Quality of Service in IP Networks

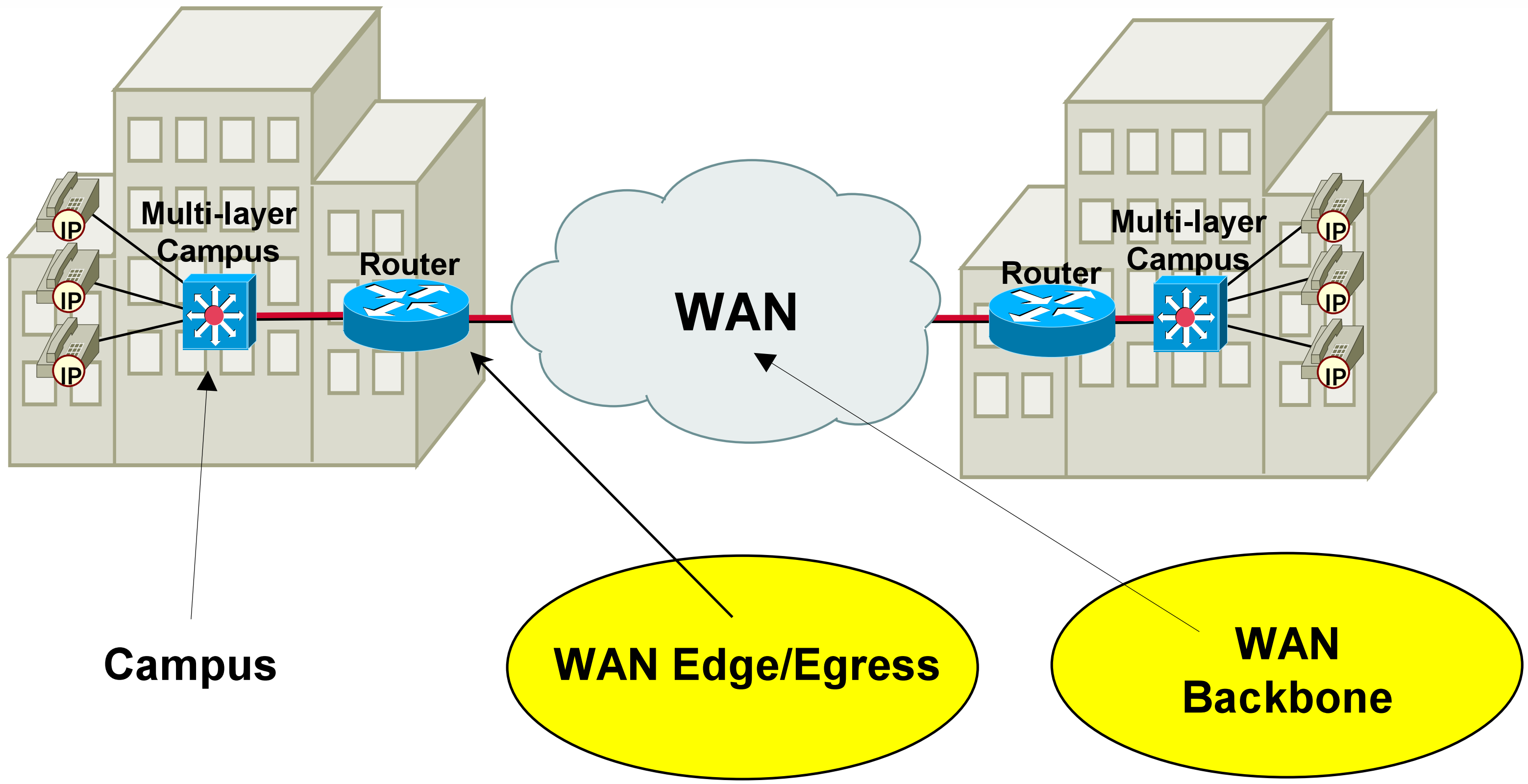
Ivo Němeček,

Systems Engineer, Cisco Systems

inemecek@cisco.com

Domains of QoS Consideration

Strong as your Weakest Link



Avoiding Loss, Delay and Delay Variation (Jitter)

VoIP Packet Format

VoIP Packet

Voice Payload	RTP Header	UDP Header	IP Header	Link Header
X Bytes	12 Bytes	8 Bytes	20 Bytes	X Bytes

- Payload Size, PPS and BPS Vendor Implementation Specific
- For Example:

Not including Link Layer Header or CRTP

Cisco Router at G.711 = 160 Byte Voice Payload at 50pps (80kbps)
Cisco Router at G.729 = 20 Byte Payload at 50pps (24kbps)
Cisco IP Phone at G.711 = 240 Byte Payload at 33pps (74.6kbps)
Cisco IP Phone at G.723.1 = 24 Byte Payload at 33pps (17kbps)

Note - Link Layer Sizes Vary per Media

Various Link Layer Header Sizes

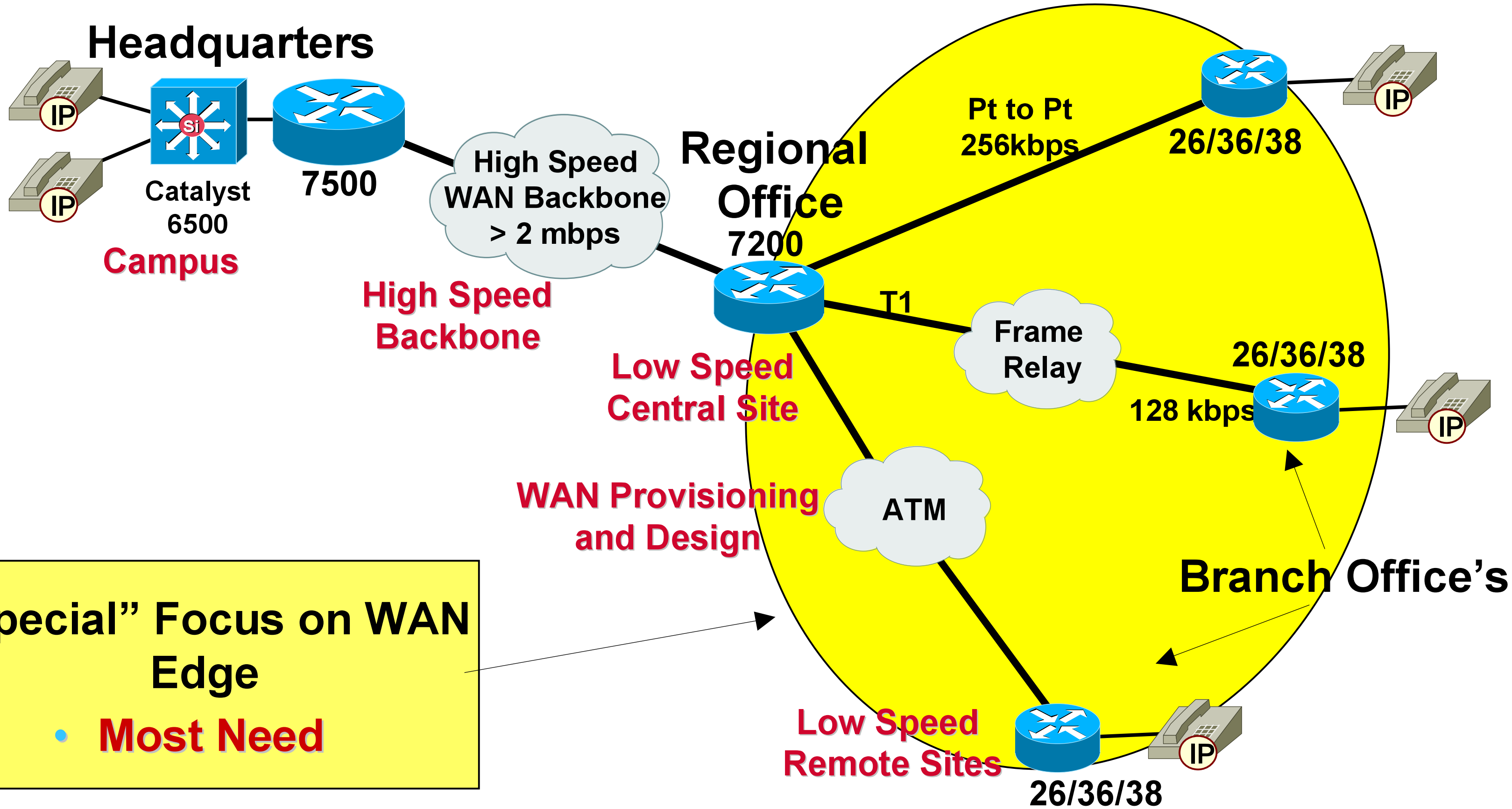
"Varying Bit Rates per Media"

Example - G.729 with 60 byte packet (Voice and IP Header) at 50pps (No RTP Header Compression)

Media	Link Layer Header Size	Bit Rate
Ethernet	14 bytes	29.6kbps
PPP	6 bytes	26.4kbps
Frame Relay	4 Bytes	25.6kbps
ATM	5 Bytes Per Cell	42.4kbps

Note - For ATM a single 60byte packet requires two 53 Byte ATM cells

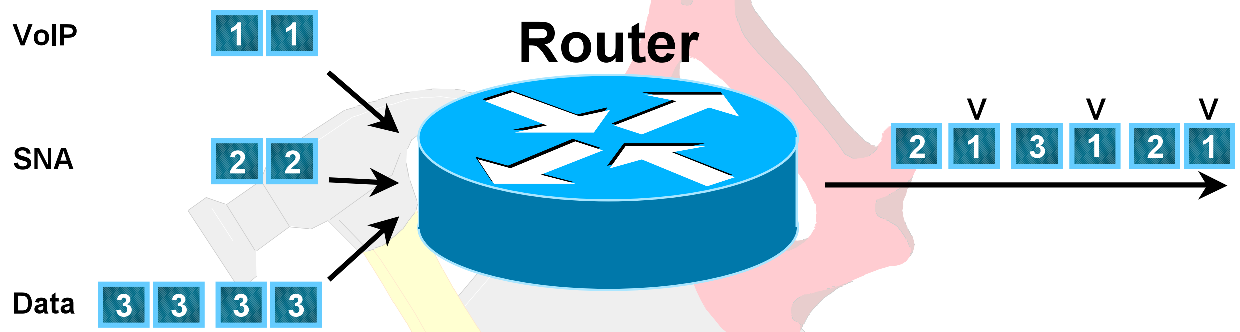
Case Study: End to End Quality of Service



Applying Proper Tools in Proper Location

Router/Switch Egress QoS Tools

"Three Classes of QoS Tools"



- **Prioritization**
  - **Low Speed WAN, High Speed WAN, Campus**
- **Link Efficiency**
  - **Fragment and Interleave, cRTP, VAD**
- **Traffic Shaping**
  - **Speed Mismatches + to avoid Bursting**

Low Speed WAN Egress QoS

2 meg or less

- **CAR, RED, WRED**
- **WFQ Based QoS Mechanisms**

  IP Precedence, RSVP
- **CBWFQ - Class Based WFQ**
- **Alternatives**

CAR

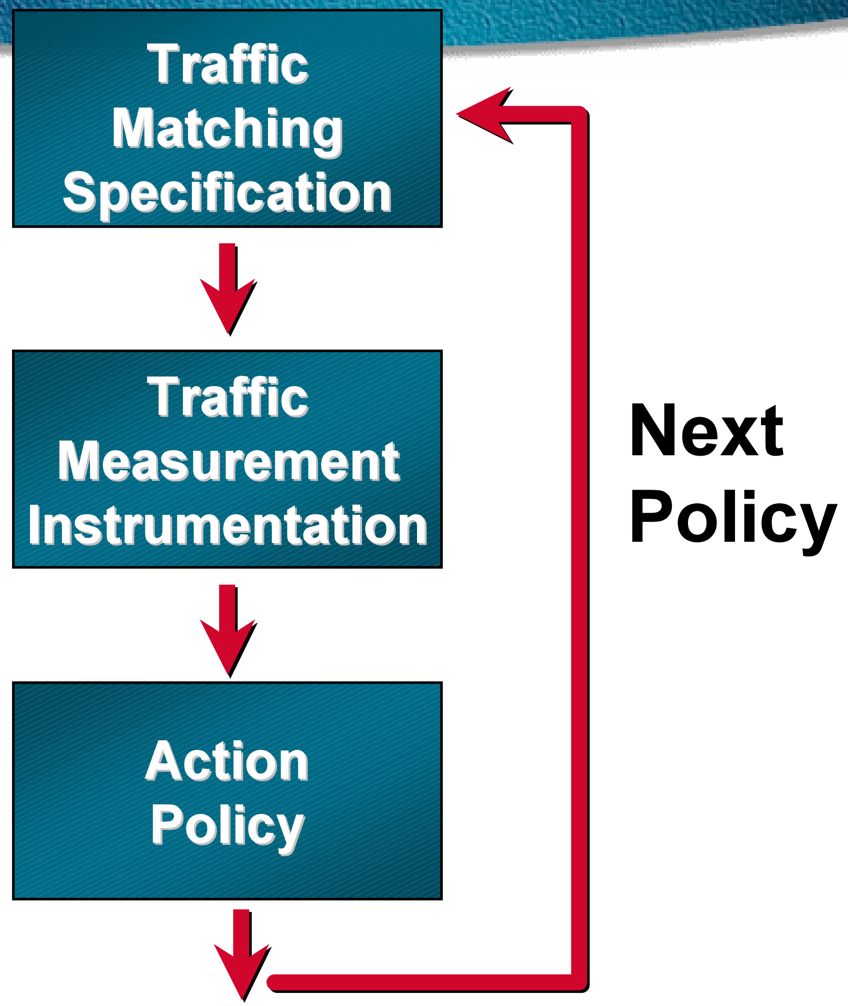
Committed Access Rate (CAR)

- CAR is IOS Feature name
- Two functions

  Packet Classification —
  sort a subset of traffic matching some complex criterion

  Traffic Conditioning
  rate measurement, rate limiting, packet marking (IP Precedence rewrite)

CAR—Overview



CAR—Traffic Matching Specification

- Identify packets of interest for packet classification or rate limiting or both
- Matching specification

    1) All traffic

    2) IP precedence

    3) MAC address

    4) QoS group

    5) IP access list—Standard and extended (slower)

CAR—Traffic Measurement

- Uses the **token bucket scheme** as a measuring mechanism
- Tokens are added to the bucket at the **committed rate** and the number of tokens in the bucket is limited by the normal burst size
- Depth of the bucket determines the burst size

CAR—Traffic Measurement

- Packets arriving with sufficient tokens in the bucket are said to **conform**
- Packets arriving with insufficient tokens in the bucket are said to **exceed**

CAR—Traffic Measurement

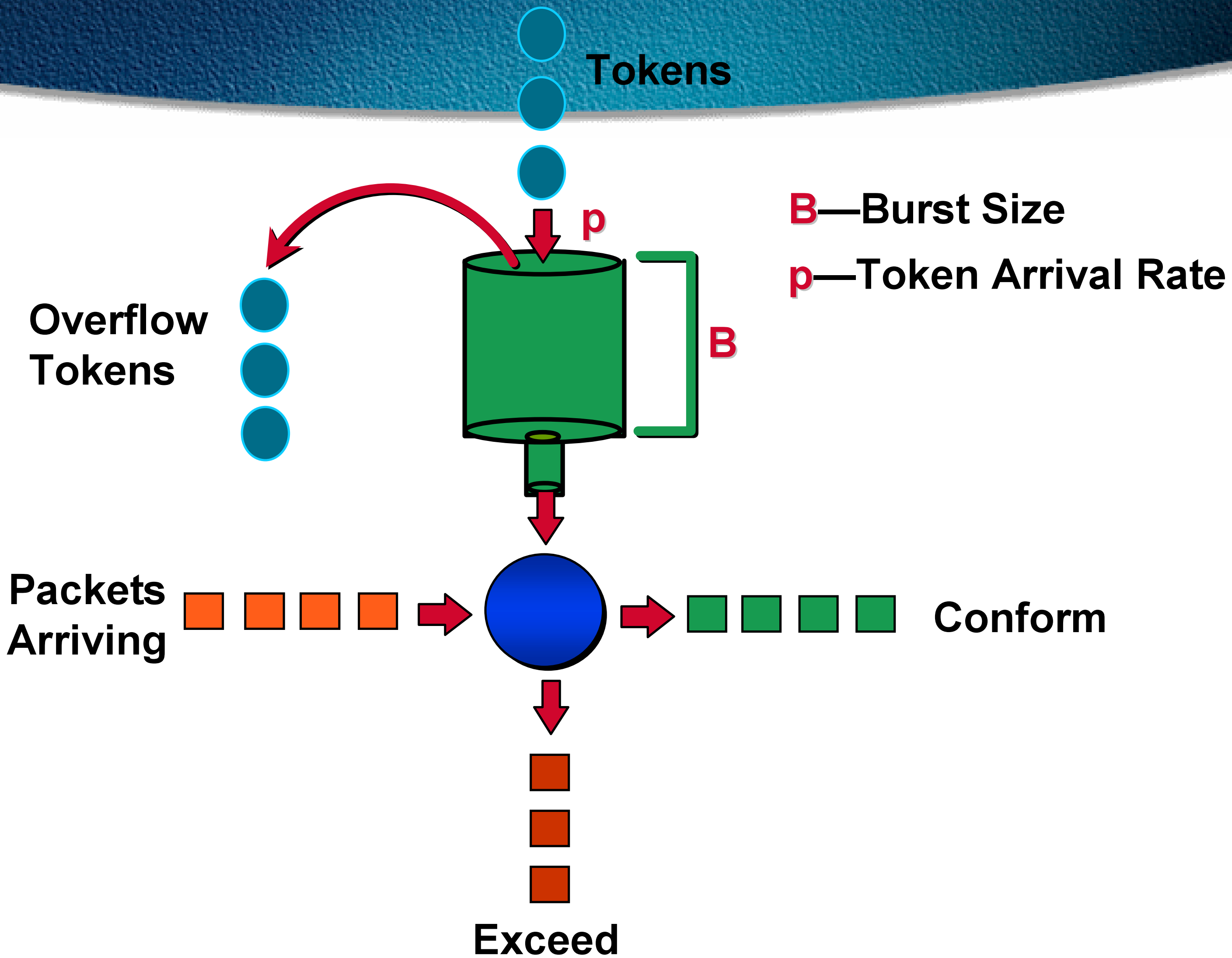
- Packets arriving exceeding the normal burst but fall within the extended burst limit are handled via a RED-like managed drop policy
- This reduces TCP Slow-Start oscillation

(When the exceed-action is to drop packets)

