INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2004; **17**:935–962
Published online 22 September 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dac.683

QoS guarantee in telecommunication networks: technologies and solutions

Enzo Fortunato[1,‡], Mario Marchese[2,*,†], Maurizio Mongelli[2,§] and Annamaria Raviola[1,¶]

[1] *Marconi Selenia Communications S.p.a., Genova, Via di Francia, 3, 16149 Genova, Italy*
[2] *CNIT, Italian National Consortium for Telecommunications, University of Genoa Research Unit, Via Opera Pia 13, 16145 Genova, Italy*

SUMMARY

The aim of this paper is to show the effect of the presence of specific management functions within a network that offers quality of service (QoS). The objective is not privileging a particular technology but to highlight the importance to know which control functions a solution may use, which performance limits the functions have and what can be a realistic user expectation.

The paper focuses on the meaning of QoS and on the applications requiring quality, then describes QoS solutions including transport technologies, QoS-oriented technologies, parameters and management functions. In more detail, the effect on QoS provision of the following issues is investigated and discussed concerning the possibility (or not) to aggregate and differentiate traffic, the implementation of call admission control and of traffic filtering to limit flows to their committed rates. Again, the conclusions should not be considered a merit mark about technology, but only an investigation about: what users and customers should expect by the technologies using specific control functions evidencing that the real limitations are not imposed by a specific technology, whose features may be changed and extended, but by the application of control functions that can guarantee requirements' matching.

KEY WORDS: telecommunications networks; quality of service (QoS); performance evaluation

1. INTRODUCTION

The recent evolution of networked multimedia applications and of the services to transport information over network backbones have highlighted the need to investigate techniques, tools and device configurations to guarantee a certain level of quality of service (QoS) to the end users.

*Correspondence to: Mario Marchese, CNIT, Italian National Consortium for Telecommunications, University of Genoa Research Unit, Via Opera Pia 13, 16145 Genova, Italy.
† E-mail: mario.marchese@cnit.it
‡ E-mail: enzo.fortunato@marconiselenia.com
§ E-mail: maurizio.mongelli@cnit.it
¶ E-mail: annamaria.raviola@marconiselenia.com

Received 31 July 2003
Revised 19 February 2004
Accepted 6 July 2004

The range of commercial applications requiring a fixed degree of QoS is very wide: from tele medicine to videoconference, from tele-education to remote control. The characteristics of th network used are extremely important to provide mechanisms to offer end-to-end Qo guarantee traversing multiple domains to support existing services and to create new dynami and flexible services. Some types of networks, i.e. asynchronous transfer mode (ATM) ones have been designed to support QoS for specific traffic flows. A statistic investigation to verify th availability of the resources to guarantee a fixed level of service is performed before accepting new call entering the network. On the other hand, the Internet is a 'best-effort' network, (sinc TCP/IP protocols have not been designed to provide guaranteed quality of service) and it onl does 'its best' [1]. It is characterized by heterogeneity both from the point of view of algorithm and management, and from the point of view of physical links. Two solutions are available t match the QoS request over IP networks: the integrated services approach and the differentiate services approach.

Multiprotocol label switching (MPLS), developed more recently, is a powerful ne technology, which may be used as a possible convergence between the IP world, which involve at the same time IP open standards and their simplicity and the ATM world, which completely oriented to the QoS and uses traffic control techniques.

The last mentioned point is a key factor. Independently of the technology chosen, QoS guaranteed only if some control mechanisms are imposed in the network. Packet marking, ca admission control (CAC), traffic shaping, scheduling, resource reservation, flow control, an QoS routing are essential to provide QoS. One of the objectives of this work is to show what the effect of various functions and that, independently of the interest towards fashionab technologies, which can detract from an appropriate analysis of QoS real needs, it is importa to have specific service functions that satisfy precise performance requirements.

This paper reports some possible definitions of QoS and tries to focus on meaningf applications concerning their QoS needs, in Section 2. Section 3 is structured into sub-sectio dedicated, respectively, to the technologies for information transport at the physical laye (layers 1 and 2, in this work), to a description of the main technologies in the market to provi QoS, and to a revision of the most meaningful management functions, already mention above. The aim of Section 4 is to show the effect of the presence of specific manageme functions within a network. Many results are reported to get to the aim.

2. QoS DEFINITION AND APPLICATIONS

2.1. QoS definition

According to ISO 8402, the word quality is defined as 'the totality of characteristics of entity that bear on its ability to satisfy stated and implied needs'. ISO 9000 defines quality the degree to which a set of inherent characteristics fulfills requirements. ITU-T (Recommend tion E.800 [2]) and ETSI [3] basically defines QoS as 'the collective effect of service performan which determine the degree of satisfaction of a user of the service'. As stated in Reference [IETF considers QoS as the ability to segment traffic or differentiate between traffic typ in order for the network to treat certain traffic flows differently from others. QoS encompass both the service categorization and the overall performance of the network for ea category.

Concerning the network viewpoint, QoS is the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements can be satisfied. QoS manages bandwidth according to application demands and network management settings.

The term QoS is used in many meanings, ranging from the user's perception of the service to a set of connection parameters necessary to achieve particular service quality. The QoS meaning changes, depending on the application field and on the scientific scope. Reference [5], starting from the terminology concerning QoS in IP networks, defines some reference points about the QoS issue. The authors, mentioning Reference [6], identify three types of QoS: intrinsic, perceived and assessed. Intrinsic QoS is directly provided by the network itself and may be described in terms of objective parameters as, for instance, loss and delay. Perceived QoS (P-QoS) is the quality perceived by the users; it heavily depends on the network performance but it is measured by the 'average opinion' of the users. Mean opinion score (MOS) methods are often used to perform the measure of the quality: users assign a rating as follows: 1—bad, 2—poor, 3—fair, 4—good, 5—excellent, to the application they are evaluating. The MOS is the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best) [7].

Even if there is a strict connection with the objective metrics provided by the network, the user does not necessarily perceive an objective performance increase (or decrease), in correspondence of an intrinsic QoS variation. Reference [8] reports a performance evaluation by using MOS values along with measures of objective metrics and shows that the correspondence is not always straightforward.

The last term reported concerns assessed QoS. It is referred to the will of a user to keep on using a specific service. It is related to P-QoS but also depends on the pricing mechanism, on the level of assistance of the provider and on other marketing and commercial aspects.

At the moment, most of QoS provision is offered in terms of intrinsic (objective parameters) QoS by using a service level specification (SLS) which is 'a set of parameters and their values which together define the service offered to a traffic' [9]. SLS is a separated technical part of 'a negotiated agreement between a customer and the service provider on levels of service characteristics and the associated set of metrics [5, 10]', which is the commonly adopted definition of a service level agreement (SLA). An example of SLS is represented by the ATM Traffic Contract [11] that is composed of traffic parameters and descriptors, along with a set of QoS parameters. SLS used in this work includes: the type of traffic (e.g. Premium VBR, Mission critical, best effort, etc.); traffic description and conformance testing (packet dimension, application peak and average rate, and, if requested, maximum burst and bucket size); performance guarantees (packet loss rate, packet transfer delay, and packet delay jitter).

2.2. Applications

Which applications need quality of service? The answer is simple: all the applications that require a specific level of assurance from the network. The answer does not give any idea about the amount of applications that need QoS. Some of them are listed below: basic services for information transfer both for backbones and access, assured database access to retrieve information, tele-medicine (transmission of clinical test, x-rays, electrocardiograms, magnetic resonance), tele-control (remote control of robots in hazardous environments, remote sensors, systems for tele-manipulation), bank and financial operations, purchase and delivery, tele-learning, telephony, videoconferences, applications for emergencies and security.

Having very different characteristics, each mentioned application deserves a specific grade of service, defined at the application layer. Several standardization bodies have tried to define service categories (also called QoS classes, to be intended at application layer).

ITU-T (in Recommendation Y-1541 [10]) suggests a definition of QoS classes (for the IP world) that is summarized in Table I.

The ETSI Project TIPHON [4] proposes an alternative QoS class definition, reported in Table II.

Concerning broadband-integrated services digital network (B-ISDN), ITU-T defined a set of service categories, reported below [12].

Derived from the general categories reported in Table III and consequently to the standardization of ATM as the technology for implementing B-ISDN, the ATM Forum defined five ATM service categories, reported in Table IV [13].

Each application mentioned at the beginning of this sub-section may be inserted in the classifications reported above but, besides the general definition reported, each application (or, in this context, more exactly, each user) needs a detailed specification in terms of traffic descriptors and intrinsic QoS parameters to allow a proper flow identification and service provision by the network.

Table I. ITU-T Y-1541 QoS classes.

QoS class	Characteristics
0	Real-time, jitter sensitive, highly interactive
1	Real-time, jitter sensitive, interactive
2	Transaction data, highly interactive
3	Transaction data, interactive
4	Low loss only (short transactions, bulk data, video streaming)
5	Traditional applications of default IP networks

Table II. TIPHON QoS classes (from [4]).

QoS class	Components	QoS characteristics
Real-time conversational (telephony, teleconference, videophony and videoconference)	Speech, audio, video, multimedia	Delay and delay variation sensitive, limited tolerance to loss and errors, constant and variable bit rate
Real-time streaming (e.g. audio and video broadcast, surveillance, graphics)	Audio, video, multimedia	Tolerant to delay, delay variation sensitive, limited tolerance to loss and errors, variable bit rate
Near real-time interactive (e.g. web browsing)	Data	Delay sensitive, tolerant to delay variation, error sensitive, variable bit rate
Non real-time background (e.g. e-mail and file transfer)	Data	Not delay and delay variation sensitive, error sensitive, best effort

Concerning the network viewpoint, QoS is the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements can be satisfied. QoS manages bandwidth according to application demands and network management settings.

The term QoS is used in many meanings, ranging from the user's perception of the service to a set of connection parameters necessary to achieve particular service quality. The QoS meaning changes, depending on the application field and on the scientific scope. Reference [5], starting from the terminology concerning QoS in IP networks, defines some reference points about the QoS issue. The authors, mentioning Reference [6], identify three types of QoS: intrinsic, perceived and assessed. Intrinsic QoS is directly provided by the network itself and may be described in terms of objective parameters as, for instance, loss and delay. Perceived QoS (P-QoS) is the quality perceived by the users; it heavily depends on the network performance but it is measured by the 'average opinion' of the users. Mean opinion score (MOS) methods are often used to perform the measure of the quality: users assign a rating as follows: 1—bad, 2—poor, 3—fair, 4—good, 5—excellent, to the application they are evaluating. The MOS is the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best) [7].

Even if there is a strict connection with the objective metrics provided by the network, the user does not necessarily perceive an objective performance increase (or decrease), in correspondence of an intrinsic QoS variation. Reference [8] reports a performance evaluation by using MOS values along with measures of objective metrics and shows that the correspondence is not always straightforward.

The last term reported concerns assessed QoS. It is referred to the will of a user to keep on using a specific service. It is related to P-QoS but also depends on the pricing mechanism, on the level of assistance of the provider and on other marketing and commercial aspects.

At the moment, most of QoS provision is offered in terms of intrinsic (objective parameters) QoS by using a service level specification (SLS) which is 'a set of parameters and their values which together define the service offered to a traffic' [9]. SLS is a separated technical part of 'a negotiated agreement between a customer and the service provider on levels of service characteristics and the associated set of metrics [5, 10]', which is the commonly adopted definition of a service level agreement (SLA). An example of SLS is represented by the ATM Traffic Contract [11] that is composed of traffic parameters and descriptors, along with a set of QoS parameters. SLS used in this work includes: the type of traffic (e.g. Premium VBR, Mission critical, best effort, etc.); traffic description and conformance testing (packet dimension, application peak and average rate, and, if requested, maximum burst and bucket size); performance guarantees (packet loss rate, packet transfer delay, and packet delay jitter).

