Brocade SIP-Intelligent Application Switching for IP Communication Services

Highlights the value of highly available and scalable application switching and details the specialized capabilities required in the application switches for deployment in a SIP service infrastructure for maximum benefit.
Session Initiation Protocol (SIP) is a key enabling technology for new-generation IP-based communication services such as voice, messaging, streaming, and video. This paper highlights the challenges of deploying these services and discusses solutions to help deliver commercially viable, always available, and highly scalable SIP-based communication services.

Application-aware Layer 4-7 load balancing switches have played a vital role in increasing service uptime and scalability for IP and Web-based data applications. Enterprise, government, service provider and e-commerce organizations rely on this technology for their business- and mission-critical applications. By embedding SIP-based service awareness and intelligence into the application switches, Brocade delivers a critical solution to enable highly reliable, scalable and secure SIP-based communication services for enterprise and service provider organizations.

OVERVIEW AND CHALLENGES
In the mid-nineties, Web technologies revolutionized commerce and business transactions by enabling a wave of Web-based application services. No other technology has since been more widely deployed and used by businesses, people and organizations. Session Initiation Protocol (SIP) and Voice over IP (VoIP) technologies are promising a similar revolution in communication services. IP-based communication services are being deployed by most major providers, and are growing in prominence with end users. Although IP-based voice communication is a small percentage of the overall communication services business today, it is bound to be the predominant service in the years ahead. New value-added services, including messaging and streaming, will migrate to an IP infrastructure using the highly-flexible and extensible SIP technologies.

When Web technologies and services became critical business drivers for most organizations, and in many cases life support systems for businesses, the limitations of traditional server farm and application models became obvious. Poor availability, application performance, security and infrastructure scalability became insurmountable hurdles to delivering business-class IP and Web services. Layer 4-7 load balancing and Web switching solutions emerged as the technology of choice to overcome these limitations. Application-aware networking devices were deployed in front of the Web server farms to build a virtualized infrastructure, which
was resilient, scalable and secure, and is totally transparent to the end user. These switches delivered service resilience from server failures, on-demand and unlimited server farm scalability with commodity servers, and optimized performance by efficient server selection and acceleration. The application switches continue to evolve, even today, to deliver even more value-added optimization and acceleration capabilities for Web applications and services.

Today, service providers, carriers and Enterprises face a similar set of challenges in deploying IP-based communication services – poor service availability due to server and application failures, lack of scalability due to server limitations, and poor performance and quality of service for voice communications. These challenges have a magnified impact on voice services due to the stringent quality and uptime requirements and expectations for these communication services. Additionally, capacity and performance demands are significantly higher for voice services because of the high volume of sessions and users. Businesses can overcome these challenges with innovative, cost-effective and commercially proven application switching solutions.

**SIP NETWORK AND SERVICE REQUIREMENTS**

**SIP Service Network**

Figure 1 shows a typical network consisting of SIP clients and SIP servers. SIP clients (user agents) are devices in the IP network that engage in communication. These devices may be IP phones or other IP devices equipped with necessary tools to help facilitate various types of communication like voice, messaging, streaming and video. Prior to engaging in a dialogue, SIP clients must locate each other and agree to communicate. Centralized IP-based SIP application servers (Proxy and Registrar servers) play a central role in facilitating control communication between the clients to help locate and connect to each other. Once the two clients accept to communicate, the dialogue starts over a separate point-to-point transport channel that does not include the servers. Once the dialogue starts, the SIP servers are no longer involved in the communication, until the clients need to terminate the conversation.

![Figure 1. SIP network overview.](image)

There are two types of SIP servers that are central to a SIP service: 1) SIP Proxy and 2) SIP Registrar. Other servers, like the media, redirect and message servers, are also relevant to the overall SIP service but, to help highlight the value proposition of the SIP application switching solution, this document will only use the Proxy and Registrar servers as examples. The Registrar server maintains a database of SIP clients and their location. The clients notify the Registrar server about their location using REGISTER messages defined in the SIP protocol. The Proxy server facilitates call processing by contacting the target client and...
establishing the transport channel for the dialogue. The Proxy server relies on the Registrar
database to identify the client location to initiate the call. These two applications may be
deployed together on a single physical server, or on separate dedicated servers for the
respective functions at the discretion of the service manager.

The SIP protocol uses the underlying IP network to exchange messages. The protocol
specification allows for SIP messages to be exchanged using the TCP or UDP transport
protocol. A standard (well-known) port 5060 is assigned for SIP, although any other port
can be defined by the network administrator. SIP services are most commonly implemented
using UDP transport, and this document assumes that the service is implemented using UDP
transport. However, neither the solution concepts nor the value proposition change materially
when applied to a TCP-based SIP service implementation.

