Overview of QoS in Packet-based IP and MPLS Networks

Paresh Shah

Utpal Mukhopadhyaya

Arun Sathiamurthi

Agenda

- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- MPLS QoS
- Summary

- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- MPLS QoS
- Summary

What is Quality of Service?

QoS represents the set of techniques necessary to manage network bandwidth, delay, jitter, and packet loss.

From a business perspective, it is essential to assure that the critical applications are guaranteed the network resources they need, despite varying network traffic load.

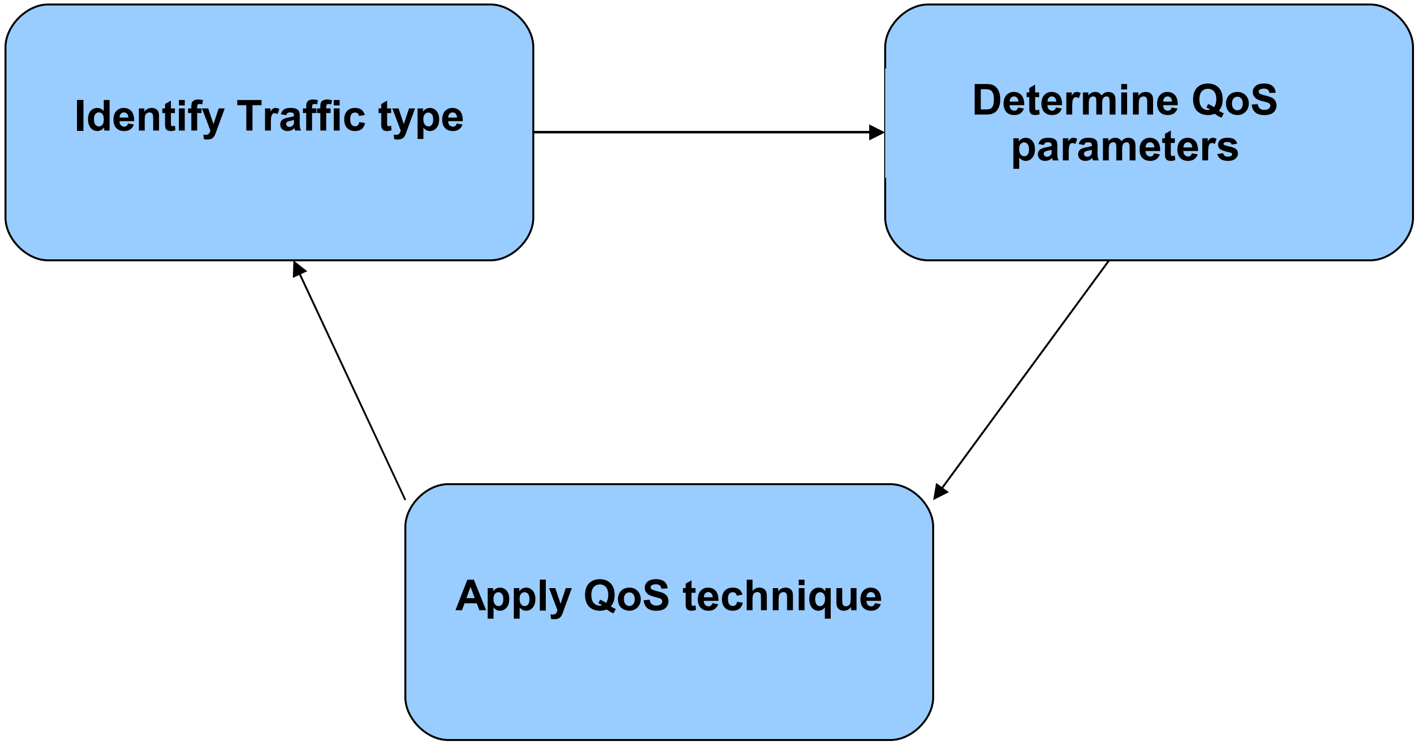
Traffic Characterization

- **Identify traffic sources and types**
- **Need for appropriate handling**
  - **Realtime and Non-realtime**
    - **Voice (Delay sensitive)**
    - **Video (Bandwidth intensive)**
    - **Data (Loss sensitive)**
      - **HTTP, FTP, SMTP**
  - **Bursty and Constant type**
  - **Multi-service traffic: IP, MPLS**
  - **Single or Multiple flows of the same type**

QoS Requirements

- Traffic influencing parameters
  - Latency, Jitter, Loss
- Management of finite resources
  - Rate Control
  - Queuing and Scheduling
  - Congestion Management
  - Admission Control
  - Routing Control Traffic protection
- Service Level Agreement (SLA)
  - per-flow
  - aggregated

QoS Triangle



QoS Approaches

- Fine-grained approach
    - flow-based (individual flows)
- Coarse-grained approach
    - aggregated (large number of flows)
- Leads to two different QoS Models

- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- MPLS and QoS
- Summary

QoS Service Models

- **Best effort (No QoS)**
- **Integrated services (Hard QoS)**
- **Differentiated services (Soft QoS)**

Best Effort Model – Traditional Internet

- “We’ll do the best we can”

  But messages may be lost en route

- Traditional datagram model
- Not a traditional telephone company model

  Pay for what you want, and get exactly that

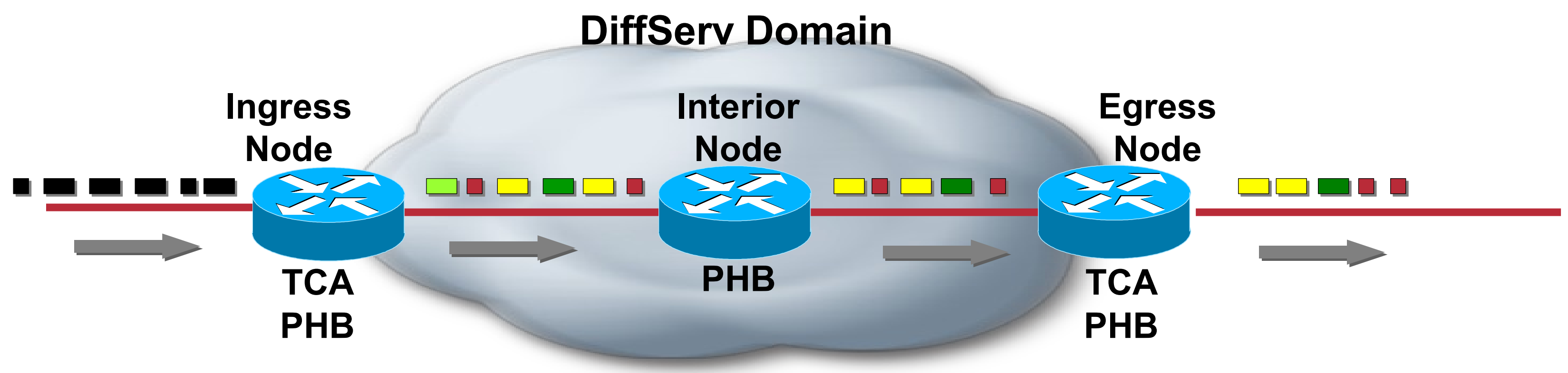
Integrated Services Model

- IntServ Architecture (RFC 1633)
- Hard QoS
- Guarantees per-flow QoS
- Strict Bandwidth Reservations
- Needs Signaling to accomplish Path Reservation
  - Resource Reservation Protocol RSVP (RFC 2205)
  - PATH/RESV messages
- Admission Control
- Must be configured on every router along the path
- Works well on small-scale
  - Has issues with scaling with large number of flows
  - Requires devices to retain state information

Differentiated Services Model

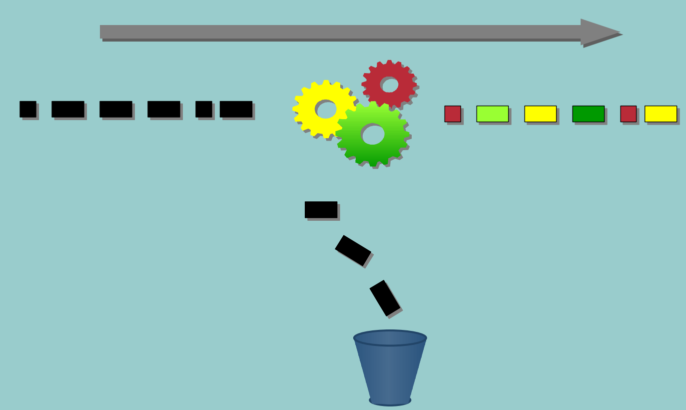
- DiffServ Architecture – RFC 2475
- Scales well with large flows through aggregation
- Creates a means for traffic conditioning (TC)
- Defines per-hop behavior (PHB)
- Edge nodes perform TC
  - Allows core routers to do more important processing tasks
- Tough to predict end-to-end behavior
  - Especially with multiple DiffServ Domains
  - DiffServ implementation versus Capacity planning

Differentiated Services Architecture



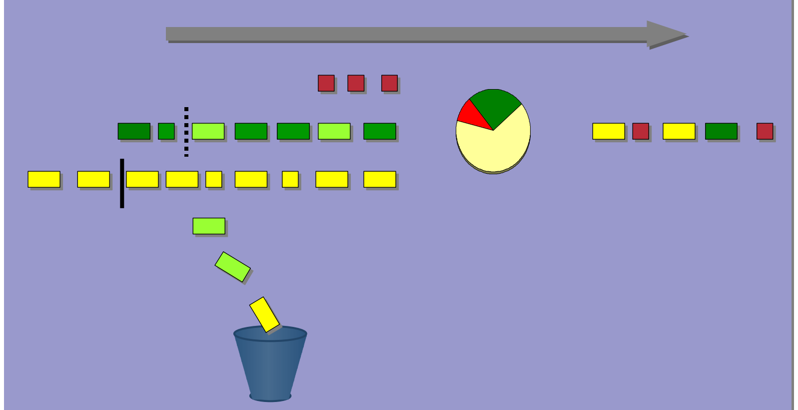
Traffic Conditioning Agreement (TCA)

Classification/Marking/Policing/Shaping



Per-Hop Behavior (PHB)

Queuing/Dropping

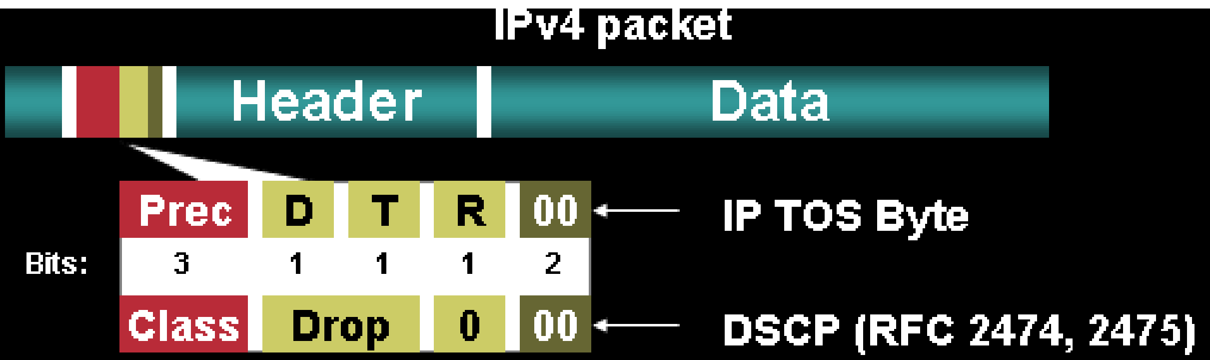


- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- MPLS and QoS
- Summary

IETF DiffServ Model

- Re-define TOS byte in IP header to Differentiated Services Code Point (DSCP)
- Uses 6 bits to categorize traffic into “Behavior Aggregates”
- Defines a number of “Per Hop Behaviors” applied to links
- Two-Ingredient Recipe:
    - Condition the Traffic at the Edges
    - Invoke the PHBs in the Core

