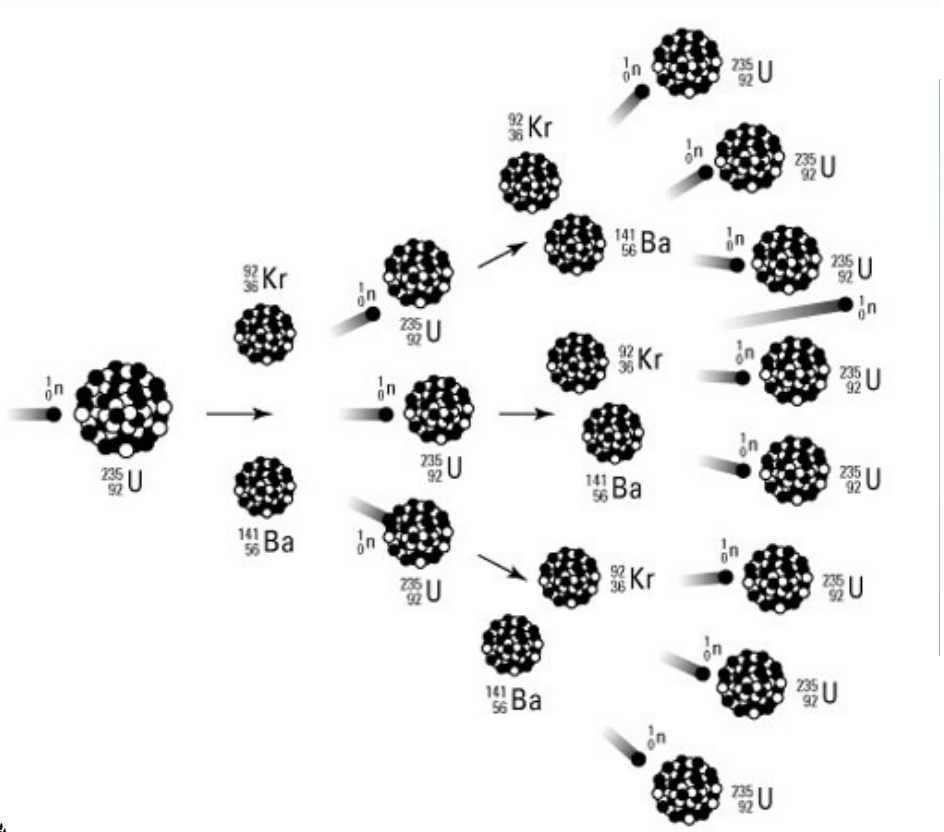
Next Generation Multicast VPN (MVPN)
PLNOG 01.2009



Bartłomiej Anszperger
Network Consulting Engineer
Cisco Advanced Services
banszper@cisco.com

Multicast dzisiaj

Multicast dzisiaj
Kto z tego korzysta?

- Rynek finansowy (Trading, Market Data, Financial SP)
  - Tibco, Hoot n Holler, Data Systems
- „Video collaborative environments”
  - Telepresence, DMS, MP/WebEx video conferencing, Video Surveillance
- Dostęp szerokopasmowy (rozrywka)
  - Cable, DSL, ETTH, LRE, Wireless
  - Video 2.0 iQuadplay
  - Broadcast TV / IPTV, VOD, Connected Home
- Operatorzy (usługi tranzytowe)
  - „Natywny” multicast v4 i v6
  - Label Switched Multicast (LSM)
  - Multicast VPNs (IP i MPLS)

Multicast dzisiaj
Mnogość rozwiązań

- Zróżnicowane wymagania = różne rozwiązania
  - Z reguły klient usługi L3VPN chce by jego mulitcast po prostu działał
  - Dostawca IPTV chciałby móc dokladniej sterować tworzeniem drzewa P2MP po którym dostarczany jest do odbiorcy ruch
- To powoduje istnienie wielu możliwych alternatyw budowy sieci multicast'owych
  - Nawet w prostym scenariuszu wdrożenia istnieje kilka możliwości, których wybór przez projektanta jest zdeterminowany często bardzo małymi różnicami w wymaganiach klienta
  - Czynniki takie jak szybkość napływu komunikatów *join* czy specyficzna topologia sieci również wymagają kompromisów
- IP multicast staje się bardzo ważną usługą dla operatorów
  - Dlatego też istnieje przekonanie, że rozwiązania czysto L2 mają ograniczoną wartość

Multicast dzisiaj
IETF

- PIM WG
  - Niezawodność
    - PIM over TCP (draft-farinacci-pim-port-00)
- MBONED WG
  - MVPN Deployment (draft-ycai-mboned-MVPN-pim-deploy-02)
  - AMT (draft-ietf-mboned-auto-multicast-08)
- L3VPN WG
  - MVPN (draft-ietf-l3vpn-2547bis-mcast-06)
  - BGP vs PIM (draft-rosen-l3vpn-MVPN-profiles-00)
- MPLS WG
  - LSM
    - MLDP / P2MP RSVP-TE
- MSEC, SOFTWIRES, BEHAVE, AVT, MMUSIC, FECFrame, ANCP, L2VPN, RMT, BMWG

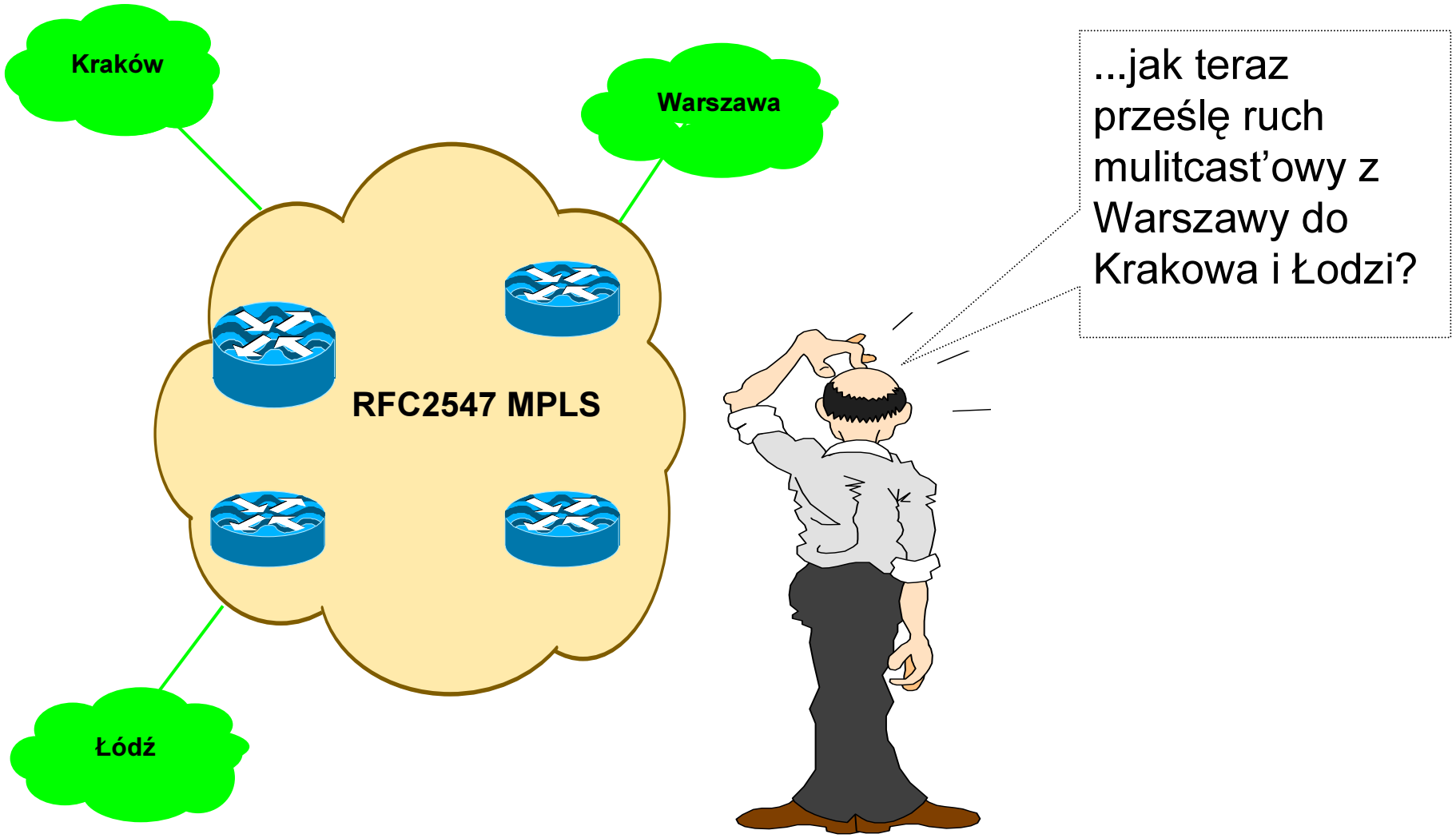
MVPN następnej generacji (aka MVPN 2.0)

MVPN 2.0

- Wstęp do MVPN
- MVPN 2.0
- MoFRR
- LSM - Label Switched Mulitcast
  - –mLDP
  - –p2mp TE
- BGP vs PIM

