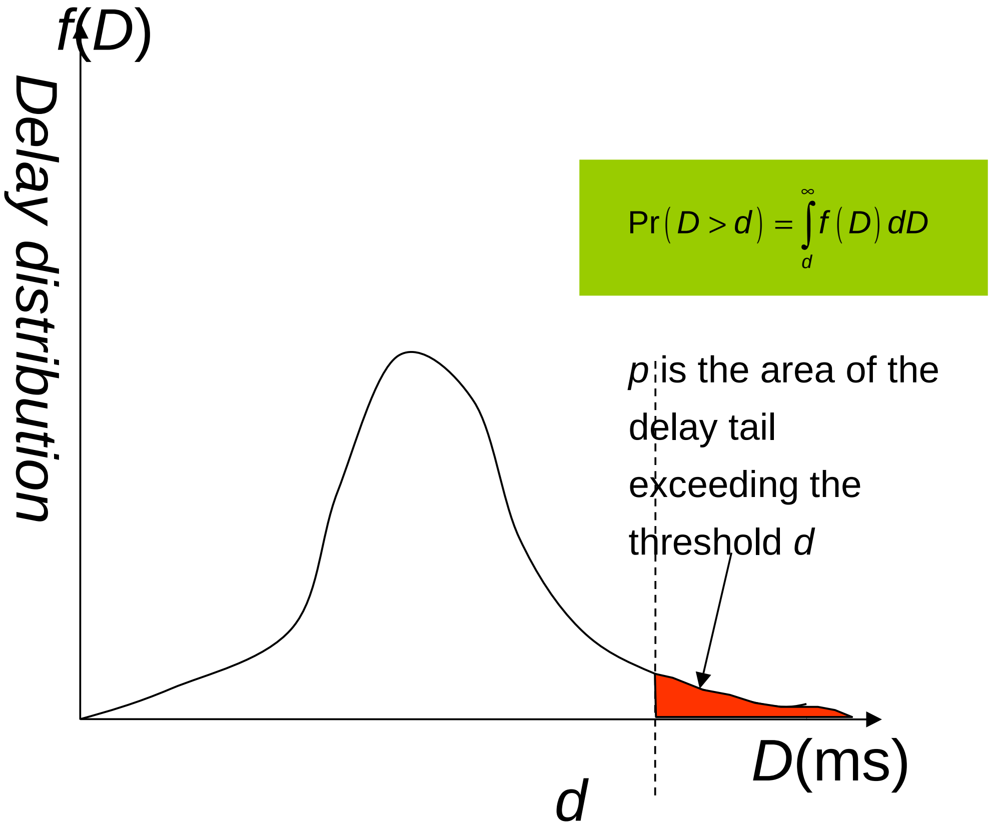
Sample delay-oriented Service level Agreement (II)

- Note tha SLAs are defined *statistically*
- A maximum delay threshold *d* is fixed
- A maximum fraction *p* of packets can be allowed to exceed the delay threshold *d*
- Formally, the SLA is defined as
  - $\Pr(D>d)<p$
  - Where *D* is the actual delay of packets

SLA (QOS)	Access link speed: 2.048 Mbit/s MAX E2E delay and maximum fraction *p* of packets that can exceed max delay	Access link speed: 4 Mbit/s MAX E2E delay and maximum fraction *p* of packets that can exceed max delay	...
Gold:	d=90 ms, p=0.001	d=80 ms, p=0.001	...
Silver:	d=120 ms, p=0.001	d=100 ms, p=0.001	...
Bronze:	d=250 ms, p=0.001	d=210 ms, p=0.001	...

