



Inżynieria ruchowa w sieciach IP

Jak dobrze i skutecznie zaplanować rozkład ruchu

Łukasz Bromirski
lbromirski@cisco.com

Agenda

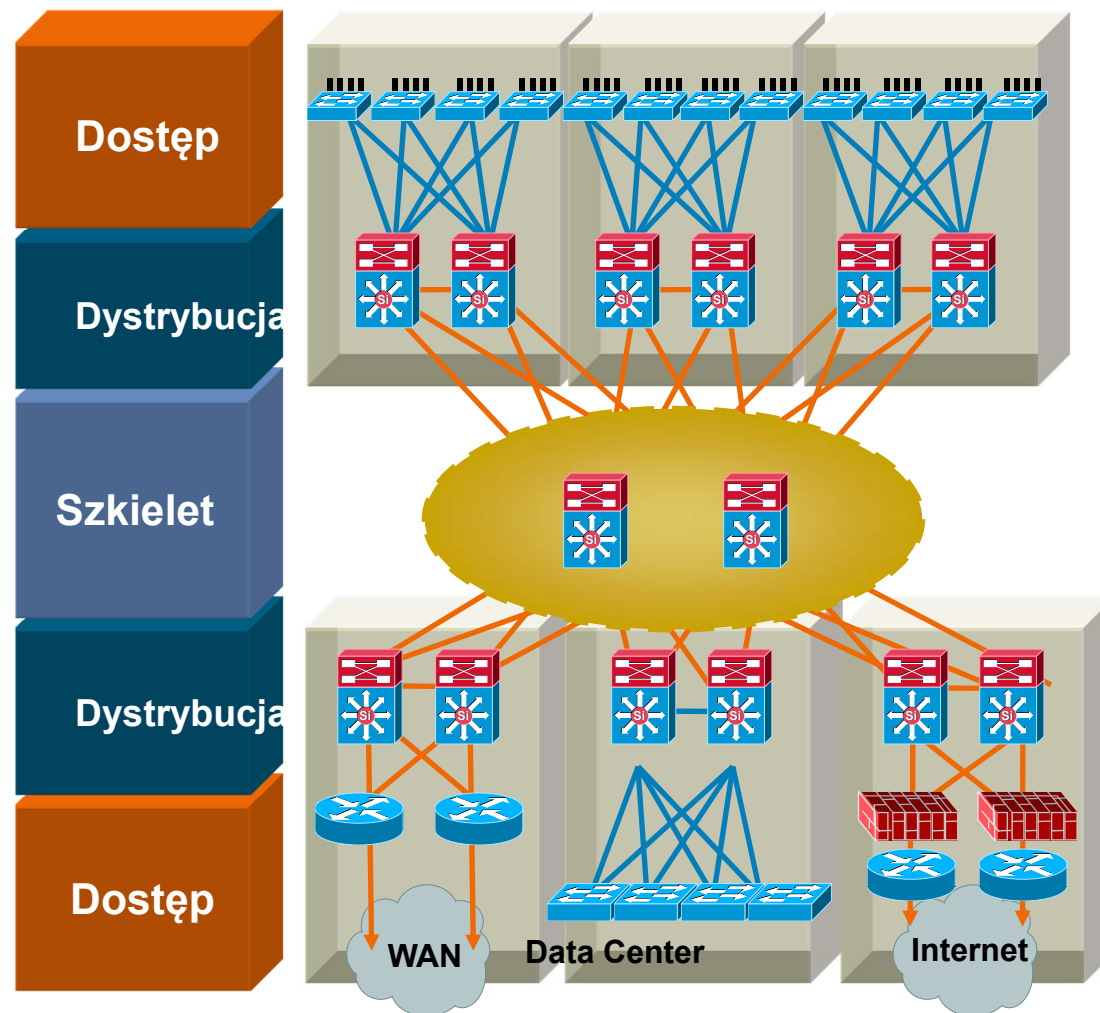
- Inżynieria ruchowa w sieci wewnętrznej
- Inżynieria ruchowa na brzegu sieci własnej i innych
- Q&A



Inżynieria ruchowa w sieci wewnętrznej

Budowa sieci i granica L2/L3

Czy jest mi potrzebny szkielet?

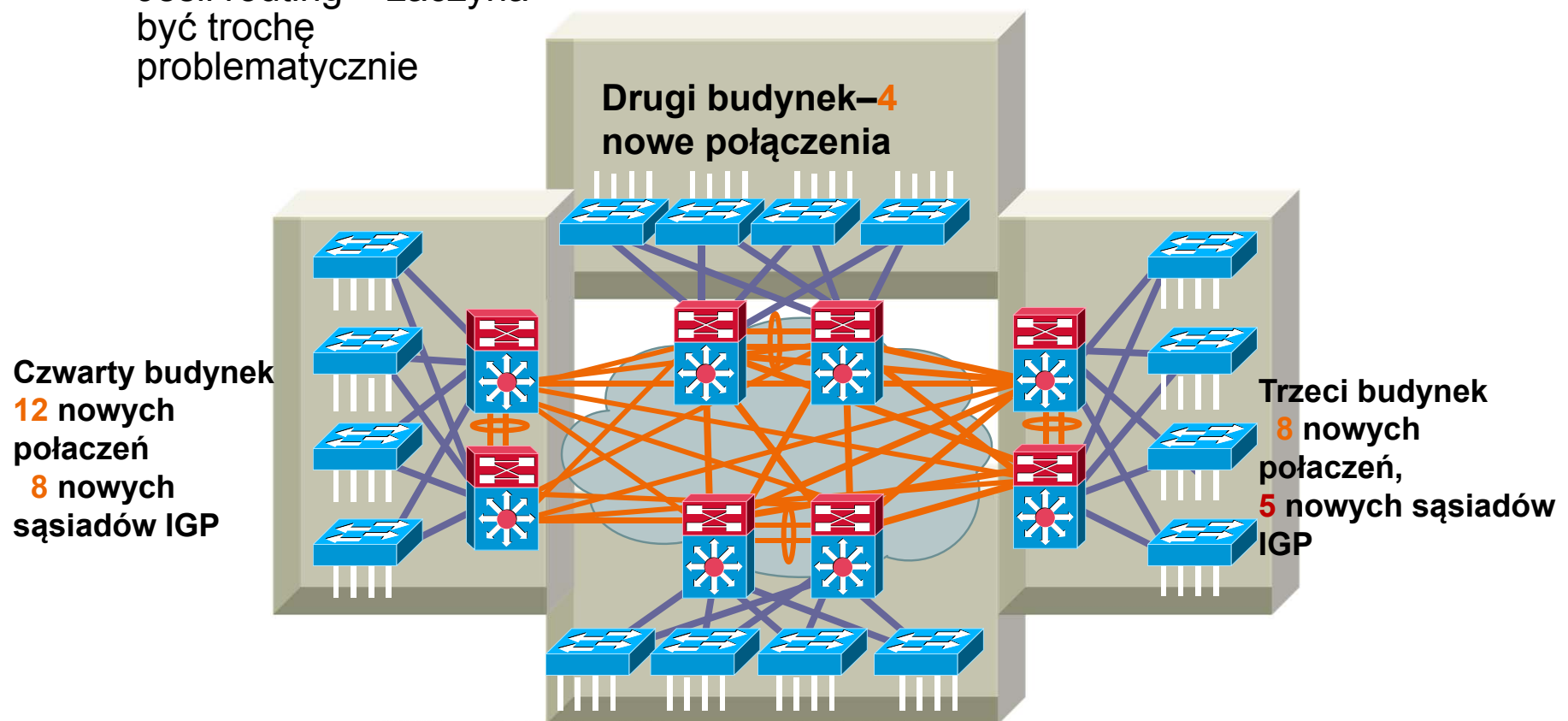


Czy jest mi potrzebny szkielet?

To kwestia skali, złożoności i konwergencji

Brak szkieletu

- Połączenia pomiędzy „warstwami” zwykle w topologii pełnej siatki
- Okablowanie fizyczne nie zawsze daje się tak prowadzić/jest dostępne
- Jeśli routing – zaczyna być trochę problematycznie

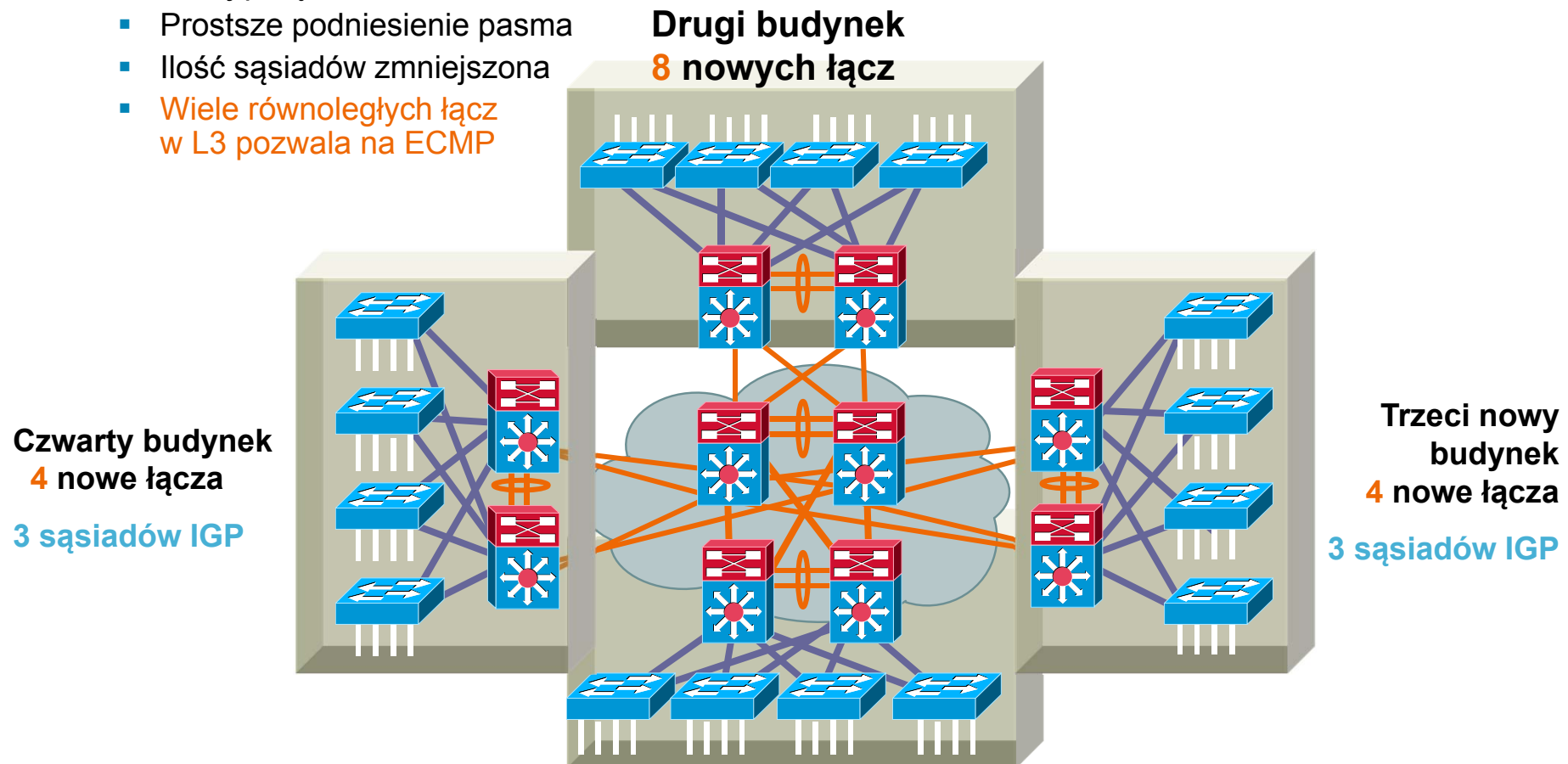


Czy jest mi potrzebny szkielet?

To kwestia skali, złożoności i konwergencji

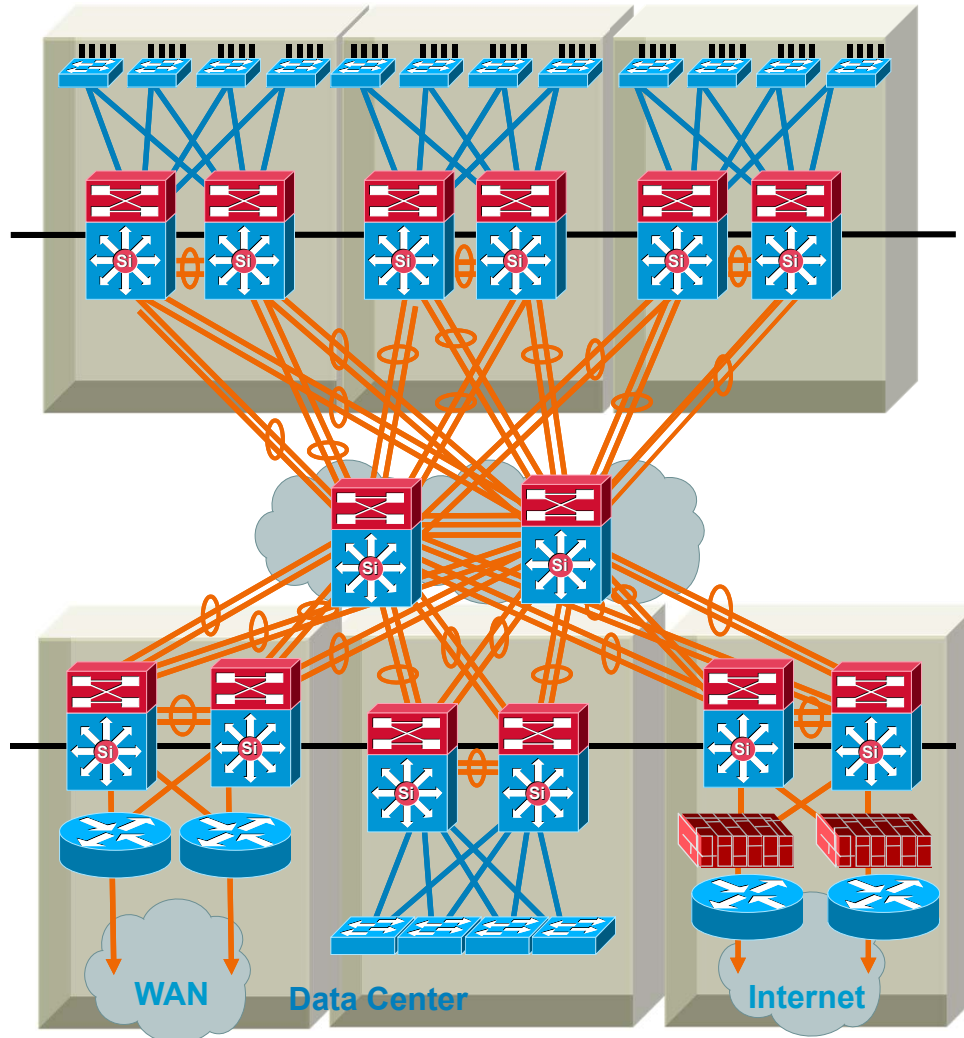
Dedykowane przełączniki szkieletowe

- Łatwiej dodać budynek/sekcję/połączenie
- Mniej połączeń do szkieletu
- Prostsze podniesienie pasma
- Ilość sąsiadów zmniejszona
- **Wiele równoległych łączy w L3 pozwala na ECMP**



Warstwa 2: Etherchannel 1GE

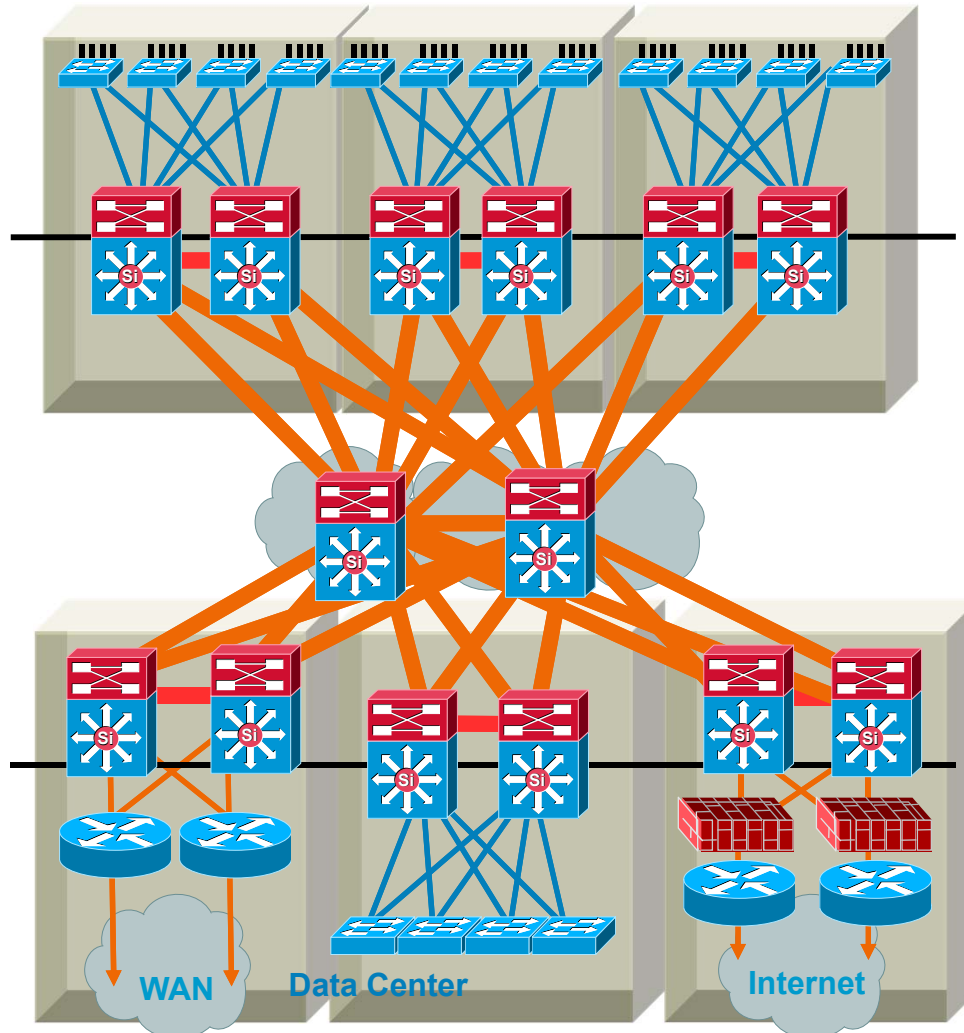
Zmniejsz poziom skomplikowania



- Więcej połączeń = więcej sąsiedztw dla protokołów routingu lub ścieżek L2
- EtherChannel pozwala zestawić jedno połączenie logiczne – w L2 lub L3
- Przepustowość jest sumą aktywnych łączy
- Dodatkowa zaleta przy wykorzystaniu różnych członków stosu 3750/3750E/3750X i kart z przełączników liniowych – awaria jednej karty/przełącznika nie powoduje zmian topologii

Warstwa 2: Etherchannel 10GE

Po co interfejsy 10GE?

