Mechanizmy QoS

Absolutnie zupełne podstawy

(dla początkujących)

Łukasz Bromirski
lbromirski@cisco.com

Rafał Szarecki
rafal@juniper.net

PLNOG, Kraków, wrzesień 2011

Zawartość (z grubsza)*

- Co to jest QoS?
- DiffServ i IntServ
- Markowanie
- Policing vs Shaping
- Profile dla PHB
- RED/WRED/MDWRR
- HQoS
- Q&A

* agenda może ulec zmianie bez ostrzeżenia, nawet w trakcie prezentacji

Co to jest QoS?

Co to jest QoS?

- Jakość Usługi – ang. *Quality of Service*
- Zestaw mechanizmów starających się zapewnić **nierówne** traktowanie ruchu sieciowego
- Celem jest zapewnienie krytycznych paramentów ruchowych dla aplikacji, przy optymalnym wykorzystaniu zasobów.
- Myśląc o QoSie mówimy głównie o zarządzaniu KPI:

  Opóźnieniem dla ruchu – (ang. delay)

  Zmiennością opóźnień dla ruchu (ang. jitter)

  Pasmem (ang. bandwidth)

  Stopą utraty pakietów (ang. packet loss/drop)

Po co nam mechanizmy QoS?

- Aplikacje są wrażliwe na opóźnienia, jitter, utratę pakietów. Na szczęście niekonieczne na wszystkie na raz
- Aplikacje wrażliwe na jitter, kompensują go kosztem zwiększenia opóźnienia, poprzez de-jitter buffer po stronie odbiornika
- Przykłady:

	Delay	Jitter	Loss-rate
VoIP	sensitive	sensitive	moderate [<1%]
IPTV	not sensitive*	not sensitive	sensitive [10E-7]
Gaming	sensitive	moderate	moderate
Translational services (e.g. SAP)	moderate	not sensitive	moderate
WEB, e-mail, ftp	not sensitive	not sensitive	not sensitive

* “Not sensitive” – oznacza relatywnie niską wrażliwość - jednak w rozsądnych granicach ☺

Czynniki wpływające na KPI

- Pewne składniki są niekontrolowane (opóźnienie propagacji, przełączania, błędy CRC, itp.)

  Pasmo to nie wszystko: 10Mbps SAT != 10Mbps FO != 10Mbps 4G

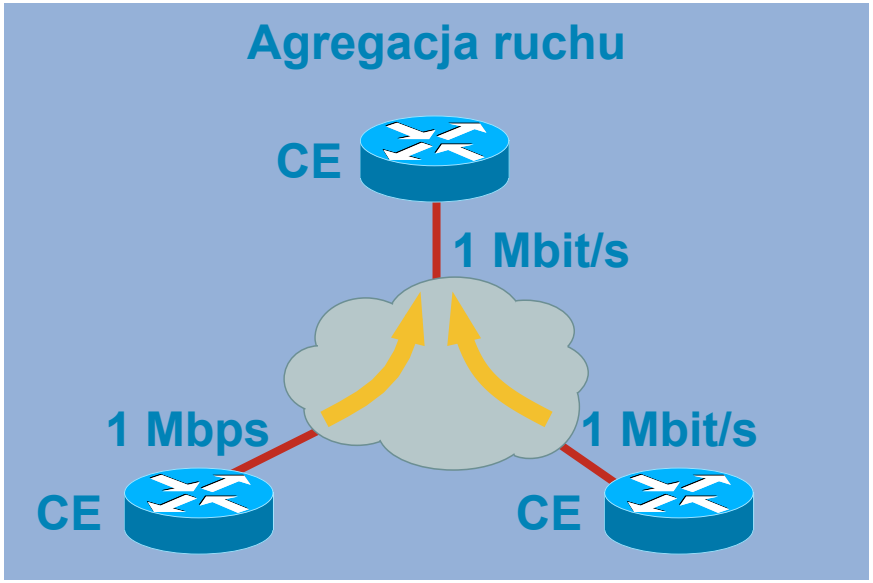
- Inne składniki możemy kontrolować

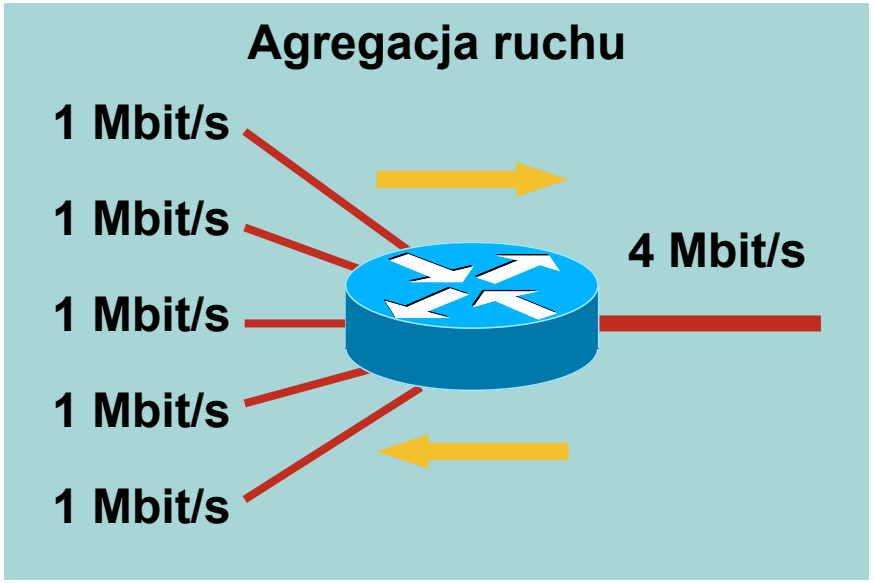
  Długość buforowania

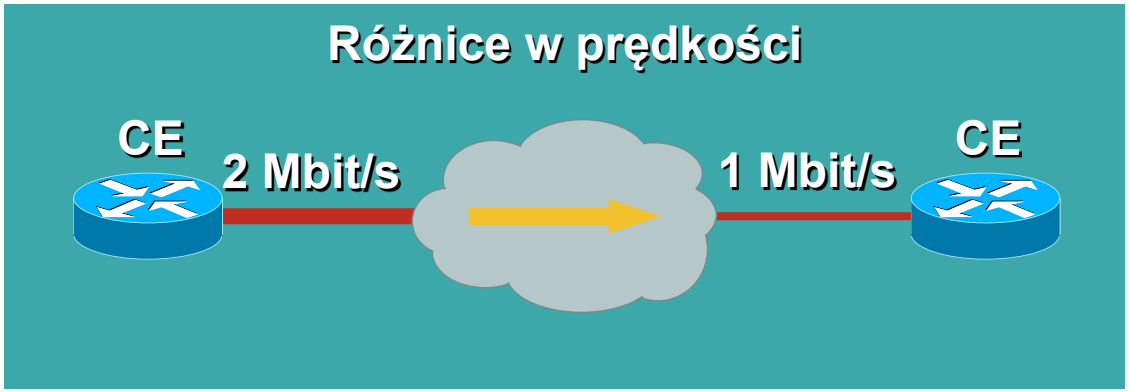
  Kolejność obsługi buforów

- W większości sieci znajdują się miejsca narażone na przeciążenia
- 10Mbit/s -> 100Mbit/s -> 1GE -> 10GE -> 100GE (-> 400GE) ->?

Scenariusze stworzone dla mechanizmów QoS







Do czego dążymy z QoS?

Klasyfikacja i oznaczanie

Kolejkowanie i odrzucanie

Wysyłanie ruchu dalej

DiffServ i IntServ

Architektury QoS

- Architektury QoS opisują sposób w jaki mechanizmy QoS działają w sieciach zapewniając gwarancje usług end-to-end
- Pierwsze architektury IP QoS nie wspominały wprost o SLA end-to-end. Podejście hop-by-hop

  IP Precedence (RFC 791), IP ToS (RFC 1349)
- Dwa modele obecnie używane to:

  Differentiated Services – Diffserv (RFC2475)

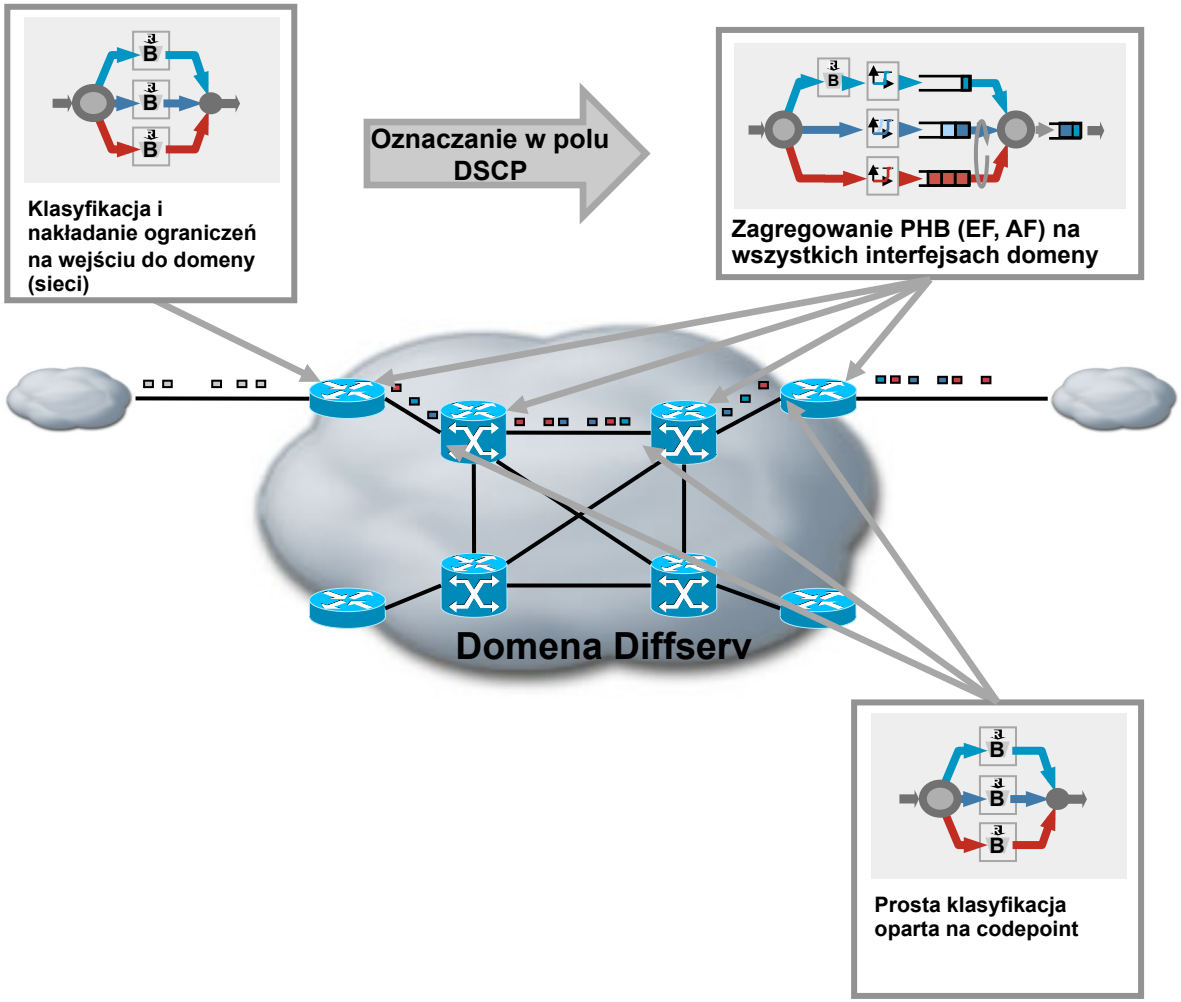
  Integrated Services – Intserv (RFC1633)
- TEORETYCZNIE, w sieciach L3 dostawców usług oba modele mogą współistnieć i uzupełniać się, choć dla „prostych” polityk QoS stosuje się tradycyjnie architekturę Diffserv (mało aplikacji obsługuje RSVP wprost)
- W sieciach lokalnych L2 stosuje się architekturę Diffserv

