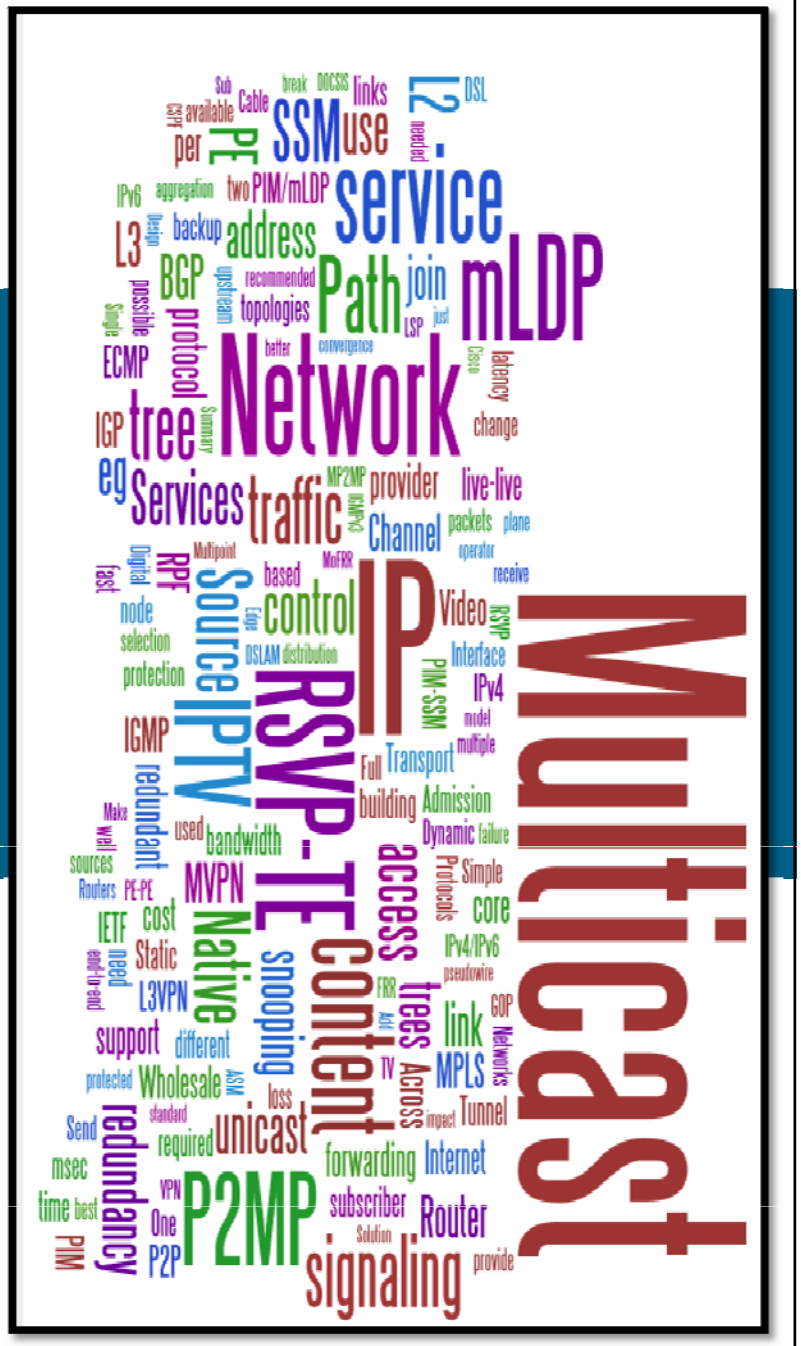
Budowa sieci multicast



Łukasz Bromirski
lbromirski@cisco.com

Kraków, 09/2009

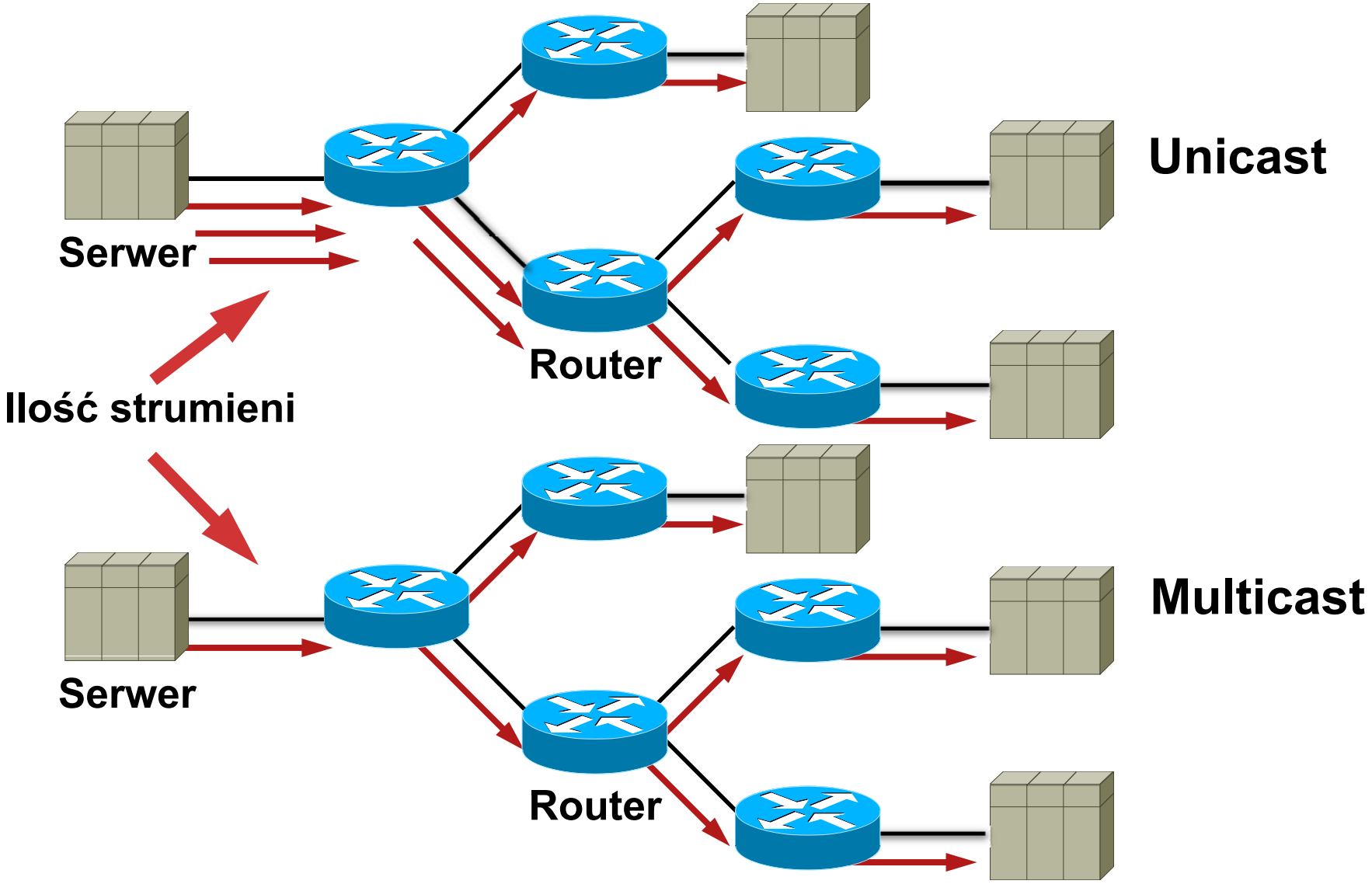
PLNOG

Agenda

- To po co jest ten multicast?
- Podstawy multicastów
- Protokół PIM
- Wybór RP
- Q&A

To po co jest ten multicast?

Unicast vs. Multicast

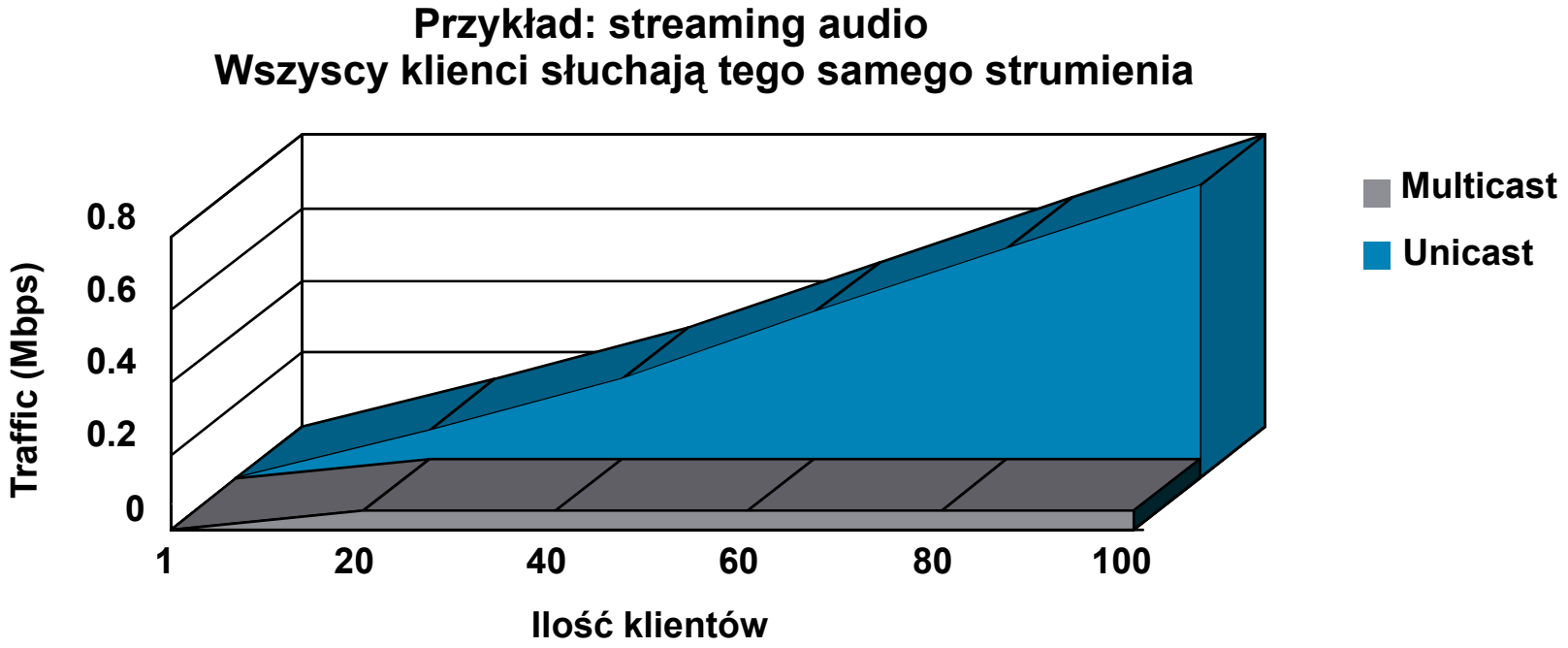


Zastosowanie multicastów

- Dowolna aplikacja która przesyła dane do wielu odbiorców
  - W modelu jeden-do-wielu lub wielu-do-wielu
- Dystrybucja wideo transmitowanego na żywo
- Oprogramowanie do współpracy grupowej
- Okresowe ‘wypychanie’ danych do różnego rodzaju aplikacji
  - Wiadomości giełdowe, wyniki sportowe, ogłoszenia (w tym webowe adsy!)
- Replikacja serwerów/stron WWW
- Odkrywanie i mapowanie zasobów w sieci
- Rozproszone symulacje interaktywne (DIS)
  - Gry wojenne
  - Wirtualna rzeczywistość

Zalety wykorzystania multicastów

- **Większa efektywność:** mniejsza ilość ruchu sieciowego w szkielecie/dystrybucji, mniejsze obciążenie CPU nadającego
- **Optymalna wydajność:** ruch nie jest wysyłany wielokrotnie
- **Rozproszone aplikacje:** aplikacje korzystające z transmisji wielopunktowej stają się praktyczne – i dostępne



Unicast a Multicast

- TCP - unicast OK, multicast nie

  TCP jest protokołem zorientowanym na połączenia

  Wymaga three-way handshake

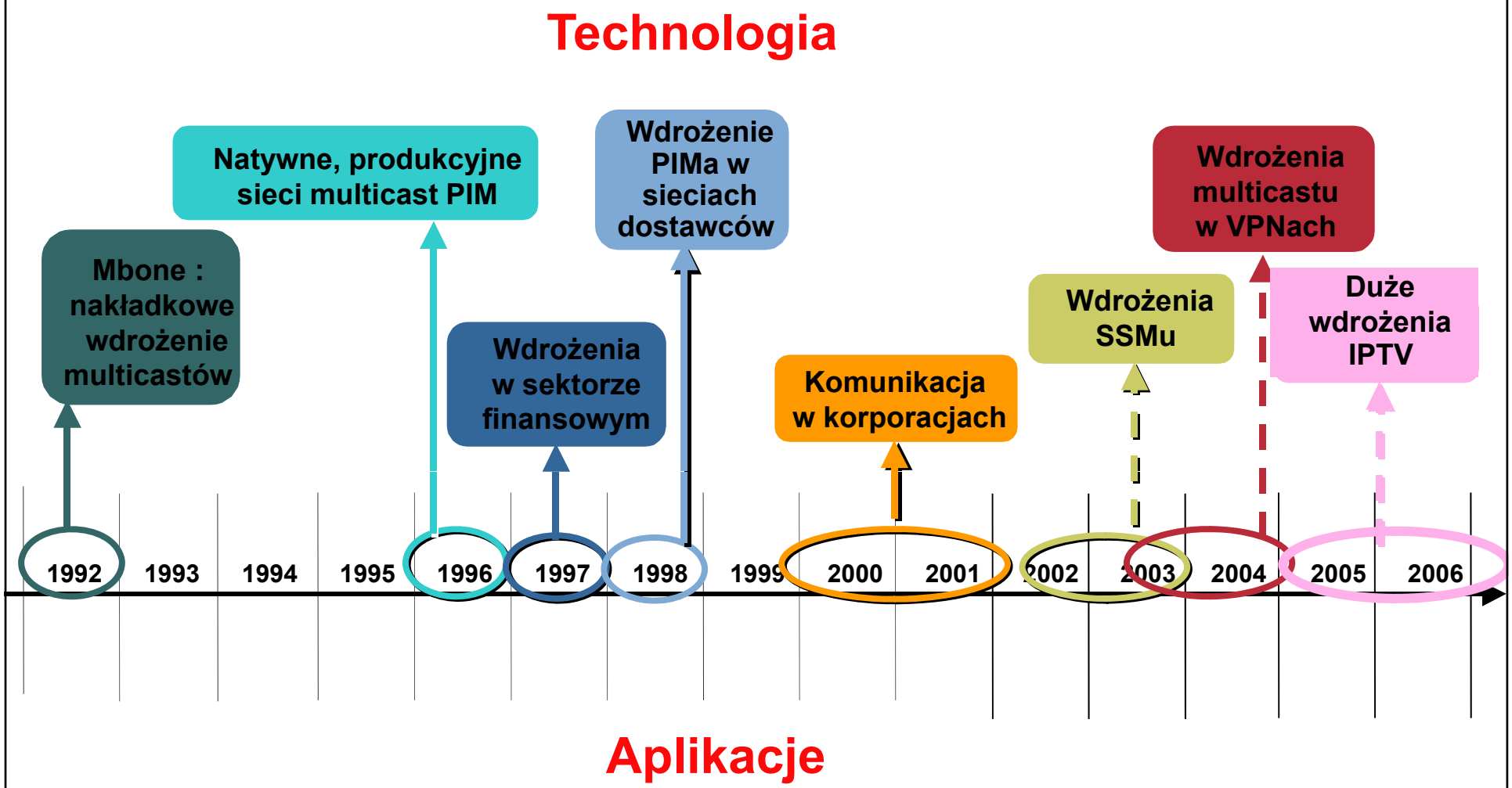
  Zapewnia obsługę kontroli przepływu danych oraz dostarczania z potwierdzeniem dzięki numerom sekwencyjnym i potwierdzeniom

- UDP - unicast i multicast OK

  Protokół bezpołączeniowy

  To aplikacja ma zapewnić funkcje takie jak kontrola przepływu czy potwierdzenia/retransmisje

Historia technologii IP multicast



Ewolucja protokołów multicastowych

BGP

IGP
OSPF

RPF-flooding

Spanning-Tree-flooding

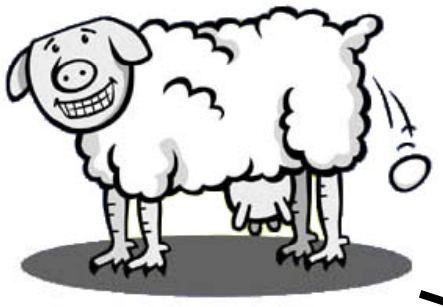
MOSPF

PIM-DM

DVMRP

CBT

ASM:
PIM-SM / MSDP
MBGP, AutoRP
Anycast-RP



“eierlegende Wollmilchsau”
(niem.)
= składająca jajka, dająca mleko i wełnę owca, będąca również źródłem mięsa i mleka

RPFv4v6:
MT-technology
IGP + BGP

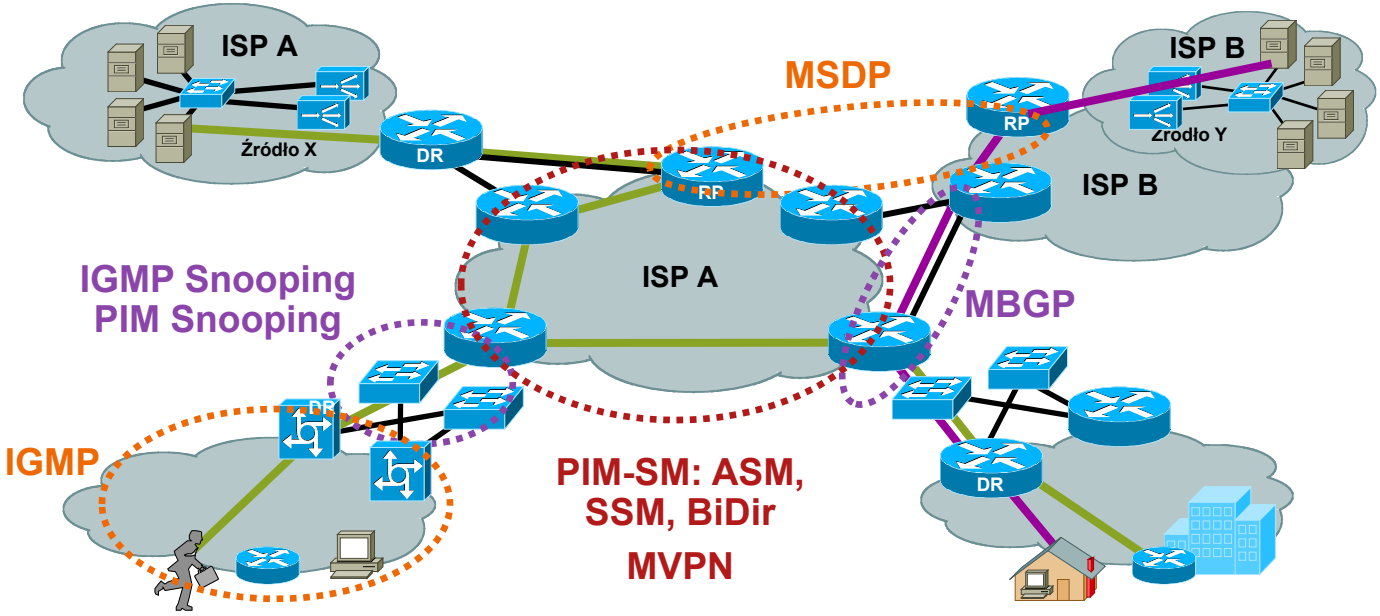
ASMv6:
PIM-SM+
Embedded-RP
Intra/Interdomain

