Call Routing Management in Enterprise VoIP Networks

Kunwadee Sripanidkulchai[†], Zon-Yin Shae[‡], and Debanjan Saha[‡]
[†]NECTEC, Thailand   [‡]IBM T.J. Watson Research Center

ABSTRACT

Voice over IP (VoIP) usage is growing at an astronomical rate, resulting in large infrastructures that are becoming increasingly challenging to manage. In this paper, we focus on the management of call routing which determines how calls are routed inside a VoIP infrastructure, particularly for large multi-site infrastructures such as enterprise networks. Call routing can be complex because of several contributing factors. First, each site may independently manage and implement their own call routing policies. Second, a wide range of policies such as minimal cost or maximum performance may be in use. Existing call routing tools help with configuring or installing the policies but do not provide a means to understand or verify the policies that have been configured. In this paper, we present an analysis and visualization tool to manage call routing policies by providing a consolidated view that facilitates reasoning about the routing policies that are in use in the system.

1. INTRODUCTION

Voice over IP (VoIP) technology is gaining popularity as an alternative to traditional telephony in homes and enterprises. Deployment in homes is fueled by lower cost services provided by home broadband Internet services providers, VoIP providers [1] and peer-to-peer networks [2], collectively boasting up to 17.4 million users [3]. Similarly, enterprises are converting their private branch exchanges (PBXs) and contact centers to VoIP. The key motivation for this transition is not just the reduced cost, but the ease of integration of voice services with other network-based applications and across multiple sites, particularly in the enterprise environment [4].

Call routing determines how calls are routed inside a voice infrastructure. Compared to the plain switched telephone network (PSTN) which uses a straightforward hierarchical call routing policy, VoIP has more flexibility. For example, a large VoIP infrastructure that spans a country will often have multiple voice gateways located at different cities that act as transit points for calls between the VoIP infrastructure and the PSTN. Calls destined to the PSTN need to exit the VoIP network at one of these gateways. The choice of gateways depends on the call routing policies implemented in the system. For example, a call may exit at the nearest voice gateway

to the destination to minimize cost assuming that using the PSTN is more expensive that using the IP network, or a call may exit at the closet voice gateway to the caller to maximize call quality because the PSTN provides guaranteed performance whereas the IP network does not. In large enterprise networks with many geographically dispersed sites, these routing choices may often be complex. Further, each site may independently manage and implement their own call routing policies, requiring some coordination amongst sites to ensure global connectivity.

In this paper, we take a step towards simplifying VoIP call routing management. We focus on policy abstraction and visualization to provide a means to understand or verify the policies that are being used in the network. Note that our tool is not meant to be used in the routing design stage, but rather in the deployment and maintenance stage when policies and routes have already been deployed. Our tool applies simple rules to abstract and group call routing policies from call detail records collected from one or more sites. We then classify the resulting policies into higher-level design abstractions that operators can then use to verify and validate against expected behavior. We apply the tool to a set of call detail records collected over a three-month period from an enterprise VoIP deployment and discuss the range of policies in use.

This paper is organized as follows: we begin with an overview of a typical VoIP infrastructure in Section 2. In Section 3 we present our analysis and visualization tool and in Section 4 we discuss the resulting analysis of call routing policies in an enterprise VoIP infrastructure conducted using the tool. We then generalize our findings and discuss the call routing design space in Section 5, present related work in Section 6 and summarize in Section 7.

2. BACKGROUND

In this section, we provide a short overview of VoIP infrastructures and call routing configurations also known as dial plans.

2.1 VoIP Infrastructure

Figure 1 depicts an example of a simplified enterprise VoIP infrastructure that provides basic voice services. Consider an enterprise with two geographically remote sites, site A and B on the left-hand side of the figure. Sites A and B are connected to the enterprise-wide IP network. Site A is in the middle of a VoIP transition and has both a VoIP infrastructure and a legacy PBX system in operation concurrently whereas site B has completed the transition and has an all-IP infrastructure.

