Survey on QoS Management of VoIP

Xiuzhong Chen[*], Chunfeng Wang[*], Dong Xuan[+], Zhongcheng Li[*], Yinghua Min[*] and Wei Zhao[#]
[*]*Institute of Computing Technology, Chinese Academy of Sciences*
Beijing, 100080, China
{chenxz, cfwang, zcli, min}@ict.ac.cn

[+]*Dept. of Computer and Information Science*
Ohio State University
Columbus, OH 43210, USA
xuan@cis.ohio-state.edu

[#]*Dept. of Computer Science*
Texas A&M University
College Station, TX 77843, USA
zhao@cs.tamu.edu

Abstract

In this paper, we present a survey on management mechanisms used for ensuring the Quality-of-Services for Voice-over-IP applications. We first address the motivations of QoS management for VoIP. We then partition the system into two management planes: data plane and control plane, and describe mechanisms in each plane respectively. In data plane, we cover several important techniques including packet classifier, buffer management, scheduling, loss recovery, and error concealment. In control plane, we describe admission control, resource provisioning, traffic engineering, and connection management etc. As we will see, admission control plays a critical role in QoS for VoIP. Thus, we will discuss methodologies used in admission control in detail. Specifically, we examine parameter-based and measurement-based admission control algorithms, and their use in VoIP. Finally, we discuss limitations of current technologies and future research issues in providing QoS to VoIP applications.

1. Introduction

VoIP (Voice-over-IP) refers to the transmission of voice using IP technologies over packet switched networks. It consists of a set of facilities and protocols for managing the transmission of voice packets using IP. Internet Telephony is one of the typical applications of VoIP. Compared to traditional resource-dedicated PSTN, IP network is resource-shared. Therefore, IP-based VoIP applications are cost-efficient. Open protocol and layer architecture of IP permit that VoIP technologies can not only realize existing quality telephone service, but also provide more and much better services, such as telephone conference and whiteboard.

However, current IP networks are based on best-effort services. They lack stringent QoS control. Congestion is inevitable in IP networks and may result in packet loss, delay, and delay jitter, which directly impact the quality of VoIP applications. So, the current IP network architecture must be enhanced by some QoS guaranteeing mechanisms in order to ensure QoS of VoIP applications. This paper will focus on QoS management mechanisms for VoIP applications.

The rest of the paper is organized as follows: In Section 2, we will describe typical VoIP architecture and present an overview of QoS management of VoIP. Then, In Section 3 we will describe QoS mechanisms in data plane, especially issues related to packet forwarding, loss recovery, and error-concealment. In Section 4, we will provide an overview of the QoS mechanisms in control plane and discuss resource provisioning, traffic engineering and call admission control (CAC). Section 5 will summarize the paper with final remarks.

2. VoIP architecture and QoS management

Current implementation of VoIP has two types of architectures, which are based on H.323 and SIP frameworks, respectively. H.323, which was ratified by ITU-T, is a set of protocols for voice, video, and data conferencing over packet-based network. SIP, which is

Proceedings of the 2003 International Conference on Computer Networks and Mobile Computing (ICCNMC'03)
0-7695-2033-2/03 $17.00 © 2003 **IEEE**

defined in RFC2543 of the MMUSIC working group of IETF, is an application-layer control signaling protocol for creating, modifying, and terminating sessions with one or more participants. Regardless of their differences, the fundamental architectures of these two implementations are the same. They consist of three main logical components: terminal, signaling server and gateway. They differ in specific definitions of voice coding, transport protocols, control signaling, gateway control, and call management.

QoS requirements of VoIP include packet loss, delay, and delay jitter. It should be pointed out, however, while the current H.323 and SIP frameworks support some kind of interfaces to QoS management (e.g., the one between H.323 and RSVP), they do not provide functional QoS management mechanisms. Consequently, products in the market now (e.g., Cisco's and Alcatel's VoIP systems) cannot provide QoS guarantees to VoIP applications [7].

