IP COMMUNICATIONS

# SIP Trunking

Migrating from TDM to IP for Business to Business Communication

Christina Hattingh
Darryl Sladden
ATM Zakaria Swapan

ciscopress.com

# SIP Trunking

Christina Hattingh, Darryl Sladden, and ATM Zakaria Swapan

Published by:
Cisco Press
800 East 96th Street
Indianapolis, IN 46240 USA

Printed in the United States of America

First Printing March 2010

Library of Congress Cataloging-in-Publication Data is on file.

ISBN-13: 978-1-58705-944-5

ISBN-10: 1-58705-944-4

## Warning and Disclaimer

This book is designed to provide information about Session Initiation Protocol (SIP) networking technology. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an “as is” basis. The authors, Cisco Press, and Cisco Systems, Inc. shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the authors and are not necessarily those of Cisco Systems, Inc.

## Trademark Acknowledgments

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

## Corporate and Government Sales

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:
**U.S. Corporate and Government Sales** 1-800-382-3419 corpsales@pearsontechgroup.com

For sales outside the United States please contact: **International Sales** international@pearsoned.com

## Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

**Publisher:** Paul Boger

**Associate Publisher:** Dave Dusthimer

**Executive Editor:** Brett Bartow

**Managing Editor:** Patrick Kanouse

**Project Editor:** Mandie Frank

**Editorial Assistant:** Vanessa Evans

**Cover Designer:** Louisa Adair

**Composition:** Mark Shirar

**Indexer:** Tim Wright

**Business Operation Manager, Cisco Press:** Anand Sundaram

**Manager Global Certification:** Erik Ullanderson

**Development Editor:** Deadline Driven

**Copy Editor:** Apostrophe Editing Services

**Technical Editors:** Maulik Shah and Vinay Pande

**Proofreader:** Jovana San Nicolas-Shirley

**Americas Headquarters**
Cisco Systems, Inc.
San Jose, CA

**Asia Pacific Headquarters**
Cisco Systems (USA) Pte. Ltd.
Singapore

**Europe Headquarters**
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

# Introduction

Today is an unusually interesting time in the field of unified communications as we stand at the dawn of a major industry transition in how communications flow between businesses. The network implementation to enable this communications flow is drastically changing in terms of technology, products, connectivity, service provider offerings, and cost. One of these changes is the transition to Public Switched Telephone Network (PSTN) trunking technology.

The industry is active with discussions of Session Initiation Protocol (SIP) trunking, but the appreciation of its real implications is just beginning. This book is written to help you understand and navigate these uncertain and fast-changing waters. Deciding when to change your network to take advantage of SIP trunking, how much of your network to change, and at what pace are all daunting decisions, and there might not be a single answer for all. The decisions you make today can have far-reaching effects on the productivity and cost-effectiveness of your company's communications. As an equipment vendor, we advise numerous customers on the available options and network designs and provide anecdotes as to what has worked and warnings to what has not in this epic journey to end-to-end rich media communications.

This book lays out charts and navigational assistance to help you sail these waters successfully. A variety of decision makers can benefit from this book, whether you are interested in the details of the industry transition from Time Division Multiplexing (TDM) to SIP, you are an executive who has to decide where to direct the communications investments of your company, or you are an IT staff member or consultant challenged with implementing these technologies.

The book is divided into three easy-to-use sections to address each of these interest groups:

- Part I, "From TDM Trunking to SIP Trunking," is for the beginner or executive reader and covers an overview of the industry transition, trends in the communications industry, an introduction to transitioning your network from TDM to SIP trunking, and a sample cost analysis.
- Part II, "Planning Your Network for SIP Trunking," is for the IT staff member or consultant responsible for the planning, architecture, design, and rollout logistics of the network transition. This section provides information on the network design considerations, dial plan operation, phases of rollout, service provider offering evaluation, and much more.
- Part III, "Deploying SIP Trunking," is for the network engineer who implements the changes and configures the equipment. This section provides information on network deployments and configuration examples including several case studies.

## Goals and Methods

The goal of this book is to help you understand, decide, and execute the transition from using TDM trunking to SIP trunking in your network for PSTN access. If you have already made the decision to go forward with this transition, this book will help you implement it.

The organization of the book arranges the content into three easy-to-use sections at successive levels of detail and an increasingly closer look at the actual implementation of the network. The first part provides guidance and overview material, the second part offers hands-on, in-the-trenches planning information, and the last part provides configuration examples for network implementation.

This book is a practical guide for migrating your network to incorporate SIP trunking for PSTN access. It addresses questions such as:

- What is a SIP trunk, and why should I care about it for my network?
- What are the components of a SIP trunk solution?
- How much can I expect a SIP trunk to save me on my communications bill?
- How do I determine which service provider's offering to go with?
- What does SIP trunking mean to the way calls flow in my network between my data centers and remote offices?
- What does SIP trunking mean to my dial plan?
- What security issues does a SIP trunk expose my network to?
- How do I configure the extensive features required to implement a successful SIP trunk?

## Who Should Read This Book?

This book is a roadmap to help you understand the path toward rich end-to-end communications using SIP.

Part I of this book is helpful if you want to learn about the communications industry trends and direction. It provides an overview of the TDM to SIP transition at a high, easy-to-grasp level with a handy cost-analysis right up front.

Part II should be read if you are responsible for the planning and design of a network migration for your company or for your customer's network. It is also helpful if you are doing the service provider evaluation for bringing in a SIP trunk.

Part III is helpful to the network engineers and implementers who configure the network equipment that carries calls to and from SIP trunks.

This book is not designed to be a general networking topic book; a certain familiarity and level of experience with SIP and unified communications are assumed.

## How This Book Is Organized

Although this book can be read cover-to-cover, it is designed with increasing levels of technical depth and practical network implementation as the chapters progress, geared toward readers of different interest levels. If you are an overview or trends reader, the chapters in Part I are for you:

- Chapter 1, "Overview of IP Telephony": Provides the introduction to IP Telephony and its essential components such as voice compression technologies, multiple voice signaling protocols, PBXs, Digital Signal Processing (DSPs) and TDM trunks, including an introduction to the latest development in IP Telephony and SIP Trunks.
- Chapter 2, "Trends in IP Telephony": IP telephony has traditionally been about connecting new IP endpoints, such as IP PBXs from businesses or IP phones for consumers, to the traditional TDM-based PSTN. The latest trend is for the service providers (SP) to offer IP-based PSTN connectivity services or trunks. This enables traditional IP PBX customers, in both large and small businesses, to purchase IP trunking services as opposed to traditional TDM trunking services. This chapter provides an overview of these industry migrations.
- Chapter 3, "Transitioning to SIP Trunks": Explains the stages that a company needs to plan or execute when transitioning its current systems from one that uses TDM or H.323 trunks to a system that uses SIP trunks, either for intra-application connections or for PSTN access connectivity.
- Chapter 4, "Cost Analysis": Focuses on the cost structure of deploying the SIP trunks, one of the major criteria to select the SIP trunk service provider.

If you are a network planner or designer, Part II's chapters are appropriate reading, with or without necessarily having read Part I, or perhaps reading them out of sequence:

- Chapter 5, "Components of SIP Trunks": Describes the major components used to deploy SIP trunks including call agents, session managers, session border controllers, billing servers, network infrastructure components, media gateways, and monitoring equipment. This chapter also describes the differences in components for SIP trunks into large enterprises, medium-sized enterprises, and small-medium businesses.
- Chapter 6, "SIP Trunking Models": Explains the centralized, distributed, and hybrid network topology alternatives for connecting a SIP trunk from an SP into an enterprise network for PSTN access.
- Chapter 7, "Design and Implementation Considerations": Focuses on network design and implementation considerations after a decision has been reached to connect to an SP via a SIP trunk, including geographic and regulatory considerations, IP connectivity options, dial plans and call routing, supplementary services, network demarcation and security, call traffic capacity and bandwidth control, and scalability and high availability.

- Chapter 8, “Interworking”: Explores an additional set of interworking network design considerations, including dual-tone multi-frequency (DTMF) conversion, fax and modem traffic, and connecting different H.323 and SIP variations.
- Chapter 9, “Questions to Ask of a Service Provider Offering and an SBC Vendor”: Explains the questions you should ask and appropriate answers you should seek from your SPs and SBC vendors before you deploy a SIP trunk into your network.

Part III is for the network implementer. If you are familiar with the material in Part II, you can dive straight into this, but for most readers it would make more sense to cover Part II first before starting on Part III’s chapters.

- Chapter 10, “Deployment Scenarios”: Covers configurations for the typical SIP trunk deployment scenarios, including enterprise to SP deployments, SMB to SP deployments, and configurations with SRST, transcoding, a collocated firewall, Tcl scripting, and many more.
- Chapter 11, “Deployment Steps and Best Practices”: Pulls together the detailed information from preceding chapters on network design considerations and deployment scenarios into an easy-to-use, step-by-step guide to implementing SIP trunking and a summary of best practices to follow.
- Chapter 12, “Case Studies”: Covers a few case studies of real-life SIP trunk deployments, including the challenges faced and overcome.
- Chapter 13, “Future of Unified Communications”: Shows some of the examples of how Unified Communications might develop in the next 40 years based on roadmaps of current products and technologies.

Chapter 7

# Design and Implementation Considerations

This chapter covers the following topics:

- Geographic and regulatory considerations
- IP connectivity options
- Dial plans and call routing
- Supplementary services
- Network demarcation
- Security considerations
- Session management, call traffic capacity, bandwidth control, and QoS
- Scalability and high availability
- SIP trunk monitoring

This chapter focuses on network design and implementation considerations after a decision has been reached to connect to a service provider via Session Initiation Protocol (SIP) trunking and the choice has been made regarding the appropriate network model (centralized, distributed, or hybrid) as discussed in Chapter 6, "SIP Trunking Models."

Considerations about the network design and implementation of SIP trunking include

- Geographic and regulatory considerations
- Internet Protocol (IP) connectivity options
- Dial plans and call routing
- Supplementary services
- Network demarcation
- Security considerations

- Session management, call traffic capacity, bandwidth control, and Quality of Service (QoS)
- Scalability and high availability
- SIP trunk monitoring

Another key area of consideration includes interworking and interoperability, which is discussed further in Chapter 8, “Interworking.”

Sample configurations of specific implementation examples to select service providers in the market are provided in Chapter 10, “Deployment Scenarios.”

## Geographic and Regulatory Considerations

If your network spans multiple geographic boundaries, continents, or countries, keep in mind both regulatory and distance considerations:

- **Regulatory:** Not all countries regard Voice over IP (VoIP) calls in the same way, and although virtually no country regulates what can be deployed inside an enterprise network, several countries regulate to varying degrees what calls can be handed off between an enterprise and a public (service provider) network. Ensure that you become familiar with country-specific regulations when deploying SIP trunking, especially if the endpoint (the site where the call originates) and the SIP trunk (where the call enters a public network) are in different regulatory jurisdictions.
- **Distance:** A second consideration is sheer distance, with the hairpinned media paths resulting from the centralized SIP trunk model discussed in the Chapter 6. If the remote office originating the call is in California and the central SIP trunk is in New York whereas the PSTN destination of the call is again in California, the media path for this call traverses the North American continent twice, adding latency to the end-to-end call.

  Adding latency to signaling paths is much less of a concern (it might add marginally to post-dial delay but does not impact voice quality on the active call), but latency of the media path directly affects voice quality and should be taken into consideration when designing and connecting a SIP trunk into an enterprise network.

## IP Connectivity Options

Several different types of service providers offer SIP trunks. Enterprises can find offers from service providers to transport just their data services or just their voice services, or both. When the data and voice services are delivered by different providers, each traffic type is typically delivered over a separate physical medium. Also, when the SIP trunk carries high traffic, for example 1000 sessions or more, a separate physical medium for the SIP trunk is often used.

