#43

A First Step in the World of IP QoS

Nicolas Simar

DANTE IN PRINT is a track record of papers and articles published by, or on behalf of DANTE. HTML and Postscript versions are available from: http://www.dante.net/pubs/dip/

For more information about DANTE or ***DANTE IN PRINT*** please contact:

DANTE
Francis House
112 Hills Road
Cambridge CB2 1PQ
United Kingdom

Tel: +44 1223 302 992
Fax: +44 1223 303 005
E-mail: dante@dante.org.uk

A First Step in the World of IP QoS

Nicolas Simar

Abstract

This paper sets out to explain the benefits of Quality of Service (QoS) in comparison with a basic service, such as Best Effort. The content is generally directed towards those who are not experts in QoS but who wish to have an overview of its benefits. The approach is mainly on a per router basis.

KEYWORDS: Quality of Service/QoS, Best Effort, IP Premium, IP Packet Delay Variation, IPDV, throughput-based service, end-to-end QoS.

1. Introduction

We will begin by defining basic notions such as the Best Effort service, which is the standard IP service. Next we will define congestion. Finally, we will give a general definition of QoS. This paper is based on the diffSERV architecture, which implies that it will investigate how the router treats the packets.

The second part of this document is based on interviews conducted with network users. The interview results showed that some applications require guarantees on parameters. Out of the required guarantees, it is possible to extract two different sets. For each of these sets, we will define their parameters. Then for each of these parameters, we will explain how the Best Effort service prevents the user from obtaining the expected guarantees and present a solution for achieving these guarantees.

Finally, we will outline the issues to be considered before the implementation of QoS.

Nicolas Simar is a Network Engineer in the Network Engineering and Planning team at DANTE and is an active participant in the SEQUIN project. His email address is <nicolas.simar@dante.org.uk>.

2. Definitions

2.1. Best Effort

By nature, the basic IP service available in most of the network is the Best effort (BE).

From a router point of view, this service could be described as follows:
When a router receives a packet:

- first it will determine where to send the incoming IP packets (the next-hop of the packet). This is usually done by looking up the destination address in the forwarding table[1] .

- Once it is aware of the next-hop, it will send the packet to the interface associated to this next-hop. If the interface is not able to immediately send the packet, it is stored on the interface in an output queue.

- If the queue is full, the arriving packet is dropped. If the queue already contains packets, the newcomer is subjected to extra delay due to the time needed to emit the older packets in the queue.

All the packets are treated equally. There are no guarantees, no differentiation and no attempt at enforcing fairness. However, the network should try to forward as much traffic as possible with reasonable quality. This leaves freedom with respect to how the reasonable quality should be optimised for everybody (FIFO output queue or Active Queue Management –Random Early Drop).

[1] *The type of routing is not a part of the quality of service definition.*

2.2. Congestion

A congestion period on an interface is when a router starts to build an output queue on this interface. This happens when an interface receives, from the router switch fabric, more traffic than it is able to forward.

During a period of congestion, (in the case of Best Effort behaviour) the packets suffer extra delay due to the time needed to transmit the packets which were previously enqueued in that interface. The queue sizes vary quite considerably. Once the output queue is full (all the output buffering capacity of the interface has been allocated), the new packet is dropped.

In order to provide the guarantees (on delays, IPDV, etc) required by applications such as voice over IP, or video-conferencing, the IP packets have to cross routers with empty or nearly empty queues. Mechanisms are thus needed in order to provide guarantees to the traffic generated by such applications.

2.3. Quality of Service

One way to provide a guarantee to some traffic is to treat the packets differently from packets of other types of traffic. This is where Quality of Service (QoS) comes into view. One definition for QoS is: the ability of a network element to have some level of assurance that its traffic and service requirements can be achieved [NfQ].

Based on user needs, we have identified four different parameters requiring guarantees. We will define these parameters and highlight what could prevent a Best Effort based service from achieving these guarantees. We will then suggest some solutions based on an IP Differentiated Services (DiffSERV) domain [RFC2475].

DiffSERV is a layer-3 framework to provide control to aggregate of flows. At the edge of a DiffSERV domain, packets are classified into flows and the flows are conditioned (marked, policed or shaped) to a traffic conditioning specification. The flows are then aggregated. A DiffSERV Codepoint (DSCP) identifies a per-hop behaviour (PHB) and is set in each packet header.

