ON THE DEVELOPMENT OF VOICE OVER IP

A Record of Study

by

XU YANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF ENGINEERING

May 2008

Major Subject: Engineering
College of Engineering

ON THE DEVELOPMENT OF VOICE OVER IP

A Record of Study

by

XU YANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF ENGINEERING

Approved by:

Chair of Committee,	Duncan M. (Hank) Walker
Committee Members,	Jennifer L. Welch
	Rabi N. Mahapatra
	Narasimha Reddy
	Rohinton Gonda
Head of Department,	N. K. Anand

May 2008

Major Subject: Engineering
College of Engineering

ABSTRACT

On the Development of Voice over IP. (May 2008)

Xu Yang, B.E., Beijing University of Posts and Telecommunications (China);

M.CS., University of Texas at Arlington

Chair of Advisory Committee: Dr. Duncan M. (Hank) Walker

This record of study documents the experience acquired during my internship at Sonus Networks, Inc. for the Doctor of Engineering Program.

In this record of study, I have surveyed and analyzed the current standardization status of Voice over Internet Protocol (VoIP) security and proposed an Internet draft on secure retargeting and response identity. The draft provides a simple and comprehensive solution to the response identity, call recipient identity and intermediate server retargeting problems in the Session Initiation Protocol (SIP) call setup process.

To support product line development and enable product evolution in the quickly growing VoIP market, I have proposed a generic development framework for SIP application servers. The common and open architecture of the framework supports multiple products development and facilitates integration of new service modules. The systematical reuse of proven software design and implementation enables companies to reduce the development cost and shorten the time-to-market.

As the development and diffusion of VoIP can never be isolated from the social sphere, I have investigated the current status, influence and interaction of three most important factors: standardization, market forces and government regulation on the development and diffusion of VoIP. The worldwide deregulation and market privatization have caused the transition of the standards development model. This transition in turn influences the market diffusion. Other than standardization, market forces including customer needs, the revenue pressure on carriers and vendors, competitive and economic environment, social culture and regulation uncertainties create both threats and opportunities. I have examined market drivers and obstacles in the current VoIP adoption stage, analyzed current VoIP market players and their strategies, and predicted the direction of VoIP business. The regulation creates the macro environment in which VoIP develops and diffuses. I have explored modern telecommunications regulation principles based on which government makes decisions on most current issues, including 911 support, mergers and acquisitions, interconnection obligation and leasing rights, rate structure and universal service fees.

To My Parents

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to my advisor, Duncan M. (Hank) Walker, for his enormous encouragement, advice and help during the program. His devotion to his students is genuine, exemplifying an ideal professor and advisor of higher education. This record of study could not have been completed without him.

I would like to thank my internship supervisors Mr. Rohinton Gonda, for his professional guidance, support and help throughout the program.

I would also like to thank my committee members, Dr. Jennifer L. Welch, Dr. Rabi N. Mahapatra, and Dr. Narasimha Reddy.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

This record of study documents the working and research experience that I acquired during my internship at Sonus Networks, Inc. in partial fulfillment of the requirements for the degree of Doctor of Engineering.

Sonus Networks, Inc. develops, markets and sells a suite of carrier class network infrastructure products, including the GSX9000 Open Services Switch, Insignus Softswitch, Sonus Insight Management System and Open Services Partner Alliance. The company was founded in 1997 and now has a $1.9B market capitalization[1] and more than 700 employees worldwide. Benefiting from the first-mover advantage and the accelerated development of worldwide Voice over Internet Protocol (VoIP) markets, the company has built a strong lead position in the next generation telecommunications equipment market. According to iLocus, in 2006, 40 percent of the national long distance traffic and about 26 percent of the international long distance traffic was routed through Sonus equipment. The internship is at the main company site in Chelmsford, Massachusetts.

During the internship, I worked as a senior software engineer and was responsible for developing features and applications related to the Session Initiation Protocol (SIP) [Rosenberg et al. 2002] on the Sonus Insignus Softswitch. The internship provided me a

This record of study follows the style and format of *ACM Transactions on Computer Systems.*

[1] As of Feb 23, 2007 on Nasdaq.

cooperative environment that facilitated my investigation of broadly based problems. In Sonus, the sophisticated engineering practice concentrates on efficient and effective product delivery that meets the needs of the rapidly evolving market. The availability of the original source code enabled me to examine design alternatives in large software systems consisting of more than three hundred thousand lines of code. Being fascinated with the software architecture and implementation, I further developed them into a generic framework to achieve fast and agile development of SIP application servers. My talented coworkers, who are active in the Internet Engineering Task Force (IETF), also provided their valuable opinions while I was developing the Internet draft on secure retargeting and response identity.

Other than technical advances, I was able to discuss questions with coworkers in the finance, marketing and legal departments. I asked questions such as “What is the impact of the FCC’s decision on E911 service?” and “Given the product life cycle, what is the current stage of VoIP products?” The ideas and data collected in this way, together with materials learned from academic journals and industry reports, formed my study of the current status and influence of standardization, market forces and government regulation on the development and market diffusion of next generation telecommunications technologies.

The Record of Study is organized as follows. Chapter II describes the final internship objectives. Chapter III briefly introduces the internship site, Sonus Networks, Inc., its product line and major applications. Chapter IV describes two of my typical assignments during the internship. In chapter V, after surveyed the best current practice and standardization status of SIP security, I propose an Internet draft on secure retargeting and response identity. In chapter VI, I propose a generic architecture and implementation framework for the SIP stack and SIP application servers. The architecture design, internal design of selected layers, process design and design patterns applied are also explained. In chapter VII, I investigate the development and diffusion of the next generation network from the standardization, market and regulation points of view. Chapter VIII concludes this Record of Study.

CHAPTER II

INTERNSHIP OBJECTIVES

Objective 1: Develop skills in telecommunication protocols and standards development and implementation.

The Session Initiation Protocol (SIP) [Rosenberg et al. 2002] is a baseline signaling protocol in the next generation telecommunications architecture. It provides such services as Internet conferencing, telephony, presence, events notification and instant messaging. SIP is continuously evolving and new SIP-related protocols have been proposed to implement new features and achieve interoperability.

The SIP stack provides parsing, encoding, transport and transaction services to its application clients through a well-defined user interface. In Sonus, I have developed SIP stack components to meet the application needs and kept the SIP stack up to date with respect to evolving standards. In this Record of Study, I have briefly described common SIP operations, surveyed the current standardization work on SIP security and proposed an Internet draft for secure retargeting and response identity. The architecture and implementation of the SIP stack are also described as a part of the generic architecture and implementation framework of SIP application servers mentioned below.

Objective 2: Develop understanding and experience in design and development of large software systems.

The SIP processing engine (SIPE) is a subsystem of the Sonus softswitch. Built on top of the Adaptive Communication Environment (ACE), the SIPE collaborates with other Sonus softswitch components to provide location and proxy services in the SIP network. I have been involved in SIPE development projects including 3xx message handling and configurable local recursion, network asserted identity handling, route handling, and automated testing. In each project, I have independently carried out the design, implementation and unit testing.

I was fascinated by the SIPE design. After talking to senior coworkers, I learned that several telecom startups also use a similar architecture. Because of limited documentation, I have gone through more than 30,000 lines of source code in an effort to grasp the essence of both the detailed implementation and the macro system architecture. The SIPE architecture could serve as a model for developing new products. It well follows object-oriented design concepts (e.g. encapsulation, inheritance and polymorphism), applies design patterns and fully uses the ACE framework to enable fast and agile development. In my Record of Study, I have further developed this model and described a generic architecture and implementation framework for the SIP stack and upper applications. The framework does not reflect the actual architecture design of the Sonus softswitch. Even so, as the design and implementation of Sonus products are strictly confidential, the implementation details have to be omitted from this Record.

Objective 3: Develop a deep understanding of the influence of social, political, institutional and other non-technical factors on the development and diffusion of new telecommunications technologies.

As part of my internship, I have investigated the influence of market forces and government regulation on the development and market diffusion of next generation telecommunications products. In the past four years, VoIP companies have solved most technical problems. VoIP is moving into the mainstream to compete with traditional circuit-switched telecom systems [Doherty 2005]. The development and diffusion of technologies, however, can never be isolated from the social sphere. The law of suppression of radical potential (i.e. existing market leaders attempt to suppress revolutionary technology) and the supervening social necessities (i.e. customer needs) influence the development and diffusion agenda of next generation telecommunications products. The technical development, market needs and government regulations interact with one another. In this Record of Study, I have explored the principle and trend of government regulation and incentives of government decisions on most current issues, including 911 support, mergers and acquisitions, interconnection obligation and leasing rights, rate structure and universal service fees, and spectrum and wireless service regulation. Along with the liberalization of telecommunications markets and the development of new telecommunications technologies are the rapid development of VoIP markets. Market forces including customer needs, the revenue pressure on carriers and vendors, competitive and economic environment, and social culture create uncertainty, threats and opportunities, influencing the development and diffusion of next generation technologies. With the help of a number of market frameworks, including the product life cycle model, product features model and market segmentation model, I have explained the evolution of VoIP market, examined market drivers and obstacles in the current VoIP adoption stage, scanned the current VoIP marketplace, and when possible, predicted the direction and trend of VoIP business. As an inseparable part of the VoIP discussion, I have also briefly analyzed the 3G value chain and the worldwide 3G market development.

CHAPTER III

INTERNSHIP SITE AND PRODUCTS OVERVIEW

This chapter is organized as follows: Section 1 describes the internship site. Section 2 describes the functions of selected Sonus network components. Section 3 describes applications of Sonus network components[2].

1. Internship Site Overview

Sonus Networks, Inc. develops, markets and sells a suite of carrier class[3] network infrastructure products including the GSX9000 Open Services Switch, Insignus Softswitch[4], Sonus Insight Management System and Open Services Partner Alliance. Sonus develops these products on the advanced Open Services Architecture™ (OSA) framework, which enables fast, scalable and efficient development of a full range of carrier applications and enhanced services. Sonus customers include communications service providers, such as long distance carriers, wireless service providers, Internet service providers (ISPs), and international telephone companies [Reuters Corp. 2004]. In

[2] To illustrate the detailed role of each Sonus network component, in appendix A, the call setup process in a traditional circuit switched network and the call setup process in a packet switched network using Sonus network components are compared and discussed.

[3] Carrier class systems ensure extremely high availability (e.g., required to be operational at least 99.999 percent of the time, known as five nines reliability), high capacity, short call set up time and high speech quality without noticeable delay and noise. The tough requirements have translated to fully redundant, self-healing, highly scalable and manageable systems.

[4] In the late 1990s, the softswitch concept was proposed to decompose the traditional gateways into efficient hardware-based media gateways and a few centralized software intensive media gateway controllers, or *softswitches* that perform the signaling and call control functions for the media gateways.

2003, Sonus reported total revenue of $93 million. Revenue rose to $170 million in 2004, $190 million in 2005, $270 million in 2006 and $320 million in 2007.

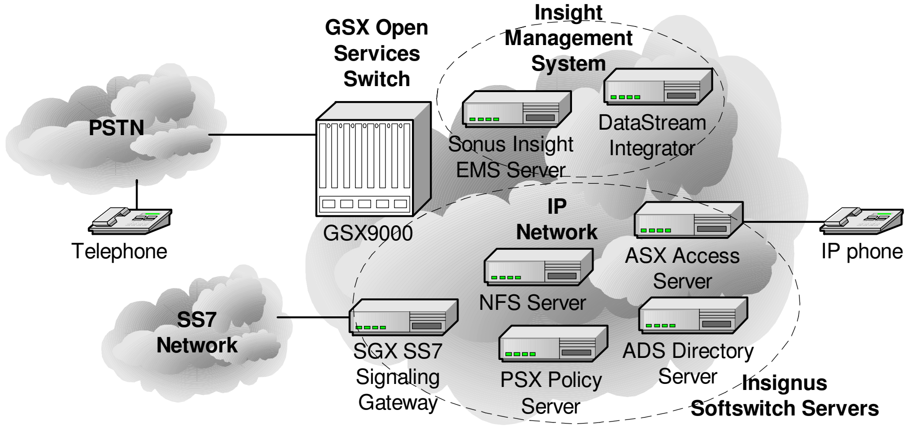

Fig. 1. Sonus VoIP network components.

2. *Products Overview*

Fig. 1 shows the Sonus VoIP network components. The components communicate with one another using IP based protocols over a 10/100 Mbps Ethernet management network.

The GSX provides carrier-class media gateway functions [Arango et al. 1999] between traditional circuit-switched networks and IP packet-switched networks, offering voice coding exchange between the two. Under the control of the PSX Policy Server, the GSX performs limited Public Switched Telephone Network (PSTN) user interactions such as announcements, tones, and digit collection.

On the circuit network side, the GSX handles the upper level of the Common Channel Signaling System No. 7 (SS7), e.g., Integrated Services Digital Network (ISDN) User Part (ISUP) and Transaction Capabilities Application Part (TCAP) [Dryburgh and Hewett 2005], and uses the SGX SS7 Signaling Gateway for the lower level of SS7 signaling, e.g., SS7 Message Transfer Part (MTP) [Dryburgh and Hewett 2005]. On the IP network side, the GSX typically uses proprietary signaling between Sonus components and uses the Session Initiation Protocol (SIP) to connect to third party media gateways.

The GSX converts call traffic over the Time Division Multiplexed (TDM) circuit into packet voice using various codes. The selected codec takes samples of the voice received from the TDM and builds packets as described in the International Telecommunication Union (ITU) standard G.711. Other supported codec standards include G.723.1, G.723.1A, G.726, G.729A, G.729A+B and T.38 [Protocols.com 2007]. These standards typically include compression techniques for reducing packet sizes and silence suppression for eliminating the unnecessary transmission of packets when call participants are not speaking. The GSX Circuit Network Server (CNS) module with onboard Digital Signal Processing (DSP) serves as a data and protocol-processing engine.

The Network File System (NFS) server serves as a storage device for the GSX, as the GSX itself does not possess a hard drive or flash memory. When a GSX boots up, it locates a NFS server, mounts to it, and starts downloading operational software and configuration information. For example, the announcement files (with file extension *.wav*) are downloaded from the NFS server to each GSX CNS. Since the CNS card cannot store all announcements, it updates *.wav* files on the Least Recently Used (LRU) basis. The GSX plays a *.wav* file when it receives instruction from the PSX Policy Server. Furthermore, the NFS server also records GSX event logs for system troubleshooting and accounting.

The Insignus softswitch is a software product that provides routing, signaling, and application services. The products include PSX Policy Server, ASX Access Server, ADS Access Directory Server and SGX SS7 Gateway. The SGX provides an interface for the GSX and PSX to communicate to the SS7 network. It implements the MTP of SS7 to transmit ISUP or TACP [Dryburgh and Hewett 2005] messages between GSX or PSX and SS7 network elements. The ASX Access Server provides line-side (Class 5) signaling to phones connected to a packet network. The phones include IP telephones and standard phones connected to integrated access devices (IADs).

The PSX Policy Server performs all call routing decisions and determines call treatment such as screening and blocking. The PSX contains a database of signaling addresses for routing calls. After received signaling information from GSX, SIP application server, or ASX, the PSX instructs the requesting system on how to establish calls. The PSX also interacts with PSTN databases via TCAP, and application servers via SIP, which enable a range of enhanced services.

The interaction of the PSX and other network components is illustrated in Fig. 2.

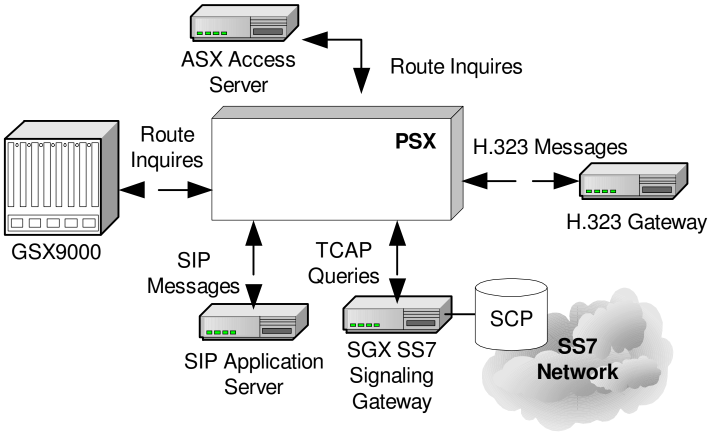

Fig. 2. The interaction of the PSX and other network components.

