VERTICAL QoS MAPPING OVER WIRELESS INTERFACES

MARIO MARCHESE AND MAURIZIO MONGELLI, UNIVERSITY OF GENOA

ABSTRACT
This article focuses on the problem of quality of service mapping between layers in a cascade. Protocol stacks in telecommunications networks are composed of functional layers. QoS provision depends on the performance achieved at each layer and is based on functions performed at layer interfaces. In practice, QoS derives from reliable physical and link layers that can offer specific transport services to upper network layers. The data flows (or bundles of flows) generated by the upper layers (e.g., the network layers) are forwarded down to a physical interface that transports the information along a channel that provides, if possible, the expected QoS to the upper layers. The action is called vertical QoS mapping and poses many challenges for a communication scientist, in particular if it is applied to wireless interfaces.

This article states the definition of vertical QoS mapping, proposes a formal separation between technology-dependent and technology-independent layers, models each functional layer as a battery of buffers, generalizes the relation between layers through a chain of buffers in a cascade, formulates the theoretical problems of vertical QoS mapping, and suggests possible solutions that use dynamic bandwidth allocation schemes.

INTRODUCTION
Modern telecommunication networks are composed of different portions and technologies. The challenge is to offer end-to-end QoS guarantees transparently over these heterogeneous networks to users.

Each layer also should have a specific role in QoS provision interacting with the adjacent layers. The overall problem of QoS interworking can be structured into two different actions [1], which together match the issues mentioned previously: horizontal QoS mapping (points 1 and 2) and vertical QoS mapping (point 3). The former, even if often linked to vertical QoS mapping when implemented in the field, is represented by the need to transfer QoS requirements among network portions implementing their own technologies and protocols. Horizontal QoS mapping is heavily linked to signaling and is beyond the scope of this article. Vertical QoS mapping is the subject of this article. The overall result depends on the QoS achieved at each layer of the network, and it is based on the services offered at the layer interfaces. The idea is to define an interface between adjacent layers through which it is possible to offer a specific QoS service. For example, if layer 3 implements efficient QoS mechanisms, it is typical that layer 2 assures a given service to layer 3; otherwise the implementation of complex QoS mechanisms at layer 3 is useless. QoS requirements flow vertically and must be received, understood, and satisfied by the layer below. For example, to guarantee the requested QoS at layer 3, given the bandwidth allocated at that layer, the bandwidth must be allocated at layer 2 to guarantee the required service to layer 3. Figure 1 shows the interface between layers 2 and 3 and also the possible enlargement of the allocated bandwidth so as to satisfy the performance requirements. The vertical interaction between layers in a cascade is defined as vertical QoS mapping. It leads to technological problems that are open standardization and research areas.

The first action is to establish a QoS-oriented formal relation between layers in cascade. A very good example is the protocol architecture proposed by the European Telecommunications Standards Institute (ETSI) [2] for the access points to a broadband satellite multimedia (BSM) network portion. The protocol stack for the upper layers is the Transmission Control Protocol/Internet Protocol (TCP/IP) suite, used as a reference in this article. The satellite-dependent (SD) layers (i.e., satellite physical, MAC and link control, strictly satellite dependent) are isolated from the satellite-independent (SI) layers (e.g., IP and the upper layers) by a satellite independent-service access point (SI-SAP), which offers specific QoS services.
The reference environment is the ETSI BSM protocol architecture, which has been studied for satellite communication but has a wider application. The ETSI BSM Technical Committee has expended great effort concerning QoS architectures, and the obtained results also can be applied to other environments.

Figure 1. QoS mapping action.

The aims of the TI-SAP are:
- To establish a common interface through the separation of technology-dependent (TD) and technology-independent (TI) layers and the definition of a generic interface called technology-independent-service access point (TI-SAP).
- To assure a QoS-based service. Currently, the architecture is formalized for satellites but can be generalized to include different physical supports, both wired and wireless. Actually, the idea is to extend the concept of functional independence between physical interfaces and upper layers through the separation of technology-dependent (TD) and technology-independent (TI) layers and the definition of a generic interface called technology independent-service access point (TI-SAP).