Architektura IntServ

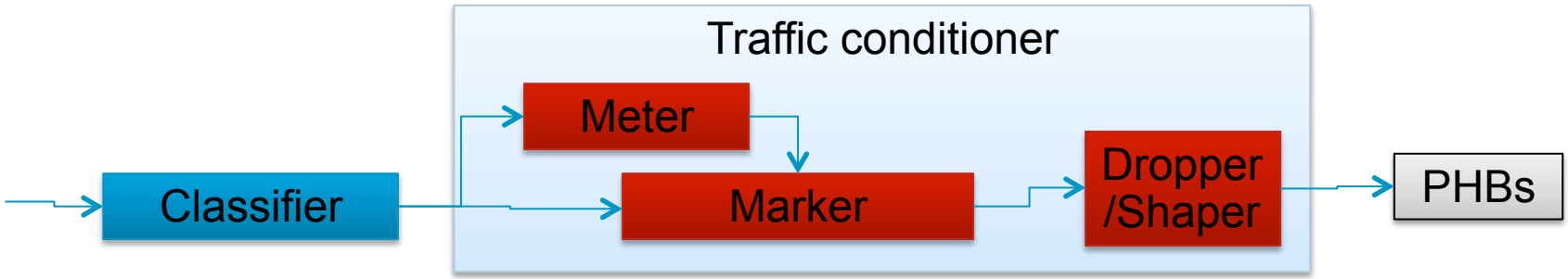
- Aplikacja końcowa żąda od sieci gwarancji zasobów dla indywidualnej sesji sieciowej
  - Wsparcie aplikacji dla RSVP
  - Sesja = Internet 5-typle
- Każdy element sieciowy musi znać stan każdej sesji przechodzącej przez niego. RSVP soft-state.
- Nie stosowane w sieciach SP/Telco
  - Skalowalność – ile sesji TCP mamy w Internecie? Jak je kontrolować?
  - Pełna implementacja – kolejka per sesja. Ten sam problem który zabił ATM
- Elementy IntServ są wykorzystywane w MPLS-TE
  - Aggregacja
  - Gwarancje tylko w control-plane
  - Nie jest to IntServ as per RFC 1633

Architektura Diffserv – RFC2475

- Mały poziom złożoności – brak utrzymania stanów czy sygnalizacji
- Usługi tworzone są przez kombinację:
  - Złożonej klasyfikacji, oznaczania, ewentualnie przycinania ruchu na brzegu sieci
  - Prostej klasyfikacji wewnątrz sieci w oparciu o DSCP
  - Konfiguracji sposobu obsługi ruchu (PHB)



Bloki funkcjonajne DS-węzła (1)



- **Classifier**

  Identyfikuje strumienie pakietów, które powinny być jednakowo traktowane – **Behavior Aggregate**. Pakiety nie mogą zmienić kolejności

  Rodzaje:

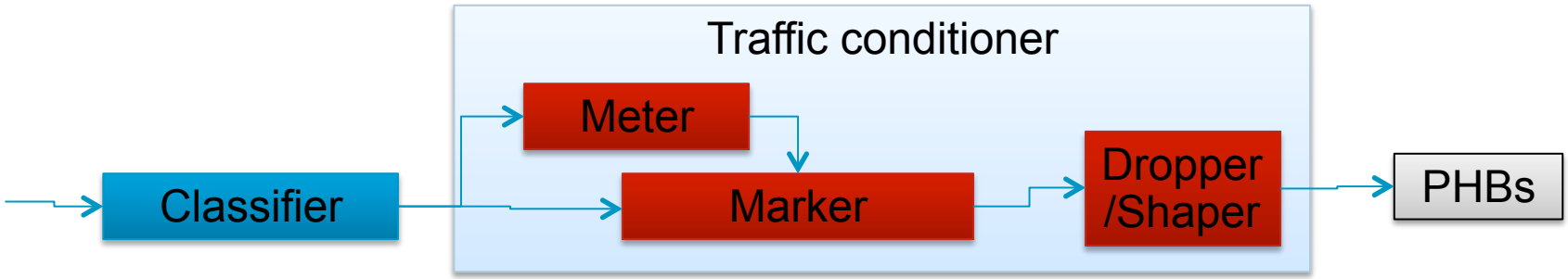
  Behavior Aggregate (**BA**) – sprawdza wyłącznie DSCP

  Multi-field Classifiers (**MF**) – sprawdza wiele pól w nagłówku. Stosowany na wejściowym DS-węźle. Z reguły bezstanowy, ale może być stanowy/aplikacyjny

  Ordered Aggregate (OA) – zbiór BA, dla których wszystkie pakiety muszą być wysłane w tej samej kolejności w jakiej zostały otrzymane

  Dany **BA** może należeć do tylko jednego **OA**

Bloki funkcjonajne DS-węzła (2)



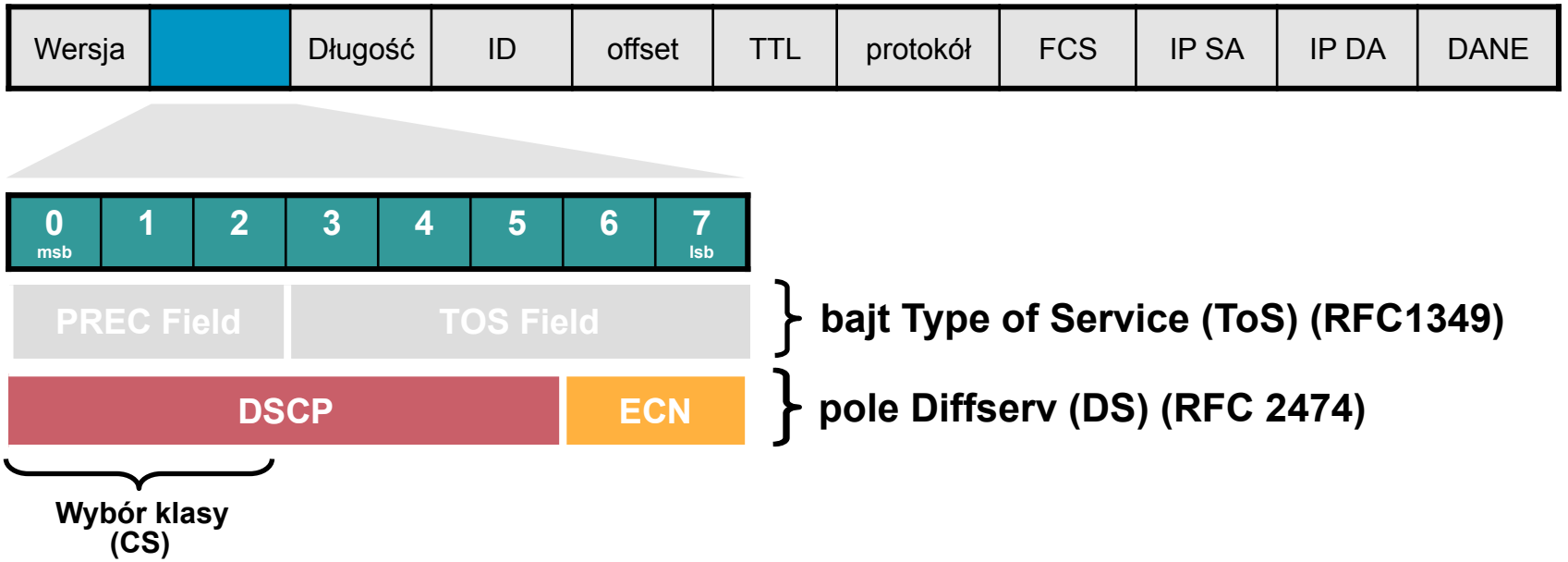
- **Meter** – Mierzy poziom ruchu dla strumienia pakietów wybranych przez classifier (BA). Meter przekazuje informację o stanie in- lub out-of-profile pakietu do innych elementów Traffic Conditioner w celu wykonania operacji
- **Marker** – wpisuje wartość DSCP w odpowiednie pole nagłówka.
- **Dropper**/shaper – odrzuca/buforuje pakiety w taki sposób, aby przyciąć ruch do założonego poziomu
- **PHB**

PHB i PSC

- PHB = Jak zasoby są alokowane dla danego strumienia BA
  - Rozmiar Bufora FIFO/kolejki
  - Pasmo na interfejsie wyjściowym
    - Gwarantowane
    - Maksymalne
    - Proporcja
  - Która kolejka kiedy - scheduler
- PHB Scheduling Class (PSC) = Jak zasoby są alokowane dla danego strumienia OA
  - Zbiór PHB, dla których pakiety nie mogą się “przeskakiwać”
  - Aby zagwarantować kolejność – jeden bufor FIFO (kolejka)