Consider two aspects of IP connectivity when connecting to a SIP trunk:

- Physical medium of delivery
- IP addressing

## Physical Delivery and Connectivity

A dedicated physical connection for a SIP trunk is not uncommon for larger enterprises. The physical delivery for these types of terminations is often optical fiber (OC-3 or higher) and through a series of multiplexing, switching, and routing equipment eventually terminates as a gigabit Ethernet connection onto the enterprise's Session Border Controller (SBC). This connection model is shown in the left panel of Figure 7-1.

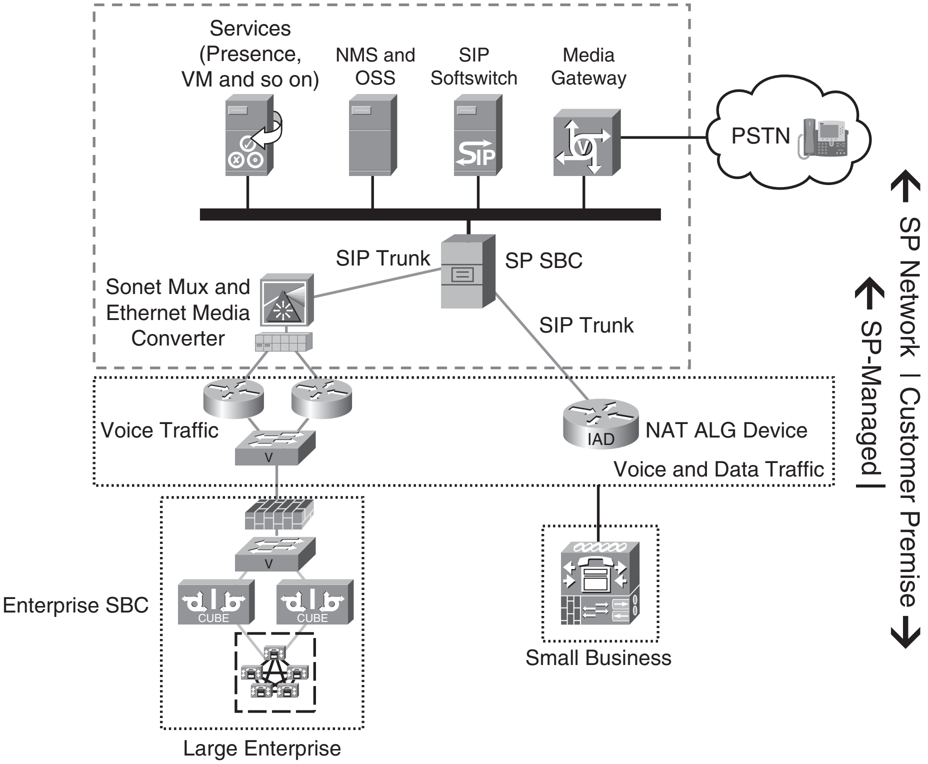


**Figure 7-1** *Physical Delivery of a SIP Trunk*

For smaller businesses this dedicated delivery model is not cost-effective and these organizations predominantly get their voice and data services from the same provider over the same physical connection. The connection in this case might be Digital Subscriber Loop (DSL), cable, Integrated Service Digital Network (ISDN), T1/E1, 3G wireless, or any other medium capable of carrying QoS-enabled IP traffic. This connection might connect directly to the customer's routing equipment, or the provider might drop off a Customer Premise Equipment (CPE) access device (for example, a DSL, cable modem, or a low-end router), which connects with Ethernet to the customer's routing or switching equipment. This connection model is shown in the "Small Business" model in Figure 7-1 using an Integrated Access Device (IAD).

It is important to note that only service providers that have complete control over the QoS of the physical connection can offer business-class voice services over a SIP trunk. VoIP services that ride on non-QoS enabled networks owned by a separate Internet provider cannot provide guarantees of quality levels because they do not control the sequencing of packets on the physical medium of delivery into your premises.

Physical connection options include Ethernet, DSL, cable, wireless (3G cellular), and traditional T1/E1.

### IP Addressing

The configuration of a SIP trunk requires coordination between the enterprise to configure its border element and the service provider's border element before starting to exchange SIP traffic. The provider allocates either explicit IP addresses or access via Domain Name System (DNS). For dedicated voice-only connections, most providers allocate two addresses per SIP trunk, whereas some offer more. If only two addresses are provided, these can often be used (per agreement with the service provider) in either a primary-secondary failover or a load-balancing algorithm.

For integrated data and SIP trunk services, there is often a single IP address. Several service providers that offer both data and voice over a single IP interface also offer Multiprotocol Label Switching (MPLS) services and require that voice be sent with an MPLS label. This setup enables the service provider to terminate voice traffic, whereas data traffic marked with a different label can be tunneled through the network.

## Dial Plans and Call Routing

Adding a SIP trunk service to your network most likely means there are service changes (accessible numbers and their associated cost), and you should optimize call routing in your network for the most cost-efficient calling patterns. This optimization, in turn, can affect current call admission control (CAC) and bandwidth-allocation policies implemented in your network.

Some specific items that might affect your dial-plan and call-routing configuration include

- If you currently have a separate dedicated Time Division Multiplexing (TDM) Public Switched Telephone Network (PSTN) voice gateway per Cisco Unified Communications Manager (CUCM) cluster or IP Private Branch Exchange (PBX), then you have a single enterprisewide SIP trunk shared between them.
- If the SIP trunk offers only long-distance (or certain types of inter-regional) calls, then your TDM PSTN gateways offers both local and long-distance calls.

- Whether the SIP trunk is going to be used by all the users in your network (all sites) or only by users colocated at the site where the SIP trunk terminates.
- Whether routing of emergency or fax, modem, Point of Service (POS), or Telecommunications Device for the Deaf (TDD) calls need to be rethought because they might not initially make use of the SIP trunk service.

Certain service providers require that a "+" be added to the front of a phone number sent on a SIP trunk. Specifically, the From field in a SIP message must be valid, as in From: +14085551212. When interconnecting through CUCM, this configuration can be accomplished by using translation rules on a Cisco Unified Border Element (CUBE) between CUCM and the SIP trunk service provider.

Certain SIP trunk providers require users to complete a registration before they can use the service. This security practice is a good one for service providers to ensure that calls originate from only well-known endpoints. CUCM does not natively support registration on SIP trunks, but this support can also be accomplished by using a CUBE. The CUBE registers to the service provider with the phone numbers of the enterprise on behalf of CUCM.

Two additional considerations regarding call routing include

- Direct Inward Dial (DID) number reachability
- Emergency call routing

## Porting Phone Numbers to SIP Trunks

When an enterprise starts using a SIP trunk for *incoming* calls, the phone number must be ported to this service. When external end users call the number, rather than ringing at the traditional TDM gateway owned by the enterprise, it rings in the service provider's core network, and the call is routed to the enterprise with the SIP trunk.

Because of the complexity of porting phone numbers, most SIP deployments find it easier to start services with *outbound* calls or with *inbound* contact center toll-free service calls (non-DID). It is important for the enterprise to understand the timelines and transition plans offered by the service provider for porting DID numbers. Enterprises' business users cannot afford to be unreachable on their primary PSTN phone numbers while this porting activity occurs.

## Emergency Calls

Emergency calling is an important consideration to account for when integrating SIP trunk access into the enterprise. Traditionally emergency calling is based on the emergency responder knowing the physical location of the TDM connection from which the call is coming. With a SIP trunk, that relationship between the physical location and the calling number no longer exists.

Options for handling emergency calling include

- Continuing to route emergency calls through your TDM PSTN gateways
- Having a small number of TDM trunks dedicated to this function at the physical location of the service provider
- Adopting a SIP-based emergency calling solution

All SIP trunk providers should provide clear explanations of their solution for providing emergency calling when an IP connection is evaluated. Some aspects to emergency calling have not been solved technologically or with the currently offered services. In the United States, the Federal Communications Commission (FCC) continues to work with the industry to define E911 operation, and a geolocation SIP header is in an Internet Engineering Task Force (IETF) draft status.

Investigate these issues in all countries and areas of the world where your network is considering a SIP trunk for PSTN access because the capabilities and regulations vary significantly.

## Supplementary Services

Cisco Unified Communications deployments offer a rich set of supplementary services. With the use of SIP trunks, how these services operate might change, and you need to evaluate how they can be maintained when a SIP trunk brings external calls into your enterprise.

Following are different areas of supplementary services to evaluate:

- Voice calls
- Voice mail
- Transcoding
- Mobility

### Voice Calls

Telephony features such as call hold/resume, call transfer, call waiting, three-way conferencing, distinctive alerting, calling line identification (CLID), calling name, and call toggle can be provided by CUCM Express and CUCM for IP phones and by a voice gateway for analog phones. IP Centrex or Class 5 type features (for example, call forwarding, call screening, call park, call return, and so on) can be provided by central SIP servers resident in the service provider's network. An analog phone in the enterprise can trigger these features with access codes (typically starting with an asterisk) provided by the service provider. Cisco Unified Voice Gateways send the access codes in the SIP INVITE message over the SIP trunk to the service provider to trigger these features.

## Voice Mail

You can provide voice mail within the enterprise network in a distributed design using Cisco Unity Express or with a centralized design using Cisco Unity or Cisco Unity Connection. You can also choose to get voice mail services from the service provider using a hosted solution (a cloud-based service).

Message Waiting Indicator (MWI) is a visual light on IP phones; it is indicated by a stutter dial tone on analog phones. MWI for enterprise-provided voice mail systems is not impacted by SIP trunking, but if a service provider hosted solution is chosen, MWI indications are provided via SIP indications (as per RFC-3842) from the service provider system and are relayed to the enterprise endpoints. Be sure to test these scenarios for SIP trunking if this is your deployment model.

Dual-tone multi-frequency (DTMF) interworking is often needed for voice mail as well. Even if all the communicating systems use SIP, there are various ways of relaying DTMF in SIP. DTMF interworking is discussed in more detail in Chapter 9, "Questions to Ask of a Service Provider Offering and an SBC Vendor."

## Transcoding

When offnet access was provided by TDM access to the PSTN, codec choices were entirely within the control of the enterprise. Codecs were configured on IP endpoints, enterprise applications, and TDM PSTN gateways. Codec choices did not affect offnet calls in any way.

With SIP trunking entering your network, this is no longer the case. Codec choices are now end-to-end on the IP segments and enterprise endpoint negotiated codecs with external endpoints and application controlled by other networks. Codecs offered by external endpoint might be against the bandwidth (call admission control) policies of your enterprise network, or your older endpoints might be incapable of supporting some of the newer codec choices, resulting in failed calls or inappropriate bandwidth use on your network.

Sometimes SIP trunking is a cost-effective choice only when G.729 is the codec chosen (for bandwidth delivery reasons), especially for high volume contact centers operations.

For all these reasons, it might be necessary to do transcoding at the border of your enterprise network to change the codec to the appropriate one before these calls enter your network. Local Digital Signal Processing (DSP) can provide transcoding for a call that uses a high-bandwidth codec such as G.711 on one side and a low-bandwidth codec such as G.729 or Internet Low-Bitrate Codec (iLBC) on the other side. Transcoding is discussed in more detail in Chapter 9.

### Mobility

Mobility of users in the enterprise is often aided by various call forwarding features. Calls between internal numbers can end up being forwarded externally due to a call forwarding mobility feature.

Another call flow to consider is an external call that arrived via the SIP trunk into your enterprise, which, in turn, is forwarded from the internal endpoint to a second external destination via the same site's SIP trunk, or if you have distributed SIP trunking deployed, perhaps a different site's SIP trunk. This call flow consumes the bandwidth equivalent to two individual calls.

