Call Admission Control: QoS Issue for VoIP

Mohammad Asadul Hoque
Database Administrator, IT Division
TM International Bangladesh Limited (AKTEL)
Dhaka, Bangladesh
asad1752@ieee.org

Farhana Afroz
Department of Mechanical Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
diba_mym@yahoo.com

***Abstract*—One of the vital key elements for providing Quality of Service (QoS) for VoIP is the Call Admission Control (CAC) capabilities of the Session Management /Call Session Control Function or gateway. Even though the network may be designed to meet a given performance and restoration objective for the engineered traffic loads, the actual traffic may be significantly higher. Without a CAC function in a VoIP network during overloads, links become congested and new calls keep getting admitted. All calls in progress, not just the new calls start dropping packets and experiencing longer delays. Contrast this to a circuit switched network, where new calls get blocked, but calls in progress experience good call quality. In the VoIP case, packet loss could become large enough that calls become unintelligible, callers hang-up their call, and most will reattempt. This paper gives an overview of potential CAC approaches, highlighting four basic alternatives; based on endpoint performance measurements, path-based bandwidth management, link-based bandwidth management, and per-call bandwidth reservation. The paper also recommends a link bandwidth management approach for its scalability and efficacy.** *(Abstract)*

Keywords- call admission control; bandwidth ; circuit-switched; RSVP; OSPF-TE; RTCP;(key words)

I. INTRODUCTION

VoIP (Voice over Internet Protocol) is growing rapidly as the technology of choice for new voice network deployments as well as conversion of existing networks. Delivering and maintaining excellent voice communications quality, comparable to what is experienced by the subscribers in the PSTN (Public Switched Telephone Network), is one of the top priorities of most service providers. VoIP introduces a number of potential impairments that can impact voice quality adversely, such as the use of lossy low-bit-rate codecs, the effects of tandem encoding/transcoding, longer delays, and packet loss [1]. Most of these impairments are either not present or are negligible in circuit switched networks. Thus new techniques for delivering and maintaining voice quality are needed for VoIP networks.

Figure 1 shows a typical VoIP network with IP endpoints going over an IP/MPLS (Multi-Protocol Label Switched) core. We assume that the network has DiffServ[2] enabled and a traffic class reserved for voice. The impairments that a voice call experiences can be classified as either architectural or load dependent. Architectural components include IP phone codecs and their configuration parameter settings as well as fixed

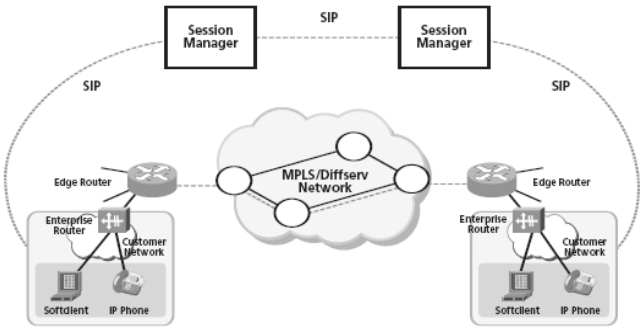

Figure 1. Typical VoIP network

components of delay such as processing delays at each network element along the path and the end-to-end propagation delay. These architectural components define an upper bound on the best voice quality that could be achieved in a given network. If the upper bound is unacceptable, then changes in equipment and configurations will be required.

In general, if the architecture is satisfactory, then low packet loss and delay are sufficient to ensure good voice quality. Load dependent impairments include packet loss, queuing delay, and jitter. As load increases, these parameters deteriorate and begin to degrade voice quality. The voice quality a user experiences depends on the behavior of the entire end-to-end connection. This connection may cross multiple network domains each with its own set of controls and management methods. Since impairments across the connection are cumulative, it is possible that each network domain delivers acceptable voice quality while the end-to-end connection does not.

This paper focuses on QoS for single domain connections and, in particular, the need for CAC (Call Admission Control). It summarizes different alternatives for providing CAC functionality, and reviews the considerations for choosing a CAC strategy for given networks and services.

II. IS CAC NEEDED?

In order to provide high levels of voice QoS, it is important to understand what happens to VoIP voice quality in a congested and overloaded network. In a circuit-switched voice network, bearer traffic of each active call is allocated with dedicated time-slots at call set up time throughout the life of the call.

