Beyond Technology: The Missing Pieces for QoS Success*

L. Burgstahler K. Dolzer C. Hauser J. Jähnert S. Junghans C. Macián † W. Payer

Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, 70569 Stuttgart, Germany
[burgstahler, dolzer, hauser, jaehnert, junghans, macian, payer]@ikr.uni-stuttgart.de

ABSTRACT

Years of research on QoS architectures for IP networks have delivered sophisticated proposals, which have nevertheless not found broad commercial use. The reasons are not lack of technical soundness or insurmountable technological complexity, but insufficient attention to other, non-QoS-specific matters. First among them is the lack of a commercialization model for the Internet together with the necessary accounting and charging architecture. Another crucial issue is the assurance of end-to-end QoS coherence in the face of multiple intervening parties (network and content providers, users). Furthermore, the practical requirements imposed by those parties to any successful QoS architecture have not been fully taken into account: Ease of management, simplicity and measurable guarantees are some of the main ones. In this paper, the overall constraints on and conditions for the successful deployment of QoS in IP networks are analyzed and some possible directions explored.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: General, Network Architecture and Design, Internetworking

General Terms

Design, Economics, Human Factors, Management, Performance, Security

Keywords

QoS, Internetworking, Next Generation Internet

1. INTRODUCTION

For at least the last 35 years, Quality of Service (QoS) in data networks has been vehemently debated. Many implementations thereof exist for all kinds of networks. With the advent of IP as the all-encompassing networking technology, in the last 10-15 years a lot of attention has been devoted to the development of QoS architectures for IP-centric networks. Despite all that, QoS is not widely applied presently and its future is uncertain. QoS, understood as the provision of differentiated levels of quality to different proportions of traffic, is nevertheless a necessity. It is our claim, that it is not possible to provide a satisfactory quality to every present and future service in every environment in a best-effort network in a sustainable way. This is especially evident, although not exclusively, in access networks, where overprovisioning is less of an option, for technical as well as economic reasons.

In principle, some form of differentiated QoS can already be provided in any environment, as is proven by the various existing architectures. It is mainly a number of other, non-QoS-specific constraints, which have not been satisfactorily solved so far: Accounting, charging and billing in the commercial Internet age, e.g., or the development of adequate business models. How to solve the interaction among network and content providers and the user in a fragmented market in order to provide consistent QoS levels is another open issue. How to deal with an increasing heterogeneity, in the number of scenarios as well as networking technologies and requirements is yet another one.

In this paper, an extensive analysis of the different constraints affecting the development, deployment and widespread usage of differentiated QoS in IP-centric networks will be presented. To that end, the meaning of QoS will be reviewed in section 2. Section 3 reviews the different mechanisms and criteria employed to implement QoS architectures, as well as evaluating if and how QoS is needed — from the point of view of the different actors involved. Section 4 points to the implications that the latest networking trends have for the development of QoS. According to the points exposed in sections 3 and 4, section 5 analyzes the critical issues for the future of QoS in IP networks, and section 6 concludes the paper.

2. WHAT IS QOS?

Many of the so-called definitions of QoS found today are merely descriptions of technologies and techniques being used to provide a better service quality to selected traffic classes. But QoS is a concept and not a technology or a

*To appear in the ACM SIGCOMM 2003 Workshops, August 25 & 27, 2003, Karlsruhe, Germany.
†Corresponding Author

ACM SIGCOMM 2003 Workshops August 25 & 27, 2003, Karlsruhe, Germany

technique. The term QoS is often used very restrictively in the sense of *high level of QoS*, which could mean high bandwidth, low packet loss, small delay, etc. Simply said, everything that provides a better quality than best effort is called QoS, disregarding that best effort does already provide some quality.

Actually, QoS describes something much broader as the two following definitions clearly show:

- "The collective effect of service performance which determines the degree of satisfaction of a *user* of the *service*." (ITU-T) [7]

- "A set of service requirements to be met by the network while transporting a flow." (IETF) [6]

ITU-T and IETF have a different point of view: ITU-T takes the capability of the network as given and the user rates the level of quality, whereas IETF takes the user's requirements as given and the network has to meet them. Despite this difference, both definitions have something in common: They do not restrict the term QoS to a high level of quality but to some measurable performance that satisfies some(one's) demand. This implies that there are low expectations that can be met by very simple means and there are very strict requirements that could only be satisfied by very advanced techniques.

As a consequence, we can state that *any* network offers a degree of QoS and this renders the question whether we need QoS inane – some level of QoS is always existing. Therefore, the question is whether there is a need for different quality levels for different services in different environments.[1] To better answer this question, we need to examine how the general definitions above fit to specific users and/or services.

QoS as perceived by users is very subjective, since everybody has own individual preferences. Yet, for most services there are levels of quality that are widely accepted as good, i.e., nobody would rate a telephony service as being bad, because the speech quality is not Hi-Fi. People also set priorities when different qualities are to be compared: Mobility, price and ease of use can be more important than speech quality or availability. Finally, there are physiological limits (as, e.g., the human hearing range), where a service's quality does not improve by crossing them.

As every user has different demands on different services, the set of service requirements mentioned by the IETF is not the same for each service. Neither is the level of quality expressed by a certain numerical value the same — even for the same criterion — if different services are considered. An example for this is delay: Interactive real-time applications have more stringent requirements in terms of delay than non-real-time applications. Still there is a level of delay where even for non-real-time applications the quality becomes unbearable. Thereby, the first task is to define values for all quality criteria that indicate the minimum level of good quality. Field trials were conducted to quantify what the majority of users considers to be good or acceptable quality for a certain service (e.g., [5]). Since users do not (and cannot) express their perception in specific numbers, it was necessary to map abstract evaluations ("long" delay, "much" noise) to exact values describing quality levels, e.g., delay and loss.

Once the different quality levels are defined, the question is who is responsible for the adherence to these levels. Whereas ITU-T does not allocate responsibilities, IETF names the network only. Normally the user is only interested in end-to-end QoS, so we have to divide the end-to-end path into several segments, at least: End-user equipment, access networks, and core network(s). The influence of end-user equipment on QoS is large: QoS-aware applications, the selection of a suitable codec that improves QoS often regardless of the underlying network, fast hardware that can process (e.g., compress) data better, and high-end speakers enhancing the sound quality, to name a few. But this remains entirely in the hand of the user. We will therefore assume that the user provides equipment that supports his demand for a certain QoS level, and exclude this issue in the following.

