Tech Brief

Meeting the Challenges of an HA Architecture for IBM WebSphere SIP

Voice and multi-media features available in IBM WebSphere Application Server enable a new generation of integrated applications but also introduce new challenges for IT staff. The inclusion of F5 BIG-IP LTM as a core component of a high availability SIP deployment addresses those challenges without increasing network complexity.

by Lori MacVittie

Senior Technical Marketing Manager

Contents

Introduction

For any network service, high availability is desirable, and it is absolutely mandatory for business continuity in many organizations. Such organizations' increasing reliance on SIP (Session Initiation Protocol) to enable communication between employees, partners, and customers elevates the importance of ensuring a high availability (HA) environment.

In recognition of that reliance, IBM introduced a SIP container into version 6.1 of its WebSphere Application Server (WAS). This included support for JSR 116—the SIP Servlet 1.0 API. JSR 116 enables developers to create services such as back-to-back user agents (B2BUA) and proxy-based applications.

An HA environment for proxies, however, involves several challenges. These challenges can be met with configurations of F5® BIG-IP® Local Traffic Manager™ (LTM) in a WebSphere SIP deployment.

F5 products and IBM WebSphere SIP proxies may be integrated into many possible architectures to deploy an HA environment for SIP-based communications. A high-level illustration of the most common architectures and configurations can be useful for discussion but should not be considered an exhaustive study of all possible integration and configuration options. Detailed configuration information is available in the respective manuals for BIG-IP LTM and IBM WebSphere Application Server.

Prerequisite Knowledge and Terminology

An HA SIP deployment with F5 technologies requires general familiarity with SIP, high-availability architectures, and IBM WebSphere Application Server as well as a basic understanding of the workings of SSL and a working knowledge of load balancing techniques, functions, and features. A few common abbreviations, general SIP terms, and product-specific references include:

Proxy: A proxy is a server or network component that forwards a request on behalf of the requestor. A proxy can be stateful or stateless. A stateless SIP proxy does not maintain the state of the SIP request or monitor the transaction once it is forwarded to the next hop (node). A stateful proxy does just the opposite, tracking and maintaining the transaction state throughout the session.

User Agent Client (UAC): A UAC creates a request or initiates a dialog.

User Agent Server (UAS): A UAS handles a request or dialog.

Back-to-Back User Agent (B2BUA): A B2BUA is a converged UAS/UAC that mediates a dialog. A B2BUA terminates SIP sessions, acting as the ultimate endpoint, and duplicates the dialog with the actual client. This is often used to enable SIP when end-points are within protected networks, that is, when NAT is in use. In service provider networks, a B2BUA would be referred to as a session border controller (SBC).

F5 BIG-IP LTM: An F5 product used to provide load balancing and application delivery services for a particular web service or other application. BIG-IP LTM manages traffic in a data center or a group of servers.

Virtual IP address (VIP): An F5 term for a host representing a resource managed by BIG-IP LTM.

Secure Network Address Translation (SNAT): SNATs are used when traffic originates inside a protected network and must be translated to a VIP or external network address to enable bidirectional communication.

