



Wysoka dostępność w sieciach operatorskich

Łukasz Bromirski
lbromirski@cisco.com



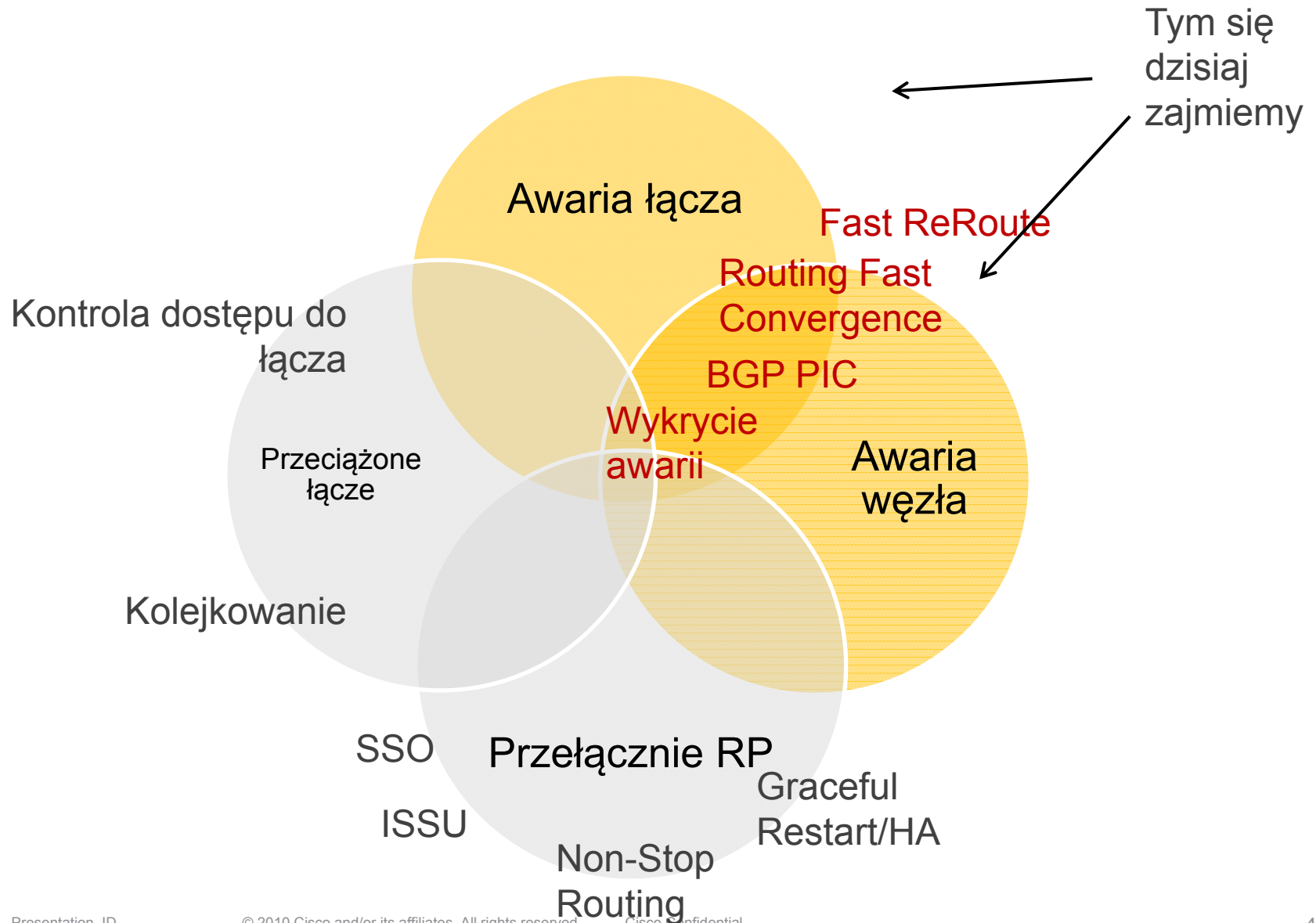
Agenda

- Wysoka dostępność
...szybka konwergencja
- Mechanizmy L1
- Mechanizmy L2
- Mechanizmy L3