IP TOS vs IP DSCP



Prec	Original Use (IP Prec)	DSCP Class
000	Routine	Best Effort
001	Priority	AF Class 1
010	Immediate	AF Class 2
011	Flash	AF Class 3
100	Flash Override	AF Class 4
101	Critical	EF
110	Inter-network Control	Inter-Network Control
111	Network Control	Network Control

DTR	Original Use (Delay, Throughput, Reliability)	DSCP Drop Probability
000	Normal, Normal, Normal	Class Selector
001	Normal, Normal, High	Reserved
010	Normal, Normal, Normal	Low
011	Normal, High, High	Reserved
100	Low, Normal, Normal	Medium
101	Low, Normal, High	Reserved
110	Low, High, Normal	High
111	Low, High, High	Reserved

Diffserv Class Selector

- Class Selector provides support for IP Prece using DSCP terminology

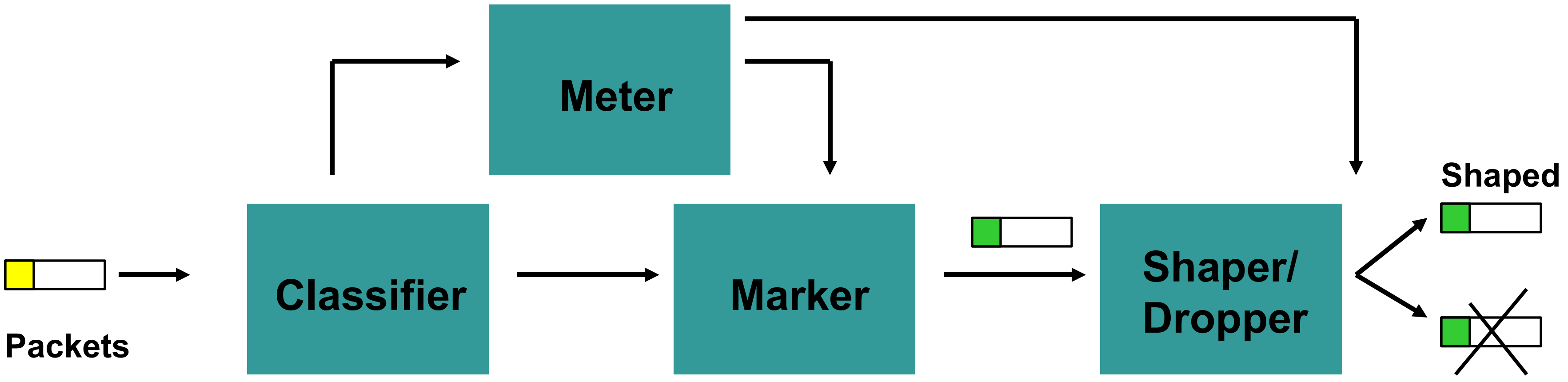
	Type	Class Selector Code Point
Prec 0	Routine	000000 (0)
Prec 1	Priority	001000 (8)
Prec 2	Immediate	010000 (16)
Prec 3	Flash	011000 (24)
Prec 4	Override	100000 (32)
Prec 5	Critical	101000 (40)
Prec 6	Inter-net	110000 (48)
Prec 7	Net-Control	111000 (56)

"match ip dscp 24"

is the same as

"match ip precedence 3"

DiffServ Traffic Conditioner



- Classifier: Selects a packet in a traffic stream based on the content of some portion of the packet header
- Meter: Checks compliance to traffic parameters (eg Token Bucket) and passes result to the marker and shaper/dropper to trigger a particular action for in/out of profile packets
- Marker: Writes/rewrites DSCP
- Shaper: Delays some packets to be compliant with a profile
- Dropper: Discards some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile

Classification and Marking

- **Classification**
  - Identification based on field(s) in a packet
  - Flow identification parameters
    - Src/Dest. Address, Source/Dest. Port, Protocol
  - IP Precedence / DSCP based
- **Marking**
  - Marking/Coloring packets to indicate class
  - Application marked or node configured
    - IP Precedence or DSCP
    - MPLS EXP
    - Other instances (FR-DE and ATM-CLP)

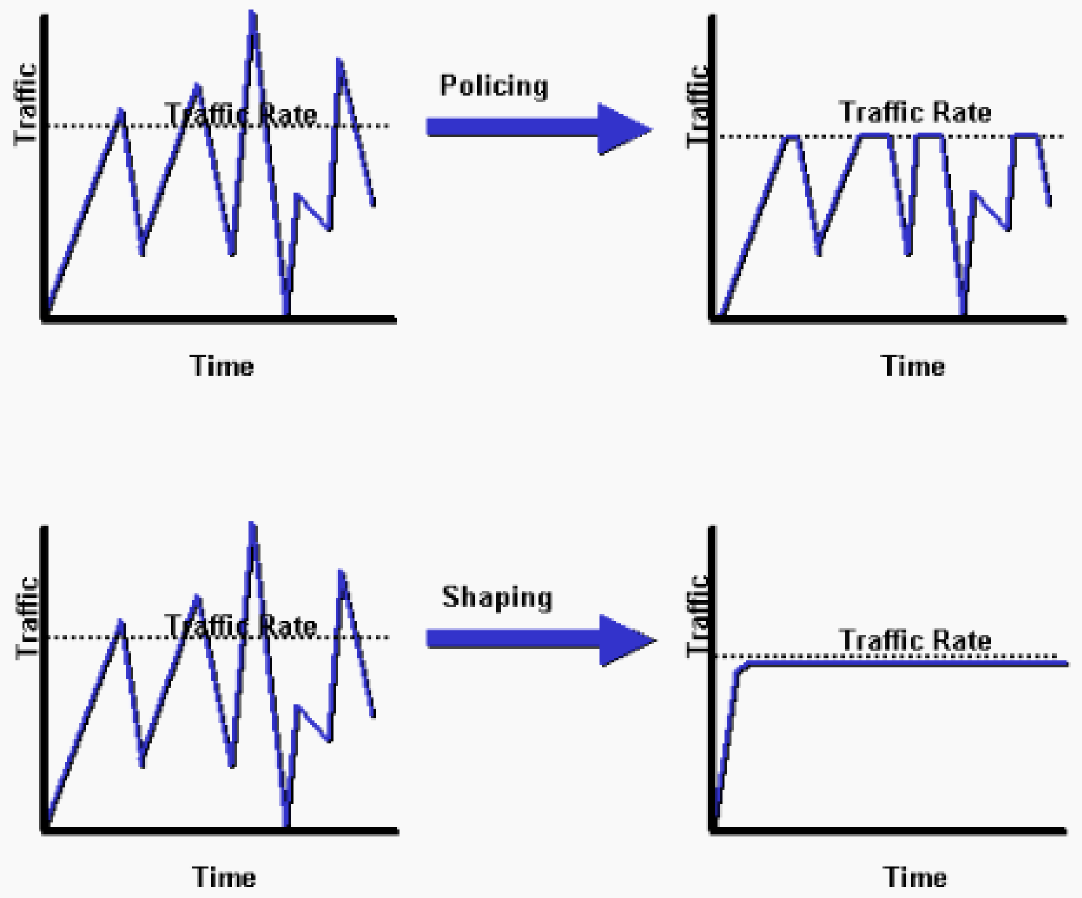
Traffic Metering

- **Traffic Rate Management in network boundary nodes**
- **Traffic Metering measures traffic**
    - **Does not alter traffic characteristics**
    - **Reports compliance results to Shaper or Dropper**
- **Uses Token Bucket Scheme to measure traffic**
    - **Mean or Committed Information Rate**
    - **Conformed Burst size**
    - **Extended Burst size**

Policing and Shaping

- **Police**
  - Sends conforming traffic and allows bursts
  - Drops non-conforming traffic (due to lack of tokens)
  - Provision for Packet re-marking
- **Shaping**
  - – Smoothes traffic but increases overall latency
  - – Buffers packets when tokens are exhausted

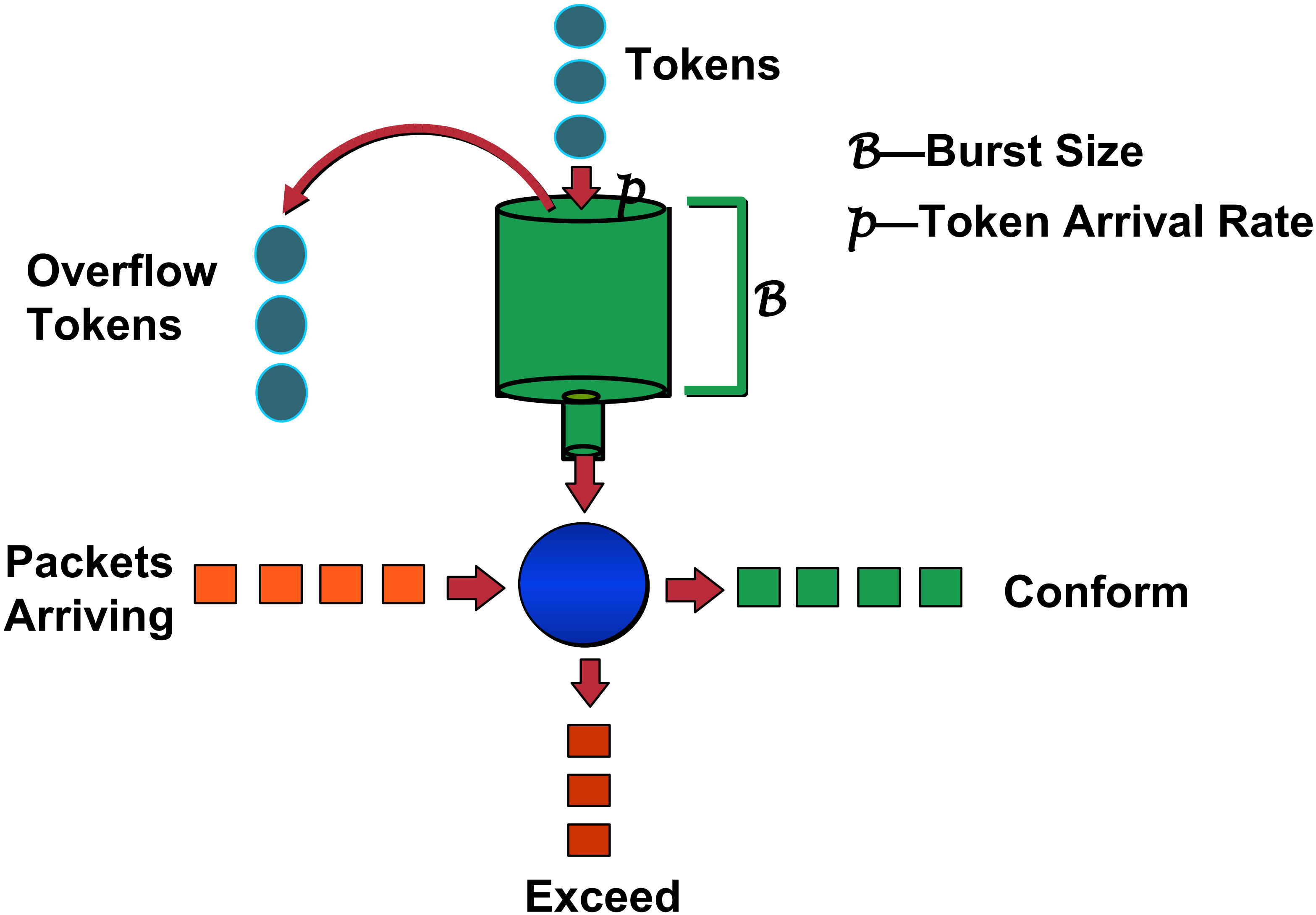
Policing and Shaping



Policing

- Uses the token bucket scheme
- Tokens added to the bucket at the committed rate
- Depth of the bucket determines the burst size
- Packets arriving with sufficient tokens in the bucket are said to conform
- Packets arriving with insufficient tokens in the bucket are said to exceed