2.2. Applications

Which applications need quality of service? The answer is simple: all the applications that require a specific level of assurance from the network. The answer does not give any idea about the amount of applications that need QoS. Some of them are listed below: basic services for information transfer both for backbones and access, assured database access to retrieve information, tele-medicine (transmission of clinical test, x-rays, electrocardiograms, magnetic resonance), tele-control (remote control of robots in hazardous environments, remote sensors,

Table III. B-ISDN services.

Categories	Applications (examples)
Conversational	Video conference, video surveillance, high speed document communication (file transfer, fax, images, sound)
Messaging	Email and video email, chat
Retrieval	High speed document retrieval (video, still images, sounds, transaction data)
Distribution *without* user-individual presentation control	Document and video distribution, Pay TV, Radio
Distribution *with* user-individual presentation control	Full-channel broadcast TV and videography

Table IV. ATM forum service categories.

ATM service categories	Representative applications	QoS characteristics
Constant bit rate (CBR)	Circuit emulation	Low cell delay variation, low loss
real time variable bit rate (rt-VBR)	Video on demand	Moderate cell delay variation, low loss
non real-time variable bit rate (nrt-VBR)	Packet traffic	Moderate loss
Available bit rate (ABR)	Adaptable rate sources	Low loss
Unspecified bit rate (UBR)	Best effort traffic	No requirements

3. QoS SOLUTIONS

The section is structured the following sub-sections: information transport technologies, QoS technologies, and QoS management functions.

3.1. *Information transport technologies*

A QoS-based service derives from reliable physical layers (including, in this case, layers 2 and 1) that can offer specific services to the upper layers. An example of it is represented by the protocol architecture proposed by ETSI [14] concerning the access point to a Broadband Satellite Multimedia (BSM) network portion. The architecture is reported in Figure 1. The physical layers (i.e. satellite physical, MAC and link control, strictly satellite dependent) are isolated from the rest by a satellite independent service access point (SI-SAP), which should offer specific QoS services to the upper layers. Using a satellite network is only an example and the architecture can be generalized to include different physical supports both wired and wireless.

The service definition offered by SI-SAP should be accurately defined to assure a QoS-based service.

Figure 2 reports a graphical model of the relation between lower (physical) layers and higher layers. The connections (or bundles of them) are forwarded down to a physical interface that transports the information along a channel. Some transport technologies, whose short description is reported in the following, also appear in Figure 2. Some technologies, as time division multiplexing (TDM) and dynamic transfer mode (DTM), may also be applied on end-to-end basis but this brief summary is aimed at concentrating on link-by-link application, where the techniques are used only to physically connect switches or routers.

Time division multiplexing (*TDM*): The channel resource is divided into portions of time called frames and each of them is structured in smaller portions called slots. Each slot (or group of slots) is assigned to a user who maintains the assignation for the all duration of the connection. The resource allocation is static but it should not be confused with the equivalent end-to-end transfer mode. In this latter case, TDM is used to install a circuit between the source and the destination that will be dropped at the end of the communication. In this context, as said above,

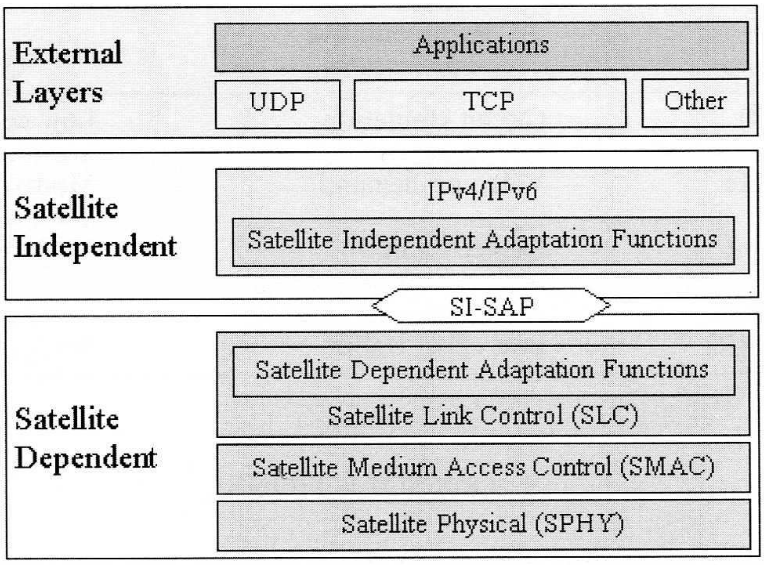

Figure 1. ETSI BSM protocol architecture.

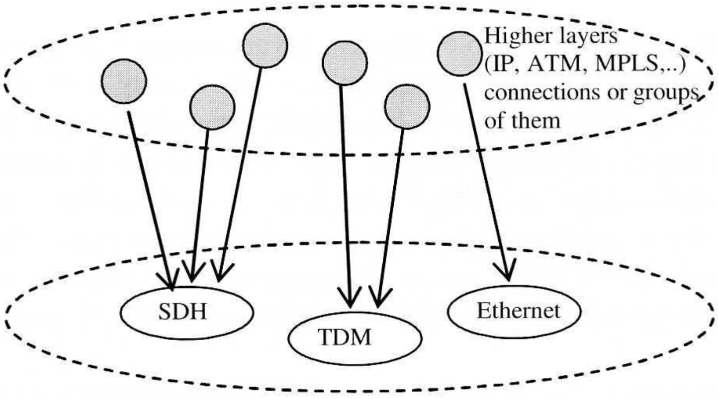

Figure 2. Higher layers over transport technology.

the focus is on link-by-link technology: using TDM means essentially having a time-slotted channel.

Dynamic transfer mode (*DTM*): The concept is the same as TDM: the channel is divided in frames (one each 125 ms) and each frame in slots (of 64 bits). The slots may be 'control slots', located at the beginning of the frame, and 'data slots'. The main difference with TDM is that time slots may be dynamically allocated even during a communication, either automatically or upon an explicit user request issued by an involved party over the signalling channel. In other words, while TDM time slots are allocated in the call set-up phase and maintained in number and identity till call clearing, DTM allows dynamic bandwidth tuning during communications.

Synchronous digital hierarchy/synchronous optical network (*SDH/SONET*): SDH is a hierarchy of synchronous time-multiplexed transmission. In North America, an equivalent technology is called SONET. SDH has been standardized in Reference [15]. The basic SDH transmission signal is called STM-1, whose bit rate is 155.520 Mbit/s; higher interfaces are obtained by scaling forward the basic rate n-times ($n = 4$, 16, 64), so obtaining STM-4 (622.080 Mbit/s), STM-16 (2488.320 Mbit/s) and STM-64 at 9953.280 Mbit/s. The SDH format is the standardized frame for ATM: the 53 bytes cells are transported over the STM payload.

Ethernet: The main interest concerning Ethernet (a technology designed for local area networks and characterized by a pure FIFO queue management) is the recent possibility of implementing classes of service (CoS) at the layer 2. CoS (being focused on traffic differentiation and related priorities) does not directly imply QoS (that requires more control functions), as should be clearer in the following, but it is very important. The extension for CoS of the basic Ethernet standard is defined in IEEE 802.1Q [16] and 802.1p [17]. In addition to the standard Ethernet frame, 802.1Q adds a 4 bytes TAG in the header that includes a priority field of 3 bits. The 3 bits of the priority guarantees a minimum traffic differentiation whose default priorities are defined in IEEE 802.1p.

ATM: ATM (and also MPLS) technology has not been defined to be a physical layer transport technology (see Figure 2) but, sometimes, it is seen as a layer 2 because it is much used in core networks [18] to transport other technologies (e.g. IP). Details about ATM (and MPLS) are provided in the following.

3.2. QoS technologies

In the following, a list, which should include the main technologies [19] available in the market to provide QoS, is briefly summarized, with particular attention to the capability of marking a specific traffic flow.

ATM: It is connection-based. Information is packetized in fixed length packets, called cells, of 53 bytes (5 for header and 48 for information). ATM identifies and assures a single user flow or an aggregate of flows by using two fields contained in each cell header and defined in the ITU-T Recommendation I.113 [20]: virtual channel (VC) and virtual path (VP).

It has been built to guarantee QoS and uses call admission and congestion control schemes properly designed for the aim. The scientific literature of these last 15 years is very rich concerning resource allocation and reservation schemes that use statistical multiplexing gain without penalizing users.

ATM strength points are listed in Table V, together with possible drawbacks. Many issues have been taken from Reference [21].

Table V. ATM strength points and drawbacks concerning QoS.

Strength points	Drawbacks
QoS: ATM has been designed for QoS, QoS managing capacity is standardized and checked also in multi-vendor commercial applications; ATM allows requesting quantitative QoS intrinsic parameters, also for each single user. *Diffusion*: Major carriers and Internet providers use ATM backbones. *Network management*: ATM has several tools, already industrially tested, to manage the network. Monitoring and fault isolation tools are standardized. *Connection-oriented*: ATM can guarantee dedicated pipes both between users and network nodes. Pipes can be used, priced and sold. *Adaptation options*: Users can choose among different adaptation layers. For example, AAL1, AAL2 and AAL5 can carry voice in dependence of the customer needs. *Signalling*: Signalling options are QoS-oriented. *Fixed and small dimension cells*: It is helpful both for statistical multiplexing in switching elements and for inserting a proper CRC in possible errored channels. Moreover, it is important in low bit rate trunks for voice applications because it is simpler to fill the cells avoiding delays needed in case of longer datagrams.	Complexity and overhead for AAL

3.2.1. QoS-IP. Native IP is connectionless and offers best-effort services. The service received by a user depends on the network load. Queuing managing within routers is, essentially, first in first out (FIFO). Concerning the tools to identify a traffic flow, IP needs to be differentiated between versions 4 and 6.

IPv4 offers two ways to mark traffic:

- A vector composed of the following fields: 'IP source address', 'IP destination address', 'Protocol', all contained within the IPv4 header; 'TCP/UDP source port' and 'TCP/UDP destination port', contained in the TCP/UDP header.
- ToS field (8 bits in the IPv4 header), whose first six bits define the DiffServ code point (DSCP) field.

Packets with the same identifier need to be treated coherently by each router.

IPv6 may use two fields directly in the IP header:

- Flow label field (20 bits).
- Traffic class field (8 bits), functionally equivalent to IPv4 ToS field and containing the DSCP field.

The difference between IPv4 and IPv6 is that IPv6 can mark a flow through the flow label, as well as an aggregate of flows (similarly to VC/VP in ATM), while IPv4 can identify either a limited number of aggregates through the DSCP field, up to a theoretical maximum of 2^8, or a specific flow (but not an aggregate of them) through the vector 'IP source address, IP destination address, Protocol, TCP/UDP source port, TCP/UDP destination port'.

Flow identification is only a starting point for QoS guarantee. Two paradigms have been proposed to match the market QoS request: integrated services and differentiated services.