The Insight Element Management System (EMS) implements operation, administration, maintenance, and configuration functions. The EMS system includes an element management server and DataStream Integrator (DSI). The element management server provides a Web-based GUI for centralized provisioning of Sonus components, managing component operations, and monitoring component performances and faults. The DSI collects Sonus proprietary call detail records (CDRs) from the GSX and ASX Access Server, and converts them into industry-standard billing records.

3. *Overview of Applications with Sonus VoIP Components*

The Sonus network components collaborate to provide access, packet switching, network border switching and enhancement services. The representative applications include:

- Long Distance: Instead of routing a call to a long distance carrier to go through multiple circuit switches (expensive), a GSX can convert the call to voice packets and send the call long distance over a packet network to another GSX. The second GSX can then switch the call back onto the PSTN to reach its final destination. An example deployment of the long distance application is shown in Fig. 3.

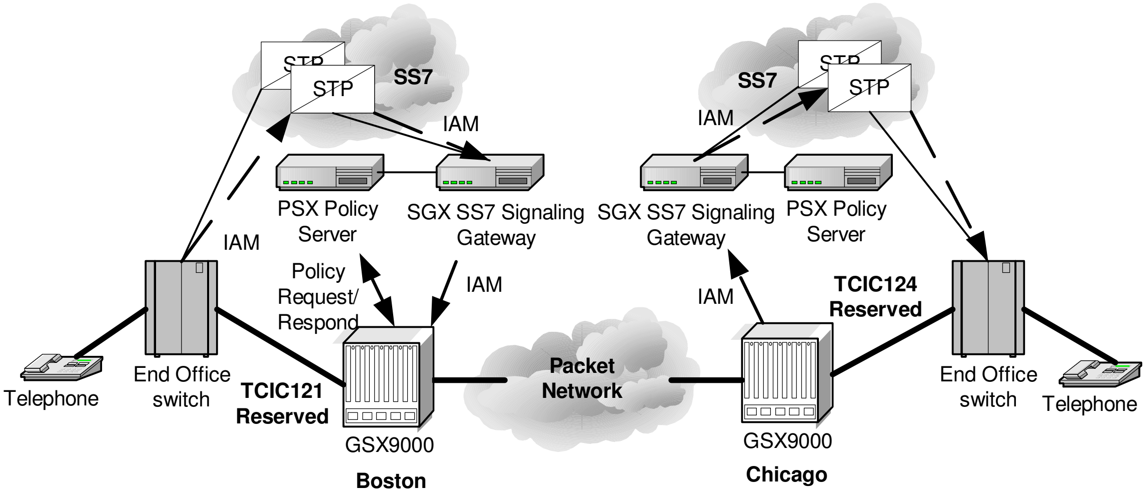

Fig. 3. A long distance application.

- Tandem Switching: Similar to long distance, the GSX can be used to replace Class 4[5] tandem switches. Instead of switching calls over fully connected TDM inter-machine trunks, multiple GSXs exchange calls over a packet network backbone.

[5] Class 5 switches are used to terminate local calls. Tandem or Class 4 switches are intermediate switches that connect other Class 4 switches or Class 5 switches. Class 4 switches are only used for long distance communications in the PSTN.

- IP Voice Termination: The combination of the ASX and GSX can be used to route calls originating from IP telephones or IADs to various destinations accessible either on packet or traditional circuit networks.
- Network Border Switching: As the adoption of packet voice technologies continues to increase, carriers need to interconnect with each other using IP connections rather than circuits. The PSX and GSX collaborate to support packet peering, in which real-time communication traffic is passed from one packet network to another network that belongs to a separate administrative or security domain, providing appropriate security and traffic controls.

CHAPTER IV

SELECTED INTERNSHIP ASSIGNMENTS

1. The Design and Implementation of the SIP Extensions for Network Asserted Identity

1.1. RFC Requirements

Network Asserted Identity is an identity that is initially derived by a SIP server as the result of an authentication process, e.g., SIP Digest Authentication. RFC 3324 [Watson 2002] describes the short-term requirements for the exchange of Network Asserted Identities within a Trust Domain. Several key terms defined in RFC 3324 are as follows:

- *Identity*: An identity is a sip:, sips: or tel: URI and optionally a Display name. The identity must be meaningful in that if used as a Request-URI in a request, it could cause the request to be routed to the user/line that is associated with the identity.
- *Trust Domain*: A Trust Domain is a set of SIP nodes in compliance with a certain set of specifications, Spect(T). It can be a set of devices of a single network operator or multiple Trust Domains joined together by bi-lateral agreements.
- *Trust*: A node α trusts node β if and only if (1) a secure connection exists between α and β, *AND* (2) β has configuration information indicating α is a member of the trust domain.

Network Asserted Identities can be transparently transported within a Trust Domain. A node α can also securely send a Network Asserted Identity to a node γ outside a trust domain, provided it conforms to the privacy requirement of the identified message originator and the specification of the trust domain. If γ trusts α, then the Network Asserted Identity may be considered as valid and used in γ. Otherwise, there is no guarantee for the Network Asserted Identity to carry a user's true identity.

Based on the requirements of RFC 3324, RFC 3325 [Jennings et al. 2002] describes SIP private extensions that enable a network of trusted SIP servers to assert the identity of authenticated users, and to convey indications of end-user requested privacy in a Trust Domain. Nodes in such a Trust Domain are explicitly trusted by its users and end-systems to assert their identities carried in the SIP extension headers, and to be responsible for withholding that identity outside of the Trust Domain when privacy is requested.

RFC 3325 proposes two SIP extension headers, the P-Asserted-Identity header and P-Preferred-Identity header, and a new privacy type *id* to the Privacy header defined in RFC 3323 [Peterson 2002]. For example:

```
P-Asserted-Identity: "Alice" <sip:alice@tamu.edu>
P-Preferred-Identity: "9798251234" <tel:9798251234@tamu.edu>
Privacy: id
```

When a proxy receives a message from a node that it does not trust, it first removes the P-Asserted-Identity if it is present in the message, and then may add a P-Asserted-Identity header from the authentication results after authenticating the message originator. When a proxy receives a message from a node it trusts, it can use the P-Asserted-Identity header as if it had authenticated the caller itself. At any time, a P-Asserted-Identity header can contain at most one SIP or SIPS URI.

When a proxy forwards a message to a node it does not trust, it first examines the Privacy header filed to determine whether to remove the P-Asserted-Identity. If the Privacy header field value is set to "id", the proxy must remove all the P-Asserted-Identity header fields before forwarding the message. If the Privacy header field value is set to "none", then the proxy must not remove the P-Asserted-Identity header fields. If the Privacy header is not present, then the action taken depends on the specification defined in Spec(T). A proxy can transparently forward a P-Asserted-Identity header to a node it trusts.

A P-Preferred-Identity serves as a hint suggesting which of the multiple valid identities for the authenticated user should be asserted during the proxy authentication process. If such a hint does not correspond to any valid identity known to the proxy for that user, the proxy can add a P-Asserted-Identity header of its own construction, or it can reject the request (for example, with a 403 Forbidden). The proxy must remove the user-provided P-Preferred-Identity header from any message it forwards.

Other than the RFC requirements, the Japan Telecommunication Technology Committee (TTC) specified concrete behaviors for transferring network asserted identity information between a carrier SIP network (a SIP trust domain) and other trusted networks, e.g., ISUP or MGCP network. The behavior includes:

- Determining contents: when a signal message arrives at a SIP trust domain from other trusted networks, a SIP boundary server checks the contents of the message, decides the network asserted identity information, e.g., TEL URI, TEL DISPLAYNAME, SIP URI, and SIP DISPLAYNAME in accordance with certain rules, and includes them as the P-Asserted-Identity header and the Privacy header of a SIP message to be transferred inside of a SIP trust domain.

- Transferring within a trust domain: inside of a SIP trust domain, P-Asserted-Identity header and the Privacy header of a SIP message are transferred transparently unless there is a specific purpose to do otherwise.
- Sending outside: At an outgoing boundary, the P-Asserted-Identity header and the Privacy header are either deleted, or mapped to other signal protocol message content to be transferred.

1.2. Project Detail

The project implements the asserted identity processing logic according to RFC 3324, RFC 3325 and TCC-1004.

A GUI is developed on the softswitch for operators to input Spec(T) rules such as whether a message is about to be routed out of the Trust Domain, whether to handle the Privacy header, the default privacy handling when no Privacy header field is present, and whether to modify the network asserted identity for specific calls. According to the message content and the Spec(T), the softswitch either (1) modifies the PAI header, (2) deletes the PAI header, (3) adds a new PAI header to the forwarded message, or (4) transparently transfers the PAI header to the next node.

In the project, the priority logic that extracts TEL URI, TEL DISPLAYNAME, SIP URI and SIP DISPLAYNAME from a received SIP request is developed. The old SIPE and SIP stack interface is replaced the with a more complex data structure. The enhanced interface enables the application to achieve more flexible manipulation of a PAI header and change both SIP PAI and TEL PAI headers at the same time.

2. The 300 Handling and Local Recursion Project

2.1. RFC Requirements

When a SIP proxy receives a request, it performs a sequence of tasks including (1) validate the request, (2) preprocess routing information, (3) determine target(s) for the request, (4) forward the request to each target, and (5) process all responses.

The target determination process includes obtaining a set of target URIs from a location service. The location service can use any information in or about the request or the current environment of the element to construct the target set, e.g., the contents or the presence of specific header fields and bodies, the time of day of the request's arrival, the interface on which the request arrived, and so on. A proxy may continue to add targets to the target set during the process of request forwarding. New targets can be obtained from a redirect response (3xx), or from further consultation with a location service. A target, however, cannot be added more than once.

Upon receiving a non-empty target set from a location service, a proxy forwards the request to each target using the following steps:

(1) Make a copy of the received request. Except for fields that are subject to modification during request forwarding, a copied request contains all the header fields from the received request.

(2) Choose a target and update the Request-URI with a target’s Request-URI. A common mechanism to choose a target from a target set is based on a target’s

qvalue parameter obtained from the Contact header field. Targets are processed from highest *qvalue* to lowest;

(3) Update the Max-Forwards header field by decrementing its value by one;

(4) Optionally add a Record-route header field value;

(5) Other steps include optionally add additional header fields; post-process routing information, e.g. mandating a request to visit a set of specific proxies by pushing Route values into the Route header field; determine the next-hop address, port, and transport; add a Via header field value before the existing Via header field values; create a new client transaction to forward the new request; and set timer C to handle the case when an INVITE request never generates a final response.

The response processing includes: (1) find the appropriate client transaction and response context, (2) update timer C for provisional responses, (3) remove the topmost Via header field, (4) add the response to the response context, e.g., update the target set with the received 3XX response, (5) check and immediately forward a provisional response (exclude 100 Trying response) and any 2XX response, (6) when necessary, choose the best final response from the response context. If no final response has been forwarded after every client transaction associated with the response context has been terminated, the proxy must choose and forward the "best" response from those it has seen so far, and (7) other processes that must be performed on each forwarded response.

2.2. Project Detail

The project implements the second step of the request forwarding process and the fourth step of the response processing, called local recursion and 300 handling. After a location service has determined that multiple routes could be used for a request URI, the proxy tries them in order to complete the call. The following rules are used in the local recursion and 300 handling:

- The initial list of routes are determined by consulting a location service, which form an initial target set;
- New targets may be added to the target set when a prior forwarded INVITE results in a 3XX response;
- Each target in the target set is tried in order;
- The SIP protocol stack processes all the 1xx provisional responses. The local recursion logic is not affected by a 1xx provisional response.

2.2.1. Target Set Processing

A target set consists of a list of targets to which calls may be routed. The proxy tries each target in the target set in order until a call is established or terminated. Targets can be derived from any of the three possible sources, stated as follows:

- Initial Targets

The initial target set comes from a query response of a location service. When the proxy receives an INVITE, it sends a route query to a location server. Based on the request URI and data in the location server database, the location server returns a list of routes to be contacted, constituting the initial target set in the proxy.

- Redirected Targets

If a routed request results a 3xx response, the contact list contained in the response forms redirected targets. The proxy could treat these redirected targets in any of three ways, including (1) accepting redirected targets recursively, (2) accepting single level redirected targets only (non-recursive), or (3) rejecting all redirected targets. Action (1) is the default.

The contact list may contain an optional parameter, called *q-value*, ranging from 0.000 to 1.000. The redirected targets are sorted based on their q-values before merged into the target set. A redirected target containing no *q-value* takes higher precedence (q=1) than one that does. The proxy must forward a request to a redirected target with higher *q-value* before the one with lower *q-value*.

Duplicate targets will be eliminated through a duplicate checking algorithm to prevent infinite redirection loops. A target counter is also used to limit the maximum target set size. The default value for the size limit is 100.

- Re-routed Targets

When the request URI of a redirected target contains the proxy address in its host portion but a different user name from the original user name in the initial INVITE request (otherwise it will be considered as a loop and discarded), the request URI is sent to the location service. The returned routes are added to the target set as the creation of the initial target set.

2.2.2. Response Processing

While the proxy attempts to contact each target in the target set until a call is established or terminated, the number of targets in the target set may grow due to insertion of redirected and re-routed targets. Once a target is contacted, the proxy decides whether to terminate or continue the local recursion based on the returned response code, e.g., if the proxy receives a ‘402 Not Found’, ‘502 Bad Gateway’ or ‘503 Service Unavailable’, the proxy retrieves the next target from the target set, forms a new Request-URI, and continues the local recursion. If there is no target is available, the proxy responds to the INVITE originator with a ‘504 Server Timeout’.

CHAPTER V

SIP AND SIP EXTENSIONS OVERVIEW, STATUS AND DEVELOPMENT

This chapter is organized as follows: Section 1 briefly introduces the standardization of SIP and SIP extensions; Section 2 outlines SIP messages, network elements and basic operations; Section 3 surveys the current practice and standardization status of SIP security; Section 4 proposes an Internet draft that addresses the secure retargeting and response identity issue in SIP.

1. Introduction

Session Initiation Protocol (SIP) is an application-layer signaling protocol that creates, modifies, and terminates multimedia sessions including Internet telephone calls, multimedia distributions and multimedia conferences [Rosenberg et al. 2002]. The core SIP specification is defined in RFC 3621, which obsoletes RFC 2543 [Handley et al. 1999].

In 1997, the Internet Engineering Task Force (IETF) Multiparty Multimedia Session Control Working Group (MMUSIC) developed the first version of SIP as part of the Internet Multimedia Conferencing Architecture. MMUSIC submitted the second version as an Internet Draft in 1998. In March 1999, IETF established the SIP working group and moved the protocol to the Proposed Standard status (named RFC 2543) to meet the growing interest in SIP. In order for an RFC to advance from proposed standard status to draft standard status, the protocol must have multiple independent implementations and achieve interoperability. The SIP interoperability test events or SIPit, have been held several times each year since 1999. SIP, together with Media Gateway Control Protocol [Arango 1999], have become the core signaling protocols of both the next generation wireline telecommunications architecture and the next generation wireless telecommunication (3G) IP Multimedia Subsystem (IMS) architecture.

SIP operates on a HTTP (Hyper Text Transfer Protocol) like client and server transaction model, where client and server exchange messages in SIP requests and responses [Johnston 2003]. The model is in conformance with the Internet model in which intelligence such as call processing logic and call states resides on end devices. It is scalable, resists a single point of failure, and is open to the implementation of new services.

SIP provides the following signaling functions [Rosenberg et al. 2002]:

- Register end user locations;
- Reach an end user based on its single, location independent address;
- Perform calling and called user agents authentication;
- Negotiate media and media parameters to be used;
- Create new sessions and manage existing sessions, including transferring or terminating sessions, modifying session parameters, and invoking services.

After a multimedia session is created, the communication itself has to be supported through other protocols, e.g. Real-time Transportation Protocol (RTP) transports real-time multimedia data across the network, Session Description Protocol (SDP) describes session characteristics, such as codes, transportation protocols and data rate on end devices. The architecture of Internet multimedia protocols is summarized and depicted in Fig. 4.