Single-number reach call flows, where a single phone number is set up to ring a user's desk phone and an alternate mobile device such as a cell phone (which is often physically resident on an external network), should also be considered and tested. These call types have unique requirements for transfer, forwarding, voice mail.

## Network Demarcation

Demarcation has to do with defining and protecting the borders between networks owned or managed by different entities while maintaining interconnectivity and interoperation of traffic and features between the two networks.

TDM PSTN gateways offered an implicit enterprise network demarcation point. Until recently VoIP has been deployed only in private enterprise and small business networks for on-net calls—calls that remain within the organization's own network. Off-net calls that went from the enterprise to (or from) the PSTN were converted between IP and TDM at the PSTN interconnect point (even though many service provider backbone networks have also been VoIP for many years).

With SIP trunking the provider-to-enterprise interconnect is now also migrating to using VoIP technology. This means you no longer need TDM PSTN gateways, but it also means you lose all the demarcation features TDM gateways implicitly provided to your network. These demarcation features include

- Compliance with service provider's User-to-Network Interface (UNI)
- Codec choice
- Fault isolation
- Statistics and voice quality reporting
- Billing and call accounting
- QoS marking
- Topology hiding (security)

These demarcation features are critical to the maintenance, security, and management of your network. An SBC, such as the CUBE, can be placed at the edge of your network to terminate the SIP trunk entry point and fulfill the needed demarcation role in an all-IP network connection. For smaller businesses CUCM Express might be deployed, which includes SIP trunk capability and border element demarcation features.

All the areas of demarcation are discussed in the remainder of this section, except topology hiding, which is further discussed in the "Security Consideration" section in this chapter.

## Service Provider UNI Compliance

SIP trunk service providers offer an explicit UNI specification of what message types, formats, and fields are valid on their service offering. For the enterprise to comply with this UNI, it is often easier to place a border element at the edge of the network to *normalize* all the variants from different enterprise applications and endpoints than to try to configure each individual application or endpoint to comply with the UNI. It is especially true if the enterprise connects to multiple different SIP trunk providers, perhaps for least-cost routing or for redundancy purposes.

For the service provider, it is often easier to drop off a validated border element CPE device to ensure that the enterprise or small business network complies with its UNI—rather than certify each possible vendor and release combination of the possible applications and endpoints, the enterprise or small business might want to connect to their service.

Different deployment scenarios result from this need for network demarcation. Often enterprises want to manage their own border element so that they can control the configuration of this device and adjust it for new applications, application upgrades, or call flows. Alternatively, the service provider might provide a border element as a CPE device as part of the SIP trunk service to ensure UNI compliance regardless of the enterprise equipment. In some cases, especially for larger enterprises, both exist, and there is a pair of border elements at the enterprise edge, one side owned by the service provider (CPE), the other by the enterprise. This is separate from the SBC that always exists at the service provider edge and is a shared device among many SIP trunk customers.

The SIPConnect forum has been established by a consortium of members as an industry organization to focus on specifying and defining the SP UNI as a standard to ease some of interconnecting and interoperability issues currently still experienced.

## Codec Choice

Codec choice was briefly discussed previously in this chapter in the "Transcoding" section. Transcoding is one of the demarcation features you might want to deploy in your network to normalize codec use at the border of your network to the choices you have engineered your network for, independent of the codec choices on the SIP trunk or those chosen by the offnet destination of the call. Newer wideband codecs, such as G.722, can be used in your network but might not yet be available on SIP trunks services. The choice of SIP delayed offer or early offer also has some influence over what codecs can be chosen for any particular call.

To control the use of codecs on your network to comply either with bandwidth engineering (call admission control) or with other enterprise policies, you have the following choices:

- Allow calls with inappropriate or incompatible codecs to fail.
- Involve a transcoder to resolve calls with incompatible codecs and to change incoming codecs to those you prefer to use on your network.
- Configure features to control codec negotiation and filtering in SIP call setups as the call passes through the border element.

## Fault Isolation

In traditional TDM PSTN access, the PSTN gateway terminated the TDM connection from the provider's network and originated a VoIP connection inside your enterprise network. If voice quality or connectivity problems existed, this demarcation point was an easy place to conduct testing and isolate whether the problem existed within your enterprise network or whether it was the service provider's problem. TDM loop testing is common, enabling the service provider to test the TDM loop to the edge of your network to determine if the problem exists on that part of the connection.

Bringing a SIP trunk into the enterprise removes this demarcation point and, therefore, also the problem isolation techniques that existed for TDM interconnection. If voice-quality problems occur, it can be difficult to isolate whether they are caused by something in the service provider's network or by an element in your enterprise network.

Using a CUBE as an IP demarcation point restores this troubleshooting capability, enabling testing within the enterprise network up to the CUBE and testing from the service provider's side to the CUBE to determine where a fault might exist. The IP *loop* can be tested in the same conceptual manner (RTP loopback capability and Service Assurance Agent [SAA] responder support) as the TDM loop to allow the service provider to determine if the service is causing a problem or whether the problem exists in the enterprise.

## Statistics

Metrics, such as delay, jitter, and voice quality scoring, help enterprises and providers monitor and control the voice quality on their networks. These metrics can typically be derived only at the endpoint (DSP) of a VoIP call and not in the middle of it. (No DSP is involved in the middle of the call.) For calls on SIP trunks, it is necessary to calculate, or estimate, some of these metrics at the border element to reflect the quality of the call on the enterprise side of the network, separate from the metrics of the call leg on the service provider side of the network.

You can use different features to derive a reading of the metrics at the network border. One is to use transcoding on the CUBE, which terminates the VoIP call leg on a DSP and re-originates it on the other side—because a DSP termination is involved, actual statistics on both call legs are available from the DSP.

Another method is snooping on the Real-Time Control Protocol (RTCP) statistics as they travel through the border and to report on some of these statistics. However, many VoIP endpoints do not support RTCP.

The IP Service Level Agreement (IP SLA) Real-Time Transport Protocol (RTP)-based VoIP operation feature provides another method to provide statistics. This Cisco IOS feature uses test calls to a DSP to determine values for voice quality metrics over different network segments. The CUBE can be either the originator or the destination of the IP SLA probes to provide readings for voice quality statistics up to your network border.

There is also the Cisco IOS Voice Performance Statistics on Cisco Gateways feature (using the command **voice csr statistics**) that collects call statistics such as active calls, failed calls, packet loss, latency, and jitter.

## Billing

Typically, service providers bill without any information from the enterprise. Call detail records (CDR) from the CUBE can provide a consolidated aggregate view of calls sent and received on the SIP trunk and can be used to validate the service provider's billing.

Drawing billing records from your border element also provides a consolidated view of SIP trunk traffic use if you share a SIP trunk among multiple CUCM clusters or IP-PBXs.

Cisco IOS Software CDRs contain calling and called numbers, local and remote node names, data and time stamp, elapsed time, call failure class fields, and some vendor-specific attribute (VSA) fields. Each call through the CUBE is considered to have two call legs. Start and Stop records are generated for each call leg. These records can be sent to a RADIUS server or retrieved with Simple Network Management Protocol (SNMP) polling using the dial-control Management Information Base (MIB).

## QoS Marking

TDM voice gateways originated IP packets and, therefore, could control the QoS markings on both the signaling and media VoIP packets entering your network for calls from the PSTN. With an end-to-end VoIP call over a SIP trunk, it's quite possible that the service provider preferred QoS markings are different from the ones you prefer, and, therefore, packets have to be remarked in both directions as the packets cross the border.

The CUBE is a back-to-back user agent and, therefore, has full control over packet marking in both directions and can be set either globally or based on destination. For example, if you have two SIP trunks to different providers and their choices of marking is different from each other and from your choice in the enterprise, the border element can remark these packets on a per-flow basis.

VoIP endpoints and call agents such as CUCM and CUCMExpress also have facilities to control and mark packets. These can be used directly if the enterprise markings are the same as the SP UNI markings, and an SBC can be used if markings need to be translated between the enterprise and the SP networks.

## Security Considerations

The security concerns of TDM trunking, primarily toll fraud, exist equally on SIP trunking. In addition, SIP trunking exposes your network to IP level threats similar to data WAN or Internet access, such as denial of service (DOS).

For a hacker to gain access to your enterprise IP network via a TDM voice trunk is virtually impossible to do unless the TDM connection is specifically configured for modem dial-up access—and most voice trunks are not. Perpetrating a DOS attack on a TDM trunk is also highly unlikely as it is both expensive to do and requires large-scale autodialer equipment the average Internet hacker does not have access to. Launching these same attacks on IP addresses is significantly easier and open to a much larger pool of perpetrators because no sophisticated equipment is necessary, and the attacks can be launched for free from any Internet access connection.

When considering security on SIP trunks, you need to take into account different aspects of security. These aspects call for a series of features and capabilities to mitigate the potential threats. Security is always best deployed in a layered architecture, rather than a single box or feature that strives to protect against all possible attacks. Areas worth exploring for SIP trunk security include

- Determine the level of exposure on the SIP trunk, which depends on how it is deployed and who the provider is.
- Limit the devices that can contact your network via the SIP trunk. Mitigation capabilities include features such as access lists, hostname validation, and voice source group definitions.
- Hide your enterprise network addressing from the outside (which could be Internet-visible) and inspect the validity of traffic that enters your network. Mitigation techniques include network address translation (NAT), topology hiding, firewalls, and intrusion protection services (IPS).
- Determine protocol and session validity. Mitigation techniques include SIP port settings, SIP protocol inspection and termination, registration, and authentication methods.
- Lock down your SIP trunk against toll fraud access using the same methods you used on your TDM gateways.
- Control the privacy of sessions on the SIP trunk. Mitigation techniques involve the control of originator information available outside the enterprise network with the use of SIP privacy headers, SIP normalization, digit manipulation, and encryption

methods of the signaling and the media streams (such as Transport Layer Security [TLS], Secure RTP, and the use of IPSec tunnels or virtual private networks (VPN) on the IP connections).

## SIP Trunk Levels of Security Exposure

The level of security exposure depends on the characteristics of how the SIP trunk connects into your network and the strength of security protection your service provider offers.

Figure 7-2 illustrates four increasing levels of exposure depending on the connectivity method of your SIP trunk:

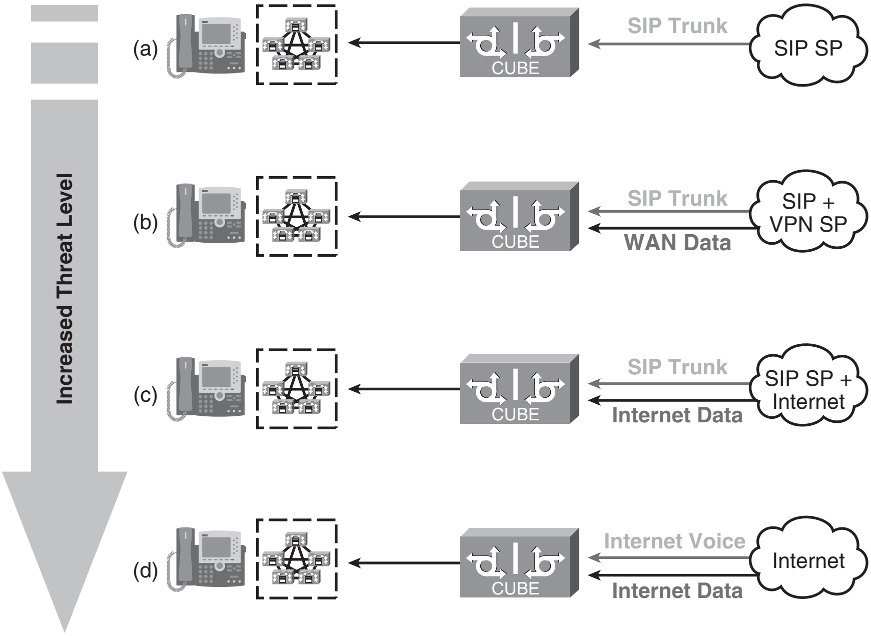


**Figure 7-2** *Increasing Levels of Security Exposure*

- In model (a) the SIP trunk connects from a Tier 1 service provider with strong security over a dedicated physical connection into your network. No data traffic traverses this connection. With this model, your security exposure is low, and you can consider not having a firewall in addition to a border element on such a connection.
- In model (b) the SIP trunk connects from a Tier 1 service provider with strong security over a physical connection that carries both your voice and your VPN WAN data connection, such as an MPLS service. No Internet data traffic traverses this connection. With this model, your security exposure is still fairly low, and you might not need a firewall in addition to a border element on such a connection.