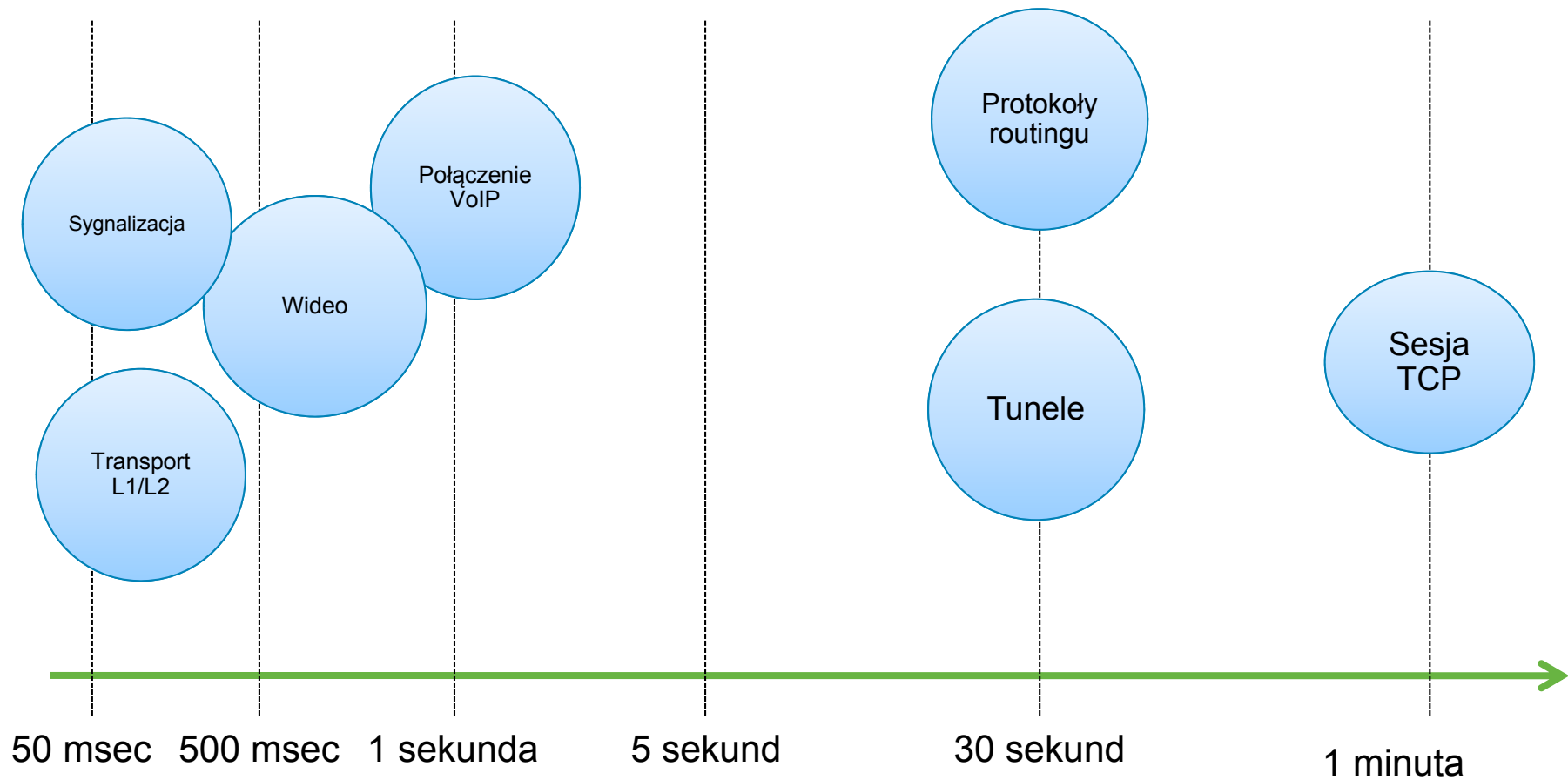


Wysoka dostępność a szybka konwergencja

Problemy i sposoby ich rozwiązywania



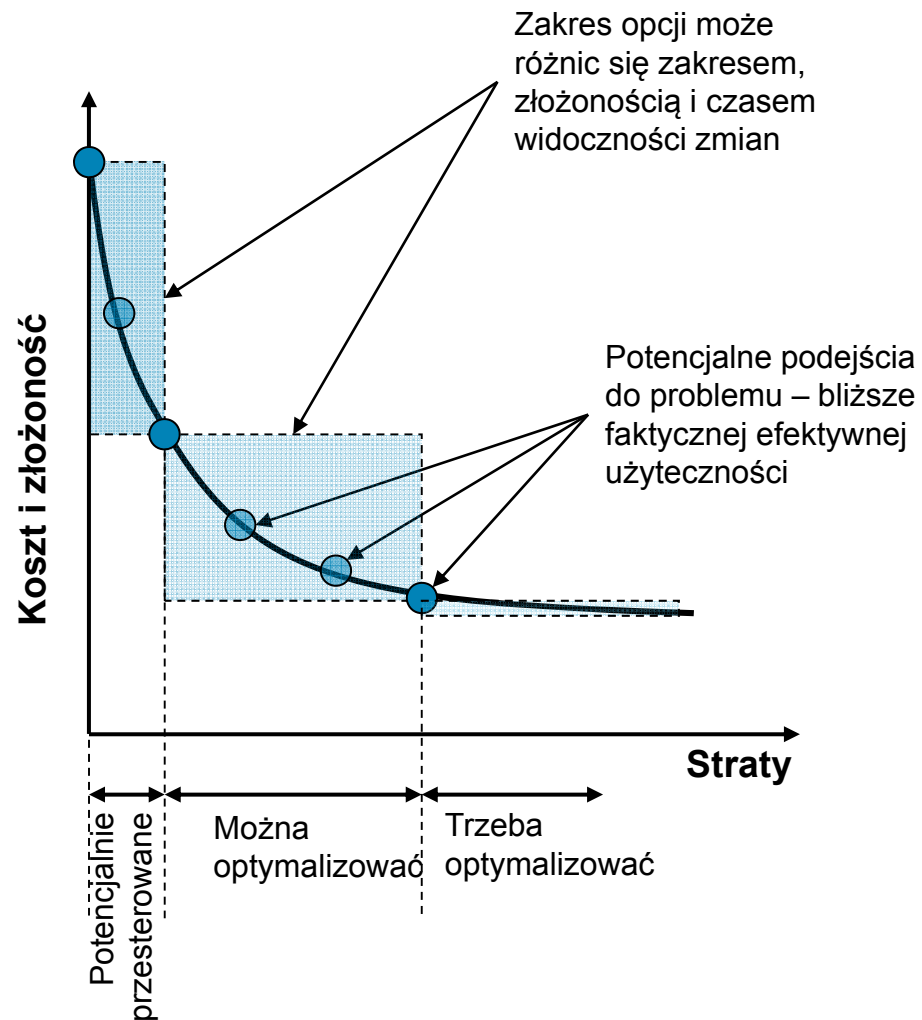
Odporność usług na przerwę w pracy sieci



Projektowanie sieci

- Szybka konwergencja to trochę więcej niż parę poleceń
- Parę warstw do rozważenia **i ich wzajemnego działania**
 - Warstwa 1 i 2 – wykrywanie awarii i zmian w topologii fizycznej
 - Warstwa 3 – zachowanie protokołów, interakcje pomiędzy protokołami
 - Warstwy 4-7 – zachowanie aplikacji i usług
- Wszelkie aspekty związane z budową platform sieciowych – czasem z dokładnością do możliwości wykrycia awarii w L1, prędkości budowania (programowania) tablic FIB i MFIB, itp. itd.

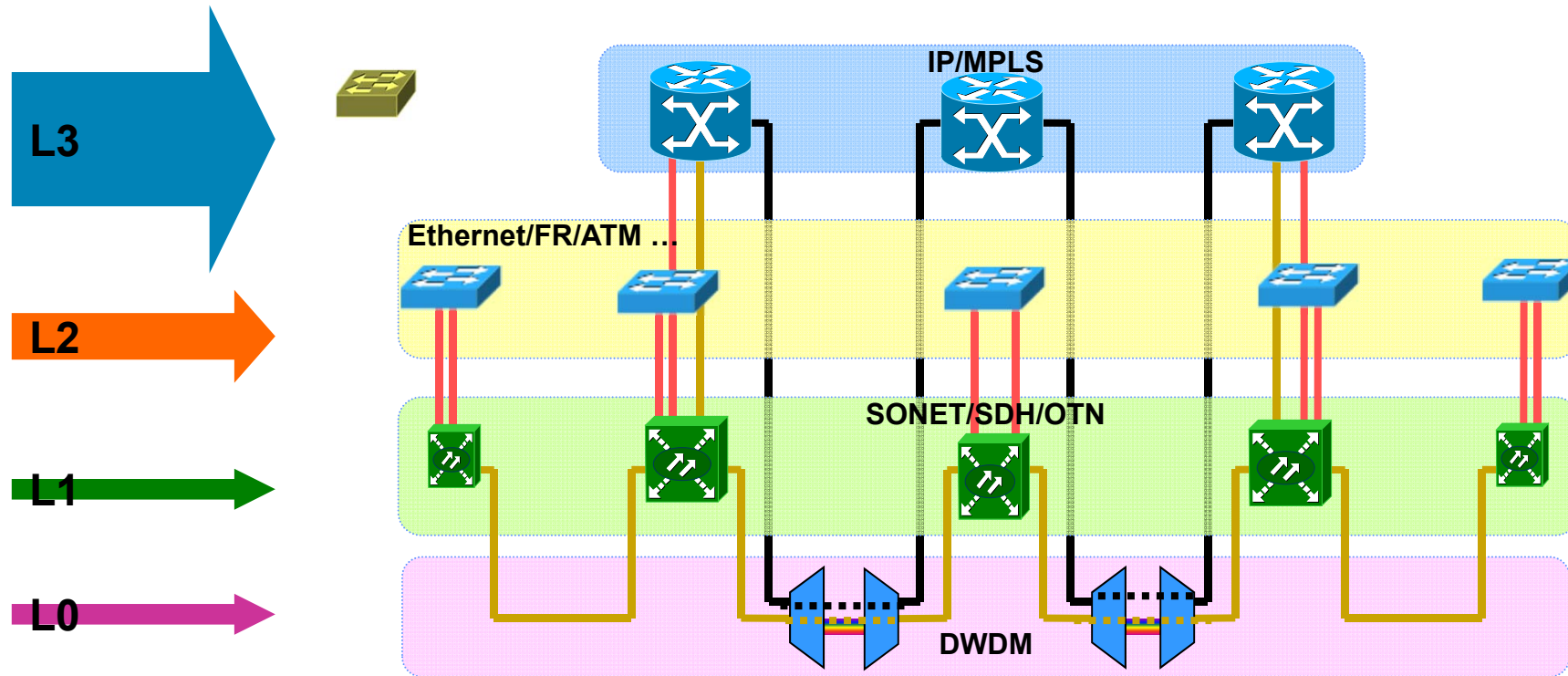
„Działa? Nie poprawiać!”