**SIP Service Requirements**

The SIP Proxy and Registrar services are extremely critical and fundamental to delivering
any SIP-based IP communication service (Voice, Video, Messaging, Streaming and Data).
These services must maintain five nines availability, scale to millions of calls, and deliver
performance in the tens of thousands of calls per second. High call volume also places
performance and capacity demands on the server farm and switching infrastructure.
The network and server farm infrastructure must be capable of supporting millions of
concurrently active SIP sessions. The servers and network infrastructure cannot be the single
points of failure, or impediments to service scalability and performance. The traditional
model of implementing Proxy and/or Registrar functions on a single server or a cluster of
servers is clearly not the answer to achieve the desired scale and availability for these critical
services. The SIP service infrastructure must meet the following key requirements:

- **Ultra High Availability.** Service must be resilient to failures of individual servers,
  applications and network devices. Downtime of a server, application or network device
  must not mean service downtime. SIP application infrastructure must be self healing,
  with rapid detection of failures and transparent redirection of clients to other available
  resources to improve service uptime and end-user experience.

- **On-Demand Scalability.** As IP-based communication services grow, the infrastructure must
  scale to accommodate growth without forklift upgrades, and without the use of expensive
  and hard to manage high-end servers. Incremental and on-demand server farm scalability
  helps avoid unnecessary capital expenditure, and eliminates the need for faulty long-term
  capacity planning.

- **High Performance.** Infrastructure must deliver consistently high service performance and
  response time independent of the call volume and traffic load in the network. The self-
  healing infrastructure must detect server load and response time dynamically, and ensure
  that client requests are sent to servers that are available and respond the fastest.

- **Security.** IP networks and services are prone to attacks due to the open-standards nature
  of the protocols, and SIP-based services are no exception. SIP servers and applications
  must be safe from attacks that threaten service uptime. Security intelligence at the SIP
  application level must be integrated into the network to block threats early and further
  upstream to increase security and resource efficiency.
SIP-INTELLIGENT APPLICATION SWITCHING

Application Switching Overview

Application switch acts as traffic cop in front of the servers, and use their application-level intelligence to ensure high service uptime and superior scalability. Application switches efficiently distribute client requests to the “best” server in the pool. The switches consider server availability, load, response time, and other performance metrics in server selection to provide the best end-user response time and availability. Using sophisticated “health checks” to servers and applications, application switches identify unavailable resources in real time and switch users to other available servers. They also help scale server farms on demand by adjusting server capacity without impacting application availability. When more capacity is needed, a new server can simply be added to the pool without impacting the end user.

Brocade is a leader in delivering application switching solutions to the most demanding customers for their business- and mission-critical IP and Web application services. The ServerIron family of Layer 4-7 application switches leads the market in intelligence, performance, security and price. With Brocade’s innovative technology, high-performance architecture and comprehensive application intelligence, ServerIron switches deliver wire-speed Gigabit rate protection against DoS attacks, and provide the highest availability and maximum scalability to IP and Web applications. The world’s most demanding Service Provider, E-Commerce, and Enterprise customers use Brocade’s solution in their most demanding application infrastructures to deliver non-stop application services at peak performance.

The ServerIron switches feature best-of-breed network and application level intelligence to maximize server utilization, protect capital investment, and allow for infrastructure growth on demand. These switches distribute user traffic to multiple servers and increase performance of business-critical services. The ServerIron switches allow organizations to leverage “bite-sized” commodity servers to build large application infrastructure to provide IP services to millions of users.
**SIP Application Switching**

Brocade’s ServerIron switches feature advanced SIP application intelligence to deliver the benefits of high availability, scalability, security and performance to IP-based communication services. The ServerIron switches front-end the SIP servers and deliver high availability, accelerated performance, on-demand scalability and robust security to IP-based communication services. These switches feature highly sophisticated and customizable SIP application intelligence to efficiently manage SIP traffic, and support advanced health monitoring at the SIP application level to ensure service availability and best response time. Each new SIP session is assigned to one of the *best* servers based on load and other metrics. All messages belonging to a given SIP session are forwarded to the same server to preserve transaction integrity. The ServerIron switches also protect the SIP servers from various types of attacks by filtering SIP messages based against user-configurable and highly-customizable content rules.

Figure 2 shows a SIP server farm with the Brocade application switches enabling load balancing and high availability to multiple SIP servers.