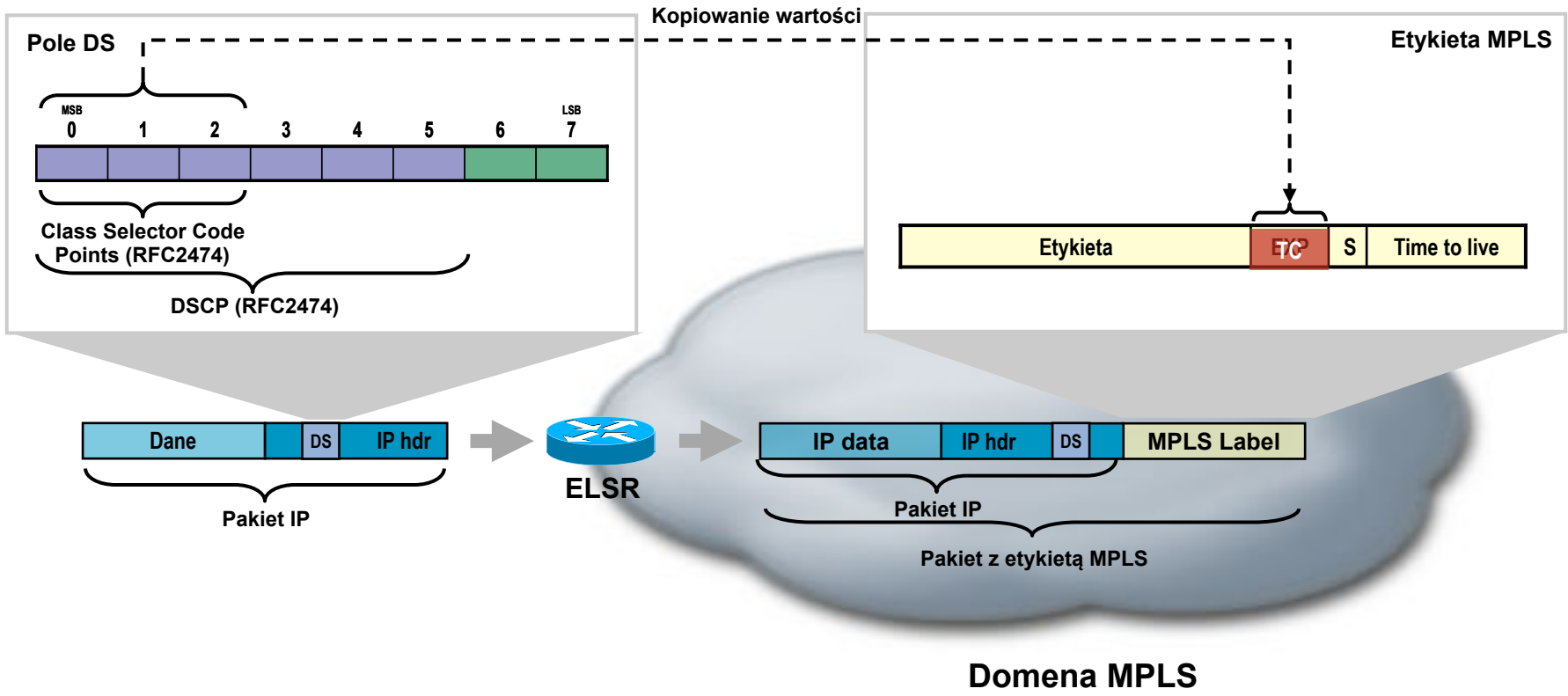
Markowanie

Markowanie – IP DSCP



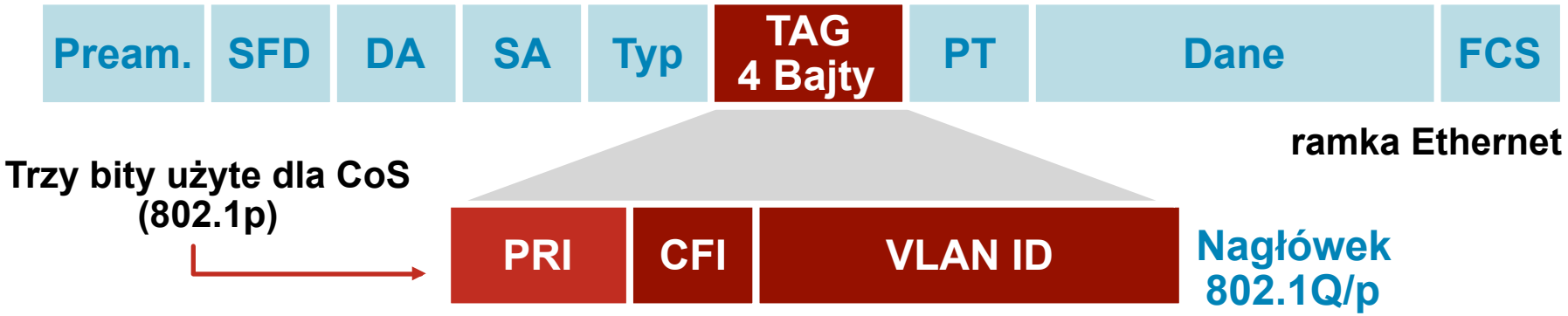
- Pole składa się z 6 bitów - Jedna wartość dla jednego **BA/PHB**
- Pierwsze 3 bity DSCP - Definiują **OA/PSC**
- Ostatnie 3 bity DSCP – Definiują zależność między zajętością kolejki, a priorytetem (prawdopodobieństwem) odrzucenia (drop) pakietu
- RFC 3168 definiuje dodatkowo wykorzystanie dwóch bitów z pola ECN (Explicit Congestion Notification)
- Jeśli ostatnie 3 bity DSCP to 000b i ECN to 00b , to Pierwsze 3 bity są selektorem clasy (CS) – funkcjonalnie odpowiednie i wstecznie kompatybilne z IP Precedence

MPLS i Diffserv



- Do „sygnalizacji” QoS używamy 3-bitowego pola TC w nagłówku pakietu MPLS
  - Kopia lub mapowanie z wartości pola DSCP oryginalnego pakietu IP
- Decyzja o trafieniu pakietu do konkretnej klasy usługowej może odbywać się na podstawie wartości pola TC, lub IP Precedence/DSCP oryginalnego pakietu
- Wsparcie dla architektury DiffServ w MPLS opisano w RFC 3270 (choć bez żadnych propozycji)

Markowanie – CoS – 802.1p



- 802.1p - pole ‘priorytet użytkownika’ zwane również Class of Service (CoS)
- Różnym typom ruchu przypisane są różne wartości CoS.
- Wartości CoS 6 i 7 są zarezerwowane dla ruchu sieciowego
- Formalnie nie związany z DiffServ

CoS	Aplikacja
7	Zarezerwowane
6	Routing
5	Głos
4	Video
3	Sygnalizacja połączeń
2	Dane krytyczne
1	Dane masowe
0	Dane „Best Effort”

IETF DiffServ a implementacje

MF Classifier	Access-list, stateless Firewall Filter, stateful firewall, IDP, etc
Meter	Policer, rate-limit profile, token bucket
Marker	Operacja ustawienia wartości
Dropper	Operacja w Policerze , Operacja w filtrze
PSC	Jedna kolejka i jej parametry
PHB	Kombinacja kolejki i profilu odrzucania pakietów

Elementy:
Policing vs Shaping

Zarządzanie przeciążeniem

- Z obsługą zbyt dużej ilości ruchu związana jest konieczność odrzucania ruchu nadmiarowego
- Jeśli jednak przeciążenie jest krótkookresowe, można obyć się bez strat. Dwa podejścia:

  policing – token bucket

  shaping – buforujemy ruch przez określoną ilość czasu

Policer – Token Bucket

- Żetony do wiaderka są dostarczane w stałym tempie
- Żetony mogą być pobrane z wiaderka w dowolnej ilości (z nieskończoną prędkością)
- Aby przesłać pakiet, dla każdego bajtu, trzeba pobrać odpowiednią ilość żetonów z wiaderka
- Jeśli żetonów brakuje, to pakiet jest odrzucany/oznaczany
- Brak opóźnień
- Po przejściu policera, ruch nadal ma charakter niestały (bursty)
- Po przejściu policera, średnia wartość ruchu odpowiada zadanej prędkości