- In model (c) the SIP trunk connects from a service provider that offers both SIP trunking and Internet access on the same physical connection. This is often a cost-effective model for smaller businesses with no WAN data service between sites or that have only a single site. Regardless of the strength of security measures in the service provider's network, you are exposed to Internet attacks on this kind of connection, and you have to firewall in addition to deploying a border element to secure this type of connection.
- In model (d) there is no SIP trunk service offering, and you use plain Internet consumer voice access and Internet data from a general Internet service provider. This model is strongly discouraged for business-class voice access because there is no quality control on such a connection, and it is extremely exposed to all kinds of voice and data Internet attacks. Firewalling and border controlling alone are still not sufficient to make this model capable of providing business-quality voice services.

Many security features on both firewalls and border elements protect against attacks on SIP trunks. The following sections discuss these techniques in more detail.

A general best practice for SIP trunk security is always to use a border element to terminate a SIP trunk coming into your network. This can be an appliance function (such as deploying a dedicated CUBE), or it can be an integrated function, such as an IAD or CUCM Express device that acts as a border element and a routing or IP-PBX device in your network.

In addition to a border element, you can choose also to deploy a firewall. Again, this might be a separate appliance, or it might be integrated into a Cisco IOS router providing multiple functions to your business. Separate, dedicated devices tend to be the norm for larger enterprise and higher volume SIP trunks, whereas integrated devices tend to be the cost-effective solution for smaller sites or small business networks.

## Access Lists (ACL)

Always strictly limit the devices that can access your SIP trunk, both from internal to your network and external to it. If you terminate your SIP trunk on a border element, you do not need all these security mitigation measures on every enterprise application, only on the border element. The border element itself should be set up to accept connections on the service provider side only from the provider's SBC, and on the enterprise side only from legitimate CUCM, IP-PBX, or other valid applications (for example, SIP proxies and meeting conference servers).

United States federal information reports that hackers are as frequently located inside your enterprise network as on the outside, and for that reason, it is imperative to lock down your border element on both sides so that rogue endpoints and applications inside your network cannot use the SIP trunk service for fraudulent calls. Similarly, rogue endpoints on the Internet should contact your SIP trunk. This configuration is illustrated in Figure 7-3.

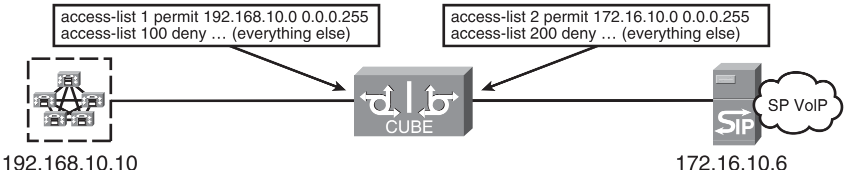

access-list 1 permit 192.168.10.0 0.0.0.255
access-list 100 deny … (everything else)
access-list 2 permit 172.16.10.0 0.0.0.255
access-list 200 deny … (everything else)
CUBE
SIP
SP VoIP
192.168.10.10
172.16.10.6

**Figure 7-3** *Locking Down a SIP Trunk with ACLs*

Additionally, voice Source IP Groups can be used with the ACLs, as shown in Figure 7-3, to provide further restrictions on the devices that might originate SIP traffic to your border element. On devices in your network that should not run SIP traffic at all, the Control Plane Policing (CoPP) feature can be used to deny all SIP traffic.

CUCM has (by default) a feature that restricts traffic on a SIP trunk to be accepted only from the IP address configured on the SIP trunk.

## Hostname Validation

You can use the hostname validation feature of the CUBE to restrict the valid hostnames that are accepted in the host portion of the SIP URI of an incoming SIP INVITE. Example 7-1 illustrates the commands used by this feature to enable calls only from the four hostnames listed.

**Example 7-1** *Hostname Validation*

```
sip-ua
  permit hostname dns:example1.sip.com
  permit hostname dns:example2.sip.com
  permit hostname dns:example3.sip.com
  permit hostname dns:example4.sip.com
```

Security features often overlap to some extent, and it is a good practice to deploy these overlapping features because they provide layered security protection. Every layer might protect you against one particular attack that might have skirted around a single layer protection to exploit a weakness in a particular appliance, device, feature operation, or configuration.

## NAT and Topology Hiding

Hiding the IP addresses of enterprise voice endpoints (such as those belonging to IP phones, call agents, and TDM voice gateways) from external view can in some cases be achieved with traditional NAT features. NAT adjusts the IP addressing of IP packet headers and some of the IP addresses appearing elsewhere in SIP packets, but generic NAT devices are Layer 3-capable only. Those that have Application Layer Gateways (ALG) have more sophisticated SIP awareness, but still, generally, might offer only suboptimal capabilities to translate deeply embedded IP addresses in SIP messaging.

It is therefore more secure to use a border element that is a full SIP back-to-back user agent (B2BUA) as the network demarcation offering 100 percent SIP packet inspection and address translation. The CUBE is a full SIP B2BUA and can therefore offer complete network address translation, usually referred to as topology hiding in this context to distinguish this function from appliance NAT devices. Both media and signaling flow through the CUBE and the service provider and off-net endpoints see only the addresses of the border element and never the addresses internal to your enterprise network.

Topology hiding is important to ensure that any attacks that might come from the service provider side can be directed only toward the border element, and the communications and call agents within your enterprise remain unaffected.

Figure 7-4 illustrates how topology hiding can be accomplished by using the CUBE.

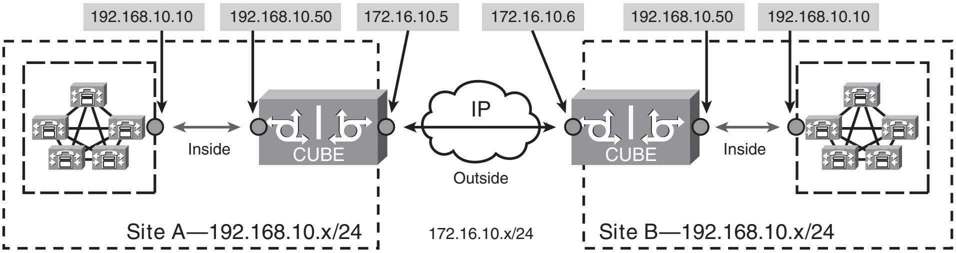


**Figure 7-4** *Topology Hiding*

## Firewalls

Many security features on both firewalls and border elements protect against attacks on SIP trunks. A certain amount of overlap occurs between the capabilities, especially true for the higher end firewalls with sophisticated SIP ALGs.

Generally you should deploy a firewall to provide generic IP protection against any kind of IP traffic, and your border element as a much more focused, voice-specific session protection function. For the least capable firewall devices, you should simply open pinholes for the traffic destined to the border element and have the border element do all the SIP inspection. For firewalls with SIP ALGs, there is some overlap in the inspection the firewall does and the inspection done by the border element. The border element always provides the most sophisticated layer of protection because it is a B2BUA whereas the firewall essentially inspects and passes through traffic but does not terminate it.

Functions that firewalls are particularly well suited to mitigate are Layers 2 and 3 inspection functions including:

- General IP DOS attacks
- Black hole routing
- TCP window control and dropping UDP packets
- Access lists, specifying what traffic is correct and allowed
- Optional SIP ALG for cursory SIP rogue and malformed packet inspection
- Optional SIP ALG protection against spikes of SIP calls (SIP-specific DOS)

More sophisticated SIP capabilities that some firewalls can have include

- Whitelist/blacklist filtering of SIP calls based on calling and called numbers
- Rate limiting of specific SIP methods to mitigate against SIP-specific DOS attacks

Firewalls are not as well suited to protecting against attacks launched from inside your network or doing session management at the level of deciding whether packets are arriving for valid sessions only, in valid sequences (or SIP dialogs), and for valid codecs or other negotiated parameters of the session. Some of the more sophisticated firewalls, such as the Cisco ASA product series or the Cisco IOS Firewall, have SIP ALGs that offer some protection services at protocol layers higher than Layer 3.

Specific functions a border element is well suited for include Layers 5 to 7 SIP inspection actions such as:

- Rejecting nonallowed calls and generating CDRs of call attempts for tracking
- Call limiting (only accept a certain number of calls)
- Codec limiting (only accept certain codecs)
- Call admission control to provide bandwidth protection
- Access lists specifying valid source and destination call agents
- Complete rogue and malformed SIP packet protection
- Digest authentication and hostname validation to ensure sessions are set up only between valid endpoints
- SIP registration to authenticate session originations
- SIP listening port configuration

Broadly, firewalls and border elements are deployed in one of two ways:

- Separate devices in series
- Integrated in a Cisco IOS device with collocated functions

Figure 7-5 provides six possible deployment models of firewalls and border elements.

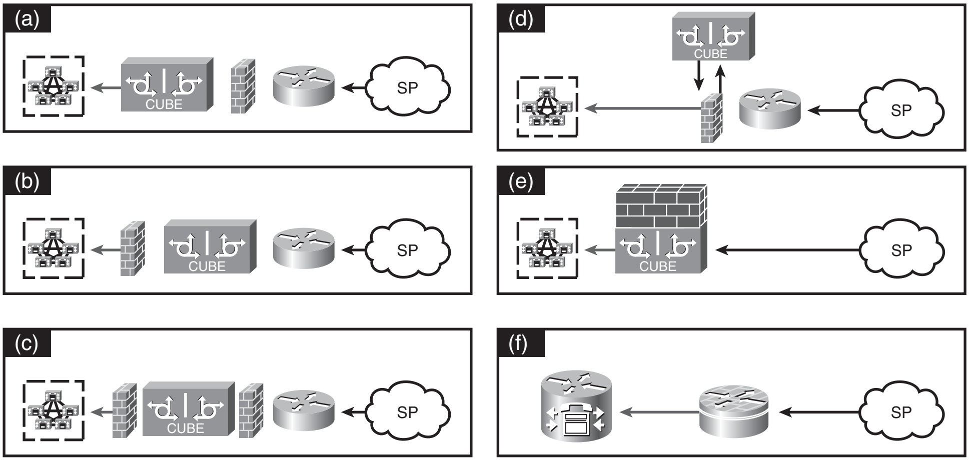


**Figure 7-5** *Possible Firewall and Border Element Designs*

Models (a), (b), and (c) shown in Figure 7-5 are better suited to medium-to-large enterprises and high volume contact centers, and models (d), (e), and (f) are better suited to smaller businesses.

- In model (a) the firewall appliance is on the outside of the border element. This is the recommended deployment model if you use separate devices for firewall services and a border element. This deployment generally makes sense for campus and data center locations where there is already a firewall present. This model also makes sense if the firewall is managed by the security team, whereas the border element is managed by the voice team. This is a mandatory model if the physical medium coming into the enterprise premises carries Internet traffic.