Consider the VoIP infrastructure at site A. End-devices or user-agents, such as hardware-based phones (hardphones) and software-based phones (softphones) are used to initiate and listen for incoming calls. End-devices register their location with a signaling proxy. Note that we do not distinguish between signaling protocols in this
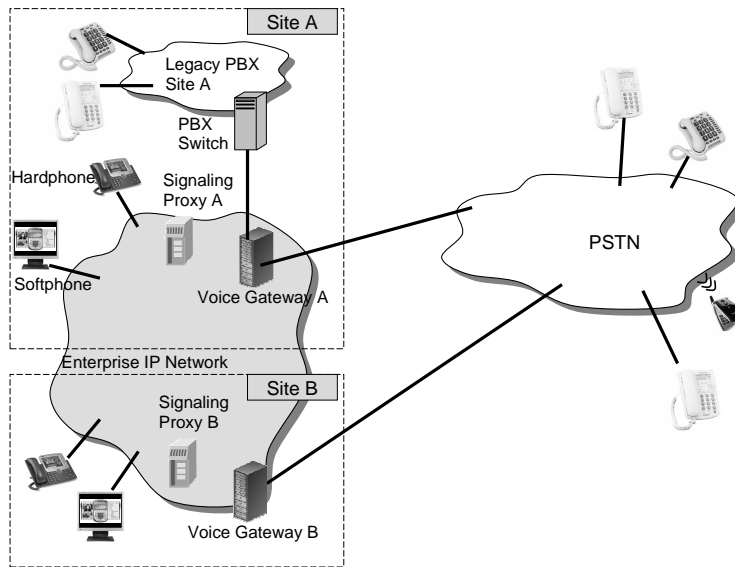
Figure 1: Simplified multi-site enterprise VoIP infrastructure.

figure. Many protocols such as SIP [5], H.323 [6], and Cisco's Skinny protocol [7] are conceptually and operationally similar. The signaling proxy participates in call signaling by relaying signaling messages such as call initiation and termination between the caller and the called party. Once signaling is established, the call's media session may be transported directly between the end-devices if both are IP-based or through the voice gateway that translates the media between the IP network and the public switched telephone network (PSTN). Site A also has a legacy PBX system that connects to the VoIP infrastructure and the PSTN through the voice gateway.

Sites A and B have their own signaling proxies and voice gateways. While it is possible to place these components at one centralized location to be shared by all enterprise sites, it is often preferable to place them at distributed locations for fault tolerance, load distribution and cost-effectiveness. For example, if we were to place all signaling and voice gateways at site A located in North America, then all calls from site B located in Asia would be forced to exit to the PSTN through voice gateways in North America. Calls that are destined to "local" phone numbers in Asia would automatically become long distance from North America making this an expensive choice. Therefore, having a distributed infrastructure is common in large enterprises.

2.2 Dial Plan

Given that a call can potentially use one of many voice gateways, it is important to design good call routing strategies. A *dial plan* controls how calls are routed to the appropriate gateway in the network. The dial plan is implemented by matching the destination telephone number to a pre-configured pattern which in turn maps to a specific media destination such as a voice gateway, hardphone or softphone. In addition to the destination phone number, information about the inbound leg of the call may also be used to help match calls. For example, an inbound call from the PBX and an inbound call from a hardphone to the same destination may be processed differently. Many VoIP vendors such as Cisco [8], Avaya [9], and open-source alternatives such as Asterisk [10] have similar dial plan treatments.

Similar to PSTN routing, the routes defined in the dial plan are semi-static in the sense that they are configured ahead of time and do not change unless a new dial plan configuration is manually

loaded. There are basic failover and load balancing options such as the ability to specify a set of media destinations instead of one for the matched phone number.

3. CALL ROUTING ANALYSIS FRAMEWORK

In this section, we present an analysis tool that operators can use to verify or validate call routing behavior. There are several challenges that we need to address in order for our tool to be useful. We need to effectively extract and reconstruct enterprise-wide call routing rules without necessarily requiring cooperation from all sites, group rules into higher-level policy abstractions, and present policies in a format that is easy to visualize and analyze.

