Quality of Service in DiffServ IP-based Networks

Zoran Bojković, Bojan Bakmaz
Faculty of transport and traffic engineering
University of Belgrade
Vojvode Stepe 305, Belgrade
SERBIA AND MONTENEGRO

Abstract: This paper deals with the problem of quality of service (QoS) in DiffServ IP-based networks. A short review open system interconnection (OSI) framework is given. Then we focus on QoS from provider's and costumer's point of view, mainly based on DiffServ concepts. Also, we discuss QoS in IP networks. These networks promise convenient QoS-enabled communication. Finally, we analyze DiffServ in IP networks with wide support for QoS and architectures concerning service management and traffic engineering functions.

Key-Words: Quality of service, DiffServ, Internet Protocol, Provider, Interaction

1 Introduction

Meeting quality of service (QoS) guarantees in multimedia systems is an end-to-end issue that is from application to application. A key observation is that for applications relying on the transfer of multimedia and, in particular, continuous media flows, it is essential that quality of service be configurable, predictable and maintainable system-wide, including the end-system devices, communication subsystem and networks. Furthermore, it is also important that all end-to-end elements of distributed systems architecture work in unison to achieve the desired application-level behavior A number of QoS principles motivate the design of a generalized QoS framework:

- The transparency principle state that applications should be shielded from the complexity of underlying QoS specification and QoS management. The benefits of transparency are that it reduces the need to embed functionality in the application, hides the detail of underlying service specification from the application, and delegates the complexity of handling QoS management activities to the underlying framework.

- The integration principle states that QoS must be configurable, predictable, and maintainable over all architectural layers to meet end-to-end QoS [1]. Flows traverse resource modules (e.g., CPU, memory, multimedia devices network, etc.) at each layer from source media devices, down through the source protocol stack, across the networks, up through the receiver protocol stack to the playout devices. Each resource module traversed must provide QoS configurability (based on a QoS specification), resource guarantees (provided by QoS control mechanisms), and maintenance of ongoing flows.

- The separation principle states that media transfer, control, and management are functionally distinct architectural activities. The principle states that these tasks should be separated in architectural QoS frameworks. One aspect of this separation is the distinction between signaling and media transfer. Flows (which are isochronous in nature) generally require a wide variety of half-bandwidth, low-latency, nonassured services with some form of jitter correction. On the other hand, signaling (which is full duplex and asynchronous in nature) generally requires low-bandwidth, assured-type services [2].

- The multiple timescales principle guides the division of functionality between architectural modules and pertains to the modeling of control and management mechanisms. It is necessitated by, and is a direct consequence of, fundamental time constraints that operate in parallel between resource management activities (e.g., scheduling, flow control, routing, and QoS management) in distributed communications environments.

After a short overview of open system interconnection (OSI) quality of service (QoS) framework, we will deal with provider's and costumer's requirements. Next, we will continue with QoS in IP networks mainly based on DiffServ in IP networks with wide support for QoS and the corresponding architecture.

2 Open System Interconnection (OSI) QoS Framework

One early contribution to the QoS-driven architecture is the QoS framework, which concentrates primarily on quality of service support for OSI communications. The OSI framework broadly defines terminology and concepts for QoS and provides a model which identifies objects of interest to QoS in open system standards. The QoS associated with objects and their interactions is described through the definition of a set of QoS characteristics. The key OSI QoS framework concepts include the following:

- QoS requirements, which are realized through QoS management and maintenance entities;
- QoS characteristics, which are a description of the fundamental measures of QoS that have to be managed;
- QoS categories, which represent a policy governing a group of QoS requirements specific to a particular environment such as time-critical communications; and
- QoS management functions, which can be combined in various ways and applied to various QoS characteristics in order to meet QoS requirements.