  In this model, the firewall provides the first line of defense on all traffic arriving from the outside, passes the voice traffic to the border element for a Layer 7 inspection on the voice traffic. If the firewall has an ALG function, there is bound to be some overlap in functionality between the firewall and the border element. It is nevertheless recommended that you turn on both to get the fullest set of inspection and protection that you can, rather than having potential security holes between the appliances.

- In model (b) the border element is on the outside of the firewall. This deployment model makes sense when the physical medium bringing the SIP trunk into your premises carries *only* SIP trunk traffic and nothing else. This means your data connections come in on a different physical path, onto different routers, and get firewalled entirely separately from the SIP trunk traffic. This model mandates that you trust your service provider's network to offer only clean SIP traffic to your enterprise.

- In model (c) two firewalls are on either side of the border element. Some refer to this model as the one for the truly paranoid, but this is the classic design of a DMZ (demilitarized zone). It is not an uncommon design, especially in large financial, educational, and government institutions, or any other business particularly attractive to hackers.
- Model (d) is a variation of model (c), where there are two virtual firewalls on either side of the border element, but one physical firewall device is used for the function, routing the unified communications (UC) traffic twice. This is a virtual DMZ design often used in video deployments where the CUBE is not only fronting a SIP trunk, but is also bringing in H.323 Internet video traffic and acting as a Cisco IOS Gatekeeper.
- Model (e) provides a more cost-effective integrated deployment model for smaller sites or businesses where a separate firewall appliance does not already exist, is not desirable, or the cost is not justified. In this model the Cisco IOS router acts as both the CUBE and the firewall. Traffic flowing through this router is inspected first by the firewall and then handed to the border element for further processing. It is therefore conceptually similar to model (a).
- Model (f) provides a lower end offering for commercial or small businesses (without IT departments) that do not want to carry the cost or the management of either a border element or a firewall. In this model, an integrated service from a service provider is purchased, and all security and demarcation issues is handled by the service provider. The service provider puts an IAD at the customer premises to connect to its IP-PBX or key system, such as CUCM Express. The IAD device will likely do NAT, perhaps basic firewalling, but essentially all the service provider's network and security are delivered as a managed service.

## Security Protection at the SIP Protocol Level

SIP is a widely used and understood protocol and simple to create because it uses straight text encoding in its messages (unlike H.323 that uses ASN.1 encoding). This makes SIP an easy target for hackers. Many of the protocol attacks can be launched against H.323 as well, but very few incidents of this were in the industry because H.323 is not as accessible as SIP.

Several ways to protect your network against a variety of SIP protocol attacks include

- Setting the SIP listening port
- Using TLS for authentication
- Using a border element B2BUA

- Using SIP normalization techniques to suppress or overwrite information in the SIP message such as the calling phone numbers, hostnames, or descriptive tags before a call enters the public network
- Using digit manipulation techniques to suppress or overwrite phone numbers before a call enters the public network
- Using SIP privacy settings to communicate the information within the SIP message that might or might not be used

Each of these areas is discussed in the following sections.

### SIP Listening Port

Every Internet hacker knows the default SIP listen ports and can sweep them from any Internet location to find an open port to launch fraudulent calls, all while your business pays for them. One way to protect against this is to change the SIP listening port to a nondefault setting. It requires the service provider to set the complementary port on the provider edge SBC. This alone can protect you against the majority of hacker attacks launched against SIP port 5060.

Example 7-2 shows the commands needed to set the SIP listening port to a nondefault setting.

**Example 7-2** *SIP Listening Port Setting*

```
voice service voip
 sip
   shutdown
voice service voip
 sip
   listen-port non-secure 2000 secure 2050
voice service voip
 sip
   no shutdown
```

### Transport Layer Security (TLS)

Another way to protect against this attack is to use TLS (specified in IETF RFC-2246). TLS uses an authentication mechanism that ensures only valid endpoints connect to your SIP trunk, and if the authentication fails, the call is refused.

Although this is a good way to mitigate fraudulent SIP calls, none of the current SIP trunk offerings in the market include TLS as an option. Hopefully this situation will change.

## Back-to-Back User Agent (B2BUA)

A B2BUA (such as the CUBE) terminates and reoriginates all calls before they enter your network. All SIP traffic passes through the SIP stack on the B2BUA twice (on ingress and egress) so that all malformed or rogue packets are dropped.

## SIP Normalization

There are certain numbers, names, or other internal information you might want to populate informative displays on the endpoints in your network. When these calls exit over the SIP trunk to external destinations, you might not want all this information to remain in the SIP messaging, especially non-DID numbers used by your organization. You can use SIP normalization features to insert, delete, or change this kind of information in the SIP messaging on your border element.

Examples 7-3, 7-4, and 7-5 show how SIP normalization can be used on the CUBE to modify the *From* header in an INVITE to a **gateway@ip-address** format and to add the **phone-context=gateway** field to the *To* header of the INVITE. Example 7-3 shows the commands needed for the configuration; Example 7-4 shows the original SIP INVITE; and Example 7-5 shows the resulting INVITE after normalization has been applied.

**Example 7-3** *SIP Normalizations Commands*

```
voice service voip
  sip
    sip-profiles 1
voice class sip-profiles 1
  request INVITE sip-header From modify "(<.*:)(.*@)" "\1gateway@"
  request INVITE sip-header To modify "<(.*)>" "<\1;phone-context=gateway>"
```

**Example 7-4** *Original SIP INVITE*

```
INVITE sip:22220000205060 SIP/2.0
Via: SIP/2.0/UDP 9.13.24.6:5060;branch=z9hG4bK1AD9E2
Remote-Party-ID: "sipp " <sip:sipp@9.13.24.6>;party=calling;screen=no;privacy=off
From: "sipp "<sip:sipp@9.13.24.6>;tag=23C3F840-99A
To: <sip:2222000020@9.13.24.7>
Date: Thu, 30 Aug 2007 07:04:36 GMT
```

**Example 7-5** *Normalized SIP INVITE*

```
INVITE sip:22220000205070 SIP/2.0
Via: SIP/2.0/UDP 9.13.24.7:5060;branch=z9hG4bK1191BFD
Remote-Party-ID: "sipp " <sip:sipp@9.13.24.7>;party=calling;screen=no;privacy=off
From: "sipp "<sip:gateway@9.13.24.7>;tag=1EDB2D94-11DD
```

*continues*

**Example 7-5** *Normalized SIP INVITE (continued)*

```
To: <sip:2222000020@9.13.32.240;phone-context=gateway>
Date: Thu, 30 Aug 2007 07:04:36 GMT
```

### Digit Manipulation

Another technique to suppress or change nonpublic numbers from exiting your network is to use digit manipulation techniques at the border of your network. For example, a non-DID number can be changed to your organization's basic public PSTN number if the call should go off-net.

### SIP Privacy Methods

Various SIP specifications control the privacy of end user information in SIP messaging such that numbers and names can travel in the messaging but still be suppressed from delivery or display to the destination endpoint. Similar methods exist in ISDN when interconnecting to the traditional PSTN.

SIP specifications (and CUBE capabilities) of interest in this area include

- The Privacy SIP header (RFC-3323) provides guidelines for withholding the identity of a person (and related personal information) from one or more parties in an exchange of SIP communications.
- The P-Asserted-Identity (PAI) and P-Preferred-Identity (PPI) (RFC-3325) headers provide extensions that enable the communication of the identity of authenticated users and the application of existing SIP privacy mechanisms to communicating these identities.

If your applications are not SIP-capable, or if they do not insert these headers, you can have your border element insert (or change) the content of these headers as a call leaves your premises over the SIP trunk. The CUBE can also convert between the widely deployed Remote-Party-ID (RPID) header to and from PAI/PPI and Privacy headers.

## Registration and Authentication

You can use SIP mechanisms to validate the originator of a SIP call and therefore provide a mechanism to reject SIP INVITEs that come from rogue endpoints. These mechanisms include

- **Registration:** Some service provider SIP trunk offerings include a registration sequence enabling the enterprise edge to register explicitly with the provider's SIP softswitch. Some SIP applications are capable of this; if not you can have your CUBE do the registration on behalf of the endpoints behind it in the enterprise network.
- **Digest Authentication (RFC-2617):** A SIP softswitch can challenge the INVITEs, and the originator must respond with credentials that are then authenticated by the SIP softswitch. Unlike a SIP registration sequence that happens once, the Digest

Authentication happens on every SIP INVITE. The CUBE can respond to Digest Authentication challenges with configured credentials.

Example 7-6 shows sample commands to configure the CUBE to do a SIP registration with credentials, and Example 7-7 shows the configuration for SIP Digest Authentication.

**Example 7-6** *SIP Registration*

```
x(config)#sip-ua
x(config-sip-ua)#credentials username 1001 password cisco realm cisco.com

sip-ua
  registrar ipv4:172.16.193.97 expires 3600
  credentials username 1001 password 0822455D0A16 realm cisco.com
```

**Example 7-7** *SIP Digest Authentication*

```
sip-ua
  authentication username xxx password yyy
```

## Toll Fraud

Toll fraud has existed for as long as telephone networks have been in operation. This constitutes making unauthorized calls that someone else pays for. The perpetrator can be inside your network (for example, an employee making personal international calls) or an external hacker using your SIP trunk to make calls that your company pays for.

Ensure that whatever measures you took to combat toll fraud in your TDM PSTN access network are also implemented on your SIP trunk PSTN access network. Some of the common CUBE tools that enable you to mitigate toll fraud attacks include

- Use ACLs to enable explicit sources of calls and deny all other traffic.
- Apply explicit incoming and outgoing dial-peers to both Border Element interfaces to control the types and parameters of calls allowed through the network border. If an incoming dial-peer is not found for a call, the system default dial-peer 0 is used enabling all calls; to avoid this, specify explicit incoming dial-peers for valid call flows and deny all other calls.
- Use explicit destination-patterns on dial-peers (try to avoid using .T if you can) to block out disallowed off-net call destinations.
- Use translation rules to ensure only valid calling/called numbers are allowed. This allows you to add access codes dialing to gain entry to certain destinations (for example, international destinations). Your employees know these access codes, but off-net hackers do not.
- Use Tool Command Language (Tcl) or Voice Extensible Markup Language (VoiceXML) scripts to do database lookups or require PINs or authorization codes

for additional validity checks to allow/deny call flows. This method protects against internal fraudulent calls.

- Change the SIP listening port to something other than the default of 5060.
- Close unused H.323 or SIP ports—if your Border Element is connected purely to a SIP trunk, there is no need for the H.323 ports to be open.
- The Class of Restriction (COR) feature restricts call attempts based on both the incoming and outgoing dial-peers matched by the call.

### Signaling and Media Encryption

Another area of security to consider is the privacy of communications, that is, how to keep hackers from recording calls or hijacking them and inserting or deleting segments. Several encryption features for voice call flows mitigate these types of attacks. Separate features for protection of the signaling traffic (TCP or UDP) and the media traffic (RTP) exist.

- Signaling encryption can be achieved by IPsec tunnels (both TCP and UDP SIP traffic) or TLS (SIP TCP). You can use TLS just for authentication or also for encryption of the signaling stream.
- You can achieve media encryption with Secure RTP (SRTP) (RFC-3711).

As the media encryption keys are exchanged in the signaling stream, there is no point in encrypting media without also encrypting the signaling. Only encrypting signaling is a valid option.

None of the current SIP trunk offerings in the market include TLS or SRTP as an option. Hopefully this situation will change. The CUBE can convert between encrypted communications (TLS/SRTP) on one side and nonencrypted (SIP/RTP) on the other side, so if your business can benefit from (or demands) encryption in the enterprise, you can still connect to a SIP trunk provider.