3.1 Dial Plan vs. Call Detail Record

As discussed in the previous section, dial plan configurations directly determine routing. Dial plans are the most direct source of data for call routing analysis. However, in a multi-site setting, reconstructing a global picture would require access to dial plans from all sites in the enterprise network. This is extremely challenging when each site is independently managed by a local team of system administrators such as our deployment environment described in Section 3.2. Therefore, our first challenge is to explore whether or not we can obtain a global view of call routing by only using local information available at one site.

We found that call detail records (CDR) provide useful information that could be used for call routing analysis. When a call is served by the system, a record containing information about the call such as caller, callee, time of call and call duration is generated and logged. CDRs are used mainly for billing and record keeping purposes. However, from a slightly different angle, caller and callee information in the CDR is essentially the input and outcome of the dial plan. Using CDRs, we can reconstruct call routing rules for the local site and for all other remote sites that made calls that reached the local site. Although this is not a complete enterprise-wide view, it is a relevant view with respect to the local site conducting the analysis. The coverage provided by CDRs is wider than what we can obtain when using only the local dial plan.

Fields in the CDR that are relevant for call routing analysis are source and destination telephone numbers, signaling IP addresses, and media IP addresses. Note that CDR formats differ from ven-

dor to vendor, but the essential information in order to conduct this analysis is available from at least two major vendors: Cisco and Asterisk. Our tool currently supports the Cisco implementation because our deployment environment is Cisco-based. However, the techniques presented in this paper can be adapted to other vendors' environments.

3.2 Deployment Environment

We use CDRs collected from the signaling proxies from one site in our deployment environment for our analysis. The VoIP infrastructure is Cisco-based, using Skinny as the signaling protocol. The data was collected over several weeks from April 18, 2005 to April 24, 2005 and from May 16, 2005 to June 28, 2005 during which the VoIP infrastructure was in the pilot phase with partial deployment covering one division of the enterprise network [1]. The particular division has several geographically remote sites, four of which are included in this study. These four sites have sufficient deployment numbers ranging from a few hundred VoIP users to over a thousand users. The sites are located in North America – East coast (NA1), North America – West coast (NA2), Europe (EU), and the Middle East (ME). Our CDR data is collected from the NA1 site. The total call volume in the data set is over 180,000 calls. A portion of these calls were not complete in that the user dialed the wrong number or hung up before the call was established (no media session). We remove those calls from our analysis and are left with over 120,000 calls to process.

3.3 Reconstructing Call Routing

In the first stage, our tool reconstructs call routing information by taking the source and destination information from each call detail record and representing them as routing rules. Each rule can be visualized as a graph where nodes in the graph are source and destination phone numbers, signaling nodes and media nodes. Directional edges from source to destination are drawn between nodes where there exists an instance of that particular route in the CDR.

As a result, we have a rule for each call detail record or over 120,000 rules to analyze. By looking at only unique rules, we can reduce the number of rules down to around 33,000.

3.4 Rule Reduction

In the second stage, we further reduce the 33,000 rules by defining equivalence classes for each type of node.

We group telephone numbers based on:

- Device type: softphone, hardphone, or PSTN/PBX.

- Location: the geographic location corresponding to the area code and country code of the PSTN telephone number or the site prefix of the PBX telephone number.

IP addresses are grouped based on:

- Device type: softphone, hardphone, signaling proxy or voice gateway.

- Location: the geographic location using a corporate mapping of IP addresses/names to geographic locations.

Note that at a particular site, there may be many instances of signaling proxies or voice gateways for load balancing purposes.

Figure 2(a) depicts an example of the resulting routing rule. Using the top graph as a legend, the rule starts with the source telephone number and location, the source signaling node and media node, the destination signaling and media node, and ends with the

destination telephone number and location. In the following graph, the source phone number is a hardphone located at the North America East coast site (NA1) making a call to a PBX phone at the site in Europe (EU). The signaling and media nodes both start at the hardphone and terminate at the voice gateway (vg) at NA1.

Applying equivalence class reduction decreases the number of rules from around 33,000 down to 1,036 rules.