Block schema of the OSI QoS framework is illustrated in Figure 1. This framework is made up of two types of management entities, layer-specific and system-wide entities, that attempt to meet the QoS monitoring requirements by monitoring, maintaining, and controlling end-to-end QoS.
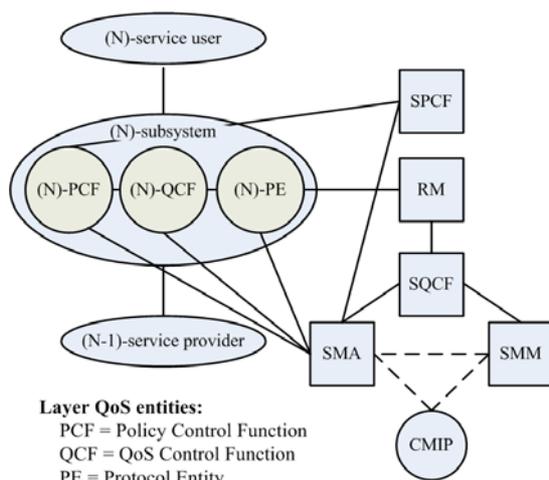
Layer QoS entities:
PCF = Policy Control Function
QCF = QoS Control Function
PE = Protocol Entity

System QoS entities:
SPCF = System Policy Control Function
RM = Resource Manager
SQCF = System QoS Control Function
SMA = System Management Agent
SMM = System Management Manager
CMIP = Common Management Information Protocol

Figure 1. OSI QoS framework

The task of the policy control function is to determine the policy that applies at a specific layer of an open system. The policy control function models any priority actions that must be performed to control the operation of a layer. The definition of a particular policy is layer specific and therefore cannot be generalized. Policy may, however, include aspects of security, time-critical communications, and resource control. The role of the QoS control function is to determine, select, and configure the appropriate protocol entities to meet layer-specific QoS goals. The system management agent is used in conjunction with OSI systems management protocols to enable system resources to be remotely managed. The local resource manager represents end-system control of resources. The system QoS control function combines two system-wide capabilities: to tune performance of protocol entities and to modify the capability of remote system via OSI systems management. The OSI systems management interface is supported by the systems management manager, which provides a standard interface to monitor, control, and manage end systems. The system policy control function interacts with each layer-specific policy control function to provide an overall selection of QoS functions and facilities.

3 QoS from Provider's and Customer's Point of View

The existing definition of QoS lacks the clarity required to express separately the service provider's and customer's viewpoints, as illustrated in Figure 2. QoS required by the customer is a statement of the level of quality of a particular service required or preferred by the customer. The level of quality may be expressed by the customer in technical or non technical language. A typical customer is not concerned with how a particular service is provided or with any of the aspects of the network's internal design, but only with the resulting end-to-end service quality. It must be recognized that the customer's QoS requirements can be sometimes subjective. These requirements are useful, although subjective. It is up to the service provider to translate them into something of objective use [3].

QoS offered by the service provider is a statement of the level of quality that is offered to the customer. This is the level of service that the service provider can achieve with the designed of the network. The level of quality is expressed by values assigned to network performance parameters, which cover the network and network support [4].

QoS achieved by the service provider is a statement of the level of quality achieved by the service provider. It is record of the level of quality that have been achieved. These are expressed by values assigned to the parameters specified for the offered QoS. These performance values are summarized for specified periods; for example, for the previous three months and/or an annual basis.
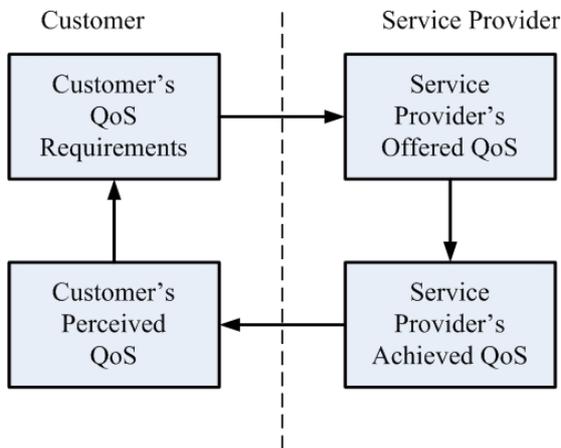
Figure 2. The four QoS viewpoints

QoS perceived by the customer is a statement expressing the level of quality experienced by the customer. The perceived QoS is expressed usually in degree of satisfaction and not in technical terms. The perceived QoS is accessed by various methods, including customer's surveys, customer's comments, and customer's complaints. Figure 3 shows how the various QoS viewpoints interrelate with one another. The service provider and the network provider need not always be the network provider.
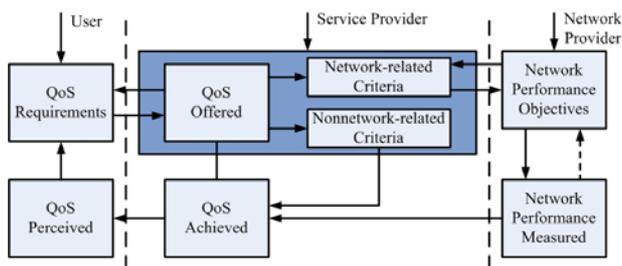
Figure 3. Various QoS viewpoints: intrarelationships

