Performance Analysis of Measurement-Based Call Admission Control on Voice Gateways

Feng Cao Hanlin Fang Mary Conlon
Packet Telephony Division, Cisco Systems, INC.
170 West Tasman Drive, San Jose, CA 95134, U.S.A.
Email: {fcao, hfang, meconlon} @cisco.com

Keywords: Call Admission Control, Quality of Service, Voice over IP

Abstract

Quality of Service (QoS) is critical for the success of real time applications over IP, such as Voice over IP (VoIP). On voice gateways, Call Admission Control (CAC) plays an important role for guaranteed QoS, for it makes decisions on whether and how to deliver the traffic based on different kinds of resources. In our study, we propose the measurement-based CAC model based on the various system resources on the local voice gateways. The end-to-end network congestion is also considered along with the configurable busy out on the voice gateways. The performance data is provided to show the improvement on a set of parameters for better Quality of Service and better serviceability of voice gateways.

1 Introduction

The traditional circuit switch infrastructure for telephony services will be augmented by packet switch infrastructure in the near future. Transport of voice and data across Internet has been integrated by both enterprises and service providers. There are many benefits for choosing Voice over IP. For example, the low cost of IP can save the customers more than the expensive circuit switches. More service can be easily created and delivered in IP than the current telephony infrastructure.

Voice gateways play an important role for carrying voice over IP. The voice comes into the ingress voice gateways through T1, E1, or POTS and is streamed into Voice over IP (VoIP) packets that is routed to the egress voice gateways.

VoIP applications are different from data services. As they are real-time and interactive, there are strict requirements on the delay and the jitter for end-to-end delivery. Therefore, Quality of Service (QoS) is critical for providing the expected behaviors of VoIP applications. In most of cases, it is impossible to imagine that voice calls or fax calls can be delivered without reasonable delay and loss for customers.

In this paper, we study the QoS issues on voice gateways through Call Admission Control (CAC). To guarantee QoS, CAC must be provided on voice gateways, allowing the voice traffic to be accepted when and only when expected performance can be assured before the voice traffic enters the voice gateways. Many factors may be considered in CAC, such as interface bandwidth, system resources of gateways, the network conditions, and configured policy control.

In the following sections, we show how system resources on voice gateways help to guarantee the QoS of voice traffic. System resource module is shown for call admission control and traffic engineering. Network conditions are important for real-time streams. We provide the end-to-end probing module to detect connectivity and congestion for delivering the traffic. All the modules discussed here can be used as a part of CAC to guarantee QoS, some procedures are recommended for integration based on the experience on H.323 VoIP calls in this paper.

2 System Resource Availability

System resources refer to the common resources on VoIP gateways in this paper. To be more specific, CPU, memory and call volumes are discussed in this study for providing better QoS for VoIP applications.

Voice gateways may be overloaded in some extreme cases. CPU utilization may go over 99% if huge bursty traffic is coming into voice gateways simultaneously, such as a large number of fax calls or Interactive Voice Response (IVR) calls. The memory consumption may be quite high if multiple IVR calls or fax calls are in process. Similarly, only a certain number of calls can be handled by voice gateways, otherwise the performance of voice gateways will be decreased.

In order to prevent the extreme cases from affecting the performance of all other processes on voice gateways, we provide the module for system resource measurement, and measure-based call admission control with per-call treatment and voice gateway busyout. The performance result is presented to demonstrate better serviceability and availability on voice gateways.

2.1 Two-threshold model

The two-threshold model is proposed here to catch abnormal cases. Namely, it has low and high thresholds. Whenever the current value is over the high one, the model remains in the unavailable state until the current value drops below the low one. For example, if CPU utilization is configured as [70%, 90%], and the current value is 92%, which is over 90%, that means that CPU is in unavailable state until the current value drops below 70%.