Having solved quality issues at his premises, the user then requests the network provider to support his requirements. This request is the driving force for a provider's decision to deploy any QoS mechanisms. Users will only select and pay for a provider, if its service is satisfying. The biggest problem for the provider remains to find an optimal solution, because a possible optimum depends on the user and the service. The many (quality) criteria open up a multi-dimensional space that usually offers not a single solution but a solution space, containing many good solutions. Thus, the implementation of a QoS architecture will always be a trade-off. An alternative for a provider would be to offer a network optimized for specific services. This is what telephony companies have done for decades.

Excluding optimized networks and heading for an integrated network, our focus is exclusively on the access and the core network's capability to provide a requested level of QoS to a user with respect to the various requirements. Within this context, our evaluation includes traffic-dependent solutions like scheduling, shaping, routing etc., security- and mobility-related issues as well as economic ones like pricing. In brief, we consider QoS to be a set of context-dependent (service, user) requirements to be met by the network end-to-end to provide a degree of satisfaction to a user of the service. A QoS architecture describes a structured solution to meet those requirements. This finally leads us to the refined question: Is there a need for mechanisms supporting differentiated QoS in the core and access networks and if so, what are they?

3. ALLOCATING RESOURCES

In recent years, much attention has been devoted to differentiated QoS because it is assumed that it is impossible to provide a satisfactory quality to every present and future service in every environment in a best-effort network in a sustainable way. Even the fiercest defenders of overprovisioning would agree that wireless networks face severe resource scarcity. On top of that, bandwidth is not the only problem: Even in lightly loaded networks, undifferentiated traffic handling can induce unacceptable delay and/or jitter for sensitive real-time applications. Hence there will always be a likelihood that the available resources do not satisfy all users' expectations. Consequently there will be users who are willing to pay more for their share of the resources but they will ask for at least relative guarantees. For the network

[1]Note that we will use the term *differentiated* in a general way and not restrict it to IETF's differentiated services specification (e.g., [2]).

providers, this means that mechanisms for resource allocation must be deployed. The first question in this context is to clarify which criteria dominate our allocation strategy.

3.1 Allocation Criteria

A variety of different criteria exist which are the basis for mechanisms being introduced in the next section. Unfortunately, they are heavily interrelated, introducing unavoidable trade-offs. One such pair of criteria are multiplexing gain on the one hand and traffic differentiation/isolation on the other. Greater differentiation/isolation generally leads to a reduced multiplexing gain. The decision on how much differentiation is really required also depends on whether relative or absolute QoS guarantees are required [8]. Scalability, another major criterion, has two counterparts, namely granularity and dynamics. Finer granularity and higher dynamics, respectively, lead to decreased scalability.

Beside these technical criteria, also more economic ones like profit maximization and maximization of network utilization can be applied. From a network provider's economic point of view, QoS provisioning and the decision of absolute versus relative guarantees makes only sense, if there is an efficient and scalable way to charge for differentiated services. The offered level, dynamics, and granularity of QoS to a service determines also the granularity at which the resource usage must be metered.

Unfortunately, a fine granular accounting scheme might impose higher costs on an operator than the additional revenue being generated out of the corresponding offer. This fact is often neglected in discussions about QoS provisioning as these are mainly user-centered and thus ignore the operators' needs.

Hence, it is important to clearly identify the criteria influencing the design of a QoS architecture. The parameters to be considered can be derived from technical, economical and social (and probably other) requirements. Thus, there is no optimal solution in this context, but always a trade-off. The challenge is to find an equilibrium, in which most relevant criteria are sufficiently satisfied.

3.2 Allocation Mechanisms

3.2.1 Technical approaches

There are several well-known mechanisms to allocate resources, a subset of which is usually combined in any network implementation: Reservation, bandwidth overprovisioning, admission control, traffic shaping, traffic separation and scheduling.

Explicit reservation of resources is the most basic mechanism. It may be static, carried out by network management or dynamic, which requires signaling. The simplest mechanism within resource reservation is bandwidth overprovisioning (over), i.e., the reserved bandwidth is somewhere between the mean and the peak bandwidth [10]. In order to be able to over bandwidth, the amount of incoming traffic has to be known and controlled. To do so, admission control was proposed [12]. Another mechanism to control the amount of data delivered to a system is traffic shaping. Hereby, in case of congestion traffic at the ingress is selectively delayed instead of rejected. Finally, traffic separation is the means for (virtually or physically) splitting up resources for different requirements, usually with the help of a scheduling algorithm [18, 4].

3.2.2 Economic approaches

Resource distribution can not only be achieved by technical means. Economical mechanisms are basically equally well suited to distribute resources efficiently. Pricing plays a central role in controlling the access to scarce goods: Price differentiation, market segmentation, service bundling, and auction are some examples of economic approaches to QoS resource allocation.

Price differentiation's basic assumption [16] is that each customer is willing to pay differently for each service. The challenge is to discover this maximum amount for each customer. For this reason, classes of customers are defined, e.g., as airlines do (students, tourists and business travelers).

The most successful resource distribution mechanism was and is the auction. Nevertheless, accomplishing an auction on the network leaves some open questions with respect to the overall scalability and product composition.

Service bundling offers a set of services for a fraction of the components' total price. It can be shown [16], that adequate bundling can increase the overall revenue significantly. This also helps to reduce the number of business partners, that a customer has to deal with. Sometimes the bundle composition is targeted to exclude the requirements of certain potential market segments. The goal is to demarcate the different market segments and product positioning by introducing exclusive offerings.

Summarizing, any QoS architecture will probably combine several of these mechanisms According to the relevant allocation criteria. It is the use of economical mechanisms for resource allocation, as well as the interworking between these and the more technical ones, which has not been sufficiently explored yet.

