Differentiated Services

Reading

- RFC 2475, "An Architecture for Differentiated Services", December 1998.
- RFC 2474, "Definition of the Differentiated Services Field in the IPv4 and IPv6 Headers", December 1998.
- X. Xiao and L. Li, "Internet QoS: A Big Picture," *IEEE Network*, Vol.13, No. 2, March/April 1999, pp. 8-18.
- C. Metz, "IP QoS: Traveling in First Class on the Internet," *IEEE Internet Computing*, Vol. 3, No. 2, March/April 1999, pp. 84-88.

Objectives

- Provide a wide variety of services and provisioning services for IP networks.
- Work with existing applications, without the need for API or host software changes.
- Scalable to large number of flows, without the requirement of per-flow or per-customer state.
  - Simple implementation in the core routers to allow for very high speed operation.
- Reasonable interoperability with current networks.
- Should accommodate incremental deployment.

Main Idea

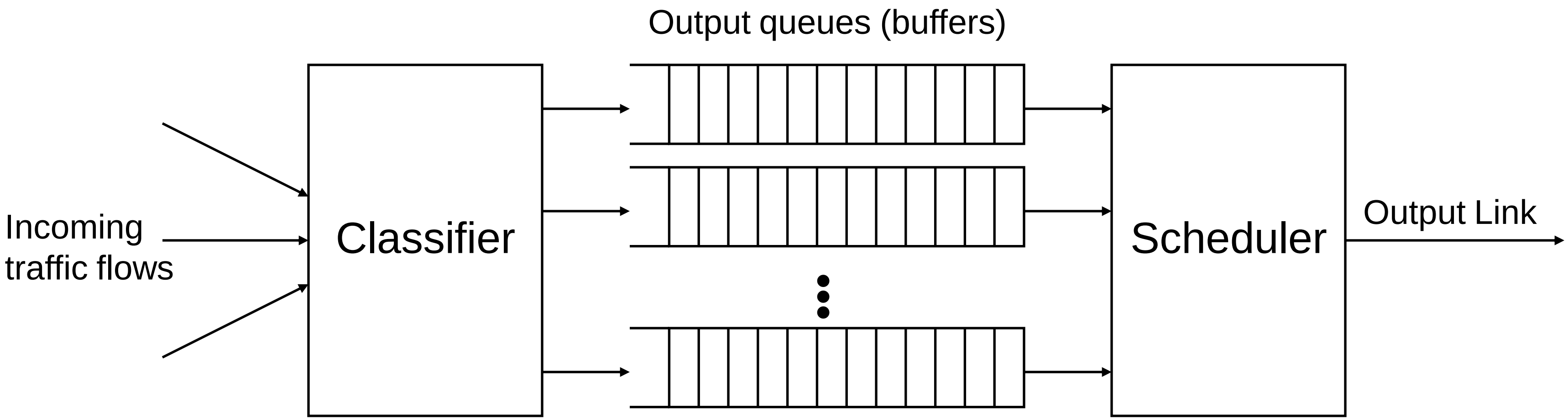
- Protocols such as RSVP and ST2 require routers:
  - Keep state per flow
  - Perform hop-by-hop signaling

  (which imposes a large burden on fast core routers)
- Basic idea:
  - Classify the microflows at the edge of the network into service classes (“gold”, “silver”, “bronze”).
  - Inside the network, all packets belonging to a class receive the same service.
  - Single field in the packet header differentiates the classes.

Ingredients

- Router management schemes
  - How to handle the different classes of services inside the router
- Protocols
  - How to mark that a packet belongs to a certain class of service in the network

Queue Management in Routers



- Routers have buffers (queues) in their output links.
- These queues are intended to "absorb" instantaneous spikes of traffic that exceed link capacity.
- Problem: how to manage these queues to provide different classes of service?

