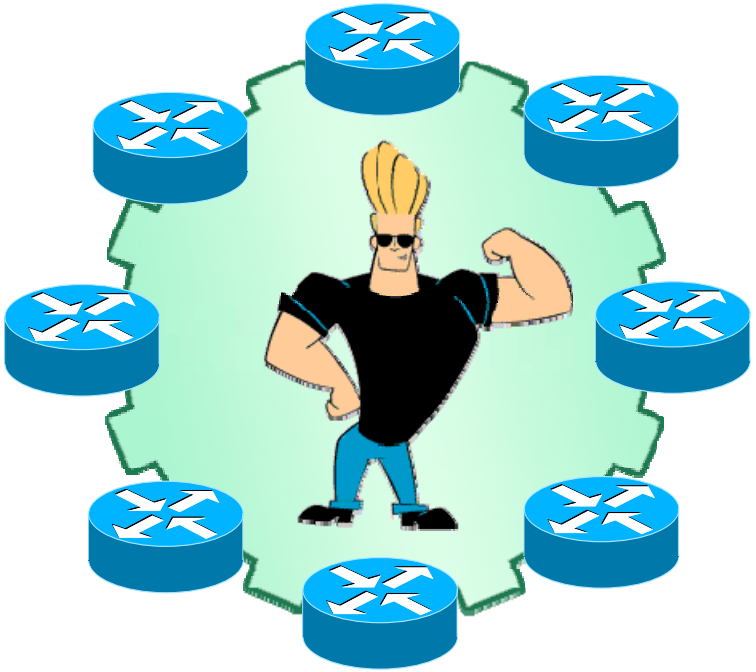
(secure)

# ROUTING WITH OSPF AND BGP FOR FUN, FUN & FUN




**Łukasz Bromirski**

lukasz@bromirski.net

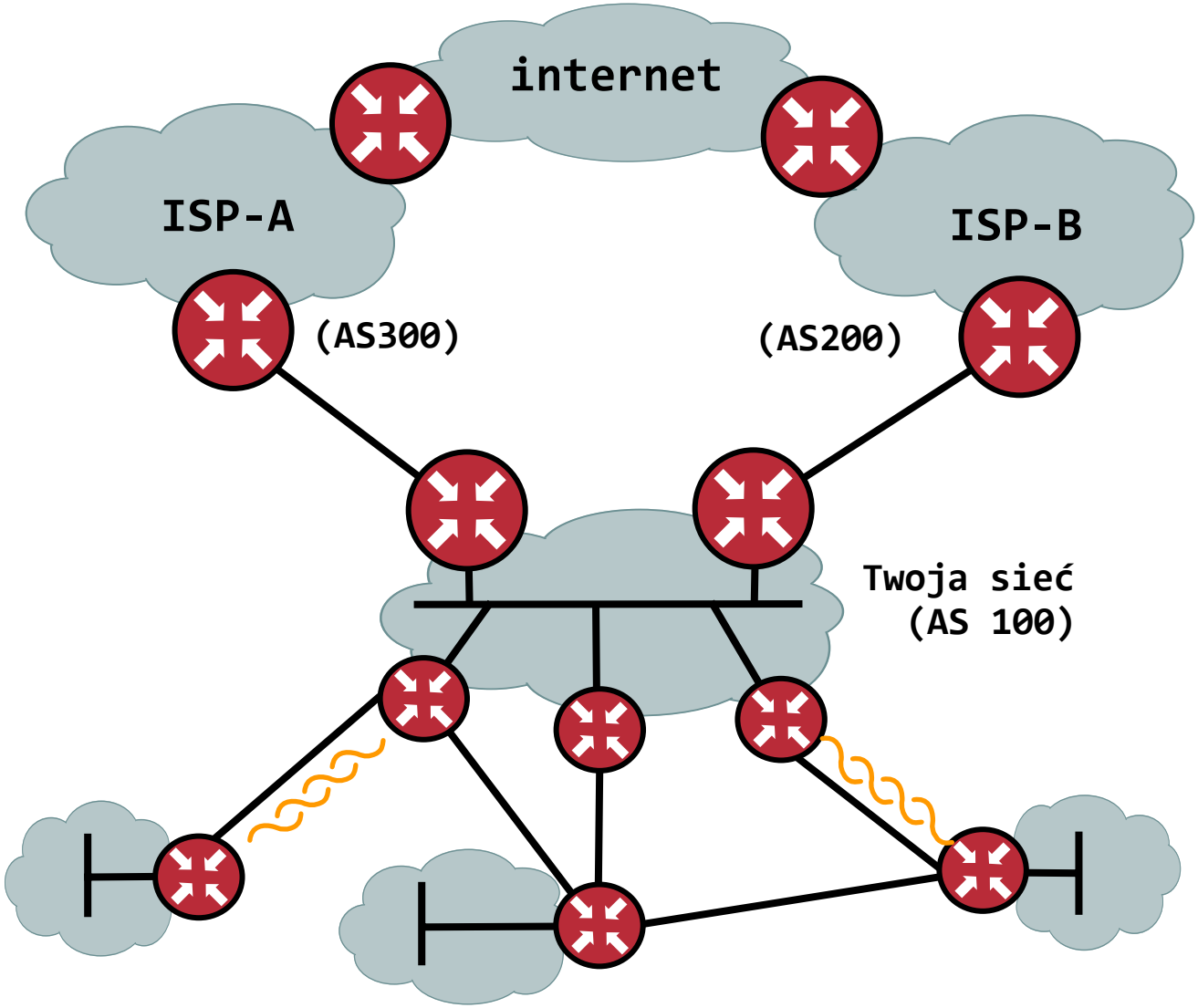
# Agenda

- Gdzie i dlaczego OSPF?
- OSPF w praktyce
- Gdzie i dlaczego BGP?
- BGP w praktyce
- Q&A

## Wymagana będzie...

- ...znajomość chociażby na minimalnym poziomie zagadnień związanych z routingiem dynamicznym i protokołów OSPF i BGP
- ...pakietu Quagga i/lub OpenOSPFd/BGPd
- Na końcu tej prezentacji znajdują się odnośniki do materiałów uzupełniających – również tych bardzo podstawowych

# O czym będzie ta sesja?



# GDZIE I DLACZEGO OSPF?

# Dlaczego OSPF?

- **Najszybszy standardowy protokół IGP**

  **IGP = wewnętrzny (ang. internal gateway protocol)**

- **Większość małych i średnich sieci ma strukturę hierarchiczną, lub da się ją stworzyć**
- **Dostosowywalny do większości scenariuszy za pomocą standardowych mechanizmów i narzędzi**
- **Dostępny ‘for free’ w implementacjach:**

  **Quagga – ospfd**

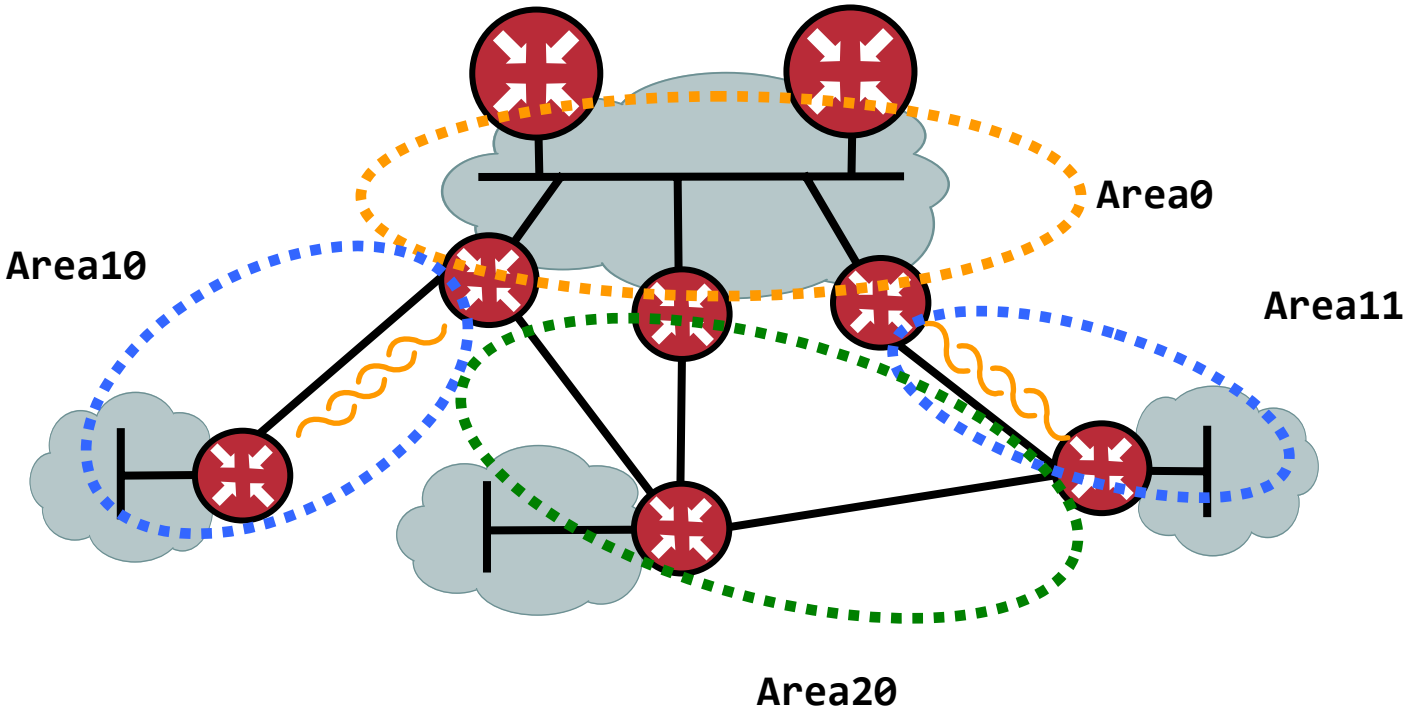
  **OpenOSPFd**

# Protokół routingu OSPF

## Jak działa?

- **OSPF posługuje się hierarchiczną strukturą sieci:**
  - **obszar backbone (area 0)**
  - **obszary podłączone różnych typów**
- **Każdy z obszarów musi być połączony do obszaru 0**
  - **jeśli nie może – przez link wirtualny**
- **Routery identyfikowane są za pomocą router-id**
  - **najwyższy adres IP ze wszystkich interfejsów**
  - **pierwszeństwo mają interfejsy loopback – użyj ich!**