Mechanizmy L1

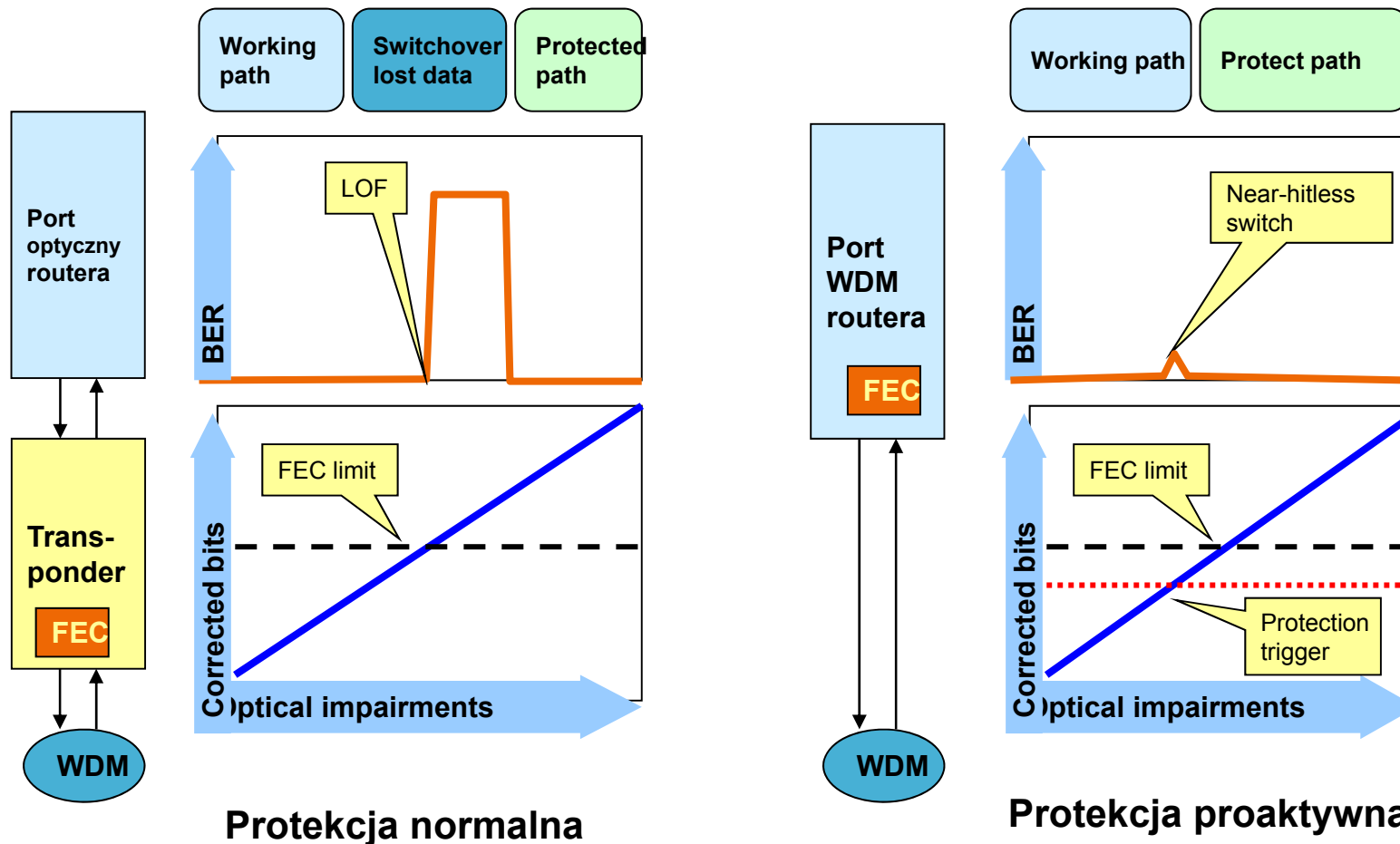
Opcje transportu IP w sieci transportowej



- IP -> Layer 2: Ethernet, EoDWDM, Frame Relay, ATM ...
- IP -> Layer 1: SONET/SDH (POS), xWDM (Transponder, EoDWDM)
- IP -> Layer 0: G.709 (IPoDWDM)

Wykrywanie awarii w L0 – IPoDWDM

- Integracja optyczna i IP wprowadza możliwość zidentyfikowania łącza gorszej jakości i automatycznego włączenia ochrony (i rekonwergencji protokołów L3) – w wielu wypadkach oznacza to że przeniesienie ruchu odbędzie się bez straty ruchu



Wykrywanie awarii w L1 – SONET/SDH

- Alarmy

Reakcja natychmiastowa, możliwa do kontrolowania poleceniem

```
pos delay-trigger
```

Należy oczywiście wziąć pod uwagę protekcję SONET/SDH przy konfiguracji wyzwolenia położenia logicznego interfejsu

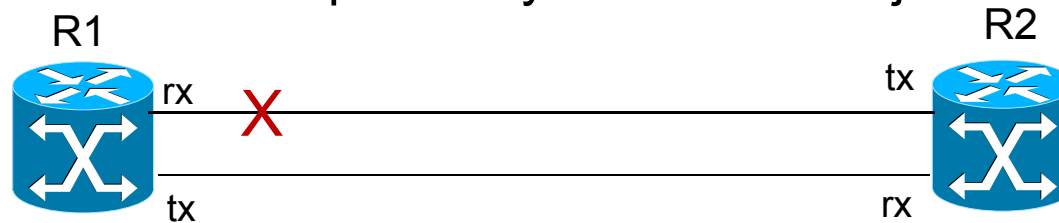
Domyślnie polecenie shutdown nie generuje alarmu – należy to wprost włączyć poleceniem

```
pos ais-shut
```

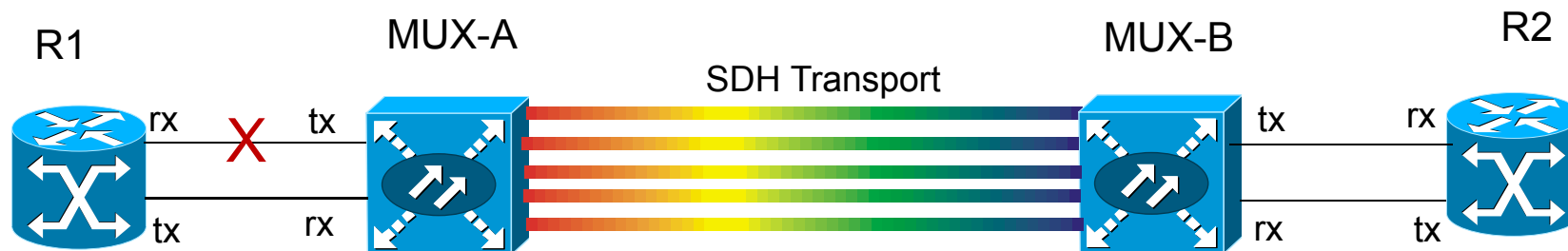
na interfejsie fizycznym

Wykrywanie awarii w L1 – światłowód/GE

- Autonegocjacja (tak jak opisano to w IEEE 802.3z/802.3ae) może sygnalizować lokalne problemy stronie zdalnej



- W przypadku połączeń Ethernet ponad SONET/SDH problemem może być przeniesienie sygnalizacji – zwykle po prostu to nie działa





Mechanizmy L2

carrier-delay
IP dampening

Wykrywanie awarii w L2 – transport

- Wykorzystanie konstrukcji protokołów L2 i różnego rodzaju odpowiedników pakietów typu „Hello”
 - pakiety keepalive w PPP i HDLC
 - LMI we Frame-Relay
 - OAM w ATMie
 - OAM w Ethernet
- Mechanizmy te nie dają możliwości osiągnięcia konwergencji na poziomie poniżej sekundy
- Obsługa pakietów keepalive w różnych protokołach (do minimalnej sekundy) nie jest zalecana – większość producentów nie optymalizuje platform do ich priorytetowej obsługi co może prowadzić do fałszywych alarmów

Carrier-Delay

- 1 i 2 warstwa przekazują sygnał o awarii łącza (LINK i/lub LINEPROTO)
- Domyślnie większość platform stosuje dodatkowy licznik zanim zareaguje – Cisco IOS domyślnie od zera (Catalyst) do 2 sekund

```
interface ...  
  carrier-delay msec 0
```

- Oczywiście czas należy możliwie zmniejszyć
- Aby zapobiec niepotrzebnemu ,klapaniu' łącz i wpływowi tego na routing, stosuje się IP dampening

```
interface ...  
  dampening
```

Carrier-Delay asymetrycznie

- Zadeklarowanie interfejsu w stan „up” można opóźnić, aby umożliwić pierwszym zapytaniom ARP zakończyć zbieranie informacji o sąsiedztwach

```
interface ...  
  carrier-delay up 2000 msec
```

- Wsparcie dla Cisco IOS - od 12.0(32)SY2, 12.2SRD, XR3.4.0
- Niektóre sterowniki mają wbudowany czas „up”
 - POS: z reguły 10 sekund
 - 7600 ES20/40 WAN: 4 sekundy



