End-to-end QoS control architectures from a wholesale and retail perspective: benefits and challenges

Ulf Bodin, Olov Schelén, Claes Vemmervik
{ulf.bodin, olov.schelen, claes.vemmervik}@operax.com

Operax

Abstract

Providing an equal or better grade of service compared to legacy networks is essential for an NGN operator as insufficient service quality severely reduces the value of the network. Also, cost-effectiveness and rapid service deployment are two important properties making the NGN attractive to both operators and their customers. Recognizing the attractive properties of a competitive market, operators must further meet regulatory requirements on offering network services to independent retailers.

The Bandwidth Manager provides end-to-end multi-service QoS control that ensures sufficient grade of service and efficient network utilization. This control facilitates rapid creation of new service agreements. This paper describes the fundamental properties of the Bandwidth Manager in a retail and wholesale perspective. Benefits and challenges in deploying a Bandwidth Manager are also explored.

Introduction

The increasingly competitive environment in telecom is eroding operators' revenues and margins. Operators now turn to value added services as a way towards additional growth and increased customer retention. Operators face however threats from service providers like Real Networks™, SF-Anytime™ and Skype™, which offer services over the operator's best-effort networks without having to recover investments in infrastructure.

On the other hand, since operators control the network resources they have the power to provide guaranteed service performance as a differentiator. An operator can leverage this control to provide guaranteed QoS to its own retail business units and to independent retailers and service providers.

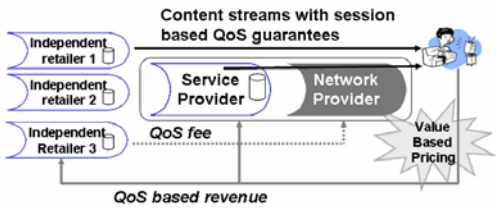

Figure 1 Multi-provider value chain

By exploring benefits of QoS control operators can convert the threat from independent service and content providers to an opportunity. Network owners can exploit their control over network resources by offering session based wholesale with guaranteed service delivery (Figure 1). In some markets the regulatory body requires this capability from the incumbent. The area of QoS control, which includes both policy control and bandwidth control, provides an important part of the service delivery architecture providing an efficient wholesale and retail solution for session based QoS control.

This paper describes a multi-service resource and admission control (RAC) architecture that enables end-to-end forwarding guaranties across multiple service providers and network operators. Throughout the paper RAC is used interchangeably with QoS control. We instantiate the RAC architecture in the form of a stand-alone RAC subsystem referred to as a Bandwidth Manager. The Bandwidth Manager follows and extends the RAC subsystem functional specification and its associated protocols that recently have been released by ETSI TISPAN as part of its NGN release 1 [7] and the RAC functions as of ITU-T [8].

Before describing the RAC architecture and its Bandwidth Manager instantiations, we discuss the role of RAC in a carrier-grade NGN and the relation to network provisioning. Benefits and challenges are discussed for providing end-to-end QoS, for supporting a wholesale and retail split of RAC, and in business models relying on network value-add QoS services.

Grade of service

Providing sufficient grade of service is essential for an NGN operator. The grade of service in an NGN must be equal to or better than legacy networks supporting applications such as PSTN, leased line, high-quality video streaming and video conferencing. If not, migrating to an NGN will be considerably less attractive to customers. Expensive discounts boosting the interest in migrating might be needed, which would severely weaken the operator's business case on moving customers to an NGN.

Another reason for avoiding periods of degraded service quality is that the operator's creditability as a premium carrier becomes damaged when expectations on service quality are not met. Hence, the key issues in

providing appropriate grade of service are costs associated with giving refunds and handling support calls as well as the overall creditability of the operator.

Reasons for degradations in service quality include network overload and service unavailability due to failing network or service delivery equipment. This means that proper network provisioning and adequate resilience for operational equipment are fundamental requirements in building a carrier-grade NGN.

