F5 White Paper

Session Initiated Protocol (SIP) and Message-based Load Balancing (MBLB)

The ability to provide new and creative methods of communications has ensured an SIP presence in almost every organization. The use of message-based load balancing offers the ability to aggregate and disaggregate SIP messaging, and prevents message-blocking asymmetric server return.

By Paul Stalvig
Technical Marketing Manager

Contents

Introduction **3**

SIP Communications **3**

Aggregation/Disaggregation 4

TLS Offload 4

Message-based Load Balancing 4

Blocking 5

Location 5

Conclusion **6**

Introduction

At the service provider level, the number of VLAN, TCP (or UDP) port, and IP triplet combinations is affecting the Session Initiated Protocol (SIP). Between servers, where the VLAN, TCP port, and IP variations are limited, the ephemeral ports are constantly being pushed and overrun. Because of active SIP subscriber counts in the millions, the potential to use various combinations of IP addresses and ephemeral ports is not realistic. Providing a method to aggregate and disaggregate these communications into a single triplet—or as an individual stream of information—enables greater scalability, performance, and reliability, and relieves the strain caused by these limitations. Even with the use of IPv6 IP addresses, the IP address limitation is often found between the SIP proxy and SIP server, because aggregating communications provides performance advantages that outweigh the relief of IPv6 triplet combinations.

Aggregation is the combining of multiple communications within a single stream connection. Disaggregation is the separation of the communications from the stream connection.

The aggregation and disaggregation of SIP communications can lead to improved user experiences and increased average revenue per user (ARPU), while simultaneously decreasing or eliminating downtime. The need to increase performance and decrease bottlenecks in IP Multimedia Subsystem (IMS), Service Delivery Platform (SDP), and Voice over IP (VoIP) architectures has grown with the increase in availability and adoption of applications that use SIP.

SIP Communications

As communications protocols converge on IP, they can create issues at the network level. SIP is no exception. With inherent asynchronous routing and the need for continual updating of long-lasting communications, the impact of SIP on the network is only now being fully understood. But, SIP broadens user capabilities and, consequently, improves the user experience. By enabling more devices (laptop, cell phone, PDA, etc.), to establish, maintain, and improve communications, user experience improves.

In order to maintain communications, most implementations of SIP are, at best, chatty. This chattiness can range from simple updates during communications to the complete flooding of networks with communications control and reporting. At first glance, it would seem that reducing the chattiness of SIP could solve many issues; however, this can ultimately degrade the user experience.

One of the reasons that SIP can become chatty is that the actual SIP communications do not always follow the same network path as their association.

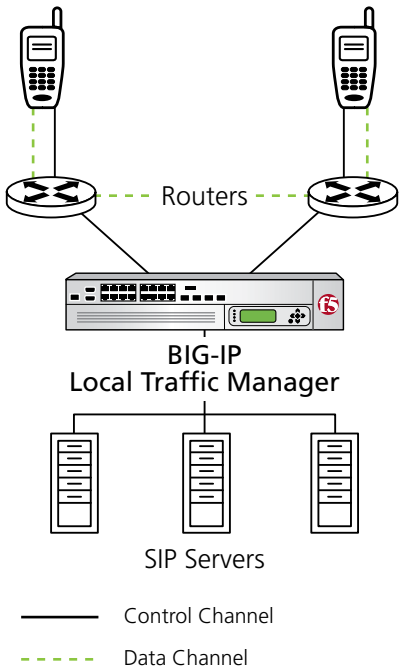

Figure 1: *Asynchronous Routing*

Asynchronous routing is often seen as a flaw in SIP, but it actually can remove control (or signaling) of information from affecting the data (or voice) channel information. By performing the separation of control information from the data channel information, the SIP servers are better positioned to provide analysis and performance information to improve the data channel stream and ultimately ensure a good user experience. Ultimately, without the SIP (control channel) chattiness, the data channel could be lost and would not reconnect.

Any message within SIP must belong to the communications between a user device and the server. When multiple SIP servers are used, then all of the corresponding messages must be sent to the correct SIP server for them to have any effect. This includes the connection performance details from the communications, such as control-channel data (as seen in *Figure 1*). The SIP and control-channel data (noted in black) would flow from the devices to the SIP server, whereas the data-channel information (noted in green) would flow between the devices and their routers without going to the server.

Aggregation/Disaggregation

F5 BIG-IP® Local Traffic Manager™ (LTM) can provide aggregation/disaggregation of the connections to the SIP servers using TCP, TLS/TCP, UDP, or SCTP. This enables the SIP servers to perform operations on the information in the connection without the overhead of numerous smaller connections. Aggregation/disaggregation improves the session's setup time and enables SIP servers and proxies to focus on their primary purpose instead of managing the numerous connections. The reduction can provide increased performance and delay the need for additional servers. When it becomes necessary to increase servers, F5 BIG-IP LTM enables health-checking of any new device to ensure proper functionality.

TLS Offload

BIG-IP LTM can also provide increased SIP server performance by terminating TLS/SSL sessions. By using hardware-accelerated encryption, the BIG-IP is better able to handle the workload of encrypting and decrypting the sessions. The transport layer security (TLS) offloading function improves SIP server performance by 50 percent.

