A dual redundant SIP service

Introduction

The Session Initiation Protocol (SIP) eco-system: a unit of interdependent protocols functioning together within the environment of the Internet to enable multimedia communications.

Matching performance in the SIP eco-system, which has grown organically, to the highly structured hierarchical system of traditional circuit switched networks, is a challenge.

SIP by its nature operates in a distributed network and benefits from basic levels of recovery inherent in IP-based networks. However, in the connected world, the need for network resilience and fault tolerance imposes a higher burden. To achieve the required level of service, the network is expected to maintain steady-state calls (connections) all the time.

To reach the level of network reliability commonly required by fixed and mobile telcos' next generation or 3G networks, there are some scenarios that must be addressed as they have been in traditional, circuit switched networks. Security and quality of service models tend to be the main focus of network architects, but as SIP-based networks mature, more emphasis must be placed on resilience in order to achieve the holy grail of high availability.

It is the need to cater for telco- or carrier-grade levels of availability whilst operating in the SIP eco-system that this white paper is addressed.

A dual redundant SIP service

Lost calls = lost customers and lost revenue

A fundamental requirement of any communications network, be it the PSTN or any 21st century IP-based network, is equipment performance at a level often expressed as 'five nines' availability (99.999% up-time). To meet that minimum criterion, systems must be designed with reliability and fault tolerance in mind. Continuity of service is the paramount factor, such that a system must remain viable when a failure occurs – whether the failure is related to hardware or software.

Above all, to achieve high levels of availability and continuity of service, systems must display resilience. That is the ability to recover from some form of adverse condition or fault. System failure is an expensive business. If part of a mobile phone network fails for several hours during a peak period, for instance, the combined cost to the operator in terms of lost revenue and damage to reputation can dwarf the cost of rectifying the actual fault.

No system can be made 100% reliable; it can only strive to approach the magic number, hence the 'five nines' compromise. Paradoxically, the more complex (and invariably useful) the system, the more potential modes of failure exist and the more failure prone it becomes.

Conventionally, there are two methods of achieving resilience: through 'prevention', by means of redundant components; and by 'cure', through using failover or backup routines.

The requirement for resilience has been addressed for decades in the PSTN, where the integrity of signalling components and preservation of an end-to-end signalling path is vital. To achieve reliable, resilient performance, network architects have employed signalling system number 7 (SS7). Multiple signalling paths to each end point and multiple links between adjacent signalling points represent the classic physical redundancy architecture used worldwide.

The basic structure of SIP and SS7 networks is quite similar. Both architectures employ separated signalling and media (bearer channels). Therefore, the principles of dual redundancy should be employed in SIP networks in order to ensure reliable operation and integrity of communications.

Dual system redundancy

There are three basic tenets of dual system redundancy: 1. physically separated signalling paths; 2. dual-chassis redundant signalling controllers; and 3. synchronisation of status information.

In an SS7 network, the fundamental approach is to duplicate the SS7 interface for a single signalling point in separate, but interconnected chassis. If one chassis fails, the other continues to operate without loss of calls. Additionally, separating the media paths from the signalling paths gives a further degree of resilience or fault tolerance. This is further strengthened by the standard practice of routing the redundant paths through separate physical paths between network nodes.

The dual redundant controllers are also configured to mirror call setup and status information through a synchronisation scheme that ensures the maximum call retention in the event of a failure at any point in the system.

Interestingly, this approach mirrors the classic design common amongst disaster recovery implementations. In fact, most highly available systems stick to this simple design pattern: a single, high quality, multi-purpose physical system with comprehensive internal redundancy running all interdependent functions paired with a second, physically separated, like system.

The overriding purpose of this design is prevention of, or rapid recovery from, a failure, which allows a system to continue to operate despite a partial or complete failure of any significant component.

The aforementioned concept of 'prevention' is employed by addressing areas of unreliability and that of 'cure' by means of built-in recovery routines. In such a way are created fault tolerant systems with redundancy, no single point of failure, and failover or recovery routines that kick in when needed. Resilience is achieved and high availability results – a win:win for all concerned.

A dual redundant SIP service

Dual redundancy with SIP

To achieve the same levels of performance – resilience and service continuity – within a SIP-based network, a similarly redundant architecture is required.

A SIP service can be deployed on two separate controllers that act together in an 'active/passive' configuration to provide a resilient system that can reduce interruptions to user applications if hardware or network issues affect one of the servers.

More particularly, a dual redundant SIP service consists of a pair of separate servers that are configured to share a well-known address to the outside world for SIP traffic. One of these servers will process call control traffic and is identified as the active server, whilst the other server, which is recognised as the passive server, mirrors the state of calls on the active, using a separate connection. When a problem occurs on the active server, the passive server performs a takeover (failover) whereby it assumes the role of the active and manages all existing connected calls, indicating to the application that a takeover has occurred.