The DSCP is carried in the DS-field, which is formed from six bits of the IP header former ToS byte [RFC2474]. The PHB is the forwarding behaviour which is to be applied to the packet in each node in the DiffSERV domain (the Best Effort paragraph describes the BE per hop behaviour).

3. Service Needs

The first step is to identify the user needs and understand their application requirements.

The SEQUIN project[2] conducted an interview to try and understand the users' needs for Quality of Service, their perception of QoS, the applications they intend to use and how their networks are used [Interv]. A questionnaire was sent to twenty pan-European groups who could make use of QoS. Ten answers were received.

From these answers, there would appear to be a clear need for two types of service in addition to the Best Effort service. A time-based service and a throughput-based service. For these two types of service, four parameters requiring guarantees were identified: delay, jitter, bandwidth and loss. The time-based service will be named "IP Premium" in accordance with the TF-NGN and SEQUIN work.

For each of these parameters, we will provide a definition of what could prevent the BE per hop behaviour from satisfying these requirements (this list is not intended to be exhaustive) and we will explain how other services could also try and fulfil the requirements.

3.1. IP Premium Service

The IP Premium service is used for real time applications, voice over IP, video-conferencing etc. These applications require low one-way delay, low jitter and low loss. The IP Premium serv-

[2] *http://www.dante.net/sequin/*

ice provides upper-bounded one-way delay, upper-bounded one-way packet delay variation, guaranteed bandwidth and zero or low loss.

3.1.1. One-way Delay

+ Definition

The one-way delay could be defined as the time between the emission of the first bit of an IP packet by the source and reception of the last bit of this packet by the receiver [RFC2679].

A distinction between the delay required by the application and the one-way delay introduced by the network has to be made. The user has to take into account that the operating systems (interruptions), the packetisation, the compression etc. can introduce some extra delays which are not taken into account by the “network one way-delay”.

The one-way delay is equivalent to the sum of the single-hop delays suffered between each pair of consecutive pieces of equipment encountered on the path.
A pair of equipment is made up of the emitting equipment and the receiving equipment.

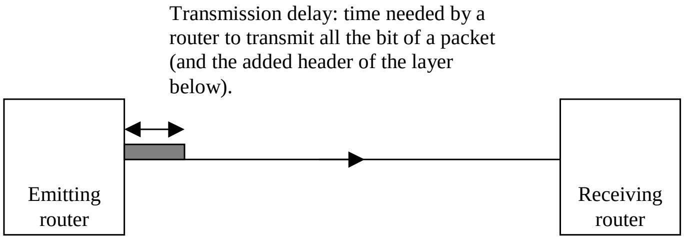
Transmission delay: time needed by a router to transmit all the bit of a packet (and the added header of the layer below).

Figure 1. Transmission delay

The single-hop delays suffered between two piece of equipment consist of:

- transmission time: time taken to transmit all the bits of the frame containing the packet, i.e the time between emission of the first bit of the frame and emission of the last bit. It is inversely proportional to the line speed. (Fig. 1)

- propagation delay: the time between emission (by the emitting equipment) of the first bit (or the last bit) of a packet and the reception of this bit by the receiving equipment. It is mainly a function of the speed of the light and the distance travelled. (Fig. 2)

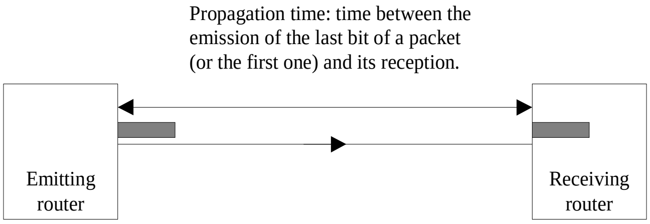
Propagation time: time between the emission of the last bit of a packet (or the first one) and its reception.

Figure 2. Propagation delay

- equipment delay: delays introduced by the receiving piece of equipment of the pair. It consists of all the delay introduced by this equipment before it becomes an emitting equipment of a pair by sending the first bit of the frame to the next piece of equipment. This delay consists of the processing time, the packet switching, the queueing delays etc.

We consider that the equipment delay of the receiver equipment (at the other end of the path) is considered to be equal to zero. The end user has to keep it in mind.