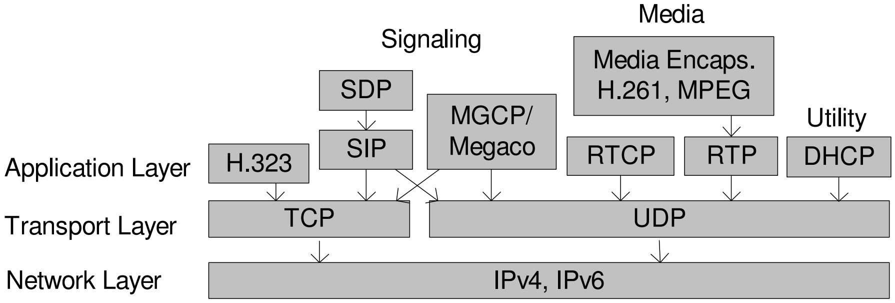

Fig. 4. The architecture of Internet multimedia protocols.

The development of SIP has led to the formation of other SIP related working groups [Johnston 2003]. The SIP Project INvestiGation (SIPPING) working group concentrates on the application of SIP and its extensions. It specifies the frameworks, requirements and common practice of SIP applications. The SIP Instance Messaging and Presence Leveraging Extensions (SIMPLE) working group works on presence and instance message applications. The 3rd Generation Partnership Project (3GPP) uses SIP in release 5 and later of the Internet Multimedia Subsystem (IMS). Other SIP related working groups include the PSTN and Internet Internetworking (PINT) working group, and the PSTN/IN requesting Internet Service (SPIRITS) working group. These working groups investigate SIP applications, extensions, interoperations with the PSTN, and publish best current practices (BCP).

At the same time, many SIP extensions and related specifications are being proposed. RFC 3265 Session Initiation Protocol (SIP)-Specific Event Notification [Roach 2002] enables SIP nodes to request notification of certain event occurrence from remote nodes. RFC 3428 Session Initiation Protocol Extension for Instance Message enables [Campbell et al. 2002] SIP to use multipart bodies to deliver instant messages. RFC 3262 Reliability of Provisional Responses in Session Initiation Protocol (SIP) [Rosenberg and Schulzrinne 2002a] describes reliable provisional response. RFC 3263 Session Initiation Protocol (SIP): Locating SIP Servers [Rosenberg and Schulzrinne 2002b] describes DNS mechanisms for locating SIP servers. RFC 3264 An Offer/Answer Model with Session Description Protocol (SDP) [Rosenberg and Schulzrinne 2002c] specifies how to use SDP within SIP to negotiate sessions.

2. SIP Overview: Messages, Network Elements and Basic Operations

SIP operates on a client and server transaction model and exchanges messages in the form of request and response. In a request, a client specifies a method name, a request-URI indicating the call recipient’s address, miscellaneous message headers providing additional information such as a unique call identifier, source and destination addresses, and message body type and length. Both the request and response can carry a message body, which usually contains a description of the session encoded in another protocol format, such as Session Description Protocol (SDP). SIP also uses the message body to transfer instance messages and events. The whole message uses a clear text-encoding scheme, inherited from Simple Mail Transport Protocol (SMTP).

The method name in a SIP request specifies the operation to perform. In RFC 3261, six methods are defined: a REGISTER method that registers a user's contact information, an INVITE method that invites another user to a session, an ACK method that facilitates reliable message exchange for INVITEs, a CANCEL method that terminates a request, a BYE method that terminates a session, and an OPTIONS method that queries servers about their capabilities. Other SIP methods defined in SIP extensions include INFO, PRACK, MESSAGE, SUBSCRIBE and NOTIFY.

The status code in a response indicates the outcome of the request execution. Six classes of response status code are defined, which are:

1xx: Provisional -- request received, continuing to process the request;

2xx: Success -- the action was successfully received, understood, and accepted;

3xx: Redirection -- further action needs to be taken in order to complete the request;

4xx: Client Error -- the request contains bad syntax or cannot be fulfilled at this server;

5xx: Server Error -- the server failed to fulfill an apparently valid request;

6xx: Global Failure -- the request cannot be fulfilled at any server.

In a 2xx response, the message body may carry the media preference of the callee.

In RFC 3261, five logical entities are defined: *user agent*, *proxy*, *registrar*, *redirector,* and *location server*. The user agent is a network end device that issues SIP requests or responses. The user agent that initiates a SIP request is called *user agent client* (UAC) and the user agent that responds to the request is called *user agent server* (UAS). Note that both the UAC and UAS are logic elements and are specific to each transaction, e.g., a user agent can act as a UAC in one transaction and act as a UAS in another.

The SIP proxy sends requests and receives responses on behalf of its clients. It routes a request to a user's registered location, authenticates and authorizes users for services, implements call routing policies, and provides features.

The redirector accepts requests, maps the request-URI into a contact list and then returns the contact list in a 3xx response. The registrar receives a user's REGISTER request and stores the provided contact list into a location server. The location server contains a database that stores such user information as registration, contact list and presence information.

A session can be created between two user agents with three SIP messages: an INVITE request, a 200 OK response and an ACK request. A session created in this three-way handshaking process represents the simplest session creation transaction. If a calling party knows the called party's IP address and port number, it can dial the called party's IP address and port number and use the IP routing mechanism to route the SIP message to the called party. In the real world, however, this is neither convenient nor feasible, e.g., the IP address reveals privacy of the called party and may no longer be valid when the called party changes its location. The SIP infrastructure facilitates locating users or services, so that a description of the session can be delivered. Therefore, a call is usually created with the involvement of SIP proxies and redirectors, corresponding two basic SIP operation scenarios: the session creation in redirect mode and in proxy mode.

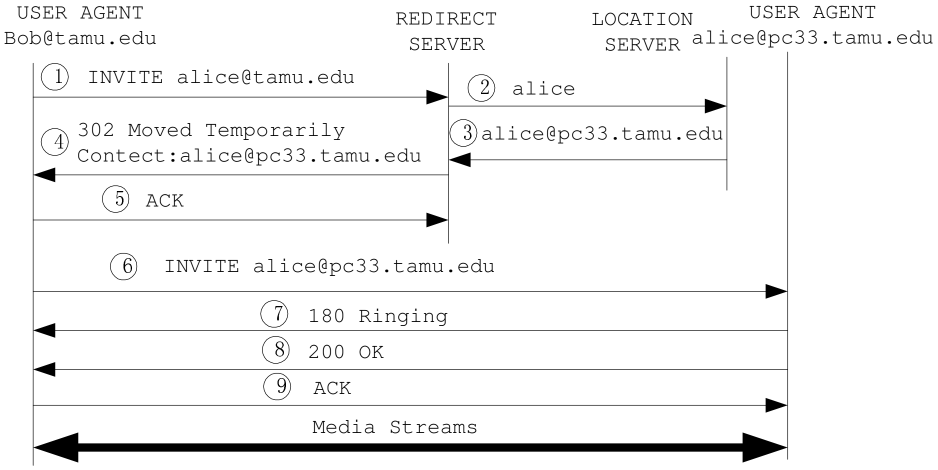

Fig. 5. Session creation in redirect mode.

Fig. 5. illustrates the redirect mode. When *Bob* wants to make a call to *Alice*, he picks up the phone and enters *Alice*'s address alice@tamu.edu. The phone forms an INVITE and sends it to the redirect server. The redirect server consults the location server for *Alice*'s current location and obtains alice@pc33.tamu.com that she has registered earlier with the registrar. The redirect server sends *Bob* a 302 Moved Temporarily message with its contact header indicating *Alice*'s new address. *Bob* confirms receiving 302 with an ACK and sends an updated INVITE to pc33.tamu.edu. The message will be resolved via DNS look up and the redirect server will no longer participate in the interaction. When *Alice* picks up the phone, it sends back a 200 OK, which will be routed back to *Bob* based on the route information collected by the request and copied to the response in the reverse order. The media streams are created after *Bob* returns an ACK. The message content is specified in appendix A.

A session can also be created in a proxy mode (shown in Fig. 6). Instead of returning the contact list to *Bob*, the proxy server directly forwards the INVITE with an updated request-URI. In practice, the proxy server usually serves as a session border controller, a billing center or a network core routing engine. The operation mode and routing scheme can be either based on user location data provisioned through user registrations or pre-configured local policies.

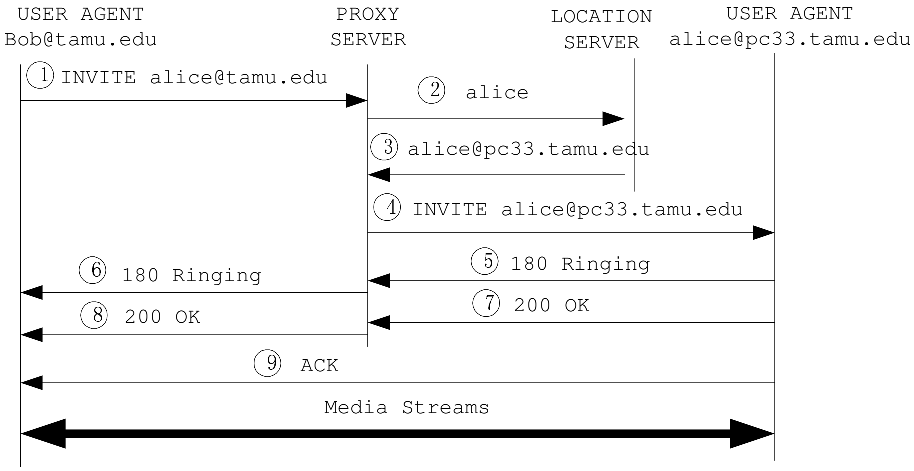

Fig. 6. Session creation in proxy mode.

In both modes, the media stream exchange is separated from the signaling exchange. The signaling may pass through several proxies or redirect servers before it reaches the destination, whereas the media stream can take a more direct path between calling and called parties. The approach is analogous to the separation of signal and media transport in the SS7 network.

3. The Current Standardization Status on SIP Security

3.1. Introduction

VoIP offers low cost and high flexibility. It also presents significant security challenges. Compared with the conventional telephone system, in which eavesdropping requires tapping a line or penetrating a switch, the standards-based VoIP shares physical network connections with the data network and presents intruders many more potential vulnerable attack points. Widely available network tools facilitate intruders to monitor and control network packets. An intruder can easily modify the message content and route calls through malicious nodes. As more service providers enter the market and accept traffic from foreign domains or from the open Internet, providers must consider how they will deal with security issues before a security catastrophe really happens.

In VoIP, the media stream can be secured through SRTP (Secure Real-Time Transport Protocol) or IPSec. Session keys for SRTP can be established/exchanged in the following three paths:

1. signaling path, e.g., the MIKEY (Multimedia Internet KEYing) protocol [Arkko et al. 2004] uses pre-shared keys, public keys or the Diffie-Hellman method to set up session keys. The integration of MIKEY with SIP/SDP is defined in [Arkko et al. 2005].

2. media path, e.g., ZRTP [Zimmermann et al. 2006] performs Diffie-Hellman key exchange during a call setup in-band in the Real-time Transport Protocol (RTP) media stream established using other signaling protocols such as SIP.
3. both the signaling path and the media path, e.g., DTLS-SRTP (Datagram Transport Layer Security – Secure Real-time Transport Protocol) [McGrew and Rescorla 2006] exchanges public keys in SDP and uses it to establish a DTLS session over the media channel. The endpoints then use the DTLS handshake to establish SRTP session keys.

The VoIP signaling message carries identities of the calling and called parties, the session ID, contact information, media stream specifications, instance messages or events, and possibly encryption keys used for the media stream. The signaling protocol needs to have secure mechanisms to preserve message confidentiality and integrity, prevent message replay attacks and message spoofing, provide authentication for session participants, and prevent denial-of-service attacks.

Contrary to high security requirements, SIP is not an easy protocol to secure. Its use of intermediaries, its expected usage between untrusted elements, and its user-to-user operation make security far from trivial. Since intermediate proxies use SIP message headers such as Request-URI, Route and Via to route messages, encrypting the whole message end-to-end is impossible.

In this section, we investigate the current and proposed security mechanisms for SIP in IETF.

3.2. Using Transport Layer Security

SIP can employ transport layer security mechanisms such as IPSec (IP Security) or TLS (Transport Layer Security) to encrypt the entire SIP message on a hop-to-hop basis. IPSec is most commonly used between hosts or domains that have existing trust relationships. It operates at the operating system level and provides confidentiality and integrity for all traffic passing on a host or security gateway. TLS operates on the application level and is most suited between hosts with no pre-existing trust association, e.g., between the UA and local proxy server [Rosenberg and Schulzrinne 2002a].

The end-to-end security might be compromised if a single proxy server along the route does not implement TLS or IPSec. SIP specifies the SIPS URI scheme to signify that each hop along the signaling path *must* forward the request and its responses over TLS connections. In practice, it is hard to guarantee that TLS usage will be truly end-to-end. It is possible that cryptographically authenticated proxy servers along the way are compromised, or noncompliant or disregard the forwarding rules associated with SIPS, downgrading the security requirement indicated by the caller.

3.3. Taxonomy of Security Mechanisms Provided in SIP

SIP incorporates digest authentication and asymmetric keys based mechanisms to achieve message confidentiality, integrity and authenticity. Based on the scope of their usage, the current standardization work of SIP security mechanisms is classified into (1) security within the same administrative domain, (2) security within the same trust domain, and (3) end-to-end security.

3.3.1. Security in the Administrative Domain

In an administrative domain, a SIP proxy receives and routes requests for user agents (UAs), and a registrar accepts user updates on their current locations. If the registrar cannot authenticate the originator of a request, an attacker can impersonate another UA to change the contact address of the UA with a REGISTER request, causing future requests for the UA to be routed to the attacker's device. The proxy uses authenticated user information for billing, caller ID, and user administration.

RFC 3261 [Rosenberg et al. 2002] borrows challenge-response based digest authentication for UAs from HTTP. When a proxy or a registrar receives a request, it can challenge the initiator of the request with a 407 'Proxy Authentication Required' or a 401 'Unauthorized' response. The 'Proxy-Authenticate' or 'WWW-Authenticate' header field of the response carries authentication parameters, among others, the authentication scheme, authentication domain (realm) and a nonce (number used once). The UA then locates credentials associated with the specified domain, either from a user's input or from an internal keyring, and re-originates the request with credentials embedded in the 'Authorization' or 'Proxy-Authorization' header field. In the forking operation, the forking proxy is responsible for aggregating all the challenges from various proxies into a single response. The re-originated request will include an Authorization value for each WWW-Authenticate value, and a Proxy-Authorization value for each Proxy-Authentication value, differentiated by the 'realm' parameter.

The digest authentication requires the proxy (or registrar) to share a secret with the UA. It is therefore often used when the proxy (or registrar) and the UA are in the same administrative domain. The mechanism authenticates messages and protects against relay attacks. It does not, however, ensure message integrity and confidentiality.

On the other hand, while UAs can authenticate themselves to servers with digest authentication, if a UA cannot authenticate a server to whom it sends a request, a malicious server can forward the request to inappropriate or insecure resources. SIP does not provide a mechanism to authenticate servers. The server uses site certificates delivered by TLS to authenticate themselves to UAs or next hop servers. RFC 3261 mandates implementation of TLS and certificate validation mechanisms on SIP proxies, redirectors and registrars. Once a UA and a registrar have mutually authenticated each other and created a TLS connection, the UA can leave the TLS connection open if the registrar also acts as a proxy server to which requests are sent for users in the same administration domain.

3.3.2. Security in the Trust Domain

RFC 3324 [Watson 2002] and 3325 [Jennings et al. 2002] extend the identity asserted in the authentication process to a trust domain. The identity is defined as sip:, sips: or tel: URI with an optional Display name. Its meaning lays in that if used as a Request-URI in a request, it could cause the request to be routed to the user/device associated with the identity. A trust domain consists of mutual trust nodes. A node α trusts node β if and only if (1) a secure connection exists between α and β, AND (2) β has configuration information indicating α is a member of the trust domain, which is often achieved through bi-lateral agreements.

In the authentication process, a proxy asserts the caller's identity and inserts a P-Asserted-ID header to carry this identity inside a trust domain. If a user registered multiple identities in a domain, the user can provide the proxy with a P-Preferred-Identity header to suggest which of the multiple valid identities for the authenticated user should be asserted. Domains that receive a request with a P-Asserted-ID header from an untrusted domain cannot use the header to assert the caller's identity. To meet the privacy requirement, before forwarding a message to servers or UAs in untrusted domains, a proxy must remove all P-Asserted-Identity headers if the caller requested that this information should be kept private.