3.3 Requirements on Allocation Strategies

3.3.1 Customer Requirements

Customers are generally interested in communication anywhere and anytime. This basic requirement holds true for every customer, be it a user at home accessing some entertainment service or a business customer on a trip needing to synchronize his groupware application. Independently of the service being in use, a large fraction of the customers does not want to be faced with too much complexity or technical detail in order to make this happen. The customer's perception of a service's quality is quite simple: The service is working satisfactorily or not. While the past has shown that customers will accept a non satisfactory service in terms of quality, if another advantage can be gained (e.g., speech quality vs. availability in mobile networks), it is also easy to realize that the customers' perception of a service depends on his familiarization and experience with available quality in general. The overall trend to mobility in its different characteristics combined with other items slipping quite slowly into customers attention, like security issues, will largely dictate what the customer requires.

From the previous discussion, it is easy to derive some customer requirements, which are basically not QoS-specific, but capture the customers' relationship to service access.

Generally all types of services should be accessible at any time and at any place. All necessary steps for installation, configuration, update and use of a service need to be simple to allow the service to be accessible and usable by people with a wide range of skills and in several situations. All

service offerings need to employ a flexible pricing structure (i.e., for each session criteria like best quality or lowest cost can be applied) while generally the overall cost has to be low and the specific cost of services needs to be transparent to and understandable for a user.

Even if the customers' perception of quality is quite simple, the predictability of the quality of a service is very important (users generally prefer lower but stable quality than higher but variable quality) which also means acceptable responsiveness and set-up time. Quality parameter expectations and users' preferences could be part of one or more user profiles, which could potentially contain additional criteria like uptime, time to repair or compensation for poor subjective QoS. This can be regarded as representing a contract based on a Service Level Agreement (SLA) for the user. Generally it is also desirable from a customers' point of view for the service access to be reliable, error tolerant, stable and to provide self-monitoring and diagnosis properties. This would allow, e.g., for a visual feedback of system status information and help define responsibilities in case of any problem. System security in terms of user authentication, encryption, remote access protection, privacy and confidentiality as well as minimized opportunity for misuse has an increasing significance.

There is no real difference in the requirements of home customers and business customers. It could be argued that a business customer in principle can gain more experience with comparable systems and hence has a more detailed notion of his QoS requirements.

3.3.2 Content Provider Requirements

The main concerns of the content provider are twofold: How to steadily and cost-efficiently develop new services, and how to reach a maximum customer base. Product development can be made easier if it is free of network operator-specific constraints and if the resulting product is portable among different technologies, which also allows to reach a broader audience more easily. Equally, a simple deployment mechanism helps to integrate the product faster and thus reduces its time-to-market. As a consequence, minimum direct interaction with the network is desired to hide its heterogeneity. This argument speaks in favor of middleware solutions and open (network interface) standards.

3.3.3 Application Requirements

There are two critical points in the relationship between applications and the delivery of a certain QoS to them: The ability of the application to adapt to changing conditions and the exchange of information (or lack thereof) between the application and the lower layers for the joint delivery of QoS. Although applications are not the main focus of this paper, both issues are shortly reviewed in this section.

Most applications have been developed under the premise of total independence from the underlying layers, especially from the network technology. Since the application cannot affect the behavior of the network, the only alternative to ensure a certain quality level by the application itself is to adapt its own behavior to the present network conditions [15]. That could become the default paradigm in QoS-aware environments such as mobile applications and (streaming) video and audio transmission over the Internet: The choice of codecs, for example, could be dependent on the partner machine's capabilities and the perceived transmission quality. The question is how that "perception" is achieved.

In any case, heterogeneous environments, in which an application uses a number of different communication channels during the life of a communication session, strongly ask for application adaptability. Mobility, specifically, introduces the fact of changing environmental conditions with time. Without dynamic application adaptability, though, mobility would imply the tear down and setup of a new communication session every time the environment changes significantly, although the application might support different QoS levels, e.g., through the use of different codecs.

One possibility for the applications to build their "perception" of network quality would be to directly communicate with the network layer through an interlayer signaling protocol. In this case, the differentiation between network and end system/application would become somewhat artificial. Nevertheless, it seems very little probable that a unified network interface, independently of network technology, will emerge. Furthermore, different networks will support different QoS architectures, mechanisms and implementations thereof. The mapping between application requirements and network capabilities will certainly not be standard. Furthermore, many applications will surely not be able, now and in the near future, to communicate their quality requirements. It must not be forgotten either, that no matter which configuration is deemed suitable by the application itself, the end user will probably want to introduce other (non-)technical conditions in the choice of quality, e.g., price considerations. As a consequence, the appearance of some form of QoS-broker/proxy/middleware [14] between the applications and the network, which can be configured by the user, presents itself as a very strong alternative. That QoS-middleware would then terminate the signaling with the different networks and together with the user requirements forward the information, possibly over an standard interface, to the application.

The question is if there will be any signaling between the middleware and the application, or if a "pure" QoS-proxy is enough, i.e., an entity taking the QoS-decisions according to certain policies without informing the application. Since application adaptability has been identified as a very desirable condition, in this scenario the application itself would have to detect the quality delivered by the network (under influence from the proxy). The application could simply measure certain technical parameters to choose the most appropriate mechanism in every circumstance. The problem with this approach is how to introduce non-measurable or non-technical parameters in the equation, like the price. Consequently, it seems very probable that QoS-middleware and not only proxies will emerge to coordinate networks, user desires and application requirements.

3.3.4 Network Provider Requirements

Before the arising of data networks and the general opening of markets for operators and providers of telecommunication networks and services, there was a "telephony culture" with strict requirements in terms of high system reliability, world-wide interoperability, guaranteed QoS (dedicated circuits and a signaling scheme to check whether sufficient end-to-end capacity was available) and the general ability of the system to identify the caller. This has changed drastically.

Data networks are very different and not only with re-

spect to the aforementioned criteria, but also in many other aspects: The traffic pattern, the very usual connectionless property, general dynamics in routing, traffic crossing many administrative zones, etc. Even more important, data networks have many more degrees of freedom in their management and configuration than telephone networks, due to their multi-service nature. However, to integrate services from both worlds in a converged network, many of these degrees of freedom need to be restricted, resulting in a much more static network. Many of the constraints can be regarded as originally being driven either by customers or content providers in order to satisfy their respective needs and to generate revenue from that.