Sample delay-oriented Service level Agreement (III)

- A statistical delay SLA is defined on the basis of the distribution of packet delay, *f*(*D*)
- Given the packet delay distribution, the probability of exceeding a delay threshold *d* is the area under the curve *f*(*D*), from *d* to infinity
- Thus, statistical delay SLAs are based on the concept of the *delay tail*, whose weight (area) should not exceed a preassigned measure, *p*
- A statistical delay SLA can be indicated with the short-hand notation (*d*, *p*)



$$\Pr(D > d) = \int_d^\infty f(D)\,dD$$

p is the area of the delay tail exceeding the threshold *d*

Sample delay-oriented Service level Agreement (IV)

- A statistical delay SLA could be implemented as follows
  - The (*d*, *p*) SLA is defined
  - The SLA is measured on the working system (for example, the number of packets exceeding the delay threshold should be counted over pressigned time intervals)
  - A summary evaluation of the received QoS should be produced on the basis of the measured data
  - The customer can establish if it has received the contracted QoS
- In real life, it is not always easy to reach an agreement on how QoS is measured

Sample delay-oriented Service level Agreement (V)

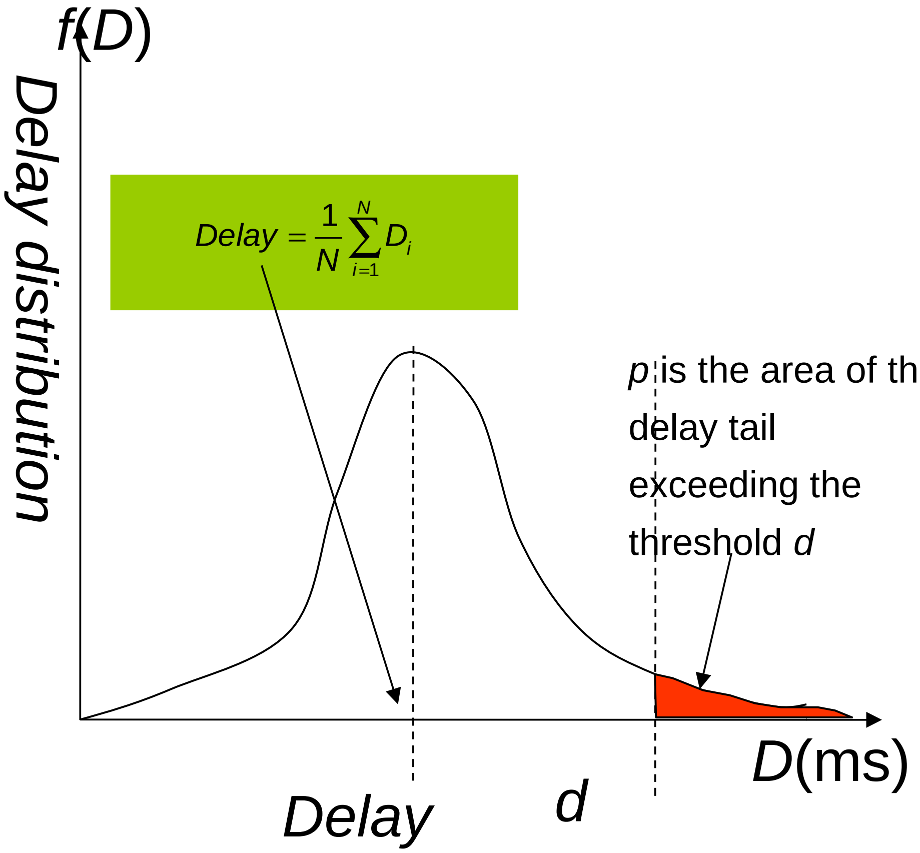
- For example, the provider TELCO proposes to its customer ACME the following method to measure quality
- Given a SLA (*d*, *p*):
  - Periodically, every Δt (s), an end-to-end ping packet is sent to measure delay
  - Let us define D_i as the delay of the *i*th ping packet
  - In a time period $T > \Delta t$, T/ $\Delta t = N$ ping packets are sent and their delay is measured
  - The following formula to calculate delay is proposed

$$Delay = \frac{1}{N}\sum_{i=1}^{N} D_i$$

- If Delay>d, the SLA has been violated
- *Is there anything strange with this method?*