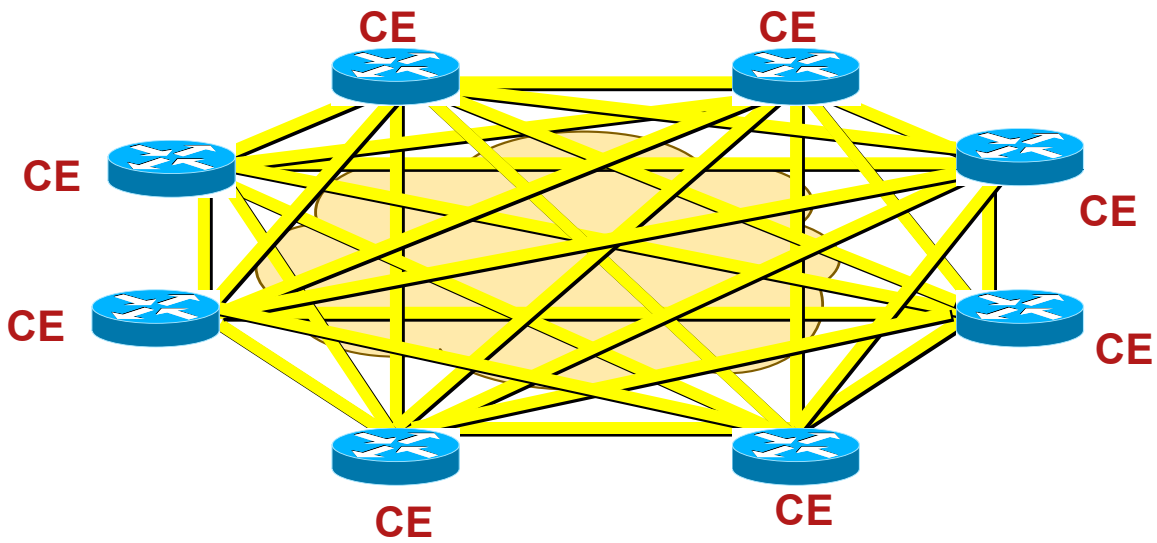
Wstęp do MVPN

Zbudowałem usługę MPLS based L3 VPN i …



Multicast VPN – najprostsze rozwiązanie

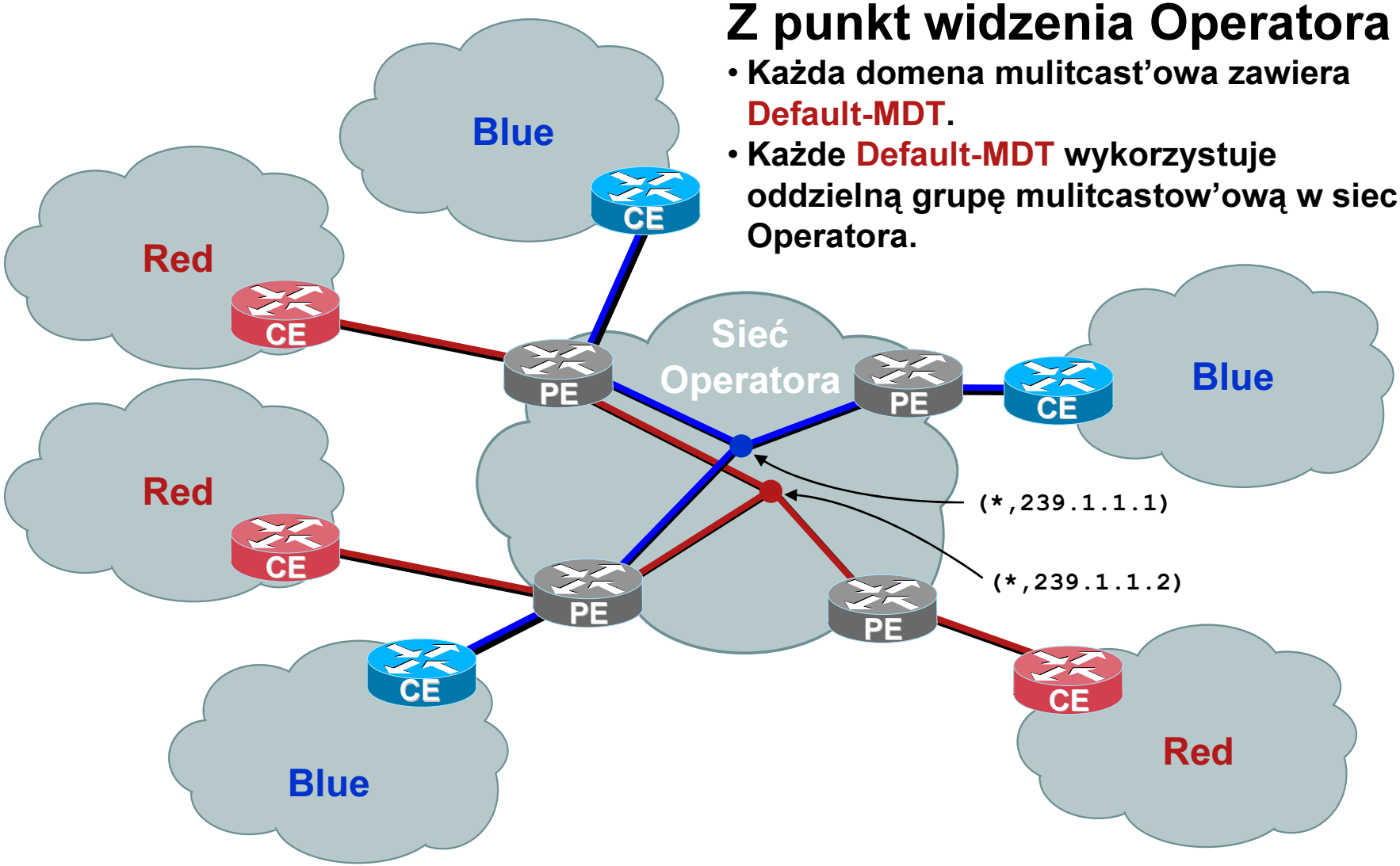
- Brak wsparcia w MPLS dla Multicast'u
- Workaround: tunele punkt-punkt w relacji CE - CE
- Rozwiązanie nieskalowalne wraz ze wzrostem ilości routerów CE!
  - –Overhead związany z ruchem
  - –Overhead związany z administracją



Multicast VPN – jak to działa?

Z punkt widzenia Operatora

- Każda domena mulitcast'owa zawiera Default-MDT.
- Każde Default-MDT wykorzystuje oddzielną grupę mulitcastow'ową w sieci Operatora.

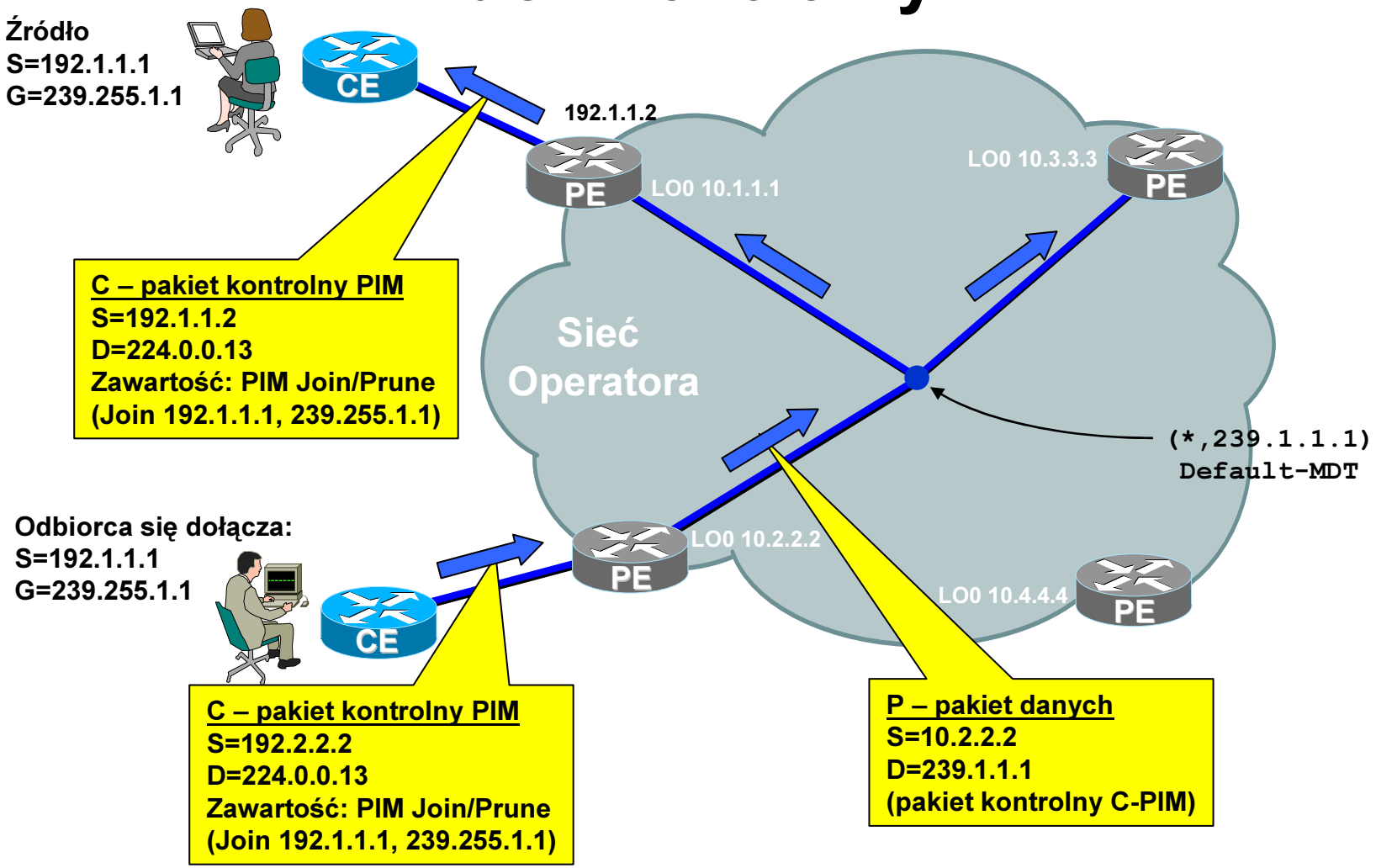


Dwa typy grup MDT

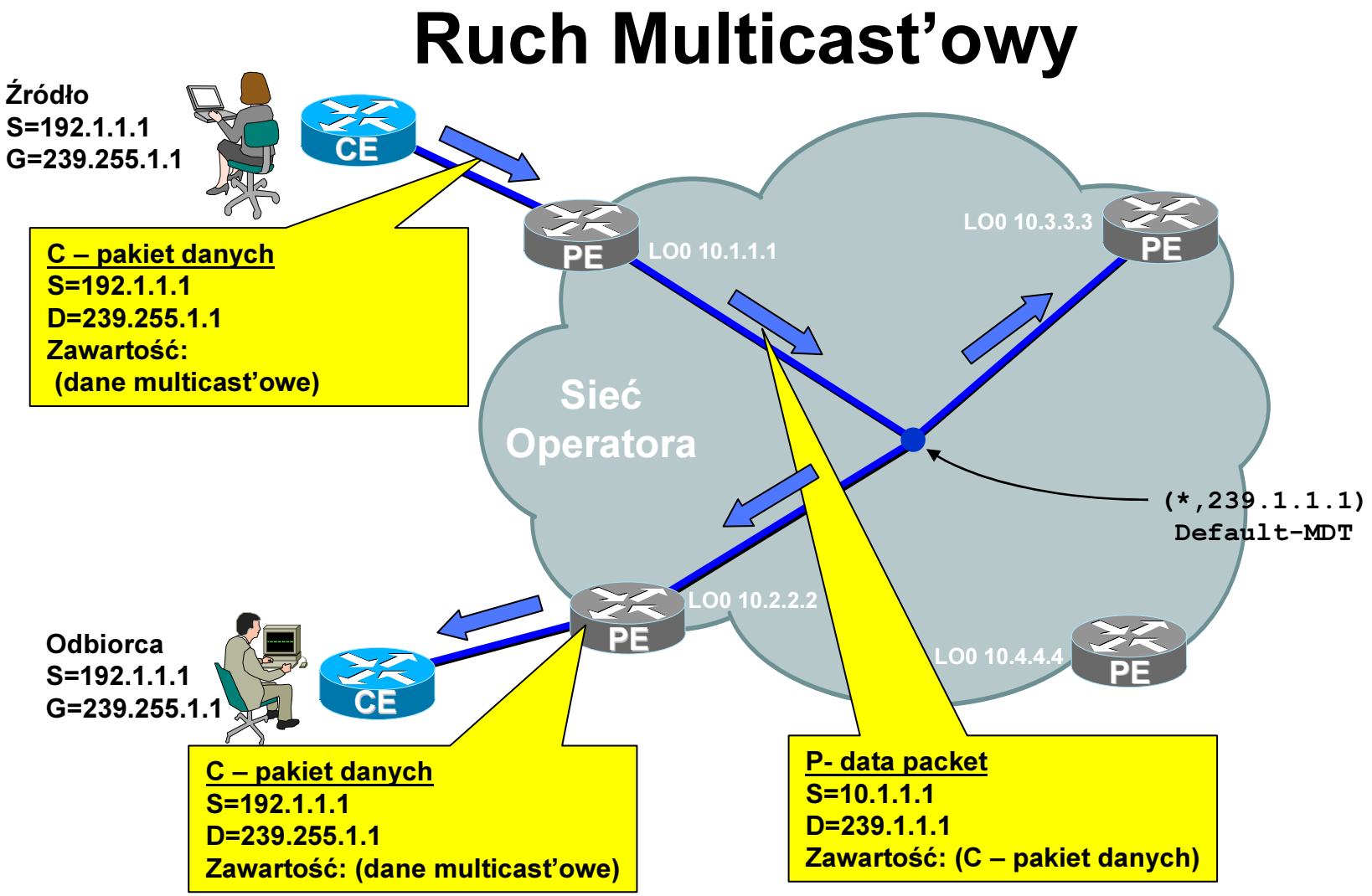
- Grupy Default MDT
  - Obowiązkowo skonfigurowana dla każdego MVRF
  - Wykorzystywana dla
    - ruchu kontrolnego PIM
    - ruchu ze źródeł o małych wymaganiach na pasmo
    - Rozsyłania ruchu typowego dla „dense-mode”
  - "Multidirectional Inclusive" PMSI (MI-PMSI wg *2547bis-mcast*)
- Grupy Data MDT
  - Opcjonalna dla danego MVPN
  - Używana przez źródła o dużych wymaganiach na pasmo tak by ograniczyć replikacje dużej ilości ruchu do PE niezainteresowanych odbiorem danej transmisji
    - "Selective" PMSI (S-PMSI wg *2547bis-mcast*)

Default MDT z bliska...