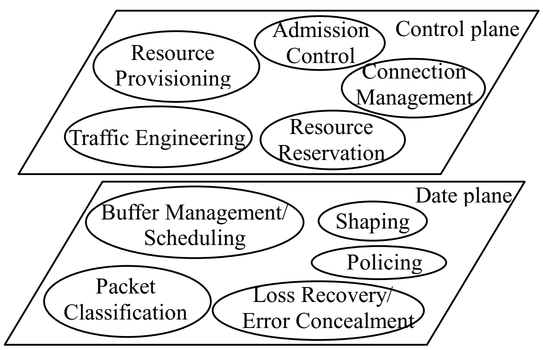

Figure 1. QoS management mechanisms of VoIP

However, many solutions have been developed in industrial and university laboratories to address this problem. Many are readily realizable within the framework of H.323 and SIP. In the rest of this paper, we survey these solutions and discuss related issues. QoS management architecture of VoIP can be partitioned into two planes: data plane and control plane [1] (Figure 1). Mechanisms in data plane include packet classification, shaping, policing, buffer management, scheduling, loss recovery, and error concealment. They implement the actions the network needs to take on user packets, in order to enforce different class services. Mechanisms in control plane consist of resource provisioning, traffic engineering, admission control, resource reservation and connection management etc. They allow the users and the network to reach a service agreement, and let the network appropriately allocate resources to ensure QoS guarantees to the calls that have been admitted. QoS management mechanisms in both data plane and control plane work together to provide QoS to voice in IP networks. At the configuration time, resource provisioning of control plane determines which portion of resources to be allocated for voice traffic. At run-time, upon a new call arrival, admission control of control plane will decide the acceptance of the new call based on the amount of provisioned resources and the usage of the resources by the admitted calls. Once the call is admitted, the end-hosts start sending voice packets to the network. QoS mechanisms of data plane will be invoked to perform traffic-enforced functions such as traffic classification, shaping, buffer management, scheduling and error control etc, directly on the voice packets.

In the following sections, we will describe QoS mechanisms in both data plane and control plane.

3. QoS Mechanisms in Data Plane

3.1. Overview

QoS mechanisms in data plane can be mapped to corresponding layers of IP protocols. QoS mechanisms of network layer consist of packet classification, buffer management, scheduling, shaping and policing. These mechanisms commonly implement the control functions of packet forwarding by controlling the per-hop behavior of packet forwarding. They are closely related, but address rather different performance issues [1]. In different networking environments, different methods may be adopted due to trade-offs among QoS service performance, operations and management complexity, and implementation cost. We will discuss the forwarding mechanisms in Section 3.2.

Different applications have different behavior characteristics and QoS requirements, so QoS guaranteeing mechanisms should be application-specific. VoIP is very sensitive to delay and less sensitive to loss (up to certain degree). Some studies have exploited this characteristic of VoIP in order to improve the voice quality of VoIP

[3,5]. Application-level QoS managements of VoIP include jitter buffer, end-to-end loss recovery and error concealment. Jitter buffer is a common means to eliminate the impact of delay variations. Loss recovery and error concealment are the required means to improve quality of voice because packet loss is inevitable in IP networks. While Admission Control provides non-stringent delay control and permits packet loss below a degree of probability, loss recovery and error concealment can work as useful complementary mechanisms. In Section 3.3, we will discuss loss recovery and error concealment.

3.2. Packet Forwarding

3.2.1. Basic Forwarding Operations

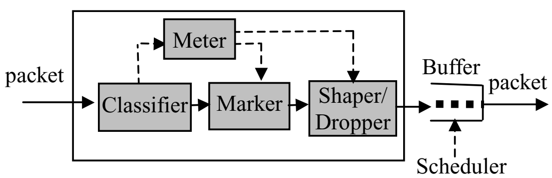

Figure 2. Basic forwarding operations

Figure 2 shows the basic forwarding operations in a router. As described in [2] and RFC 2475, when a packet is received, a packet classifier determines which flow or class it belongs to. All packets belonging to the same flow/class obey a predefined rule and are processed in a similar manner. For VoIP applications, the basic criteria of classification could be IP address, TCP/UDP port, protocol, input port, IP precedence, DiffServ code points (DSCP), or Ethernet 802.1p class of service (CoS). Cisco supports several additional criteria such as access list and traffic profile.