The services provider must always take full responsibility for the QoS offered to the customer. From the intrarelationships between the QoS viewpoints, it can be concluded that both the customer's and service provider's quality interest must be in a state of equilibrium in order to have successful business relationships. Therefore, it is necessary to manage the activities and relationships

associated with QoS viewpoints to obtain the optimum quality levels in accordance with the price the customer is willing to pay.

The effect on end-to-end image quality of packet loss is not yet well defined. In early MPEG reference models, cell loss rates lower then 10^{-9} were proposed, but rates of 10^{-4} are currently being considered as acceptable. The effect of cell loss is not dependent only on the average cell loss rate, but also on the distribution of cell loss over time. Periods of high cell loss due to network congestion can have a serious detrimental impact on image quality [5].

Delay requirements vary depending on the application. For interactive video applications, a maximum end-to-end delay of some 100 ms is appropriate, while a much longer delay would be tolerable for a user simply watching a recorded clip or movie in a video playback application. Delay requirements have a strong impact on the type of the network service to be provided [6, 7]. For example, in the case of video playback, a large buffer in a settop box can absorb considerable variation in network delays of successive calls. On the other hand, the tight delay constraints for real-time communication limit the possibility of dealing with the congestion on network lines by cell buffering. Coding delays must be included in the overall delay budget, thus limiting the scope for rate smoothing in a closed-loop coder producing constant bit rate (CBR) output.

4 Quality of Service in IP Networks

The Internet Engineering Task Force (IETF) started working on providing the QoS in IP networks in the mid 1990s. Two different approaches have been introduced: integrated services (IntServ) in 1994 and differentiated services (DiffServ) in 1998. IntServ was introduced in IP networks in order to provide guaranteed and controlled services in addition to the already available best-effort service. It is an extension to the Internet architecture to support both non-real-time and real-time applications over IP.

IntServ and reservation protocols such as resource reservation protocol RSVP have failed to become an actual end-to-end QoS solution, mostly because of the scaling problems in large networks and because of the need to implement RSVP in all network elements from the source all the way to the destination.

DiffServ came to remedy the disadvantages of IntServ in providing QoS in IP networks. DiffServ aims to provide simple, scalable, and flexible service differentiation using a hierarchical model.

That is, resource management is now divided into two areas: interdomain and intradomain.

In DiffServ all of the costumer's local network requirements for QoS are aggregated, and then a service level agreement (SLA) is made with the network service provider. The SLA may be static (negotiated and agreed on a long-term basis, e.g., monthly) or dynamic, changing more frequently. The local network is then responsible for providing differentiated services to end users with the network. This is usually done through marking packets with specific flags used in the type of service field of IPv4 or the traffic class field of IPv6. Here IPv4 and IPv6 means Internet protocol version 4 and Internet protocol version 6, respectively.

Given the background information, we are ready to analyze causes of QoS problems. They can generally be divided into two categories: non-networked-related and network-related causes [8].

Non-network-related causes include:

Overloaded servers (e.g., Web or e-mail) users are trying to access: In this case, common ways to improve QoS are to upgrade the servers, or to add servers and use a better load-balancing scheme among them.

Network operation errors: Configuring routers/switches is a complex and error-prone process. For example, duplicate IP addresses can be mistakenly configured and cause routing problems.

Network-related causes include:

Equipment problems: Routers/switches are complex systems with sophisticated software and hardware that are required to process millions of packets per second. Equipment vendors are compelled to deliver products as early as possible. Therefore, it is not uncommon that routers/switches have hardware and software problems.

Lack of access capacity: For economic reasons, there are always customers with slow access links (e.g., dialup modems) or oversubscribed uplinks. The technical solution for this kind of problem is clear:

- Adding capacity
- Classifying traffic and marking it differently for subsequent treatment using policing, shaping, and so on. This will be further discussed in our approach.

  However, it should be pointed out that providing QoS may not make economic sense here if users are not willing to pay for it.