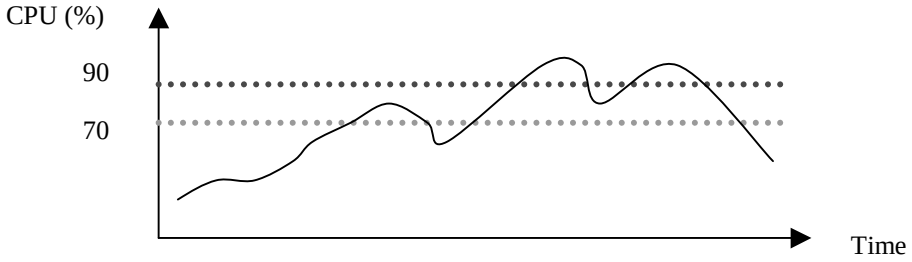


Figure 1: Two-threshold example

There are many advantages for this model. First, it is generic for modeling different resources. For call volume, the thresholds can be used as the number of calls. For memory, the thresholds can be either percentage or the absolute byte numbers. Another advantage of using two-threshold model is to avoid the spiking condition of some system resources. Moreover, the two-threshold model is a super set of one-threshold model. For a example, if CPU utilization is configure as [90%, 90%], it is the same as one-threshold model with the threshold defined as 90%.

On the other hand, it needs careful configuration for different gateways if resource-based call admission control is enable in the rest of this section. The proper values for low threshold and high threshold should depend on the administrator's requirement and expectation. They may be different among configured resources on different gateways, and have respective impact on the performance of voice gateways.

2.2 More features on call volume

Besides the option for extreme cases, call volume may also be used for traffic engineering. For example, the users can specify the different two thresholds for call volumes for the multiple access Voice gateways with the busyout enabled on each of them. This will help the switch to send the calls to the available access gateways

instead of continuously delivery of calls to the unavailable ones. Better load-balancing can be achieved if multiple gateways connected to the same switch have CAC enabled.

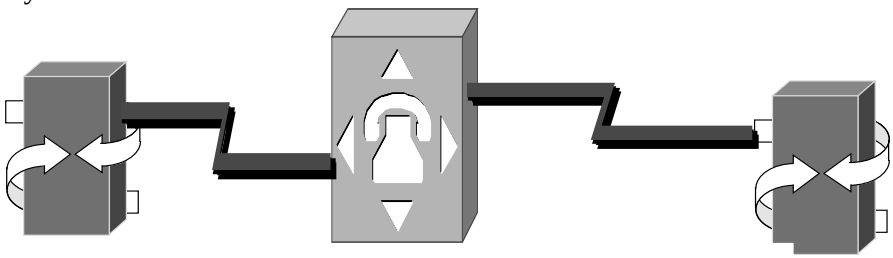


Figure 2: load-balancing example by call volume control

2.2 Resource-based CAC

Whenever the high thresholds are crossed on the configured system resources, including CPU average utilization, memory consumption and call volume, this will trigger the admission control module. We provide two options:

- *per-call treatment*: the new calls will not be accepted and be treated as the configured behavior, such as playing message saying "Please try another number ...." or playing different tones.

- *system denial*: the PSTN interfaces of the ingress gateways will be busied out to inform the PBXs or the switches to not send new calls to them.

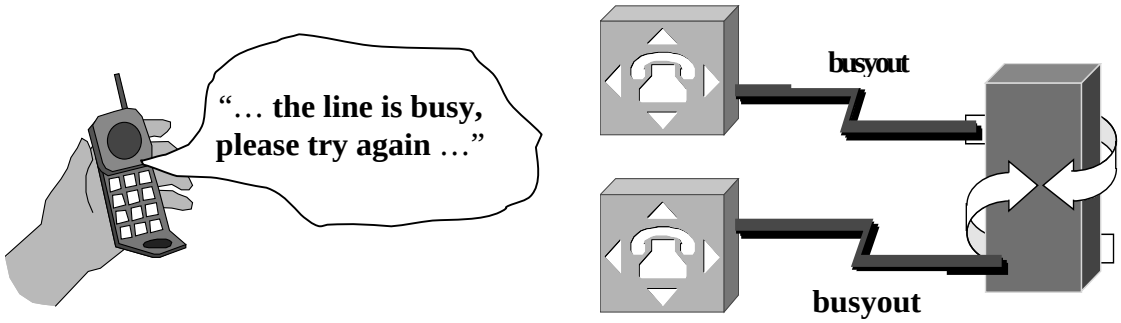


Figure 3: per-call treatment vs. system denial

By adding these options, system resource availability becomes a part of call admission control to guarantee the performance of VoIP applications. By per-call treatment or system denial, PBXs or the switches may reroute that call through the paths with good quality.