The interface is illustrated in Fig. 1. Obviously, a TD layer can satisfy or not satisfy the QoS requests of TI layers through specific actions, but this should be done transparently to the TI layers through dynamic bandwidth adaptation methods. The description of the vertical QoS mapping as a chain of buffers in a cascade of wide area networks (WANs). Figure 2 illustrates an example to help in understanding the real meaning of the TI-SAP. The wireless portion is located in the middle between two generic WANs called WAN1 and WAN2.

Although the technological examples may help in understanding, it is important to state that vertical QoS mapping is not limited to this illustration but has a general application when there are two or more layers in a cascade, and the lower layer must offer a QoS guarantee to the upper layer. For instance, this occurs within each single QoS-oriented network node, as shown in Fig. 3 without making any reference to specific physical-transport technology. This is one of the ideas that we wish to communicate. In addition to the technological application, the general concepts of QoS mapping described here also can be applied in a broader environment where layers act in a cascade.

**The Technology-Independent Service Access Point**

As mentioned in the introduction, the reference environment is the ETSI BSM protocol architecture, which has been studied for satellite communication but has a wider application. The ETSI BSM Technical Committee (TC) has expended great effort concerning QoS architectures, and the obtained results also can be applied to other environments. Much of the material contained in this article was inspired by the ETSI BSM studies contained in [3–5] and to the book documented in [1]. Reference [3] defines the QoS functional architecture. Reference [4] is an open specification to enable QoS services over IP-based multimedia systems with reference to the integrated services (IntServ) paradigm. Its focus is on the mapping of IP QoS functionalities over satellite-specific QoS features. In the same general QoS context, [5] refers to differentiated services (Diffserv) and specifies QoS mapping between Diffserv code points (DSCPs) and SD services.

This section generalizes the vertical QoS mapping issue without ignoring the large amount of work developed within the ETSI BSM TC.

The layers whose features depend on the technology are identified here as TD (SD, in the BSM architecture), whereas the layers whose characteristics are independent of the transport technology are classified as TI (SI, in the BSM architecture). The interface between them is defined as the TI-SAP, which is the SI-SAP in the ETSI BSM architecture.

Figure 2 shows the localization of the TI-SAP within a wireless portion that is part of an overall IP-based heterogeneous network composed of wide area networks (WANs). Figure 2 illustrates an example to help in understanding the real meaning of the TI-SAP. The wireless portion is located in the middle between two generic WANs called WAN1 and WAN2.

**TI-SAP Model**

It is very important to have a model to describe the TI-SAP or, more generically, the action of vertical QoS mapping that the TI-SAP helps formalize. The
The idea is to model the TI and TD layers, as well as the TI-SAP, through blocks of queues, in a manner that is similar to the modeling for the SI-SAP in [3], where each single SI-SAP queue is individuated by a specific identifier called a queue identifier (QID). Each identifier represents an abstract queue available at the TI-SAP interface and identifies a specific QoS level that is used to transfer packets from the TI to the TD layer. Vertical QoS mapping is described by using a chain of queues in a cascade. TD layers are responsible for assigning the required capacity to each abstract queue so that the QoS levels, which the TI layers must receive from the TD layers, can be satisfied. Using the concept of abstract queues at the TI-SAP allows the decoupling of the mapping problem into two different problems: TI queues mapped over abstract queues and abstract queues mapped over TD queues. The assumption about the existence of a battery of buffers at the TI-SAP, actually modeling the TI-SAP, is very important. It means that any network node where vertical QoS mapping is implemented provides different levels of QoS within a transport bearer through a number of queues that govern the service offered by the TI-SAP to upper layers. The TI layers can access and modify the abstract queues so as to request and possibly receive a different QoS service.