ASMv4/v6:
Bidir-PIM
Intradomain

SSMv4v6:
PIM-SSM
Intra/Interdomain

Podstawy multicastów

Elementy architektury multicastowej



- Stacje końcowe – router-do-hosta
  - IGMP

Multicast w sieci LAN/WAN

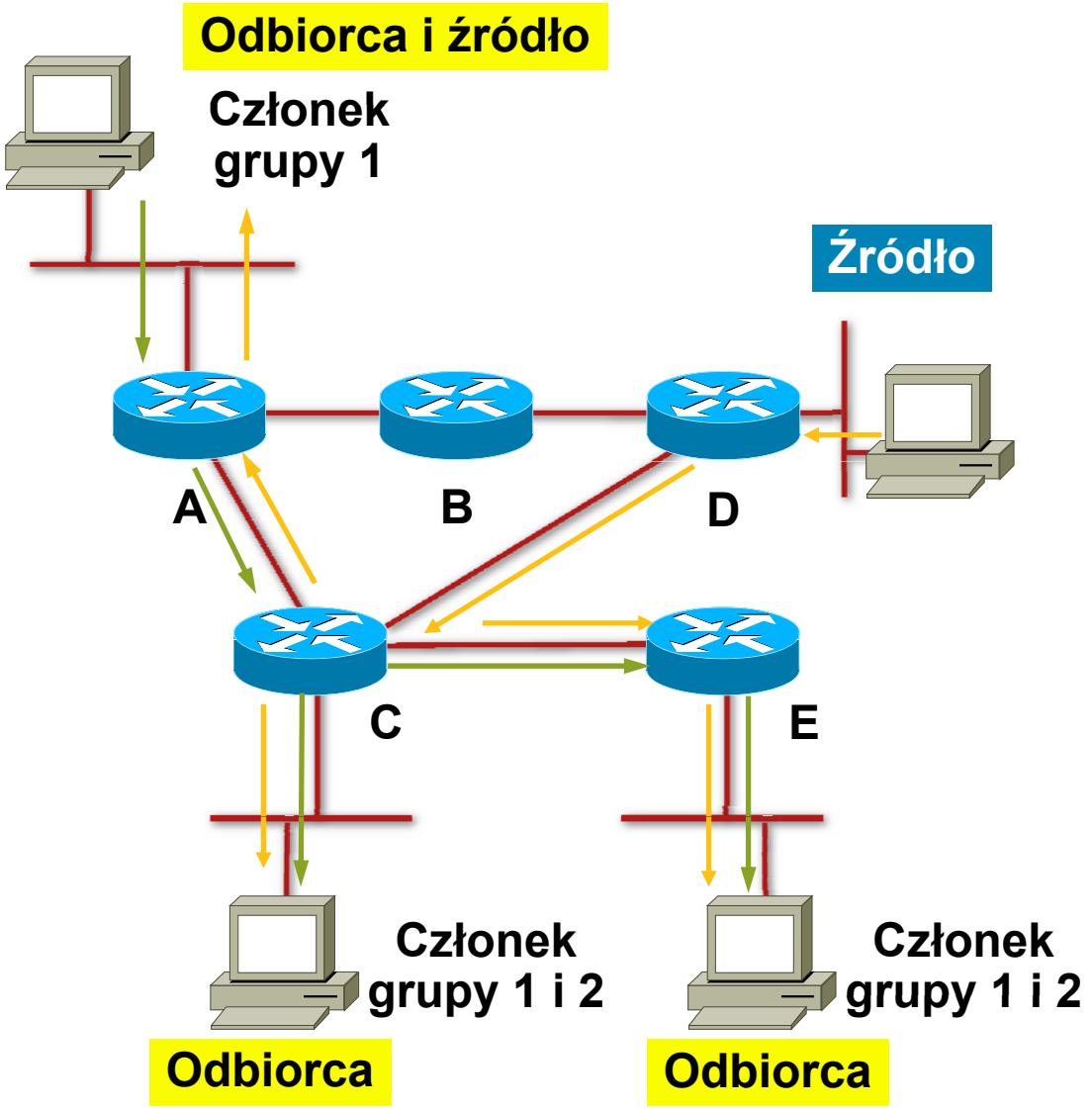
- Przełączniki (optymalizacja L2)
  - IGMP snooping i PIM snooping
- Routery (protokół przekazywania)
  - PIM sparse mode lub bidirectional PIM

Multicast między domenami

- Multicast pomiędzy domenami
  - MBGP
- Multicast source discover
  - MSDP with PIM-SM
- Source Specific Multicast
  - SSM

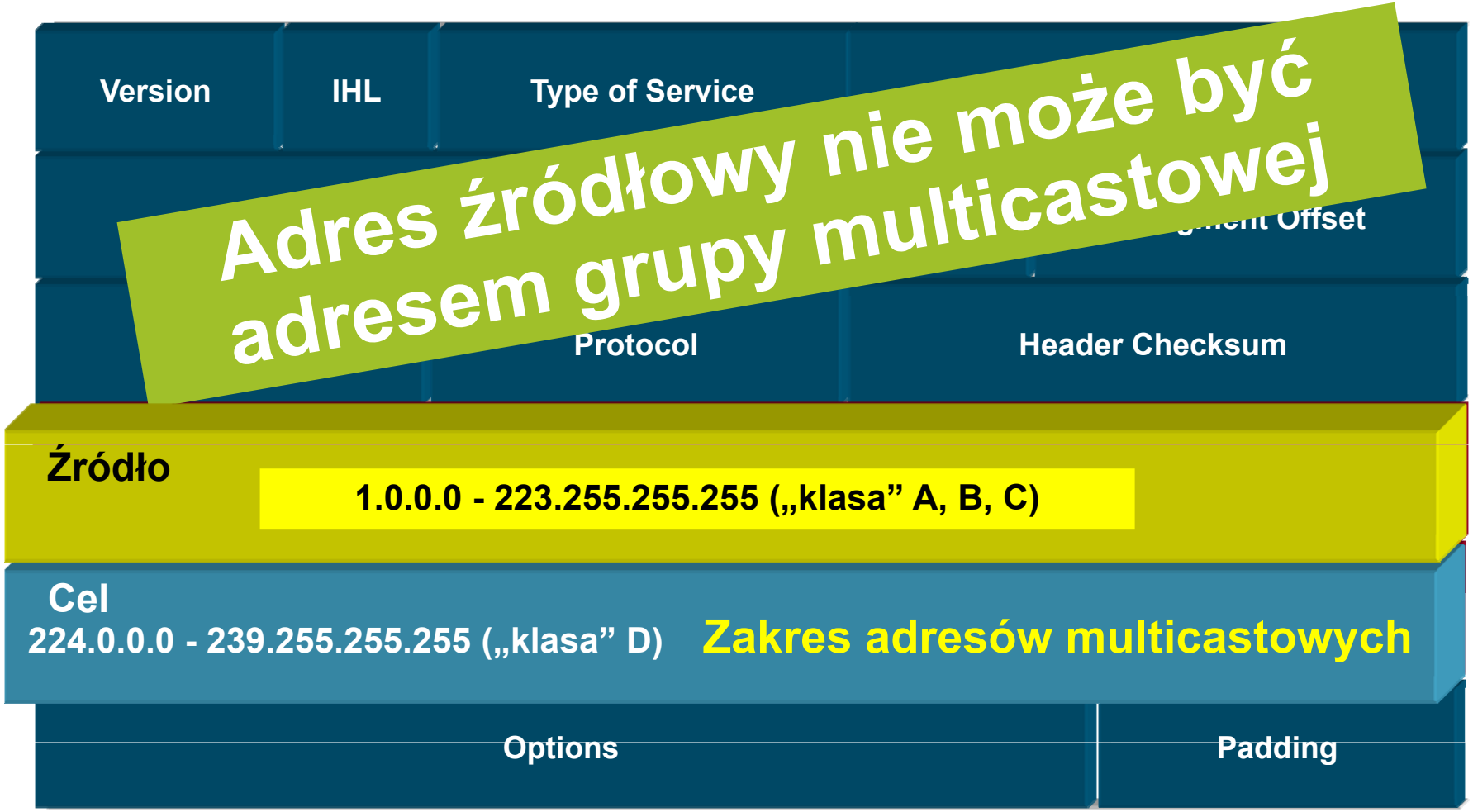
Koncepcja grupy multicastowej

1. Musisz być „członkiem” grupy aby otrzymywać adresowane do niej dane
2. Wysyłając dane na adres grupy – wysyłasz ruch do wszystkich jej członków
3. Nie musisz być członkiem grupy by móc wysyłać do niej dane

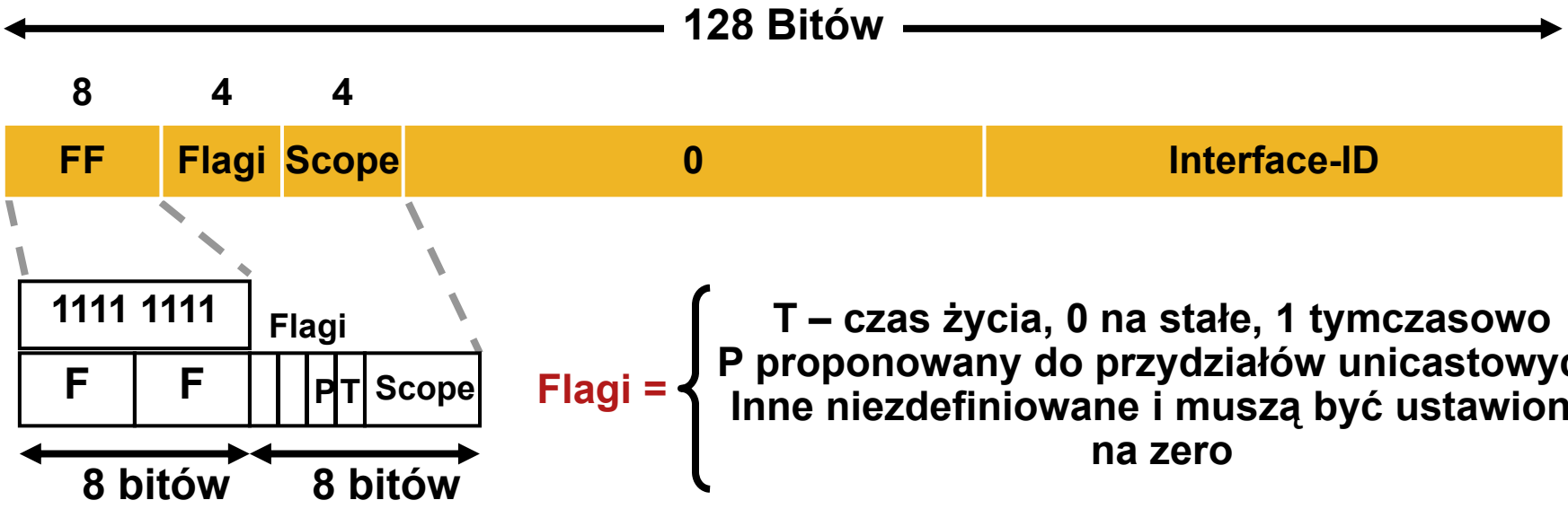


Adresacja multicastów

Nagłówek IPv4



Adresacja multicastów w IPv6 (RFC3513)



Flagi =

- T – czas życia, 0 na stałe, 1 tymczasowo
- P proponowany do przydziałów unicastowych
- Inne niezdefiniowane i muszą być ustawione na zero

Scope =

- 1 = interface-local
- 2 = link
- 4 = admin-local
- 5 = site
- 8 = organization
- E = global

Multicast w IPv4 i w IPv6 - różnice

Cecha	IPv4	IPv6
Zakres adresów	„klasa” D	128 bitowy (grupy 112-bitowe)
Routing	Niezależny od protokołu Wszystkie IGP i BGP4+	Niezależny od protokołu Wszystkie IGP i BGP4+ z SAFI v6 Mcast
Forwarding	PIM-DM, PIM-SM: ASM, SSM, BiDir	PIM-SM: ASM, SSM, BiDir
Zarządzanie przynależnością do grup	IGMPv1, v2, v3	MLDv1, v2
Kontrola nad domeną	Zakres/Brzeg	Identyfikator scope
Wykrywanie źródeł w innych domenach	MSDP pomiędzy domenami PIM	Jeden RP w globalnie współdzielonych domenach

Adresacja multicastów—224/4

- Zarezerwowane adresy łącza lokalnego

  224.0.0.0–224.0.0.255

  Wysyłane z TTL = 1

  przykłady

224.0.0.1	Wszystkie systemy w tej podsieci
224.0.0.2	Wszystkie routery w tej podsieci
224.0.0.5	routery OSPF
224.0.0.13	routery PIMv2
224.0.0.22	IGMPv3

- Inne zarezerwowane adresy

  224.0.1.0–224.0.1.255

  Adresy „zdalne” – transmitowane z TTL>1

  Przykłady

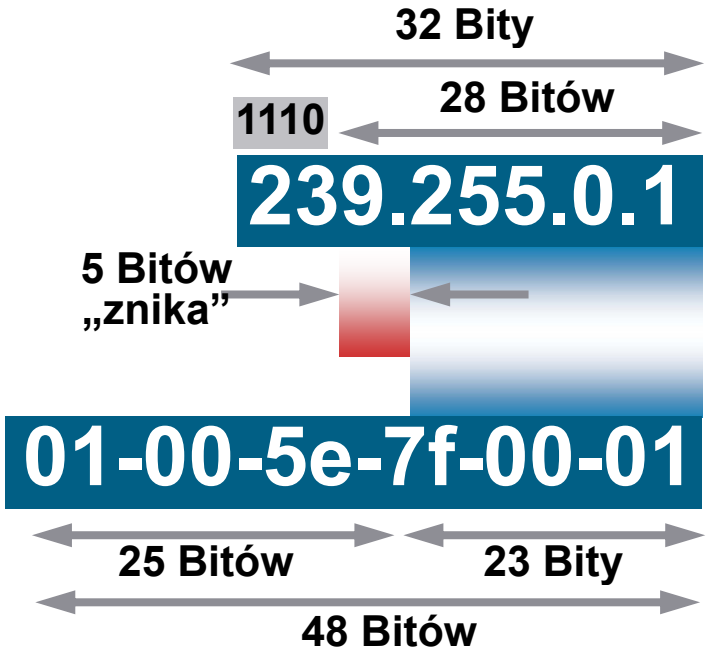
224.0.1.1	NTP (Network Time Protocol)
224.0.1.32	routery mtrace

Adresacja multicastów—224/4

- Adresy przydzielane przez administratora
  - 239.0.0.0–239.255.255.255
  - Przestrzeń prywatna
    - Koncepcja analogiczna do adresów z RFC1918
- Zakres SSM (Source Specific Multicast)
  - 232.0.0.0–232.255.255.255
  - Koncepcja przeniesienia ‘broadcastu’ do Internetu
- GLOP
  - 233.0.0.0–233.255.255.255
  - Przydział grupy /24 per ASN

Adresacja multicastów

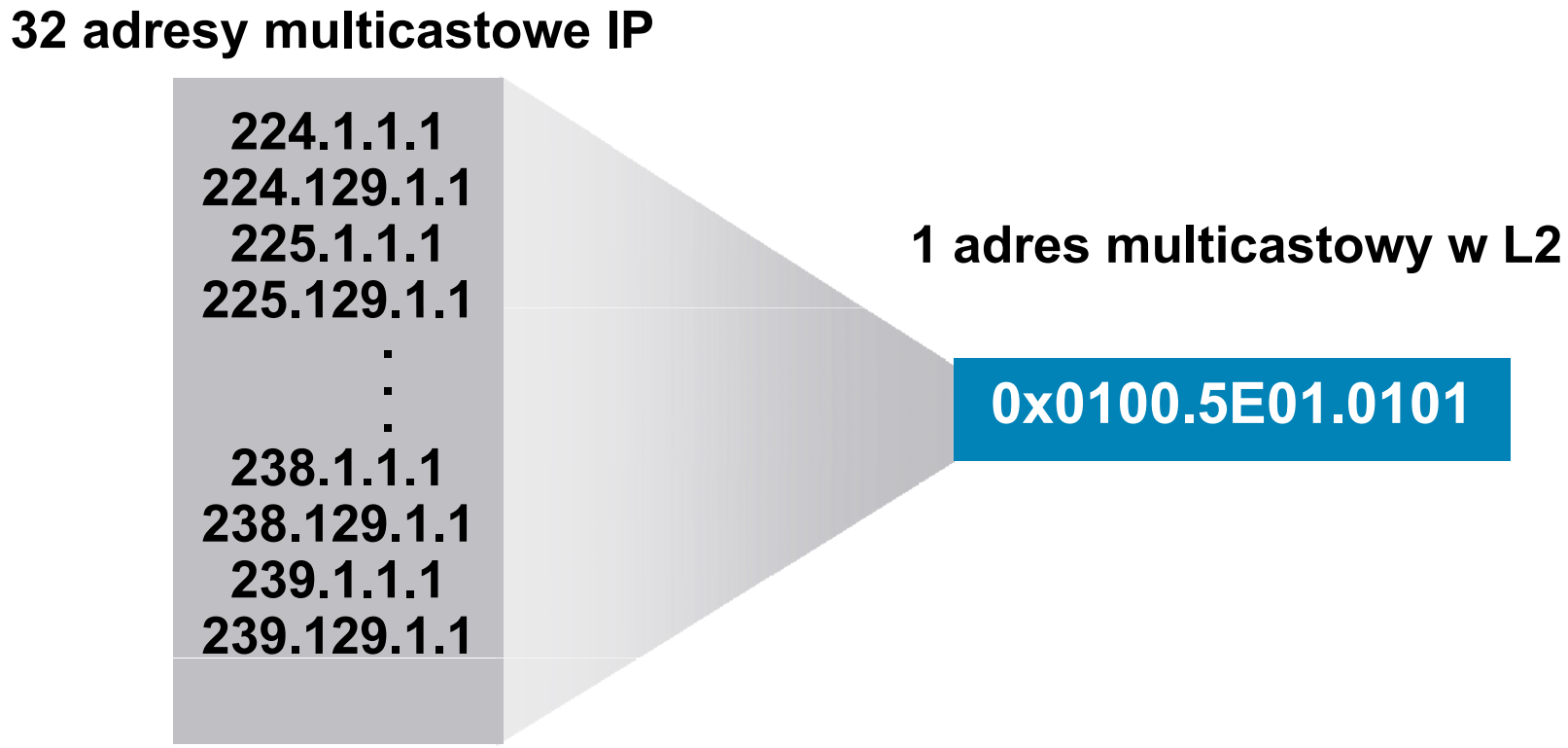
Mapowanie adresów MAC na IP



Adresacja multicastów

Mapowanie adresów multicastów IP na MAC

Nakładanie 32:1



Sygnalizacja host-router

Sygnalizacja host-router: IGMP

- Sposób poinformowania routera przez host o chęci przynależenia do konkretnej grupy
- Routery zbierają informacje z podłączonych bezpośrednio hostów
- RFC 1112 opisuje wersję 1 protokołu IGMP