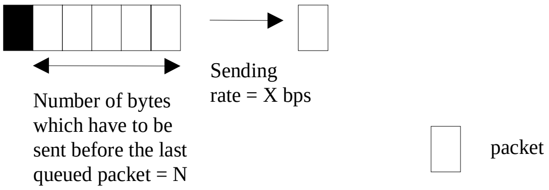
Sending rate = X bps
Number of bytes which have to be sent before the last queued packet = N
packet

*Figure 3. Congested interface output queue. The last packet in the queue (in black) has to wait (N*8) / X sec before being emitted by the interface.*

+ Best Effort Inadequacies

The first two delays are fairly constant. The third one is more important in case of congestion and increase with the load[3].

The queueing delay is the most “dangerous” delay for the application requiring an upper-bounded delay as the queue length varies quite considerably. A congested STM-1 interface with a 55 Megabyte First In First Out queue can in-

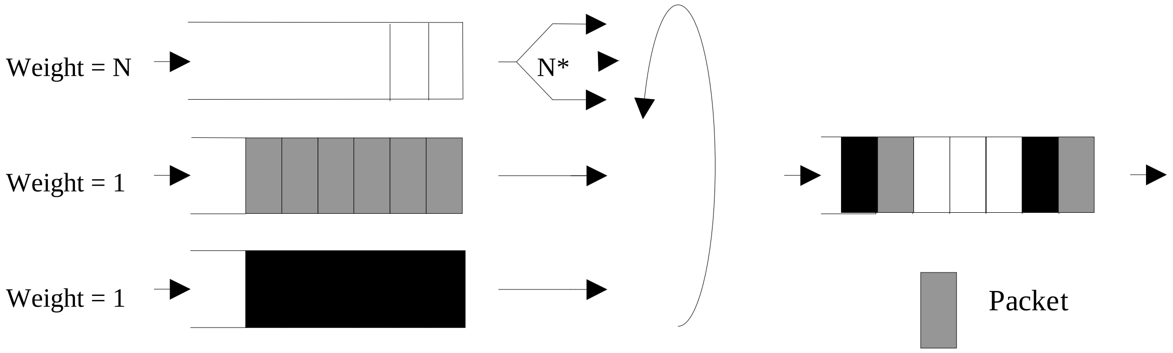

Figure 4. Congested interface with a Round Robin like scheduling algorithm. The IP Premium packets (white) are sent into a separate output queue to the other types of traffic. This output queue has a weight of N which allows it to be emptied more quickly than the others. During the next round, the N first packets of the white queue will be de-queued (in this example, all the packets have the same size) and one packet of each other queue will be de-queued. N could be 98, so we are serving 98 packets of the white queue, this means a rate of 98% of the link capacity (the packets having, in this example the same packet size).

troduce three seconds of queueing delay. On a path between end-users, you can easily encounter several congested interfaces (with different "level" of congestion). Now imagine the impact of such heavily congested interfaces on an end-to-end application requiring an upper bounded one-way delay of 150ms.

+ Scheduling Mechanism

One solution is to give priority, in the output queue, to the packets requiring an upper-bounded delay over the other packets. This can only be done if the load of the upper-bounded one-way delay packets is bounded so there is no queueing or loss in this high priority class. In such a way, the queueing delay of these packets can be minimised and kept under a certain value, even in the case of interface congestion.

To reach this goal, we should be able to differentiate these packets from the other one. This can be achieved by:

- writing a different DSCP value other than the BE one in IP header.
- classifying these packets in an output queue other than the BE ones.
- using scheduling mechanisms such as:
  * Strict Priority queueing, where the IP Premium service packets are assigned to a queue which has an absolute priority over all other queues. The queue is served until it is empty.
  * Weighted Fair Queueing (WFQ [Keshav97]), where IP Premium service packets are assigned to a queue, which has a relatively high weight in comparison to all the other ones.
  * Deficit Round Robin (DRR [Sheer95]), Modified Deficit Round Robin (MDRR [Sreen]), Weighted round Robin (WRR), where all the other queues are restricted to a small round-robin time-shared where the IP Premium service queue is served more frequently.

  The goal is to try to have the IP Premium service queue as empty as possible.
- the packets are able to go straight to the high Priority queue without risk of starving the other queues (and the other traffic types they contain), because a strict policing is applied on those packets at all the ingress points.. This is very important in case of use of strict priority queues.