RFC 3325 has enjoyed widespread deployment [Rosenberg 2006]. The technique is built into many proxies, application servers, and end user devices. Many IP phones or adaptors use the P-Asserted-ID header field as the source for secure caller ID. The technique, however, depends on the underlying trust infrastructure and mutual trust agreement between providers or trust domains. It may suffer from the same vulnerability as calls created over hop-to-hop TLS connections.

3.3.3. End-to-End Security

- *S/MIME (Secure Multipurpose Internet Mail Extensions)*

The first work in end-to-end security is within RFC 3261 [Rosenberg et al. 2002] itself, which specifies SIP support for S/MIME. To provide end-to-end authentication and integrity, the sender can sign a SIP message and attach the signature as an "application/pkcs7-mime" body. Since proxies on the signaling path can legitimately modify certain SIP headers, including Request-URI, Via, Record-Route, Max-Forward and Proxy-Authorization, the sender has to replicate some header fields that the sender wishes to secure in a “message/sip” MIME body, forming an “inner message.” If confidentiality is desirable, the inner message can be encrypted with the public key of the intended recipient and becomes an “application/pkcs7-mime” MIME body. In practice, since a plaintext version of certain SIP headers, including To, From, Call-ID, Cseq, Contact, is always required in requests and responses, the general use of encryption with S/MIME is to secure message parts like SDP and other header fields that have an end-to-end semantic, such as Subject, Reply-To, Organization, and Supported. In the end, the SIP message creates a “multiple part/signed” body that contains (1) either a plaintext “message/sip” body or an encrypted “application/pkcs7-mime” body for the “inner message,” (2) an “application/pkcs7-mime” body for the signature on the “inner message,” and (3) a certification bearing the public key necessary to verify the signature.

The S/MIME certificate associates the Address of Record (AoR) with keys, endorsed by certificate authorities. When a UAS receives a request that contains a certificate, it validates the certificate with available root certificates of certificate authorities. If the certificate is self-signed, or signed by an unknown authority, the use of the certificate needs the user’s consent. Verified or explicitly authorized certificates are added to the local keyring that consists of AoR and certificate pairs. The same processing is also applied for UAs that receive responses containing certificates.

Although the S/MIME mechanism is very secure, it has seen little implementation and no deployment, since it depends on the existence of end user certificates and there is virtually no consolidated central authority today.

Similar to S/MIME is the Authenticated Identity Body (AIB) specified in [Peterson 2004], which requires a UA to sign the identity in the body of a request or response. Since it also requires a public key infrastructure to support using private keys and certificates in every UA, it has little deployment.

- *Enhanced SIP Identity*

RFC 4474 [Peterson and Jennings 2006] provides a signature-based technique to deliver authenticated identities and message bodies to the caller recipient securely. To better scope the problem, it suggests using the domain certificate instead of individual certificate for each UA. Once Alice sends an INVITE over a TLS connection to an authentication service proxy in her domain, the authentication service authenticates Alice via digest authentication and validates that she is authorized to assert the identity populated in the From header of the request. The authentication service then computes a hash on a canonical string generated from certain components of the SIP request, including the header fields of From, Date, Call ID and the message body. The hash is then signed with the domain certificate and inserted in the 'Identity' header. The authentication service also inserts an 'Identity-Info' header to inform the call recipient Bob about where to acquire the domain certificate.

When Bob's domain receives the request, it first retrieves and verifies Alice's domain certificate if no local verified version is available. With the domain certificate and the signature in the Identity header, Bob can validate whether the authentication service in the host portion of the AoR in the From header authenticates the user, and permits the user to assert the From header field value.

If a received Identity-Info header contains a URI that cannot be dereferenced or the referenced certificate cannot be validated, the call recipient responds with a 436 ‘Bad Identity-Info header’ or a 437 ‘Unsupported Certificate’ respectively. If a received request does not have an Identity header or the Identity signature does not correspond to the hash of the digest string, the call recipient responds with a 428 ‘Use Identity Header’ or a 438 ‘Invalid Identity Header’ respectively.

The SIP identity does not enjoy as widespread deployment as the P-Asserted-Identity. First, the SIP identity is more complex and requires many more updates on the network element. Second, since the identity mechanism generates a signature over key parts of a SIP request, including the message body, the widespread usage of back-to-back user agents (B2BUA) and other elements on the request path that modify the body will essentially invalidate the signature, and consequently the mechanism. [Rosenberg 2006] proposes a mechanism for coexistence of the SIP identity and P-Asserted-Identity, which suggests using the SIP identity mechanism between proxies and to use P-Asserted-Identity for transfer of asserted identity within a domain.

Other than the security mechanisms mentioned above that are in the RFC status, several Internet drafts draw much attention in the IETF. These include connected identity and end to middle security.

- *Connected Identity*

The SIP Identity only provides the called party with the calling party’s identity, but not in the reverse direction. As retargeting commonly occurs, calls can be transferred and forwarded to a different AoR other than the one specified in the original request, transparent to the caller. In practice, both the calling party and the called party can change during a call. [Elwell 2007] addresses this issue and uses mid-dialog request, e.g., an UPDATE method or re-INVITE method, to transfer the updated calling and called party's identity according to RFC 4474. The solution involves changing the URI in the To and From header fields of the mid-dialog requests and their responses, compared with the corresponding values in the original request and response that form the dialog – a practice prohibited in RFC 3261.

Other than end-to-end security, [Ono and Tachimoto 2007] proposed end-to-middle security for securely accessing specific servers on the signaling path while keeping some message content confidential to other servers.

In section 4 of this chapter, we propose an Internet draft that addresses the request retargeting and response identity issue.

3.3.4. Other Security Issues

Other than the problems and solutions addressed in this document, a whole class of problems are expected to receive further study in ongoing SIP work. A comprehensive taxonomy of VoIP security and privacy can be found in [VoIPSA 2005]. While some issues can be addressed in SIP, others can only be solved in a systematic approach involving things such as product implementation, security applications, policy and legal actions.

As SIP servers are designed to accept traffic from worldwide IP endpoints, they face distributed denial of service attacks. Some common examples include sending a large number of invalid, malformed or random SIP messages to trick a SIP server into parsing or setting up a larger number of transactions, causing the server to crash or exhaust resources. The attacker can also use the Via header, Record-Route header and Route header to route requests or responses to some vulnerable target hosts, and amplify the attack with forking proxies. Comprehensive lists of possible SIP related Dos attacks can be found at the VoIPSA website [VoIPSA 2005]. Recent work that addressed VoIP Dos attacks includes [Oulu University 2006] that provided a security test suites called PROTO for malformed messages, [Chen 2006] that used a modified SIP transaction state machine to detect transaction anomalies, [Geneiatakis et al. 2005] that proposed a framework for detecting malformed SIP messages, and [Wu et al. 2004] that proposed an abstract intrusion detection framework called SCIDIVE that detects anomalies such as inconsistency between the signaling channel protocol and the media channel protocol.

SIP messages may frequently contain sensitive information about their senders - not just what they have to say, but with whom they communicate, when they communicate and for how long, and from where they participate in the session. Many applications and their users require that this sort of private information be hidden from any party that does not need to know it [Rosenberg 2002a]. The message requires the personal information in the header fields to be concealed - not only in the From and related headers representing the originator of the request, but also in the To header.

There are also less direct ways in which private information can be divulged. If a user or service chooses to be reachable at an address that is guessable from the person's name and organizational affiliation (which describes most addresses-of-record), the traditional method of ensuring privacy by having an unlisted "phone number" is compromised. A user location service can infringe on the privacy of the recipient. An implementation consequently should be able to restrict, on a per-user basis, what kind of location and availability information is given out to certain classes of callers [Rosenberg and Schulzrinne 2002a]. The current work in IETF is limited to removing certain privacy sensitive headers before a message leaves a trust domain.

In [Rosenberg and Jennings 2008], the problem of call, instance message and presence Spam and the solution space are studied.

4. Retargeting Security and Response Identity in the Session Initiation Protocol (SIP)

As a SIP request is processed along its route to the destination, the initial request-URI can be altered without the callers' notice or consent. The caller may concern both the identity of the final call recipient and the authorities of the SIP intermediaries that alter the request-URI. Especially when the caller does not know the final call recipient, simply giving his/her identity to the caller will not help the caller to decide the legitimacy of the call. Without a secure retarget mechanism, the end-to-end security of SIP cannot be guaranteed. In the section, we propose a security mechanism to provide the caller with credentials of SIP intermediaries that retarget a request and the final recipient's identity through response.

4.1. Introduction

As a SIP request is processed in intermediaries, the initial request-URI can be altered with one or more targets identified via a location service. This process, so-called retargeting, is often done without the callers' notice or consent. Since the current standards do not provide a mechanism for a UAC to constrain or authorize SIP intermediaries as to what should be performed, and to authenticate the final call recipient's identity through SIP response, the UAC does not know where a request goes, how a request reaches a particular UAS and who this UAS is [Peterson 2005].

In some circumstance, users are more interested in how a request reaches a particular UAS, e.g., when Alice calls Bob and Bob redirects calls to Carol, Alice wants to make sure that it is Bob who designated the delegation agent. This is also useful in the calling center when a call is redirected to a special handling agent, especially when the agent is outside the original domain. The UAC can determine the URI that a request has eventually reached and determine whether the chain of trust is broken during request retargeting, e.g., the request is retargeted in some suspicious domains.

Although connected identity [Elwell 2007] has proposed to use mid-dialog requests, e.g., an UPDATE method or re-INVITE method, to transfer the updated calling and called party's identity based on RFC 4474, several issues are still not resolved:

(1) The response identity problem. The handling of responses such as '493 Undecipherable' and 3xx is fraught with risks if the identity of the sender of the response cannot be identified. Consider the following scenario mentioned in [Peterson 2005]: If Alice's request to Bob is retargeted to Carol and Carol does not possess the private key corresponding to Bob's public key, she would send some sort of failure response code (perhaps a 493 Undecipherable). According to the manner suggested in RFC 3261, Alice might re-initiate the session using Carol's certificate received in the body of the 493 response. Here, Alice has no

way of knowing if Carol is actually an attacker who sends a 493 in order to bid-down the security for the ensuing RTP session.

(2) If (1) can sometimes be avoided with connected identity [Elwell 2007], it means more rounds of message exchange. For example, if a session only consists of an INVITE and a 3XX, [Elwell 2007] seems over weighted.

(3) If the target is redirected to an unknown domain, then secure retargeting is more important because of the unanticipated respondent problem, and the caller trusts the initial call recipient's domain and its retargeting process more than the new respondent.

To achieve end-to-end security, we feel that the response identity problem cannot be omitted, and it has to be solved with the secure retargeting. It is essential for the protocol to provide a mechanism to feed the caller with (1) the retargeting information, (2) the credential of the intermediate server that retargets a request to a different Request-URI, and (3) the final recipient identity. This document proposes such a mechanism.

4.2. Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

Related response: final responses that define the security attributes of existing or future dialogs. The related responses include the 2xx response that carries the callee's identity, the 3xx response that carries the callee's new contact addresses, and the 496 or 493 response that carries the security key for future dialogs.

4.3. Overview of Solution

The fundamental functionality provided by the secure retargeting mechanism is the ability to collect credentials of intermediate servers that retarget requests and capture the associated request-URI change. The original request-URI, modified request-URI and the Identity of the server that does the request-URI modification are recorded in a new header for SIP messages: Target-Info. The signature used for validating headers including the Target-Info header is conveyed in the Identity header, and the reference to the certificate of the signer is conveyed in the Identity-Info header, as described in RFC 4474. An additional index parameter is added to each of the above header to group related information for a single retargeting server.

In applications that concentrate on sending callers target change for successfully established dialogs, the Target-Info header is added to 2xx, 3xx and responses that would change the secure attribute of a future dialog, such as the 496 and 493 responses. In this specification, these responses are called related responses.

If a caller wants intermediaries to provide credentials of retargeting on related responses, she/he MUST insert a new option tag “target-info” in the request to initiate a session.

To shield the network configuration and reduce computation overhead, proxies on the border of a trusted network SHOULD eliminate intermediate retargeting process information for routing and other purposes. There, the Authentication Service proxy SHOULD be logically configured on the network border. When the Authentication Service proxy received an incoming request with “target-info” in the Supported header and the related response indicating that request-URI is changed within the trusted network, the proxy MUST insert Identity and Identity-Info headers in the response before forwarding the response to an untrusted network. Given secure connections exist between trusted network elements, the proxy SHOULD merge multiple Target-Info headers inserted within the trusted network into a single Target-Info header, which only records the last changed request-URI and the original request-URI received by the trusted network.

4.4. Behavior

4.4.1. User Agent Behavior

When issuing an INVITE request, a UAC that wishes to learn the intermediate target change MUST include a "target-info" option tag in the Supported header filed.

When receiving a response with Target-Info, Identity and Identity-Info header, the UAC inspects the signature in the Identity headers and validates related header fields and the message body.

Since each Target-Info provides an old and changed request-URI, and the last Target-Info provides the identity of the sender of the response, the Target-Info headers can form a trace of request-URIs when request is routed to the destination. For adjacent Target-Info pairs, the changed request-URI in the prior Target-Info MUST equal to the old or current request-URI in the next Target-Info.

The criteria for judging a request-URI change or for detecting a missing request-URI change segment are specified in section 6.

4.4.2. Proxy Behavior

For proxies that do not retarget requests, no behavior change is required.

The following is the behavior of a proxy that has performed or is about to perform retargeting.

When a proxy server receives a request with a “target-info” option tag in the Supported header filed, if the proxy server is about to change the request-URI but is not able to provide authentication service for the future related response, the proxy SHOULD return a 420 ‘Not Supported’ response. If the authentication service is provided in a centralized server, the proxy MUST be able to create a secure connection with the central authentication service.

When a proxy server receives a redirect response, before retargeting the request to the request-URI extracted from the contact header, the proxy server MUST first verify whether the redirect response is directly received from a trust domain, or whether the contact header of the response is verifiable from the Identity and Identity-Info header. Here, the proxy delegates the process of authentication of the response to the caller. A proxy server MUST not forward any related response that comes from an untrusted domain and does not have an Identity and Identity-Info header.

If the response comes from an untrusted domain but has an unverifiable Identity and Identity-Info header, the proxy SHOULD forward the response upstream to the caller. If the response is a result of retargeting performed at the proxy, the proxy MUST insert a Target-Info header, and then use the domain key to sign the hash of the canonical string generated from certain components of the response before forwarding.

If the proxy performs a sequential or parallel search, the proxy SHOULD exhaust verifiable contact headers first.

If several proxies within a trust domain perform retargeting, then each of these proxies SHOULD insert a separate Target-Info header. If network privacy is enforced, e.g., the consent framework [Rosenberg et al. 2007] conceals the detailed user location, the border proxy MUST omit privacy sensitive request-URI changes within the domain. In a transition domain, only the original request-URI received by the domain and the last changed request-URI when the request left the domain are kept in the Target-Info header. In the destination domain, only the original request-URI received by the domain is left in the Target-Info header. While security always causes overhead, the proper network configuration can significantly reduce it. Centralized authentication service on a border proxy is one example.

The Target-Info header MUST be signed before sending the related response out of the trust domain.

4.4.3. Redirector Behavior

If the redirect service only serves the proxy in the trust domain, then there is no behavior change.

Otherwise, when a redirector receives a request with the "target-info" option tag in the Supported header filed, it SHOULD insert a Target-Info, Identity and Identity-Info header for the redirect response, or do so through an authentication service.

4.5. Criteria for Recording and Checking Request-URI Change

The criteria of justifying a request-URI change depend on the request-URI scheme and the portion of the request-URI involved in a change.

If a GRUU [Rosenberg 2007] request-URI is used, each request-URI change MUST be recorded.

If the tel URI scheme is used, adding or deleting international or area code MAY be considered as a target change.

The username change in a sip URI MUST be considered as a target change.

In the same trust domain, the host portion of a request-URI may be changed several times. In the destination domain, the host portion of a request-URI is often detailed to a specific host address. As specified in section 4, the authentication service MAY choose to conceal such detailed retargeting information. In the same trust domain, only receiving last modified host portion of the request-URI is recorded. In the destination domain, only the receiving user name and host portion of the request-URI is recorded.

The user's involvement MAY be required for some ambiguous target change. The UA can list suspicious target changes via GUI.

4.6. Header Syntax

The Target-Info header carries the following information, with the mandatory parameters required.

- target change: A mandatory parameter contains either the current target or a pair of targets that reflect the target change.