## Session Management, Call Traffic Capacity, Bandwidth Control, and QoS

Managing simultaneous voice call capacity and IP bandwidth use is essential for providing consistent quality in enterprise communications. Areas regarding session management and CAC to be considered in the design of your network include

- Trunk provisioning
- Bandwidth adjustments and consumption
- Call admission control

- QoS metrics, such as packet marking, delay, jitter, and echo
- Voice-quality monitoring

## Trunk Provisioning

The capacity of a SIP trunk is normally defined by the number of simultaneous calls supported and the bandwidth provided for the trunk. An enterprise uses the same Erlang calculations traditionally used in a TDM environment to determine the number of simultaneous calls required on a SIP trunk.

Generally service providers offer a tiered service based on capacity. One of the major benefits of a SIP trunk is that as an enterprise's needs expand, the number of simultaneous calls can be readily expanded without changing the physical interconnection, or even without an increase in provisioned bandwidth, provided excess bandwidth is already available.

## Bandwidth Adjustments and Consumption

Bandwidth consumption for IP call traffic inbound from the PSTN on a TDM gateway is easily predicted and controlled because the codec assignment is done by the gateway (or by the enterprise call agent such as CUCM). The use of a CUBE can ensure that this capability is maintained when an enterprise adds a SIP trunk to its communications infrastructure.

CAC policies and features are deployed in the enterprise network based on predictable patterns of codec use by calls (that is, typically G.711 for calls within a site on the LAN and G.729A for calls that traverse the WAN between sites). The bandwidth consumption of inbound SIP trunk calls is partly based on the service provider's configuration, but an enterprise can use a CUBE to influence codec selection (also called codec filtering or stripping) or to transcode streams in the codec selections the enterprise prefers to use.

## Call Admission Control (CAC)

Gateways connecting to the PSTN through a TDM interface provide an implicit form of CAC in both directions (inbound and outbound) by virtue of the limited number of channels (or timeslots) physically available on the analog, BRI, T1, or E1 interface. No more calls can simultaneously arrive from the PSTN into the enterprise than there are timeslots available on the gateway TDM trunks, providing implicit call admission control.

With a SIP trunk entering your network on a physical GE connection (possibly fiber or OC3 transport within the service provider's network before hand-off to your network), nothing physical limits the number of calls that could enter or exit your network at any one time.

Top-tier service providers exert CAC control in their networks, and how much protection this offers your enterprise network depends on who your service provider is and how well the controls are implemented. But there is virtually no physical limit, and it is strongly recommended that you protect your own network with your own CAC controls at your Border Element (especially if you are considering a SIP trunk offering without an explicit SLA). This protects against occasional unplanned bursts or surges in legitimate traffic and against potential malicious Dos attack traffic. Lack of CAC control could overrun bandwidth on your network and adversely impact network operations.

One general problem with CAC implementations is that many policies are often based on simple *call-counting* mechanisms (such as the CUCM Locations CAC feature) as opposed to bandwidth-based mechanisms (such as Resource Reservation Protocol [RSVP]). It is therefore important to control not only the number of calls arriving through the SIP trunk, but also the codec assigned to the calls.

In addition to transcoding and codec filtering, a CUBE can support the CAC policy of the enterprise in the following two ways:

- Limiting calls per dial peer (per destination)
- Limiting calls based on memory and CPU

### Limiting Calls per Dial-Peer

You can configure the **max-conn** command on both the inbound and outbound dial-peers of the CUBE to ensure that no more than the configured number of calls connects at one time. Each call, regardless of codec or the direction of the call, counts as one call.

When a call arrives at a dial-peer and the current number of calls in the connected state exceeds the configured amount, the SIP INVITE request is rejected with a 503 result code to indicate that the gateway is out of resources.

Example 7-7 shows how to configure CAC per dial-peer.

**Example 7-7** *Using Dial-Peer CAC Mechanisms*

```
dial-peer voice 1 voip
  max-conn 2
```

### Global Call Admission Control

The CUBE can also be configured to monitor calls on a global basis; that is, without regard of which dial-peer the call might be active on. This global CAC control can be done based on:

- A global system count of calls
- A CPU threshold (as a percentage)
- A memory threshold (as a percentage)
- Any combination of the preceding three metrics

The CUBE checks these configurations and metrics before it completes the processing of a SIP INVITE request. If system resources used exceed the configured amount, the CUBE returns a result code in the SIP INVITE request, indicating that the gateway is out of resources.

Example 7-8 shows how to configure global CAC.

**Example 7-8** *Using Global CAC Mechanisms*

```
call threshold global total-calls low 20 high 24
call threshold global cpu-avg low 68 high 75
call threshold global total-mem low 75 high 80
call threshold interface Ethernet 0/1 int-calls low 5 high 2500
call treatment cause-code no-resource
call treatment on
```

The **call threshold global total-calls** command controls the total number of calls to be supported on the CUBE. The command tracks the number of calls, rejecting the 25th call and not accepting calls again until the total number of calls falls below 20. The **cpu-avg** and **total-mem** options rejects the calls if the CPU or memory of the border element exceed the given thresholds regardless of the actual active call count. The **call threshold interface** command limits the number of calls over a specific IP interface.

The **call treatment cause-code no-resource** command correlates (by default) to a SIP 503 Service Unavailable message sent when calls are rejected.

## Quality of Service (QoS)

Cisco provides many methods of measuring and ensuring QoS in an enterprise IP network. You should always use these methods internally when designing a UC system, and you should also extend them to the interconnect point when using a SIP trunk to connect to a service provider. Consider several areas of QoS including:

- Traffic marking
- Delay and jitter
- Echo
- Congestion management

### Traffic Marking

QoS on IP networks depends on the QoS marking on the IP packets. As with codec settings, QoS markings on voice signaling and media IP packets on IP call traffic inbound from the PSTN on a TDM gateway is easily predicted and controlled by the configuration on your gateway. On SIP trunks, the default packet markings are whatever the service provider sets them to and this might not be in line with your enterprise policies.

The CUBE can re-mark all media and signaling packets that enter you network or exit your network to comply with the SP UNI specification. Re-marking can be done on a per-dial-peer basis (that is, per voice call destination) or per interface (either ingress or egress or both).

Example 7-9 shows how to mark packets per dial-peer.

**Example 7-9** *Marking QoS on a Dial-Peer*

```
dial-peer voice 40800011 voip
  destination-pattern 408.......
  session protocol sipv2
  session target ipv4 :10.10.1.1
  dtmf-relay rtp-nte
  ip qos dscp ef media
  ip qos dscp cs4 signaling
  no vad
```

## Delay and Jitter

The telephone industry standard specified in ITU-T G.114 recommends the maximum desired one-way delay be no more than 150 milliseconds. With a round-trip delay of 300 milliseconds or more, users might experience annoying talk-over effects.

When using SIP trunks, you should consider the IP delay of *both* the enterprise and service provider networks. In some cases, centralized SIP trunk services cannot be effectively deployed because of the resulting increase in latency. A border element device at the customer premises is required to ensure that latency in the service provider network and enterprise network can be independently measured and controlled.

## Echo

An echo is the audible leak-through of your own voice into your own receive (return) path. The source of echo might be a TDM loop in the call path or acoustic echo that applies to all-IP calls. Acoustic echo can come from improper acoustic insulation on the phone, headset, or speakerphone (all Cisco IP Phones have an acoustic echo canceller) and is common on PC-based softphones.

A border element demarcation point at a customer site can help you determine if a problem with echo is occurring at the customer premises or in the service provider's network.

## Congestion Management

When using a single connection for both voice and data, you should carefully consider congestion management (for example, queuing techniques such as Low-Latency Queuing [LLQ]) and bandwidth allocation to prevent data traffic from affecting the voice quality of SIP trunk calls.

The end-to-end voice quality experience of your SIP trunk calls depend on congestion management techniques in both your network and in the service provider's delivery network to your premises. A enterprise border element can help you determine in which network jurisdiction a problem lies.

## Voice-Quality Monitoring

To ensure business class voice quality within the enterprise network and to determine if a service provider is meeting an agreed-upon SLA, your enterprise should monitor some metrics. Each enterprise might choose to monitor different metrics, but an effective method of collecting the metrics independent from the service provider is important.

Table 7-1 describes some of the important metrics you can monitor. These metrics can be gathered by using various features previously discussed in this chapter in the "Statistics" and "Billing" sections. You can use these basic metrics from the network to calculate the more typical voice quality measurements such as Mean Opinion Score (MOS) or Perceptual Evaluation of Speech Quality (PESQ) to quantify with a single number the voice quality attained by the network.

**Table 7-1** *Voice-Quality Monitoring Attributes*

| | | | |
|---|---|---|---|
| Round-trip delay (RTD) | 100–300 ms | The RTD is the delay for a packet sent from the originating endpoint at the customer location to the terminating endpoint at the service provider and back | This metric can be monitored through the RTD metric in Cisco IOS Software; it is provided per call and is also available through IP-SLA probes. |
| Jitter | 50–100 ms | Jitter is a measurement of the change in the delay of one packet to another during a call. | Jitter is measured in the per-call statistics; the maximum jitter detected during the call is |
| Packet loss | 1 percent or lower | Packet loss is the number of packets lost during any given call, including UDP and TCP packets. | This metric can be monitored by SNMP in Cisco IOS Software; it is provided per call and can also be tracked with IP- |
| Uptime | 99.999 percent | Uptime is the percentage of time that a path is available for the customer to complete a call to the PSTN. | When uptime is measured, planned outages should be accounted for, and it should be measured as the number of unplanned minutes of outages and monitored with trouble |

*continues*

**Table 7-1** *Voice-Quality Monitoring Attributes (continued)*

| Metric | Goal | Definition | Method to Monitor |
|---|---|---|---|
| Answer seizure rate (ASR) or call success rate (CSR) | Varies | The ASR can be recorded as the number of calls made divided by the number of calls that complete a voice path. This number varies greatly because of calling numbers that are unassigned or busy. CSR is the percentage of calls successfully completed through a service provider. The CSR rate should be more than 99 percent. The ASR rate is typically approximately 60 percent. | ASR can be measured by a summary of call activity at the end of the month. The specific value of ASR is not as important as whether there are large swings in the ASR from one month to another that might indicate a problem with end-to-end network connectivity. |

## Scalability and High Availability

One of the attractive cost benefits of SIP trunking is the technical ability to centralize PSTN access for the enterprise into a single large pipe. Doing so, however, creates several design considerations, including both scalability and high availability:

- **Scalability:** Routing all calls from the entire enterprise over a single or a small number of centralized SIP trunk access points means that you are looking at a SIP trunk capacity of several hundred to several thousand connections for all enterprises except the really small ones.

  This implies border handling session capacity equipment that often far outstrips any single TDM gateway that exists in the typical enterprise. Most enterprise gateways are in the 1 to 16 T1/E1 range that equates up to between 384 to 480 sessions. Even a T3 gateway, a relative rarity in the average enterprise, presents only 672 sessions.

  Some of the redundancy schemes covered in the remainder of this section simultaneously address scalability mechanisms including higher-capacity equipment and load balancing over clusters of individual boxes.

- **High Availability:** The more sessions that are concentrated into a single physical pipe, the larger the business impact to your organization of this single point of failure. For this reason few enterprises truly deploy a single SIP trunk entry point into their networks; there are almost always multiple points.

Redundancy also becomes a much more pressing consideration because of the potentially large session capacity of SIP trunks. TDM gateway redundancy amounted to alternative routing over a different gateway when there was a failure. But when a single failure can now easily impact more than a 1000 calls, and potentially the routing of all PSTN-destined calls, the need for mitigation of such a failure escalates.