# Jak podzielić sieć na obszary OSPF?

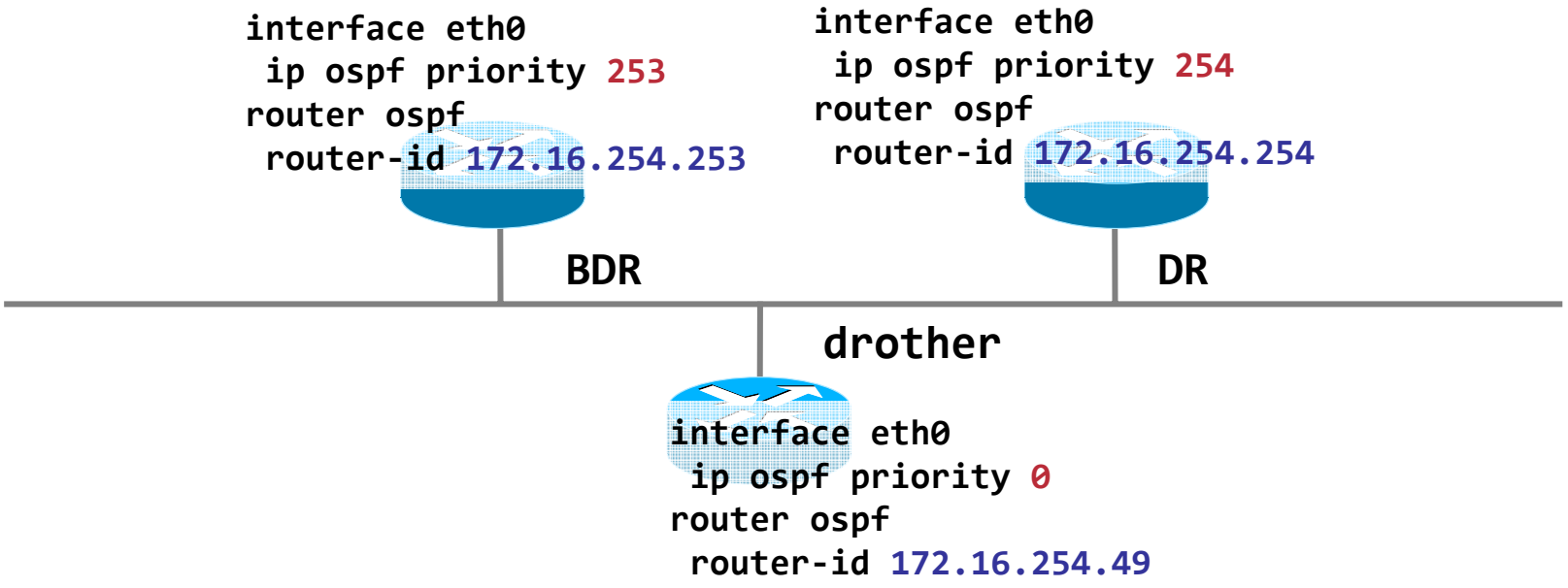


# Jak właściwie wybrać DR/BDRa dla segmentu?

- **We wszystkich topologiach, w których wiele routerów łączy wspólny segment Ethernet**

  **priorytet routera (0-254, 254 najwyższy, 0 – nie zostanie DR)**

  **wyższe router-id (zalecane stabilne rozplanowanie numeracji interfejsów loopback!)**



# Protokół OSPF

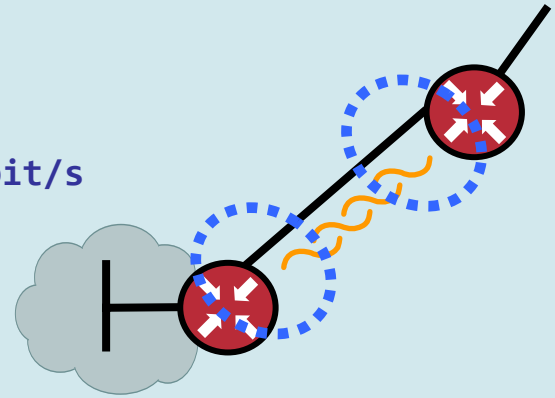
## Kiedy point-to-point?

- **Zalecane w przypadku sieci używających jako warstwy transportowej 802.11:**

  **oszczędzamy czas potrzebny na elekcję DR/BDR**

- **Dwa interfejsy:**

```
interface eth0
 ! interfejs podłączony do mostu 802.11
 ip ospf network point-to-point
 ip ospf cost 20 ! interfejs 100Mbit/s pracuje jak 5Mbit/s
interface eth1
 ! interfejs podłączony do linku 100Mbit/s
 ip ospf network point-to-point
 ip ospf cost 1 ! interfejs 100Mbit/s
```



# Protokół OSPF

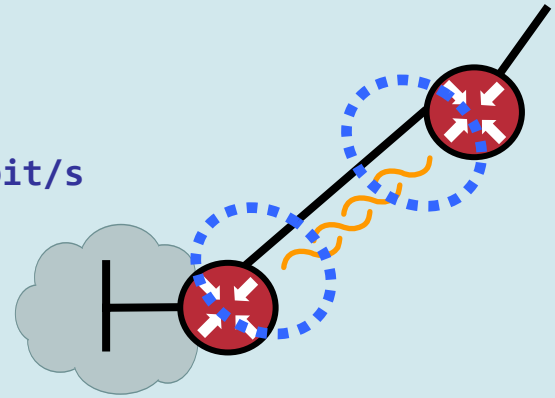
## Kiedy point-to-point?

- **Zalecane w przypadku sieci używających jako warstwy transportowej 802.11:**

  **oszczędzamy czas potrzebny na elekcję DR/BDR**

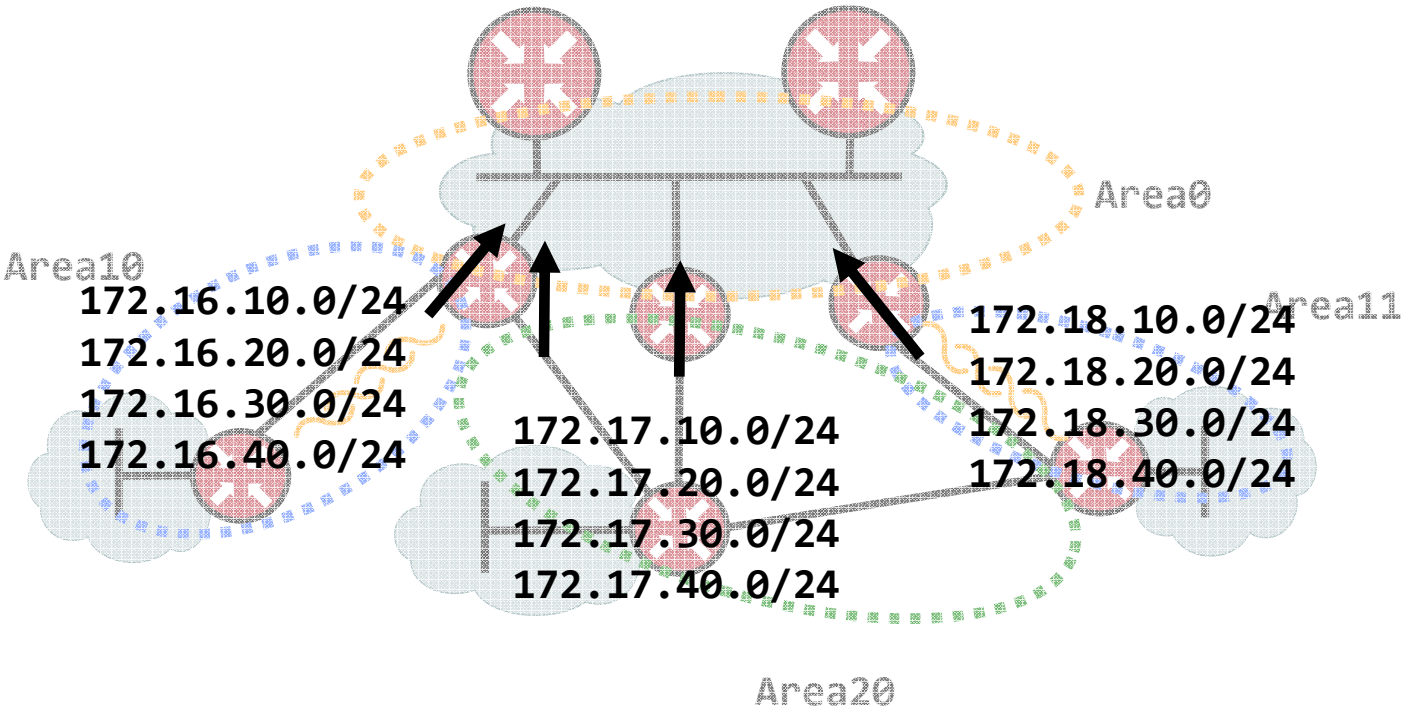
- **Jeden interfejs do bridge'a i łącza 100Mbit/s:**

```
interface eth0.10
 ! interfejs podłączony do mostu 802.11
 ip ospf network point-to-point
 ip ospf cost 20 ! interfejs 100Mbit/s pracuje jak 5Mbit/s
interface eth0.20
 ! interfejs podłączony do linku 100Mbit/s
 ip ospf network point-to-point
 ip ospf cost 1 ! interfejs 100Mbit/s
```