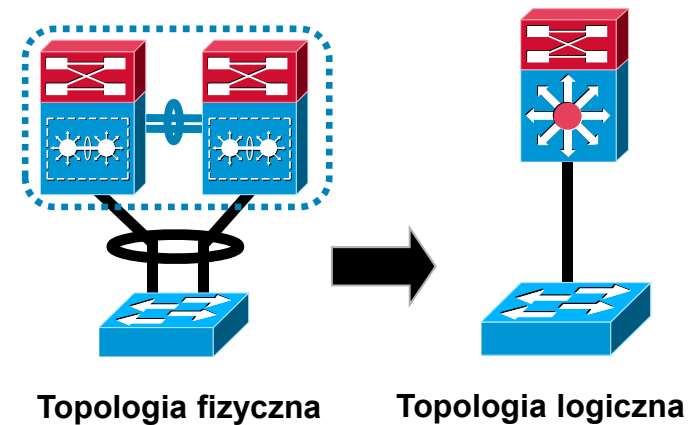


- 8x1GE \neq 1x10GE
- Uproszczona topologia, uproszczone rozwiązywanie problemów
- Potencjalnie uproszczone uszkodzenie topologii 😊

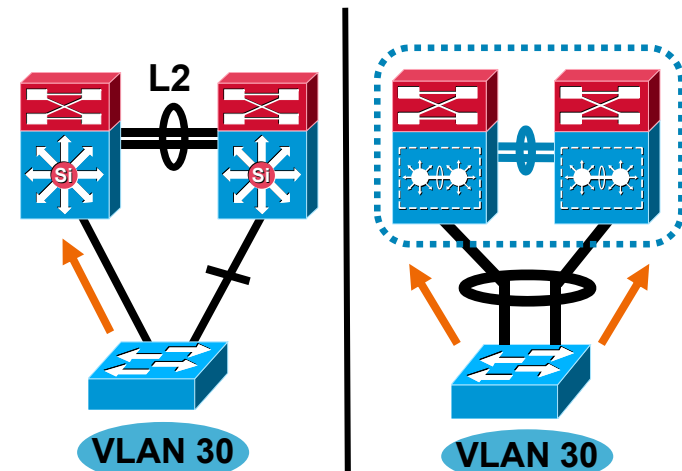
Virtual Switching System

Multichassis EtherChannel (MEC)

- MEC pozwala stworzyć agregację portów należącą do dwóch członków VSS
- Z punktu widzenia przełącznika po drugiej stronie agregacji, domena VSS jest jednym urządzeniem
- Jedno urządzenie – STP nie zamyka jednego z połączeń (uplinków) w ramach unikania pętli



Multichassis EtherChannel



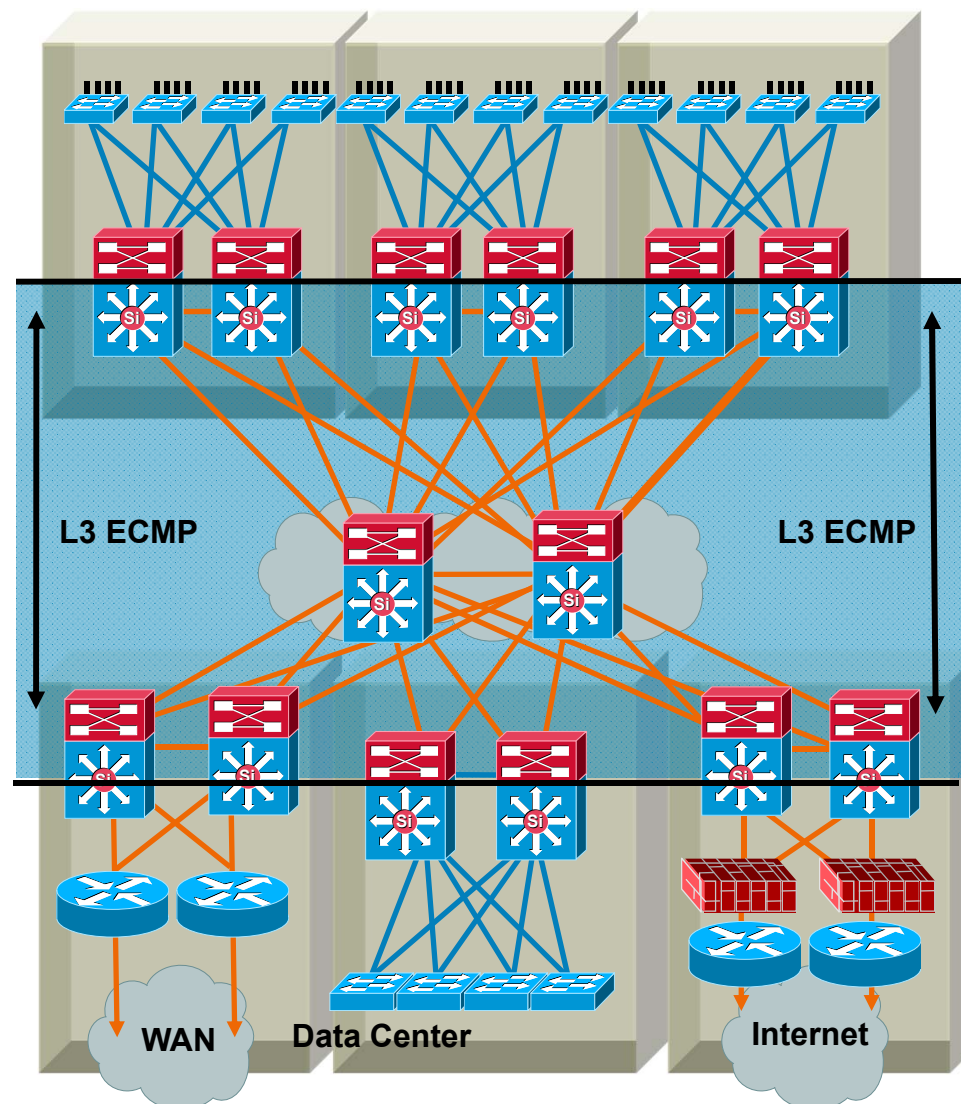
Bez MEC

MEC

Pasmo efektywnie dostępne do wykorzystania przy MEC

Warstwa 3: routing IPv4/IPv6

- Zwykle w połączeniach do szkieletu z dystrybucji, rzadziej bezpośrednio z dostępu
- Wraz z rozpowszechnieniem się routingu IP wykonywanego tak szybko jak switching, zacierą się różnice w czasie konwergencji, ale troubleshooting L3 jest zwykle dużo prostszy niż L2
- Topologia trójkąta jest zwykle lepsza dla przewidywalnej konwergencji niż kwadrat
- Zapewnij redundantne połączenia L3 pomiędzy urządzeniami aby uniknąć „odrzućania” ruchu
- Dostosuj równoważenie obciążenia w ramach CEF w L3/L4 by maksymalnie wykorzystać łącza i uniknąć efektu polaryzacji





Inżynieria ruchowa w sieci wewnętrznej Routing na podstawie bliskości

Proximity routing

Wprowadzenie

- Zapewnianie usług pamięci podręcznej, replikacji i lokalizacji zasobów i usług staje się coraz ważniejsze dla efektywności prowadzenia działalności komercyjnej w internecie

najbliższa kopia filmu, najbliższy serwer

- Różne konteksty treści i usług

usługi zarządzane, aplikacje

sieci P2P

serwery DNS

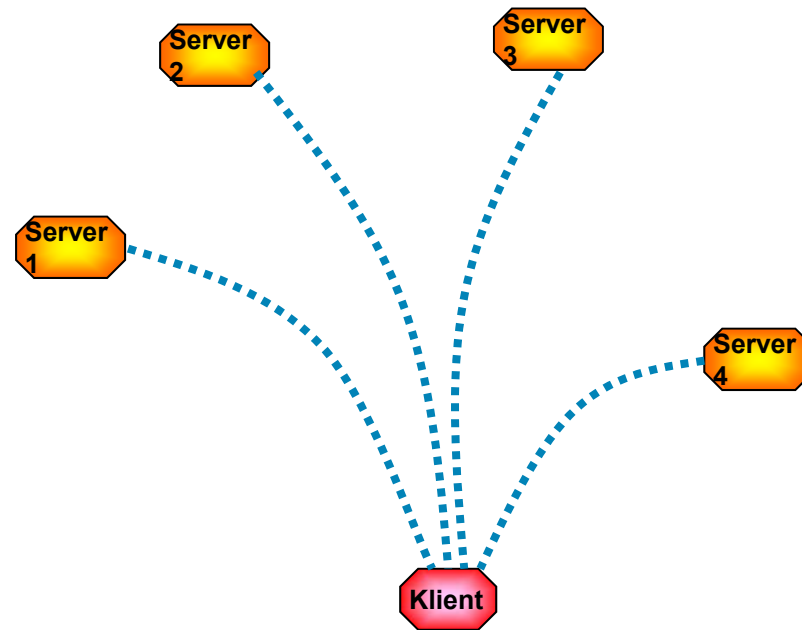
Proximity routing

Dzisiejsze problemy

- Sieciowane aplikacje nie zakładają ograniczeń w miejscu umieszczenia usług lub elementów – w tym elementów samych aplikacji
 - serwery, pamięci cache, hosty, aplikacje sieciowe, (...)
- Zaobserwowano, że aplikacje typu P2P bardzo często przemieszczają ten sam plik (lub jego fragment) wielokrotnie przez to samo łącze, ponieważ mechanizm wyboru ‘peera’ nie bierze pod uwagę topologii sieci
 - ponad 75% ruchu w sieciach ‘osiedlowych’ to ruch P2P
 - gdy prognozy że do 2012 roku video zajmie 90% łącz, sytuacja będzie jeszcze gorsza

Proximity routing

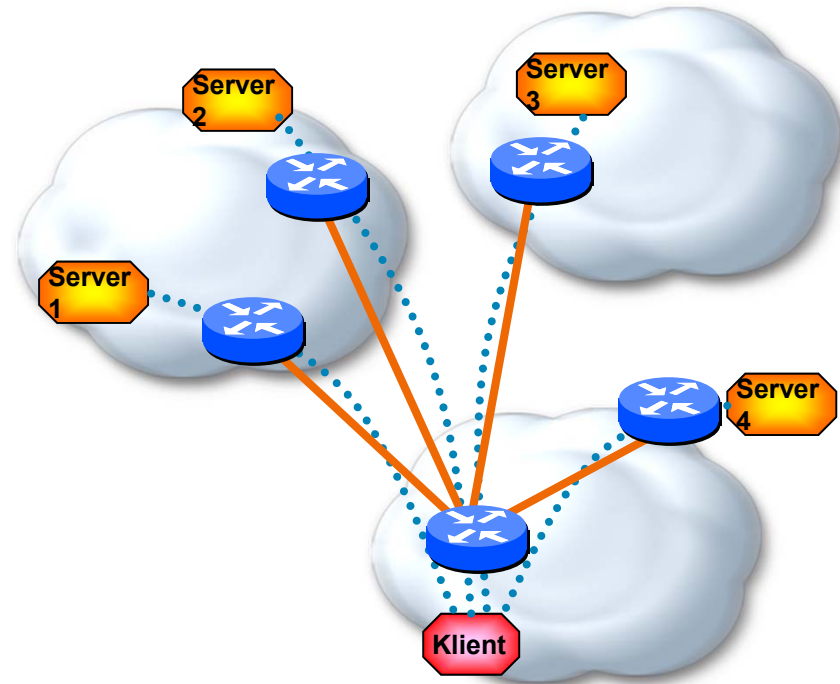
Przykład problemu



- Film „Armageddon 741” dostępny jest z różnych serwerów. Który z nich wybrać? Na podstawie jakich kryteriów?
- Z punktu widzenia aplikacji: minimalne opóźnienie, minimalne obciążenie, maksymalna moc serwerów
- Z punktu widzenia sieci: minimalne zużycie zasobów ‘cennych’, maksymalne tych o możliwie niskim koszcie utrzymania

Proximity routing

Czy mamy narzędzia?



- Pomiar RTT *może* być wskazówką
 - Niezbyt efektywny – bazuje na jednej składowej (np. kiedy różnica jest znacząca – i co to znaczy ‘znacząca’?)
 - Co więcej, RTT jest zmienne
- Widoczność z punktu widzenia routingu pokazuje topologię
 - Kierunek ruchu, zmiany tras, połączenia pomiędzy ASami

Proximity routing

Istniejące rozwiązania

- Oparte o DNS
- Oparte o pomiary RTT
- Oparte o konfigurację i rekonfigurację ręczną (ew. 'oskryptowaną')
- Oparte o pomiary z GPS
- Oparte o wirtualną topologię (tworzone zwykle firmowo i niedostępne powszechnie)

Proximity routing

Jak działa mechanizm?

- Zapytanie o 'bliskość' w swojej ogólnej postaci wygląda następująco:

Kto z poniższych jest najbliżej 192.168.10.1?

192.168.20.44 / 192.168.43.32 / 192.168.65.76

- „Bliskość” może mieć dynamiczne i definiowalne znaczenie:

Koszt routingu?

Informacja o stanie (np. RTT, czy jitter)

Polityka?

Inne?

Proximity routing

Elementy mechanizmu

- Cisco proximity routing zapewnia informacje o lokalizacji – przez system ‘ocen’
- Do wyliczenia ‘bliskości’ wykorzystywana jest znana baza o topologii (baza routingu budowana przez protokoły OSPF, IS-IS i BGP)
- Protokół jest możliwie prosty – zapytanie-odpowiedź
a’la DNS
- ...ale dzięki temu łatwo go rozbudować o dodatkowe usługi
np. pomiary z warstwy aplikacji (IP SLA, serwery aplikacji, etc)

Proximity routing

Warstwa routingu?

- Stabilność wyboru 'bliskości' – dane służące do przekazywania ruchu służą do wyliczenia optymalnych tras
- Informacja zmienia się dynamicznie na podstawie danych z warstwy routingu
- Wykorzystywane są istniejące rozszerzenia, np.:
 - RFC5029 – ISIS Link Attribute
 - RFC5130 – ISIS Route Tags
 - BGP community / extended community
 - draft-ietf-isis-genapp – ISIS Generic Application TLV

Proximity Routing – gdzie jesteśmy?