Sample delay-oriented Service level Agreement (VI)

- Yes, there is something strange!
- *Delay* is actually the average value of end-to-end delay
- It is practically impossible that the average delay is greater than a delay threshold *d*, if *d* is set greater than average delay
- In this way, the result of the measures would be that the SLA is always fulfilled, even in the case in which the fraction of packets exceeding the delay threshold d is greater than *p*
- Note also that the *p* parameter is not considered in the proposed method


$f(D)$
Delay distribution
$Delay = \frac{1}{N}\sum_{i=1}^{N} D_i$
p is the area of the delay tail exceeding the threshold *d*
D(ms)
Delay
d

Sample delay-oriented Service level Agreement (VII)

- The customer, ACME, complains with the provider TELCO about this way of measuring the SLA
- ACME asks for a better method and TELCO replies with an alternative proposal
    - The previous system of ping packets is maintained
    - The 10% of highest delays is removed
    - Moreover, if a delay measure *Di* occurs when a router (any router) or a link (any link) is overloaded, that measure is removed
- Clearly, also with this system the SLA results as always fulfilled
- In fact, a high delay is necessarily a consequence of a loaded router and/or link and, with this definition, all highest delays are removed

Sample delay-oriented Service level Agreement (VIII)

- How a thorough method for measuring a statistical delay SLA could be defined?
- For example:
- Given a SLA (*d*, *p*):
  - Periodically, every Δt (s), an end-to-end ping packet is sent to measure delay
  - Let us define D_i as the delay of the *i*th ping packet
  - In a time period $T > \Delta t$, T/ $\Delta t = N$ ping packets are sent and their delay is measured
  - Let N_1 be the number of packets exceeding the delay threshold d
  - If $N_1/N > p$, the SLA has been violated

Sample delay-oriented Service level Agreement (IX)

- This experience suggests that guaranteeing end-to-end statistical delay SLAs is difficult
- Provider may want to seek for some protection in the contract with customers
- The correct path towards a thorough QoS guarantee is to acquire the knowhow on how to manage bandwidth in order to assign to each service class the amount of resources necessary and sufficient to obtain the required QoS

Quality of service: basic building blocks

In this chapter we introduce the basic buiding blocks of a network architecture for the provisioning of QoS. We will deal with traffic regulators (policers, shapers), admission control and resource provisioning, traffic conditioning, scheduling, packet classification and active queue management

Traffic Conditioning Agreement and Service level Agreement

In order to guarantee QoS, the provider and the customer must preliminarily stipulate a *Traffic Conditioning Agreement* (TCA) and a *Service Level Agreement* (SLA)

The SLA specifies the target QoS that the provider is committed to deliver to the customer for a specified set of traffic flows

However, it is neither possible nor reasonable that the provider is obliged to fulfill the SLA independently on the amount of traffic that the customers offers to the network

Some upper limit to the amount of customer's traffic for which the provider must meet the SLA must be specified

This "upper limit" is specified in the TCA

The TCA specifies a *traffic profile*

For a given reference flow, the part of traffic complying with the TCA is called *IN traffic* or *conformant traffic*

The part of traffic exceeding the TCA is called *OUT traffic* or *non-conformant traffic*

Traffic Conditioning Agreement and Service level Agreement

The SLA must meet the SLA only for IN traffic, while a number of actions can be made on OUT traffic

In fact, the provider must protect other established SLAs from excess traffic that a customer may offer to the network

Examples of actions on OUT traffic

Policing: OUT traffic is dropped

Shaping: OUT traffic is delayed until it is possible to send it complying with the TCA

Marking: OUT traffic is marked and offered to the network; in case of congestion, PUT traffic is dropped first

Traffic Conditioning Agreement and Service level Agreement

The rationale of this strategy is that in order to guarantee a SLA for a traffic flow or a flow aggregate, the provider must reserve a suitable amount of resources

