



Protokoły warstwy 2

Przegląd dostępnych opcji



Łukasz Bromirski
lbromirski@cisco.com



PLNOG, Warszawa, marzec 2011

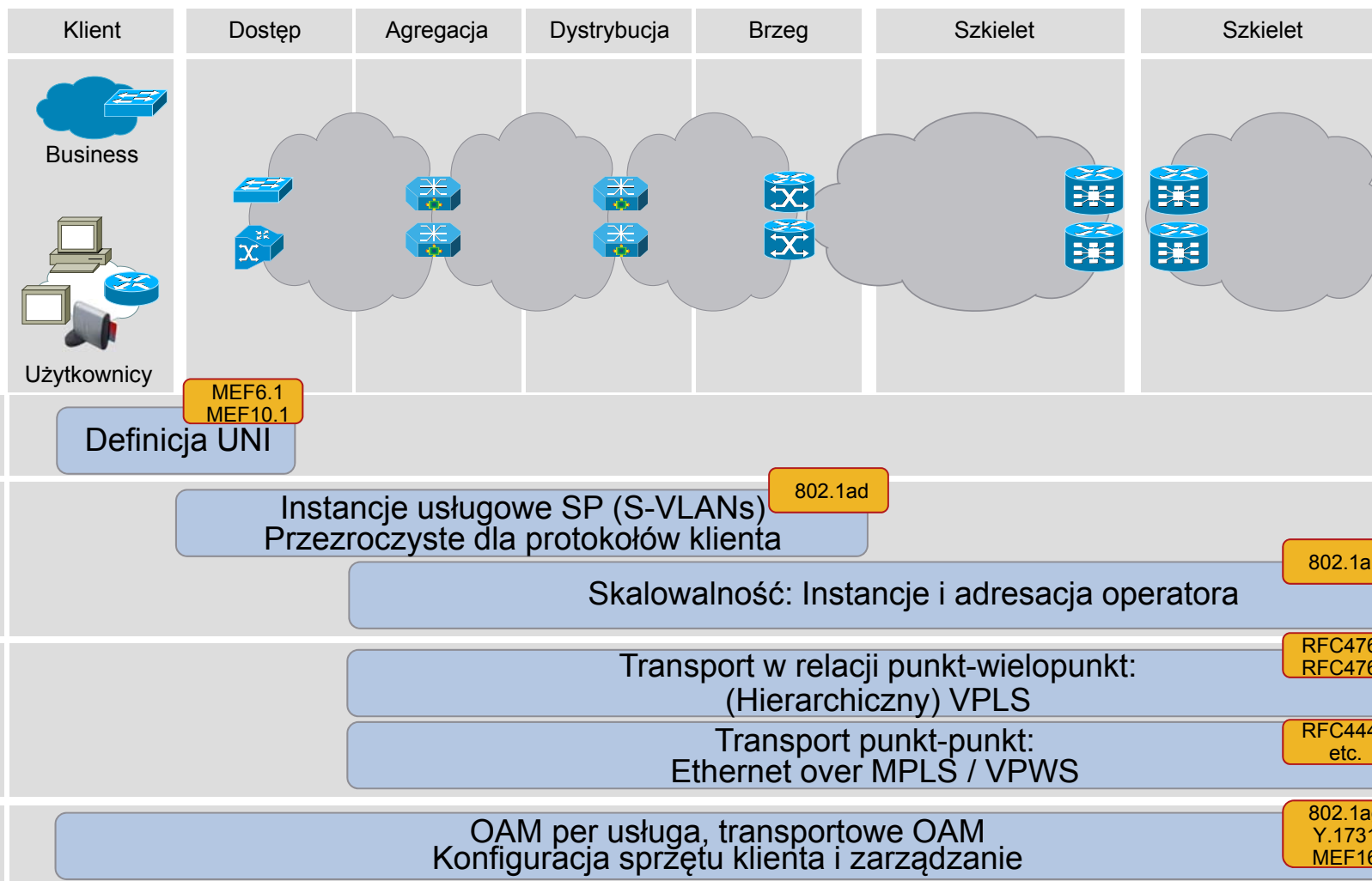
Agenda

- Ethernet
- Spanning Tree
- Praca warstwy drugiej w pierścieniach
- Transport ruchu L2 przez operatora
- Nadzór sieci i diagnostyka
- Nowe standardy
- Q&A

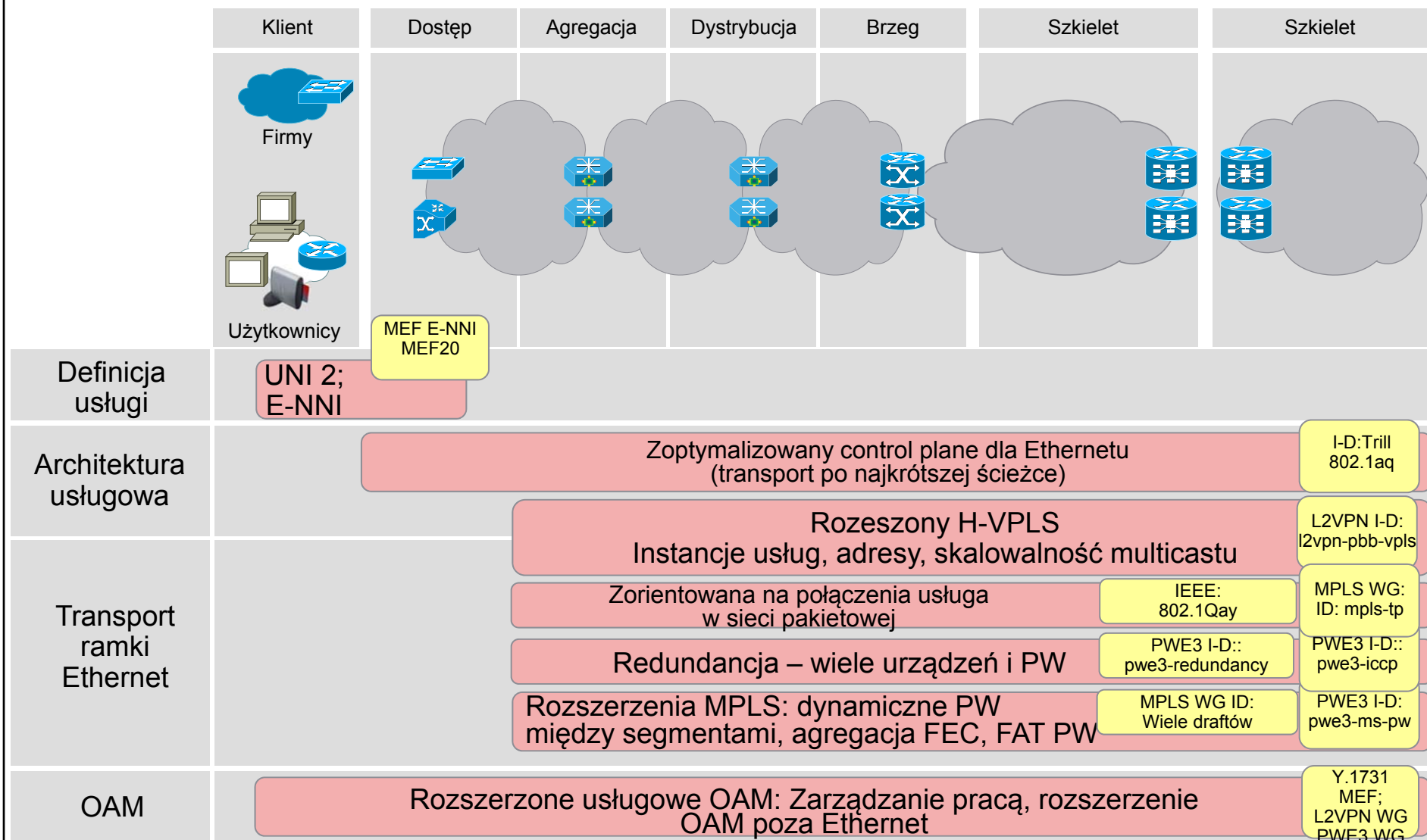
Ethernet

Co się dzieje obecnie i jak to wygląda?