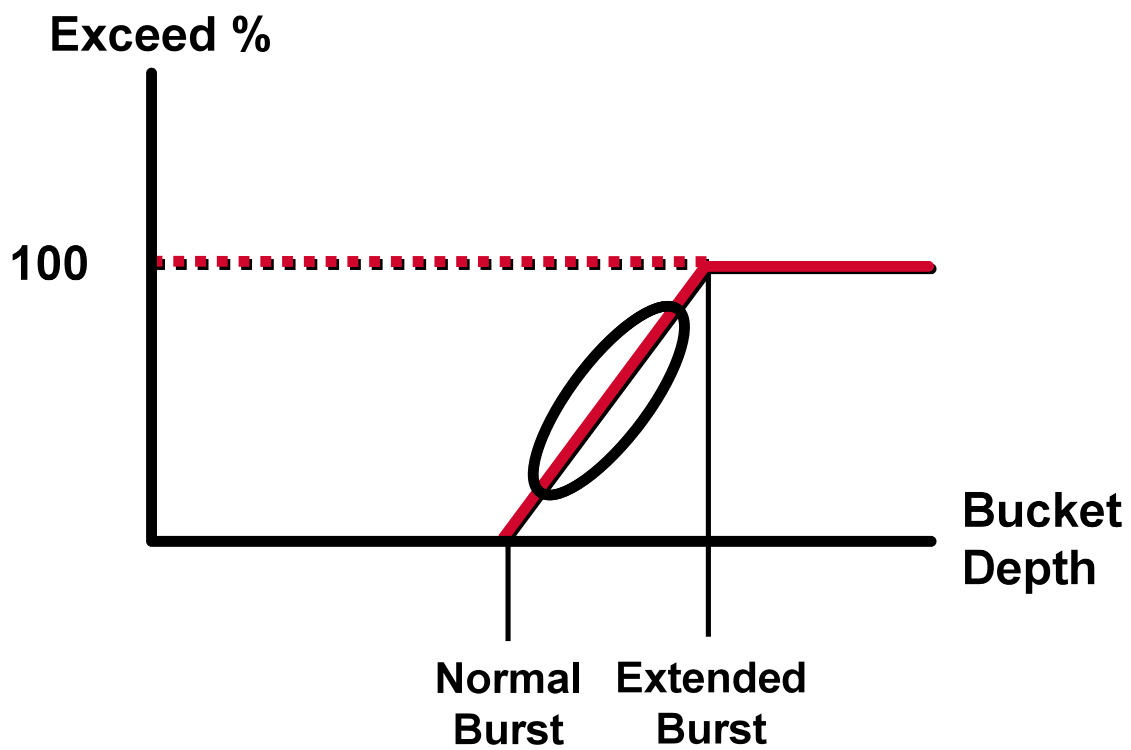
CAR—Traffic Measurement

- **Token bucket configurable parameters**
  - Committed rate (bits/sec)
    - Configurable in increments of 8Kbits
  - Normal burst size (bytes)
    - To handle temporary burst over the committed rate limit without paying a penalty
  - Extended burst size (bytes)
    - Burst in excess of the normal burst size

Token Bucket



Extended Burst



CAR—Action Policies

- Configurable actions

  Transmit

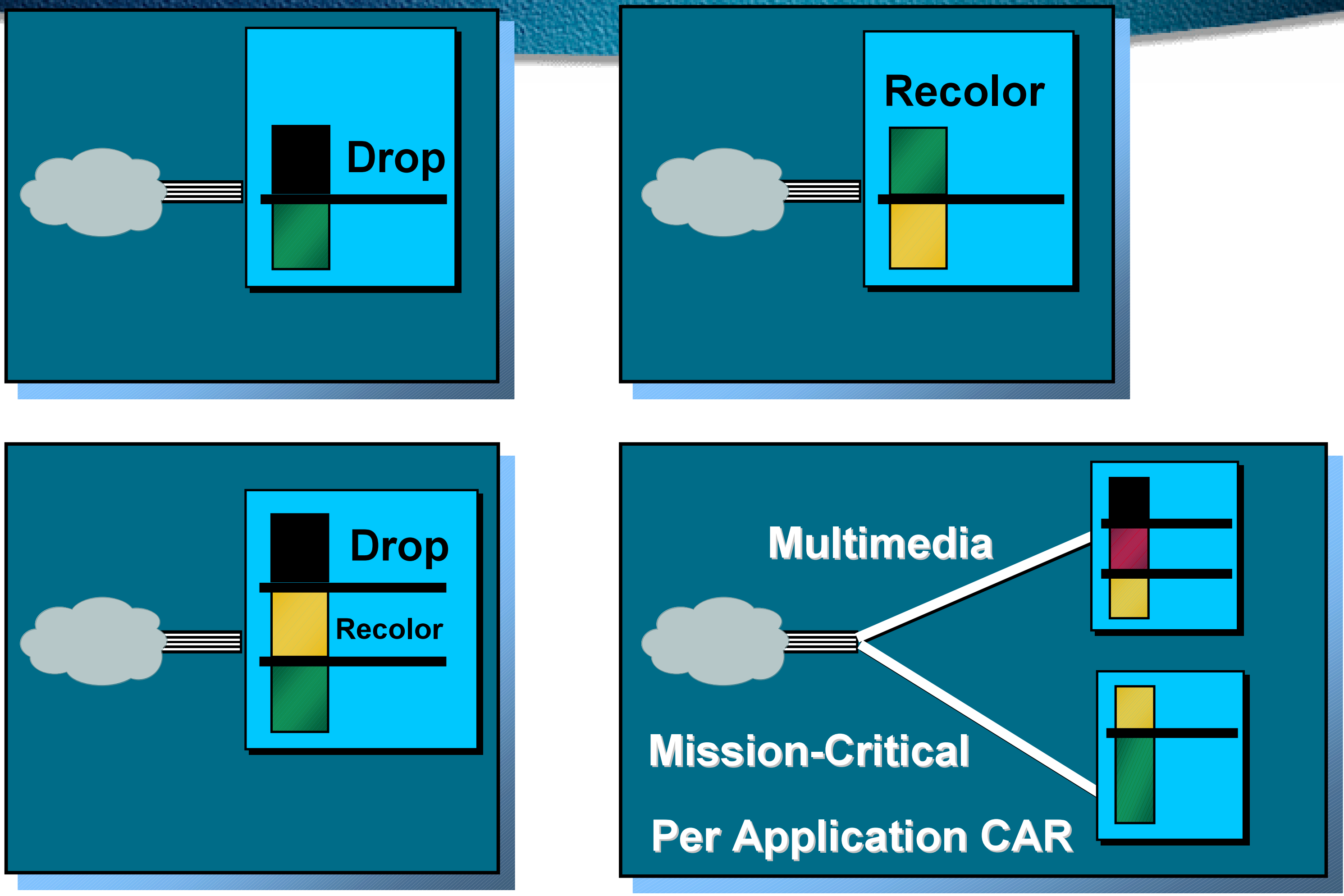
  Drop

  Continue (go to the next rate-limit in the list)

  Set precedence and transmit
  (rewrite the IP precedence bits and transmit)

  Set precedence and continue
  (rewrite the IP precedence bits and go to the next rate-limit in the list)

CAR—Policy Examples



Packet Marking

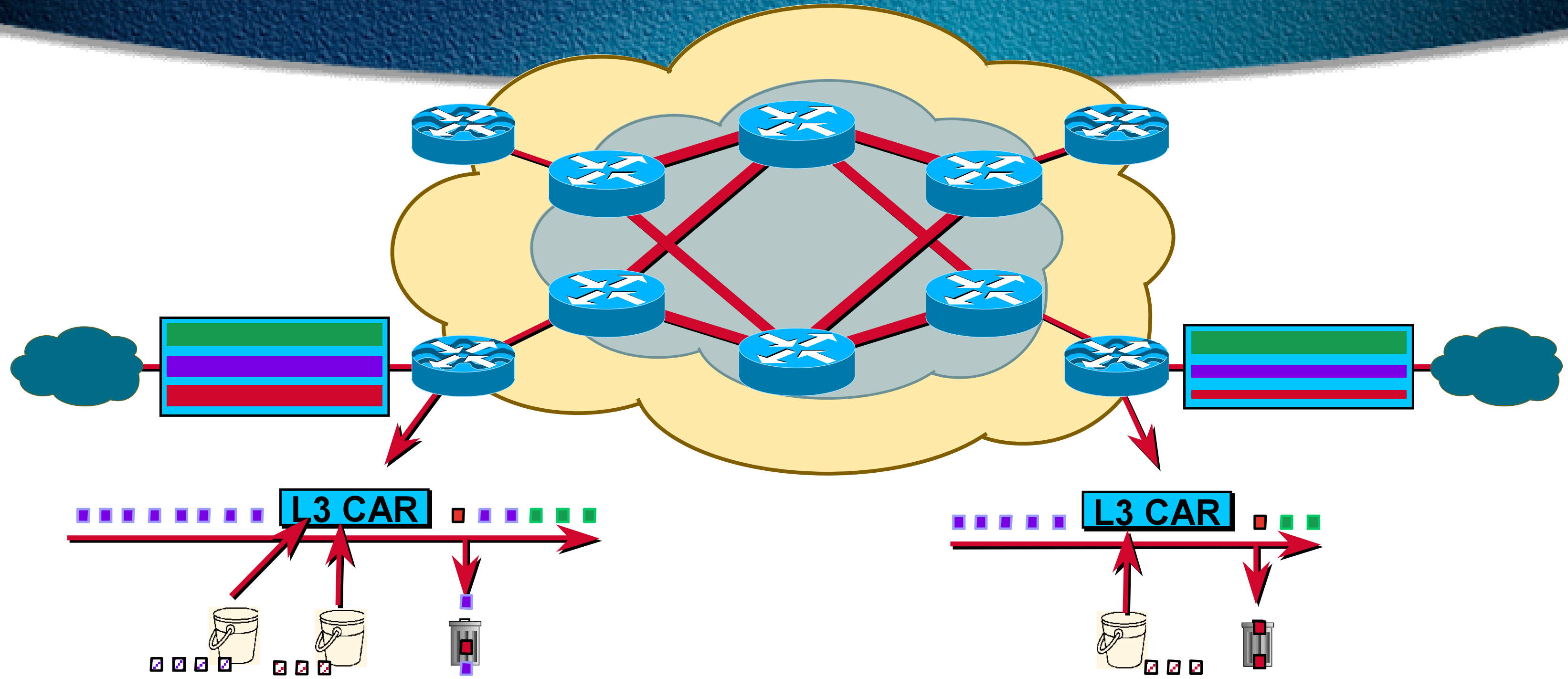
- Also known as colouring or labeling of packets
- Partition network traffic into multiple traffic classes or Class of Service (CoS)
- Marking can be done using several methods

    CAR

    QoS Policy Propagation via BGP

    Policy routing

CAR



Ingress Router

Packet Classification
Rate Limiting
  Committed Rate
  Burst
  RED-like Managed Drop

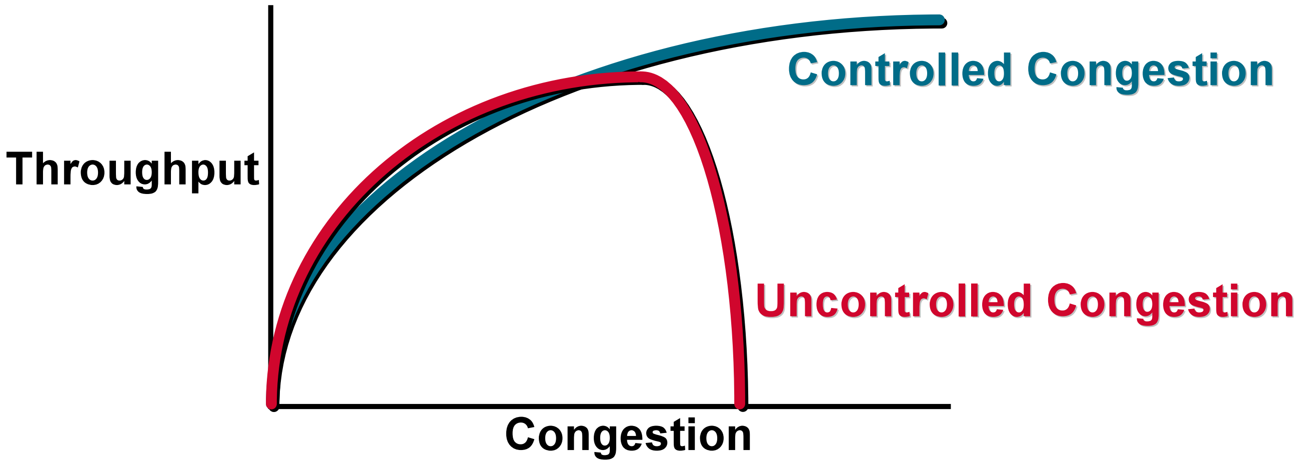
Egress Router

Packet Classification
(Reset Precedence Bits)
Rate Limiting
  Committed Rate
  Burst
  RED-like Managed Drop

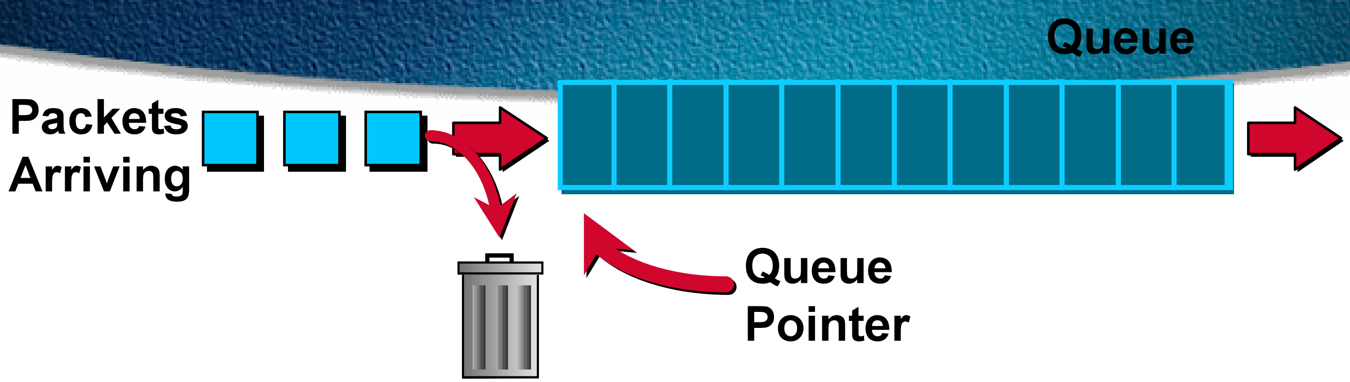
RED

The Problem of Congestion in TCP/IP

- Uncontrolled, congestion will seriously degrade system 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 until the system collapses



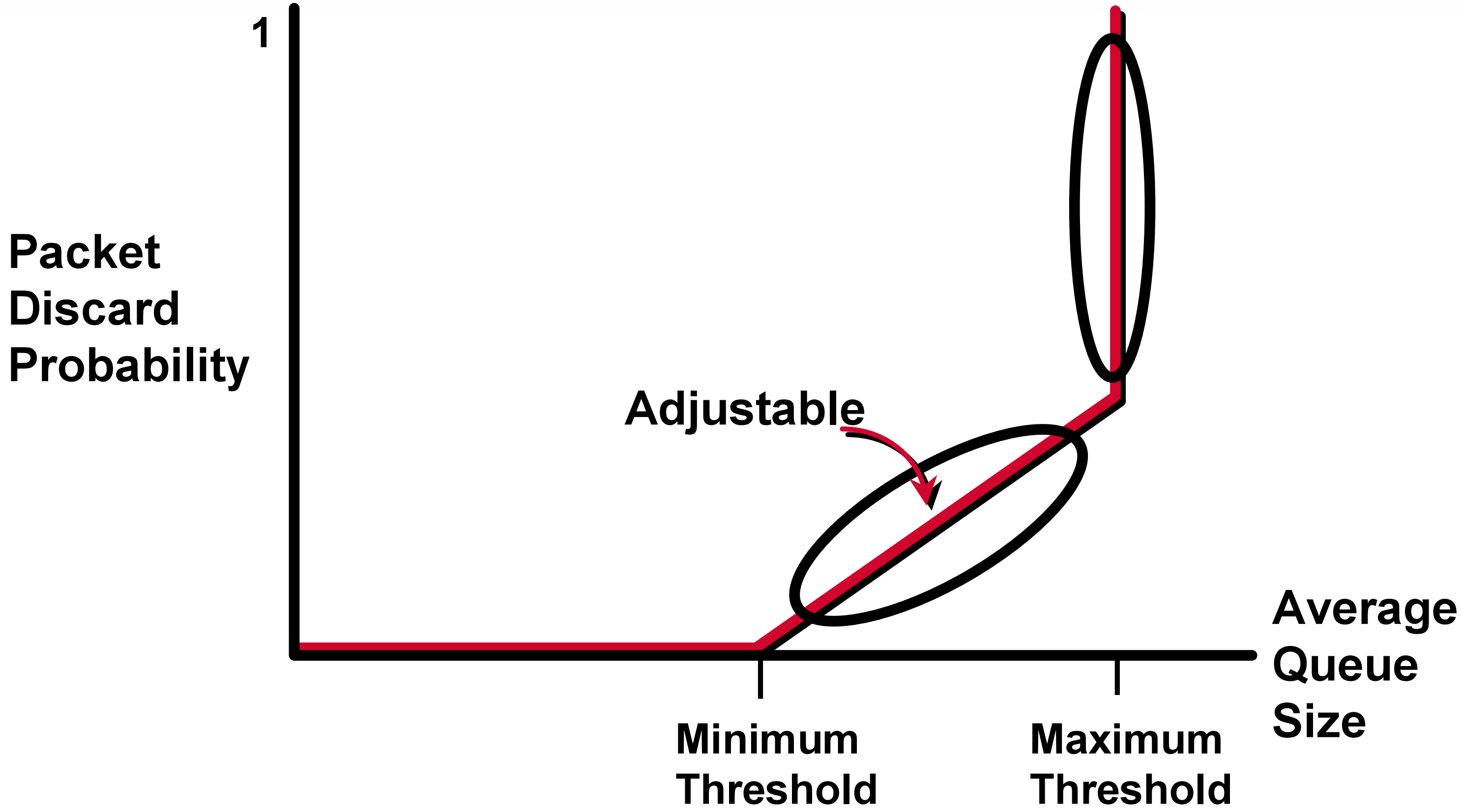
RED



- **Without RED when the queue fills up all packets that arrive are dropped—Tail drop**
- **With RED as opposed to doing a tail drop the router monitors the average queue size**

  using randomization chooses flows to notify that a congestion is impending

RED


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

RED—Average Queue Size

- Used to determine the degree of burstiness that will be allowed in the queue
- The average queue size is calculated based on the previous average and the current size of the queue