However, in IP networks, the voice packets for different calls are multiplexed and share the packet forwarding capacity at each router. When new calls are admitted into an already congested network, both new and existing calls will experience increased packet loss and delay. Voice quality will degrade and eventually conversation will become impossible. Therefore it is critical to introduce mechanisms to mitigate the possibility of congestion. An option often quoted for small networks and initial deployments with a limited number of subscribers is to overprovision the network – i.e., to provide a network capacity much beyond the traffic loads expected. While such an approach may be sufficient initially, it is not scalable or viable in the long run when users demand quality voice services. Figure 2 shows a simple example of a VoIP network in the U.S with PSTN access through voice gateways. Forecasted point-to-point traffic was routed through the network on the shortest path and the links were sized for .1% packet loss. The green number on each link shows the required bandwidth based on this traffic engineering rule. In addition, the number of TDM (Time Division Multiplexed) ports on each gateway was sized for .5% blocking. To overprovision, we examine what would be the worst case traffic distribution in terms of congesting a given link. The brown number indicates the bandwidth required on each link to handle the worst case traffic load, taking into account the natural limits placed by the TDM ports. On average, the network needs to be over provisioned by a factor of 3.3, with some links over provisioned to significantly higher ratios.

In fact, this level of over provisioning increases as the number of nodes in the network increases, and is not a scalable solution. In addition, if IP endpoints are to be supported, the natural throttle defined by the TDM gateway ports is not available and the situation can be much worse. Also, this example does not include the additional capacity required to handle node or link failure.

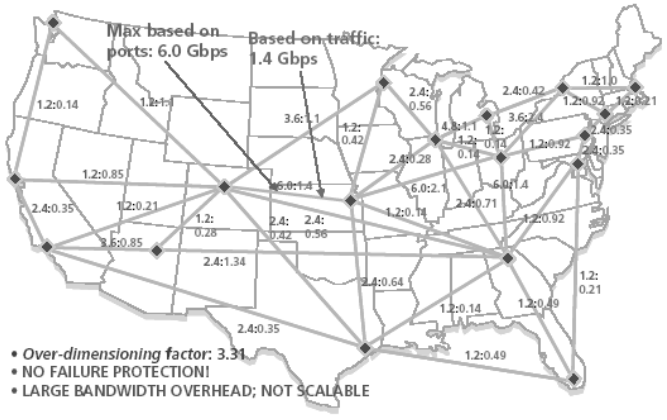

Figure 2. Over provisioning with no CAC

III. CAC ALTERNATIVES

There are several major alternatives that can be used to provide CAC functionality in a network.

A. *Per Call Bandwidth Reservation Method*

One natural approach is to reserve bandwidth hop-by-hop in the network on a per call basis using a signaling protocol. The CAC function is implemented locally on each router with little service provisioning required. Operations costs are reduced, and bandwidth on each link can be fully shared so there is excellent statistical multiplexing and efficient use of the bandwidth. In IP networks, the guaranteed service of the IntServ model[1] supports this capability. Unfortunately, this model does not scale well to large IP networks as routers are not designed to handle large volumes of signaling messages. Also, with RSVP (Resource Reservation Protocol), refresh messages are required periodically to maintain existing connections. Network design and network stability are two other significant issues associated with per call bandwidth reservation techniques and related to routing. When bandwidth is reserved dynamically, a routing protocol such as OSPF-TE (Open Shortest Path First – Traffic Engineering) may not always take the shortest path. If this is part of the plan, then new network design algorithms must take this feature into consideration. If the routing protocol is allowed to take arbitrarily longer paths to take advantage of idle capacity elsewhere in the network, there are cases where, under congestion, the network becomes bi-stable and all calls may be routed on long paths and total network efficiency reduced. This is a known phenomenon in alternate routed circuit-switched networks [4,5]. Nontrivial stable QoS routing algorithms must be designed and implemented to alleviate this problem.

B. *Call Path Measurements Method*

Another method to provide QoS for VoIP calls is to use call path performance measurement statistics to determine whether a call should be admitted. There are several variations on this idea. One is to send a stream of test packets (pings) at call set-up time and measure the round-trip response time, its variability, and possible packet loss. If these are below some threshold, then admit the call. Otherwise, the call is either blocked or routed to the PSTN, if such an option is available. One advantage with this method is that it makes no assumptions about the network that the call is being carried over. If an enterprise customer is sending traffic over the Internet or even an IP/VPN (Virtual Private Network), they have little control over the other traffic in the network and can only measure performance and act accordingly. There can be a non-negligible increase in call set-up time with this variation, caused by waiting for the test packets to return.