After classification, the packet is passed on to a traffic conditioner, which may contain meter, marker, shaper, and dropper. A meter is to decide whether the packet is in a traffic profile. This information may be used by other elements to trigger a particular action. In-profile packets are put in different service queues for further processing. A shaper or a dropper delays or drops out-of-profile packets in a packet stream in order to bring the stream into compliance with its traffic profile. The function of a dropper is known as traffic policing. A marker marks the certain field in the packet, such as DS field, to label the packet type for differential treatment later. After the traffic conditioner, a buffer is used to store packets that wait for transmission. A scheduling policy is used to select packets for transmission in a link.

3.2.2. Buffer Management and Scheduling

Buffer management and scheduling are closely related and ensure that service guarantees of applications are met [1].

As described in [2], buffer is designed to absorb short-term packet bursts. Traditionally, packets are dropped only when the queue is full. This policy may keep queues at or near maximum occupancy and cause unfair resource usage. Active queue management, such as Random Early Detection (RED) and Fair RED [1, 2], which drops packets before a queue becomes full, can avoid these problems.

Scheduling policy is primarily to control queuing delay and bandwidth sharing [2]. The aggregate bandwidth of a link can be shared among multiple entities. There are varieties of scheduling disciplines, such as First Come First Serve (FCFS), Static Priority (SP), Weighted Fair Queuing (WFQ), Earliest Deadline First (EDF).

In general, all routers support FCFS while many new devices are now able to support WFQ and Class-Based WFQ (CBWFQ). In some of Cisco products, policies of FCFS, WFQ, Custom Queuing, Priority Queuing, CBWFQ, Priority Queuing WFQ, Low Latency Queue (LLQ) are all supported.

There is a trade-off between simplicity and performance in different scheduling policies. Ideally, a scheduler should provide predictable queuing delay and bandwidth sharing to the traffic, which naturally asks for a per-flow based treatment [2] to the individual flows. Some schedulers follow this line, for example the per-flow based WFQ. With this scheduling policy, the flows are put into different queues and treated individually, hence the predictable queuing delay and bandwidth sharing for the traffic can be achieved. However, schedulers of this type are not scalable. Their overhead increases as the number of the on-going traffic flows increases. They may not be able to function at a

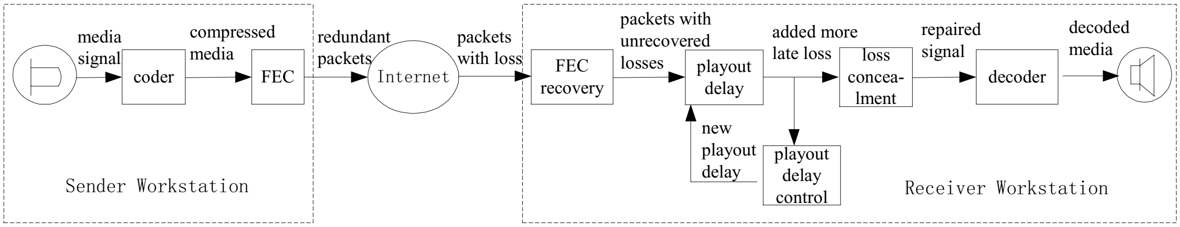

Figure 3. Loss recovery and error concealment in packet audio [5]

high-speed link. For example, at OC-48 rate, a scheduler only has 100ns per packet to make a scheduling decision. In this sense, a scheduler needs to be simple and scalable. Class-based schedulers such as Static Priority (SP) and CBWFQ meet this requirement. They schedule the traffic in a class-basis fashion. Traffic is aggregated into different classes and put into different queues. The overhead of scheduling is not related to the number of on-going traffic flows. The downside of the class-based schedulers is that it is difficult for the individual flow to get the predictable delay and bandwidth sharing. Care has to be taken in applying this type of schedulers in voice applications which have strict delay requirements.