- Nad zagadnieniem optymalizacji ruchu pracują dwie grupy robocze:
 - IETF Working Group: ALTO (Application Layer Traffic Optimization)
 - IRTF Research Group: P2PRG
- Nowa grupa robocza w IETF zajmująca się proximity routingiem
 - standaryzacja protokołu
- Cisco ma już zaimplementowane dwie części rozwiązania:
 - CDS / CDS/IS
 - SAF



Inżynieria ruchowa na styku sieci własnej i innych

Mechanizmy „od brzegu”

- Stan interfejsów – wykrywanie i „ochrona” przed szybkimi zmianami
- Sesje routingu BGP – czy w ogóle używamy tego protokołu, jeśli tak – jak wykorzystać go maksymalnie?
- Pozostałe problemy projektowe to zwykle adresacja:

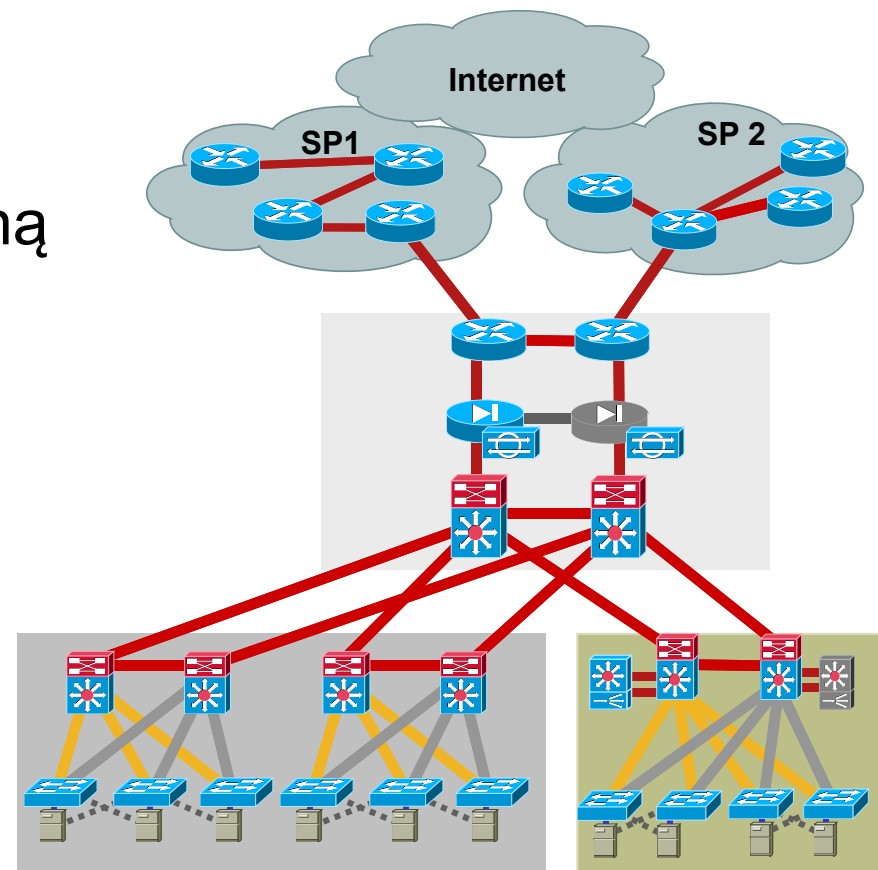
adresacja publiczna zależna od operatora (PA) – usługi widoczne na świecie mogą mieć różne IP, strefy DNS powinny zawierać oba adresy, z punktu widzenia ruchu do Internetu – nie ma to dużego znaczenia (choć może mieć – np. VPNy)

adresacja publiczna niezależna od operatora (PI) – usługi na widoczne na świecie mogą pochodzić ze stałego bloku adresów identyfikowanego jednoznacznie z naszą firmą

RIPE pozwala na multihoming IPv6:
<http://www.ripe.net/ripe/docs/ripe-466.html#PIAssignments>

Jak wygląda brzeg sieci?

- N dostawców internetowych
- Redundantne routery internetowe stanowią fizyczną granicę sieci
- Redundantna warstwa bezpieczeństwa pozwala na bezpieczne dołączenie kolejnych bloków DMZ
- Połączenie do szkieletu zapewnia możliwość rozbudowy bez zmian w infrastrukturze





Inżynieria ruchowa na styku sieci własnej i innych

Mechanizmy podstawowe

Podstawowe założenia

- Rozkładanie ruchu per-pakiet lub per-cel
- Routing domyślny przez wiele interfejsów
- Policy Based Routing
- IP SLA i śledzenie obiektów
- Performance Routing (Cisco PfR)

Konfiguracja interfejsu L3

- Polecenia fizycznego interfejsu kontrolują w jaki sposób mechanizmy routingu będą mogły wykonywać równoważenie ruchu:

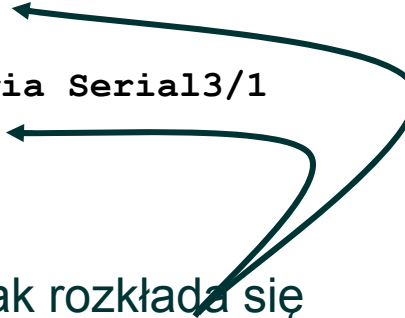
```
router (config-if) # ip load-sharing per-packet  
router (config-if) # ip load-sharing per-destination
```

- Ma sens jeśli wszystkie interfejsy przez które odbywać może się ruch do danego IP docelowego skonfigurowane są tak samo, oraz klasa adresowa używana jako źródłowa akceptowana jest przez operatora nadrzędnego przez wszystkie interfejsy (uRPF!)

Konfiguracja interfejsu L3

Równoważenie ruchu

```
router# sho ip route 192.168.239.0
Routing entry for 192.168.239.0/24
  Known via "eigrp 100", distance 170, metric 3072256, type external
  Redistributing via eigrp 100
  Last update from 192.168.245.11 on Serial3/1, 00:18:17 ago
  Routing Descriptor Blocks:
  * 192.168.246.10, from 192.168.246.10, 00:18:17 ago, via Serial3/0
    Route metric is 3072256, traffic share count is 1
    ....
  192.168.245.11, from 192.168.245.11, 00:18:17 ago, via Serial3/1
    Route metric is 3072256, traffic share count is 1
    ....
```



Parametr 'udział w ruchu' określa jak rozkłada się podział przekazywania ruchu do danego prefiksu

Jak jest wyliczany?

Konfiguracja interfejsu L3

Równoważenie ruchu

```
router# sho ip route 192.168.239.0
Routing entry for 192.168.239.0/24
  Known via "eigrp 100", distance 170, metric 3072256, type external
  Redistributing via eigrp 100
  Last update from 192.168.245.11 on Serial3/1, 00:18:17 ago
  Routing Descriptor Blocks:
  * 192.168.246.10, from 192.168.246.10, 00:18:17 ago, via Serial3/0
    Route metric is 3072256, traffic share count is 1
    ....
  192.168.245.11, from 192.168.245.11, 00:18:17 ago, via Serial3/1
    Route metric is 3072256, traffic share count is 1
    ....
```

Metryka każdej z tras dzielona jest przez
najwyższą z metryk dla danego prefiksu

$$3072256/3072256 == 1$$

Wynik jest następnie używany do określenia
udziału w ruchu

Konfiguracja interfejsu L3

Równoważenie ruchu

```
router# sho ip route 192.168.239.0
Routing entry for 192.168.239.0/24
  Known via "eigrp 100", distance 170, metric 3072256, type external
  Redistributing via eigrp 100
  Last update from 192.168.245.11 on Serial3/1, 00:18:17 ago
  Routing Descriptor Blocks:
  * 192.168.246.10, from 192.168.246.10, 00:18:17 ago, via Serial3/0
    Route metric is 1536128, traffic share count is 2
    ....
  192.168.245.11, from 192.168.245.11, 00:18:17 ago, via Serial3/1
    Route metric is 3072256, traffic share count is 1
    ....
```

Jeśli metryka będzie mniejsza – udział w ruchu może wzrosnąć (trasa będzie atrakcyjniejsza)

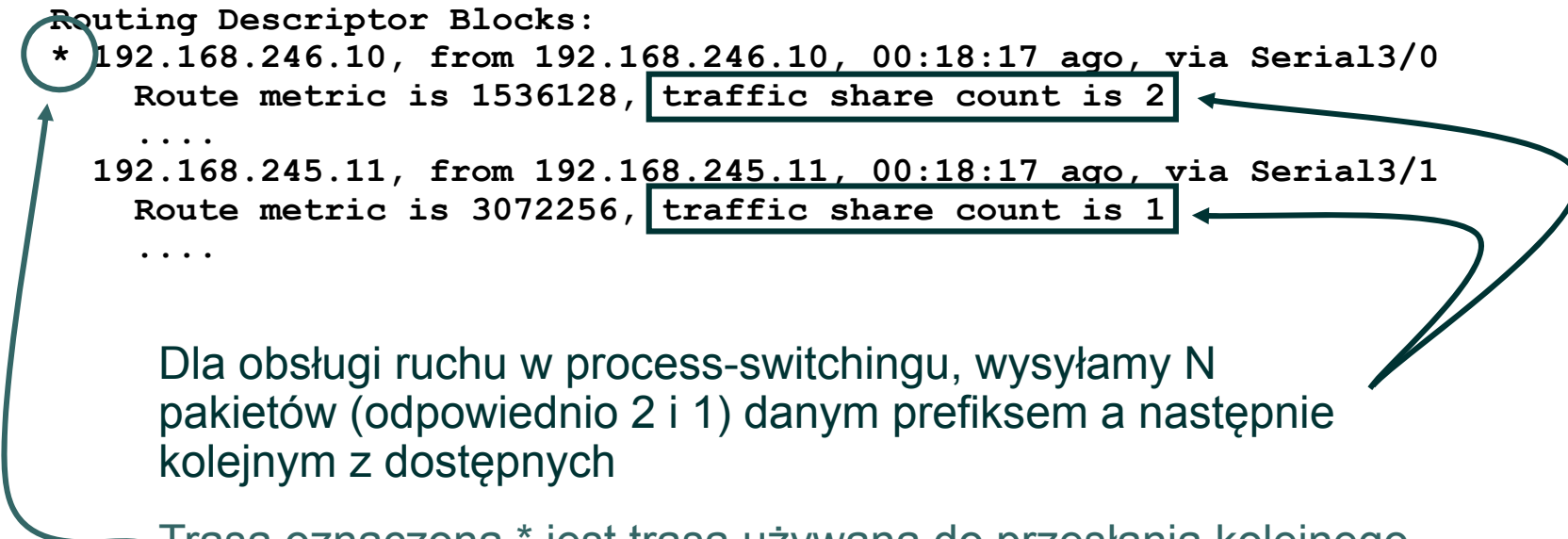
$$3072256/3072256 == 1$$

$$3072256/1536128 == 2$$

Konfiguracja interfejsu L3

Równoważenie ruchu

```
router#sho ip route 192.168.239.0
Routing entry for 192.168.239.0/24
  Known via "eigrp 100", distance 170, metric 3072256, type external
  Redistributing via eigrp 100
  Last update from 192.168.245.11 on Serial3/1, 00:18:17 ago
  Routing Descriptor Blocks:
  * 192.168.246.10, from 192.168.246.10, 00:18:17 ago, via Serial3/0
    Route metric is 1536128, traffic share count is 2
    ....
  192.168.245.11, from 192.168.245.11, 00:18:17 ago, via Serial3/1
    Route metric is 3072256, traffic share count is 1
    ....
```



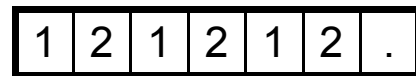
Dla obsługi ruchu w process-switchingu, wysyłamy N pakietów (odpowiednio 2 i 1) danym prefiksem a następnie kolejnym z dostępnych

Trasa oznaczona * jest trasą używaną do przesłania kolejnego pakietu

Konfiguracja interfejsu L3

Równoważenie ruchu

```
router#sho ip route 192.168.239.0
Routing entry for 192.168.239.0/24
  Known via "eigrp 100", distance 170, metric 3072256, type external
  Routing Descriptor Blocks:
    * 192.168.246.10, from 192.168.246.10, 00:18:17 ago, via Serial3/0
      Route metric is 1536128, traffic share count is 2
      ....
    192.168.245.11, from 192.168.245.11, 00:18:17 ago, via Serial3/1
      Route metric is 3072256, traffic share count is 1
      ....
```



Tablica ECMP

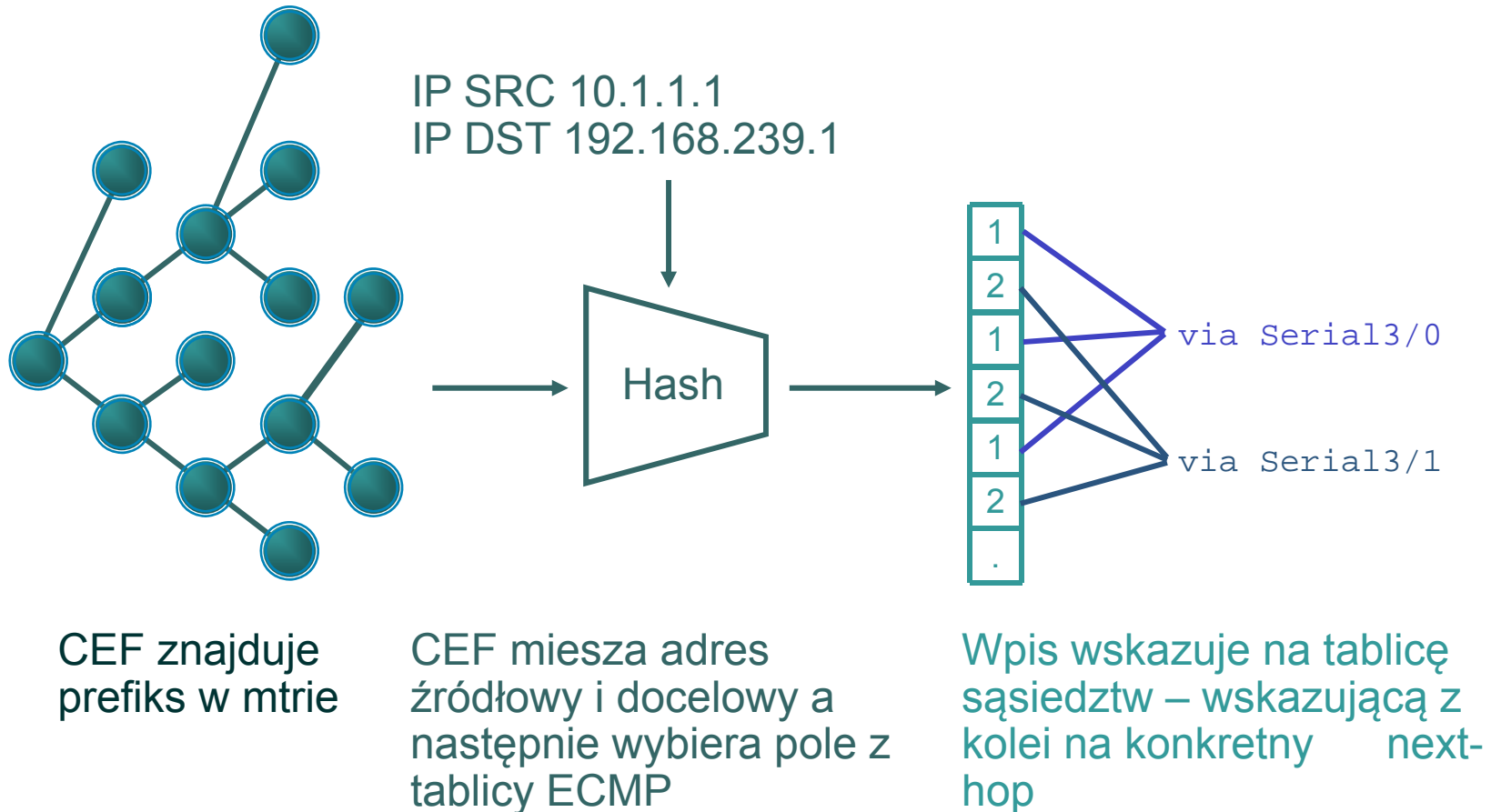
CEF używa wartości z pól do wypełnienia tablicy ECMP

Każdy z prefiksów jest wstawiony odpowiednią ilość razy w zależności od wartości udziału w ruchu – aż do wypełnienia tablicy

Rozmiar tablicy zależy od platformy

Konfiguracja interfejsu L3

Równoważenie ruchu



```
router# show ip cef exact-route 10.1.1.1 192.168.239.1
10.1.1.1          -> 192.168.239.1      : Serial3/0 (next hop 192.168.246.10)
```

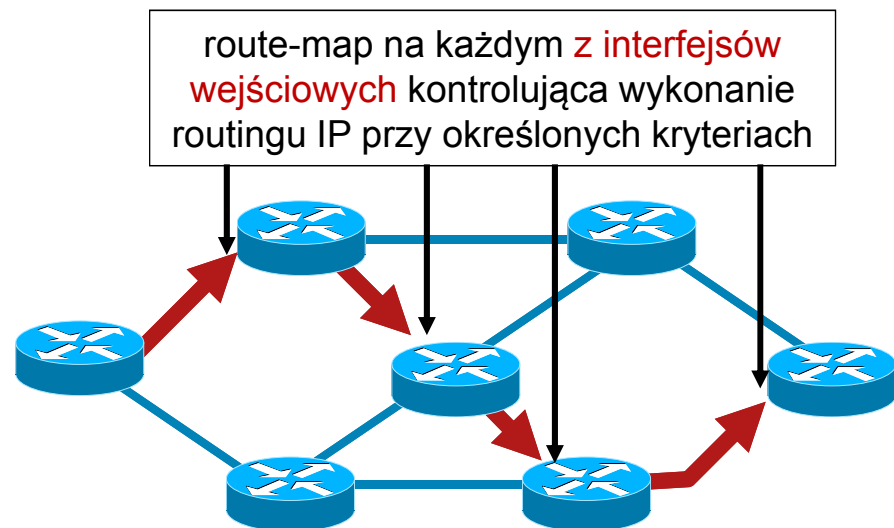
Policy-Based Routing

- Po co używać PBR?