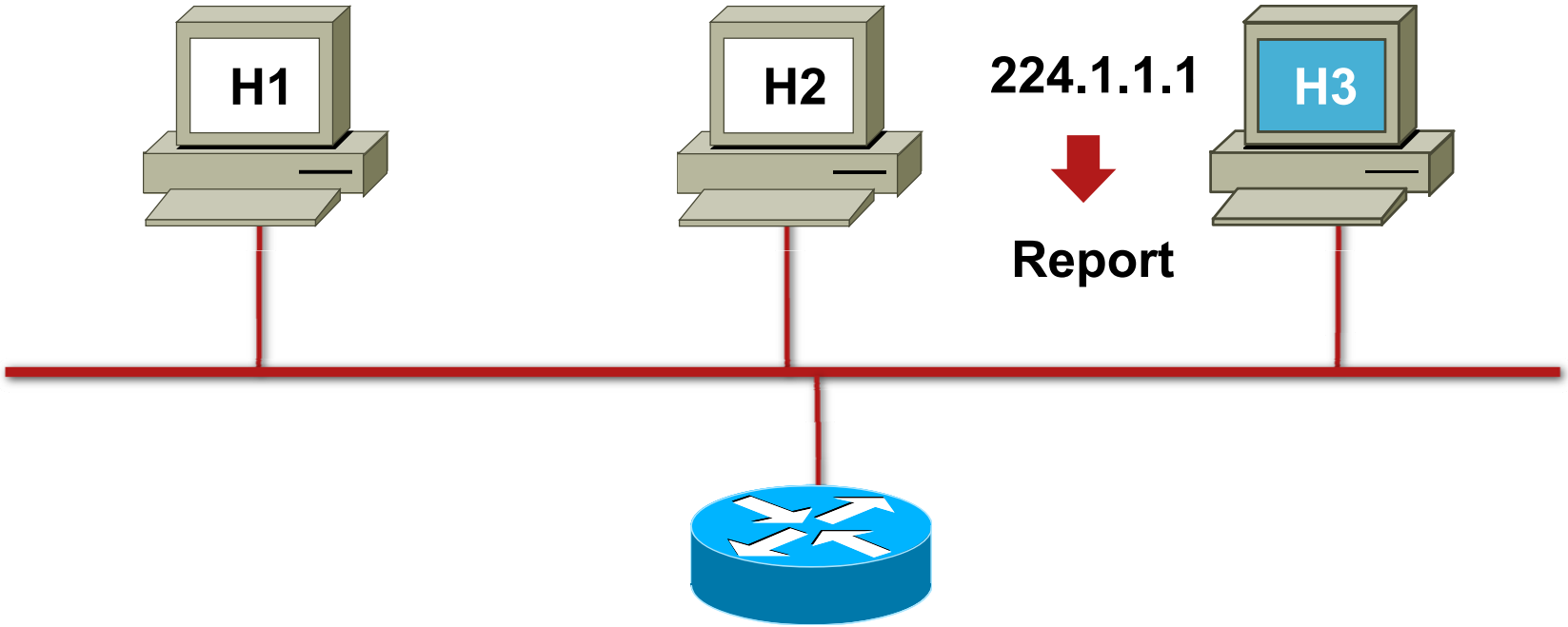
  wspierane na Windows 95
- RFC 2236 opisuje wersję 2 protokołu IGMP

  wspierane na Windows i większości systemów UNIXowych
- RFC 3376 opisuje wersję 3 protokołu IGMP

  wspierane w Window XP i większości systemów UNIXowych

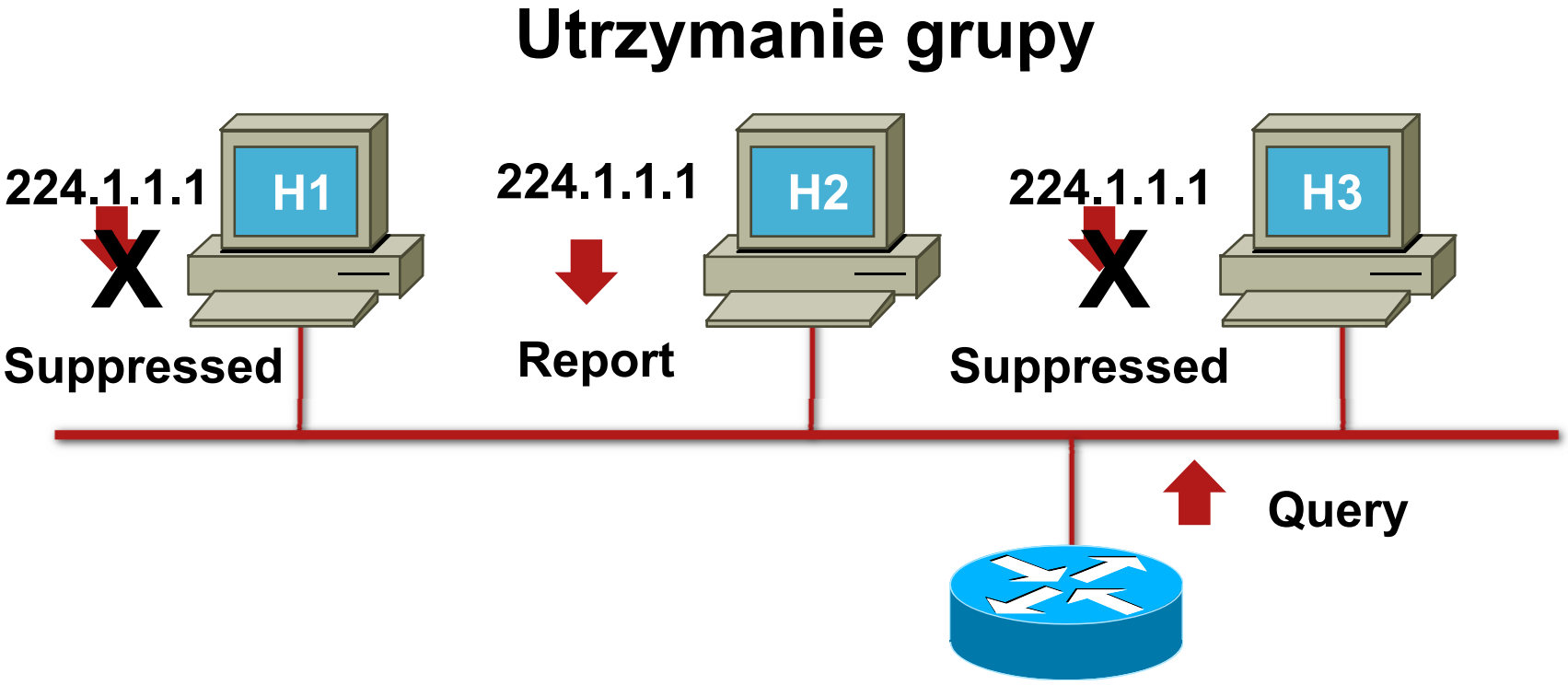
Sygnalizacja host-router: IGMP

Dołączenie do grupy



- Host wysyła komunikat IGMP report by dołączyć się do grupy

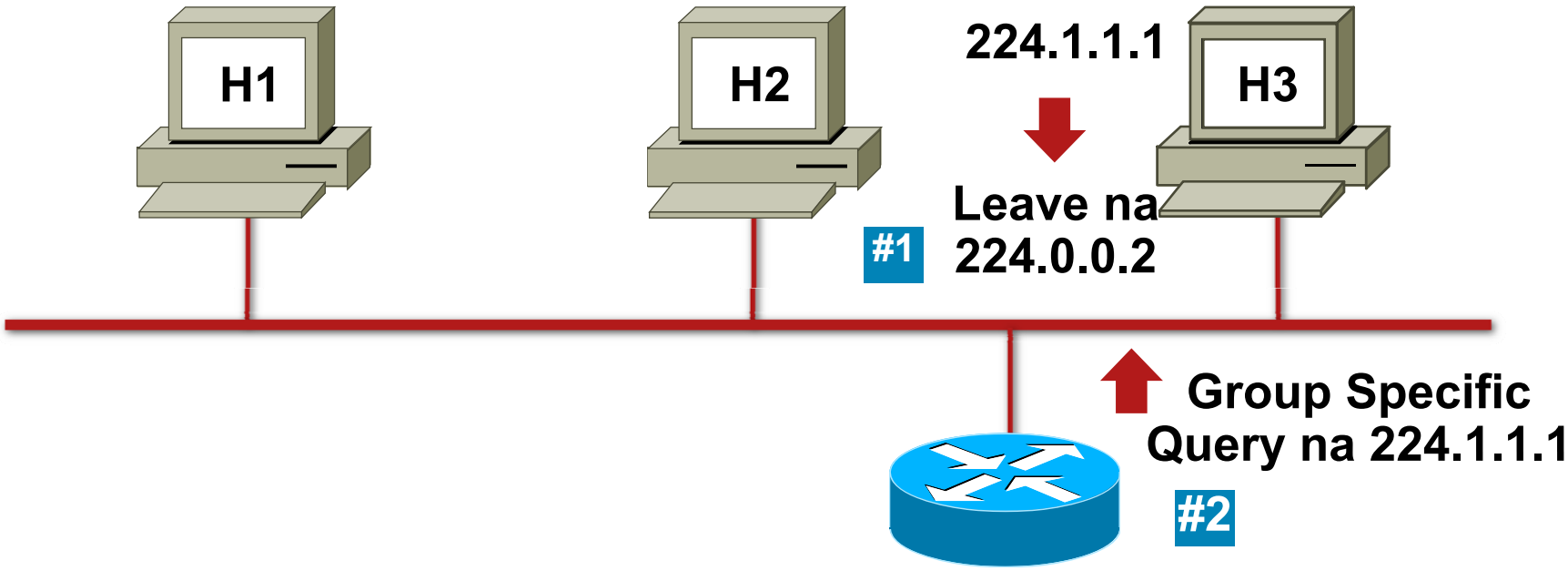
Sygnalizacja host-router: IGMP



- Router wysyła okresowe zapytania na adres 224.0.0.1
- Jeden z członków na podsieć odpowiada
- Pozostali członkowie wstrzymują się z odpowiedzią

Sygnalizacja host-router: IGMP

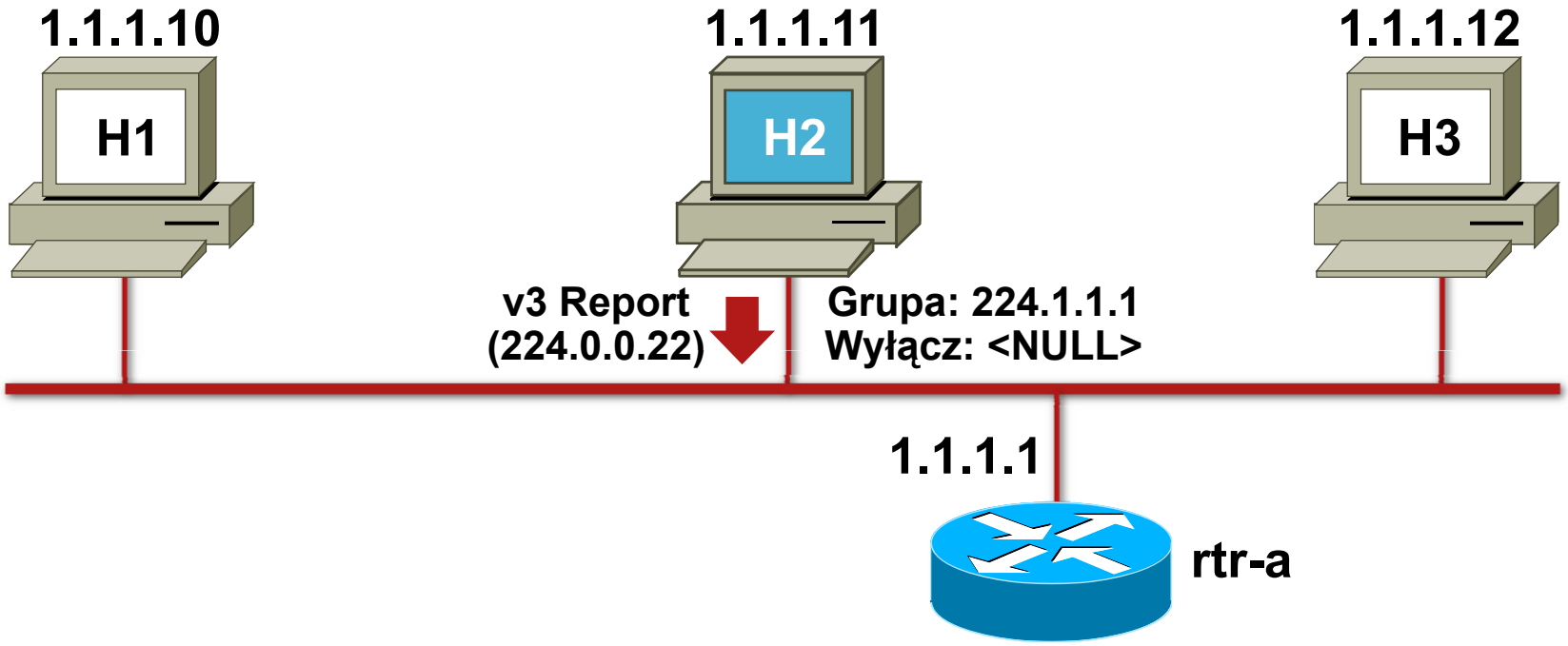
Opuszczenie grupy (IGMPv2)



- Host wysyła komunikat ‘leave’ na adres 224.0.0.2
- Router wysyła zapytanie specyficzne dla grupy na 224.1.1.1
- Jeśli brak raportu IGMP w ciągu ~ 3 sekund...
- ...grupa 224.1.1.1 wygasa

Sygnalizacja host-router: IGMP

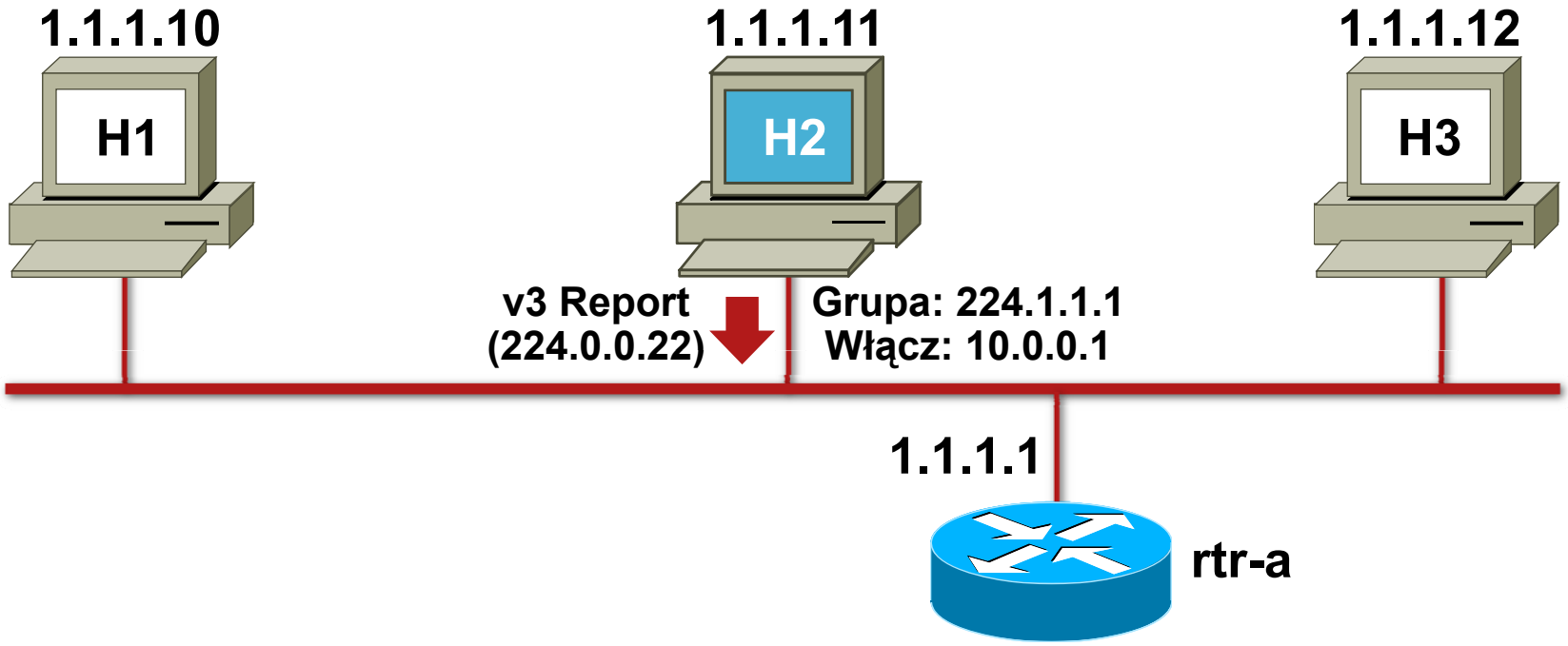
IGMPv3 – dołączenie do grupy



- Członek wysyła raport IGMPv3 na adres 224.0.0.22

Sygnalizacja host-router: IGMP

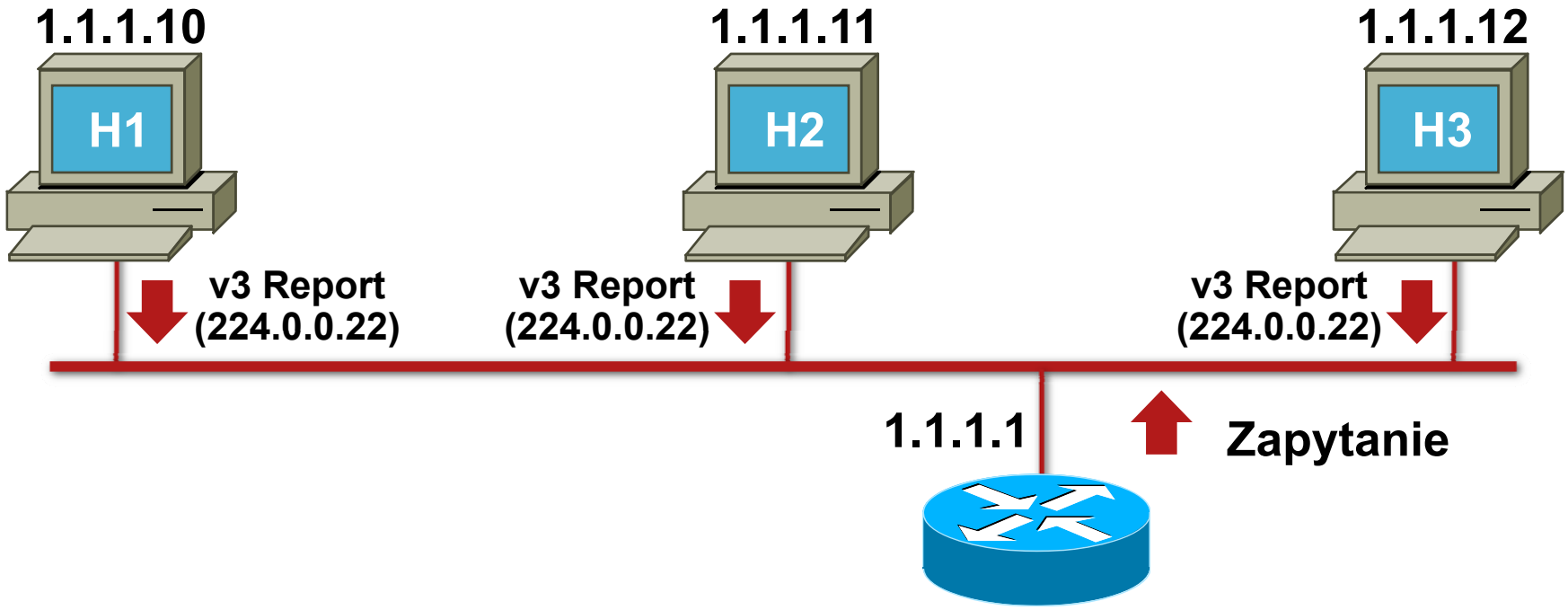
IGMPv3 – dołączenie do konkretnej grupy



- Raport może zawierać konkretny adres grupy – tylko ona jest następnie ‘subskrybowana’

Sygnalizacja host-router: IGMP

IGMPv3 – utrzymanie subskrypcji



- Router odpowiedzialny jest za okresowe odpytanie segmentu
- Wszyscy członkowie IGMPv3 odpowiadają na zapytanie

Multicast Listener Discover—MLD

- MLD jest odpowiednikiem IGMP dla IPv6
- Komunikaty MLD transportowane są przez ICMPv6
- Numeracja wersji MLD nie jest równa IGMP!

  MLDv1 odpowiada IGMPv2

  RFC 2710

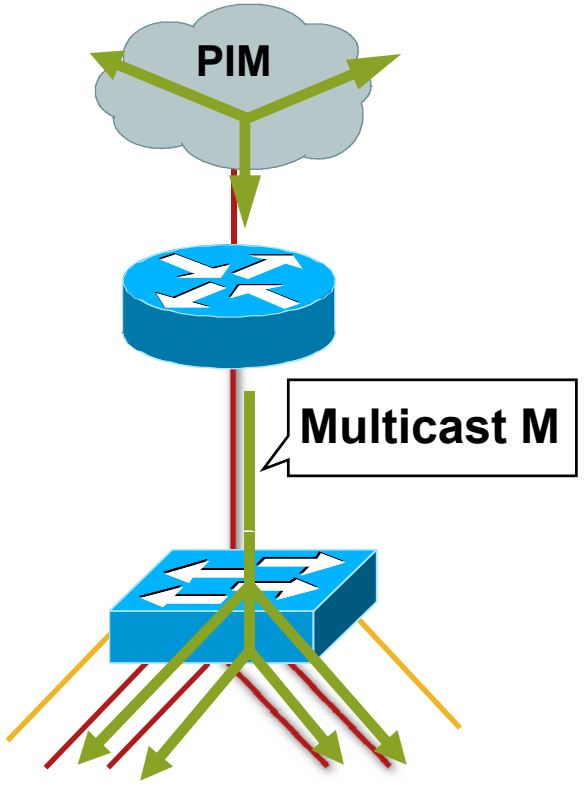
  MLDv2 odpowiada IGMPv3, jest wymagana dla SSM

  RFC 3810

Switching ruchu multicast w L2

Problem: ramki multicast w L2 traktowane są jak broadcast

- Typowy przełącznik L2 potraktuje ruch multicastowy L2 jako nieznany lub broadcast – i roześle jego kopię do wszystkich portów w danej domenie rozgłoszeniowej (VLANie)
- Oczywiście można posłużyć się definicją statyczną by wskazać porty zainteresowane konkretnymi grupami – słabo się to jednak skaluje (choć jest stabilne ☺)

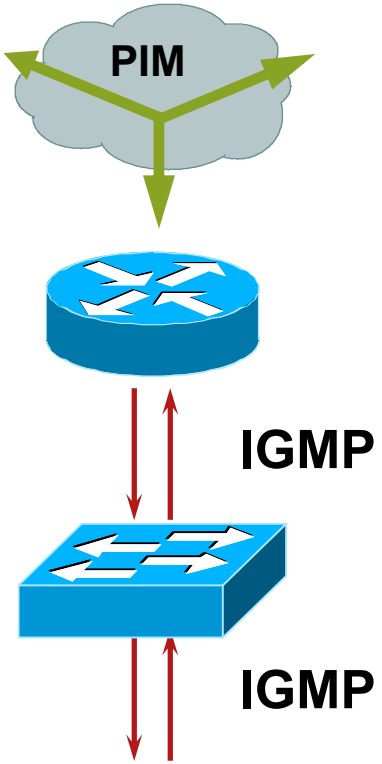


Switching ruchu multicast w L2

IGMPv1–v2 Snooping

- Przełączniki rozumieją IGMP – pakiety IGMP są przechwytywane przez RP lub specjalizowany ASIC urządzenia
- Przełącznik musi potrafić zrozumieć komunikat IGMP i skojarzyć port z potencjalnym ruchem do grup multicastowych

  Raporty IGMP i komunikaty o opuszczaniu grup
- Tanie, lub używające źle zaprojektowanych ASICów przełączniki w tym momencie umierają – cały ruch L2 multicast przechodzi przez (zwykle) wolne CPU



Switching ruchu multicast w L2

Co zmienia IGMPv3 w procesie snoopingu?