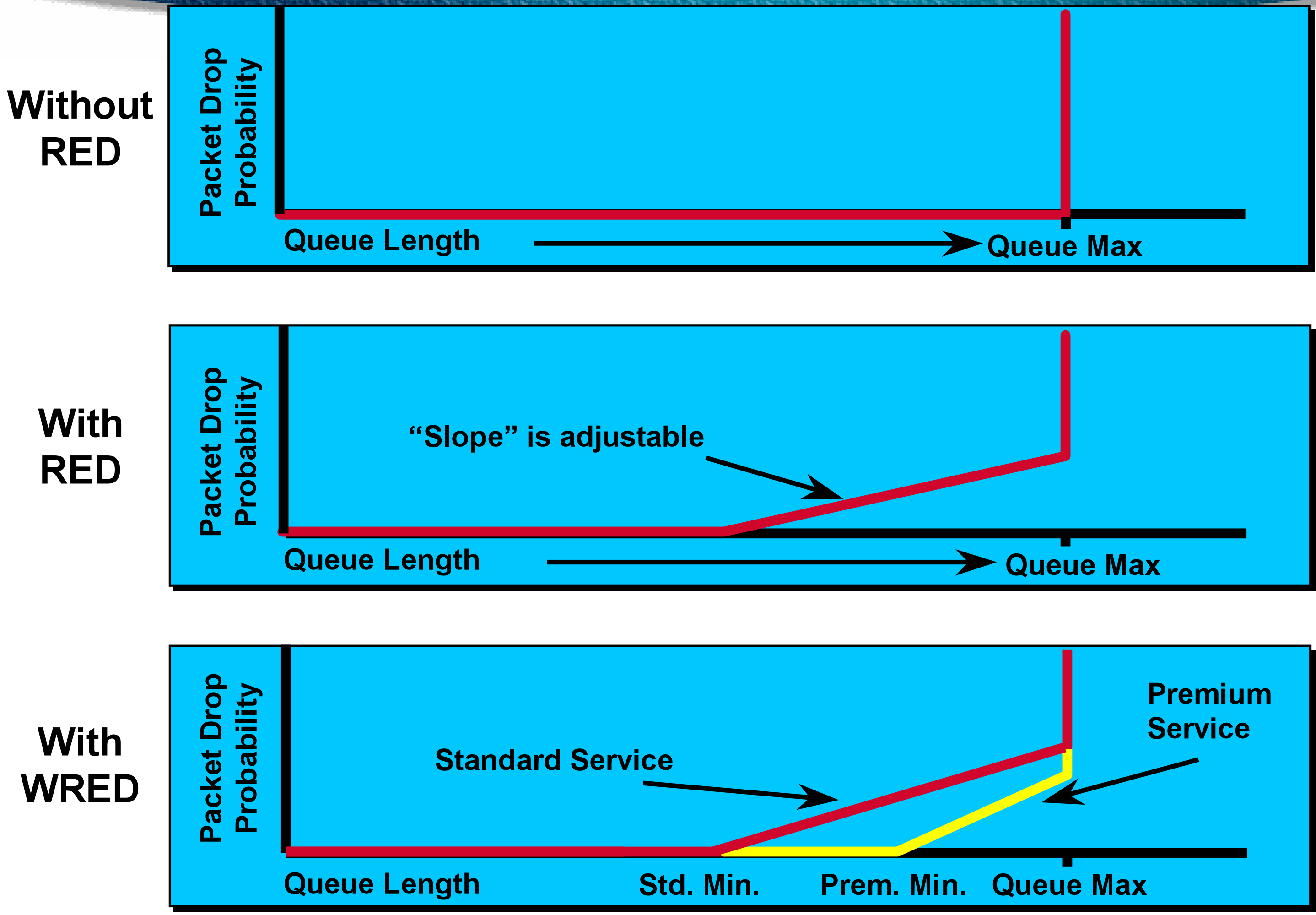
  Avg = (old_avg * (1 - 1/2^weight)) +
  (current_queue_size * 1/2^weight)

- Where weight is the exponential-weight-constant

Weighted RED (WRED)

- WRED combines RED with IP Precedence to implement multiple service classes
- Each service class has a defined min and max threshold, and drop rates
- In a congestion situation lower class traffic is throttled back first before higher class traffic
- RED is applied to all levels of traffic to manage congestion

Weighted RED



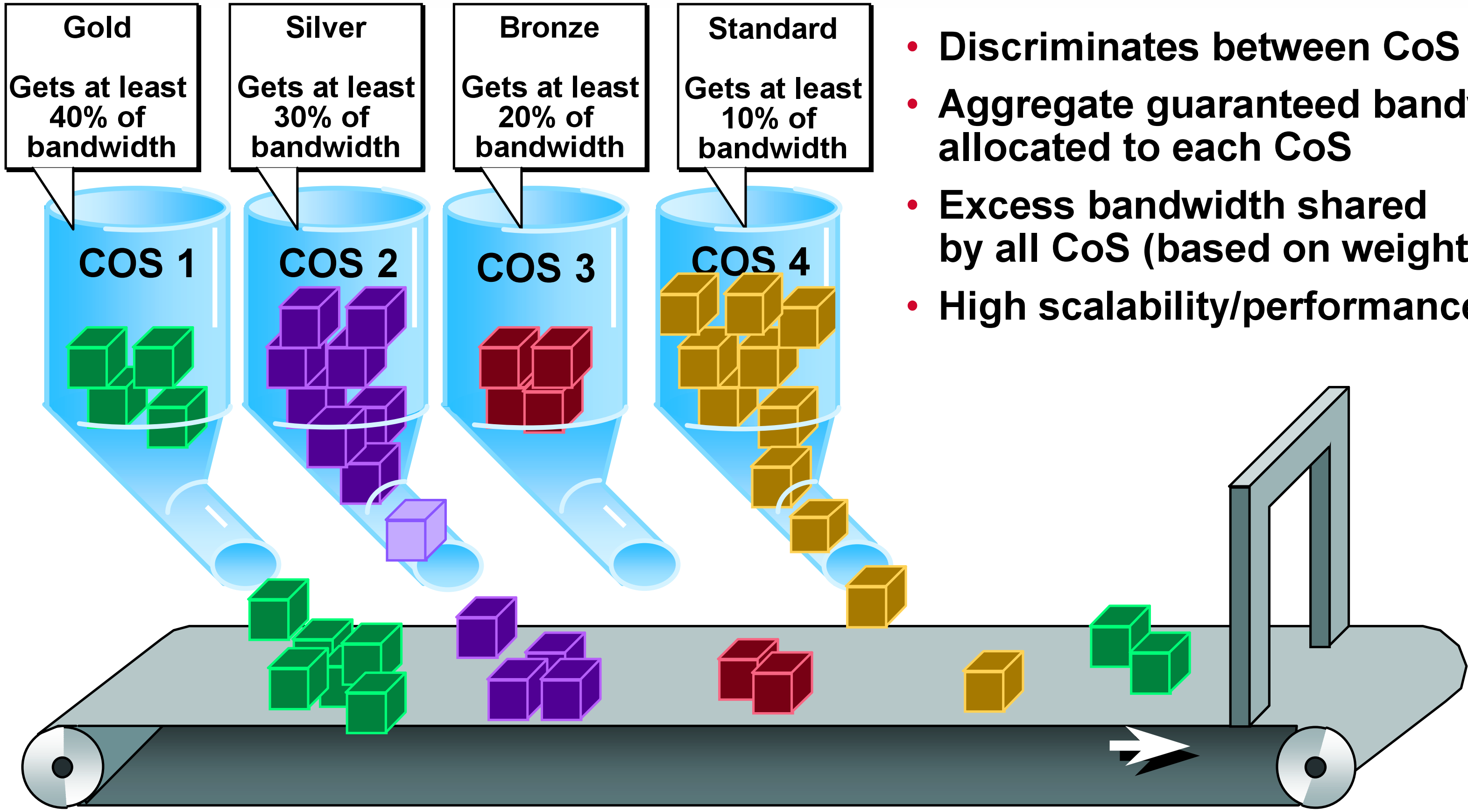
Where/When Should I Use WRED?

- Congested long-haul links (e.g., trans-oceanic links)
- Where the bulk of your traffic is TCP as oppose to UDP

  Remember only TCP will react to a packet drop UDP will not

WFQ

Weighted Fair Queuing



- Discriminates between CoS
- Aggregate guaranteed bandwidth allocated to each CoS
- Excess bandwidth shared by all CoS (based on weight)
- High scalability/performance

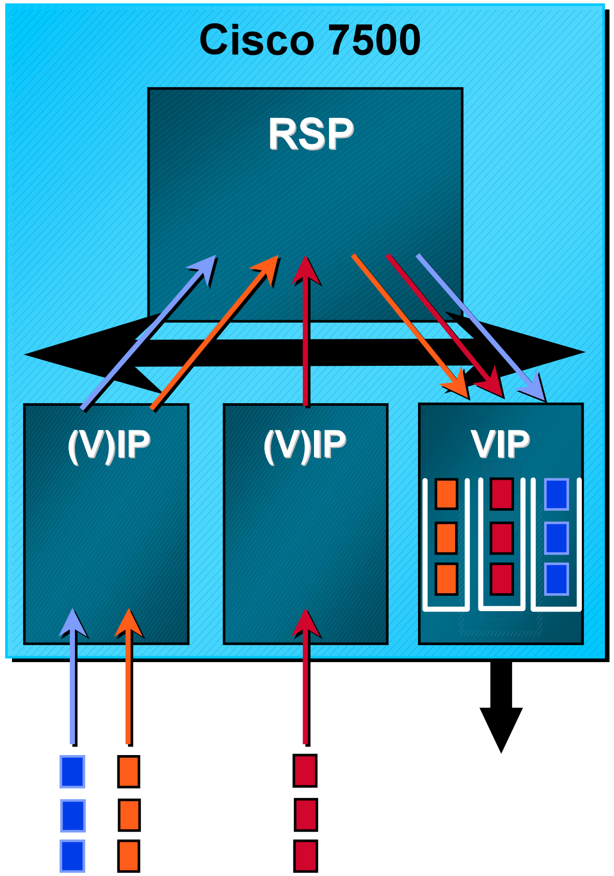
Packet Scheduling

- **An algorithm that determines the order in which packets are sent out to the transmission link**
- **Examples of packet scheduling schemes**

  FIFO

  Fair Queuing

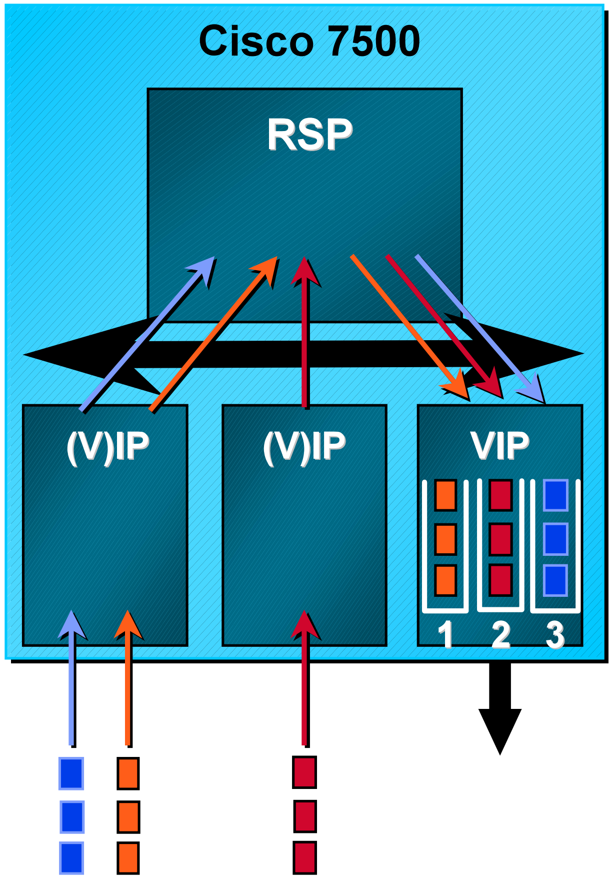
  Weighted Fair Queuing



Weighted Fair Queuing (WFQ)

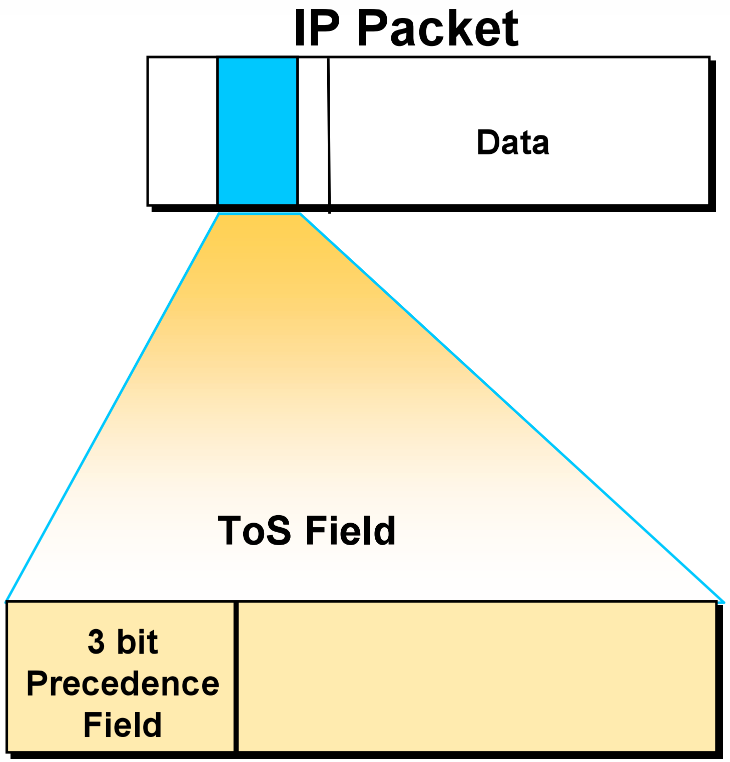
- Multiple “logical” queues
- Assign a weight for each queue
- Backlog queues are served in proportion to their weight
- Ideal WFQ

  packet of a queue transmitted at least at same time it would be transmitted if it had its own interface with speed of queue service rate



IP Precedence

"Controlling WFQ's De-queuing Behavior"



$$\text{Weight} = \frac{4096}{(1 + \text{IP Prec})}$$

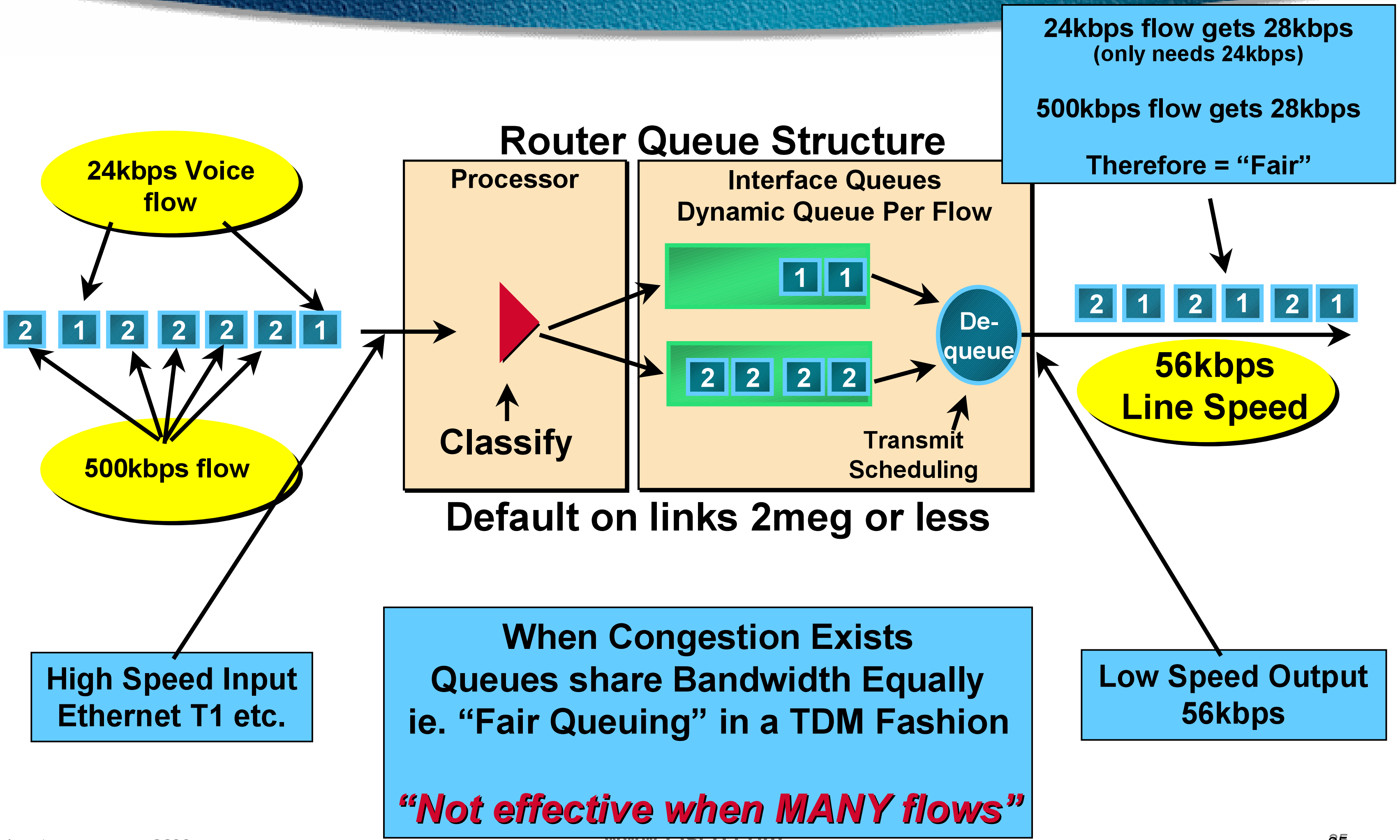
IP Prec	Weight
0	4096
1	2048
2	1365
3	1024
4	819
5	682
6	585
7	512

IP Precedence

Not a QoS Mechanism turned on in the router

"In Band" QoS Signaling - Set in the End Point

Weighted Fair Queuing (WFQ)
Treats Flows with same IP Precedence Equally



Why Use WFQ?

- **Provides relative bandwidth guarantees**

  Fair Queuing (FQ) **allocates equal share of bandwidth to each active queue**

  Weighted Fair Queuing (WFQ) **allows for unequal allocation of bandwidth**

Why Use WFQ?

- **Provides absolute bandwidth guarantees**

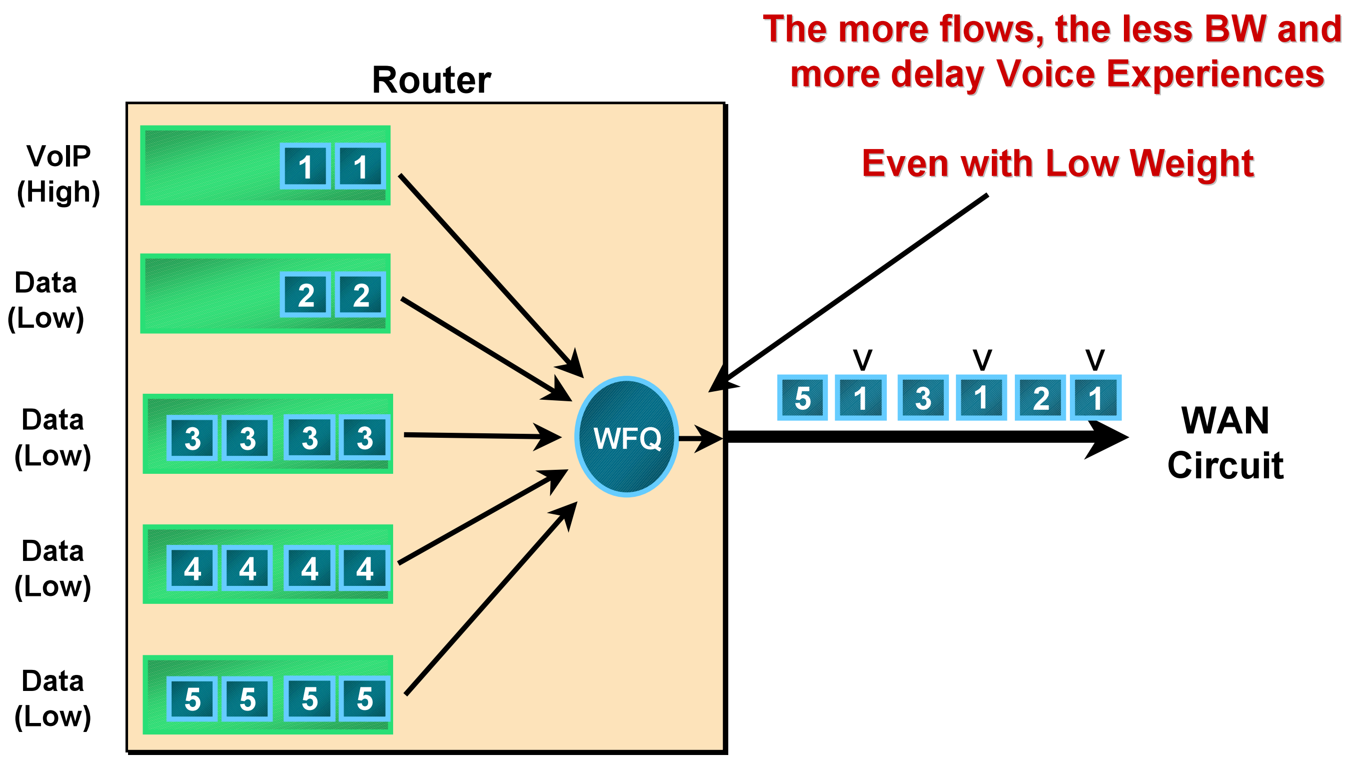
  Good for real-time applications (e.g., audio/video) and bandwidth provisioning