This amount of resources depends on both TCAs and SLAs of connections sharing the capacity of a link

Resources are allocated for IN traffic

Once resources are allocated, accepting OUT traffic is a risk, because OUT traffic may consume resources allocated for other flows, whose SLA may be in turn violated

Accepting OUT traffic in the network is a delicate issue; it can be done, but marking such traffic in order to drop it if it degrades other SLAs

Traffic Conditioning Agreement and Service level Agreement

However, the provider in general would like to transport as much traffic as possible (for example, billing could be based on the volume of transported traffic, among other billing metrics)

Thus, if the network is not congested, the provider may want to transport OUT traffic in order to use momentarily free resources

OUT traffic may be

Marked (i.e., transported with a higher dropping priority in case of congestion)

Transported with lower service priority

Traffic conditioning agreement

The TCA can include a variety of parameters in order to characteriza IN and OUT traffic

Such parameters usually include

Peak rate of traffic

Average rate of traffic

Maximum length of bursts (i.e., the maximum number of consecutive packets transmitted at the peak rate of traffic)

Maximum length of packets

Minimum length of packets

The TCA specifies the statistical profile of IN traffic, that is, of the traffic for which the SLA must be fulfilled

Traffic conditioning agreement and regulators

Once the TCA and the associated SLA are established, the user-generated traffic is examined, at the network ingress, by a *regulator*

The regulator splits the user's traffic into (at least) two logically separated flows

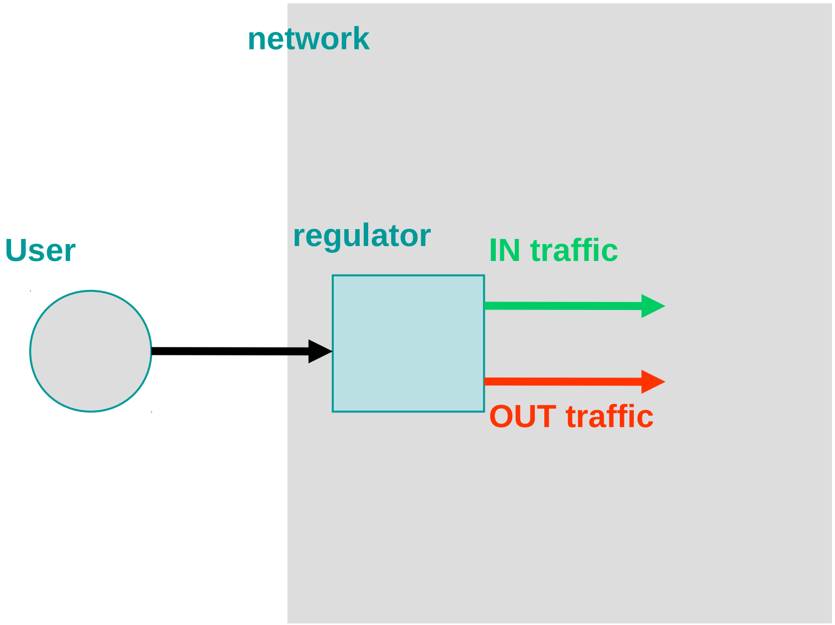
IN traffic (referred to also as green)

OUT traffic (referred to also as red)

There exists also regulators splitting traffic into three logically separated flows, i.e., green, yellow, and red