A key question is whether or not proper network provisioning alone is sufficient to preserve operator creditability and to keep refund and support costs associated with network overload low enough. A problem of network overload is that users are likely to terminate their ongoing sessions when experiencing degraded forwarding quality and then immediately try to reinitiate these sessions. This behavior can extend and worsen periods of network overload and increase the number of users affected.

Many users being exposed to network problems at the same time will probably overload support centers. However, even more troublesome is when the users realize that many of them have suffered from degraded quality. When this happens it is highly likely that they will notify the public press about the service takeout, which make the problem end up in unwanted publicity, effectively and brutally working against any advertising campaign of the operator.

Network provisioning

Worst case dimensioning is extremely expensive as the network then need to be designed to handle the case where all users are simultaneously active with all the services to which they subscribe. Clearly, such scenarios are likely to be rare. Dimensioning based on a calculated risk for temporary overload is therefore the only viable option in practice. The question is more of the grade of how much excess capacity compared to the average need that should be provisioned.

Reasons for network overload include social extremes and synchronized user behavior resulting in unusually high network load and slightly uncommon usage patterns. Examples of events resulting in synchronized user behavior include new years eve/day and adverse weather. Media stimulated events such as televoting, and lotteries are other more predictable examples but that still can cause network overload. Disaster scenarios caused by environmental catastrophes (high rates of regular and emergency calls) are examples of unpredictable events.

Random failures reducing network capacity can not be completely ruled out, although a lot effort is put into making NGNs reliable. The resilience design of a carrier grade NGN should make it unlikely that terrorist attacks could take out both network equipment and make trunk capacity unavailable. However at some point, the available capacity of the NGN may be reduced and together with increased demand for network connectivity boosted by peoples' need to communicate in such a frightening situation, the risk of fatal network overload is evident.

Differentiated forwarding

A single service best-effort network cannot easily be over-dimensioned to meet requirements of different subscriber groups (e.g. residential and enterprise) and applications such as data, IPTV, conversational voice/video and gaming.

A fist step in trying to mitigate the dimensioning problem is to deploy class-based differentiation at the forwarding plane (e.g., DiffServ [3][4][5], IEEE 802.1p, ATM VP/VC). Further differentiation of resources can be provided by using bandwidth tunnels based on MPLS [6]. These methods make the dimensioning problem more tractable, but they do not protect from overload within the different classes of services that may occur due to social extremes, media stimulated events, unpredicted and random failures and disaster scenarios.

Using service differentiation in the forwarding plane is an important piece of the total solution. However, if it is taken to an extreme, where separation between all end user services and subscribers with special requirements, it creates operational overhead which grows with the number of services and with the size of the network. Most importantly, the need for dynamic reconfiguration due to dynamic business growth and changing user demand becomes a challenge with such extreme separation using service differentiation.

QoS control

As discussed in previous sections, QoS control, or Resource and Admission Control (RAC), constitutes a key function needed to offer strict guaranties on forwarding quality in the provisioned classes of service and bandwidth tunnels in the forwarding plane. Without admission control there is a risk that some resources become overloaded with failing QoS as a result.

Without QoS control (i.e. non-blocking networks), many users suffer the risk of being exposed to degraded and insufficient forwarding quality. With QoS control (i.e. networks capable of prioritizing or blocking sessions through admission control) sufficient quality is given to a maximum number of existing sessions and important arriving sessions in situations when the network would be otherwise overloaded without QoS control.

QoS control through call count

There are several approaches to QoS control. The usage of call counters within application frameworks such as

call servers or IMS signaling proxies constitutes a simplistic solution to QoS control that may be sufficient for initial deployments. This approach has however clear drawbacks with regard to operational cost, resource efficiency and support for multiple services and service providers.

A counter based QoS control solution requires a tight coupling to the resources provisioned for the service in question. For example, a mesh of MPLS bandwidth pipes can be provisioned in the network for a voice service. Call servers can then provide call counting into each of those pipes to protect them from overload. Application specific counters do however require quite complex and operational intensive network provisioning. Also they do not scale well to multiple services as separate call counters would have to be implemented in all application control devices.