3.3.4.1 Access/ISP.

The main interest of a network provider in the access is to optimize his coverage and to minimize the corresponding costs. Connecting all potential customers via fiber is generally regarded as being far too expensive, while the use of existing wiring (like twisted-pair from POTS or coaxial cable from the CaTV) or mobile access in general introduces restrictions in terms of bandwidth or possible bridgeable distance to the customer. This is why access networks currently are the major bottleneck. Technological trends let foresee that this will be the case for some more years to come.

Here the need for QoS support is especially evident. The major issue is the quite large number of available access technologies with very different characteristics. This has prevented the development of a unified access architecture, including a unified QoS architecture. Whatever technology and mechanisms are chosen they need to conform to the operators' requirements. One major item is the provisioning of authentication, authorization and accounting mechanisms in combination with simple and reliable metering, charging and billing schemes for all types of services. Scalability in combination with moderate processing requirements is essential. This also means a moderately generated signaling load, semi-static customer profiles and only a minimum of dynamics in the network.

One possibility to provide the predictable and coherent end-to-end service quality that users expect would be to host most services within the access network. Services like ASP, content distribution, presentation adaptation, content reformatting, server based games or applications, web hosting and others could be locally available. Additionally, the increasing demand of mobility by customers needs to be supported. Finally all that should be easy to manage and operate.

Current approaches like, e.g., PacketCable aim at developing architectures to integrate services like Internet telephony. While these architectures usually contain operation support systems for AAA and security, they only cover a subset of services, are extremely complex, require cooperation of the customers and do not solve the issue of interaction to users, core network operators and content providers in order to provide QoS in a persistent form.

3.3.4.2 Core.

The main requirements of a core network provider are robustness, resilience and predictability of the network. Also due to the speed of operation, especially when moving to optical networks, all mechanisms being applied in the core network itself need to be simple and scalable. Moreover, data storage and the amount of available processing time per packet is a general issue. The trend to move most of the necessary processing to the edge of the network, where the conditions are not that strict, will persist.

Ever since, it has been in the interest of the operator to keep core networks very static with changes appearing in the order of tens of minutes to weeks in order not to endanger network predictability. Also the core operator does not want to be faced with individual traffic streams but only with aggregates, which in turn increases scalability. This aggregation or separation of traffic can be done according to a number of different criteria like service classes, application types, origin or destination of traffic, VPNs, and more. Currently and in the near future, bandwidth is not being regarded as an issue in core networks, which paves the way for quasi-leased-line approaches.

3.4 The QoS Story So Far: Scanning the Past

The application of QoS architectures can be traditionally subdivided into telephone networks and data networks. Hereby, the telephone scenario is rather simple as traffic in such networks is characterized by almost constant bit rate and QoS can be achieved by signaling-triggered reservation of the maximum bandwidth. This is true for all networks which are dimensioned for voice traffic like the current telephone network or mobile networks of the 2nd generation.

For data networks, QoS provisioning is much more complicated because the traffic is sporadic and bursty, i.e., reservation of the maximum bandwidth results in an enormous waste of bandwidth. The multi-service nature of IP networks, and the fact that different applications have different QoS requirements, intensify this problem. LAN/MAN technologies introduced QoS considerations in their architectures very early: FDDI, Token Ring, etc. Nevertheless, they were not very successful, since it soon became evident that bandwidth scarcity was not a problem in the LAN/MAN domain. Their unnecessary complexity (and superior price) strongly pushed for easier (and cheaper) solutions, like Ethernet, without QoS-support.

Starting in the 80ies, ATM was developed as the future broadband integrated network. However, after developing mechanisms for every contingency, industry realized that the implementation of all these features was much too costly and the management complexity far too high. Furthermore, caused by the relative small cell size as well as the mapping of IP packets in those cells, a very high overhead ("cell tax") was introduced by ATM.

The failure of ATM as the broadband convergence layer led to the development of QoS-supporting IP-based network architectures. The first one was Integrated Services (IntServ), which — in principle — is similar to ATM and also provides QoS by resource reservation-triggered signaling. However, because soft states were to be kept in intermediate nodes, the scalability of this proposal for large networks was deemed insufficient. Therefore, Differentiated Services (DiffServ), was proposed which does not require signaling. In order to additionally reduce the complexity, QoS is not provided on a per-flow basis like in ATM or in IntServ, but for traffic aggregates. By doing so, relative QoS differentiation can be achieved between different so-called behavior aggregates. However, caused by the lack of explicit reservation, signaling and admission control, no absolute guarantees can be given without applying complex extensions

to the model, which is a problem for some services. It is also difficult to price, since the quality delivered by relative priorities is difficult to grasp and not constant over time. In our opinion, this IP-based QoS architecture also will not be a great success beyond very basic implementations (2–3 classes).

A parallel evolution went towards applying ATM just as a simple and dumb forwarding technology and apply all intelligence within IP. In this context, Classical IP (CLIP), multi-protocol-over-ATM (MPOA), as well as multi-protocol label switching (MPLS) were proposed. Whereas CLIP and MPOA also were not considered broadly, MPLS gained increasing interest because it provides fast and efficient forwarding and QoS only with the mechanism of (physical) traffic separation in a quasi-leased-line approach. Another advantage of this approach is its relative low management complexity. This idea is now being extended in GMPLS (Generalized MPLS) and integrated in the Automatically Switched Optical Network (ASON) umbrella in order to totally skip ATM and directly control the optical layer — which offers much more bandwidth — by IP. Thus, IP is the convergence layer with respect to services and control, but the convergence layer with respect to QoS is located in lower layers, see also section 4.

4. TRENDS IN NETWORKING

4.1 Access

Even with growing rates on access lines, the last mile is still a bottleneck today and in the near future. Therefore QoS support is necessary to provide fairness on these technologies. The installation of new network cables is a costly procedure, so Internet providers tend to use already installed resources like telephone lines or TV cable networks or wireless technologies to provide access to their network.