3.5 Rule Merging

Next, we merge similar rules together to form groups of rules representing higher-level policy abstractions. The similarity algorithm that we use to merge rules is a simple iterative different-by-1 merging. Rules that are different in just one field are merged together. Subsequent iterations take previously merged results and again perform different-by-one-field merging until all fields are visited. For example, in the first iteration we merge together all rules that have the exact same fields except for the DstPhone,Loc field. Then, in the next iteration we take these already merged rules and merge them again by looking for rules that have all of the same fields except for a second field, say the SrcSignal.Loc/SrcMedia.Loc field, and so on.

In order to determine whether or not the ordering of fields matter in the merge, we conduct an experiment where we run all combinations of ordering and look at the resulting number of merged rules after all iterations. The results indicate that we obtain the most effective merging when the ordering is as follows: (i) destination phone number, (ii) source signaling/media node, and (iii) destination signaling/media node. Using this ordering we get 99 merged rules, the fewest possible using this type of similarity algorithm. If we were to swap the ordering as (iii), (ii), (i), we get 141 merged rules instead.

We manually inspect the resulting merged rules and validate some of these rules against dial plan configurations obtained from all four sites and found them to be a good match. We plan to conduct more formal validation as part of future work.

3.6 Cost vs. Performance Policy Abstractions

Next, we take the merged rules and analyze them based on cost vs. performance abstractions. We define the routing cost and performance for each route using a general guideline that makes the cost of using the IP network much lower than using the PSTN network, but the performance of the PSTN network better than the IP network. We assign values as follows:

- IP-based: The cost of an IP-IP edge is defined to be 0 and the performance is determined by the average round trip time between the two end-points. Examples of IP-IP edges are routes between two IP-based phones or routes between an IP-based phone and a voice gateway.

- PSTN-based: The cost of PSTN-voice gateway edge is determined using the long distance calling rate offered between the two locations whereas the performance is the route-trip propagation delay between the two locations.

For example, consider the graph of the rule depicted in Figure 2(a) where a call from a hardphone in the NA1 site to a PBX phone at the EU site exits to the PSTN early at the voice gateway at NA1. Note that when conducting cost/performance analysis, we only consider the media IP address pairs because they are the end-points that transport the voice data. We ignore the signaling IP address pairs because they are not involved in the media transport. The cost/performance of the call is equal to the sum of the IP-based cost/performance between the hardphone at NA1 to the voice gateway at NA1 and the PSTN/PBX-based cost/performance between
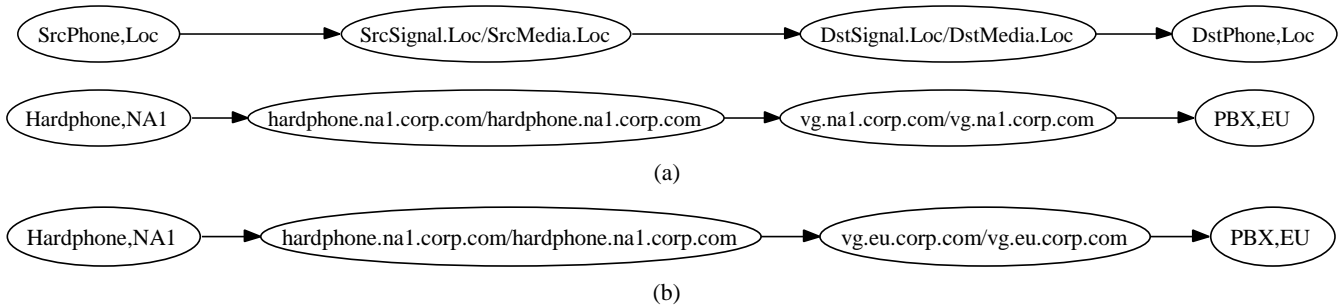

Figure 2: Example reduced routing rule.

the NA1 site and the EU site. Assuming that the cost/performance for the first leg inside the NA1 site is [0/2] and second leg outside to the PSTN/PBX is [100/50], then the resulting cost/performance is [100/52]. Contrast this with a route in Figure 2(b) where instead of exiting to the PSTN early, the call uses the IP network all the way to the EU site and exits to the PBX there. The cost/performance is now the sum of the first leg between the hardphone to the EU voice gateway [0,100] and the second leg from the EU voice gateway to the EU PBX [0,2] which amounts to [0,102]. Compared to the second rule [0,102], the first rule [100/52] has a higher cost, but faster performance.