Mechanizmy L2

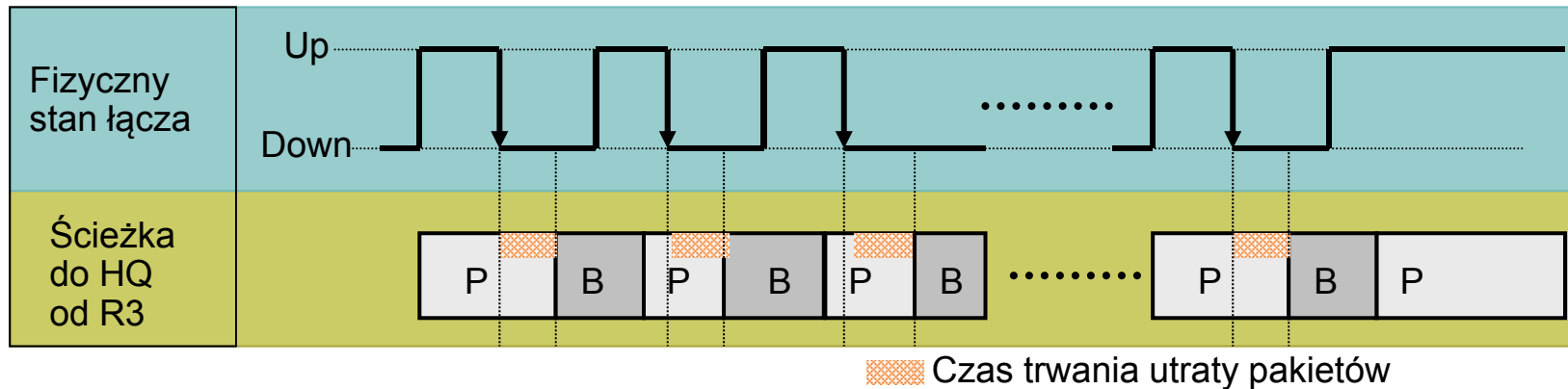
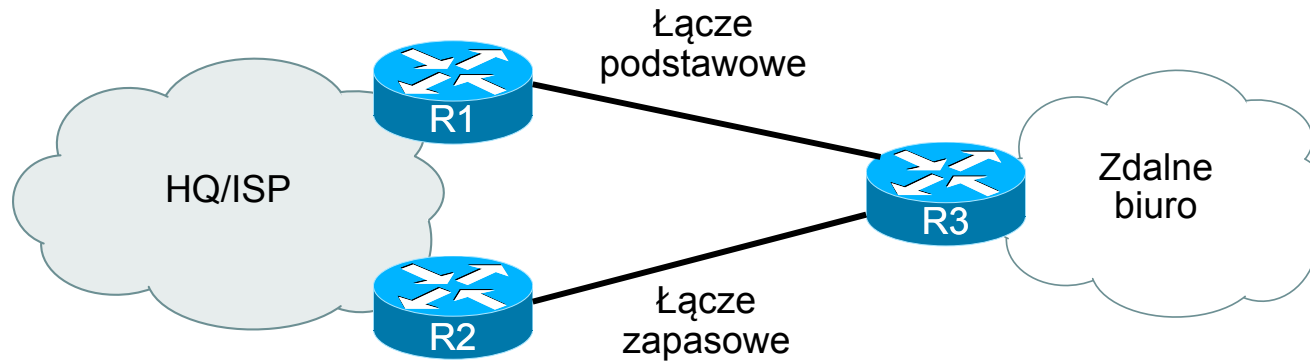
carrier-delay
IP dampening

IP Event Dampening

- Zapobiega ciągłym zmianom informacji z protokołów routingu
- Wspiera wszystkie protokoły routingu, oraz:
Routing statyczny, RIP, EIGRP, OSPF, IS-IS, BGP
HSRP i routing CLNS
- Dostępny od 12.0(22)S, 12.2(13)T

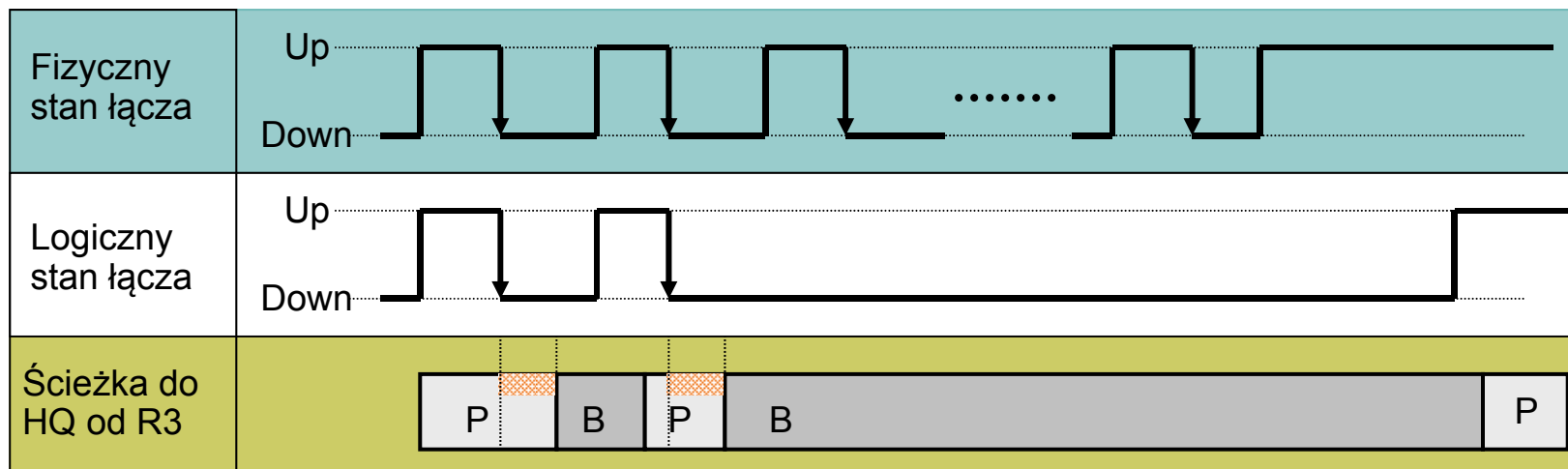
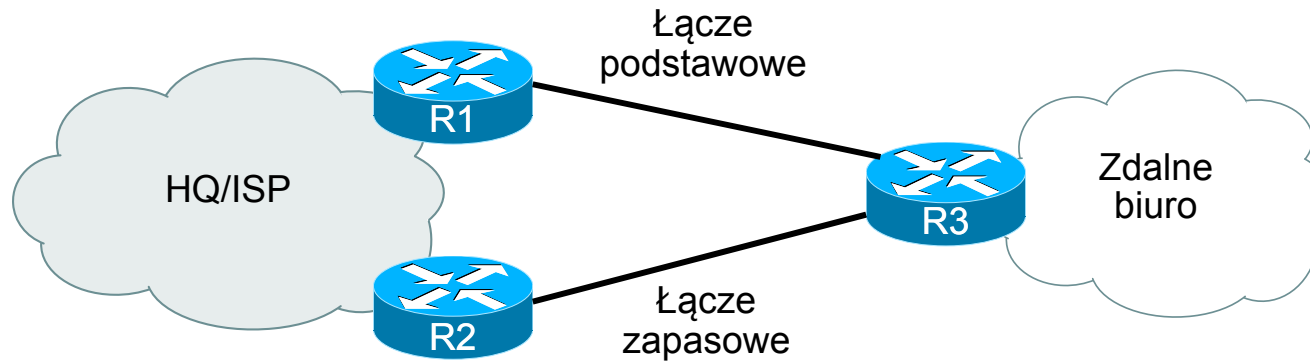
IP Event Dampening