Independent QoS control

The Bandwidth Manager is an independent system for QoS control. It replaces traditional QoS control functions that were vendor specific and tightly integrated with either application frameworks or network element management systems. To avoid vendor specific vertical solutions and to bring the full potential of a converged network, QoS control must be implemented as an stand-alone entity with standardized interfaces for inter-working with multiple application and session control entities, with multiple networks technologies and with multiple vendors (Figure 2).

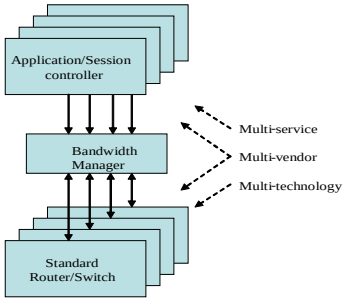

Figure 2 Independent QoS control

Separating QoS control from applications and session control entities removes the need for complex transport-related functions within each application and permits controlled sharing of network capacity between applications. This enables independent optimization of application and network architectures.

Standards for QoS control

The NGN and IMS define control plane functions in the network architecture. In Figure 3 we have categorized these functions into Service Control and Network Control, showing a high level picture. Service control functions based on SIP are supported in all standards. In NGN networks there are also other service control protocols. The network control functions include resource and admission control, service policy decision, and network attachment control (dynamic IP address mapping, subscriber line policies). There is an objective to converge the standards as much as possible.

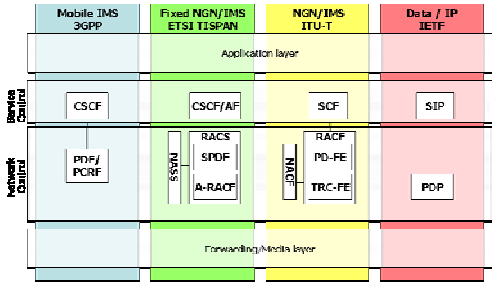

Figure 3 Standards for Next Generation Networks

A Bandwidth Manager implements the standards of the Network Control layer in Figure 3. Individual Bandwidth Managers may be targeted for a particular standard or cover interfaces/functions from all of them. Figure 4 shows the different functions of the Bandwidth Manager using ETSI TISPAN terminology (i.e. the Bandwidth Manager implement all functionality covered by RACS [7]).

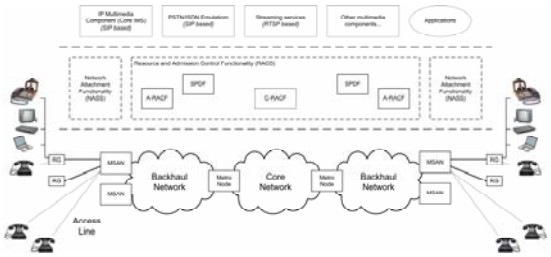

Figure 4 The Bandwidth Manager as RACS

The role of the Bandwidth Manager varies slightly depending on which service it handles. For IMS services there is a Home Subscriber Server (HSS) in the service control plane that handles user policies to support roaming users. For IMS services the role of the Bandwidth Manager is admission control based on service policies, subscriber line policies and transport network resources. For non IMS services the Bandwidth Manager may handle user policies as well.

Requesting resource reservations

Resource reservations need to be explicitly requested for QoS control to protect network resources from overload. However, the mechanism used to issue such resources may differ between networks and applications. The path-coupled and path-decoupled models represent two fundamentally different approaches to requesting resources, each approach with its own benefits and challenges. The differences

between these approaches have been discussed within the scope of the NSIS working group of the IETF [1].

In the path-decoupled model, applications and session control push admission requests to the Bandwidth Manager. In the path-coupled model forwarding level mechanisms pull admission requests concerning policies and subnet resources from the Bandwidth Manager. Consequently, the independent Bandwidth Manager can support both push and pull models. They may be used in different network domains respectively to provide an end-to-end solution.

