On The Reliability of Voice Over IP (VoIP) Telephony

Sumalya Pal, Raviteja Gadde and Haniph A. Latchman
University Of Florida
Department of Electrical and Computer engineering

Abstract

Voice over Internet Protocol (VoIP) for telephony is becoming important and widespread with the use of the Internet for multimedia traffic and particularly voice communication. The Public Switched Telephone Network (PSTN) system is widely recognized as having set a very high bar in expected telephony performance - an overall availability of 0.99999 ("five nines"). For VoIP telephony to replace traditional systems they should be able to provide a comparable level of availability to the users as on a PSTN system - a formidable challenge for VoIP systems. This paper reviews the "five nines" availability for PSTN systems and explores how emerging VoIP systems may achieve this level of performance. We propose a Kamailio and Freeswich combination (an open-source SIP Router and VoIP application server, respectively), operating in a Linux environment in conjunction with a load balancer and ENUM to provide high availability VoIP service.

Key Words: Alpine Linux, Availability, reliability, redundancy, Five-nines, Kamailio, FreeSwitch, LVS, Ultra Monkey, Virtual Servers. VoIP.

1. INTRODUCTION

IN THE world of traditional telephony, the providers of the particular telephone services often quote a service level of "five nines" or an availability of 0.99999 as a standard signifying quality and assurance. This is sometimes used as a bragging point about their individual products and service offerings. It also means that any new competitors in the same field have to achieve this goal in order to compete effectively. Hence emerging Voice over Internet Protocol (VoIP) telephony – both for long distance or toll calls, as well as for local inter- or intra-office calls (an alternative for traditional telephone switches or Private Branch exchanges (PBX's)) – faces the "five nines" hurdle.

Reliability, a concept that generally deals with the continuous operation of a service, depends on the hardware and software elements of the system, while availability, is a measure of fraction of time that the service is usable. Reliability can be defined as the calculated value, which represents how often the system fails as compared to the percentage of time the system, is available. Availability on the other hand depends largely on the *probability* of the failure of a hardware component. It's calculated by counting the number of components in a system and then calculating the overall *mean time between failures (MTBF)*. It is represented by this formula [1]:

Availability = MTBF / (MTBF+MTTR)

Where MTTR = mean time to repair. Typically MTBF increases as the number of components in the system increases. The most important factor is the MTTR [1], which literally decides the availability of a certain IP telephony connection. For example let's take MTBF = 400,000 Hours and MTTR = 1 Hour. Then the Availability will be approximately 99.9997%, which is excellent from any industrial standard. But let's take MTTR to be 96 hours then the Availability will be 99.7% approximately. .

Traditional telephony achieved the impressive "five-nines" level of availability mainly from hardware redundancy and reliability of components. Several organizations have been working towards making the availability of IP telephony approach that of the PSTN. Some of the most prominent IP telephony companies working towards this goal are Cisco, ShoreTel, Nokia and AT&T.

In Section 2 we explore in greater detail the implications of the nines' criteria, while in Section 3 we will discuss some reasons why VoIP still lags behind the PSTN when it comes to availability of the communication system. Section 4 presents several recent positive developments in IP telephony systems from some of the companies working in this field and in Section 5 we examine the VoIP frameworks that are being created for providing high availability. Section 6 identifies some of deficiencies in present approaches to VoIP systems and in Section 7 we propose a framework for highly availably VoIP systems based on open-source and an Internet inspired approach. Section 9 presents our conclusion and anticipates further work towards high availability open-source based VoIP systems.

2. THE NINES' CRITERIA

There are different figures of 'nines' that are used in demarcating the availability of a device. The one nine represents a 9.0 % availability, which translates on an annual basis to 332 days of downtime in a year, which means that the system will function properly on average for only one month in a year. The two-nine reliability or 99%, availability implies 3 days and 15 hours seconds of downtime in a year. The three-nines or 99.9% availability, signifies downtime of eight hours and forty-six minutes in a year. The four-nines criteria or 99.99% availability means that there will be a downtime of fifty two minutes and thirty six seconds in a year. Now in today's age of telecommunications the five-nines criteria has become the holy grail of telecommunications industries.