Ruch kontrolny PIM

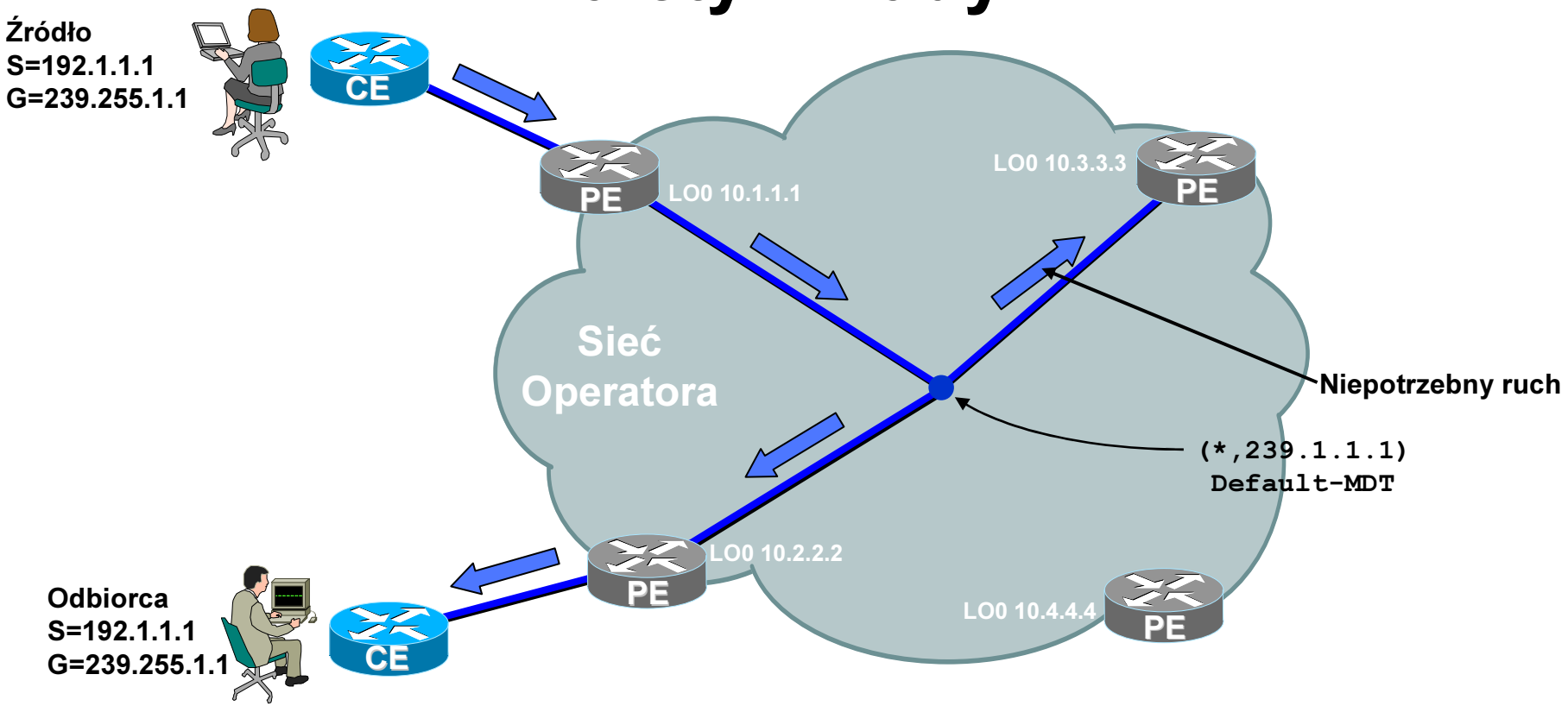


Default MDT z bliska...



Default MDT z bliska...

Zalety i Wady

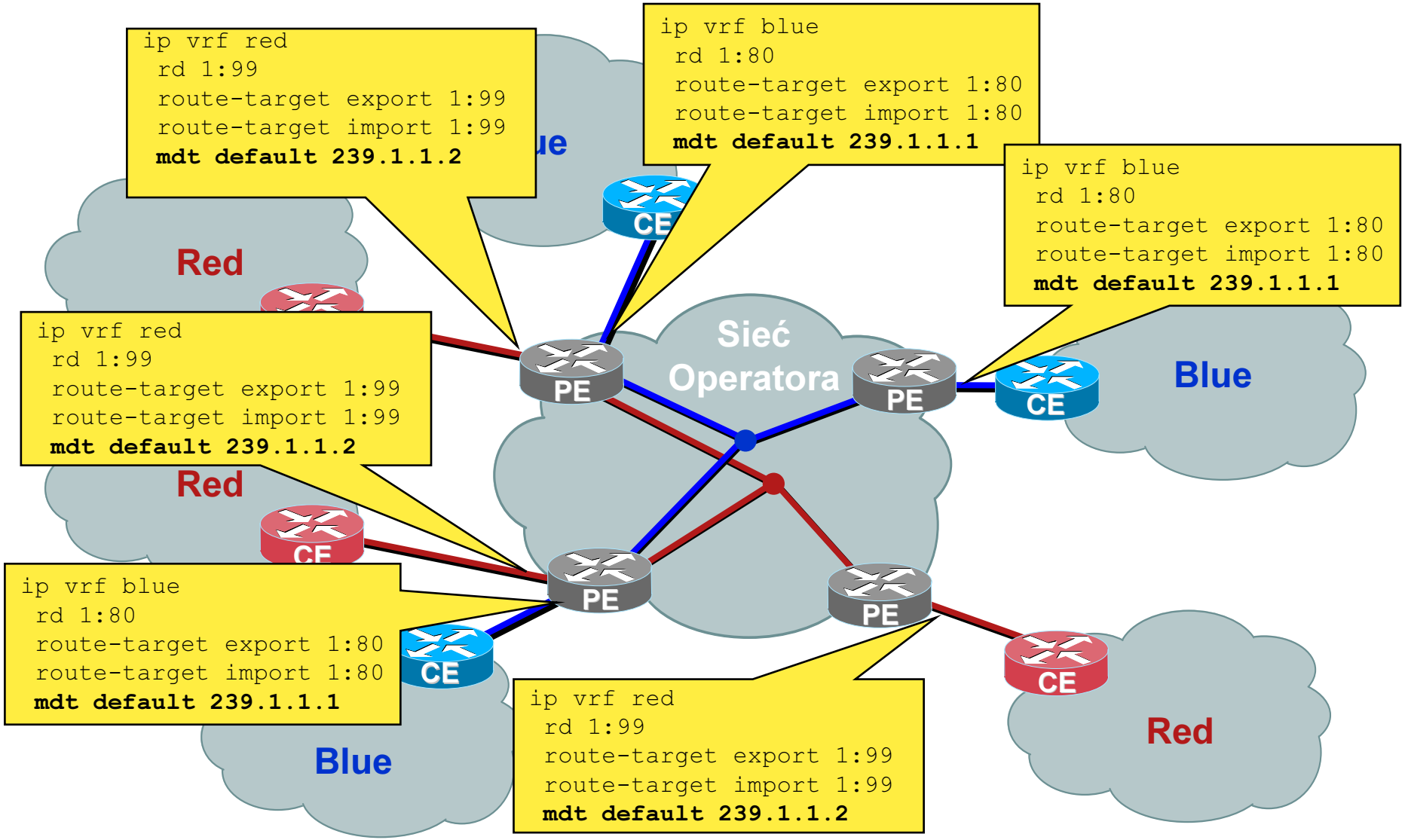


Zaleta : Redukcja stanów mulitcast'owych w routerach P

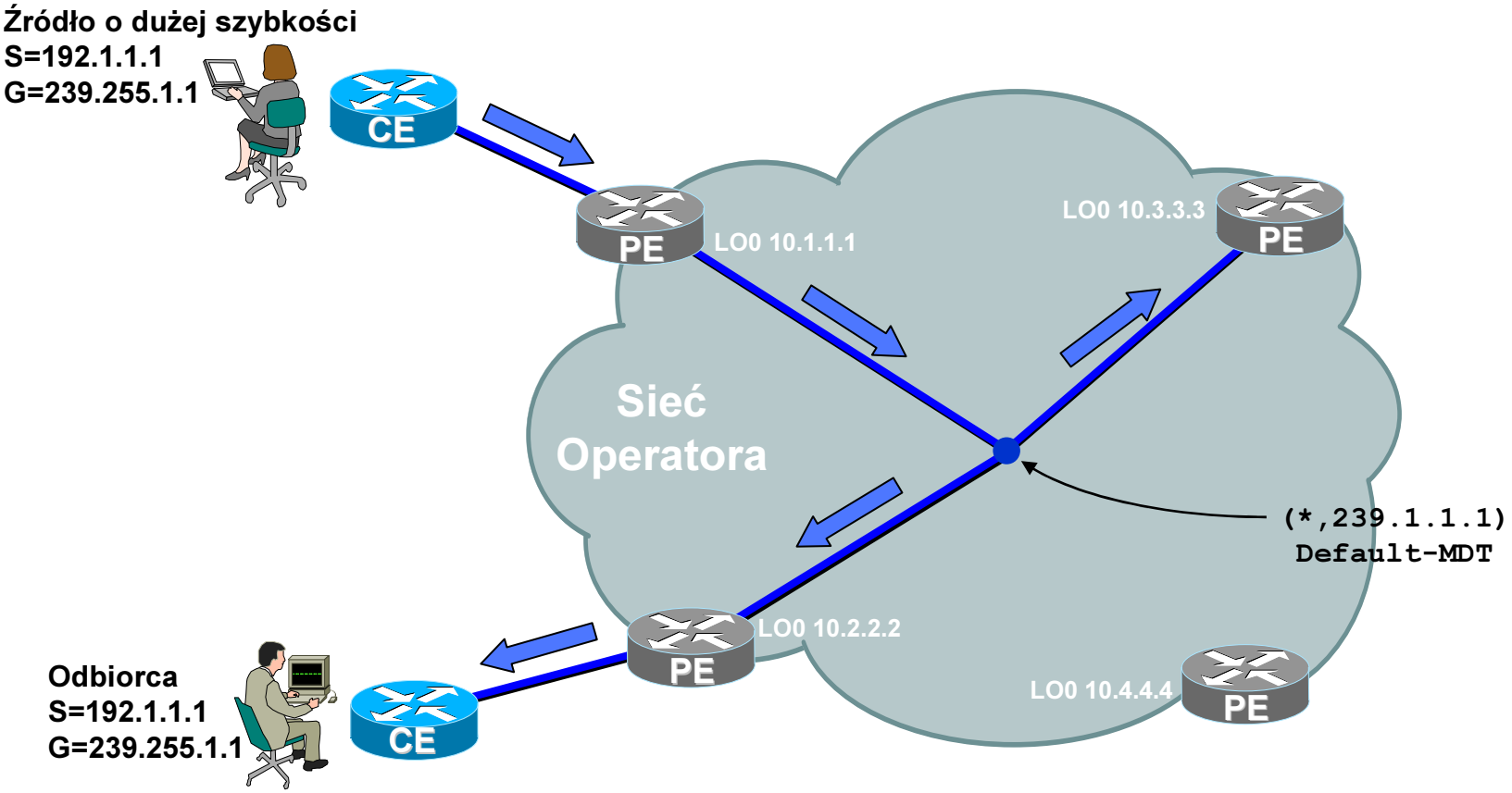
Wady : Marnotrawstwo pasma.