In the figure, a *two-color* regulator is depicted

If also yellow is distinguished, the regulator is called *three-color*



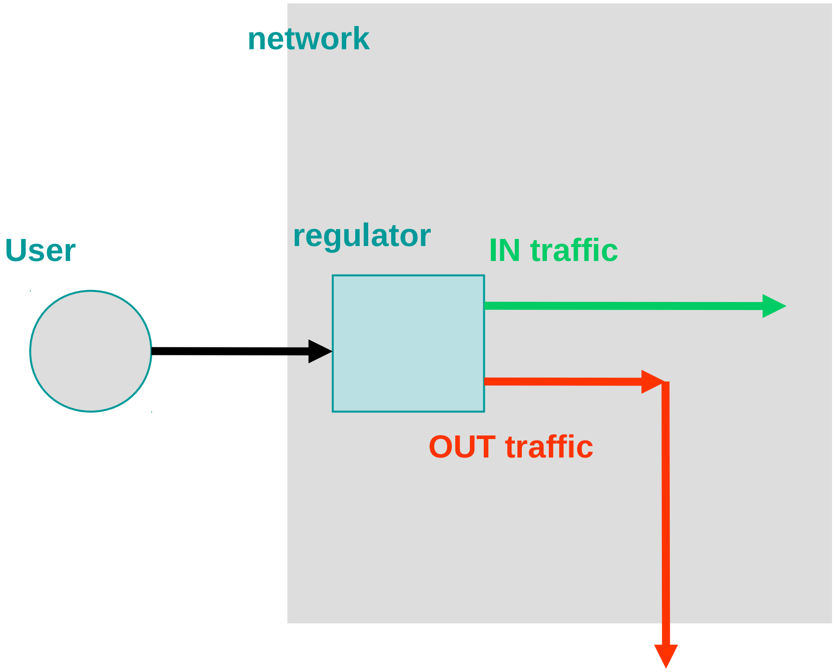
Policers

Policing regulators (policers) drop OUT traffic

Only green traffic proceeds into the network

In this way, other already established traffic flows are always protected from excess traffic of other users

However, if the network has spare resources, these resources are not used



Shapers

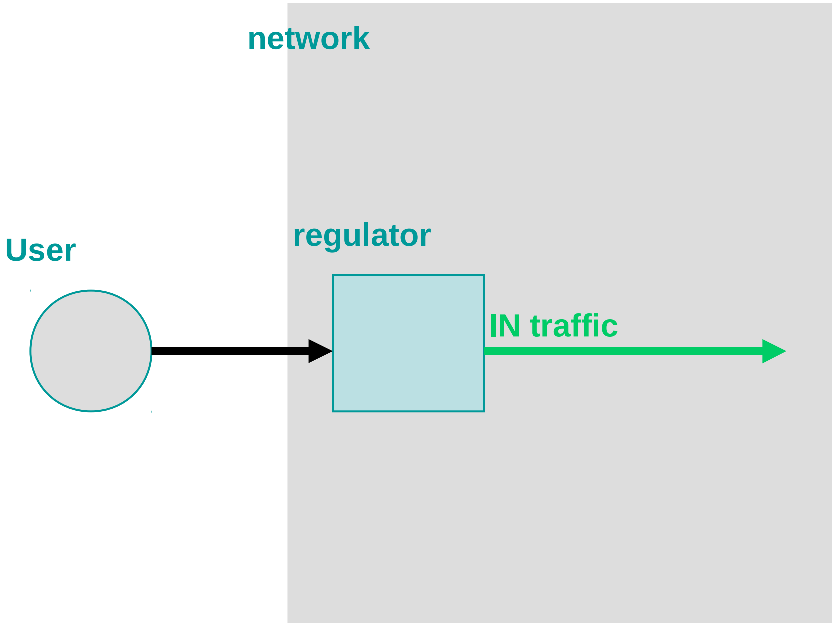
Shaping regulators (shapers) delay OUT traffic in a buffer, inside the regulator, in such a way it is transmitted into the network only when it is possible to do it without exceeding the TCA

Traffic entering the network is always green

Also in this case, other already established traffic flows are always protected from excess traffic of other users

Also in this case, if the network has spare resources, these resources are not used as OUT traffic is delayed, however, resources are in general used more efficiently than with a policer