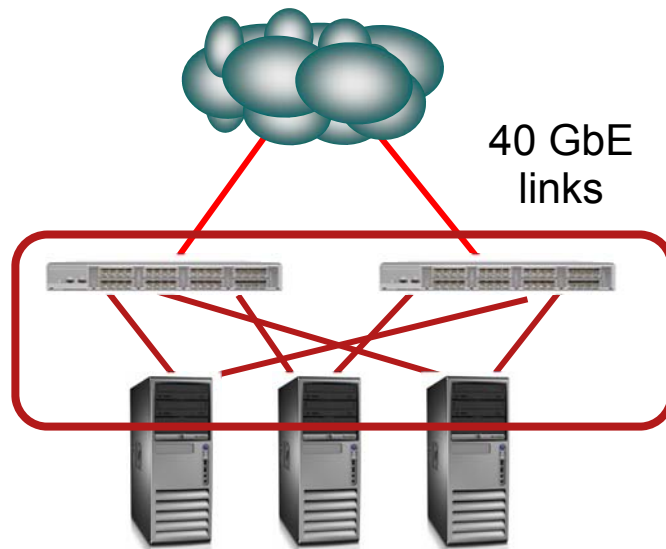
Standardy Carrier Ethernet



Standardy Carrier Ethernet - nowe

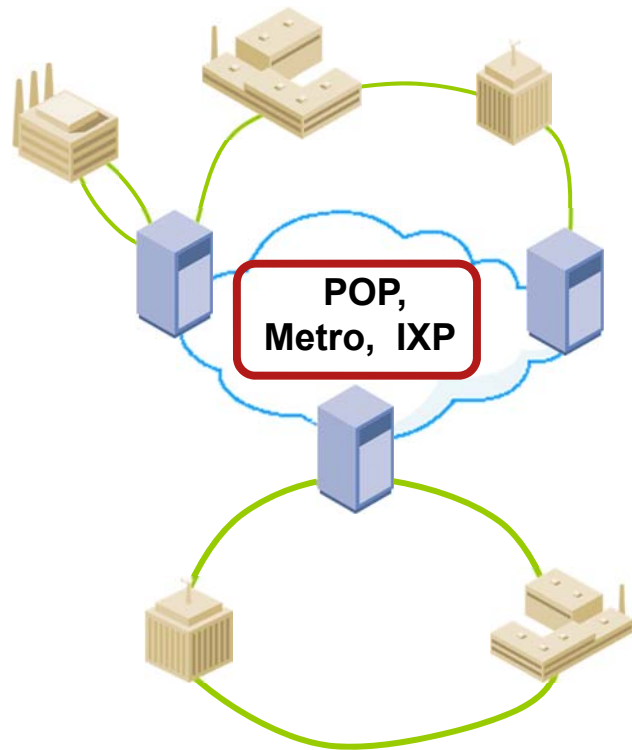


IEEE 802.3ba: 40Gbit/s Ethernet



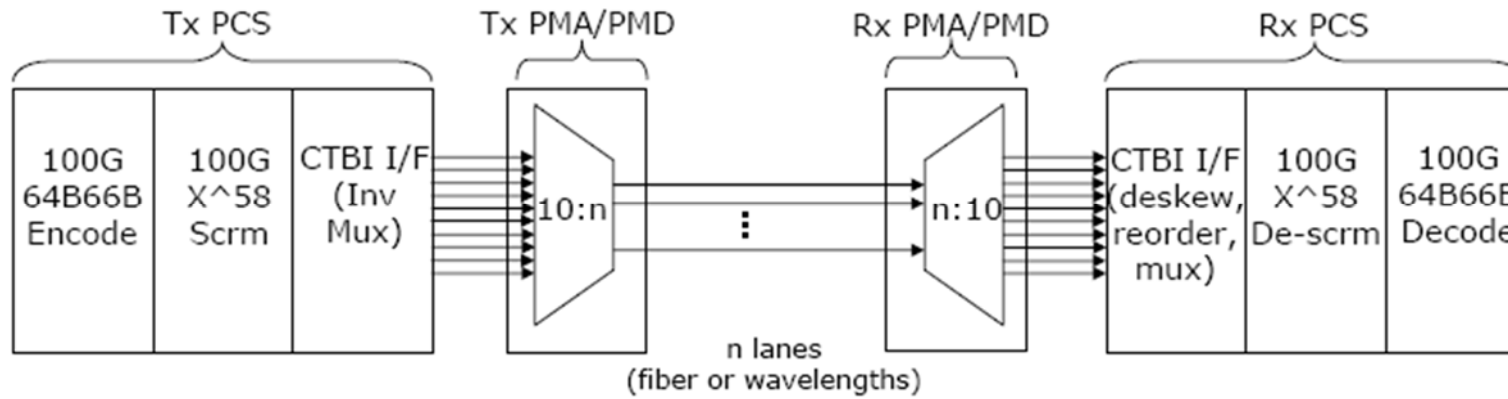
- Wsparcie do pracy wyłącznie w trybie full-dupleks
- Wykorzystuje format ramki Ethernet oraz limity wielkości
- Zachowuje BER na poziomie 10^{-12}
- Transmisja ustandaryzowana dla
 - 1m dla połączeń wewnątrz matrycy
 - 10m na kablu miedzianym
 - 100m światłowod OM3
 - 10km światłowod SM

IEEE 802.3ba: 100Gbit/s Ethernet



- Wsparcie do pracy wyłącznie w trybie full-dupleks
- Wykorzystuje format ramki Ethernet oraz limity wielkości
- Zachowuje BER na poziomie 10^{-12}
- Transmisja ustandaryzowana dla
 - 10m na kablu miedzianym
 - 10km SM dla sieci Metro
 - 40km SM long-haulWsparcie dla transmisji w sieci OTN

IEEE 802.3ba: 100 GE and 40 GE

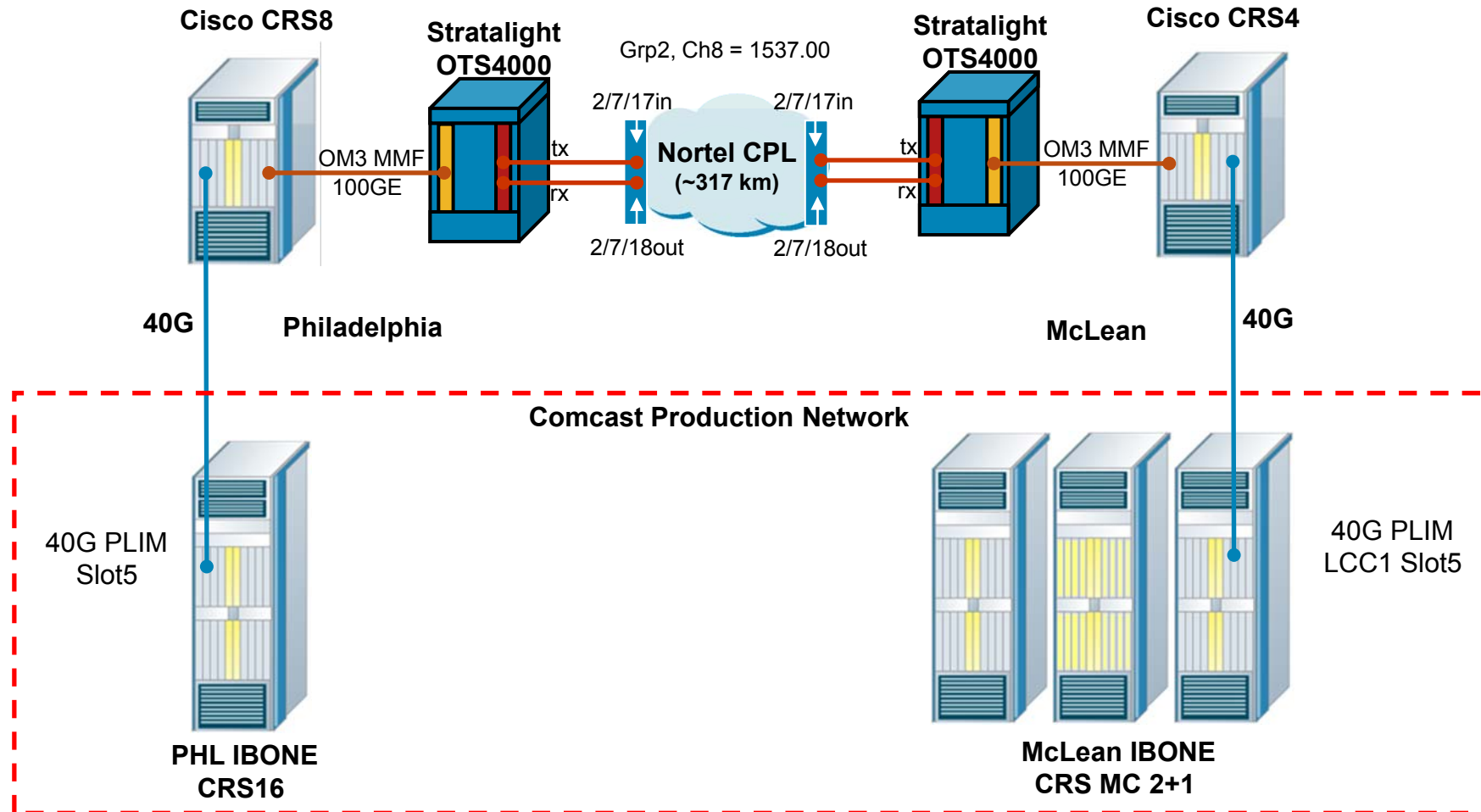


Distance	40G Ethernet	100G Ethernet
At least 1m backplane	☑	
At least 10m cu cable	☑	☑
At least 100m OM3 MMF	☑	☑
At least 10km SMF		☑
At least 40km SMF		☑

Target standards completion: 2010

Cisco 100GbE demo – Czerwiec 2008

100GE DEMO Network



http://newsroom.cisco.com/dlls/2008/prod_062608c.html

Did You Know that Copper Takes Longer than Fiber for Failure Detection?

- Direct point to point fiber provides for fast failure detection
- IEEE 802.3z and 802.3ae link negotiation define the use of **Remote Fault Indicator & Link Fault Signaling mechanisms**

Spanning Tree

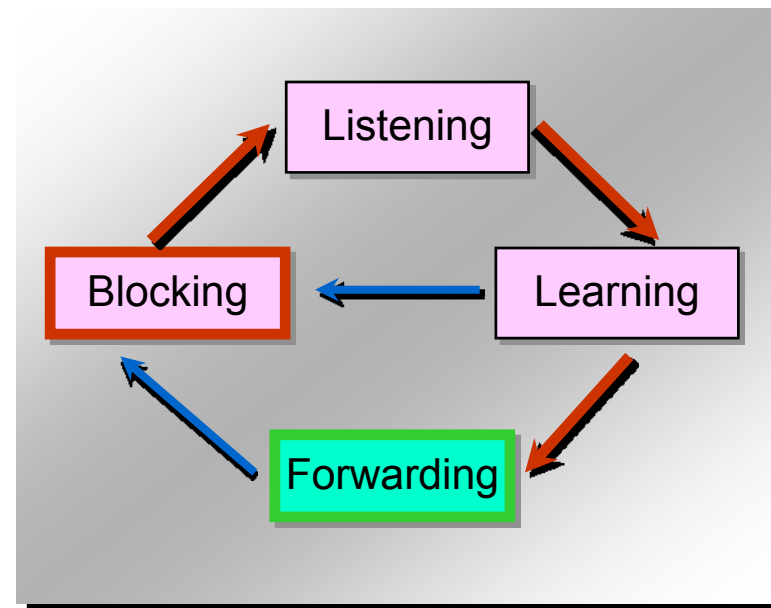
802.1d, 802.1s, 802.1w i inne standardy

Co w 802.1 piszczy?