# Protokół routingu OSPF

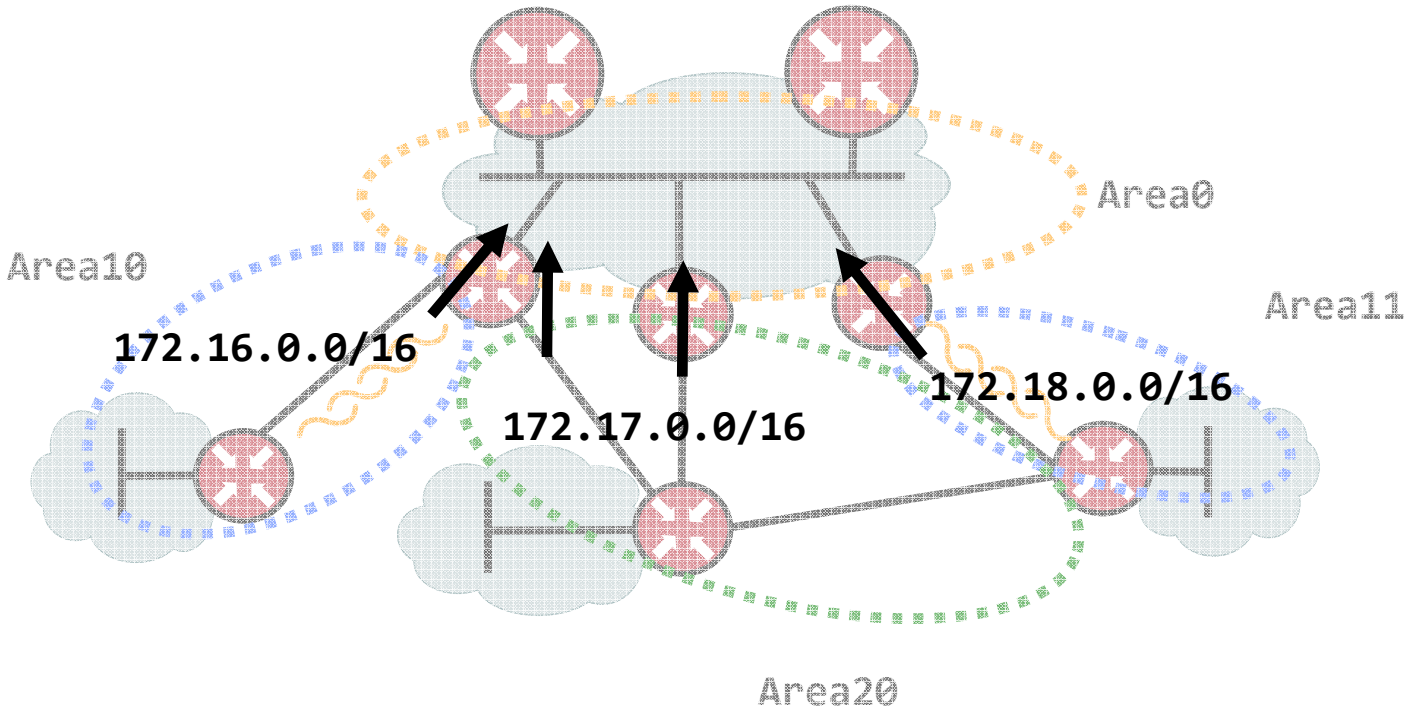
## Sumaryzacja - brak




- Wiele prefiksów w Area 0
- Dodatkowe obciążenie dla algorytmu SPF i routerów backbone
- „up/down prefiksu” – obciążamy CPU routerów w A0

# Protokół routingu OSPF

## Sumaryzacja - brak




- Pojedyncze prefiksy w Area 0
- Zmniejszamy obciążenie CPU, nie propagujemy problemów w warstwie dostępowej do szkieletu/rdzenia

# Protokół routingu OSPF

## Sumaryzacja - konfiguracja

- **Obszar 10 rozgłasza sieć 172.16.0.0/16, zamiast:**

  **172.16.10.0/24**

  **172.16.20.0/24**

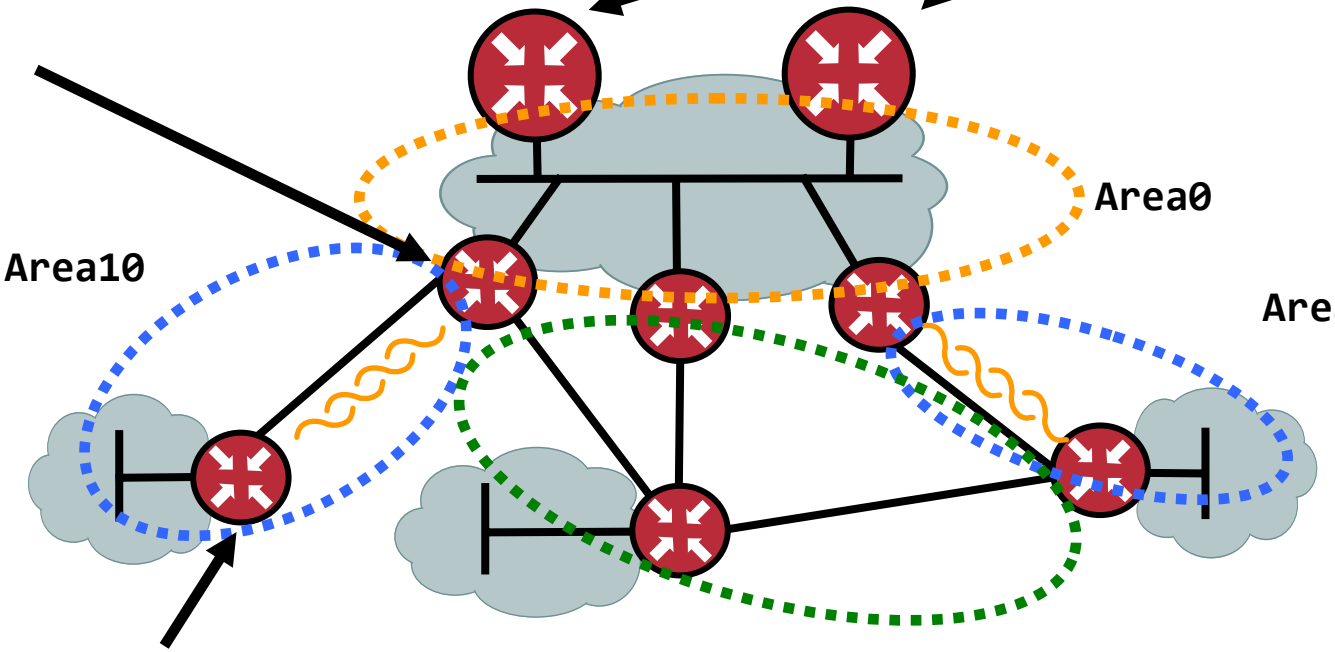
  **…**

```
router ospf
  ospf router-id 172.16.254.11
  [...]
  area 0.0.0.10 range 172.16.0.0/16
```

# Trasa domyślna

```
router ospf
 default-information originate [always]
```

```
router ospf
 area 0.0.0.10 stub no-summary
```




```
router ospf
 area 0.0.0.10 stub no-summary
```

# GDZIE I DLACZEGO BGP?

# Protokół BGP

## Wstęp i rozwinięcie

- **Protokół routingu używany do wymiany informacji o osiągalności sieci (prefiksów) pomiędzy systemami autonomicznymi**

  **klasy EGP – Exterior Gateway Protocol**

- **Do niedawna RFC1771 + rozszerzenia, od stycznia 2006 obowiązuje RFC4271:**

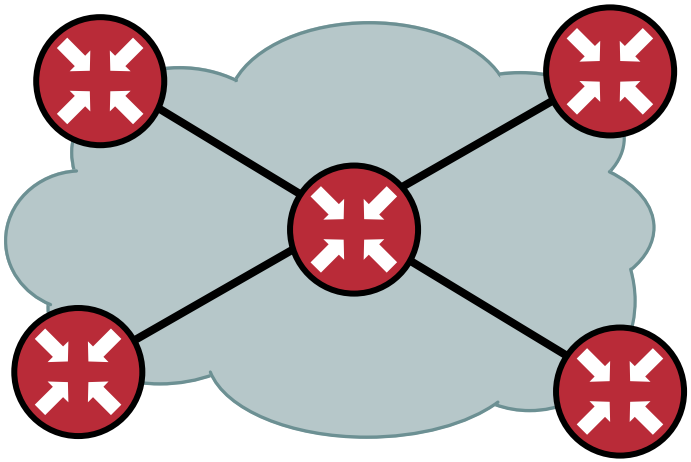
  **http://www.ietf.org/rfc/rfc4271.txt**

  **RFC4276 opisuje raport implementacyjny RFC4271**

  **RFC4277 dodatkowo opisuje doświadczenia z różnymi implementacjami**

# Protokół BGP

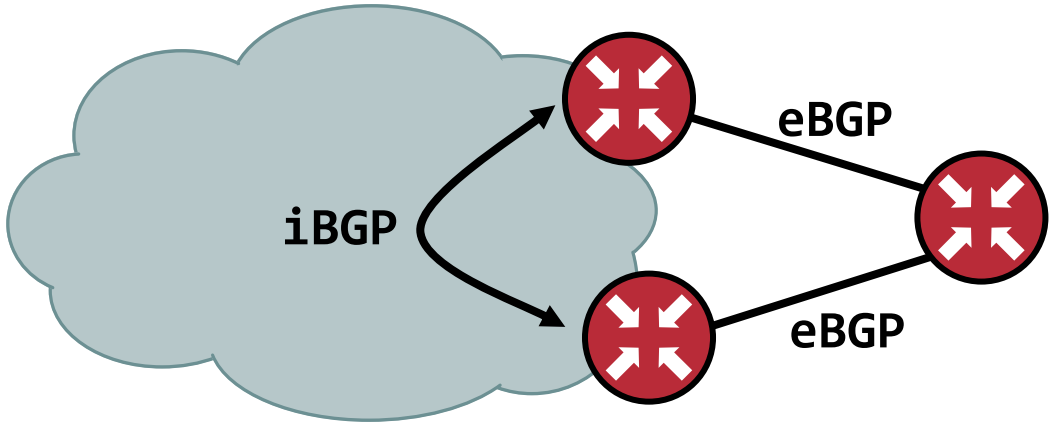
## System Autonomiczny



**AS100**

- Zwykle jedna firma/organizacja, zarządzana przez jedną grupę ludzi
- Jedna polityka routingu wewnętrznego i zewnętrznego
- ASN 0 i 65535 zarezerwowane, 23456 również (do przejścia na 32-bitową notację), natomiast 1-64511 zarządzane centralnie (publiczne)
- 64512-65534 do prywatnego użytku
- Obecnie zarejestrowano trochę ponad 39935 ASN, z czego około 21000 jest widocznych w globalnych tablicach routingu