3.3. Packet Loss Recovery and Error Concealment

A recovery process can be divided into two stages: loss recovery and error concealment [3, 4] (Figure 3 [5]). Loss recovery is to recover the original content of a lost packet. Loss recovery can only recover a single lost packet, or work well under special network scenario. Recent researches show that packet loss can exhibit temporal dependency or bursts, which degrade its effectiveness [5]. Therefore, error concealment is needed to conceal the remaining loss in voice streams after loss recovery. Error concealment complements loss recovery.

3.3.1. Loss Recovery

Loss recovery mechanisms may be split into two major classes: active retransmission and passive channel coding [3] (Figure 4). Retransmission increases the latency of packets and may not be suitable for VoIP. Passive schemes mainly use forward error correction (FEC). FEC adds redundancy information into voice streams for aiding the loss correction. FEC can be either media-independent or media-dependent.

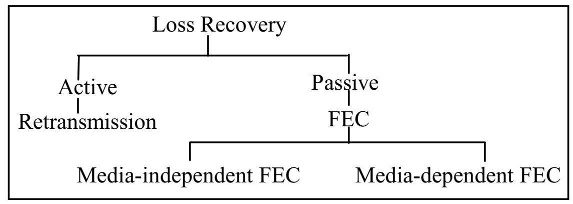

Figure 4. A taxonomy of loss recovery

Media-independent FEC [3] uses voice block, or algebraic, codes to produce additional repair packets. Parity coding and Reed-Solomon coding are two common schemes of block coding. They are relatively simple and easy to implement. The disadvantages are the additional delay imposed, increased bandwidth, and difficulty in decoder implementation. RFC2733 and RFC2198 define how to apply these schemes in RTP streams and most of VoIP products have implemented them.

Media-dependent FEC [3] protects against packet losses by transmitting each unit of voice in multiple packets, i.e. primary encoding and secondary encoding. The secondary usually uses a lower-bandwidth, lower -quality encoding than the primary. This method results in low latency, with only a single-packet delay being added. But this results in increased overhead. The repair efficiency varies with the overhead. RFC2198 describes the standard RTP payload format for media-specific FEC.

3.3.2. Error Concealment

Error concealment schemes [3, 4] produce a replace-ment for a lost packet, which is similar to the original lost packet. This is possible because voice signals exhibit large amounts of short-term self-similarity. Depending on interactions with source encoding

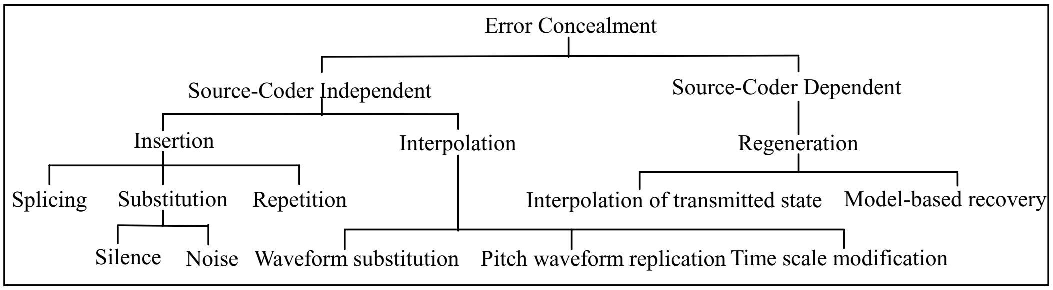

Figure 5. A taxonomy of error concealment

schemes, error concealment schemes can be divided into source-coder independent and source-coder dependent schemes (Figure 5). The former does not exploit the knowledge of the underlying coding algorithms, and only creates a replacement by simple interpolation. The latter regenerates a replacement by exploiting features in individual coders [4].