Rozwiązanie : Użyj oddzielnych Data-MDT dla źródeł o dużych wymaganiach

Default MDT - przykład konfiguracji

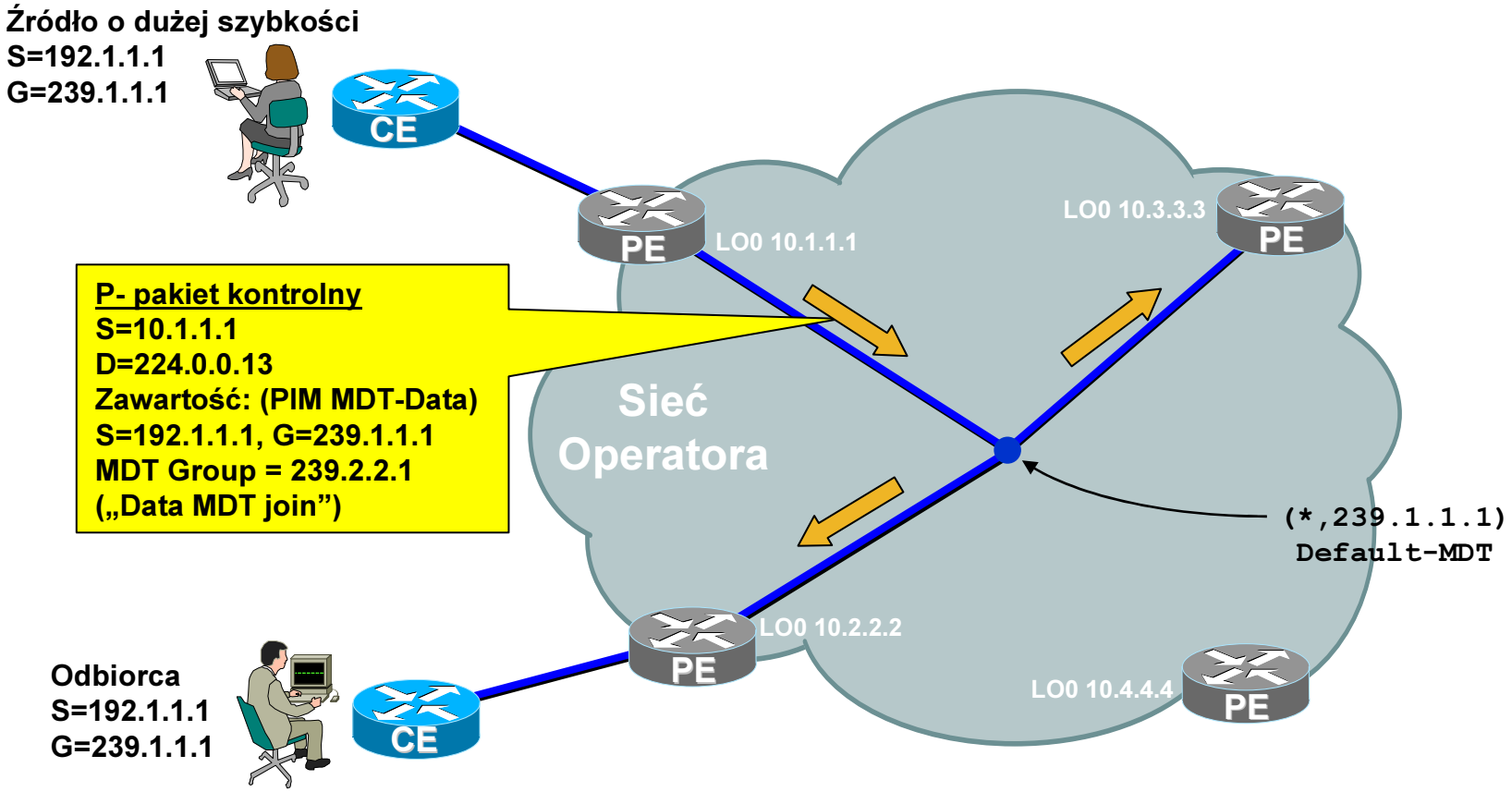


Data MDT z bliska...



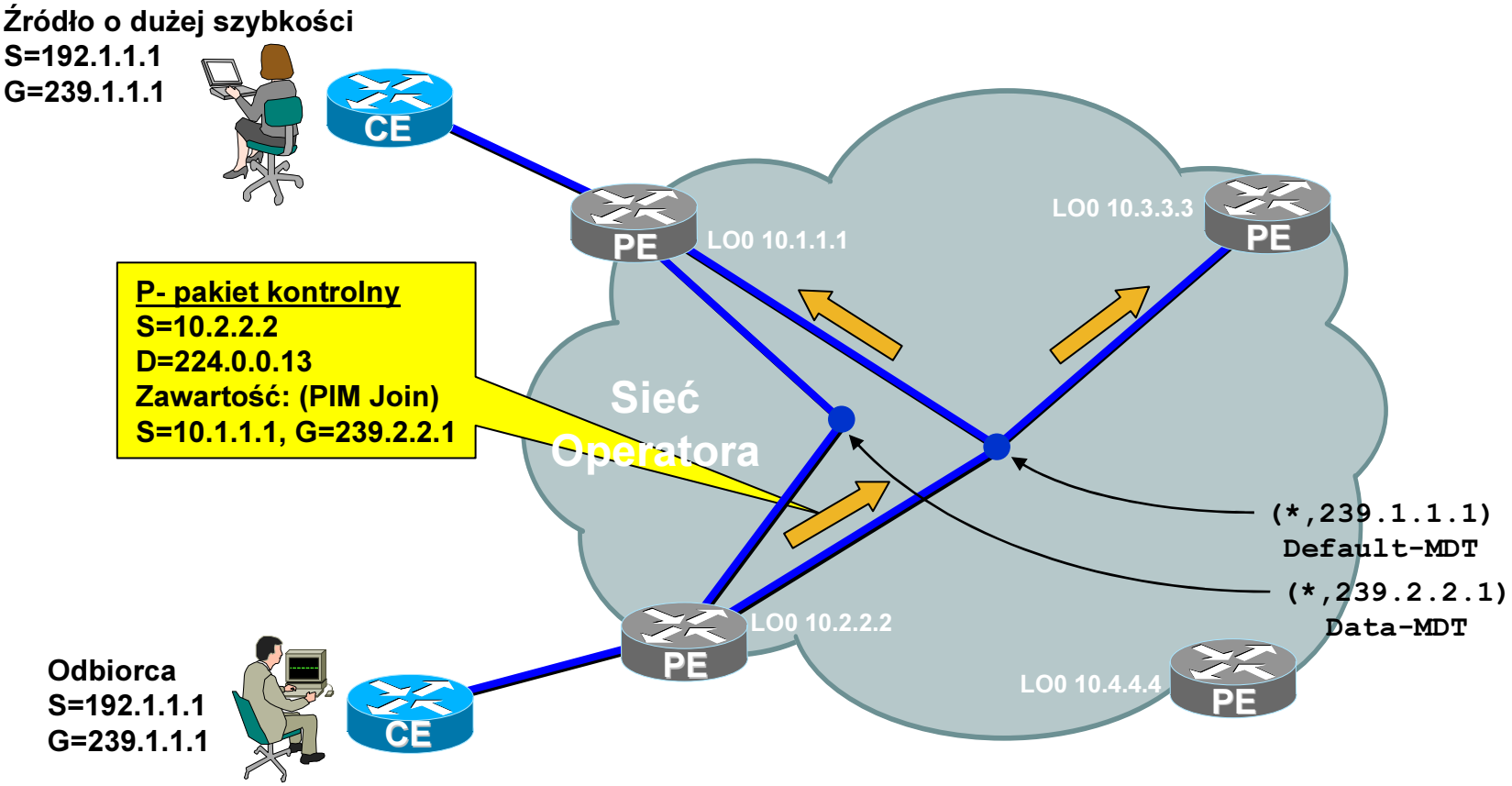
- **Ruch wykraczający ponad poziom skonfigurowany dla Data-MDT na PE.**

Data MDT z bliska...



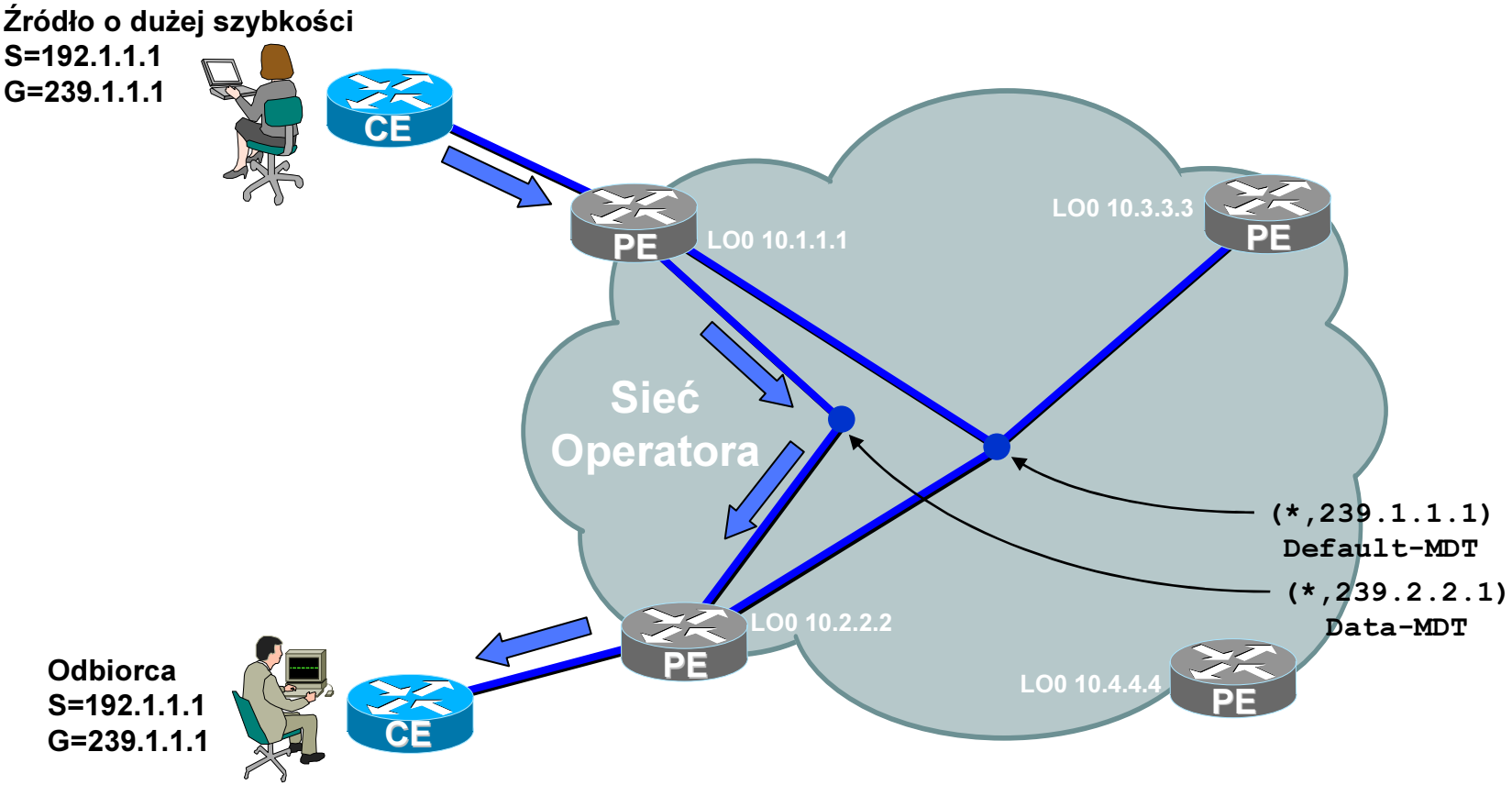
- **Router PE sygnalizuje konieczność przełączenia na Data-MDT z wykorzystaniem nowej grupy 239.2.2.1**

Data MDT z bliska...