routing przez konkretny interfejs

ustawienie pól QoS w ruchu

routing na podstawie adresu
źródłowego



Policy-Based Routing

Konfiguracja

- route-map jest konstrukcją sekwencyjną – podobnie jak ACL:

```
router(config)# route-map rm-fe00 permit 10
router(config-route-map)# [...]
router(config)# route-map rm-fe00 deny 20
router(config-route-map)# [...]
router(config)# route-map rm-fe00 permit 30
router(config-route-map)# [...]
```

- Ruch pasujący do reguł match wpisu z 'deny' jest ignorowany
- Ruch który dotrze do końca route-mapy jest ignorowany i podlega normalnym regułom routingu

Policy-Based Routing

Konfiguracja – kryteria klasyfikacji

```
router(config)# route-map rm-fe00 permit 10
router(config-route-map)# match ?
  as-path           Match BGP AS path list
  community         Match BGP community list
  extcommunity      Match BGP/VPN extended community list
  interface         Match first hop interface of route
  ip                IP specific information
  ipv6              IPv6 specific information
  length            Packet length
  local-preference  Local preference for route
  metric            Match metric of route
  [...]
  route-type        Match route-type of route
  source-protocol   Match source-protocol of route
  tag               Match tag of route
```

Policy-Based Routing

Konfiguracja – dodatkowe opcje

```
router(config)# route-map rm-fe00 permit 10
router(config-route-map)# match ip ?
    address          Match address of route or match packet
    next-hop         Match next-hop address of route
    route-source     Match advertising source address of route

router(config-route-map)# match length ?
    <0-2147483647>   Minimum packet length
router(config-route-map)# match length 50 ?
    <0-2147483647>   Maximum packet length
router(config-route-map)# match length 50 100
```

Policy-Based Routing

Konfiguracja – możliwość oznaczenia ruchu

```
router(config)# route-map rm-fe00 permit 10
```

```
router(config-route-map)# set ?
```

as-path	Prepend string for a BGP AS-path
automatic-tag	Automatically compute TAG value
community	BGP community attribute
dampening	Set BGP route flap dampening params
default	Set default information
extcommunity	BGP extended community attribute
interface	Output interface
ip	IP specific information
ipv6	IPv6 specific information
level	Where to import route
local-preference	BGP local preference path attribute
vrf	Define VRF name
weight	BGP weight for routing table

Policy-Based Routing

Konfiguracja – przypisanie do interfejsu

```
router(config)# interface fastEthernet 0/1
router(config-if)# ip policy route-map rm-fe00
```

```
R2# sh ip interface fastEthernet 0/1
FastEthernet0/1 is up, line protocol is up
Internet address is 99.99.99.1/24
Broadcast address is 255.255.255.255
[...]
IP fast switching is enabled
IP fast switching on the same interface is disabled
IP Flow switching is disabled
IP CEF switching is enabled
IP CEF Feature Fast switching turbo vector
[...]
Policy routing is enabled, using route map rm-fe00
```

Enhanced Object Tracking

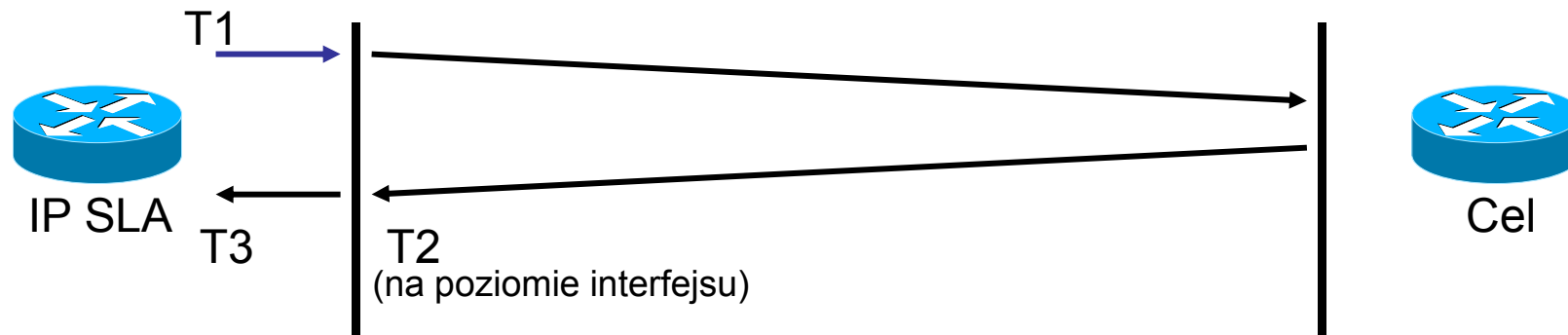
- Mechanizm „śledzenia” osiągalności/dostępności zasobów w oparciu o mechanizm IP SLA i inne zdarzenia dostępne w Cisco IOS

Opcja	Składnia
Status protokołu	track object-number interface type number line-protocol track 1 interface serial 1/1 line-protocol
Status routingu IP (+IPCP)	track object-number interface type number ip routing track 2 interface ethernet 1/0 ip routing
Osiągalność adresu	track object-number ip route IP-Addr/Prefix-len reachability track 3 ip route 10.16.0.0/16 reachability
Waga (0-255) trasy protokołu	track object-number ip route IP-Addr/Prefix-len metric threshold track 4 ip route 10.16.0.0/16 metric threshold
Operacja IP SLA	track object-number ip sla type number state track 5 ip sla 4 state
Osiągalność hostów IP SLA	track object-number ip sla type number reachability track 6 ip sla 4 reachability

IP SLA – wsparcie dla protokołów

- **UDP** echo, jitter i jitter dla VoIP
- **ICMP** echo, path echo, path jitter
- **RTP** (Real-Time Transport Protocol)
- **VoIP** – zwłoka w zarejestrowaniu do gatekeepera
- **VoIP** – zwłoka w zestawieniu połączenia
- **HTTP, FTP, DNS, DHCP**
- **TCP**
- **DLSw+**
- **MPLS** – badanie całej LSP
- **OAM** w sieciach ME

ICMP Echo (pomiar)



Czas obsługi zdarzenia na routerze wysyłającym: $T_{proc} = T3 - T2$

RTT:

$$T = T3 - T1 - T_{proc}$$

$$T = T3 - T1 - (T3 - T2)$$

$$T = T3 - T1 - T3 + T2$$

$$T = T2 - T1$$

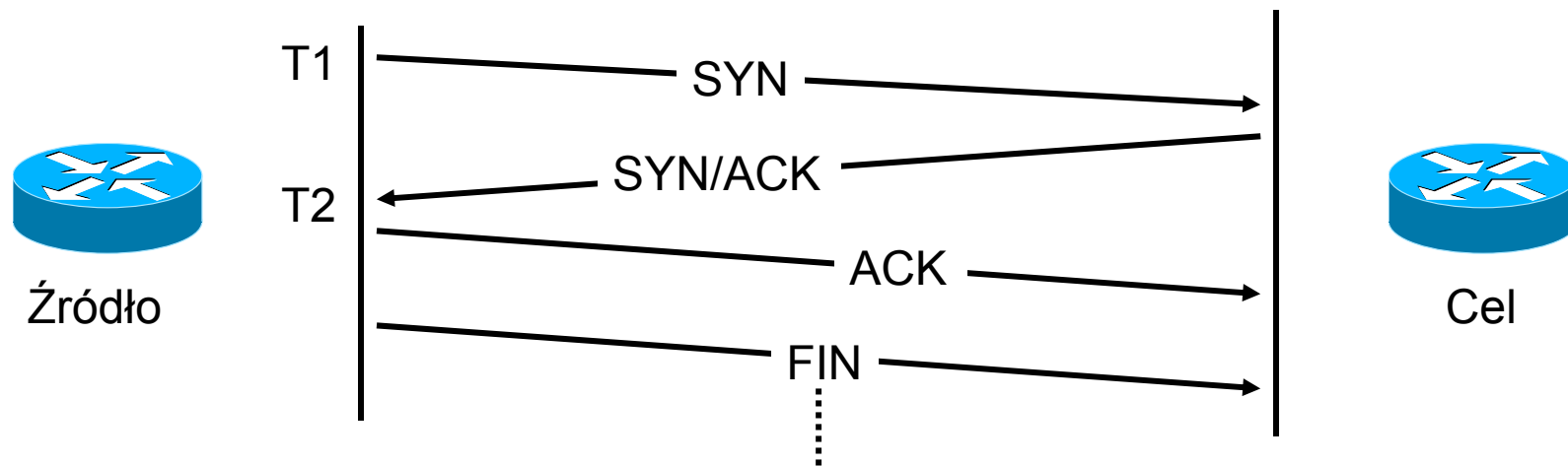
ICMP Echo (przykład)

```
ip sla monitor 2
  type echo protocol ipIcmpEcho 10.32.130.2
  tos 32
  frequency 120
ip sla monitor schedule 2 life forever start-time now
```

```
router# show ip sla stat 2
```

```
Round Trip Time (RTT) for          Index 2
      Latest RTT: 100 ms
Latest operation start time: *17:32:53.315 CET Tue
Feb 21 2006
Latest operation return code: Timeout
Number of successes: 0
Number of failures: 1
Operation time to live: Forever
```

TCP Connect (pomiar)



Czas mierzony jest jako różnica pomiędzy wysłaniem pakietu z flagą SYN a otrzymaniem pakietu ACK, w tym przypadku = $T2 - T1$

TCP Connect

Przykład

```
ip sla monitor 123  
type tcp-connect 10.52.132.68 9 control disable  
ip sla schedule 123 start-time now
```

Połączenie do 10.52.132.68 na port 9

Jeśli na hoście nie pracuje IP SLA,
należy wyłączyć protokół kontrolny

TCP Connect

Dostępne statystyki

```
router# sh ip sla monitor statistics 123 detail
```

```
Round trip time (RTT)      Index 123
```

```
    Latest RTT: 1 ms
```

```
Latest operation start time: 14:20:26.272 CET Mon  
Mar 13 2006
```

```
Latest operation return code: OK
```

```
Over thresholds occurred: FALSE
```

```
Number of successes: 24
```

```
Number of failures: 0
```

```
Operation time to live: Forever
```

```
Operational state of entry: Active
```

```
Last time this entry was reset: Never
```

Wykorzystanie IP SLA w śledzeniu stanu

Definicja próbek śledzących

```
track 10 ip sla 1 reachability
  delay down 10 up 15
track 20 ip sla 2 reachability
  delay down 10 up 15
!
ip sla 1
  icmp-echo 1.1.1.2 source-interface FastEthernet1/0
  timeout 100 ! milisekundach
  frequency 5 ! w sekundach
ip sla schedule 1 life forever start-time now
!
ip sla 2
  icmp-echo 1.1.2.2 source-interface FastEthernet2/0
  timeout 100 ! w milisekundach
  frequency 5 ! w sekundach
ip sla schedule 2 life forever start-time now
```

Wykorzystanie IP SLA w śledzeniu stanu

Przykład 1: wykorzystanie PBR – routing statyczny

```
route-map pbr-ipsla permit 10
  set ip next-hop verify-availability 10.1.1.2 10 track 10
  set ip next-hop verify-availability 10.1.2.2 20 track 20
!
interface FastEthernet0/0
  ip address 192.168.10.1 255.255.255.0
  ip policy route-map pbr-ipsla
```

Wykorzystanie IP SLA w śledzeniu stanu

Przykład 1: wykorzystanie PBR – routing statyczny

- route-map pokaże osiągalność dla każdego ze zdefiniowanych testów

```
pbr#sh route-map
route-map host, permit, sequence 10
  Match clauses:
  Set clauses:
  ip next-hop verify-availability 1.1.1.2 10 track 10      [up]
  ip next-hop verify-availability 1.1.2.2 20 track 20      [down]
Policy routing matches: 4107 packets, 468198 bytes
```

Wykorzystanie IP SLA w śledzeniu stanu

Przykład 2: wykorzystanie routingu

```
ip route 0.0.0.0 0.0.0.0 FastEthernet1/0 1.1.1.1 track 10
ip route 0.0.0.0 0.0.0.0 FastEthernet2/0 1.1.2.2 track 20
```

```
router#sh track
```

```
Track 10
```

```
IP SLA 10 state
```

```
State is Up
```

```
1 change, last change 1d09h
```

```
Delay up 15 secs, down 10 secs
```

```
Latest operation return code: Success
```

```
Tracked by:
```

```
STATIC-IP-ROUTING 0
```

```
Track 20
```

```
IP SLA 20 state
```

```
State is Up
```

```
1 change, last change 1d09h
```

```
Delay up 15 secs, down 10 secs
```

```
Latest operation return code: Success
```

```
Tracked by:
```

```
STATIC-IP-ROUTING 0
```

Wykorzystanie IP SLA w śledzeniu stanu

Przykład 3: wykorzystanie routingu z łączem zapasowym

```
ip route 0.0.0.0 0.0.0.0 FastEthernet1/0 1.1.1.1 track 10
ip route 0.0.0.0 0.0.0.0 Dialer1 250 ! interfejs zapasowy - aktywowany gdy
                                       ! śledzona sesja zostaje wyrzucona z tablicy
                                       ! routingu
```

```
router#sh track
Track 10
  IP SLA 10 state
  State is Up
    1 change, last change 1d09h
  Delay up 15 secs, down 10 secs
  Latest operation return code: Success
  Tracked by:
    STATIC-IP-ROUTING 0
```

Wykorzystanie IP SLA w śledzeniu stanu

Przykład 3: wykorzystanie routingu z łączem zapasowym

- A co z NAT po przełączeniu łącza?
- EEM i route-map w NAT

```
ip nat inside source route-map nat-1 interface FastEthernet1/0 overload
ip nat inside source route-map nat-2 interface Dialer1 overload
```

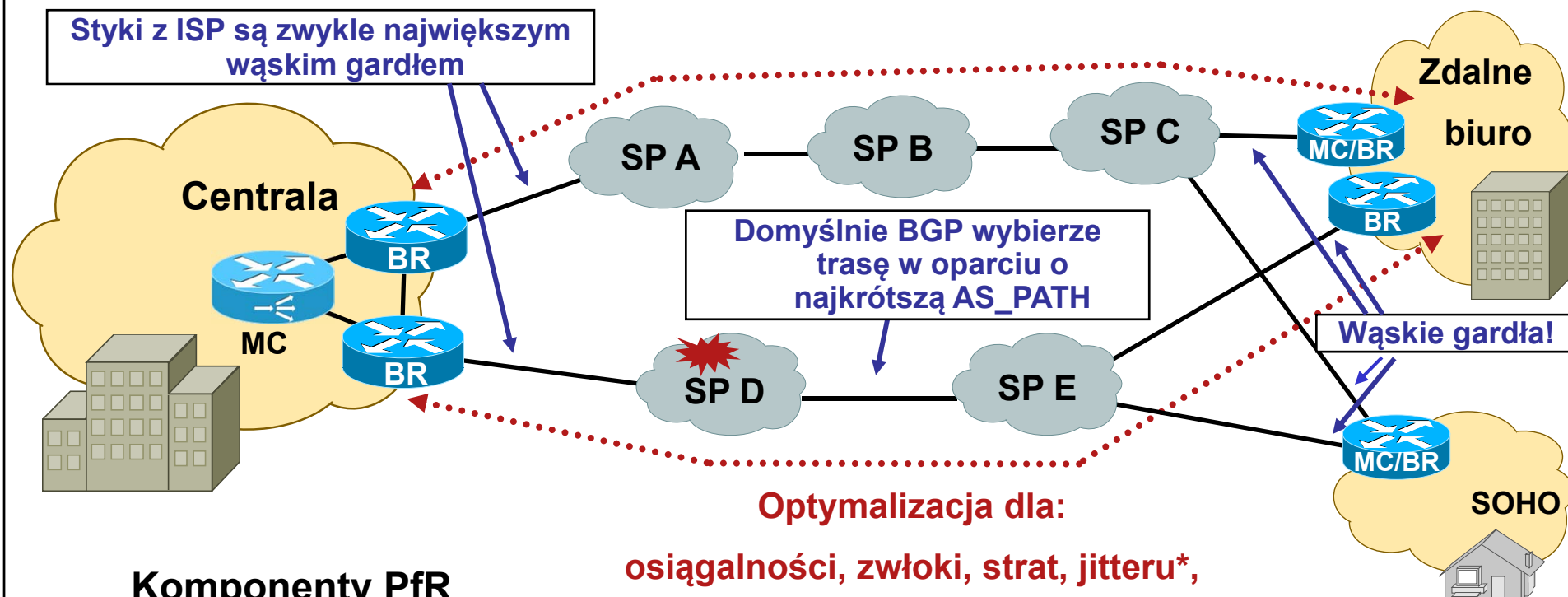
```
route-map nat-1 permit 10
  match ip address acl-nat
  match interface FastEthernet1/0
```

```
route-map nat-2 permit 10
  match ip address acl-nat
  match interface Dialer1
```

```
event manager applet TRACK_INTERNET
  event syslog pattern "%TRACKING-5-STATE"
  action 1.0 cli command "enable"
  action 1.1 cli command "clear ip nat translations *"
  action 1.2 syslog msg „Wyczyszczono sesje NAT"
```