Insertion-based repair schemes [3] derive a replacement by inserting a simple fill-in, i.e. silence, background noise, comfort noise, or repeating packet of the last received packet. The simplest method is to splice together the voice on either side of the loss. Splicing disrupts the timing of the stream and is not an acceptable repair technique. The implementation of these schemes is simple. All of them but repetition, generally result in poor performance. Comfort noise is commonly used in most of VoIP products.

Interpolation-based schemes [3] account for the changing characteristics of a voice signal, and use some pattern matching algorithms and packets surrounding a loss to create a replacement. Depending on the algorithms used, they are subdivided into waveform substitution, pitch waveform replication, and time scale modification. Compared to insertion-based schemes, they are complex, difficult to implement, and require intensive processing. But they can achieve good performance.

Regeneration-based schemes synthesize voice of a lost packet by using voice compression algorithm to derive codec parameters [3]. Interpolation schemes can interpolate the states of codec based on transform coding or linear prediction, e.g. ITU G.723.1. Model-based schemes use a model to generate the replacement. The majority of low-bit-rate speech coders use an autoregressive model and an excitation signal [3]. These schemes typically are computationally intensive and have high cost for keeping the state information, although their performance is proven to be effective.

In summary, there are trade-offs among cost and quality, complexity and performance in applying recovery schemes. [5] has compared the performance between traditional FEC and low bit-rate redundancy (LBR) under bursty loss.

4. QoS Mechanisms in Control Plane

4.1. Overview

VoIP applications are delay-sensitive and loss -insensitive. Studies show that the end-to-end delay is the primary parameter affecting the quality of voice in the Internet. ITU G.114 defines that the maximum tolerable end-to-end delay is 150ms.

The existing Internet service cannot satisfy the QoS requirements of VoIP, primarily because of network congestion due to insufficient network resource. Many mechanisms in control plane, including resource provisioning, traffic engineering, admission control, resource reservation, and connection management, have been proposed to resolve this problem. These mechanisms offer QoS in one of the following ways: either (1) by predicting the traffic and engineering the network to make violations of the committed QoS highly unlikely or (2) by restricting the total amount of traffic competing for the same resource. Generally speaking, resource provisioning and traffic engineering follow the first way. They are always conducted at the configuration time, and naturally focus on the coarse-granularity traffic, e.g. the whole body of the potential traffic of the network. On the contrary, QoS mechanisms such as admission control and resource reservation are always performed at run-time, and work for the fine granularity traffic, e.g. an individual voice flow. In the following section, we will first give a brief description about resource provisioning and traffic engineering mechanisms, and then focus on the discussion about various admission control mechanisms.

4.2. Resource Provisioning and Traffic Engineering

Resource provisioning refers to the configuration of resources for applications in the network. The amount of provisioning depends on the previous aggregate user demand. This mechanism allocates network resources at the network configuration time.

In industry, the main approach of resource provisioning is over-provisioning. The overprovision mechanism is to abundantly provide network resources so that applications behave as if the network is unloaded. It is the simplest but most expensive method. Currently, bandwidth requirement of Internet applications will vary highly, which means overprovision requires a huge increase in the capacity architecture of the link and would result in low average utilization levels. However, there are several factors that make this solution attractive: the cost of bandwidth in the backbone is decreasing, the network planning is simple, and provisioning can be planned. When the bandwidth is sufficient, over-provisioning is an attractive solution.

Traffic Engineering is concerned with the performance optimization of traffic handling in operational networks. Its main focus is minimizing over-utilization of a particular portion of the network while the capacity is available elsewhere in the network. It is concerned with the design, provisioning, and tuning of operational Internet. Multi-protocol Label switching (MPLS) and constraint-based routing[1] provide powerful tools for traffic engineering. Constraint-based routing is an extension to the basic topology based routing. It routes packets based on multiple constraints, such as bandwidth or delay requirement, network topology, network resource availability information (mainly link available bandwidth), and policy constraints. Constraint -based routing is an important tool for making the traffic engineering process automatic. It helps to provide QoS guarantees by carefully determining the routing paths. The resultant paths by constraint-based routing can be pinned by MPLS.