Bez skonfigurowanego



IP Event Dampening

Po skonfigurowaniu



Czas trwania utraty pakietów



Mechanizmy L3

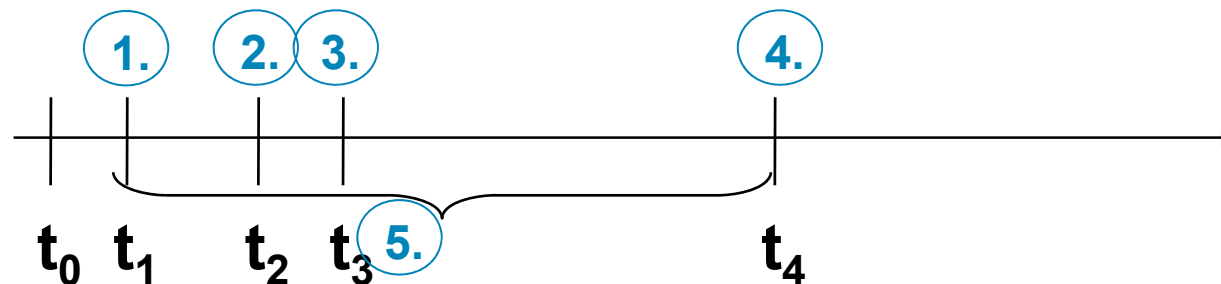
Tuning protokołów routingu IGP

Tuning protokołu BGP

BGP PIC Edge i Core

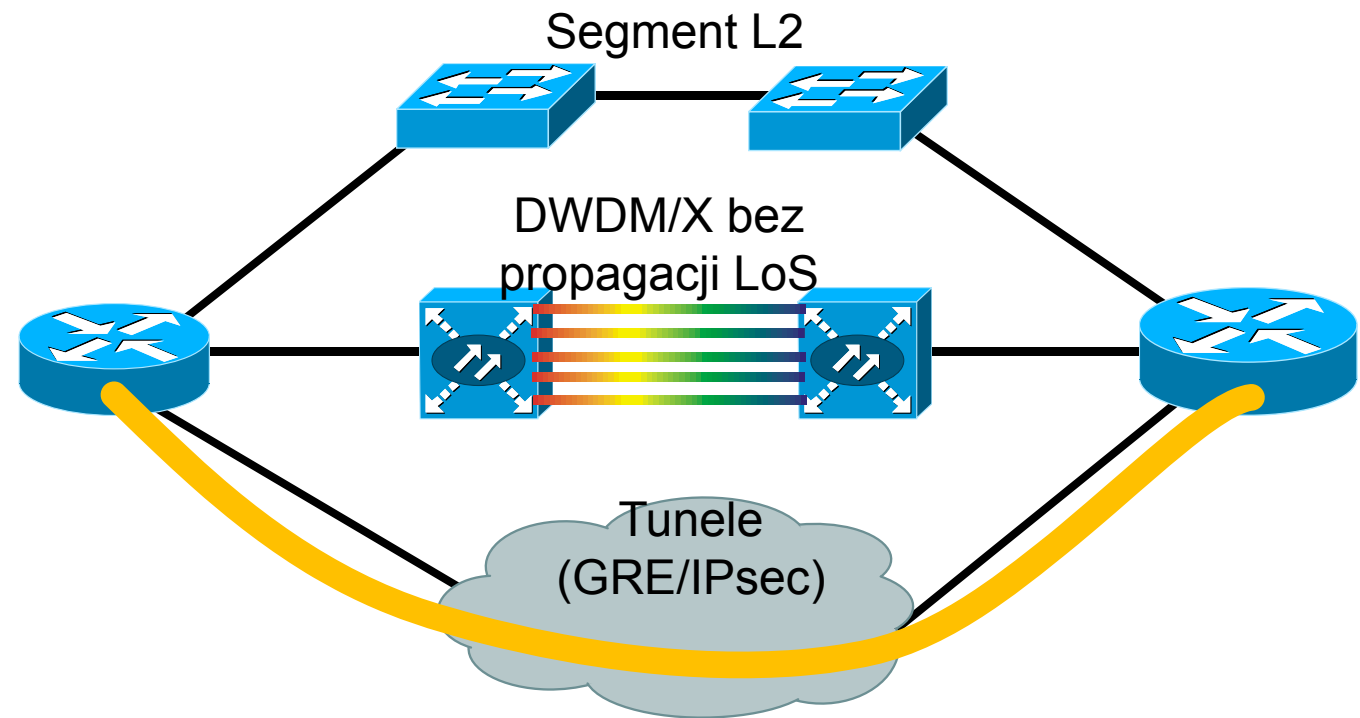
Komponenty konwergencji w L3

1. Wykrycie awarii
2. Propagacja informacji o awarii (flooding, etc.)
3. Przeliczenie topologii
4. Uaktualnienie tablic przechowujących informację o routingu (RIB & FIB)
5. Wydajność warstwy kontrolnej węzła sieciowego



Wykrycie problemu w L3

- Nie wszystkie problemy da się wykryć za pomocą L2 – czasami sygnalizację zdalnej awarii musi zapewnić L3



Warstwa routingu

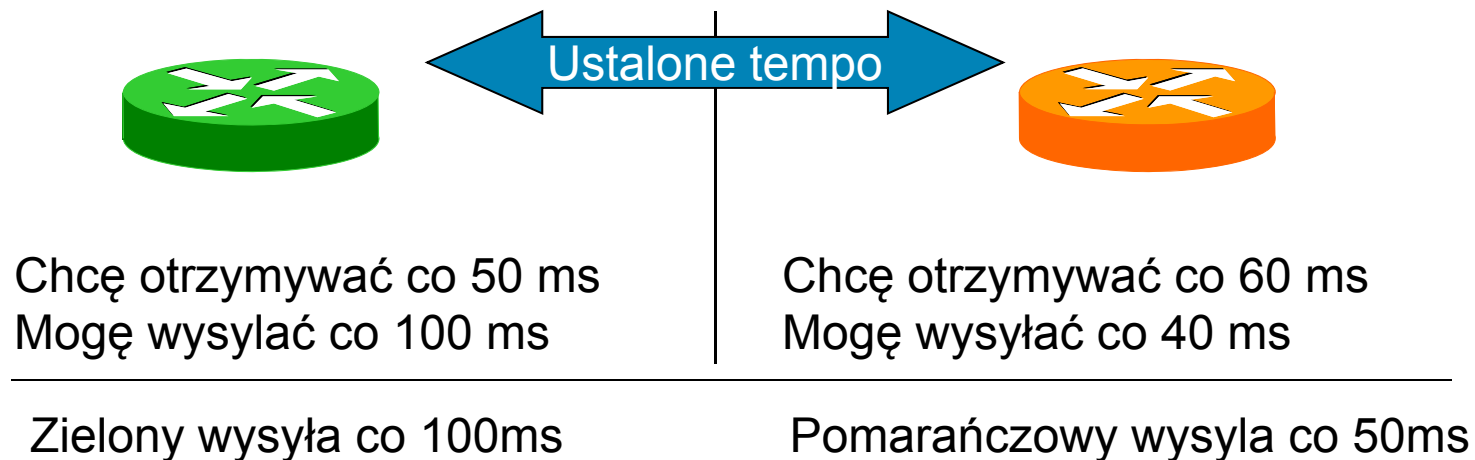
- Wszystkie protokoły IGP (EIGRP, OSPF i ISIS) używają pakietów HELLO aby utrzymać sąsiedztwa i sprawdzić osiągalność sąsiadów
- Liczniki hello/hold można odpowiednio dostosować w dół aby osiągnąć czasy wykrycia awarii na poziomie poniżej sekundy, jednak:
 - Nie skaluje się to dobrze dla dużych ilości sesji (wiele setek, tysiące)
 - Duże obciążenie CPU może spowodować niechciane próby rekonwergencji sieci wokół problemu który w ogóle nie wystąpił
 - Można jednak to robić – w sieciach na małą skalę
- Lepszym rozwiązaniem jest coś uniwersalnego

BFD (Bi-directional Forwarding Detection)

- Lekki, prosty protokół z niskim narzutem
- BFD może być przetwarzany w sposób rozproszony (np. na kartach liniowych routerów GSR, CRS czy ASR9k) a zatem daje się go w przewidywalny sposób skalować dla większej ilości sesji
- Dowolna „zainteresowana” aplikacja taka jak protokół routingu (OSPF, BGP, EIGRP) czy mechanizm (HSRP) może „zarejestrować” chęć bycia poinformowaną przez BFD o utracie osiągalności ścieżki

Konfiguracja BFD i negocjacja pracy

- Sąsiedzi mogą stale renegecować parametry pracy
- Wolniejszy system będzie dyktował parametry zestawienia sesji



```
interface <name>  
  bfd interval <msec> min_rx <msec> multiplier <n>
```

Propagacja zdarzeń

OSPF

- Pierwsze opóźnienie związane z generowaniem LSA domyślnie ustawione jest na 500ms (dotyczy tylko Router/Network)
- Kolejny opóźniający timer – czas pomiędzy rozgłoszeniem konkretnego LSA – domyślnie 5 sekund
- Odbiór LSA – w trakcie odbioru LSA router zakłada domyślnie, że nowe instancje LSA mogą spływać co MinLSArrival – domyślnie co 1 sekundę

```
timers throttle lsa all <lsa-start> <lsa-hold> <lsa-max>  
timers lsa arrival <timer>
```