Figure 4 shows the vertical-QoS-mapping cascade of queues model and the related control modules. The number of queues should be large enough to support the desired QoS. The abstract queues are the means to define the TI-SAP independently of the real technology used at the TD layer. So the abstract queues are fundamental from a technological viewpoint. Nevertheless, from a strictly theoretical point of view, the problems are related to the existence of a cascade of queues and do not change if TI queues are mapped directly on TD queues. As assumed in the following sections of this article.

The queuing architecture shown in Fig. 4 must be connected to the related control plane. Following the lines drawn in [4] and [5], the control plane is composed of the following control blocks:

- TI resource management entity, which allocates and manages the resources at the TI layer (IP, in this article).
- TD resource management entity, which physi-
cally allocates the required resources at the TD layer. It can work in parallel with a local bandwidth allocator, which is often called the network control center (NCC) and is used widely in the case of radio and satellite networks where the bandwidth to be allocated is shared among different remote stations.
- QoS mapping management entity, which receives the resource allocation requests from the TI resource management entity. A request may concern resource reservation, release, and modification. The communication between the QoS mapping management and the TI resource management entity is established through a proper interface and by a group of primitives. A proposal for satellite interfaces (SI-SAPs) is provided in [4], and a proposal for TI-SAPs is contained in [1]. After receiving a resource allocation request, the QoS mapping management entity maps it on the lower layer. In other words, it translates the request into reservation, release, and modification actions to be applied at the TD layer.

The QoS mapping action includes the use of control algorithms and addresses different embedded problems, described explicitly in the next section.
In this framework, there are three problems arising from the action of layers in a cascade. The first two may be applied generically. The third one is related to the time-varying-channels characteristic of wireless links.

**VERTICAL QoS MAPPING PROBLEMS**

After identifying the technological tools to establish communication between the TI and TD layers, now we explain the problems of QoS mapping between two (or more) layers in a cascade and generalize the concepts further. Abstract queues, for example, are very powerful tools in practice because they allow for the decoupling of the QoS mapping problem; however, they can be ignored without affecting the nature of the problem, as mentioned regarding the previous actions. This introduces a direct QoS mapping between the TI and TD layers. In other words, both the direct TI over TD mapping and the two decoupled problems can be modeled as queues in a cascade. The former is used in the reminder of the article. In this framework, there are three problems arising from the action of layers in a cascade. The first two may be applied generically. The third one is related to the time-varying-channels characteristic of wireless links.

**CHANGE OF INFORMATION UNIT**

The consequence of a TI-traffic transport over a TD portion implements a specific technology. At each layer, the information coming from the upper layer, which is called a service data unit (SDU), is encapsulated within a new frame composed of the SDU and of a header and/or trailer (called protocol control information [PCI]). The newly created entity is called a protocol data unit (PDU). This means that the TI-layer packet accesses the TD queue after being encapsulated in a new frame. It is intuitive that the service bandwidth provided at the TD layer must consider the additional bits of the header/trailer to retain a fixed level of service. The imposed additional information is called overhead.

**HETEROGENEOUS TRAFFIC AGGREGATION**

As outlined for satellite communications in [6], “it is accepted in the BSM industry that at the IP level (above the SI-SAP interface) between 4 and 16 queues are manageable for different IP classes. Below the SI-SAP these classes can further be mapped into the satellite dependent priorities within the BSM which can be from 2 to 4 generally.” More generically, the number of queues implemented within each single layer can vary. Typically the number of queues decreases from the upper to the lower layers due to efficiency and speed requirements. This means that traffic might be required to be aggregated when it flows down from one layer to the adjacent one. Consequently, the bandwidth at the TD layer must be adapted.

**FADING EFFECT**

Finally, but importantly, many transmission environments, such as satellite and wireless links, must handle time-varying-channel conditions due to fading.