When pushing for admission decisions, the Gq' DIAMETER application can be used [9]. It allows applications and session control entities to request resources on behalf of end-users, which only need to request an application service without bothering about network QoS.

A request for QoS made using Gq´ can be made on a per-session basis, or on a per-aggregate basis. In the latter case, the requestor makes one reservation to be used for an aggregate of sessions/flows as controlled by the requestor. Aggregate reservations can be used to reduce reservation signaling provided that the requestor is capable of grouping sessions into aggregates.

Distributed intra-operator QoS control

The Bandwidth Manager can be distributed over an arbitrary number of physical devices that may be geographically separated or located in computer blades on the same chassis. Additional hardware can thereby be added to the distributed Bandwidth Manager making it scale with increasing demands on performance for handling reservation requests and increasing network size. The distributed nature of the Bandwidth Manager is recognized by the Multi-Service Forum (MSF) [2].

A distributed Bandwidth Manager will not only communicate with service control frameworks and underlying network devices, but also in between its internal nodes. **Error! Reference source not found.** illustrates such communication between physical instances of the Bandwidth Manager maintaining separate network resource maps of the end-to-end path. Internal reservation requests within the Bandwidth Manager may cover either the complete path of the reservation originally requested by an application, or just part of that path.

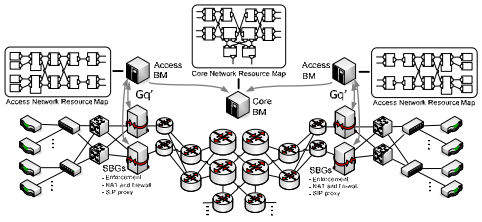

Figure 5 End-to-end QoS for IMS, intra-operator

As for reservation requests issued over Gq´, reservation requests internal to the Bandwidth Manager can be made on a per-session basis, or on a per-aggregate session basis. For aggregate requests an algorithm is needed to decide the amount of resources that should be pre-allocated at given point in time. To optimize resource utilization this algorithm needs at a minimum to account for how often each individual aggregate reservation needs to be updated the speed at which it can be updated, and the amount of resources that can be expected to be allocated for each update. Figure 6 shows the distributed Bandwidth Manager within which aggregate reservations can be used between RACF instances to server per-session requests from the SPDF layer.

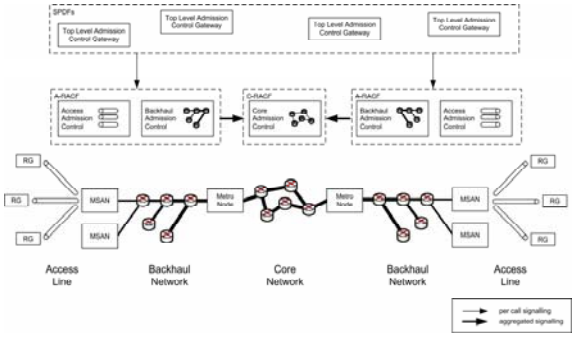

Figure 6 The distributed Bandwidth Manager

The distributed Bandwidth Manager can offer a single point of contact to applications and session control devices for issuing all admission requests, independently of which physical location in the network that is concerned by the requests. Applications provide IP addresses or contractual identifiers to the Bandwidth Manager, which matches this information to the physical location of the end-points of the requested reservation.

A benefit of the distributed Bandwidth Manager is that it supports gradual deployment. Starting with point solutions for selected access networks and services the Bandwidth Manager can be extended to cover additional access networks and one or more core networks connecting them together. When handling all potential contention points at a given end-to-end path, the Bandwidth Manager can provide end-to-end QoS to applications.

End-to-end inter-operator QoS control

Peering relations with several other network providers can enrich the offerings of a network provider to its attached service providers. End-to-end resource and policy based admission control is such an offering.

A key solution for end-to-end QoS control is inter-operator communication between Bandwidth Managers.