Cisco Performance Routing (PfR)

- Dynamiczna **optymalizacja tras** per prefix dla dwóch lub większej ilości tras równoległych

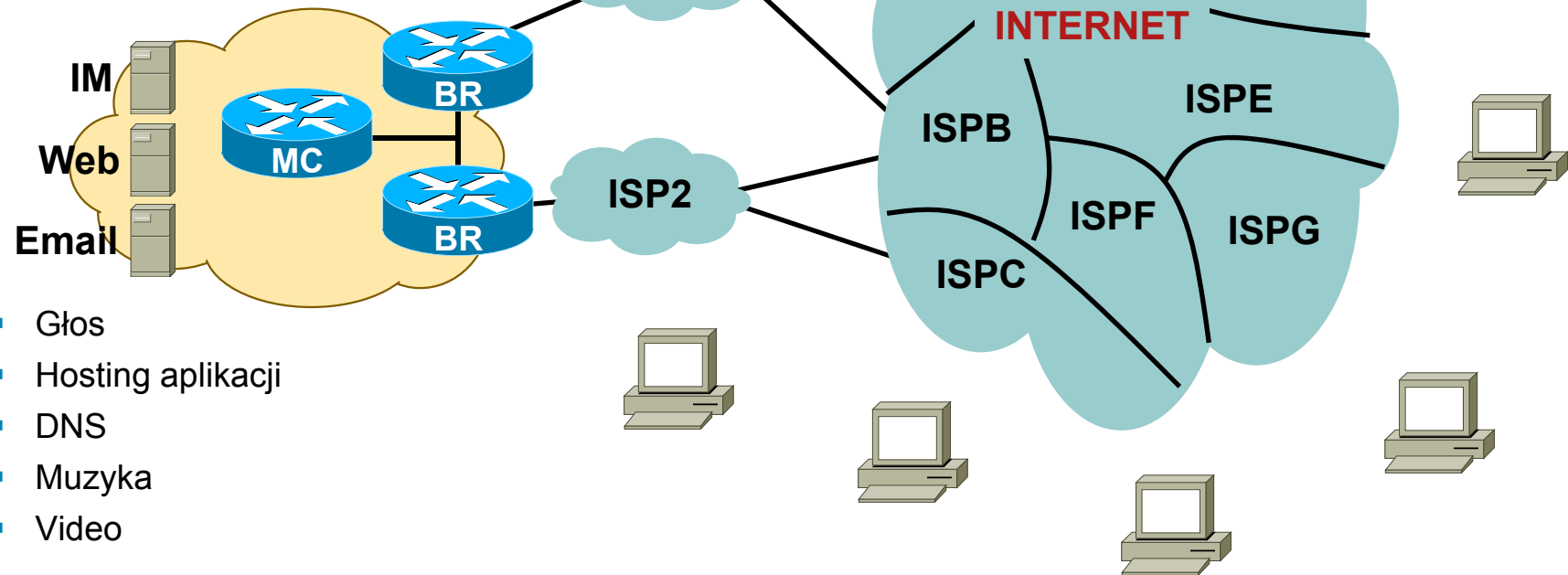


Komponenty PfR

- BR—Border Router
- MC—Master Controller (podejmuje decyzje)

Zakres usług którym pomoc może PfR

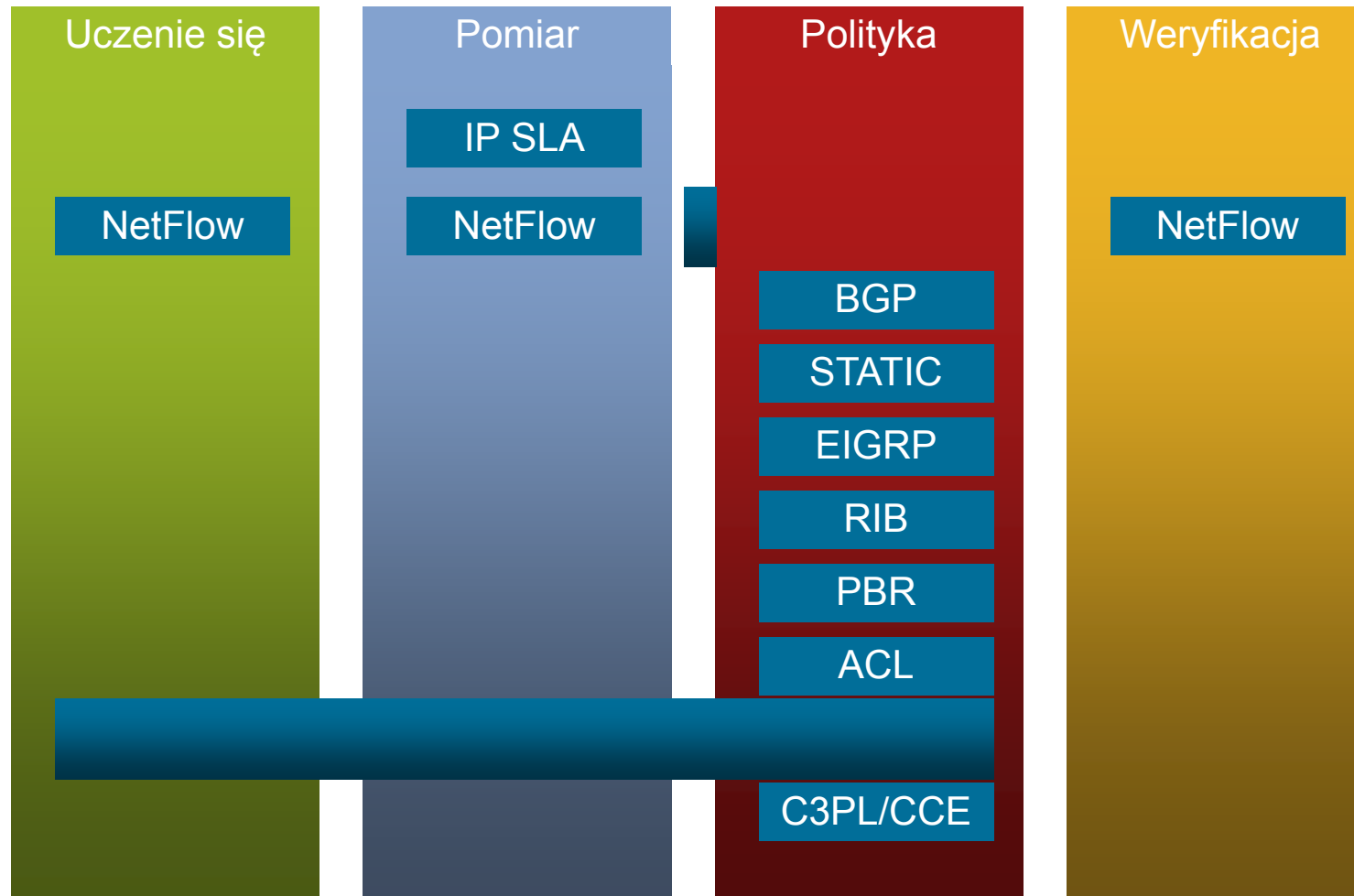
- Banki online
- Hosting email
- Systemy biletowe
- Instant Messaging
- Katalogi online
- Nowości/pogoda



- Głos
- Hosting aplikacji
- DNS
- Muzyka
- Video

BR—Border Router, MC—Master Controller

Narzędzia używane przez PfR



Firmy udostępniające usługi online

1. Interfejsy E3

Ser12/0, Ser13/0, ...

2. Routery ISR, ISR G2, 7200, ASR 1k

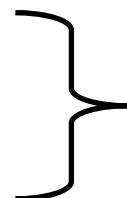
3. Routing BGP

BR muszą być peerami iBGP

domyślny routing

częściowe tablice

pełne światowe tablice



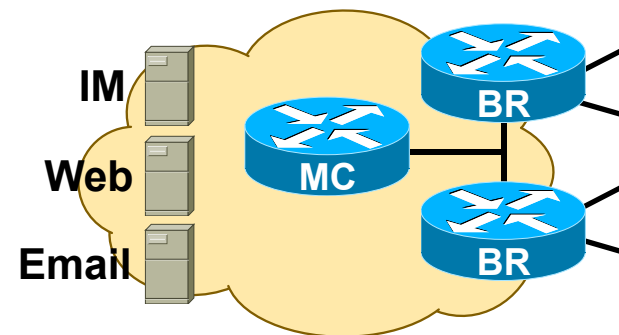
Taka sama konfiguracja PfR

4. Do 20000 prefiksów

12.4(15)T i wyższe, 7200 z NPE-G2

5. Klienci różnią się priorytetem polityki

6. Polityka uczenia się prefiksów to pasmo i później opóźnienie



BR—Border Router, MC—Master Controller

Firmy udostępniające usługi online

Domyślnie: wydajność i obciążenie

```
key chain key1
```

```
key 1
```

```
key-string oer
```

```
oer master
```

```
logging
```

```
mode route control
```

```
mode select-exit best
```

```
backoff 90 3000 300
```

```
periodic 600
```

```
border 10.1.1.2 key-chain key1
```

```
interface Ethernet8/0 internal
```

```
interface Serial12/0 external
```

```
interface Serial13/0 external
```

```
border 10.1.1.3 key-chain key1
```

```
interface Ethernet 8/0 internal
```

```
interface Serial12/0 external
```

```
interface Serial13/0 external
```

```
learn
```

```
throughput
```

```
delay
```

```
monitor-period 1
```

```
periodic-interval 0
```

```
prefixes 500
```

```
expire after time 240
```

MC 10.1.1.1

**Najlepsze wyjście
niezależnie od kierunku**

**Sprawdzamy co
10 minut**

**Uczymy się
500 prefiksów**

```
key chain key1
```

```
key 1
```

```
key-string oer
```

```
oer border
```

```
logging
```

```
local loopback 1
```

```
master 10.10.10.1 key-chain key1
```

```
interface ser12/0
```

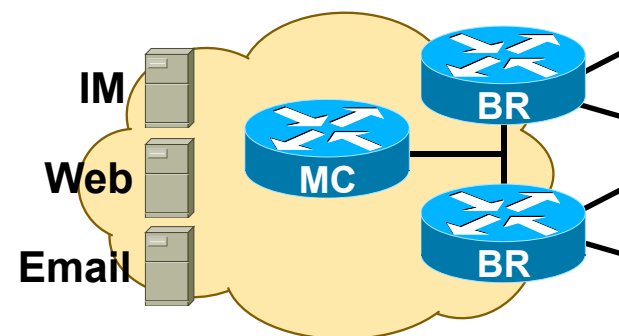
```
load-interval 30
```

```
interface ser13/0
```

```
load-interval 30
```

BR 10.10.10.2

BR 10.10.10.3



**Prefiks, który nie został
ponownie nauczony po 240
minutach zostaje usunięty**

Firmy udostępniające usługi online

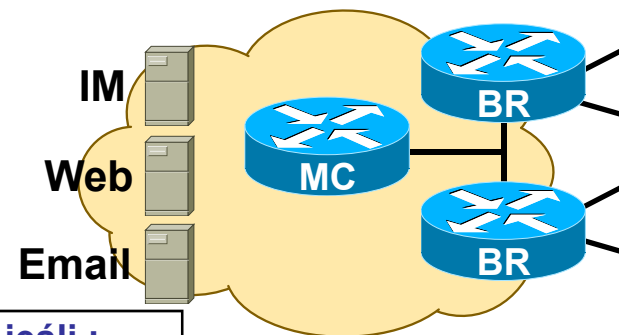
Tylko rozkładanie obciążenia

- Do domyślnej polityki dodajemy:

Wyłączenie okresowej
ewaluacji prefiksów

```
oer master
no periodic
resolve utilization priority 1 variance 5
resolve range priority 2
no resolve delay
no resolve loss
max-range-utilization percent 50
border 10.1.1.2
interface Serial12/0 external
max-xmit-utilization percent 90
interface Serial13/0 external
max-xmit-utilization percent 90
border 10.1.1.3
interface Serial12/0 external
max-xmit-utilization percent 90
interface Serial13/0 external
max-xmit-utilization percent 90
```

MC 10.1.1.1



OOP jeśli :
% util > najniższe
+ 50
% util > 90

- Wyjścia kandydujące mają:
 $\% \text{ util} \leq \text{lowest util} * (100+5) / 100$
- Ograniczamy kandydatów do:
 $\% \text{ util} < \text{lowest util} + 50$

Bardziej skomplikowany przykład #1

Przekierowanie ruchu na łącze spełniające warunek dla głosu: MOS na poziomie 4.00

Polityka głosowa

```
ip prefix-list BRANCH permit 10.1.1.0/24
ip prefix-list BRANCH permit 10.1.2.0/24

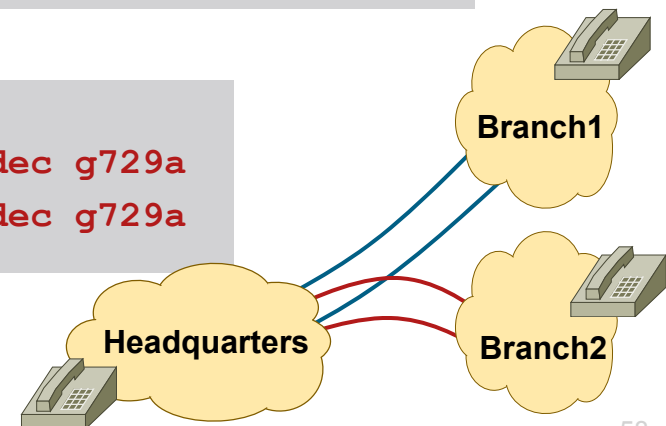
oer-map MAP 20
match traffic-class application nbar rtp-audio prefix-list MYBRANCH
set unreachable threshold 5
set mos percent 20 threshold 4.00
set resolve mos priority 1
set mode monitor fast
set probe frequency 2
```

Konfiguracja próbek

```
oer master
active-probe jitter 10.1.1.1 target-port 2000 codec g729a
active-probe jitter 10.1.2.1 target-port 2000 codec g729a
```

Responder dla próbek na zdalnych routerach

```
ip sla responder
```



Bardziej skomplikowany przykład #2

Przekierowanie ruchu na łączy zapewniające mniejsze opóźnienia dla ruchu interaktywnego

- Aplikacje wrażliwe na opóźnienia—telnet, ssh
- Inne aplikacje

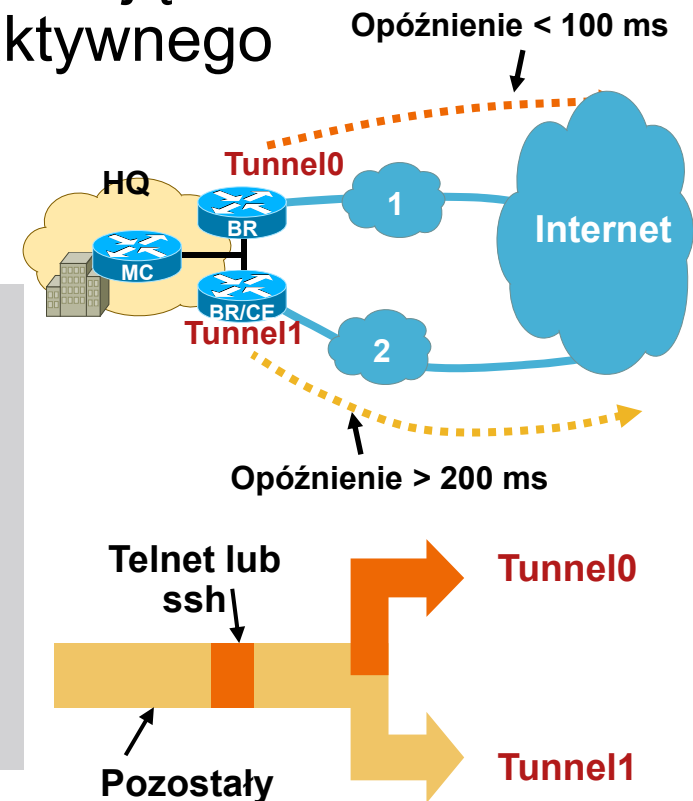
Uczenie się ruchu - aplikacje

```
ip prefix-list BRANCH_PFX permit 10.1.0.0/16
!
oer master
learn
list sequence 10 refname BRANCH_APPL
traffic-class application telnet ssh filter BRANCH_PFX
throughput
list sequence 20 refname BRANCH_PFX
traffic-class prefix-list BRANCH_PFX
throughput
```

Definicja polityki

```
oer-map MAP 10
match oer learn list BRANCH_APPL
set delay threshold 100
set resolve delay priority 1 variance 5
```

```
oer-map MAP 20
match oer learn list BRANCH_PFX
set delay threshold 400
set resolve range priority 1
```



A co gdy MC i BR to jeden router? ...i robimy NAT?