2.3 Performance Analysis

In our performance analysis, we chose system denial as the method for busying out the calls from coming into the voice gateways. Theoretically, given that the two-threshold model is used and the switches are blocked from sending new calls to worsen the unavailable situation, this should improve the call success ratio after the calls are connected.

2.3.1 Test topology

Figure 4 and 14 show the test topology on system resources concentrate on the originating voice gateways.

2.3.2 Overall CAC overhead

With the measurement turned on, the load for providing the measurement data and configured actions for busying out the trunks must be considered to ensure the introduced CAC load doesn't worsen the performance of voice gateways. The data in Figure 5 shows the load of CAC overhead can be ignored.

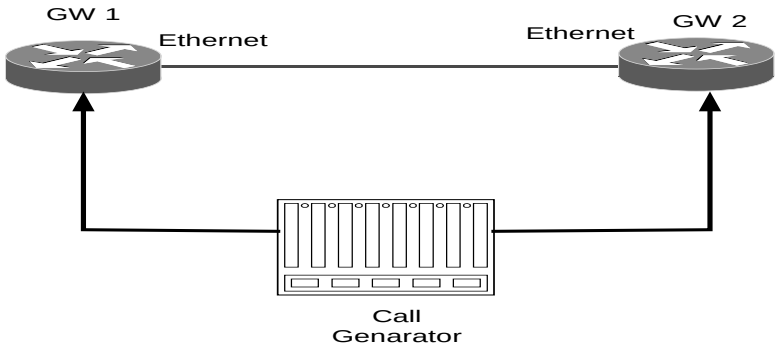


Figure 4: Overall CAC overhead Topology

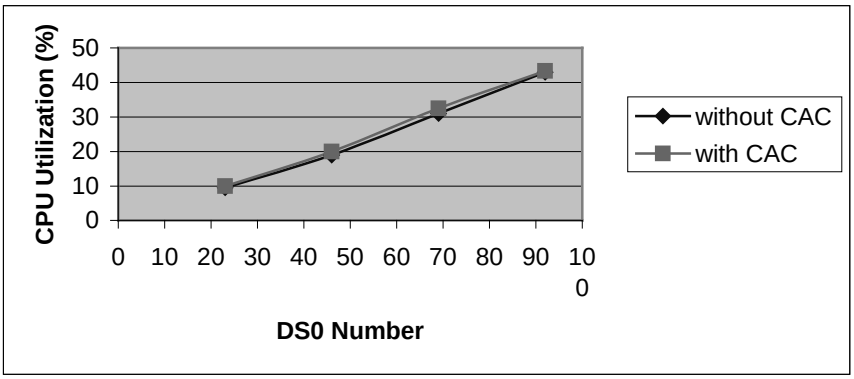


Figure 5: the CPU overhead for CAC

2.3.3 Call Success Ratio, delay, …

There are parameters that demonstrate that CAC based on system resources can provide the better serviceability and the availability. One of them is Call Success Ratio (CSR), which is the ratio of the final successful calls to the calls with successful setup.

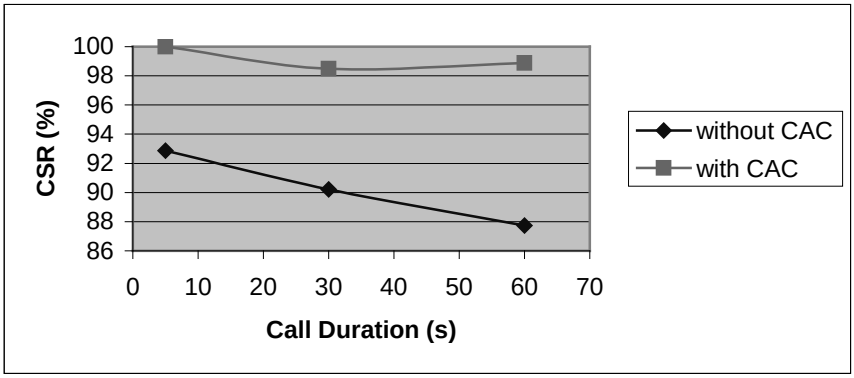


Figure 6: CSR comparison with different call durations

The tests here focus on the parameters on the originating gateways with the bursty data. Figure 6 demonstrates CSR is consistent and acceptable with CAC enabled. Without CAC enabled, more calls fail after successful call setups and CSR is dropping to unsatisfactory level.