The integrated services [19, 22]: It is based on the concept of flow defined as a packet stream that requires a specified QoS level and it is identified by the vector 'IP source address, IP destination address, Protocol, TCP/UDP source port, TCP/UDP destination port'. Actually, the same scheme may be used in IPv6 by using the flow label field but there is no standardization about it. QoS can be reached by an appropriate tuning of different blocks: resource reservation, admission control, packet scheduling, and buffer management. A status concerning the different incoming flows must be maintained in the routers. It is very different from the best-effort approach provided by the Internet. Information about flows must be periodically updated and a specific resource reservation signalling system (RSVP [22, 23]) is used for this aim. The IntServ approach defines two service classes: guaranteed service (GS) [24] and controlled-load service (CLS) [25]. GS is a service characterized by a perfectly reliable upper bound on end-to-end packet delay; CLS has a much looser specification, which is supposed to offer a service that is comparable to best-effort service in a 'lightly loaded' network. The drawback of this approach is scalability. IntServ needs to detect each single flow and both packet scheduling and buffer management act on per-flow basis. Although IntServ is ATM-like, it has not the tools provided by ATM because it is derived from an intrinsically best-effort and connectionless technology. The cost and the complexity of the control system increase with the number of flows. Moreover, the signalling RSVP starts a connection on request of the source-host; it does not have any release message, so a periodic refresh message is needed to confirm the resource (bandwidth) request; a connection is dropped only if the refresh message has not been received for some time. The drawbacks are:

- even if the IP flow is terminated, the resources are not immediately released;
- refreshing signalling is bandwidth consuming.

The Differentiated Services (*DS or Diffserv*) [26, 27]: have been proposed to cope with the scalability problem faced by integrated services. The solution uses the DSCP field of the IP packet header by using the first 6 bits either of IPv4 ToS or of IPv6 Traffic Class. The 6 bits field DSCP specifies the forwarding behaviour that the packet has to receive within the DiffServ domain of each operator. The behaviour is called per hop behaviour (PHB) and it is defined locally; i.e. it is not an end-to-end specification (as for RSVP) but it is strictly related to a specific domain. The same DSCP may have two different meanings in two different domains. Negotiations between all adjacent domains are needed to assure a correct end-to-end forwarding behaviour.

The DiffServ approach does not distinguish each user flow throughout the network. The traffic is classified and aggregated in different traffic classes, each of them individuated by a label provided by setting bits in the DSCP field. The identification is performed at the network edges. Within the network core, packets are managed according to the behaviour associated to the specific identification label. Figure 3 shows an example of aggregation: after the DiffServ router the flows are aggregated depending on their label (light or dark grey in Figure 3) and information about each single user (A, B, C or D, in Figure 3) is completely lost.

The class selector PHB offers three forwarding priorities:

- Expedited forwarding (EF) characterized by a minimum configurable service rate, independent of the other aggregates within the router, and oriented to low delay and low loss services.
- Assured forwarding (AF) group recommended in Reference [28] for 4 independent classes (AF1, AF2, AF3, AF4) although a DS domain can provide a different number of AF classes. Within each AF class, traffic is differentiated into 3 'drop precedence' categories. The packets marked with the highest drop precedence are dropped with lower probability than those characterized by the lowest drop precedence.
- Best effort (BE), which does not provide any performance guarantee and does not define any QoS level.

It is clear that DiffServ offers a limited set of priorities and may be strongly limited by the need of traffic aggregation. 'Therefore, while DiffServ architecture solves the scalability problems of QoS provisioning, it fails to be the solution for end-to-end provisioning' [29] because no quantitative values can be provided but only a qualitative QoS.

A possible improvement of the services provided by DiffServ architecture is represented by a mixed IntServ/DiffServ approach. Figure 4 shows an example of it.

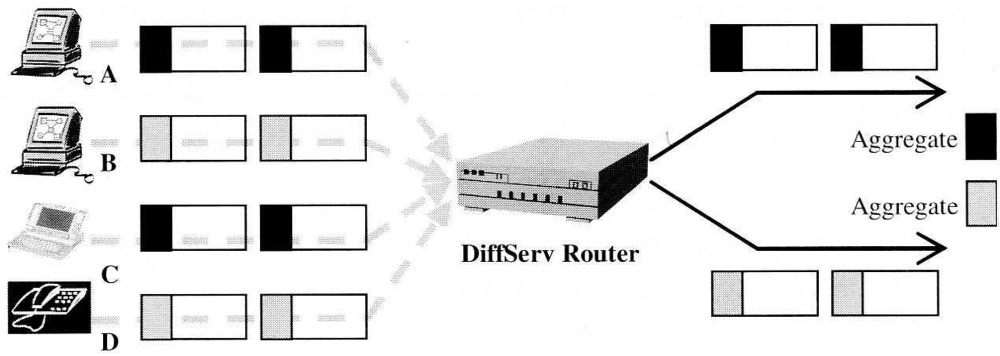

Figure 3. DiffServ aggregation behaviour.

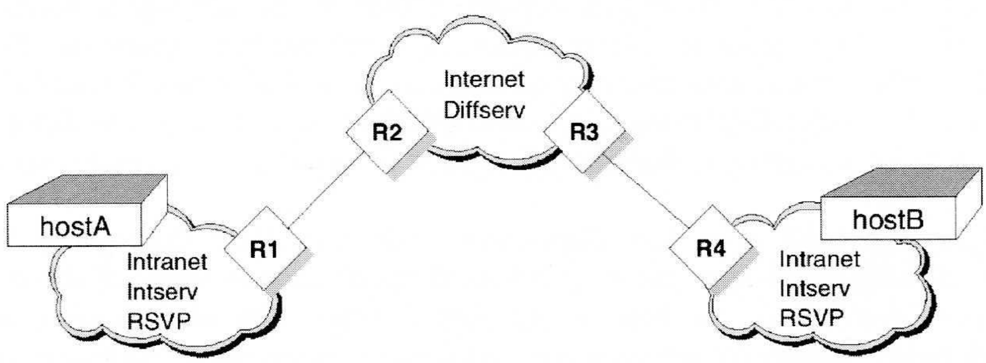

Figure 4. IntServ/DiffServ architecture. Taken from 'TCP/IP Tutorial and Technical Overview' http://www.redbooks.ibm.com.

The IntServ approach may be reserved to small private Intranets where there are no scalability problems and the DiffServ approach may be applied to the interconnection backbone. Nevertheless, it means that no quantitative QoS can be provided over the backbone. It is acceptable having the Internet best-effort service in mind but it cannot be applied to a commercial network offering QoS-guaranteed services. The indication perceived in international conferences and magazines is that the Internet Community itself recommends IntServ for small private networks for the access segment.

A comparison between IntServ and DiffServ is reported in Table VI (derived from [30]). Advantages and drawbacks of QoS-IP solutions are summarized in Table VII.

The envisaged problems may be mitigated by using IPv6, which can take the best of IP technology. Concerning flow identification, IPv6 offers the same capabilities of ATM but, to use all the features, it is necessary to import most of the control functions (part of the ATM

Table VI. IntServ versus DiffServ.

Feature	IntServ	DiffServ
QoS assurance	Per flow	Per aggregate
QoS assurance range	End-to-end (application-to-application)	DiffServ domain (edge-to-edge)
Resource reservation	Controlled by application	Configured at edge nodes based on service level agreement
Resource management	Distributed	Centralized within DiffServ domain
Signalling	Dedicated protocol (RSVP)—Refreshing needed	Based on DiffServ code point (DSCP) carried in IP packet header
Scalability	Limited by number of flows	Limited by number of classes of service
QoS classes	GS, CLS, best-effort	EF, AF, best-effort
Complexity	High	Low
Availability	Yes	Yes

Table VII. QoS-IP Solutions strength points and drawbacks.

Strength points	Drawbacks
Diffusion: IP is widespread: terminal user, LAN, network access, backbones, due to the Internet development.	*QoS*: IntServ (RSVP) and DiffServ can manage QoS within the IP world but each of them has limitations referring to the possible user performance requirements.
Simple web integration: Web interfaces are much used, both over the Internet and over private networks.	*IntServ*: It is considered too complex and not scalable to be used within backbones and multi-user networks. RSVP uses a refreshing mechanism that brings resource management inefficiencies.
Header compression: Header compression is very important to reduce the transmission overhead (in particular the stack RTP/UDP/IP for voice).	*DiffServ*: It can guarantee only qualitative performance requirements by using a limited number of priorities. It is scalable but it cannot provide end-to-end solutions.

Label Value – 20 bits	Exp – 3 bits	S – 1 bit	TTL – 8 bits

Figure 5. MPLS label format.

approach) within IP. In practice, it would be interesting to keep IPv6 format to simplify interworking but to use control mechanisms (e.g. CAC, filtering and signalling) for each flow, as done in ATM.

3.2.2. MPLS. MPLS derives from the convergence between the IP world, which involves open standards and simplicity, and the ATM world, which is completely oriented to the QoS and uses traffic control techniques, as well as QoS aware routing. MPLS would like to 'summarize' the best of both technologies. A label of 20 bits identifies the traffic. The label switching mechanism should be associated with control modules that allow guaranteeing a fixed level of quality. In principle, MPLS offers the same capabilities to mark flow as in ATM and IPv6. Figure 5 shows the MPLS label format, which is composed of 32 bits consisting of the following elements: the already mentioned label value; the field Exp, dedicated to future extensions; the bit *S*, aimed at indicating the presence of more labels; the time to live.

The entrance and the exit of a MPLS domain are governed by label edge routers (LERs) that generate and apply the labels when information enters the domain and remove them at the exit. The basic technology inside the domain is represented by label-switching routers (LSRs), which switch the traffic in dependence of a specific path, called label switched path (LSP), associated with the MPLS label. The label switched path defines the sequence of nodes where the traffic of a connection flows within the MPLS domain. The definition is performed through QoS-based traffic engineering techniques [31] aimed, for instance, at minimizing the number of hops, meeting bandwidth requirements, supporting precise performance requirements, bypassing potential points of congestion, directing traffic away from the default path selected or simply forcing traffic across certain links or nodes in the network. MPLS is still under analysis and standardization but it has a great potential to unify and implement QoS-based interworking among network portions implementing different technologies.

LER and LSR are MPLS-aware routers. The introduction of MPLS is aimed at creating a QoS-guaranteed network suited also for IP packet format. In other words, MPLS-enabled technology should implement the ATM QoS in a smooth way for other technologies (e.g. IP packets, which should only add a small label). Some literature embeds ATM in the MPLS world. In this context, the VCI and VPI fields of the ATM header represent the MPLS Label. A similar observation may be done for the DLCI field of the frame relay environment.

3.3. QoS management functions

'*There is a common misconception that purchasing an oversupply of bandwidth will solve all service-quality challenges. Throwing bandwidth at the problem is sometimes perceived as a simpler solution than QoS management*' [32]. This approach ignores not only bandwidth optimization but also the possible future trends and requirements of new services. It is not a solution, QoS management is strongly necessary and QoS management functions are aimed at offering the necessary tools to get a certain level of quality. One of the objectives of this work is to show the effect of the various functions and, independently of the technology used, the importance to have specific service functions to satisfy precise performance requirements. The effect of most functions will be evaluated in the results. Meaningful QoS management functions are reported in the following.

Packet marking: The identification of packets so that they may receive a different treatment within the network is topical to guarantee QoS, ranging from a minimum priority-based service to quality assurance for a specific user. Single QoS-oriented technologies, presented before, show different methods to classify packets, as flow label and traffic class in IPv6, ToS and vector 'IP source address, IP destination address, Protocol, TCP/UDP source port, TCP/UDP destination port' in IPv4, VPI/VCI in ATM, labels in MPLS.