- retarget-server: A mandatory parameter captures the server name that performs the retargeting.
- index: A mandatory parameter that groups related Target-Info, Identity and Identity-Info headers. The index starts at one. Each subsequent index increases by one.

```
Target-Info = “Target-Info”  HCOLON ( target-change |  current-target )
            COMMA retarget-server COMMA index
target-change = previous-target COMMA changed-target COMMA index
previous-target = “previous” EQUAL request-URI
changed-target = “changed” EQUAL request-URI
current-target = “current” EQUAL request-URI
retarget-server= “server” EQUAL name-addr
index = “index” EQUAL 1*DIGIT
request-URI = name-addr
```

The Identity and Identity-Info header defined in RFC 4474 are also updated with the additional index parameter.

```
Identity = "Identity" HCOLON signed-identity-digest COMMA index
Identity-Info = "Identity-Info" HCOLON ident-info *( SEMI ident-info-
params ) COMMA index
```

The signed-identity-digest is a signed hash of a canonical string generated from certain components of a SIP response. To create the content of the signed-identity-digest, the authentication service MUST use the elements of a SIP message placed in a bit-exact string specified in RFC 4474, and the added Target-Info header specified in this document, separated by a vertical line, "|" or %x7C, character:

```
digest-string = digest-string = addr-spec "|" addr-spec "|" callid "|"
1*DIGIT SP Method "|" SIP-date "|" [ addr-spec ] "|" message-body "|"
Target-Info
```

The Target-Info above refers to the local added Target-Info.

4.7. Message Examples

It is expected that most retargeting cases happen in the destination domain, in which the authentication service signs and forwards the response from the final call recipient backward to the caller.

In the following example (Fig. 7), we describe a simple case when UA Alice initiates an INVITE to Bob and the INVITE is redirected in the destination domain. We assume the destination proxy, the destination redirector, and the final call recipient Bob are all in the same trust domain.

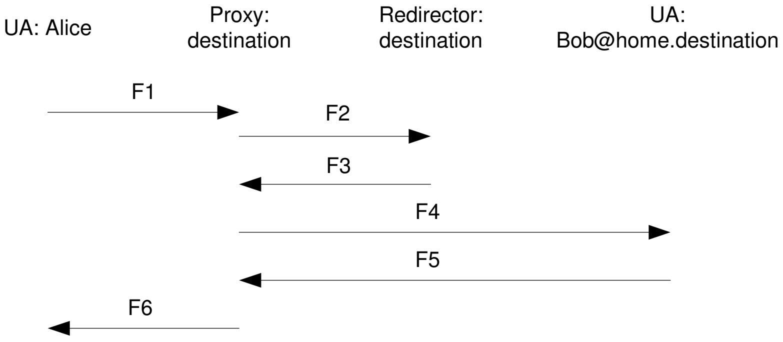

Fig. 7. Message flow.

F1: UA Alice -> Proxy destination

```
INVITE sip:bob@destination.com SIP/2.0
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Date: Thu, 21 Feb 2002 13:02:03 GMT
Supported: target-info
Contact: <sip:alice@pc33.atlanta.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=UserA 2890844526 2890844526 IN IP4 pc33.atlanta.example.com
s=Session SDP
c=IN IP4 atlanta.example.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

F2: Proxy destination -> Redirector destination

```
INVITE sip:bob@destination.com SIP/2.0
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Supported: target-info
      Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:alice@atlanta.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=Alice 2890844526 2890844526 IN IP4 atlanta.example.com
s=Session SDP
```

```
c=IN IP4 atlanta.example.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

F3: Redirector destination

Since both the destination proxy and redirector are in the same trust domain, no security-retargeting headers are generated. Otherwise, the redirector MUST insert security-retargeting headers and the proxy MUST verify these headers before retargeting the request to contact addresses specified in the returned 3xx response.

```
302 Temporally Moved
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:bob@home.destination.com>
Content-Type: application/sdp
Content-Length: 0
```

F4: Proxy destination -> UA Bob

The destination proxy changes the request-URI to bob@home.destination.com.

```
INVITE sip:bob@home.destination.com SIP/2.0
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Supported: target-info
```

```
Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:alice@atlanta.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=Alice 2890844526 2890844526 IN IP4 atlanta.example.com
s=Session SDP
c=IN IP4 atlanta.example.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

F5: UA Bob -> Proxy destination

```
200 OK SIP/2.0
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:bob@home.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=Alice 2890844526 2890844526 IN IP4 atlanta.example.com
s=Session SDP
c=IN IP4 atlanta.example.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

F6: Proxy destination -> UA Alice

Assume secure communications exist between the destination proxy and UA Bob and the destination proxy verifies UA Bob’s identity through HTTP change and response.

Also assume that the privacy policy allows the proxy to disclose the user location information to the caller.

```
200 OK SIP/2.0
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Target-Info: previous=bob@destination.com,
changed=bob@home.destination.com,current=bob@home.destination.com,
server=proxy.destination.com,index=1
Identity:
”PonWJMGvQTBDqghoWeLxJfzB2a1pxAr3VgrB0SsSAaifZYNBbHC00VMZr2kZt6VmCvsRdi
OPoQZYOy2wrVghuhcsMbHWUSFxI6p6q5TOQXHMmz6uEo3svJsSH49thyGnFVcnyaZ++yRlB
YYQTLqWzJ+KVhPKbfU/pryhVn9Yc6U=”, index=1
Identity-Info:   <https://desination.com/destination.cer>;alg=rsa-sha1,
index=1
Max-Forwards: 70
Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:bob@home.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=Alice 2890844526 2890844526 IN IP4 destination.com
s=Session SDP
c=IN IP4 destination.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

If we assume that the privacy policy does not allow the proxy to disclose the user location information to the caller, F6 should look like this:

F6: Proxy destination -> UA Alice

```
200 OK SIP/2.0
Via: SIP/2.0/UDP proxy.destination.com;branch=z9hG4bKnashds8
```

```
Via: SIP/2.0/UDP pc33.atlanta.source.com;branch=z9hG4bKnashds8
To: Bob <sip:bob@destination.com>
From: Alice <sip:alice@atlanta.source.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Target-Info:                current=               bob@destination.com,
server=proxy.destination.com,index=1
Identity:
”grB0SsSAaifsRdiOPoQZYOy2wZYNBbHC00VMZr2kZt6VmCvPonWJMGvQTBDqghoWeLxJfz
B2a1pxAr3VrVghuhcsMbHWUSFxI6p6q5TOQXHMmz6uEo3svJsSH49thyGnFVcnyaZ++yRlB
YYQTLqWzJ+KVhPKbfU/pryhVn9Yc6U=”, index=1
Identity-Info:   <https://desination.com/destination.cer>;alg=rsa-sha1,
index=1
Max-Forwards: 70
Date: Thu, 21 Feb 2002 13:02:03 GMT
Contact: <sip:bob@home.source.com>
Content-Type: application/sdp
Content-Length: 147