In VoIP networks, to implement resource provisioning efficiently it is necessary to consider the distributed model of VoIP users and the behavior model of the call. Every call is often regarded as a CBR flow. Multi-path-constraint-based routing is still an open issue. Delay-bandwidth-constrained routing can be helpful to improve the performance of VoIP systems.

As stated above, resource provisioning and traffic engineering are usually performed at the configuration time. With these two mechanisms, a certain amount of network resources can be reserved for the potential voice traffic along the paths. These paths may be determined by constrained-based routing algorithms or just by the traditional shortest path routing algorithms. Admission control mechanism, to be discussed in the following section, will limit the resource usage of voice traffic within the amount of the provisioned resources.

4.3. Admission Control

4.3.1. Admission Control Basics

The traditional telephones control the user access via call admission control. However, most of the current IP networks have no admission control and can only offer best effort service. That is, new traffic may keep entering the network even beyond the network capacity limitation; consequently make both the existing and the new flows suffers packet delay and loss. To prevent these occurrences, admission control mechanisms should be at place.

Figure 6 illustrates the architecture of admission control. Admission control unit makes admission decision to the new request. Admission Criteria is a set of conditions used to determine if an incoming call is to

[1] Routing, i.e. path determination, can be done at the configuration time, or at run-time, e.g. during admission control.

Proceedings of the 2003 International Conference on Computer Networks and Mobile Computing (ICCNMC'03)
0-7695-2033-2/03 $17.00 © 2003 IEEE

be accepted. Network QoS state[2] and flow information are necessary and useful information to admission control unit.

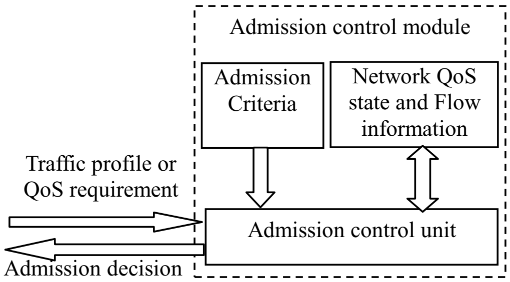

Figure 6. Components of admission control

Upon a new call arrival, admission control unit will make admission decision based on admission criteria, network QoS state and flow information. Once a call is admitted, connection management will be invoked to take care the establishment and tear-down of the call, and the resource will be reserved along the path[3] of the voice traffic following the call.

In the following, we will introduce two main admission control approaches: parameter based and measurement based.

4.3.2. Parameter-based Admission Control

This is a traditional admission control scheme, which computes the worst case delay or packet loss, using the information of traffic profile of the existing and new flows, and then makes admission decision.

Parameter-based Admission Control uses explicit traffic descriptors. A typical descriptor is a token bucket, which limits the amount of traffic over any period. All flow information and available network resource status are stored in the database. Admission control unit computes the worst-case delay, and compares with the required delay requirement to make admission decision accordingly.

In the following, we will describe some typical algorithms used in parameter based admission control for VoIP, which include Cisco's resource reservation based [6], utilization-based [11], per-flow end-to-end guaranteed delay service and class-based admission control [8].

Cisco's resource reservation based admission control uses RSVP along the path from source gateway to destination gateway to pre-allocate resources. The intermediate routers only conduct the simple bandwidth summation. This algorithm can be applied only to support the network where routers support RSVP.