# Protokół BGP
## internal BGP vs external BGP




- **iBGP – sesje pomiędzy routerami w tym samym AS**

  **wszystkie routery wewnątrz AS muszą nawiązać sesje każdy z każdym (można to obejść przez konfederacje/klastry)**

- **eBGP – sesje pomiędzy routerami w różnych AS**

  **domyślnie połączenie bezpośrednie, należy wprost wskazać że połączenie jest multihop**

# 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

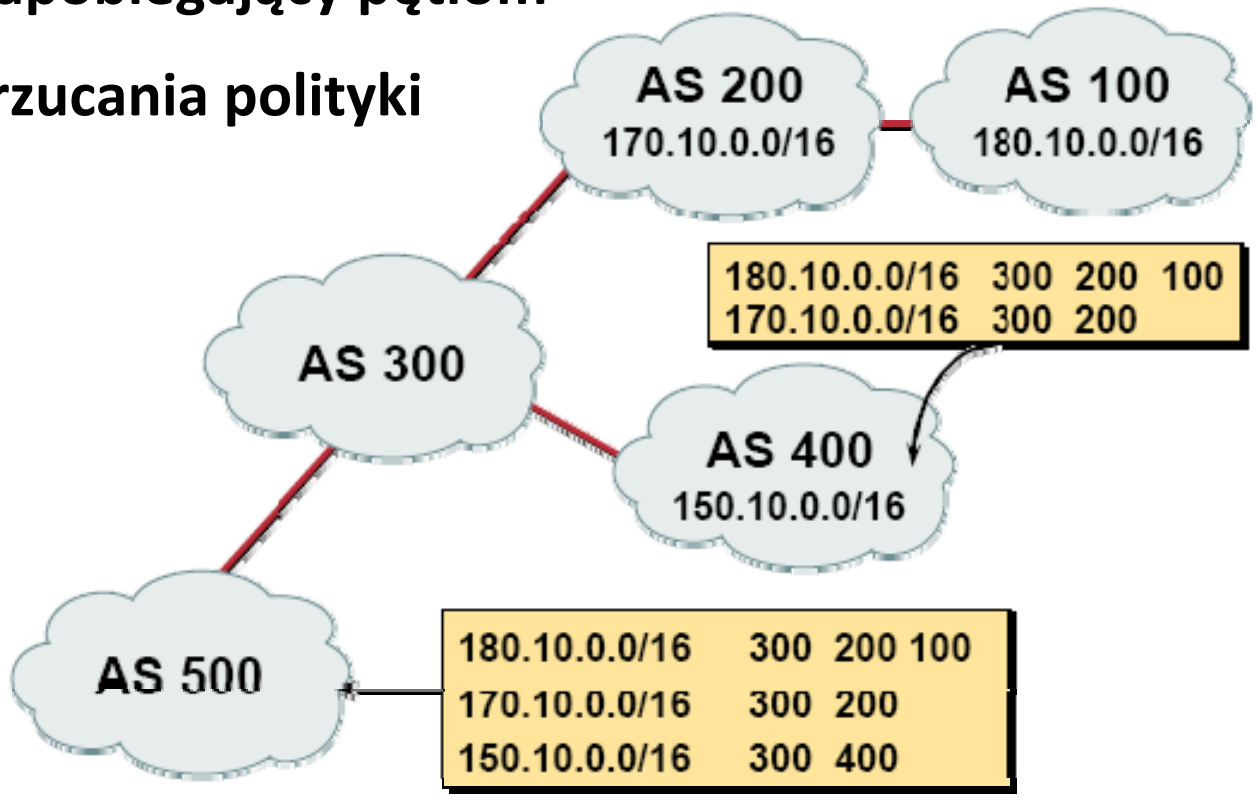
14: MP_REACH_NLRI

15: MP_UNREACH_NLRI

# Protokół BGP

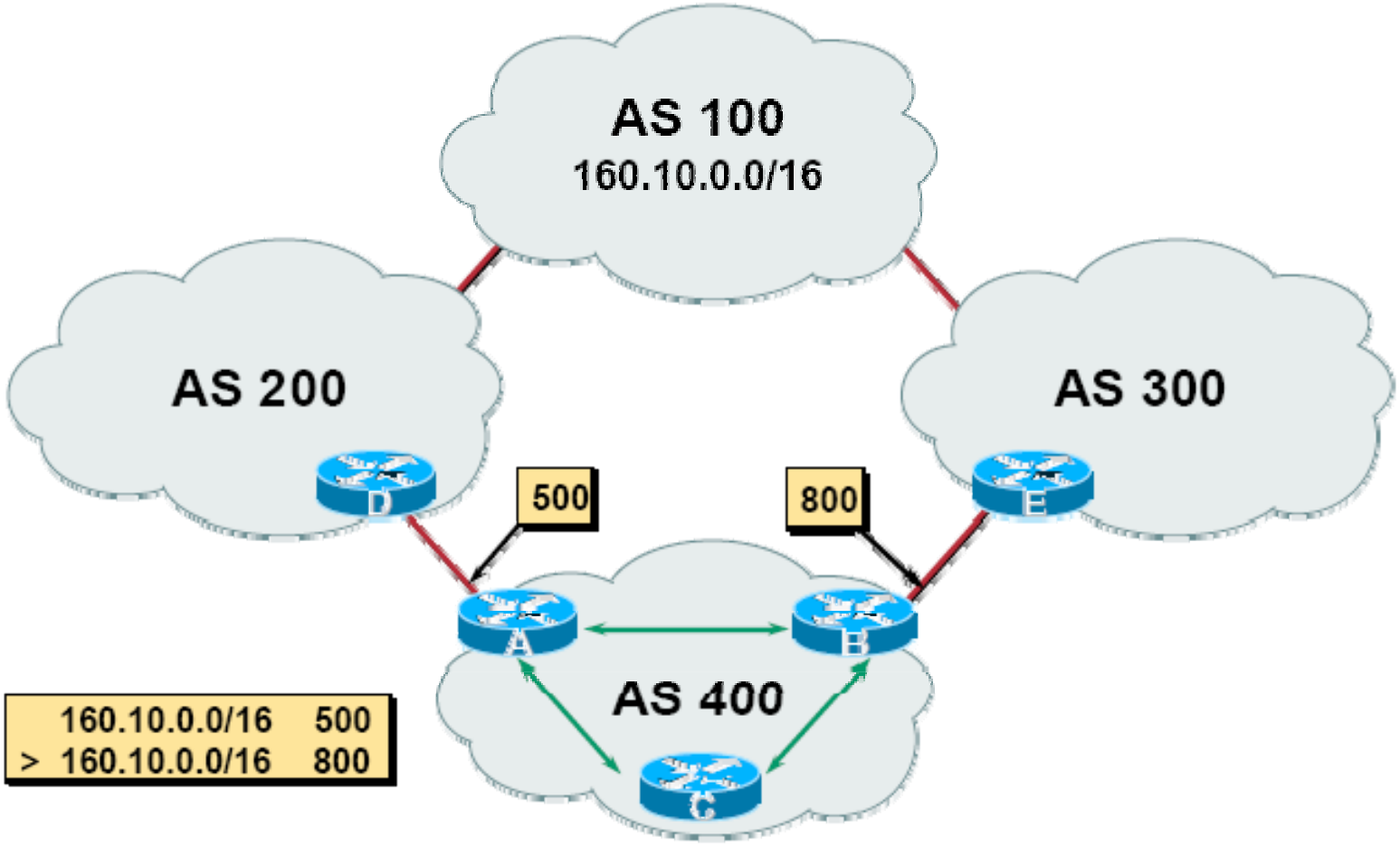
## Atrybuty – AS-Path

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



# Protokół BGP

## Atrybuty – local preference



# Protokół BGP

## Atrybuty – community

- Opisane w RFC1997, mogą być przekazywane pomiędzy ASami (transitive) i są opcjonalne

  32 bitowa wartość dodatnia

- Dla wygody zapisywana jako dwie 16-bitowe wartości rozdzielone dwukropkiem

  Standardowo <NUMER-AS>:XXXX

  Na przykład: 64999:666

- Bardzo przydatne i często spotykane w publicznych peeringach do sterowania polityką dotyczącą oznaczonego prefiksu

# Protokół BGP

## Atrybuty – community

- Opisane w RFC19
  pomiędzy ASami (
  - 32 bitowa wartość
- Dla wygody zapisy
  wartości rozdzielo
  - Standardowo <NU
  - Na przykład: 6499
- Bardzo przydatne
  publicznych peeri
  dotyczącą oznaczo

```
Internet Partners BGP community support
e-mail contact: <bgp4@ipartners.pl>
-------------------------------------
Communities to control traffic (settable by peers):

8246:2000 Do not announce to GTS CE (AS5588)
8246:2001 Prepend +1 when announcing to GTS CE
8246:2002 Prepend +2 when announcing to GTS CE
8246:2003 Prepend +3 when announcing to GTS CE

8246:2011 Prepend +1 when announcing to T-Systems
8246:2012 Prepend +2 when announcing to T-Systems
8246:2013 Prepend +3 when announcing to T-Systems

8246:2100 Do not announce to TPNET (AS5617)
8246:2101 Prepend +1 when announcing to TPNET
8246:2102 Prepend +2 when announcing to TPNET
8246:2103 Prepend +3 when announcing to TPNET

8246:2200 Do not announce to NASK (AS8308)
8246:2201 Prepend +1 when announcing to NASK
8246:2202 Prepend +2 when announcing to NASK
8246:2203 Prepend +3 when announcing to NASK

8246:2300 Do not announce to POL34 (AS8501)

8246:2400 Do not announce to Sunsite ICM (AS8664)
8246:2401 Prepend +1 when announcing to Sunsite ICM
8246:2402 Prepend +2 when announcing to Sunsite ICM
8246:2403 Prepend +3 when announcing to Sunsite ICM

8246:2411 Prepend +1 when announcing to WP
8246:2412 Prepend +2 when announcing to WP
8246:2413 Prepend +3 when announcing to WP
```