Round Trip Delay (RTD) in this paper is measured by the call generator on the delay for voice path confirmation. Figure 7 shows the improvement of RTD is much better when CAC is enabled.

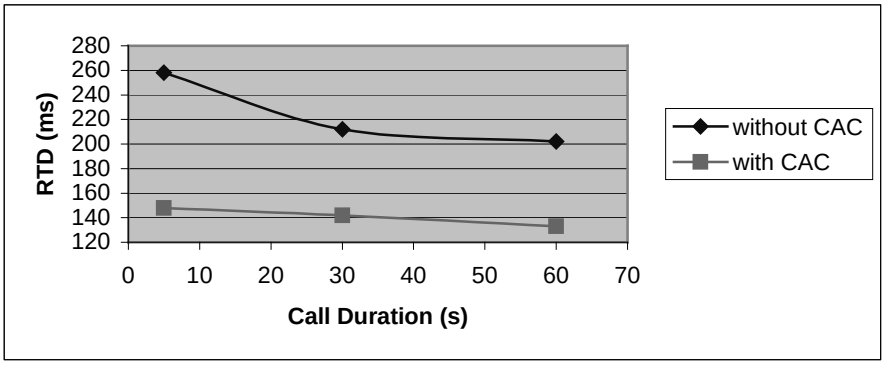


Figure 7: RTD comparison with different call durations

3 Network congestion measurement

Given the concern about the scalability of RSVP and some existing old Internet infrastructure, RSVP hasn't been deployed in all the routers in the current Internet. That implies that VoIP cannot fully rely on RSVP in many scenarios, at least at the current time.

QoS could be based on measurements for VoIP. Two of important factors in choosing the VoIP routes for QoS are network connectivity and availability.

If the network is down along the paths from the source to the destination, it's better to stop all the incoming VoIP calls and reroute them through traditional PSTN if there is no IP route. The same strategy applies the congested network. If the congested network cannot guarantee the QoS of VoIP calls, it's better stop new incoming VoIP calls and wait for the recovery from network congestion.

The most used parameters in determining QoS for VoIP applications are loss, delay, jitter and ICPIF (ICPIF short for Calculated Planning Impairment Factor, see ITU-T G.113). Many service providers need this kind of measurement as a part of CAC for VoIP applications.

3.1 Probe-based measurement

There are many ways for network measurement and management. One of them recently provided by Cisco is specific for VoIP applications through most of Cisco gateways. Parameters such as jitter, delay, and ICPIF are obtained through Response Timer Response (RTR) probes.

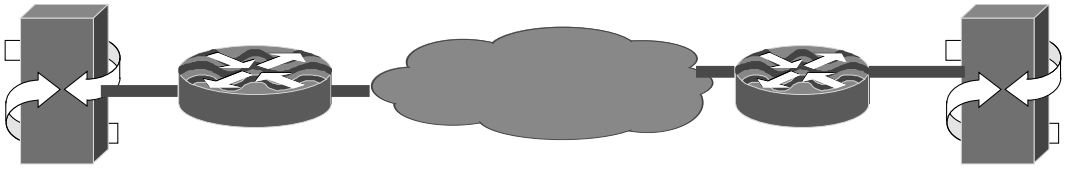


Figure 8: end-to-end network condition

RTR is the enhanced module for point-to-point probes. In addition to the traditional ICMP echo probes, RTR provides more features. One of them is to use configured IP ToS fields for ICMP echo probes. Another is to support new probes for UDP and TCP response time measurement. For the later, the far end must enable a RTR responder to listen to a UDP port (port 1976) for RTR control message authentication.

To make the probes more accurate, RTR probes allow the users to define the packet size, the number of packets and the interval between consecutive packets. First, VoIP packets are RTP/UDP, which can use RTR's

UDP response time. Second, based on the VoIP codec, we can use the proper packet size to simulate the actual voice packets. As loss, delay and ICPIF may rely on the codec, RTR probes provide a better way to discover QoS and the voice quality.

Based on the RTR probes that simulate the VoIP packets, there are usually two scenarios for service providers. One is the probes show the network is disconnected from the source to the destination. The reasonable action in the VoIP gateways is to busy out its proper PSTN interfaces, which prohibits the switches to send more VoIP calls to the same destination. As soon as the RTR probes indicate the network is connected again, the PSTN interfaces should be brought up again.