You can deploy several strategies to protect against the business impact of a SIP trunk failure:

- Local and geographical SIP trunk redundancy
- Border element redundancy
- Load balancing and clustering
- PSTN TDM gateway failover

The handling for emergency calls that you decide on (see the “Emergency Calls” section earlier in this chapter) might affect considerations for the redundancy mechanisms discussed next.

## Local and Geographical SIP Trunk Redundancy

For redundancy purposes there are almost always multiple SIP trunk entry points into an enterprise network even in a largely centralized design. This ensures that calls have alternative routing points if an equipment or building power failure occurs or a natural disaster in a particular region occurs. The only realistic alternative to multiple SIP trunk entry points is to have a single SIP trunk and maintain TDM gateway access to the PSTN for failover, a scenario discussed later in this section. For small, single-site businesses, cellular phone access might be a realistic alternative to a single SIP trunk, but this is rarely practical for a multisite enterprise of any size.

Consider three different areas of SIP trunk redundancy:

- **Local redundancy:** Most SIP trunk services offer at least two IP addresses. For local redundancy the physical medium is most likely shared and terminates into the same building on your premises. Local redundancy protects against equipment failure or power failure to a single piece of equipment. These two IP addresses should ideally terminate onto two redundant border elements. Most providers offer either a primary/secondary or a load-balancing scheme that the enterprise can choose from.
- **Geographic redundancy:** Most medium-to-large enterprises prefer to bring in the two IP addresses or perhaps two different SIP trunks (that is, four IP addresses, each SIP trunk with local redundancy) into two separate buildings, likely data centers, in two different geographies. This protects against natural disasters and buildingwide power or other outages.

- **Service provider redundancy:** Some enterprises and contact centers get SIP trunks from two different providers, both for least-cost routing opportunities and for redundancy purposes. If one provider is having problems, the other provider's facilities can carry all traffic. This scheme is easy to implement for outbound traffic but harder (due to DID mapping) for inbound traffic.

## Border Element Redundancy

SIP trunks terminate on the session border controller, or border elements, in the enterprise. These elements have to be redundant for high session capacity SIP trunks, both for scalability and high availability reasons. You can use various ways to provide redundancy for a particular border element platform (in addition to the local and geographic redundancy schemes already previously discussed):

- In-box hardware redundancy
- Box-to-box hardware redundancy
- Clustering

### In-Box Hardware Redundancy

In-box redundancy means duplicate processing components exist contained within the platform itself so that if one hardware component fails, another immediately takes over. In-box redundancy often includes components such as the CPU card, possibly the memory cards, I/O interface cards, and control and data plane forwarding engines.

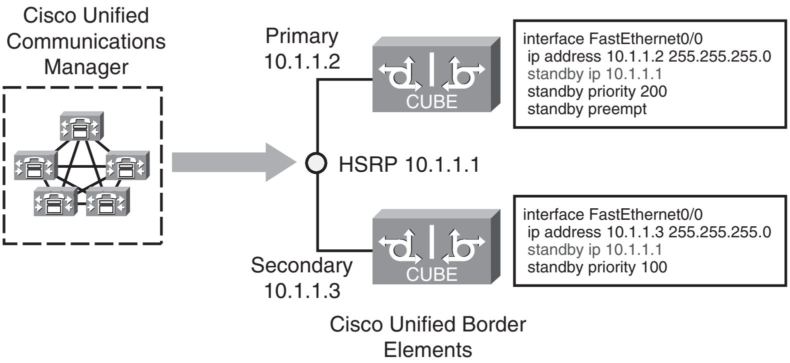
The level of hardware redundancy the CUBE provides depends on the hardware platform on which the function is installed. The higher-end platforms offer more hardware redundancy than the lower-end platforms. In-box hardware redundancy is almost invariably seamless, also called stateful failover, so sessions are not dropped and end users on active calls are generally unaware that a hardware failover has occurred.

### Box-to-Box Hardware Redundancy (1+1)

Box-to-box redundancy, or 1+1 redundancy, means there are duplicate platforms, acting and configured as a single one, in an active/standby arrangement with a keepalive mechanism between them. If the active hardware platform fails, the standby platform takes over.

One such method is the Hot Standby Router Protocol (HSRP) supported on Cisco IOS routers. With HSRP transparent hardware failover is possible while maintaining a single SIP trunk (that is, a single visible IP address) to the service provider. How well HSRP works in a particular deployment depends on the service provider IP addressing rules and the release of software deployed on the CUBE.

HSRP redundancy is not inherently stateful but can support stateful failover if the higher layers of software support application-level checkpointing and the basic router keepalive. The operation of this mechanism is shown in Figure 7-6.


Cisco Unified Communications Manager
Primary 10.1.1.2
CUBE
HSRP 10.1.1.1
CUBE
Secondary 10.1.1.3
Cisco Unified Border Elements

```
interface FastEthernet0/0
 ip address 10.1.1.2 255.255.255.0
 standby ip 10.1.1.1
 standby priority 200
 standby preempt
```

```
interface FastEthernet0/0
 ip address 10.1.1.3 255.255.255.0
 standby ip 10.1.1.1
 standby priority 100
```

**Figure 7-6** *Using HRSP for Redundancy*

Enterprise TDM gateways do not offer stateful failover redundancy because the session capacity per gateway is limited; therefore, the impact of a failure is limited. If an individual CUBE carries no more sessions than the average enterprise TDM gateway, there might not be a reason to expend the cost on deploying high-end hardware with stateful failover capability on the border element either. Instead, border element clustering can provide effective redundancy, as it does for TDM gateways.

### Clustering (N+1)

Redundancy via clustering, or N+1 redundancy, means there are duplicate platforms independent of each other and each carries a fraction of the traffic, together providing a high session count SIP trunk. There is no state sharing or keepalives between the components, and if a single element is lost, some calls drop, but it is not the entire SIP trunk that goes down.

The CUBE can be deployed in a clustering architecture with load balancing over the individual components managed by the attached devices or by a SIP proxy element. (Load balancing methods are explored further in the next section.) A clustering architecture has the advantage of a pool of smaller elements, each of which can be taken out of service and upgraded without affecting the entire SIP trunk. The cluster can also be spread out over several buildings or geographic locations to enhance redundancy concerns about the impact of a power loss or a natural disaster on a building or data center.

## Load Balancing

SIP trunks from providers usually come with two (sometimes more) IP addresses. As previously discussed, you might want to have multiple border elements fronting this SIP trunk for both redundancy and scalability benefits. If you choose a load-balancing algorithm (as opposed to a primary/secondary active/standby arrangement) for the multiple platforms forming the network border, some network entity is required to do load balancing across the possible destinations.

You can use multiple ways to implement SIP trunk load balancing:

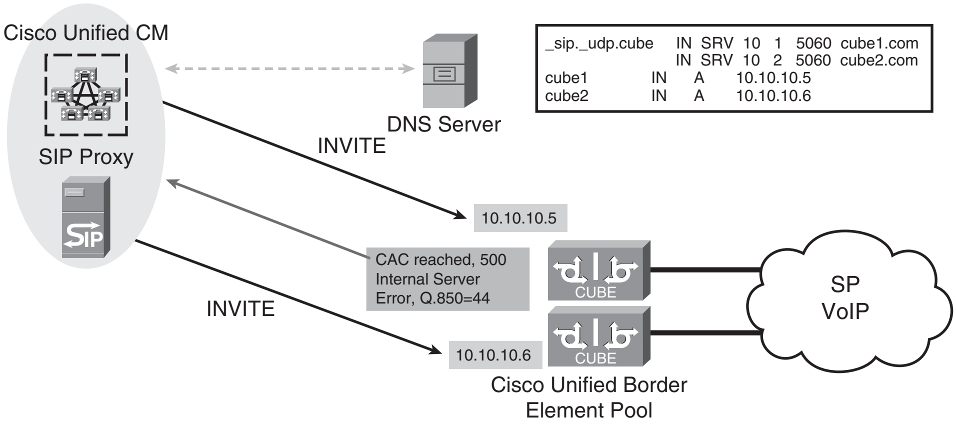
- Service provider load balancing
- DNS
- CUCM route groups and route lists
- Cisco Unified SIP proxy

## Service Provider Load Balancing

Many SIP trunk providers offer a choice of primary/secondary or load-balancing algorithm to the enterprise customer. If load balancing is chosen, this is implemented either on their SIP softswitch or their provider edge SBC.

## Domain Name System (DNS)

You can use DNS SRV records (RFC-2782) to provide multiple IP address resolutions for the same hostname. In this way, the individual platforms in the border element cluster can be addressed dynamically using the information returned by DNS. The operation of this mechanism is shown in Figure 7-7.



```
_sip._udp.cube    IN  SRV  10   1   5060  cube1.com
                  IN  SRV  10   2   5060  cube2.com
cube1           IN     A      10.10.10.5
cube2           IN     A      10.10.10.6
```

**Figure 7-7** *Using DNS SRV for Load Balancing*

The attached SIP softswitch (this can be used either on the service provider side or on the enterprise side) queries DNS for the IP addresses of the border element. The originating softswitch uses these addresses to load balance traffic. If a call is presented to a CUBE that is overloaded (its configured CAC threshold has been reached), it returns a SIP 503 Internal Server Error, and the softswitch can use the next available address in the DNS SRV record.

DNS is not offered by all service provider SIP trunk offerings, but when it is, this is generally a good method of load balancing. Even when it is not offered, this mechanism can still be used to good effect on the enterprise side of the network border. This method is dependent on a predictable design of DNS server response time to ensure that post-dial delay (PDD) is minimal.

The DNS SRV mechanism can also be used for load-balancing calls outbound from the CUBE to an attached softswitch. If DNS is used for this call path, the SIP INVITE retry timer might need to be tuned to constrain PDD for outbound calls, as shown in Example 7-10.

**Example 7-10** *SIP Retry Timers*

```
sip-ua
  retry-invite 2
```

## CUCM Route Groups and Route Lists

When connecting a CUCM to a cluster of border elements for PSTN SIP trunk access, its Route Group and Route Lists constructs can be used to implement a load balancing algorithm for presenting calls outbound from the enterprise to the PSTN. Other SIP softswitches and IP-PBXs most likely have similar alternative routing capabilities that can be used in a similar manner. The operation of this mechanism is shown in Figure 7-8.

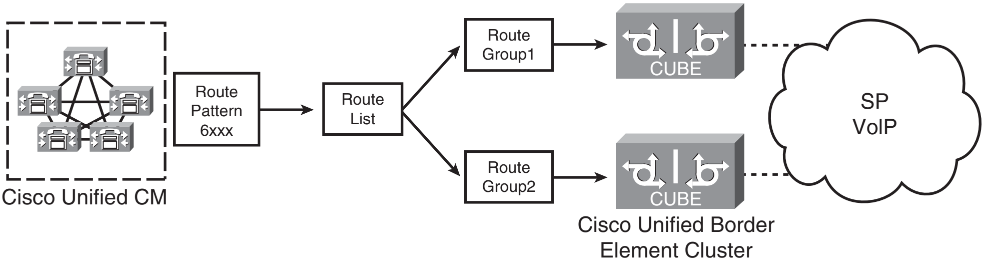


**Figure 7-8** *CUCM Route Groups and Route Lists*

Configure a Route Group on CUCM pointing to each individual border element. Aggregate these Route Groups into a Route List that points to the SIP trunk. Configure a Route Pattern in the CUCM dial-plan to route calls of the appropriate dialed number patterns to this Route List. Configure CAC on the individual CUBEs to refuse calls under overload conditions, forcing CUCM to reroute to the next Route Group in the Route List.