[2] One should be careful with traffic engineering techniques not to increase the delays by choosing a non optimal path. This means other link metric than those proportional to the propagation delays (or transmission delay in case of short distance between equipment). If pure shortest-path routing would lead to congestion on some paths, then traffic engineering is one possible way to reduce the "queueing delay" by distributing traffic to less congested paths. However it degrades the service by making traffic take a longer path than would be necessary if the network was adequately provisioned along the shortest path.

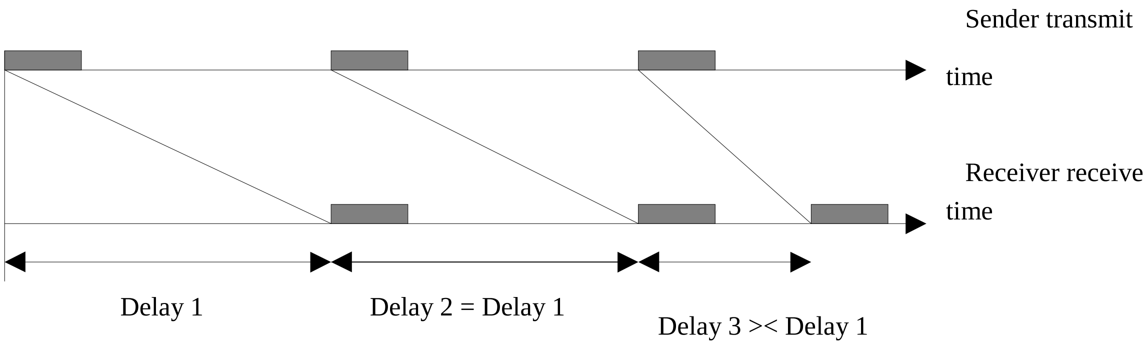

Figure 5. IPDV is the difference between the one-way delay of two packets. IPDV between packets 1 and 2 is null. But the IPDV between packet 2 and 3 is different of zero. This implies that the receiver will not receive packets with a regular interval.

These mechanisms allow us to reduce the one-way delay in case of congestion and to quantify the worst case delay queueing.

3.1.2. IP Packet Delay Variation

+ Definition

According to the IETF IPPM group, the IP Packet Delay Variation (IPDV) of packets within a stream [IPDV] is defined as:

"The IP Packet Delay Variation of a pair of packets within a stream of packets, defined for a selected pair of packets in the stream, going from measurement point to another one, is the difference between the one-way-delay of the first of the selected packets and the one-way-delay of the second of the selected packets."

This is a measure of the packet inter-arrival time. Some applications (such as voice) required a regular arrival of packets.

+ Best Effort Inadequacies

As the IPDV is defined as the difference between the one-way delay of two packets, its value is influenced by the variation of the one-way delay. The main factor influencing the one-way delay is the queueing delay. Itself influenced, in case of congestion of the interface, by the output queue length.

The problem is to guarantee an upper-bounded limit on the IPDV, this means an upper-bounded variation of the one-way delay.

+ Scheduling Mechanism

It is important to avoid too huge a variation of the output queue (of the queueing delay). Sending IP Premium packets to a different output queue than the other packets and trying to minimise the IP Premium queue length can achieve this. The queue length can be minimised by using a scheduling algorithm such as that listed above and by allocating to the queue a weight proportionally higher than the other queue. In order to retain as few packets as possible, we can reduce and minimise the queueing delay variation. The solution proposed to upper-bound the one-way delay is also a solution which can be used to upper-bound the IPDV.

3.1.2. IP Packet Delay Variation

+ Definition

The packet loss can be defined as the percentage of packets sent but not received, or received in error. Its measurement is detailed in RFC 2680 [RFC2680].

The packet losses are due to failures, problems with equipment configuration, bad behaviour of the line or the equipment encountered on the path, congestion on the path etc. Some losses due to failure can be avoided by introducing redundancy.

+ Best Effort Inadequacies

In case of congestion, the congested interface starts to build its own output queue. Once the buffer space allocated to the queue is completely used (the queue is full), the interface starts dropping packet. These losses are unpredictable and a BE packet can be lost as well as a packet requiring guarantees.