The traffic backlog and delay inside the regulator may become large



Markers

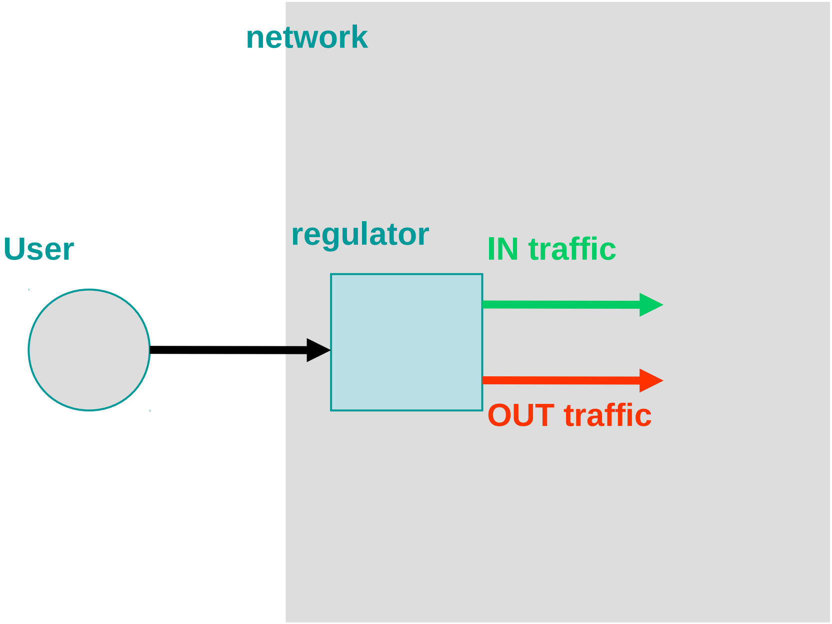
Markers let OUT traffic proceed, but this traffic is marked

Marking can be done in such a way to increase the dropping priotity of this traffic component

Alternatively, the SLA of this traffic may be downgraded, for exampled, it could be forwarded as Best-Effort

If the network has spare resources, they can be used more efficently

However, the issue of congestion management becomes important



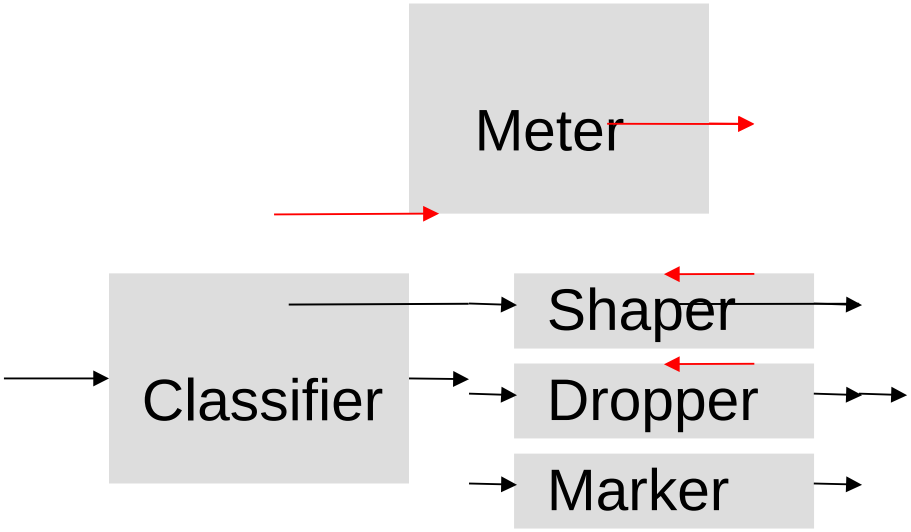
Traffic conditioner

A traffic conditioner may contain the following elements: meter, marker, shaper, and dropper

A traffic stream is selected by a classifier, which steers the packets to a traffic conditioner

A meter is used to measure the traffic stream against a traffic profile and the result of metering can affect a marking, dropping, or shaping action

