RC22822 (W0306-089) June 17, 2003
Computer Science

IBM Research Report

Hybrid Model for VoIP Embedded Application

Zon-Yin Shae, Xiping Wang
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Hybrid Model for VoIP Embedded Applications

Zon-Yin Shae, Xiping Wang
IBM Watson Research Center
{zshae, xiping}@us.ibm.com

Abstract

The enterprise telephone system, regardless it is a regular Telco based phone system or the VoIP phone system, currently is independent of the enterprise applications. It is very desired to closely integrate the voice communication as an integral part into enterprise applications. The combination of VoIP softphone and SIP protocol has been promised to provide a link for such purpose, but with a limited acceptance of enterprise usage. Besides the voice quality (delay, jitter, echo), user interface and user experience reasons (talk to PC), it is mainly due to the current IT infrastructure (e.g., NAT, VPN and firewall) unfriendly to both SIP control messages and media data, which is posing challenges to its pervasive availability. This paper proposes a hybrid system architecture for seamlessly integrating the voice communication into the enterprise application using the standard SIP infrastructure and the existing telephone infrastructure. This hybrid system embeds the voice communication control in the enterprise applications using the standard SIP protocol, enabling enterprise applications to closely integrate regular telephones and VoIP hard phone devices for voice communication. Built on this hybrid architecture, an illustrated VoIP enabled instant messaging enterprise application is also prototyped and reported.

1. Introduction

In the past few years, SIP (session initiation protocol) [1] based IP phone has increasingly been adopted by enterprises for cutting the telephone cost and introducing new applications. Many corporations have built a network infrastructure for supporting IP phones and further bridging regular phones and IP phones. With such an infrastructure and high bandwidth of today's enterprise intranet, a reliable voice IP communication can be achieved within a corporate.

Although the combination of VoIP softphone and SIP protocol has long been promised to provide a required bridge for integration of the voice communication into the enterprise applications, only a very limited usage has been adopted within enterprise. It is mainly due to the voice quality, user not used to talk to PC, and the IT infrastructure issues. First, voice communication is very sensitive to the round trip delay and echo. According to ITU studies [4], an acceptable round trip delay for voice communication must be less than 300 ms. A PC based voice communication system, in addition to voice data transmission delay, includes delays for voice signal capture/playback, encoding/decoding, packetization/depacketization and data transformer for processing. If these real time operations are all handled in PC software, a much longer delay usually occurs compared to a dedicated DSP based IP phone sets or regular phones. Also, the dedicated hard phone devices already have proved comfortable solution to echo. Second, the subtle user interface is another issue. People are used to make calls with a regular phone rather than PCs. It gives them an uncomfortable experience when speaking to a PC. Lastly, work from home and traveling users need to remotely access their corporate Intranet normally via Virtual Private Network (VPN). A VPN [2] channel is secure and nowadays is the most popular way to create a tunnel through firewalls for enterprise to support their employee working from home or traveling. VPN functions like NAT [3] to translate the end point's local IP address to a routable IP address within the corporate Intranet domain, which creates great difficulty for a large numbers of Internet applications. This NAT problem applies to SIP protocol as well. SIP makes use of control messages to communicate between end points about its IP addresses and port numbers that will be used later for both its control signaling and media data. NAT creates problems for completing the SIP protocol and forbids its associated media applications. The problem of NAT to SIP can be categorized into two parts: control path and media data path.

Since SIP application needs to complete both paths, all the current proposals all try to provide solutions allowing both control and media data paths through the NAT. However, the problem to allow both paths through NAT is so complicated (as described in section 2), and there is not yet a clean and universal solution exists. This paper proposes a practical solution to overcome these problems and enables the seamlessly integration between voice communication and enterprise applic ations.