+ Scheduling Mechanism

Once again, the solution is to classify the BE packets and the IP Premium packets into two different output queues. We have to try to keep the IP Premium queue as short as possible in order to avoid it using all its allocated resources and dropping packets. This can be achieved by allocating to it a proportionally bigger weight than to the other output queues. The packet loss is then avoided and the one-way delay is still upper-bounded.

3.1.4. Bandwidth

Bandwidth, in this context, is the amount of data, in bits per second and at an IP level, which is transferred from one end of a path to the other (one user to the other one). The bandwidth (capacity) can be specified by a maximum burst size, a peak bandwidth, a minimum assured bandwidth and an average bandwidth value.

We will focus mainly on a leased line type of capacity. Leased line type of capacity means that the value of the peak bandwidth will be equal to the minimum bandwidth and to the average bandwidth. The burst size will be equal to one MTU.

The losses on the path between the two end-users, together with the high round-trip time (eg for TCP), (thus delay) prevent the ability to provide a throughput guarantee. In order to guarantee the capacity, we have to minimise the losses (cfr previous parameter) and limit the queueing delays.

It is not possible to guarantee high (proportionally to the link capacity encountered on the path) IP Premium bandwidth. If the Bandwidth used by the IP Premium traffic is too high, it could badly damage its own performances and that of the other traffic in times of congestion, as the IP Premium traffic has some precedence over the other types of traffic. We have to bear in mind that, to get good performance for one type of traffic, you must, in congestion time, reduce the performance of other types of traffic.

This behaviour corresponds to the Expedited Forwarding Per Hop Behaviour [RFC2598] and provide a IP Premium service.

3.2. Throughput Based Service

The second service required by the users is a guaranteed throughput service with no commitment on the one-way delay and IPDV. This service can be applied to projects which want to ensure a certain amount of connectivity between two sites.

The throughput can be seen, for congestion-aware transport protocol such as TCP, as Bulk Transport Capacity = data_sent / elapsed_time where “data_sent” represents the unique “data” bits transferred (i.e., not including header bits or emulated header bits).

It should also be noted that the amount of data sent should only include the unique number of bits transmitted (i.e., if a particular packet is retransmitted the data it contains should be counted only once). [RFC3148] This definition of throughput for congestion-aware transport protocol is user oriented because the network still has to transport the packets which have been lost (and which have to be retransmitted). We must take them into account and it is necessary to make some adaptations between this definition and what the network has to carry in reality.

With the BE service, the throughput guarantee cannot be assured because of losses induced by congestion due to other traffic. So, it is necessary to differentiate between the guaranteed throughput packets and the other ones and assure them proportionally less loss than the Best

Effort service packets. The throughput is also limited by the value of the round-trip time.

The use of an Assured Forwarding Per Hop Behaviour (AF PHB) [RFC2497] can achieve this. We will suggest one solution out of several. The proposed solution is to assure some bandwidth and, if there is some extra free capacity, the assured throughput flows can use it but this extra bandwidth will get the same guarantees as the Best Effort traffic.

Throughput guarantee can be achieved by:

- writing a DSCP value different to the other defined service ones in IP header for all the packets conforming to a certain envelope (average rate, burst rate etc). The packets which are not conforming to this envelope are marked with a different value.
- classifying the conforming and non-conforming packets in the same output queue as the BE ones.
- configure WRED on the output queues.

The BE and the out-of-profile packets should have a more aggressive drop policy than the guaranteed bandwidth packets. Their WRED minimum and maximum thresholds are smaller than the guaranteed bandwidth packets to cause the drop of these first packets before the drop of the others, in case of congestion. The guaranteed bandwidth packets are assured to have a very low packet loss and out-of-envelope packets have the same drop probability as the BE ones.

This will give a guaranteed bandwidth, and a bit more, even in times of congestion and this behaviour provides a guaranteed bandwidth service.

4. Issues to Consider for an End-to-End QoS

In the previous chapter, we have taken a per hop behaviour point of view. But the users' requirements are end-to-end. This chapter will mainly present different issues which could be encountered in the deployment of such service within a single domain or across several domains. These issues are mainly related to the way in which we are implementing the services.

Should a user wish to use a different service than Best Effort, what should they do and from whom should it be requested? Will the QoS reservation across multiple domains be placed via a web interface, via a single phone call or via a bunch of e-mails, phones calls etc? What type of information will the user have to give, their own IP address, the IP address of the destination, the bandwidth they will use to send and receive? The answers to these questions will mainly depend on how the service are to be implemented.