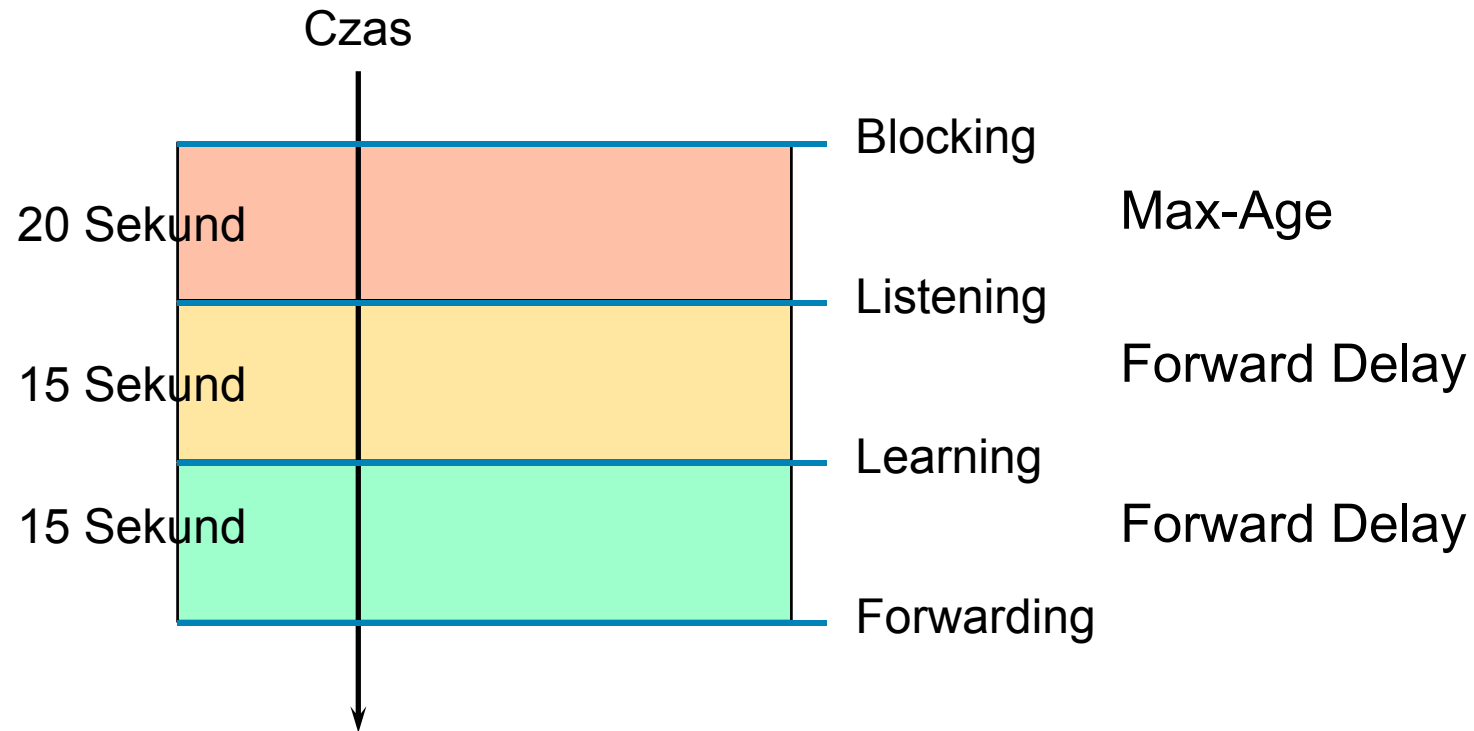
- 802.1D: MAC Bridges (Spanning Tree Protocol)
- 802.1w: Rapid Spanning Tree Protocol (RSTP)
- 802.1s: Multiple Spanning Tree Protocol (MST)
- 802.1t: 802.1d Maintenance
- 802.1Q: VLAN Tagging (trunking)

Stany portów w STP

- Blocking
- Listening
- Learning
- Forwarding
- Disabled (wyłączony)



Spanning Tree i liczniki



Głównym ograniczeniem konwergencji tradycyjnego STP jest przywiązanie do zależności czasowych

Nowe funkcje w Rapid STP

- Nowe role i stany portów
- Zmodyfikowane BPDU
- Nowa obsługa i zachowanie BPDU
- Nowy mechanizm zmiany topologii
- Szybkie przejście do stanu „forwarding”
- Możliwość migracji – kompatybilność z 802.1D

Role portów w Rapid STP

- RSTP definiuje 4 role portów:

Root port

Designated port

Alternate port

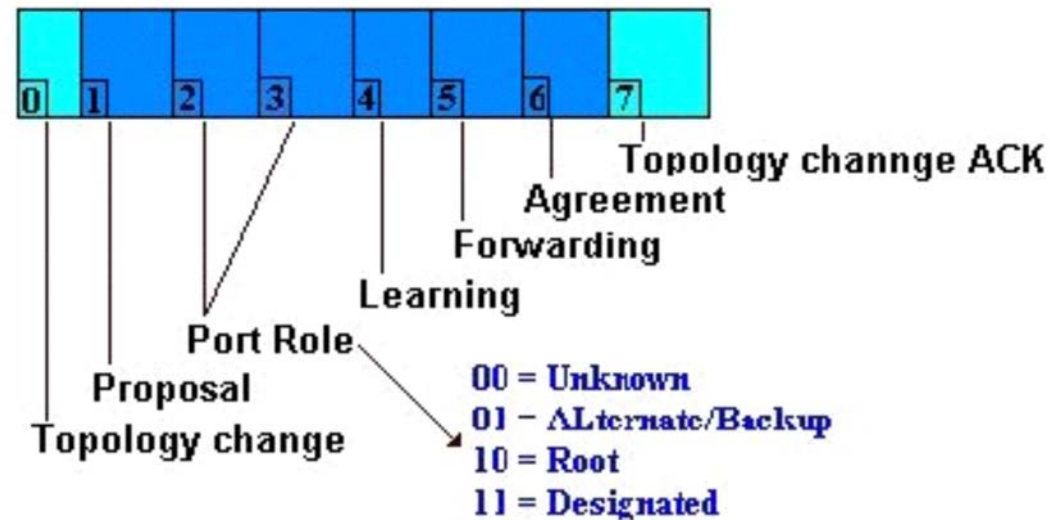
Backup port

} **blokujące**

- Stan portu może być ustalony bez względu na jego rolę: blocking, forwarding, learning (listening = blocking)

Zmodyfikowane BPDU

- Pole „wersja protokołu” ustawione na 2 (było 0)
- Znika TCN BPDU
- Zmiana w polu flag



- Mosty 802.1D odrzucają BPDU z RSTP

Nowa obsługa i wykorzystanie BPDU

- BPDU są wysyłane jako ramki „keepalives”

most wysyła BPDU co czas „hello” (domyślnie 2 sekundy)

informacja o porcie zostaje unieważniona po upływie 3x czasu bez otrzymania BPDU

Szybkie przejście portu do „forwarding”

- Dotyczy tylko połączeń P2P

Port domyślnie jest P2P jeśli pracuje w duplexie oraz nie jest portem brzegowym

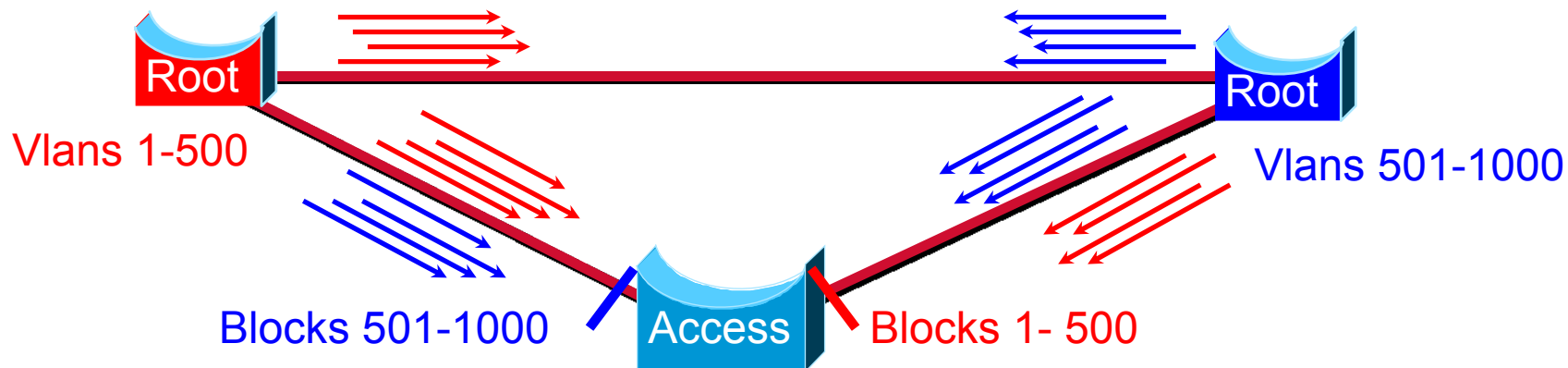
- Po otrzymaniu lepszej propozycji, most od razu wykonuje jednocześnie trzy rzeczy:

Przyjmuje lepszą propozycję

Blokuje port w stronę dotychczasowej sieci

Wysyła propozycję nowego root'a w stronę reszty sieci, jednocześnie akceptuje propozycję lepszego root'a i przesuwa porty w stan „forwarding”

Po co MST?

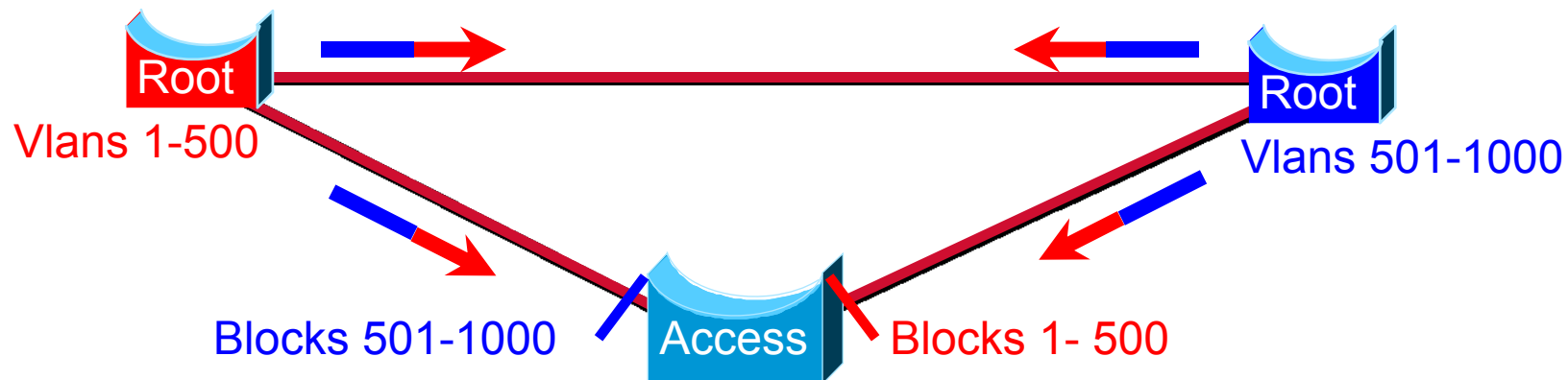


Jeden VLAN to jedna instancja STP

Możliwość zrealizowania „ręcznego” rozkładania ruchu przez podział VLANów

CPU obsługuje 1000 instancji mimo tylko dwóch topologii fizycznych

Co daje MST?