Call control applications currently deployed on media processing platforms such as the Prosody family from Aculab can readily take advantage of the resiliency provided by these dual redundant SIP features. Applications do not need to be modified, however, new events and routines are included to make management of the system easier, such as querying the state of the dual redundant system. Also, additional configuration options must be set in applications and individual SIP servers to ensure that the system operates correctly.

How it works

Each application is configured to connect to a pair of dual redundant SIP servers. These SIP servers are in turn configured to accept connections from such remote applications (see figure 1 below). Each dual redundant SIP server is then also configured with information pertaining to its corresponding peer SIP server, including details of how the servers will connect to each other. In addition, each SIP server is programmed with a well-known IP address, known as the 'floating IP address' (see table), that it must configure when it is in active mode. This 'floating IP address' must be the only IP address used to deliver SIP traffic to the active server.

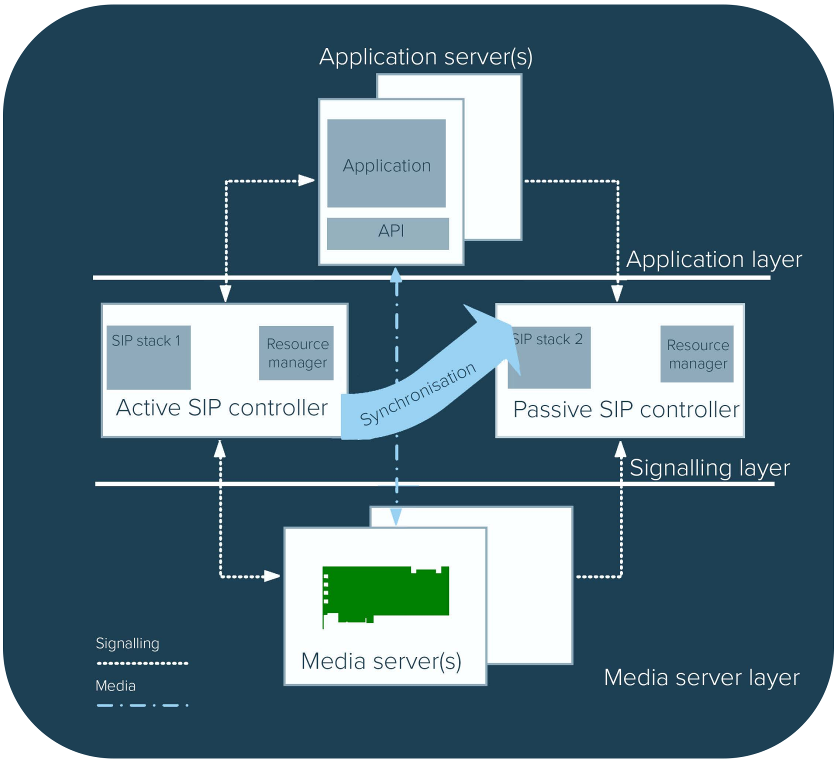

Figure 1 – Dual redundant SIP architecture

A dual redundant SIP service

When an application starts, the SIP servers are designated to operate in either the active role or the passive role. The active server configures the 'floating IP address' and advertises that it owns this address. The passive server remains quiescent, in terms of SIP signalling activity, however, it handles synchronisation messages from the active server.

During normal operation, the active server sends synchronisation messages to the passive server, allowing the passive server to mirror the state of calls on the active server. In the event of a failure, the passive server will assume the active role and it will configure the 'floating IP address' and resume call processing for existing calls and new traffic.

In the event of a failure resulting in a takeover, all calls in an active, established state will remain connected after the failover. It is also possible for an application to force a takeover to occur using maintenance commands, therefore, enabling a server to be shut down gracefully.

Terminology used

Active server

The active server handles incoming and outgoing calls while providing the passive server with data that allows the passive server to takeover in the event of a problem. The active server advertises itself as the owner of the floating IP address, meaning SIP traffic will be directed towards it.

Takeover (failover)

The takeover process results in the passive server becoming an active server when the system discovers a problem with the current active server. The passive server takes on the floating IP address, advertises that it is now the owner of that IP address and becomes the active server.

Passive server

The passive server builds a mirror of the current state of the active server allowing it to takeover in the event of an error. The passive server only handles SIP traffic when it becomes the active server.

Maintenance

When a user wishes to modify the current active server or invoke maintenance functions, the active server can be shut down gracefully allowing calls to migrate to the passive server.

Floating IP address

The floating IP address acts as the public address for the resilient SIP service. The active server will configure one of its interfaces to use this as a virtual address or alias. The system uses ARP to advertise the current owner of the floating IP address to ensure that traffic reaches the correct destination.