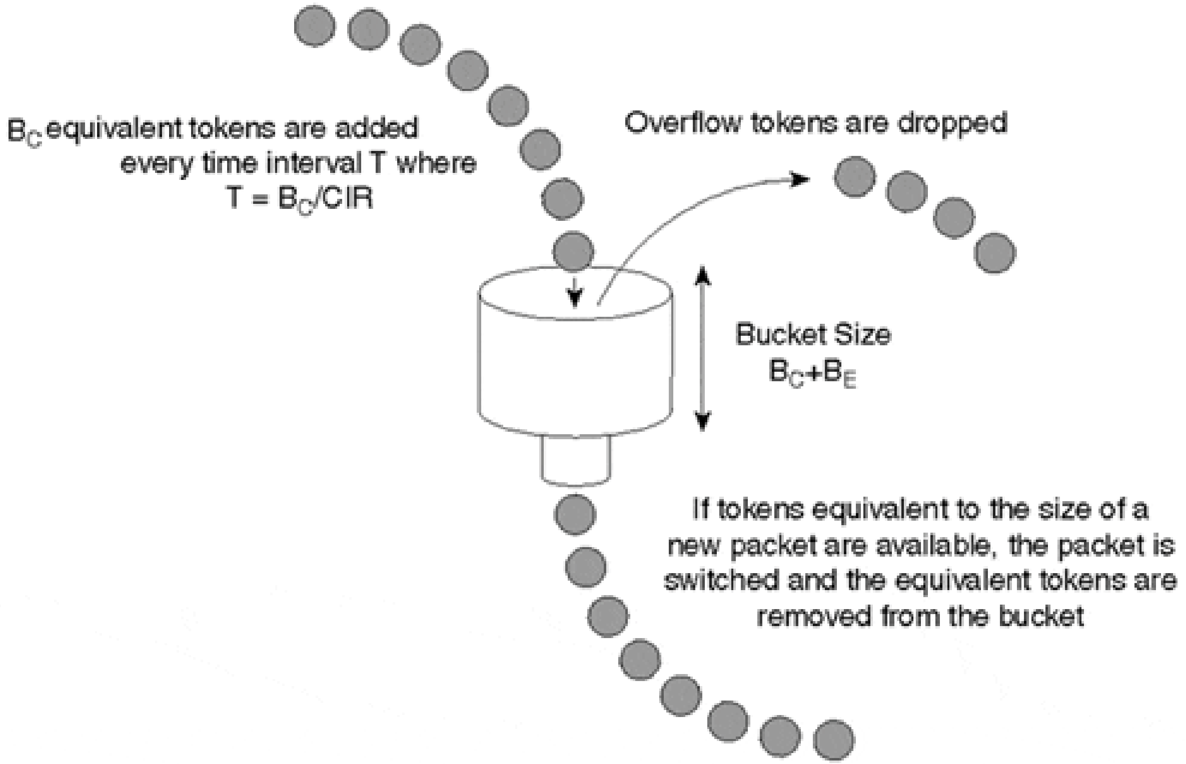
Token Bucket in Policing



Shaping

- Uses the **token bucket scheme**
- Smoothes bursty traffic to meet CIR through buffering
- Queued Packets transmitted as tokens are available

Token Bucket in Traffic Shaping



Per-Hop Behavior (PHB)

- PHB relates to resource allocation for a flow
- Resource allocation is typically Bandwidth
- Queuing / Scheduling mechanisms:
  - FIFO / WFQ / MWRR / MDRR
- PHB also involves determining a packet drop policy
- Congestion avoidance schemes – primary technique

  RED / WRED

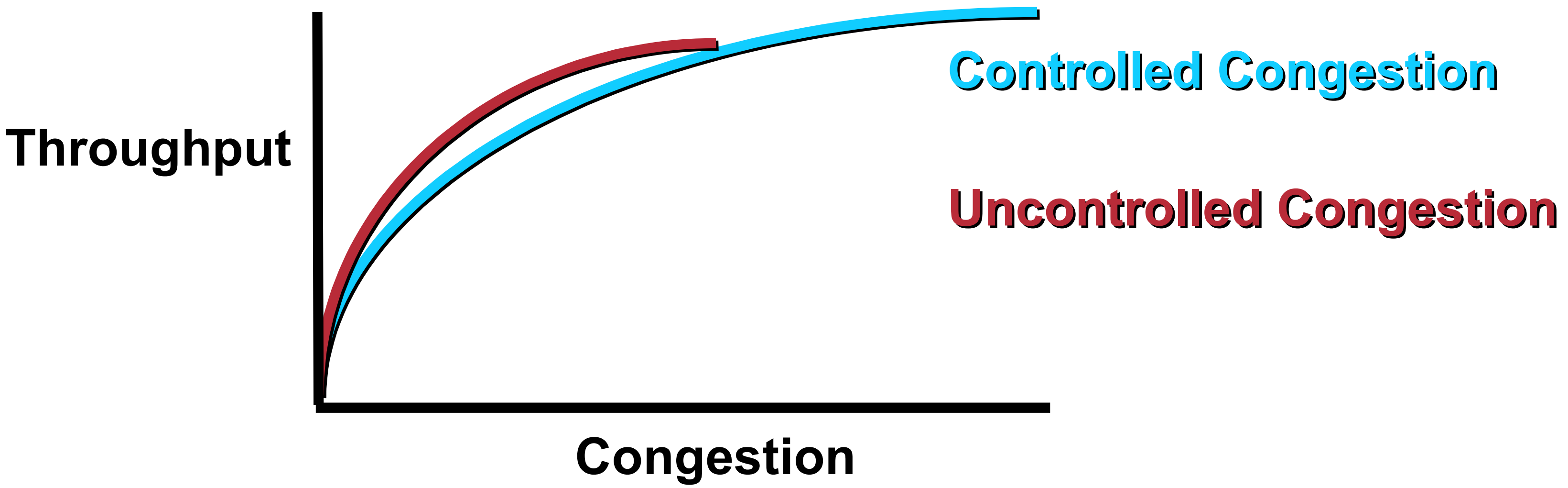
Queuing/Scheduling

- **Scheduling mechanisms guarantee BW for flows**
- **More bandwidth guarantee means dequeue more from one queue or set of queues.**
- **De-queue depends on weights allocated to queues**

Congestion Avoidance/Management

- **When there is congestion what should we do?**
  - **Tail drop i.e. Packets dropped due to Max Queue Length**
  - **Drop selectively but based on IP Prec / DSCP bit**
- **Congestion control mechanisms for TCP traffic**
  - **Adaptive**
  - **Dominant transport protocol**

The Problem of Congestion



- **Uncontrolled, congestion will seriously degrade performance**
  - **The system buffers fill up**
  - **Packets are dropped, resulting in retransmissions**
  - **This causes more packet loss and increased latency**
  - **The problem builds on itself**

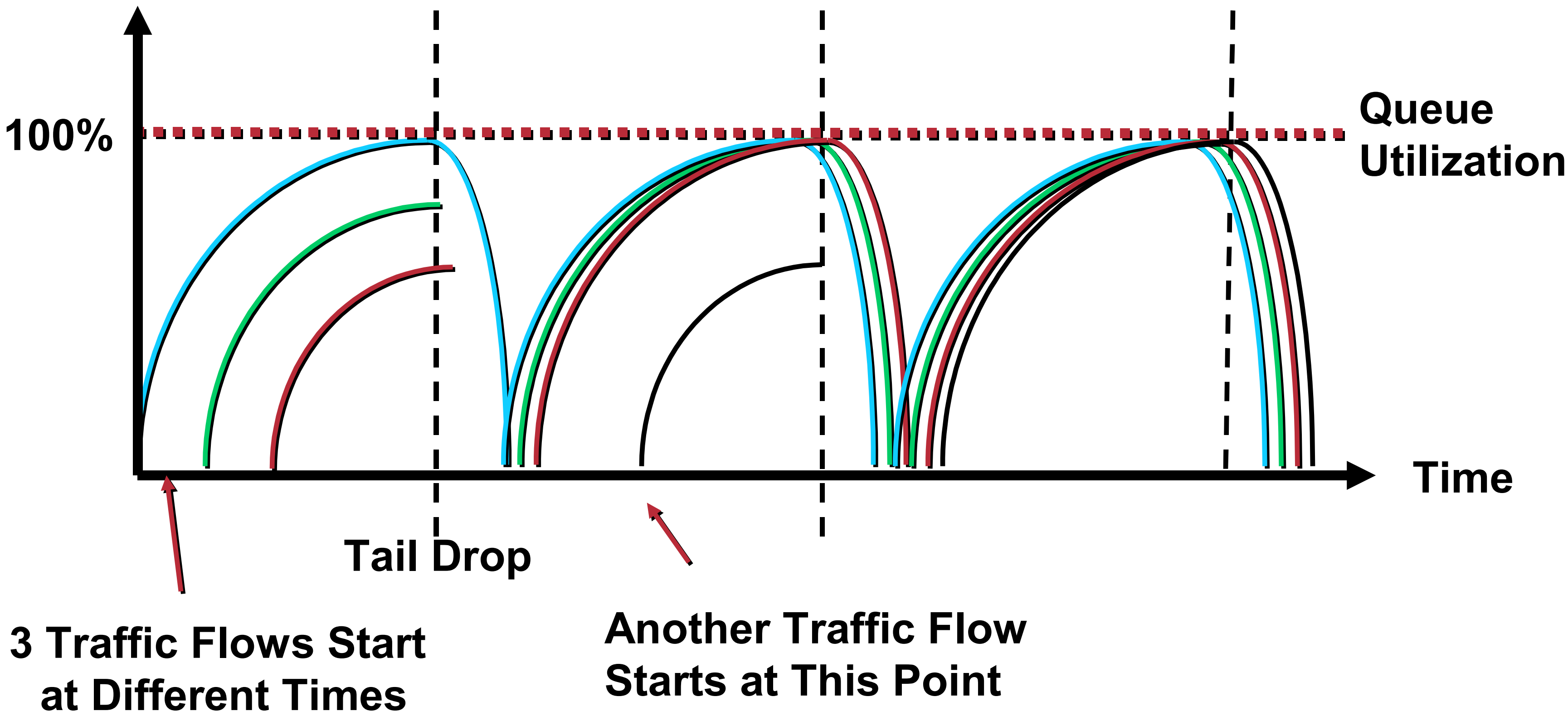
TCP traffic and Congestion

- **Congestion window based on slow-start**

  Sender / Receive negotiation

- **Packet loss indicator of congestion**
  - **Congestion window re-sizing**
  - **Source throttles traffic**

Global Synchronization



- Global synchronization is many connections going through TCP Slow-Start mode at the same time

