Distributed Agent Architecture to Measure Availability of IP Telephony Services

Shiva Shankar. J
Cisco Systems

Vishal Kumar Singh
Cisco Systems

Muktevi Srinivas
Cisco Systems

Abstract-- **A system is developed that measures the Availability and Performance of the IP Telephony infrastructure and the underlying services. IP telephony changes the proprietary traditional telephony deployed in enterprises into standards based application that runs on top of IP. There is a requirement to measure the reliability of IP Telephony to assess its ability to run business critical applications that were based on traditional PBX's that came with five 9s availability. Availability of an IP Telephony network is defined as a function of availability of the network, call server, other resources (transcoding resources, bandwidth) and human errors. The performance metrics of the IP Telephony network is measured as a function of the Call quality metrics and the Voice quality metrics. The failures are categorized based on whether they were signaling failures leading to failures during call setup or media failures that are noticed when the actual media is transmitted. Further classification to determine which factor in the availability function was the cause of the failure is attempted. In some cases the sub categorization might not be entirely accurate but can provide an overall direction to the study. This information is used for proactive fault management, capacity planning, network design (redundancy) and fault diagnosis by higher level management systems.**

***Index Terms*—IP Telephony, SIP, Distributed system, Availability**

I. Introduction

Enterprises have adopted IP telephony at a fairly rapid pace. Enterprise IP telephony unlike Internet based Voice over IP is run on a controlled infrastructure with service level agreements and higher reliability expectations. Enterprises run business critical applications such as Conferencing, ERP and Stock broking over the IP Telephony network. Non availability of the IP Telephony system could mean serious consequences. Management applications rely on protocols such as SNMP or ping to retrieve statistics from the various devices and correlate them to provide details of availability of a single platform. IP telephony on the other hand requires the orchestration a large number of components in a resonant fashion to provide the required set of services. The integration of large number of components leads to the possibility of failures at various legs that could affect the availability and performance of the system. In such a system where multiple devices interact there is a larger probability for device level misconfiguration which affects the availability. For example addition of an access list on a router to block certain ports might cause failures in calls. Creating an access-list is a valid operation from a device perspective but when this affects the end users capability to make calls it makes the service unavailable.

The focus of this paper is to build a system that can measure the availability and performance of the Enterprise IP telephony system as perceived by an end user. The system requires no additional instrumentation on the network infrastructure or the call server to do this measurement. The capability exposed by this technique is uniform and is vendor independent. The system was built for a network that used Session Initiation Protocol (SIP) [4] as the signaling protocol however the concepts apply to other protocols such as H.323 or MGCP as well. Media is assumed to transmitted using Real Time Transport Protocol (RTP).

Calls are run from different locations and the statistics of call failures and successes are collected over a period of time and analyzed for purposes of Capacity planning and Network design. This system can be plugged into a fault/alarm management framework that can either notify the user or correct the error. A well known example of error correction is routing the call through a PSTN line if the WAN

Shiva Shankar J. is with the Cisco Systems e-mail: shaj@cisco.com
Vishal Kumar Singh is with Cisco Systems e-mail: vishalkr@cisco.com
Muktevi Srinivas is with Cisco Systems e-mail: muktevi@cisco.com

link between two locations fails.

In the next section we describe the network topology that was used to test the infrastructure. In sections III and IV we describe the various factors that affect availability and a framework that can be used to measure this in a network. We describe the reference implementation of the architecture using SIP in Section V .

II. Experimental Topology

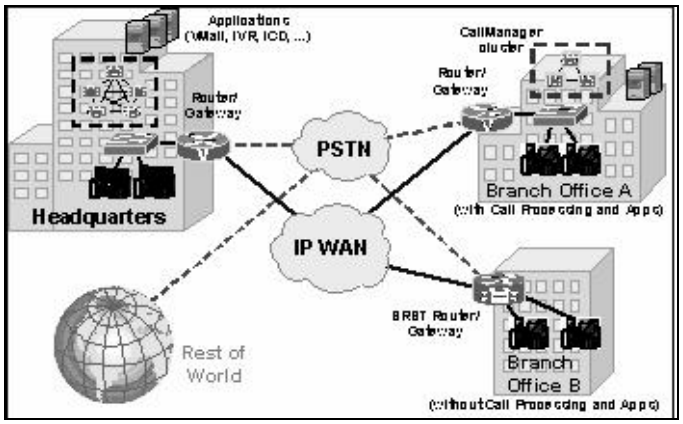

Figure 1: Topology for Availability Measurement

The headquarters consists of the complete call server infrastructure as shown. The Branch office A is a reasonably big branch and has a call server of its own to satisfy the requirements. Branch office B is a small office and so does not have an actual Call server infrastructure and all phones there use the Call server infrastructure in headquarters.