**JOINT PROBLEMS**

The three problems presented previously can be viewed jointly. The overall buffer model is shown in Fig. 5 with three buffers at the TI layer (h, i, and j) and one TD layer. The bandwidth assigned to each buffer, so as to provide a given QoS to the flows entering the buffer, is identified as $R_{id}^{TI}$ at the TI layer (where $id = h, i, j$ identifies the buffer, in Fig. 5) and as $R_{TD}^T$ at the TD layer. The fading effect can be modeled as a reduction of the bandwidth actually “seen” by the TD buffer through a multiplicative stochastic process whose values are real numbers ranging from 0 (total outage) to 1 (free error channel).
In other words, the real bandwidth that serves the TD buffer is decreased over the time of a stochastic variable contained between 0 and 1.

The model can be iterated if there are more than two layers in a cascade. The key point is bandwidth adaptation, which is very challenging, both from a theoretical and a practical viewpoint. Figure 5 illustrates how to dimension the bandwidth \( R_{TD}(t) \) at the TD layer so that the service is transparently guaranteed to the TI-layer queues. Bandwidth allocation is treated widely in the literature. Most of the schemes are based on the concept of equivalent bandwidth (EqB), which is defined as the minimum service rate to be provided to a traffic buffer to guarantee a certain degree of QoS in terms of objective parameters (e.g., packet loss, delay, jitter). EqB techniques are usually obtained analytically for homogeneous traffic trunks; concerning both traffic characteristics and QoS requirements, they use a single QoS constraint and are heavily based on the knowledge of traffic features, which are mathematically modeled. The complexity of the overall input flow process that enters the TD layer in the vertical QoS mapping model described previously makes the bandwidth allocation algorithms that use mathematical models of the flow process almost non-applicable. The flow that accesses the TD queue comes from the actions of format change and traffic aggregation, which implies the generation of heterogeneous trunks from the point of view of both traffic sources and QoS requirements. These actions modify the original features of the flows that entered the TI layer. In other words, the resulting flow is so complex that it hardly can be modeled analytically. Additionally, fading can affect the overall bandwidth availability, making it variable over time. The same observations are still true when applied to a generic layer within a layer in the cascade model, which allows extending the scope and generality of this section.

In this context, it is clear that bandwidth variability and problem complexity require automatic bandwidth management through proper control schemes, which do not use a priori information about traffic flows and buffers, as well as closed-form expressions of involved quantities. Also, other solutions based on off-line measures and on bandwidth overprovision can give good results, but the dynamics of the allocation can improve the performance. The problem is stimulating from a scientific point of view.

The next section presents a possible reference scheme in this framework for dynamic bandwidth allocation at the TI-SAP and a solution from the current research of the authors, which seems to be well suited for the described environment.

**Reference Scheme for Dynamic QoS Mapping over TI-SAP Interface**

The basic concept is to allocate the bandwidth periodically at the TD layer after receiving the QoS constraints through the TI-SAP primitives. This can be applied generically to any decisional scheme. The bandwidth is decided at fixed-time instants — the bandwidth \( R_{TD}(t_k) \) allocated at the instant \( t_k \) might depend on the bandwidth allocated at previous instants, up to a given depth of the allocation-scheme memory and on an information vector. The latter can be composed of information about the TD buffer or simply be the error \( e(t_k) \) as here, which is defined as the difference between the minimum bandwidth that guarantees the QoS constraints in the interval \([t_{k-1}, t_k)\] that is known at \( t_k \) and the bandwidth allocated at \( t_{k-1} \), which has given origin to the performance in the interval \([t_{k-1}, t_k)\]. In practice, \( e(t_k) \), if above 0, is the minimum additional amount of bandwidth that would have enabled the satisfaction of QoS constraints in the interval \([t_{k-1}, t_k)\]; if below 0, \( e(t_k) \) is the over-provisioned bandwidth, that is, the maximum amount of bandwidth that can be dropped without violating QoS constraints. A meaningful subclass is represented by the schemes where the memory depth is one. In short, \( F(\cdot) \) being a generic function, a possible generic representation of the allocated bandwidth is: \( R_{TD}^{F}(t_k) = F(R_{TD}^{TD}(t_{k-1}), e(t_k)) \). A corresponding bandwidth allocation update is reported in (1), where \( w_k \) is a weight described in the following.