We present in this paper a hybrid architecture to separate the SIP control path and media data path, and to process these two paths in two separate infrastructures. SIP control path will need to traverse through VPN/NAT, but not the media data path. The media data path will go through the traditional Telco phone switch. This way, the separated SIP control signal can be seamlessly embedded inside enterprise applications to enable the voice communication. Currently, the enterprise telephone system is independent of the enterprise applications. It is very desirable to closely integrate the voice communication as an integral part of enterprise application. This hybrid architecture using SIP protocol provides a sound basis for embedding a reliable voice communication for enterprise applications, also results in higher voice quality and better user experience provided by existing Telco infrastructure. As an example, this paper also presents an experimental enterprise collaboration application built on the proposed hybrid infrastructure, for illustration purpose, using IBM SameTime [12], SIP infrastructure, and existing Telco telephony infrastructure supported within IBM Research.

The paper is organized as follows. Section 2 describes the complexity of VoIP phone over NAT. Section 3 describes the proposed hybrid architecture for voice embedded application. Section 4 describes a collaboration system using the proposed hybrid architecture. Section 5 discusses technical issues and solutions, followed by conclusion section given in Section 6.

2. Complexity of VoIP Phone over NAT

There has been a profound study for the various mechanisms on the topic of NAT traversal for SIP. The control path traversal through NAT can be achieved much easier than the traversal of media path. Numerous proposals for SIP control path and media path traversal exist, e.g., Simple Traversal of UDP through NATs (STUN) [5], Traversal Using Relay NAT (TURN) [6], Realm Specific IP [7], Application Layer Gateways (ALG) [8], Middlebox Control Protocol (MIDCOM) [9]. Each proposal can solve only within its own particular network environment, e.g., STUN does not enable incoming UDP packets through certain type of NAT and requires SDP extensions, ALG serious scalability and is application dependent, MIDCOM does not work with the existing NAT. Since the entire network scenarios are so complex, even IETF has created a document to describe the scenarios and its possible solution for a given scenario [10]. Recently, an Interactive Connectivity Establishment (ICE) [11] is proposed to integrate all the above proposals into a single solution. The result is a very complex system and better solution is still in need.

The basic issue is to maintain a set of opened NAT connections and reuse the connections for carrying both SIP signaling and associated media. A NAT connection can be opened both by TCP and UDP from the residential user behind the NAT, and not from external. If it is opened by TCP, the connection is persistent as long as the TCP remains connected. If it is opened by UDP, it will be closed normally within one minute [10]. Residential NAT normally perform not only IP address translation, but also port address translation (PAT). This is required due to the IP address assigned by ISP to this NAT is very limited, usually only single one. The residential NAT needs to make use of PAT to support multiple internal IP addresses and single external IP address

Some mechanisms are previously proposed to open and maintain NAT connections for SIP usage. One approach is to install a special server, e.g., STUN, at the ISP side. Special STUN client running in PC will make a proprietary query the STUN server, and the server will report the gateway IP address and port number back to the client. The other approach is the double registrations. At the first registration, the mapping is obtained from the registration response. Then it can be used in the contact header for the second registration. Very high periodical register request rate (about once every minutes) is required if UDP is used to maintain the NAT connection open. Both above mentioned mechanisms need to make use of "rport" parameter in the "Via" header of SIP NAT extensions. If the Proxy supports it, this opened NAT connection can be reused for SIP traffic across NAT. If the SIP Proxy does not support this extension, then both mechanisms are invalid. Moreover, the STUN protocol still needs to open two additional NAT connections (for RTP and RTCP) for each media steam created during the call signaling process. RTCP is optional as call can proceed without it. In general, all proposed solutions require multiple STUN like out of band queries, and also requires SDP extensions.

3. Hybrid Architecture for Voice Embedded Applications

We propose in this paper to separate the SIP control path and media data path. SIP control path will need to traverse through NAT, but not the media data path. The media data path will go through the traditional Telco phone switch. SIP control path enables the voice control signaling to be embedded inside enterprise applications. This architecture presents a very nature transition solution for integration of VoIP into enterprise applications until the IP phones completely replace the PSTN phones and the IT infrastructure (VPN, NAT, and firewall) completely accommodates SIP protocol and its Voice over IP (VoIP) UDP media traffic. Our emphasis is on the mechanism for traversing VPN as it has become the most popular configuration for enterprises to support their mobile employee working from home and on the business trip. VPN gateway establishes a secure tunnel (e.g. IP Sec) between user's machine and the VPN gateway through corporate firewall. The VPN tunnel hides the residential NAT from the user. Normally, from the user perspective, the VPN gateway (e.g., AT&T VPN) presents a NAT of persistent connection for both TCP and UDP traffic with only address translation and no PAT.