The other is the probes show loss rate, delay or ICPIF is too bad to provide the expected QoS for new traffic. An option is provided for the users to drop the new incoming VoIP calls to that destination whenever the configured thresholds are crossed for loss, delay or ICPIF.

There are some drawbacks for RTR probes. One is that they are asynchronous probes, which may not reflect the dynamic changes of network performance. Aother is that the probing packets are not the voice packets, which may not reflect the real treatment of voice packets from the source to the destination. Probe packets add extra load on the network, especially when the network is already congested. Therefore, this approach makes more sense for detecting network connectivity and then triggers the system denial on some voice interfaces.

3.2 RTR-based CAC

Whenever the probing results to the desired destinations are below the configured expections (such as delay, loss, or icpif), this will trigger the admission control module. We provide two options:

- *per-call rejection*: the new calls will not be accepted and be denied with configured cause code, such as no QoS available.

- *Selected system denial*: the users can select certain PSTN interfaces of the ingress gateways to be busied out to inform the PBXs or the switches to not send new calls to them until the probing results turn good.

.

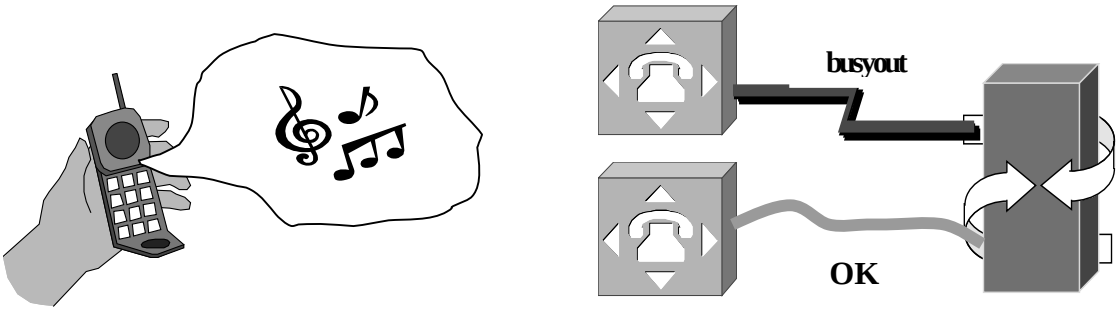


Figure 9: per-call rejection vs. configurable trunk busyout

Note that the selected system denial provides the flexibility of sharing the gateways for different ISPs or different users. For example, user A owns ISDN interface One and delivers calls to New York, and user B owns ISDN interface Two and delivers calls to Los Angels. A can configure to busy out ISDN interface One preventing calls from the switch if the network condition to New York is below expectation.

By adding these options, system resource availability becomes a part of call admission control to guarantee the performance of VoIP applications. By per-call rejection or selected system denial, the PBX or the switch may reroute that call through the paths with good quality.

3.3 Performance Analysis

In our performance analysis, we choose selected system denial as the method for busying out the calls from coming into the voice gateways.

3.3.1 Test topology

The tests on probing the network condition concentrate on the originating voice gateways.