# 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)

# CZY MUSZĘ UŻYWAĆ BGP?

# Czy na pewno muszę mieć BGP?

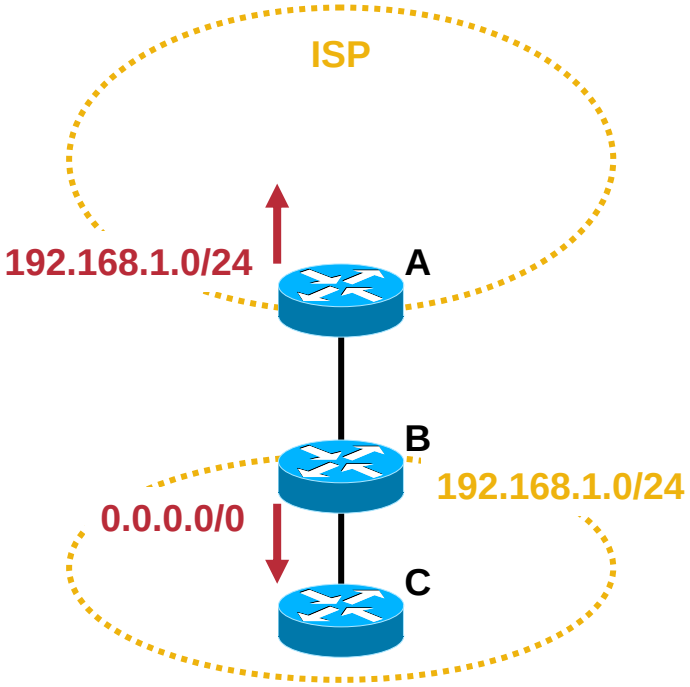
- **Pojedyncze połączenie do Internetu - NIE**

  **w sieci używasz domyślnej trasy (statycznie lub protokół routingu)**

  **Twój ISP zapewnia widoczność i osiągalność przydzielonej Ci adresacji IP**

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

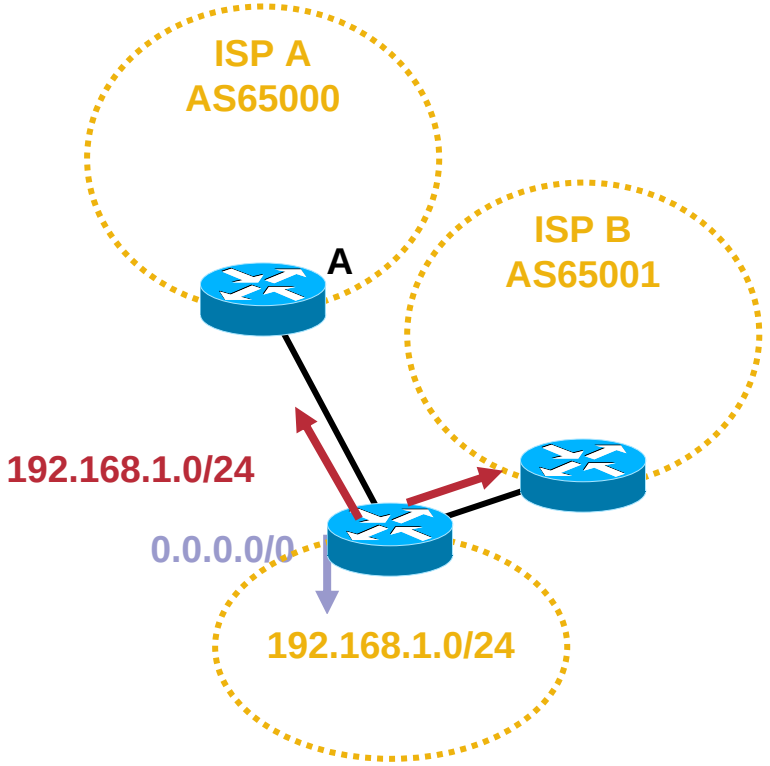
  **wiele tras domyślnych na różne IP po stronie Twojej i ISP**



# Czy na pewno muszę mieć BGP?

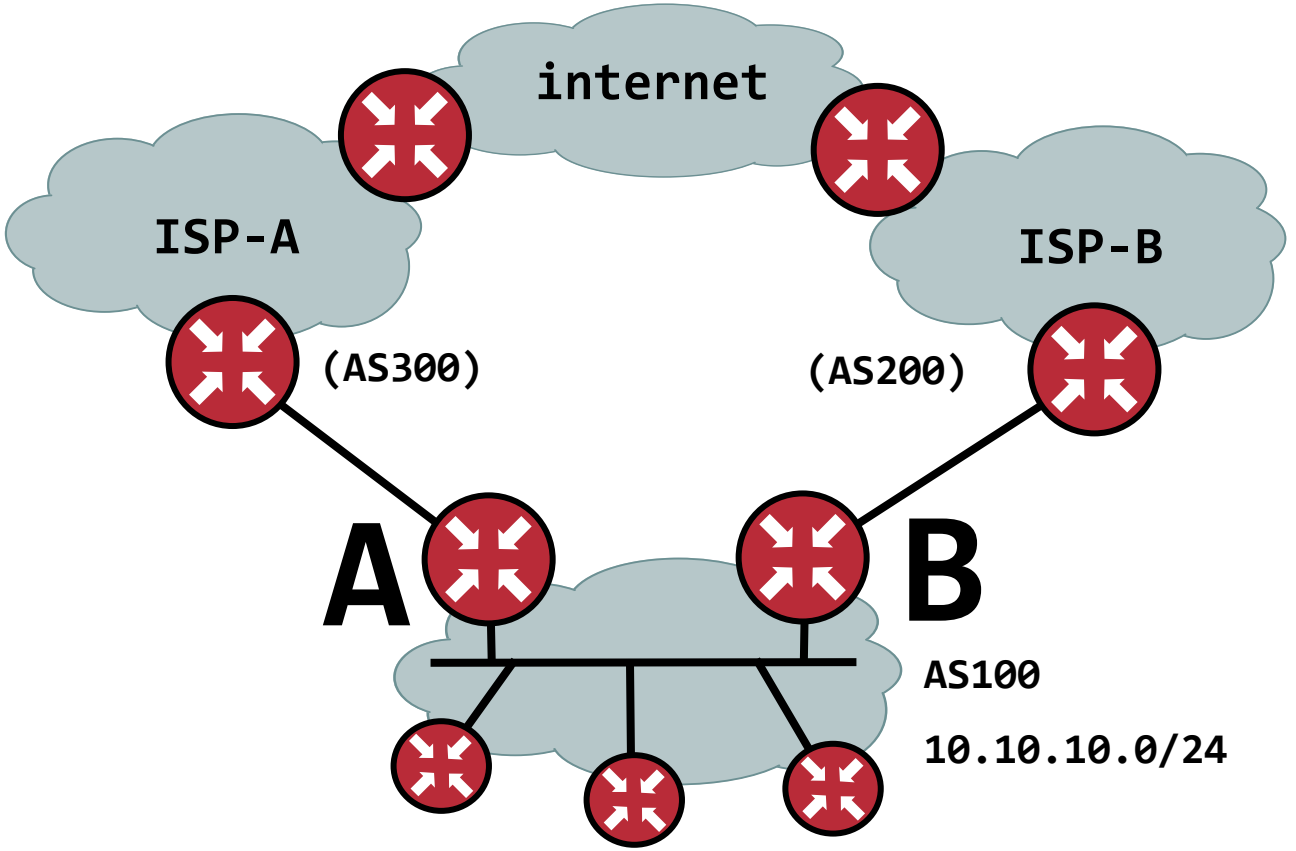
- Jeśli jesteś połączony do dwóch różnych ISP – BGP jest pożądane – rozgłasza do obu swoją pulę adresów, Internet widzi ją przez obu ISP
- Nie oznacza to, że musisz pobierać wszystkie światowe prefiksy

  ...pozwala to jednak ‘świadomie’ kształtować własną politykę routingu dużo dokładniej