This includes inter-Bandwidth Manager communication between both service providers and network providers as well as between multiple network providers for end-to-end QoS control. Such Bandwidth Manager instantiations correspond to the value chains in which aggregates of network resources are offered between networks in potentially several steps before being offered to subscribers and booked for individual traffic streams.

The concept of allocating bandwidth for sessions or for aggregates of traffic via inter-RAC communication makes the RAC architecture scale to support the multi-provider value chain that can exist over end-to-end paths between communicating subscribers or between subscribers and the content to which they are to be given access.

Inter-Bandwidth Manager communication supports the NGN objective of separating service control from network control. In this separation the primary objective of service control is to provide session setup signaling to identify the end-points that are to be intercommunicating. When the endpoints are identified, session control requests from network control the necessary resources for the media stream.

The basic IMS signaling identifies the current location of a roaming subscriber by signaling though the home domain of that subscriber in order to find his/hers current location.. The media stream on the other hand should be routed along a more optimal path to avoid unnecessary QoS problems. If optimization of media route would not be provided, a roaming user would suffer the delay of having the media streams being routed through the home domain. This would be a problem especially when two end-points to communicate are both roaming far away from home. The methods of inter-Bandwidth Manager communication (path-decoupled or path-coupled) support media route optimization independently of the route of session control.

The wholesale and retail split

Service providers using RAC services offered by a network provider can rely on the network provider's RAC to enforce different subscriber policies, which are potentially related to multiple application provider agreements. Ideally, a network provider should be agnostic about subscriber policies and equipment controlled by the service provider. This is allowed by the RAC architecture as it supports inter-RAC communication, which enables the service provider to deploy its own RAC, creating a two-tier RAC physical instantiation (Figure 7).

The service provider (SP) RACF handles subscriber policies and service provider equipment such as CPEs (Customer Premises Equipments). This RACF may also enforce policies for different application providers on behalf of the subscribers. In addition, the SP RAC may perform admission control to network resources such as access lines under the control of the service provider (e.g. local loop unbundling). The network provider (NP) RAC handles policies for the different service providers and network equipment in the network provider domain. The NP RAC typically performs admission control to all network resources not being retailed by any service provider. Examples include access and core transport resources.

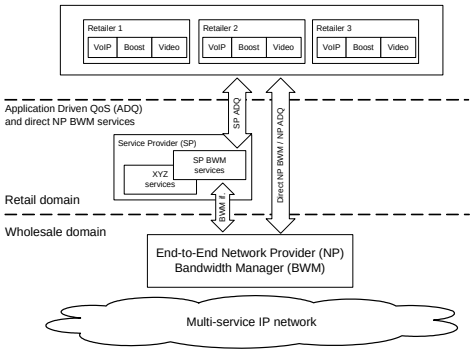

Figure 7 Retail and wholesale split

A benefit of chaining RAC for deployment in both the server provider domain and the network provider domain is that policies as well as control of resources and devices can be clearly divided into separate physical instantiations. The alternative of having the network provider RAC perform policy, resource and device control on behalf of different service providers precludes such separation of responsibilities and requires service providers to expose sensitive subscriber information to the network provider.

Another benefit of chaining RAC is that a core network provider may host some services even though the network access is provided by an independent access provider (e.g. local loop unbundling). For example an access provider may provide broadband access connectivity and some selected application services, while IMS services are hosted by the core network provider. In that case the IMS based call session control functions hosted by the core network provider request resources from the core Bandwidth Manager, and through inter-RAC communication with a Bandwidth Manager of the access service provider, end-to-end QoS is provided.

Benefits of the independent Bandwidth Manager in a wholesale/retail perspective

When the complexity of the service offering increases it will be increasingly difficult for the network operators to guarantee service performance relying on network provisioning or call counters. The end-to-end RAC provides the flexibility to introduce a multi-service and

QoS guaranteed business model across networks and service providers.