- Raporty IGMPv3 wysyłane są do konkretnej grupy (224.0.0.22)

  Przełącznik może słuchać tylko ruchu do tej konkretnej grupy

  Tylko ruch IGMP – brak ruchu użytkowego

  Obniża obciążenie IGMP

- Jesteśmy w stanie śledzić poszczególnych użytkowników – brak ‘powstrzymywania’ raportów

Routing multicastowy

Routing multicastowy

Routing multicastowy to ‘przeciwieństwo’ unicastowego

- Routing unicastowy zajmuje się zagadnieniem gdzie skierować pakiet aby dotarł do miejsca przeznaczenia
- Routing multicastowy zajmuje się zagadnieniem skąd pakiet przyszedł
- ...czyli wszystko jest odwrotnie...prawie

Routing unicastowy a multicastowy

Routing unicastowy

- Docelowy adres IP wprost wskazuje gdzie przekazać pakiet
- Przekazywanie odbywa się hop-by-hop

  Tablica routingu wskazuje wyjściowy interfejs i adres następnego routera

Routing unicastowy a multicastowy

Routing multicastowy

- Docelowy adres IP (grupy) nie wskazuje wprost gdzie przekazać pakiet
- Przekazywanie ruchu jest zorientowane połączeniowo

  Odbiorcy muszą się najpierw ‘podłączyć’ do drzewa multicastowego zanim zaczną odbierać dane

  Komunikaty o podłączeniu (PIM join) wysyłane są zgodnie z unicastową tablicą routingu

  Następnie budowane są drzewa dystrybucji, określające jak przekazywać ruch multicastowy

  Drzewa dystrybucyjne mogą być dynamicznie przebudowywane wraz ze zmianą topologii sieci

Reverse Path Forwarding (RPF)

Przeliczenie RPF

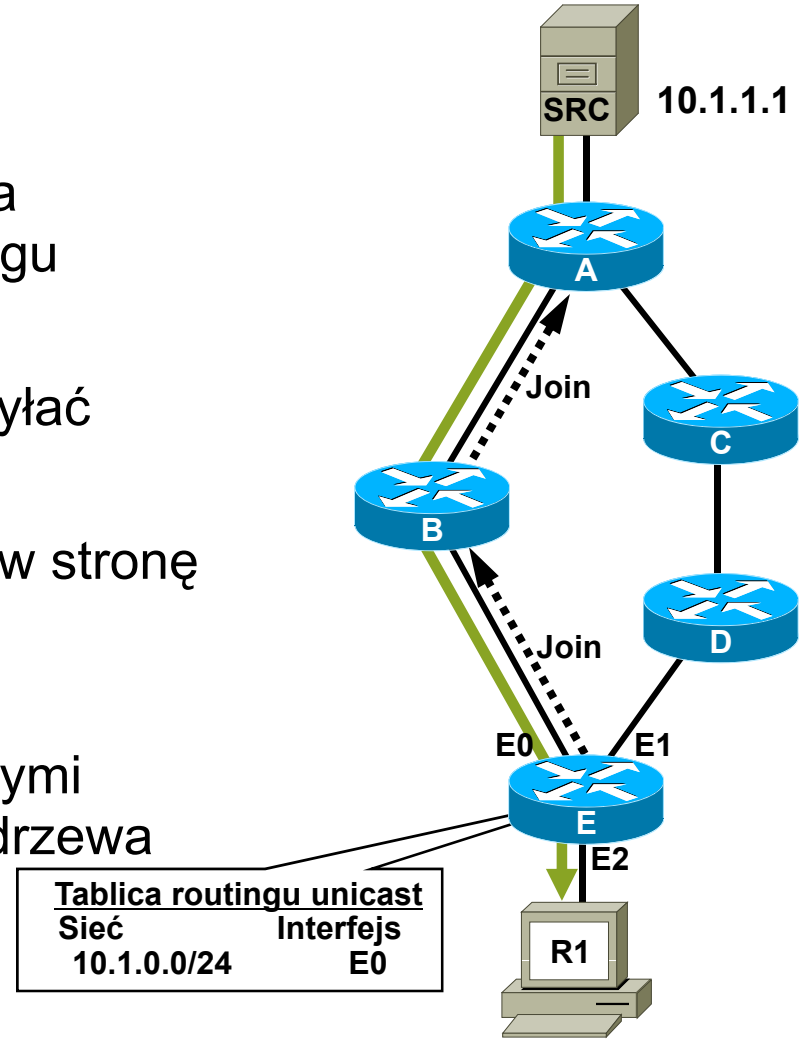
- Adres źródłowy pakietu multicastowego jest sprawdzany w oparciu o tablicę routingu unicastowego
- Określa to interfejs i router upstream w kierunku źródła do którego wysyłane są wiadomości PIM join
- Interfejs ten staje się ‘Incoming’ RPF

  Router przekazuje datagram multicastowy tylko jeśli odbiorca jest osiągalny przez interfejs wskazany przez RPF

Reverse Path Forwarding (RPF)

Przeliczenie RPF

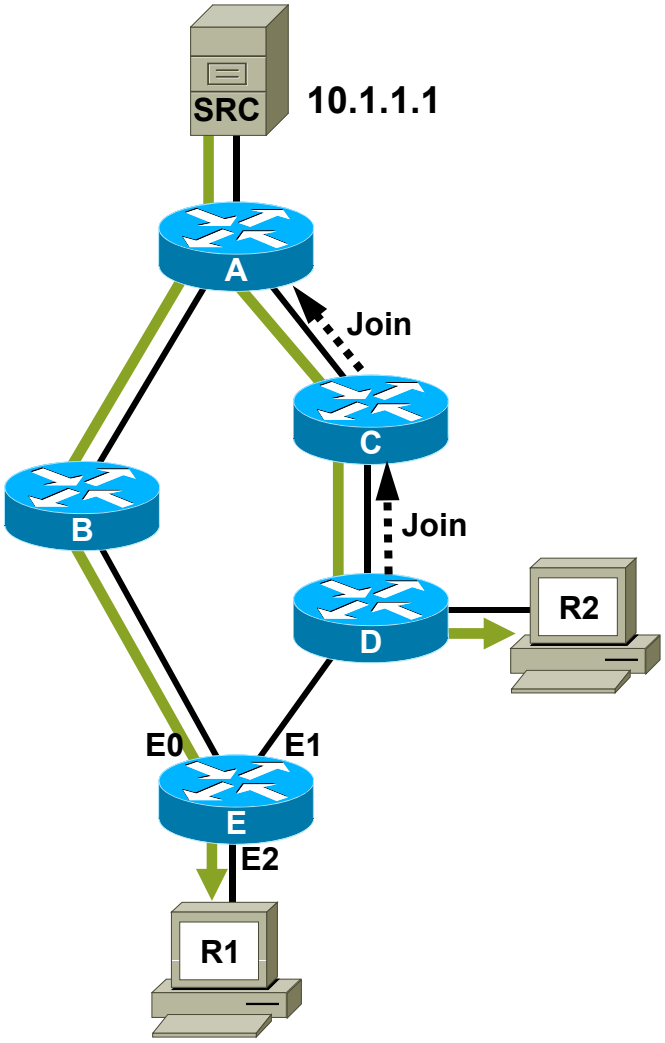
- Oparte o adres źródłowy
- Najlepsza ścieżka do źródła pochodząca z tablicy routingu unicastowego
- Określa w którą stronę wysyłać komunikaty ‘join’
- Pakiety ‘join’ kierowane są w stronę źródła – budując ‘drzewo’ multicastowe
- Pakiety multicastowe z danymi przesyłane są w ‘dół’ tego drzewa



Reverse Path Forwarding (RPF)

Przeliczenie RPF

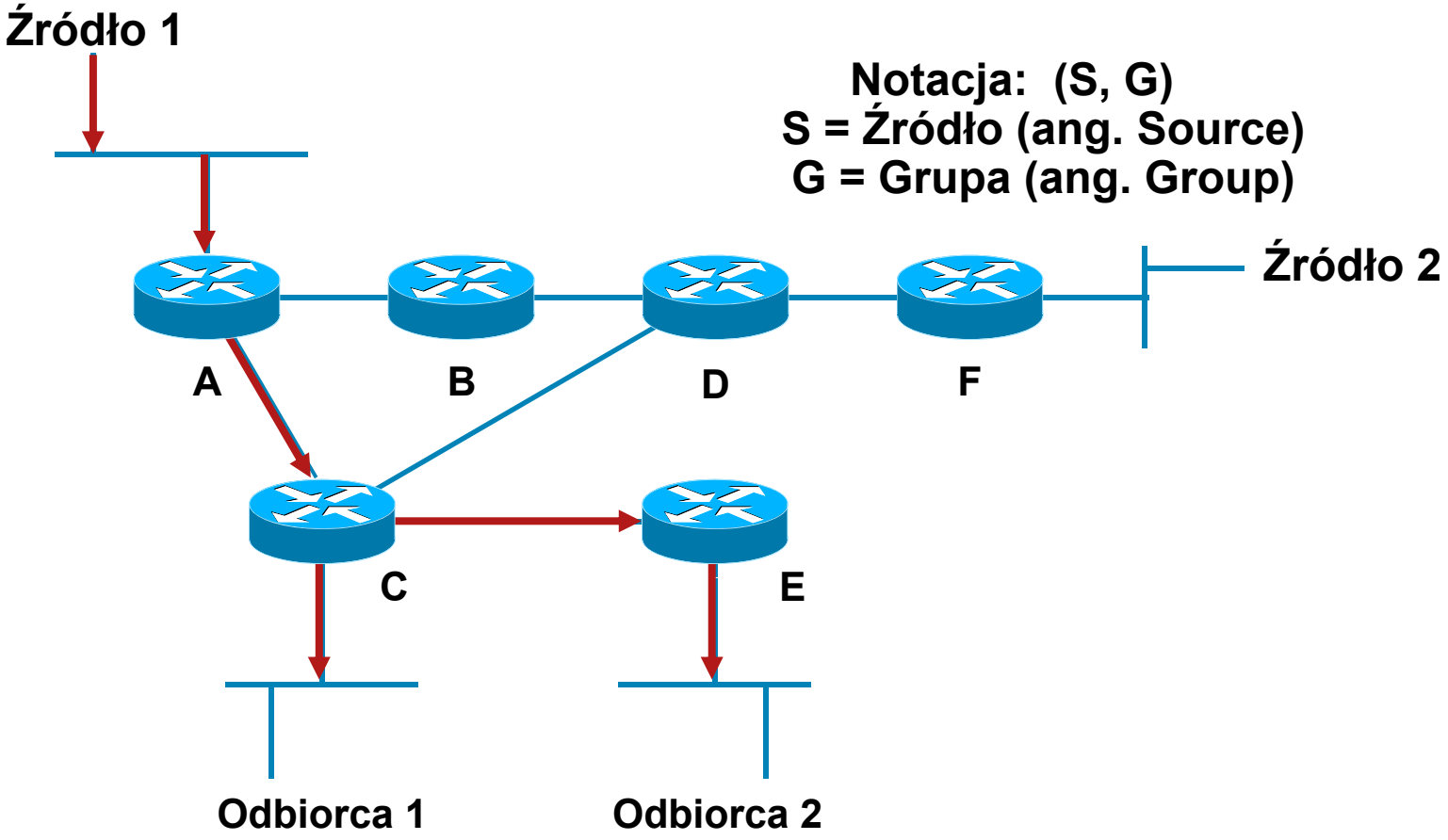
- Oparte o adres źródłowy
- Najlepsza ścieżka do źródła pochodząca z tablicy routingu unicastowego
- Określa w którą stronę wysyłać komunikaty ‘join’
- Pakiety ‘join’ kierowane są w stronę źródła – budując ‘drzewo’ multicastowe
- Pakiety multicastowe z danymi przesyłane są w ‘dół’ tego drzewa
- Powtarzamy dla każdego nowego odbiorcy