- Wygodne rozkładanie ruchu i oszczędność CPU – tylko dwie różne topologie STP
- Dostyc zložone – we wdrożeniu a potem przy ewentualnym rozwiązywaniu problemów oraz współpracy z innymi protokołami

Praca w topologii pierścienia w warstwie drugiej

REP, EAPS, RRPP, EPSR, MRP, RRP i ERPS

Trendy i problemy w L2

- Duże domeny SPT

 - Coraz więcej węzłów w sieci

- Większa ilość klientów

 - Zwiększamy ilość VLANów i adresów MAC dla domeny L2

- W oczywisty sposób powoduje to zwiększenie problematyczności rozwiązywania problemów i diagnostyki

- Naturalne aspiracje dla usług „klasy operatorskiej”

 - Co z szybką konwergencją?

 - STP/rSTP/MSTP nie jest szczególnie „klasy operatorskiej”

Topologia gwiazdy i pierścienia

Pierścień

- Współdzielone pasmo pomiędzy urządzeniami U-PE.
- Bardziej skomplikowana inżynieria ruchu. Kolejki wyjściowe współdzielone z innymi U-PE.
- Wyższe opóźnienie z powodu większej ilości przeskoków.
- Dłuższy czas zbieżności.
- Utrudniona kontrola procesu uczenia się adresów MAC.

Różnice

Gwiazda

- Każdy U-PE posiada własne połączenie 1GE do n-PE.
- Prostsza inżynieria ruchu. Kolejki wyjściowe dedykowane dla lokalnego urządzenia U-PE.
- Mniejsze opóźnienie.
- Krótszy czas zbieżności.
- Łatwiejsza kontrola adresów MAC.

Podobieństwa

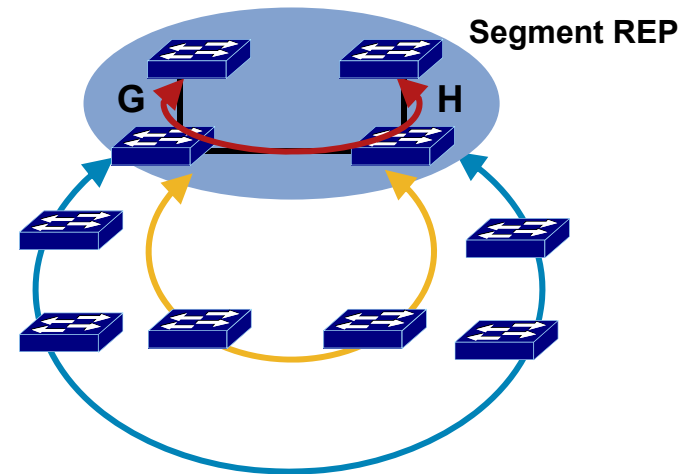
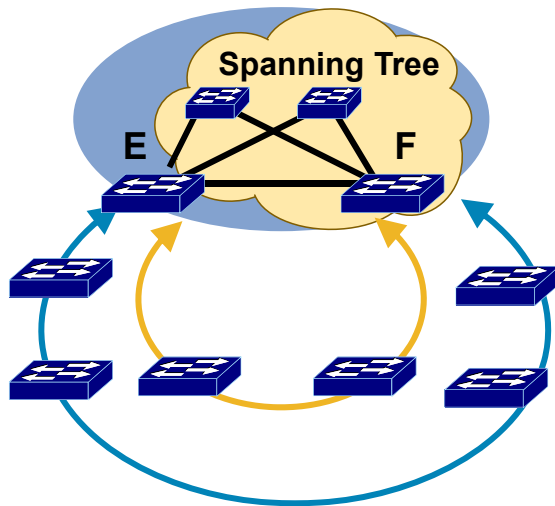
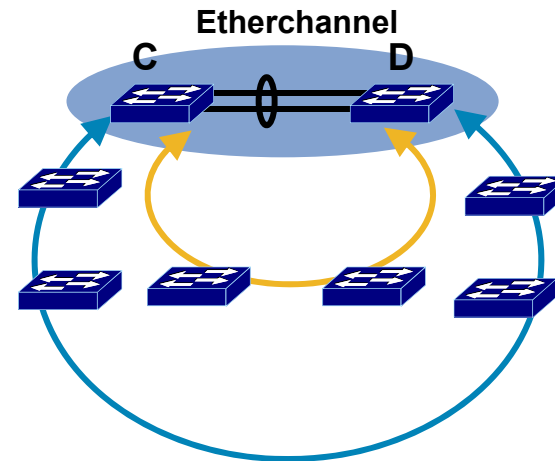
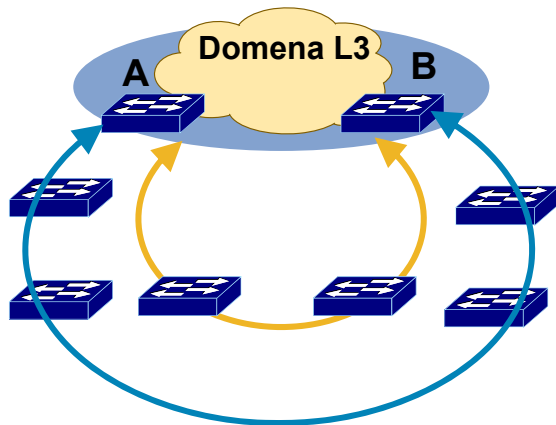
- Wszystkie usługi są dostępne w obydwu topologiach.
- Dostępne te same funkcjonalności bezpieczeństwa, dostępności oraz QoS na urządzeniu N-PE.

Wszyscy potrafimy to robić

- Cisco REP
- Extreme EAPS
- 3COM (Huawei H3C -> HP) RRPP
- Allied Telesyn EPSR
- Brocade MRP / Force10 RRP

- **Standard:** ITU G.8032 ERPS

REP - Topologia pierścienia



Ethernet dla operatorów

802.1ad, 802.1ah i 802.1Qay

IEEE 802.1ad – Provider Bridges

- Przezroczyste dla VLANów klienta

IEEE 802.1ad powinny dostarczyć ustandaryzowaną wersję “QinQ”

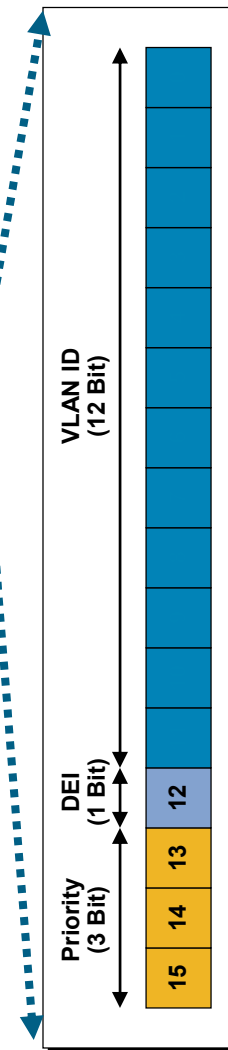
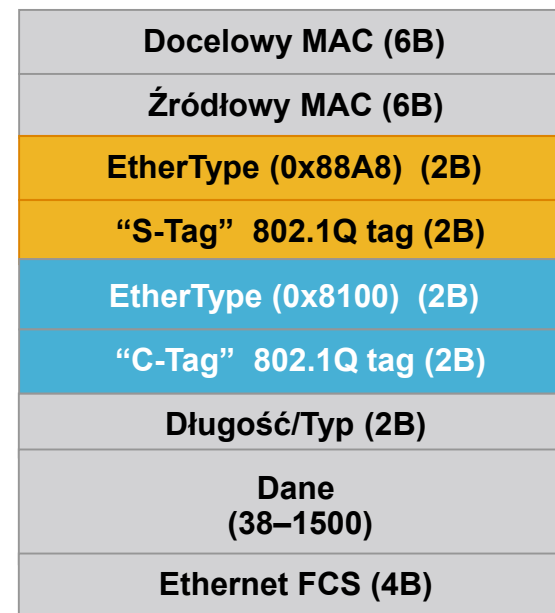
Standard ma zawierać pewne rozszerzenia

- Format ramki zgodny z “QinQ”

Nowy EtherType: 0x88A8

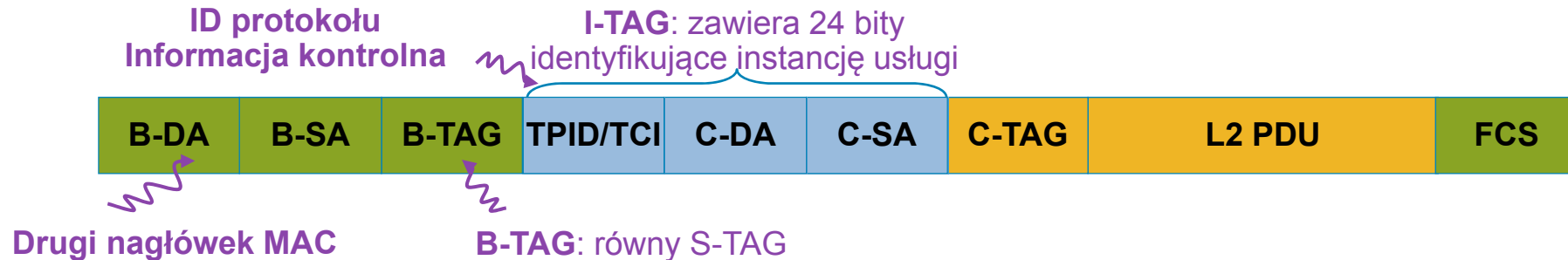
- Technicznie zatwierdzony już dawno – 8 grudzień 2005

“Opcjonalne”



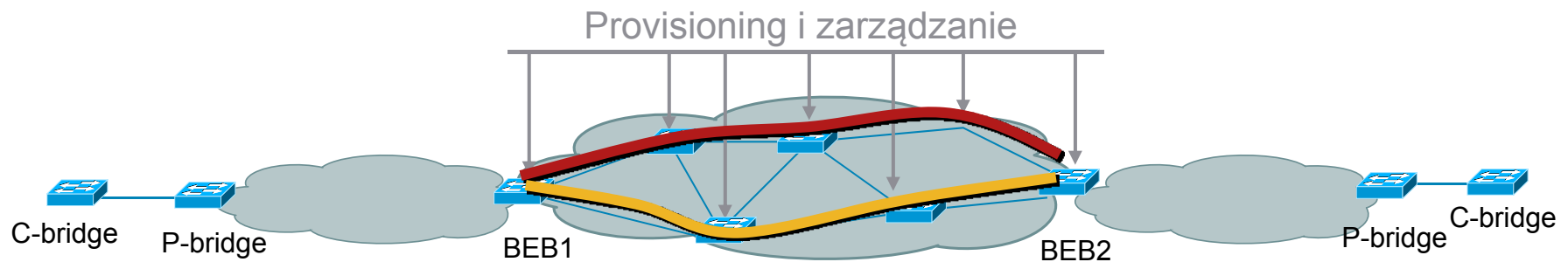
- Więcej na: <http://www.ieee802.org/1/pages/802.1ad.html>