4.1.1 Wired

The two-wire telephone lines are initially dimensioned for low frequency voice transmission. With growing bandwidth demand, new technologies use new frequency bands on these lines. Those Digital Subscriber Line (xDSL) technologies are available as a whole family with different data rates from 144kbit/s to 52Mbit/s and in symmetric or asymmetric rates. Unfortunately, the maximum data rate of xDSL links is strongly dependent on distance. While xDSL covers several kilometers with line rates of about 1Mbit/s, the distance decreases to 1km for symmetric transmission of 10Mbit/s or 2.5km in the asymmetric 10/1Mbit/s case. For higher maximum rates, the distance constraints get even tighter. To achieve high bit rates, the provider must bring his line termination equipment as close as possible to the customer, which implies high cost and complexity. The xDSL technologies provide a layer-1 transport mechanism and serve as a "bit-pump". ATM is used as an additional layer-2 protocol, which could provide a large number of QoS provisioning functions.

TV cable networks are designed for high data rates as needed for TV broadcasting and therefore would also be suitable for broadband network data transmission. As these networks were designed as broadcast media not providing bi-directional communication channels, they must be upgraded with bi-directional amplifiers. Cable networks are characterized by a tree and branch topology. At the root of the tree, a head-end entity controls the traffic. Bi-directionality transforms the cable into a shared medium. In order to control access to this shared medium, especially in the upstream direction, a MAC protocol is needed, as well as mechanisms to guarantee privacy. The deployment of an effective distributed MAC as introduced in [3] can solve the QoS challenge, which is especially necessary for the traffic from the users to the core network.

Powerline technology was seen as a third simple approach for broadband access with many characteristics quite similar to CaTV. While a lot of research and development has been done in this area, it seems now as if powerline technology had been set aside.

Current access technologies like xDSL or TV cable provide bit rates, which allow the transport of several parallel multimedia streams. In conjunction with ATM for the xDSL case or a suitable MAC protocol for TV cables the provisioning of QoS on the last mile underlies no technological limitations.

4.1.2 Wireless

With the increasing penetration of networked applications in users' lives, especially in leisure time, the Internet should be ubiquitously available. Therefore, network access will have to support more and more user mobility.

Current mobile networks (GSM, GPRS) are not yet fully Internet compliant and can not provide sufficient bandwidth to large audiences simultaneously. Also UMTS (as well as EDGE/IMT-2000), although providing an IP-packet service using a tunneling mechanism, still employs all the circuit-switched mechanisms of 2nd Generation Networks [1]. While UMTS extends the capacity of a wireless access cell significantly, it is still expected to be a future bottleneck.

The approach of supporting packet switching over the existing connection-oriented network is mostly considered as an intermediate step towards a pure IP-based solution [11], which finally will be available in the fourth Generation mobile communication (4G) networks. In 4G networks, heterogeneous mobile networks incorporating different technologies such as UMTS and IEEE 802.11 will be integrated aiming at seamless so-called vertical handovers between them. Thereby, the advantages of several wireless access technologies are exploited and combined. IEEE 802.11 is a short-range Ethernet-like access technology that provides 10Mbit/s in current standards and even higher in future versions. Nevertheless, a global coverage by IEEE 802.11 is not suitable; therefore an integration with the broad coverage UMTS is aimed at. More generally, this approach includes all possible access technologies by handling the mobility management on the network layer.

These different integrated networks are not necessarily provided and operated by one single network provider. Thus multiple operators can be involved in a communication path, whereby interaction among them and security become an issue, see 5.3 and 5.5. There will certainly be roaming agreements among providers, easing the interaction with the user. Nevertheless, the complexity of interaction among providers themselves remains. Furthermore, thinking on the multitude of possible operators, it is arguable whether there will exist treaties among all of them.

The special characteristics of the air interface have triggered a further trend aiming at solving the QoS problem not only on the network layer. The main reason behind it is the

higher error rate on the air interface compared with modern wired networks and its effect on TCP behavior. TCP will interpret any packet loss as congestion, which is mostly the reason for packet loss on the core network. As a result, TCP decreases the sending rate, provoking an unnecessary reduction on network performance. To increase the exploitation of the wireless access network, a movement can be observed where higher network layers tackle these effects, so that in terms of QoS the user will not notice these network limitations. Issues currently being addressed include robustness against errors in mobile networks: As long as the communication context is kept uncorrupted and vital control information contained in the packet headers arrives correctly, the application layer could handle payload errors. Passing of erroneous SDUs from a link layer through the error-ignorant IP layer to the application layer would then be more efficient than discarding the packets. A task still to be solved is the problem of how to achieve the required interlayer cooperation between application and link layer and possibly even lower.

4.2 Core

As already indicated in Section 3.3.1, a technological trend in the core is from an IP-over-ATM-over-SDH-over-WDM network architecture towards an IP-controlled optical backbone, also known under the keyword IP-over-WDM where IP controls a network which all-optically forwards data from network ingress to network egress. The main enablers of this transition towards an IP-over-WDM architecture are (i) the progress in optical components which allows to start thinking about optical switching as well as (ii) ultra long haul transmission of optical signals which allow to transmit a signal for 1000s of kilometers without regeneration.

Realizations of IP-over-WDM architectures can be classified by increasing granularity as optical circuit switching, optical burst switching, and optical packet switching, with the latter being not foreseen for many years to come.

Main drivers for this evolution are from the data plane point of view (i) the strongly increased amount of bandwidth which is provided by WDM, (ii) an increase in costs (CapEx and OpEx) of SDH-based networks, which is faster than an increase in revenue, and (iii) the need for faster provision of bandwidth.

From the control point of view, the ASON builds an umbrella also including the standardization efforts in the context of GMPLS of the IETF as well as the user network interface (UNI) which has been standardized by the Optical Internet Forum (OIF). This umbrella allows the control of all layers in a future core network and thus is the basis for cost efficiency and fast provisioning of bandwidth. The approach again is pseudo-circuit switched, pseudo-static and simple, as is the QoS being offered by such networks.