  But requires cooperation of :

  –admission control

  –traffic engineering

Problem:
WFQ Cannot provide Strict Prioritization



More evident the slower the link

IP Precedence

Calculating given Flow BW based on IP Precedence under congestion

$$\textit{Flow A BW} = \left(\frac{\textit{Flow A "Parts"}}{\text{Sum of all Flow "Parts"}}\right) \times \text{Circuit BW}$$

***Individual Flow "Parts"* = 1 + IP Precedence**

IP Precedence	Flow "Parts"
0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8

IP Precedence

Flow BW Calculation Example

$$\text{Flow A BW} = \left(\frac{\text{Flow A “Parts”}}{\text{Sum of all Flow “Parts”}}\right) \times \text{Circuit BW}$$

Example A

56kbps link

2 - VoIP Flows A+B at 24kbps (IP Prec 0)
2 - FTP flows at 56kbps (IP Prec 0)

$$\text{14kbps} = \left(\frac{1}{4}\right) \times \text{56kbps}$$

14kbps **NOT** suitable for a 24kbps flow
Example of many Flows with WFQ and equal Precedence flows

Weighted “Fair” Queuing

Example B

56kbps link

2 - VoIP Flows A+B at 24kbps (IP Prec 5)
2 - FTP flows at 56kbps (IP Prec 0)

$$\text{24kbps} = \left(\frac{6}{14}\right) \times \text{56kbps}$$

24kbps **SUITABLE** for a 24kbps flow

WFQ preferring IP Precedence

IP Precedence
No Admission Control

Moral of the story: Know your environment, Voice traffic patterns etc.
Recommendations for certain Bandwidth's to Follow

Example C

56kbps link
2 - VoIP Flow's at 24kbps (IP Prec 5)
4 - FTP flows at 56kbps (IP Prec 0)

21kbps = $\left(\frac{6}{16}\right)$ X 56kbps

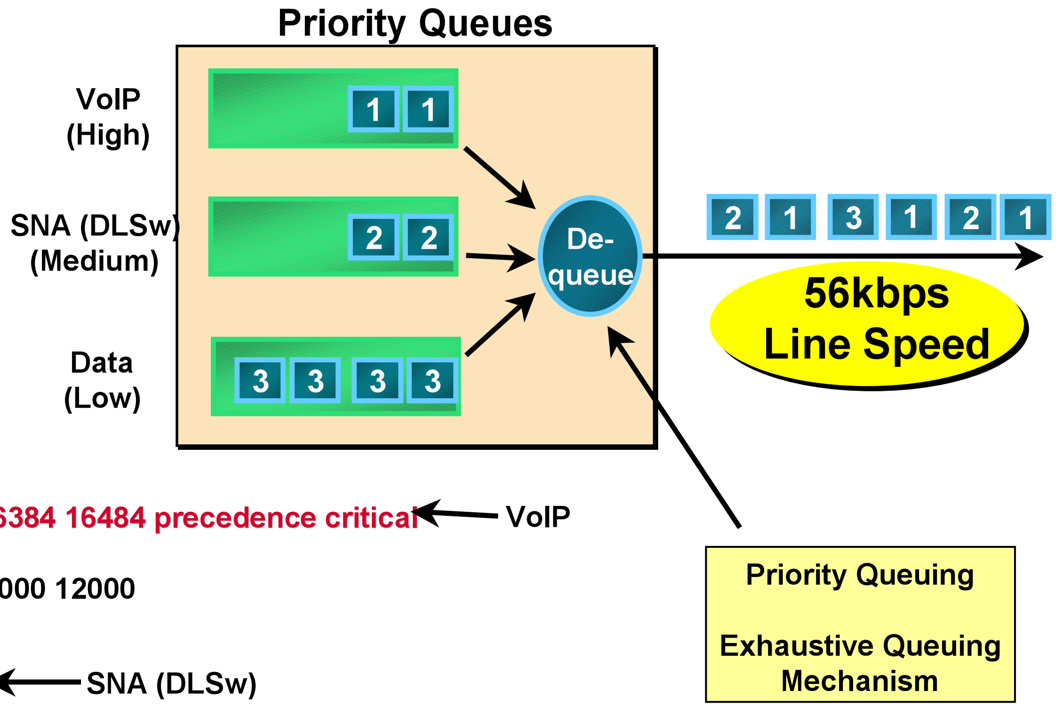
21kbps NOT suitable for a 24kbps flow

RTP Header Compression would help since it would reduce VoIP flow to 11.2kbps
Also RSVP or CBWFQ

Options - Priority Queuing

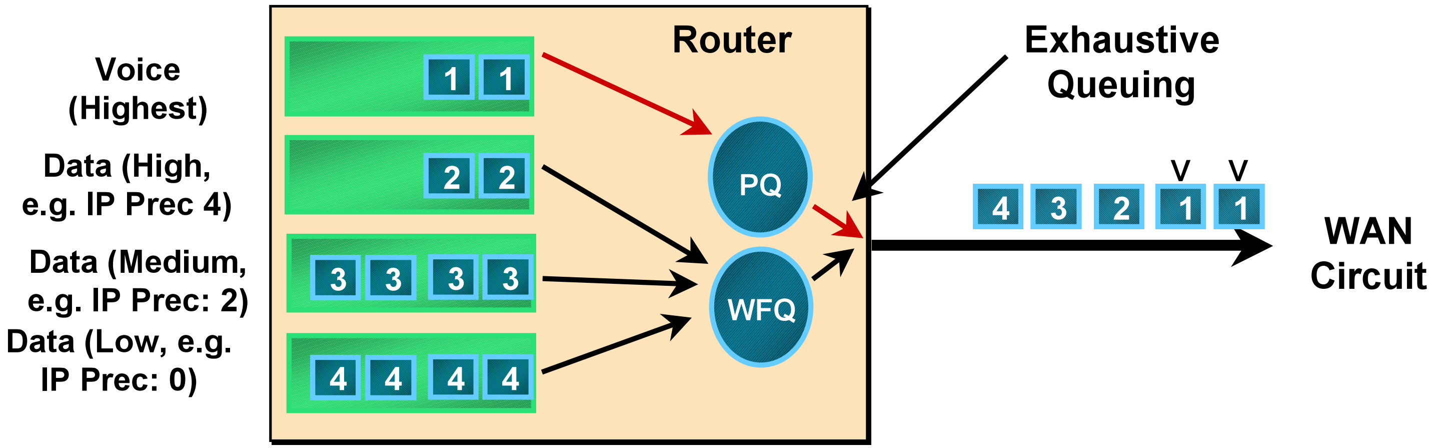
Issues - Cannot run with FRF.12 or MLPPP
Warning - IP MTU Size Reduction Breaks Things
Less than 600bytes breaks IP Phone TFTP download
other protocols still can cause delay

```
interface Serial0/0.1 point-to-point
 ip address 10.1.1.1 255.255.255.0
 ip mtu 300
 frame-relay interface-dlci 100
 class voice64
!
map-class frame-relay voice64
 frame-relay cir 56000
 frame-relay bc 1000
 no frame-relay adaptive-shaping
 frame-relay priority-group 1
!
access-list 101 permit udp any any range 16384 16484 precedence critical   <-- VoIP
access-list 101 permit tcp any any eq 1720
access-list 101 permit tcp any any range 11000 12000
access-list 106 permit icmp any any
priority-list 1 protocol ip low list 106
priority-list 1 protocol ip medium tcp 2065   <-- SNA (DLSw)
priority-list 1 protocol ip high list 101
```



PQ-WFQ (IP RTP Priority)

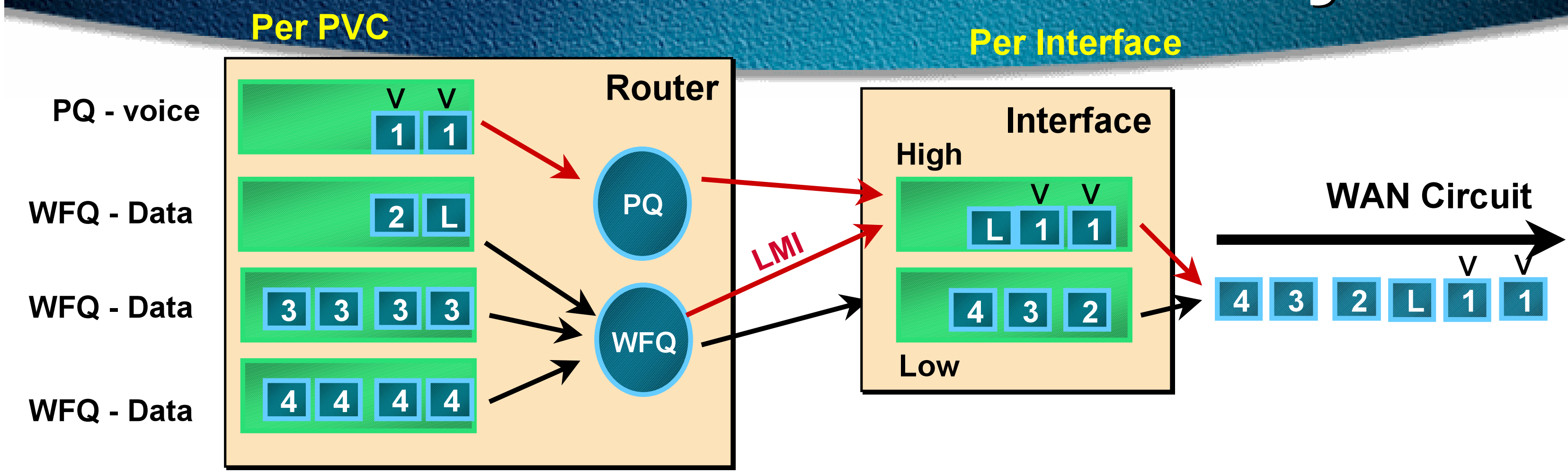
- Queue-limit for PQ is 64
  - No CLI to change
- Packets exceeding the allocated BW are dropped
- WFQ for:
  - non-RTP traffic
  - RTP traffic outside given port range



Obsoletes/Replaces the use of IP RTP Reserve

Prioritization - Queuing

PQ-WFQ for Frame Relay



- Dual Interface FIFO is turned on automatically when FRTS is configured
- Interface Queue operation before PQ-WFQ in 12.0(5)T:
  - LMI and unfragmented packets (VoIP, VoFR) to Hi queue
  - All fragmented packets (data) to Lo queue
- Interface Queue operation with PQ-WFQ in 12.0(5)T:
  - LMI and PQ contents (VoIP and VoFR) to Hi queue
  - Everything else, regardless of fragmentation (data) to Lo queue

Class Based Weighted Fair Queuing CBWFQ - 12.0(5)T

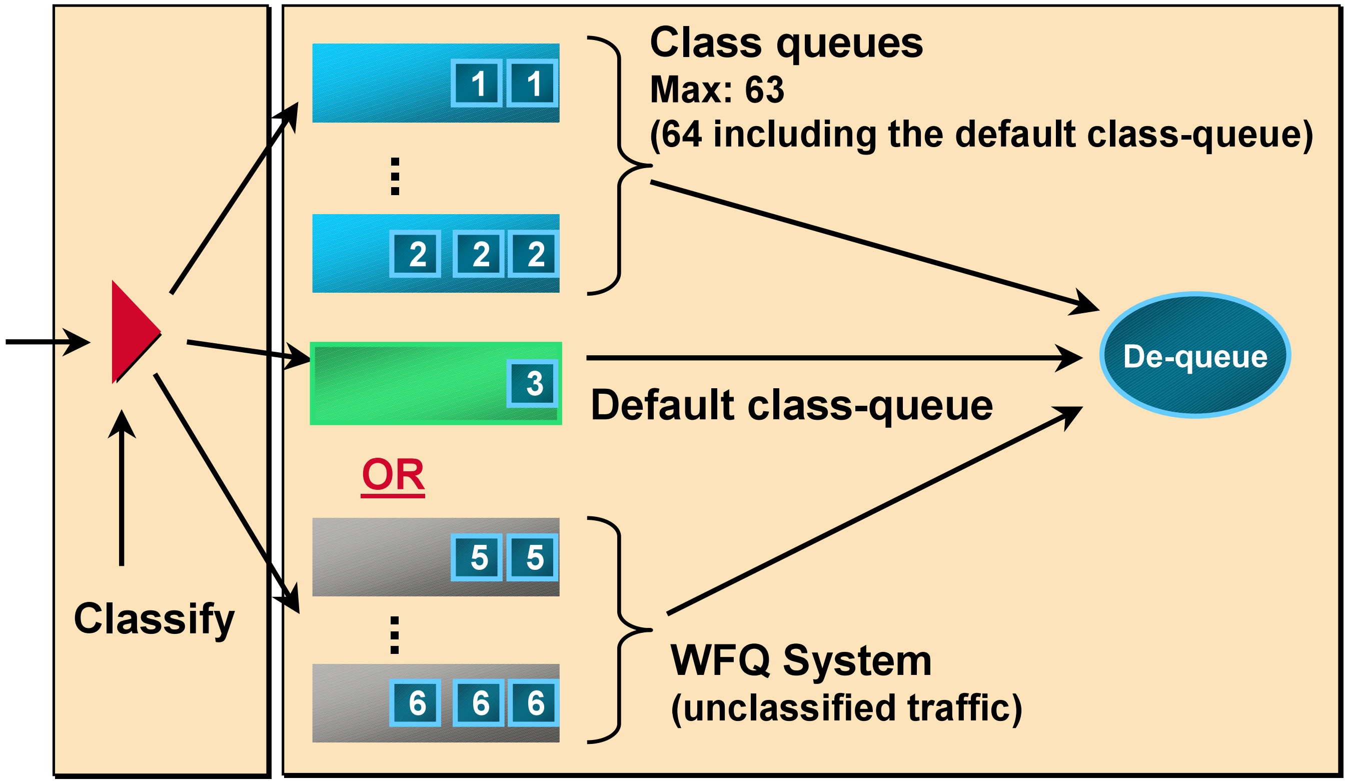
Queues represent "Classes" that have
an associated minimum bandwidth in kbps

Traffic assigned to classes via a "policy-map"

Max 64 classes which support:
WFQ between classes
RED per class

Class-Based WFQ (CBWFQ)

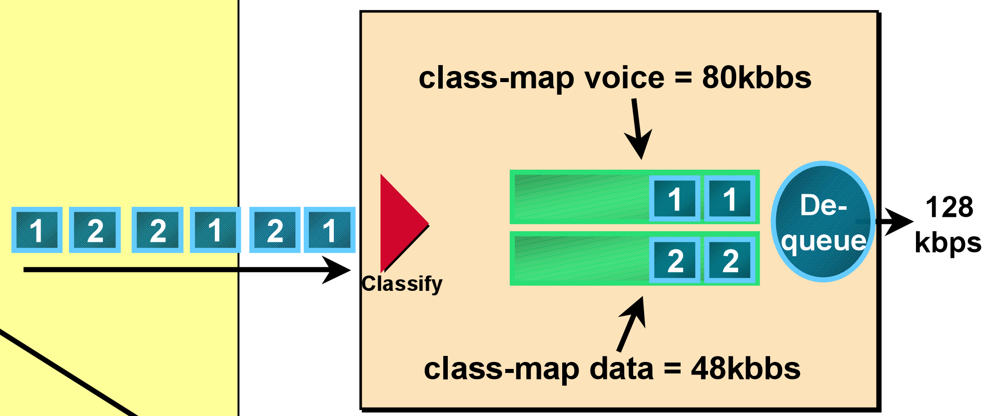
12.0(5)T



Class Based Weighted Fair Queuing
CBWFQ

```
class-map data
  match input-interface Ethernet0/0
class-map class-default
  match any
class-map voice
  match access-group 101
!
!
policy-map WAN
  class voice
   bandwidth 80
  class data
   bandwidth 48
!
interface Serial0/1
 ip address 10.1.6.2 255.255.255.0
 bandwidth 128
 no ip directed-broadcast
 service-policy output WAN
!
access-list 101 permit ip any any precedence critical
```

Class Based WFQ



Any packet with IP Precedence = 5 gets assigned to a class that will get a minimum of 80kbps on a 128kbps circuit

NOTE: Does not work on sub-interfaces and with Frame Relay Traffic Shaping

WFQ vs. CBWFQ

WFQ

IP Prec	Weight	Num Flows	BW %	BW (K)
6	585	1	30.4%	19.5
5	682	2	52.2%	33.4
0	4096	4	17.4%	11.1
		23*	100%	64K

- Tail-drop if queue fills up
- Weights given; BW derived
- RSVP gives weight of 4
- No BW guarantee
- No limit on incoming traffic

CBWFQ

Class	Weight	Num Flows	BW %	BW (K)
A	385	1	9.4%	6
B	1536	2	37.5%	24
C	2175	4	53.1%	34
			100%	64K

- Tail-drop/RED for excess traffic
- BW given; weights derived
- No RSVP Support yet
- BW guarantee
- Default: 75% of BW allocatable
- Classification on ACLs
- Policing on incoming traffic

* ↗ flow-parts = (1x(6+1)) + (2x(5+1)) + (4x(0+1)) = 23

** Calculations done for 4096 base weight

Varies / Derived

Fixed / Configured

Flow-Based DWFQ

- A flow ID is computed for each packet

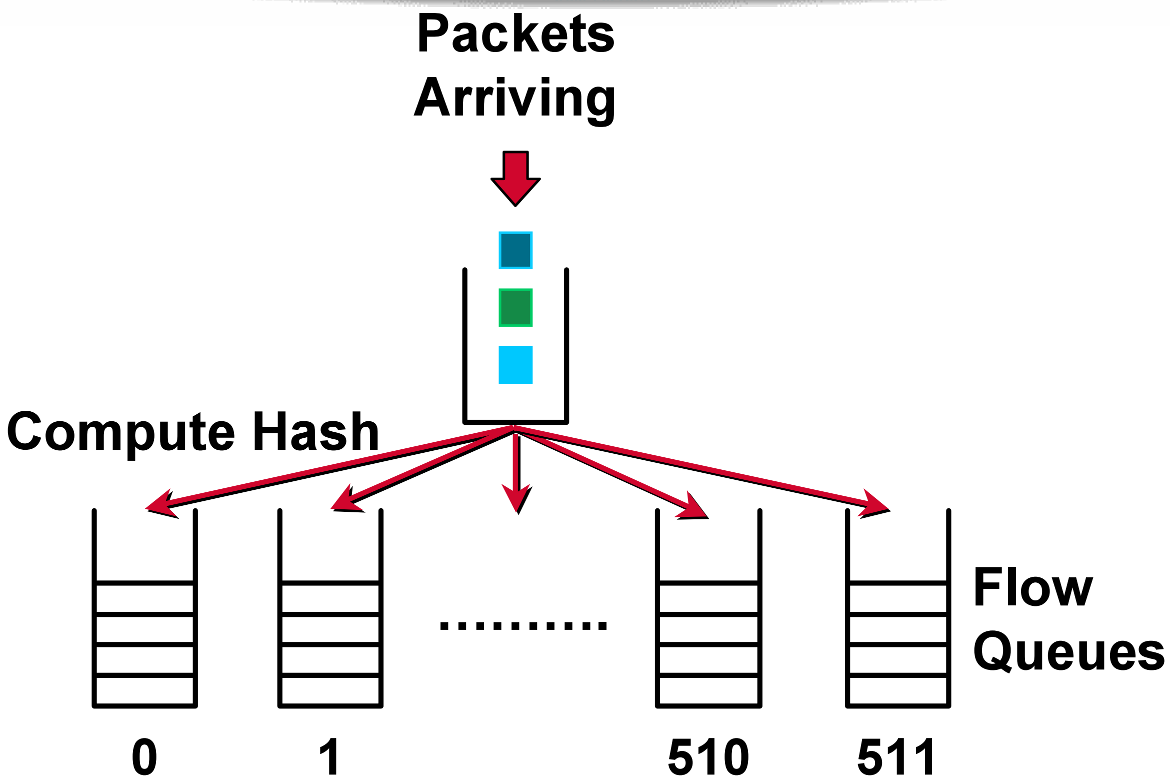
  The flow ID is a hash computed on source and destination IP address, source and destination TCP/UDP port, protocol and ToS field