**SIP Load Balancing and Persistence**

The clients (phones, hosts or other end-user devices) connect to the SIP service using a virtual IP address in the network, which resides on the ServerIron switch. When the SIP clients communicate with the ServerIron switch using the virtual address, the ServerIron switch selects the *best* server based on the server farm state and load conditions. Additional content-level criteria like message type may be used in server selection to increase the granularity of traffic management, and therefore further optimize server utilization and capacity. For example, REGISTER messages can be distributed only among those servers that are dedicated for this service. The ServerIron switch load balances “INVITE” requests to multiple SIP Proxy servers based on availability and other metrics. Subsequent messages, after the initial INVITE, belonging to the same session are identified by the unique “CALL-ID” field in the SIP header, and are sent to the same server selected at the time of initial INVITewith a different CALL-ID is subject to load balancing, which assigns it to the *best* server in the pool.

Growth is a natural assumption to any business or service, and a network and server farm infrastructure delivering critical communication services must be designed to scale with subscriber and call volume growand also the evolution of the service to carry higher amounts of traffic. When a ServerIron application switch isfront-ending the server farm, adding
capacity simply requires the network manager to add another server to the resource pool and configure the ServerIron switches to use the new server. Similarly, the network manager can remove servers from a pool to perform maintenance and/or upgrade operations while the ServerIron switch uses the remaining servers for user traffic. The end users are totally transparent to physical resource changes in the infrastructure. The clients always communicate to the virtual address and get connected to one of the available servers. On-demand and end user transparent scalability minimizes service disruption due to capacity and infrastructure changes.

**SIP Server and Application Health Monitoring**

One of the key functions of an application switch is its ability to monitor the health of the servers and applications in real time and react to failures or performance degradation by redirecting clients to alternative resources. The ServerIron switch provides highly customizable SIP-specific health monitoring to rapidly determine any degradation or failure of SIP servers and application functions. The switches send REGISTER and OPT messages to the SIP Proxy and Registrar servers. When an error-free response is received, the ServerIron switch marks the server as being available and starts assigning new SIP sessions to the available servers. Health monitoring messages are user configurable per server and per application port. The switches periodically sends health monitoring messages at a user-configured frequency, which can be as low as once a second per server.

Brocade’s unique system architecture includes a dedicated processor for health monitoring and device management, which significantly increases the reliability and efficiency of health monitoring and therefore improves the overall service availability. Below are samples of healthcheck messages sent by ServerIron switches to SIP servers. Responses to these health monitoring messages from the servers with “200 OK” indicate a healthy server and application. Network managers can tune the switch to accept other response codes to indicate a healthy and available server. Additionally, content matching rules can be applied to the response messages to detect server and application status at finer granularity using information embedded deep in the response message.

**REGISTER Health Monitoring Message**

- ServerIron request message to each individual server
  Request: REGISTER sip:companyname.com SIP/2.0
  Via: SIP/2.0/UDP 172.16.1.2:1034;branch=z9hG4bKh123456
  Max_Forwards: 70
  To: <sip:dummy-user@companyname.com>
  From: Server Iron <sip:dummy-user@companyname.com>; tag=123456
  Call-ID: 25875172.16.1.155@172.16.1.2
  CSeq: 101 REGISTER
  Contact: <sip:172.16.1.2>
  Expires: 0
  Content-Length: 0

- Expected response message from a *healthy* server
  Response: SIP/2.0 200 OK. 1034;branch=z9hG4bKh123456
  Via: SIP/2.0/UDP 172.16.1.2:1034;branch=z9hG4bKh123456
  Max_Forwards: 70
  To: <sip:dummy-user@companyname.com>;tag=118429
  From: Server Iron <sip:dummy-user@companyname.com>; tag=123456
  Call-ID: 25875172.16.1.155@172.16.1.2
  CSeq: 101 REGISTER
  Contact: <sip:172.16.1.155>
  Expires: 0
  Content-Length: 0
OPTIONS Health Monitoring Message

- ServerIron request message to each individual server
  Request: OPTIONS sip: companyname.com SIP/2.0
  Via: SIP/2.0/UDP 172.16.1.2:1670;branch=z9hG4bKh
  Max_Forwards: 70
  To: <sip: companyname>
  From: Server Iron <sip: dummy-user@companyname.com>;tag-123456
  Call-ID: 25875172.16.1.155@172.16.1.2
  CSeq: 735 OPTIONS
  Contact: <sip:172.16.1.2>Accept: application/sdp
  Content-Length: 0