4.3 Security

While there are still many networks without data protection, there is a trend towards securing networked communication. This is driven also by a trend towards e-commerce and context-aware applications both handling a lot of sensitive user data. One result from this trend is that security and privacy increasingly will influence the user's rating of a service which by now mostly is dominated by factors like bandwidth, delay, or price. Therefore, it must be assumed that in the future more and more traffic will be encrypted. Users will also be increasingly careful regarding trust, e.g., towards network operators. This results firstly in the need of considering the effect of cryptographic delay when regarding overall QoS as sensed by the user and secondly in the necessity of securing the QoS mechanisms themselves (see section 5.4).

4.4 Service Models and Service Introduction

Networks evolve. There is a constant push to extend their functionality and the services supported, since this is seen as the way to increase revenue by the network operators and to increase utility (in the economic sense) by the customers. Especially in the field of service models several trends support this claim, as will be explained shortly.

Application Service Providers (ASP) and Managed Service Providers (MSP) are closely related [13]. They offer to reduce the management complexity and internal know-how requirements for the operation and maintenance of Information Technology (IT) and communications infrastructure to other companies by taking over those tasks. In the process, they also help to reduce personnel and equipment costs at the outsourcing companies. This business model can nevertheless only be successful, if the end users experience the same working conditions as if the applications were locally installed in the company's LAN. This is especially the case for interactive applications, typical of office environments (including intra-company phone calls, usually through a PBX). Furthermore, the transfer of company internal information to and from the ASP/MSP servers imposes the need for secure transmission. This combination of ASP/MSP and VPN technology clearly necessitates some sort of QoS architecture to be able to deliver that "private network feel", i.e., to separate the "private" traffic from the rest of the backbone traffic and give special treatment to the most sensitive information, like interactive traffic. This paradigm also applies to global private networks, in which independently of distance and interconnecting network technology, members of the same team should share a common work environment.

The given examples of special network instances, with their particular functionality and requirements, are concrete occurrences of the more general overlay network concept [17]. Overlays present nowadays a very relevant trend, for they simplify the introduction of new functionality in parts of the network without forcing an update of the whole of it. They also allow to connect "functionality islands", thus facilitating the spreading of new services. In order to guarantee the correct functioning of such overlays, a QoS architecture, as has been argued, would be very helpful.

Another important aspect of network evolution is the transitional phase between legacy and emerging technologies. In this respect, middleware and proxies represent a powerful help. Horizontal and vertical middleware (i.e., between layers or between network elements) and proxies provide an entry point into the network to realize a functionality that the applications or even the network itself can not handle. Policies and profiles are the basis of such systems. As was discussed in the section 3.3.3, they will probably play an increasingly important role in the future, also as supporting entities of QoS architectures.

4.5 Other evolutionary trends

As of now there is no visible trend allowing a forecast if

the market in terms of network and content providers will truly consolidate. Some years ago there was a considerable consolidation among content providers, but this was well before communication needs became as self-evident as today. While due to the economical development, in recent months a similar development among network providers in general (and especially mobile and core network operators) has occurred, this is not necessarily true for content providers as well. However, independent of market consolidation or fragmentation, any trend will have an influence on daily communication in terms of the number of parties participating in a communication. Therefore, it will also influence the evolution of the ways of interacting between providers to offer a consistent (end-to-end) QoS.

Another relevant issue here is the number of ISPs/access networks and content providers that a customer wants or needs to be involved with. While there is evidence that customers would want to have as many offers and options as possible, at the same time the number of points of interaction to any of these providers should be as small as possible. While, again, a forecast is hardly possible, the topic is of high relevance as issues like multihoming in combination with addressing, routing, and route pinning, the heterogeneity of network access and the quantity of changeovers between different networks due to user and terminal mobility as well as signaling and dynamics in networks in general are closely interlinked.

5. CRITICAL ISSUES

5.1 The Impact of Complexity

Reliability is the principal asset of network operators, as discussed previously (see section 3.3). Management complexity, derived from the introduction of complex QoS architectures, endangers reliability [2]. It increases the risk of malfunctioning by defective configuration and makes problem resolution more complex: Tracing the real cause of misbehavior is more difficult. Unexpected feature interaction is a prevalent danger in modern, complex multi-service networks [9]. As a consequence, network operators have to invest heavily on network maintenance and acquire the necessary manpower and expertise. This goes to such length as endangering the increase in profit expected from the added functionality. That is why reluctant network operators would welcome easy to manage QoS architectures.

5.2 Commercialization

Technology has reached a position in which it can translate most QoS proposals into reality, but certainly not for free. This economic dimension is accompanied by a deep concern on copyright protection among content providers, who see digital content delivery over the network as a potentially dangerous distribution channel. Furthermore, a pervasive all-free mentality still dominates the Internet, which predisposes customers against online paying services. These issues shape a rough business market for network-based QoS for which few (successful) business models have been developed. Without them, the technological possibilities will never be translated into product offerings.

Virtually all research is considering IP as the final means for integrating access networks from any technology — wireless or wired — and the equally IP-based core network. This poses imminent problems regarding commercialization. Whereas the packet-based Internet has not been designed for commercial purposes, the traditional circuit-switched network infrastructure was designed for commercial purposes from the early beginning. Hence, migration adds considerably to the pressure to provide commercial services in this migrated IP-network.

While packet-switched voice and data communications are currently the key drivers for the development of new communication systems and technologies and thus the main cost factor, circuit-switched voice communication still dominates the telecommunications market with respect to revenue. The circuit-switched network has a well-established business model already describing in detail the relationship between the customer and the network operator, and the users widely accept the pricing model offered by the operator. However, in the Internet, there is not yet a well-established business model also considering QoS in place, which is widely accepted by the users and able to generate a significant revenue.