- Based on the flow ID the packet is then classified to the appropriate queue

  there are a total of 512 queues for each interface

- Each active queue is allocated an equal share of the bandwidth

Flow-Based DWFQ



ToS-Based DWFQ

- Packets are classified into 4 queues based on IP precedence

  Follows directly from the precedence value

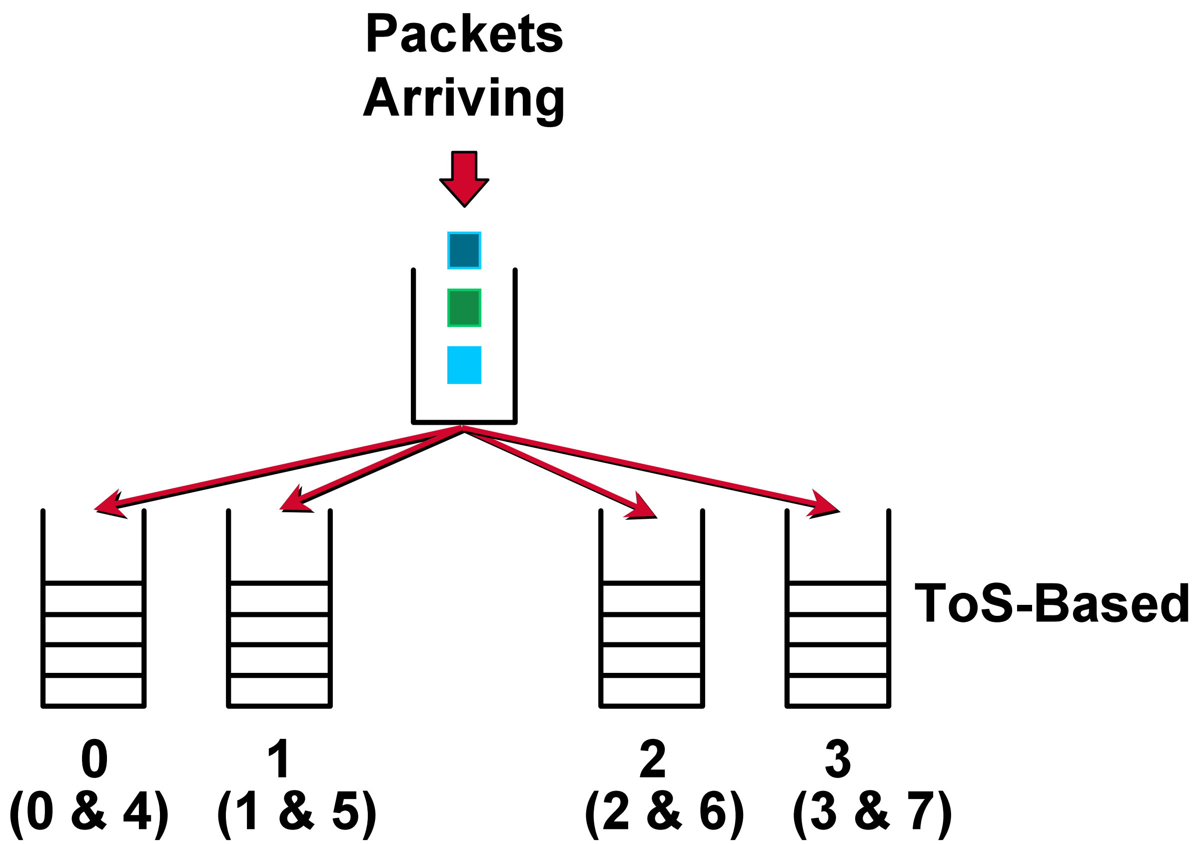
  but MSB of Precedence ignored for queue selection, used as “in/out” bit

- Each queue is weighted

  Expressed in percentage (%)

- The weight determines the amount of bandwidth that each active queue is allowed to consume during periods of congestion

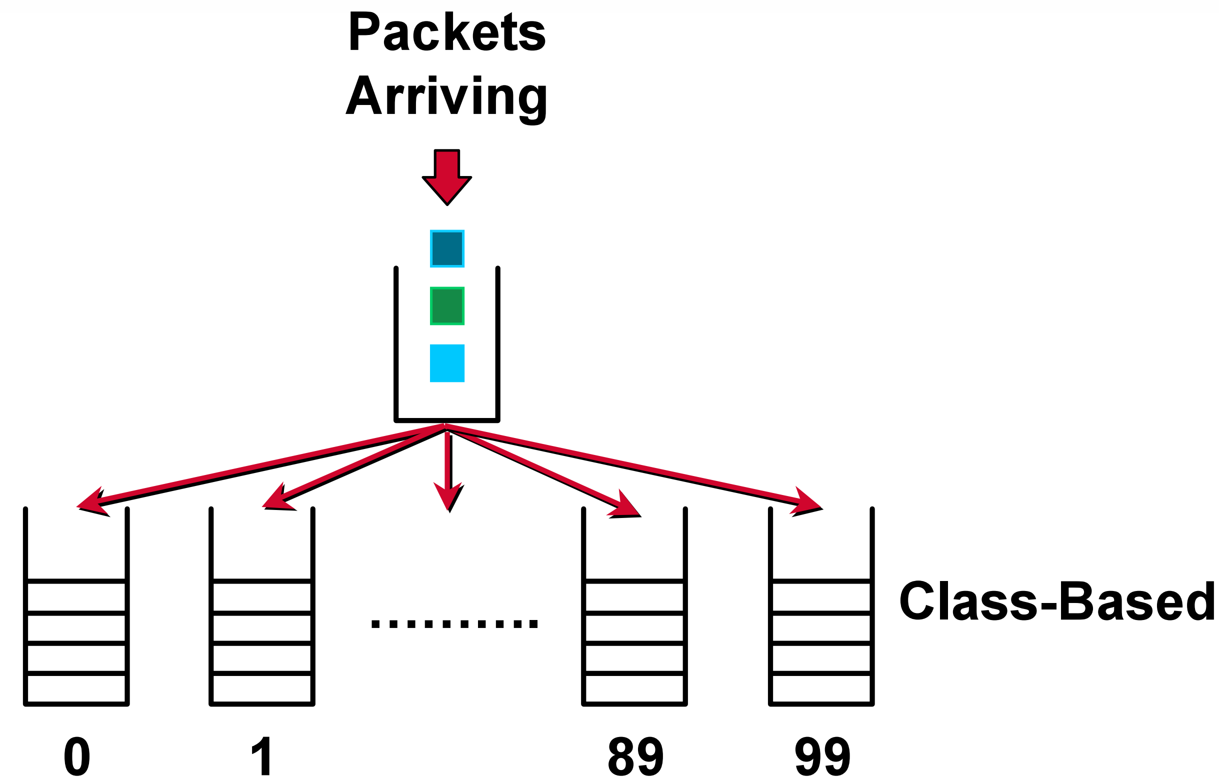
ToS-Based DWFQ



QoS-Group-Based DWFQ

- Packets are classified into different queues based on their QoS group
  - QoS group is set using CAR or QoS policy progation via BGP
- Each queue is weighted
  - Expressed in percentage (%)
- The weight determines the amount of bandwidth that each active queue is allowed to consume during periods of congestion

QoS-Group-Based WFQ



DWFQ queue size parameters

- The user has the option to set

  aggregate queue depth

  individual queue depth

- During periods of congestion the individual queue depth limit is enforced

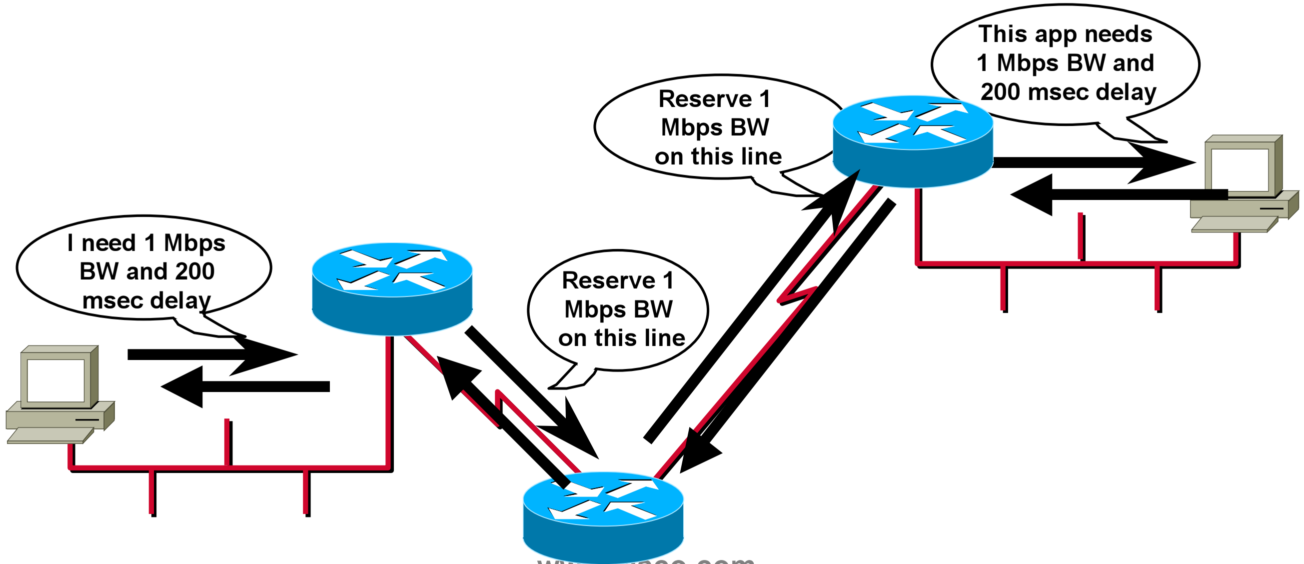
- With ToS and QoS group-based WFQ the user also has the option to set the queue depth for each ToS or QoS group queues

Combination of WFQ / WRED

- Both mechanisms can work together and in conjunction
- Guaranteed delay for real time applications (UDP/RTP, H.323)
- Differentiation on drop probability and drop threshold for bursty data traffic (WWW, FTP, SMTP…)

RSVP

- Allows users of multimedia apps to reserve network resources and guarantee end-to-end quality of service
- Controlled from router configuration and host applications that implement RSVP
- Enables coexistence of multimedia applications (real-time) with sporadic applications
- Supports both unicast and multicast flows

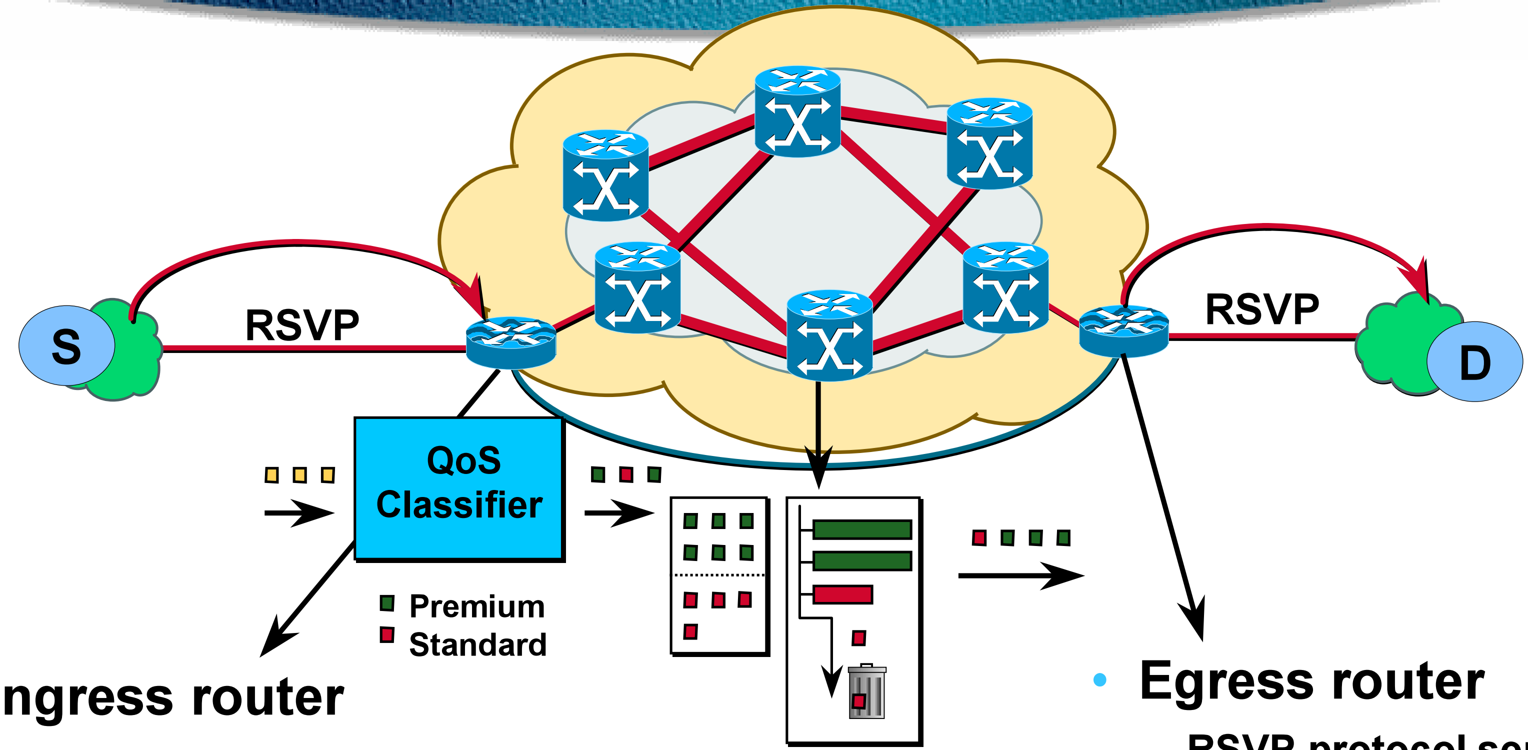


“Integrated Service” model
Scaling RSVP

- **See IETF “RSVP Applicability Statement”**
- **RSVP not scalable to large Internet networks**
- **Requirement for Vendor specific aggregation schemes :**

Intserv-Diffserv integration

IntServ-DiffServ Integration



- **Ingress router**
  - RSVP protocol
  - Mapped to QoS level (DS)
  - Forwarded to egress
- **Backbone routers**
  - WFQ and/or WRED applied on CoS
- **Egress router**
  - RSVP protocol sent on to destination

Link Efficiency

Low Speed WAN QoS Tools

- Fragmentation and Interleave (LFI)
- RTP Header Compression (CRTP)

Fragmentation and Interleave

Only needed on slow links

Real-Time MTU | Elastic Traffic MTU

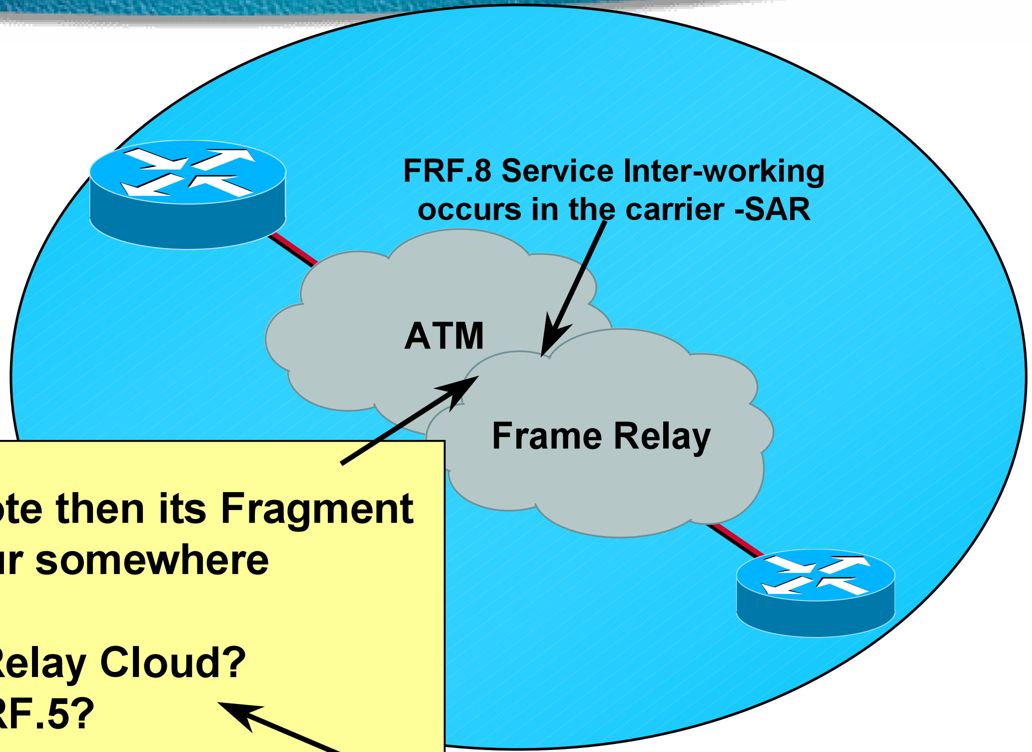
214ms Serialization Delay
for 1500 byte Frame at 56 kbps

Elastic MTU | Elastic MTU | Real-Time MTU | Elastic MTU

Mechanisms

<u>Pt to Pt Links</u> - MLPPP with Fragmentation and Interleave
<u>Frame Relay</u> - FRF.12 (Voice and Data can use Single PVC)
<u>ATM</u> - (Voice and Data need separate VC's on Slow Links)

Issue - FRF.8 ATM/FR Service Inter-working



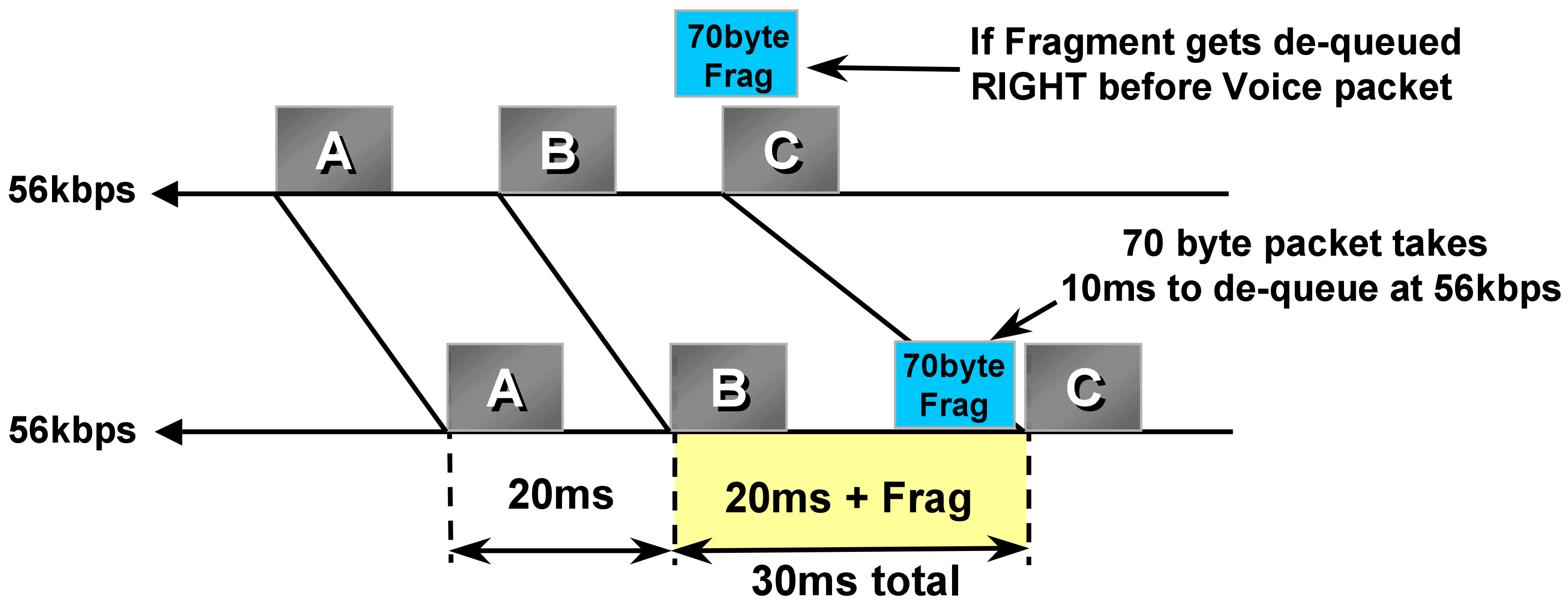
Note: If FRF.12 needed at remote then its Fragment re-assembly must occur somewhere

FRF.12 in the Frame Relay Cloud?
What about FRF.5?