IEEE 802.1ah – Provider Backbone Bridges



- Skalowalność usługi – 24 bity (I-SID) wskazujące usługę
- Izolacja domen, skalowalność adresów MAC
 - Na wejściu do domeny operatora, nagłówek MAC klienta jest wkładany w nagłówek MAC operatora
- Kompatybilny wstecz z 802.1ad
 - Zewnętrzny nagłówek jest normalnym nagłówkiem 802.1ad
 - 802.1ah zakłada wykorzystanie istniejących mechanizmów L2 – takich jak SPT i mechanizmy uczenia się/floodowania
 - Inne opcje wdrożenia są również dostępne - 802.1aq, 802.1Qay, ukrywanie topologii z wykorzystaniem VPLS/MPLS oraz redundancja PW
- Standard 802.1ah zaakceptowany 12 czerwca 2008

IEEE 802.1Qay – PB + TE

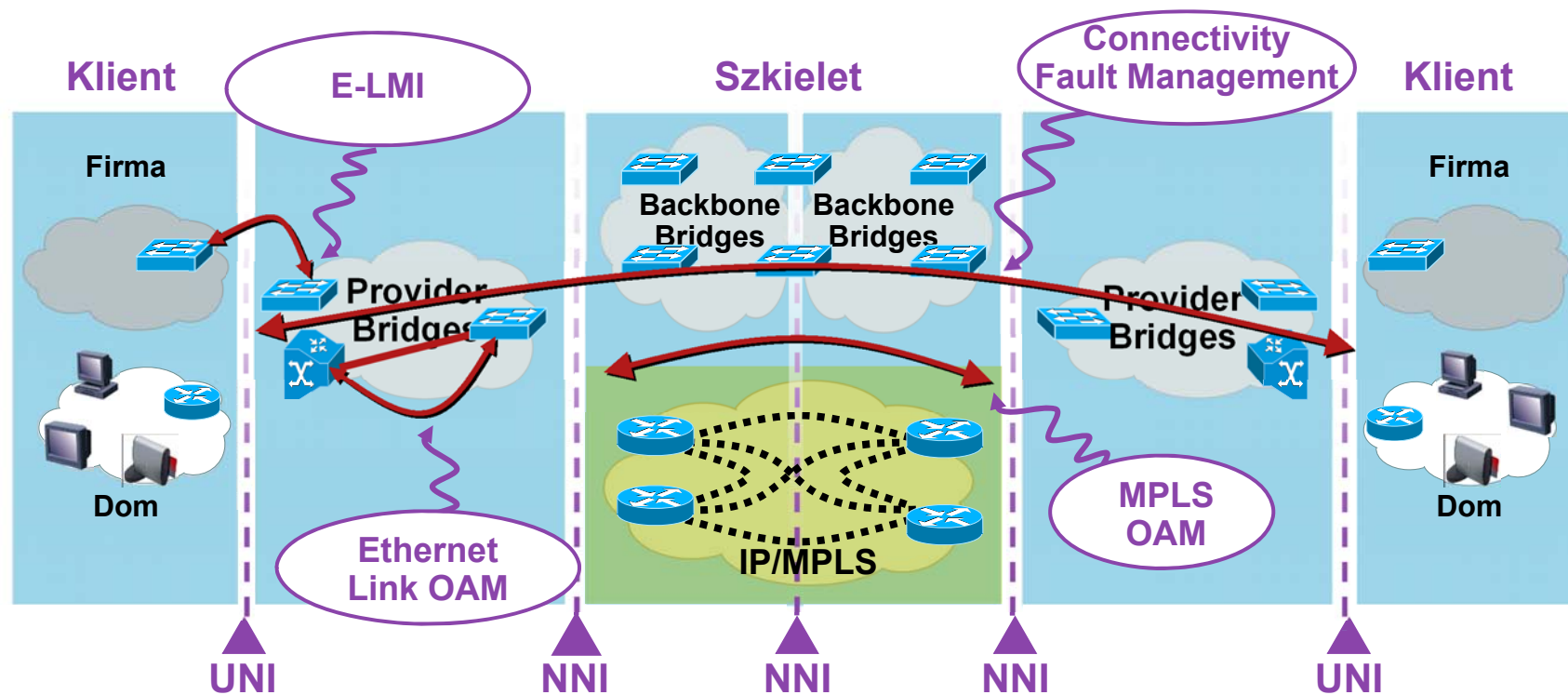


- Standaryzacja mechanizmów inżynierii ruchowej dla sieci składających się z połączeń punkt-punkt (w stylu PVC ATMowych)
- Bazuje na PBB z ograniczonymi opcjami
 - Wyłączone uczenie (jest w 802.1Q) i flooding
- Statyczny control plane
 - Provisioning odbywa się przez NMS, wykorzystywany jest MIB 802.1
 - Osobne prace nad dynamicznym procesem provisioningu (802.1aq) – poza standardem

OAM

Operations, Administration & Management

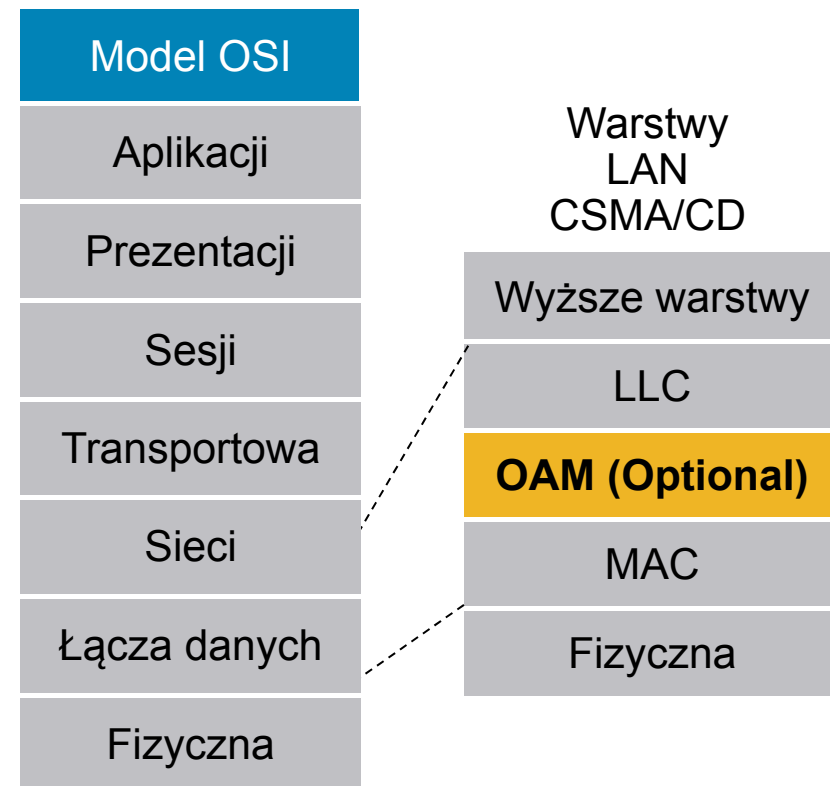
Ethernet OAM – co i gdzie?



- E-LMI—User to Network Interface (UNI)
- Link OAM—dowolne łącze punkt-punkt 802.3
- CFM—End-to-End UNI do UNI
- MPLS OAM—w chmurze MPLS

Link OAM (IEEE 802.3ah, klauzula 57)

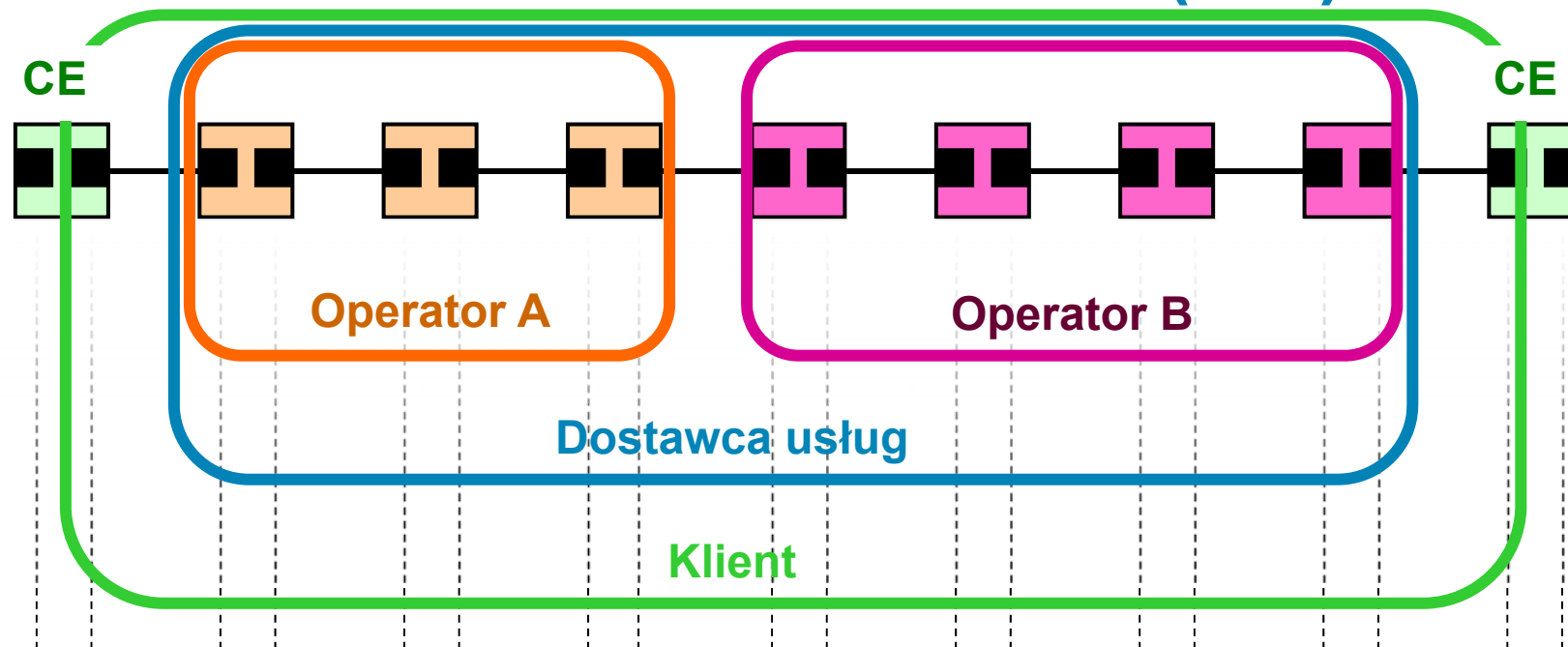
- Mechanizm dla „monitoringu pracy łącza” czyli między innymi do:
 - Monitoringu samego łącza
 - Wskaźnika zdalnej awarii
 - Kontroli zdalnej pętli lokalnej
- Dodaje opcjonalną podwarstwę OAM
- Pracuje na łączach punkt-punkt 802.3
- Używa ramek OAMPDU nie przekazywanych dalej przez klientów MAC wysyłanych relatywnie wolno
 - do 10 ramek na sekundę