In [7], a framework is proposed to integrate utilization based admission control with Cisco's gatekeeper, aiming to provide delay guarantees to VoIP networks. Utilization-based admission control adopts some configuration steps to simplify the real-time delay test to a simple utilization check. As long as the utilization along the path is not beyond the pre-determined threshold, the worst case delay of calls can be guaranteed. The threshold is determined and verified at the configuration time. The key step with this algorithm is the worst-case delay computation at the configuration time. [11] proposed CPIDC (connection -population-insensitive delay computing) algorithm to remove the dependence of the general delay formula on the dynamic information. Once the safe utilization -bound is verified at the configuration time, utilization -based admission control can admit the new request based on the utilization at the run time. This type of admission control is very efficient and straightforward.

In [9], a bandwidth broker architecture was proposed to support VoIP applications, which uses two admission control algorithms, per-flow end-to-end guaranteed delay service and class-based admission control in a core stateless network proposed in [8]. It use bandwidth broker to store and manage QoS reservation state, and excuse admission control and resource management. The core router need not maintain per-flow information. Upon a new request arrival, admission control unit first computes the equivalent bandwidth requirements of the new flow; and then compares with available resource to make decision.

To summarize, parameter-based admission control can provide delay guaranteed service to applications, which can be accurately described, such as VoIP.

[2] It includes the information about the amount of resources provisioned for the voice traffic at the configuration time.

[3] The path can be determined at the configuration time or at run-time by the traditional shortest path routing algorithms or constraint-based routing algorithms.

However, when the traffic is very bursty, it is very difficult to describe the traffic characteristics; and then parameter-based admission control may result in overbooking the network resource and hence lowering network utilization.

4.3.3. Measurement-based Admission Control

Measurement based admission control was first proposed in [12], and further studied in [13,14]. It consists of two parts: measurement, which is used to estimate the current network load, and admission control based on estimated network load.

Measurement-based admission control algorithms can be classified into two types, local load measurement -based and end-to-end load measurement-based. With local load-based measurement algorithms [12,13,14] each router along the flow path measures its local load status and conducts admission control. End-to-end measurement based admission control algorithms [10,15] rely on the end hosts or edge nodes to estimate the available network resources. The end hosts and edge nodes may actively send probe messages to get the network status. The core routers usually do not involve in the estimation. Compared with the traditional parameter-based admission control algorithm, the measurement-based admission control algorithm suffers a certain degree of QoS degradation, and it can only provide the so-called soft QoS guarantees.

Several measurement-based admission control algorithms have been used in existing VoIP products: Cisco's measurement-based call admission control algorithm [6] and traffic matrix-based admission control algorithm (TAMC) [10].

Cisco's measurement-based call admission control algorithm has two types: Advance Voice Busyout (AVBO) and PSTN Fallback [6]. They both use Security Assurance Agent (SAA) to conduct measurement. SAA adopts Response Time Reporter (RTR) method to measure the network performance, such as delay, packet loss, delay jitter and ICPIF (Calculated Planning Impairment Factor, ITU-T G.113), etc., and feeds the measurement results back to source gateway for admission control. When the network performance is below the pre-determined threshold, admission control unit will be activated. The difference between AVBO and PSTN Fallback is that PSTN Feedback doesn't configure IP addresses statically; and uses cached addresses. Those two mechanisms can only be used in the network with the support of Cisco's SAA.

Traffic Matrix admission control (TMAC) provides QoS guarantees within a single domain or multi-domains. It uses Clean House architecture [10] to provide intra- and inter- QoS guarantees to Voice over IP networks. It moves resource configuration and admission control away from the router to the layered Clean House architecture. TMAC uses four metrics, demand metrics, upper-bound Matrix, Node-level Traffic Matrix and POP-level Traffic matrix to describe the total bandwidth requirements, the resource threshold, and measured load of each ingress-egress pair or pop pair respectively. Consequently, admission control only checks whether the sum of the peak rates of the new flow and the measured traffic of the same path exceeds the corresponding upper-level item.