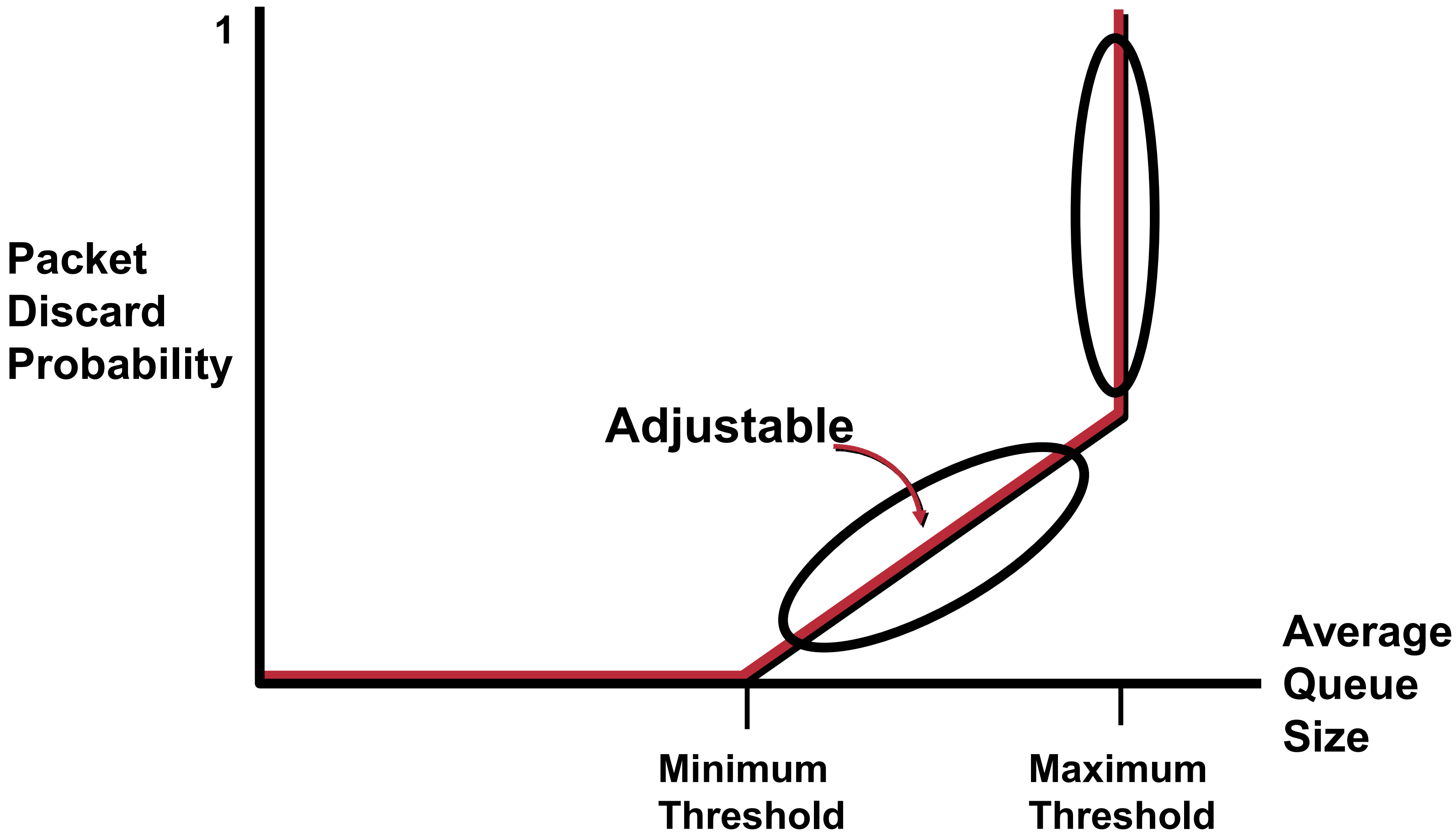
Random Early Detect (RED)

- A congestion avoidance algorithm
- Designed to work with a transport protocol like TCP
- Minimize packet delay jitter by controlling average queue size
- Uses Packet drop probability and Avg. Queue size
- Avoids global synchronization of many connections

RED—Packet-Drop Probability

- Packets are dropped sufficiently frequently to control the average queue size
- The probability that a packet is dropped from a connection is proportional to the amount of packets sent by the connection

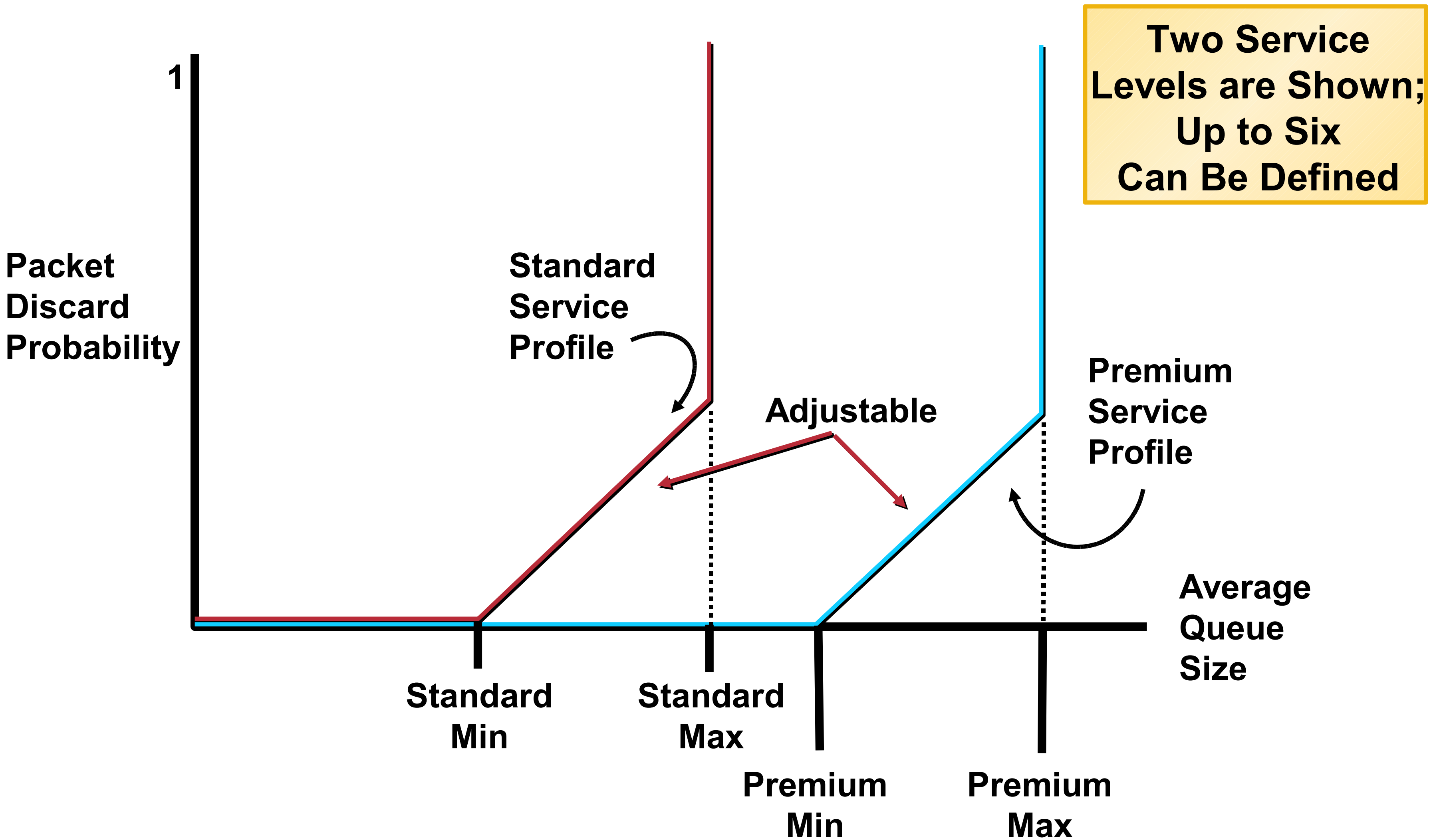
RED


1
Packet
Discard
Probability
Adjustable
Average
Queue
Size
Minimum
Threshold
Maximum
Threshold

Weighted RED (WRED)

- WRED combines RED with IP Precedence or DSCP to implement multiple service classes
- Each service class has a defined min and max thresholds, and drop rates

WRED Service Profile Example



When Should WRED be Used?

- **Where the bulk of your traffic is TCP as oppose to UDP**

  Only TCP will react to a packet drop; UDP will not

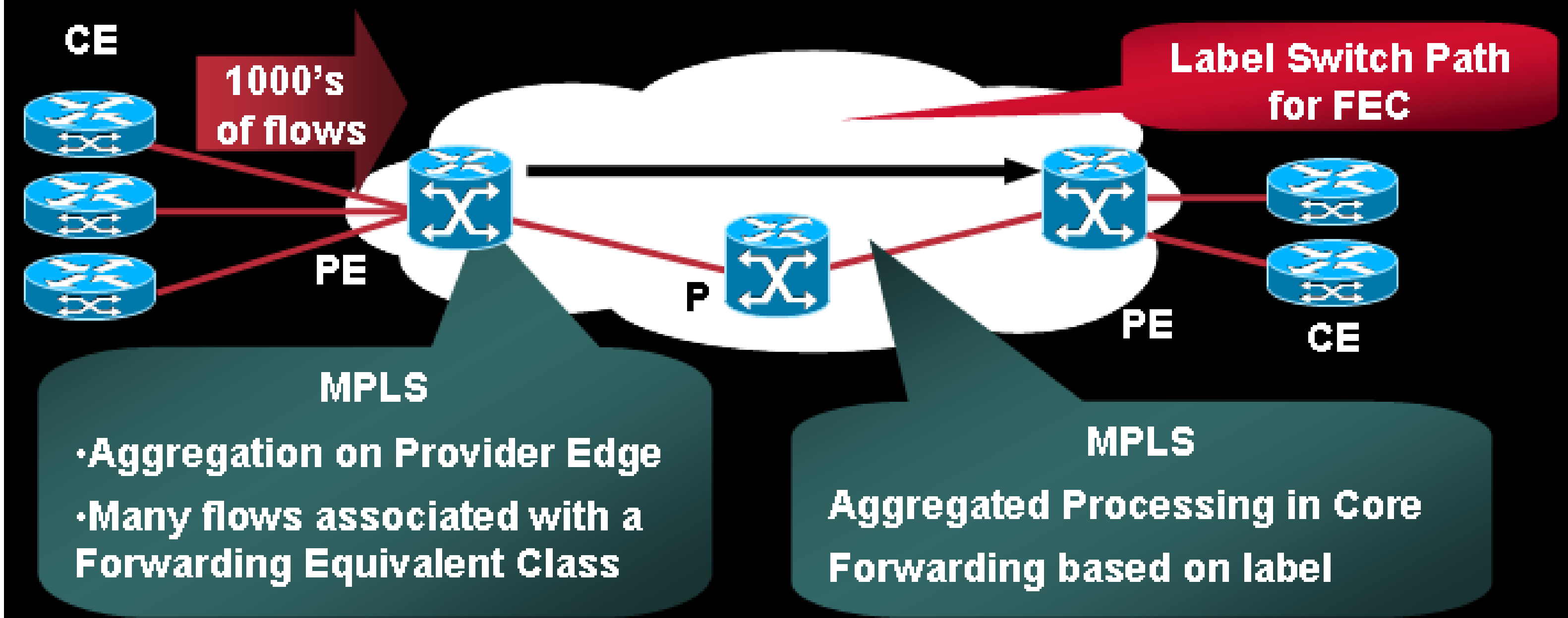
- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- **MPLS and QoS**
- Summary

MPLS Diffserv

MPLS DiffServ Architecture

- MPLS does NOT define new QoS architectures
- MPLS QoS uses Differentiated Services (DiffServ) architecture defined for IP QoS (RFC 2475)
- MPLS DiffServ is defined in RFC3270