Currently, concepts on how to describe, define, and detect IP service usage at a finer grain and finally how to charge for this kind of service usage in an efficient manner is still an open issue in both the wired Internet and even more in the future wireless packet-based mobile Internet. Pursuing this goal, but often focusing on different aspects, several working groups in the IETF and other consortia such as 3GPP/2 or MWIF have identified both the key concepts and the missing components and have proposed complementary and sometimes competitive approaches. Along with the IETF and IRTF AAA work, a metering architecture has been developed which provides a promising base for the missing functions and mechanisms. Although the basic mechanisms required are available and widely understood, their efficient and scalable integration is still an open point. This integration however, has to consider not only technology-driven aspects, but also the basic principles of a commercial network, mostly coming from marketing and other economic-based disciplines. Unfortunately the development of mechanisms and concepts has only started recently and lacks far behind the already available mechanisms to provide QoS. This leads to the unhappy situation of the QoS researchers missing a clear indication about which parameters and concepts will be charged for and thus must be supported by the network. The implementation of a QoS architecture itself would not be a problem. Finally it should be concluded that the deployment of QoS in a mobile and highly dynamic environment can not be treated as a stand-alone solution since the definition of the requirements for such a scenario is a complex issue and must respect both, potential technical mechanisms as well as non-technical correlations. So it seems that all required mechanisms are available, however there exists no clear vision how to combine them in order to satisfy both, network provider and end user.

5.3 Viability

In order to successfully deploy QoS architectures, not only the necessary technology must be available. Backwards compatibility with existing technology and migration strategies are equally important. Such plans have not been thoroughly developed, among other reasons due to the lack of consen-

[2]This is true independently of the level of reliability offered by the technology itself.

sus on the most probable QoS architecture. Especially in the first steps of QoS introduction, though, such plans are indispensable.

Networked applications require a certain quality level to be provided end-to-end. This is especially the case for the most sensitive ones, like interactive and real-time applications. Solutions that enhance some network segments may help provide a better service, but absolute guarantees, in the sense provided by the telephone network, ATM, or RSVP can only be achieved with a consistent end-to-end solution. This issue is further complicated by the problematic interaction among the several parties involved in a communication session (network and content providers, application, user). Nevertheless, consistent quality levels would strongly enhance the possibilities of QoS success.

A related issue in the quest for end-to-end coherence is that communication is, per definition, bi-directional. Hence, many networked applications have similar QoS requirements in both directions, others do not. Two problems arise: If two or more parties are involved, e.g., in a telephone conference, how is the overall quality level chosen, in the case of conflicting interests? Furthermore, the nature of IP data transport is unidirectional. This implies that the network does not establish any relationship between both directions of one and the same flow, let alone among several flows simultaneously active for the same communication. Filtering and/or signaling in the network and/or between network and application could help tackle this issue. Proxies and filters are extensively used nowadays for similar purposes without any kind of explicit signaling. To establish a relationship between flows or flow directions in modern multimedia applications, the decoding and interpretation of data contained in several protocols is needed (IP, TCP and application header and even payload). This is a computationally very intensive task that would have to be realized on a per-packet basis. A more scalable solution speaks for the use of signaling and/or flow descriptors to ease the filtering problem.

The independence between routing and QoS decisions in IP networks also influences the coherence of QoS-providing measures. Under certain circumstances, it can provoke the sending of data along a different path than the one chosen by the QoS decision-taker. This well-known issue can be tackled by "pinning" a flow to a route, once a QoS decision has been taken. Although it somewhat violates the IP routing approach, several proposals have shown its feasibility. GMPLS, e.g., completely integrates routing and QoS in a common framework, thus solving this issue.

A very important issue of general viability of any consistent end-to-end QoS is the ability to solve the problem of interaction among different (network and content) providers and possibly also the customers, all of which are involved in most communication relationships. The world-wide trend to open markets in telecommunications has also increased the problem of separate provisioning of network access, content and core transport. While in the past this has been addressed by offering transport, content and services through only one provider ("wholesale provider"), the current issue is to solve this interaction for many providers and an increasing number of services. This issue becomes especially critical, if it comes to traffic crossing many administrative domains between the origin and the destination.

While technically it is possible even now to do all that either in one network or for a certain service or in a controlled environment, and it is even possible to evaluate the advantages and disadvantages of different mechanisms in this type of simple scenario, the problem of interaction on a large scale has hardly been addressed. It is an extremely difficult task to define or even identify the necessary points of interaction and the necessary data and mapping of parameters without at the same time imposing restrictions on services and more generally constraining the independent evolution of service, content, and network providers.

5.4 Security

Providing QoS mechanisms in networks implies several critical issues regarding security. First of all, from an operational point of view there are questions of accountability and non-repudiation. Thinking of a user having a traffic contract with a network operator, the operator must have means to authenticate the user properly in order to prove, e.g., that a user has violated the contract. Moreover, the operator needs means to prove that he provided the service in the negotiated manner if, e.g., a user claims not to be treated fairly. The other way around, the user needs means to prove that he did not send more or different data than allowed. Furthermore, if no absolute guarantees on QoS are given, it can happen that users want to identify or even to determine network operators on the path of their packets in order to decide whether they will trust the providers or not. This is not a question of trust regarding user data — as these can be encrypted — but rather a question of trust in provided QoS, as the same class of service can mean different absolute QoS at different providers.

From a technological view, the control data of QoS mechanisms needs to be secured. On the one hand an attacker must not be able to change the control data (e.g., the class of service field) in order ,e.g., to flood an operator with prioritized traffic, as this could make network behavior even worse than without QoS mechanisms. Another possible attack would be the degradation of a packet resulting in worse QoS sensed by the user, thus causing damage to another operator's reputation. On the other hand, users may wish to protect the class of service against eavesdroppers in order not to disclose the type of traffic. When designing protection mechanisms, it must be taken into account that encryption is not necessarily always the best solution, since it takes significant amounts of time. This could degrade performance in high-speed core networks and increase complexity, e.g., due to a required Public Key Infrastructure (PKI). This could advise to use organizational protection means such as policies among providers (e.g., changing the packets' class of service is not allowed). In this case, mechanisms must exist to observe policies' violations and penalties must be enforced.