The WAN link between the headquarters and the Branch Office A has a bandwidth of 2Mbps or greater. The link between the headquarters and the Branch office B has a bandwidth of 256 Kbps or greater.

III. PRELIMINARIES

Availability in IP Telephony is represented as a function of successful signaling functions S and successful media transfers M. The basic elements of Availability in a VOIP network are described in [1]. This brings us to one of the important differences in IP telephony over traditional telephony which is signaling and voice can take different paths and is handled by different equipment.

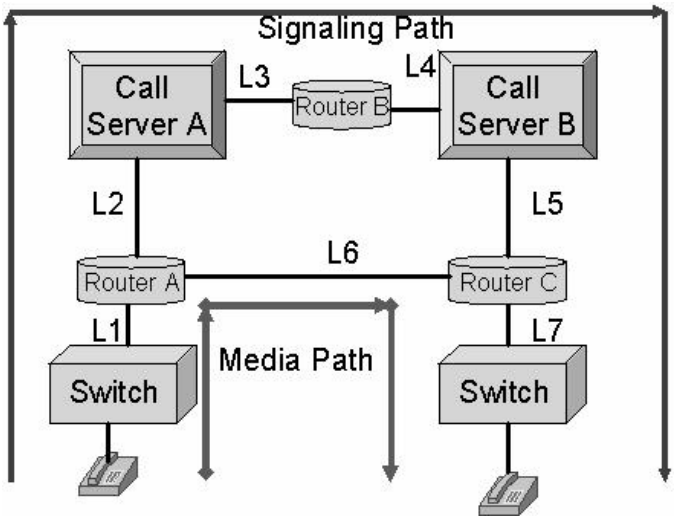

Figure 2: Signalling and Media Path

Availability = $f_1(S) * f_2(M)$

Where $f_1(S) = f(SP_{AV}, CS_{AV}, R_{AV}, S_q)$

SP_{AV} = *Signalling Path Availability*
CS_{AV} = *Call Server Availability*
R_{AV} = *Resources Availability (RSVP etc)*
S_q = *Signalling quality factor*

And $f_2(M) = f(MP_{AV}, MR_{AV}, V_q)$

MP_{AV} = *Media Path Availability*
MR_{AV} = *Media Resources Availability (Transcoder etc)*
V_q = *Voice quality factor*

Signalling Path Availability (SP_{AV}): A call signaling path is considered to be available if the network path to the Call server and any intermediate Call servers that are involved in the call setup are available. A multiplicity of call routing agents can be involved in routing a call. In cases where a Call server is not involved in the call setup this represents the availability of the network between the two end agents to exchange signaling information. Signaling path availability implies that an end to end network path (connectivity) exists between all entities involved in the call setup. In the figure this is a function that is defined by

$SP_{AV} = f(L1_{AV}, L2_{AV}, L3_{AV}, L4_{AV}, L5_{AV}, L7_{AV})$

$L1_{AV}$ = Availability of link L1
$L2_{AV}$ = Availability of link L2
$L3_{AV}$ = Availability of link L3

$L4_{AV}$ = Availability of link L4
etc …

Call Server Availability (CS_{AV}): A call server is a policy based call routing agent that is involved in setting up media sessions between end stations. The availability of call server to route calls depends on current load on call servers as well as policies configured on it. A call server can fail to route a call because of high load on server, hardware or software failures and improper policies configured on it.

Resources Availability (R_{AV}): As a part of the call setup there might be a need for certain resources such as DNS servers, protocol specific entities such as a Location server in SIP. This factor accounts for the availability of these entities at the time of establishing the call. Note that these resources are not used once the media path is setup.

Signaling quality factor (Sq): This factor determines availability based on call signaling parameters and accounts for dynamic network behavior changes. The call signaling parameters like time interval between going off hook and receiving dial tone or dialing a number and receiving ring tone can vary as a result of changes in network conditions. IP telephony service can become unavailable temporarily because of dynamic network factors like link down, route convergence, network outages, congestion etc. The degradations in quality of call set up can result in a user dropping a call even before media session is established. Well defined thresholds are used to determine if the call has failed for these reasons. For e.g. If the time between offhook and dial tone is greater than 5 seconds then it is considered as a failure as in the real world the user would drop the call or retry by going on -hook and offhook again.

