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.

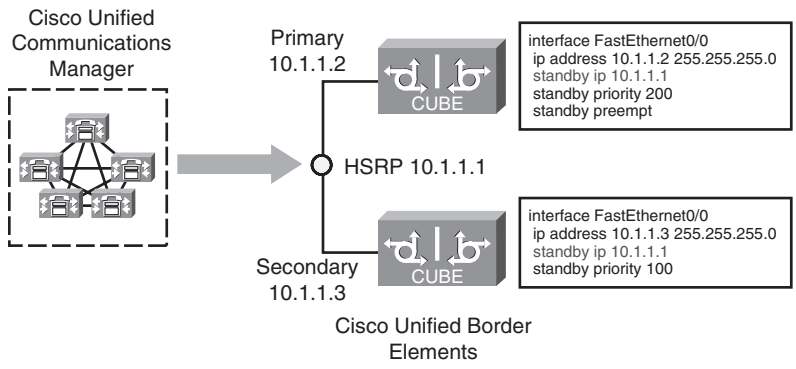


**Figure 7-6** *Using HRSP for Redundancy*

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.

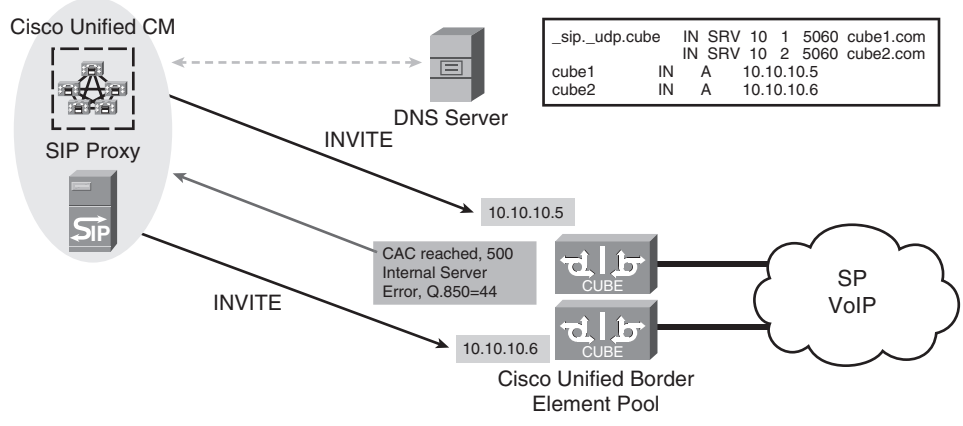


**Figure 7-7** *Using DNS SRV for Load Balancing*

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.

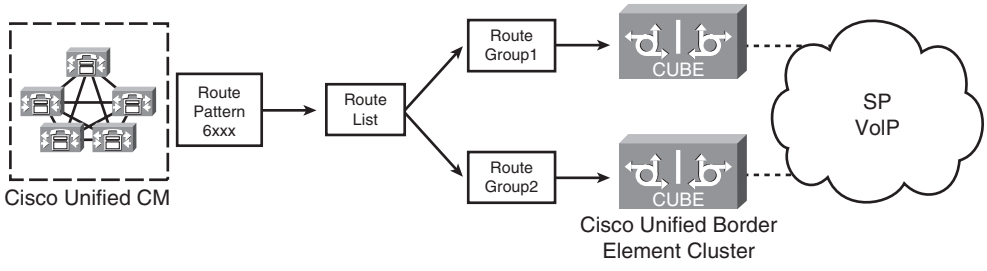


**Figure 7-8** *CUCM Route Groups and Route Lists*

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.