DiffServ Scalability via Aggregation



MPLS scalability comes from

- aggregation of traffic on Edge
- processing of aggregate only in Core

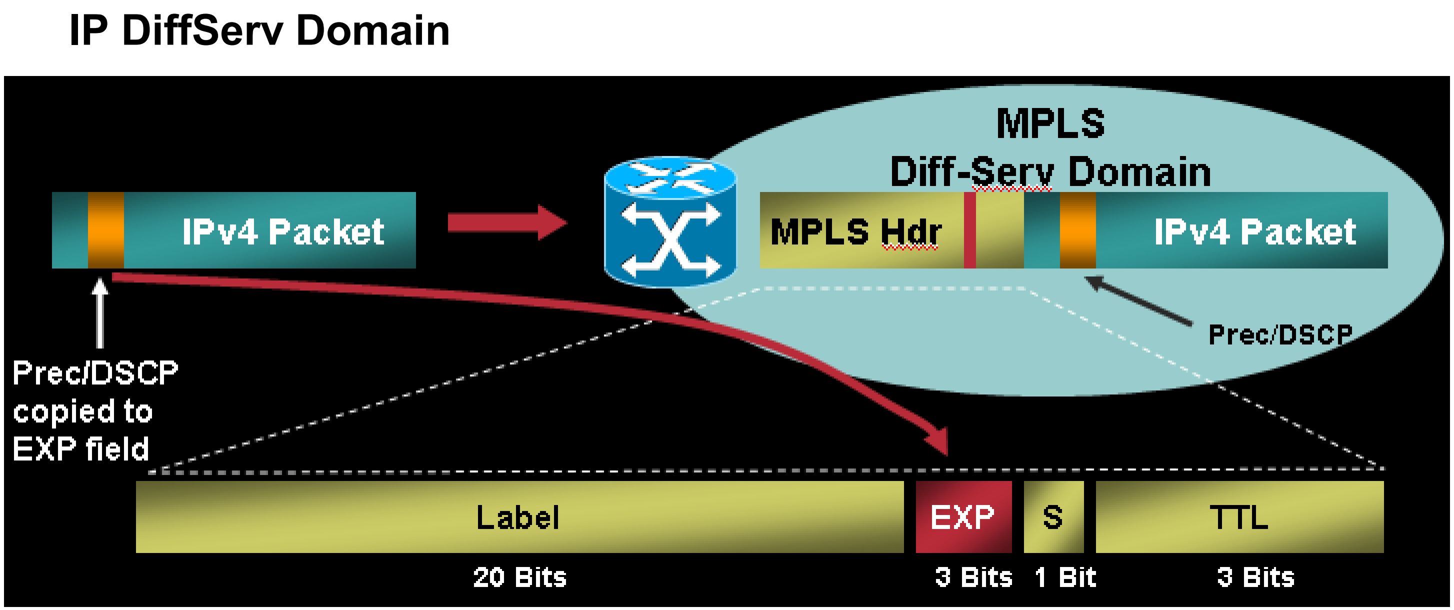
What’s Unchanged in MPLS DiffServ

- **When Compared to IP DiffServ**
  - Functional components (TCA/PHB) and where they are used

    Classification, marking, policing, and shaping at network boundaries

    Buffer management and packet scheduling mechanisms used to implement PHB
  - PHB definitions
    - EF: low delay/jitter/loss
    - AF: low loss
    - BE: No guarantees (best effort)

What's new in MPLS DiffServ ?



- Prec/DSCP field is not directly visible to MPLS Label Switch Routers (they forward based on MPLS Header and EXP field)
- Information on DiffServ must be made visible to LSR in MPLS Header using EXP field / Label.
- How do we map DSCP into EXP ? Interaction between them.

DSCP to EXP Mapping

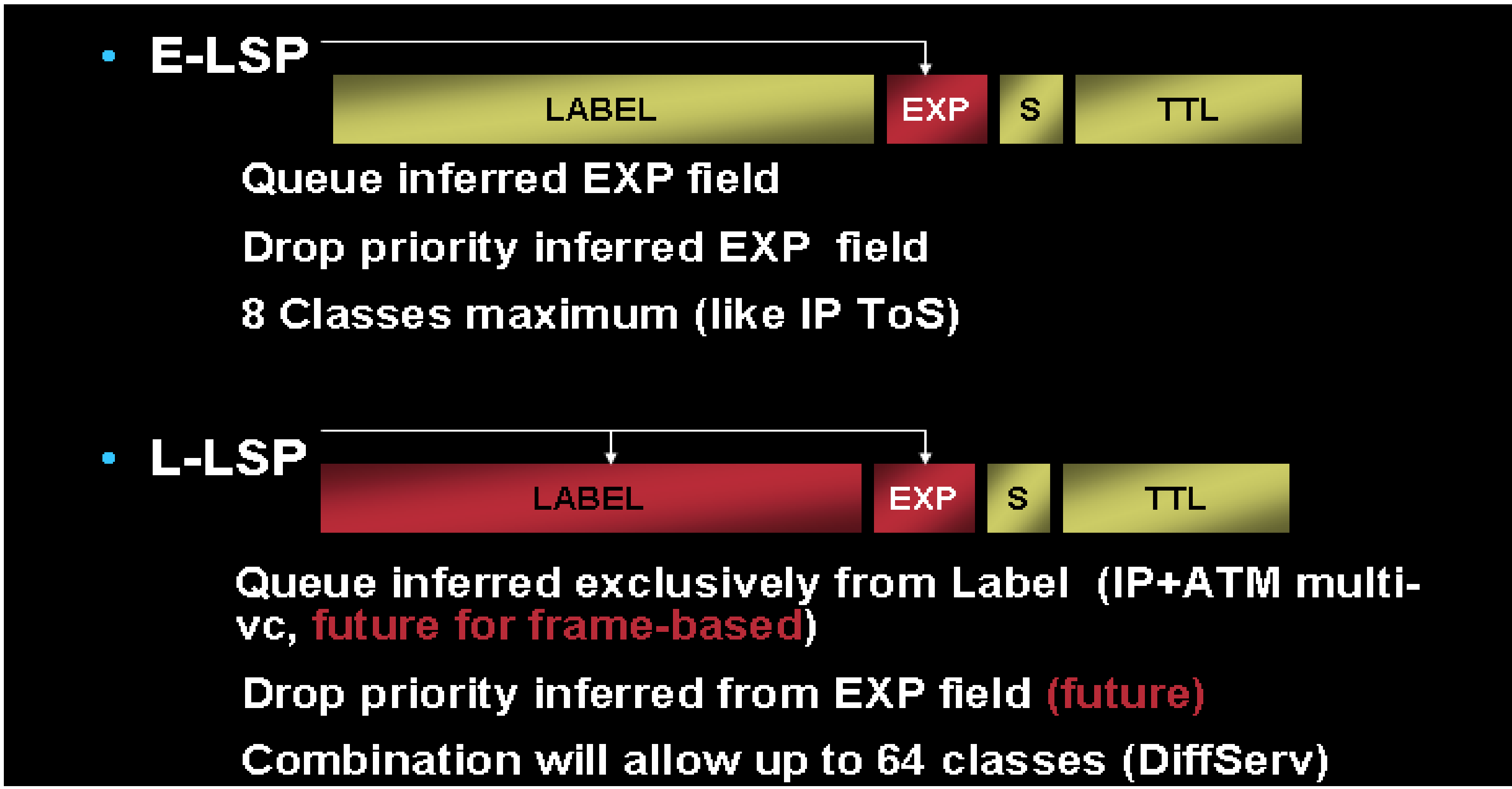
		DSCP Value (6 Bits)				EXP Value (3 bits)
Expedited Forwarding	EF	101110			→	101
Assured Forwarding						
Class 1	AF1	001010	001100	001110	→	001
Class 2	AF2	010010	010100	010110	→	010
Class 3	AF3	011010	011100	011110	→	011
Class 4	AF4	100010	100100	100110	→	100
Best Effort		000000			→	000

RFC3270 does not recommend specific EXP values for DS PHBs (EF/AF/CS)

MPLS DiffServ – RFC 3270

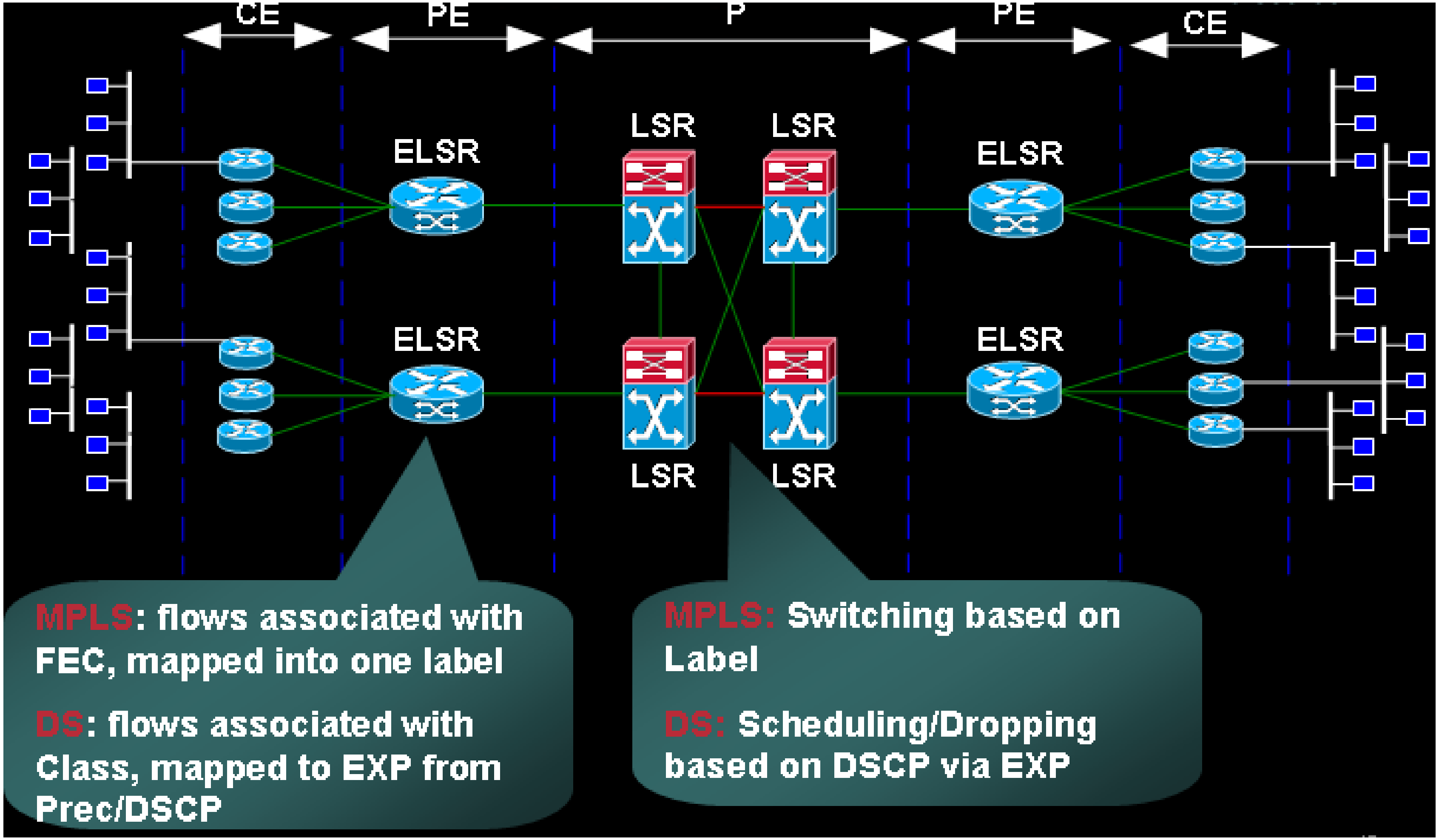
- **Problem: IP DSCP = 6 bits while MPLS EXP = 3bits**
- **Solution: where 8 or less PHBs are used, those can be mapped into EXP field➔use "E-LSPs with preconfigured mapping"**
- **Solution: where more than 8 PHBs are used in core, those need to be mapped in both "label and EXP" ➔"L-LSPs" are needed**

