Scaling SIP

This paper addresses the challenges and solutions when providing high-availability and scalability for SIP services.

White Paper
by Lori MacVittie

Introduction

Since Session Initiation Protocol (SIP) was originally defined as a signaling protocol for multimedia sessions, SIP has grown beyond its targeted use as an IP-based telephony messaging medium to conferencing and instant messaging.

The extensibility and flexibility of SIP, combined with its use of standards-based, simple transport protocols, brings value to IP-based telephony service providers. That extensibility, as is often the case, can also cause of challenges for providers. This extensibility gives providers the ability to offer additional services to customers, but can also cause problems if they need to support a wide variety of user and premise-based devices that might not recognize extensions to the protocol.

Additionally, the explosive growth of IP-based telephony services accessible from a wide variety of devices has created a very real need for reliable scalability. Like PSTN service before it, IP-telephony requires two channels: one for signaling and control and one for the exchange of actual data. SIP is the signaling and control plane for voice-based communications, and it must be able to affect a second, completely different channel that uses a different set of transport and application layer protocols. Thus, a service provider infrastructure focused on supporting IP-based telephony services is necessarily more complex than simple, web-based services. This makes ensuring reliable, scalable services more difficult.

The Session Border Controller

While SIP and HTTP share many characteristics?text-based, easily readable by humans, connection-oriented?there are unique challenges associated with handling of SIP. SIP-based communications combine TCP- and UDP-transported data for signal control and data exchange respectively, and do so as separate streams. Yet those separate streams must be correlated and used together to provide a high quality of communication for those participating in SIP-based dialogs.

The architecture necessary to support SIP-based communication and the need for high-availability further complicates SIP communications. The use of Session Border Controllers (SBCs) at the edge of such deployments provides a number of functions related to SIP and its flexible nature. SBCs occupy a unique place in a service provider?s architecture, acting as the gateway between the access and core networks. This makes SBCs an obvious choice for providing many functions associated with SIP and network layer protocols.

SBCs are typically deployed for the purposes of:

- Providing interoperability between protocols

- Overcoming challenges associated with NAT (network address translation)
- Enforcing quality of service (QoS) policies
- Acting as a point of regulatory compliance
- Offering core network security

Successful deployments of SIP rely heavily on the reliability and scalability of the SBC. This is because the SBC provides much of the functionality that makes IP-based telephony services work.

Scaling SBCs and SIP services in general, however, is not a trivial task. As noted earlier, SIP and HTTP share many characteristics. While a simple, layer 4 load balancer might more than adequately provide for the availability needs of HTTP-based applications, it will not provide high-availability services for SIP.

Scaling SIP

In order to provide high-availability services and scale SBCs, a solution must first be SIP aware. This means that any high-availability solution has to be capable of inspecting and acting upon layer 7 (application) data. This is because SIP carries much of the relevant information regarding client, server, capabilities, and its connection in its headers and payload. Solutions incapable of inspecting the SIP payload will not be able to extract the information necessary to maintain SIP sessions.

Special Considerations for SIP

SIP has some unique characteristics that need to be considered when implementing a highly available, scalable environment.

1. SIP requires that responses be returned in the same order that they are sent 1. by the SIP servers?not necessarily in the order the load balancing device received it. This applies to both TCP and UDP connections, which requires that the load balancing solution be able to handle?and correctly deliver in order?both transport protocols at the same time.
2. SIP can combine multiple sessions in a single, long-lived TCP connection that 2. then must be load balanced across multiple long-lived TCP connections. Because of this, the load balancing solution must be able to disaggregate messages and distribute them to the appropriate server-side TCP connection.
3. The Diameter base protocol has been adopted as the primary signaling protocol for AAA and mobility management in IP Multimedia Subsystem (IMS) and other service provider-focused environments. SIP services in these environments are dependent upon the availability of the Diameter servers; high-availability services for SIP depend on highly available Diameter servers.
4. Similar to application session persistence, SIP maintains information relative to a dialog for the duration of a session. In a highly-available architecture, failure

F5 BIG-IP devices are capable of message-based load balancing.

F5 BIG-IP devices are support SIP persistence using a SIP persistence profile or by manually creating an F5 iRule if there is a need to combine fields or create unique persistence values.

to route all relevant, single-session requests to the same server can cause the loss of state and thus a loss or degradation of service. It is important, therefore, that any high-availability solution be capable of providing SIP persistence to ensure session integrity.

5. Some functions such as QoS and rate shaping, though traditionally based on 5. network layer parameters, require inspection of application (SIP) data and headers. This means that the application of QoS and rate shaping policies that need to occur at the edge of the network must to be applied and/or enforced by a high-availability solution that is application-layer aware.

Persistence

One of the reasons SIP separates signaling from media is help enable a non-disruptive way to adjust call session parameters during the session. Potential ? communications? that might occur on the text-based, signaling side of a SIP session include negotiation of codecs, additional communication capabilities, and QoS.