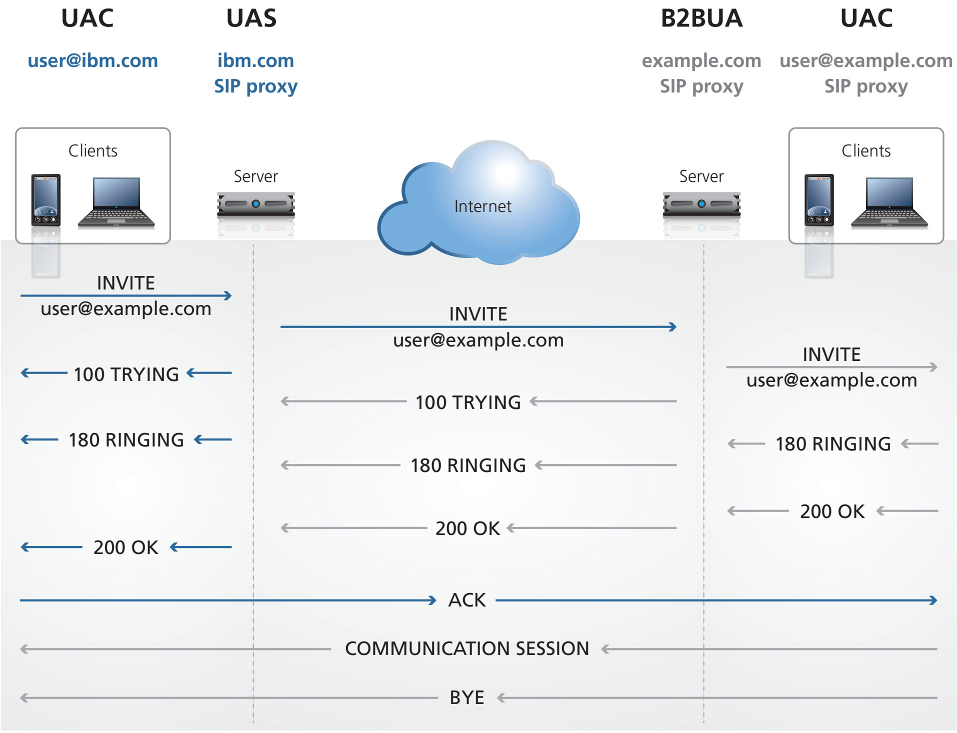


Figure 1: SIP sessions use proxies to mediate communication between clients across protected network boundaries.

HA SIP Challenges

The challenges associated with enabling an HA environment for proxies are three-fold. First, the HA service provider must be able to accurately monitor and subsequently detect an outage. For SIP communications, this requires a more specific data exchange than is often used in simple HTTP monitoring.

Second, because there may be more than one active SIP proxy, the HA provider must accurately route existing communications to the SIP proxy through which the session was originally established. This requires the insertion of persistence into the architecture at the load balancing service, which in turn necessitates the ability to perform deep content inspection on both secured and unsecured SIP protocol traffic.

Finally, SIP, like many of its UDP-based predecessors, introduces unique challenges for proxies and load balancing services. SIP services assume direct, bidirectional communication between parties. Network data (IP and port) must be known prior to traversal through upstream routing devices, because they are encoded in the payload. This requirement is often met by the use of SNAT to ensure traffic is routed properly back through the network. The challenge with this technique is an ephemeral port constraint, limiting SNAT options to fewer than 65,000—a limit that is often exceeded even in modest SIP deployments because of the protocol's heavy reliance on multiple ports per session. Use of the Via header, which specifies a trail of upstream proxies, provides one solution by allowing proper reverse traversal of the network without imposing the capacity-limiting constraints of SNAT.

With the appropriate architecture and configuration, BIG-IP LTM in a WebSphere SIP deployment can overcome these challenges to provide high availability while supporting secured and unsecured communications.

General HA Architecture

A high-availability environment for SIP when using IBM WebSphere Application Servers and proxies requires the configuration of replication domains to enable memory-to-memory replication between SIP containers. This ensures a non-disruptive failover of proxy to proxy or server to server—what is often known more generally as a stateful failover.

IBM SIP proxies, as noted earlier, are stateless. Therefore the pair of BIG-IP LTM devices providing load balancing services is also charged with providing persistence to ensure proper routing of communication to the proxy and server on which the session was originally established.