Five nine criteria is generally used to represent a target availability figure. Some people even go to the extent of calling it a catchall since with high availability we get high reliability and vice versa. Five-nine availability (or 99.999%) means a downtime of five minutes and fifteen seconds or less per year. The PSTN service is now shooting for six nines availability, which means an availability of 99.9999%. This means a downtime of thirty-two seconds or even less, per year.

3. REASONS WHY VOIP STILL LAGS BEHIND PSTN

Consumers expect same level of reliability as they expect from traditional PSTN service in VoIP. PSTN services have matured into a highly reliable system providing 99.999% of availability. It has been noted that the PSTN system can be characterized as an upper bound of dependability that a distributed computing system can achieve.

The major factors that affect the VoIP system are network outages and SIP server outages. The service availability of the system drops to 98% when network outages interrupt calls [12]. In the following table we can see the service availability of VoIP systems in various networks.

Network/path type	Call success probability
All	99.53%
Internet2	99.52%
Internet2+	99.56%
Commercial	99.51%
Domestic (US)	99.45%
International	99.58%
Domestic Commercial	99.39%
International Commercial	99.59%

Table 1 CALL SUCCESS PROBABILITY ON FIRST CALL ATTEMPT WITH RESPECT TO NETWORK/PATH TYPE [12]

Also Packet Loss causes considerable quality degradations for the users, which can be equated to a dropped call. Even when the system has packet loss below 0%, it was observed that service availability that can be achieved is a maximum of 97.7%. [12]

Network Outages are not a fleeting occurrence but a common place in present Internet. The advantage the PSTN is that once a call is established it is ensured quality but on other hand Internet provides a best effort service.

There is also a overall call abortion probability of 1.5% giving 98% service availability which is falls far short of achieving 99.999% availability [12]. Given all the above reasons we can see that achieving five-nines in a normal Internet conditions is difficult.

Many VoIP systems often feature a central Session Initiation Server (SIP) or other type of VoIP Server. Server outages may occur because a software problem or a failure in the server hardware. To achieve Five Nines availability the system needs to have MTBF of 2,400,000 hours and MTTR of 24 hours. Even if MTTR is reduced to 4 hours, we need to have a MTBF of 400,000 hours. The state of the art system provides a MTBF of 100,000 hours and MTTR of 24 hours [1].

This means we cannot design a high availability system with a single component. Highly available systems are designed by adding a redundant similar component as a stand-by. When the primary fails, the system falls back to secondary. A similar approach is needed for high availability VoIP.

4. IP TELEPHONY RELIABILITY FRAMEWORKS

Accessibility, continuity and fulfillment are the main factors in IP telephony. Accessibility is the ability to initiate a voice call when desired; continuity is the ability to finish the call successfully without jitter upon successful access; and fulfillment is the desired call quality by the customer upon successfully establishing the call [3].

The basic function of IP network is to carry data and traffic without experiencing disruptions due to increased delay and packet loss. IP backbone routers inherently lack carrier grade reliability [3]. In order to overcome this weakness in IP telephony networks, a complete redundant connection between backbone routers, which are grouped in pairs, needs to be established.

A network architecture having a fully redundant backbone guarantees network stability and minimum delay for the rerouted traffic. This is because in the event when one backbone router fails, the other backbone routers, which are grouped together, will be able to reroute any traffic. However such network architecture involves high expenditure, which is a tradeoff that the architecture imposes for high network availability. In essence the high availability is to a large extent dependent on the design of the network infrastructure.

The architectures that we use for network applications follows IP distribution routing protocols like that of OSPF (open shortest path first). Under such conditions the rerouting delay is pretty high, and so these are not ideal for robust Internet telephony applications. Cognitive networking is becoming a necessity and newer ways to reroute traffic are already under research and development. One such protocol is MPLS (multiprotocol label switching). In this protocol architecture, rerouting paths are assessed in advance and whenever there is a failure in network, immediately this alternate path is invoked [6].