In summary, measurement-based admission control can only provide soft (rather than deterministic) delay guarantees for voice traffic. It relies on the measurement period to reflect the dynamics of network status. Shorter measurement period, which means measurement is conducted more frequently, can reflect the networks better, but consumes more network resources; and that longer measurement period cannot reflect the network dynamics well, but cost less network resources. Also, in some cases, the probe messages are not payload messages themselves, which may not be able to reflect how the network treats the real voice packets.

5. Conclusion

In this paper, we have surveyed QoS management mechanisms for VoIP applications. While much progress has been made in VoIP related QoS mechanisms, several issues remain to be addressed in order to fully deploy the technology: for example, signaling protocols of QoS controlling need to be further studied and analyzed; multi-level VoIP services are yet to be considered. Some mechanisms in data plane and control plane, such as flow-based classifier and scheduler, signaling, admission control, bandwidth broker and inter-domain QoS provisioning require further refinements for a scalable and efficiency implementation of QoS guaranteed VoIP system.

References

[1] R. Guerin and V. Peris. Quality-of-service in packet networks: Basic mechanisms and directions. *Computer Networks*, 31(3):169--189, February 1999.

[2] Weibin Zhao, David Olshefski and Henning Schulzrinne. Internet Quality of Service: an Overview. Columbia Technical Report, CUCS-003-00. Columbia University, Computer Science Department, Feb. 2000.

[3] Colin Perkins, Orion Hodson, and Vicky Hardman. A survey of packet loss recovery techniques for streaming audio. IEEE Network, 12(5):40-48, September 1998.

[4] Benjamin W. Wah, Xiao Su, DongLin, A survey of error-concealment schemes for real-time audio and video transmissions over the Internet. Proc. Int'l Symposium on Multimedia Software Engineering, IEEE, Taipei, Taiwan, Dec. 2000, pp. 17-24.

[5] Wenyu Jiang, Henning Schulzrinne, Modeling of packet loss and their effect on real-time multimedia service quality. NOSSDAV 2000, Chapel Hill, NC, June 2000.

[6] VoIP Call Admission Control. http://www.cisco.com /univercd/cc/td/doc/cisintwk/intsolns/voipsol/cac.htm

[7] S. Wang, Z. Mai, W. Magnussen, D. Xuan and W. Zhao. Implementation of QoS-Provisioning System for Voice over IP. Proceeding of IEEE Real-Time Technology and Applications Symposium 2002 (RTAS 2002).

[8] Zhi-Li Zhang, Zhenhai Duan, Lixi Gao, and Yiwei T. Hou. Decoupling QoS Control from Core Routers: A Novel Bandwidth Broker Architecture for Scalable Support of Guaranteed Services. Proc. ACM SIGCOMM'00, Sweden, August 2000.

[9] Zhi-Li Zhang, Zhenhai Duan, and Yiwei T. Hou. On Scalable Design of Bandwidth Brokers. IEICE Transaction on Communications, E84-B (8), Aug. 2001

[10] Chen-Nee Chuah. A Scalable Framework for IP-Network Resource Provision Through Aggregation and Hierarchical Control. PH.D. Thesis, Fall 2001.

[11] B. K. Choi, D. Xuan, C. Li, R. Bettati, and Wei Zhao. Utilization-Based Admission Control for Scalable Real-Time Communications. In Journal of Real-Time Systems, accepted in October 2000.

[12] S. Jamin, P. B. Danzig, S. Shenker and L. Zhang. A Measurement--based Admission Control Algorithm for Integrated Services Packet Network. SIGCOM M'95, Cambridge, MA, USA, August 1995.

[13] S. Floyd. Comments on Measurement-based Admissions Control for Controlled-Load Service. Submitted to Computer Communication Review, 1996.

[14] S. Jamin, S. Shenker, and P. Danzig. Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service. Proc. of INFOCOM '97, April 1997.

[15] V. Elek, G. Karlsson, and R. Ronngre. Admission control based on end-to-end measurements. In Proc. of the IEEE INFOCOM, Tel Aviv, Israel, Mar. 2000.