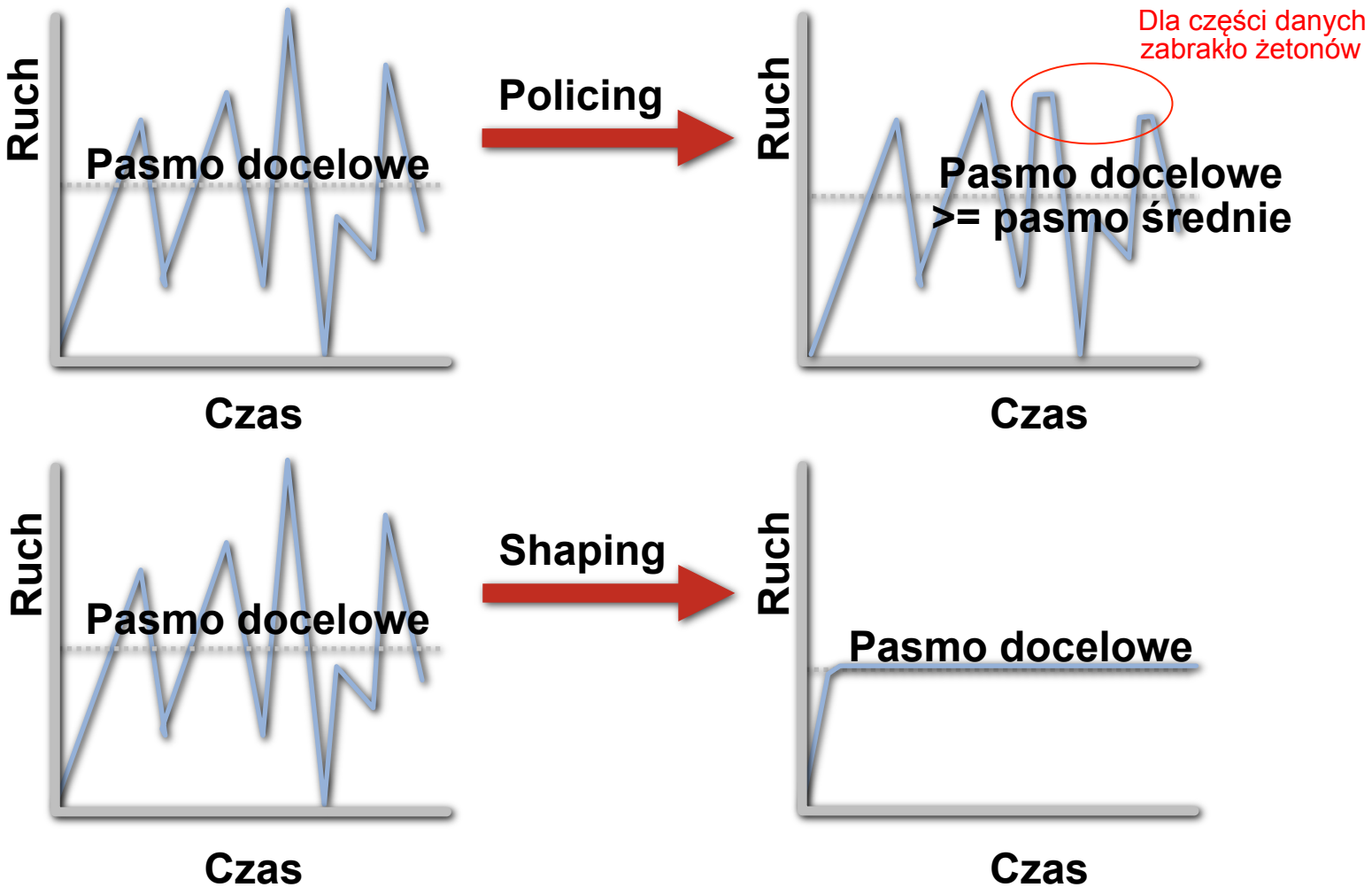
Shaper – Leaky Bucket

- Żetony wyciekają z wiaderka i są dostarczane w stałym tempie
- Aby przesłać pakiet, dla każdego bajtu, trzeba poczekać aż z wiaderka wycieknie dostateczna ilość żetonów
- W oczekiwaniu na żetony pakiety są składowane w buforze pamięci typu FIFO (kolejka)
- Pakiety są opóźniane w buforze
- Jak zabraknie bufora, pakiet trzeba odrzucić

  Head drop – najstarszy pakiet jest wyrzucany, aby zrobić miejsce nowemu

  Tail drop – najnowszy pakiet jest odrzucany, bo się nie mieści
- Po przejściu shapera, ruch nadal ma charakter stały
- Po przejściu shapera, średnia wartość ruchu odpowiada zadanej prędkości

Policing vs shaping



PHB Scheduling Profiles – parametry kolejki

Kolejki

- Kolejka ma konfigurowalne atrybuty:

  Rozmiar – wielkość pamięci, wyrażana w Bajtach, milisekundach, %, pakietach

  Kolejka ma priorytet

  Kolejka ma przypisane pasmo

    CIR (udział w paśmie interfejsu). Minimalne gwarantowane pasmo dla strumienia **OA** wyrażone w % lub kbps

    PIR Maksymalne pasmo dla strumienia **OA** wyrażone w % lub kbps

    Wagę lub udział w paśmie interfejsu – relatywny udział względem innych kolejek

- Dynamiczny stan kolejki – pozytywne/negatywne saldo
- Na interfejsie jest wiele kolejek – typowo 4 lub 8
- Scheduler – process decydujący która kolejka bedzie obsługiwana

Scheduler

- Wiele algorytmów
  - WFQ
  - WRR
  - PQ
  - …
- **MDWRR** Najbardziej popularny na urządzeniach operatorskich i high-end
  - wszystkie routery Junipera
  - CRS, ASR, 7600

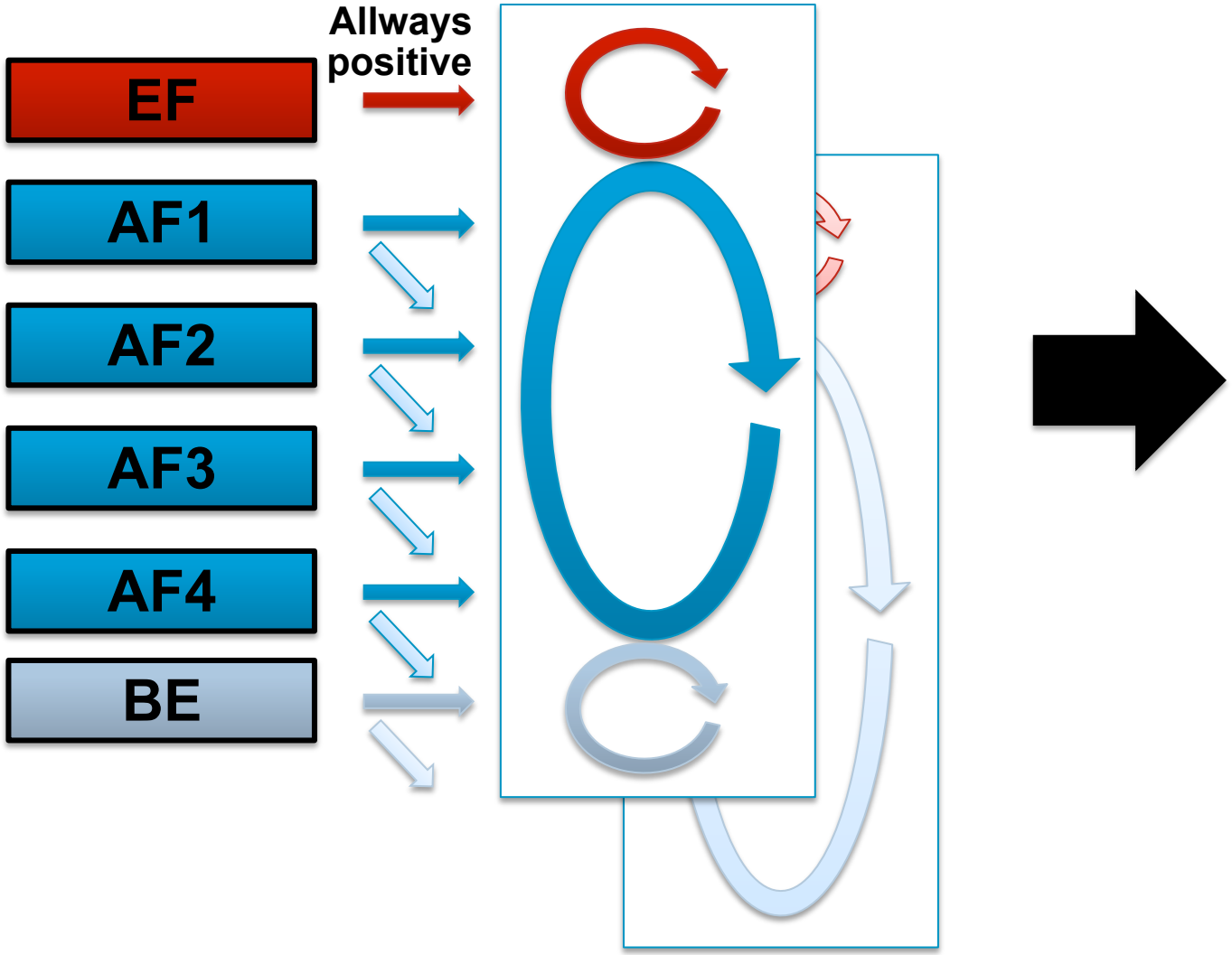
MDWRR (1)

- 2 token bucket dla każdej kolejki
  - Pierwszy – napełniany w tempie odpowiadającym CIR
  - Drugi – napełniany w tempie odpowiadającym PIR
- Jeśli interfejs nie jest przeciążony, ruch z kolejki może być nadawany szybciej niż CIR
  - Ujemne żetony w pierwszym TB (Deficyt)
  - Kolejka jest w stanie negatywnego salda