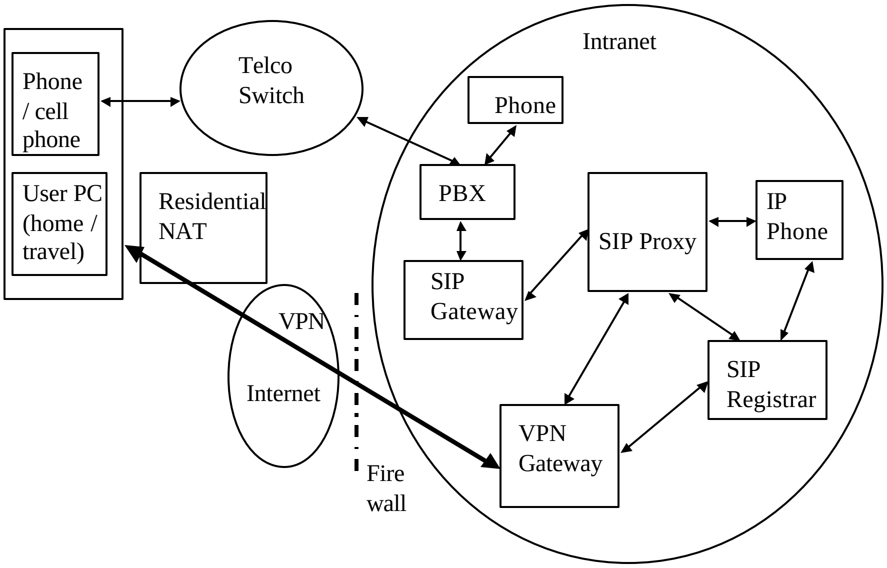


Figure 1. Hybrid Architecture for Embedde d Voice in Enterprise Applications

Figure 1 shows the hybrid architecture which consists of SIP service infrastructure (SIP proxy, SIP gateway, SIP registrar), VPN infrastructure (VPN gateway, VPN client agent), Telco phone switch infrastructure (Telco Switch, PBX), IP network infrastructure (Internet, Intranet, private network behind residential NAT), and end devices (regular phones, IP phone, and PC). All of these infrastructures are integrated together seamlessly by SIP and IP protocol. With this hybr id architecture and SIP protocol the voice communication services can be easily embedded in the applications.

3.1 Simplified SIP Registration Process

Every SIP end device needs periodically to register itself (e.g., once every hour) to the SIP Registrar. The purpose of this registration is to keep the up to date mapping between user's sip uri and its current IP address. The SIP Registrar will remember the mapping between the "From" (the user's sip uri) and "Contact" (the machine's IP address) headers. This register operation enables the registered user being able to be reached by his/her registered sip uri.

If the user's PC is behind a residential NAT, in a normal SIP registration process, its SIP user agent (UA) will register the "private" un-routable IP address only meaningful behind the residential NAT. When a VPN tunnel is established between user and VPN gateway, an IP address will be assigned by gateway to represent the "routable" address for that user within the Intranet. In order to make SIP control signaling go across VPN and communicate with others end devices within the Intranet, it is required for the user's SIP UA behind residential NAT to find out this assigned IP address and register it as its own contact address in the SIP protocol. This can be achieved by doubled registrations. The UA first send a regular registration message to the SIP registrar with it "un-routable" local IP address behind residential NAT. It is required by the standard SIP protocol to include a "Via" header in the registration message where the user's un-routable local IP address is set. This message might travel through a set of SIP proxies before it reaches the SIP registrar. The first proxy receiving the message observes it comes from a different IP address (due to VPN) that the "Via" header reported. This proxy will add a "received" tag to record the IP address that this message is received from before it forward along the message. Note that the IP address recorded in the “received” tag is the VPN gateway IP address assigned for this connection. The message will finally reach SIP registrar. The Registrar then sends a response back to the user (note that the “Via” mechanism in SIP protocol guarantees this response can traverse in an identical but reverse direction along all the proxies back to VPN gateway, and then tunnels to the user). After the user receives the registration response and observes that the “received” tag has a different IP address, the UA will know it is behind some sorts of NAT. The UA will send out a second registration message with this “received” IP address as its own contact address. Note that the procedure described here is completely based on the standard SIP protocol specification and requires no SIP extension. More importantly, it can work more pervasively in various environments.