\[
R_{TD}^{F}(t_k) = R_{TD}^{TD}(t_{k-1}) + w_k \cdot e(t_k) \quad (1)
\]
If the requirement is that the bandwidth allocation algorithm does not use a priori information about traffic statistical properties, any assumption about buffer dimensions, and any closed-form expression of the involved variables, a possible solution is to use only measures of the ongoing processes. For example, if there is only one performance metric (e.g., the information loss; however, more metrics can be used such as delay and jitter; also metrics can be used jointly) with its corresponding performance threshold, it is possible to measure the number of arrived and lost bits at the TD-layer buffer during a given interval to:

- Compute the loss rate that can be tolerated
- Check the bandwidth underprovision or over-provision (i.e., the mentioned error $e(t_k)$), which is an estimation of the bandwidth requirement
- Allocate the bandwidth in the next interval consequently

The weight $w_k$ acts either as a reducer or as an amplifier of the bandwidth requirement estimation and can be dynamic. A numerical example might help explain: the performance requirement flowing through the TI-SAP by a proper primitive is that only 1 percent of the information packets can be lost; the time interval $[t_k - t_{k-1}]$ is 10 s; the number of arrived packets in the interval is $2 \times 10^5$; the number of measured lost packets is $3 \times 10^3$. As a consequence, the packet loss rate is 300 packets/s, whereas the packet loss rate that can be tolerated is 200 packets/s. The difference between the two quantities is 100 packet/s, which is the missing bandwidth, that is, the additional amount of bandwidth that would have allowed the matching of the performance constraint in the interval $[t_k - t_{k-1}]$. It does not mean that the allocated additional bandwidth must necessarily be 100 packet/s. It could be less or more, depending on the applied control strategy. The key parameter is $w_k$. Its computation can be very complex and dependent on additional information such as the previously mentioned information vector. The authors of this article have proposed a control scheme, which uses the sensitivity of the system performance (e.g., of the information loss in the case presented above) to variations of the allocated bandwidth to decide the value of the weight $w_k$ dynamically over time. It is called a reference chaser bandwidth controller (RCBC), and the results in the following section are obtained from it. The details can be found in [1, 7]. What is important here is to focus on the requirement to have a bandwidth allocation scheme that acts only with local knowledge (e.g., measures of the arrived and lost traffic at the TD layer) and also reacts to traffic and channel variations. The example results, reported in the next section, aim to show evidence of these characteristics.

**Example Results**

An aggregate trunk of 50 voice over IP (VoIP) on-off sources composes the traffic at the TI layer flowing through the TI-SAP interface. Asynchronous transport mode (ATM) encapsulation is used at the TD layer. There is only one IP queue and one ATM queue (e.g., expedited forwarding...
Packet loss and average packet delay are the two performance metrics for the VoIP traffic. The performance thresholds are set to: $2 \times 10^{-2}$ for the packet loss and $20 \text{ ms}$ for the packet delay. The buffer size is $1600 \text{ bytes}$ ($20 \text{ VoIP packets}$) at the TI layer and $3710 \text{ bytes}$ ($70 \text{ ATM cells}$) at the TD layer. The bandwidth is reallocated every minute. Time-varying-channel degradation affects the TD buffer service rate. The reduction factor for process is modeled through the process $\phi(t) = \{0.0, 0.15625, 0.3125, 0.625, 0.8333, 1.0\}$. The values come from [8]. Figure 6 shows the behavior of the reduction factor over time in part (a); the packet loss probability (PLP) at the TI and TD layers, together with the loss performance threshold $2 \times 10^{-2}$, in part (b); the average delay (AD) at the TI and TD layers, together with the delay performance threshold, $20 \text{ ms}$, in part (c); and the bandwidth allocation at the TI and TD layers by using RCBC in part (d). The process generates peaks of channel degradation, especially in the time interval $[4800, 6000]$. There are only four peaks of performance degradation (where PLP and AD are above the QoS thresholds) and only in correspondence of the reduction factor changes. It is important to note the quick reaction and bandwidth adaptation to fading variations. Figure 6 highlights the dynamic behavior of bandwidth allocation and its ability to track high time-varying targets. Because the method is based on real measures, the length of the reallocation period is important. On the one hand, it must be long enough to filter that the measured metrics are representative of average values with a sufficient degree of confidence; but on the other hand, it must be short enough to assure quick reaction to traffic and network changes. The solution is represented by a compromise that in this article is reached through heuristics. Further research can be dedicated to find the ideal compromise using analytical tools. From a numerical viewpoint, by averaging the confidence interval measured for each reallocation decision over the overall test duration, the shown loss values have an average error of 24 percent for 95 percent of the cases. Similar values were obtained for the delay.