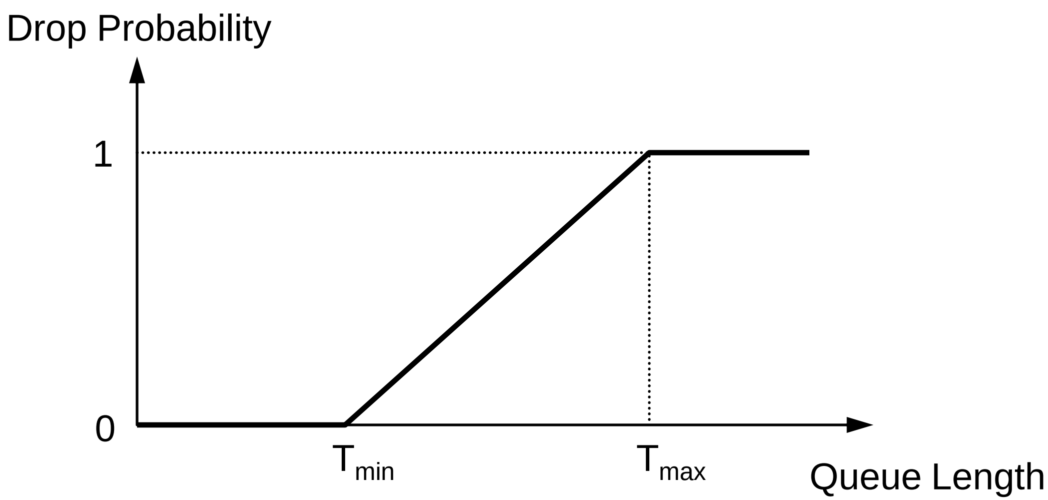
Simplest Case: Single FIFO Queue

- No classifier, no scheduler.
- Known as “tail drop”.
- Packets are served in a FIFO order.
- If a packet arrives and the queue is full, it is discarded.
- Problem: global synchronization of TCP connections:
  - TCP interprets packet loss as a sign of congestion
  - When the queue overflows, packets will be dropped from all connections
  - All TCP connections will throttle down
  - Bandwidth becomes available
  - All TCP connections ramp up, starting another cycle

Random Early Detection (RED)

- Basic idea: start dropping packets at random before the queue overflows
  - Define two thresholds, T_{min} and T_{max}.
  - If the queue length is less than T_{min} , no packets are dropped.
  - If the queue length is between T_{min} and T_{max}, incoming packets are dropped at random with increasing probability.
  - If the queue length is T_{max}, all incoming packets are dropped.

RED (cont.)



- Improves fairness and eliminates synchronization.
- Only provides better performance over tail drop in specific situations. May even reduce performance.

RED with In and Out (RIO)

- Designed to work with two classes of service: the “in profile” class (higher priority) and the “out of profile” class (lower priority):
  - All packets are added to the same queue.
  - There are two threshold lengths in the queue.
  - If the queue is below the first threshold, all packets are sent.
  - If the queue is in between the two thresholds, random “out” packets are discarded.
  - If the queue is over the second threshold, both “in” and “out” packets are randomly discarded, but the “out” packets are hit more aggressively.
  - If the queue is full, all incoming packets are discarded.

Scheduling Mechanisms

- Priority Queuing:
  - Each queue has a defined priority
  - A queue is only served when all queues of higher priority are empty
  - Minimizes queuing delay for high priority traffic
  - Could starve low priority classes if there is too much traffic in the higher priority class
  - Packets can be classified by: IP Precedence, Source and/or Destination IP Address, protocol, port, etc.
- Weighted Fair Queuing (WFQ)

Weighted Fair Queuing (WFQ)

- Ideal model: “Fluid Fair Queuing”(FFQ)
  - Each queue is given a positive weight w_i, $i = 1..N$
  - At time t, let $B(t)$ denote the set of queues that are non-empty
  - If the link capacity is denoted by C, the service rate for non-empty queue i at time t will be:

$$\frac{w_i}{\sum_{j \in B(t)} w_j} C$$

  - Cannot be applied to actual systems because traffic is not divisible - one packet is sent at a time

WFQ (cont.)