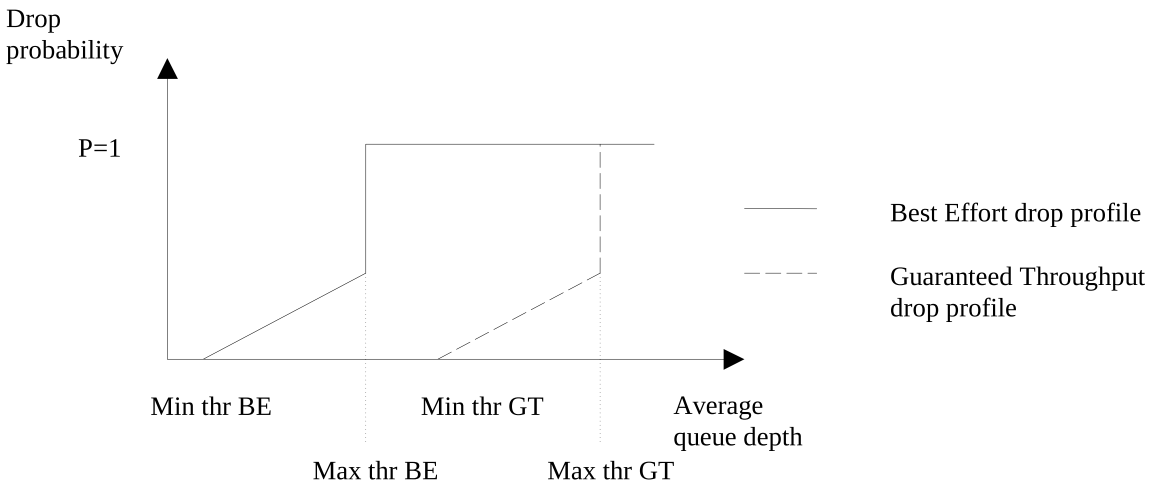

Figure 6. Best Effort Drop profile and the in-profile Guaranteed Throughput drop profile. When a queue start to be built, the BE packets and out-of-profile have a higher drop probability than the in-profile Guaranteed throughput.

4.1. Destination Aware or Destination Unaware.

Two different types of service implementation can be used: destination aware and destination unaware.

The first one, destination aware, is close to a point to point service. The service has been requested from point A to point B. It allows precise dimensioning of the requirements in the network. The maximum resources used at each node can be evaluated. This implies an admission control based on DSCP value, IP source and destination addresses. This requires very good collaboration between domains and quite a bit of administrative work if it is not automated.

For the second model, destination unaware, the user has requested a certain amount of service, independently of the destination. An admission control, based on a single envelope at the ingress, has to be done per DSCP value. The capacity used at each egress point of the network cannot be known in advance and, in the worst case, it could be equal to the sum of the ingress. The ingress is sending the traffic to the same egress, which has some capacity limitation. There is a need to limit the total ingress capacity accordingly to the egress capacity. This is easier to implement than with the previous model but could imply that less capacity for the service can be reserved.

4.2. Packets Marking and Policing

The user has to mark their packets with the DSCP value corresponding to the service required. Can we trust other users not to use this service without authorisation (for boosting their internet connection or for DoS attacks)? The answer is no. Can we trust a user or domain not to send more traffic than what has been subscribed for? In general, the answer is yes, but we have to prevent any accidents which could damage performance for the other users.

A check of “who uses what” must be done by the first router encountered on the end-to-end path. This check has to be done based on the source address, destination address, and a traffic envelope. All excess traffic coming from an authorised source has to either be re-marked or be dropped (policing). Traffic originating from an unauthorised source having a different DSCP value to the Best Effort one has to be re-marked as BE (or be dropped).

Each ingress router for a domain has to check the aggregate that other domains are sending to it. The check is based on the DSCP value and an envelope (a description of the traffic behaviour). The excess packets can either be dropped or be remarked as Best Effort. These drops/ remarks have to be monitored and a signal must be sent to the domain indicating the submission of non-conforming traffic.

The main issue is the ability of the ingress routers to perform such large scale policing (e.g access list in hardware).

4.3. Remarking

In order to allow differentiation of the packets by the network and to treat them in the appropriate way, the packets have to be marked as being part of one class of service.