3.2 Simplified SIP Call and Media Data Process

The call and media process in our hybrid architecture is very simple and quite universal as it makes use of the existing Telco switching infrastructure for the media. The SIP user agent at the PC sends the INVITE message to a phone that user would like to use for voice communication. This INVITE message might go through a SIP gateway and PBX to enter the Telco switch network if a regular PSTN phone (including cell phone) is used. The phone will ring, and a “200 OK” response will be sent back after user pick up the phone. By the SIP protocol design, this “200 OK” SIP response will traverse back to user behind the residential NAT (as described in previous section). After receiving the “200 OK” response, the PC user agent will send a “REFER” message to this answered phone and further direct it to make a call to the other phone chosen by application. As such, a voice communication is established and it is embedded inside enterprise applications. The complexity of media through NAT is therefore avoided.

4. Voice Enabled Enterprise Application Prototype

Communication is critical in today's competitive business world. Instant Messaging (IM) has transformed the way we communicate over the Internet. In the enterprise environment, IM has increasingly been adopted for routine business needs. For example, IM brings people at different locations of an enterprise to a unique virtual text chat environment so that they can share their thoughts and conduct businesses. By increasing corporate communications, businesses are able to anticipate market changes and respond in an effective fashion, being proactive rather than reactive. Consequently, reduce in person meetings, reduce meeting travel, and improve problem resolution time. Overall, cut meeting and communication delay expenses and hence increase business productivity.

During the course of text-based IM, from time to time, people in the serious business chat needs to have real time voice communication as well if more complicate thought exchange is required. The requirement of the dual modes, text IM and real time voice communication, is so strong that virtually every existing IM product (e.g., AOL, MSN, Microsoft NetMeeting, IBM SameTime) supports it. However, these products support voice communications by using PC multimedia capability, it is limited to a very small group of people due to long voice latency and poor voice quality. Voice communication is very sensitive to the round trip delay. Since these operations are all handled in software, a much longer delay usually occurs compared to DSP based IP phone sets. Especially in case of multiparty chat, all the participants' voice signal is sent to the central server, and the final fixed voice signal is then broadcasted to each participant after signal mix processing. This introduces a long delay that makes the quality of voice communication unacceptable. For this reason, these products only allows for a small group people to have voice communication. Even though, the quality of voice is still too poor for practical use. In practice, therefore, those systems are often used in a corporate environment only as text-based chat.

Manually dialing a separate PSTN telephone call is required in parallel for meeting participants to communicate with each other quickly and accurately.

4.1 Application System Configuration and Call Service Flow

Figure 2. shows the configuration of the experimental system consisting of a SameTime server, SIP proxy server, SIP/PSTN gateway, call service server, and a LDAP directory. A SIP infrastructure consisting of SIP proxy and SIP gateway has been built and become part of standard IT infrastucture within IBM Research, which can provide voice communication with a good quality. A SIP/PSTN gateway is installed to provide communication between VoIP and PSTN phones.