- WFQ tries to approximate FFQ in a discrete system:
  - Let *t* be the time when the server is ready to transmit the next packet
  - Between all the packets present in the system at time *t*, select the one that would complete service first in the ideal FFQ system if no more packets arrived after *t* and send it.
- Effect of WFQ:
  - Bandwidth is effectively distributed between classes
  - No classes starve

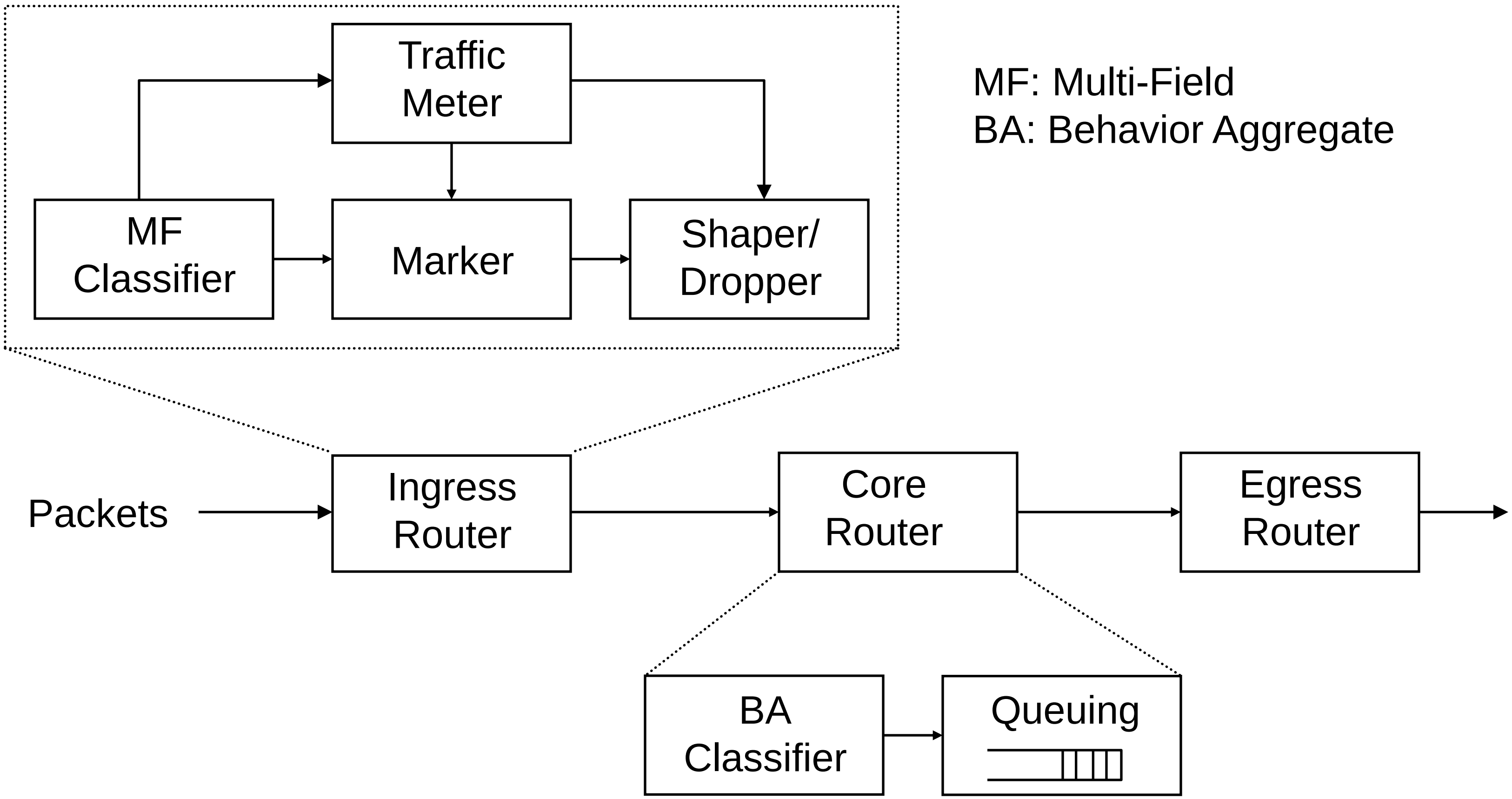
DiffServ Architecture Components

- **DS-Field** is a header field that identifies the service each packet should receive as it is forwarded on each hop.
  - Uses the TOS byte in IPv4 and the Traffic Class field in IPv6.
- **Per-Hop Behavior (PHB)** defines the service each class of packets receive as it is forwarded through the network.
- **Behavior Aggregate (BA)** is a group of packets with the same DS Code Point (DSCP).

Architecture, cont.