  Beyond the standardized functionality at the IP layer, a large body of work has been devoted to architectures and functions necessary to deliver end-to-end QoS. These functions can be categorized into traffic engineering (TE) functions and service management (SrvMgt) functions [9].

TE functions are mainly concerned with the management of network resources with the purpose of accommodating offered traffic in an optimal fashion. SrvMgt functions deal with the handling of customer service requests, trying to maximize incoming traffic, in terms of number of contracts and throughput, while respecting the service provider's (SP's) commitments. Also, SrvMgt needs to avoid overloading the network beyond loads it can gracefully sustain. SrvMgt functions that deal with the latter task are referred to as admission control and are the key focus of this article.

A hierarchical service model is adopted in Figure 4 which spans from service level agreements (SLAs) to per hop behaviors (PHBs), the basic QoS building blocks in IP DiffServ networks SLAs describe all aspects of a service contract. The technical aspects of a service contract are described by the so-called service level specifications (SLSs). For QoS-based IP connectivity services SLSs are modeled on the basis of standard templates. The service hierarchy introduces the notion of QoS classes to link SLSs with PHBs. QoS classes depict the elementary QoS transfer capabilities of an SP domain, consisting of an ordered aggregate (OA) and associated QoS parameters such as one-way delay and packet loss. Each service corresponds to a number of QoS classes. Therefore, given a service, its QoS is completely defined through the QoS classes of its constituent SLSs [10].
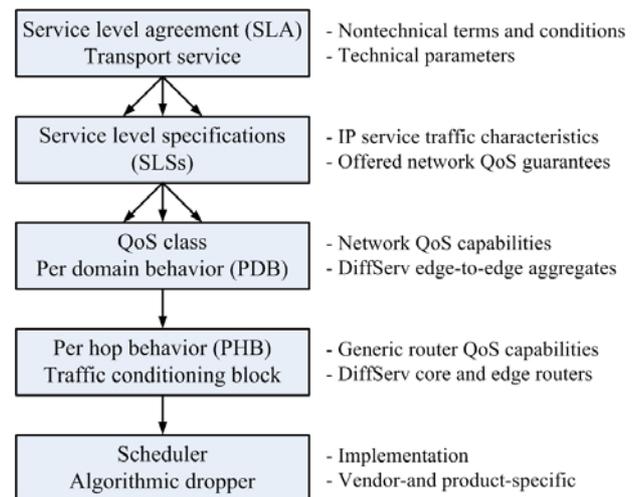
Figure 4. The hierarchy of the service model

5 DiffServ in IP Networks

The enlargement of the Internet user community has generated the need for IP-based applications requiring guaranteed quality of service (QoS) characteristics. The integrated services

(IntServ) and differentiated services (DiffServ) frameworks have been proposed to address QoS. While IntServ operates on a per-flow basis and hence provides a strong service model that enables strong per-flow QoS guarantees, it suffers from scalability problems. On the other hand, DiffServ keeps per-flow information only at the edge of a domain and aggregates flow into a limited set of traffic classes within the network, resolving the scalability problem at the expense of looser QoS guarantees [11].

Beyond the standardized functionality at the IP layer, a large body of work has been devoted to architectures and functions necessary to deliver end-to-end QoS. These functions can be categorized into traffic engineering (TE) functions and service management (SrvMgt) functions. TE functions are mainly concerned with the management of network resources with the purpose of accommodating offered traffic in an optimal fashion. SrvMgt functions deal with the handling of customer service requests, trying to maximize incoming traffic, in terms of number of contracts and throughput, while respecting the service provider's (SP's) commitments on the agreed QoS guarantees. SrvMgt mechanisms for service offering, agreement, and activation need to be in place. In addition, in order to guarantee the agreed QoS requirements, SrvMgt needs to avoid overloading the network beyond loads it can gracefully sustain. SrvMgt functions that deal with the latter task are referred to admission control [12].

SrvMgt and TE functions do not act in isolation as shown in Figure 5. TE functions provide the grounds on which the SrvMgt functions operate, while SrvMgt functions set the traffic-related objectives for the TE functions to fulfill. Specifically, SrvMgt establishes subscriptions based on which a Traffic Forecast function produces the traffic matrix (TM), which specifies anticipated QoS traffic demand between network edges. Traffic demand is forecast from historical data and/or SP expectations (e.g., sales targets). Based on the forecasted traffic demand, the network is appropriately dimensioned by the TE functions, in terms of PHB configuration parameters and QoS route constraints. In turn, TE functions produce the resource availability matrix (RAM), which specifies estimates of the availability of the engineered network to accommodate QoS traffic between network edges. Based on the availability estimates, SrvMgt functions, also utilizing network state information, handle the admission of service requests so as not to overload the network.