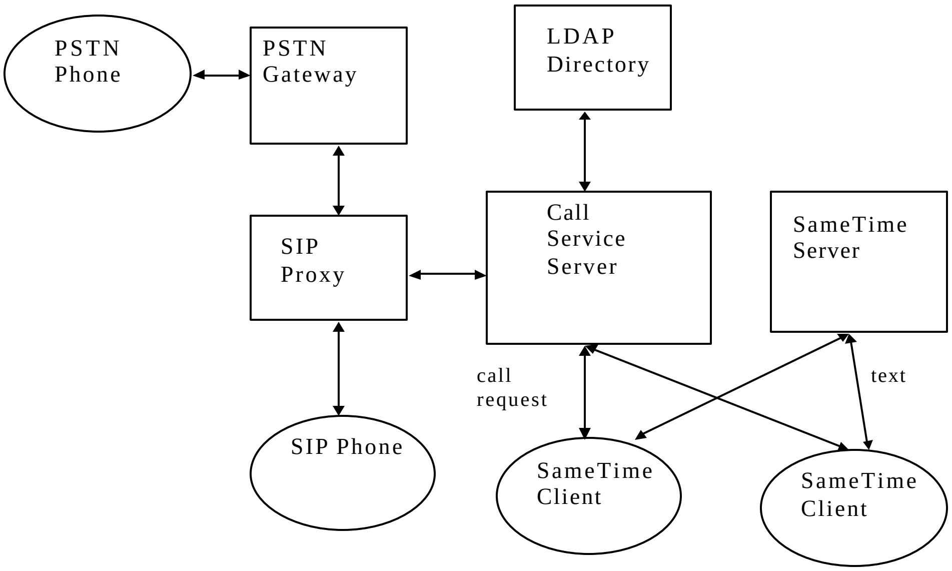


Figure 2. Application System Configuration

We implemented a special SameTime client and call service server to enable the hybrid mode of IM. The call service flow is shown in Figure 3. The SameTime server provides a virtual collaboration environment where multi-users can gather together for IM text chat. During the course of text chat, participants can initiate a voice call by sending the call service request to the call service server which performs the third party call service. Although the display name of participant in the instant messaging can be arbitary, it is required in the corporate environment that it would be unique and usually is the participant's email address for security and management reasons. Consequently, only the participants' email addresses are available to the call service server. The call service server can look up the company's LDAP directory for the participant's telephone number (more complex scenarios in supporting the mobile users is described in the section 4.3). The call service server then sends an INVITE to the SIP proxy to establish a call to his own telephone. If the phone is a PSTN phone, the SIP gateway will create a peer SIP UA on behalf of the PSTN phone. After the first phone connection has been established, the call service server then send REFER to the first phone requesting it to establish the call with the second phone. Consequently, it integrates the functionality provided by the infrastructure of the PSTN and SIP phone network into a virtual collaboration environment provided by SameTime seamlessly and gives the user a "click-to-talk" on-demand voice expirence.