Types of Label Switched Paths



- **Both E-LSP and L-LSP can use LDP or RSVP for label distribution**

MPLS DiffServ Topology



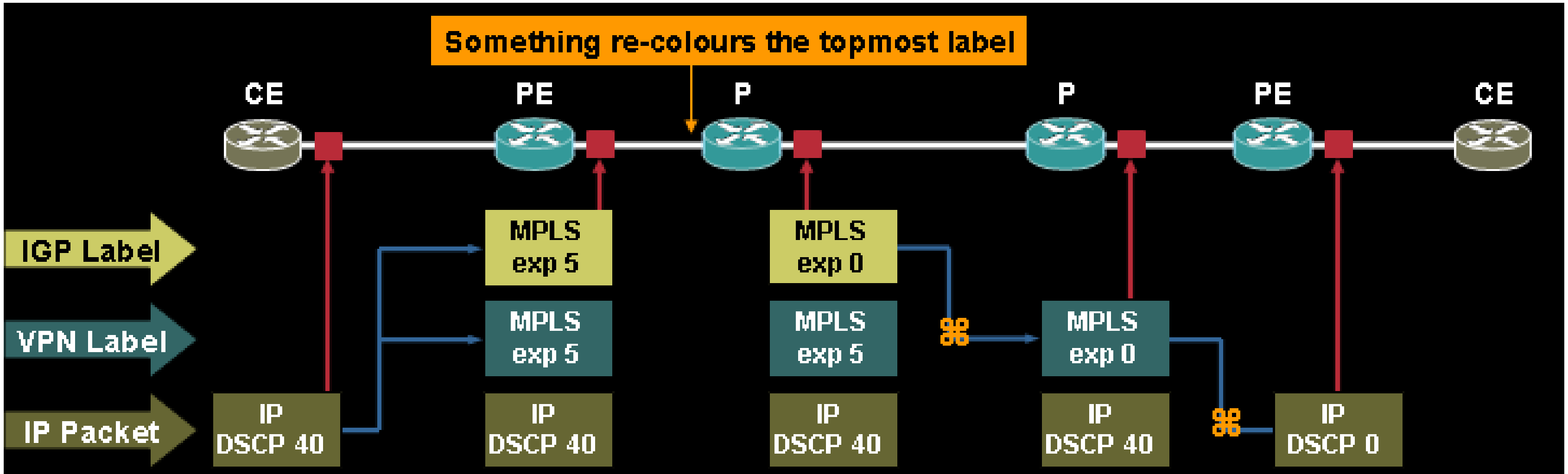
MPLS DiffServ Tunneling Modes

- Based on RFC 3270
- Modes
  - Uniform
  - Short-Pipe
  - Pipe

Uniform Mode

- **Assume the entire admin domain of a Service Provider is under a single DiffServ domain**
- **Then, it is likely a requirement to keep the colouring information uniform (keep it when going from IP to IP, IP to MPLS, MPLS to MPLS, MPLS to IP).**

Uniform Mode



⌘ In both the MPLS-to-MPLS and the MPLS-to-IP cases, the PHB of the topmost popped label is copied into the new top label or the IP DSCP if no label remains.

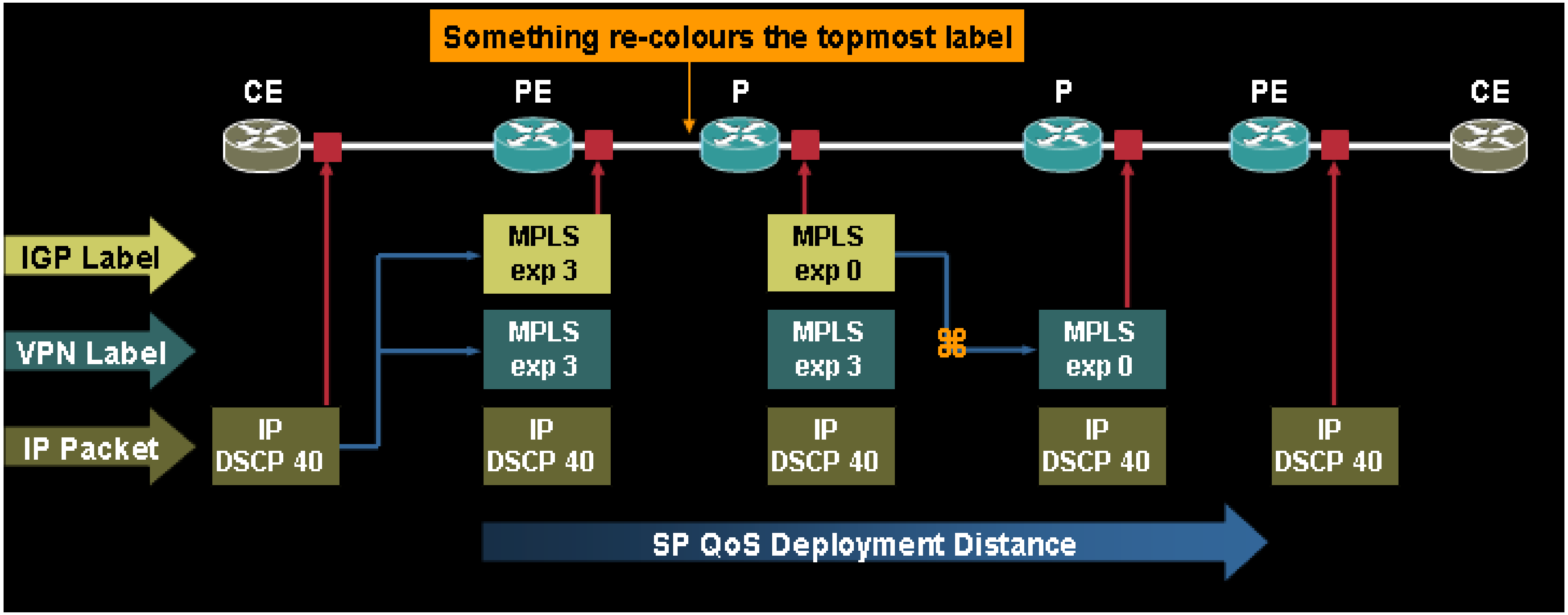
Short-Pipe Mode

- Assume an ISP network implementing a DiffServ Policy
- Assume its customer network implementing another policy
- Requirement:

  Transparency: the customer wants to preserve its DSCP intact

  Uniformity: within the IP/MPLS backbone, the SP wants to have a uniform diffserv domain

Short-Pipe Mode

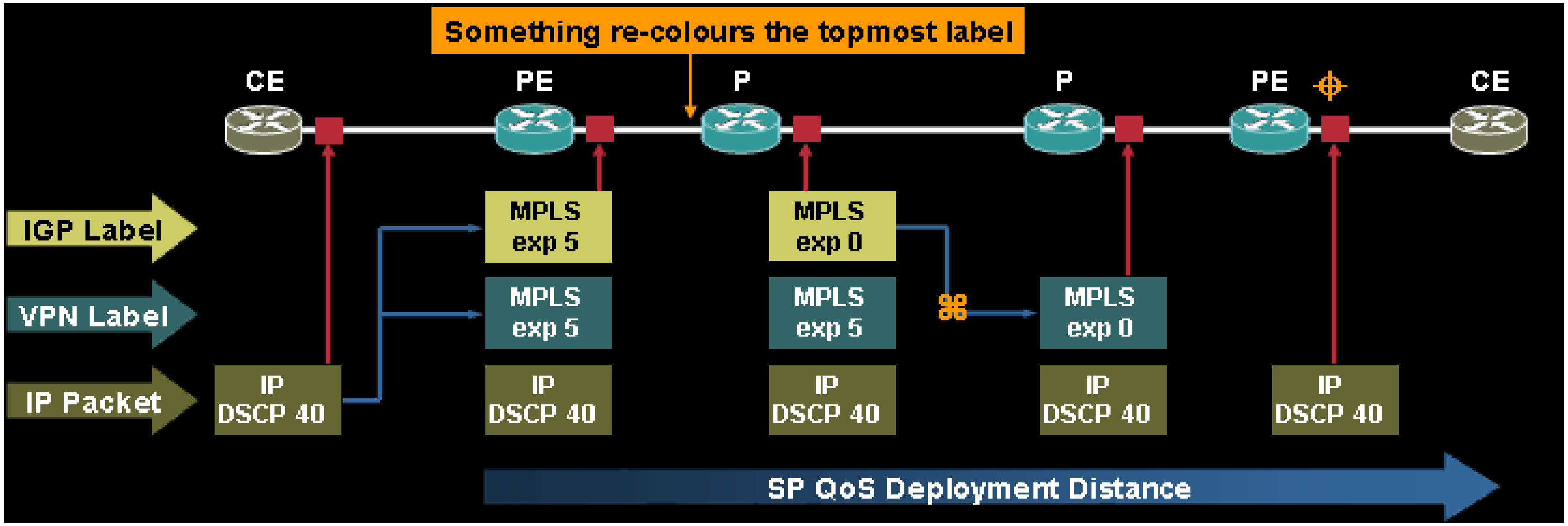


- ⌘ The PHB of the topmost popped label is copied into the new top label
- Note that policy applied on the egress interface of the egress PE is based on the DSCP of the customer, hence the ‘short-pipe’ naming.

Pipe Mode

- Exactly the same case as Short-Pipe
- However, the SP wants to drive the outbound classification for WFQ/WRED on the egress interface from a PE to a CE based on its DiffServ policy (EXP)

Pipe Mode



- ⌘ The PHBs of the topmost popped label is copied into the new top label
- ⌖ Classification is based on mpls-exp field (EXP=0) of the topmost received MPLS frame