**CONCLUSIONS**

This article introduces the problem of QoS mapping over wireless interfaces through the presentation of the vertical QoS mapping concept and, operatively, through the definition and localization of a proper interface, called the TI-SAP. The article proposes a queuing model to describe the TI-SAP as a cascade of buffers and states the main problems of QoS mapping (change of information unit, heterogeneous aggregation, and fading effect) by using this queuing model. The queues in the cascade model enable the description of the three issues as a joint problem that implies the solution of a dynamic bandwidth allocation problem, which is currently an object of research. It seems recommendable to use dynamic schemes based on measures that can react quickly to changes in traffic and performance parameters and that do not make use of complex mathematical traffic models, often unsuitable for real network conditions.

Future research will be dedicated to investigating implementation details that are topical for the practical development of the bandwidth adaptation mechanisms such as the length of the reallocation period and the dynamic variation of the weight $w_j$ over time. The final aim is to obtain a real prototype that implements the RCBC within a TI-SAP-based architecture.

**REFERENCES**


**BIOGRAPHIES**

**MARIO MARCHES** [‘594, M’97, SM’04] (mario.marchese@unige.it) 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 Ph.D. (Italian “Dottorato di Ricerca”) degree in telecommunications at the University of Genoa in 1996. From 1999 to 2004, he worked with the Italian Consortium of Telecommunications (CNT), by the University of Genoa Research Unit, where he was Head of Research. From February 2005 he has been Associate Professor at the Department of Communication, Computer and Systems Science (DIST), University of Genoa, Italy. He is the founder and still the technical responsible of CNIT/DIST Satellite Communications and Networking Laboratory (SCNL) by the University of Genoa. He chaired the IEEE Satellite and Space Communications Technical Committee from 2006 to 2008. He is author and co-author of about 200 scientific works, including international magazines, international conferences, and book chapters, and the book Quality of Service over Heterogeneous Networks, John Wiley & Sons, 2007. His main research activity concerns satellite and radio networks, transport layer over satellite and wireless networks, quality of service and data transport over heterogeneous networks, and simulation and validation of telecommunication networks and satellite components.

**MAURO MONGEL** 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 Ph.D. (Italian “Dottorato di Ricerca”) degree in electronic and computer engineering at the University of Genoa in 2004. His Ph.D. was funded by Selex Communications S.p.A., Genoa, Italy. He worked for both Selex Communications and the Italian Consortium of Telecommunications (CNT), by the University of Genoa Research Unit from 2000 to 2004. He is now a member of the research staff of the Telecommunication Networking Research Group by the Department of Communication, Computer and Systems Science (DIST), University of Genoa, Italy, with a post-doctoral scholarship funded by Selex Communications S.p.A. His main research activity concerns are QoS architectures, resource allocation, and optimization algorithms for telecommunication systems.