However further investigation is needed to verify that such protocols can be used in a more cost effective way. This is because it might not be possible to implement such fast routing protocols in every class of networks due to cost or may be due to security considerations.

5. AN EVALUATION OF THE APPROACHES TOWARDS “FIVE-NINE” PERFORMANCE FOR VOIP

There have been several attempts by industries and academic institutions to find a robust and cost effective mechanism for making VoIP five-nines available. Some of the approaches include server redundancy (1+1) mechanism by industries, software based approaches using SIP protocol and different software tools like Ultra Monkey [25], Piranha and other similar technologies based on Unix based systems or even using peer-to-peer (P2P) as a mechanism for effective VoIP usage.

In the following discussion we will evaluate these approaches as to how effective they are in the search for a five-nine-availability VoIP solution.

Session Initiation Protocol (SIP) based approach

a. The SIP Architecture

The Session Initiation Protocol (SIP) is a more flexible and simpler solution compared to H.323 architecture [23], for handling multimedia sessions of VoIP. SIP is defined by IETF signaling protocol as a text based protocol (that uses UTF-8 encoding) which incorporates several elements of the Hyper Text Transfer Protocol (HTTP) and Simple Mail Transfer Protocol (SMTP) and is widely used for controlling multimedia communications such as voice and video calls over Internet Protocol (IP). SIP is an application layer protocol and is independent of the underlying transport layer. SIP can run efficiently on Transmission Layer Protocol (TCP), User Datagram Protocol (UDP) and also on Stream Control Transmission Protocol (SCTP). It uses port 5060 for UDP as well as for TCP [24]. SIP acts as an enabling protocol for VoIP telephony services.

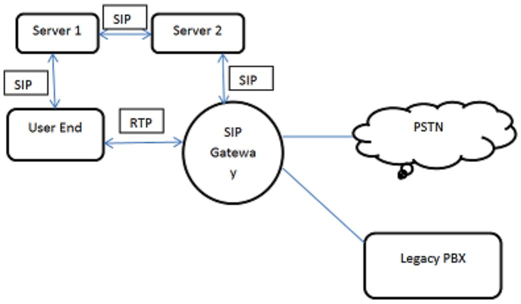

Fig -2 Basic SIP architecture for VoIP

As an alternative to classical PSTN, SIP based telephony services are being considered since it offers a number of advantages over the former.

There are many impressive features of SIP that made it a possible choice for VoIP. Some of these features are:

1. SIP makes sure that the call made by the user reaches its destination, without the restriction of the respective party’s location.
2. This takes into account that not all party (in a conference call) can support all the features, for example video. Hence negotiating feature is highly regarded in SIP.
3. Call cancellation and call hold is completely supported by SIP. A user can add more users to the call or terminate any user. Call transferring is also supported in SIP.
4. Flexibility in call features is one of the major advantages of SIP. The user can start using another program while maintaining the call. For example a user can start using video while on phone.

5. Less advanced devices can be incorporated to make a call since SIP allows usage of several different codecs, which enable negotiation of the media in a call.

b. SIP availability

Wenyu Jiang and Henry Schulzrinne of Columbia University [12] conducted an experiment in which they set up a group of test clients in different networks and simulated automatically random calls between these clients. The overall service availability as calculated by them was 0.98.

They stated that in order to calculate the correct measure of availability in VoIP system, we must consider the call abortion probability of the users. They concluded that although SIP does not really provide five-nines availability like the PSTN, it does come very close to mobile networks' availability, which is 0.97 to 0.99.

N+1 Redundancy approach

This method is followed by several networking/hardware industries in order to make their products five-nines available. Ciscos, Nokia, ShoreTel are some of the companies who have invested considerable amounts in this technique.

Cisco has made considerable advancement in achieving five-nines availability in their IP telephony systems. Taking a leaf from the definition of legacy PBX connections, which over all neglects the influence of non-redundant components of an IP telephony system, Cisco claims to have theoretically proven that their IP telephony solution meets the five-nines standard. They tested their IP telephony solution for hardware reliability, software reliability, link/carrier reliability, power/environment reliability and network design reliability. For this they largely included the parts count method by Telcordia, which is used, for MTBF calculations [2]. Hardware reliability was assessed using the CI infrastructure model with redundant Catalyst 6509 chassis, power supplies and supervisor modules for Cisco Call Manager access switches, which were all redundant in nature [2]. Based on the estimated software forced reloads the theoretical software reliability for the IP telephony product was measured [2].