The figure shows the block diagram of a classifier and traffic conditioner



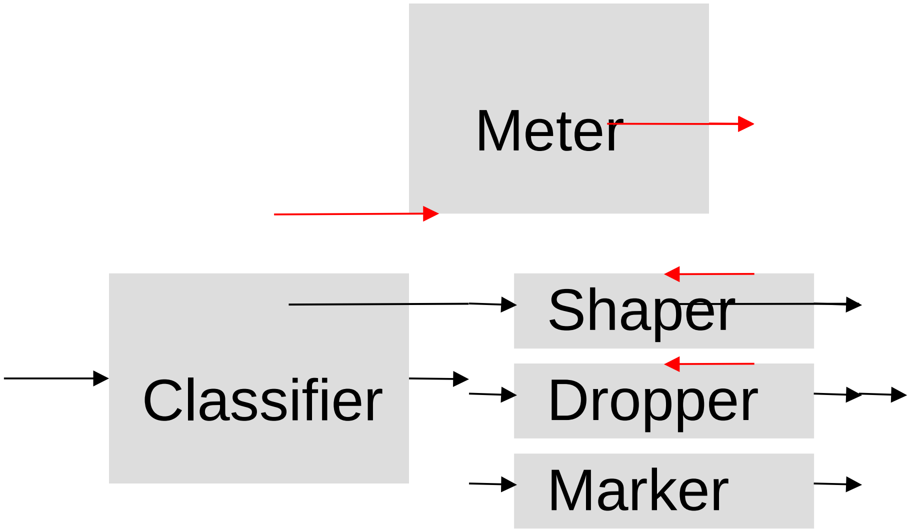
Black lines represent the flows of packets

Red lines represent control information

Traffic conditioner

The traffic meter measures the temporal properties of the stream of packets selected by a classifier against a traffic profile specified in a TCA

The meter passes state information to other conditioning functions to trigger a particular action



Black lines represent the flows of packets

Red lines represent control information

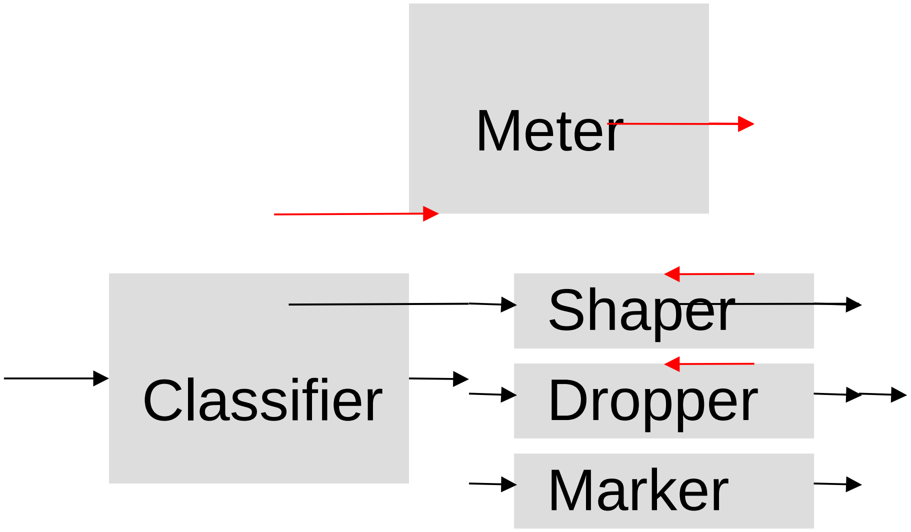
Traffic conditioner

The marker sets the DS field of a packet to a particular codepoint, adding the marked packet to a particular DS behavior aggregate

The marker may be configured to mark all packets which are steered to it to a single codepoint

Alternatively, it may be configured to mark a packet to one of a set of codepoints used to select a PHB in a PHB group, according to the state of a meter

