane w bazach OSPF i IS-IS – topologii (LSDB) oraz TE (TED)



© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP Link State

Przykład konfiguracji

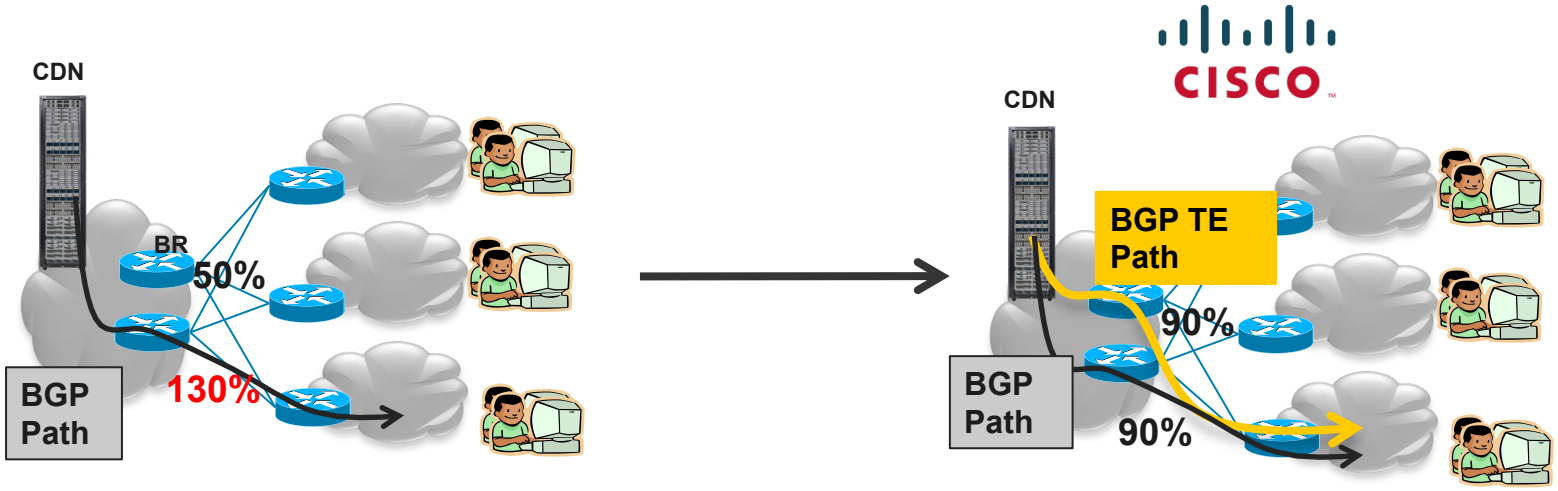
```
router isis DEFAULT
 is-type level-2-only
 net 49.0000.1720.1625.5001.00
 distribute bgp-ls level 2
 address-family ipv4 unicast
  metric-style wide
  mpls traffic-eng level-2-only
  mpls traffic-eng router-id Loopback0
 !
 […]
  !
 !
!
router bgp 65172
 address-family link-state link-state
 !
 neighbor 172.31.0.1
  description Controller
  remote-as 65172
  update-source Loopback0
  address-family link-state link-state
  !
 !
!
```

Dystrybucja informacji z ISIS L2 do BGP-LS

Nowa rodzina protokołów – i konfiguracja sąsiada

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Zastosowanie BGP LS do inżynierii ruchowej