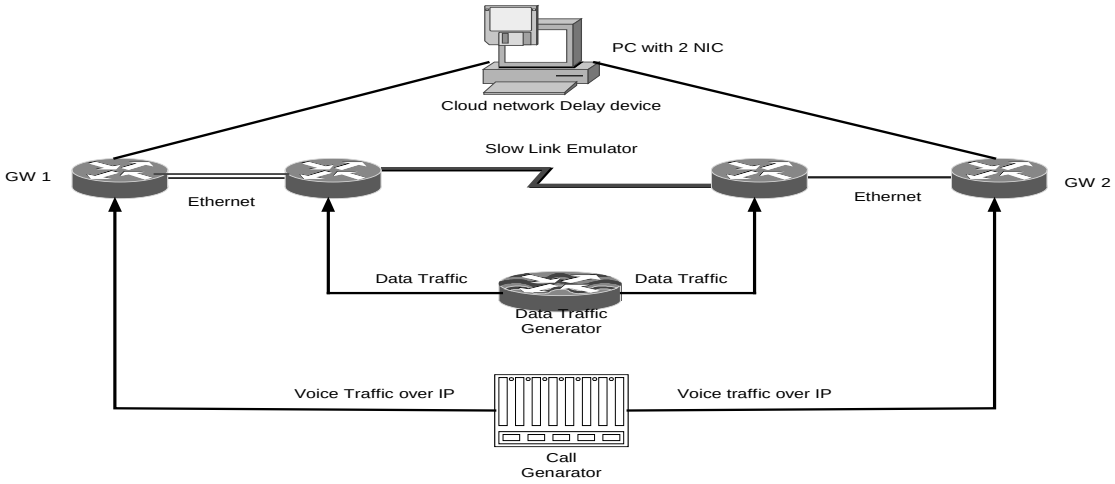


Figure10: Congested Network Topology

3.3.2 Overall CAC overhead

With the measurement is turned on, the load for providing the probes and configured actions for busying out the trunks must be considered to ensure the introduced CAC load doesn't worsen the performance of voice gateways.

It's not surprising that CPU utilization and memory consumption are actually lower with CAC enabled. As the system denial is functioning when the network condition is unsatisfactory, the switches are blocked from sending new voice calls into the gateways. Therefore the gateways use less systems resources.

The concern for the probing approach is the extra probing traffic introduced. This could worsen the congested networks or could not scale if the gateways want to deliver calls to a large number of destinations belonging to different domains. The extra load also depends how large the probing packet is. For example, if the probes, in G729, are updated by 10 packets every 10 seconds, then the bandwidth required is (20+12+8+20)*8*10/10 = 480 bps.

3.3.3 Call Success Ratio, delay, …

In this subsection, we demonstrate the RTR probes can provide the better serviceability and the availability through preventing voice traffic from entering congested networks.

In Figure 12, we compare the CSR in different traffic patterns in the tests. The percentage mentioned below is about the bandwidth of the slow link in the test topology. In Traffic pattern 1, the network traffic through the common link is 400 kbs voice and 400 kbs data. In Traffic pattern 2, 800 kbs voice and 800 kbs data. In Traffic pattern 3, 1100 kbs of voice and 1400 kbs of data. In Traffic pattern 4, 1100 kbs of voice and 1800 kps data. Note that voice traffic means the traffic the switches want to send to the voice gateways . With the CAC enabled, the channels are busied out and the voice traffic cannot enter the voice gateways.

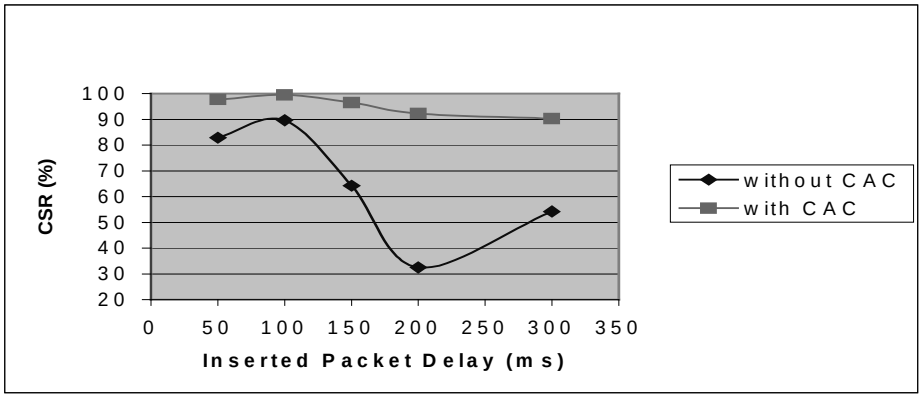


Figure 11: CSR comparison with packet delays

Figure 12 demonstrates that the CSR is quite good and consistent under different traffic patterns with CAC enabled. The reason for that is that voice gateways inform the switches of backing off whenever the network is congested. On the other hand, without CAC enabled, the voice calls keep entering the IP world even if the network is congested, which will introduce a lot of call failures due to packets drop such as no voice path failures.