For example, OUT packets may be re-marked and assigned to an “inferior” PHB



Black lines represent the flows of packets

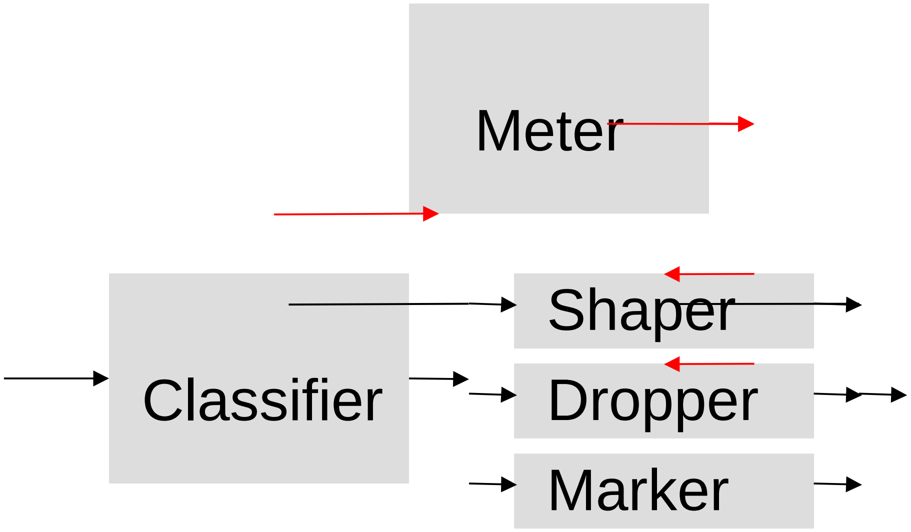
Red lines represent control information

Traffic conditioner

The shaper delays some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile

A shaper usually has a finite-size buffer, and packets may be discarded if there is not sufficient buffer space to hold the delayed packets

Actually, practical shapers can be referred to as “shapers-droppers”



Black lines represent the flows of packets

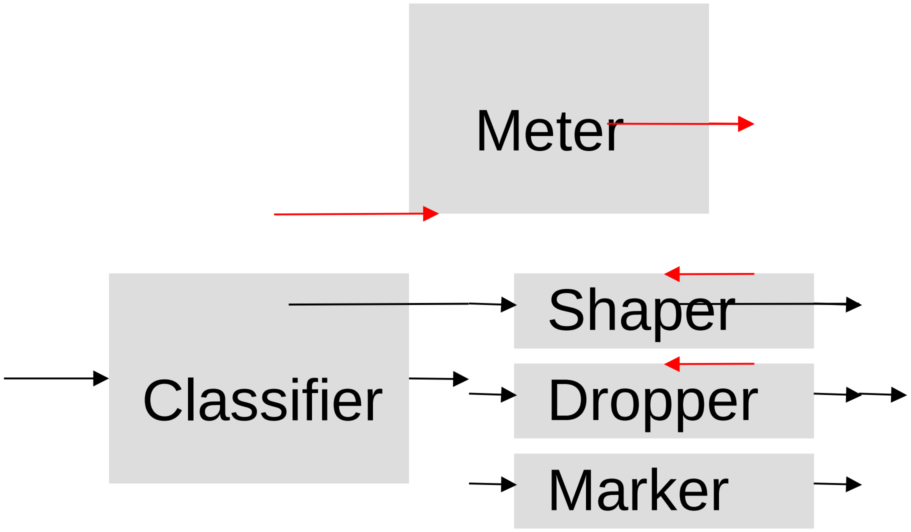
Red lines represent control information

Traffic conditioner

Droppers discard some or all of the packets in a traffic stream in order to bring the stream into compliance with a traffic profile

This process is know as "policing" the stream

Note that a dropper can be implemented as a special case of a shaper by setting the shaper buffer size to zero (or a few) packets



Black lines represent the flows of packets

Red lines represent control information