Ruch poddawany procesowi NAT

```
access-list 1 permit 10.1.0.0  
0.0.255.255
```

```
route-map isp-1 permit 10  
match ip address 1  
match interface Se1/0  
route-map isp-2 permit 10  
match ip address 1  
match interface Se2/0
```

```
interface Eth3/0  
ip nat inside  
interface Se1/0  
ip nat outside  
interface Se2/0  
ip nat outside
```

Interfejs
wewnętrzny
PfR

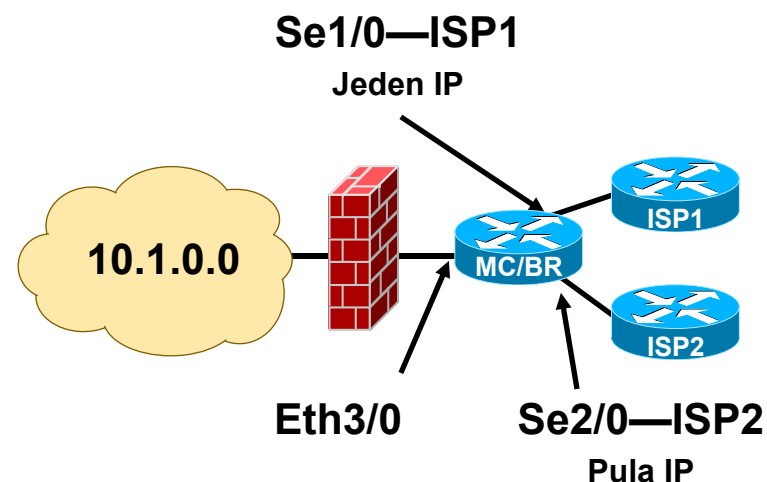
Interfejsy
zewnętrzne
PfR

Jeden IP

```
interface virtual-template 1  
ip nat inside source route-map isp-1 interface  
Virtual-Template1 overload oer
```

Pula adresów IP

```
ip nat pool ISP-2 <min-ip-addr> <max-ip-addr>  
prefix-length <len>  
ip nat inside source route-map isp-2 pool ISP-2  
oer
```



Mechanizm PfR dla ruchu przychodzącego

Wewnętrzny prefiks:

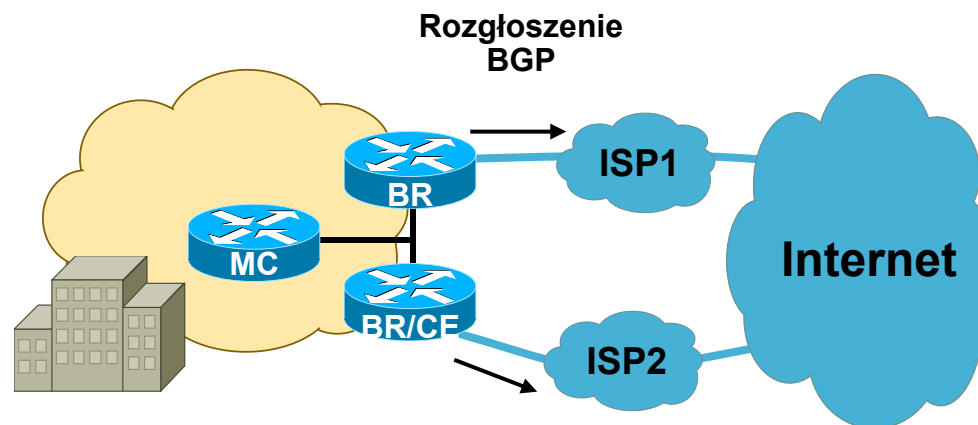
```
oer master  
learn  
inside bgp  
oer-map MAP 10  
match PfR learn inside
```

Konfiguracja wprost

```
ip prefix-list INSIDE permit 10.1.1.0/24  
oer-map MAP 10  
ip address prefix-list INSIDE inside
```

Sposób modyfikacji rozgłoszeń BGP

AS prepend - automatycznie



BGP Community

```
oer master  
border 10.1.1.1 key-chain PfR  
interface ethernet1/0 external  
downgrade bgp community 3:2
```



Inżynieria ruchowa na styku sieci własnej i innych

Protokół BGP

Czy na pewno muszę mieć BGP?

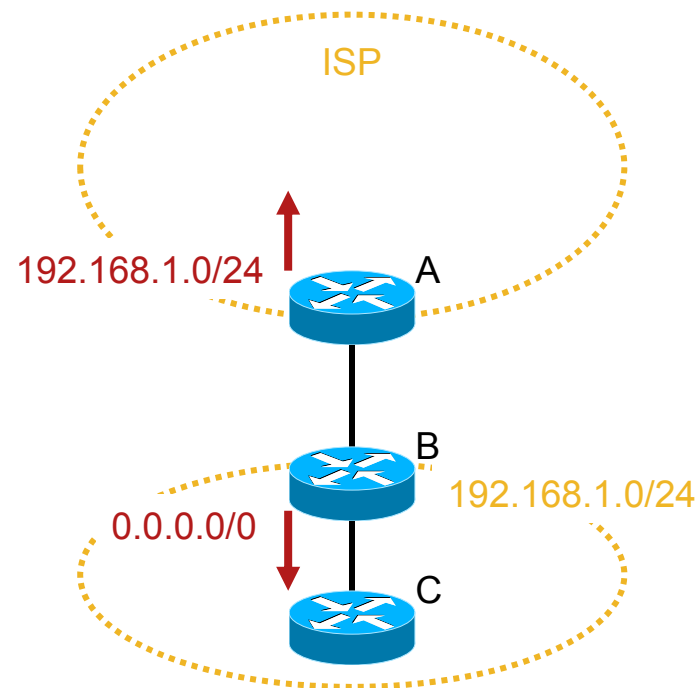
- Pojedyncze połączenie do Internetu - **NIE**

w sieci używasz domyślnej trasy
(statycznie lub protokół routingu)

Twój ISP zapewnia widoczność i
osiągalność przydzielonej Ci adresacji
IP

- Nawet jeśli łączy są dwa lub trzy do
tego samego ISP, BGP nie jest
potrzebne

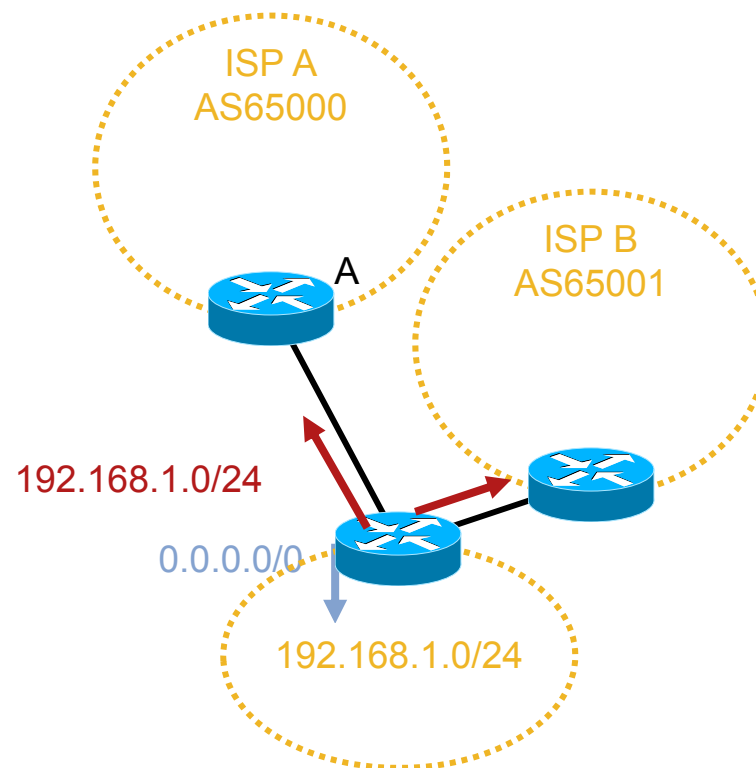
wiele tras domyślnych na różne IP po
stronie Twojej i ISP



Czy na pewno muszę mieć BGP?

- Jeśli jesteś połączony do dwóch różnych ISP – BGP jest pożądane – rozgłasza do obu swoją pulę adresów, Internet widzi ją przez obu ISP
- Nie oznacza to, że musisz pobierać wszystkie światowe prefiksy

...pozwala to jednak 'świadomie' kształtować własną politykę routingu dużo dokładniej



Protokół BGP

Atrybuty pozwalające wpływać na trasę

1: ORIGIN

2: AS-PATH

3: NEXT-HOP

4: MED

5: LOCAL_PREF

6: ATOMIC_AGGREGATE

7: AGGREGATOR

8: COMMUNITY

9: ORIGINATOR_ID

10: CLUSTER_LIST

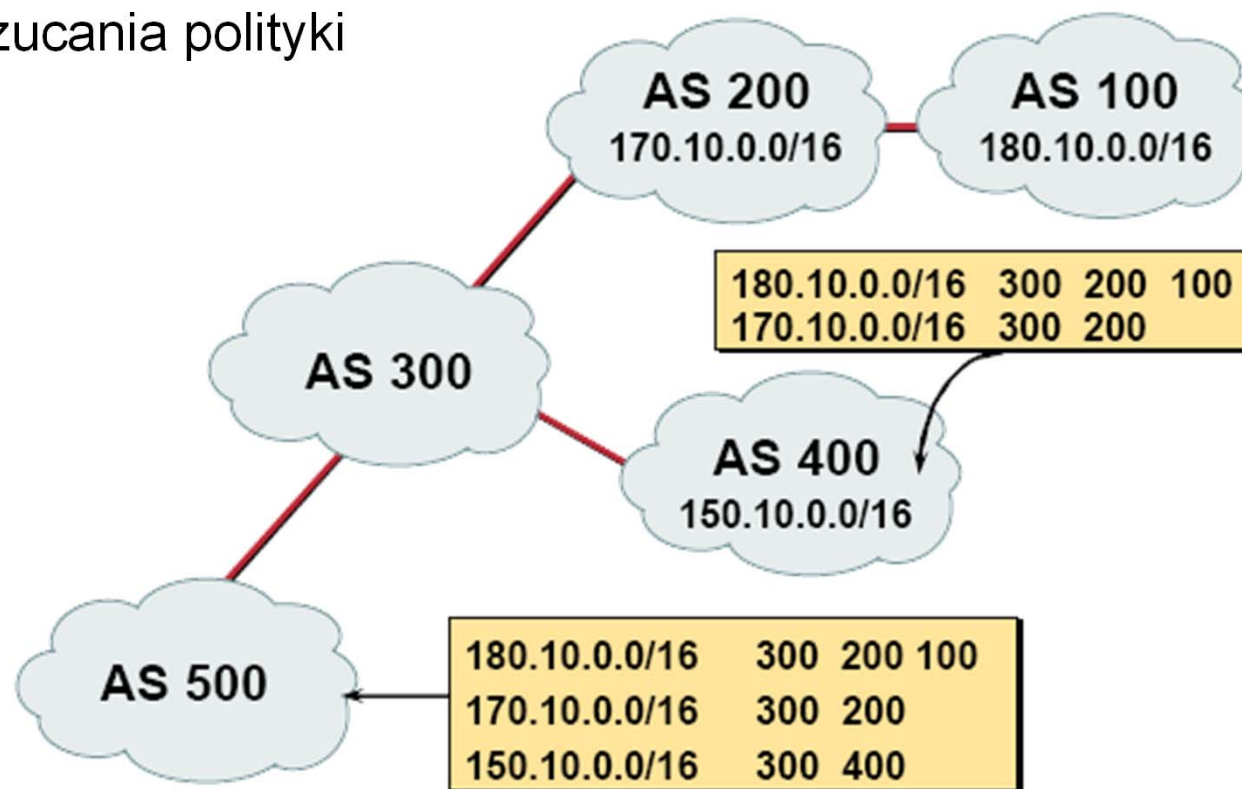
14: MP_REACH_NLRI

15: MP_UNREACH_NLRI

Protokół BGP

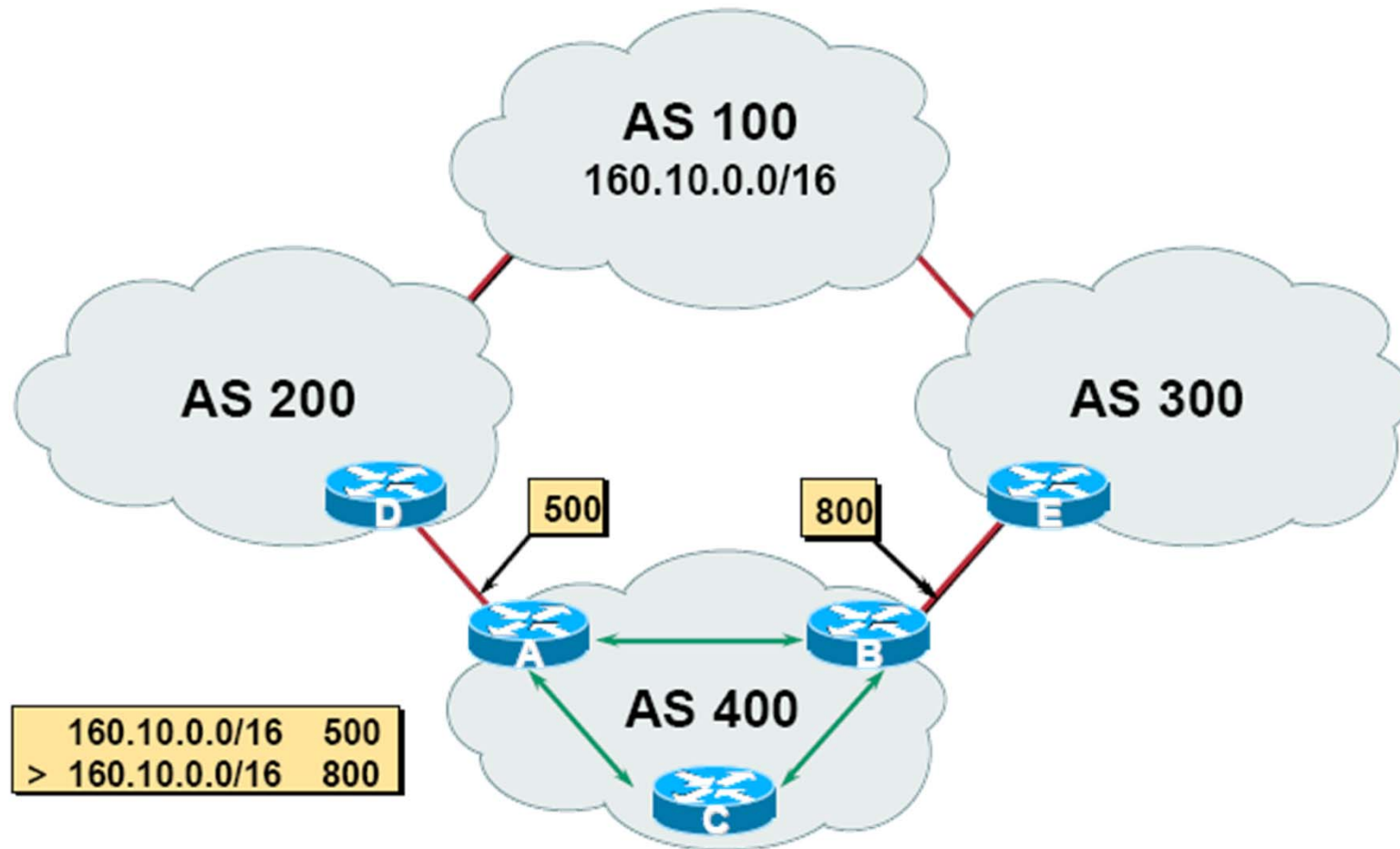
Atrybuty – AS-Path

- Trasa, jaką prefiks przeszedł
- Mechanizm zapobiegający pętlom
- Narzędzie narzucania polityki



Protokół BGP

Atrybuty – local preference



Protokół BGP

Atrybuty – community

- Opisane w RFC1997, miedzy ASami (transit) 32 bitowa wartość dodatnia
- Dla wygody zapisywana rozdzielone dwukropkiem Standardowo **<NUMER-AS;** Na przykład: **64888:777**
- Bardzo przydatne i częste w peeringach do sterowania oznaczonego prefiksu

```
aut-num: AS5617
as-name: TPNET
descr: Polish Telecom's commercial IP network
descr: Telekomunikacja Polska S.A.
descr: ul.Nowogrodzka 47A
descr: 00-695 Warszawa
descr: POLAND
import: from AS1299
       accept ANY
export: to AS1299
       announce AS-TPNET
import: from AS5511
       accept ANY
export: to AS5511
       announce AS-TPNET
admin-c: TPRG
tech-c: TPRG
remarks: TPNET peers upstream with Telia and OpenTransit
remarks: and with customers in Poland
remarks:
remarks: =====
remarks: Communities used in AS5617
remarks: =====
remarks:
remarks: 5617:103 - prefix is not announced to foreign peers
remarks:
remarks: We do not accept all RFC1997 communities (no-export etc.)
remarks:
remarks: communities for specific link are 5617:ab0x :
remarks: x=0 for "do not advertise"; x=1,2,3 for "prepend 1,2,3 times"
remarks:
remarks: for AS1299 Telia 5617:110x
remarks: for AS5511 OpenTransit 5617:120x
remarks:
remarks: for AS8501 POL34: 5617:210x
remarks: for AS16283 Lodman: 5617:220x
remarks: for AS8664 ICM: 5617:240x
remarks: not used 5617:250x
remarks: for AS8890 Kampus Ochota:5617:260x
remarks: for AS8364 Pozman: 5617:270x
remarks: for AS8508 Silweb: 5617:280x
remarks: for AS8970 WASK: 5617:290x
remarks:
remarks: =====
remarks:
remarks: 5617:997 - BLACKHOLING COMMUNITY
remarks:
remarks: Blackhole community for \32 routes available for peers
remarks: advertising only one AS
remarks:
remarks: =====
```

Protokół BGP

Jak BGP wybiera najlepszą trasę? (wersja skrócona)

- Najwyższy local preference (w ramach AS)
- Najkrótsza ścieżka AS-Path
- Najniższy kod pochodzenia (Origin)
 - IGP < EGP < incomplete
- Najniższa wartość MED (Multi-Exit Discriminator)
- Lepiej trasa z eBGP niż z iBGP
- Najpierw trasa z niższym kosztem wg. IGP do next-hop
- Najniższy router-id routera BGP
- Najniższy adres peer'a

(pełna lista w RFC4271)

Multihoming?