Cisco typically enforced N+1 redundancy technique to their IP telephony environment.

The overall availability has been calculated using the equation:

$$\text{System Availability} = \pi^{n}_{i=1} = \text{availability (i)}$$

Where i = number of components from 1 to n. This number is then multiplied by the availability of each component to give the overall system availability [2][4].

Now in case there is another failure while the repair for the first failure is still under way then calculation of availability of a redundant parallel system has to be conducted. Cisco performed this by using this equation:

$$\text{Parallel availability} = 1 – [\ \pi^{n}_{i=1} (1 – \text{component availability(i))}]\ [2]$$

In essence Cisco, Nokia and Shore-Tell calculated the parallel availability/reliability of each component in a redundant system thus ignoring the non-redundant parts of the system and then calculated the overall system reliability by adding up the parallel reliability of the redundant system.

This process of redundancy is very effective in five-nines availability assessment of any system, but on the flip side, it increases the cost of manufacturing the system.

6. PROBLEMS WITH CURRENTLY AVAILABLE APPROACHES IN ACHIEVING FIVE-NINES AVAILABILITY

The redundant mechanism followed by most industries in order to raise the service availability of traditional VoIP by making use of some redundant service hosts coordinated by central load balancer, in which each component is coupled with a stand by partner is not a very cost effective mechanism.

Another approach is to make use of a layer-4 switch in addition to the available switching device between SIP clients and a set of SIP servers to dispatch service load without any modifications to the SIP service components. An example of such an approach is the development of a hardware-based mechanism like BIG-IP [18]. Such an approach might be able to improve the availability of a system but will cost a lot in implementation.

Software based approach developed on a Linux Virtual Server (LVS) [19], can achieve the required availability, but may cause delay as the SIP messages traverse through intermediate nodes.

There exists a Peer-to-Peer (P2P) approach [20] [21] towards making VoIP more accessible and available. In such a mechanism, there exists no centralized manager and the peers themselves act as servers as well as clients. However there are numerous security concerns with such an approach. Any peer with malicious intent can alter the SIP messages by other peers. Then again there is a problem of anonymity, as the peers might not want to share SIP messages, which can be read by other peers. Such a mechanism also suffers from SIP message transversal delays, which might cause additional packet loss and poor connectivity.

7. KAMAILIO – FREESWITCH BASED VOIP FRAMEWORK

Here we propose a mechanism for attaining five-nines availability of VoIP telephony using RTP and SIP. The basic setup consists of Kamailio or SIP Router [26], which is an open source SIP server, and FreeSwitch [27], which is an open-source telephony platform, designed to facilitate the creation of multimedia-messaging (voice video and text) driven products.

The basic setup consists of two or more Kamailio virtual servers and one or more FreeSwitch virtual server running on Linux platform. Each of the Kamailio virtual servers has a pre-determined number of supportable clients based on the CPU and memory limits. The X-Lite [28] softphone or any number of SIP Compliant IP telephones are used as the primary IP Telephony devices. Call are routed from the clients to the Kamailio servers using a distributed DNS-based ENUM (E.164 Number to URI Mapping) system with priority settings. If at any moment a particular Kamailio virtual server is down or running at its full capacity then it will reroute the calls to other Kamailio virtual server 2 or to the FreeSwitch virtual server(s) which will primarily serve the purpose of voice mail and music on hold, an automated attendant.