Another variation to obtain call path measurements is to use RTCP (Real Time Control Protocol) measurements of existing calls in progress. If a new call request is going to the same destination as some of the existing calls in progress, the most recent RTCP report for those calls will give an excellent indication of current performance. If the RTCP reports are available, this method is preferable to the per call method described above since there is negligible impact on call set-up delay, and the estimate of current performance is more robust. Results for such a method are described in a paper by Houck and Uzunalioglu [6].

[2] IntServ refers to the IETF standard on Integrated Services (RFC 1633, "Integrated Services in the Internet Architecture: an Overview").

In general, the call path measurements method does not guarantee a prescribed VoIP QoS level, although it can offer a much "better than best effort" performance. The method relies on performance statistics at or before call set-up time, which does not always translate to performance during the rest of the call. For example, VoIP and data could be sharing bandwidth via a weighted fair queuing scheduling policy that guarantees each one a certain amount of bandwidth. If data traffic is light for a period of time, the voice packets can borrow from data and not be dropped or delayed significantly. As a result, voice calls could continue to be admitted even during voice overload conditions. If the data traffic picks up and reclaims its share, all voice calls in progress would experience significant packet loss. Before deployment in a live-network, it is important to study how to implement these measurements, such as active polling versus RTCP measurements, based on the network architecture. In addition, a decision has to be made as to how often and how long to poll the network if the active polling is used.

C. *Per Call Reservation with Path-based Bandwidth Allocation*

A third approach to provide VoIP QoS is to allocate appropriate bandwidths along each path in an IP network. For example, this can be done in IP/MPLS networks between pairs of edge routers using RSVP-TE and DiffServ (or between pairs of Voice Gateways). The bandwidth reserved path is referred to as a "Virtual Trunk Group" (VTG)[8] and guaranteed QoS is provided through distributed accounting based CACs implemented at the session managers. That is, the session manager (SoftSwitch) will keep a count of the active calls or equivalent bandwidth usage on the path and start blocking when the path load reaches the allocated capacity. Figure 3 depicts an example of this algorithm with VoIP gateways attached to the edge routers. There is a set of VTGs from every edge router to the other edge routers. Each VTG is allocated a certain amount of bandwidth. A bandwidth counter at the session manager keeps track of used and available bandwidth for each VTG. This way, a call set up attempt is blocked when there is no more bandwidth available at the respective VTG, which is chosen based on the endpoints of the call set-up request.

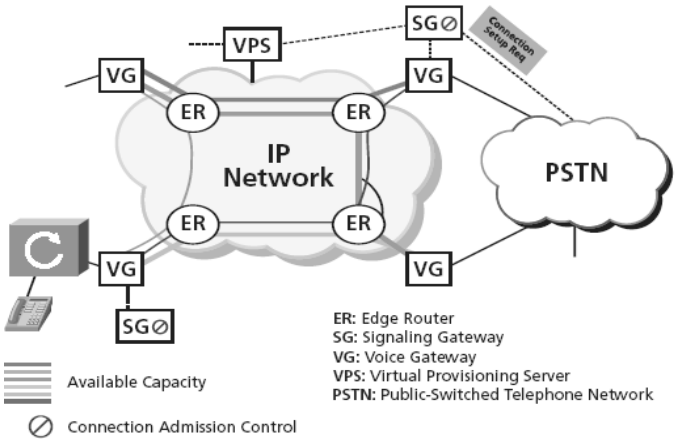

Figure 3. Per Call Reservation with Path-based Bandwidth Allocation

One of the key advantages of this approach is that control can be distributed to the session manager that controls the ingress to each VTG and little global network knowledge is necessary. In addition, the approach takes advantage of the infrastructure present in today's circuit switched networks. This method also enables delivery of guaranteed QoS so that each call is guaranteed the bandwidth it needs – bandwidth allocation of VTGs can be engineered to the prescribed call blocking level. The challenges to this approach include scalability, link efficiency, and operational costs. The number of VTGs grows as n^2 where n is the number of edge routers in the network. VTGs capacity allocation may need to be re-engineered as traffic changes. One solution to the n^2 problem is to create a hierarchical architecture where a core network is fully meshed and is fed by the access VTGs. A path then consists of three segments: access VTG, core VTG, and egress VTG. There are now three separate admission control decisions. The VTG approach has been patented and was recognized by the *MIT Technology Review*[9] as one of the top five patents of the year in 2003.