Connectivity Fault Management

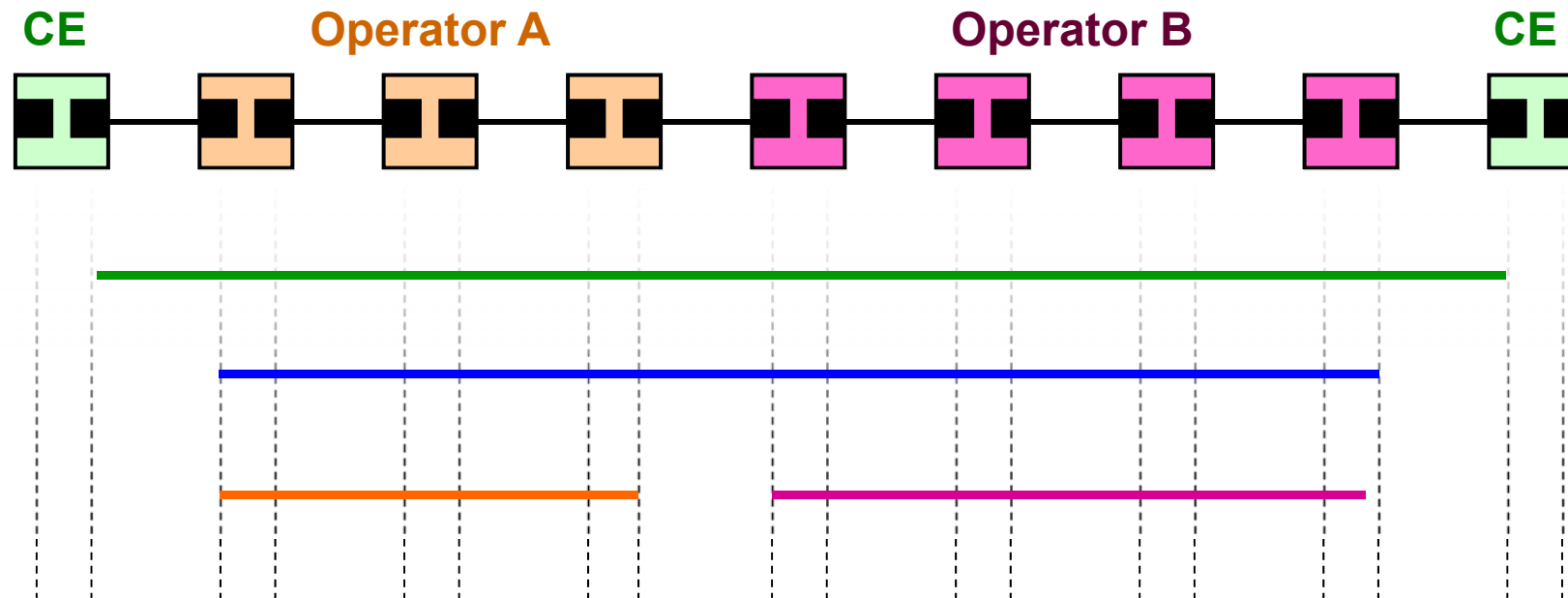
- Rodzina protokołów zapewniająca możliwość wykrywania, weryfikacji, izolacji i raportowania problemów w transporcie ruchu na całej trasie od źródła do celu
- Tradycyjne ramki Ethernet, z ustawionym polem EtherType na 0x8902 i docelowym adresem MAC multicast
- Ustandaryzowany przez IEEE pod koniec 2007

CFM – Maintenance Domain (MD)



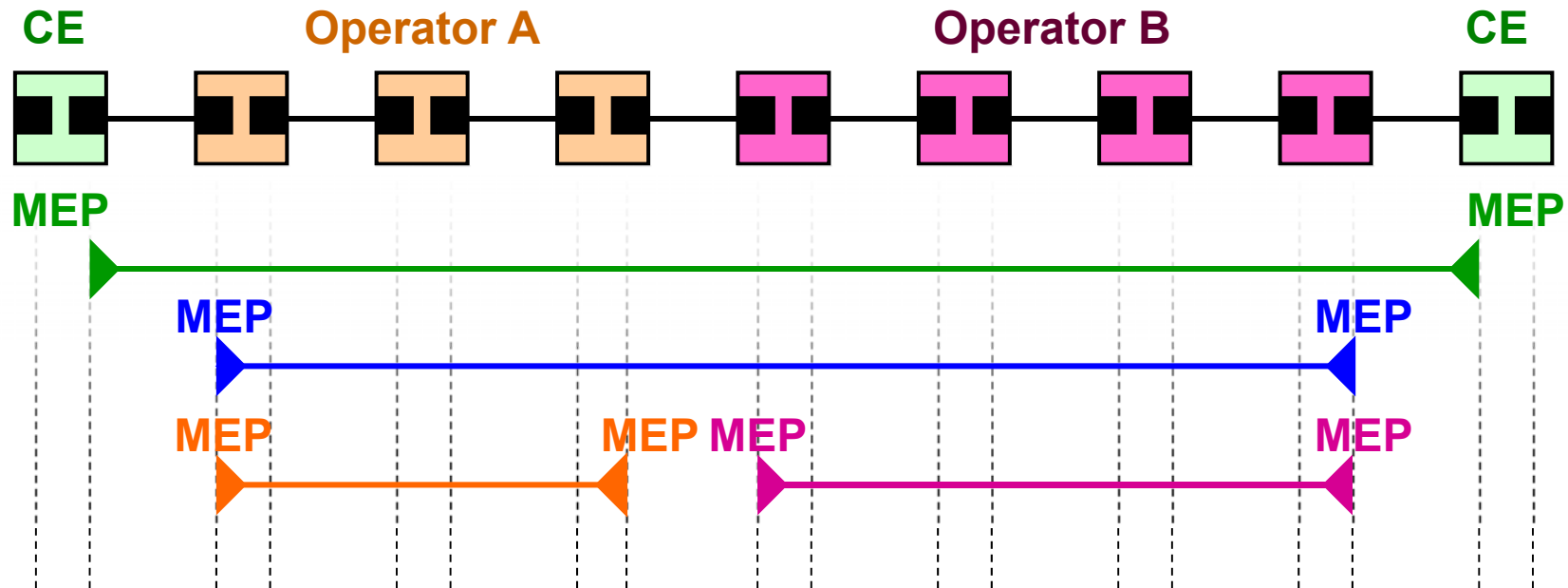
- Definiowana przez granice operacyjne/kontraktu
- MD mogą być zagnieżdżone lub się stykać, ale nie mogą się przenikać
 - do 8 poziomów zagnieżdżenia (0...7) – czym wyższy poziom tym większy zasięg
- Nazwa MD: pusta, adres MAC, DNS lub ciąg znaków

CFM – Maintenance Association (MA)



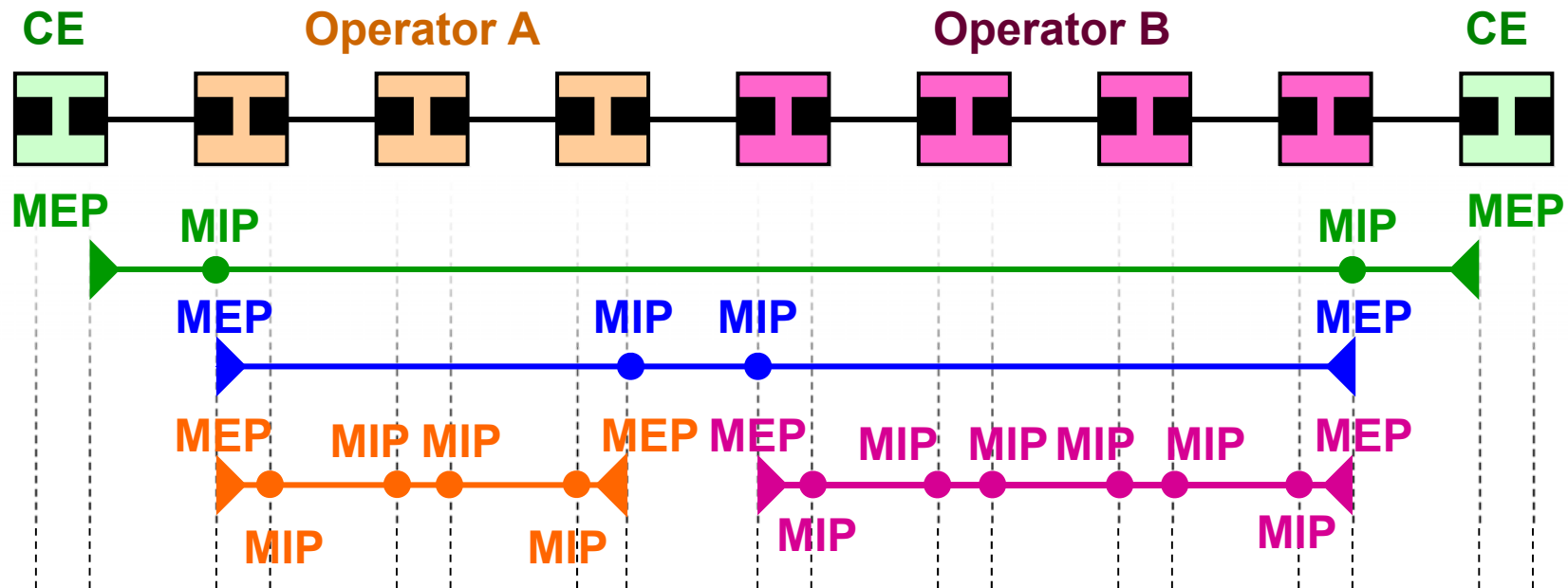
- Monitoruje łączność w danej instancji usługi w danym MD (np. 1 usługa przechodząca przez cztery MD = 4 MA)
- Definiowana przez zestaw Maintenance End Points (MEP) na brzegu domeny
- Identyfikowana przez MAID == „Krótka nazwa MA” + nazwa MD
- Krótka nazwa MA: ID VLANu, VPN, liczba dodatnia lub ciąg znaków