Not all applications are able to mark the IP packets with the DSCP value corresponding to a service. That is why the first router encountered has to be able to re-write the DSCP value of an IP packet.

The DSCP value characterising a service can vary from one domain to another. The ingress or egress routers of a domain, depending on the router capabilities of the two domains, have to be able to re-write it.

We have to keep in mind that, when a DSCP value is attributed to a service, a few routers cannot perform some operations based on the DSCP value, only on precedence field. The precedence field is formed from the three first bits of the DSCP field.

4.4. Shaping

In order to reduce the burstyness of the traffic, the shaping has to be done as close to the source as possible. This could also be done at the egress of the network to avoid forwarding a bursty aggregate to the next domain. The main use of a shaped traffic is to avoid receiving traffic burst on an outgoing interface. This could result in too quick a build up of the queue and potentially drop packets of a source irrespective of its envelope.

4.5. Monitoring

Monitoring is needed in order to provide capacity to cope with the traffic evolution and to prove to the end-user that the domain respects its part on the end-to-end path performances.

Ideally, the main parameters of a service (e.g. one-way delay, IPDV, throughput and loss for IP Premium) have to be monitored on an end-to-end basis by the end-users. If the performances are worse than those expected, the end user has to report it and should be able to check the contribution to this value of each encountered network/domain on the path.

Each domain should be able to produce some graphs of each parameter for each ingress-egress pair of access/connection. They should be able to write a DSCP value.

Between two measurement boxes/software there should be two “measurement connections” (one in each direction) to measure the one-way delay, IPDV and one-way losses for each used DSCP value. A measurement box/software should be placed in each PoP of the domain where an access/connection is landing. Each box/ software tool should be connected with a full mesh of “measurement connections” per used DSCP value to the other box/software of the domain (one-way delay, jitter loss). A protocol also has to be defined to allow two measurement boxes to talk to each other from one domain to another.

Depending on the accuracy and topology required, GPS or NTP will be needed to synchronise the measurement boxes.

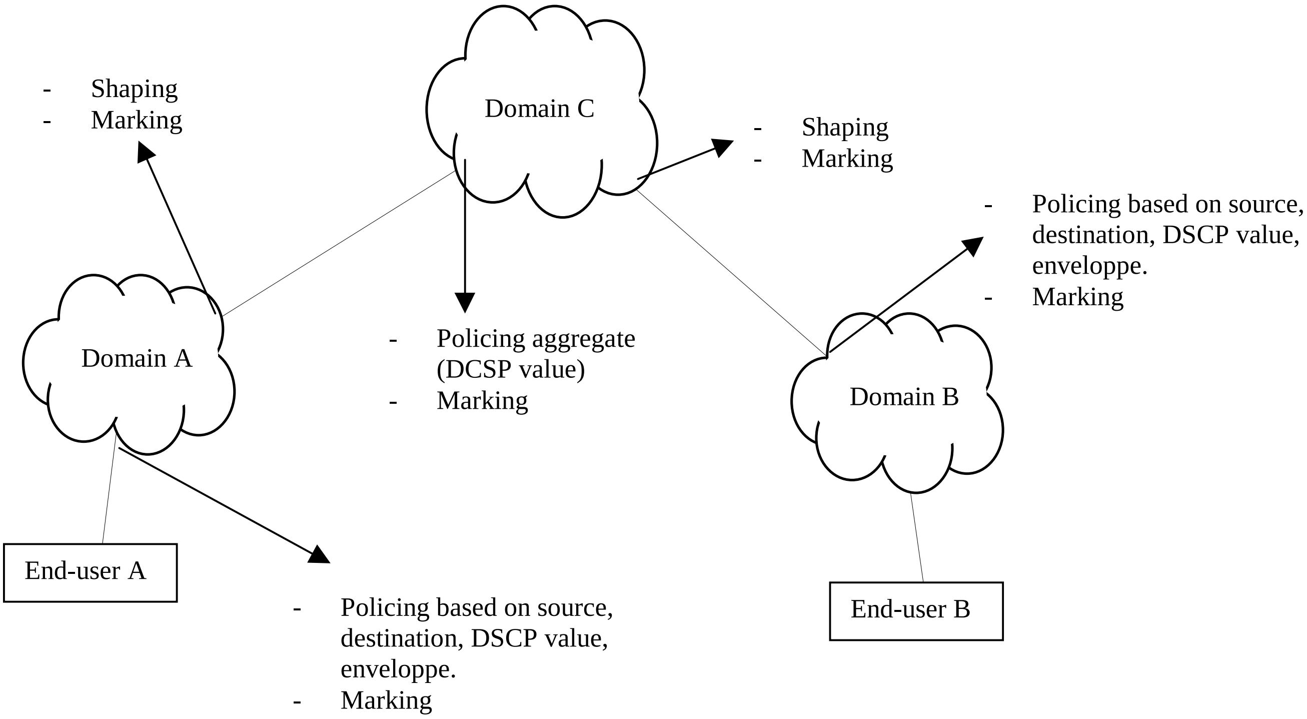
- Shaping
- Marking
Domain C
- Shaping
- Marking
- Policing based on source, destination, DSCP value, enveloppe.
- Marking
- Policing aggregate (DCSP value)
- Marking
Domain A
Domain B
End-user A
End-user B
- Policing based on source, destination, DSCP value, enveloppe.
- Marking