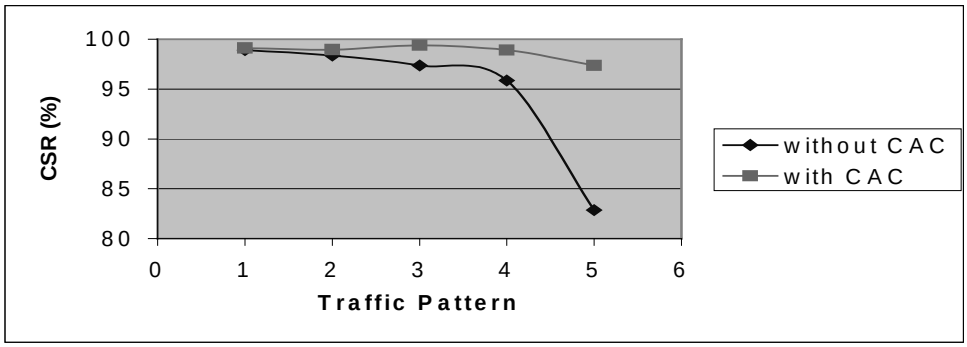
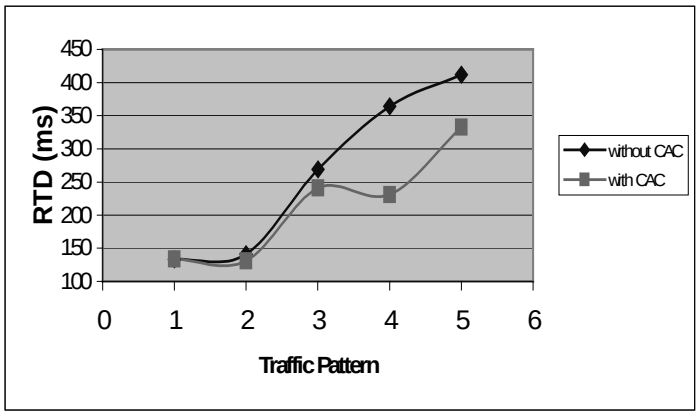


Figure 12: CSR comparson with different traffic patterns

Figure 11 shows the similar results when the certain number of delay is introduced in the network. Without CAC, the call failure rate is unacceptable even if the calls are connected. With CAC, the call success ratio is quite better.



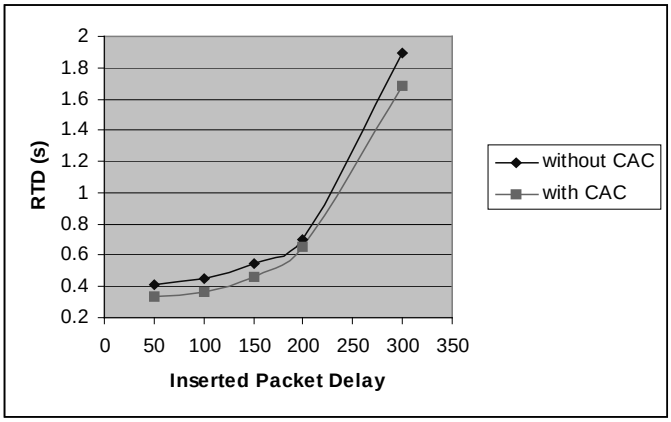


Figure 13: Round Trip Delay comparison with different traffic patterns

Figure 13 provides the same information about the improvement of RTD with CAC enabled. The reason is that calls are blocked when the network is congested. So RTD is better for calls when the network condition gets better.

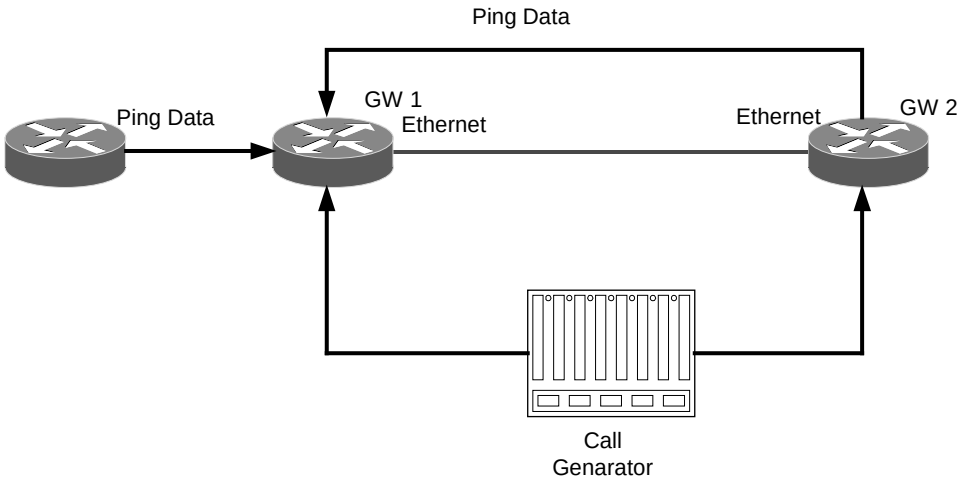


Figure14: Mixed Traffic Topology

4 Integration for CAC

In this paper, we demonstrate some modules for CAC on Voice gateways. With the help of these module, improved QoS will be provided for end-to-end VoIP applications. Each module is independent of each other, and can be used in different places. Based on our experience on VoIP H.323 calls, one procedure we recommend is