CAC: CAC decides whether a new connection request may be accepted or not. It is a powerful tool to guarantee quality because it allows limiting the load entering the network and verifying if enough resources are available to satisfy the requested performance requirements of a new call without penalizing the connections already in progress. A connection provides traffic descriptors (e.g. the set of traffic parameters of an ATM source) and QoS requirements. The network evaluates if there are sufficient resources in terms of bandwidth and buffer to match the request and decides either to accept or reject the connection.

Traffic control (*shaping*): It is very important to guarantee performance requirements. Shaping policies limit flows to their committed rates (e.g. the flows need to be conformant with their traffic descriptors). If flows (or also single connections) exceed their bandwidth consumption specifications, the network, which has dimensioned resources in strict dependence with the declarations, cannot guarantee any specified QoS requirement. An example of traffic shaping action is shown in Figure 6, where two, out of four, connections are not conforming with their committed rates set to 16 kbit/s. The two non-conformant flows that generate 64 kbit/s are cut down to the committed rate.

Scheduling: Packet scheduling specifies the service policy at a queue within a node (for example, an IP router, an ATM switch). In practice, scheduling decides the order to be used to pick the packets out of the queue and to transmit them over the channel. It is an important issue, concerning QoS. It has a strong impact on different QoS parameters as delay, jitter and loss. The main problem arises from the impossibility of assigning the committed bandwidth to a specific flow at each time instant. In general, bandwidth is allocated in average and most scheduling policies may guarantee the respect of this condition. Unfortunately, even respecting the average bandwidth may be not sufficient to assure the guarantee of other QoS parameters,

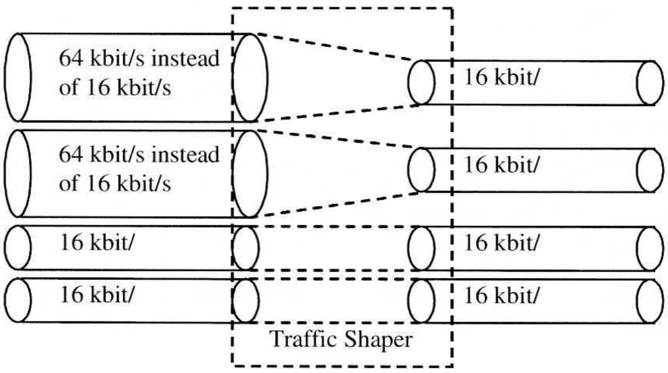

Figure 6. Action of traffic shaping.

as, for example, the delay. The only scheduler that allocates the whole outgoing bandwidth to all flows (or users) in progress in strict proportion with the bandwidth allocated, for each time instant, is the generalized processor sharing (GPS). It applies an ideal policy that supposes to serve all the users in the same time and to split infinitesimally the bandwidth, which is allocated to the users exactly as requested. GPS is often used in simulations. Real systems have to use alternative schemes with the aim of performing as closest as possible to GPS.

A clear and complete revision of the most interesting schedulers is reported in [19]. It includes weighted fair queuing (WFQ), virtual clock (VC), self-clocked fair queuing (SCFQ), deficit round robin, WF^2Q, and WF^2Q+, as well as a comparison of the mentioned schemes concerning latency, fairness and time complexity.

Resource reservation: An accurate resource reservation to guarantee that traffic flows receive the correct service is strictly needed. Resource means, in this context, bandwidth and buffer. Its allocation is strictly related with CAC mechanisms and the acceptance/rejection of a new connection is performed subject to a check (that may be statistical) about the availability of network resources in consequence of specific requirements. After that, if enough resources are available, they are reserved. Resource reservation may act also on a larger time scale concerning network planning and link dimensioning.

Flow control: In some cases, the bit rate entering the network may be ruled according to a congestion notification (ECN–explicit congestion notification). Some protocols (e.g. TCP) consider packet loss as a congestion indication. Generally flow control is implemented end-to-end at the transport layer (but some mechanism are implemented at the application layer); even if it may help avoiding network saturation, it cannot guarantee, used alone, a specific QoS requirement. It may be used mostly to improve the performance of best-effort traffic.

QoS routing: Packet routing decisions are often taken with little or no awareness of network status and resource availability. This is not compatible with QoS provision. QoS routing needs to identify end-to-end paths where there are enough available resources to guarantee performance requirements in terms of metrics as loss, delay, call blocking, number of hops, reliability as well as bandwidth optimization.

The intervention time scale of the mentioned functions is different. Figure 7 shows a possible time mapping for them.

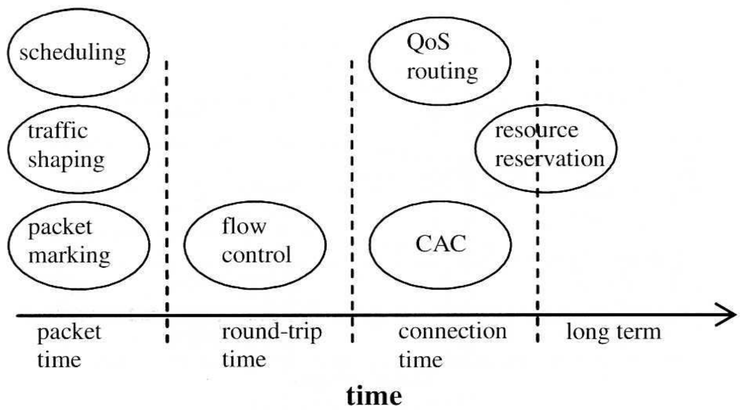

Figure 7. QoS management functions versus time.

4. PERFORMANCE EVALUATION

The aim of this section is to show the effect of the presence of some specific management functions. More than privileging a technology versus another one, it is important to know which functions a solution may use, which performance limits the functions have and what the user expectation can be. The following issues will be investigated by showing some performance results: traffic aggregation, CAC, and traffic shaping. The real limitations are not imposed only by the technology, whose features may be changed and extended, but by the application of control functions that can guarantee a degree of service that supports the application needs in terms of QoS. The investigation has been carried out without any reference to current fashion trends (as, for instance, ATM some years ago and IP now). It should allow also to individuate what can be done to improve QoS provision in the future.

4.1. Traffic aggregation

We now take a closer look at the effect of traffic aggregation on the network performance.

Figure 8 shows two traffic flows that, in one case, may be distinguished by using the packet label, while, in the other case (Figure 9), they are totally aggregated and no distinction can be done even if the original performance requirements are different.

If traffic requiring different performance is joined in one flow, it is necessary to investigate the additional bandwidth required to keep the same performance level. An example may be

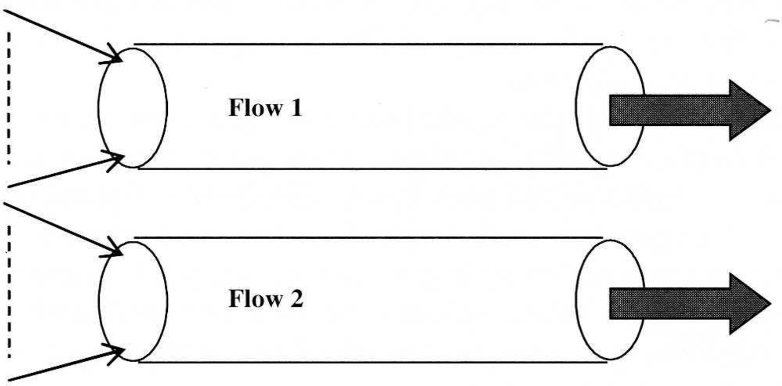

Figure 8. Distinguished traffic flows.

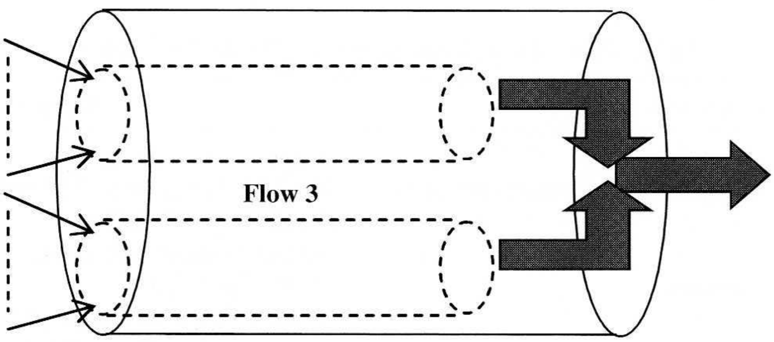

Figure 9. Aggregate flows.

represented by the environments that use a limited number of classes with respect to the technologies where a very large number of traffic classes are available. In practice, due to the limited number of traffic classes, non-homogeneous traffic flows (i.e. flows requiring diverse QoS) need to be aggregated and conveyed together.

In this perspective, the following simulation results regard the effect on performance of aggregation for traffic requiring different QoS constraints, in terms of packet loss, packet delay and delay jitter. They allow to have also a measure of the bandwidth necessary to guarantee the required performance.

If traffic needs to be aggregated, the choice of the bandwidth to be assigned to guarantee the QoS is topical. The relevant metric, in this case, is the measure of the addition (or reduction) of bandwidth necessary to keep the same level of service when traffic flows are aggregated in comparison with a complete separation. The parameter used in this work is the gain, defined as the percentage difference between the overall bandwidth necessary to satisfy the requirements if the flows are kept separated (Figure 8) and the bandwidth needed by the complete aggregation (Figure 9). For example, if 'Flow 1' needs 1.0 Mbps to satisfy the requirements and 'Flow 2' 2.0 Mbps, when kept separated, if the aggregation of the two flows requires 4.0 Mbps to offer the same performance, the defined gain is

$$100 \cdot \frac{(1+2)-4}{(1+2)} = -33.33\%$$

It means that, in this example, aggregation of non-homogeneous traffic is not convenient and that 33% of more bandwidth is necessary to guarantee the fixed requirements. A detailed investigation is reported in the following. Many studies confirm the efficiency of aggregating homogeneous traffic but the performance of non-homogeneous (from the QoS requirement viewpoint) trunks is still an open issue.

Concerning the tests reported, the application level generates on-off sources whose traffic descriptors are: *Peak bandwidth* (Mbps or kbps), *Mean burst duration* (s), *Mean silence duration* (s). The burst and silence durations are both Pareto distributed. The packet length is fixed to 53 bytes. A buffer of 5.3 kbytes for each single flow has been chosen. When the flows are aggregated, the buffer length is scaled up (e.g. in case of two flows, as used in this work, is set to 10.6 kbytes). The buffer and the server, whose capacity is the bandwidth assigned to each flow (or aggregate), may represent both a single network node and the overall end-to-end path, where the QoS requirements are satisfied.

The first part of the tests have been performed with the traffic flows appearing in Table VIII and supposing that the two flows need to be aggregated because there are not enough classes to be assigned. They differ only for the Packet loss rate parameter.

Table VIII. Two flows based on PLR: 10^{-4}–10^{-2}.

Service level specification	Range
Premium VBR	VBR
Traffic description and conformance testing	Packet dimension: 424 bit
	Peak rate: 1.0 Mbps
	Average rate: 500 kbps
Performance guarantees	PLR: 10^{-4}–10^{-2}
	Packet transfer delay (PTD): not specified
	Packet delay jitter (PDJ): not specified

The result heavily depends on the composition of the aggregate trunk.
Figure 10 contains the aforementioned bandwidth gain by varying:

(1) the number of connections within the aggregate trunk;
(2) the percentage of connections belonging to the two flows requiring, respectively, a PLR of 10^{-2} and 10^{-4}. For instance, the percentage 33% and 66% stand for 1/3 and 2/3, respectively, so to get 100% of traffic and so on;

The performance value (i.e. the PLR) for the aggregate trunk is set to 10^{-4} in order to assure that the overall trunk is guaranteed. It is worth noting that packets of the two flows are no longer distinguished within the trunk.