wszystkie wartości czasu w ms

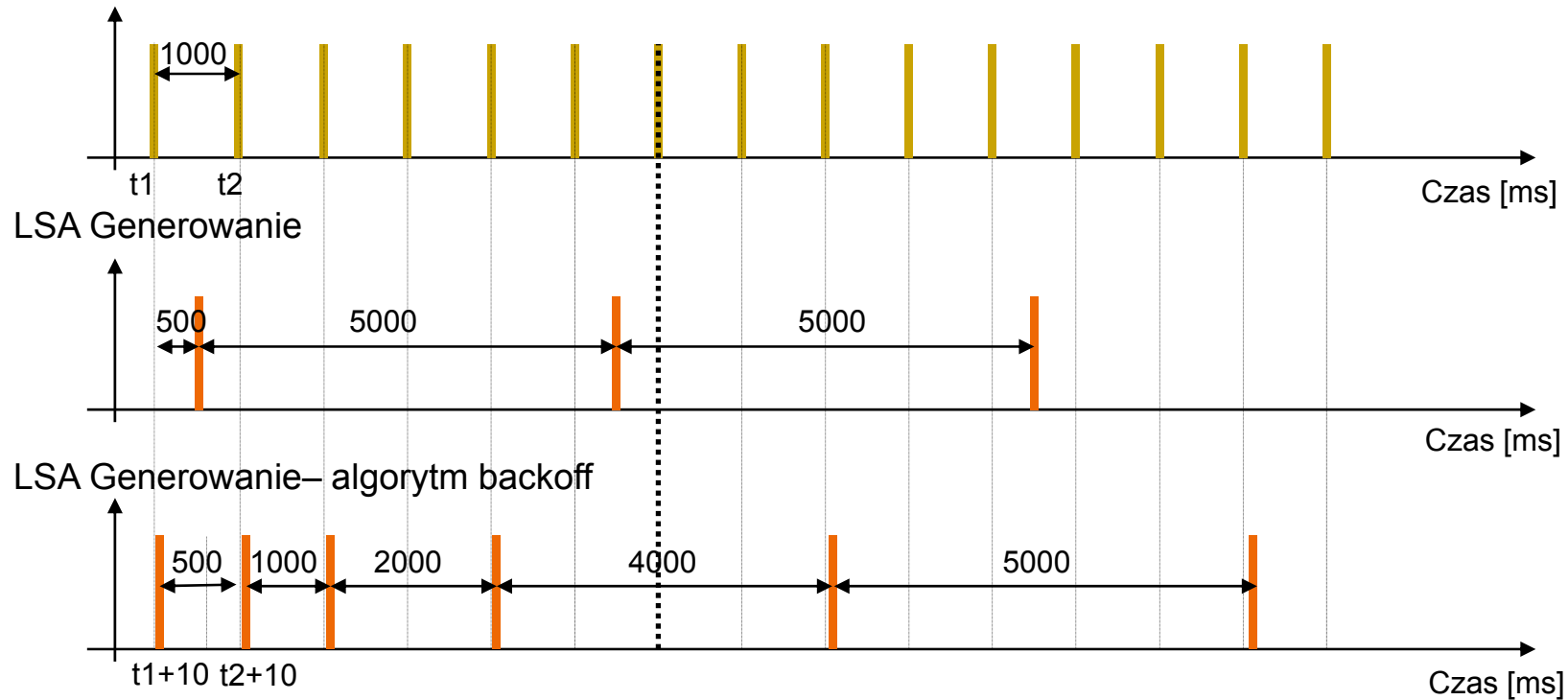
Propagacja zdarzeń

OSPF

```
timers throttle lsa all 10 500 5000
```

poprzednie generowanie LSA w t_0 ($t_1 - t_0$) > 5000 ms

Zdarzenia powodujące generowanie LSA



Propagacja zdarzeń

OSPF

- W domyślnej konfiguracji każdy z węzłów sieciowych może dodać do 33ms do momentu wysłania zdarzenia
- Zmiana:

```
timers pacing flood <timer>  
timers pacing retransmission <timer>
```

Propagacja zdarzeń

OSPF

- Odłożenie w czasie przeliczenia SPF chroni router przed zbyt dużym obciążeniem, ale będzie miało negatywny wpływ na konwergencję
- Zmiana

```
timers throttle spf <spf-start> <spf-hold> <spf-max>
```

Priorytetyzacja prefiksów

OSPF i IS-IS

- Sieć posiada zwykle zbiór prefiksów ważniejszych – w szczególności adresów interfejsów loopback (używane jako router-id, używane do nawiązania sesji BGP)
- Przy konwergencji sieci w której znajdują się setki tysięcy prefiksów, czas programowania wpisów może mieć istotne znaczenie dla szybkiej konwergencji

Priorytetyzacja prefiksów

Konfiguracja

- Priorytety dla prefiksów

Krytyczny, Wysoki, Średni, Niski

/32 dla IPv4 i /128 dla IPv6 automatycznie trafiają do Średniego

Pozostałe prefiksy domyślnie trafiają do Niskiego priorytetu

- Dopasowanie

poleceniem `spf prefix-priority`

```
interface GigabitEthernet0/1 ! do bramki VoIP
 ip router isis
 isis tag 17
router isis
 ip route priority high tag 17
```



Mechanizmy L3

Tuning protokołów routingu IGP
Tuning protokołu BGP
BGP PIC Edge i Core

Co można poprawić w BGP?

- Skaner BGP
- ATF/NHT – Address Tracking Filter/Next Hop Tracking
- FSD – Fast Session Deactivation
- MRAI – Minimal Route Advertisement Interval
- TCP PMTUD/SACK

Skaner BGP

- Skaner (domyślnie) co 60 sekund wykonuje pełne przejście tablicy BGP

`bgp scan-time x`

- Co 15 sekund działa skaner importu

...importuje prefiksy VPNv4 do VRFów

`bgp scan-time import x`

- Pełne przejście wykonuje między innymi:

sprawdzenie osiągalności routerów next-hop

sprawdzenie poprawności wyboru najlepszej trasy

zaktualizowanie tablicy po zmianach w redystrybucji i wydaniu poleceń network

sprawdzenie rozgłoszeń warunkowych

obsługę route dampening

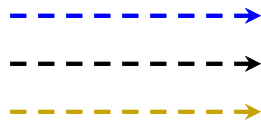
wyczyszczenie bazy BGP

ATF / NHT

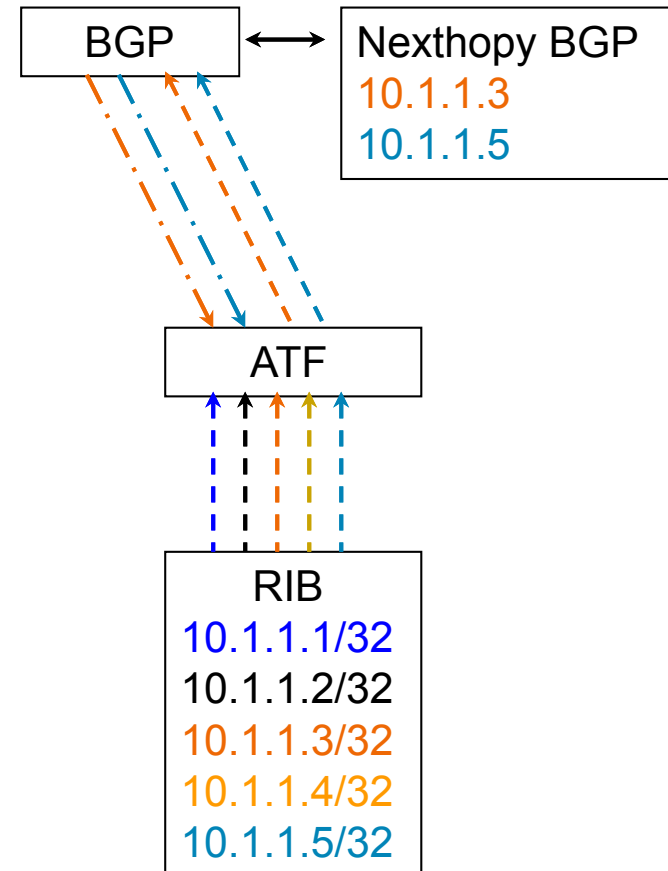
- BGP rejestruje w ATF prefiksy 10.1.1.3 i 10.1.1.5



- ATF nie informuje BGP o zmianach dla pozostałych prefiksów – np. 10.1.1.11/32, 10.1.1.2/32 i 10.1.1.4/32



- ATF informuje BGP o zmianach dla prefiksów zarejestrowanych



NHT

- Mechanizm BGP Next Hop Tracking dba automatycznie o rejestrowanie wszystkich adresów next-hop w ATF

włączony domyślnie (od 12.0(29)S i 12.3(14)T):

```
[no] bgp nexthop trigger enable
```

lista zarejestrowanych adresów:

```
show ip bgp attr nexthop
```

- Informacja z ATF uruchomi 'lekką' wersję BGP skanera:

wyliczenie najlepszych tras

...pozostałe operacje będą czekać na normalny cykl skanera, skaner nie weryfikuje już jednak osiągalności routerów next-hop oraz najlepszych tras

NHT

- Domyślnie BGP czeka 5 sekund po otrzymaniu informacji z ATF o zmianie dla prefiksu
- Do zmniejszenia wpływu dużej ilości zmian sygnalizowanych przez ATF, używany jest dampening

```
bgp nexthop trigger delay <0-100>
```

```
show ip bgp internal
```

wyświetla informacje kiedy ostatnio odbył się proces NHT i kiedy odbędzie się następny

Fast Session Deactivation

- Zarejestrowanie adresu next-hop dla prefiksów przez BGP w ATF pozwala bardzo szybko podjąć decyzję o osiągalności partnerów BGP
- Po utracie trasy do partnera sesji BGP, natychmiast można zdezaktywować sesję BGP
 - BGP nie czeka na upływanie czasu wynikającego z licznika hold
- Przydatne dla sesji eBGP
- **Bardzo niebezpieczne dla sesji iBGP**
 - IGP może nie mieć trasy do sąsiada przez ułamek sekundy...
 - FSD natychmiast rozłączy sesje...
- Domyślnie wyłączone
 - `neighbor x.x.x.x fall-over`
 - `neighbor x.x.x.x fall-over bfd ! FSD z BFD`

“MRAI [...] określa minimalny okres czasu który musi minąć pomiędzy rozgłoszeniem i/lub wycofaniem trasy do konkretnego prefiksu. Mechanizm działa z dokładnością do prefiksu, ale wartość `MinRouteAdvertisementIntervalTimer` ustalana jest dla sąsiada BGP.”