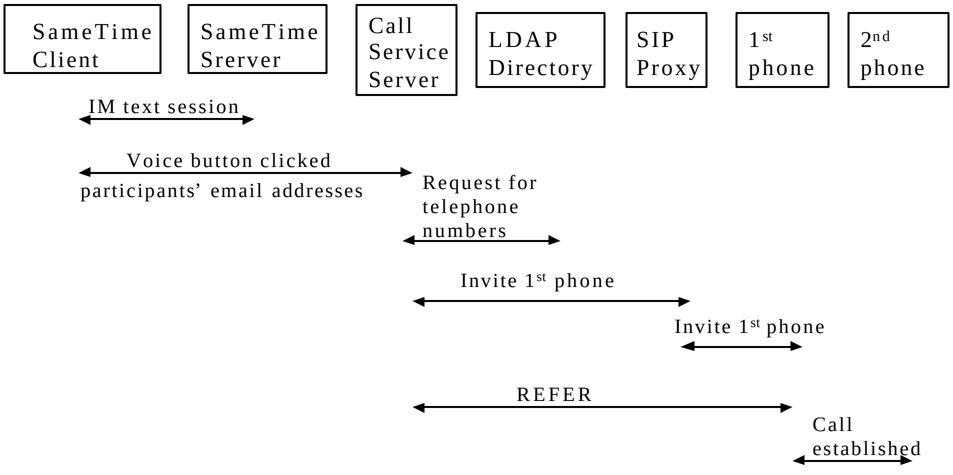
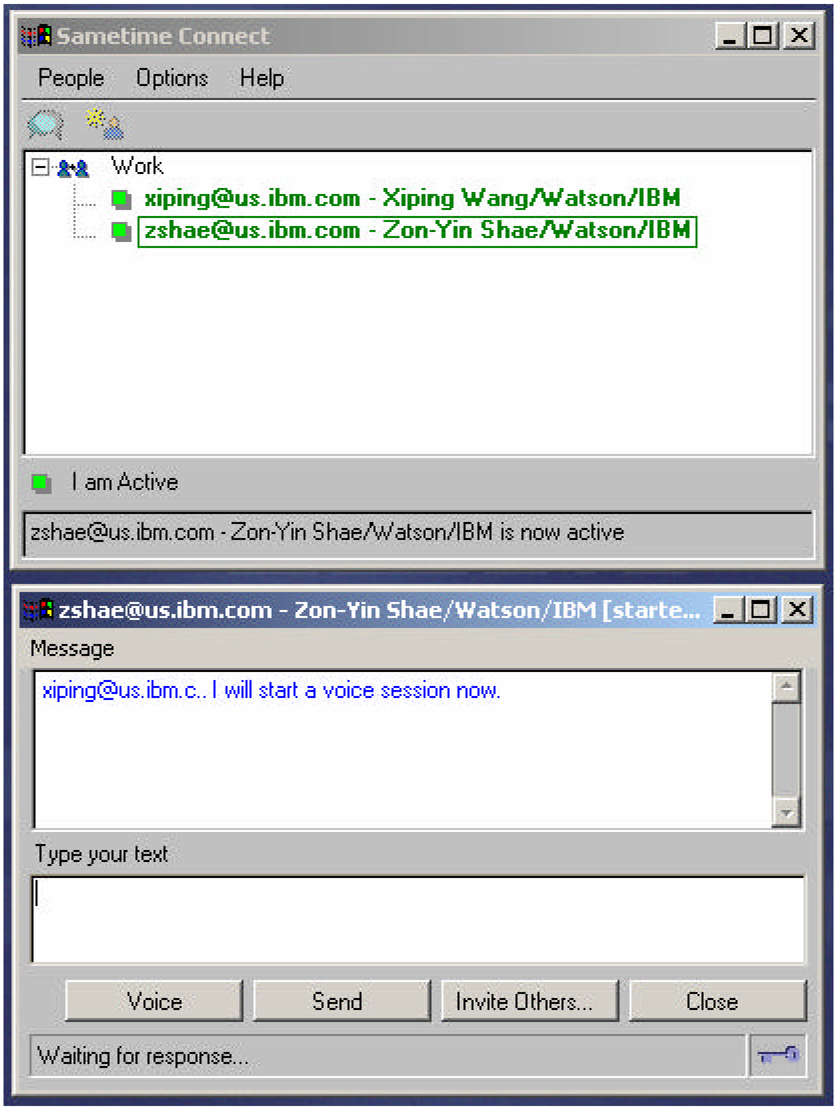


Figure 3. Hybrid Model Call Service Flow

4.2 User experience

Figure 4 shows the user interface of our voice enabled instant message experimental system. The top window in Figure 4 shows the buddy list and presence state, from where an IM session can be initiated by clicking a user name in the buddy list. An IM window (the bottom window) shows up for users to exchange information after an IM session is established. A “voice” button is added to the IM window to achieve a “click-to-talk” friendly user experience.



xiping@us.ibm.com - Xiping Wang/Watson/IBM

zshae@us.ibm.com - Zon-Yin Shae/Watson/IBM

I am Active

zshae@us.ibm.com - Zon-Yin Shae/Watson/IBM is now active

zshae@us.ibm.com - Zon-Yin Shae/Watson/IBM [starte...

xiping@us.ibm.c.. I will start a voice session now.