# PRZYKŁAD ZASTOSOWANIA BGP

# ...nasza sieć:

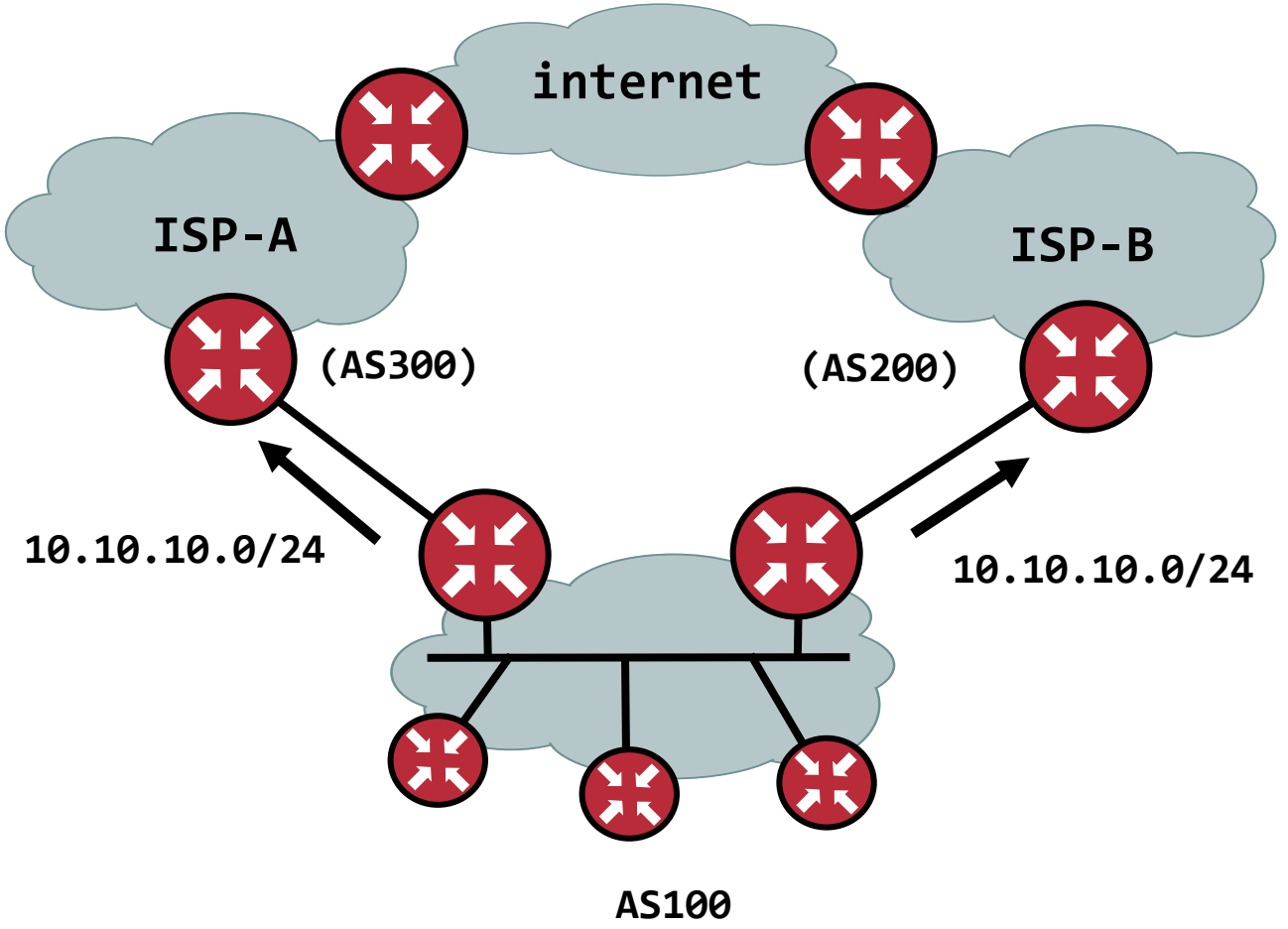


# Protokół BGP

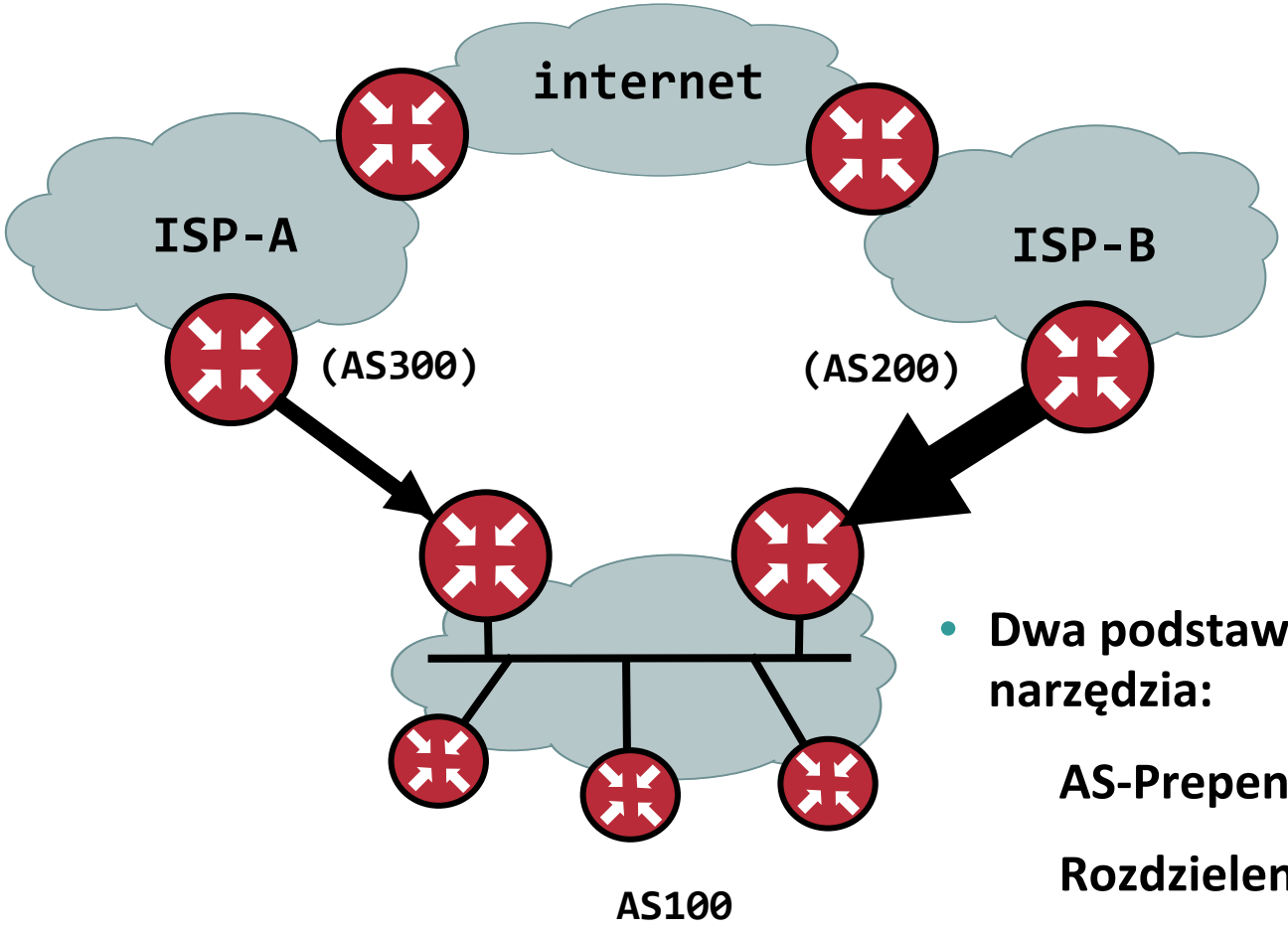
## Przykładowa konfiguracja – router A – prefix oraz sesje eBGP

```
router bgp 100
 bgp router-id 10.10.10.253
 network 10.10.10.0 mask 255.255.255.0
 aggregate-address 10.10.10.0 255.255.255.0 summary-only

 neighbor 172.16.10.1 remote-as 200
 neighbor 172.16.10.1 description AS200
 neighbor 172.16.10.1 version 4
 neighbor 172.16.10.1 password AJAX*PERSIL*E

 neighbor 10.10.10.2  remote-as 100
 neighbor 10.10.10.2  description router-B
 neighbor 10.10.10.2  version 4
 neighbor 10.10.10.2  password BEZ*MD5*JEST*ZIA-ZIA
```

# ...nasza sieć:



# Przykładowy problem:
## łącze z ISP-B jest przeciążone




- Dwa podstawowe narzędzia:
  - AS-Prepend
  - Rozdzielenie prefiksów

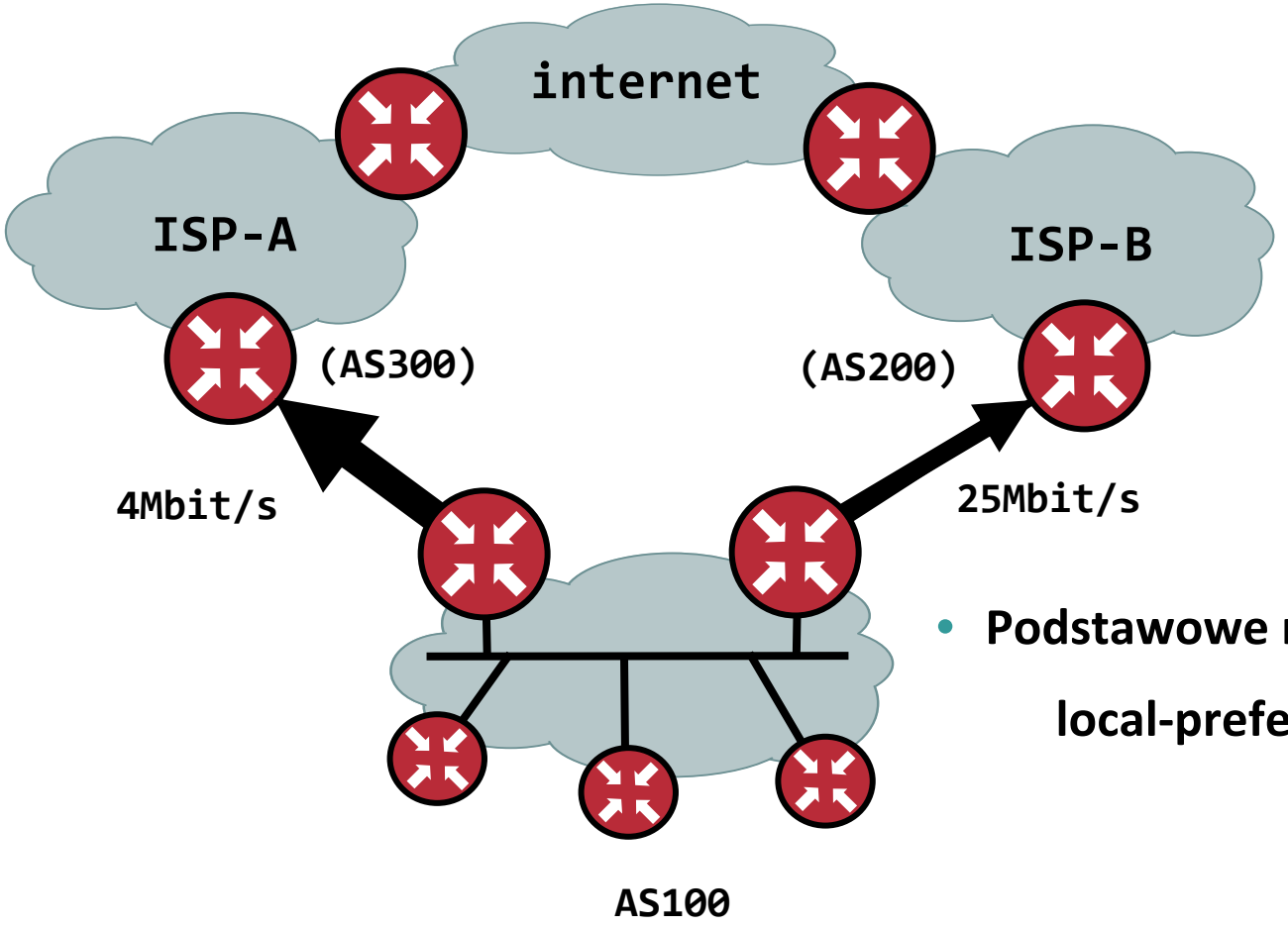
# Przykładowy problem:
## łącze z ISP-B jest przeciążone