Two PVC's required for Interleaving
ATM can not interleave cells from different packets

No one is doing this
High Risk Solution

Setting Fragment Size
Based on Minimum Desired Blocking Delay



Note: Blocking delays are always present

When is Fragmentation Needed?

Frame Size

Link Speed	1 Byte	64 Bytes	128 Bytes	256 Bytes	512 Bytes	1024 Bytes	1500 Bytes
56kbps	143us	9ms	18ms	36ms	72ms	144ms	214ms
64kbps	125us	8ms	16ms	32ms	64ms	128ms	187ms
128kbps	62.5us	4ms	8ms	16ms	32ms	64ms	93ms
256kbps	31us	2ms	4ms	8ms	16ms	32ms	46ms
512kbps	15.5us	1ms	2ms	4ms	8ms	16ms	23ms
768kbps	10us	640us	1.28ms	2.56ms	5.12ms	10.24ms	15ms
1536kbs	5us	320us	640us	1.28ms	2.56ms	5.12ms	7.5ms

Depends on the Queuing delay caused by large frames at a given speed - Fragmentation generally not needed above 768kbps

Fragment Size Matrix

Assuming 10ms Blocking Delay per fragment

Link Speed	Frag Size
56kbps	70 Bytes
64kbps	80 Bytes
128kbps	160 Bytes
256kbps	320 Bytes
512kbps	640 Bytes
768kbps	1000 Bytes
1536kbs	2000 Bytes X

$$\text{Fragment Size} = \frac{10\text{ms}}{\text{Time for 1 Byte at BW}}$$

Example: 4 G.729 calls on 128kbps Circuit
Fragment Blocking Delay = 10ms (160bytes)

Q = (Pv*N/C) + LFI

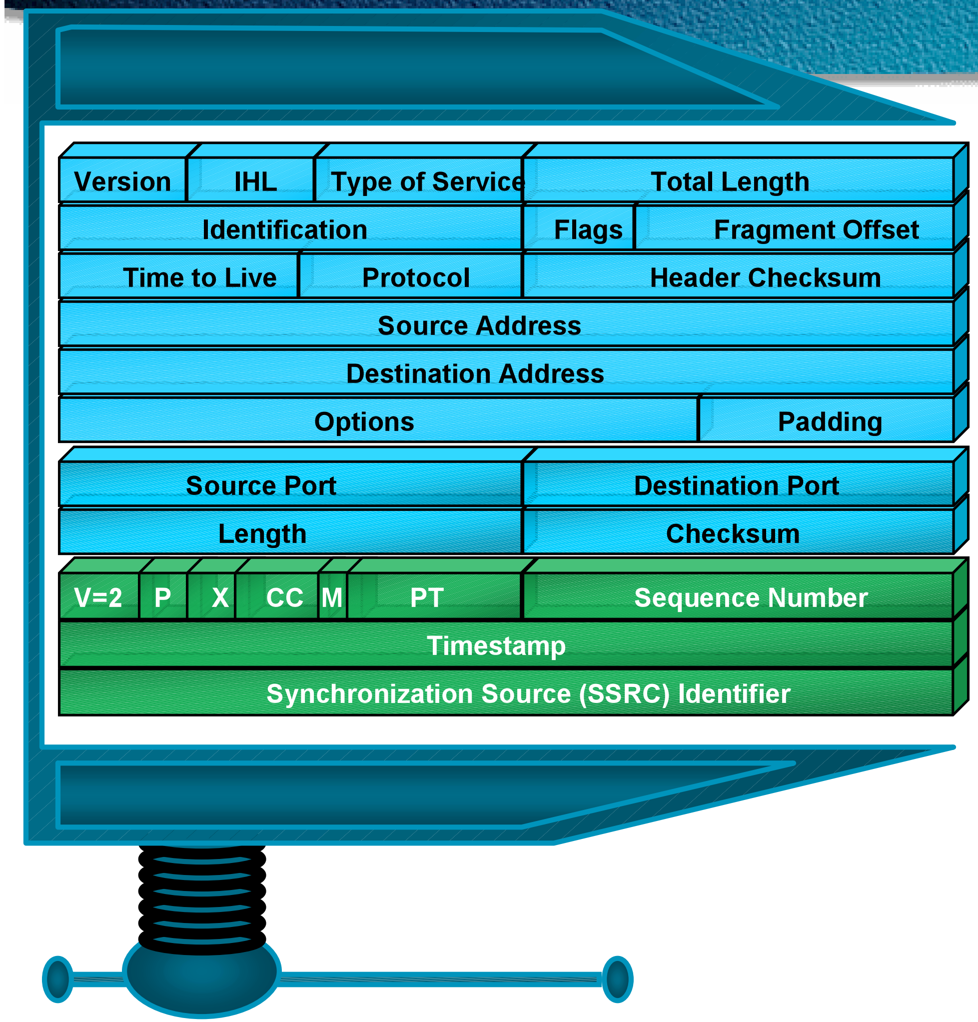
Q = (480bits*4/128000) + 10ms = 25ms

Worst Case Queuing Delay = 25ms

Q = Worst Case Queuing Delay of Voice Packet in ms
Pv = Size of a voice packet in bits (at layer 1)
N = Number of Calls
C = Is the link capacity in bps
LFI = Fragment Size Queue Delay in ms

RTP Header Compression

WARNING - Process Switched (will be fast switched later this summer)



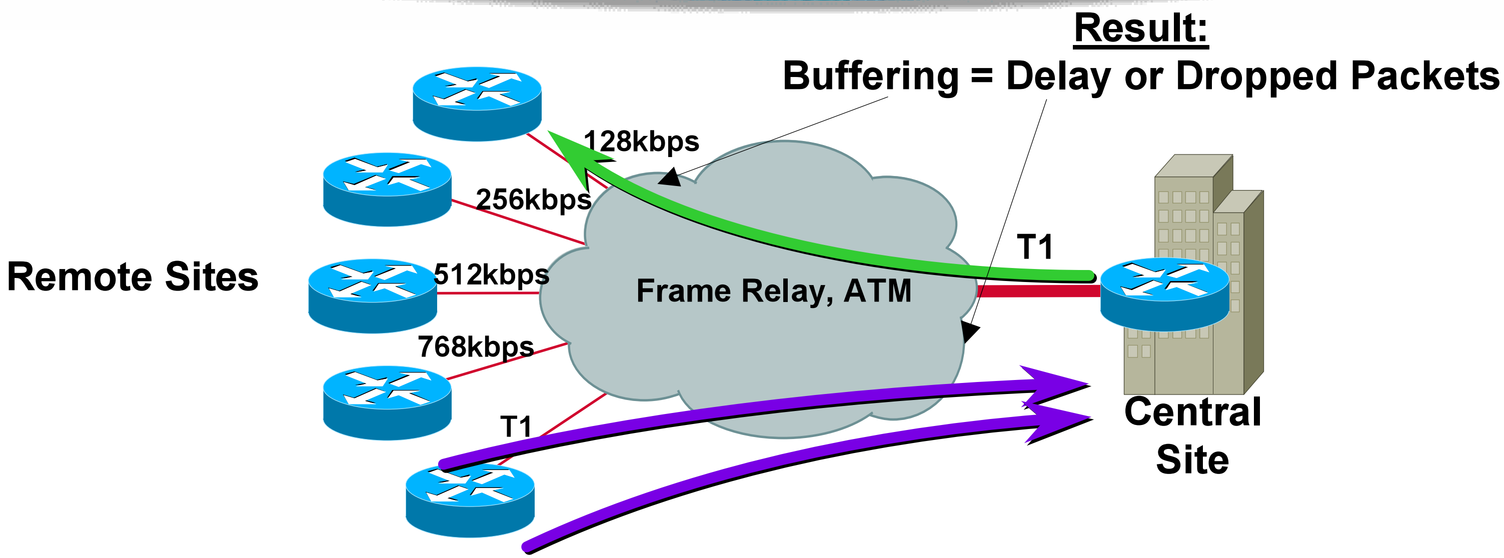
Overhead

- 20ms@8kb/s yields 20 byte payload
- 40bytes per packet
  IP header 20; UDP header 8; RTP header 12

  2X payload!!!!!!!!
- Header compression 40Bytes to 2-4 much of the time
- Hop-by-Hop on ***slow links***
- CRTP—Compressed Real-time protocol

Traffic Shaping

Why + What to Avoid?



- **Central to Remote Site Speed Mismatch - Avoid**
- **Remote to Central Site Over-subscription - Avoid**
- **Prohibit bursting Voice above committed rate**

  What are you guaranteed above you committed rate?

Understanding Shaping Parameters Frame Relay

Traffic Shaping

"AVERAGE" Traffic Rate out of an Interface

Challenge - Traffic Still clocked out at line rate

CIR (Committed Information Rate)

Average Rate over Time, Typically in bits per second

Bc (Committed Burst)

Amount allowed to Transmit in an Interval, in bits

Be (Excess Burst)

Amount allowed to transmit above Bc per Interval

Interval

Equal Integer of time within 1 sec, Typically in ms. Number of Intervals per second depends on Interval length Bc and the Interval are derivatives of each other

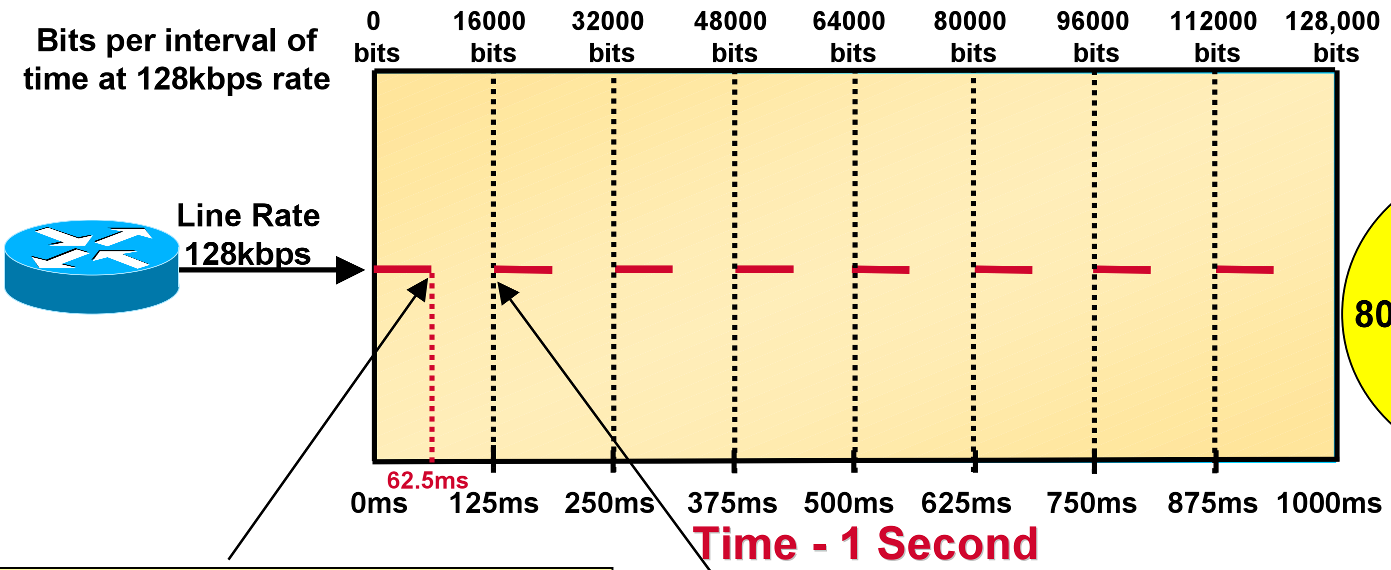
$$\text{Interval} = \frac{\text{CIR}}{\text{Bc}} \xrightarrow{\text{Example}} 125\text{ms} = \frac{64\text{kbps}}{8000\text{bits}}$$

Example - Traffic Shaping in action

High Volume Data flow towards a 128kbps Line Rate Shaping to 64kbps

$$\text{Interval} = \frac{Bc}{CIR} \longrightarrow 125\text{ms Interval} = \frac{8000\text{ bits}}{64000\text{bps}}$$

Cisco Default Bc=1/8 CIR = 125ms Interval



Net Result:

8000 X 8 = 64bkps

When 8000bits (Bc) Transmitted Credits are exhausted and no more Packet flow in that Interval This happens at the 62.5ms point of the interval

When a new Interval begins Bc (8000bit) Credits are restored and transmission may resume. Pause in transmission is 62.5ms in the case

Bc setting Considerations for VoIP

Set Bc lower if Line rate to CIR ratio is high
Example: T1 Line rate shaping to 64kbps

Bc = 8000

$$125\text{ms Interval} = \frac{8000 \text{ Bc}}{64\text{kbps CIR}}$$

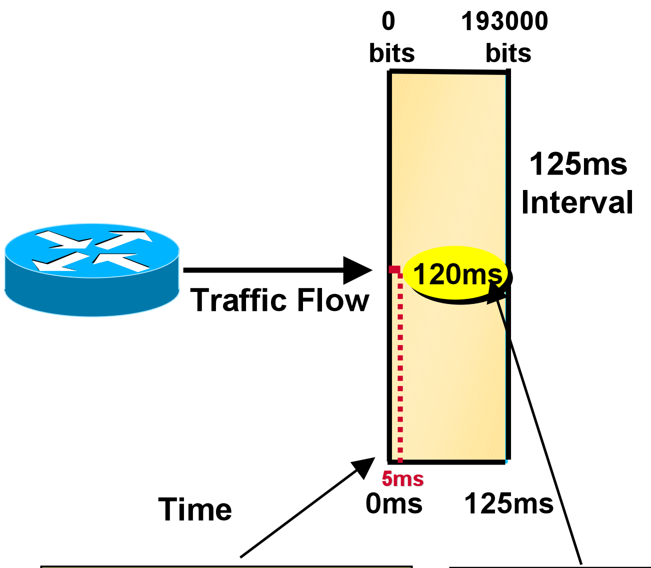
T1 can transmit 193,000 bits in 125ms

→

Bc = 1000

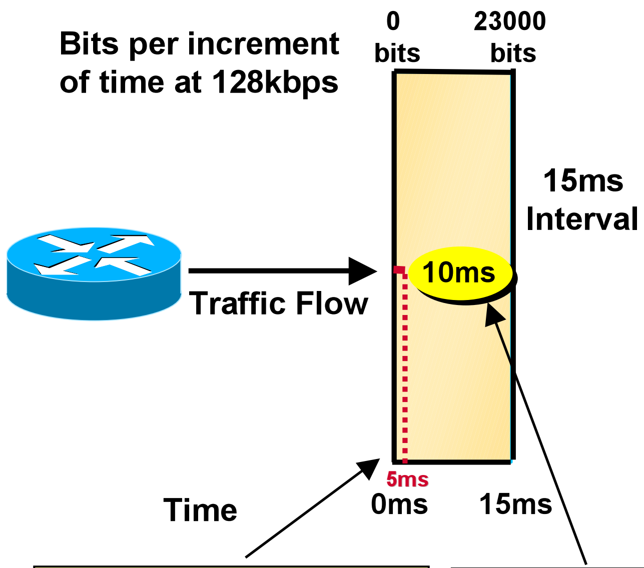
$$15\text{ms Interval} = \frac{1000 \text{ Bc}}{64\text{kbps CIR}}$$

T1 can transmit 23,000 bits in 15ms



At T1 rate 8000 bits (Bc) are exhausted in 5ms. Halting traffic flow for that PVC for the rest of that interval. Even for Voice!

120ms of potential delay for voice until new interval begins and Bc credits are restored



At T1 rate 1000 bits (Bc) still are exhausted in 5ms. Halting traffic flow for that PVC for the rest of that interval. Even for Voice!