Multicast Distribution Trees

Drzewa dystrybucji informacji multicastowych

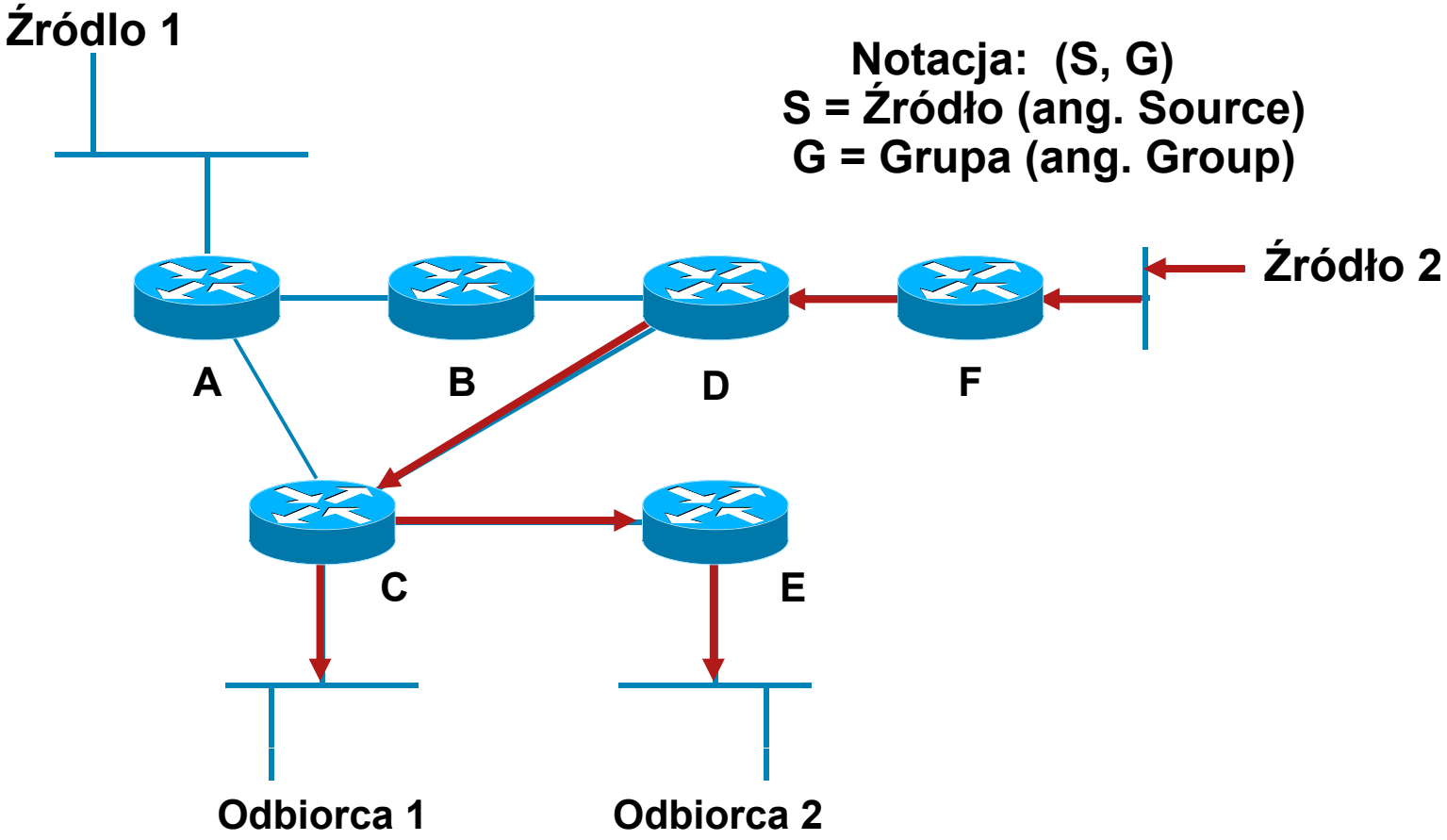
Drzewo najkrótszej ścieżki lub źródłowe



Multicast Distribution Trees

Drzewa dystrybucji informacji multicastowych

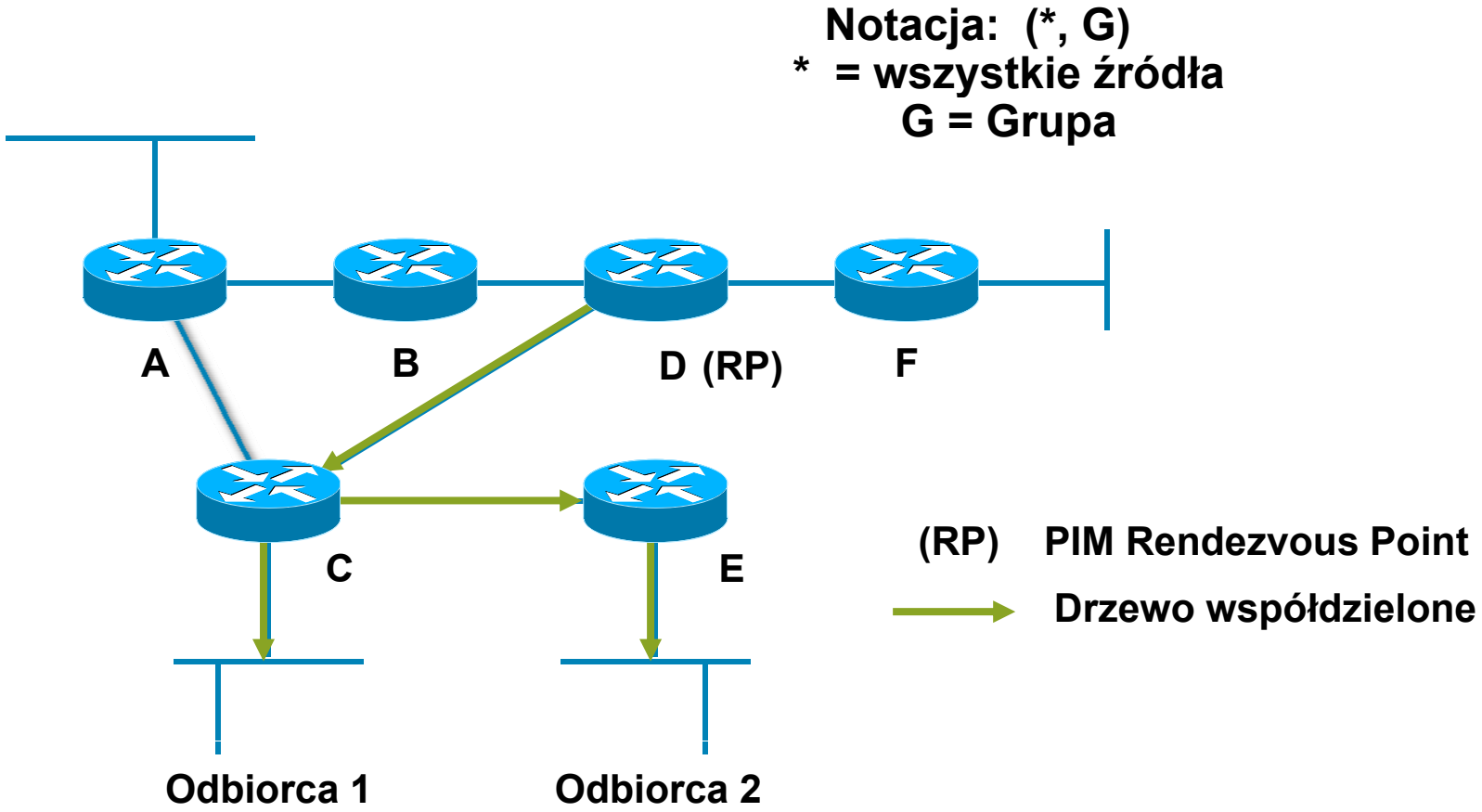
Drzewo najkrótszej ścieżki lub źródłowe



Multicast Distribution Trees

Drzewa dystrybucji informacji multicastowych

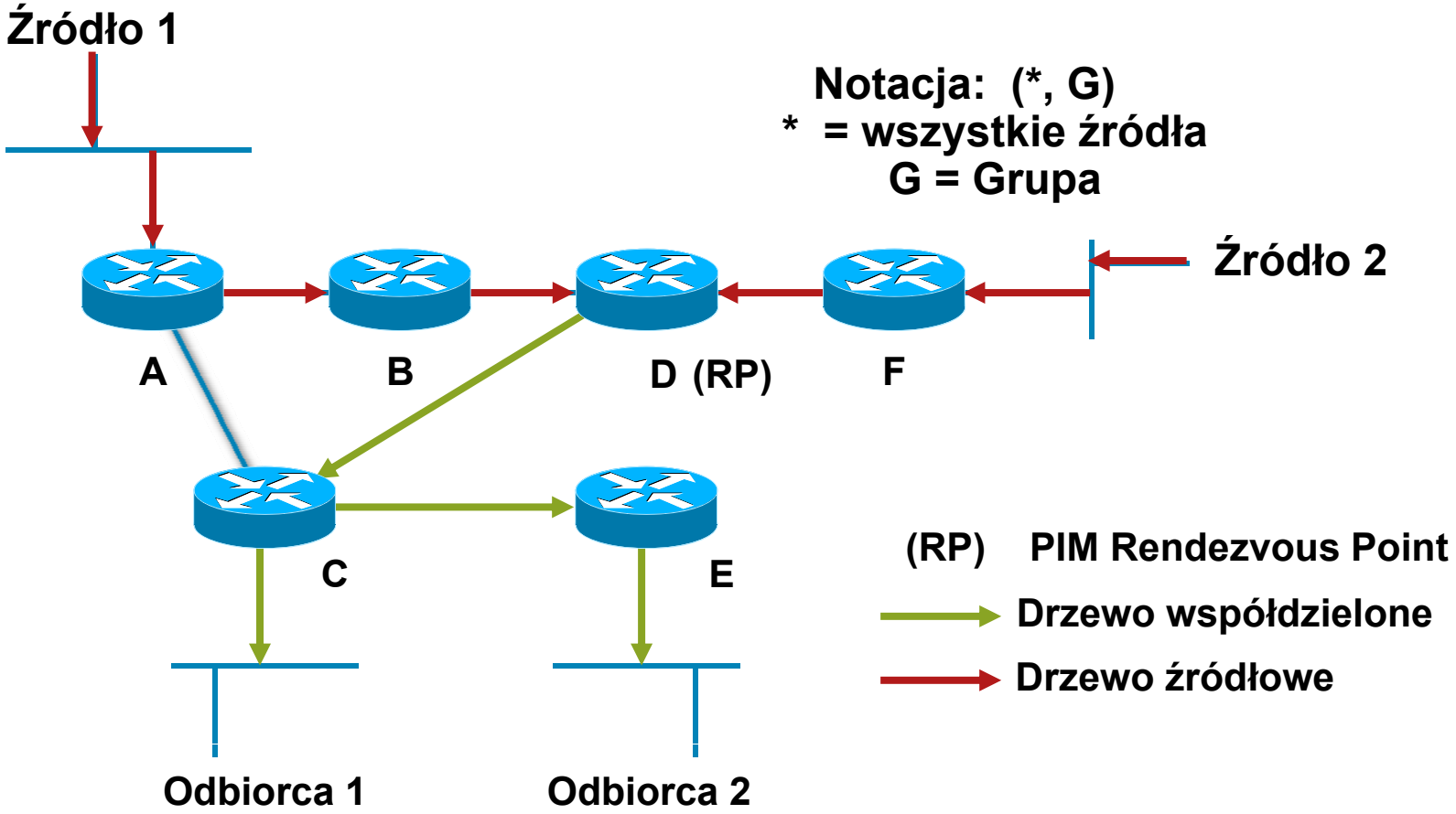
Drzewo współdzielone



Multicast Distribution Trees

Drzewa dystrybucji informacji multicastowych

Drzewo współdzielone



Multicast Distribution Trees

Charakterystyka drzew dystrybucji

- Drzewa źródłowe/najkrótsze

  Wykorzystują więcej przestrzeni w pamięci O(S*G), ale otrzymujemy optymalne trasy do wszystkich odbiorców, zmniejszając opóźnienia w transporcie ruchu multicastowego

- Drzewa współdzielone

  Wykorzystują mniej pamięci O(G), ale trasy wyznaczone przez sieć mogą nie być optymalne – może to spowodować dodatkowe opóźnienie

Tworzenie drzewa multicastowego

- Komunikaty kontrolne join/prune protokołu PIM
  - Używane do tworzenia/usuwania drzew dystrybucji
- W drzewach źródłowych
  - Komunikaty kontrolne PIM wysyłane są w stronę źródła
- ...natomiast w drzewach współdzielonych
  - Komunikaty kontrolne PIM przesyłane są do RP

Protokół PIM

Warianty PIM spotykane „na wolności”

PIM-SM

- ASM

  Any Source Multicast/RP/SPT/shared tree

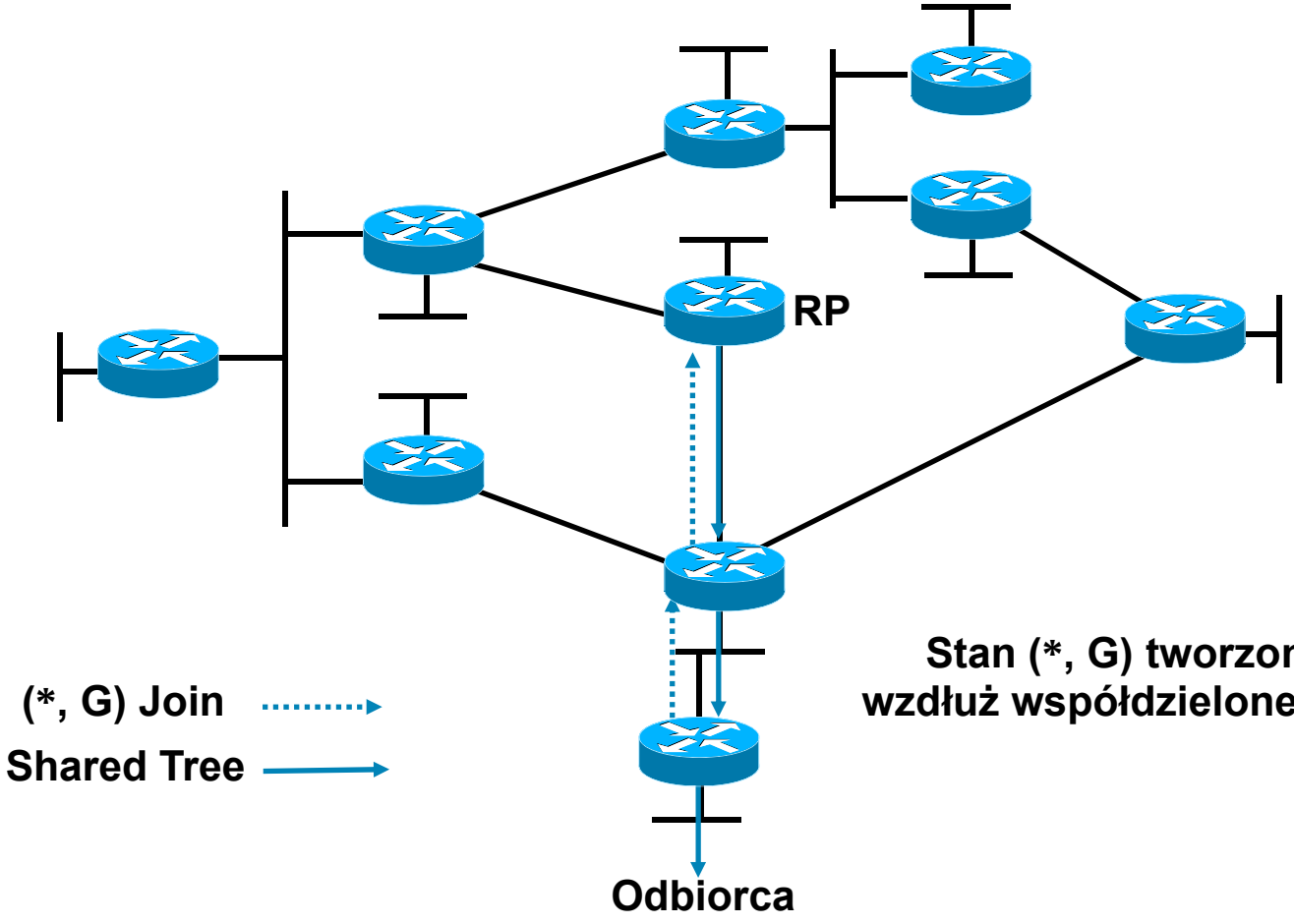
- SSM

  Source Specific Multicast, brak RP, tylko drzewo źródłowe (SPT)

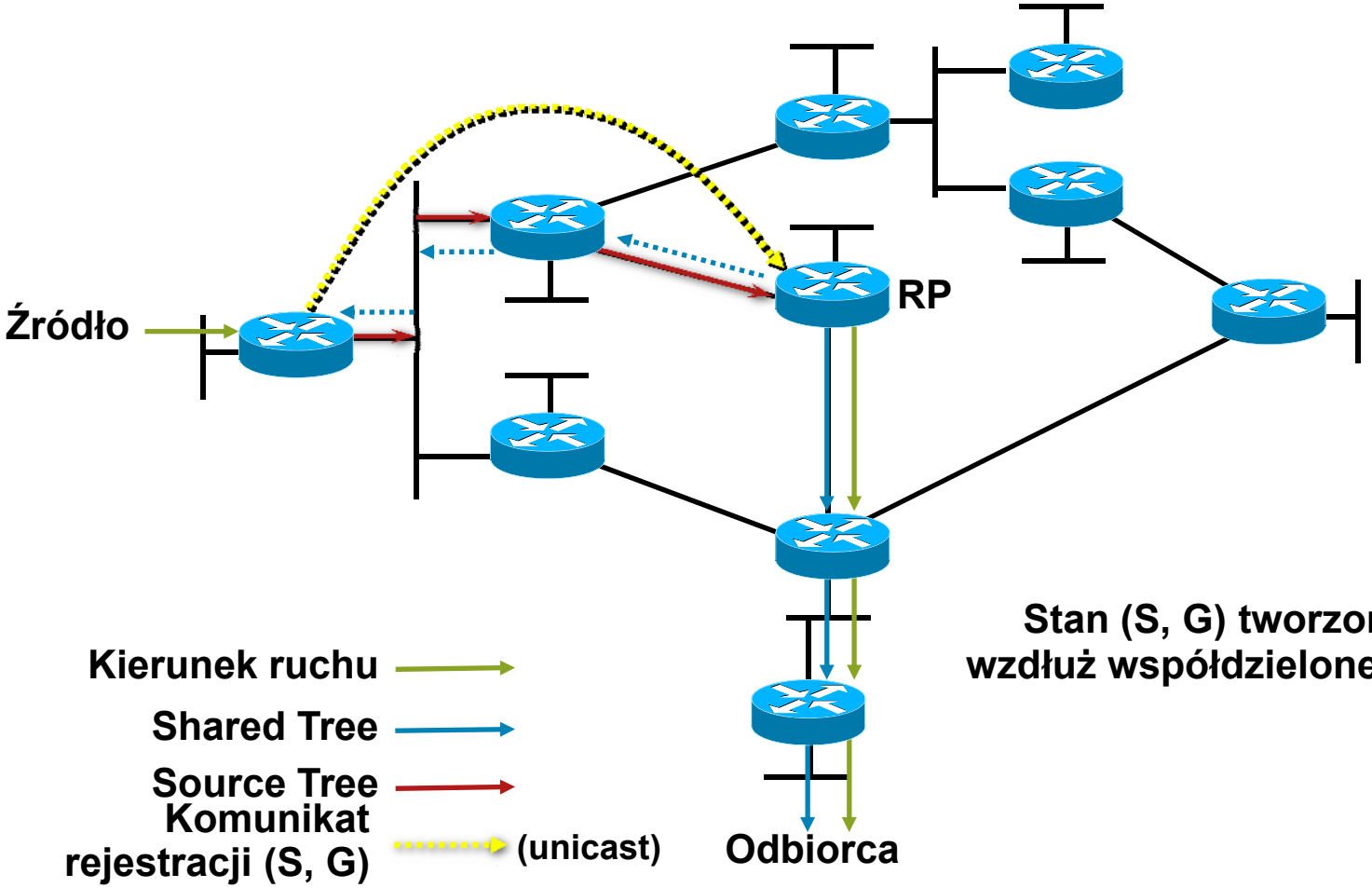
- BiDir

  Bidirectional PIM, brak SPT, tylko drzewo współdzielone

Dołączenie do drzewa PIM-SM

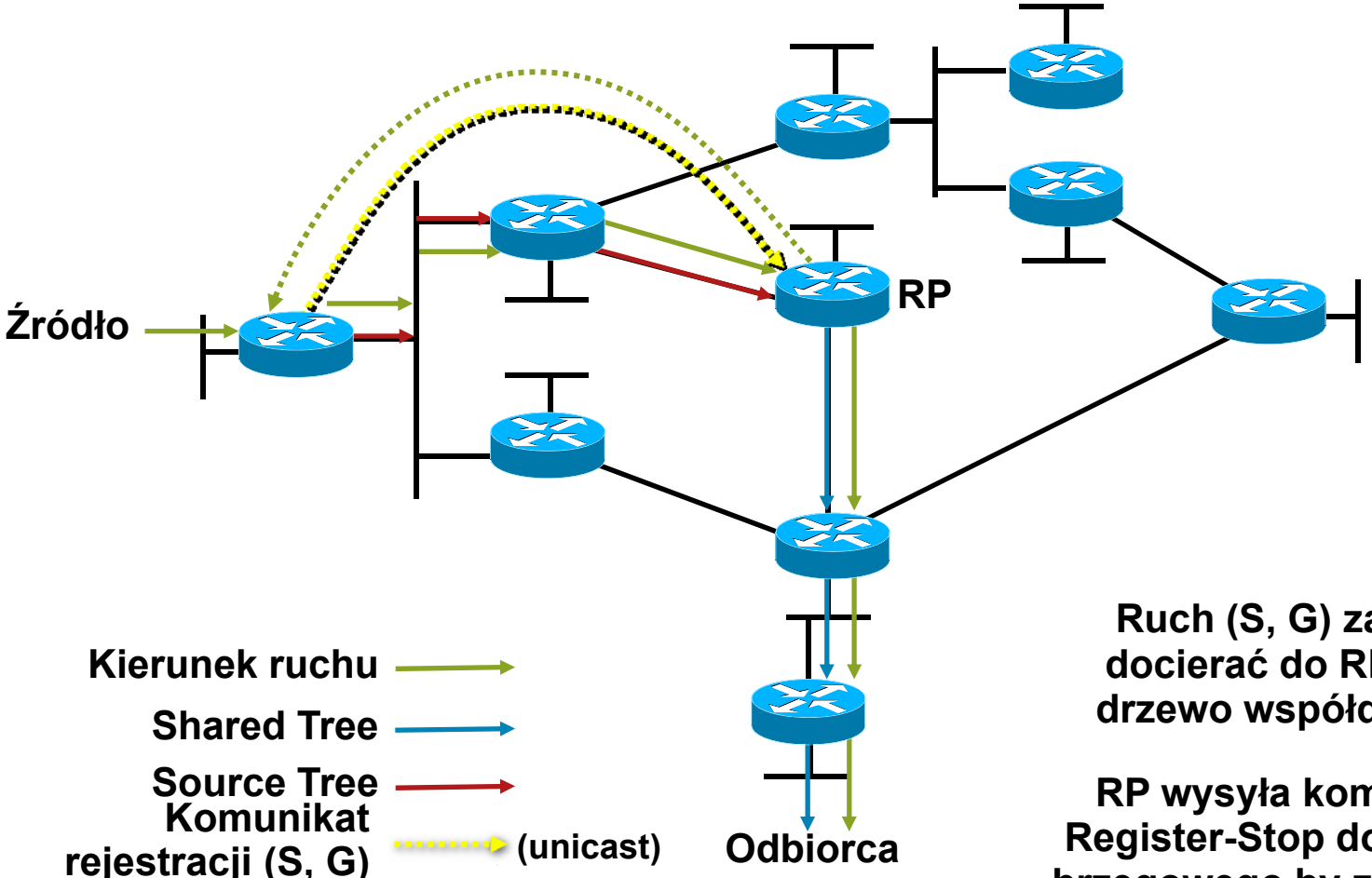


Rejestracja nadawcy w PIM-SM



Stan (S, G) tworzony tylko wzdłuż współdzielonego drzewa

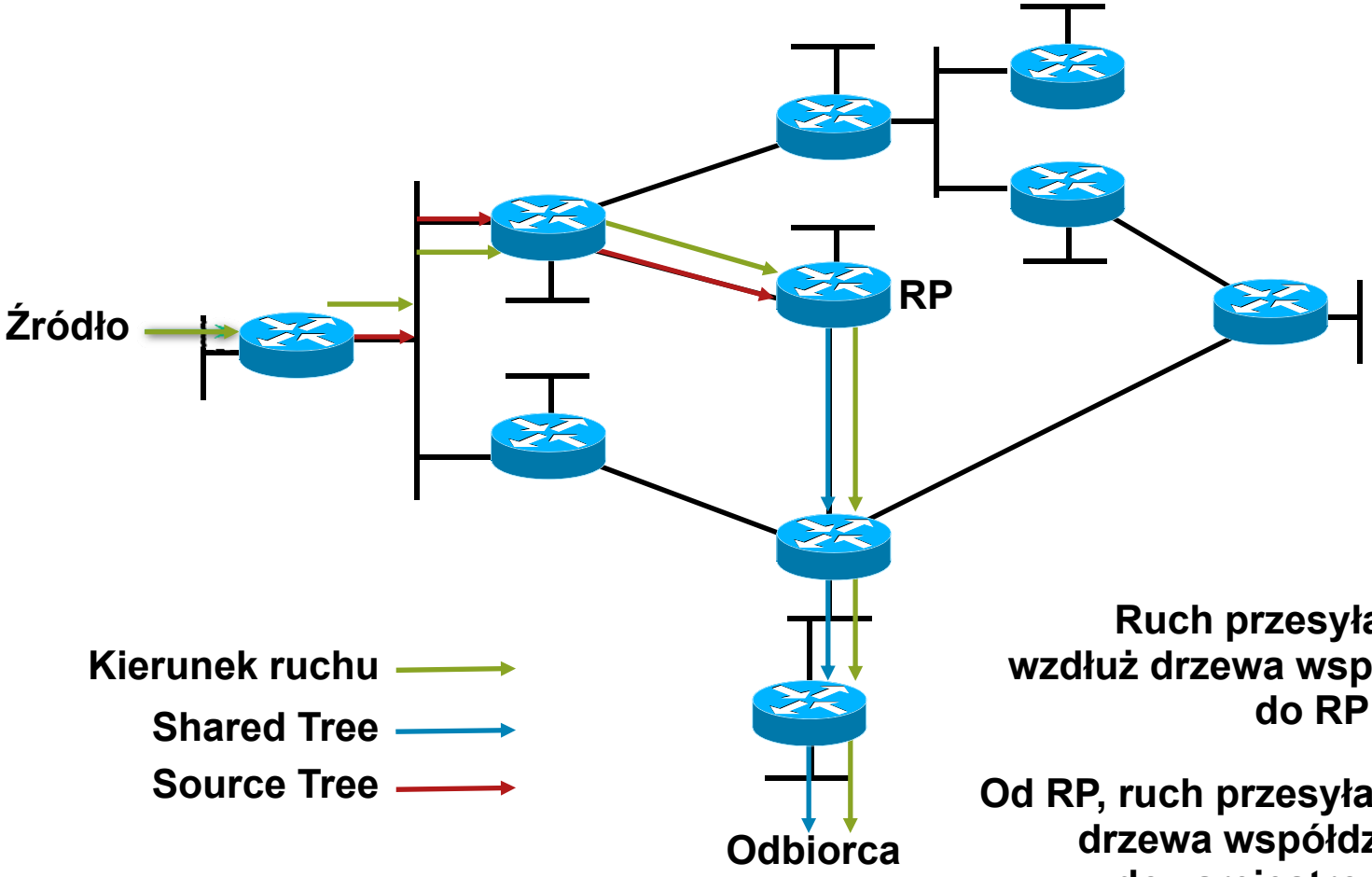
Rejestracja nadawcy w PIM-SM



Ruch (S, G) zaczyna docierać do RP przez drzewo współdzielone

RP wysyła komunikat Register-Stop do routera brzegowego by zatrzymać proces ciągłej rejestracji