CFM – Maintenance Point (MP) - MEP



- **Maintenance Association End Point (MAEP)**
- Definiuje granice MD
- Obsługuje automatyczne wykrycie problemów z połączeniem pomiędzy dowolną parą MEPów w MA
- Kojarzone per MA i identyfikowane przez MEPID (1-8191)
- Może inicjować i odpowiadać na PDU CFM'owe

CFM – Maintenance Point (MP) - MIP

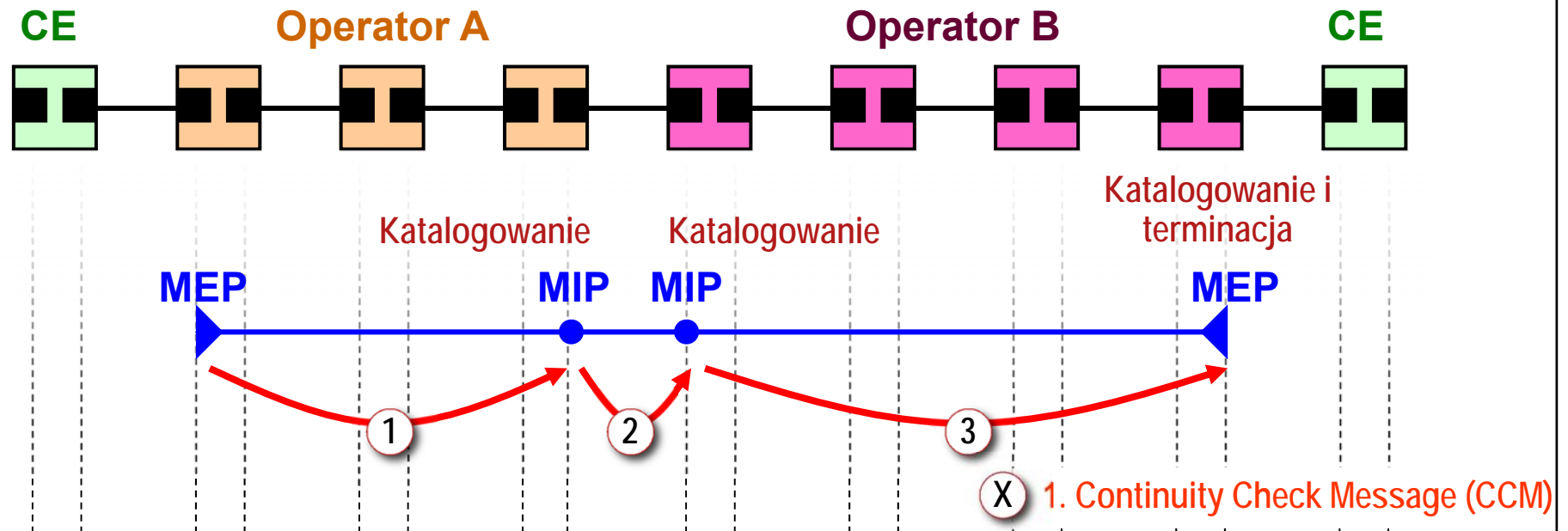


- **Maintenance Domain Intermediate Point (MIP)**
- Obsługuje wykrywanie ścieżek pomiędzy MEP i lokalizację awarii na tych ścieżkach
- Może być kojarzony per MD i VLAN / EVC (ręcznie lub automatycznie)
- Może dodawać, sprawdzać i odpowiadać na PDU CFM'owe

Protokoły CFM

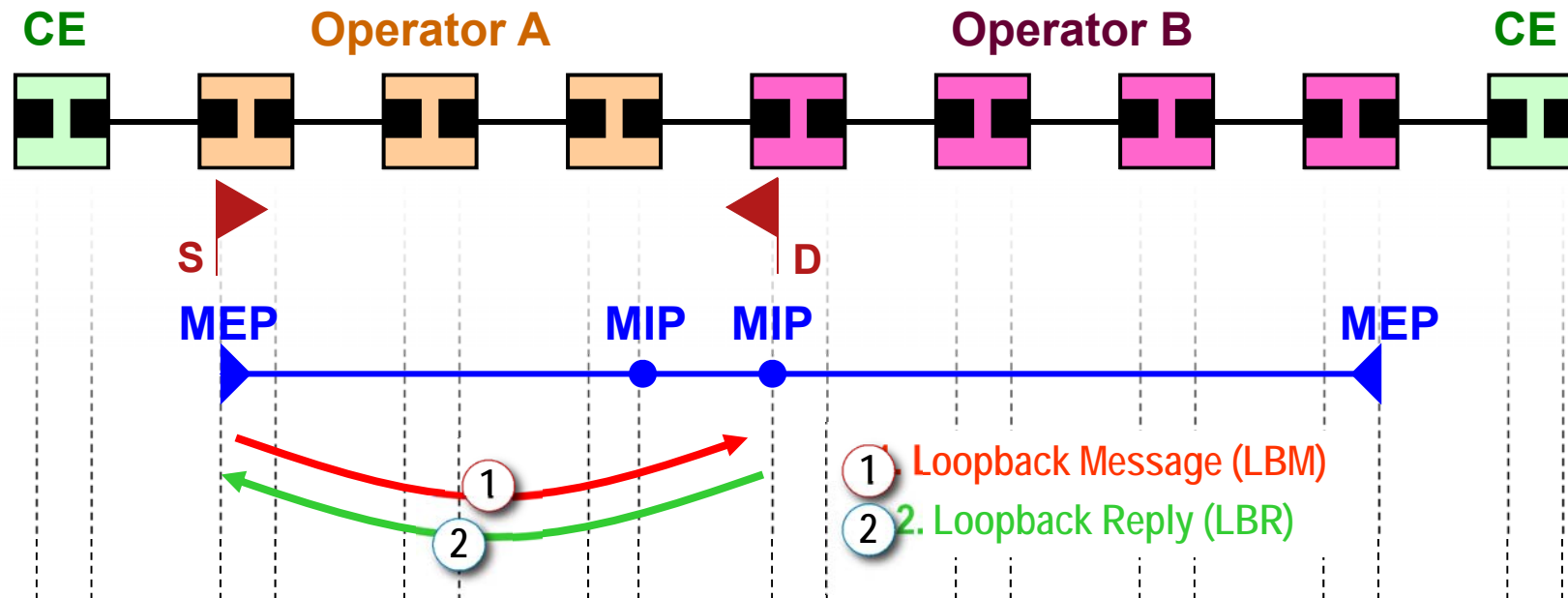
- CFM definiuje obecnie trzy protokoły
- Continuity Check Protocol
 - Wykrywanie i Notyfikacja o awarii
- Loopback Protocol
 - Weryfikacja awarii
- Linktrace Protocol
 - Izolacja awarii

CFM – protokół Continuity Check



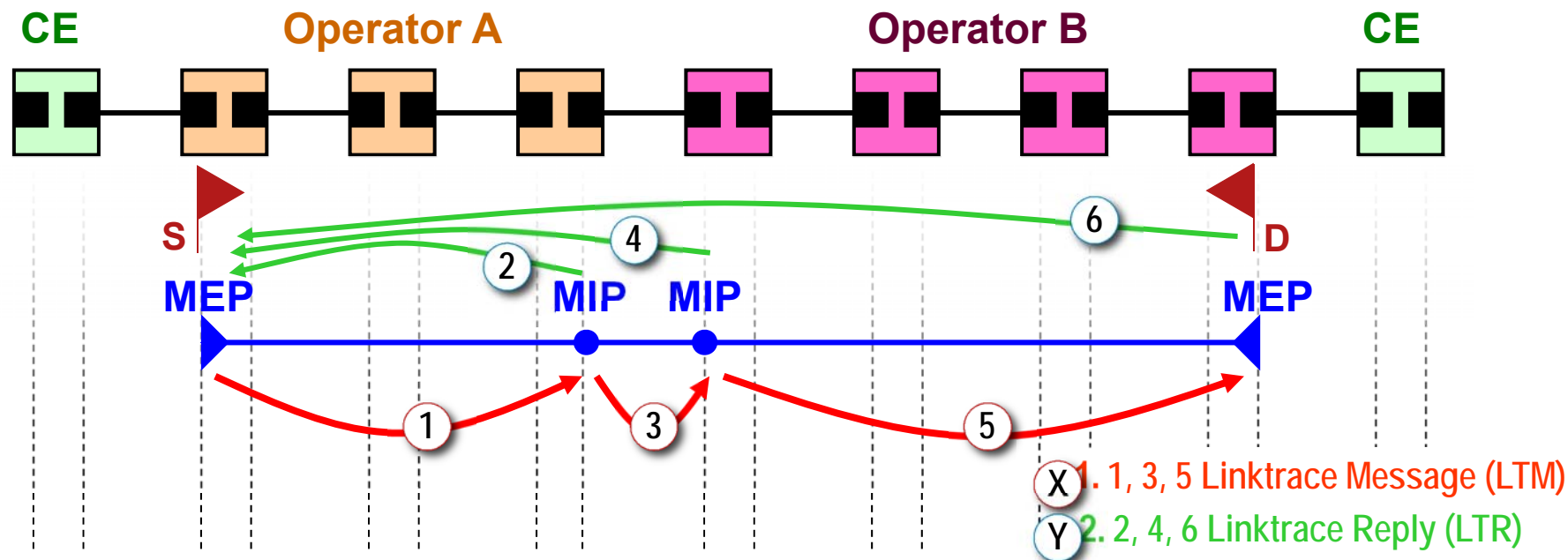
- Wykrywanie i informowanie o awariach
- Skojarzenie per-MA – multicastowa wiadomość „keepalive”
 - Wysyłana przez MEP co konfigurowalny interwał czasowy (3.3ms, 10ms, 100ms, 1s, 10s, 1m, 10m)
 - Jednokierunkowa
 - Przenosi status portu na którym skonfigurowano MEP
- Katalogowane przez MIP na tym samym poziomie MD, terminowane przez zdalny MEP w tym samym MA

CFM – protokół Loopback



- Używane do weryfikacji awarii
- MEP może transmitować unicastowe LBM do MEP lub MIP w tym samym MA
- MP otrzymujący komunikat odpowiada unicastowym LBR do źródłowego MEPa

CFM – protokół Linktrace



- Odkrywanie ścieżek i izolacja awarii
- MEP wysyła wiadomość multicastową (LTM) w celu wykrycia MP i ścieżki do MIP lub MEP w tym samym MA
- Każdy MIP na trasie pakietu i terminujący MP zwracają unicastowo wiadomość LTR do źródłowego MEPa