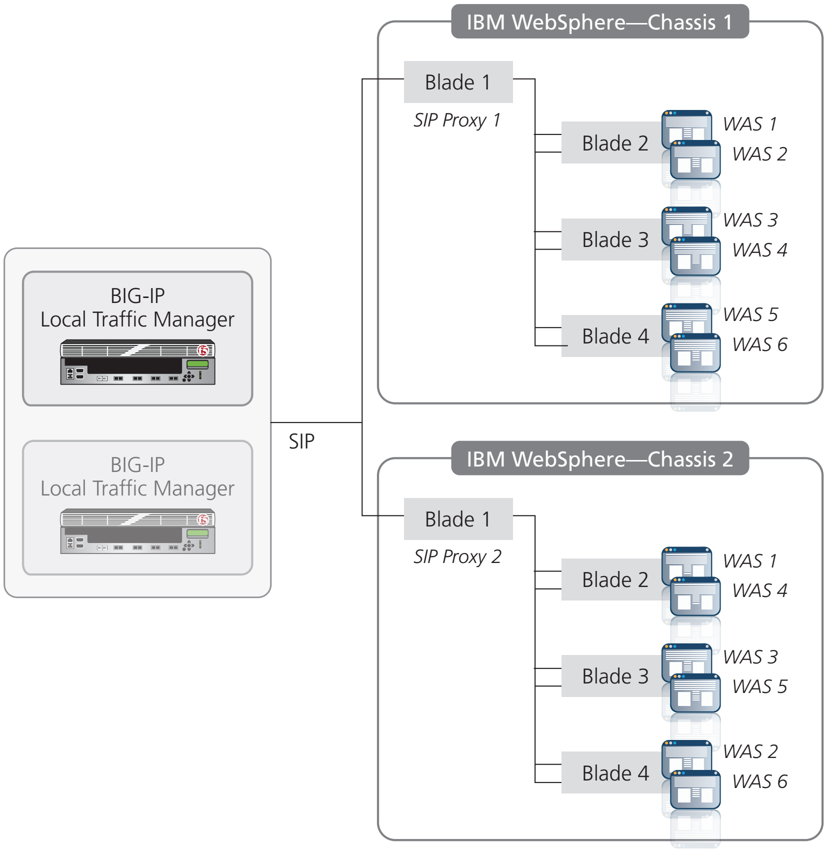


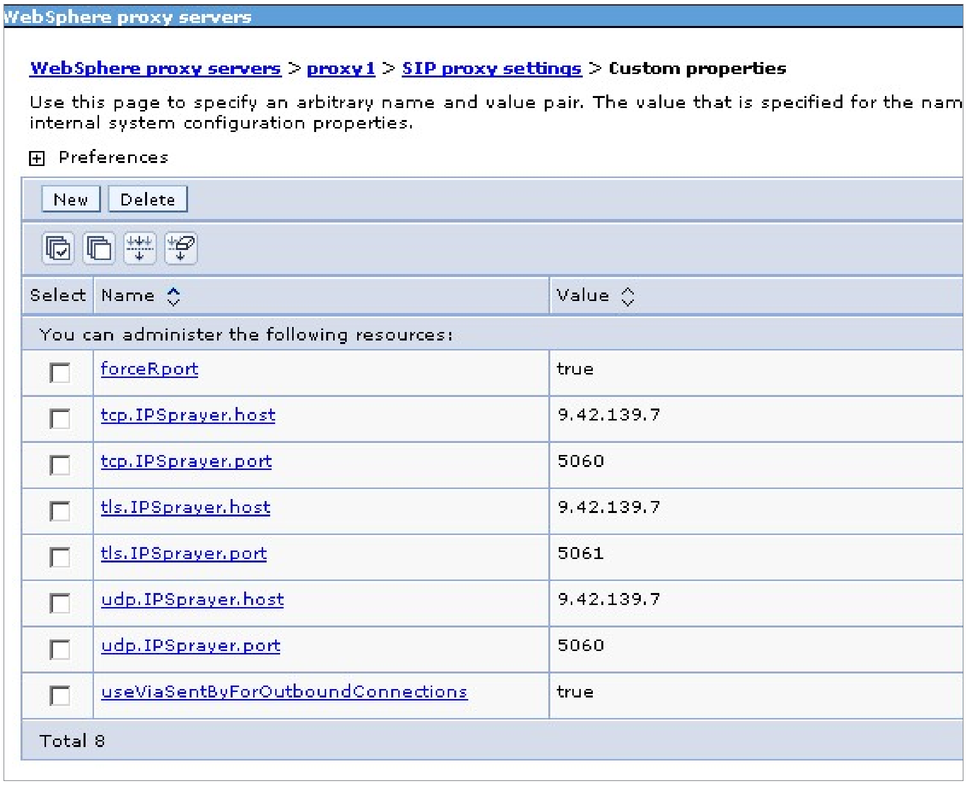
Figure 2: A high-availability architecture with BIG-IP LTM and IBM WebSphere.