- **Boundary Router** is at the edge of the DiffServ-capable network and is responsible for:
  - packet classification
  - packet marking, metering
  - traffic conditioning.
- **Interior Nodes** are core switches that apply the PHB as a function of the contents of the DS-Field.
  - Different PHBs are implemented with various queue management disciplines.
  - Examples: “Random Early Detection” (RED), “Weighted Fair Queuing” (WFQ).

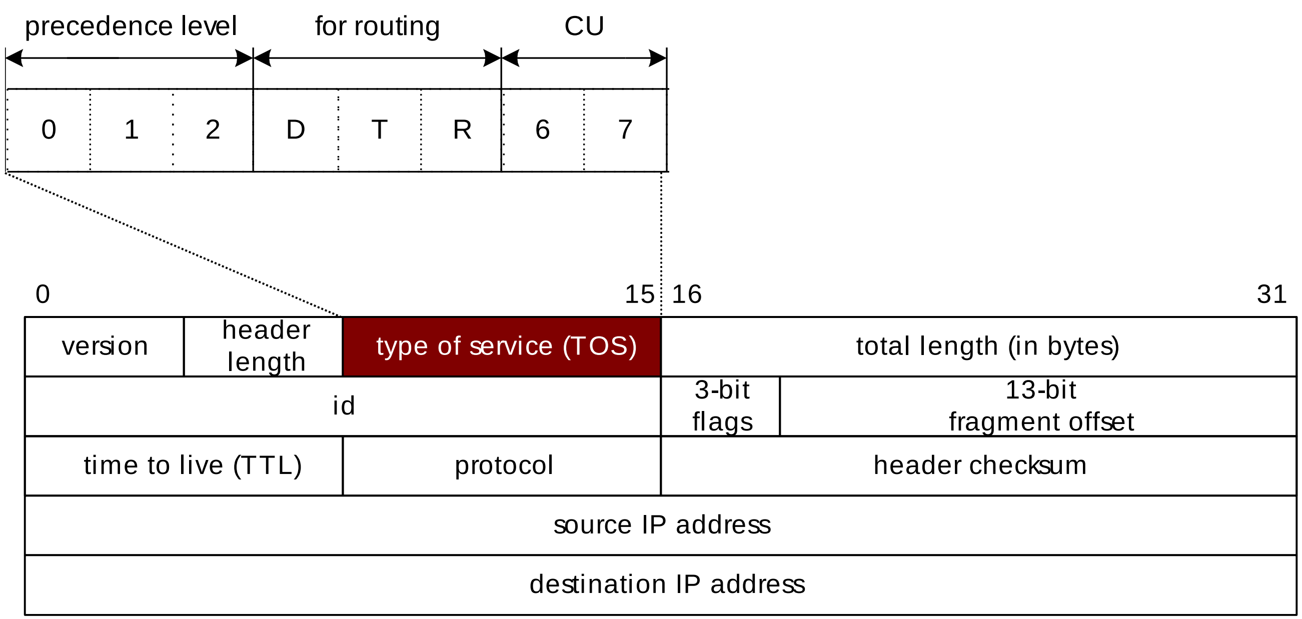
DiffServ Elements



MF: Multi-Field
BA: Behavior Aggregate

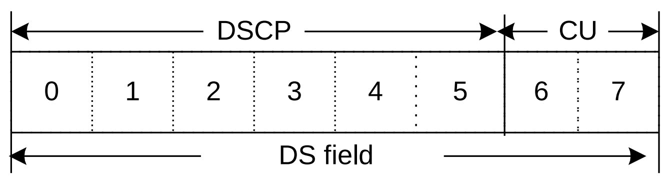
The TOS field

The ***TOS field*** of the IPv4 header:



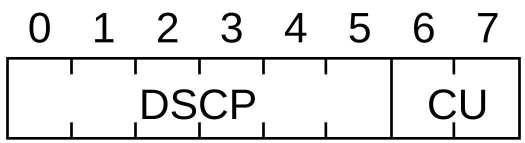
DS field

- In DiffServ, each packet carries a marking indicating which ***behavior aggregate*** it belongs to. This marking determines the treatment the packet receives at each hop.
- The field used to carry the marking is the former TOS field, renamed the ***DS field.***



- Only the 0-5 bits are currently used and called the DiffServ CodePoint or ***DSCP***. Bits 6-7 are reserved for future use.