- „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 KuLepszejPrzyszlosci out

route-map KuLepszejPrzyszlosci permit 10
 set as-path prepend 100 100
```

```
AS200rtr# 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
```

# Przykładowy problem:
## większość ruchu wychodzi przez ISP-A




- **Podstawowe narzędzie:**

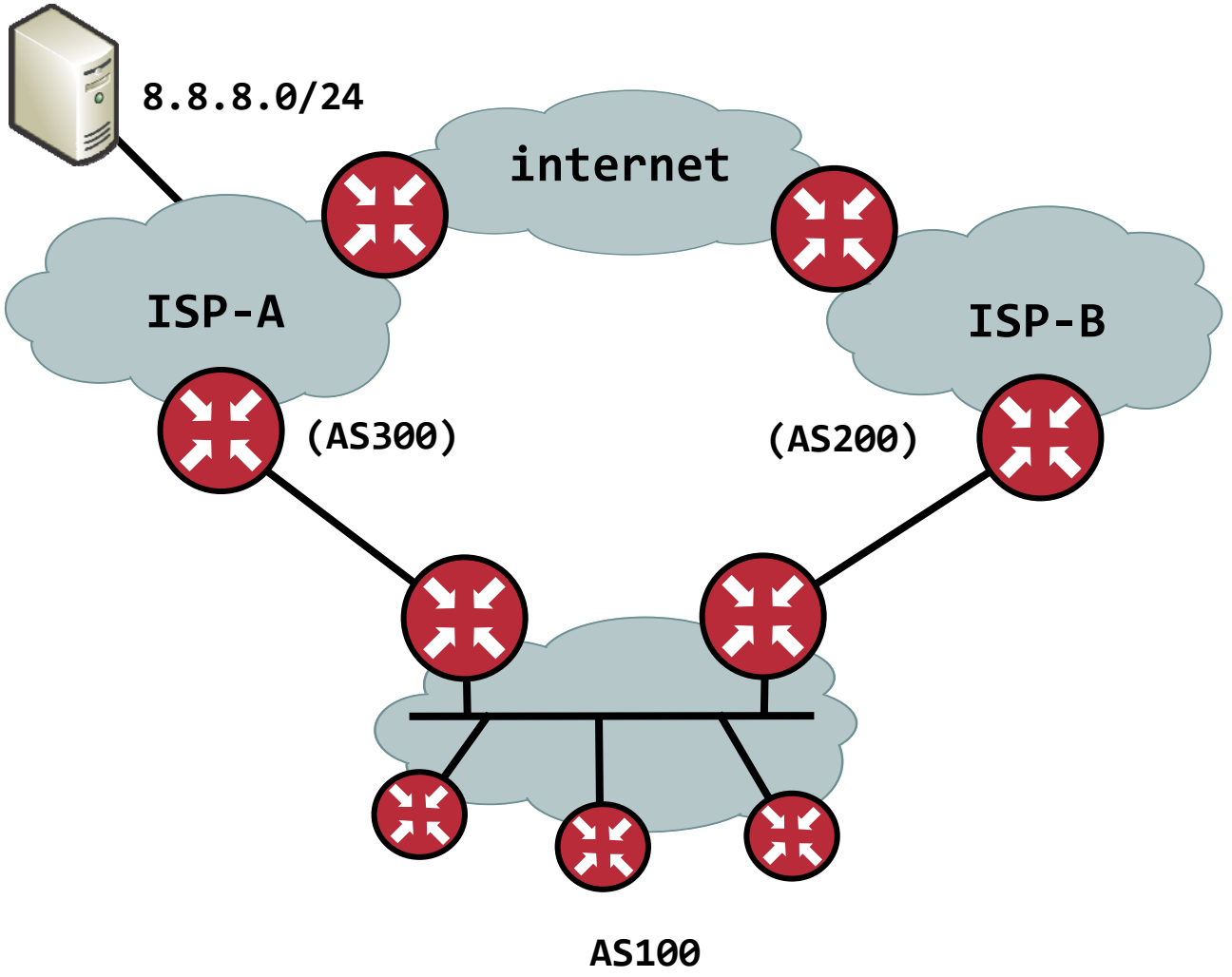
  **local-preference**

# Przykładowy problem:
## większość ruchu wychodzi przez ISP-A

- **Wszystkie** prefiksy otrzymane z AS200 są lepsze niż inne, z domyślnym (100) local-preference

```
router bgp 100
 neighbor 172.16.10.1 remote-as 200
 neighbor 172.16.10.1 route-map KuLepszejPrzyszlosci in

route-map KuLepszejPrzyszlosci permit 10
 set local-preference 5000
```

# Przykładowy problem:
## do sieci 8.8.8.0/24 zawsze lepiej przez ISP-A



# Protokół BGP

## Prefiks 8.8.8.0/24 w pierwszej kolejności przez ISP-A

- Tylko prefiks 8.8.8.0/24 lub bardziej dokładny jest zawsze lepszy przez AS300

```
router bgp 100
 neighbor 172.16.20.1 remote-as 300
 neighbor 172.16.20.1 route-map KuLepszejPrzyszlosci in