- Więcej niż jedno połączenie zewnętrzne do lokalnej sieci

Dwa lub większa ilość łącz od tego samego ISP

Dwa lub większa ilość łącz od różnych ISP

- Zwykle dwa routery brzegowe

Jeden router zabezpiecza tylko przed awarią ISP do niego podłączonych

Scenariusz: jedno łącze jako backup

- Na obu łączach akceptujemy tylko trasę domyślną (default)

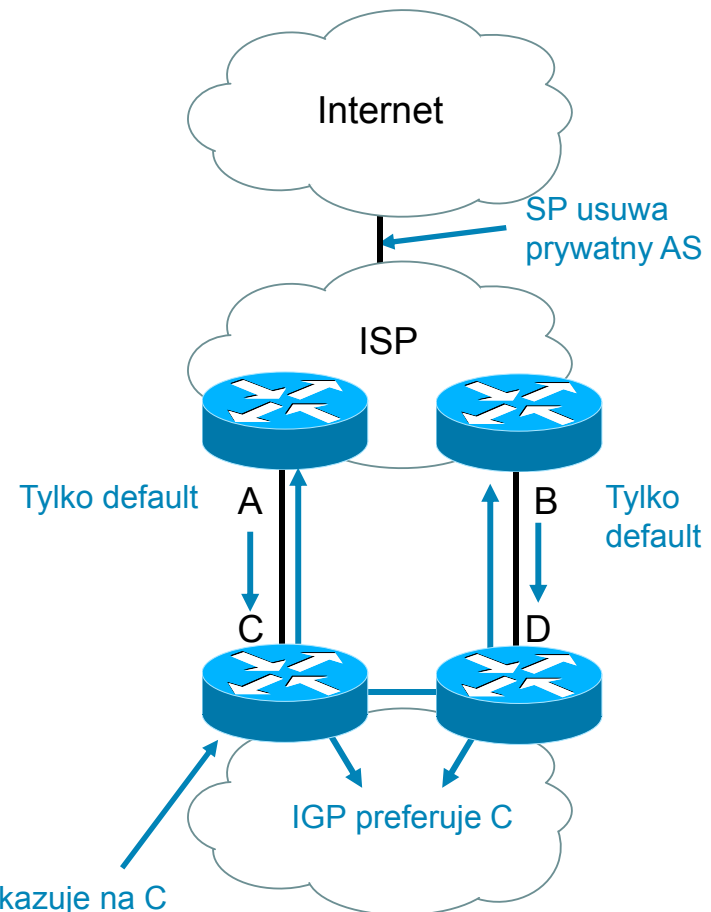
Za pomocą local preference kierujemy ruch wyjściowy przez łącze „podstawowe”

...a protokołem IGP kierujemy ruch w stronę routera

- Rozgłaszamy tą samą klasę adresową przez oba łącza

To SP będzie preferował jedno łącze

Inna możliwość to rozgłoszenie warunkowe

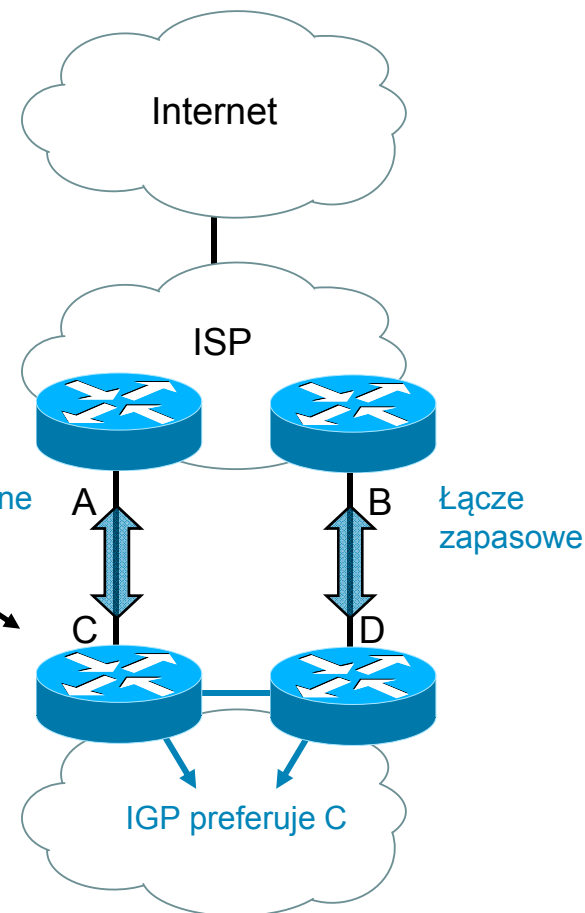


Scenariusz: jedno łącze jako backup

```
router bgp 65534
 network 121.10.0.0 mask 255.255.224.0
 neighbor 122.102.10.2 remote-as XXX
 neighbor 122.102.10.2 description primary-link
 neighbor 122.102.10.2 prefix-list aggregate out
 neighbor 122.102.10.2 prefix-list default in
 !
 ip prefix-list aggregate permit 121.10.0.0/19
 ip prefix-list default permit 0.0.0.0/0
 !
 ip route 121.10.0.0 255.255.224.0 null0
```

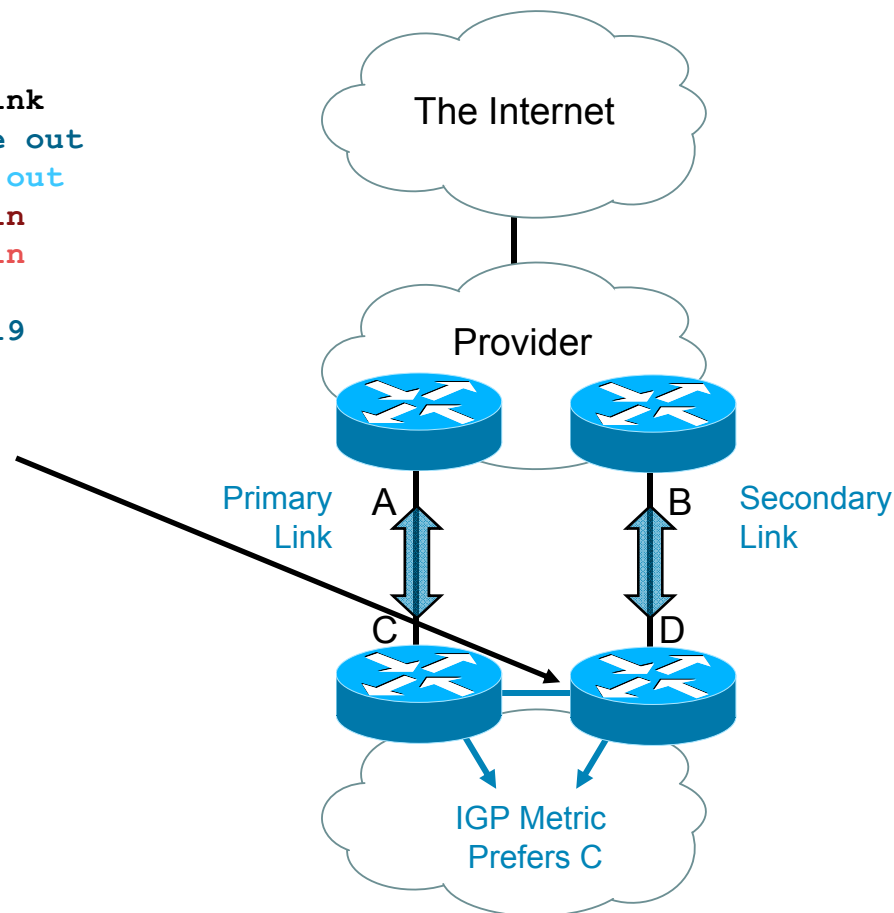
Łącze główne

Łącze
zapasowe



Scenariusz: jedno łącze jako backup

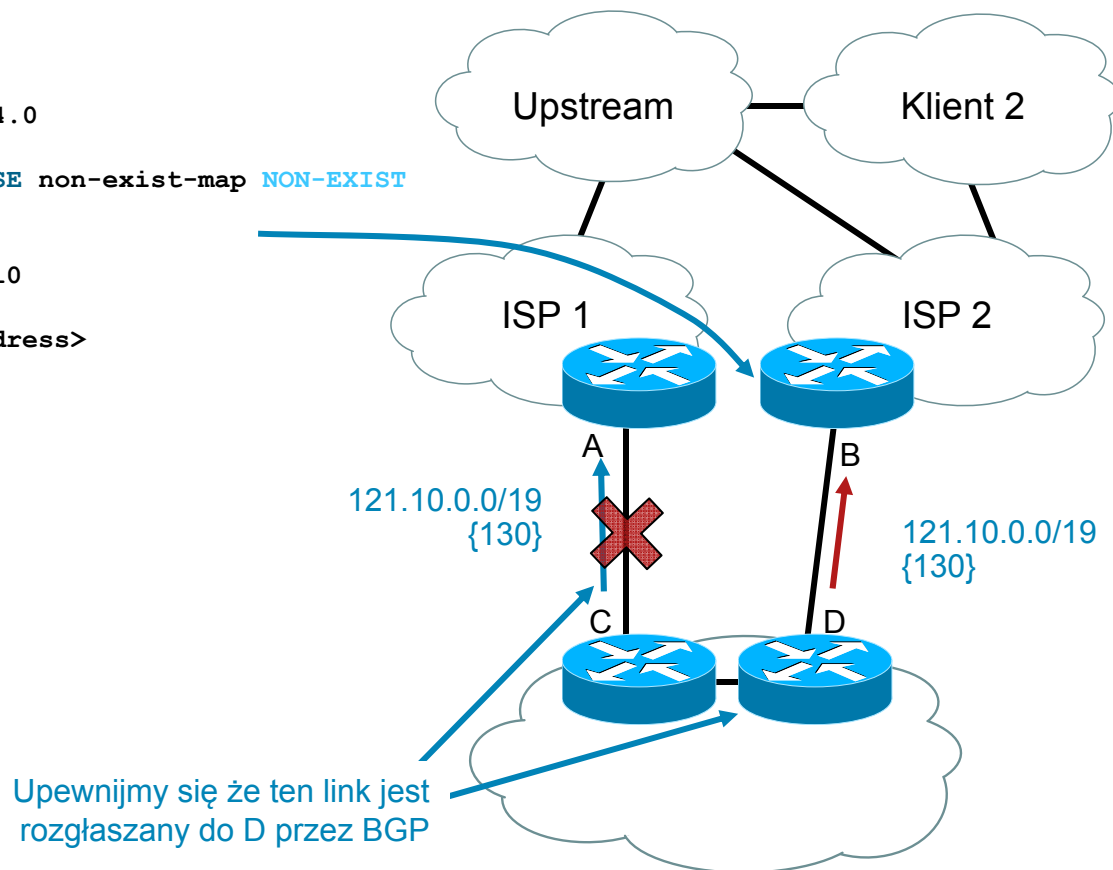
```
router bgp 65534
network 121.10.0.0 mask 255.255.224.0
neighbor 122.102.10.6 remote-as XXX
neighbor 122.102.10.6 description backup-link
neighbor 122.102.10.6 prefix-list aggregate out
neighbor 122.102.10.6 route-map backup-out out
neighbor 122.102.10.6 prefix-list default in
neighbor 122.102.10.6 route-map backup-in in
!
ip prefix-list aggregate permit 121.10.0.0/19
ip prefix-list default permit 0.0.0.0/0
!
ip route 121.10.0.0 255.255.224.0 null0
!
route-map backup-out permit 10
match ip address prefix-list aggregate
set metric 10
route-map backup-out permit 20
!
route-map backup-in permit 10
set local-preference 90
!
```



Scenariusz: jedno łącze jako backup

Rozgłoszenie warunkowe

```
router bgp 130
  bgp log-neighbor-changes
  network 121.10.0.0 mask 255.255.224.0
  neighbor <B> remote-as XXX
  neighbor <B> advertise-map ADVERTISE non-exist-map NON-EXIST
  neighbor <C> remote-as 130
  !
  ip route 10.1.0.0 255.255.240.0 null0
  !
  access-list 60 permit <a->c link address>
  access-list 65 permit 121.10.0.0
  !
  route-map NON-EXIST permit 10
  match ip address 60
  !
  route-map ADVERTISE permit 10
  match ip address 65
```



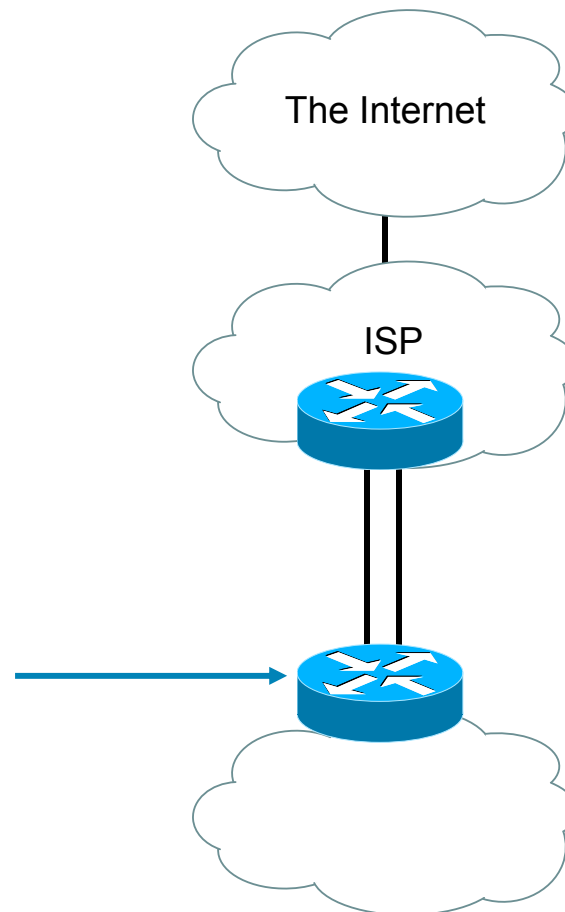
Scenariusz: rozkładanie obciążenia

- eBGP multihop – jeśli masz więcej niż jedno łącze pomiędzy parą routerów

eBGP do adresów loopback

prefiksy eBGP uczone są z adresem interfejsu loopback routera dostawcy

```
router bgp 65534
  neighbor 1.1.1.1 remote-as XXX
  neighbor 1.1.1.1 ebgp-multihop 2
  neighbor 1.1.1.1 update-soure Loopback0
!
ip route 1.1.1.1 255.255.255.255 serial 1/0
ip route 1.1.1.1 255.255.255.255 serial 1/1
```



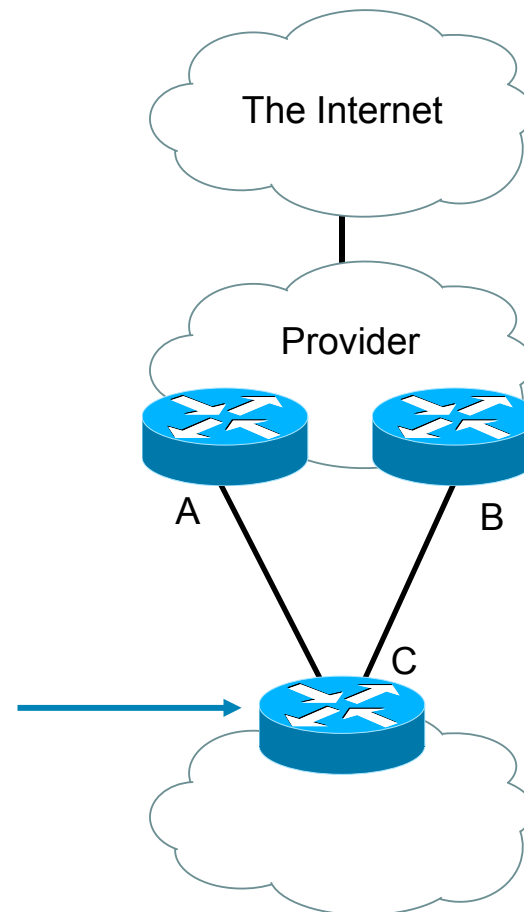
Scenariusz: rozkładanie obciążenia

- Jeśli wiele łącz prowadzi do różnych routerów – eBGP multipath

Wiele sesji eBGP do tego samego dostawcy (ASN)