Figure 2 shows a high-availability architecture with F5 BIG-IP LTM devices and two IBM SIP proxy servers using IBM blade chassis, each housing four blade servers. In the first chassis, the WebSphere ND Deployment Manager runs on the same blade as the SIP Proxy. The remaining three blades contain two application servers each. In the second chassis, the first blade contains a proxy server and the remaining three blades contain two application servers each.

In this configuration, BIG-IP LTM tracks availability of the SIP proxy servers by sending a stream of SIP OPTIONS messages at regular intervals. The proxy is expected to respond with a 200 OK message.

For this configuration to provide the high availability required, the SIP proxy must be configured to accept and handle all incoming traffic from BIG-IP LTM. If the proxy server does not respond in a pre-determined amount of time, BIG-IP LTM interprets this lack of response as an acknowledgement of an inactive or disabled proxy server. BIG-IP LTM discontinues forwarding client traffic to the proxy, but continues to send OPTIONS requests to the server. If and when the proxy server responds with a 200 OK message, BIG-IP LTM will re-enable the proxy and again route requests to it.

For this process to work, the WebSphere SIP Proxy must be configured to expect OPTIONS messages from BIG-IP LTM. In addition, a set of SIP proxy-specific custom properties must be present and configured to enable interoperability between BIG-IP LTM and the WebSphere SIP proxy server.



WebSphere proxy servers > proxy1 > SIP proxy settings > **Custom properties**

Use this page to specify an arbitrary name and value pair. The value that is specified for the nam
internal system configuration properties.

Select	Name	Value
You can administer the following resources:		
□	forceRport	true
□	tcp.IPSprayer.host	9.42.139.7
□	tcp.IPSprayer.port	5060
□	tls.IPSprayer.host	9.42.139.7
□	tls.IPSprayer.port	5061
□	udp.IPSprayer.host	9.42.139.7
□	udp.IPSprayer.port	5060
□	useViaSentByForOutboundConnections	true
Total 8		

Figure 3: The WebSphere SIP proxy settings required to communicate with BIG-IP LTM.

Technical Architecture

The architecture required to implement a high-availability SIP environment comprises multiple virtual servers, pools, and monitors on BIG-IP LTM. This is necessary to support use of both TCP and UDP with a secured (TLS) and unsecured traffic mix.

The SIP proxies are instances of WebSphere with its "Proxy Mode" installed and configured. The application servers are also WebSphere servers with the "SIP Persona" enabled. BIG-IP LTM is configured as a SIP proxy to fully participate in the flow while providing HA and load balancing services. See Figure 4.