Waiting for response...

Figure 4. “Click-to-talk” user interface

If a voice conversation is needed, the user just needs to press the "voice" button. The voice conversation built on the proposed hybrid model will then be established between users' phones (either VoIP or PSTN or cell hard phones) by the call service server. The phone number of the callee is automatically retrieved from the corporate LDAP directory by the call setup module upon caller's click, so that the caller has no need to remember and dial the callee phone number. This, therefore, simplifies the call making process and gives the user a very friendly on-demand "click-to-talk" experience.

4.3 Participants phone number discovery

People can join the IM chat session from any place (office, home, hotel, or on the road) and any device (Laptop, PC, or PDA). This flexibility of user mobility provided by IM imposes the need for a good mechanism to locate a correct phone number for best possible voice communication and improve the phone call reach-ability of participants. There are some scenarios for finding the correct telephone number for making calls. (1) Find out the user's location (e.g., via the user's current IP), then associate the location (office, home, traveling location) with phone number. This approach requires some kinds of location service. (2) Make use of user ID (e.g., email address), looking for enterprise database for phone number. This is the mechanism that we implemented in our prototype. (3) Provide a GUI for user to input his/her phone number to use. This phone number needs to be delivered (possibly via IM with special mine type) to other participants in the chat session. Besides using the proprietary IM channel, the standard SIP protocol can also be employed to transmit user's current phone number to other participants (for privacy reason) then register the current IP address of the PC to the SIP registrar. After getting the invite call, the SameTime client will respond with 302 (redirect) code and user's phone number as contact to SIP proxy in the SIP calling process to redirect the call from PC to the user's phone.

5. Summary

The enterprise telephone system, regardless it is a regular Telco based phone system or the VoIP phone system, is currently independent of the enterprise applications. This paper has proposed a hybrid infrastructure to seamlessly bridge the gap between them. As a result, it enables the close integration of voice communication into enterprise applications. In fact, the voice control is embedded into the applications and becomes an integral part. The problems encountered with SIP control signaling traversal across NAT have been solved and simplified, also the complicated media NAT traversal problem is smoothly bypassed. Based on this hybrid infrastructure, a useful voice-enabled instant messaging application has been successfully built and demonstrated. Our experiment has shown the effectiveness of integrating voice communication capability into enterprise applications.

References

1. J. Rosenberg, H. Schulzrinne, G, Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002
2. S. Vaarala, "Mobile Ipv4 Traversal across Ipsec-based VPN Gateways", draft-ietf-mobileip-problem-solution-01, April 2003.
3. D. Senie, "Network Address Translator (NAT) – Friendly Application Design Guidelines", RFC 3235, Jan. 2002.
4. ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning."
5. J. Rosenberg, J. Weinberger, C. Huitema, R. Mahy, "STUN – Simple Traversal of User Datagram Protocol (UDP) through Network Address Translators (NATs)", IETF RFC 3489, March 2003.

6. J. Rosenberg “Tranversalcusing relay NAT (TURN), draft-rosenberg-midcom-turn-01, March 2003

7. M. Borella, D. Grabelsky, J. Lo, K. Taniguchi, “Realm Specific IP: protocol Specification”, IETF RFC 3264, Jun2 2002.

8. B. Biggs, “A SIP application level gateway for network address translation” draft-biggs-sip-nat-00.txt, March 2000.

9. P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, A. Rayhan, “Middlebox Communication Architecture and Framework”, IETF RFC 3303, August 2002.

10. J. Rosenberg, R. Mahy, S. Sen “NAT and Firewall Scenarios and Solutions for SIP”, draft-ietf-sipping-nat-scenarios-00.txt, June 2002

11. J. Rosenberg, “Interactive Connectivity Establishment (ICE): A Methodology for Network Address Translator (NAT) Traversal for the Session Initiation Protocol (SIP)”, draft-rosenberg-sipping-ice-00, Feb 2003.

12. “Sametime”, http://www.lotus.com/products/lotussametime.nsf/wdocs/homepage, IBM collaboration product.