10ms of potential delay for voice until new interval begins and Bc credits are restored

Todays IP QoS Solutions

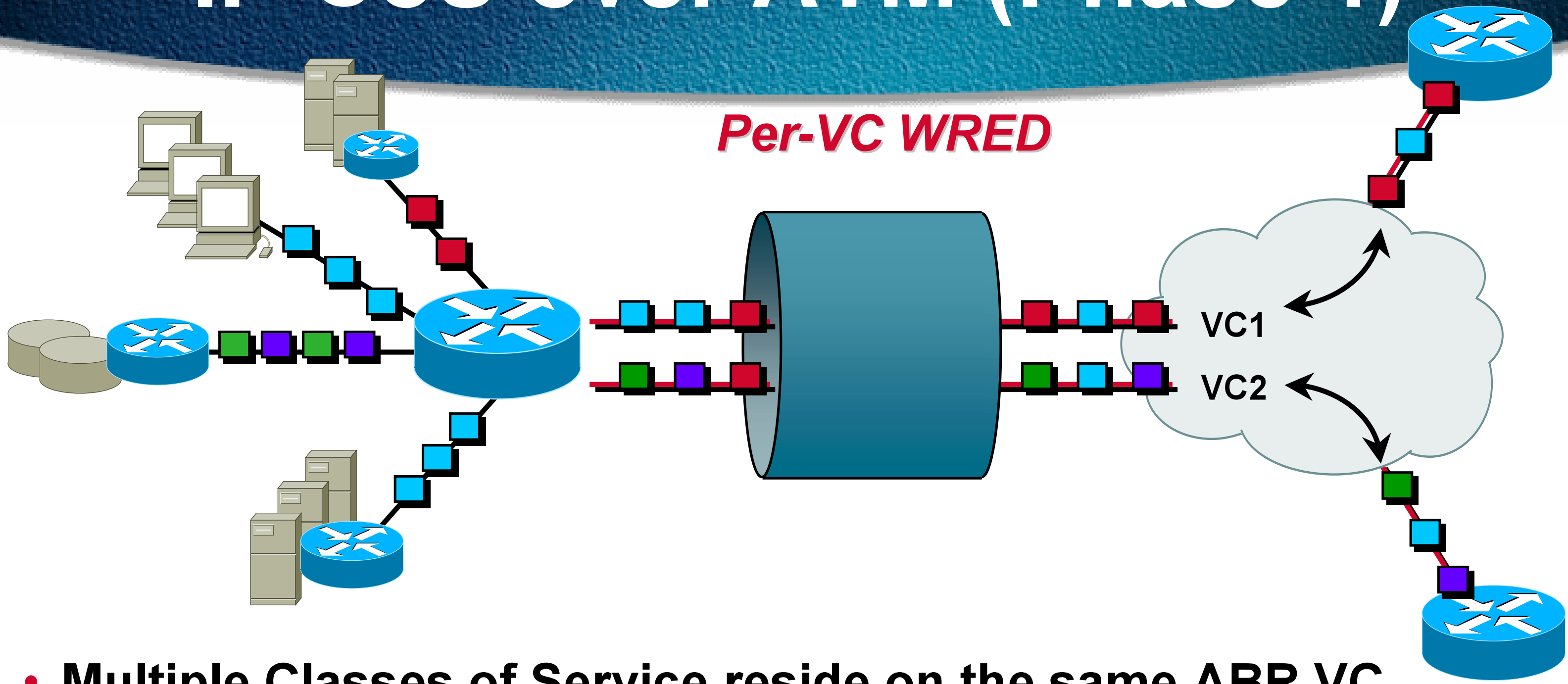
Technology	• Function
IP Precedence	• prioritization (in IP header) • indicates service class
Committed Access Rate (CAR)	• packet classification by application, protocol, etc. • sets precedence • bandwidth management: discard or change service class
WRED	• Weighted Random Early Detection • congestion avoidance • service class enforcement
WFQ, CBQ	• Weighted Fair Queuing • Class-Based Queuing • queuing policies (e.g. latency)
IP/ATM interw.	• WRED on top of a shaped ATM PVC (ABR) • CBQ on ATM PVC bundles
MPLS	• IP+ATM QoS integration • traffic engineering

Cisco Systems®

Empowering the Internet Generation℠

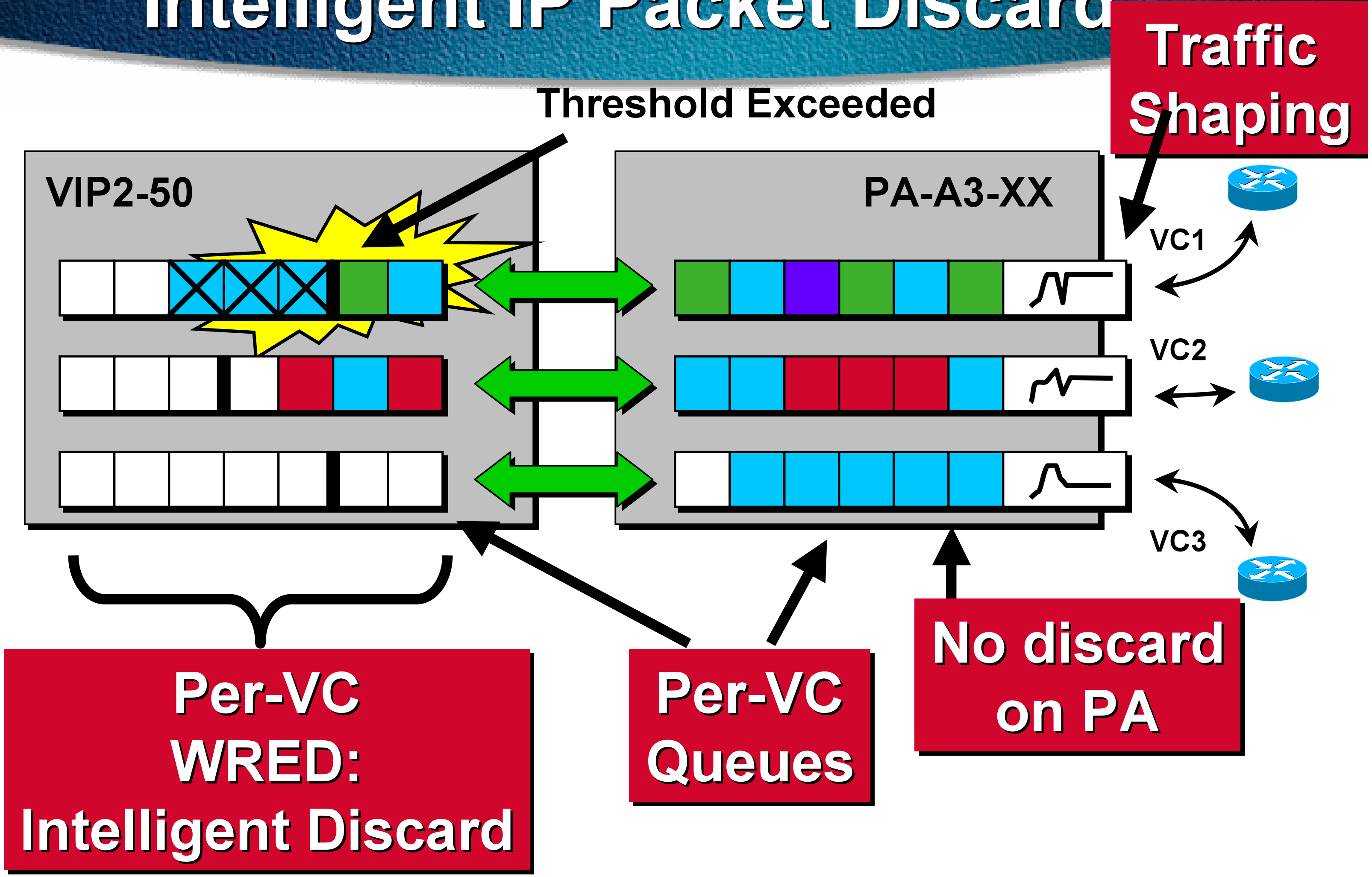
CoS with
IP ATM Overlay

IP CoS over ATM (Phase 1)

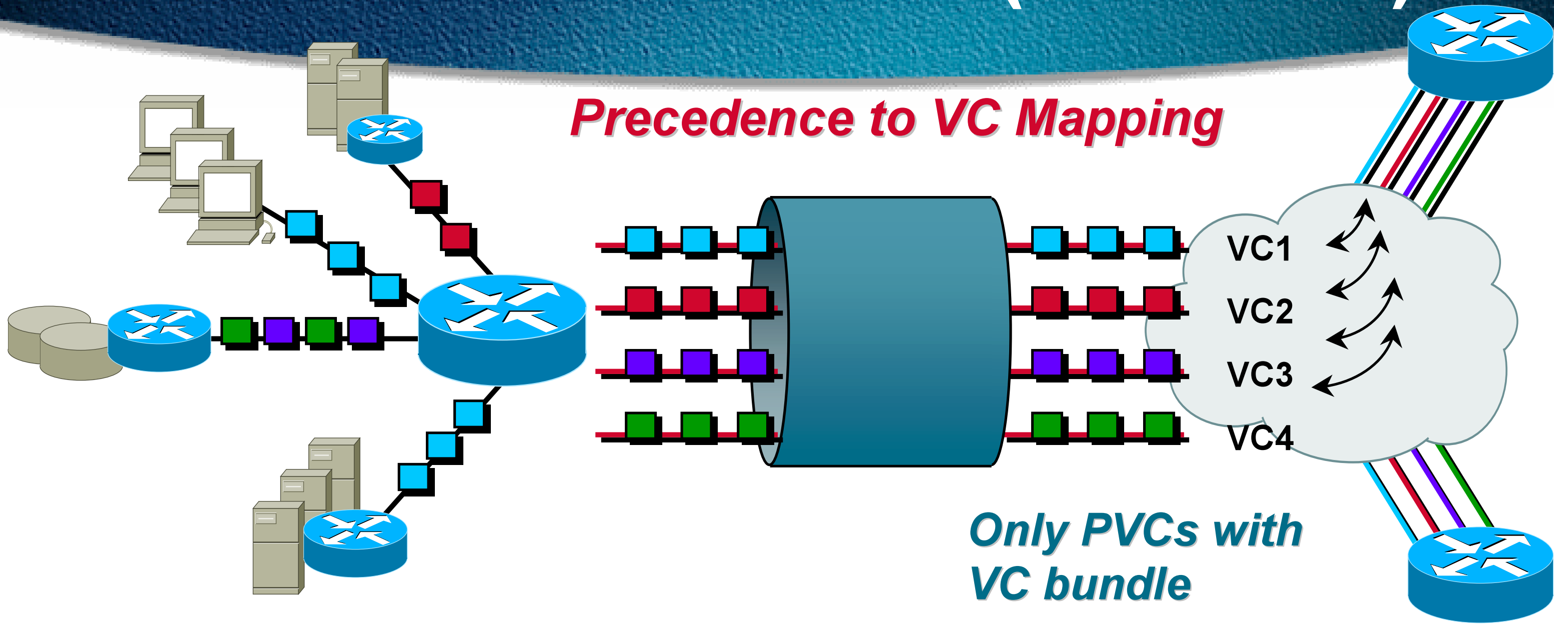


- Multiple Classes of Service reside on the same ABR VC
- Requires a single (ABR) VC for each destination
- Packets (cells) of any class can reside on the same VC
- WRED is run on each VC queue to ensure low loss for higher class service when feedback (RM cells) indicates congestion

Per-VC WRED : Intelligent IP Packet Discard



IP CoS over ATM (Phase 2)



- Separate VCs (with ATM QoS) for each Class of Service
- Requires multiple VCs for each destination
- Packets (cells) of the same class reside on the same VC
- RED may be run on each VC queue

CoS with
IP ATM Peer Model
MPLS / Tag

MPLS allows efficient Resource Allocation and COS support

- Taking the example of Cisco Tag Switching:

  Tag Classes of Service has key advantage in terms of Resource allocation (over any Overlay model like MPOA)

    thanks to allocation at the Class level instead of connection level

  Tag Classes of Service allows use of IP-friendly/optimal scheduling mechanisms in ATM switches

    WEPD + dynamic buffer control

TAG/MPLS DiffServ Model

- In non-ATM MPLS, DS bits are mapped into TAG CoS bits (3 bits)

  queuing/scheduling and intelligent discard algorithms (WFQ, WRED)

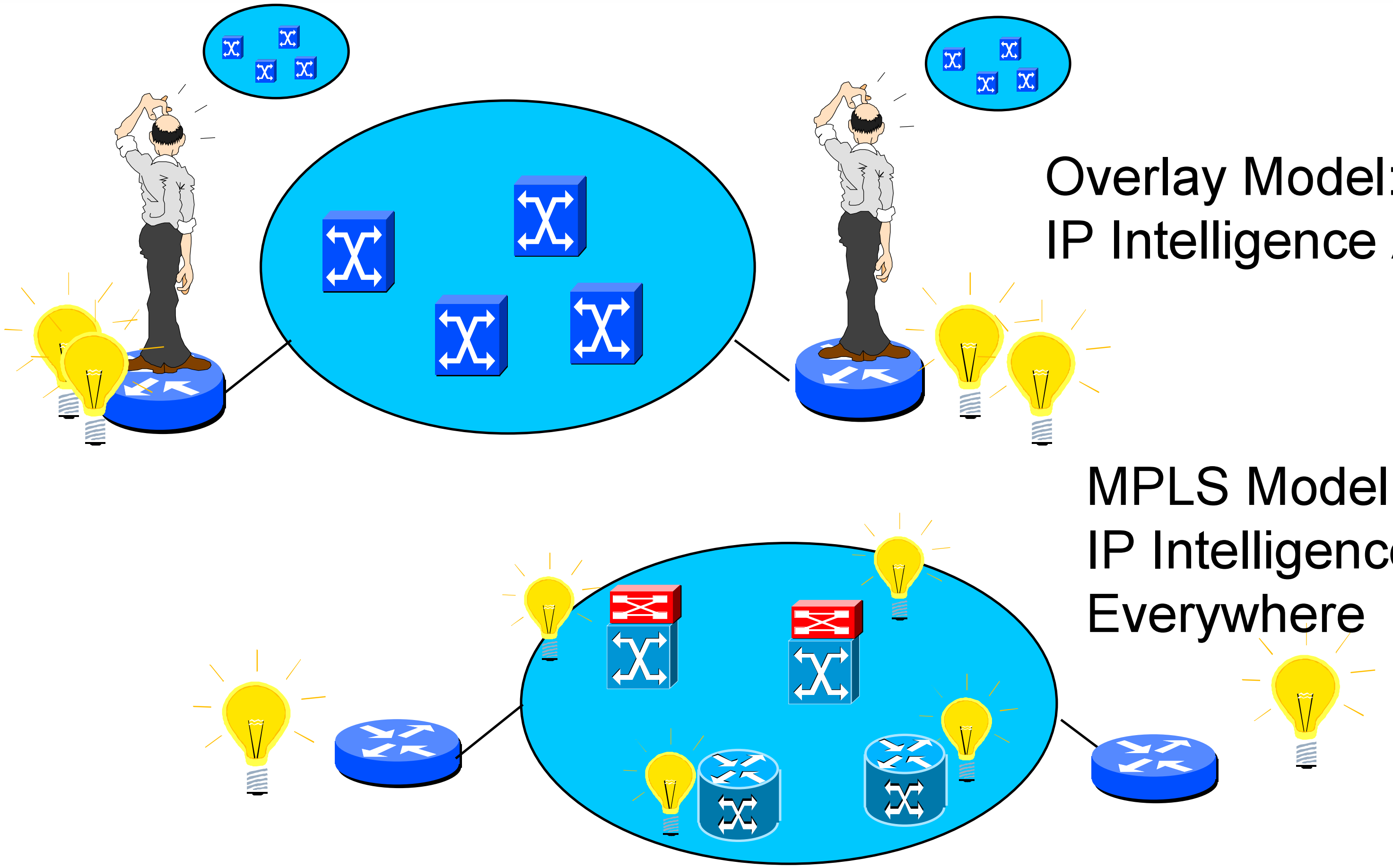
- In ATM-MPLS, LDP associates a CoS to a VC

  MPLS-ATM nodes are performing

    per class queuing/scheduling (WFQ) and

    intelligent discard (WEPD)

MPLS COS vs ATM (overlay)



Overlay Model:
IP Intelligence Around

MPLS Model:
IP Intelligence
Everywhere

Differentiated Services
vs.
Integrated Services

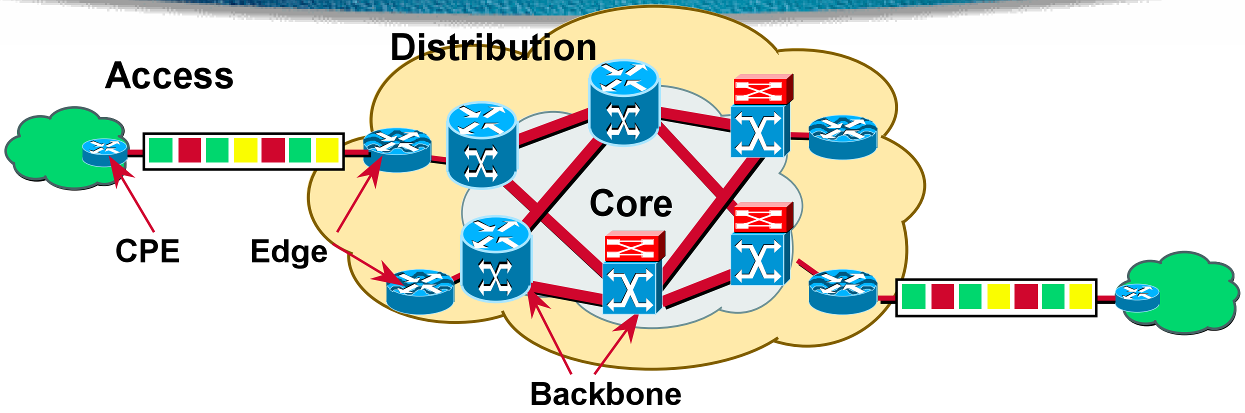
DIFFSERV Working Group Charter

There is a clear need for relatively simple and coarse methods of providing differentiated classes of service for Internet traffic, to support various types of applications, and specific business requirements.

….

A small bit-pattern in each packet, in the IPv4 TOS octet or the IPv6 Traffic Class octet, **is used to mark a packet to receive a particular forwarding treatment, or per-hop behaviour, at each network node**. A common understanding about the use and interpretation of this bit-pattern is required for inter-domain use, multi-vendor interoperability, and consistent reasoning about expected service behaviours in a network. **Thus, the Working Group will standardise a common layout to be used for both octets, called the 'DS byte'. A standards-track document will be produced that will define the general use of fields within the DS byte (superseding the IPv4 TOS octet definitions of RFC 1349)**

The *Diffserv* Model



IP QoS Edge Functions

- IP Packet classification
  - •Traffic classification and DS byte setting based on multiple criteria
  - •application
  - •address source/destination
  - •Measured bandwidth or burst
- Policing
  - •control that received traffic conforms to contract CoS of received traffic

Core/Backbone Functions

- High-speed switching
- Per Class Queuing/scheduling
  - •WFQ
- Per Class intelligent discard
  - •WRED, WEPD

“Differentiated Service” Model

- **Each Packet is coloured with the class it belongs to (DS colouring)**
- ***IP “Diffserv” Working Group at IETF*** **, has reshuffled the IPv4 TOS and IPv6 *Traffic Classes***

  into a 6 bits of common DS byte value - DSCP

- **Today 3 “precedence” bits (IPv4TOS) allow 7 classes**
- **Treat the classes differently in the network elements according to “Per Hop Behaviour” values**

Goal is Scalability

- Complex classification/conditioning at edge
- Resulting in a per-packet *DSCP *)* color
- No per-flow/per-app state in the core
- *Core only performs 'simple' 'Per Hop Behavior - PHB' on traffic aggregates*

**) DSCP - DiffServ Code Point*

Additional Requirements

- Wide variety of services and provisioning policies
- decouple service and application in use
- no application modification
- no hop-by-hop signaling
- interoperability with non-DS-compliant nodes
- incremental deployment

4 Kings

And a joker!

- **The service provided to a traffic aggregate**
- **The conditioning functions and per-hop behaviors used to realize services**
- **The DS field value (DS codepoint) used to mark packets to select a per-hop behavior**
- **The particular node implementation mechanisms which realize a per-hop behavior**

Provisioning

Why Provisioning?