DS Field Definition



DSCP: Differentiated Services Code Point
CU: Currently Unused

- Codepoint 000000 is assigned to the “default” PHB (best-effort).
- Codepoints xxx000 are assigned to the Class Selector PHB. The eight codepoints should map into at least two classes of traffic, with better service for 11x000 in relation to 000000.
- Assignment space:

Pool	Codepoint Space	Assignment Policy
1	xxxxx0	Standards Action
2	xxxx11	Experimental/Local
3	xxxx01	Experimental/Local (*)

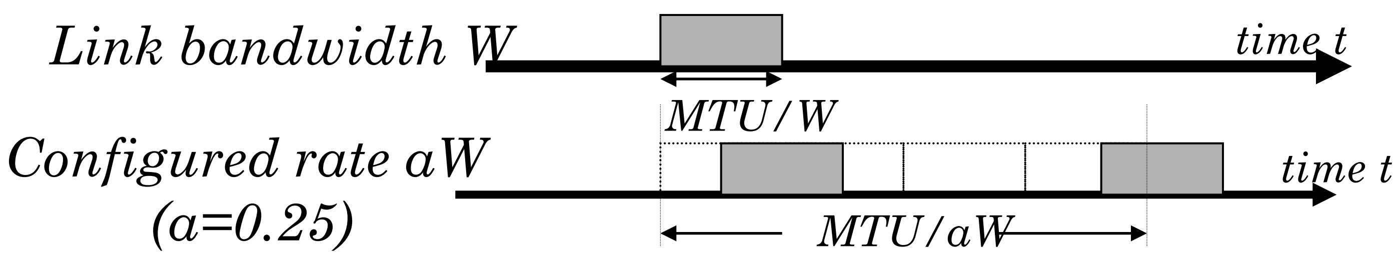
(*) Will be used for standards if Pool 1 runs out.

Assignment of DSCPs

Used for	**bits 0,1,2** (class)	**bits 3,4** (drop precedence)	**bit 5**
EXP/LU	xxx	xx	1
BE	000	00	0
CSC	xxx(≠000)	00	0
AF1	001	01,10,11	0
AF2	010	01,10,11	0
AF3	011	01,10,11	0
AF4	100	01,10,11	0
EF	101	11	0

Expedited Forwarding (EF)-I

- RFC 2598 specification:
  - "the departure rate of the aggregate's packets from any DS node MUST always equal or exceed a configured rate"
  - "it SHOULD average at least the configured rate when measured over any time interval equal to or longer than the time it takes to send an output link MTU sized packet at the configured rate"



$$\frac{\#bits\ Tx\ during\ MTU/aW}{MTU/aW} >= aW$$

Expedited Forwarding (EF)-II

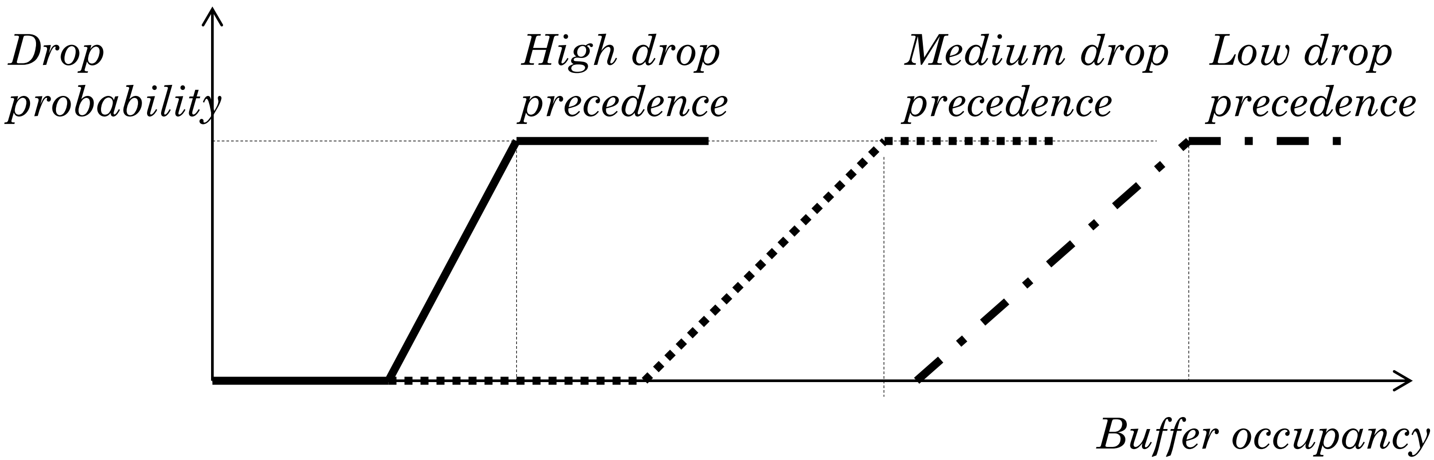
- Possible implementations of EF PHB:
  - Priority queuing
  - Weighted Fair Queuing
- Services based on EF:
  - Require control at the ingress: "conditioning the EF aggregate (via policing or shaping) so that the arrival rate at any node is always less than the configured minimum departure rate."
  - Designed for applications that require low delay, low jitter, low loss, eg. IP telephony.
  - Eg. "Virtual Wire" emulates a dedicated line: users are restricted to transmit at a configured rate but experience very low delay.