Message-based Load Balancing

SIP communications use a message-based context of communications. The ability to perform load balancing of individual messages is referred to as message-based

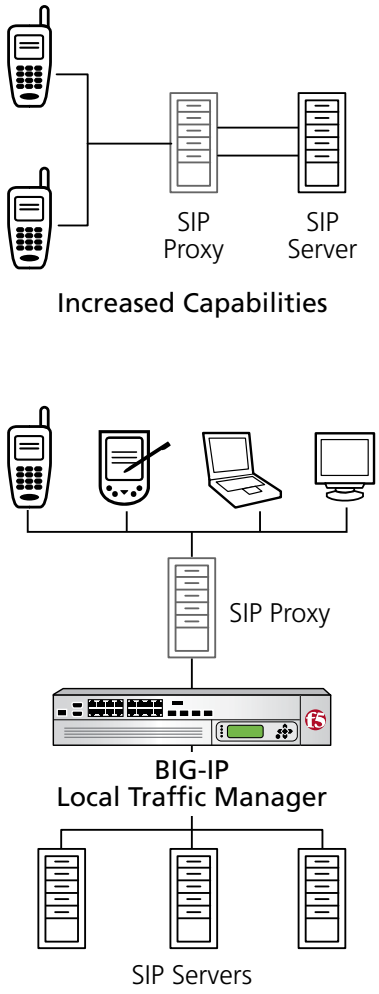

Figure 2: *Aggregation*

load balancing (MBLB). Because SIP is used for voice communications, sessions must be persisted for the duration of the conversation. This requires that the connection between the client device and the SIP server must be persisted or the communication channel will fail.

SIP control channel connections can pass very little data from the server to the client during high-quality communications. If the data channel communications experience some form of degradation, then the control channel communications will pass messages back and forth to re-establish the proper quality of service. This quality is directly related to the user experience.

Figure 3: *One TCP connection containing multiple SIP sessions*

BIG-IP LTM SIP message-based load balancing provides a message-by-message look within the SIP communications to determine and maintain the correct client-server connections. In smaller organizations this may not be a problem; The BIG-IP LTM can simply persist based on a source IP address. In the service provider space there can be thousands of connections originating from a single IP address. BIG-IP LTM load balances SIP communications accurately—even if they are aggregated into a TCP, UDP, or SCTP stream—by reading the SIP messages one at a time.

Blocking

Normally associated with TCP, "queuing" or "call-flow blocking," occurs when one piece of information needs to be retransmitted and causes all subsequent information to wait, pending the receipt and acknowledgement of the lost information. While in most network communications environments this does not cause negative issues, the delays caused by blocking in SIP can mean the difference between a positive and a negative user experience.

By using a multi-stream construct, within the SIP aggregation, as shown in *Figure 3*, BIG-IP LTM is able to continue processing communications without waiting for a response from the first session. Only the communications stream that is affected would be blocked; the other communications streams would continue to flow.

Location

Each service provider will have different requirements for the architectural location for the new capabilities. However, with the increased advantages of using BIG-IP LTM SIP message-based load balancing to control the performance of critical connectivity, the location in the architecture becomes easier. For example, one carrier may need to provide the benefits between the Call Session

Control Function (CSCF) and the Session Border Control (SBC). Another carrier needs to be able to provide the functionality within the CSCF between the "Interrogating," "Proxy," and "Serving" CSCFs. Both carriers could also use the SIP message-based load balancing and BIG-IP LTM capabilities to provision the key benefits between the control layer CSCF and the application servers in the IMS infrastructure.

Conclusion

The F5's message-based load balancing ensures a good user experience by maintaining the connection from the user's device to the server. If the direct communications are not performing correctly, the server can immediately step in and re-establish itself as the hub for the communications, ensuring that the user experience and performance levels are maintained.

F5 BIG-IP Local Traffic Manager can help organizations analyze the SIP communications to determine current and future network and system performance needs. BIG-IP LTM can also ensure that the SIP server is running properly by implementing various health checks. Before an additional SIP server is activated, the server must pass a health check. This enables SIP services to expand without downtime.

The ability to provide new and creative methods of communications has ensured a SIP presence in almost every organization. The use of message-based load balancing offers the ability to aggregate and disaggregate SIP messaging, and prevents message blocking asymmetric server return. BIG-IP LTM can improve user experience, increase the user's device lists and capabilities, and provide a direct link to average revenue per user. BIG-IP LTM enables growth and performance analysis while eliminating downtime.

Key Benefits

- Increase user capabilities
- Increase user experience
- Increase Average Revenue Per User (ARPU)
- Provide a measurement into user experience
- Provide performance statistics
- Increase performance of SIP servers using (dis)aggregation with TCP, TLS over TCP, UDP, and SCTP
- Increase capabilities without downtime
- Increase SIP server capacity
- Decrease SIP session setup/ teardown time
- Integrate into existing and future routing methods, using RIP, BGP, OSPF, and IS-IS—whether it is IPv4 or IPv6

F5 Networks, Inc.
Corporate Headquarters
401 Elliott Avenue West
Seattle, WA 98119
+1-206-272-5555 Phone
(888) 88BIGIP Toll-free
+1-206-272-5556 Fax
www.f5.com
info@f5.com

F5 Networks
Asia-Pacific
+65-6533-6103 Phone
+65-6533-6106 Fax
info.asia@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
+44 (0) 1932 582 000 Phone
+44 (0) 1932 582 001 Fax
emeainfo@f5.com

F5 Networks
Japan K.K.
+81-3-5114-3200 Phone
+81-3-5114-3201 Fax
info@f5networks.co.jp

WP-SIP-MBLB 05/08