Media Path Availability (MP_{AV}): Media path is the network path that the exchanged RTP packets traverse from the source to the destination. The path may range from being a simple Layer 2 path in an enterprise to a complex layer 3 path that could traverse various autonomous systems. One of the important things to note is that the path in the forward and reverse direction need not be symmetric. Also an additional level of complexity is introduced if load balancing across multiple links is used while transmitting RTP. In the figure this represented by

$MP_{AV} = f (L1_{AV}, L6_{AV}, L7_{V})$

Media Resources Availability (MR_{AV}): In some cases additional resources are required so that the media can be understood by both endpoints or to provide additional functionality. An example of this would be if the two endpoints cannot use the same codec. In such a case a transcoder is used to perform the media conversion and unless this is available it will cause the call to fail. Other examples include mixers for conferencing and gateway resources while calling analog endpoints. High load, software or hardware failure can make these resources unavailable.

Voice quality factor (Vq): Poor voice quality can be result from high packet loss, delay, jitter and echo. These factors depend on network conditions at any instance of time and are non-deterministic in nature. If the voice quality degrades to a level that it becomes annoying or unintelligible there is a high likelihood that the user will drop the call. So voice quality is measured at the end of the conversation using the E-Model [8] and PESQ [7] to determine the extent and duration of the degradation. If the call does not have the required quality the call is marked as having failed. Hence, the mere fact of the media path and transcoding resource being up and running does not guarantee IP telephony service availability.

In this section we have defined the various factors that go into calculation of IP telephony availability. We have derived availability to be function of signaling and media transmission availability and quality. We also explained how each of parameters maps to availability of network infrastructure, resources and network dynamics.

IV. FRAMEWORK

Availability is measured by setting up calls between various agents spread across various locations. The Agents send back the results of the call to a server that uses it to make extensive computations over time. It

would be difficult to configure each agent at the location where it is installed. So we have a central server component that allows us to configure the agents remotely. The results that are obtained need to be post processed to produce reports. In fact results from agents involved in a call need to be correlated to produce the aggregate result. So the result post processing is performed in a single server that provides external interfaces to retrieve the post processed data. We now define the framework that is used to measure the availability. The basic elements of the framework are depicted in Figure 3.

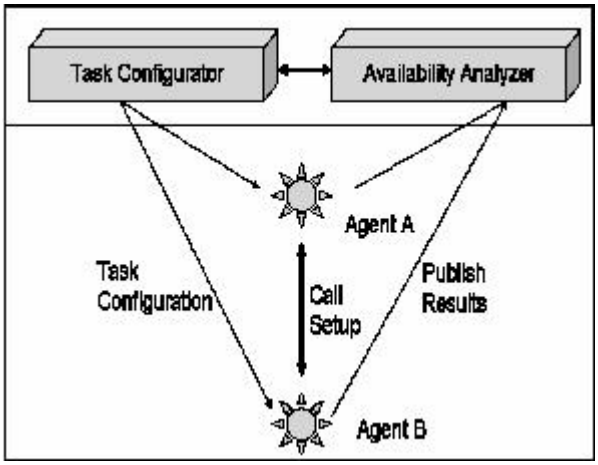

Figure 3: IP Telephony Availability Measurement Framework

The various components in the architecture are

Agent – The Agent executes one or more tasks each of which is essentially a call to another agent on the network. Any configured call involves two agents. One plays the role of the initiator while the other plays the role of the receiver.

Task Configurator- This is a frontend componet that resides on the server. This is used to configure the tasks on the various agents. The Task Configurator may be a User interface or a script that reads a configuration file. The information passed to the user agent includes role definitions and task specific parameters such as URL to use to register etc.

Availability Analyzer- This component is a background service that collects results from the various agents and correlates them. It makes a final analysis of success and failure and computes the availability and performance metrics. It uses a persistent store top store these results and has a web based interface that can be used to access the results. Graphing applications can use the interfaces exposed by the Availability Analyzer to provide detailed statistics and trends.

The following sequence of steps are involved in applying the framework

- ⊕ Task Configurator setup
- ⊕ Setting up agents at distributed geographical locations
- ⊕ Setting up tasks for each of the agents from the Task Configurator
- ⊕ Agents run the task and publish the result back to the Availability analyzer
- ⊕ Post processing and analysis of results by Availability analyzer.

V. Implementation Architecture Overview

The framework described in the previous section was implemented to study the availability of the experimental deployment of an IP telephony Network.