Non-homogeneous aggregation is often convenient (see the '20% 10^{-2}; 80% 10^{-4}' and '33% 10^{-2}; 66% 10^{-4}' cases) but, if traffic is unbalanced towards the less restrictive traffic (the '66% 10^{-2}; 33% 10^{-4}' and '80% 10^{-2}; 20% 10^{-4}' cases), a bandwidth portion is wasted to guarantee the specified performance. Results reported above give an operative solution to operate bandwidth dimensioning.

The trend is even clearer if the QoS differentiation stands in the PTD constraint (Table IX). The performance constraints fix the PLR to 10^{-2} and differentiate traffic flows by imposing a delay transfer of 10 and 50 ms.

Even in this case, the more restrictive constraint is chosen for the overall trunk (i.e. 10 ms). Figure 11 contains three cases: 20% of traffic requires 50 ms of packet transfer delay and 80% requires 10 ms; traffic requiring 10 and 50 ms equally shares the buffer; traffic whose constraint is 50 ms is dominant representing 80% of the load. The percentage of additional bandwidth required to guarantee the fixed performance for the aggregate flow is relevant in the last two cases: also when most demanding (10 ms) traffic is 50% of the overall load, more than 10% of additional bandwidth is required when there are 100 connections, more than 15% with 150 connections and 20% with 200 connections.

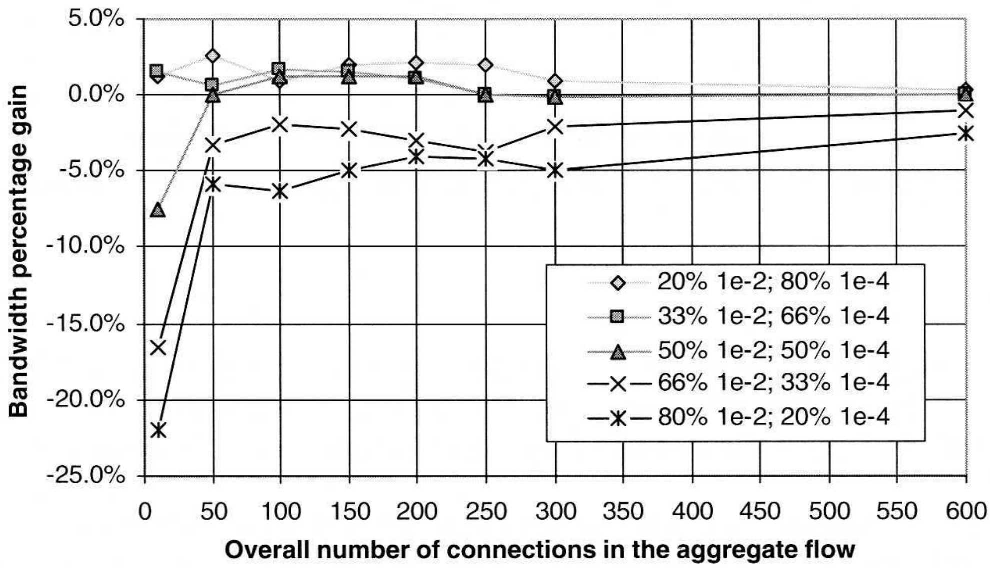

Figure 10. Bandwidth percentage gain in traffic aggregation: the PLR case.

Table IX. PLR 10^{-2} and PTD 50 and 10 ms.

Service level specification	Range
Premium VBR	Variable bit rate (VBR)
Traffic description and conformance testing	Packet dimension: 424 bit Peak rate: 16.0 kbps Average rate: 8.0 kbps
Performance guarantees	PLR: 10^{-2} PTD: 50–10 ms PDJ: not specified

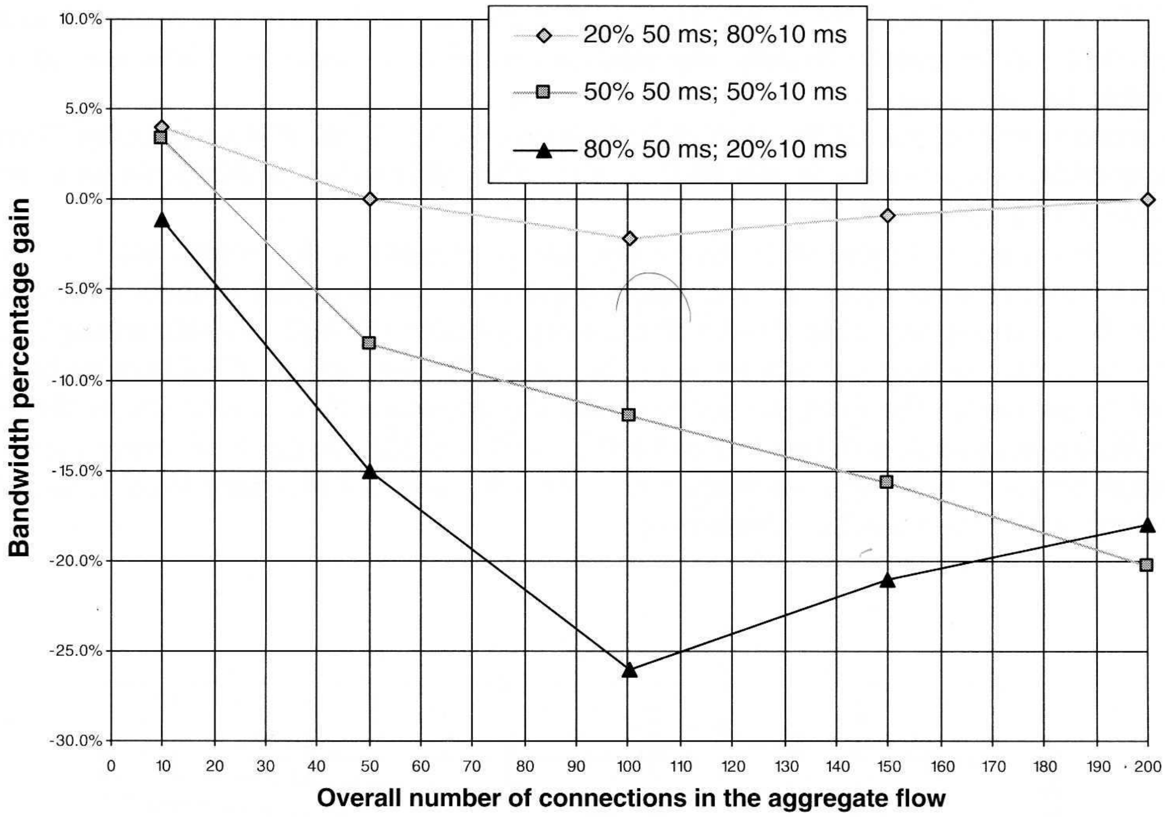

Figure 11. Bandwidth percentage gain in traffic aggregation: the PTD case.

Similar comments may be made if a further constraint on jitter is added (Table X). A packet loss of 10^{-2}, a transfer delay of 50 ms and a delay jitter of 9 and 5 ms are imposed. Figure 12 reports the results by changing the percentage of 9 and 5 ms traffic in the overall load.

The behaviour is similar to the transfer delay case but it is worth noting that, in this case, the quality differentiation (5 and 9 ms) is really minimum. It means that jitter is a very sensible parameter for bandwidth reservation and a very critical issue for aggregation.

Table X. PLR: 10^{-2}, PTD 50 ms, PDJ 9 and 5 ms.

Service level specification	Range
Premium VBR	VBR
Traffic description and conformance testing	Packet dimension: 424 bit Peak rate: 16.0 kbps Average rate: 8.0 kbps
Performance guarantees	PLR: 10^{-2} PTD: 50 ms PDJ: 9–5 ms

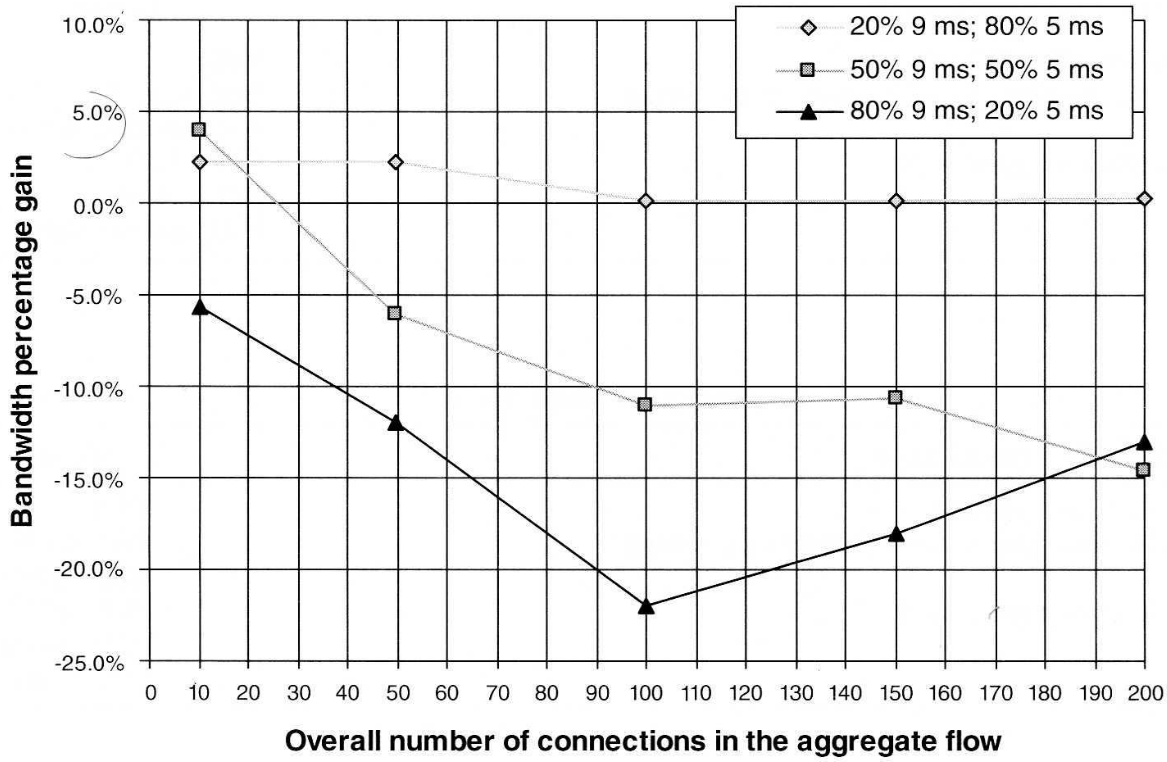

Figure 12. Bandwidth percentage gain in traffic aggregation: the PDJ case.

2. *CAC*

ɔ point out the importance of this function, we have tested the effect of implementing (or not) CAC. The following three data traffics are used. Traffic model and buffer dimension are the .me as used in the traffic aggregation analysis. Data traffic 1 imposes a constraint only on PLR ˋ 10^{-4} (Table XI). Data traffic 2, other than packet loss (10^{-2}), puts a constraint also on ansfer delay, set to 10 ms (Table XII). Data traffic 3 constraints the delay jitter below 5 ms, as ell as loss and delay, set, respectively, to 10^{-2} and 10 ms (Table XIII).