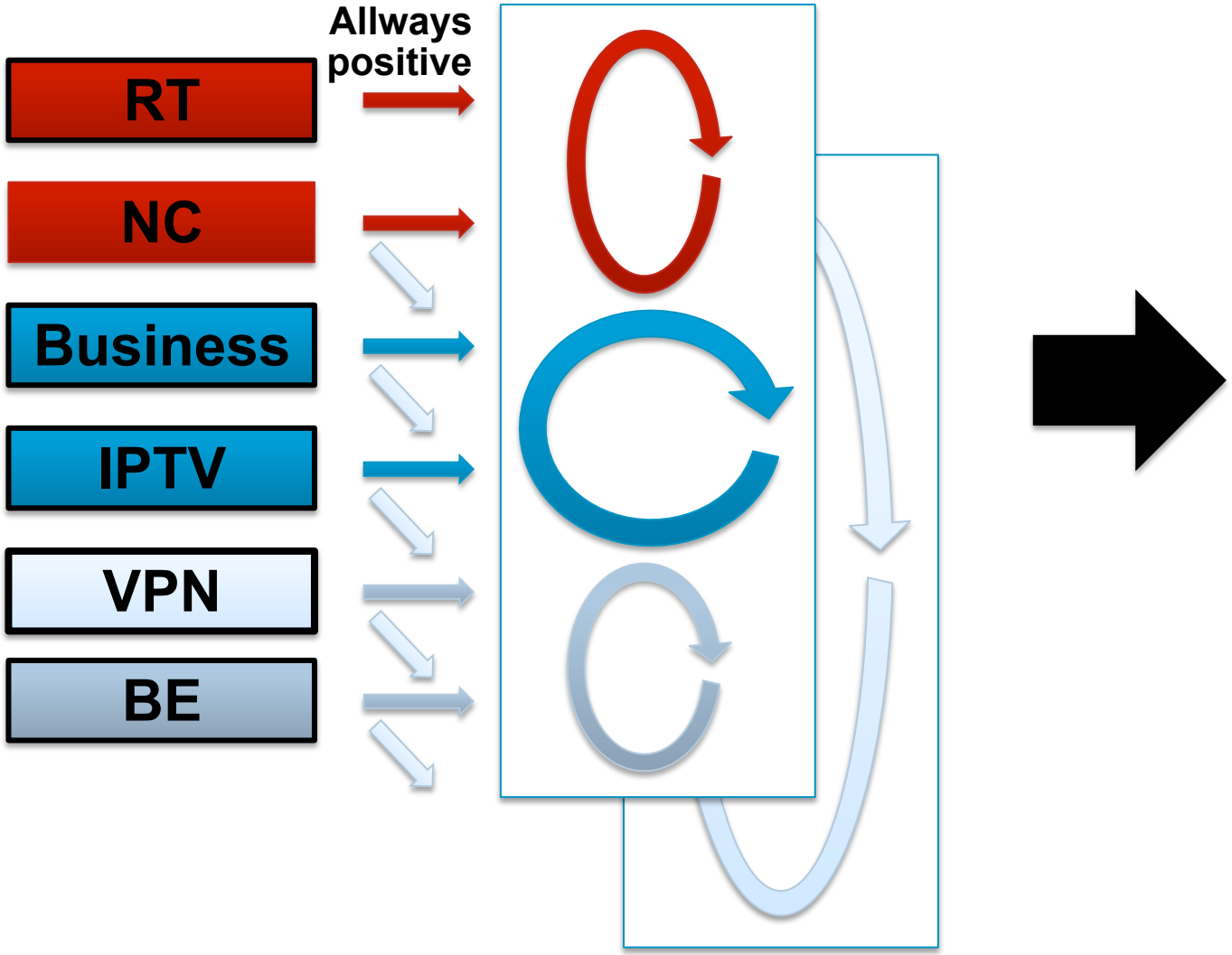
MDWRR (2) - przykłady

1. Obsługa kolejek, które mają **wysoki** priorytet, i równocześnie **pozytywne** saldo. WRR – wagą jest CIR/waga.
   - Strict-High priority – kolejka o **wysokim** priorytecie, która **zawsze** ma **pozytywne** saldo
2. Obsługa kolejek, które mają **średni** priorytet, i równocześnie **pozytywne** saldo. WRR – wagą jest CIR/waga
3. Obsługa kolejek, które mają **niski** priorytet, i równocześnie **pozytywne** saldo. WRR – wagą jest CIR/waga
4. Obsługa kolejek, które mają **wysoki** priorytet, i równocześnie **negatywne** saldo. WRR – wagą jest CIR/waga
5. Obsługa kolejek, które mają **średni** lub **niski** priorytet, i równocześnie **negatywne** saldo. WRR – wagą jest CIR/waga

MDWRR a standardowe PSC

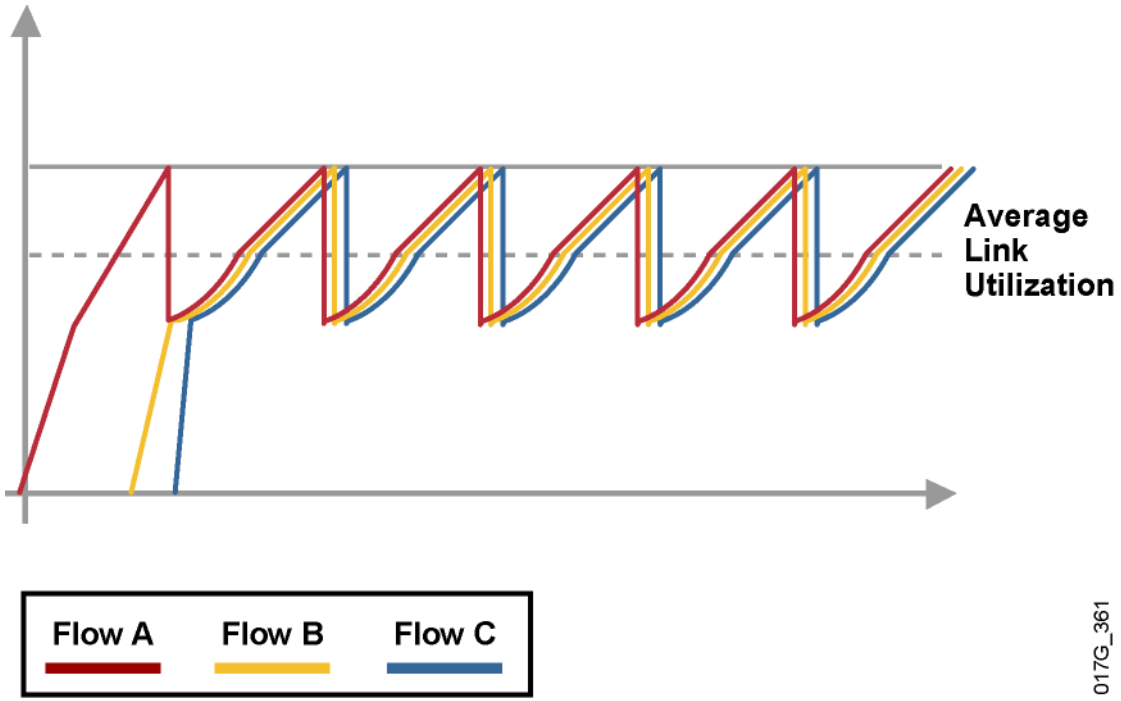


MDWRR – niestandardowe PSC



Random Early Discard

Policing lub przeciążenie – efekt synchronizacji TCP



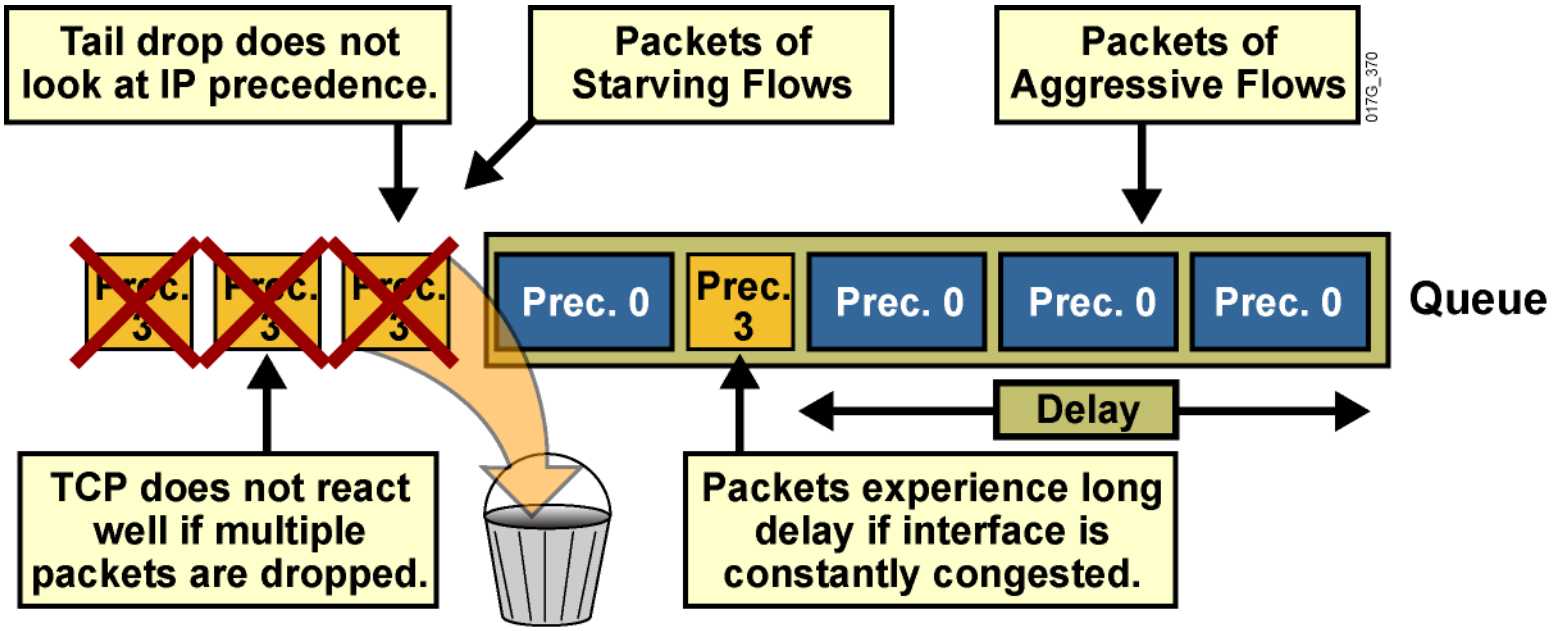
Wiele sesji TCP startuje w różnym czasie

Rozmiary okien TCP zwiększają się

Bufory/kolejki wypełniają się

Odrzucanie ruchu powoduje zmniejszenie okna TCP w tym samym momencie – a następnie ponowne zwiększanie go...

Opóźnienia, jitter i „umieranie” pakietów



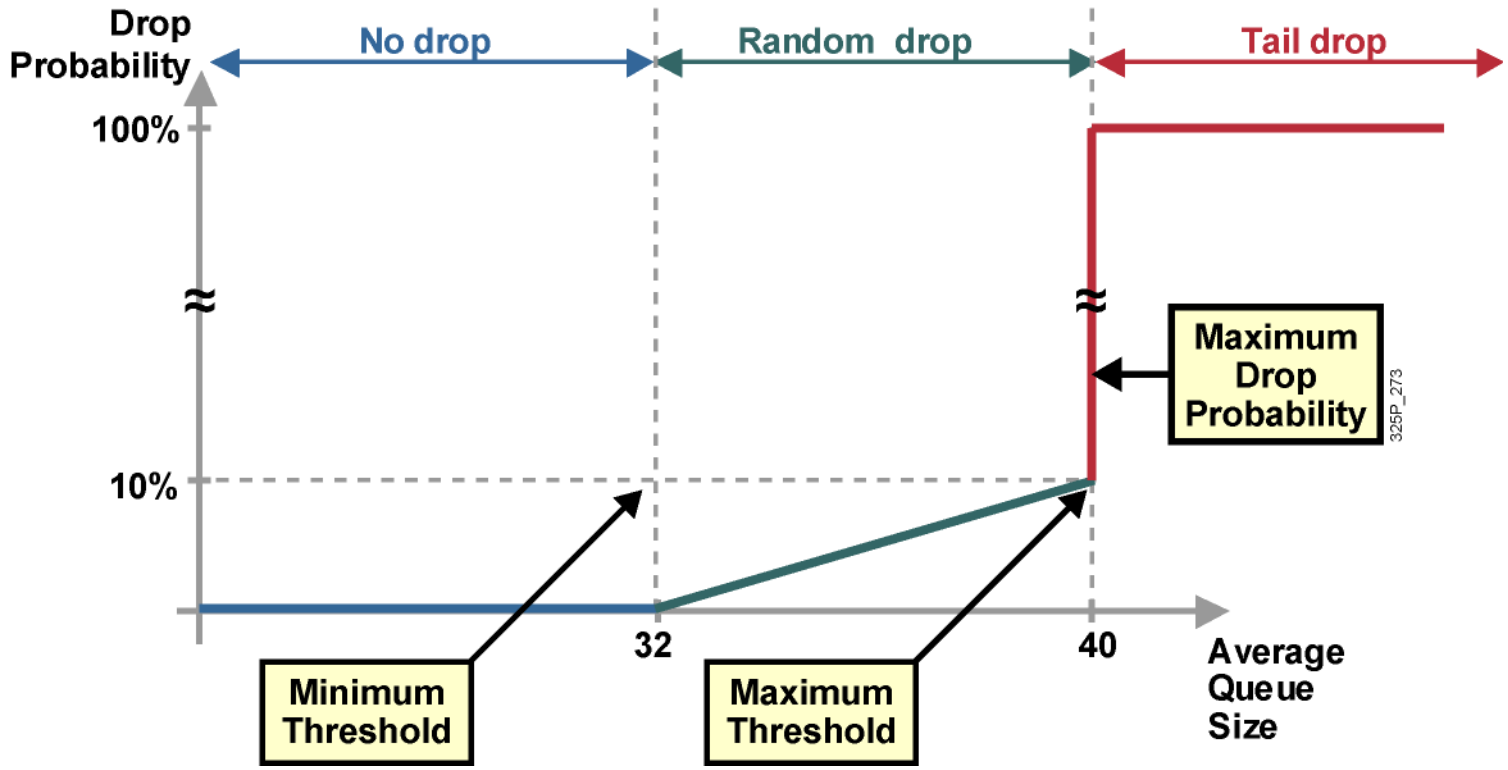
Stałe ciągłe wykorzystanie buforów powoduje opóźnienia