These options and features are session based, and while most capability information is carried in the protocol itself, the control and data channels must match up in order to apply those parameters to the data-side of any communication. This generally requires access to session tables on the server. This means that once a call is established, it is important that subsequent interaction with the sessions is directed to the same server.

Persisting an SIP dialog to the appropriate server or SBC is mandatory in a high-availability environment. However, there are no standards that designate how this persistence is to be achieved. Thus, the high-availability solution must be flexible enough to support provider-configurable persistence on any SIP field or combination of fields.

Message-Based Load Balancing

Active SIP subscriber counts are now in the millions. This puts an additional strain on the infrastructure due to limitations on the possible combinations of IP addresses and ephemeral ports. Finding a method to aggregate and disaggregate communications into a single triplet fulfills the need for greater scalability, performance, and reliability.

To achieve this, SIP might combine multiple sessions in a single, long-lived TCP connection that must be load balanced across multiple long-lived TCP connections. This requires the ability to disaggregate SIP messages and distribute them to the appropriate server-side TCP connection.

Disaggregation separates individual messages out of a single, shared TCP connection. This means that a high-availability solution must be able to inspect application layer data in order to split out individual messages from the TCP connection, distribute them appropriately, and maintain persistence.

Health Monitoring

A real-time, high-availability environment requires constant monitoring to determine where and how best to route communications. This monitoring must be more than simply measuring RTT between nodes in a network; network availability and the availability of the underlying operating system is not an indicator of SIP service or SBC availability.

Due to the differences in SIP implementations across vendors, simple ?standard? SIP monitoring systems might not support every environment. Thus, any high-availability solution must include a health monitoring sub-system that is flexible enough to work with a variety of implementations in the event of subtle differences in SIP services.

Solutions must also be able to monitor the availability of SIP-dependent services, such as Diameter. If a session cannot be initiated due to unresponsive authentication and authorization services, rendering available SIP services ineffective.

Compression

Because SIP is text-based, it can take advantage of the benefits of compression. Offloading compression from SIP servers and SBCs to a high-availability solution affords greater benefits in the performance of the entire infrastructure, as most high-availability solutions provide hardware-based compression.

Another method of ?compression? takes advantage of SIP?s mechanism for representing common header field names in an abbreviated form. This is useful when messages might be too large to be carried on the transport available, such as when exceeded the MTU using UDP. The abbreviated form of a header can be substituted at any time, and can appear in both long and short form within the same message.

This header field name variability?some of which might be used to provide information to high-availability (HA) solutions?can cause problems if the HA solution cannot recognize the different variations. It is either necessary for the HA solution to be natively capable of recognizing either forms of the headers or to provide the means by which a mapping can be created, ensuring that persistence is not impeded.

Record-Route Rewriting

The Record-Route attribute in SIP represents the SIP proxy. In a high-availability architecture, however, the SIP proxy is located behind a load balancing solution. Thus, the Record-Route value should reflect the load balancing virtual IP address, not the SIP proxy itself. Any high-availability solution must be capable of rewriting the Record-Route attribute before it is finally returned to the client; any subsequent communications will be routed to the appropriate device.

In addition, this rewriting capability needs to consider the existing Record-Route attribute and not arbitrarily remove the proxy?s address, as proxy wishes might need to remain in the path for messages sent during a dialog.

IPv6/IPv4

The depletion of IPv4 addresses impacts the service provider environment more than any other. But a ?rip and replace? migration strategy to IPv6 is often not possible due to the complexity of network and application infrastructures needed to support SIP. In addition, not all clients and vendor solutions support IPv6 natively. In order to migrate smoothly and without a service disruption, an intermediate solution is necessary. One solution is the use of an IPv6/IPv4 gateway. Such gateways must provide complete translation and load balancing between IPv4 and IPv6 networks, and be able to direct traffic across mixed IPv6 and IPv4 devices.

Any high-availability solution, then, should be capable of acting as such a gateway, as they are positioned at the edge of the network. They direct traffic multiple various user and server-side SIP components that may use a mix of IPv4 and IPv6 during the transition to a full Pv6 network.

Conclusion

More and more, service providers and organizations are adopting SIP to supply their converged communications needs. The flexibility and extensibility inherent in SIP makes it an excellent choice but also introduces a variety of technical challenges, making a highly available and reliable SIP-supporting infrastructure necessarily complex.

In order to ensure quality, reliable services it is often necessary to introduce a high-availability solution for both SIP servers and Session Border Controllers. Such solutions must be SIP-aware, and flexible enough to meet the unique needs of each environment as well address the complexity inherent in differences between SIP solution vendor implementations.

By choosing the right high-availability solution for SIP environments, organizations can ensure reliable services while maintaining the ability to rapidly support new functionality and infrastructure without sacrificing uptime or quality.

F5 Networks, Inc.
401 Elliott Avenue West, Seattle, WA 98119
888-882-4447 www.f5.com

Americas
info@f5.com

Asia-Pacific
apacinfo@f5.com

Europe/Middle-East/Africa
emeainfo@f5.com

Japan
f5j-info@f5.com