RFC 4271

Sekcja 9.2.1.1

MRAI - podstawy

- Liczniki MRAI utrzymywane są per sąsiad
 - iBGP – domyślnie 5 sekund
 - eBGP – domyślnie 30 sekund
 - `neighbor x.x.x.x advertisement-interval <0-600>`
- Zalety
 - pozwała uprościć/usystematyzować wymianę rozgłoszeń
- Wady
 - może spowolnić konwergencje
 - wartości określone w standardzie są bardzo konserwatywne i nie pozwalają zapewnić szybkiej konwergencji

MRAI – co można zmienić i jak?

- Zmiany w osiągalności prefiksów w internecie oznaczają falę zmian

Jeden „klepiący” prefiks może spowolnić znacznie konwergencje dla pozostałych prefiksów (np. nowych lub zmienianych)

W internecie mamy do czynienia z 2-3 zmianami na sekundę (zmiana w stosunku do 2006 roku – z 1-2 zmian):
na podstawie pracy Geoff Hustona:

<http://www.potaroo.net/presentations/2006-11-03-caida-wide.pdf>

- Dla połączeń iBGP i połączeń eBGP na styku PE<>CE sensowne wydaje się:

```
neighbor x.x.x.x advertisement-interval 0
```



Mechanizmy L3

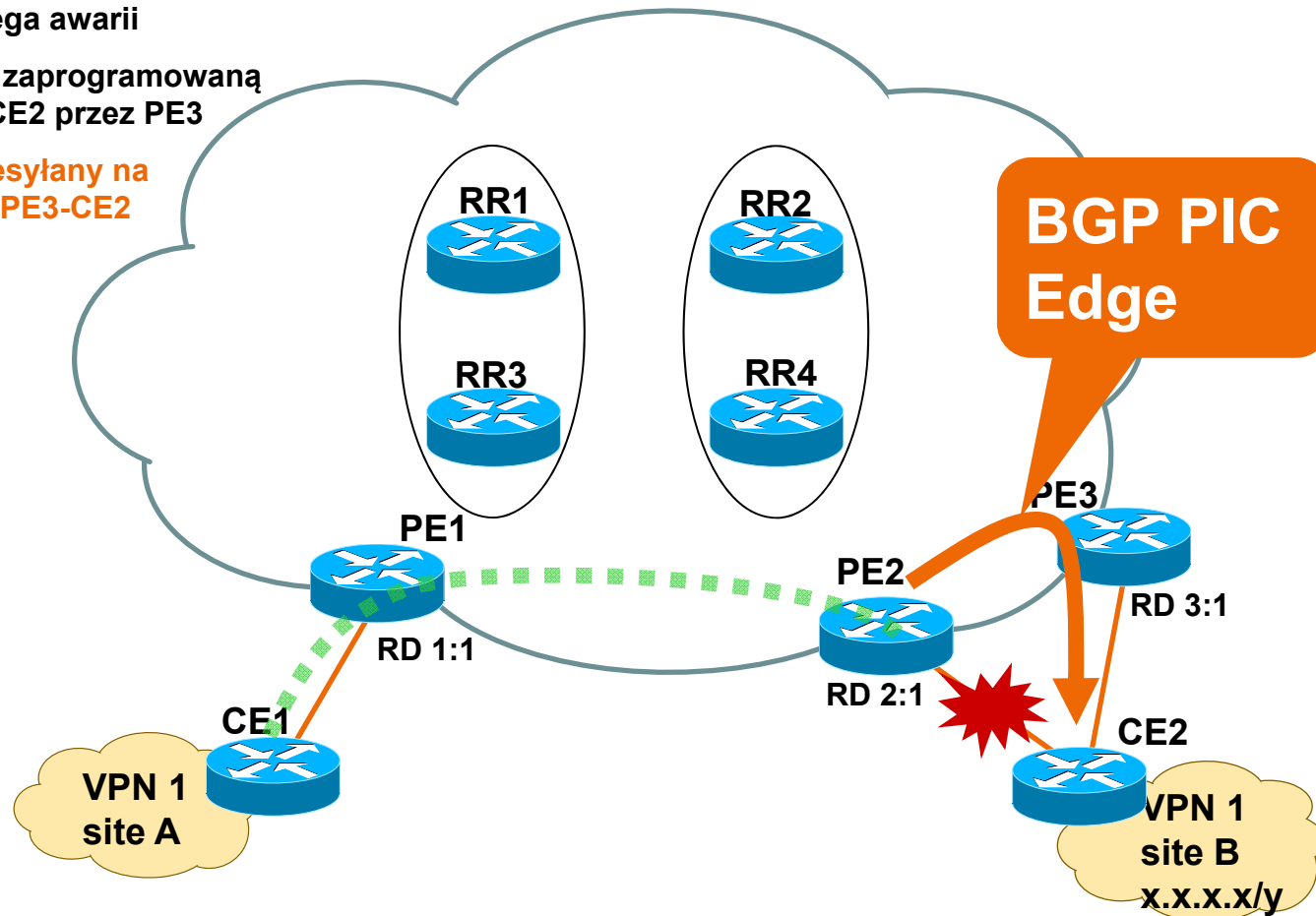
Tuning protokołów routingu IGP

Tuning protokołu BGP

BGP PIC Edge i Core

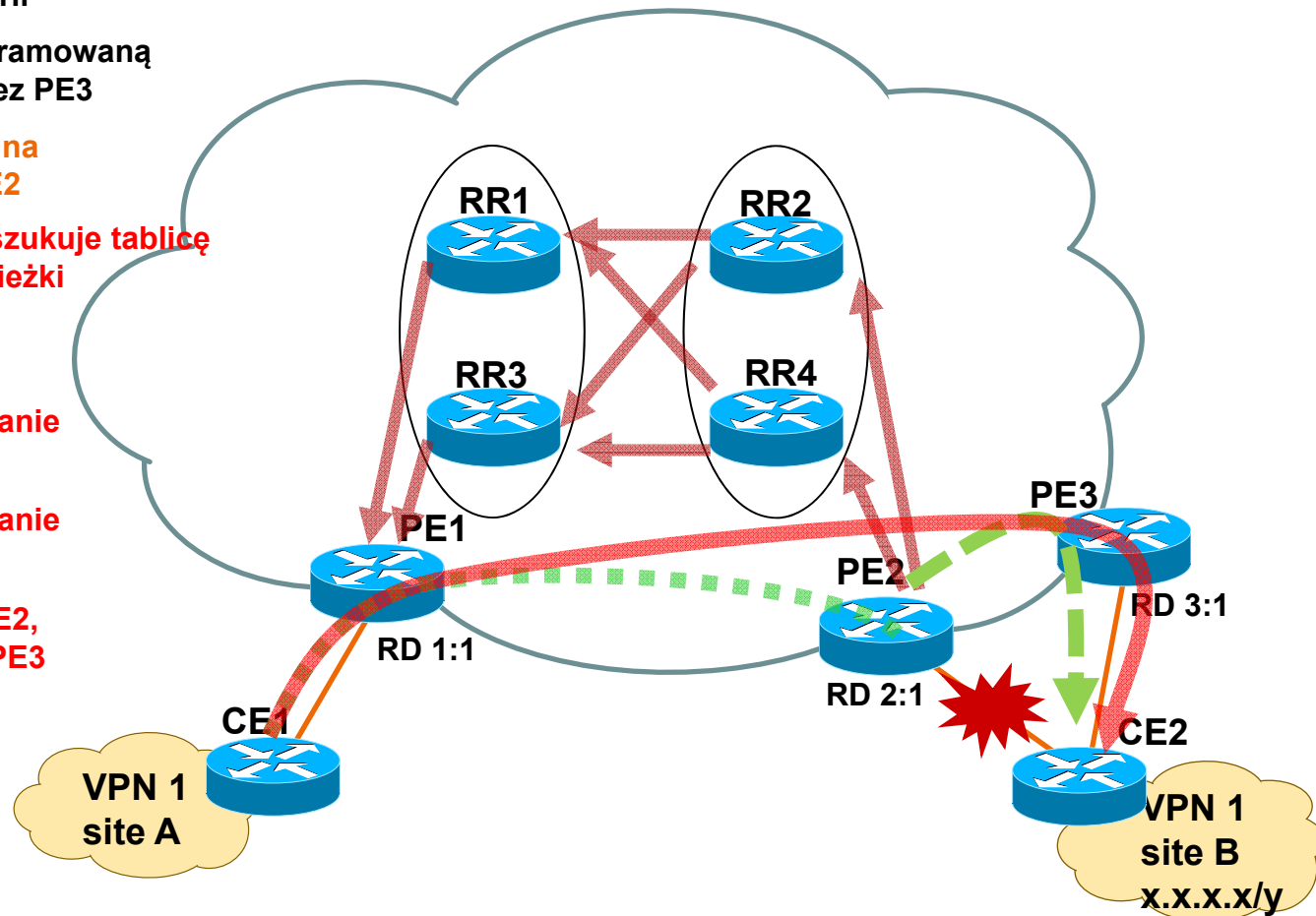
BGP PIC Edge – awaria linku PE-CE

1. Łącze PE2<>CE2 ulega awarii
2. Router PE2 posiada zaprogramowaną zapasową trasę do CE2 przez PE3
3. Ruch z CE1 jest przesyłany na trasie CE1-PE1-PE2-PE3-CE2



BGP PIC Edge – rekonwergencja

1. Łącze PE2<>CE2 ulega awarii
2. Router PE2 posiada zaprogramowaną zapasową trasę do CE2 przez PE3
3. Ruch z CE1 jest przesyłany na trasie CE1-PE1-PE2-PE3-CE2
4. Fast External Fallover przeszukuje tablicę BGP szukając najlepszej ścieżki
3. PE2 wycofuje ścieżkę
4. RR2 i RR4 propaguje wycofanie rozgłoszenia
5. RR1 i RR3 propagują wycofanie rozgłoszenia
6. PE1 kasuje ścieżkę przez PE2, trasa prowadzi teraz przez PE3