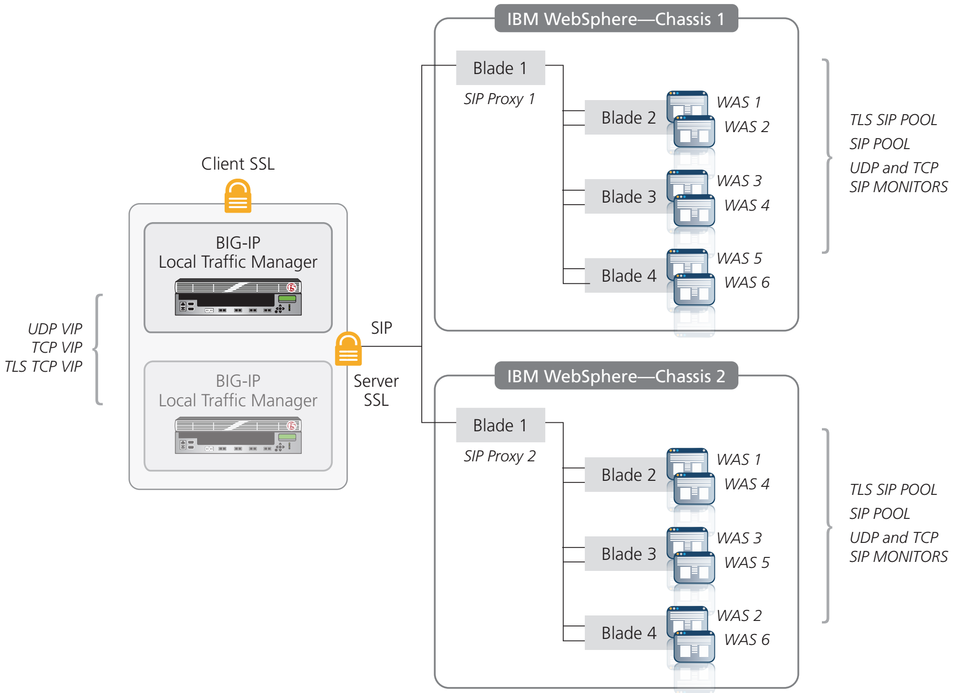


Figure 4: This BIG-IP LTM configuration enables an HA environment with WebSphere SIP proxies.

In this architecture, BIG-IP LTM serves as the initial SIP proxy—a B2BUA—to which clients connect. In addition to monitoring the IBM SIP proxy instances for availability, BIG-IP LTM is also responsible for managing SIP sessions to ensure proper routing throughout the life of the session. On an initial connection (handshake), BIG-IP LTM will select an appropriate SIP proxy, insert a Via header, and store a mapping of Call-Id to the SIP proxy in its session state table to ensure all subsequent messages related to that call session are consistently routed to the same SIP proxy.

The decision on which SIP proxy is appropriate can be as simple as selecting a node based on an industry standard load-balancing algorithm or it can be highly complex and based on both the ability of BIG-IP LTM to extract SIP variables and its monitoring of the health, current capacity, and performance of the SIP proxies. For example, BIG-IP LTM can be configured to route high-priority calls to the SIP proxy with the lightest load to ensure the fastest possible response.

When managing encrypted communications, BIG-IP LTM terminates the secure session and performs the same inspection and insertion process. This effectively offloads the cryptographic processing burden from the SIP proxies and improves overall performance by executing that processing in a hardware-accelerated environment. This offloading has a cascading effect on the performance and capacity of the SIP proxies by freeing up resources that can be used to service other clients and perform SIP-specific processing. If required by operational policy, BIG-IP LTM will re-encrypt the session before forwarding it to the appropriate SIP proxy.

In the event a SIP proxy becomes unavailable based on the configured monitoring, BIG-IP LTM will stop forwarding requests to the disabled proxy until it has responded appropriately. In the interim, BIG-IP LTM routes requests to available proxies and, by virtue of the session replication configured at the application server layer, conversations will continue uninterrupted. Unlike many load balancing services for SIP that monitor only the availability of the port on the SIP proxy, BIG-IP LTM can ensure that SIP services are not only available but executing correctly by understanding SIP and inspecting responses to ensure their correctness.

Conclusion

The deployment of multi-media features such as VoIP on traditional platforms will continue to increase as development platforms broaden their support for the protocols required to develop integrated applications. The challenges associated with these often more complex communication protocols can arise when high availability architectures are introduced. The nature of stateless proxies requires a mechanism by which sessions can be appropriately routed during both normal operation and a failover event. This mechanism requires sufficient capabilities to inspect and correctly interpret protocol-specific payloads as well as intelligent monitoring of downstream services for availability. The sensitivity of SIP to interruptions is high, causing call quality to degrade.

Incorporating BIG-IP LTM as part of a WebSphere SIP implementation helps avoid degradation in the quality of communications that are increasingly critical to business operations. BIG-IP LTM helps safeguard the ability to connect with customers by intelligently monitoring the health of WebSphere SIP proxies and re-routing communications when necessary. By inspecting both unsecured and secured communications, BIG-IP LTM ensures proper routing of sessions and further eliminates limitations imposed by the use of SNAT at the network layer without increasing complexity at that layer.

BIG-IP LTM is an integral component in a WebSphere SIP deployment, providing the necessary intelligence and control required to scale seamlessly and handle effortlessly, without disruption, what would otherwise be highly noticeable failures.

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

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com