Peering BGP

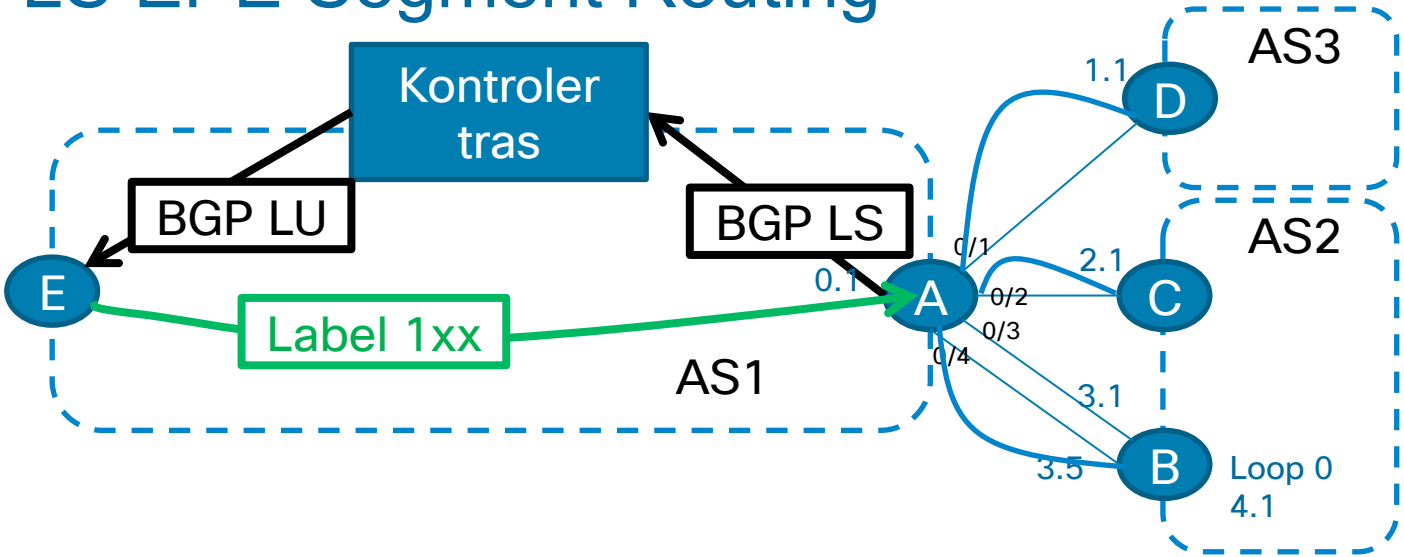
- Ruch podąża najlepszą ścieżką dla BRP

Inżynieria ruchowa dla ruchu wyjściowego

- Architektura oparta o połączenie Segment Routingu i informacji z BGP-LS

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

BGP LS EPE Segment Routing



Interfaces	PeerAdj SID	PeerNode SID	PeerSet SID
Eth0/1 (1.0)	101	111	121
Eth0/2 (2.0)	102	112	122
Eth0/3 (3.0)	103	113	
Eth0/4 (3.4)	104		

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Optymalizacja BGP na brzegu

“Pamięć ma ograniczoną pojemność” (aksjomat #1)

(w szczególności szybka pamięć w sprzęcie sieciowym)

- RIB – przechowuje pełną wiedzę routera o routingu
- FIB – przechowuje wersję RIB zoptymalizowaną do przekazywania ruchu
  - hierarchiczny FIB (pola next-hop)
- Różne kategorie sprzętu sieciowego mają różne wielkości pamięci dla RIB i FIB
  - małe przełączniki, w tym sprzęt MetroEthernet – 16-32k FIB, ~50-100k RIB
  - routery dostępowe i agregacyjne – 128-512k FIB, ~1-2M RIB
  - routery szkieletowe – 1-2M FIB, 4-16M RIB

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Tzw. “problem” 512k prefiksów IPv4

Domyślne limity FIB na platformach operatorskich

- ASR 9k z kartami Trident (IOS-XR): http://www.cisco.com/c/en/us/support/docs/routers/asr-9000-series-aggregation-services-routers/116999-problem-line-card-00.html
- Catalyst 6500/7600 Sup720 B/C i BXL/CXL - rekomendacje: http://www.cisco.com/c/en/us/support/docs/switches/catalyst-6500-series-switches/117712-problemsolution-cat6500-00.html

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

“Pamięć ma ograniczoną pojemność” (aksjomat #1)

(w szczególności szybka pamięć w sprzęcie sieciowym)

- Jak zapewnić pełną wiedzę o routingu w całej sieci bez ogromnych inwestycji w duże platformy?
  - routing nakładkowy (MPLS, Segment Routing)
  - rozdzielenie RIB i FIB (“Selective Route Download”)
  - optymalizacja (“wirtualna agregacja”* i LISP)
- Programowanie FIBu potrafi być wooooooolne (2-3 minuty? 10?)
- Zatem skoncentrujmy się nad rozdzieleniem RIB i FIB ☺

- https://www.nanog.org/meetings/nanog47/presentations/Monday/Francis_VirtualAgg_N47_Mon.pdf i http://tools.ietf.org/html/rfc6769

© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Public

Selective RIB download

IOS-XR

```
route-policy block-into-rib
 if destination in (...) then
   drop
 else
   pass
 end-if

router bgp 1
 address-family ipv4 unicast
  table-policy block-into-rib
```

IOS/IOS-XE

```
router bgp 1
 address-family ipv4
  table-map block-into-rib filter

route-map block-into-rib deny 10
```

© 2015 Cisco and/or its affiliates. All rights reserved.

Nowości w protokole BGP
Optymalizacja routingu

Łukasz Bromirski
lukasz.bromirski@cisco.com
CCIE #159292 R&S/SP, CCDE #2012::17

PLNOG