Although the VTG/accounting-based methods to monitor bandwidth utilization on a path are analogous to counting calls on a trunk group, there are a few additional items that need to be incorporated with VoIP. The bandwidth per call depends on the voice codec and sample size used, the protocol used for layer 2, and the voice activity factor (if silence suppression is enabled), etc. In addition, the accounting needs to keep track of mid-call codec changes such as those that might occur for a fax call. If a message gets lost, then the accounting is off by one and network resource may be lost forever. Similar to a circuit-switched network, an auditing method needs to be run on a periodic basis to reconcile any discrepancies.

D. *Link Bandwidth Reservation with Measurement-based CAC*

Yet another attractive alternative for VoIP QoS is a variation of the VTG approach. For IP/MPLS networks, paths can be set up as before, but bandwidth is managed differently. In this approach the LSPs (Label Switched Paths) all have zero bandwidth reserved for them. Instead, a DiffServ[1] class is reserved for voice on each link in the network and all LSPs that use a given link fully share the bandwidth on that link that was allocated for voice. This approach has several potential benefits. It simplifies operational costs because individual paths no longer need to be sized and re-sized on a regular basis – instead, only the link bandwidth needs to be managed. More importantly, since the paths all share the link bandwidth, statistical multiplexing is naturally accounted for at the links and allows for a more efficient use of the bandwidth. Houck and Uzunalioglu[10] describe some models indicating as much as a factor of two differences in bandwidth requirements between the two approaches. In addition, the link-based approach automatically takes care of non-coincidence of traffic, which can especially be a factor in long distance networks, or when business and residential traffic coexist. Further, measurement-

[1] DiffServ refers to the IETF (Internet Engineering Task Force) standard on Differentiated Services (RFC 2475, "An Architecture for Differentiated Services".

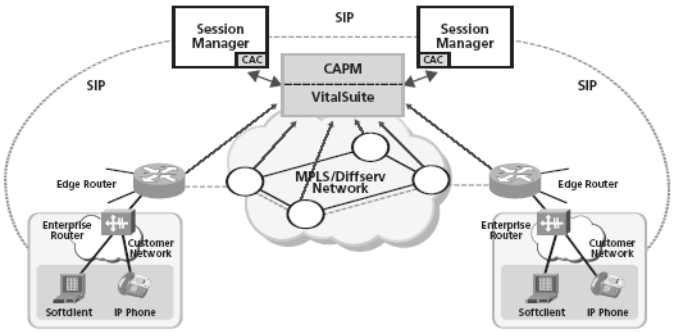

Figure 4. Link Measurement-based Management with the CAPM

based CAC is self correcting. Effects of errors or lost messages are automatically incorporated at the next measurement point.

While link-based management would be more efficient, a method must be defined to make use of the measurements in a way that scales to large networks. One such method is to measure the utilization of the voice traffic class on each link periodically. If congestion is detected, then all the LSPs that use that link are identified and blocking policies for each of those paths created. These policies will be of the form block **X** percent of the calls on this path. These path policies are then distributed to the appropriate session managers to implement the policy on a call-by-call basis. Houck et. al [11]. describe the method in more detail and give some simulation modeling results that show how the algorithm performs under different overload scenarios.

Figure 4 illustrates one implementation with IP endpoints. The Call Admission Policy Manager (CAPM), a Lucent Vital Suite Performance Management add-on module, is displayed. CAPM discovers the LSP routing through the network and the DiffServ bandwidth allocation on each link. On a periodic basis, it polls each link for utilization of the voice traffic class. Simulation has shown that monitoring a link every 30-60 seconds is sufficient to provide excellent voice quality while blocking a near minimum number of calls. This approach scales well compared to the path-based method described earlier. The algorithm parameters can be optimized to achieve the best performance depending on the service provider's policy. Engineering and tuning of the link bandwidth allocation can be applied to achieve the prescribed call-blocking ratio.