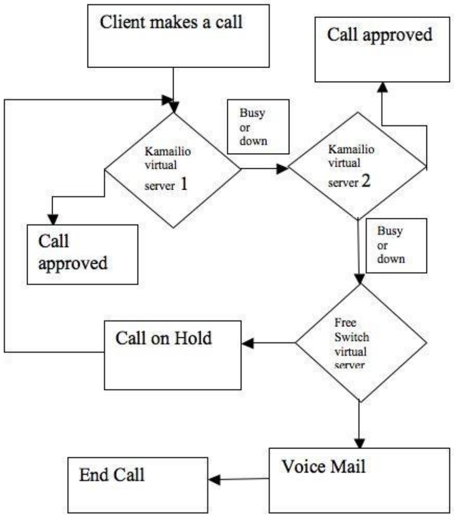

Fig 4: Flowchart of our proposed setup.

To give hardware redundancy, it is proposed to use Ultra Monkey as a load balancer on the Linux Virtual Server (LVS), which is essential for creating highly available network services.

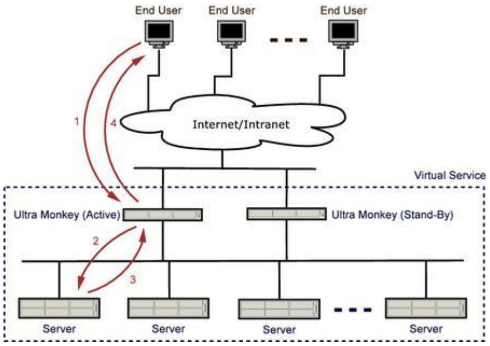

Fig 5 – Ultra Monkey setup.

Ultra Monkey is an open source project to provide flexible high availability frameworks. This can be used for developing both single server based high availability systems as well as multiple servers based high availability systems. It uses the Heartbeat protocol [29] to monitor the servers.

Heartbeat is a protocol that monitors messages sent at regular interval between two servers and at any point of time if messages are not received from one server, it is assumed that the server in question has failed and some form of evasive mechanism is followed in order to rectify this. Heartbeat protocol can send heartbeat messages over both serial links and Ethernet interfaces. It uses an IPfail plugin [30] that helps in determining which nodes should be active.

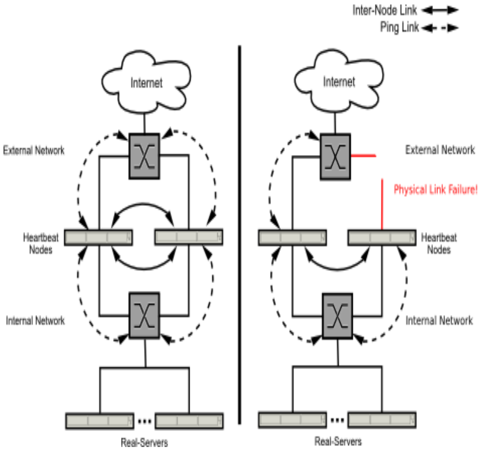

Fig 6 – Ultramonkey with ipfail plugin [30]

After configuring Heartbeat, we designate a master node which when Heartbeat starts up, designates an interface for a virtual IP address which may be accessed by external end users. Under any situation if this node fails then in order to ensure that this machine receives all traffic bound for this address, another node in the heartbeat cluster will start up an interface for this IP address. This process is carried out using Gratuitous ARP [31]. This process is called IP failover takeover.

Ultra monkey makes use of Heartbeat to manage IP addresses on the host on which Linux Virtual Server runs. This also monitors the ultimate destination of a connection made to a virtual service using IPaddr2 resource.

FreeSwitch is used as a voicemail, music-on-hold, conferencing and automated attendant server. From Figure 6 we can see how the whole setup works. The Kamailio servers if unable to handle the calls at any time will redirect the calls to the FreeSwitch server, which will either offer an appropriate service.

8. CONCLUSION AND IMPLICATIONS

The main hurdle in providing five-nines-availability over VoIP networks is that a server failure at any one point, shuts down the system for a certain amount of time in order to recover from the failure. This typically decreases the MTBF. If we can somehow can take the load off the failed server and divide it to several working servers, the MTBF will certainly increase; thus increasing the availability of the VoIP network.