ITU Y.1731 – rozszerzenia do 802.1ag

- **ETH-LB:** Ethernet Loopback—unicast i multicast
- **ETH-AIS:** Alarm Indication Signal
- **ETH-Test:** Test Signal

Test przepustowości, rekolejkowania ramek, błędów bitowych itp
Dwu lub jednokierunkowy

- **ETH-USR:** Ethernet Maintenance Channel

Zdalne zarządzanie

Przykład zastosowania: sprawdzenie VC ATMowego na DSLAMie ethernetowym na użytek ethernetowego BNG

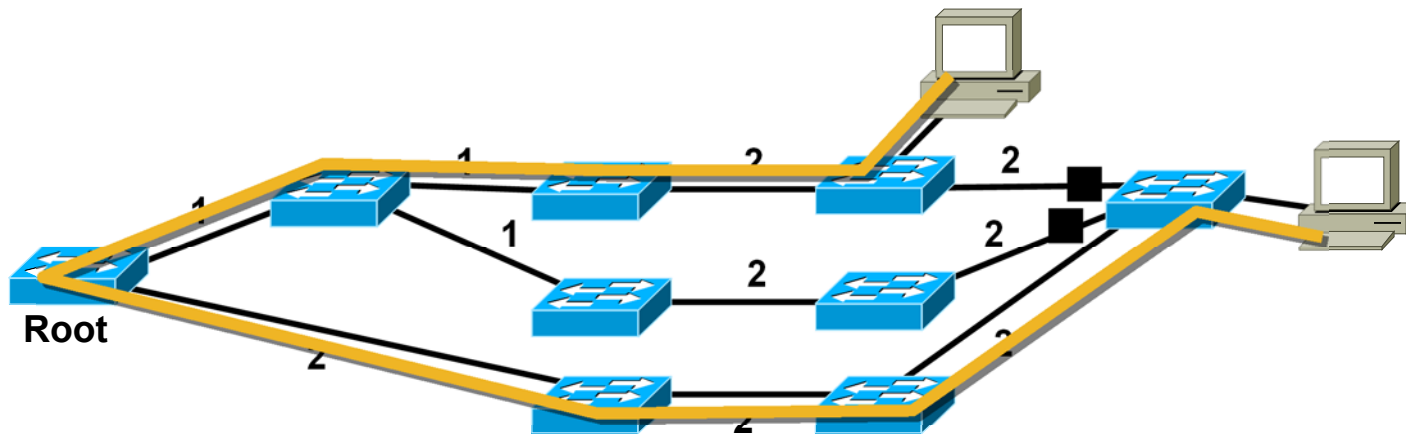
- **ETH-APS:** Ethernet Automated Protection Switching

G.8031 (wykorzystuje ETH-AIS) ; G.8032

Nowości w L2

TRILL i 802.1aq

„W poszukiwaniu nowego protokołu L2”



- Obecne Spanning Tree

 - Potencjalnie nieoptymalne ścieżki przekazywania ruchu

 - Nie można wykorzystać ścieżek równoległych

 - Problemy życia codziennego w sieciach

- Propozycje

 - IETF TRILL

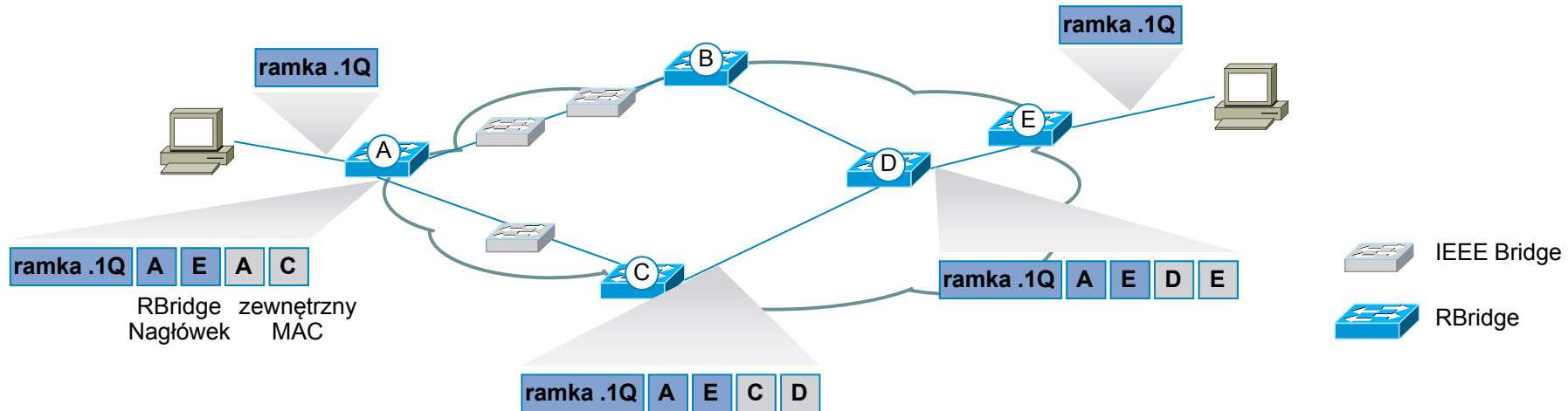
 - Shortest Path Bridging 802.1aq

 - Oba wykorzystują rozszerzenia IS-IS → grupa robocza ISIS w IETF zdefiniuje jeden zestaw rozszerzeń dla obu protokołów

TRILL – podejście IETF

- TRILL (TRansparent Interconnect of Lots of Links)
nazywane również Routing Bridges lub po prostu Rbridges
<http://www.ietf.org/html.charters/trill-charter.html>
- Główne obszary adresowane przez TRILL:
 - Wybór najkrótszej ścieżki
 - Zapewnienie tras równoległych
 - „Plug’n’Play”
- RBridges pracują „na” sieci 802.1 – model warstwowy
 - Sieć 802.1 może być wykorzystywana przez hosty by dostać się do RBridge
 - RBridge może wykorzystać sieć 802.1 do przenoszenia ruchu pomiędzy sobą
 - RBridges nie uczestniczą w procesie xSTP i odrzucają BPDU gdy je otrzymają

TRILL – jak to działa?



- Ramka otrzymuje adres Rbridge wyjściowego, a następnie dodatkowy adres Rbridge next-hop
 - Trochę jak „MAC-in-MAC” – ale pola różnią się od 802.1ah
- RBridges uczą się adresów MAC obecnych na portach brzegowych i **mogą** rozgłaszać je przez IS-IS do innych RBridge’y
 - Wybór między uczeniem się zdalnych mapować w data lub control plane
- Nieznane ramki unicastowe są rozlewane zgodnie z drzewem którego korzeń znajduje się na wejściowym RBridge

TRILL—Ethernet Data Encapsulation

Outer Ethernet Header (link specific):

Outer Destination MAC Address (RB2)	
Outer Destination MAC Address	Outer Source MAC Address
Outer Source MAC Address (RB1)	
Ethertype = IEEE 802.1Q	Outer.VLAN Tag Information

TRILL Header:

Ethertype = TRILL	V	R	M	Op-Length	Hop Count
Egress (RB2) Nickname	Ingress (RB1) Nickname				

Inner Ethernet Header:

Inner Destination MAC Address	
Inner Destination MAC Address	Inner Source MAC Address
Inner Source MAC Address	
Ethertype = IEEE 802.1Q	Inner.VLAN Tag Information

- **Outer-VLAN Tag Information:** This is used only if two RBridges communicate across a standard 802.1Q network

- **V:** Version

- **M:** Multi-destination; indicates if the frame is to be delivered to a single or multiple end stations

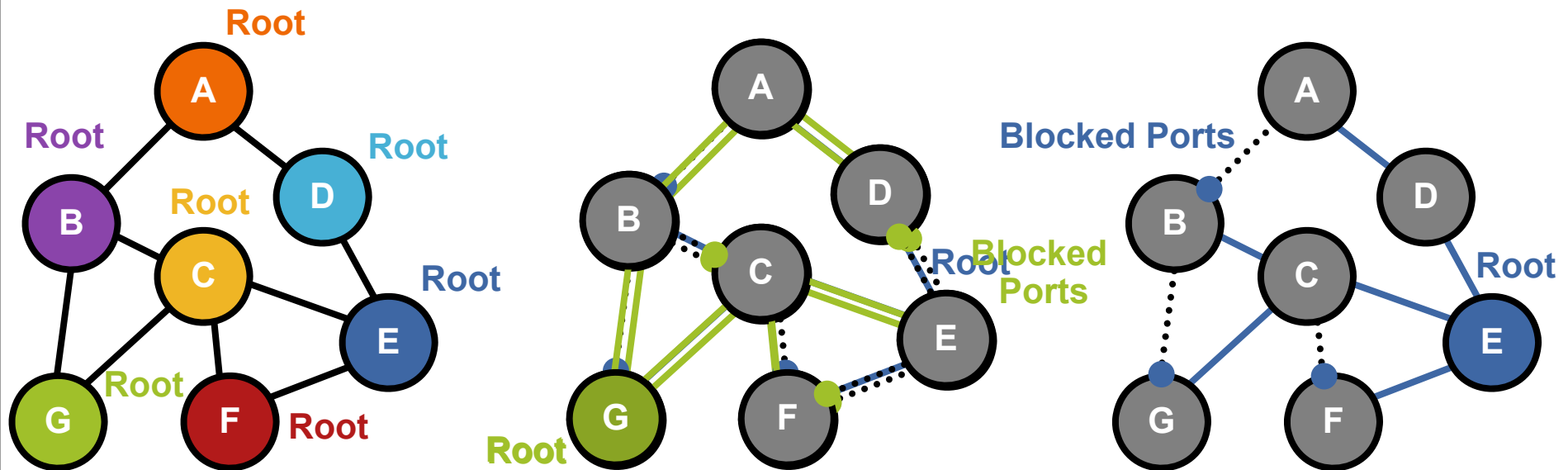
- **Opt-Length:** >0 if an Option field is present

- **Hop Limit:** Similar to TTL

- **RBridge Nickname:** Not the MAC address of the Rbridge, but the a TRILL ID for the RBridge (Egress Nickname used differently if M = 1)

Source: [draft-ietf-trill-rbridge-protocol](#)

802.1aq – najkrótsza ścieżka per bridge



- Każdy most jest korzeniem osobnej instancji najkrótszej ścieżki
- Most G jest korzeniem dla drzewa zielonego
- Most E jest korzeniem dla drzewa błękitnego
- Oba drzewa pracują aktywnie i symetrycznie

Potrzebne w Ethernetie do jednolitej obsługi multicastu i unicastu

Q&A