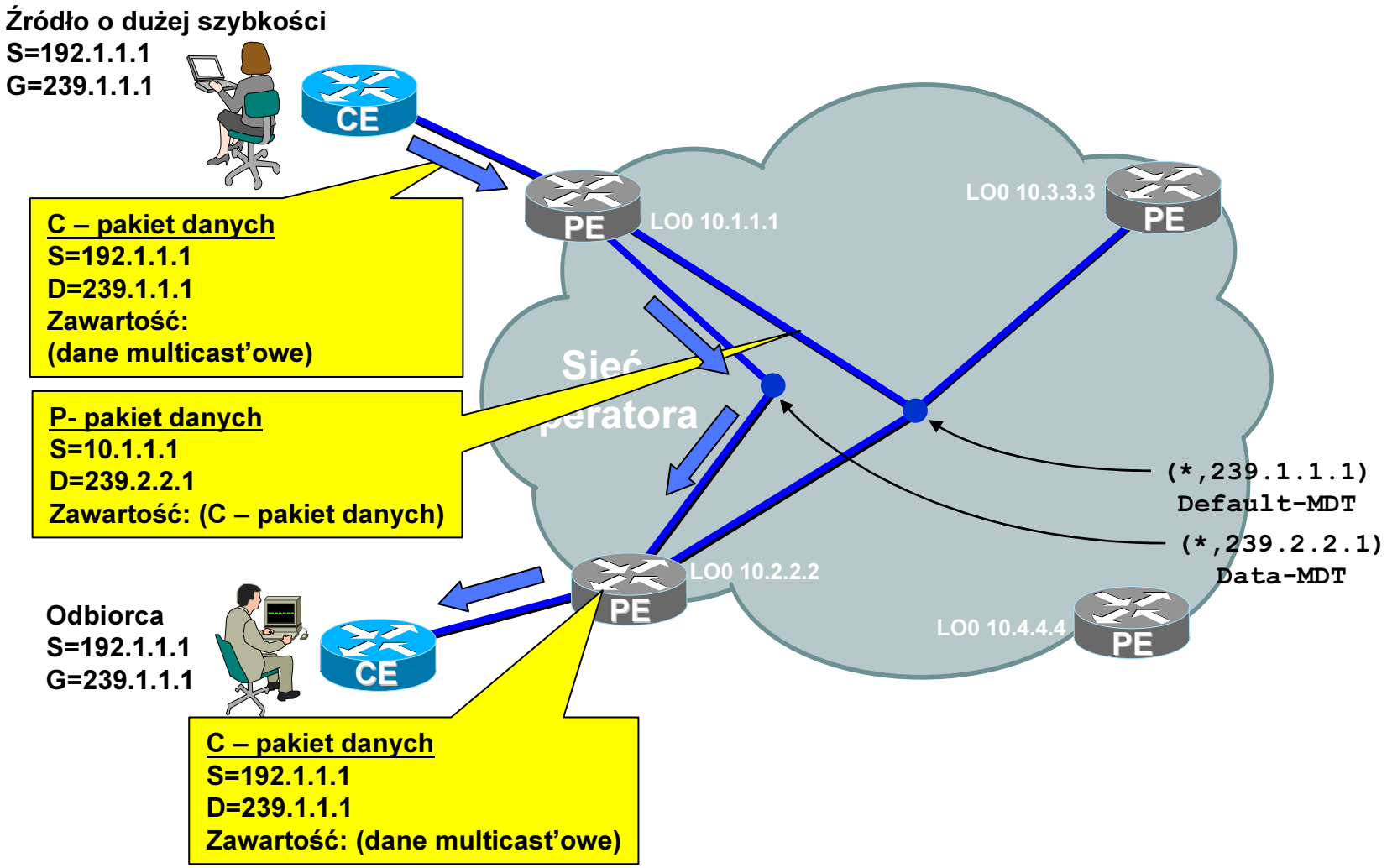
- **Router PE zainteresowany odbiorem ruchu wysyła „Join” na adres nowej grupy 239.2.2.1.**
- **Powstaje Data-MDT w oparciu o grupę 239.2.2.1.**

Data MDT z bliska...



- **Ruch ze źródła o duzej szybkości zaczyna płynąć poprzez Data-MDT.**
- **Dane docierają jedynie do PE które maja zainteresowanych odbiorców.**

Data MDT z bliska...



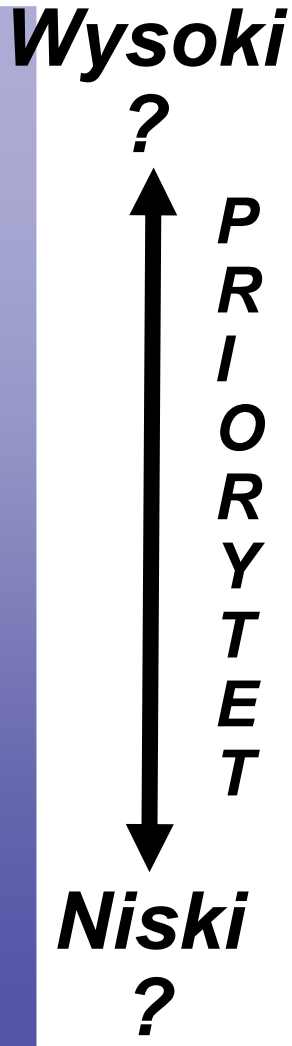
MVPN 2.0

MVPN dzisiaj i jutro

- MVPN-GRE (Data/Default-MDT) szeroko stosowane przez różnych vendor'ów na różnych platformach
- Główne obszary rozwoju:
  - MVPN-GRE: ciągle aktualny wybór! Kontynuowane prace nad lepszym dopasowanie do potrzeb klientów
  - LSM (Label Switched Multicast) - ścieżki LSP typu P2MP, MP2P, MP2MP
  - MLDP dla „typowego" MVPN: alternatywa dla IPv4-multicast(+GRE) w MPLS
  - RSVP-TE: ograniczone zastosowanie (skalowalność!), głównie w środowiskach gdzie wymagane jest wsparcie dla TE, możliwość łączenia z MLDP.
  - Inne mechanizmy „usprawniające" wynikające ze specyficznych wymagań klientów
- Usprawnienia dotyczące niezawodności
- L2VPN ?! (chęć konkurencji w „broadband edge")
- BGP zamiast PIM ?

MVPN 2.0 – obszary prac

- Poprawa elastyczności polityk mapowania ruchu do I-PMSI i S-PMSI
  - „explicit mapping”, anycast, Bidir-PIM, …
- Przesyłanie multicast za pomoca eytkiet MPLS
  - MLDP – natywnie i VPN
- Większa niezawodność
  - “sub 100..50 msec” / “TE” / MoFRR/live-live / FRR…
- L2VPN?
- RSVP-TE P2MP
  - + rozszerzenia…
- “Better scalability for signaling” ?
  - brak PIM Hello, „niezawodny” PIM
- Zamian sygnalizacji z PIM na BGP “BGPim”?!



MVPN 2.0
Lepsza elastyczność polityk mapowania ruchu

„Explicit tree to S-PMSI mapping”

- Możliwość mapowania grup mulitcastowych z VRF do docelowych grup MDT w sposób deterministyczny
- Obecnie proces ten zachodził *„dosyć losowo”*
- Łatwiejszy troubleshooting i traffic engineering

Data MDT = S-PMSI
Stara Notacja = Nowa Notacja

„Limitless S-PMSIs”

- Obecnie istnieje ograniczenie 255 MDT – jeśli istnieje taka konieczność MDT są wykorzystywane ponownie.
- Możliwość mapowania wymaganej ilości usług multicastowych do ustalonego MDT i dostarczenia ich jedynie do PE które tego wymagają
- Rozważane jest użycie 1024 group MDT jako rozwiązanie krótkoterminowe

Data MDT = S-PMSI
Stara Notacja = Nowa Notacja

„Mapping of shared trees on S-PMSI”

- Cały ruch przenoszony za pomocą (*,G) z wykorzystaniem albo *SPT threshold* ustawionym na nieskończoność bądź jako grupa BiDir przesyłany jest obecnie jedynie po Default-MDT
- To powoduje, że ruch przesyłany po MVPN jest rozsyłany do wszystkich PE
- Nowy *feature* dający możliwość wysyłania ruchu (*,G) za pomocą Data-MDT po przekroczeniue ustalonego i zdefiniowanego poziomu

Data MDT = S-PMSI
Stara Notacja = Nowa Notacja

„Inhibit reuse of dynamic mapped S-PMSI”

- Obecnie wiele ruch kliencki jest arbitralnie mapowany na dostępne Data-MDT
- Możliwość dokładnej kontroli ilości pasma dostępnego w szkielecie sieci dla ruchu klienckiego poprzez zabronienie ponownego używania raz wykorzystanych Data MDT
- Poprawia kontrole na wielkością ruchu transmitowanego poprzez szkielet sieci

Data MDT = S-PMSI
Stara Notacja = Nowa Notacja

„Immediate use of S-PMSI”

- Możliwość natychmiastowego użycia Data MDT – powstrzymanie początkowego flood’owania ruchu do wszystkich PE poprzez Default-MDT (I-PMSI)
- Zabezpieczenie przed krótkotrwałymi przeciążeniami w sieci

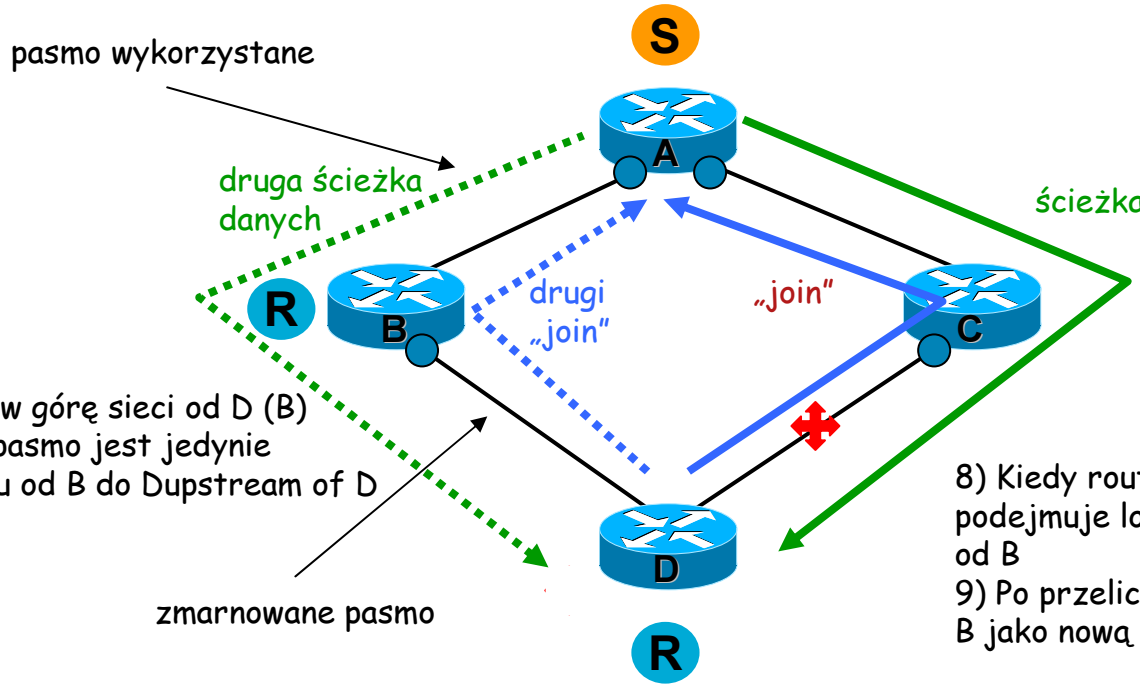
Data MDT = S-PMSI
Stara Notacja = Nowa Notacja

MoFRR

MoFRR

- Rozwiązanie typu „make-before-break”
- Routing multicast’owy nie musi „czekać” aż routing unicast’owy się ustabilizuje
- MoFRR moży być rozpatrywane jako alternatywa dla redundancji źródeł, ale
  - Nie trzeba konfigurować zapasowych źródeł ;-)
  - Nie ma problemu synchronizacji strumienia multicast’owego pomiędzy źródłami
  - Odbiorcy nie dostają zduplikowanego ruchu
- Brak tuneli zapasowych
- Żadnych nowych protokołów do zestawiania drzew
- Brak zmian w sprzęcie

MoFRR – jak to działa?