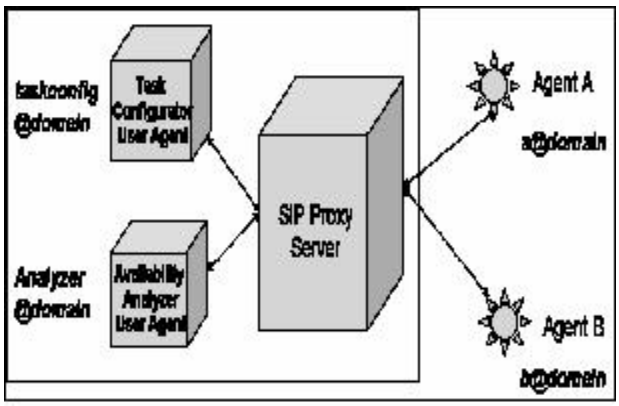

Figure 4: Implementation View

One of the considerations that were used while implementing the framework was to make the agents as lightweight and independent as possible. After an analysis of a number of distributed technologies we chose SIP [4]as the signalling protocol for the application. SIP while primarily a protocol for telephony applications is designed in a manner that is generic enough to find applicability in other areas. Using SIP provided us with a very distinct advantage. We were free to distribute the Task Configurator and Availability analyzer across different machines or collocate them on a single

machine them without having to change the configuration of the agents.

Second, addition of new services was extremely simple as all one needed was a new User Agent that provided the new functionality. Based on this all the entities described in the framework are modeled as User Agents that register with the SIP proxy server. Our implementation uses the Vovida SIP proxy server with a minor modification as the SIP proxy server. The agents, Task Configurator Availability analyzer use modified versions of the Vovida SIP User Agent implementation. The minor modification that was made in the SIP proxy server was to inform the Task Configurator of the list of agents that are registered with the proxy server so that it could display the list in the User Interface for Task configuration.

To understand how the system works depicted below is the sequence diagram of for the various messages and events that are exchanged in the system.

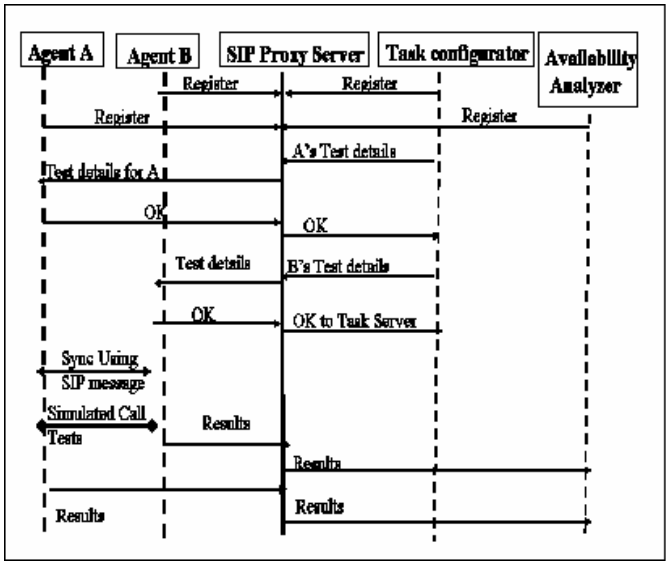

Figure 5: Operational Sequence

The sequence of operations depicted in Figure 5 is as follows

- ⊕ The Agents, Task Configurator and availability analyzer register with the SIP proxy server.
- ⊕ At this point of time if the user invokes the GUI he will see all the registered agents for which the tasks can be configured.
- ⊕ The tests are configured in the Task configurator which sends SIP Messages with XML payload on configuration tests. The agent is sent configuration information from the Task configurator that includes
  - o *Role in a call (Initiator/Receiver)*
  - o *Interval at which a call must be initiated*
  - o *Task Description (Codec, Call Duration)*
  - o *Call Processor that the agent must use to place the call*
  - o *Additional configuration or authentication information to contact the call processor*
  - o *URL of the Availability Analyzer*
- ⊕ The Agent receives the details and creates a test and sends back an acknowledgment to the server.
- ⊕ The Agent runs the test at the defined interval.
- ⊕ The results are sent back to the Availability analyzer using a SIP message. These are sent in a SIP Message [5] in XML format.

VI. Synchronization

One of the issues with the distributed architecture is that if one of the agents went down due to some error or the machine is being rebooted it would affect the availability statistics in a negative manner. Various synchronization mechanisms were introduced to minimize and remove the effect of failures from the calculations.

- o The agents exchange synchronization messages to determine if they are ready to run the test. If any of them indicates that it is not then the test is abandoned with no result. If the remote agent does not respond, a ping test is executed to check for reachability of the machine and the default gateway. If the default gateway is reachable while the machine is not, the

result is abandoned. If the remote gateway is not reachable the result is considered as it is a network outage.