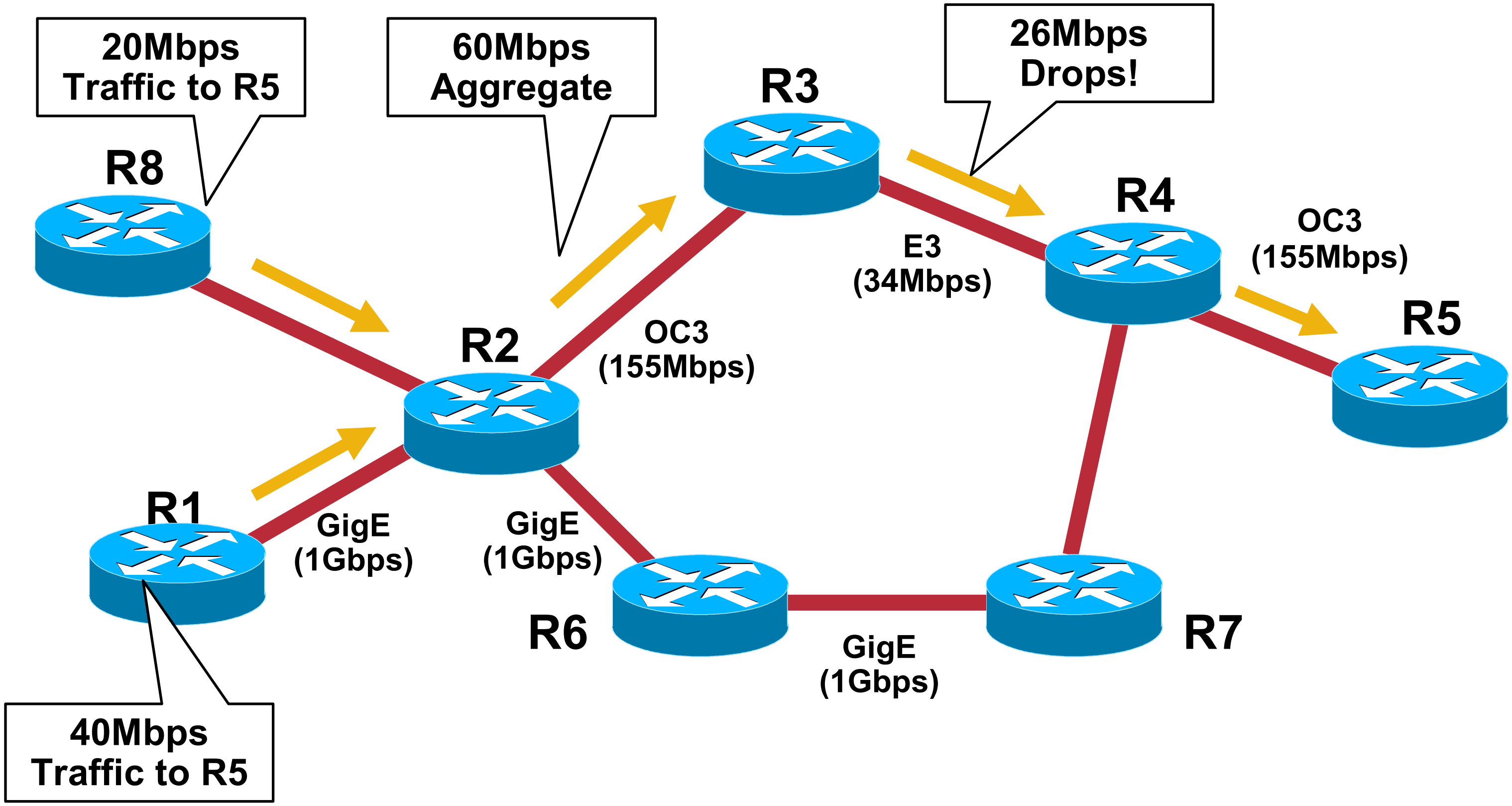
MPLS TE and Diffserv

BW Optimization and Congestion Mgmt. in Parallel

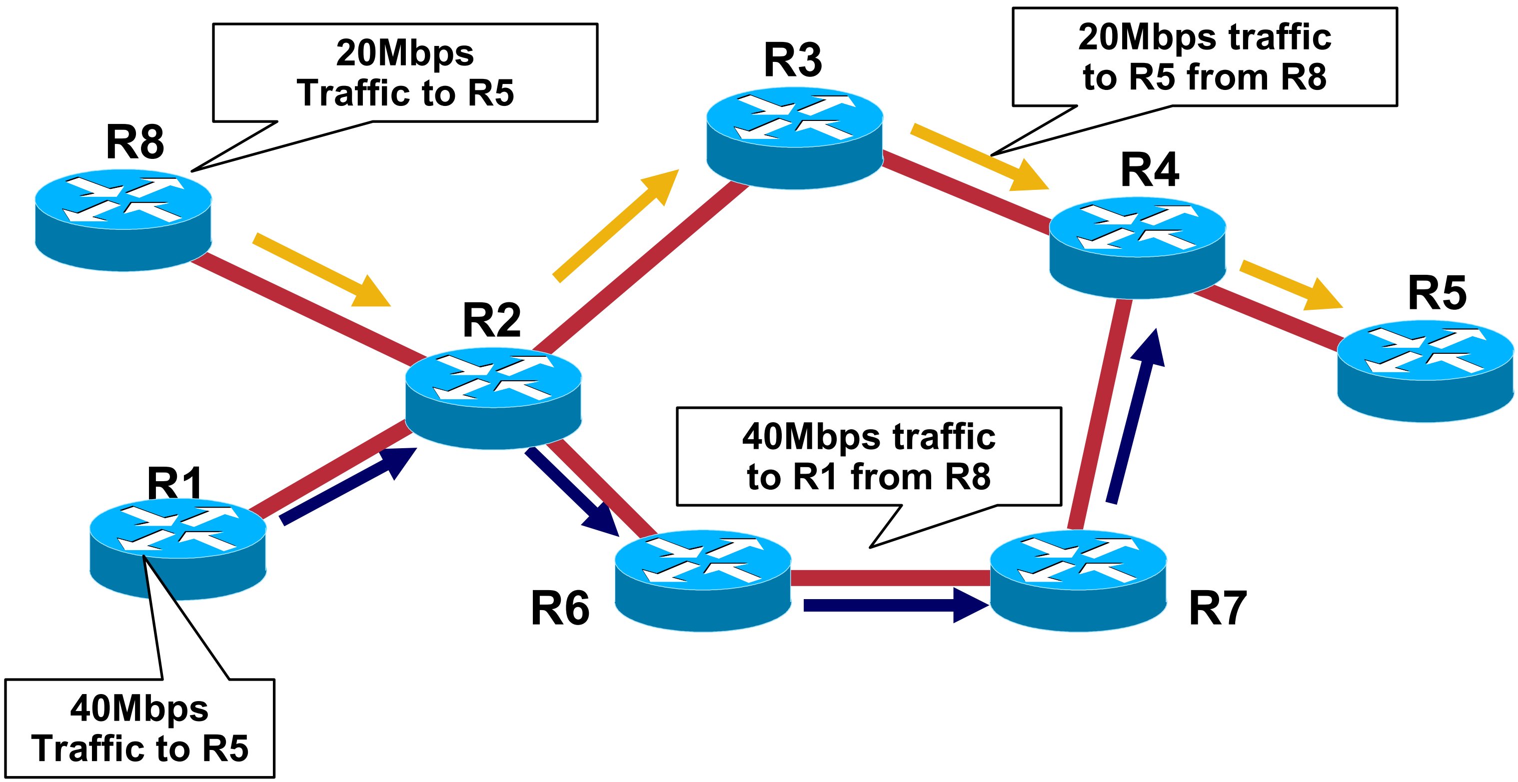
TE + DiffServ

- Spread Traffic around with more flexibility than the IGP Offers
- Reserve per-class bandwidth, sort of
- Manage Unfairness During Temporary Congestion

Why TE: Shortest Path and Congestion



The TE Solution

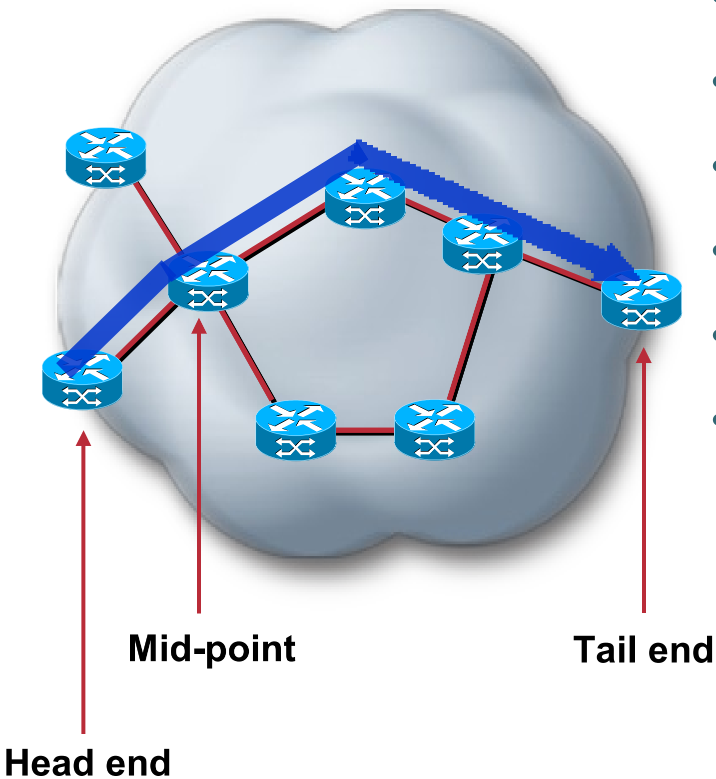


- MPLS Labels can be used to engineer explicit paths
- Tunnels are UNI-DIRECTIONAL

Normal path: R8 ➔ R2 ➔ R3 ➔ R4 ➔ R5

Tunnel path: R1 ➔ R2 ➔ R6 ➔ R7 ➔ R4

How MPLS TE Works

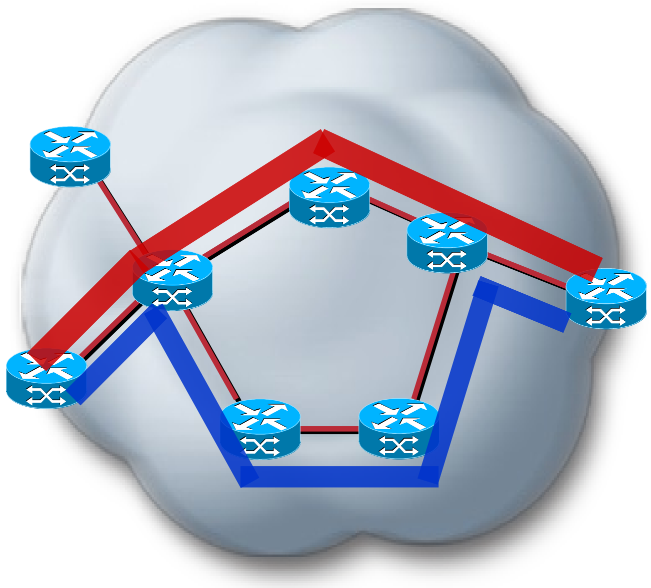


- Explicit routing
- Constrained-based routing
- Admission control
- Protection capabilities
- RSVP-TE to establish LSPs
- ISIS and OSPF extensions to advertise link attributes

DiffServ-Aware TE (DS-TE)

- **Regular TE allows for one reservable bandwidth amount per link**
- **DS-TE allows for more than one reservable bandwidth amount per link**
- **Brings per-class dimension to TE**
- **Basic idea: connect PHB class bandwidth to DS-TE bandwidth sub-pool**

DiffServ-Aware TE



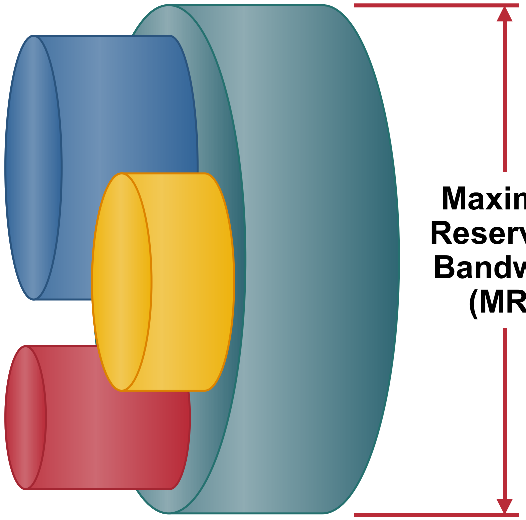
- **Per-class constrained-based routing**
- **Per-class admission control**