7) Jeśli router w górę sieci od D (B) ma odbiorców, pasmo jest jedynie tracone na łączu od B do Dupstream of D

8) Kiedy router C bądź link CD pada, D podejmuje lokalną decyzję i akceptuje pakiety od B
9) Po przeliczeniu unicast routing wskazuje B jako nową ścieżkę RPF

- ścieżka danych
- ścieżka rpf (RPF join)
- Ścieżka drugiego „join" (nie RPF join)
- Interfejsy w liście oif
- Odrzucanie pakietów/pad linku

1) D ma ścieżkę ECMP {BA, CA} do S
2) D wysyła „join" po ścieżce RPF przez C
3) D może wysłać kolejny „join" po ścieżce BA
4) A ma dwa interfejsy na swojej liście oif do tego samego odbiorcy
5) Kiedy ścieżka RPF jest aktywna D otrzymuje ruch zduplikowany
6) D odrzuca pakiety (RPF!) od B

Label Switched Mulicast (LSM)

LSM a MVPN

- MVPN z założenia odróżnia usługę (PMSI) od jej instancji (tuneli czyli mechanizmu przesyłania ruchu)
- Każda PMSI może posiadać zbiór jednego lub kilku tuneli
- Tunele mogą zostać zbudowane w oparciu o:
  - PIM (dowolna odmiana)
  - MLDP p2mp lub mp2mp
  - RSVP-TE p2mp
  - Kombinacje tuneli unicast'owych i replikacja na wejściowym PE
- Możliwość mapowania kilku PMSI w jeden tunel (agregacja)
- Enkapsulacja jest funkcja tunelu a nie usługi
- Pojedynczy operator może wykorzystywać tunele o dowolnym typie

Status LSM

Protokoły LSM	Podstawowe Własności
MLDP draft-ietf-mpls-ldp-p2mp-05	Dynamiczna budowa drzew multicastowych dla różnych typów aplikacji multicastowych Opcjonalna możliwość FRR Drzewa budowane od strony odbiorcy („*receiver driven dynamic tree building*”)
P2MP RSVP-TE RFC 4875	Deterministyczna gwarancja pasma w całym drzewie Drzewa budowane od strony nadawcy („*head end defined trees*”) „wbudowane” FRR Użyteczne dla mniejszej choć ważnej części aplikacji multicastowych np. broadcast TV z ograniczeniami pasma

MVPN:
PIM vs P2MP-TE vs MLDP

PIM+GRE

- **Szeroko stosowane podejście**
- **Wymagany upgrade jedynie routerów brzegowych**
- **Multivendor**
- **Dynamicznie zastawiane drzewa przez odbiorcę**
- **p2mp i mp2mp**
- **Skalowalność O(N), N=#ilość PE**
- **Wsparcie dla scenariuszy Inter-AS**
- **Wsparcie dla extranet**

MVPN:
PIM vs P2MP-TE vs MLDP

P2MP-TE

- **sub-50ms FRR**
- **Konieczny upgrade większej ilości routerów**
- **Jawnie specyfikowane źródło drzewa**
- **Wyłącznie p2mp**
- **Mniejsza skalowalność O(N^2), N=#Zródeł/Ujść**
- **Periodyczne update'y**

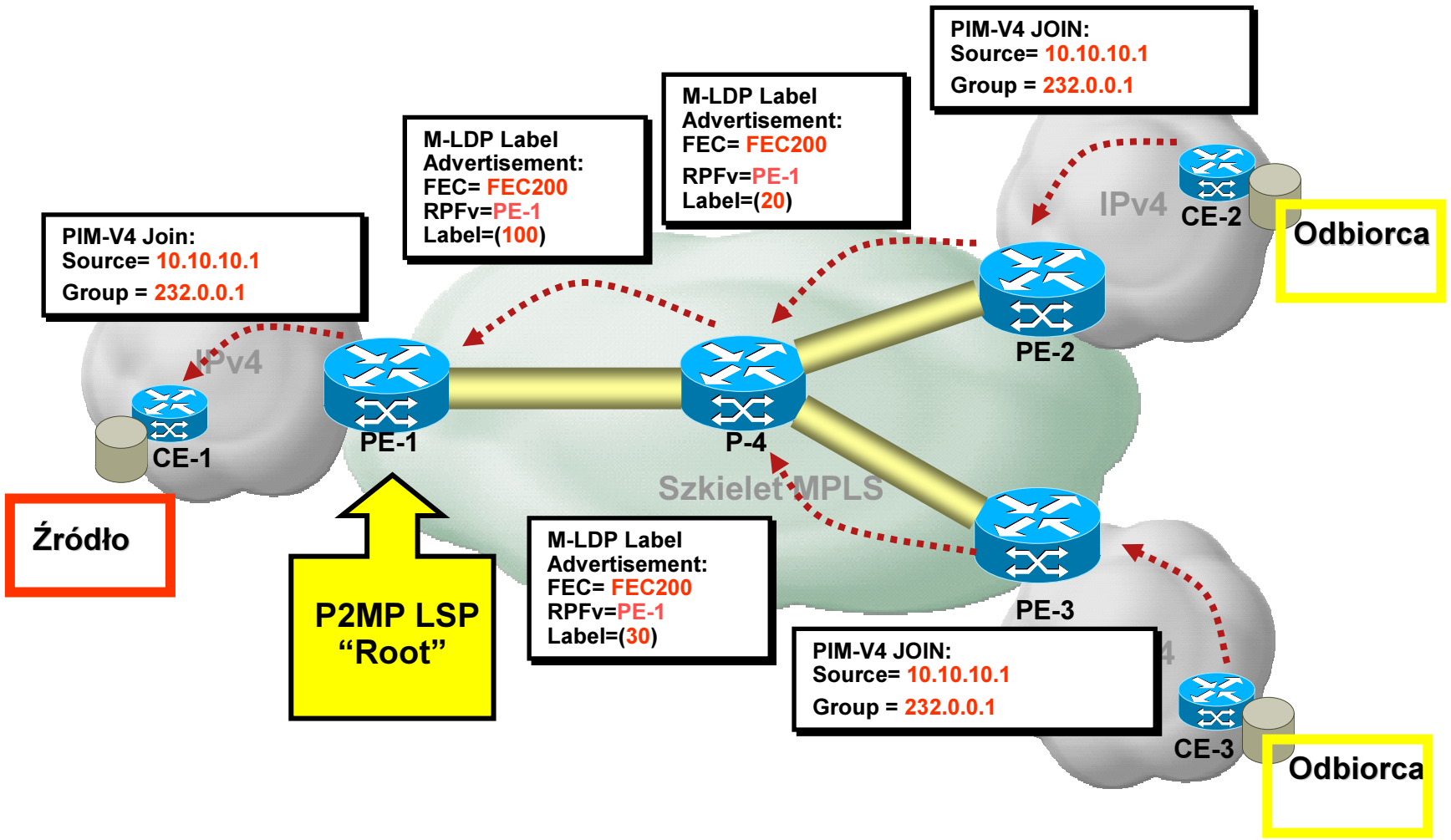
MVPN:
PIM vs P2MP-TE vs MLDP

MLDP

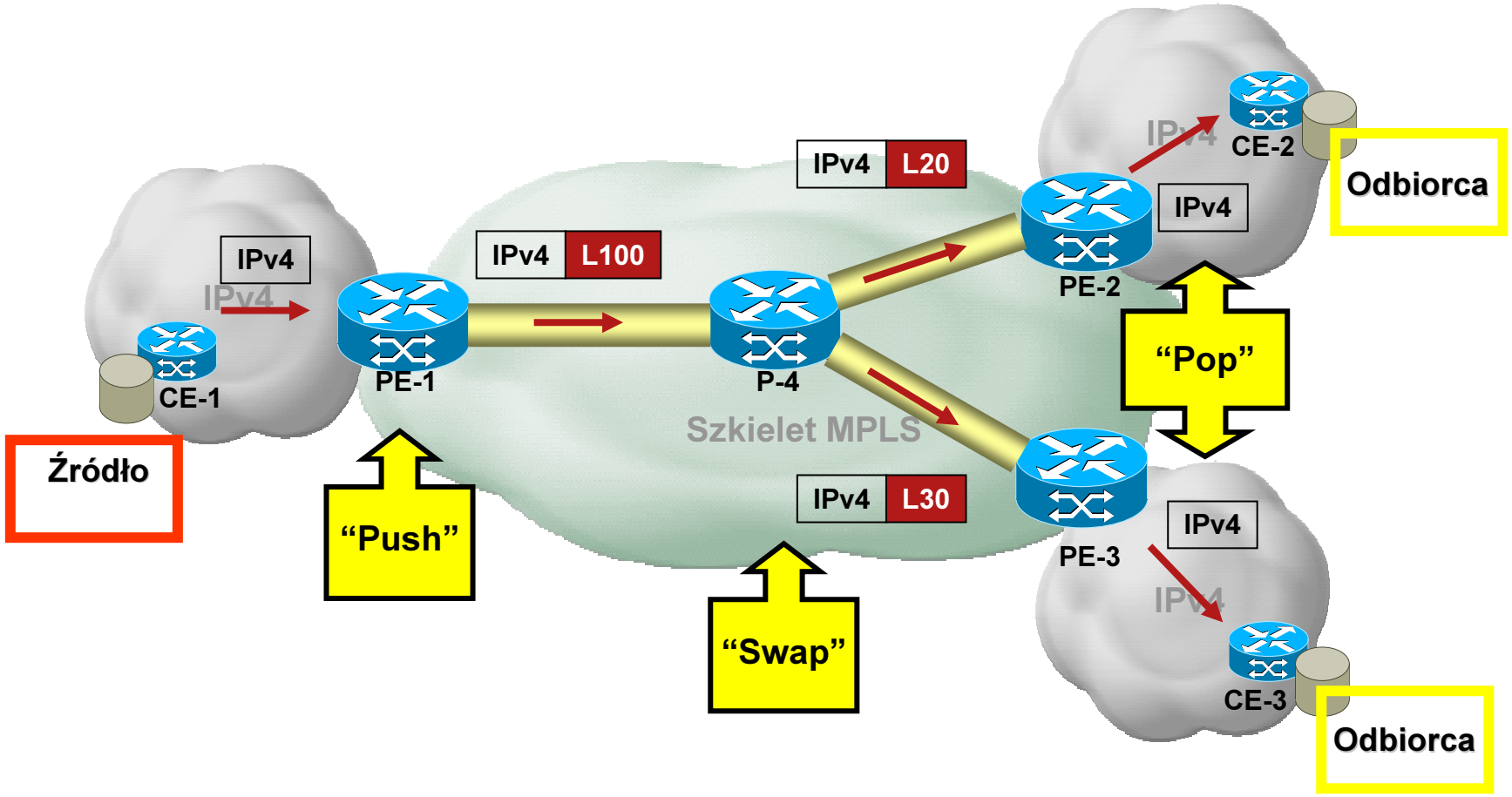
- **Brak periodycznych update'ow**
- **Konieczny upgrade większej ilości routerów**
- **Dynamiczne zestawianie drzewa**
- **„Aggregated Tree Model" – współdzielenie drzewa multicastowego przez kilka źródeł**
- **FRR**
- **p2mp i mp2mp**

MLDP

MLDP
Klasyczny SSM (IPv4 bez VPNów)