## Cisco Unified SIP Proxy

The Cisco Unified SIP Proxy can be used with a cluster of border elements as a logical large-scale SIP trunk network border interface to the attached softswitches. That is, the attached softswitches on both the service provider and enterprise sides are unaware of the individual elements, or the number of them, in the CUBE cluster. This is a handy mechanism when:

- You build large-scale SIP trunks where the number of border elements exceed the two IP addresses given by your provider.
- You want to grow the SIP trunk capacity over time without affecting the configurations of the attached softswitches on either side of the border.

The Cisco Unified SIP Proxy is responsible for the load balancing over the individual border elements, keeps track of their loads, and reroutes traffic when a particular element is overloaded or unavailable. The operation of this mechanism is shown in Figure 7-9.

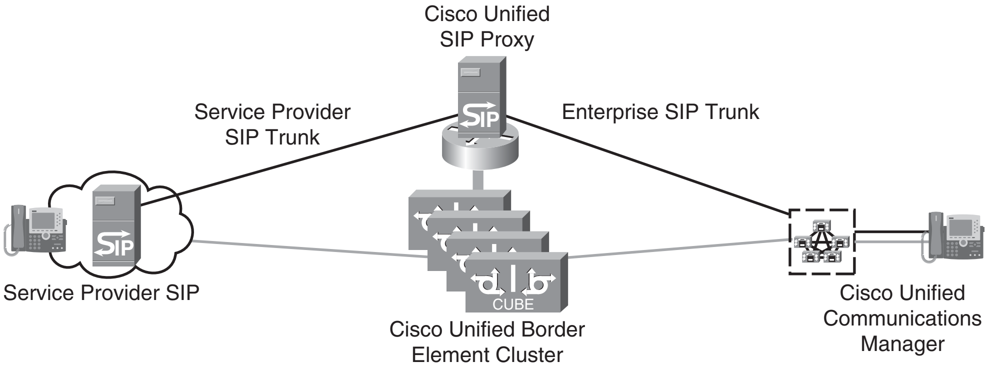


**Figure 7-9** *Cisco Unified SIP Proxy and Border Element Cluster*

In addition to load balancing, the Cisco Unified SIP Proxy offers many benefits to the SIP trunk interconnect:

- Hides the size of the border element pool from the attached softswitch configurations.
- Offers policy-based SIP trunk call routing such as time-of-day and least-cost routing.
- Offers powerful SIP Normalization capabilities.
- Offers graceful service degradation for upgrades or maintenance of the border elements.
- Offers an easy way to expand the capacity of your SIP trunk when your needs grow.
- Offers intrinsic redundancy because there isn't a single border element but a cluster of them. (The SIP proxy itself must, of course, be deployed in a redundant configuration; otherwise, it becomes a single point of failure.)

## PSTN TDM Gateway Failover

An easy and cost-effective way to provide redundancy and failover for a SIP trunk is simply to reroute calls to your already existing TDM gateways when the SIP trunk is not available or overloaded. This method provides a ready migration path while you ramp up SIP trunk traffic to full production and enables you more time to design and implement some of the other SIP trunk redundancy mechanisms in preparation for a future state where your network might no longer have TDM connectivity. The operation of this mechanism is shown in Figure 7-10.

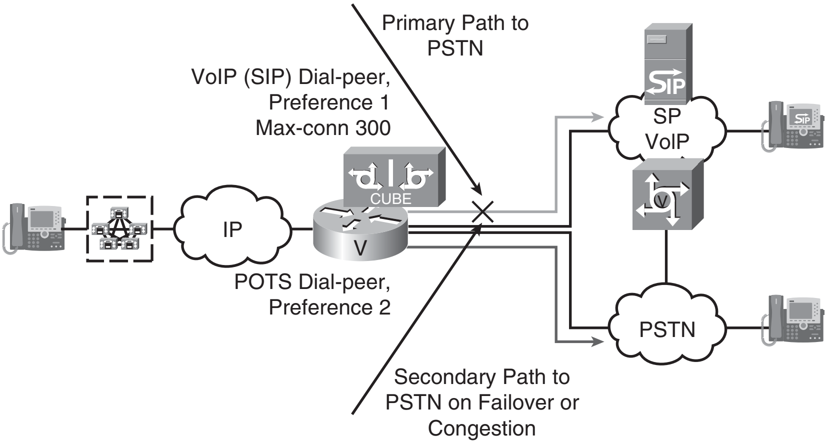

Primary Path to PSTN
VoIP (SIP) Dial-peer, Preference 1 Max-conn 300
SIP
SP VoIP
CUBE
IP
V
V
POTS Dial-peer, Preference 2
PSTN
Secondary Path to PSTN on Failover or Congestion

**Figure 7-10** *SIP Trunk to PSTN Failover*

Configure call routing to use the SIP trunk as the primary method of access (using a higher preference dial-peer) and the TDM gateway as the secondary path (using a lower-preference dial-peer). You can use the same physical Cisco platform for both functions so that adding a SIP trunk to your PSTN gateway does not mean adding equipment to the network.

## SIP Trunk Capacity Engineering

Part of the scalability assessment for your network is to determine how many concurrent sessions should be supported on the SIP trunk service offering that you get from a service provider. If you have current PSTN traffic statistics on your TDM gateways, this assessment is somewhat easier as the ratios of phones to trunks do not change with SIP trunking. But many enterprise networks do not have detailed current statistics of these call patterns.

SIP trunk session sizing is also affected if you choose a centralized model, as opposed to the distributed model of traditional TDM trunking where there is often oversubscription at each site. This oversubscription can be consolidated with a centralized SIP trunk facility, but you still have to engineer with some level of bursting of call traffic for unusual situations.

As a ballpark assessment, you can use the same method of estimating trunk (which is equivalent to a SIP trunk session) capacity as you used in the traditional voice traffic engineering exercises. An average enterprise business can use a 5:1 trunking ratio, meaning for every five phones, provision one trunk (SIP session). Enterprises that are primarily internally focused (for example research facilities or engineering departments) can use a 10:1 ratio. Contact center deployments should use a 1:1 ratio, and *phones* in this context include both live agents and automated ports serving Interactive Voice Response (IVR) front-end applications.

## SIP Trunk Monitoring

Several generic IP mechanisms can monitor the health of a network element, such as an Internet Control Message Protocol (ICMP) Ping. Although these are useful, they provide only Layer 3 health. The SIP protocol specifies an Out-of-Dialog (OOD) Options Ping method in RFC-3261 that provides a Layer 7 health indication of a SIP endpoint.

The OOD Options Ping method can provide a health check for a SIP trunk and enables attached devices to reroute traffic upon a failure of any one element in the path. Note that it is a per-hop method and that several Pings might need to be configured to provide end-to-end failure detection on a SIP trunk. This method is illustrated in Figure 7-11.

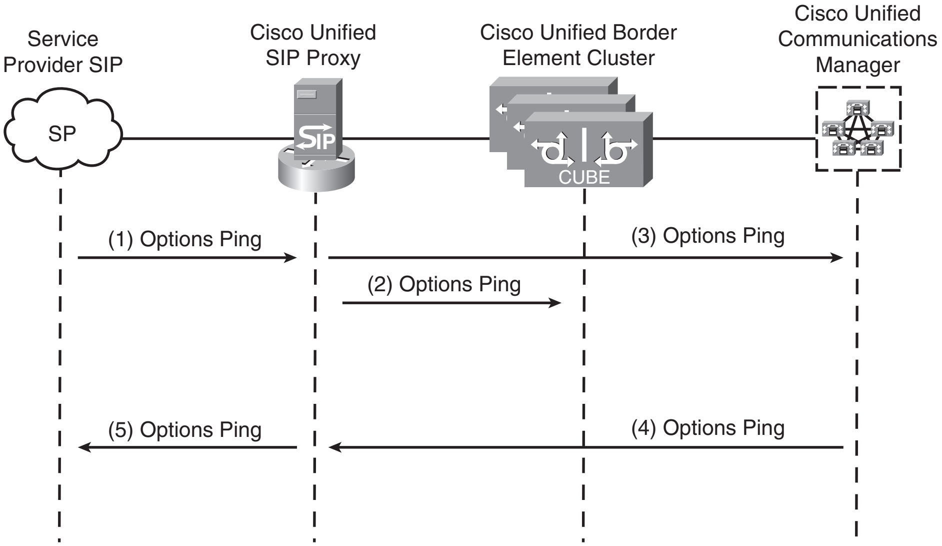


**Figure 7-11** *SIP Trunk Monitoring Using Options Ping*

If the Options Ping between the elements fails (in the direction indicated in Figure 7-11), the following actions are taken:

**Step 1.** The service provider fails over to the secondary IP address for the SIP trunk, if available, or reroutes calls destined to the enterprise.

**Step 2.** The Cisco Unified SIP Proxy marks a border element as down and reroutes calls to alternative border elements in the cluster until it comes back up.

**Step 3.** The Cisco Unified SIP Proxy marks CUCM as down and rejects incoming calls from the service provider.

**Step 4.** When supported (a future capability), this path allows a CUCM to mark the SIP trunk as down and use its alternative routing logic to place outgoing calls.

**Step 5.** The Cisco Unified SIP marks the SIP trunk to the service provider as down and rejects incoming calls from CUCM, enabling it to use its alternative routing logic to place outgoing calls. In the absence of (4), this is the method that indicates to the CUCM that the service provider SIP trunk is down.

## Summary

SIP trunks are becoming an increasingly viable option for enterprises wanting to deploy IP-based PSTN access. This chapter highlighted many of the network design and implementation considerations you should work through while planning or installing a SIP trunk for production purposes in your network. Migrating to SIP trunking is a fundamental network change that should be accompanied by the appropriate level of planning and configuration and can require several phases of deployment.

In Chapter 8, another key area of network consideration—interworking and interoperability—is explored in further detail to round out the discussion of network design considerations regarding SIP trunking.

## Further Reading

The following documents and references provide additional information on the topics covered in this chapter.

### General

SIPConnect Forum: Focused on defining SP UNI compliance as a standard to ease interop requirements. http://www.sipforum.org/sipconnect.

### Cisco IOS and Unified Border Element Documents

More information on TLS configuration for the CUBE can be found on Cisco.com. www.cisco.com/go/cube > Configure > Configuration Examples and TechNotes > Unified Border Element SIP TLS Configuration Example.

More SIP Normalization examples for the CUBE can be found on Cisco.com. www.cisco.com/go/cube > Configure > Configuration Examples and TechNotes > Unified Border Element (CUBE) Session Initiation Protocol (SIP) Normalization with SIP Profiles Configuration Example.

Voice Performance Statistics on Cisco Gateways.
www.cisco.com/en/US/docs/ios/12_3t/12_3t4/feature/guide/gt_th.html.

## IETF RFCs

Transport Layer Security (TLS) RFC-2246.
http://www.ietf.org/rfc/rfc2246.txt?number=2246.

A Privacy Mechanism for the Session Initiation Protocol (SIP) (RFC-3323).
http://www.ietf.org/rfc/rfc3323.txt?number=3323.

Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks (RFC-3325). http://www.ietf.org/rfc/rfc3325.txt?number=3325.

HTTP Authentication: Basic and Digest Access Authentication (RFC-2617).
http://www.ietf.org/rfc/rfc2617.txt?number=2617.

The Secure Real-Time Transport Protocol (SRTP) (RFC-3711).
http://www.ietf.org/rfc/rfc3711.txt?number=3711.

A DNS RR for specifying the location of services (DNS SRV) (RFC-2782).
http://www.ietf.org/rfc/rfc2782.txt?number=2782.

SIP: Session Initiation Protocol (RFC-3261).
http://www.ietf.org/rfc/rfc3261.txt?number=3261.

A Message Summary and Message Waiting Indication Event Package for the Session Initiation Protocol (SIP) (RFC-3842). http://www.ietf.org/rfc/rfc3842.txt?number=3842.

# Index