Low-Latency TE LSP with Reserved BW

Best-Effort TE LSP

DiffServ-Aware TE

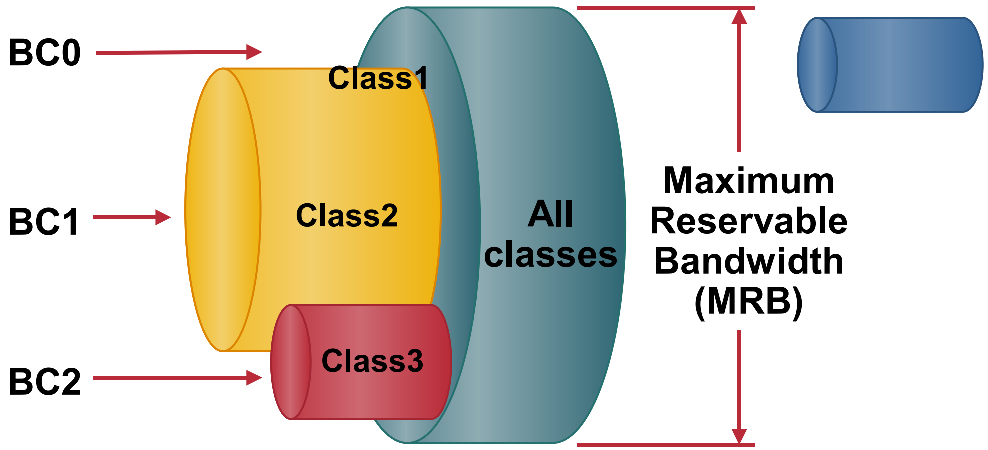
DS-TE BW Allocation



Maximum Reservable Bandwidth (MRB)

- Link BW distributed in pools or BW Constraints (BC)
- Up to 8 BW pools
- Different BW pool models

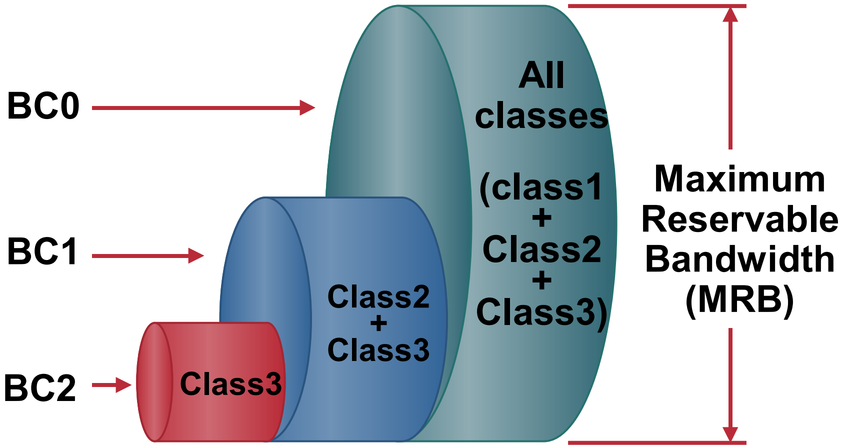
DS-TE BW Pools – Maximum Allocation Model (MAM)



BC0: 20% Best Effort
BC1: 50% Premium
BC2: 30% Voice

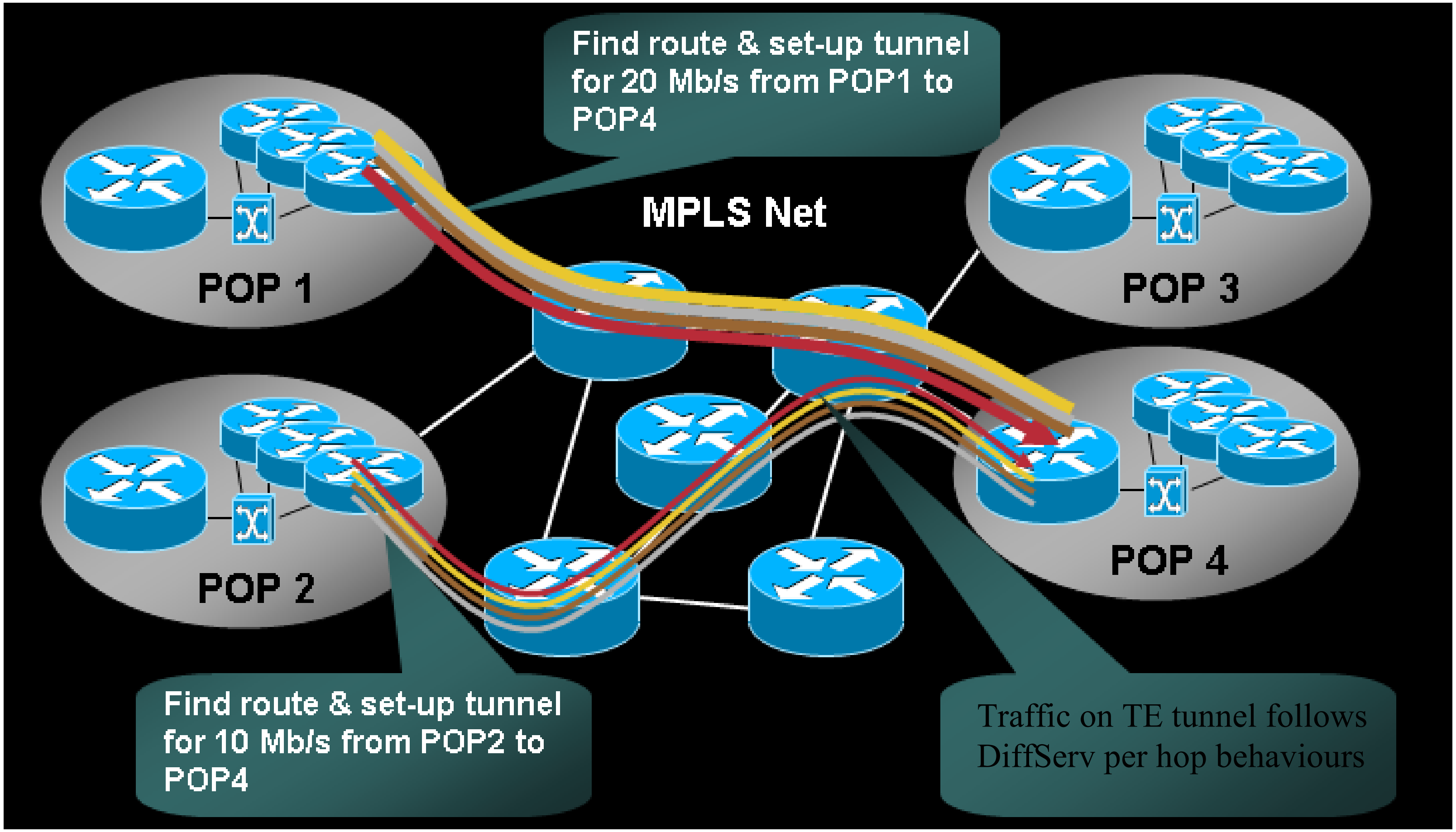
DS-TE BW Pools – Russian Dolls Model (RDM)

- BW pool applies to one or more classes
- Global BW pool (BC0) equals MRB
- BC0..BCn used for computing unreserved BW for class n

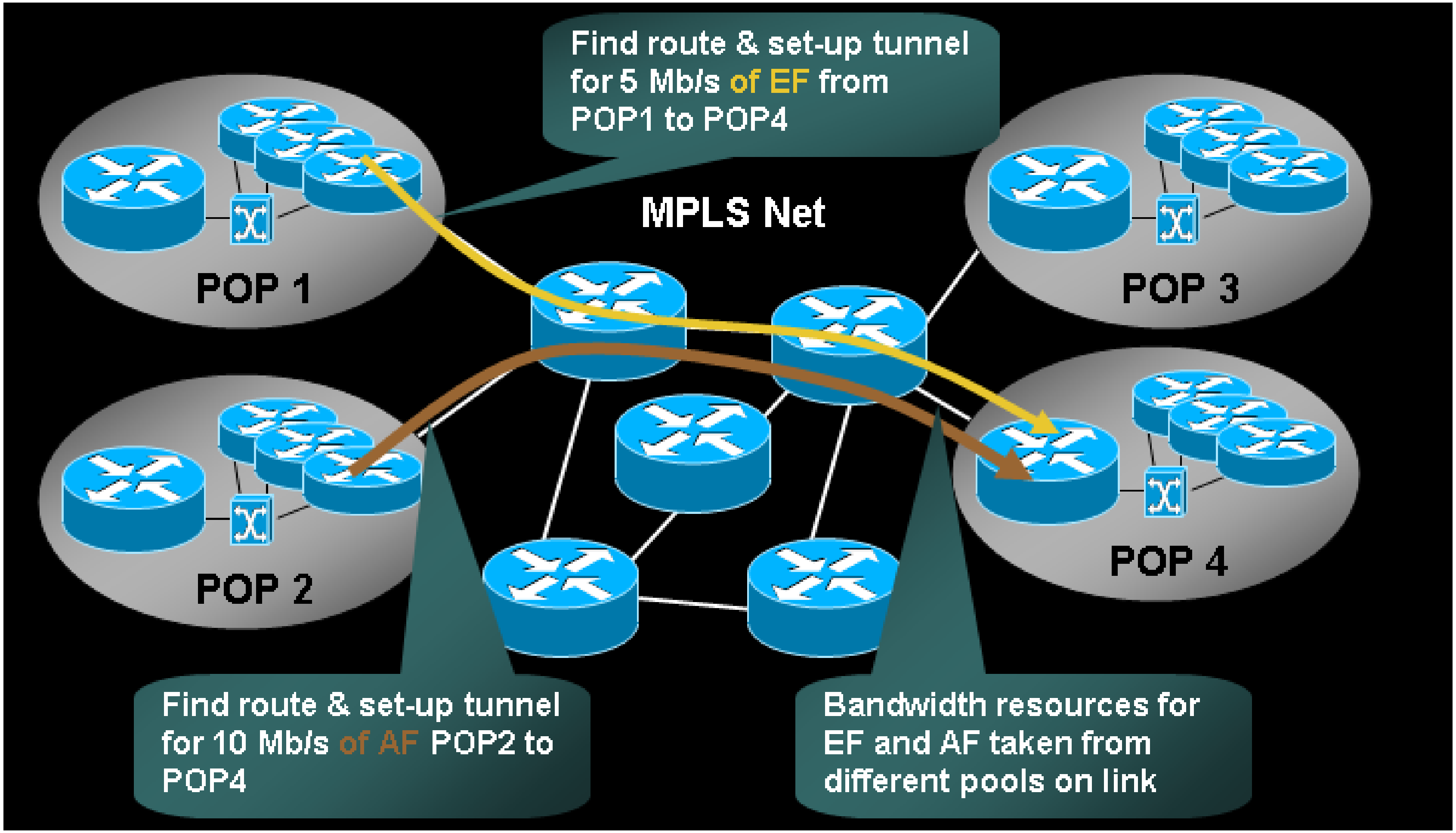


BC0: MRB Best Effort + Premium + Voice
BC1: 50% Premium + Voice
BC2: 30% Voice

Aggregate TE in DiffServ Network



DiffServ TE


Find route & set-up tunnel for 5 Mb/s of EF from POP1 to POP4
MPLS Net
POP 1
POP 2
POP 3
POP 4
Find route & set-up tunnel for 10 Mb/s of AF POP2 to POP4
Bandwidth resources for EF and AF taken from different pools on link

DS TE and QoS

“DiffServ TE <u>does not</u> preclude the necessity of configuring PHB QoS in the TE path. DiffServ TE operates in conjunction with QoS mechanisms”

- Introduction
- QoS Service Models
- DiffServ QoS Techniques
- MPLS QoS
- Summary

Summary

- QoS techniques
  - Effective allocation of network resources
- IP Diff Serv
  - Service Differentiation
- MPLS & Diff Serv
  - Builds scalable networks for SP
- DiffServ Tunneling Modes
  - Scalable and flexible QoS options
  - Supports Draft Tunneling Mode RFC
- Diff Serv TE
  - Provides strict point-to-point guarantees
  - Pipe Model

Q & A