- **QoS does not create bandwidth!**

  zero-sum game

  give better service to a well-provisioned class

  with respect to other classes

CISCO SYSTEMS

EMPOWERING THE INTERNET GENERATION℠


DiffServ
Terminology
Presentation_ID
© 1999, Cisco Systems, Inc.
www.cisco.com
90

Packet Terminology

DSCP field: 6bits	Unused: 2bits

Former ToS byte = new DS field

- **DS Field**

  the IPv4 header TOS octet or the IPv6 Traffic Class octet

  when interpreted in conformance with the definition given in [DSFIELD]

  The bits of the DSCP field encode the DS CodePoint

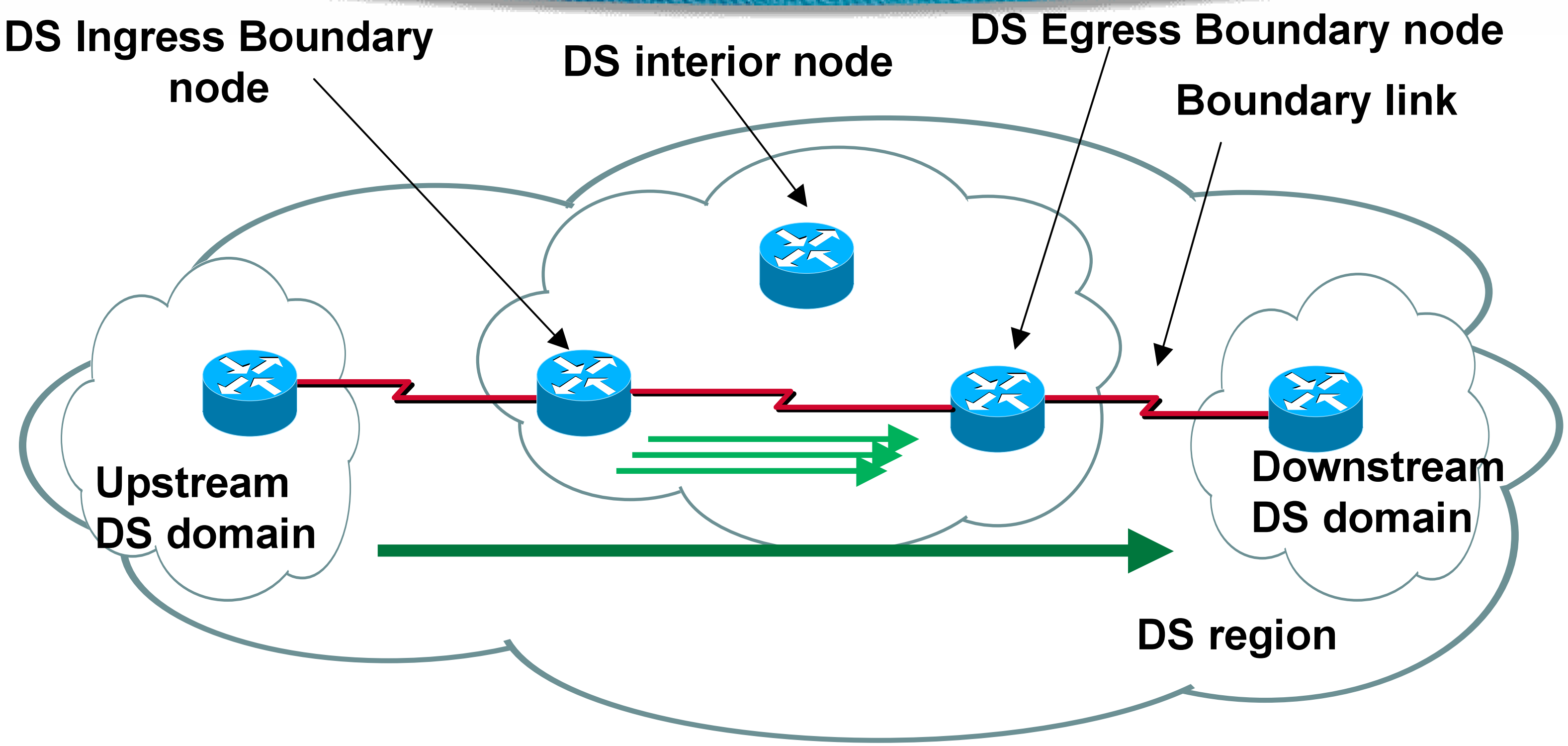
  while the remaining bits are currently unused.

- **DS CodePoint**

  a specific value of the DSCP portion of the DS field

  used to select a PHB

Topological Terminology



Traffic Stream = set of microflows

Behavior Aggregate (microflows with same DSCP)

Traffic Terminology

- **Microflow**

  a single instance of an application-to- application flow of packets

  identified by source address, source port, destination address, destination port and protocol id.

- **Traffic stream**

  an administratively significant set of one or more microflows which traverse a path segment.

  A traffic stream may consist of the set of active microflows which are selected by a particular classifier.

- **Traffic profile**

  a description of the temporal properties of a traffic stream such as rate and burst size.

- **DS Behavior Aggregate = Behavior Aggregate (BA)**

  a collection of packets with the same DS CodePoint crossing a link in a particular direction.

Actions

Classification

- **Classifier**

  an entity which selects packets based on the content of packet headers according to defined rules

- **BA classifier**

  a classifier that selects packets based only on the contents of the DS field

- **MF Classifier**

  a multi-field (MF) classifier which selects packets based on the content of some arbitrary number of header fields

  typically some combination of source address, destination address, DS field, protocol ID, source port and destination port

Actions (2)

Metering

- **Meter**

  a device that measures the temporal properties of a traffic stream selected by a classifier

Dropping

- **Dropper**

  a device that performs dropping

Marking

- **Marker**

  a device that sets the DSCP in a packet based on defined rules

- **Pre-marking**

  marking prior to entry into a downstream DS domain

Shaping

- **Shaper**

  a device that delays packets within a traffic stream to cause them to conform to some defined traffic profile

Actions (3)

Policing

- **Policer = Dropper**

  a device that discards packets (dropper) within a traffic stream in accordance with the state of a corresponding meter enforcing a traffic profile

Traffic Conditioning

- **Traffic Conditioner**

  an entity that may contain meters, markers, droppers and shapers

  Typically deployed in boundary DS nodes only

  A traffic conditioner may re-mark a traffic stream or may discard or shape packets to alter the temporal characteristics of the stream and bring it into compliance with a traffic profile

- **Traffic Conditioning Agreement (TCA)**

  an agreement specifying classifier rules and any corresponding traffic profiles and metering, marking, discarding and/or shaping rules which are to apply to the traffic streams selected by the classifier

PHB Terminology

- **Per-Hop-Behavior (PHB)**

  the externally observable forwarding behavior applied at a DS-compliant node to a DS behavior aggregate

- **PHB group**

  a set of one or more PHBs that can only be meaningfully specified and implemented simultaneously, due to a common constraint applying to all PHBs in the set such as a queue servicing or queue management policy

  a PHB group provides a service building block that allows a set of related forwarding behaviors to be specified together (e.g., four dropping priorities)

  a single PHB is a special case of a PHB group

- **Mechanism**

  a specific algorithm or operation (e.g., queueing discipline) that is implemented in a node to realize a set of one or more per- hop behaviors

Service Terminology

- **Service**

  defines some significant characteristics of packet transmission in one direction across a set of one or more paths within a network

  these characteristics may be specified in quantitive or statistical terms of throughput, delay, jitter, and/or loss

  or may otherwise be specified in terms of some relative priority of access to network resources

- **Service Level Agreement (SLA)**

  a service contract between a customer and a service provider that specifies the forwarding service a customer should receive

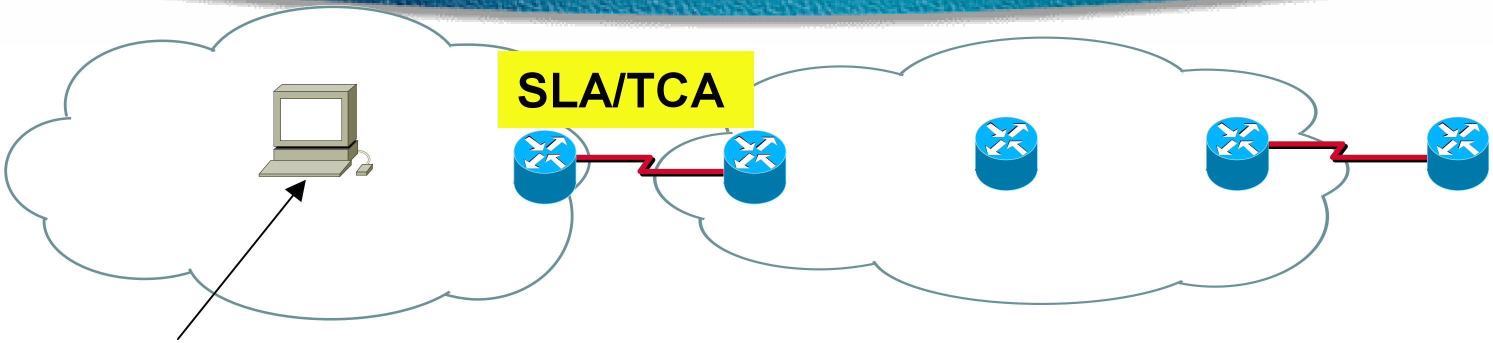
  a customer may be a user organization (source domain) or another DS domain (upstream domain)

  SLA may include traffic conditioning rules which constitute a TCA in whole or in part.

- **Service Provisioning**

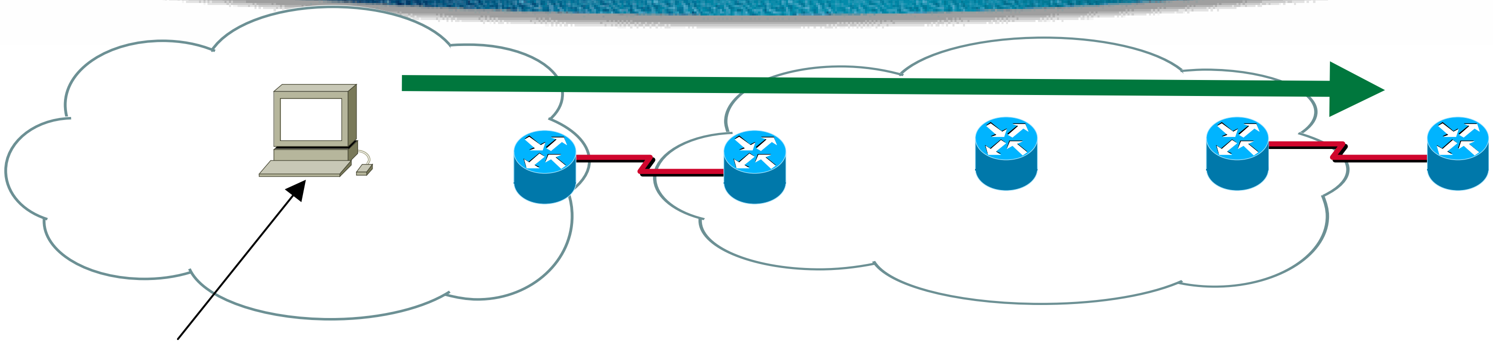
  a policy which defines how traffic Policy conditioners are configured on DS boundary nodes and how traffic streams are mapped to DS behavior aggregates to achieve a specified range of services

DiffServ Architecture



0. Negociation and agreement of an SLA/TCA

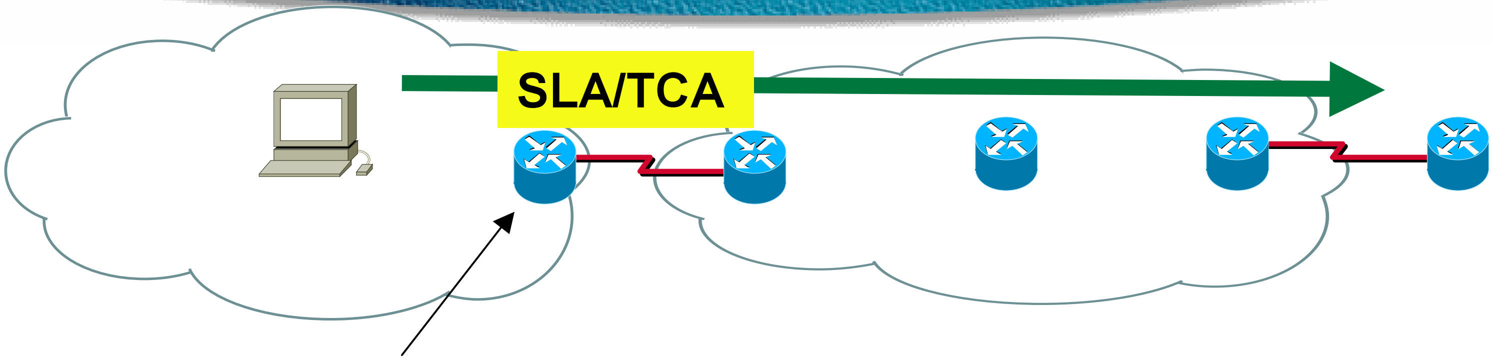
DiffServ Architecture



1. Pre-marking in the source domain

- **per-application/host basis**
- **per-default-gateway basis**

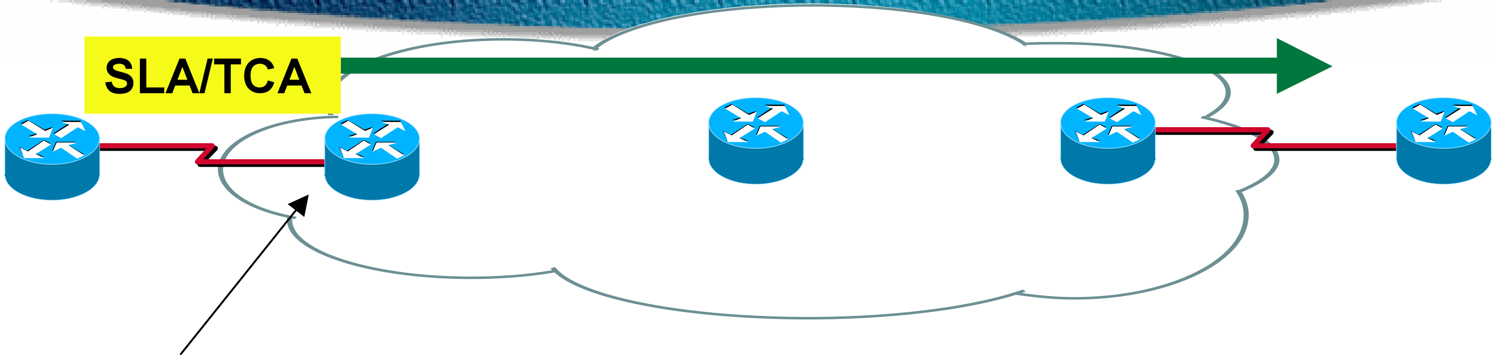
DiffServ Architecture



2. Egress Boundary DS Node of source domain applies traffic conditioning to ensure SLA/TCA compliance

hence causing possible re-marking, dropping and shaping

DiffServ Architecture

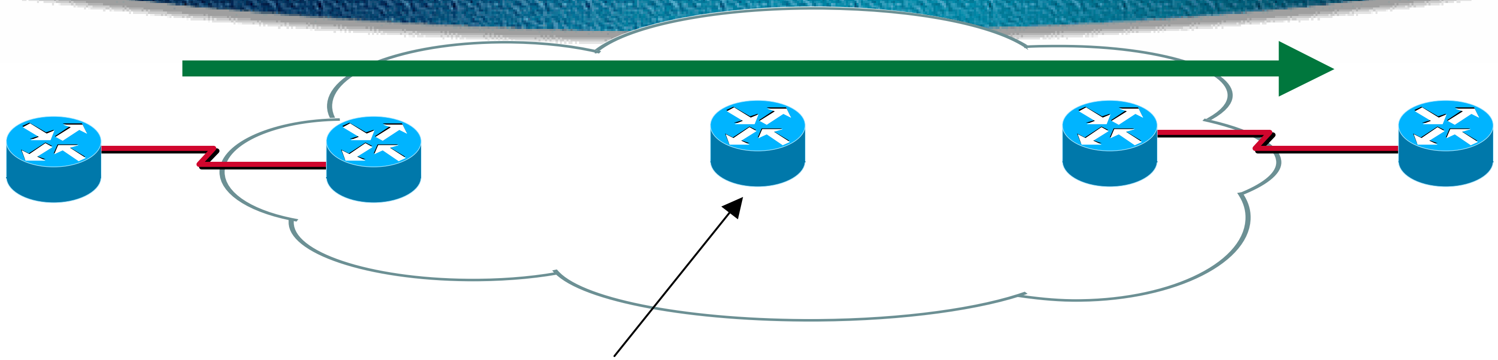


3. Classification according to SLA

4. Conditioning according to TCA

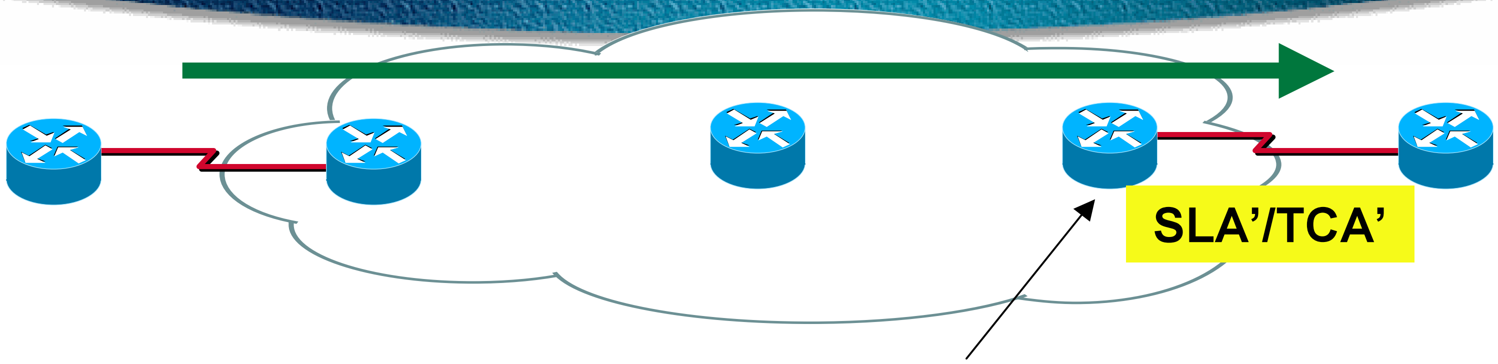
5. Assignment to a BA (DSCP setting)

DiffServ Architecture



6. Forwarding according to PHB mapped to set DSCP

DiffServ Architecture



If downstream DS domain support same service provisioning policy, same PHBs and DSCP/PHB mappings

Then 7: No-op

Else 7'a: SLA'/TCA' negotiation

7'b: Conditioning according to TCA'

DiffServ Architecture

Functional Blocks

Classifier	Conditioner	Forwarding	Queueing	Conditioner
	Metering Dropping Marking		Scheduling Dropping	Shaping

DiffServ Architecture

Metering
Dropping
Marking

Classifier	Conditioner	Forwarding	Queueing	Conditioner

- QPPB

  Based on source or destination address

  AS-path or community or access-list

  Scalable Return direction

DiffServ Architecture

Classifier	Conditioner (Metering, Dropping, Marking)	Forwarding	Queueing	Conditioner

- CAR

  mac-address

  precedence

  std/extended ACL

  all traffic

DiffServ Architecture

Classifier	Conditioner (Metering, Dropping, Marking)	Forwarding	Queueing	Conditioner

- CAR

  token bucket metering

  commited rate, normal burst (bucket depth), exceed burst (tcp-friendly behavior)

  set prec/tx, set prec/cont, set qos/tx, seq qos/cont, drop, tx, cont

DiffServ Architecture

Classifier	Conditioner	Forwarding	Queueing	Conditioner

- **EF**

DiffServ Architecture

Classifier	Conditioner	Forwarding	Queueing	Conditioner
			Scheduling Dropping	

- **Scheduling**

  “Which packet first?”

  FIFO, PQ, CQ, WFQ (CB, FB), DWFQ (CB, FB), MDRR

  on physical interface

  ATM: special case

DiffServ Architecture

Classifier	Conditioner	Forwarding	Queueing	Conditioner

Scheduling
Dropping

- Dropping

  "When/how should I drop?"

  Tail-Drop, Fair-Drop, WRED

  Physical interface

  ATM: special case

DiffServ Architecture

Classifier	Conditioner	Forwarding	Queueing	Conditioner

- GTS/FRTS

DiffServ Architecture

Classifier	Conditioner	Forwarding	Queueing	Conditioner

- CAR

  token bucket metering

  commited rate, normal burst (bucket depth), exceed burst (tcp-friendly behavior)

  set prec/tx, set prec/cont, set qos/tx, seq qos/cont, drop, tx, cont

  per subintf

Cisco Systems

Empowering the Internet Generation℠