E. Considerations in CAC Selection

There are many factors that must be considered when deciding what type of CAC makes sense in a given service provider network. The particular signaling protocol used in the network may have some impact on how and where a CAC is applied. The capabilities available in vendor products may also limit what CAC options can be used. In some cases, the right choice is to start with an initial solution, based on network size and vendor capabilities, and evolve the architecture to a more robust and optimum solution over time.

One important factor is to determine network performance requirements. Are the performance requirements based on measurements averaged over some long interval or short interval? Are there tail probabilities associated with these measurements? For example, is it permissible to have a solution that works well 99.99 percent of the time, but fails completely that other .01 percent? Remember, overloads in packet networks can cause every call in progress to become incomprehensible, whereas in circuit networks, new call setups are blocked. The average and worst-case performance requirements will impact the CAC chosen. Traffic loads, voice/data mix, and network scale will influence the selection of the CAC approach. The estimated traffic loads and, in particular, the worst case traffic loads that might occur as compared to the network capacity are important. The load mix of voice and data, along with the QoS requirements for the data applications, will make a difference. The scale of the network, including the number of edge routers and the total traffic load in the network, impact the kind of CAC approach selected. The traffic growth forecasts will influence the initial, as well as the evolution of the CAC architecture.

There are also a number of issues that should be investigated after deployment. Once a CAC approach is deployed, the performance of the approach should be monitored through network management systems to tune and optimize the parameters of the CAC function and architecture.

Another issue is support of other real-time services, such as video, and the need to provide a CAC for these services. On a longer time-scale, these CAC details need to feed into a capacity management system so that the necessary bandwidth is available when needed. Although we have focused on single domain QoS, frequently there will be the need to handle inter-carrier VoIP traffic. In these cases there must be agreement on signaling protocols and QoS/SLA requirements. When crossing network boundaries, it usually makes sense to assume that each domain will be responsible for its own QoS and that each domain has implemented its own CAC. For SLA agreements to provide end-to-end QoS, requirements, such as one-way delay and packet loss, need to be allocated among the domains.

In summary, CAC is required in a network to provide optimum QoS. Since there are many CAC alternatives and issues that interact in complex ways, a rigorous methodology must be used to evaluate the alternatives and to select the best approach.

REFERENCES

[1] D. Houck and W. Yang, "VoIP – Voice Quality of Service,"

[2] B. Doshi, D. Eggenschwiler, A. Rao, B. Samadi, Y. T. Wang, J. Wolfson, "VoIP network architectures and QoS strategy," *Bell Labs Technical Journal,* Volume 7, Issue 4, 2003: 41-59.

[3] J. Deshpande and D. Houck, "VoIP Network Design for Service Providers,"

[4] D. Houck and G. Meempat, "Call Admission Control and Load Balancing for Voice over IP," *Performance Evaluation 47* (2002): 243-253.

[5] R. S. Krupp, (1982). "Stabilization of alternate routing networks" *IEEE International Conference on Communications*. Conference Record, volume 2 (of 3): 3I.2.1-3I.2.5.

[6] D. Houck, E. Kim, H. Uzunalioglu, L. Wehr, "A measurement-based admission control algorithm for VoIP," *Bell Labs Technical Journal,* Volume 8, Issue 2, 2003: 97-110,

[7] B. Doshi, E. Hernandez-Valencia, K. Sriram, Y.T.Wang and O.C.Yue, "Protocols, Performance, and Controls for Voice over Wide Area Packet Networks," *Bell Labs Technical Journal*, Volume 3, Issue 4, 1998: 297-337

[8] B. Doshi, D. Eggenschwiler, A. Rao, B. Samadi, Y. T. Wang, J. Wolfson, "VoIP network architectures and QoS strategy," *Bell Labs Technical Journal*, Volume 7, Issue 4, 2003: 41-59.

[9] D. Houck and H. Uzunalioglu, "VoIP: Link-Based Management in MPLS Networks," *MPLS World Congress* (Paris, France, February, 2005).

[10] D. Houck and G. Meempat, "Call Admission Control and Load Balancing for Voice over IP," *Performance Evaluation* 47, (2002): 243-253;

[11] D. Houck and H. Uzunalioglu, "An Architecture and Admission Control Algorithm for Multi-Priority Voice over IP (VoIP) Calls," *Proceedings of the 9th International Conference on Intelligence in Service Delivery Networks* (Bordeaux, France October, 2004).