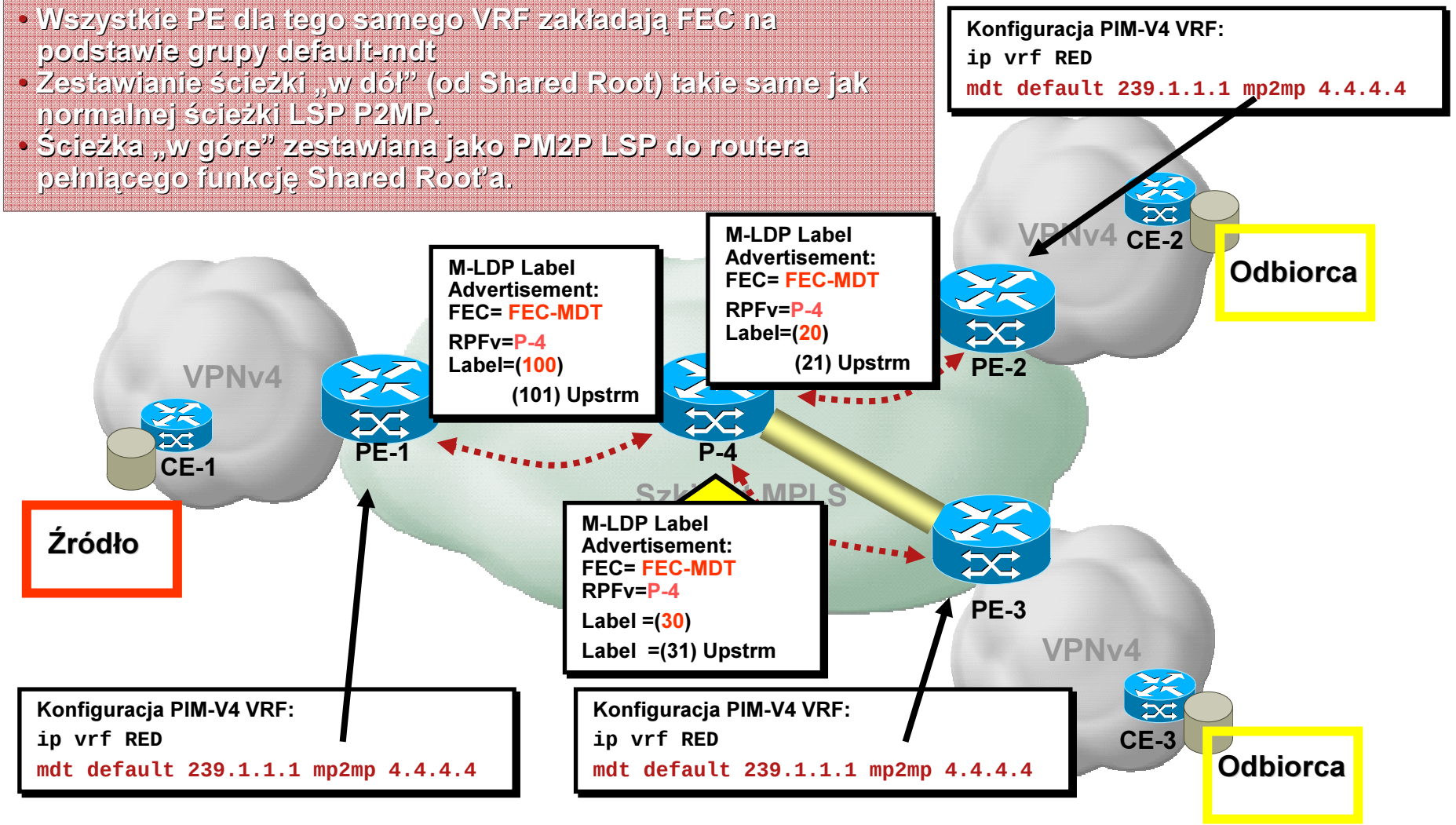
MLDP
Klasyczny SSM (IPv4 bez VPNów)



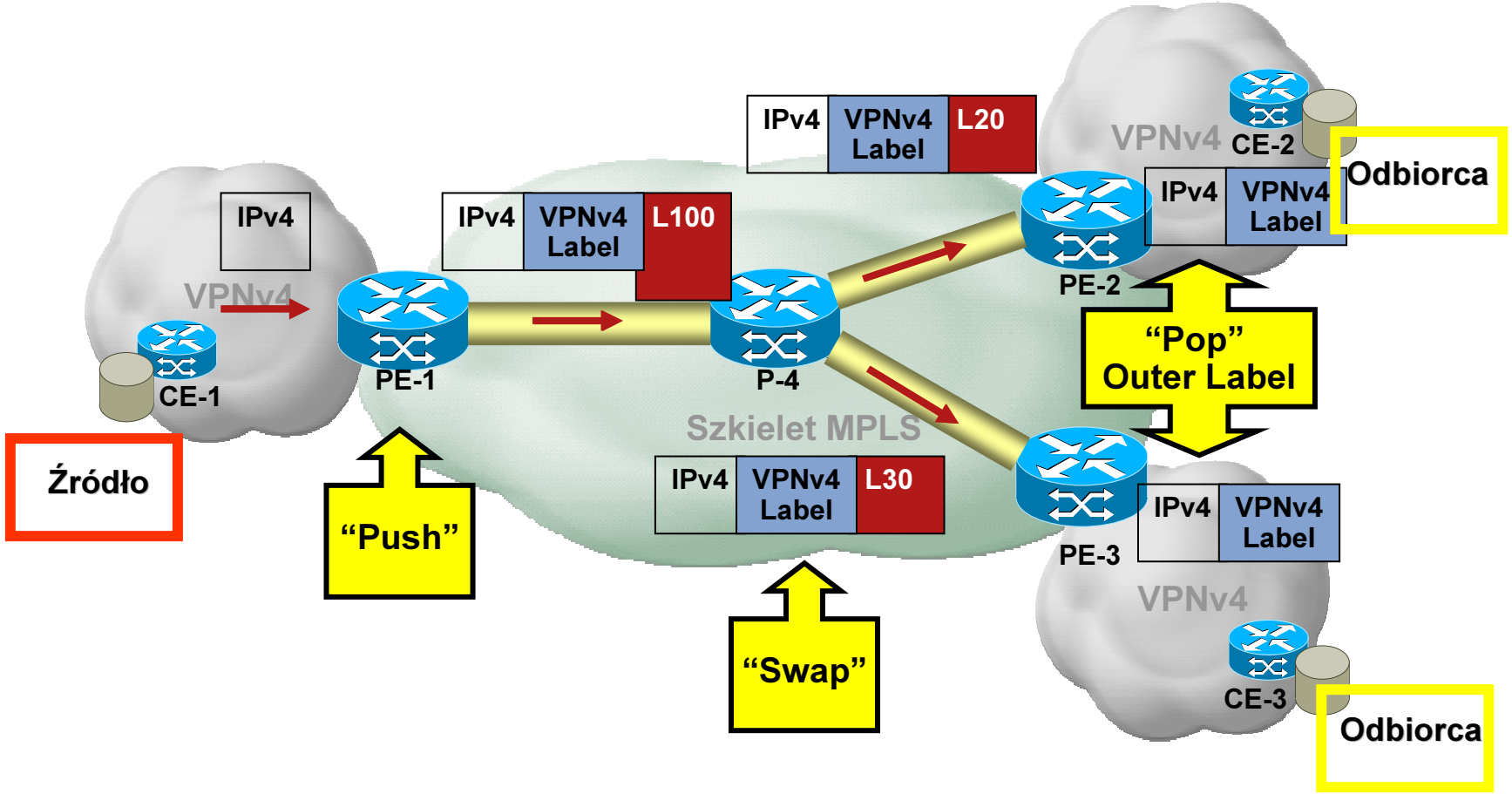
MLDP
MVPN (*Default-MDT*)

Zestawianie drzewa MP2MP

- Wszystkie PE dla tego samego VRF zakładają FEC na podstawie grupy default-mdt
- Zestawianie ścieżki „w dół” (od Shared Root) takie same jak normalnej ścieżki LSP P2MP.
- Ścieżka „w góre” zestawiana jako PM2P LSP do routera pełniącego funkcję Shared Root'a.



MLDP
MVPN (*Default-MDT*)

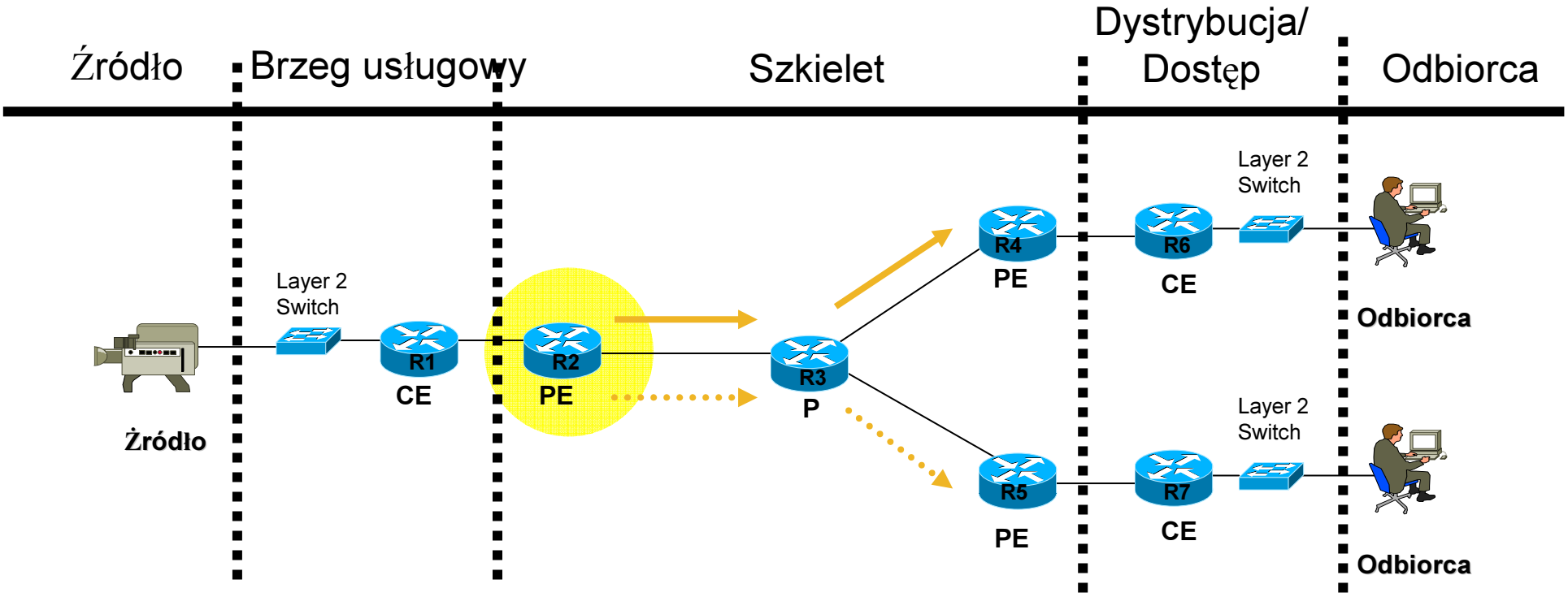


P2MP RSVP-TE

P2MP RSVP-TE

- Pamiętacie P2P RSVP-TE ? ;-)
  - –P2P LSP budowana od źródła do ujścia
- P2MP RSVP-TE
  - –Źródło buduje odpowiednia ilość P2P RSVP-TE LSP i sygnalizuje całość jako należącą do jednej P2MP LSP
  - –Routery P i PE wiedzą że syngalizowane uprzednio „pod-LSP” należą do jednego drzewa, więc je łączą
  - –Synalizowana jest tylko jedna etykieta do router poprzedzającego („upstream”) dla wszystkich „pod-LSP” danej P2MP LSP
- Wszystko inne dokładnie tak samo jak w przypadku P2P
  - –ERO, CSPF, link protection
  - –ale node protection zdecydowanie trudniejsza do osiągnięcia