The results show what happens to the performance guarantees if no CAC is implemented. The ference '0%' in Figures 13–19 identifies the situation where traffic entering the network is ocked by CAC and, consequently, the bandwidth has been properly dimensioned and the

Table XI. Data traffic 1.

Service level specification	Range
Mission critical data	VBR
Traffic description and conformance testing	Peak rate: 1 Mbit/s Average rate: 500 kbit/s
Performance guarantees	PLR: 10^{-4} PTD: not specified PDJ: not specified

Table XII. Data traffic 2.

Service level specification	Range
Mission critical data	VBR
Traffic description and conformance testing	Peak rate: 16 kbit/s Average rate: 8 kbit/s
Performance guarantees	PLR: 10^{-2} PTD: 10 ms PDJ: not specified

Table XIII. Data traffic 3.

Service level specification	Range
Mission critical data	VBR
Traffic description and conformance testing	PR: 16 kbit/s AR: 8 kbit/s
Performance guarantees	PLR: 10^{-2} PTD: 10 ms PDJ: 5 ms

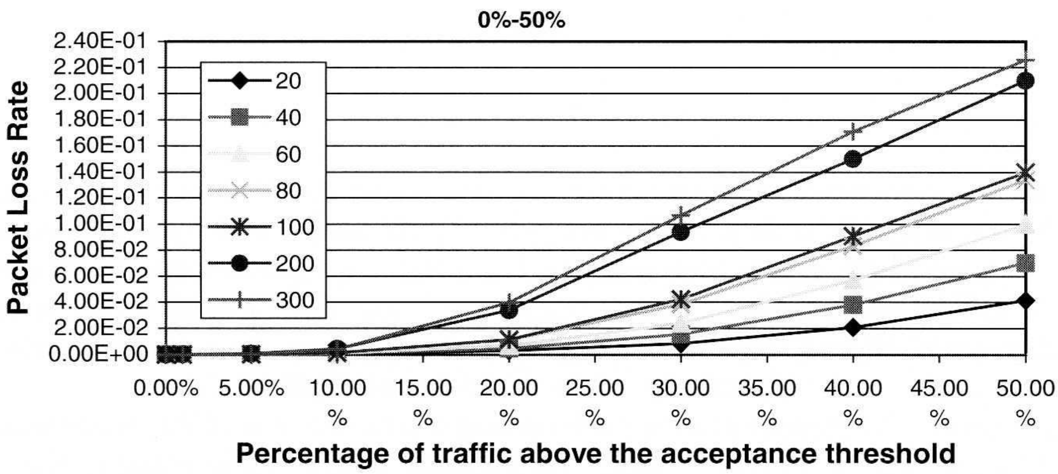

Figure 13. Data traffic 1, 0–50%.

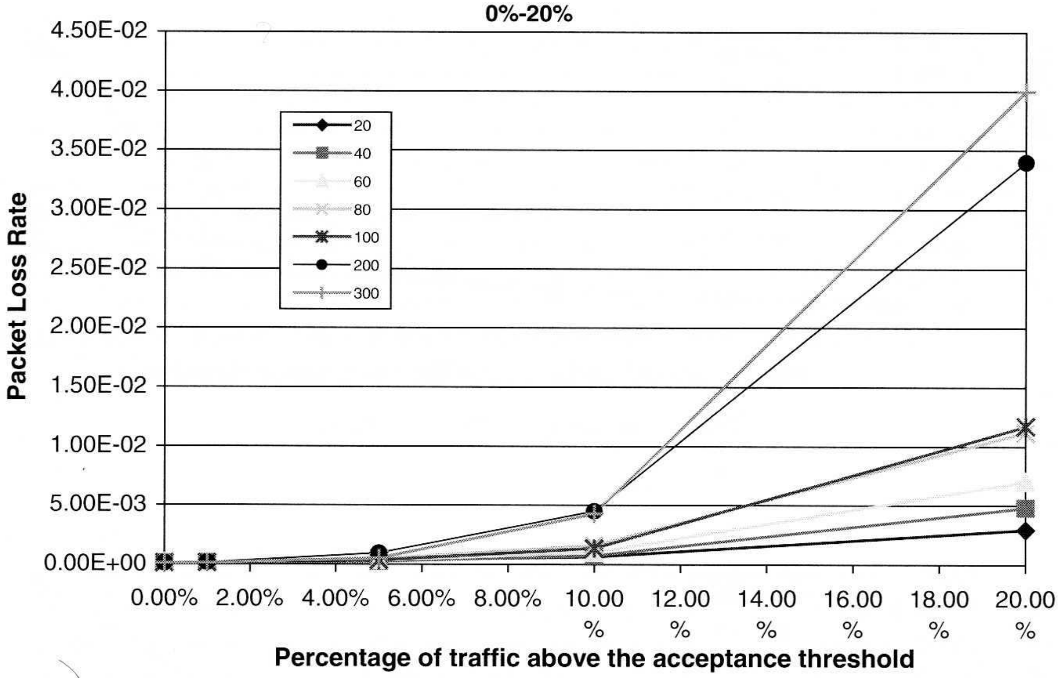

Figure 14. Data traffic 1, 0–20%.

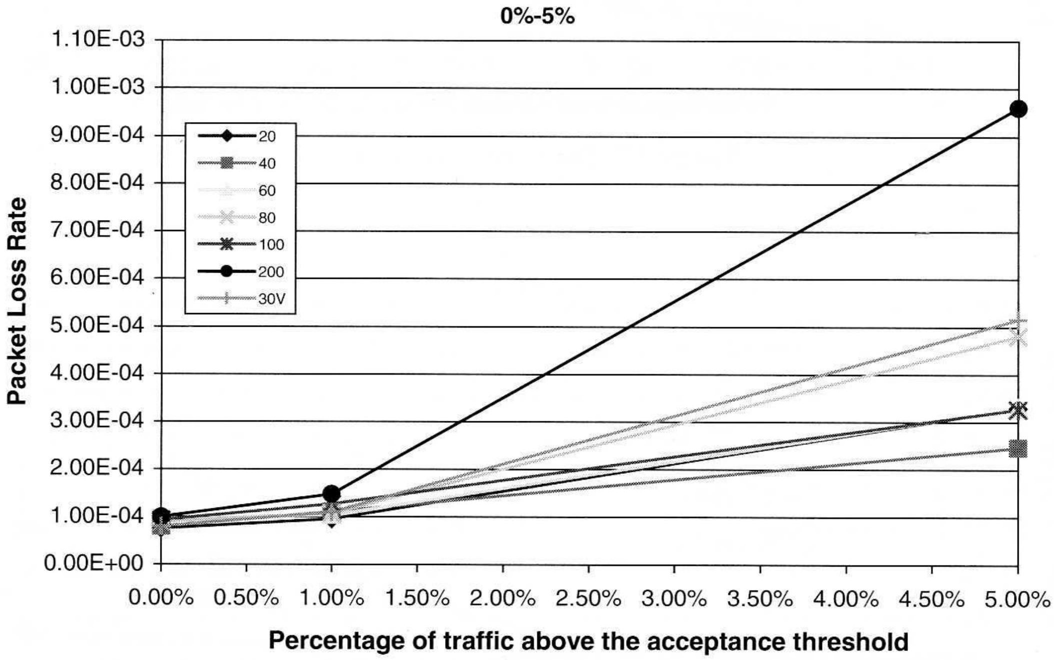

Figure 15. Data traffic 1, 0–5%.

ality guaranteed. The x-value in the figures identifies the percentage of traffic above the ceptance threshold (i.e. the percentage of connections that would have been dropped in case CAC) up to 50% of unbalance. For all the tests the number of connections range from to 300.

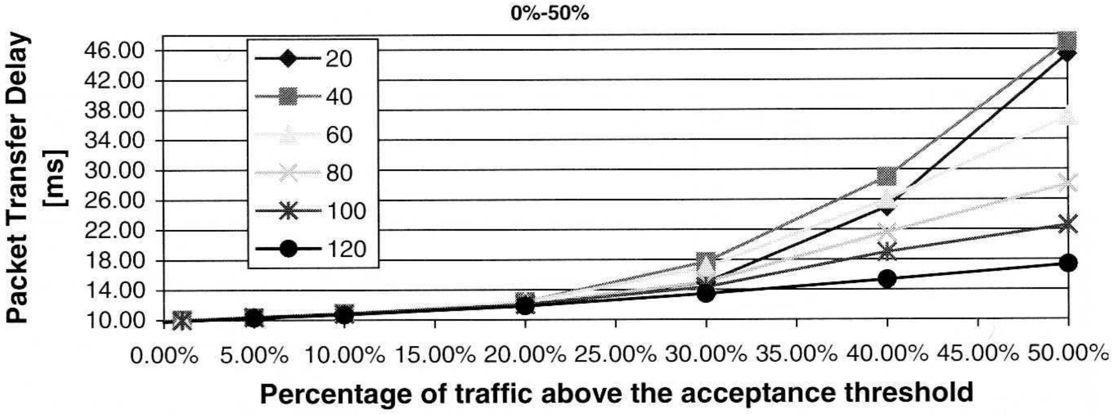

Figure 16. Data traffic 2, 0–50%.

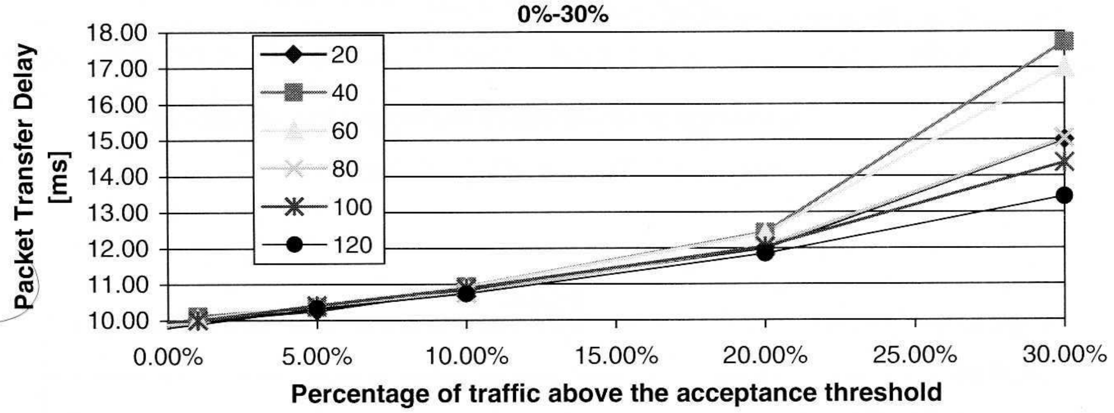

Figure 17. Data traffic 2, 0–30%.