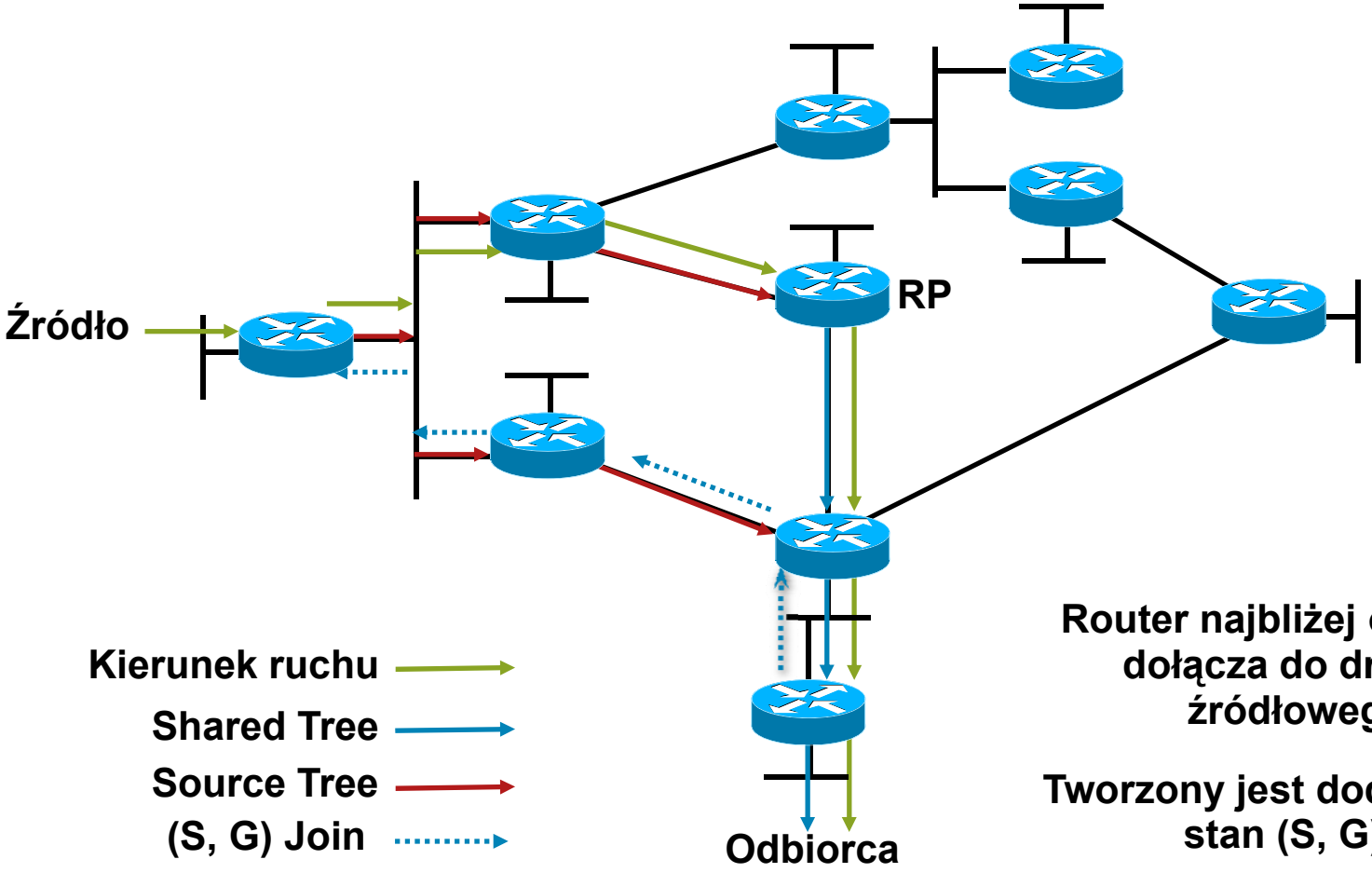
Obsługa ruchu po rejestracji



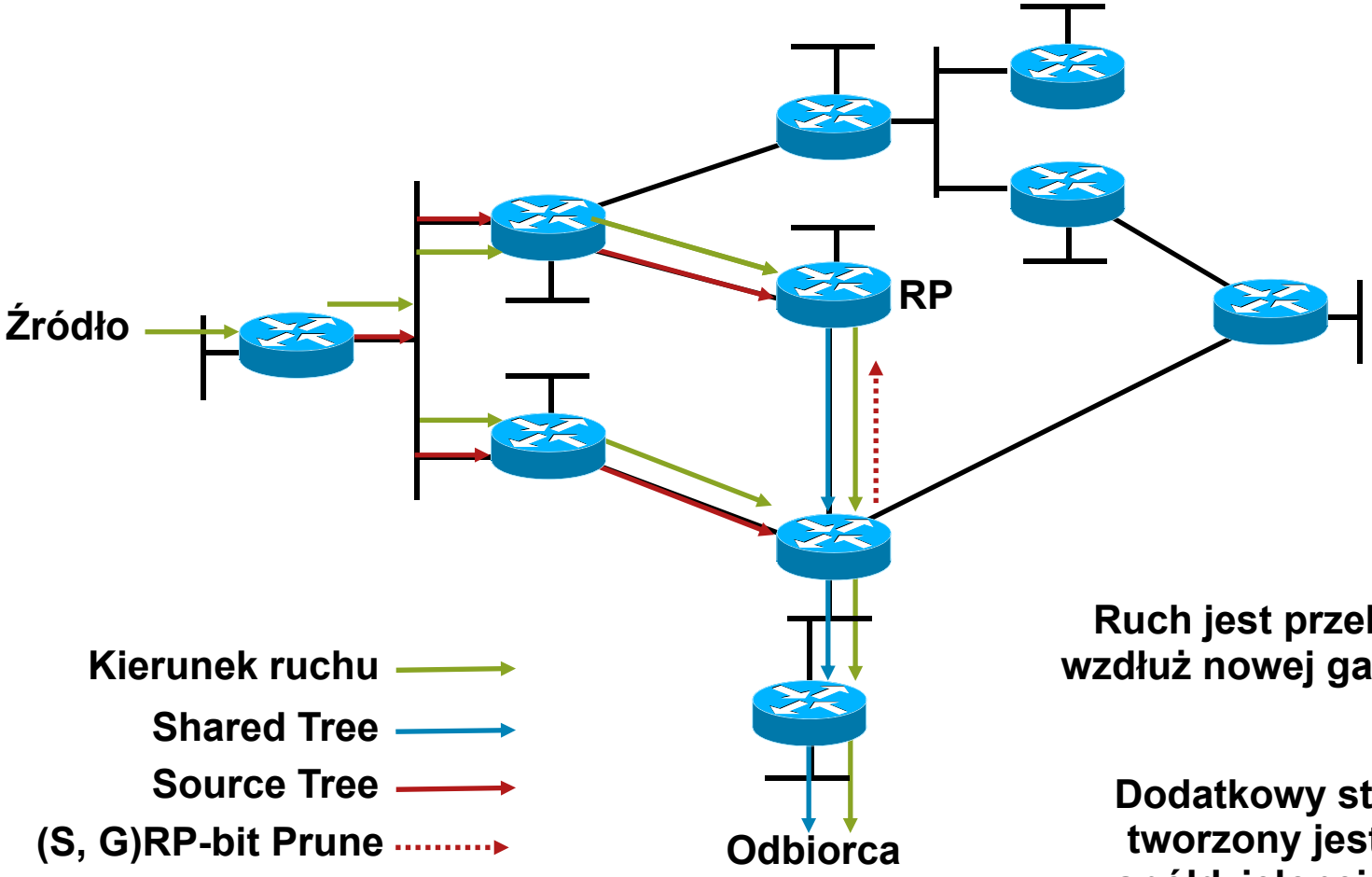
Ruch przesyłany jest wzdłuż drzewa współdzielonego do RP

Od RP, ruch przesyłany jest wzdłuż drzewa współdzielonego do zarejestrowanych odbiorców

Przełączenie drzewa PIM-SM



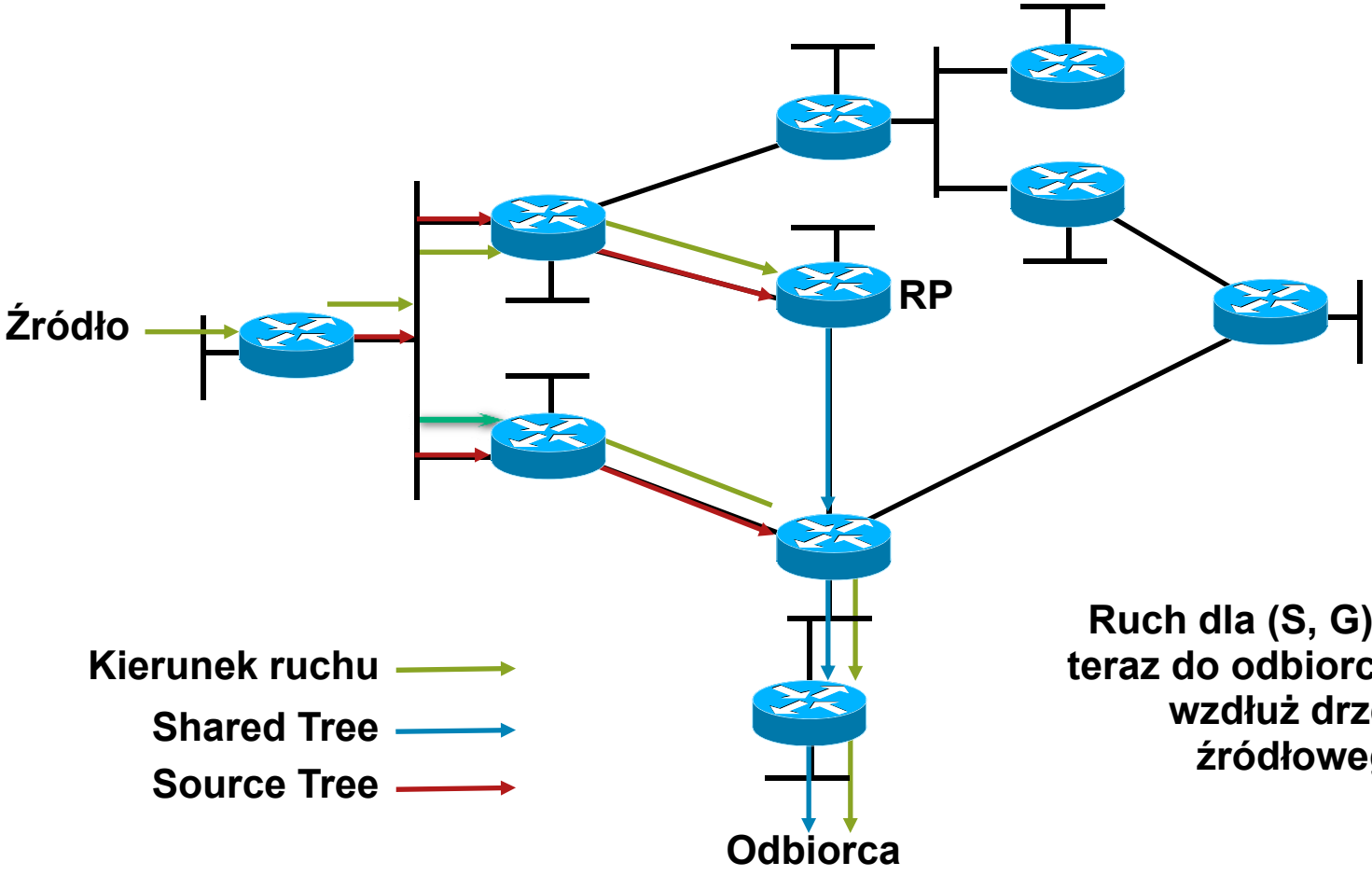
Przełączenie drzewa PIM-SM



Ruch jest przekazywany wzdłuż nowej gałęzi drzewa

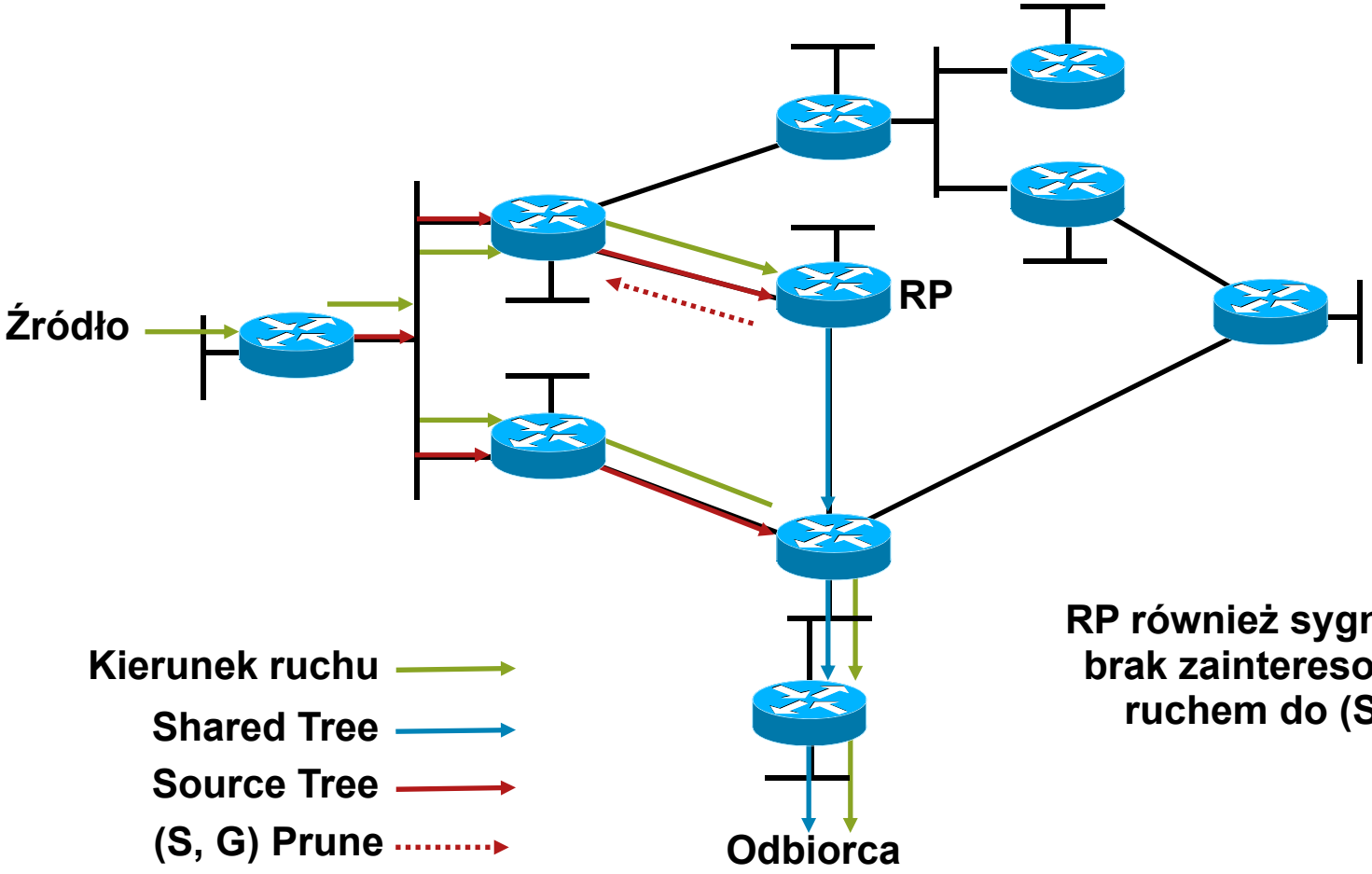
Dodatkowy stan (S, G) tworzony jest wzdłuż współdzielonej ścieżki by zatrzymać niepotrzebny już ruch (S, G)

Przełączenie drzewa PIM-SM



Ruch dla (S, G) dociera teraz do odbiorcy jedynie wzdłuż drzewa źródłowego

Przełączenie drzewa PIM-SM



To zachowanie można zablokować: **ip pim spt-threshold infinity**

PIM-SM—Podsumowanie

- Efektywny do dystrybucji multicastów dla rzadko rozproszonych w sieci odbiorców
- Zalety

  Ruch wysyłany jest tylko do odbiorców którzy wprost wyrazili zainteresowanie treścią – oszczędzamy pasmo

  Może przełączyć sposób konstrukcji drzewa (a zatem – ścieżkę przez sieć) dla źródeł nadających dużą ilość danych

  Niezależny od protokołu routingu unicastowego

Source Specific Multicast

- Przyjmijmy model jeden-do-wielu

  np. rozgłośnie video/audio

- Po co w modelu ASM potrzebujemy współdzielonego drzewa?

  Hosty i router do którego są podłączone muszą móc nauczyć się gdzie znajduje się aktywne źródło ruchu dla grupy

- A co gdybyśmy od razu to wiedzieli?

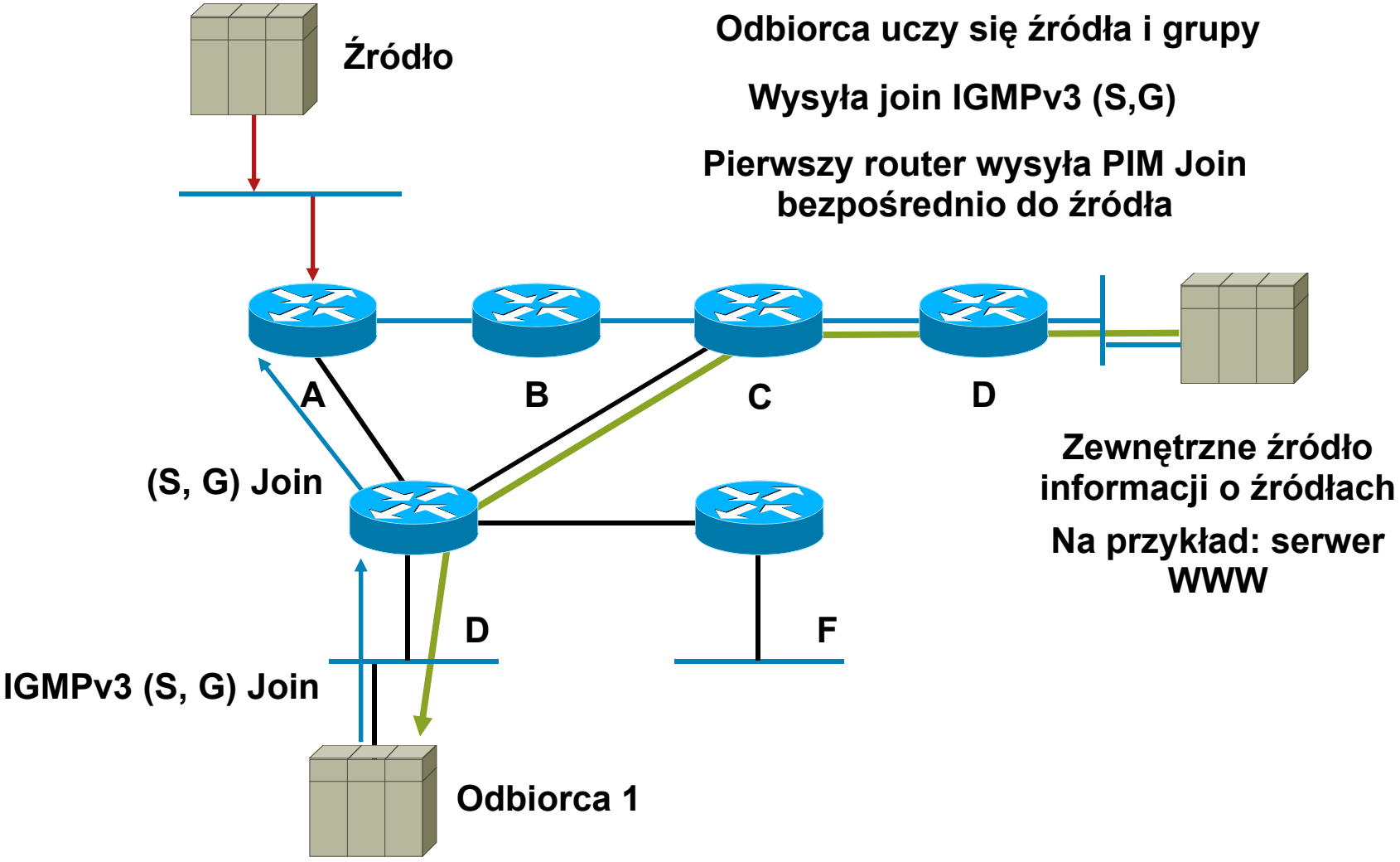
  Host może użyć IGMPv3 by zasygnalizować zapotrzebowanie dla konkretnej pary (S,G)