Given this framework, we conduct the cost vs. performance analysis on the resulting routes. Figure 3 depicts the number of call routing instances and their performance in terms of delay. The x-axis is the actual delay for the route, and the y-axis is the best possible delay one could achieve for the given source and destination pair. Each circle represents a group of routes with that particular delay. The size of circle corresponds to the number of routes in the group. Most routing instances fall on the 45 degree line and are already optimal in terms of delay. Note that there are no routes above the 45 degree line because it is not possible to have a route that outperforms the minimum delay route. However, there are several routing rules that fall below the 45 degree line indicating that they are suboptimal. For example, the large cluster of routes that have actual latency of 70 ms in the current call routing configuration could have only 40 ms latency if they were configured to be performance-optimal.

4. COST VS. PEFORMANCE CASE STUDIES

We use the framework described in the previous section to analyze call routing policies across multiple sites. Of the 99 merged rules, we examine 50 rules that are cross-site in these case studies. We focus on the cost vs. performance tradeoff in configuring a VoIP network. For example, to reach a PSTN phone, a path that uses PSTN trunks as far as possible can achieve better voice quality but also has higher cost compared to one that uses IP trunks.

4.1 North America 2 Site

NA2 is located on the opposite coast from NA1 on the North American continent. While long distance costs inside the continent is relatively low, there is still a cost. NA2 optimizes for cost over performance. When an IP phone from NA2 calls a PBX phone at NA1, the call always exits to the PSTN/PBX at NA1 using the enterprise IP network as far as possible. However, when a phone at NA1 calls a PBX phone at NA2, it alternates between optimizing for cost and optimizing for performance. Half of the time it chooses cost and exits to the PSTN/PBX at NA2, using the IP network for

as long as possible. For the remaining half, it chooses performance and immediatly exits at NA1 to use the PSTN.

4.2 Europe Site

Calls made between Europe and North America are an interesting case. The long distance cost between the two continents is relatively high and the round trip time is in the mid-range at around 100 ms. EU optimizes for cost. Almost all calls destined to North America (to toll free numbers, internal PBX numbers except for those destined to NA2 and external PSTN numbers) made by IP phones at the EU site exit to the PSTN at NA1. These calls use the IP network to the extent possible and bypass high long distance costs between Europe and North America. While this is a cost-effective solution for the EU site and an effective cost-savings solution from the enterprise-wide view, the NA1 site is paying more long distance costs inside North America because it is paying on behalf of the EU site.

In the reverse direction, calls to the EU site from NA1 always exit to the PSTN at the NA1 site. The NA1 site is optimizing for performance and is avoiding using a long IP path. It is interesting to see that there exists call routing policy asymmetry.

4.3 Middle East Site

The parameters in the cost vs. performance tradeoff for calls made between the Middle East and North America are roughly similar to the above Europe site case study. Similar to the EU site, the ME site also routes all calls to North America except for calls to NA2 through the IP network all the way to the NA1 site. However, in the reverse direction, calls from NA1 to the ME site are routed using the IP network all the way to the ME site. Both directions use the same minimal cost policy.

To summarize, call routing policies differ from site to site. From NA2, EU and ME to NA1, a cost-effective policy is used. All calls destined to NA1 use the IP network all the way to NA1. However, from NA1 to NA2, a policy that alternates between optimizing for cost and performance is used. From NA1 to EU, a performance-optimal policy is used. Lastly from NA1 to ME, a cost-optimal policy is used.

5. DISCUSSION

Given the preliminary policy analysis in the previous section, we expand our discussion to the design space for call routing in VoIP networks. Our goal in this section is to raise the issues and discuss the alternatives without necessarily deciding on a particular solution.