Ultramonkey and ENUM provides load balancing to the servers using heartbeat and priority routing. Hence at any time when there is a load on one server (in our case kamailio servers), it will aptly reallocate the load to the server, which is comparatively free. Heartbeat continuously monitors the messages sent between the two servers, and at any point when it finds that there has been a communication break down, it takes evasive action to rectify the failure.

A working model of the Kamailio-FreeSwich Utltra Monkey and ENUM-based VoIP system has been built and work is continuing on formal theoretical study of the overall availability as well as on generating empirical results as the system is scaled in terms of the number of supportable clients as well as in geographical and server count scope.

9. Reference

[1] Ensuring Reliability in IP Telephony – Shore Tel-IP Telephony from A-Z e-book

[2] IP Telephony: The Five Nine Story – Cisco systems white paper

[3] VOIP Reliability: A Service provider's perspective - Carolyn R. Johnson, Yakov Kogan. Yonatan Levy, Farhad Saheban and Percy Tarapore, AT&T Labs

[4] High Availability Solutions For SIP Enabled Voice-Over-IP Networks – Cisco Systems whitepaper.

[5] Exploring the challenges to powering the future as telecommunications transitions to IP based networks – Nicholas Osifchin International power strategies, USA

[6] IP Telephony: Reliability You can count on – Shore Tel white paper

[7] Power and Cooling for VOIP and IP telephony Applications – Viswas Purani

[8] Convergence: the business case for IP Telephony – Bob Emmerson.

[9] VOIP and IP Telephony: Planning for convergence in state government – Nascio Whitepaper

[10] Self-Admission Control for IP Telephony using Early Quality Estimation - Olof Hagsand1, Ignacio M´as1, Ian Marsh2, and Gunnar Karlsson1-1 Department of Microelectronics and Information Technology Royal Institute of Technology (KTH) S-16440 Kista, Sweden 2 Swedish Institute of Computer Science Box 1263 SE-164 29 Kista, Sweden

[11] IP Telephony Security: an overview -Cisco Systems.

[12] Assessment of VoIP Service Availability in the Current Internet –Wenyu Jiang, Henning Schulzrinne

[13] Design and Implementation of a Low Cost DNS-based Load Balancing Solution for the SIP-based VoIP Service - Jenq-Shiou Leu, Hui-Ching Hsieh, Yen-Chiu Chen, and Yuan-Po Chi

[14] High-Availability Solutions for SIP Enabled Voice-over-IP Networks –CISCO

[15] Design and Implementation of a High Availability SIP Server Architecture- Diplomarbeit , Nils Ohlmeier

[16] Lessons from the PSTN for Dependable Computing -Patricia Enriquez, Aaron Brown, David Patterson

[17] Network Performance Analysis of Internet Telephony on SIP in ENUM Implementation - Yudha Indah Prihatini, Adi Permadi, Wahyu Novian Condro Murwanto, Rendy Munadi

[18] "BIG-IP" - http://www.f5.com/products/big-ip/

[19] "IP Virtual Server" - http://www.linuxvirtualserver.org/

[20] D. Bryan, B. Lowekamp, C. Jennings, "A P2P Approach to SIPRegistration and Resource Location," draft-bryan-sipping-p2p-02, IETF, March 5, 2006.

[21] D. Bryan, B. Lowekamp, C. Jennings, "SOSIMPLE: A Serverless, Standards-based, P2P SIP Communication System,"International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications (AAA-IDEA), June 2005.

[22] http://www.networkdictionary.com/Telecom/VOIP-Architecture-based-SIP.php

[23] SIP- Session Initiation Protocol - J. Rosenberg, H. Schulzrinne , G. Camarillo, A. R. Johnston ,J. Peterson, R. Sparks , M.Handley, and E. Schooler.

[24] http://www.voip-info.org/wiki/view/SIP

[25] http://www.ultramonkey.org/

[26] http://www.Kamailio.org/w/

[27] http://www.freeswitch.org/

[28] http://www.counterpath.com/x-lite.html&active=4

[39] http://www.linux-ha.org/wiki/Heartbeat

[30] http://www.ultramonkey.org/3/ipfail.html

[31] http://wiki.wireshark.org/Gratuitous_ARP