Assured Forwarding (AF)-I

- RFC 2597 specification:
  - "an AF compliant node MUST allocate some minimum resources to each AF class, sufficient to achieve at least the configured service bandwidth over large and small time scales."
  - 4 independent AF classes and within each AF class traffic is further marked with one of 3 levels of drop precedence.
  - an AF compliant node MUST NOT reorder AF packets regardless of their drop precedence.
  - 4x3=12 allocated codepoints (see assignment of DSCPs).
- Input rate is not strictly limited - how to handle excess traffic MAY be specified.

Assured Forwarding (AF)-II

- Implementation:
  - no reordering requirement implies one queue per AF class
  - buffer management like RIO, but with 3 thresholds instead of one: one for each drop precedence:



Assured Forwarding (AF)-III

- Comparison of AF-EF.

	AF	EF
Reserved bandwidth	Yes	Yes
Limited input rate	No	Yes
Drop precedences inside one class	Yes	No
Implementation using	WFQ	PQ, WFQ

- Example uses of AF:
  - Idea: reserved bandwidth & graceful degradation during congestion 1st Mbps of traffic from a user has drop precedence A, 2nd Mbps has drop precedence B, 3rd Mbps has drop precedence C
  - Layered video: user transmits each layer at a different drop precedence level

Service Level Agreements

- In order to receive differentiated services from an ISP, the customer must have a “Service Level Agreement” (SLA).
- The SLA specifies the service classes supported and the amount of traffic on each class.
- SLAs can be static or dynamic; dynamic SLAs need to use a protocol such as RSVP to request service.
- Packets are normally shaped and marked at the ingress of the network.

Service Metrics

- Quantitative
  - specific delay, jitter, bandwidth and other metrics

  e.g., 90% of the packets will receive 75 msec delay
- Qualitative
  - no quantifiable metrics

  e.g., “low loss” service
- Proportional
  - metric is specified relative to another class

  e.g., AF1 will receive twice the bandwidth of AF2

Traffic Profiles

- Specifies characteristics of incoming traffic
  - e.g., peak rate, token bucket, effective bandwidth
  - If traffic exceeds its negotiated profile it may be marked, dropped, or reshaped
- Specifies the geographic scope
  - Destinations to which the user will transmit traffic
  - From source A to destination B
  - From sources A to destinations {B1,B2,B3}
  - From source A to all destinations

Example

- Two levels of service:
  - Assured Service: for customers that need reliable service from their ISP, even in case of network congestion.
  - Premium Service: low-delay and low-jitter services for customers that generate fixed peak rate traffic (e.g., telephony).
- Create two queues: AQ (assured queue) and PQ (premium queue).
- Traffic is shaped before being added to the PQ.
- PQ is always served before AQ.
- AQ is served with RIO.
- SLA will keep PQ to a small percentage of the link capacity.

Open problems

- What types of service differentiation can be provided with DiffServ mechanisms
- How to specify SLAs
  - What service metrics are important to users
  - What are appropriate traffic profiles (e.g. token bucket)
- How to compose end-to-end services
  - What are characteristics of aggregate traffic
  - What SLAs should networks negotiate between them
  - Networks may offer different services
  - What happens when traffic needs to be de-aggregated