5.1 Static vs. Dynamic Call Routing

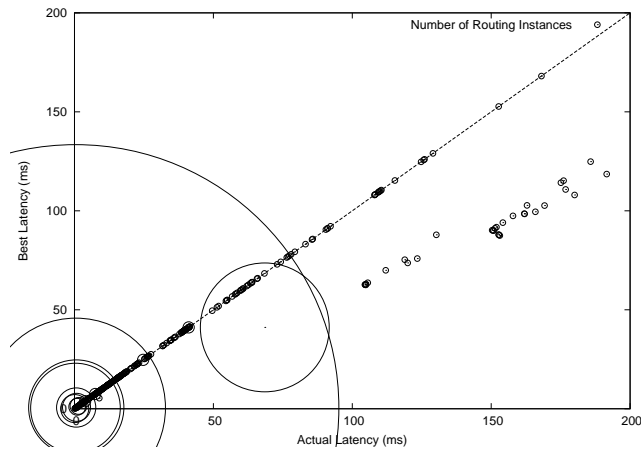The current state of the art in call routing is based on the design

Figure 3: Performance distribution of routing rules.

of its predecessor – the PSTN that uses static pre-defined routes with static failover. While routes changes in the PSTN are infrequent, is this the case for IP networks? Is such a static routing model appropriate for VoIP particularly if the IP network itself is best effort in terms of reachability and performance? VoIP routing is done at the application-level and similar to any application-level overlay should theoretically be able to select dynamic routes for robustness and performance. However, dynamic routing based on performance may come at a cost such as route flaps, routing loops, and additional protocol overhead. Any noticeable delay, reachability or performance problem may be particularly damaging for voice services because users have high expectations for voice reliability.

5.2 Centralized Call Routing

The VoIP network we study in this paper has multiple management sites each of which configures its own call routing policies. Unlike the PSTN, there is no single central entity responsible for configuring and monitoring the network as a whole. The lack of network-wide view might lead to sub-optimal routing configuration or even misconfiguration. Also, changing routing policy on one site will likely affect many other sites. The issue is further complicated by the fact that there are no protocols or feedback mechanisms that allow the distributed sites to exchange their policies and coordinate with each other. Is there a need for a call routing protocol? On the other hand if call routing were to be centralized, then who decides what the policy ought to be? Perhaps for an enterprise network, having one central authority may be a feasible solution. If so, then one needs new mechanisms to convert global enterprise policies to local dial plan configurations ready to be installed at every site. One would also need to integrate billing to the policies so that when a site uses a remote voice gateway, the site gets charged accordingly.

5.3 Mobile Users

Another important issue is supporting mobile users in a VoIP infrastructure. One key advantage of VoIP is the support for number portability where a user can be contacted using the same contact number regardless of his or her physical location. For example, an employee at site A but may be assigned to work at site B for an extended period of time. Her telephone number belongs to site A but she is physically located at site B. While at site B, she simply uses a softphone or a hardphone configured with her site A telephone number allowing her to be contacted seamlessly. Usually, this is implemented by having her softphone or hardphone register with her home site A and use its dial plan. This solution works fine from

a correctness point of view. However, from a performance point of view, it may be suboptimal. Should the user be allowed to adopt site B's dial plan while physically present at site B? If so, how would billing work in such scenarios?

5.4 Gateway Placement

The last issue which we wish to raise is the relationship between call routing policies and infrastructure design, particularly in terms of voice gateway placement. In large multi-site deployments, it is preferable to have multiple voice gateways at many physical locations so that one may maximize the use of the IP network by routing calls as far as possible on the IP network to the egress voice gateway that is closest to the destination PSTN phone number. But, voice gateways are typically the most expensive piece of equipment in a VoIP infrastructure. Clearly, there is a tradeoff between number or cost of voice gateways, the PSTN cost of calls, and infrastructure resilience. Further, the management costs of having multiple gateways should also not be overlooked. As a reference design point, consider a global infrastructure that places voice gateways at all of its sites in every country around the world against a global infrastructure that places huge clusters of gateways in a small number of locations. In the former design, gateways are close to the user reducing access time and increasing resilience. However, managing gateways at many different locations is a challenge. Consolidating servers into a smaller number of locations may help simplify management but is likely to have poorer performance, poorer resilience, and higher PSTN costs.