  Współdzielone drzewo i RP nie są w tej sytuacji potrzebne

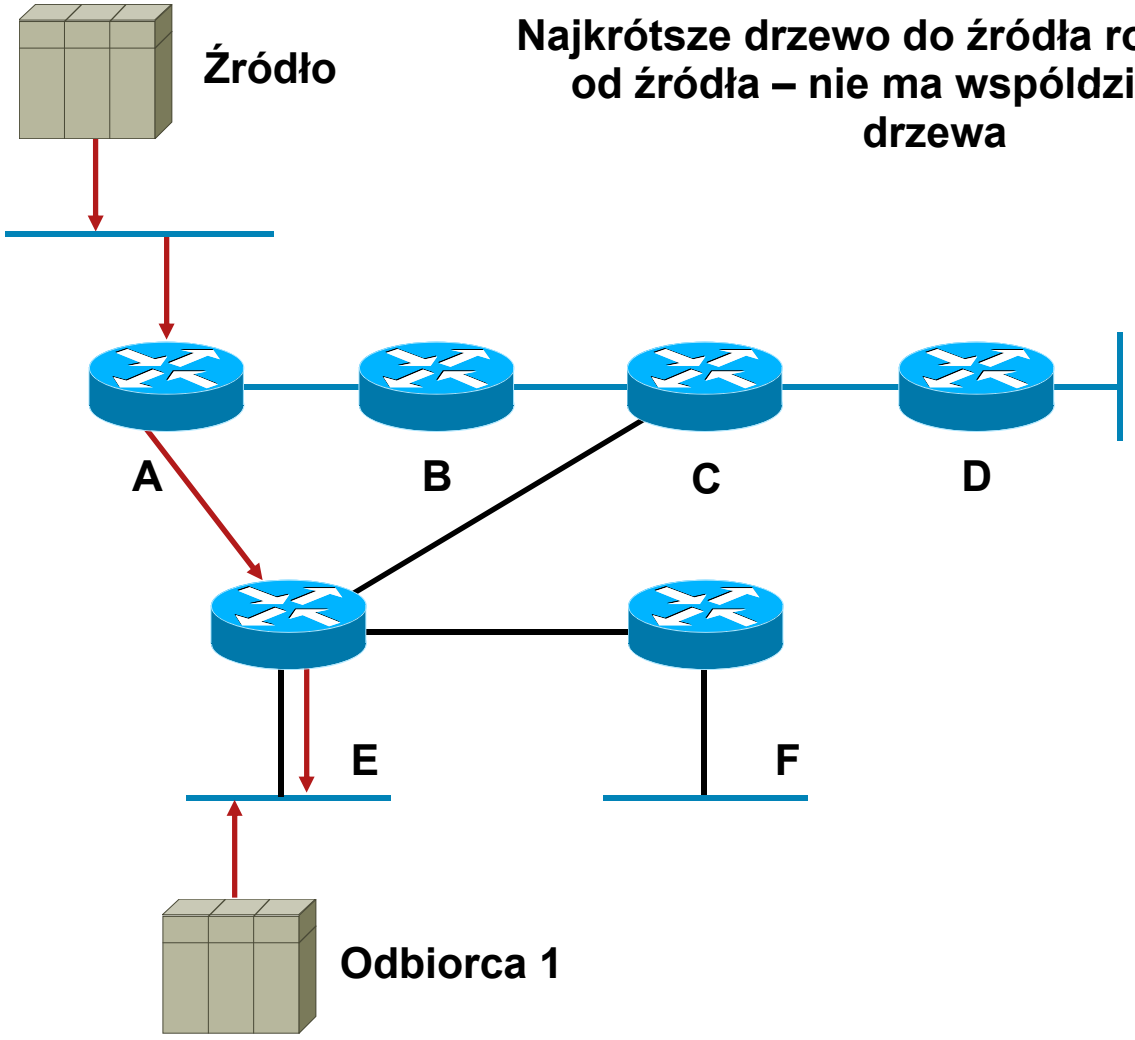
  Różne źródła mogą współdzielić ten sam numer grupy i nie przeszkadzać sobie na wzajem

- ...i tak powstał: Source Specific Multicast (SSM)
- RFC 3569: An Overview of Source Specific Multicast (SSM)

Tryb PIM Source Specific



Tryb PIM Source Specific



Najkrótsze drzewo do źródła rozpięte jest od źródła – nie ma współdzielonego drzewa

SSM—Podsumowanie

- Idealny dla aplikacji w których jedno źródło wysyła do wielu odbiorców
- Używa uproszczonego protokołu PIM-SM
- „Rozwiązuje” problem przydziału adresów multicastowych

  Ruch danych oddzielony zarówno przez źródło jak i grupę – nie tylko przez grupę

  Dostawcy treści mogą używać tych samych numerów grup – (S,G) jest unikalne
- Pozwala niejako przy okazji zapobiegać atakom DoS

  „Fałszywe” źródło ruchu nie może nadawać do grupy – nawet jeśli będzie nadawać, być może nie trafi w konkretną parę (S,G)

Problem sytuacji wiele-do-wielu

- Tworzy ogromne tablice (S,G)

  Problem z utrzymaniem dużych tablic staje się niebanalny nawet dla platform z szybką pamięcią podręczną – a nawet w szczególności dla nich

    duża ilość interfejsów wyjściowych (OIL) pogarsza problem w platformach sprzętowych

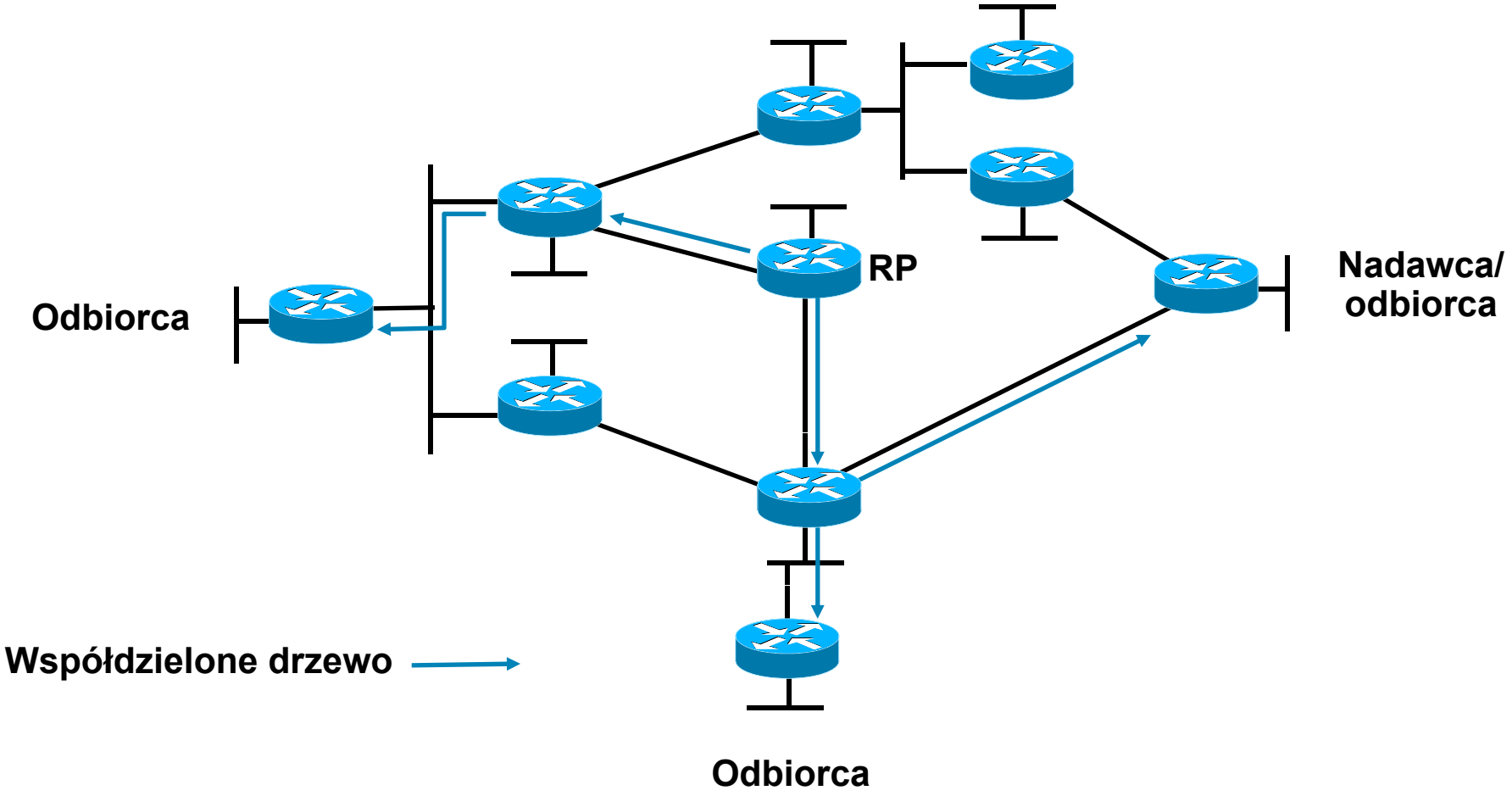
- Użycie drzew współdzielonych nieco łagodzi problem

  Redukcja ilości stanów (S, G)

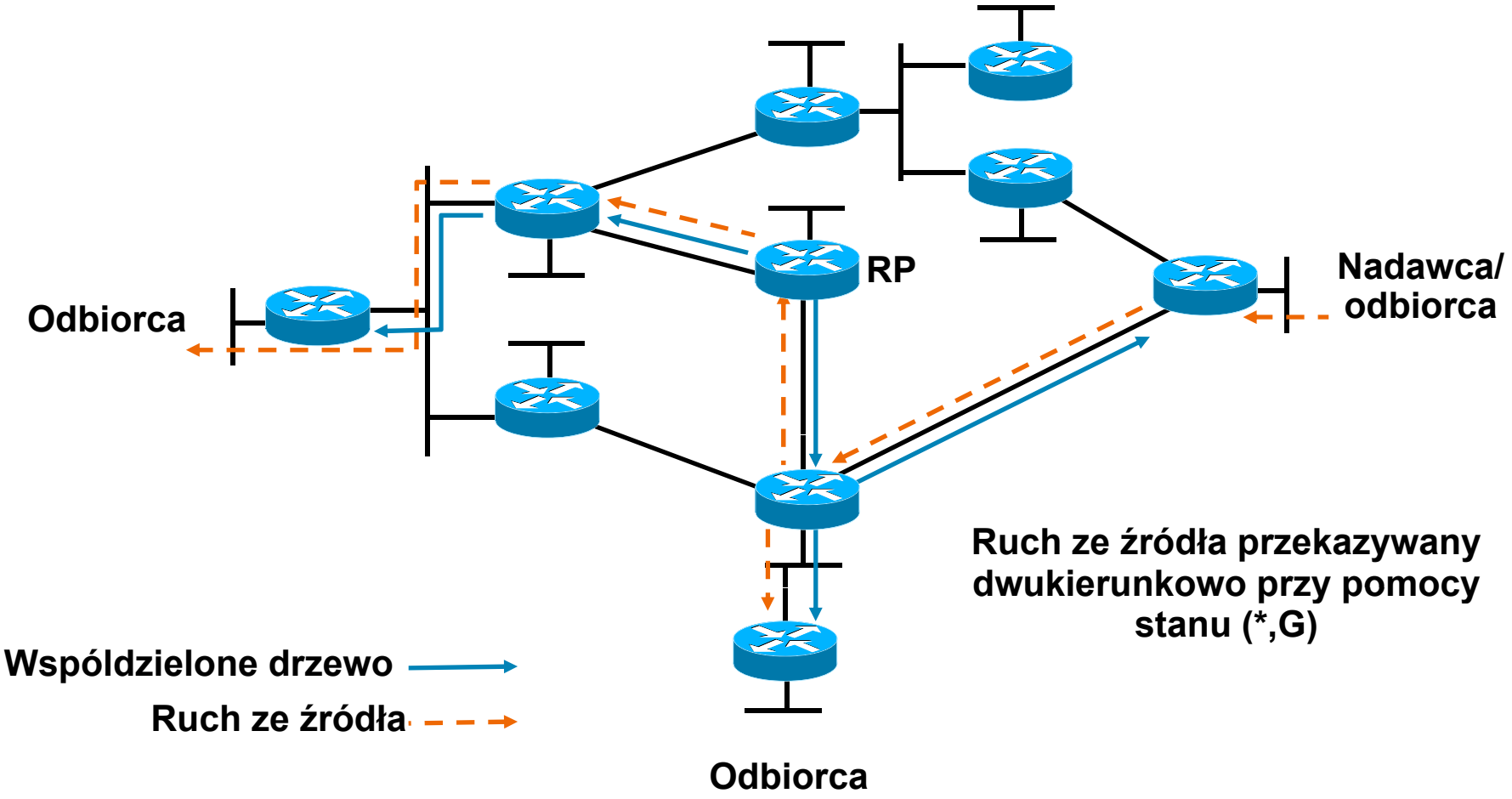
    Stan (S, G) tylko wdłuż SPT do RP

    Niestety nadal zwykle oznacza to zbyt dużo wpisów (S, G)

Bidirectional PIM—Jak to wygląda?



Bidirectional PIM—Jak to wygląda?



Bidir PIM—Podsumowanie

- Drastycznie minimalizuje ilość wpisów w tablicy multicastowej
  - Eliminuje **wszystkie** stany (S,G) w sieci
    - drzewa współdzielone między źródłami i RP również są wyeliminowane
    - ruch ze źródła trafia zarówno „w górę” jak i „w dół” drzewa współdzielonego
  - Aplikacje wiele-do-wielu mogą się w tym momencie skalować

Wybór RP

Skąd sieć wie gdzie jest RP?

- Konfiguracja statyczna
  - Ręcznie na każdym routerze w domenie PIM
- AutoRP
  - Rozwiązanie firmowe Cisco
  - Pozwoliło rozpowszechnić zastosowanie PIM-SM
- BSR
  - draft-ietf-pim-sm-bsr – „standard”

Statyczna konfiguracja RP

- Skonfigurowany na stałe adres RP

  Należy skonfigurować na każdym RP

  Wszystkie routery muszą mieć ten sam adres RP

  Failover jest niemożliwy – poza wykorzystaniem anycastu

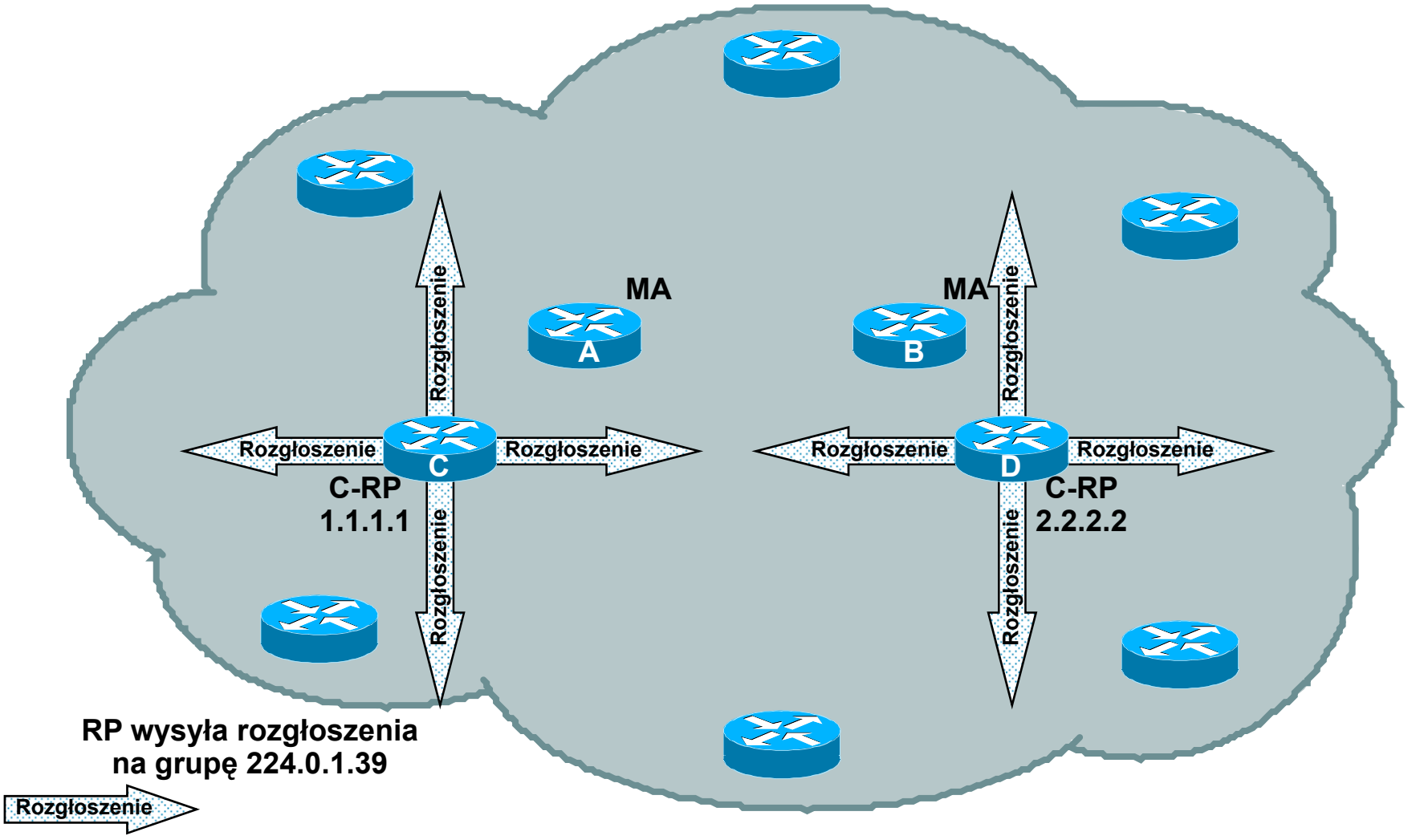
- Polecenie:

  `ip pim rp-address <adres> [group-list <acl>] [override]`

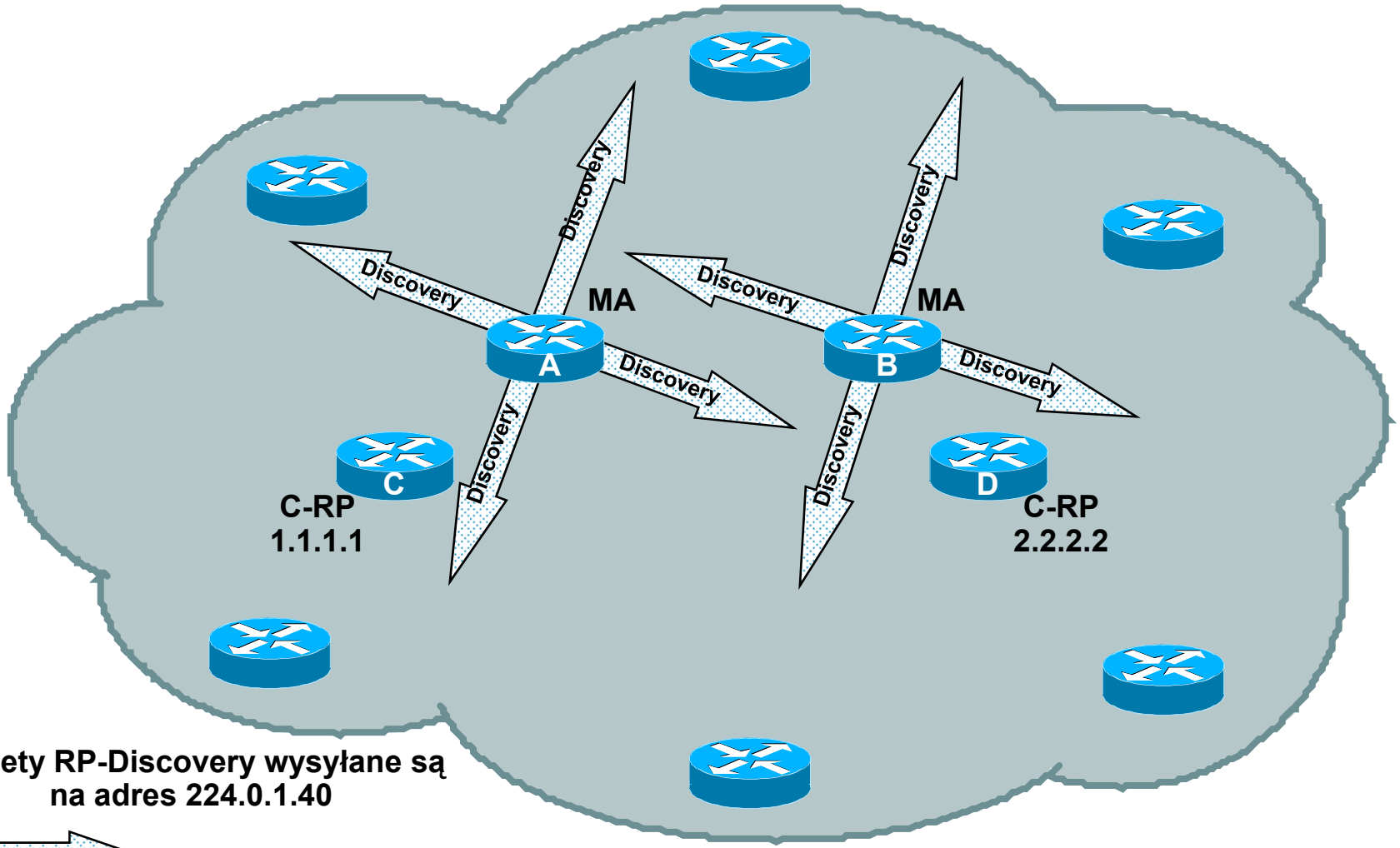
  group-list pozwala wskazać zakres grup