5.5 Mobility and QoS

The spreading of the diverse forms of mobility presents critical challenges for the consistent provision of QoS. There are several factors involved: First, mobility presents the user and the application with a quickly changing environment. On one side, the conditions of the network (load, loss ratio, etc.) change, e.g., due to a change of the physical channel or even of the infrastructure (ad-hoc networks). But on the other side, even the networks at the disposal of the user change. As stated in section 4, diverse access technologies will be present, especially in so-called hot-spots: From Blue-

tooth over 802.11 to UMTS and GSM. These technologies have very different, even incompatible QoS capabilities. The resulting dilemma is how to ensure a consistent QoS-level in such a scenario. A number of possibilities arise: Transactions could be delayed until a certain access technology is available, in order to fulfill the user's expectations. Or maybe a transaction could be interrupted and resumed afterwards when conditions change too drastically. But even with the help of such mechanisms, mobility possibly implies the necessity, especially for interactive applications, of having to accept a fluctuating QoS level or aborting the communication. Furthermore, some sort of mechanism is needed to choose among the networks at the user's disposal, especially in the face of conflicting trade-offs: E.g., Network A might be more economical, but Network B might deliver the data faster. Again, user-defined profiles and/or QoS-middleware present themselves as powerful candidates.

Another related question is the associated signaling burden. Generally, there is a trade-off between the overhead of the QoS signaling mechanism — and the resulting limitation on performance — and mobility to be solved. Whereas overhead is often discussed in the context of additional messages or bigger packets, in this context the timely delivery of signaling messages for mobility management is more important. When considering small access networks — or cells — and users on the move, the sojourn time in a network can be quite short. On each network change, mobility management has to signal the new network to the user's home network. If this update is delayed for too long, the user is no longer reachable. Thus, QoS messages that need to be exchanged before the exchange of mobility signaling, should not affect the user's reachability.

To achieve this goal in a highly dynamic environment in which the behavior of the user is not predictable is the crucial point. Talking about user behavior there are two overlaying issues to be considered. First, derived from the multiservice nature of the Internet, there is no signaling in place informing the network in advance about the kind of applications being used, and derived from this the relevant parameters for QoS provisioning, in order to meet the hard timely requirements coming from mobility management. The second even more important issue is the prediction of the user's movement. There probably exists a possibility to predict the movement of one mobile user, however, future communication scenario must be prepared to support QoS between two moving users.

6. CONCLUSIONS

Every user and every application has very different QoS requirements. It is not viable to provide enough resources for every possible need in every possible environment. This is especially true for access networks, as has been shown. Hence, QoS architectures have long been studied, in order to provide differentiated quality levels to different traffic proportions. Many proposals have emerged, which have had only very restricted success beyond the laboratories. The main reason is certainly not technical: Almost all proposals could be implemented with today's technology. It is the lack of attention to other, non-QoS-specific issues, which has slowed down the deployment of QoS. First among them is the lack of a commercialization model appealing to both (network and content) operators and users. In this context, accounting, charging and billing architectures are a missing crucial element. Another critical issue is the assurance of end-to-end QoS coherence in the face of multiple intervening parties with heterogeneous characteristics. Especially relevant for both operator and user is the ease of management of the QoS mechanisms in the face of such heterogeneity. The necessary mechanisms to signal QoS requirements and other constraints (like price thresholds) by the application and the user have also not been integrated sufficiently in an overall architecture. Only by solving such open issues could QoS achieve wide acceptance.

7. ACKNOWLEDGMENTS

Carlos Macián would like to thank Mr. Bernd Gloss at IKR for the various fruitful discussions during the preparation of this paper.

8. REFERENCES

[1] 3GPP. *Technical Specification TS 23.002*, v5.0.0: Network Architecture (Release 5), October 2000.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. RFC 2475, IETF, December 1998.

[3] J. Angelopoulos, D. Boettle, P. Christ, J. Jaehnert, H.-C. Leligou, and S. Wahl. Design and implementation of a DiffServ enabled HFC system offering strict QoS support. *European Transactions on Telecommunication Journal*, 6, 2002.

[4] S. Bodamer, K. Dolzer, M. Lorang, W. Payer, and R. Sigle. Scheduling and bandwidth allocation for flow aggregates in multi-service networks. Internal Report no. 30, Institute of Communication Networks and Computer Engineering, University of Stuttgart, August 1999.

[5] B. Bostica and M. Krampell. QUASIMODO - Summary of QUASIMODO findings on QoS. EURESCOM P906 Technical Information 6. Technical Report, EURESCOM, January 2001.

[6] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. A framework for QoS-based routing in the Internet. RFC 2386, IETF, August 1998.

[7] ITU-T. *Recommendation E.800: Terms and definitions related to quality of service and network performance including dependability*, August 1994.

[8] K. Dolzer, W. Payer. On aggregation strategies for multimedia traffic. In *Proceedings of the 1st Polish-German Teletraffic Symposium (PGTS 2000)*, Dresden, September 2000.

[9] D. O. Keck. Erkennung von Wechselwirkungen zwischen Mehrwertdiensten durch Analyse ihrer Konfigurationen und Protokollabläufe. Monographie, Institute of Communication Networks and Computer Engineering, University of Stuttgart, 2002.

[10] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. B. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, Oxford, September 2000.

[11] M. Liebsch, X. Pérez, R. Schmitz, A. Sarma, J. Jaehnert, S. Tessier, M. Wetterwald, and I. Soto. Solutions for IPv6-based mobility in the EU project Moby Dick. World Telecom Congress, Paris, September 2002.

[12] M. Lorang. Skalierbares Verkehrsmanagement für diensteintegrierende IP-Netze mit virtuellen Verbindungen und verbindungslosen Routen von Datagrammen. Monographie, Institute of Communication Networks and Computer Engineering, University of Stuttgart, 2002.

[13] R. Mejia. MSPs: This Year's Model?. *Network Magazine*, pages 69–72, CMP Media LLC, October 2001.

[14] K. Nahrstedt and J. Smith. The QoS Broker. *IEEE Multimedia*, 2(1), 1995.

[15] T. R. Schmidt. Adaptionsalgorithmen zur Erhöhung der Diensgüte verteilter interaktiver Multimedia-Anwendungen in IP-basierten Netzen. submitted as PhD thesis, Institute of Communication Networks and Computer Engineering, University of Stuttgart, 2002.

[16] C. Shapiro and H. R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. ISBN 0-87584-863-X. Harvard Business School Press, 1999.

[17] J. D. Touch, editor. Theme Issue: Overlay Networks. *Computer Networks*, 36(2-3), July 2001.

[18] H. Zhang. Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE*, 83(10), October 1995.