Sesje terminowane na tym samym routerze

```
router bgp 201
 neighbor 1.1.2.1 remote-as XXX
 neighbor 1.1.2.5 remote-as XXX
 maximum-paths 2
```

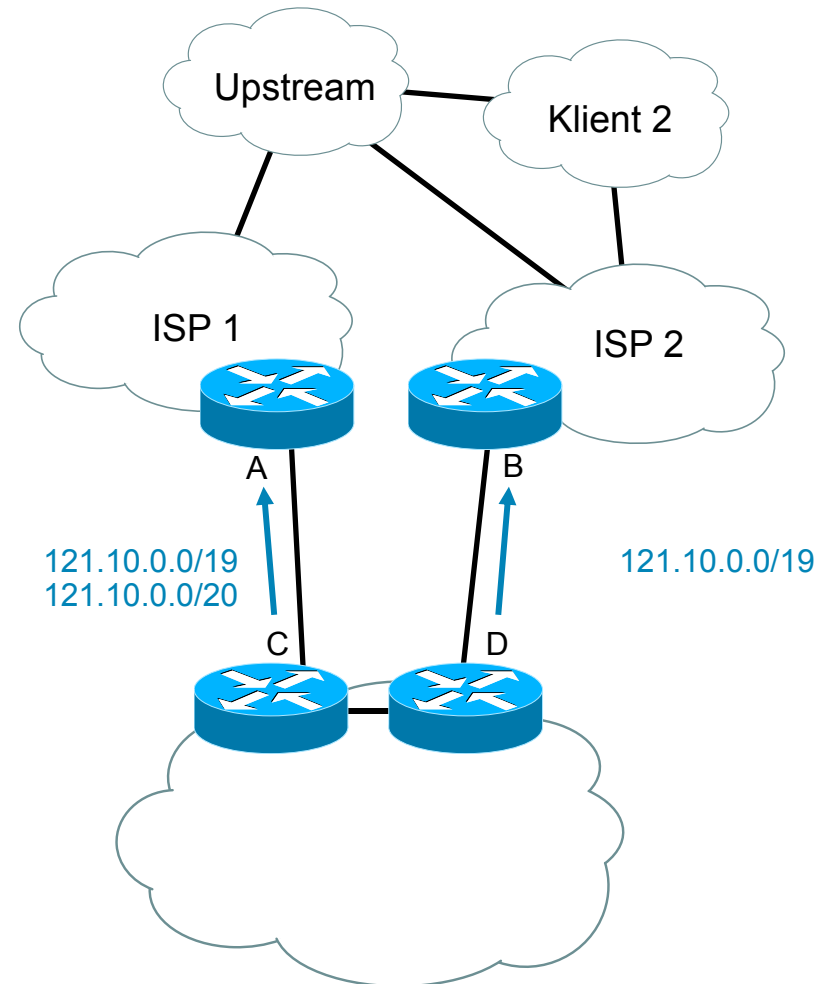


Równoważenie ruchu przychodzącego

- Dokładniejsze prefiksy

Rozgłoś 121.10.0.0/19 i
121.10.0.0/20 przez jedno
połączenie

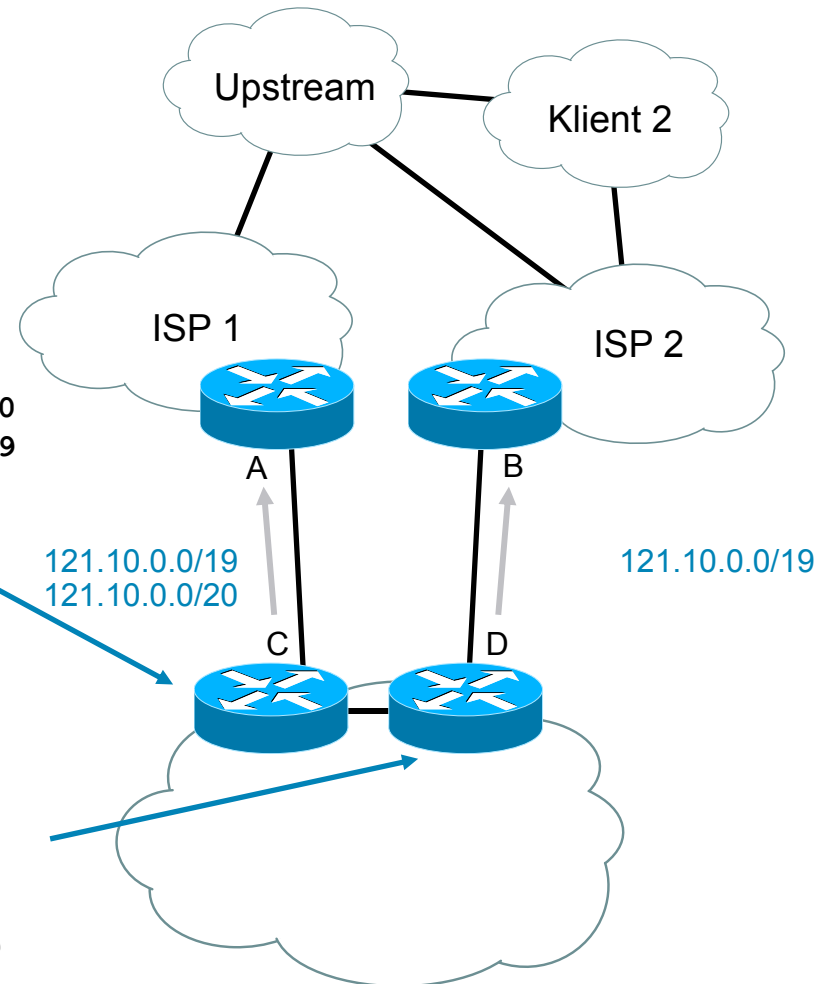
Rozgłoś 121.10.0.0/19
przez drugie połączenie



Równoważenie ruchu przychodzącego

```
router bgp 65555
 network 121.10.0.0 mask 255.255.224.0
 network 121.10.0.0 mask 255.255.240.0
 neighbor x.x.x.x remote-as <provider 1>
 neighbor x.x.x.x prefix-list firstblock out
!
ip prefix-list firstblock permit 121.10.0.0/20
ip prefix-list firstblock permit 121.10.0.0/19
```

```
router bgp 130
 network 121.10.0.0 mask 255.255.224.0
 neighbor x.x.x.x remote-as <provider 2>
 neighbor x.x.x.x prefix-list secondblock out
!
ip prefix-list secondblock permit 121.10.0.0/19
```



Równoważenie ruchu przychodzącego

- „Pogorszenie” atrakcyjności własnego prefiksu 10.10.10.0/24 przez AS200

```
router bgp 100
  neighbor 172.16.10.1 remote-as 200
  neighbor 172.16.10.1 route-map bgp-prepend-2x out
```

```
route-map bgp-prepend-2x permit 10
  set as-path prepend 100 100
```

```
rtr-AS200# show ip bgp 10.0.10.0
```

	Network	Next Hop	LocPrf	Path
*	10.0.10.0/24	172.16.10.2	100	100 100 100 i
*>		172.16.99.1	100	300 100 i



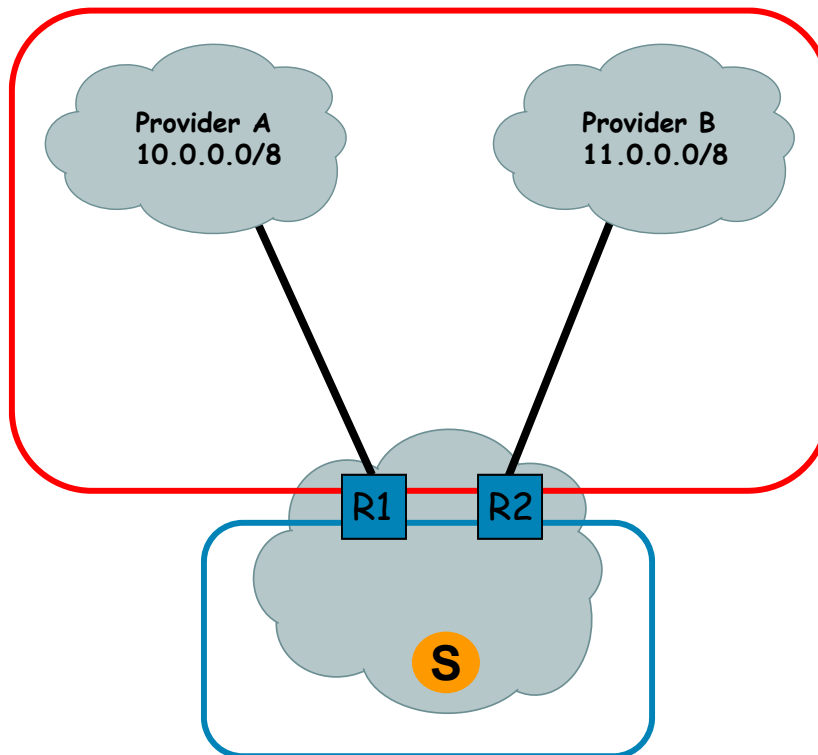
Inżynieria ruchowa na styku sieci własnej i innych

LISP

LISP?

- Problem który LISP stara się rozwiązać, to skalowalność tablic routingu międzyplanetarny internet?
- Wnioski ze spotkania Internet Architect Board w październiku 2006 opisano w RFC4984... i już 😊
- LISP jest próbą stworzenia technologii typu OTN – niezależną od protokołu warstwy trzeciej – IPv4 i IPv6

LISP – podział adresacji



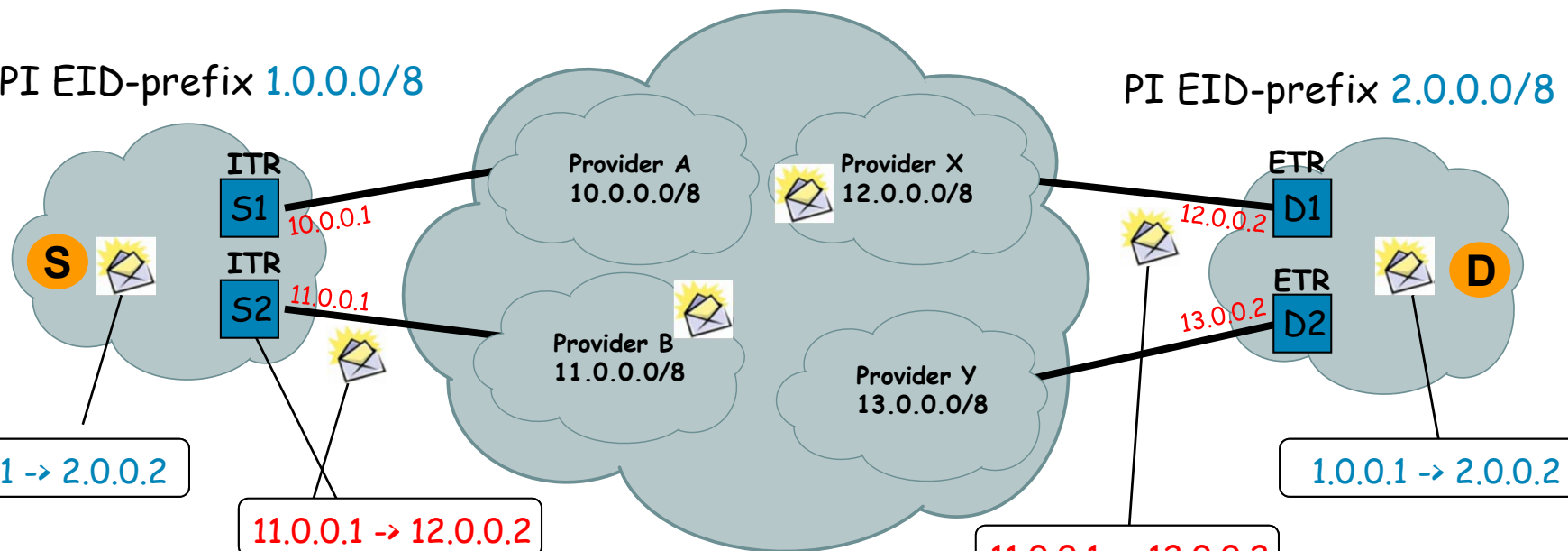
W szkielecie („Internecie”)
wykorzystywane są adresy typu
‘Locator’

Hosty używają ID/EID

Przekazywanie ruchu unicastowego

PI EID-prefix 1.0.0.0/8

PI EID-prefix 2.0.0.0/8



1.0.0.1 -> 2.0.0.2

1.0.0.1 -> 2.0.0.2

Wpis DNS:

D.abc.com A 2.0.0.2

Legenda:

EID -> Niebieskie

Locator -> Czerwone

Wpis
mapujący

EID-prefix: 2.0.0.0/8

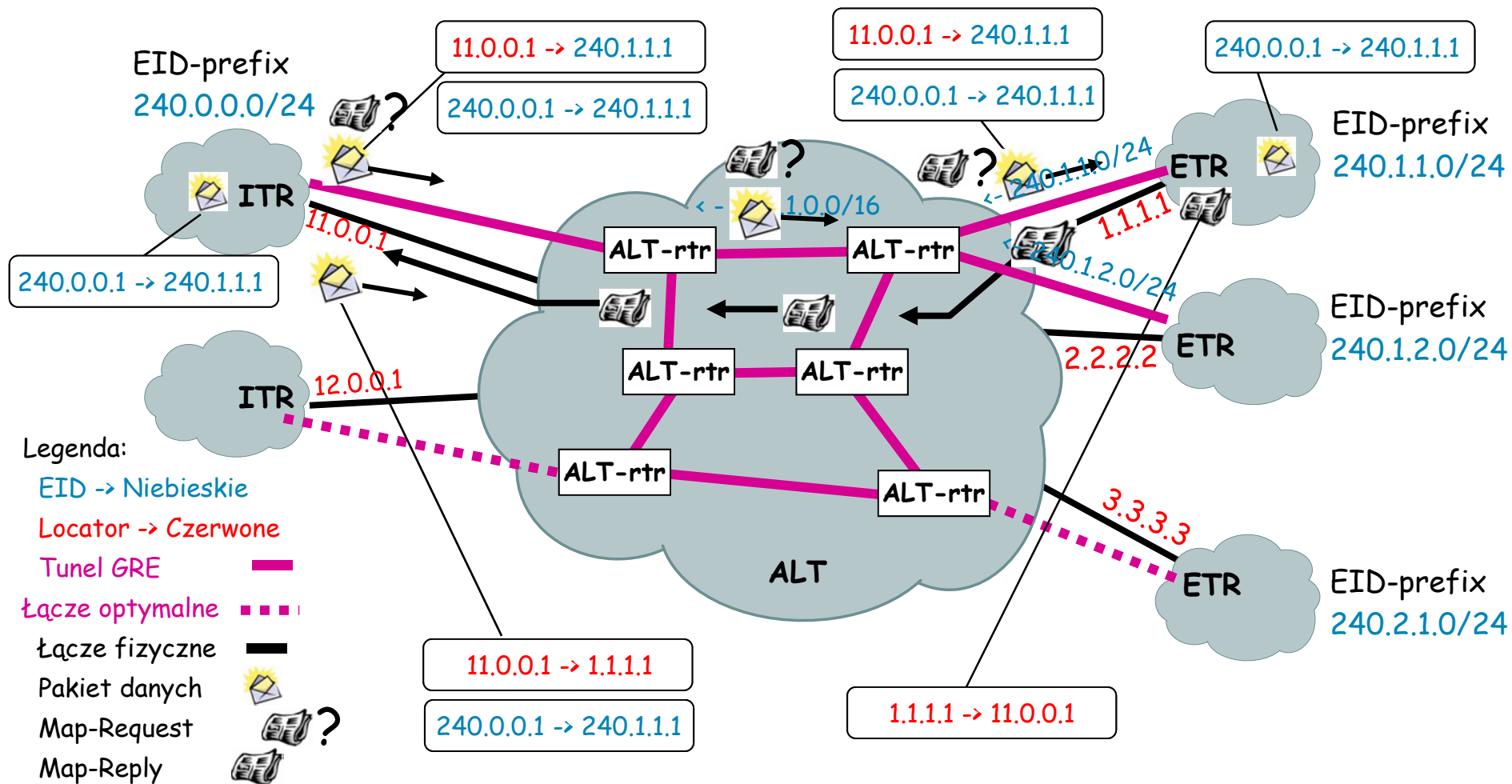
Locator-set:

12.0.0.2, priority: 1, weight: 50 (D1)

13.0.0.2, priority: 1, weight: 50 (D2)

Polityka
Kontrolowana
przez hosty
docelowy

Jak to działa razem?





Q&A

Łukasz Bromirski, lbromirski@cisco.com



CISCO