Figure 7. Shows where the policing, shaping and marking can be done. All these actions do not need to be done. It mainly depends on how the service is implemented and on the router capabilities.

4.6. Interconnection Between Different Technologies

Another issue is the interconnection between different technologies (ATM, MPLS, tunnelling, LAN etc). How can two implementations of services be matched? For example, we have two users exchanging Best Effort traffic and IP Premium traffic. One user in an ATM network, the other one on a DiffSERV network. At the edge of the two networks, how can we match the two users' IP Premium traffic into a CBR PVC and their best effort one with the best effort traffic?

This type of question will arise each time that two networks, deploying different technologies, are interconnected and for each type of material used in this interconnection.

5. Conclusion

At times of congestion, if applications require some guarantees, the Best Effort service is not sufficient and QoS mechanisms have to be introduced in order to provide some IP Premium or guaranteed throughput services. Therefore, it is interesting to deploy the QoS mechanism where congestion is expected. The implementation of such services implies a need to have appropriate equipment.

Finally, we must bear in mind that an "application-to-application" guarantee depends on the network but also on the equipment of the end-user.

References

[NfQ] -- QoS forum, *The need for QoS,* White Paper, July 1999

[Interv] -- SEQUIN, *Deliverable D2.1 – Quality of Service Definition*, Report, March 2001 - http://www.dante.net/sequin/QoS-def-Apr01.pdf

[RFC2474] -- *Definition of the Differentiated Service Field (DS Field) in the Ipv4 and Ipv6 Headers,* K. Nichols, S. Blake, F. Baker and D. Black, December 1998.

[RFC2475] -- *An Architecture for Differentiated Services*, S. Blake, D. Black, M. Carlson, E. Davie, Z. Wang and W. Weiss, December 1998.

[RFC2597] -- *Assured Forwarding PHB Group,* J. Hinanen, F. Baker, W. Weiss and J. Wroclawski, June 1999.

[RFC2598] -- *An Expedited Forwarding PHB,* V. Jacobson, K. Nichols and K. Poduri, June 1999.

[RFC2679] -- *A One-way Delay Metric for IPPM,* G. Almes, S. Kalidini, M. Zekauskas, September 1999.

[RFC2680] -- *A One-way packet Loss Metric for IPPM*, G. Almes, S. Kalidindi, M. Zekauskas, September 1999.

[RFC3148] -- *A Framework for Defining Empirical Bulk Transfer Capacity Metrics,* M. Mathis and M. Allman, July 2001.

[IPDV] -- IETF, *Instantaneous Packet Delay Variation Metric for IPPM* – http://www.ietf.org/internet-drafts/draft-ietf-ippm-ipdv-07.txt

[GÉANT] -- http://www.dante.net/geant/

[NREN] -- List of NRENs, http://www.dante.net/geant/connect.html

[Keshav97] -- *An engineering Approach to Computer Networking*, S. Keshav, Addison Wesley, January 1997.

[Sheer95] -- *Efficient FairQueueing Using Deficit Round Robin*, Sheerhar, M. and G. Varghese, SIGCOMM 1995, pages 231-242.

[Sreen] -- *Implementation and evaluation of support for Differentiated Services mapping to ABR service in an Edge/Core Network*, Sreenivasamurthy, Thesis, University of Texas.