Zmieniające się obciążenie buforów może powodować jitter

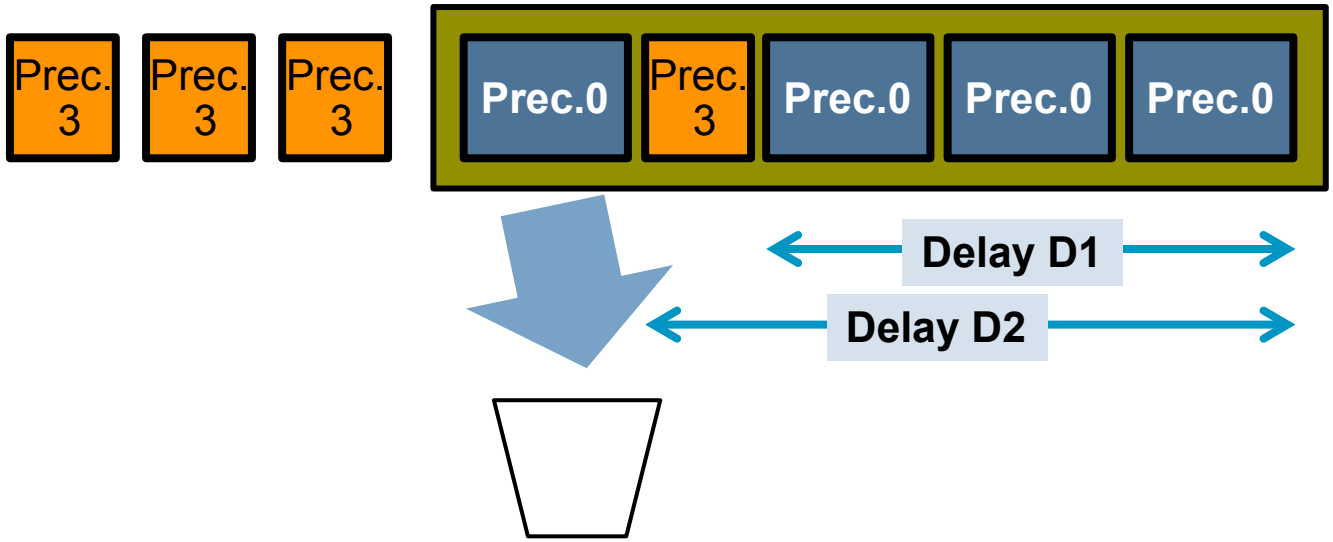
Bardziej agresywne aplikacje generujące ruch mogą powodować „umieranie” bardziej spokojnych aplikacji

Mechanizm RED

Random Early Detection

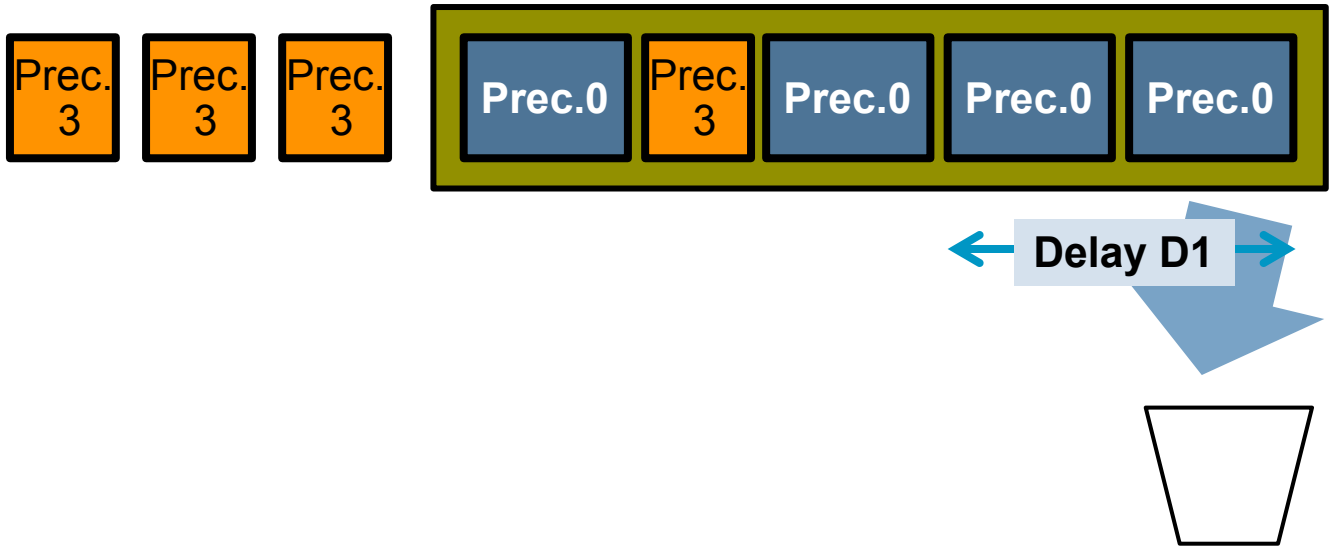


Tail RED



- Nie wpływa na opóźnienie
- Preferuje dostarczenie pakietów pomarańczowych (np. ruch kontrolny sieci + klientów sieci)
- Pakiety niebieskie to TCP - informacja o przeciążeniu w postaci brakujących pakietów jest dostarczana z opóźnieniem D2

Head RED

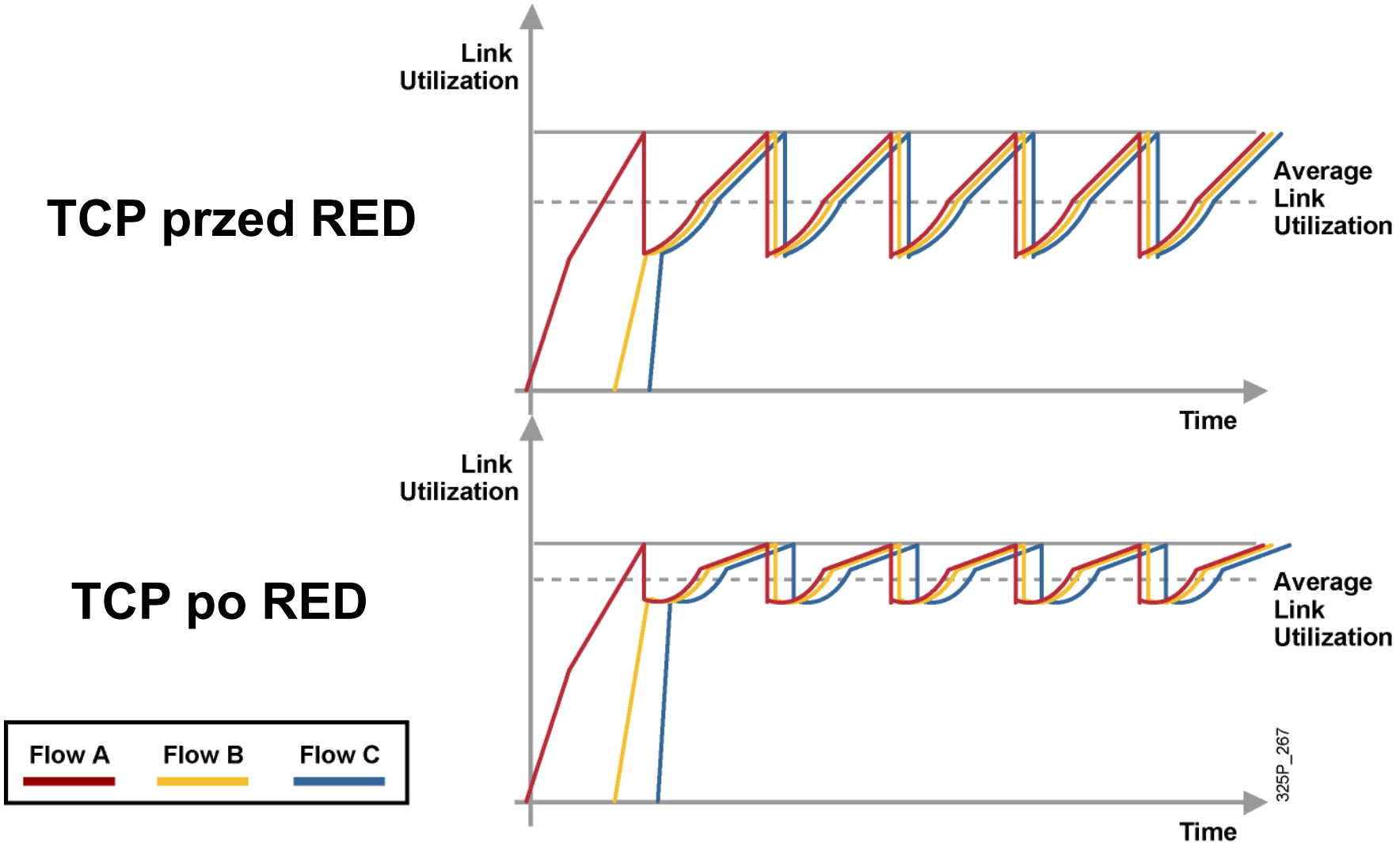


- Zmniejsza opóźnienie pakietów pomarańczowych
- Preferuje dostarczenie pakietów pomarańczowych
- Pakiety niebieskie to TCP - informacja o przeciążeniu w postaci brakujących pakietów jest dostarczana bez opóźnienia