- Expected response message from a *healthy* server
  Response: SIP/2.0 200 OK
  Via: SIP/2.0/UDP 172.16.1.2:1670;branch=z9hG4bKh123456
  Max_Forwards: 70
  To: <sip: companyname.com>;tag=118429
  From: Server Iron <sip: 172.16.1.100@companyname.com>;tag-123456
  Call-ID: 25875172.16.1.155@172.16.1.2
  CSeq: 735 OPTIONS
  Contact: <sip:172.16.1.155>
  Accept: application/sdp
  Content-Length: 0

New INVITE requests are not sent to servers that fail the real-time health checks. Health monitoring to these unavailable servers continues in the background. When the servers start responding again, they are added back to the pool of available resources and new SIP sessions are assigned to these servers.

High Availability SIP Switching

Brocade application switches are a critical foundation for ultra high availability IP application infrastructure, which is resilient to failures of servers, applications and network devices. The ServerIron switches feature advanced modes of high availability to extend service resilience across the failures of a ServerIron switch. In the high availability mode, two ServerIron switches are deployed to back each other up during failures. During normal operation, the active ServerIron switch synchronizes its sessions state and SIP CALL-ID persistence information to the second backup device in real-time. When the backup switch takes control in the event of a failure, this switch has all the state information to continue forwarding traffic on existing sessions in addition to load balancing new SIP sessions. The switches provide rapid, hitless, and transparent failover of client sessions.

- Hot-Standby Mode. In this mode, two switches are configured to manage the same Virtual IP (application) and act as primary and backup switches. Hot standby mode provides the fastest failover.

- Active-Active Mode. This mode maximizes switch utilization by having both the switches actively serve client traffic during normal operations while offering backup to one another in the event of a failure.
**SIP Server Farm Security**

The ServerIron application switches act as a first line of defense to the SIP servers and applications by preventing illegal users and traffic from flowing through to the servers. Attacks against IP services are extremely common, and because the servers and applications are poorly prepared to defend, the attacks severely impact service availability. The ServerIron switches validate SIP messages for protocol conformance, including specific protocol version, prior to forwarding the request to a server. Additionally, the switches filter illegal SIP message types. By filtering and discarding illegal messages early in the network, the ServerIron switches help offload servers from processing illegal messages and increase the available server capacity for legitimate SIP traffic. The Brocade solution also supports various rate limiting capabilities, preventing application and server abuse by enforcing “normal” service behavior by end users. With support for wire-speed access control and high-performance software-based access policy enforcement against millions of IP addresses/prefixes, the ServerIron switch acts as a “traffic cop” in a service provider environment by blocking traffic from all illegitimate IP hosts.

**ABOUT BROCADE**

Brocade connects the world’s most important information—delivering proven networking solutions for today’s most data-intensive organizations. From the data center to high-performance Ethernet networks, Brocade is extending its near-fifteen-year heritage as a leading innovator of advanced storage and networking technology.

The unique Brocade approach to high-performance networking protects IT investments, reduces costs, and minimizes risk by providing:

- Highly reliable products that reduce TCO and OpEx
- A competitive advantage through innovative and first-to-market products
- The best alternative choice to Cisco in terms of overall price/performance value

The world’s largest enterprise networks, government entities, and global service providers rely on Brocade to maximize the business return on their data. It’s no wonder 90 percent of the world’s most critical business information flows through Brocade solutions. Quite simply, Brocade enables today’s complex businesses to run. Where other vendors produce networking that’s ordinary, Brocade is committed to delivering the extraordinary.

**SUMMARY**

Service providers and traditional carriers are increasingly deploying IP-based voice infrastructure to deliver high-value revenue-generating communication services to their business and end user customers. SIP-based applications are at the foundation of delivering these services, and these applications must maintain high availability and superior scalability to match the voice service expectations and demands.

Brocade is integrating advanced SIP application intelligence into the industry-leading, high-performance ServerIron family of application switches to deliver ultra-high availability, superior scalability, and robust security to the IP-based communication service infrastructure. Brocade ServerIron switches are the industry’s only proven high-performance solution designed to meet the demands of large service providers and carrier customers. These switches scale to millions of calls, and deliver performance of tens of thousands of calls per second. With built-in advanced security features, the Brocade switches also help fortify the SIP server farm, which increases service availability.

For more about SIP (Session Initiation Protocol) RFC: [http://www.ietf.org/rfc/rfc3261.txt](http://www.ietf.org/rfc/rfc3261.txt)