The above interactions occur in resource provisioning cycle (RPC) epochs. Should anticipated traffic demand significantly change, a new TM is produced, the network is appropriately redimensioned, the RAM is in turn produced, and new RPC starts.
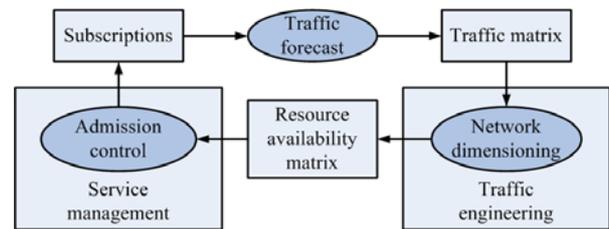
Figure 5. Resource provisioning cycle interactions between service management and traffic engineering functions

Admission control is an integral part of an IP QoS delivery solution. Hence:
- It is dependent on the informational model used to describe QoS-based services. From the perspective of admission control, service models should include specification of the QoS parameters as well as user traffic conformance for receiving the specified QoS.
- Its operation is bound to the QoS capabilities available in the network.

Furthermore, admission control should be policy-driven in order to adjust to the specific service provisioning strategies of the particular SP.

6 Conclusion
Traffic generated by the different services will not only increase traffic loads on the networks, but will also require different quality of service (QoS) requirements (e.g., cell loss rate, delay, and jitter) for different streams (e.g., video, voice, data). Delivering multiple QoS to different types of traffic while maintaining high utilization of the bandwidth is the objective of efficient traffic management, which encompasses technologies like call admission control, policing, scheduling, buffer management, and congestion control. DiffServ-aware traffic engineering can be used to prevent concentration of high priority traffic at any time so that high priority traffic will have low delay and jitter, and can be treated preferably at the expense of other classes of traffic if necessary. While DiffServ architecture solves the scalability problems of QoS provisioning it fails to be the solution for end-to-end provisioning. We expect this area to grow in the next years, with wider support for QoS and architectures by network operators.

Reference:

[1] A. T. Campell, et al., *Integrated quality of service for multimedia communications,* Proc. IEEE INFOCON'93, San Francisco, CA, pp. 732-739, April 1993.

[2] A. A. Lazar, *A real-time control management and information transport architecture for broadband networks,* Proc. Int. Zurich Sem. Digital Communications, pp. 281-295, May 1992.

[3] D. J. Wright, Assessment of alternative transport options for video distribution and retrieval over ATM in residential broadband, *IEEE Comm. Magazine*, vol. 35, pp. 78-87, Dec. 1997.

[4] N. Ohta, *Packet video: moduling and signal processing*, Norwood, MA: Artech House, 1994.

[5] M. Homidi, J. W. Roberts and P. Rolin, Rate control for VBR video coders in broad-band networks, *IEEE J. Selected Areas in Comm.*, vol.15, pp. 1040-1051, Aug. 1997.

[6] R. Steinmetz and K. Nahrstedt, *Multimedia computing, communications and applications*, Englewood Cliffs, NJ: Prentice-Hall, 1995.

[7] D.E. Mc Dysand and D. L. Spolin, ATM theory and applications, New York, NY: Mc Graw-Hill, 1995.

[8] X. Xiao et al., A practical approach for providing QoS in the Internet backbone, *IEEE Comm. Magazine*, vol. 40, pp. 56-62, Dec. 2002.

[9] K. R. Rao, Z. S. Bojković, D. A. Milovanović, *Multimedia communication systems*, Englewood Cliffs, NJ: Prentice-Holl, 2002.

[10] K. R. Rao, Z. S. Bojković, D. A. Milovanović, *Introduction to multimedia communications: applications, middleware, networking*, John Wiley, Hoboken, NJ, 2004.

[11] S. Blake et al., Admission control for providing QoS in diffserv IP networks: the TEQUILA approach, *IEEE Comm. Magazine*, vol. 41, pp. 38-44, Jan. 2003.