v=0
o=Alice 2890844526 2890844526 IN IP4 destination.com
s=Session SDP
c=IN IP4 destination.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000
```

4.8. IANA Considerations

This specification registers a new SIP header and a new option tag.

- *Header*

This specification registers a new SIP header, according to the guidelines in Section 27.1 of RFC 3261.

Name: Target-Info

Description: This new header captures the request-URI change and the current request-URI.

- *Optional Tag*

This specification registers a new optional tag, according to the guidelines in Section 27.1 of RFC 3261.

Name: target-info

Description: This option tag is used to indicate that a UA requires intermediate proxies that perform retargeting to add Target-Info, Identity and Identity-Info headers in the response.

CHAPTER VI

A SIP SERVER DESIGN FRAMEWORK

1. Introduction

The SIP specification RFC 3261 defines five logical servers: user agent, registrar, redirector, proxy, and location server. Each SIP server has a different role and provides different services. In practice, a single network element often integrates one or more SIP logical servers, e.g., a location server collocates with other SIP servers to provide location and enhanced services, as illustrated in the following example.

In the session initiation process, when a server α receives a SIP request, it consults a local location server for the callee's contact information based on the request-URI. The contact information is often created in the registration process, in which a registrar accepts a user's REGISTER message and creates a binding between the Address of Record (AoR) URI in the *To* header and the device URI in the *Contact* header of the message. The binding is further mapped into database operations to store the data at the location server for further access.

Once server α retrieves the callee's contact information from the location server, it can either act as a redirector that sends the callee's contact information to the caller in a redirect class response, or act as a proxy that directly routes the request to the callee. The location service can be further upgraded to realize value-added services, among others, including calling and called number translation, 1-800 service, ENUM, configurable sequential multi-destination dialing, and load balanced routing. The differentiated services can form a line of products, e.g., softswitches, network core routing engines and application servers.

In order to adapt to the rapid markets and standards evolution and meet different customer requirements from different market segments, an equipment vendor desires a product line with common and open service architecture and great reuse of time-proven software design and implementation.

In this chapter, we introduce a generic SIP server design framework that supports multiple products development. The framework follows a product line approach [Gannod and Lutz 2000] and provides the following features:

- *Extensibility*: The framework adapts to the fast standards development and enables software evolution. Small firms can reduce time-to-market with a base line product. As new SIP protocol extensions and service features are proposed, the framework facilitates integration of new service modules (called *feature boxes*), and extension and modification of the existing service logic.
- *Modularity*: The low cohesive module design enables autonomous development of independent function modules. In the current software development practice, due to the fact that software developers often make implementation changes without further updating the design document, the software implementation and design document often lack conformance [Jazayeri et al. 2000]. Meanwhile, the magnitude of the software continues to

grow. Because of dramatically increased complexity and lack of proper documentation, it is common that software developers make changes to the implementation without fully understanding the system. Modularizing independent functions and avoiding unforeseen couplings is critical to localize change.

- *Flexibility*: Instances of feature boxes can be loaded at run time. Services can be configured to interpret different state transition logic and actions of different SIP servers based on common criteria, among other things, including the calling URI, called URI and last hop server.
- *Performance*: The framework provides efficient data path and multithread processing with no expensive mutual lock. The modularity and flexibility do not sacrifice performance.
- *Systematic reuse*: In the literature, software reuse can be achieved at three levels: (1) the object oriented design and implementation concentrate on reusable object attributes and efficient algorithms; (2) design patterns [Gamma et al. 1995] provides a scheme for refining software elements and their relationship, and describes common structure of communicating elements that solves a general design problem within a particular context, preventing developers from traps and pitfalls often found later in the software development; and (3) domain specific architectural frameworks capture common abstractions of an application domain – both their structure and control mechanisms [Schmidt et al. 2000], which facilitates development of a

software product line, a collection of systems sharing a common set of core software components and base architecture. The framework systematically reuses software design and implementation at each level to achieve efficient and effective delivery of software products.

The chapter is organized as follows: in section 2, we describe the architecture design and functions of each architecture component; in section 3, we describe the call processing layer design with a comprehensive and architecture significant SIP call scenario - SIP 3xx message handling and local recursion; section 4 describes the SIP stack implementation architecture; section 5 describes design patterns that we employed to facilitate our design of component configuration, message passing and state machine; section 6 describes parallel architecture alternatives and a process view of our proposed framework; based on the logic and process view derived in the previous sections, and section 7 concludes with the class design. Section 8 describes related work done in the AT&T lab.

2. Architecture Design

The architectural description of the framework follows the *plane* concept introduced in the ISDN and later applied in GSM (Global System for Mobile Communications) and UMTS (Universal Mobile Telecommunications System). The solution domain is divided into the *management plane*, *provisioning and management interface plane*, and *service control plane* as shown in Fig. 8.

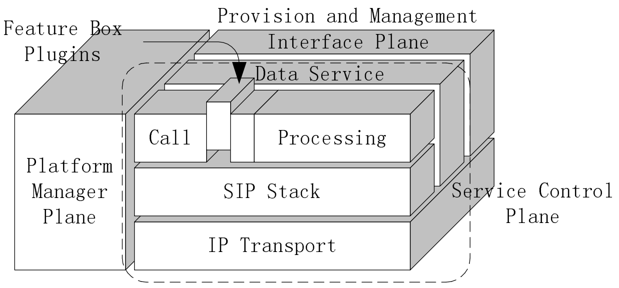

Fig. 8. Architecture components.

2.1. The Service Control Plane

The *service control plane* performs call control functions including call initialization, supervision and release. In order to structure the solution domain and promote a clean separation between reusable protocol-generic parts and protocol-specific parts, the service control plane is further decomposed into four functional groups including a system-wide data service, and a IP transport layer, SIP stack layer and call processing layer that each represents a particular level of abstraction.

The design facilitates software reuse, subdividing tasks among a team of developers and concealing services implementation details. Multiple transport components can be developed to support the SIP stack through a uniform interface, and the SIP stack can be developed to support multiple next generation products. If simple and stable interfaces are enforced between adjacent layers through engineering guidelines or policies, future code changes can be confined to the local component.

The data flow pattern consists of *requests* that send application data downwards from higher layer to lower layer and *notifications* that send network data and events upwards in the opposite direction. The communication between adjacent layers is achieved through function calls, in which data or data reference is passed as function call arguments. To decouple adjacent layers, the upper layer registers callback functions with the lower layer to accept notifications sent up. The upper layer directly calls the lower layer service interface to send down requests. In this way, the upper layer is aware of the next lower layer, but the lower layer does not have to know its upper layer users and does not rely on the data structures defined in the upper layer, achieving one-way coupling. In object-oriented design, the *Reactor* pattern can be used to implement callbacks and event de-multiplexing.

As a request moves downstream, it is usually translated into one or several subtasks, e.g., timer tasks and message format tasks, some of which are further sent downwards. The status of the request execution is passed back in the reverse direction when a function call returns. This chain of actions can also start at the lowest layer, e.g., when a network device detects a new packet, the data moves up from the lowest layer to the highest layer. There are situations where requests or notifications only travel a subset of the layers and return. In this case, the intermediate layer often maintains state information, e.g., detecting duplicate packets without further action, re-sending pending requests when timer fires.

The function of each layer is described below.

(1) *The IP Transport Layer*

The IP transport layer provides communication primitives to the SIP stack layer through a set of well-defined interfaces, e.g., Berkeley sockets API.

UDP is first transport protocol mandated for implementation in RFC 2543. As a result, most vendors start SIP application development on top of UDP. As new features are proposed and new SIP headers and contents are added, the SIP message size continues to grow. Although SIP is transport layer independent, it becomes an issue while sending a large SIP message over unstable network connections such as UDP. If a SIP message is broken into multiple fragments in the transport layer, the chance of lost fragments and retransmissions due to lost fragments will greatly increase, which seriously degrades the signaling performance. Therefore, RFC 3261 mandates implementation of "an RFC 2914 congestion controlled transport protocol, such as TCP" in addition to UDP mandated in the obsolete RFC 2543. Other transport mechanisms include Stream Control Transmission Protocol (SCTP) specified in RFC 4168 for exchanging a large number of messages between SIP entities, and Transport Layer Security (TLS) specified in RFC 4366 for secure communications over the Internet.

(2) *The SIP Stack Layer*

The SIP stack layer provides SIP transport, parsing and encoding, transaction and session service to its application clients.

The SIP transport service uses IP transport layer communication primitives to communicate other SIP entities. It also maintains and manages connections for sessions that use connection-oriented IP transportation.

The parsing service parses raw SIP packets forwarded by the transport service and saves message content into an internal SIP call data and event structure - SIP Message Block (SIP MB) that is suitable for upstream message delivery and service processing. The encoding service transforms SIP MBs traveling downstream from the application into compliant SIP raw packets and passes the raw packets to the SIP transport service.

The SIP stack transaction service manages SIP transactions and executes client and server transaction state machines. A SIP transaction consists of one request and multiple responses to the request. The transaction service manages transaction records, supports transaction validation and updates transaction state machines in realizing automatic progress update, best effort message delivery, and reliable provisional response acknowledgment (PRACK).

The SIP session service implements the *dialog* and *session* concepts defined in RFC 3261. A dialog consists of one or several SIP transactions. It creates a peer-to-peer relationship between two user agents, which facilitates sequencing messages and routing requests between user agents. A session consists of one or several dialogs. For instance, in a parallel search, when a proxy forwards an incoming request to several possible user locations and receives non-failure responses, several dialogs can be created. The session service manages these dialogs and sessions, creating context for the SIP transactions.

Callback interfaces are also provided for applications to receive upstream call data and events. The detailed design of the SIP stack is further discussed in section 4.

(3) *The Call Processing Layer*

The call-processing layer consists of a call-processing manager that manages application level transactions and feature boxes implementing call features.

The application transaction is created either when a new SIP session notification arrives, or when the application initiates a new SIP request. The data service allocates space to hold transaction data, among other things, including routing information, feature data, and references to one or multiple sessions in the SIP stack. The transaction is deleted either after a final response is issued to the SIP stack or when all referenced sessions are terminated. For example, in redirect mode, an application transaction is created when an initial INVITE is received, and is deleted after the transaction issues a 3xx response to the stack. The SIP stack is then responsible for the reliable transmission of the 3xx response to the client until an ACK is received. If a transaction resides in the system for a long time without updates, a garbage collection function contained in the call-processing manager will delete the transaction and free the memory space.

A transaction goes through a sequence of feature boxes for feature processing. Upon transaction creation, the call-processing manager ensures that the transaction passes a set of standard feature boxes based on local configurations and calling information carried in the SIP message. Additional feature boxes can be decided after calls go through feature box *Route*, through which specific features subscribed by individual caller and callee are retrieved from the data service.

A feature box can be loaded and unloaded at run time. It carries out functions independent of other feature boxes. The transaction is pushed to the next feature box either through function calls or message queues using a uniform interface. If a feature is not needed for a transaction, the related feature box just falls through. The precedence of feature boxes is pre-determined to produce correct feature interactions.

The design achieves flexibility and extensibility and does not compromise performance. New feature boxes can be developed and plugged into the system for proposed calling features. To eliminate the overhead of moving data between feature boxes, call transactions are created in the data service space and are protected as flow of control moves between function modules. Since no bulk data is moved, context switches due to call stacks overflow are omitted. The simulation shows that falling through additional feature boxes does not add noticeable overhead for call processing. The simple and uniform filter interface makes feature box recombination and reuse possible.

The design can be extended to accommodate multithreaded processing to increase system performance. The flexible thread configuration ensures optimal distribution of processing resource and eliminates computation bottlenecks, achieving maximum performance. The detailed call processing layer design is illustrated in section 3.

(4) *The Data Service Module*

The data service module provides a central data repository service for the system. According to a data lifetime, the data is classified into run time data and semi-permanent feature data.

- *Run-time data*, including protocol transaction data and application data.

- *Feature data*, loaded either from a local database created via the provisioning and management interface, or from a remote data service center, such as a location server.

In both cases, a sophisticated caching schema is required to cache feature data in system memory to reduce expensive database connection and SQL operations. The caching schema is critical to system performance. The feature data can be cached either when the application starts or when the data is first referenced. The least recent used feature data can be removed for new retrieved data.

The run-time data can be implemented as hash tables. In a multithreaded environment, data access through hash tables is protected between multithreads.

Since the interface between a SIP server and a location server is not defined in the SIP protocol, a SIP server can communicate with a location server through ODBC and other protocol interfaces. Special transport interfaces can be implemented to transport complex data structures over network packets.

2.2. *The Provisioning and Management Interface Plane*

The provisioning and management interface plane enables network operators to provision feature data stored in the data service module. It also exchanges management information with the platform manager via monitoring agents.

The interface communicates with the platform manager through network management protocols, e.g., SNMP or other socket based data structure transfers that forward such management commands as suspending/resuming services, turning on/off debug logs, and getting/resetting call statistics counters. In return, the interface receives command execution status and statistics information from the platform manager.

2.3. The Platform Manager Plane

The platform manager plane performs application configuration, initialization and run time environment management. It loads and initializes layers and feature box modules into the system according to the service configuration file. A typical module initialization process may include parsing module initialization arguments, spawning specific numbers of threads, initializing network sockets and message queues, and registering callback functions with adjacent layers.

The platform manager also initiates the logging service to collect system performance information from loaded modules. It executes platform management commands sent from the management interface plane.

3. Call Processing Layer Design

3.1. Session Initiation Protocol

The Session Initiation Protocol (SIP) is an application-layer signaling protocol that creates, modifies, and terminates multimedia sessions including Internet telephone calls, multimedia distributions and multimedia conferences [Rosenberg et al. 2002]. The core SIP specification is defined in RFC 3621, which obsoletes RFC 2543 [Handley et al. 1999]. SIP operates on a client and server transaction model, in which client and server exchange messages in SIP transactions that each consists of one request and one or multiple responses. The signaling functions of SIP can be classified as [Rosenberg et al.

2002]: (1) Register end user locations; (2) Reach an end user based on its single, location independent address; (3) Perform calling and called user agents authentication; (4) Negotiate media and media parameters to be used; (5) Create new sessions and manage existing sessions, including transferring or terminating sessions, modifying session parameters, and invoking services. After a multimedia session is created, the multimedia session has to be supported through other protocols, e.g. Real-time Transportation Protocol (RTP) that transports real-time multimedia data across the network, and Session Description Protocol (SDP) that describes session characteristics including codes, transportation protocols and data rate on end devices.

In RFC 3261, five logical entities are defined: *user agent*, *proxy server*, *registrar server*, *redirect server* and *location server*. In a SIP transaction, a *user agent client* (UAC) is a network end device that initiates a SIP request and a *user agent server* (UAS) is a network end device that responses to the request. Both the UAC and UAS are logic elements and transaction specific since a user agent can act as a UAC in one transaction and act as a UAS in another.

A SIP *proxy server* routes requests to a user's registered location, authenticates and authorizes users for services, implements call routing policies, and provides features. A *redirect server* accepts SIP requests, maps the called address into zero or more new contact addresses, and returns the contact address list to the request originator. When a *registrar server* receives a user's REGISTER request, it stores the provided contact addresses into a location server. The *location server* is essentially a database server that contains SIP registration, presence, and other information about a user.

The current VoIP markets include such products as media gateways, session border controllers, softswitches, media servers and application servers. Most of them support SIP and each consists of multiple SIP logical elements, e.g., a softswitch usually consists of a proxy server, redirect server and location server.

In the proposed framework, the unique functions of each logical server are abstracted as feature boxes and the common functions are implemented in the SIP stack. The application constitutes its usages through feature box configurations. In the following, we employ the 3xx response handling and local recursion scenario – one of the most comprehensive call creation scenarios for a SIP proxy to illustrate the call processing layer design in detail.

3.2. The 3xx Handling and Local Recursion Scenario

When a softswitch receives a SIP request, it performs a sequence of tasks including:

(1) validating the request,

(2) preprocessing routing information,

(3) determining target(s) for the request,

(4) forwarding the request to each target, and

(5) processing all responses.

In the second and third steps, the softswitch obtains a set of target addresses from the location service based on the information in or about the request in the current environment, e.g., the content or presence of specific header fields and bodies, the request arrival time, and the interface at which the request arrived.

The softswitch can continue to add targets to the target set during request forwarding. New targets can be obtained from a redirect response (3xx), or from further consultation with a location service. A single target, however, cannot be added more than once. In step four, the softswitch selects a target with the highest *q-value* from the target set and sends down service request to the SIP stack.

The SIP stack sends a request in the following steps:

(1) update the peer address or request-URI with the target address;

(2) update or add additional header fields, e.g., decreasing the value of *Max-Forwards* header, adding *Route* headers to mandate a request to visit a set of specific servers, and adding a *Via* header in front of the existing *Via* headers;

(3) create a new client transaction for the request;

(4) set timer C to clear the transaction if the request never generates a final response.

The forwarded request would cause one or multiple responses sent back from the remote target. The response processing tasks often include:

(1) find the appropriate client transaction and response context;

(2) update timer C if received a provisional response;

(3) add the response to the response context, e.g., updating the target set if a 3xx response is received;

(4) immediately forward non-100 provisional response and 2xx response;

(5) for other responses, decide whether to terminate or continue local recursion, e.g., most often, 3xx responses, '402 Not Found', '502 Bad Gateway' or '503 Service Unavailable' will trigger local recursions;

(6) if no final response has been forwarded and no target remains in the target set, choose and forward the "best" response from the response context;

(7) perform other processes such as removing the topmost *Via* header for each forwarded response.

In step five, the softswitch attempts to contact each target in the target set until the call is either established or terminated. Other than the *initial targets* from the location service, *redirected* targets from 3xx responses and *re-routed* targets from further location service lookups continue to be merged into the target set. The 3xx response contains a contact list of the callee's current possible locations. If a prior forwarded INVITE request results a 3xx response, the softswitch will extract the contact list and transform them into redirect targets. If a redirected target contains the softswitch address in its host portion but a different user name from the original user name in the initial request, that is, a non-duplicate of the initial request, a second location service lookup is required for the redirect target. The returned targets are called re-routed targets and are merged into the target set.

3.3. Call Processing Design

In this section, we use an example of 3xx handling and local recursion to illustrate our call processing design. The discussion on related SIP stack message and transaction processing is postponed to the next section. Here, we assume that the SIP stack has successfully parsed a SIP message and created a new session.