  domyślnie 224.0.0.0/4

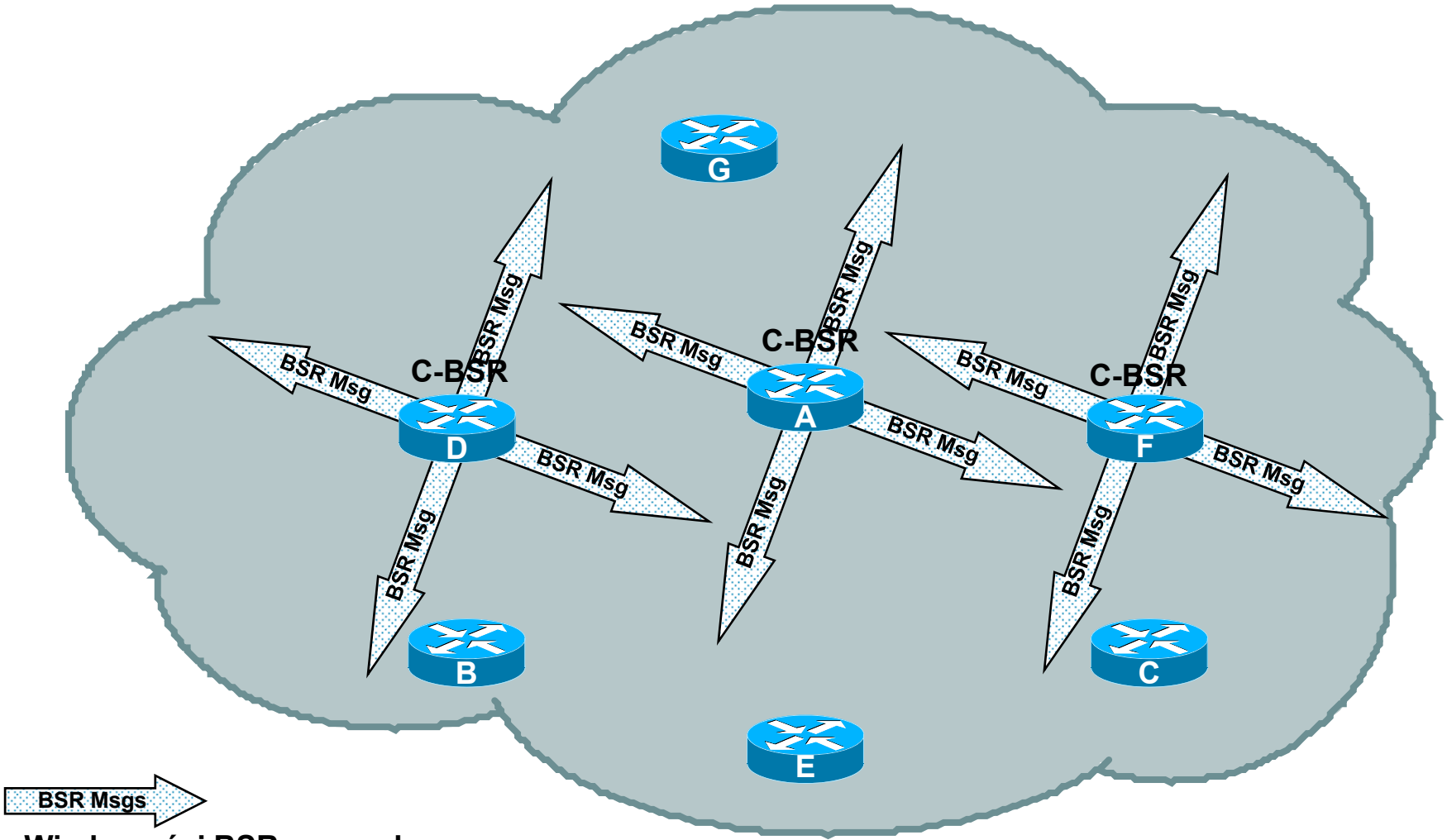
Auto-RP—z 10,000 metrów



Auto-RP—z 10,000 metrów



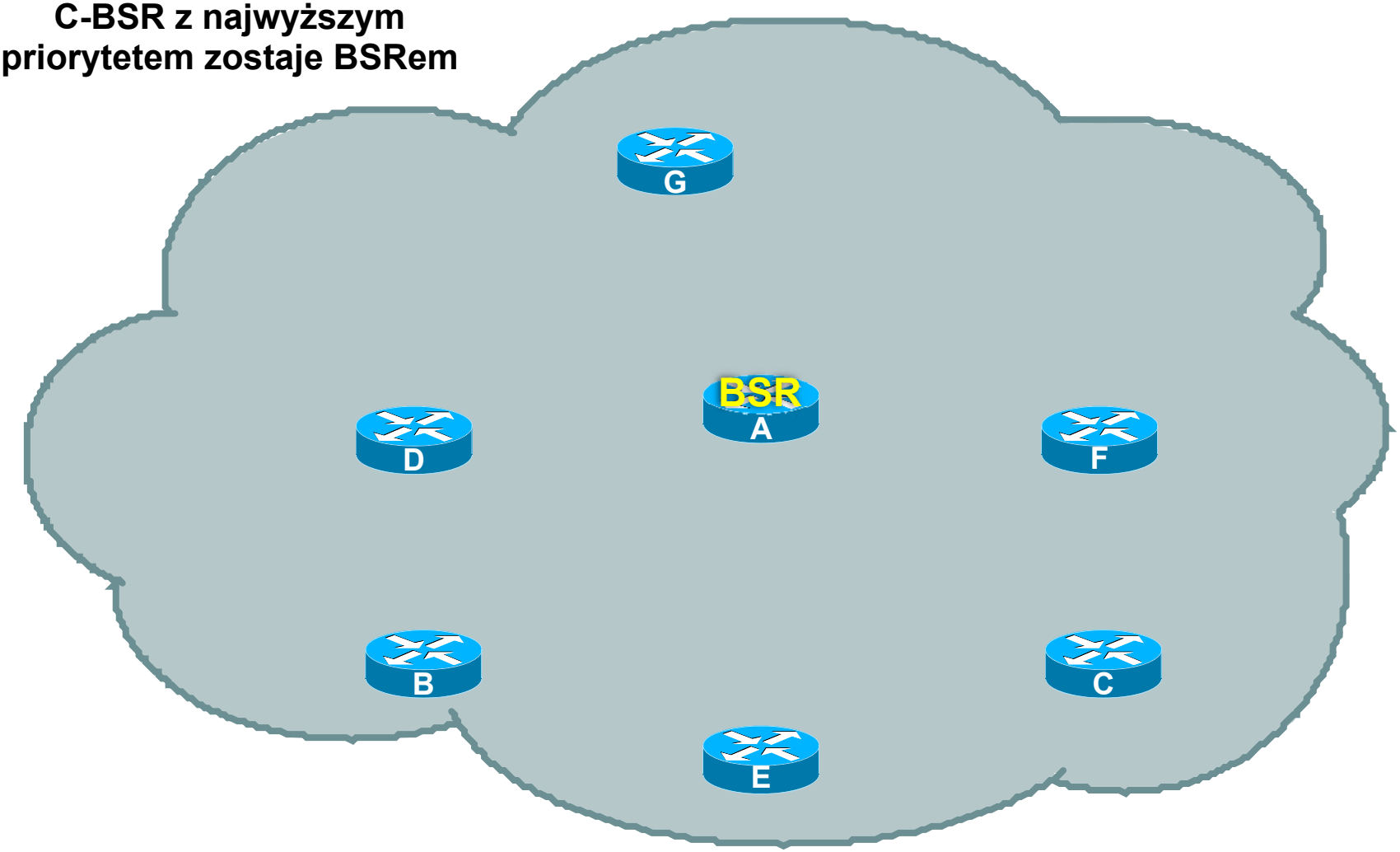
BSR—z 10,000 metrów



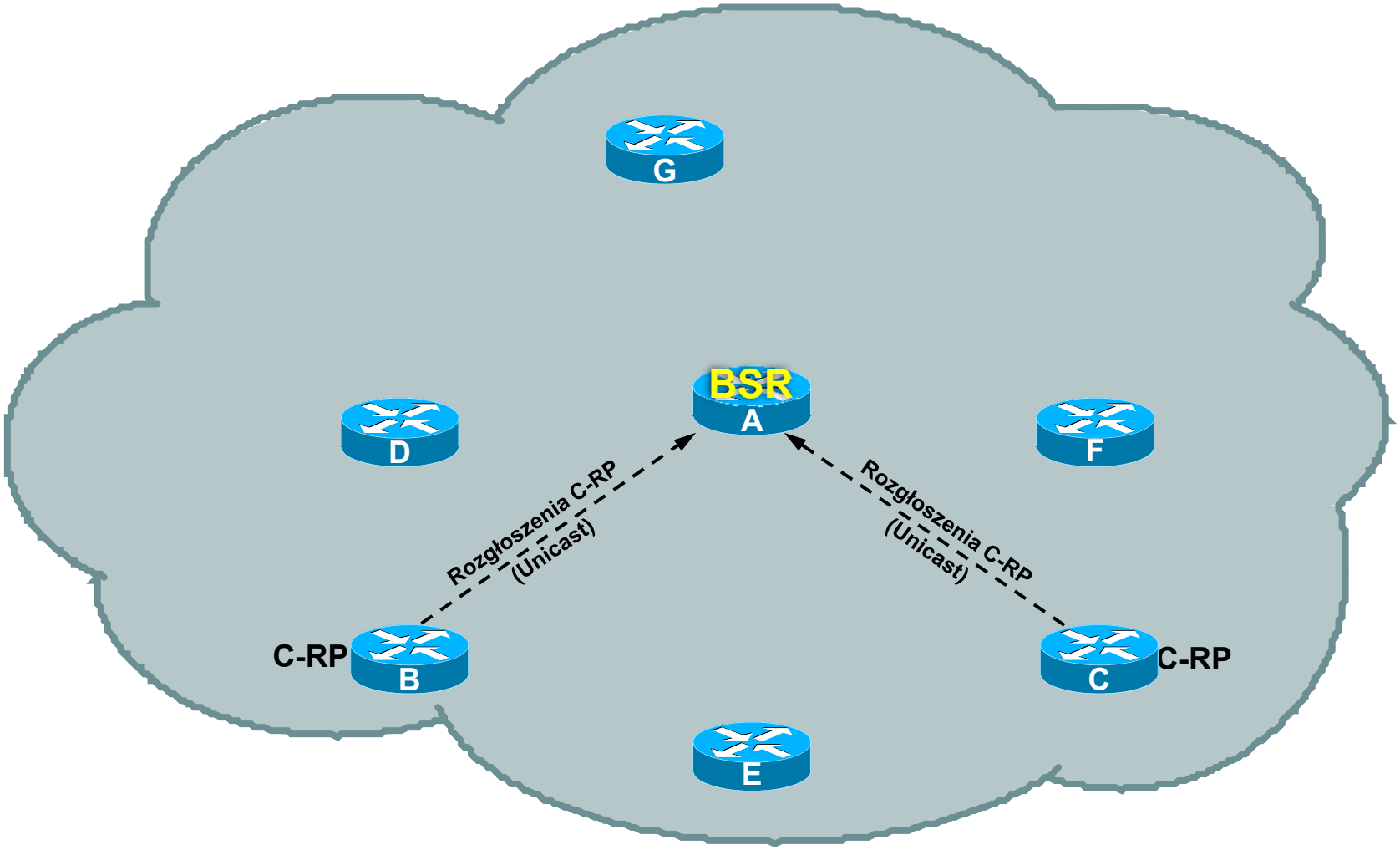
Wiadomości BSR są przekazywane hop-by-hop

BSR—z 10,000 metrów

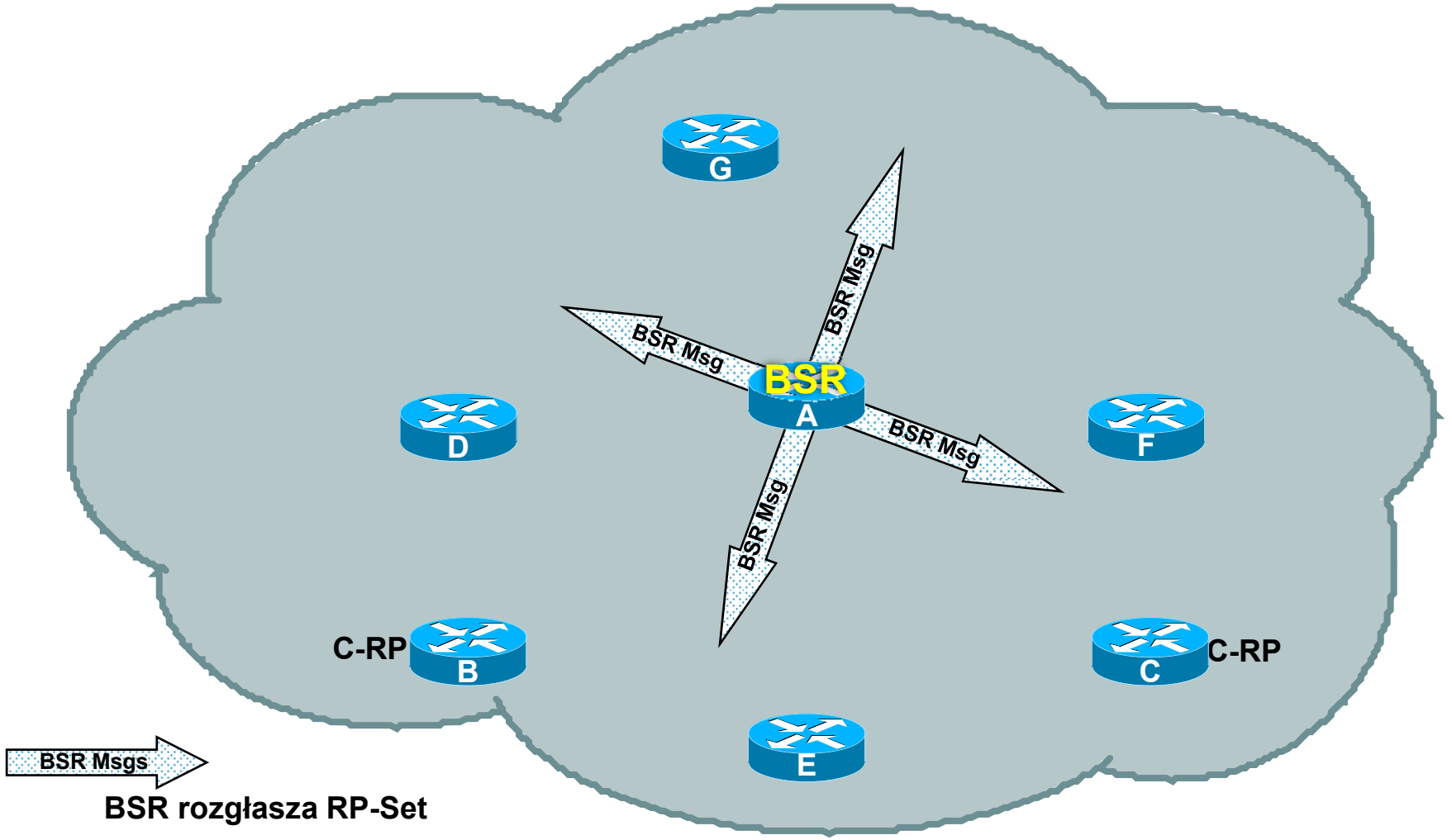
C-BSR z najwyższym priorytetem zostaje BSRem



BSR—z 10,000 metrów



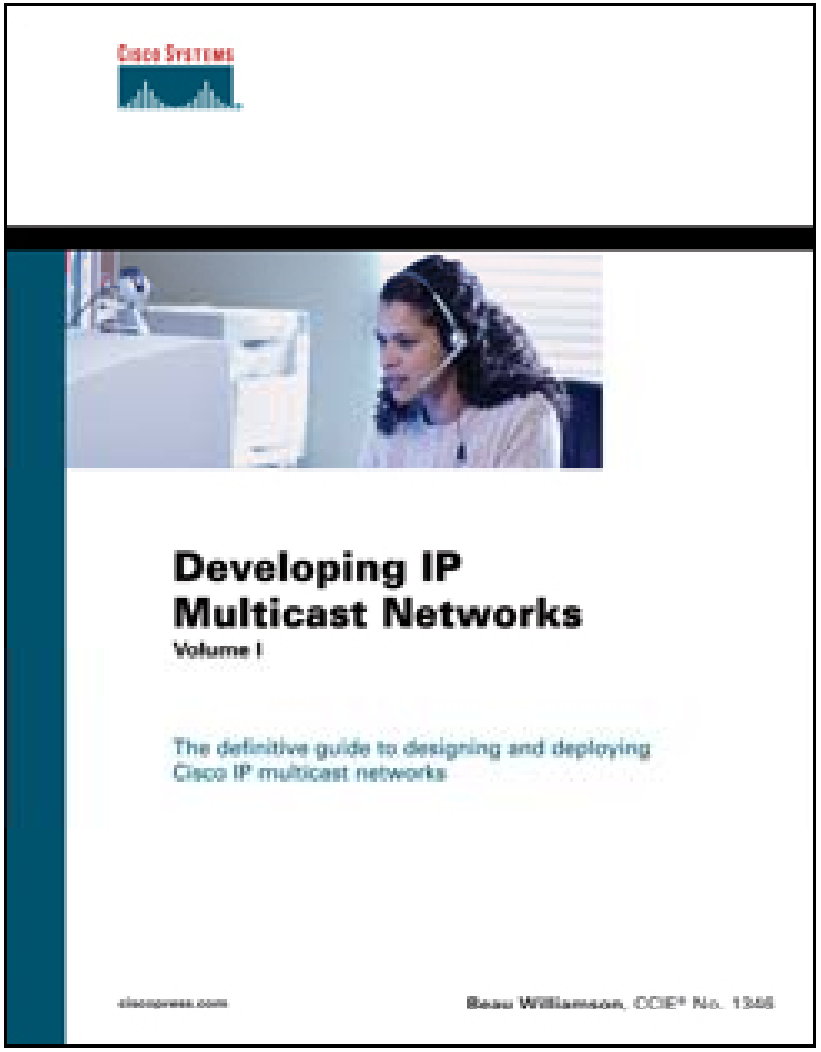
BSR—z 10,000 metrów

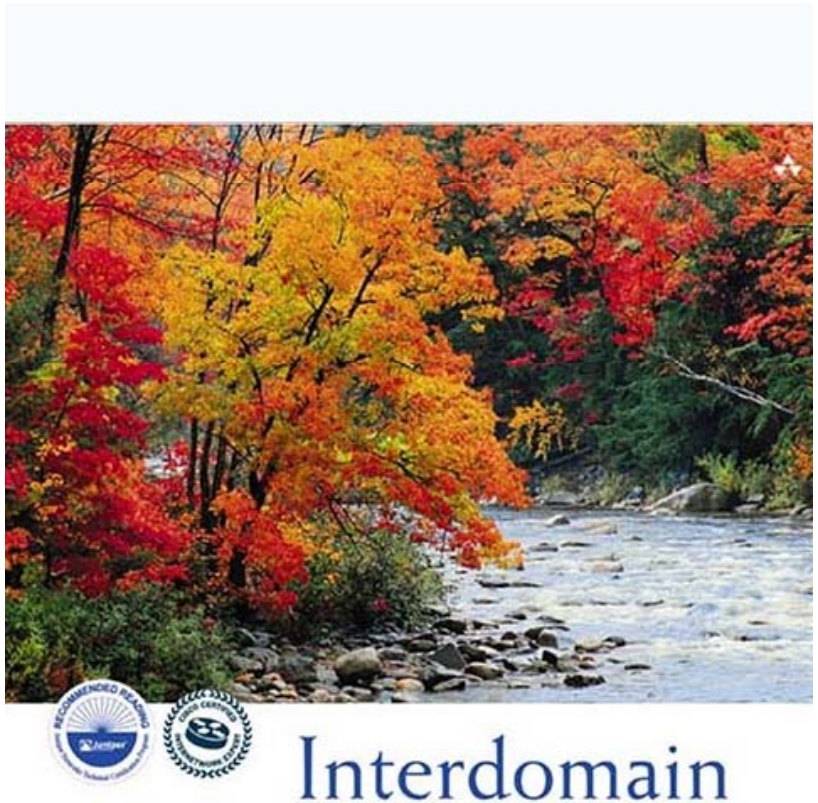


BSR rozgłasza RP-Set

Pytania?

Książki





I inne miejsca do poczytania...

- Cisco Configuring Mulitcast, 12.4T:

  http://www.cisco.com/en/US/docs/ios/ipmulti/configuration/guide/12_4t/imc_12_4t_book.html

- Juniper JunOS 9.6 multicast configuration guide:

  http://www.juniper.net/techpubs/en_US/junos9.6/information-products/topic-collections/config-guide-multicast/config-guide-multicast.pdf