when the call is from telephony side,

1. The system resource should be checked first to decide if enough resource is available on the gateways. If yes, continue. If no, do call treatment if configured.
2. The outgoing interface resource should be checked by LCAC module for bandwidth and call volume. If yes, continue. If no, reject the call.
3. If RSVP isn't enabled, use end-to-end network congestion measurement module. If yes, goto 5.
4. If RSVP is enabled, use RSVP synchronized module to reserve the bandwidth for both directions.
5. allow the call to continue.

Similarly, when the call is from IP side,

1. the incoming interface resource should be checked by LCAC module for bandwidth and call volume. If yes, continue. If no, reject the call.
2. system resource should be checked to decide if enough resource is available on the gateways. If yes, continue. If no, do call treatment.
3. If RSVP is enabled, use RSVP synchronized module to reserve the bandwidth for both directions.
4. allow the call to continue.

This procedure shows how these modules work together to provide better QoS through call admission control. Other procedures may be used based on the signaling protocol and vendors' favors, and more modules can be added for call admission control.

5 Conclusion and Future Work

QoS is critical for the success of real time applications over IP, such as Voice over IP. On voice gateways, call admission control plays an important role for guaranteed QoS, for it makes decisions on whether and how to deliver the traffic based on different kinds of resources.

In this paper, we study the QoS issues on VoIP gateways through measurement-based CAC. To guarantee QoS, CAC must be provided on VoIP gateways, allowing the voice traffic to be delivered when and only when expected performance can be assured at the time voice traffic enters the VoIP gateways. Many factors may get involved in CAC, such as interface bandwidth, gateway system resources, the network conditions, ...

We show how system resources on voice gateways help to guarantee the QoS of voice traffic. System resource module is shown for call admission control and traffic engineering. Network conditions are considered by several approaches to detecting connectivity and congestion. All the modules discussed here can be used as a part of CAC to guaranteed QoS, some procedures are recommended for integration in our paper.

As the rapid progress is being made in providing QoS, there are a lot of issues for better Call admission control. For example, with RSVP enhancement for aggregation, CAC should be provided based on the aggregation policy instead of per call requirement. How to integrate some of the above CAC modules with DiffServ and MPLS is another interesting topic.

References

1. F. Cao, H. Salama, and D. Shah, “Approaches to Providing Guaranteed Quality of Service for VoIP through Call Admission Control on Voice Gateways”, Proceedings of “International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet” (SSGRR 2000), July 31 – August 6, L’Aquila, Italy, 2000

2. S.Charttjerjee and J. Strosnider, A Generalized Admission Control Strategy for Heterogeneous, Distributed Multimedia Systems, Proceedings of the Third ACM International Multimedia Conferences, San Fransisco, 1995

3. J. Huang, Y. Wang, and F. Cao, On Developing Distributed Middleware Services for QoS- and Criticality-Based Resource Negotiation and Adaption, Journal of Time-Critical Computing Systems, 16, pp 187-221, 1999

4. J. Huang, Y. Wang, S. Vaidy, and F. Cao, GRMS: A Global Resource Management System for Distributed QoS and Criticality Support, Proceedings of the IEEE International Conference on Multimedia Computing and Systems (MMCS'97), Ottawa, Canada, 1997

5. D. Hutchison, G. Coulson, A. Campell, and G.S. Blair, Quality of Service Management in Distributed Systems, Distributed Systems Management, Ed. Morris Sloman, Imperial College London, 1995.

6. H. Kaneko, J.A. Stankovic, S. Sen, and K. Ramamritham, Intergrated Scheduling of Multimedia and hard-real-time tasks, Proceedings of the IEEE Real-time Systems Symposium, December, 1996.

7. A.M. van Tilborg and G. Koob, Foundations of Real-time computing-Scheduling and Resource Management, Kluwer Academic, 1991.