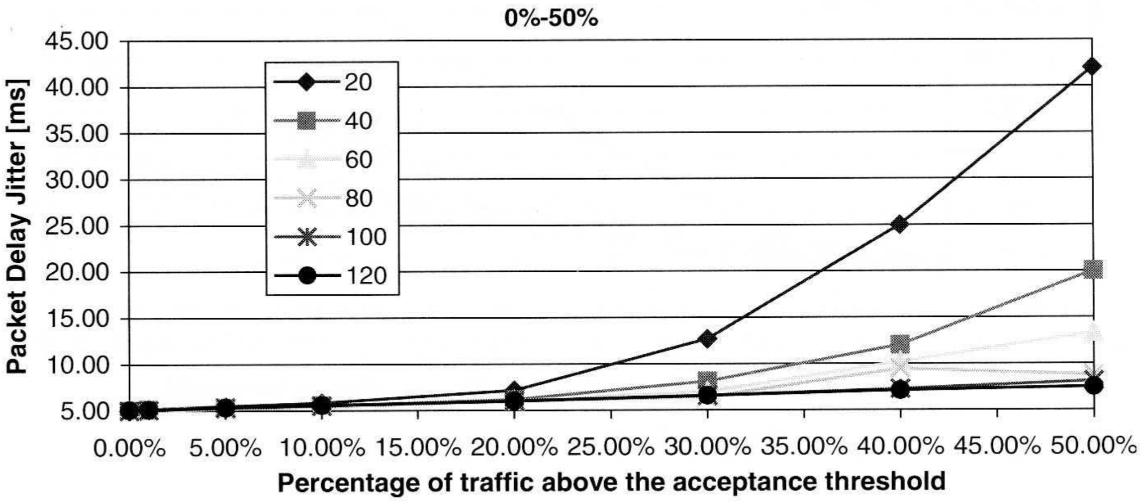

Figure 18. Data traffic 3, 0–50%.

Figure 13 reports the values of the measured PLR for 'Data traffic 1', as well as Figures 14 and 15 that contain the same values but concentrate on a limited portion of the results: 0–50% for Figure 13, 0–20% for Figure 14 and 0–5% for Figure 15.

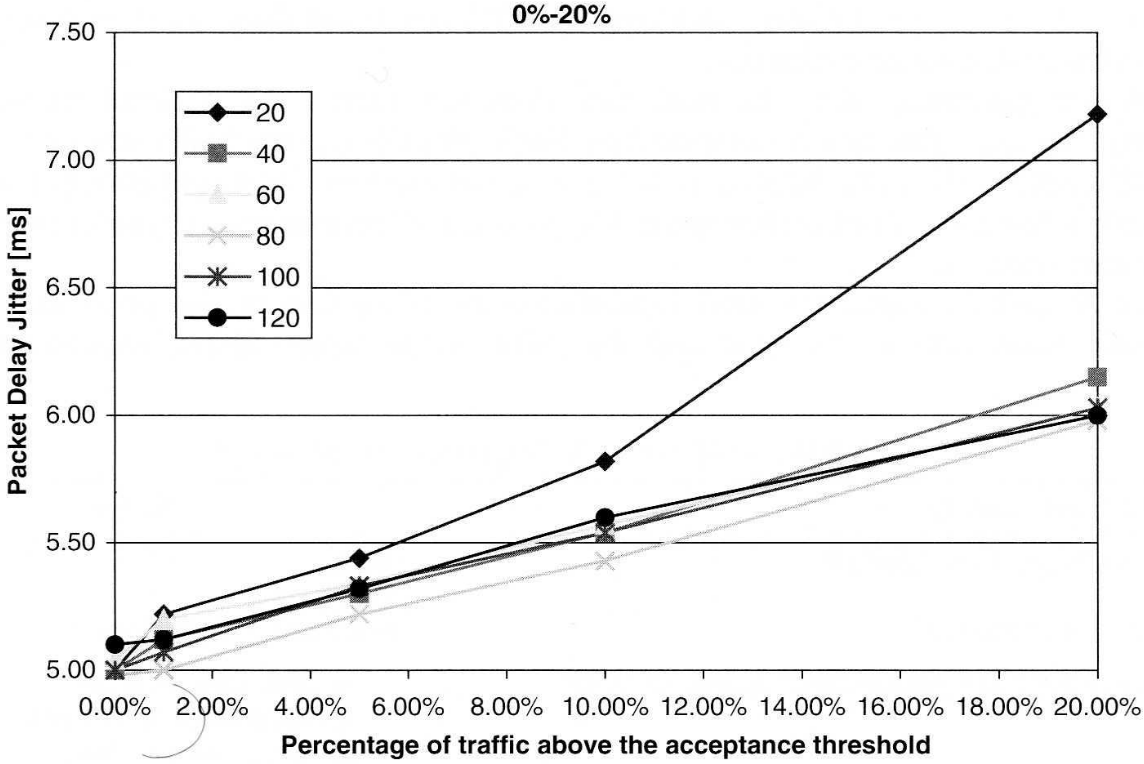

Figure 19. Data traffic 3, 0–20%.

Figures 16 and 17 show the tests for 'Data traffic 2', focusing, respectively, on 0–50% and 0–30% of non-controlled traffic and reporting the measured PTD.

Figure 18 concerns 'Data traffic 3' and shows the PDJ, as well as Figure 19 that contains the 'magnifying lens' for the 0–20 % range.

Values reported are outstanding: small variations of the number of accepted calls with respect to the reference value (that guarantees the required QoS) have a heavy effect on the traffic flow performance. Values reported allows measuring the performance decrease. The conclusion is that no performance guarantee is possible if no CAC is implemented.

4.3. *Shaping*

We emphasize the importance of traffic shaping to keep the traffic conformant with the traffic contract. The reference framework is shown in Figure 6.

We consider a *Voice over IP* (VoIP) traffic flow as in Table XIV. Taking [33] as a reference, each VoIP source is modelled as an exponentially modulated on-off process. The mean on and off times, as in ITU-T recommendation P.59 [34], are 1.008 and 1.587 s, respectively.

All VoIP flows are modelled as 16 kbps voice over RTP/UDP/IP. Without header compression, RTP + UDP + IP headers impose 40 bytes of overhead to each VoIP packet. The 40-byte overhead can be reduced to 4 bytes by using header compression. We take an IP packet size of 80 bytes as a reference. The required end-to-end performance objectives of a VoIP flow (shown in Table XIV) are: a PLR below 2% of and a PTD below 150 ms. We fixed the PTD constraint for each hop to 50 ms, in order to respect the end-to-end PTD along a route supposed of no more that 3 nodes. Reducing the PTD constraint does not change significantly the following performance evaluation.

The shaper guarantees that the peak rate does not exceed the declared threshold (i.e. 16.0 kbps). The aggregate flow is composed by 100 VoIP calls and the buffer size set to 8 kbytes (100 VoIP packets). If traffic filtering is not active, not expected 'non-conformant traffic' may significantly affect the overall performance. The measure of performance degradation is the aim of the results reported.

Figures 20 and 21 depict the QoS degradation as a function of the percentage of non-conformant traffic, taking the PLR and the PTD as the target values, respectively. Three

Table XIV. VoIP *service level specification* (SLS) [34].

Service level specification	Range
Premium VBR for *Voice over IP*	VBR
Scope	End-to-end
Connection identification	Sequence of identifiers
Traffic description and conformance testing of VoIP	• Peak rate: 16 kbps • Mean rate: 14.87 kbps • Packet size: 80 bytes • Maximum burst size: 1.008 s • Bucket size for peak rate: 16 kbps
Performance guarantees	PLR: 2% PTD: 50 ms PDJ: not applicable

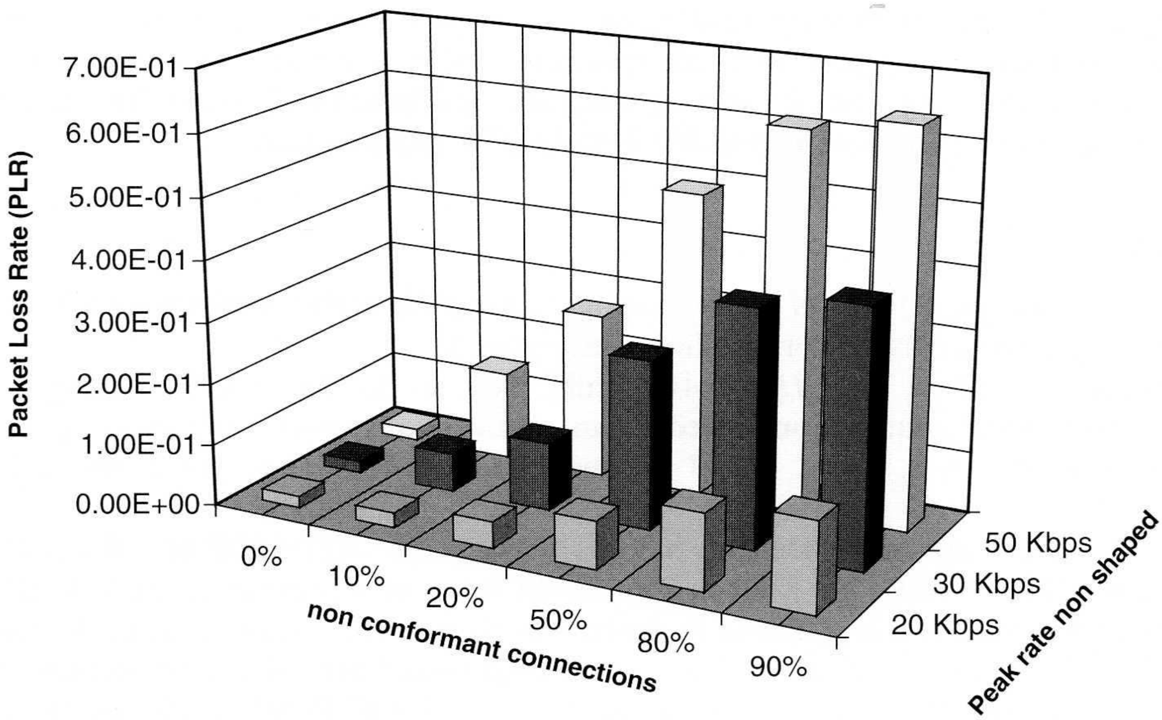

Figure 20. VoIP traffic, active shaper vs non-active shaper: PLR performance decrease.

ferent degradations scales are considered, since the non conformant traffic is supposed to ıerate packets at the rates: 20.0, 30.0 and 50.0 kbps.

Гhe ideal situation (active shaper) corresponds to the '0%' case, where non-conformant ınections are shaped before entering the network. In this case, the QoS requirements are ıranteed. Figures 20 and 21 highlight the performance decrease in case of no shaper ıilability. Precise numerical values are depicted in Tables XV–XVII. If non-conformant traffic ık rate is 20 kbps, 10% of not expected traffic is sufficient to violate PLR performance ıuirement, while if not expected traffic is more than 20% of the overall load, the impact on formance is dramatic, both for packet loss and for delay. When non-conformant peak rate is kbps, 20% of not expected load generates a performance decrease of an order of magnitude, PLR, and of 25 ms, for 'per-hop' PTD, out of a constraint of 50 ms. When non conformant ık rate reaches 50 kbps, only 10% of non conformance decreases the performance, ıcerning PLR, of more than an order of magnitude and, concerning PTD, of 27 ms (over given per-hop threshold).