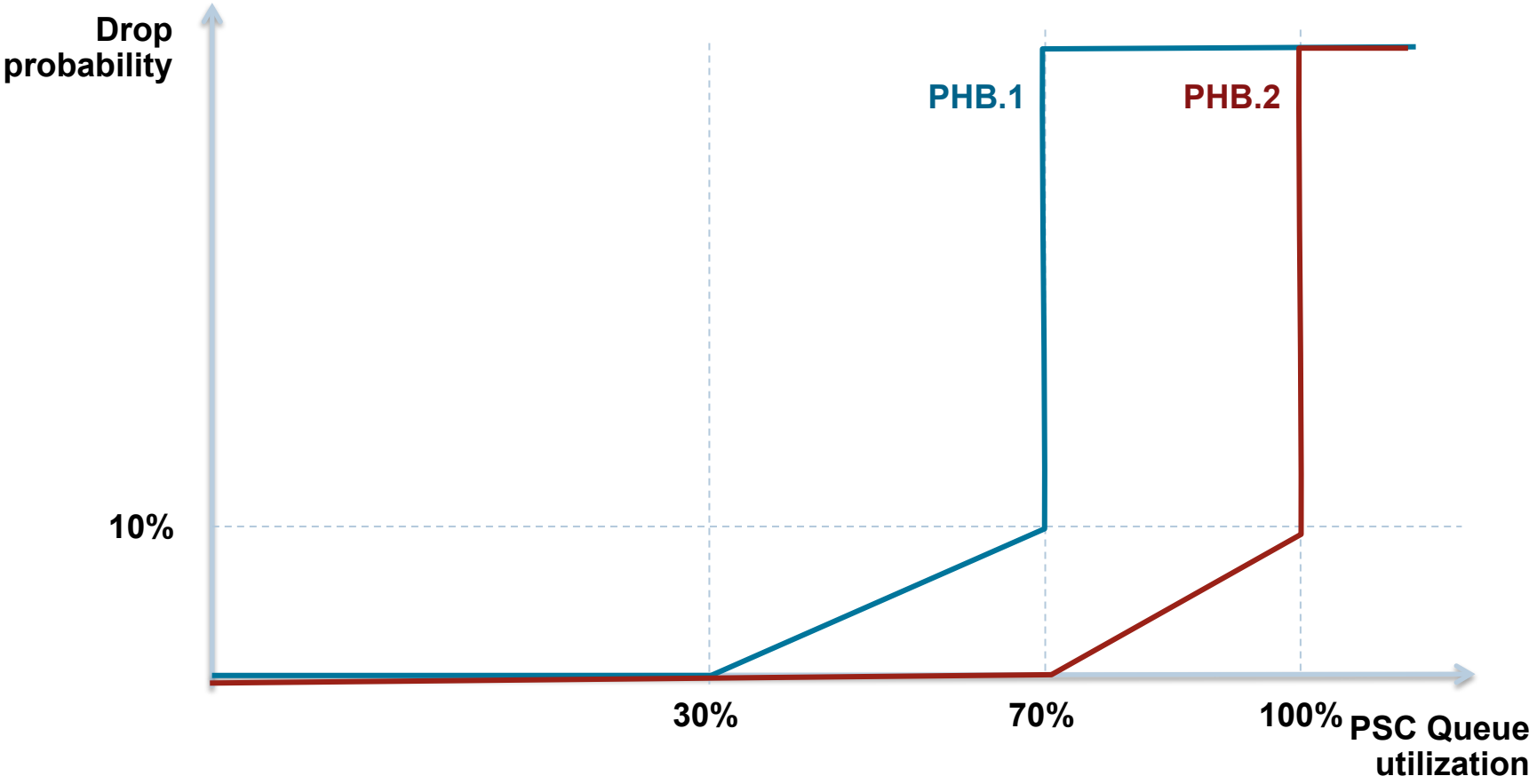
Ruch TCP i RED

TCP przed RED

TCP po RED



Mechanizm RED a PHB



KPI vs mechanizmy QoS

Wymagania aplikacji a możliwosci konfiguracji

KPI

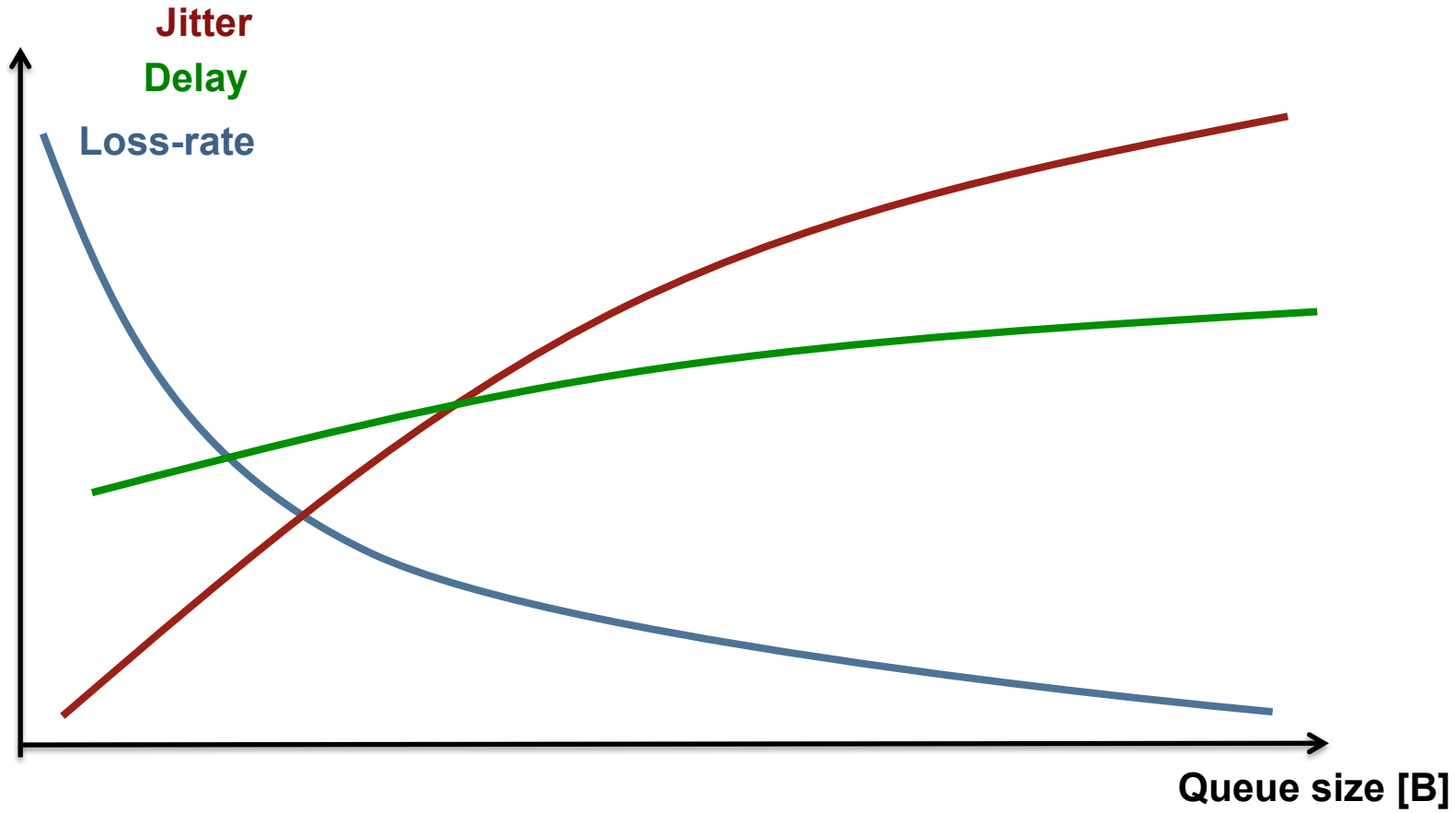
- Delay
- Jitter
- Loss-rate

Konfiguracja

- Priorytet kolejki
- Rozmiar kolejki
- CIR/PIR/Waga kolejki
- Profil RED

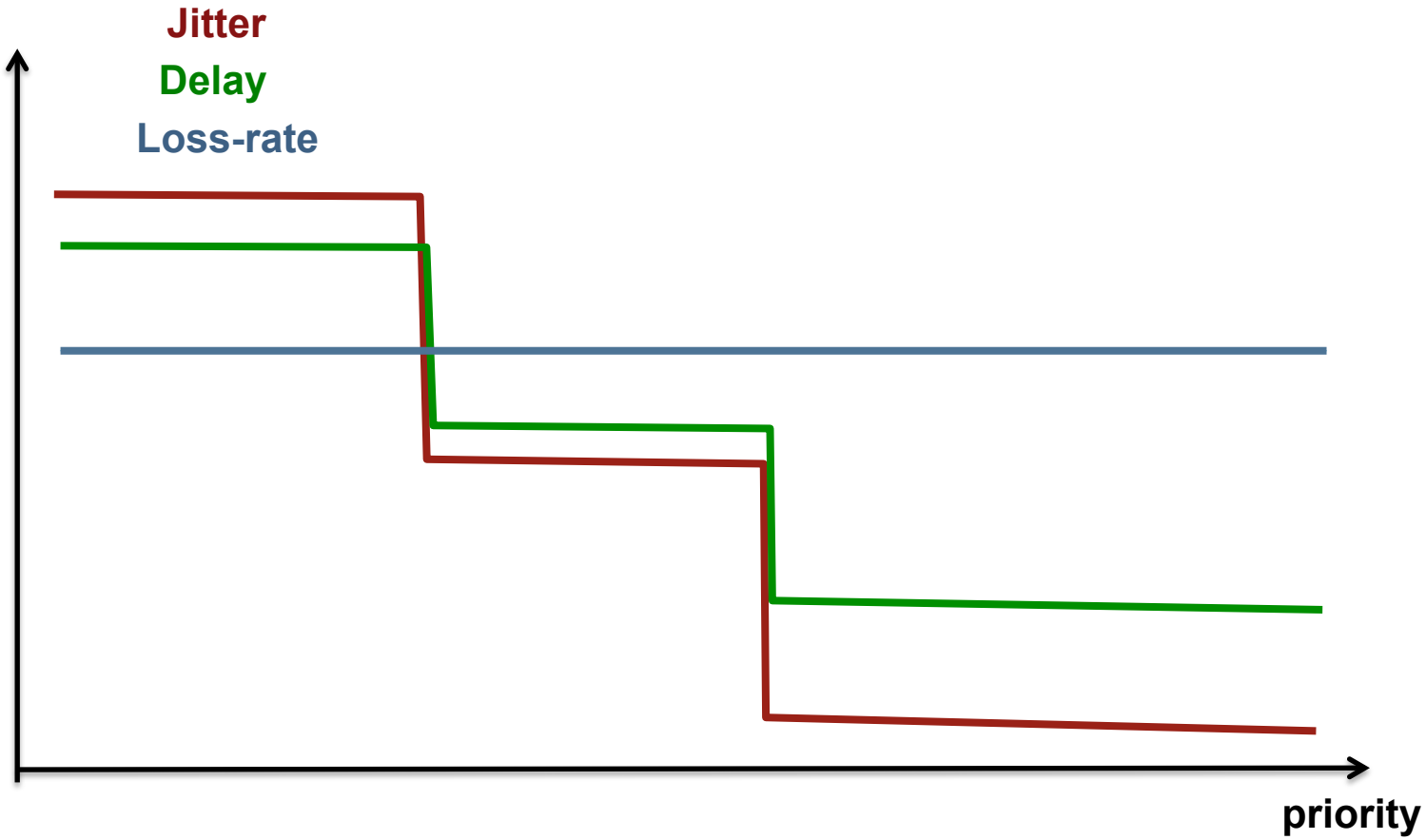
- Upgrade łącza ☺

Zależności – rozmiar kolejki

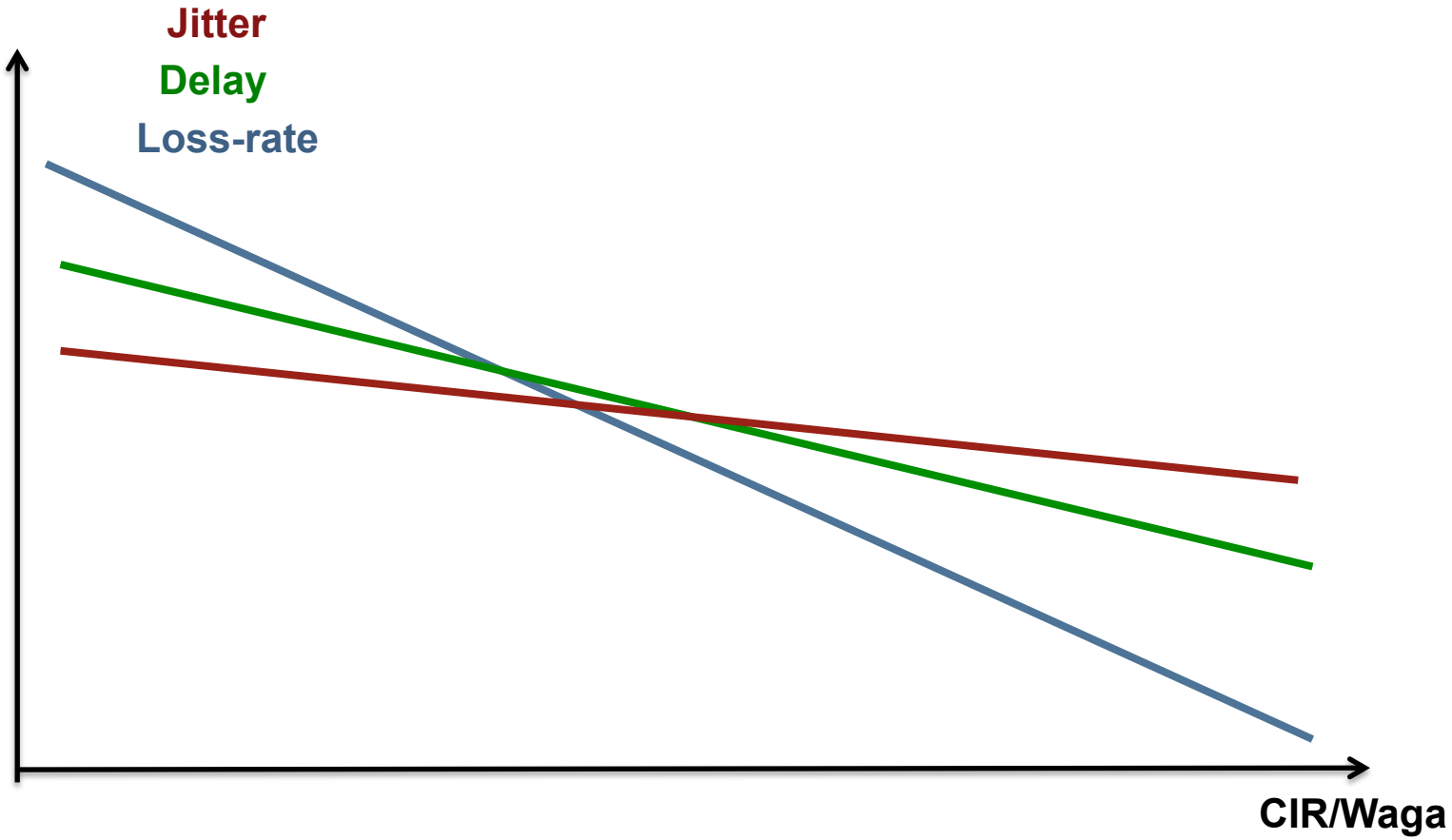


- Ups... Albo małe starty, albo mały jitter…

Zależności – priorytet kolejki



Zależności – CIR/Waga kolejki



- Zwiększenie CIR/Wagi – zawsze kosztem innej kolejki

Ale tak praktycznie…

- CIR/PIR/Waga – wynika z przewidywanego ruchu
- RED profile – max. drop probability 10%-20%. Powyżej i tak wszystkie okna TCP są resetowane
- Rozmiar kolejki

  Zależy od tego jaki jitter jest akceptowalny

  Dla TCP – przepływność * opóźnienie propagacji. (LFN – Elephant) dla optymalizacji wypełnienia łacza