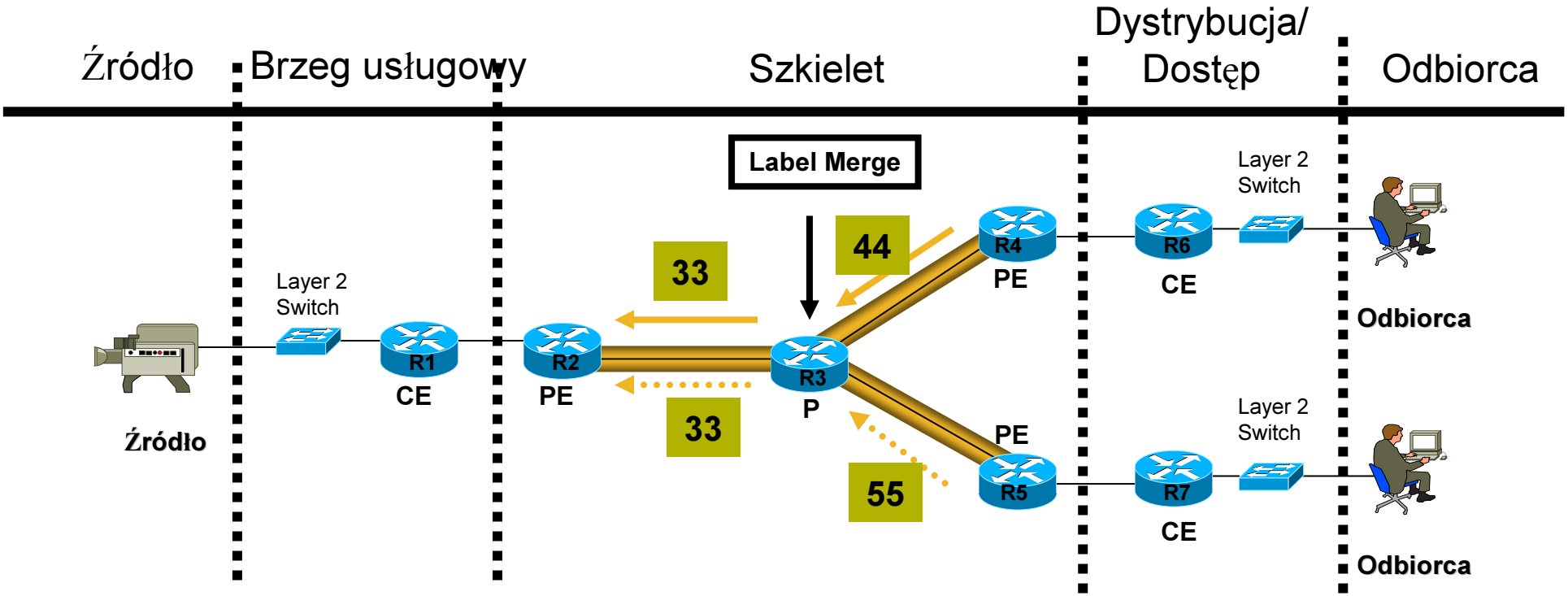
P2MP RSVP-TE – sygnalizacja w dół



Źródło wysyła jedną wiadomość PATH per odbiorca

PATH Message : ERO -> R2-R3-R4
PATH Message : ERO -> R2-R3-R5

P2MP RSVP-TE – sygnalizacja w górę



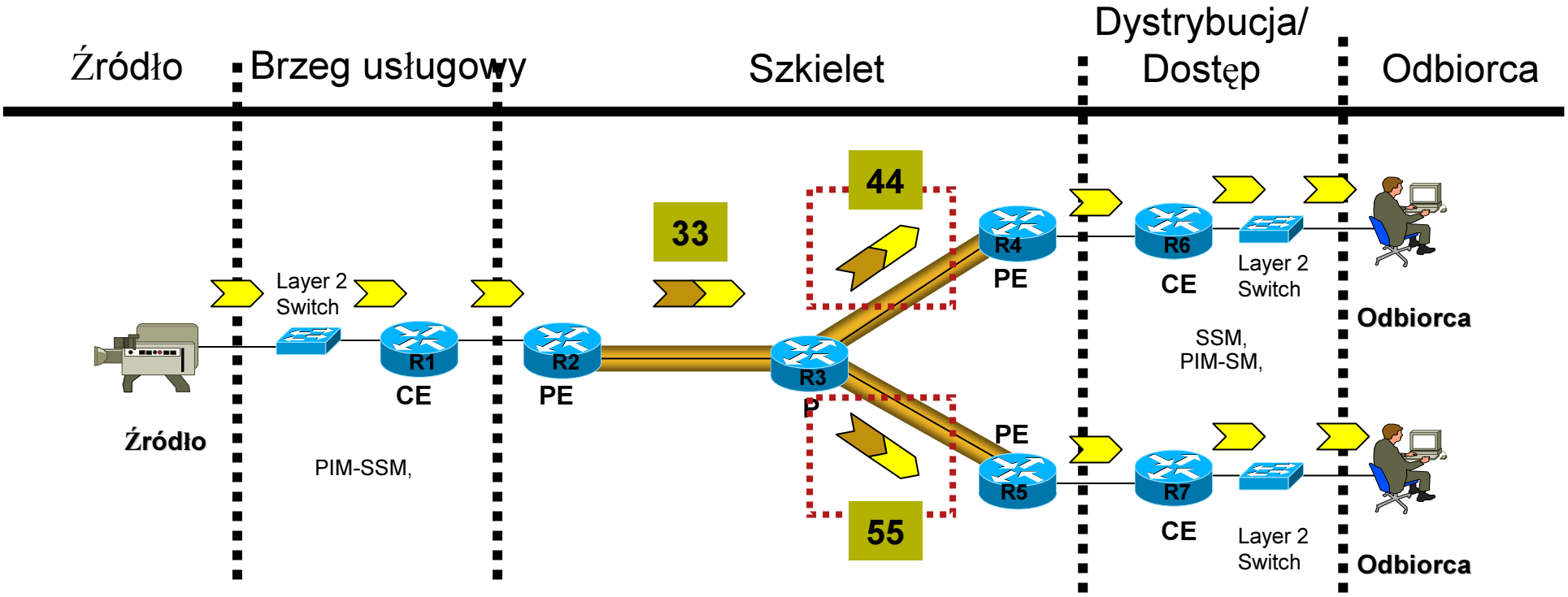
<u>Wiadomości RESV</u> są wysyłane przez routery typu „tailend” Zawierają etykiety i zarezerwowane pasmo na każdym z linków

Wiadomość RESV incjowana przez R4

Wiadomość RESV incjowana przez R5

55 Label Advertisement przenoszony w wiadomości RESV

P2MP RSVP-TE – przesyłanie ruchu



Brak PHP ! Etykieta potrzebna na routerach PE typu „tailend” do identyfikacji drzewa

Pakiet Multicast'owy

Pakiet zaetykietowany

BGP vs PIM

Problemy z PIM

- Konieczność utrzymywania dużej ilości drzew do przesyłania pakietów kontrolnych (obowiązkowy Default-MDT per każdy VRF)
- PIM to protokół utrzymujący „soft-state”
  - –Duży nakład spowodowany koniecznością retransmisji informacji o rzeczach które się wogóle nie zmieniły
  - –“oczywiście” nieskalowalny (#VPN * #grup/VPN)
- Dobry do przesyłania klasycznego multicast – a co z VPNami?

“BGP to the Rescue”

- PE nadal korzystają z PIM per VRF, ale:
  - Redystrybucja tras mulitcastowych odbywa sie do BGP
  - Brak sąsiedztw PE-PE PIM
  - Mechanizmy kontrolne PE-PE zunifikowane z mechanizmami kontrolnymi klasycznych unicast VPN
- Mechanizmy kontrolne PE-PE nie wymagają default-MDT:
  - Drzewa są tworzone tylko wtedy kiedy jest ruch do przesłania (skalowalnośc ;-)
  - Lepiej „pasuje” do drzew P2MP budowanych za pomoca P2MP RSVP-TE (czy potrzeba wówczas połączeń full-mesh wykorzystywanych przez PIM?)
- BGP = “hard state”
  - Bardzo wydajny jeśli niezbyt wiele się zmienia
  - Łatwa kontrola wymiany informacji
- Możliwość wykorzystania wszystkich mechanizmów kontroli wymiany informacji typowych dla BGP
- draft-ietf-l3vpn-2547bis-mcast-bgp-05

Autodiscovery

- Proces wykrywania wszystkich PE będących członkami danego MVPN
- Podobne do RFC4364, ale:
  - Nowa *address family* MCAST-VPN
  - Zawiera adres źródłowego PE
  - Przenosi informację o typie tuneli (instancji PMSI) używanych przez dane PE (np. PIM-SSM, MLDP itp)
  - Informuje o agreagacji VPNów w pojedyncze drzewo poprzez routery P
- Może byc wykorzystywany również do wykrywania zbioru PE zainteresowanych daną grupą multicastową w ramach VRF (czyli do tworzenia S-PMSI)
- Ale skoro wykorzystujemy PIM do budowy drzew w sieci operatora to może wykorzystać go również do autodiscovery? ;-)

Przenoszenie tras multicastowych z VRFów

- Niby łatwe, ale:
  - Jak rozwiązać problem zwiększonego opóźnieniem dla komunikatów typu Join/Prune (BGP ze swej natury nie zadziała tak szybko jak PIM)?
  - Co z wymaganą szybkością zmian stanu protokołów?
    - BGP jest fajne jak „jest spokojnie”
    - A co jak nie jest? ;-)
  - Co ze Sparse Mode?
  - Jak rozwiązać problem operacji w PIM które są operacjami „transakcyjnymi” np. Join. Użycie BGP w sposób dla niego nietypowy – zdarzenia są powodowane przez użytkowników końcowych a nie zmiany topologii!
  - Jak to będzie działać w rzeczywistym środowisku, gdzie BGP przenosi również inne informacje routing’owe?
- Gdzie przenoszenie tras mutlicastowych z VRF przez BGP sprawdza się najlepiej?
  - Multi-provider inter-AS (terminowanie tuneli na brzegu AS)
  - RSVP/TE P2MP w szkielecie
  - Przypadki w których ilość CE per VRF jest mała a PE wspiera agregacje komunikatów Join/Prune

BGP czy PIM?

- Jeśli odpowiedź byłaby oczywista to mielibyśmy jedno rozwiązanie ;-)
- Co na to operatorzy?:
  - –BGP wydaje się atrakcyjne, ze względu na wspólny *control plane* z unicast'owymi VPNami
  - –Wiele obecne działających sieci wykorzystuje PIM i nikt nie porzuci go z dnia na dzień
  - –PIM posiada ograniczenia związane ze skalowalnością, ale praktycznie, obecnie mało kto się na nie natknał
  - –NTT: "PIM is working just fine, no changes needed"
- Co na to eksperci?
  - –Ci od multicast'ów uważają rozwiązanie z BGP za ryzykowne
  - –Ci od BGP uważają dodawanie multicast'u do BGP za niepożądane

MVPN NG
Podsumowanie

- MPLS ma bogaty zbiór opcji do świadczenia usług typu mulitpoint
- Nie ma idealnego rozwiązania pasującego dla wszystkich przypadków
- Istnieje wiele możliwości deployment'u MVPN:
  - Tworzenie drzew w sieci operatora za pomocą PIM, RSVP-TE lub MLDP
  - Autodiscovery PE biorących udział w MVPN za pomocą PIM lub BGP
  - Wymiana tras multicastowych z VRF za pomocą PIM lub BGP
- Wybór sposród tych opcji nie jest prosty:
  - Wymaga dokładnego zrozumienia potrzeb, topologii i wymagań klienta
  - Best Practises powstają wraz z doświadczeniem z wdrożeń