A QoS guaranteed business model increases the value of the network by enabling quality based price discrimination with performance guaranties and, as a consequence, full exploitation of the consumer surplus (Figure 8). The RAC solution proposed in this paper enables service providers that also own networks to implement session based QoS guaranties both as a retailer and a wholesaler. This solution also provides the possibility to capitalize on QoS guaranties to third party content providers.

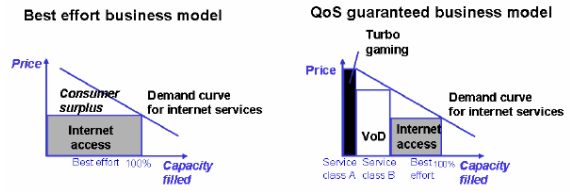

Figure 8 QoS value-add revenue generation

QoS control also reduces the CAPEX of forwarding capacity. The reason is that QoS control reduces the amount of excess capacity needed compared to a network without QoS control. As illustrated in Figure 8, even with a small safety margin - considerably more excess capacity is needed without QoS control compared to when such control is present.

Figure 9 Excess capacity with and without RAC

QoS control also has operational implications. New services with accountable QoS can be deployed quickly, without requiring immediate network level re-provisioning or an exact understanding of overall bandwidth consumption. Automated QoS control simplifies planning and dimensioning activities and thereby improves overall operational efficiency.

Guaranteed QoS also reduces operational overhead since customer complaints due to insufficient service performance are minimized. For critical and high-quality applications, customers tend to immediately call support to complain if the service quality falls below their expectations. Each single support call imposes a cost on the operator, often between $30 and $100. Periods of degraded service quality should hence be avoided or at least be kept as short as possible.

Overload situations are handled gracefully so that sufficient quality of service is provided to a maximum number of sessions at any time, gradually serving everyone. Without admission control a large number of users suffer, retry and keep the network in persistent overload.

Conclusion

This paper presents benefits of deploying QoS control in next generation networks supporting multiple services. These include guaranteed QoS, less critical network dimensioning and less operational expenditures for handling customer complaints.

The paper further presents the Bandwidth Manager, a stand-alone distributed system for QoS control. The main benefits of the stand-alone Bandwidth Manager are end-to-end guaranteed QoS for multiple services across multiple network technologies and service providers, cross service resource sharing, and less operational overhead for dynamic re-provisioning of the network due to new applications and changing user demand.

The distributed nature of the Bandwidth Manager allows it to be gradually deployed starting with a limited number of access networks and supported services. By adding hardware and network coverage the Bandwidth Manager can be extended to support end-to-end QoS, independent service providers, and advanced business scenarios such as a retail and wholesale split of the access network.

References

[1] **Schelén O., Couturier A., Bless R. and Dugeon O., "Path-coupled and Path-decoupled Signaling for NSIS", Expired Internet DRAFT, November 2002, URL; http://www.tm.uka.de/~bless/draft-schelen-nsis-opopsig-01.txt**

[2] **Gallon and Schelén: "Bandwidth Management in Next Generation Packet Networks", Aug 2005, www.msforum.org/techinfo/ reports/MSF-TR-ARCH-005-FINAL.pdf**

[3] **Blake et al: "An Architecture for Differentiated Services", Dec 1998, RFC2475**

[4] **Davie et al: "An Expedited Forwarding PHB (Per-Hop Behavior)", March 2002, RFC 3246**

[5] **Heinanen et al: "Assured Forwarding PHB Group", June 1999, RFC 2597**

[6] **Rosen et al: "Multiprotocol Label Switching Architecture", January 2001, RFC 3031**

[7] **ETSI TISPAN: "Resource and Admission Control Subsystem (RACS) Functional Architecture", ETSI ES 282 003**

[8] **ITU-T: Resource and Admission Control Functions. TD 81, rev 2 (work in progress)**

[9] **ETSI TISPAN: "Protocol specification, Gq' interface stage 3", ETSI TS 183 017**