- The Availability analyzer correlates the results from both ends before using it in any computations.

VII. Test Methodology

The test methodology is to provide maximum coverage of all the possible call scenarios and measure the availability as a whole.

- We established calls between the branch office and the main office and also between branch offices that have direct connectivity based on the amount of traffic.
- The tests are run on a production network with normal call volumes as expected on a daily basis. On a test network one would have to run load simulators that have approximately the same call distribution that is expected on a real network.
- Any failures are evaluated for possible reasons and are categorized into the categories described in Section III
- The categorization is done based on a combination of return codes or post task diagnostics that are run on failure. An example would be if the Call Agent server returned a code indicating "No more calls can be accepted" it would be categorized as a Call Agent Server failure. If the RTP stream does not reach the destination it would be a case of a network failure which would be determined by a post task ping diagnostic to the server.
- These values were computed and archived over a period of time such as 24 hours, 72 hours, a week and 30 days and this was used as the basis for determining the availability of the IP telephony system.

VIII. Results

The framework was deployed in a lab network with more than 150 agents and running 110 calls every minute across the network. In the lab implementation the server components which include the Task Configurator, Availability analyzer and the proxy server were run on a single machine. The Agents had a footprint of 600Kb and were running on windows clients across the network across the LAN and WAN.

The results showed a very scalable and fault tolerant system in place. Using SIP based mechanisms for redundancy even failures in the Proxy server caused no disruptions in the measurements. Extending the framework by networking SIP proxies was not experimented as a part of this exercise but prior experience in other areas shows that this can be extremely scalable.

IX. Related Work

Availability and performance measurement on the Internet has been an interesting and well researched topic. Some of the best work originates from the works of *Vern Paxson* [3]. However these measurements apply only in part to the IP Telephony scenario that is described here.

Wenyu Jiang and Henning Schulzrinne [1] have talked of measuring availability from an internet perspective in terms of Call success rate. Our solution includes factors such as availability of transcoding resources and voice quality statistics which are not accounted for in [1] .

There is some work from *Avaya Research* [6] in the QOS and IP Telephony readiness. The solution is focused on initial deployments and does not provide a mechanism to measure the availability of the IP Telephony solution on an ongoing basis. Additionally factors that affect user perception of service availability in the area of Signaling quality are not considered. An example is a user dropping a

call when he does not get dial tone or large post dialing delay. Our solution accounts for these factors and provides a solution that can be used to determine the exact nature of problems and availability in the IP telephony deployments.

X. Future Work

The proposed architecture can be used to determine the exact nature of problems and availability of the elements in an IP Telephony network. Critical business applications that rely on IP Telephony such as Stock broking can evaluate strategies and designs to provide better service. Further work that we are working on is the definition of Service Level Agreements in the case of an IP Telephony network and how this framework can be integrated to provide information about the compliance. Applying the metrics that are used for data services will not work effectively for IP Telephony networks.

In addition we are working to provide comprehensive failure analysis that will not only categorize each failure into the various components shown above but provide trends on how frequently and regularly a certain failure occurs and search for possible patterns in the failure.

XI. Conclusion

In this paper we have described a framework that can be used to measure availability of the IP Telephony network. The model of centralized task configuration and distributed task execution makes it a simple yet scalable and highly performance driven solution that can provide valuable statistics.

We have provided the description of a reference implementation that is based on SIP and have verified its operation.

XII. References

[1] Wenyu Jiang and Henning Schulzrinne. Assessment of VoIP Service Availability in the Current Internet. Passive and Active Measurement Workshop (PAM), La Jolla, California, April 2003.

[2] Wenyu Jiang and Timothy F. Williams. Detecting and Measuring Asymmetric Links in an IP Network. In IEEE Global Communications Conference (Globecom), Rio de Janeiro, Brazil, November 1999.

[3] Vern Paxson Measurements and Analysis of End-to-End Internet Dynamics April 1997

[4] RFC No. 3261 SIP: Session Initiation Protocol

[5] RFC No. 3428 Session Initiation Protocol (SIP) Extension for Instant Messaging

[6] M. Bearden, L. Denby, B. Karacali, J. Meloche, D. T. Stott, Assessing Network Readiness for IP Telephony September 2001

[7] ITU-T Recommendation P.862 Methods for objective and subjective assessment of quality.

[8] ITU –T Recommendation G.107 .The E-Model, a computational model for use in transmission planning