HQoS

Hierarchiczny QoS – gdzie i w jakim celu

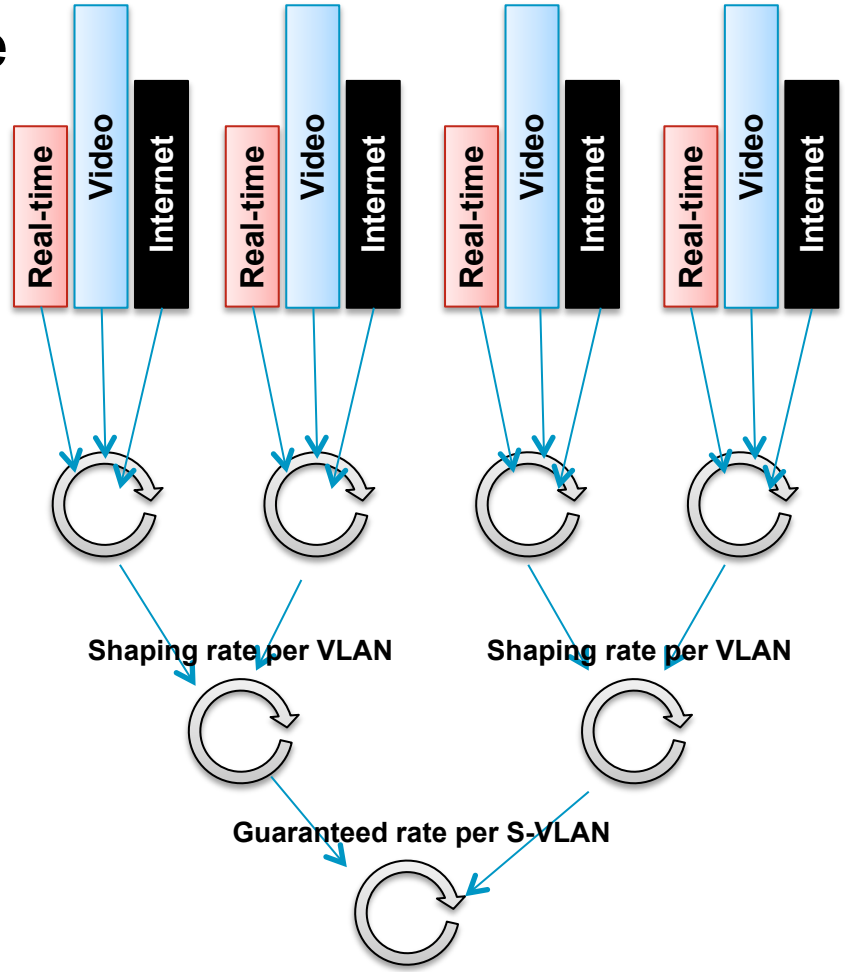
- Spotykany w środowisku masowego klienta – BNG
- Wielu abonentów, każdy z kilkoma klasami QoS
- Trzeba zapewnić każdemu abonentowi minimalne zasoby, niezależnie of klasy ruchu

  Abonent musi mieć zapewnioną możliwość realizacji kontraktu (20Mbps) w klasie BE, nawet jeśli inni abonenci są przeciążają łącze ruchem EF

- Potrzebujemy zapewnić minimalne pasmo dla każdego z urządzeń dostepowych, nawet jeśli są połączone w szeregu, tzw. łańcuszek (ang. daisy chain)

HQoS

- Outer VLAN = Access Node
  - CIR
- Inner VLAN = subscriber
  - waga
  - PIR
- Per VLAN queues
  - Real-time – CIR=PIR, S-HP, mały bufor
  - Video – CIR, MP, wielki bufor
  - Internet – CIR, LP, średni bufor, RED



Pytania?


ODERWIJ SOBIE
KARTECZKĘ!
SĄ ZA DARMO!
DZIĘKUJĘ! ☺
DZIĘKUJĘ! ☺
DZIĘKUJĘ! ☺
DZIĘKUJĘ! ☺


PLNOG