On receiving a new session notification from the SIP stack, the application transaction manager decides whether to create a new application transaction or retrieve an existing transaction from the transaction hash table based on application context. In most cases, an application transaction corresponds to one session and one or multiple protocol transactions, and a new session notification often means an application transaction creation.

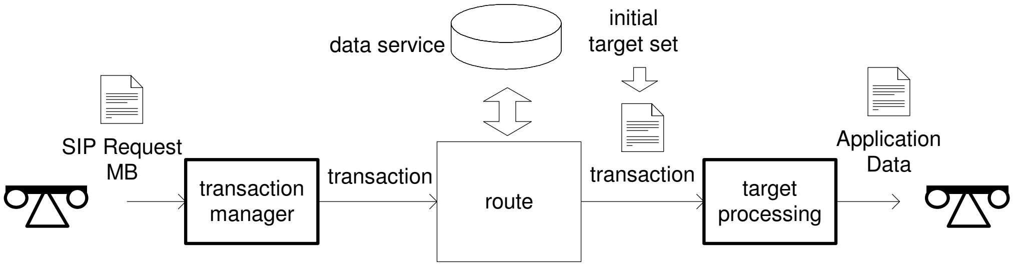

Fig. 9. SIP request processing.

In Fig. 9, the request is processed in three feature boxes. The transaction manager and target processing box are interfaces to the SIP stack and are required for each message processing sequence. The route feature box collaborates with the location service to translate the request URI into a set of target addresses and determines a set of calling features associated with the caller and callee based on pre-configured local policies or registration information that a callee registered earlier with the registrar. The route feature box inserts an initial target set and feature data into the transaction record, and hands control to the target-processing box, which fills in an application data structure with the first target address in the target set and issues forward request command to the SIP stack.

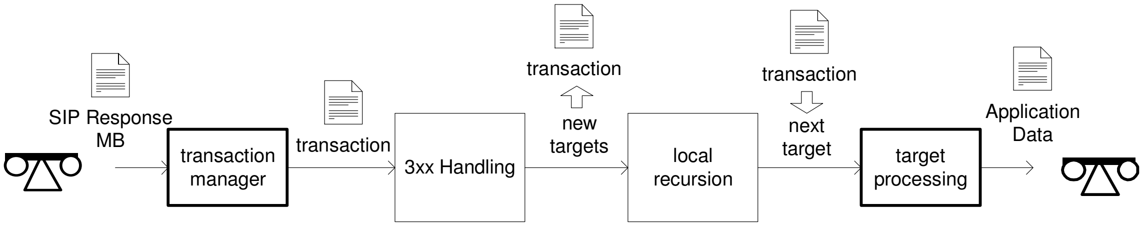

Fig. 10. SIP response and feature processing.

The forwarded request results in responses sent back from the target or in case of no response, a time out notification generated in the SIP stack. In Fig. 10, a 3xx response is received. Since we assumed that the caller subscribed to the 3xx handling and local recursion feature, after the transaction manager box retrieved the transaction record from the data service, it transfers control to the 3xx handling feature box, which extracts redirected targets from the *Contact* header of the message and adds them to the transaction target set. The control then moves to the local recursion feature box, where the response code is checked against a local recursion response list. Since 3xx is configured to be local recursion, the local recursion feature box selects the next unprocessed target with the highest precedence, e.g., a target with the largest *q-value*, and hands the control to the target-processing box.

The call processing repeats the above procedure until it receives a 2xx or a non-local recursion response, reaches the maximum request forwarding limit, or exhausts its target set, in which case the target processing picks the best response received so far from the transaction record and issues an ‘forward response’ command to the SIP stack (Fig. 11).

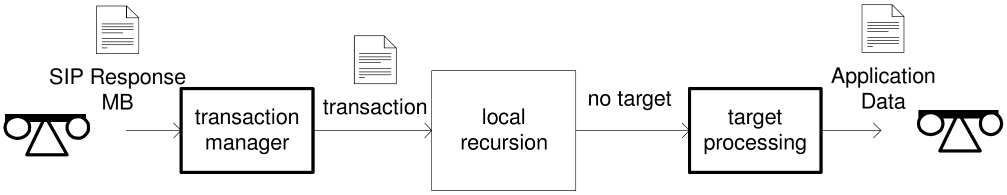

Fig. 11. Forward best response back to the request originator.

4. *SIP Stack Layer Design*

4.1. *SIP Stack Architecture*

In RFC 3261, SIP is specified as a layered protocol and the protocol behavior is described as a set of loosely coupled processing stages. The lowest layer, SIP parsing and encoding layer, decodes a SIP message into an internal data structure and encodes vice versa according to an augmented BNF grammar. The second layer, SIP transport layer, defines how a SIP client and server send and receive messages. It manages connections if a connection oriented protocol such as TCP or TLS is used and processes transport related message content, e.g., the *sent-by* and *received* parameter. The third layer, transaction layer, implements client and server SIP transaction state machines. It manages message retransmissions, matches responses to requests and executes several timers.

In the architectural view, the SIP parsing and encoding function is more properly abstracted as a utility component rather than a layer under the SIP transport layer. Since the SIP transport layer is responsible for receiving raw SIP packets, processing transport related SIP parameters and managing connections based on a parsed SIP message, placing parsing functions in a middle layer will cause flow of control to bounce back and forward between the parsing layer and the SIP transport layer, complicating the intra-layer message passing design. In addition, because the SIP transport layer will interact with both the SIP parsing service and IP transport service, and in systems that implement lazy syntax parsing, upper layers need to invoke the parsing service further, to design the SIP parsing and encoding service as an intermediate layer between IP transport layer and SIP transport layer will break good design conventions of a layered system in which each layer will only interact with at most the layers below and above. The limited interaction increases maintainability, as changes to the function of one layer affect at most two other layers.

As such, the software architecture does not have to match the protocol architecture. Compared with the protocol architecture that concentrates on functional specifications, the definition of software architecture cannot be limited to specifying the structure and functions of components; it also has to address different concerns of various stakeholders and non-functional requirements, such as implementation complicacy, extensibility and maintainability. The software architecture attributes have major effects on functional and non-functional qualities of the software. Fig. 12 shows the SIP stack architecture.

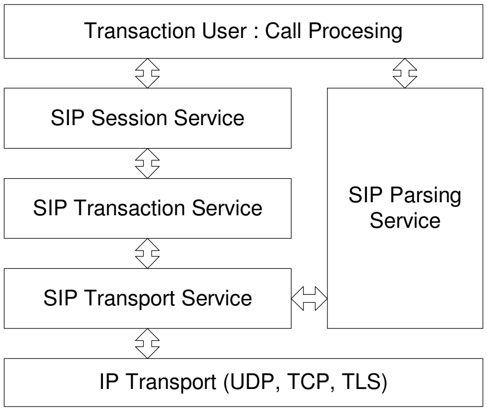

Fig. 12. SIP stack architecture.

4.2. SIP Stack Transport Service

The SIP transport service communicates with other SIP elements by invoking IP transport service primitives. An arriving message at a pre-configured port unblocks a thread, which allocates message space to hold the message and invokes the parsing service to transform the message content into an internal data structure, called message block stored in the data service. Once control returns from the parsing service with an index of the message block, the transport service thread triggers a chain of upstream services through tandem function calls, sending control upwards to the transaction user. In the downstream, the transport service calls the encoding service to transform the message block into a network packet and invokes the IP transport service to send out the packet.

The SIP transport service also manages network connections if connection-oriented protocols such as TCP and SCTP, or TLS over those are used. Since a TCP connection involves 3-way handshakes, and TLS involves expensive asymmetric key generation and authentication algorithms, the transport service ensures that responses to a request and multiple requests from the same originator reuse existing connections. The transport service uses the remote end IP address, port and transport protocol to index connections.

4.3. SIP Stack Parsing and Encoding Service

The SIP stack parsing and encoding service receives a SIP packet index from the SIP transport service and transforms the referenced packet into an internal message block structure. The message block contains elements that either point to NULL terminated strings or contain decimal representation of SIP header data.

The SIP stack parsing service architecture is similar to the front-end syntax analysis in the traditional compiler. It contains:

- Scanning component that groups input characters into tokens including (1) SIP keywords, e.g., SIP method keyword, header keyword, parameter names, (2) special characters, e.g., @, =, and (3) identifiers, decimals and strings.
- Parsing component that recognizes *sequences* of tokens according to the SIP message grammar and stores the SIP message content in the message block.
- Semantic analysis component that performs basic message validity checking, e.g., a Cseq header filed should not be larger than 999.

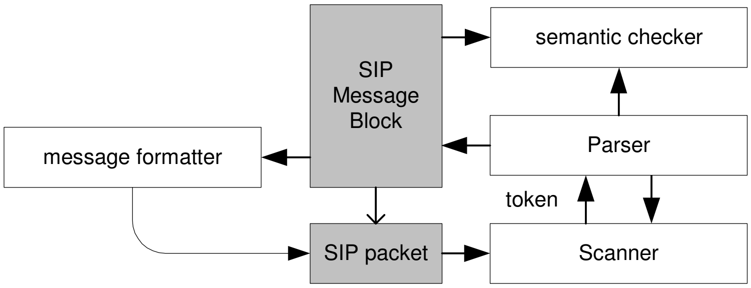

Fig. 13. SIP parsing and encoding architecture.

The parsing service architecture is shown in Fig. 13. The scanner performs pattern matching on input messages. It specifies token formats in regular expressions and corresponding actions when a specified token is met. Most often, the action just returns the token to the parser.

The parser collects tokens from the scanner and then matches them against SIP grammar rules given in RFC 3261 section 25. Once a grammar rule is matched, it triggers corresponding actions defined in executable program statements. The parser component drives the parsing service: whenever it needs a token, it issues a request to the scanner. Once the scanner reads sufficient characters from the input stream to construct a single token, it returns the token to the parser. The scanner then suspends until the next request coming from the parser.

In addition to syntax errors and grammar errors caught in the scanner and parser, the parser also invokes the semantic checker from time to time for semantic errors. Depending on the error attributes, the parsing service decides either to abandon or to continue processing. Sometimes it is important to skip errors to extract important information needed to form an error response message. Special *error* rules need to be incorporated into grammar reductions. The *error* rule enables the parser to abandon the current grammar rule and use the *error* rule when the parser recognizes faults. As such, the parser can continue processing new lines of input while skipping those that are illegally formatted. If an error rule is not located, the parser returns failure, since otherwise it will continue to call the scanner until it returns a match for the token defined after the error token, which is not desirable.

The formatter accepts requests from the SIP stack transport component and uses pre-defined SIP header templates to create a SIP message.

On the implementation, several GNU tools can be used to generate the scanner and parser. Fig. 14 shows a diagram to use FLEX, a GNU automatic scanner generator and Bison, a GNU parser generator to generate the scanner and parser.

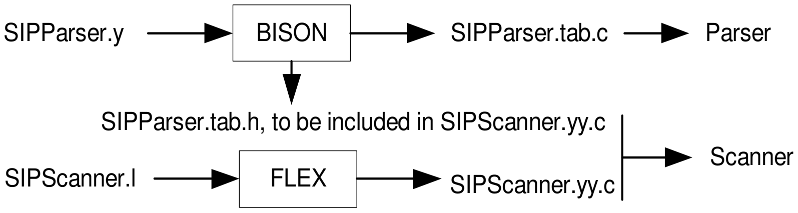

Fig. 14. Generate scanner and parse executables.

SIPScanner.l specifies token pattern and action pairs. The token pattern expressed in regular expressions describes tokens required identification in a SIP message, such as SIP method names, SIP headers and specific parameters. Once a match is determined, the text corresponding to the match is made available in the global character pointer yytext, and global length integer yyleng. The action corresponding to the matched pattern is also executed, which typically returns the token type (e.g., INTEGER), and if appropriate, returns the token value to the parser. The token types are declared in SIPParser.y and enumerated in SIPScanner.tab.h.

SIPParser.y specifies a Context Free LALR(1) grammar and associated actions. The generated parser parses input strings with a sequence of reduction and shift actions, taking a bottom-up parsing approach. As the parser retrieves tokens from the scanner and pushes them onto the stack, a reduction happens when the first few symbols at the top of the stack match the right-hand side of a certain rule. The matched symbols are then popped out from the stack and the left-hand side of the matched rule is pushed onto the stack. A shift happens when no handle is found, the parser continues to push the current token into the stack and read the next token. Code fragments are inserted in the right-hand side of the rule to perform actions with specific reduction and shift action. The most common actions are functions or macros that perform message checking or assign values to SIP message elements in the SIP message block.

The process continues until the entire message is traversed. The reduction and shift process stops when the parser derives the start symbol. To minimize the reduce-reduce conflict and shift-reduce conflict, LALR(1) looks one additional token ahead and matches it against the expected token set following each reduction rule.

Besides grammar rules and actions, SIPParser.y also contains utility functions that invoke lexical analyzer function yylex() defined in the scanner, the error reporting function yyerror(), and a main function that calls yyparse() to implement parsing.

When SIPParser.y is compiled, the command line –d option instructs BISON to generate header file SIPScanner.tab.h, which contains macro definitions for the token types defined in the grammar, the semantic value type `YYSTYPE` and a few extern variables. These definitions enable token collection function yylex() defined in SIPScanner.l to refer to token type codes and token semantic value `yyval`. The outputs of BISON and FLEX (SIPParser.tab.c and SIPScanner.yy.c) are regular C source code files that can be compiled by GNU *make* utility to generate the scanner and parser executable.

4.4. SIP Stack Transaction Service

SIP is a transaction based client and server protocol. As defined in RFC 3261, a SIP transaction consists of a single request and one or more responses to that request. On a user agent client, while receiving command from an application client to initiate a new SIP request, the stack creates a *client transaction* and enters the initial “calling” state. If an unreliable SIP transport is being used, the client transaction starts timer A with a value of T1 of 500 ms, or a time value that equals to the Round Trip Time (RTT) between the client and server transactions. When timer A fires, the client transaction retransmits the request and resets timer A with a new value 2*T1. A retransmission occurs if timer *A* fires 2*T1 seconds later. The process continues and the retransmission interval is doubled after each retransmission. These retransmissions are only done while the clients transaction is in the “calling” state and will terminate when timer B fires, which controls transaction timeout with a default value 64*T1. When the client transaction sends out an ACK, it enters the “complete” state and starts timer D, with a value of at least 32 seconds for unreliable transport. Any retransmissions of the final response will cause retransmission of ACK before timer D fires, which leads the transaction to the “terminated” state.

On the server side, upon receiving a request, the transaction service creates a *server transaction* to deliver the request to the application client. The transaction state is initiated from the “proceeding” state to “complete”, then to “confirmed” and terminated at the “terminated” state. During the transition, timer G, timer H and timer I controls retransmission and service timeout. The detailed transition logic is specified in RFC 3261 and will not be restated here.

The transaction service manages these client or server transactions and implements transaction state machines. Since the direct implementation of state machines is complicated, the solution domain has to be further decomposed. A simple and efficient solution of decomposition is to break the state machine into an upstream transaction service component that handles data traversing to the application, and a downstream component that handles data traversing to the network. Fig. 15 shows a decomposed client INVITE transaction and server INVITE transaction. In the figure, in addition to transaction management tasks such as transaction creation, retrieve and deletion, the identification of other upstream and downstream component functions is straightforward.

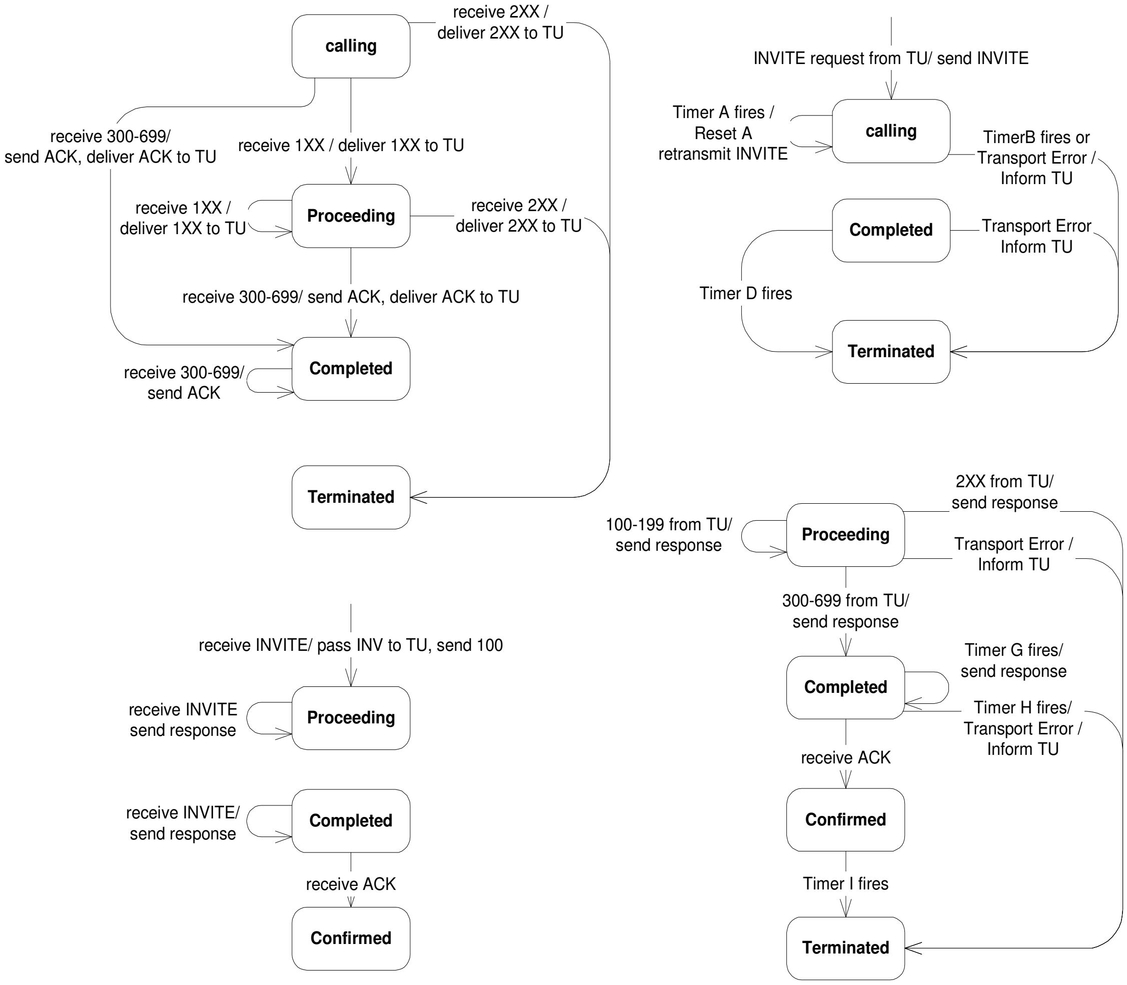

Fig. 15. Decomposed client and server INVITE transactions.

The upstream functions include:

- Accepting incoming SIP messages and delivering notifications to the transaction user.
- Issuing requests of sending responses or sending an ACK to the downstream counterpart.

- Validating SIP messages in a given transaction context, e.g., (1) ensuring a SIP message has a valid SIP method, SIP version, and IP address in the To and Request-URI header, and (2) identifying illegal cases such as receiving an ACK prior to sending a final response to an INVITE, and a BYE does not match any existing transactions.

The downstream functions include:

- Accepting request of sending messages from the transaction user or the upstream counterpart.
- Implementing various timers and sending messages as a timer fires.
- Monitoring the status of the transport service and informing the transaction user if error occurs.

The events that trigger the upstream state transition are the received SIP message, and the events that trigger the downstream state transition are either timers or requests of sending messages. In the upstream, after control is returned from the parsing service, the SIP transport service calls sipUpStreamTransServ(messageBlock *pMsg), an upstream transaction service dispatcher in the transaction service component, to invoke other upstream service functions including transaction creation and retrieve, calling transaction users to pass event notification or issuing a sending message request to the downstream counterpart. In the downstream, sipDownStreamTransServ(messageBlock *pMsg) serves as a downstream service dispatcher to receive requests from the transaction user or the upstream counterpart. It inserts transactions into the timer queue and invokes the transport service to send messages to the network.

5. Design with Patterns

In today’s time-to-market driven environment, systematic reuse of successful software designs and implementations breaks the expensive cycle of rediscovering, reinventing, and revalidating common software artifacts, ensuring efficient and cost-effective software development.

Crucial to the systematic reuse are patterns and frameworks. Design patterns abstract common static structure and dynamic interactions of communicating elements, solving a general problem in a particular context. Organized design patterns in a special domain forms a design pattern language. The pattern language for network applications such as online financial services, telecommunications, and remote access service often includes patterns that manage service component configuration, inter-process communication, event handling, concurrency and synchronization. Some well-known pattern languages include the enterprise application architecture design patterns [Fowler 2002], server application design patterns [Volter et al. 2002] and networked and concurrent computing patterns [Schmidt et al. 2000]. Some pattern languages are implemented as off the shelf software frameworks, capturing both the static structure and control mechanism of a special application domain.

In this section, we bridge the gap between the proposed architecture and implementation with design patterns by employing a framework for network applications called Adaptive Communication Environment (ACE).

As we have indicated in the previous section, the design and implementation of VoIP products involve constant change as new protocols and protocol extensions are proposed to accommodate new application requirements. This is even more prominent when developers choose to implement a subset of RFC 3261 at the initial development stage and then add more features and logic as the project develops. The feature boxes proposed in section 2 provide a simple solution. It, however, requires an open service configuration environment and message passing mechanism to integrate existing feature boxes into one or more service processes and incorporate future feature boxes. The service module configuration design and message passing design have to meet the following minimum requirements:

1. As new service modules and new feature extensions of existing modules are developed, the configuration and message parsing design should ease integration of new modules and facilitate encapsulation of changes within the new or modified service module.
2. As different customers impose different service requirements, the frameworks can load, initiate and configure selected service modules of an application.

5.1. Module Configuration Design

In [Jain and Schmidt 1997], a component configurator pattern was proposed to enable an application to load, initialize and manage its composing components dynamically at run time. Besides the minimum requirement described above, another advantage is that when the most effective solution for distributing service modules into processes and host machines is not known at the time the application is developed, and the configuration is subject to constant change, e.g., platform upgrades or rebalanced work loads of hosts and networks that require redistribution of certain modules to other processes and hosts, the configurator pattern can defer resource allocation until run time.

The pattern contains four classes: *component configurator, component repository, component* and *concrete component*. The class diagram of the component configurator pattern is shown in Fig. 16.

Class *Component* defines a uniform interface for configuring and controlling services implemented at each *concrete component*. Common control operations include service initialization, suspension, resumption, and termination. *Component configurator* reads and interprets *configuration file* and then configures/reconfigures *concrete components* of an application via *component repository*, which maintains and manages *concrete components* configured into the application at run time.

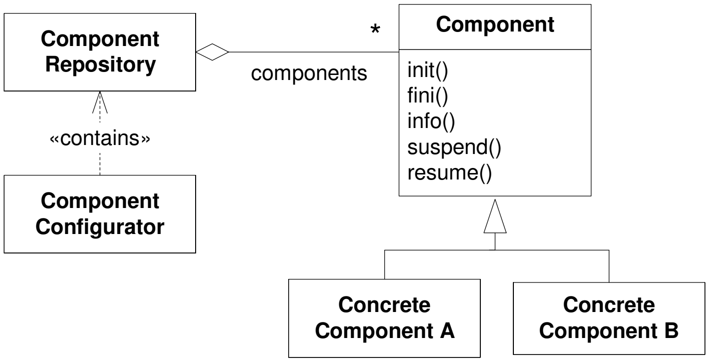

Fig. 16. The class diagram of the component configuration pattern.

The Adaptive Communication Environment (ACE) implements the component configurator pattern. Class *ACE_Service_Object* implements *component*. Class

ACE_Service_Config implements *component configurator*. The feature boxes can be implemented as *concrete component*, or service objects. The service objects override the following virtual functions of its parent class *ACE_Service_Object*:

1. `virtual int init(arguments …)`, which is called by an *ACE_Service_Config* object to initialize services. The arguments for the service initialization are passed via function arguments.
2. `virtual int fini()`, which is called by the *ACE_Service_Config* object to clean up allocated resources before terminating.

Invoked in the platform manager, an *ACE_Service_Config* function open() reads the service configuration file, and loads and executes initialization procedures encapsulated in the init() function of each service objects. The loaded service objects are then inserted into an *ACE_Service_Repository* object that implements *component repository*. The termination procedure is similar to the startup procedure. The dynamic diagram is shown in Fig. 17.

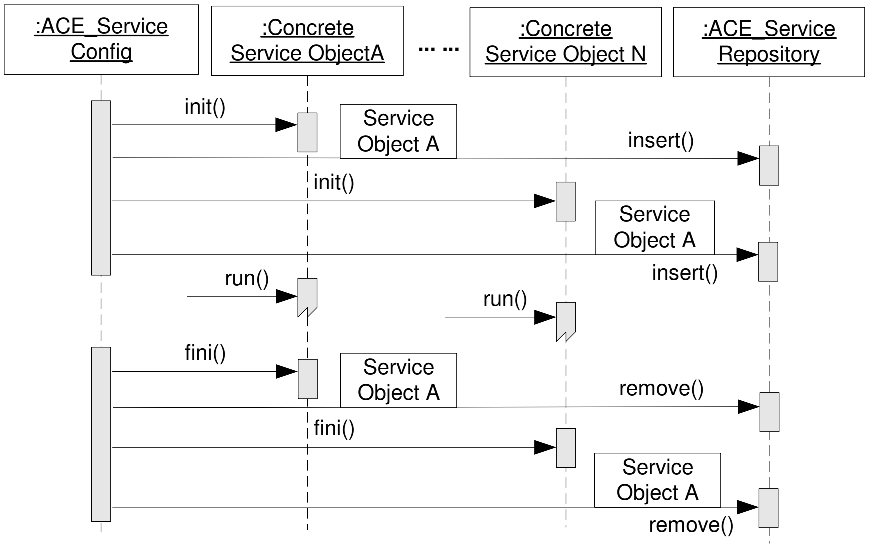

Fig. 17. Dynamic diagram of module configuration design.

5.2. *Message Passing Design*

The message passing mechanism assembles service objects into a stream to achieve message passing between them, even an object does not know which the next object is.

In [Shaw and Garlan 1996], the Pipes and Filters pattern was proposed to process a stream of data in a sequence of processing stages, with each processing stage encapsulated in an independent filter component. The pipe components move data between consecutive filter components. In some implementations, no explicit pipes exist and filters push data through direct function calls from the active to the passive component. In a multi-thread environment, a separate pipe mechanism provides queues, e.g., FIFO buffers, and synchronizes queue access among multiple filter threads. The Pipes and Filters pattern eliminates the need of intermediate files. The filter interface should be simple, which facilitates filter recombination, reuse and reducing data movement overhead. In a multi-thread environment, the pattern enables flexible thread configurations and facilitates distributing processing resources according to filter task size, which eliminates computation bottlenecks and achieves maximum performance.

In ACE, class *ACE_Task* implements the *filter* with an optional message queue. Class *ACE_Stream* implements the *pipe* that either calls the *ACE_Task* service function to pass service data as function arguments or inserts service data into the *ACE_Task* message queue. *ACE_Task* objects can be multithreaded to increase throughput. The use of message queues implies different thread parallel architectures and will be discussed in the section on process design.

In the class design phase, each feature box is abstracted as a service object - a subclass of *ACE_Task*, and a unidirectional stream is instantiated as a subclass of *ACE_Stream* to pass messages between service objects. The stream carries transaction information formed in the transaction manager in the downstream to the target processing feature box. In Fig. 18, a unidirectional stream connects service objects.

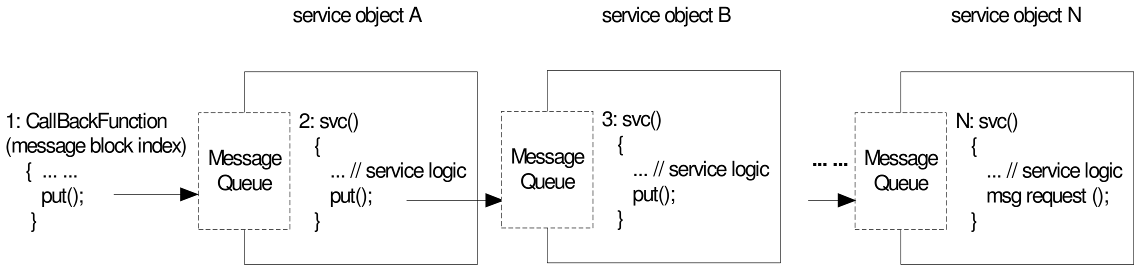

Fig. 18. A unidirectional stream connects service objects.

The integration of service objects and associated message stream is accomplished in the configuration file, which contains references of service object construction functions, enclosed in a reference to the stream object construction function. The order of service objects represents the message flow direction.

In Fig. 19, a stream *sipStream* is instantiated and will traverse service objects *sourceProxObject*, *routeObject* and *targetProxObject* in order.