ip prefix-list JedynaDroga permit 8.8.8.0/24 le 32

route-map KuLepszejPrzyszlosci permit 10
 match ip prefix-list JedynaDroga
 set local-preference 5000
```

Dobry opis jak działają prefix-listy:
http://www.groupstudy.com/archives/ccielab/200404/msg00539.html

# HACKING ROUTING FOR DUMMIES (FUN, FUN, FUN?)

# First-Hop Redundancy

- Standardy: VRRP
- Cisco-made: HSRP i GLBP
- Inne: CARP
- Idea: podstawić jeden wirtualny router w miejsce wielu fizycznych i zapewnić obsługę routingu

# Protokół OSPF

## Problem redundancji first-hop

- **VRRP – standard (RFC3768)**

  **VRRPd: http://off.net/~jme/vrrpd/**

  **FreeVRRPd**

- **CARP/UCARP – analogiczne dla OpenBSD/innych systemów**

  **http://www.ucarp.org/project/ucarp**

- **Integracja aplikacji i systemu operacyjnego:**

  **LVS: http://www.linuxvirtualserver.org/**

  **ct_sync (netfilter) + keepalived + LVS/coś innego:**

  **http://svn.netfilter.org/cgi-bin/viewcvs.cgi/trunk/netfilter-ha/**

  **http://www.netfilter.org/projects/libnetfilter_conntrack/index.html**

# Route Flap Dampening

- **Co to jest „klapnięcie” trasy?**

  **zniknięcie i ponowne pojawienie się trasy w krótkim okresie czasu**

- **Takie „klapnięcie” propaguje się przez cały Internet**

  **...obciąża wiele różnych CPU, powoduje niepotrzebny ruch**

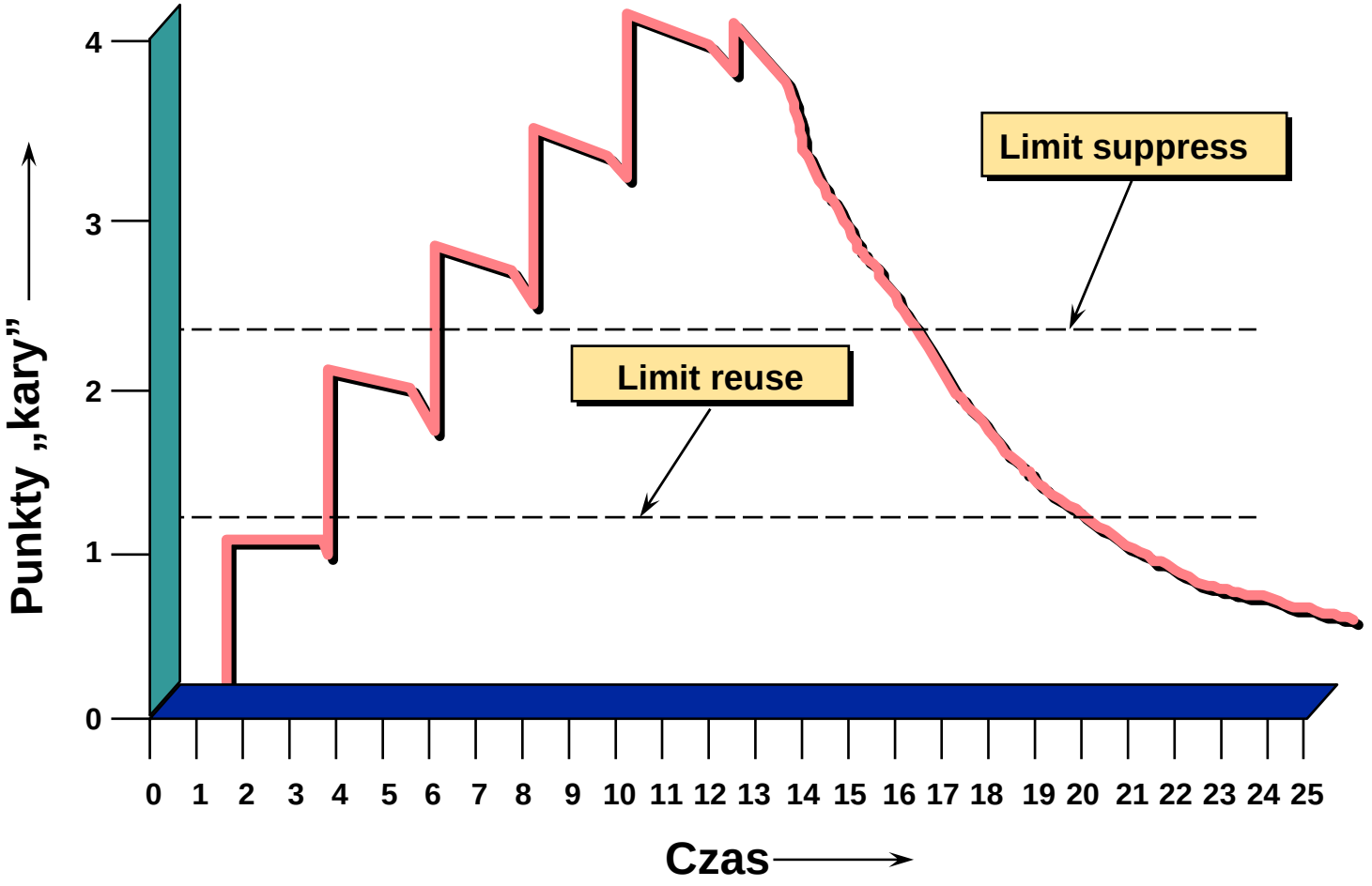
- **Rozwiązanie: ograniczmy zasięg propagacji informacji o zmianie stanu osiągalności prefiksu**

# Route Flap Dampening

- Za każde „klapnięcie” dodajemy punkty kary

  prefiksy bez ‘klapnięć’ przez czas określany parametrem ‘half-time’ czas mają zmniejszane punkty o połowę
- Jeśli ilość punktów wzrośnie powyżej limitu suppress – przestajemy rozgłaszać prefiks dalej
- Jeśli ilość punktów z czasem spadnie poniżej limitu reuse – zaczynamy ją rozgłaszać
- Dampening dotyczy tylko tras zewnętrznych
- Inne trasy równoległe pozostają do wykorzystania

# Route Flap Dampening

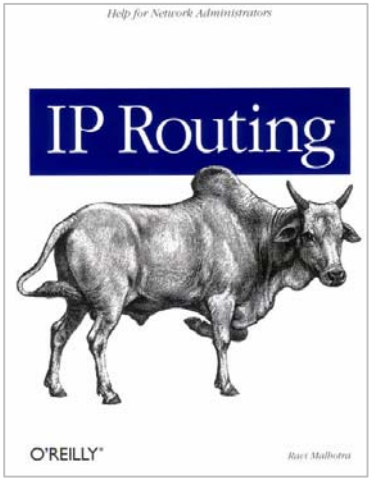


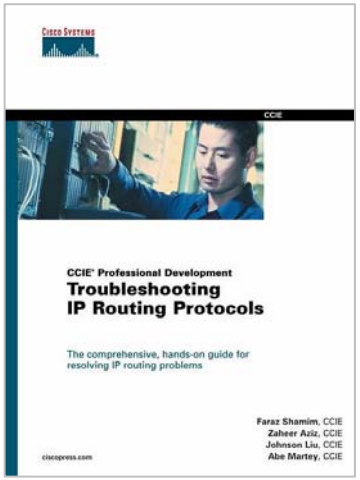
# Route Flap Dampening

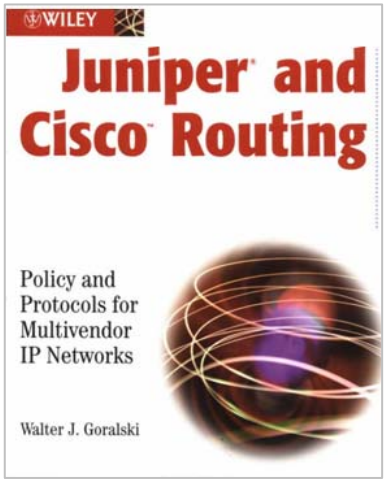
```
router bgp 100
 bgp dampening A B C D

! A = <1-45>      czas half-life
! B = <1-20000>   wartość od której rozgłaszamy prefiks
! C = <1-20000>   wartość od której wstrzymujemy prefiks
! D = <1-255>     ile maksymalnie czasu blokować prefiks

 bgp dampening 5 1500 2500 50
```

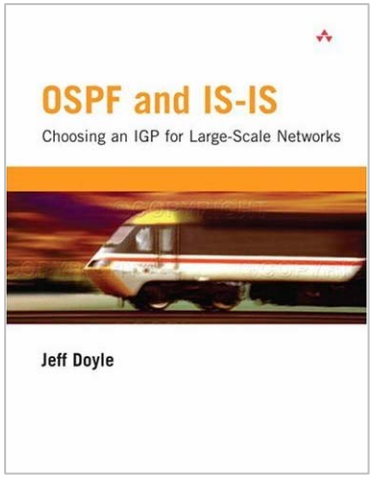
# GDZIE WARTO RZUCIĆ OKIEM

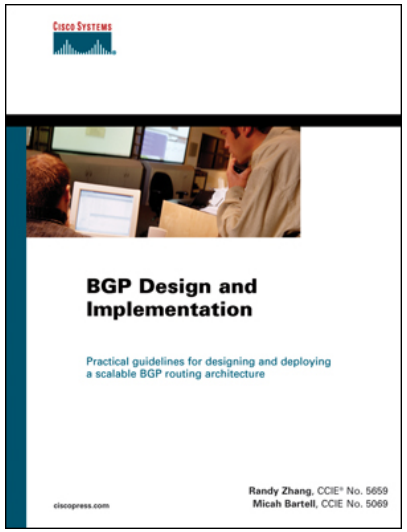
# Książki

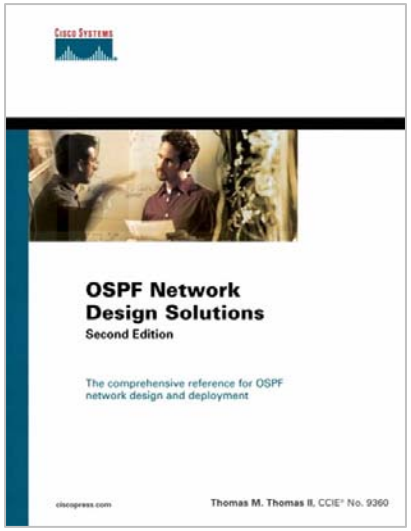

Help for Network Administrators
IP Routing
O'REILLY

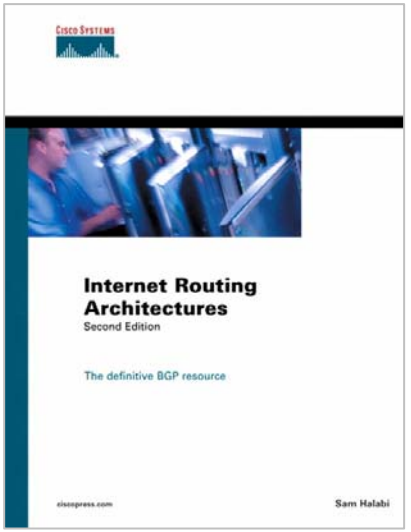

CCIE
CCIE Professional Development
Troubleshooting
IP Routing Protocols
The comprehensive, hands-on guide for resolving IP routing problems
Faraz Shamim, CCIE
Zaheer Aziz, CCIE
Johnson Liu, CCIE
Abe Martey, CCIE

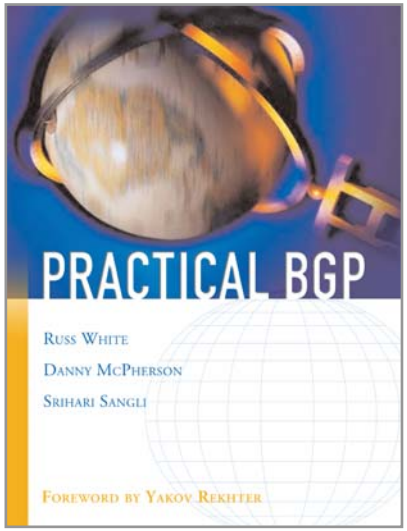

WILEY
Juniper and Cisco Routing
Policy and Protocols for Multivendor IP Networks
Walter J. Goralski


OSPF and IS-IS
Choosing an IGP for Large-Scale Networks
Jeff Doyle


BGP Design and Implementation
Practical guidelines for designing and deploying a scalable BGP routing architecture
Randy Zhang, CCIE No. 5659
Micah Bartell, CCIE No. 5069


OSPF Network Design Solutions
Second Edition
The comprehensive reference for OSPF network design and deployment
Thomas M. Thomas II, CCIE No. 9360


Internet Routing Architectures
Second Edition
The definitive BGP resource
Sam Halabi


PRACTICAL BGP
Russ White
Danny McPherson
Srihari Sangli
Foreword by Yakov Rekhter

# Zasoby WWW

- **Pakiet Quagga:**

  http://www.quagga.net

- **Pakiet XORP:**

  http://www.xorp.org

- **Demon OpenOSPFd/OpenBGPd:**

  http://www.openbsd.org

- **Einar – LiveCD wykorzystujące Xen i Quaggę:**

  http://www.isk.kth.se/proj/einar/

- **IP routing protocols home page (Cisco):**

  http://www.cisco.com/en/US/tech/tk365/tsd_technology_support_protocol_home.html

- **RFC/standardy dla protokołów routingu:**

  http://lukasz.bromirski.net/links/routing_ip.html

# Q&A


NO
QUESTIONS
VIOLATORS
WILL BE SHOT
SURVIVORS
WILL BE SHOT
AGAIN