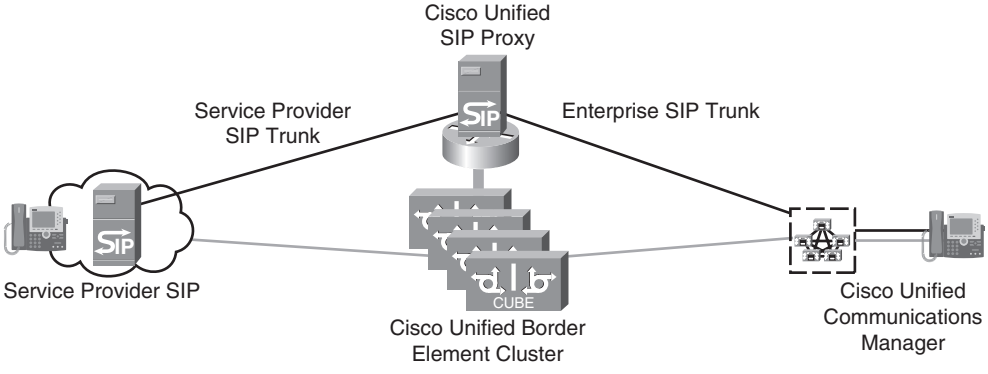


**Figure 7-9** *Cisco Unified SIP Proxy and Border Element Cluster*

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.

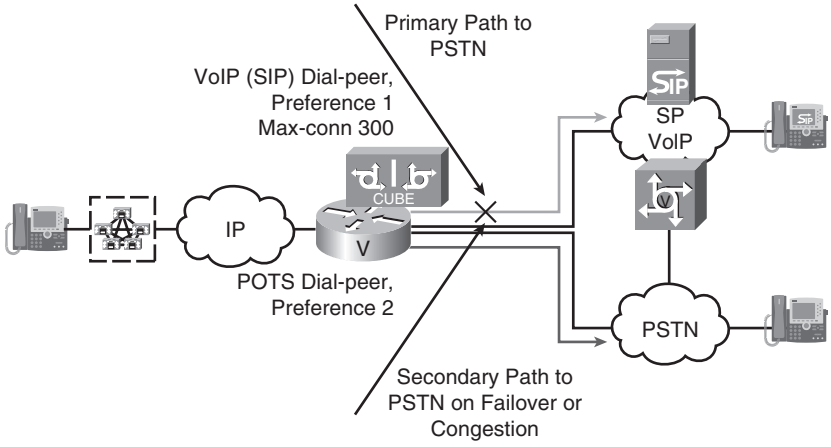


**Figure 7-10** *SIP Trunk to PSTN Failover*

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.

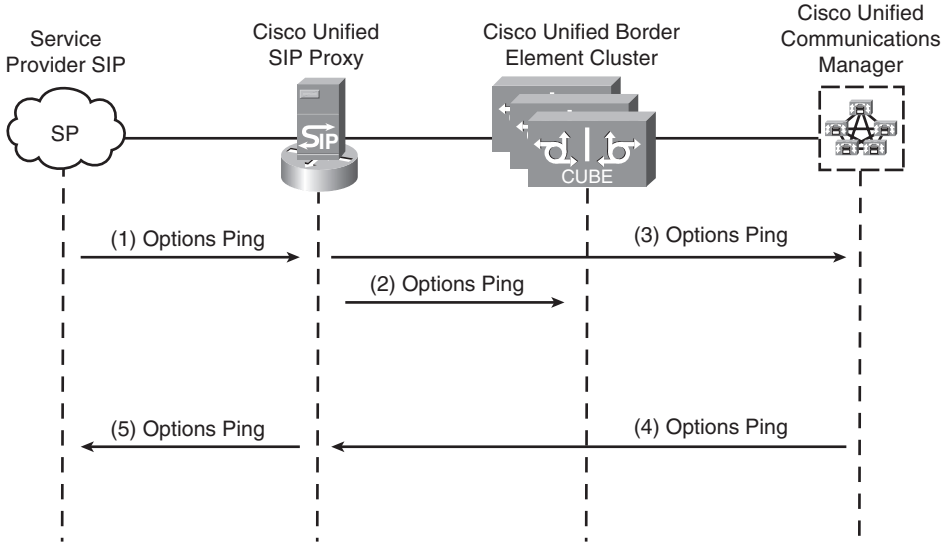


**Figure 7-11** *SIP Trunk Monitoring Using Options Ping*

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.