Awaria linku PE<>CE

- Czas konwergencji będzie zależał od

D: czas do wykrycia awarii

S(p): czas do przeskanowania tablicy BGP

przejście per-RD dla VPN4 a potem dla IPv4

B(p): czas do obliczenia najlepszych ścieżek i zaprogramowanie FIB

Wtx(p): czas generowania/propagowania wycofywania ścieżek

RR(p): czas na odbicie RR

Wrx(p): czas na otrzymanie i przetworzenie ścieżek wycofywanych

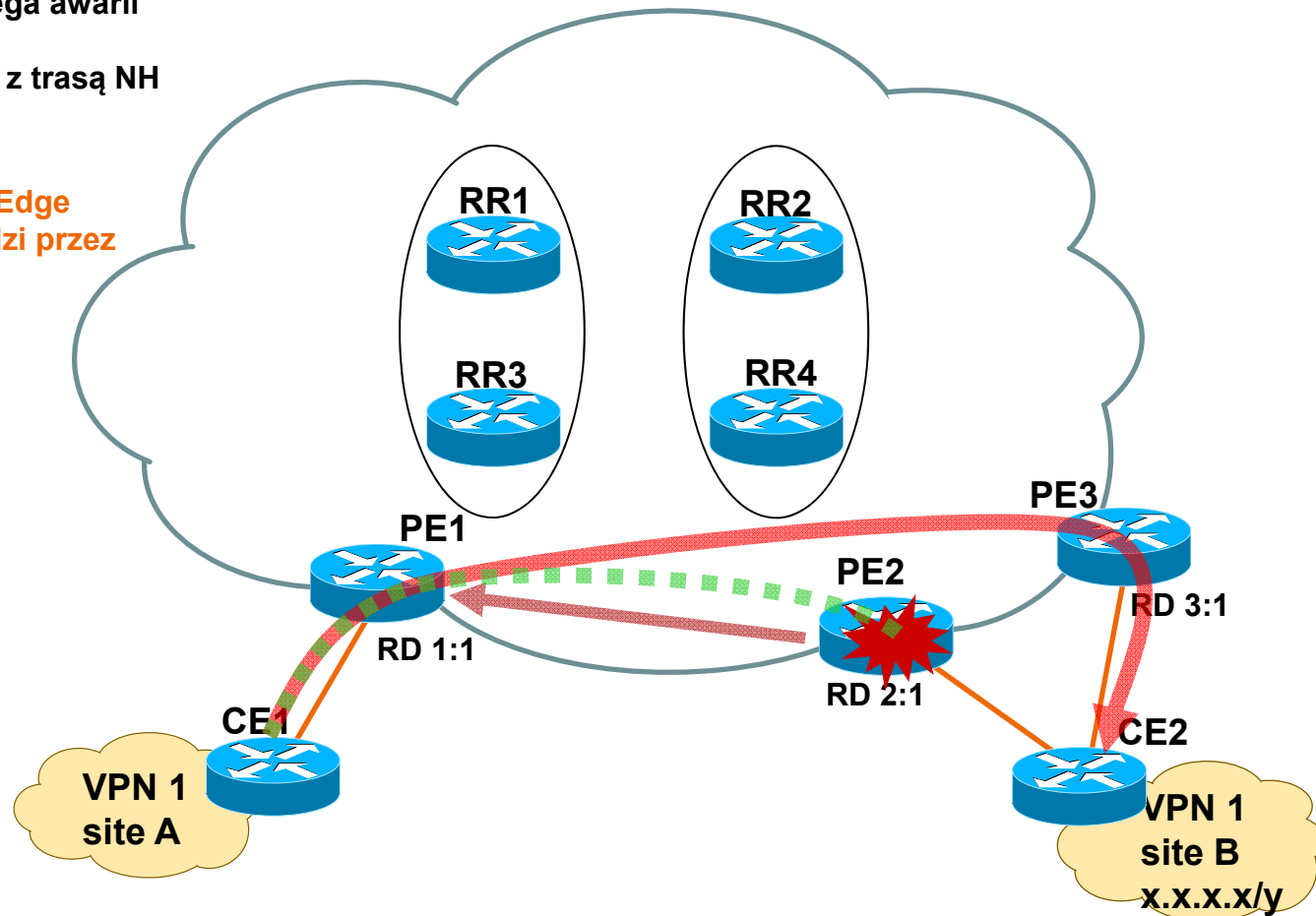
B(p): czas do obliczenia najlepszych ścieżek i zaprogramowanie FIB

X(p) oznacza że czas zależy wprost od rozmiaru tablicy

**Eliminowane
po
zastosowaniu
BGP PIC**

BGP PIC Edge – awaria węzła PE

1. Łącze do PE2<>CE2 ulega awarii
2. IGP propaguje problem z trasą NH
3. PE1 wycofuje ścieżki
4. Po włączeniu BGP PIC Edge zapasowa trasa prowadzi przez PE1,PE3,CE2



Awaria węzła PE

- Czas konwergencji będzie zależał od
 - D: czas do wykrycia awarii
 - IGP: czas na konwergencję IGP

**Eliminowane po
zastosowaniu
BGP PIC**

S(p): przeskanowanie tablicy BGP

Przejrzenie tablicy VPNv4 a potem IPv4

B(p): czas do obliczenia najlepszych ścieżek i
zaprogramowanie FIB

Konfiguracja BGP PIC

- W konfiguracji procesu BGP

```
address-family {ipv4 unicast | vpnv4}  
    bgp additional-paths install
```

- Jak to wygląda w tablicy RIB?

```
r# show ip bgp 10.0.0.0 255.255.0.0  
BGP routing table entry for 10.0.0.0/16, version 123  
Paths: (4 available, best #3, table default)  
Additional-path  
Advertised to update-groups:  
 2 3  
Local  
 10.0.101.2 from 10.0.101.2 (10.1.1.1)  
  Origin IGP,localpref 100,weight 900,valid, internal, best  
Local  
 10.0.101.1 from 10.0.101.1 (10.5.5.5)  
  Origin IGP,localpref 100,weight 700,valid, internal, backup/repair
```



Q&A

Łukasz Bromirski, lbromirski@cisco.com



CISCO