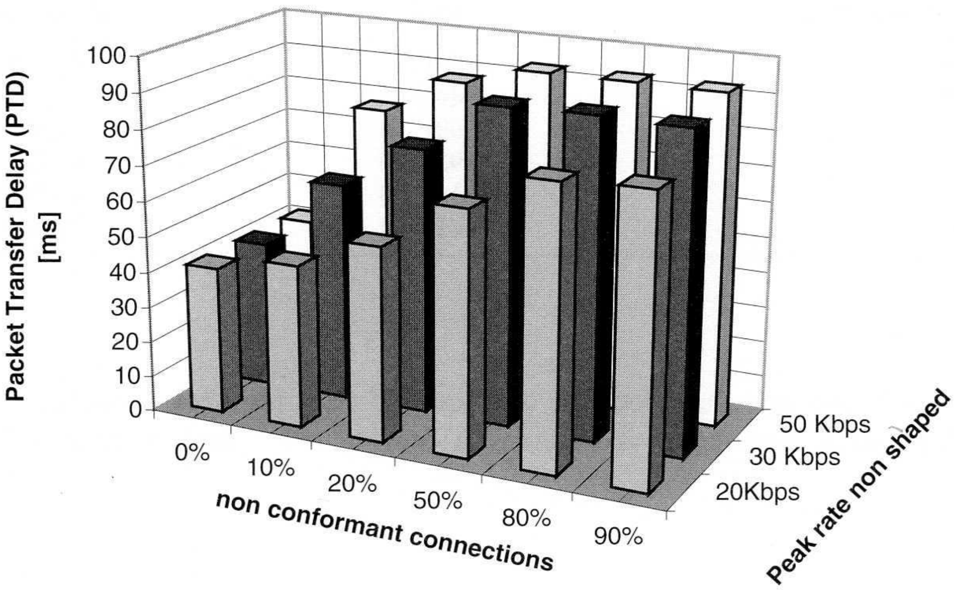

Figure 21. VoIP traffic, active shaper vs non-active shaper: PTD performance decrease.

Table XV. VoIP traffic, active shaper vs non-active shaper, non-conformant peak rate: 20 kbps.

Non-conformant connections (%)	PLR	PTD (ms)
0	1.86E-02	41.366
10	2.38E-02	45.815
20	4.25E-02	54.622
50	7.72E-02	68.181
80	1.23E-01	78.395
90	1.45E-01	79.464

Table XVI. VoIP traffic, active shaper vs non-active shaper, non-conformant peak rate: 30 kbps.

Non-conformant connections (%)	PLR	PTD (ms)
0	1.86E-02	41.366
10	6.26E-02	61.715
20	1.11E-01	74.831
50	2.73E-01	89.064
80	3.83E-01	89.648
90	4.11E-01	89.136

Table XVII. VoIP traffic, active shaper vs non-active shaper, non-conformant peak rate: 50 kbps.

Non-conformant connections (%)	PLR	PTD (ms)
0	1.86E-02	41.366
10	1.44E-01	77.447
20	2.69E-01	87.998
50	4.92E-01	93.221
80	6.15E-01	93.189
90	6.41E-01	93.003

It is simple to figure out the effect on a voice call that requires a packet loss of 10^{-2} as well as a delay below 50 ms, if it receives a loss of 10^{-1} and a delay of about 80 ms per hop.

5. CONCLUSIONS

This paper has described QoS solutions available in the market, transport technologies, parameters and management functions. The effect on QoS provision of traffic aggregation, CAC, and traffic shaping have been investigated and discussed. The study was aimed at evidencing that the real QoS provision is not totally dependent of the technology used, but heavily depends on the adoption of specific control functions, as packet marking, CAC and filtering that are topical for QoS provision. The results reported allow: to properly dimension non-homogeneous traffic trunk not penalizing the overall performance; to show the effect on performance metrics if CAC and traffic filtering are not implemented. The analysis has been carried out with simulations and many examples and results have been reported to justify the conclusions.

Other control components, such as routing and scheduling, are fundamental for QoS. Their impact on the performance will be evaluated in future research.

Floyd S, Estrin D, Minshall G, Paxson V, Zhang L, and others. Internet Research: comments on formulating the problem. 1998; 1–5, *manuscript in progress*, http://www-nrg.ee.lbl.gov/nrg-papers.html, last access June 2004.
ITU-T Recommendation. Terms and definitions related to quality of service and network performance including dependability. *ITU-T Recommendation E.800*, August 1994.
ETSI. Network aspects (NA); General aspects of quality of service (QoS) and network performance (NP). *ETSI Technical Report, ETR 003* (Second ed). October 1994.
ETSI. TC-SES, Satellite earth stations and systems (SES). Broadband Satellite Multimedia IP. IP Interworking over Satellite; Performance, Availability and Quality of Service. *ETSI Technical Report, TR 102 157 V1.0.0*, March 2003.
Gozdecki J, Jajszczyk A, Stankiewicz R. Quality of service terminology in IP networks. *IEEE Communications Magazine* 2003; **41**(3):153–159.
Hardy WC. *QoS Measurement and Evaluation of Telecommunications Quality of Service*. Wiley: Chichester, England, 2001.
Mean Opinion Score, Networking Definitions. http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci786677,00.html, last access June 2004.
Adami D, Marchese M, Ronga LS. TCP/IP based multimedia applications and services over satellite links: experience of an ASI/CNIT project. *IEEE Personal Communications Magazine* 2001; **8**(3):20–27.
Gellens R. The SYS and AUTH POP Response Codes. *IETF, RFC 3206*, February 2002.
ITU-T Recommendation. Support of IP based services using IP transfer capabilities. *ITU-T Recommendation Y.1241*, March 2001.
ATM Forum. Traffic Management Specification Version 4.0. *AF-TM-0056.000*, April 1996.
Onvural RO. *Asynchronous Transfer Mode Networks: Performance Issues*. Artech House: Boston, USA, 1994.
McDysan DE, Spohn DL. *ATM Theory and Application*. McGraw-Hill: New York, USA, 1999.
TC-SES, Satellite earth stations and systems (SES). Broadband Satellite Multimedia, IP over Satellite. *ETSI Technical Report, TR 101 985, V1.1.2*, November 2002.
ITU-T Recommendation. Network node interface for the synchronous digital hierarchy (SDH). *ITU-T Recommendation G.707*, March 1996.
IEEE 802.1 Working Group. Virtual LANs. IEEE standards for local and metropolitan area networks: virtual bridged local area networks. *IEEE 802.1Q*, 2003, http://www.ieee802.org/1/pages/802.1Q.html, last access June 2004.
IEEE 802.1 Working Group. Traffic class expediting and dynamic multicast filtering published in 802.1D. IEEE Standards for MAC bridges. *IEEE 802.1p*, 1998, http://www.ieee802.org/1/pages/802.1D.html, last access June 2004.
Bocci M, Guillet J. ATM in MPLS-based converged core data networks. *IEEE Communications Magazine* 2003; **41**(1):139–145.
Chao HJ, Guo X. *Quality of Service Control in High-Speed Networks*. Wiley: Chichester, England, 2002.
ITU-T Recommendation. Vocabulary of terms for broadband aspects of ISDN. *ITU-T Recommendation I.113*, June 1997.
Wright DJ. *Voice over Packet Networks*. Wiley: Chichester, England, 2001.
Braden R, Zhang L, Berson S, Herzog S, Jamin S. Resource Reservation Protocol (RSVP)—Version 1 Functional Specification. *IETF, RFC 2205*, September 1997.
Wroclawski J. The use of the resource reservation protocol with the integrated services. *IETF, RFC 2210*, September 1997.
Shenker S, Partridge C, Guerin R. Specification of guaranteed quality of service. *IETF, RFC 2212*, September 1997.
Wroclawski J. Specification of the controlled-load network element service. *IETF, RFC 2211*, September 1997.
IETF Differentiated Services Working Group. http://www.ietf.org/html.charters/diffserv-charter.html, last access June 2004.
Blake S, Black D, Carlson M, Davies E, Wang Z, Weiss W. An architecture for differentiated services. *IETF RFC 2475*, December 1998.
Heinanen J, Baker F, Weiss W, Wroclawski J. Assured forwarding PHB group. *IETF RFC 2597*, June 1999.
Giordano S, Salsano S, Van den Berghe S, Ventre G, Giannakopoulos D. Advanced QoS Provisioning in IP Networks: the European Premium IP Projects. *IEEE Communications Magazine* 2003; **41**(1):30–37.
Kota S, Marchese M. Quality of service for satellite IP networks: a survey. *International Journal of Satellite Communications* 2003; **21**(4–5):303–349.
Crawley E, Nair R, Jajagopalan B, Sandick H. A framework for QoS-based routing in the Internet. *IETF RFC 2386*, August 1998.
Montanez M. Deploying QoS in the enterprise. *Packet—Cisco System Users Magazine* 2002; **14**(4):30–34.
Byungsuk K, Sebüktekin I. An integrated IP QoS architecture—performance. *Proceedings of MILCOM 2002—IEEE Military Communications Conference*, Boston, MA, USA, October 2002; 1189–1193.
ITU-T Recommendation. Artificial Conversational Speech. *ITU-T Recommendation P.59*, March 1993.

Enzo Fortunato was born in Rome, Italy, in 1969. He got a degree in 'Electronics and Telecommunications' at the 'Istituto Industriale Galileo Galilei'I, Genoa, Italy, in 1988. From 1990 to November 2000, he was with 'Telecom Italia', involved in the design and management both of infrastructures for telephony and of telecommunication networks for multimedia services. Since December 2000 he has been working with Marconi Selenia Communications, Genoa, Italy, involved in the definition of new products and systems for tactical and strategic networks for Defense. Within his activity, he works with the Italian National Consortium of Telecommunications, University of Genoa Research Unit, about the study of new solutions for TCP/IP protocols for satellite and radio networks, ATM networks and Quality of Service.

Mario Marchese was born in Genoa, Italy in 1967. He got his 'Laurea' degree cum laude at the University of Genoa, Italy in 1992 and the Qualification as Professional Engineer in April 1992. He obtained his PhD (Italian 'Dottorato di Ricerca') degree in 'Telecommunications' at the University of Genoa in 1996. After a short industrial experience with Marconi S.p.A., he worked as a member of the research staff of the Telecommunication Networking Research Group by the University of Genoa with a post-doctoral scholarship. From 1999 he has been working with the Italian Consortium of Telecommunications (CNIT), by the University of Genoa Research Unit, where he is now Head of Research. He is author and co-author of more than 80 scientific works, including international magazines, international conferences and book chapters. He is the Official Representative of CNIT within the European Telecommunications Standard Institute (ETSI), Secretary of the IEEE Satellite and Space Communications Technical Committee and Member of the IEEE. His main research activity concerns: TCP/IP protocols, satellite networks, transport protocols for satellite links, ATM networks and related topics, best-effort multimedia networks.

Maurizio Mongelli was born in Savona, Italy in 1975. He got his 'Laurea' degree cum laude at the University of Genoa, Italy in 2000 and the Qualification as Professional Engineer in April 2002. He obtained his PhD (Italian 'Dottorato di Ricerca') degree in 'Electronic and Computer Engineering' at the University of Genoa in 2004. His PhD was funded by Marconi Selenia S.p.A. He worked for both Marconi Selenia S.p.A and the Italian Consortium of Telecommunications (CNIT) from 2000 to 2004. He is now a member of the research staff of the Telecommunication Networking Research Group at the University of Genoa, with a post-doctoral scholarship. His main research activity concerns: QoS protocol architectures and management functions, QoS inter-working, resource allocation, optimization algorithms and pricing for telecommunication systems.

Annamaria Raviola got her 'Laurea' degree Degree in mathematics at the University of Genoa, Italy in in 1986. She is currently studying to get a Master Degree in International Technology Management at Warwick University, Warwick, U.K. She is now the Evolutive Platform Director of the Terrestrial, Naval and Satellite Business Unit, of Marconi Selenia Communications, Genova, Italy, which is part of the Finmeccanica Group. She is responsible to design complex solutions in the field of ICT and tactical communications and she has the responsibility of a group of people for the introduction of new technologies and the definition of new products for the defense market. She is also involved in the Business Development activities for the Utilities and Government Market for the Homeland Security Project.