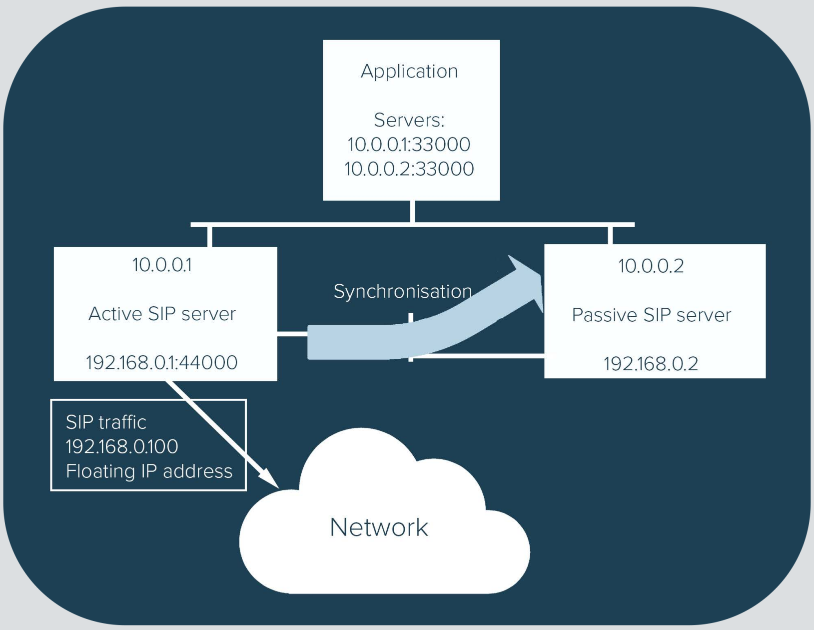

Figure 2 – Dual redundant SIP architecture block diagram

A dual redundant SIP service

Configuration

Each application should be configured identically in order for the system to operate correctly in a dual redundant mode. If multiple applications are installed on different servers the designation of the SIP servers should be the same for each. In addition, the SIP servers must be configured to listen for connections from remote applications and it is necessary for these settings to correspond with the settings that have been chosen for the applications.

The initial classification (active or passive) of the SIP servers is designated when an application connects to the SIP server. To enable intra-system communication, each dual redundant SIP server must use at least two separate network interface cards, one for communication between the application and the SIP service and another for synchronisation between the active and the passive servers.

In a system based on Aculab's enabling technology products, redundancy can be implemented at three key levels, matching service oriented architecture models and creating excellent resilience for an entire system.

1. Taking full advantage of the distributed architecture and remote application capabilities of Prosody X or Prosody S, redundancy at the telephony or media server can be readily achieved

2. Further resilience is assured at the application layer if each application that uses the system resides on a separate server to those used by the SIP servers; Aculab's API includes a mechanism to facilitate such redundancy and failover between applications

3. The dual redundant SIP service provides the third, critical layer of redundancy.

   Of course, both application and media resources can be located on the same servers used for the resilient SIP service, albeit in such designs, adequate arrangements need to be made to ensure that the media connections are maintained after failover. Such arrangements are outside the scope of this white paper. Note that Aculab's API functionality provides a means of indicating the location of resources.

Conclusion

As SIP arrives at a certain level of maturity as a voice delivery technology in major enterprise or business communications service provider networks, the concept of dual redundancy in the SIP network is becoming increasingly critical. Physical distribution of SIP across separate signalling controllers is a real step forward in communication solutions technology. Leveraging the benefits of distributed network architectures and best-practices network design (proven by decades of SS7 deployment) provide for reliable and consistent performance in an increasingly complex network environment.

Dual redundant SIP deployments will clearly differentiate the offerings of platform or equipment vendors and systems integrators who design and build VoIP solutions for the enterprise or service provider market. These companies have a straightforward opportunity to expand their business by adopting this innovative architecture.

Aculab's Prosody family of media processing platforms offers innovative organisations the ability to create high density, scalable and redundant solutions that deliver broad functionality, provide high ROI and in this way fulfil the evolving market's needs.

About Aculab

Aculab provides deployment proven telephony products to the global communications market

Whether you need telephony resources on a board, on a host server processor or from a cloud-based platform, Aculab ensures that you have the choice. We are an innovative, market leading company that places product quality and support right at the top of our agenda. With over 35 years of experience in helping to drive our customers' success, our technology is used to deliver multimodal voice, data and fax solutions for use within IP, PSTN and mobile networks – with performance levels that are second to none.

For more information

To learn more about Aculab Cloud and Aculab's extensive telephony solutions visit:

www.aculab.com

Contact us

Phone
+44 (0) 1908 273800 (UK)
+1 (781) 352 3550 (USA)

Email
info@aculab.com
sales@aculab.com
support@aculab.com

Social
@aculab
aculab