6. RELATED WORK

Most of the related work on VoIP call routing focuses on performance and call quality. Techniques based on multihoming and overlay networks manipulate the IP-level path that a call takes to reach its destination to maximize performance [11, 12, 13]. Note that path changes are implemented at a different level and are complementary to call routing policies discussed in this paper. First, all of the studies above consider pure-VoIP environments without the presence of voice gateways while we are interested in VoIP-PSTN environments. Second, while multihoming and overlay networks change the IP-level path to a given IP destination, call routing policies influence the choice of destination gateway for a given destination phone number.

Call capacity management using admission control can also influence the choice of voice gateways for a given call. When there

is not enough network capacity to a given voice gateway, one may change the destination gateway to support new calls [14]. Such features are supported by vendors such as Cisco as semi-static failover options. However, there is no discussion of how routes should be designed.

A design for call routing policy and gateway placement in a testbed deployment spanning several university sites presents an environment similar to our enterprise setting [15]. However, one design is in use: a voice gateway is placed at each site making it possible to have a full-meshed policy whereby any site can access any other site's gateway. There are no discussions on the actual policies that are in use.

7. SUMMARY

In this paper, we present a framework and tool to analyze call routing policies for multi-site enterprise networks. We apply the tool to a real deployment and find that call routing policies differ from site to site. While most of the policies are optimized for performance or delay, there are a few policies that have suboptimal delay and are optimized for cost. In addition, call routing is not necessarily symmetric. Routing decisions are made by local operators at each site without any global policy or formal agreement amongst sites. Lastly, we discuss the design space for call routing: should call routing be dynamic, should routing decisions be centralized, should mobile users be supported differently than other users, should gateways be placed at each enterprise site? We are exploring these issues as part of future work.

Acknowledgements

8. REFERENCES

[1] Vonage. http://www.vonage.com.

[2] Skype. http://www.skype.com.

[3] Nicole Willing. VOIP Subscriber Numbers Soar. http://www.lightreading.com, July 2005.

[4] Infonetics Research. Enterprise VoIP adoption in North America will more than double in 2010. http://www.infonetics.com/resources/purple.shtml? upna06.ipv.nr.shtml, May 2006.

[5] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261 (Proposed Standard), June 2002.

[6] International Engineering Consortium. H.323. Web proforum tutorial.

[7] Cisco Systems, Inc. Skinny Call Control Protocol (SCCP). http://www.cisco.com/en/US/tech/tk652/ tk701/tk589/tsd_technology_support_sub-protocol_home.html.

[8] Cisco Systems, Inc. http://www.cisco.com/.

[9] Avaya Inc. http://www.avaya.com/.

[10] Asterisk: An Open Source PBX and Telephony Toolkit. http://www.asterisk.org/.

[11] Shu Tao, Kuai Xu, Antonio Estepa, Teng Fei, Lixin Gao, Roch Guerin, Jim Kurose, Don Towsley, and Zhi-Li Zhang. Improving VoIP Quality through Path Switching. In *IEEE Infocom*, March 2005.

[12] R. K. Rajendran, S. Ganguly, R. Izmailov, and D. Rubenstein. Performance Optimization of VoIP using an Overlay Network. In *IEEE Infocom*, April 2006.

[13] Y. Amir, C. Danilov, S. Goose, D. Hedqvist, and A. Terzis. An Overlay Architecture for High Quality VoIP Streams. *IEEE Transactions on Multimedia*, December 2006.

[14] P. P. Mishra and H. Saran. Capacity Management and Routing Policies for Voice over IP Traffic. *IEEE Network*, Mar/Apr 2000.

[15] Falko Dressler. Advantages of VoIP in the German Research Network. In *5th IEEE International Conference on High Speed Networks and Multimedia Communications (IEEE HSNMC)*, 2002.