```
stream dynamic sipStream STREAM * ./platformManager:makeSipStream()
{
  dynamic sourceProxObject Module *
       ./ platformManager::makesourceProxObject ()

  dynamic routeObject Module *
       ./platformManager:makeRouteObject()

  dynamic targetProxObject Module *
       ./platformManager:makeTargetProxObject()
}
```

Fig. 19. Code of an example configuration file.

5.3. *State Machine Design*

In the design, state machines are used in the SIP transaction service and feature boxes. Continuous feature evolution results in constant logical change to the state machine implementation. This is even more prominent when developers choose to implement a subset of RFC 3261 at the initial development stage, and add more actions and event handling logic as the project develops. Therefore, an efficient and flexible state machine design is critical to the overall project development and will significantly reduce its implementation and expansion effort. In the literature, several state machine design and implementation techniques exist, among others, including the nested switch case statements and state machine design pattern. In this section, we describe our experience with these state machine design and implementation techniques.

To facilitate our discussion, we use a simple state machine as an example. The state machine has two states – an initial state Off and a state On, three possible input events: sigOn, *sigOff* and *sigReset*, and three possible actions: *actionOn*, *actionOff* and *actionReset*.

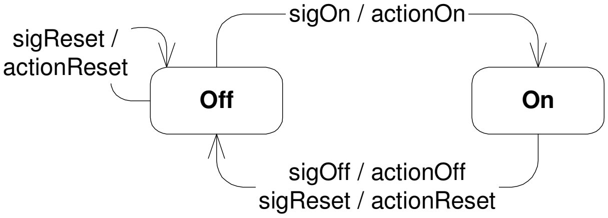

Fig. 20. A simple state machine.

The nested switch statements approach is the most common practice (pseudo-code shown in Fig. 20. It meets one of our most important implementation criteria - simple. In dealing with a large number of states and events, however, the nested switch/case statements become difficult to read and understand. The code often contains large conditional statements, similar to long procedures, which are undesirable [Gamma et al. 1995]. The code could extend for pages and it all looks the same [Martin 1998]. Fig. 21 shows the pseudo-code when we use switch/case statements to implement the example state machine depicted in Fig. 20.

```
enum State {On, Off};
enum Event {sigOn, sigOff};

void actionOn();
void actionOff();
void actionReset();

static State state = Off;

void stateMachine( event inputEvent )
{
 switch (state)
 {
   case Off:
   switch (inputEvent)
   {
     case sigOff:
       actionOn();
       state = On;
       break;
     case sigReset:
       actionReset();
       break;
     default:
       error input event;
   }

   case On:
   switch (inputEvent)
   {
     case sigOff:
       actionOff();
       state = Off;
       break;
     case sigReset:
       actionReset();
       state = Off;
       break;
     default:
       error input event;
   }
}
```

Fig. 21. A state machine implementation using switch/case statements.

It is then important to decompose the state machine into several relatively small state machines with fewer states and events. In the previous sections, we have introduced such techniques as decomposing a complex state machine into an upstream state machine and a downstream state machine, or into a sequence of processing stages or components, with each component handling a specific state.

In C++, we take an object-oriented approach and use the State pattern to implement a state machine. The State pattern offers a better solution to structure state specific code. The state transition logic is partitioned among the State subclasses, which makes its intent clearer [Gamma et al. 1995]. Fig. 22 shows the class diagram for the example state machine according to the State pattern.

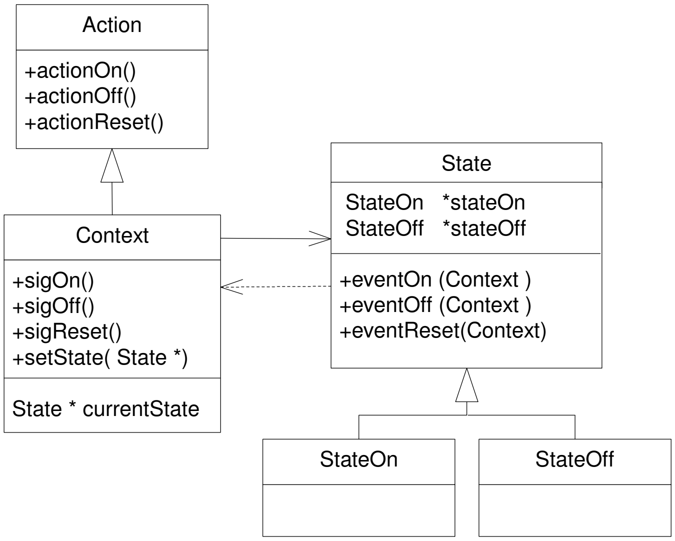

Fig. 22. State machine class diagram.

Class *Action* abstracts actions taken during state transitions. Class *State* declares a common interface to all sub-classes, each representing a specific state and encapsulating state-specific behavior. Class *Context* contains a private class *State* pointer - *currentState*. At run time, pointer *currentState* is downcast to a *State* subclass to indicate the current transaction state. Class *Context* also delegates all events to the current state object. Class *State* has two static pointers *stateOn* and *stateOff* that are pointing to its subclasses: *StateOn* and *StateOff*.

In the initial state *off*, pointer *currentState* points to class *StateOff*. The application passes *sigOn* event to the state machine through class *Context* member function *sigOn*, which in turn calls the *eventOn* function of *State*. At run time, the *StateOff* version *eventOn* function is executed (*polymorphism*). In function *eventOn*, a pointer that points to the static instance of *StateOn* is passed to function *setState* to set the new transaction state *StateOn*, and action function *actionOn* is also executed. The state machine algorithm is shown in Fig. 23.

```
class Action {
public:
virtual void actionOn();
virtual void actionOff();
virtual void actionReset();
};

class Context : public Action {
public:
void setState( State* s) { state = s;}
void sigOn() { state->eventOn(this);}
void sigOff() { state->eventOff(this);}
void sigReset() { state->eventReset(this);}
private:
State *currentState;
};

class State {
public:
virtual void eventOn(Context*) = 0;
virtual void eventOff(Context*) = 0;
virtual void eventReset(Context*) = 0;
protected:
static StateOff stateOff;
static StateOn stateOn;
};

class StateOff : public State {
public:
virtual void eventOn(Context* t) {
   t->setState(&stateOn);
   t->actionOn(); }
virtual void eventReset(Context* t) {
   t->actionReset(); }
};

class StateOn : public State
{
public:
virtual void eventOff(StateMachine* t) {
   t->setState(&stateOn);
   t->actionOn(); }
virtual void eventReset(Context* t) {
   t->setState(&stateOn);
   t->actionReset();
};
```

Fig. 23. A state machine implementation based on the state pattern.

The state pattern localizes state-specific behavior and partitions behavior into different state subclasses [Gamma et al. 1995]. A state subclass encloses all behavior associated with a particular state. Since all state specific code lives within a single subclass, new states and transaction logic can be added or modified when new subclasses of *State* are created and event-processing functions are modified. The state pattern also makes state transitions explicit. Compared to the case when an object uses internal data to represent internal state, its state transitions take the form of changing the state variable's value, which both lacks explicit representation and is prone to inconsistent internal states. Furthermore, scanning a State Transition Table (STT) or scripts written in Call Processing Language (CPL) [Martin, 1998] can automatically generate the above state machine code.

Despite the advantages, the use of design patterns imposes a longer learning curve for software engineers. Since part of the state transition logic is executed behind the Object Oriented polymorphism mechanism, it is more difficult for a software engineer to follow the program control flow.

6. Process Design

In most network applications, processes are designed to handle multiple requests at the same time. Although there are multiple ways to achieve this, multithreaded programming, which enables an application to take advantage of a multiprocessor environment, is the most common practice. In this section, we concentrate on the process design, which fits static structures and functional abstractions described in the previous sections into the process architecture [Kruchten 1995], addressing such issues as performance, concurrency, synchronization and distribution.

6.1. High Level Thread Parallel Architecture

In network applications, tasks are short-lived, e.g., a service thread receiving a message returns to the thread pool after forwarding the message to the next hop. Instead of spawning a new thread for each request, a server often pre-spawns a certain number of threads – a thread pool to handle incoming tasks. The thread pool avoids spawning a large number of threads when a server receives a large spurt of requests in a short period, which causes service degradation for all requests or resource allocation failure. In addition, it reduces overhead associated with getting a thread started and cleaning it up after it dies. In the thread pool model, when a request arrives, a thread is chosen from the thread pool to handle the request. If no thread is available when a request arrives, the request will be enqueued until one worker thread returns to the pool.

In designing the thread parallel architecture, two architectures can be used in our context:

- *Message parallel architectures*, which attach a separate thread to each incoming or outgoing message or both. Once the network transport component de-multiplexes messages and assigns a message to an available thread, the thread escorts the message through a sequence of protocol and application tasks.
- *Pipelined or layered parallel architectures*, which attach a separate thread to each architectural layer. Independent messages can be processed in parallel in

different layers. Buffering and flow control are needed if layers execute tasks at different speeds.

6.2. Low Level Synchronization Mechanisms

In the low level, since each thread has its own execution stack and registers, *automatic storage class* variables stored in the stack and *register storage class* variables stored in registers are private to each thread. *Static* and *extern static storage class* variables stored in heap and other single resources, such as message queues, are shared by multithreads. Synchronization methods are needed to protect the access to these shared resources. Among others, *Mutex* and *Condition* are the most often used methods.

Mutex ensures the integrity of the shared resource. It serializes execution of multiple threads by defining a critical section and limiting its access to one thread at a time. Mutex are often implemented via adaptive spin-locks. A thread waiting to enter the critical section waits in a loop that repeatedly checks the lock until the lock is explicitly released, or the thread that is holding the lock goes to sleep.

Condition allows one or more cooperating threads to suspend their execution until the shared data enters a particular state. In this process, a cooperating thread operating on the shared data signals other threads if the shared data has entered a particular state. One of waiting threads wakes up and re-evaluates the state and resumes processing if the shared data enters an appropriate state. Since a thread waiting on spin-locks involves busy-waiting and is not performing useful tasks, condition is implemented via sleep locks instead of spin locks to avoid the excessive resource consumption caused by a spin lock. The sleep lock, however, can trigger expensive context switches.

Condition is much more expensive than mutex because of context switches. On a SUN OS, the time to acquire and release a mutex object is about 4 μs when no other threads contend for the lock. In the same circumstances, using condition objects requires about 300 μs, almost two orders of magnitude more expensive than mutex objects [Schmidt and Suda 1994]. The results are similar in multiprocessor environments where multiple threads are contending for shared objects.

The low-level synchronization mechanisms cause different performances of these different thread parallel architectures. Compared with the message parallel architecture that only uses less expensive mutex, the pipelined parallel architecture uses both condition and mutex. Furthermore, excessive operations on message queues in the pipelined parallel architecture further downgrade the system performance. In this framework, we mostly use the message parallel architecture because of performance concerns. In case a feature box involves long execution time tasks, e.g., while consulting remote network elements, we separate the task into a different thread, and use the pipelined parallel architecture to enqueue unprocessed messages. The best performance can be achieved through quantitative simulation. The proposed framework enables flexible thread configurations, which facilitate simulation and postponing thread configuration to a later stage of product development.

6.3. Process View

In ACE, class *ACE_Task* and associated classes *ACE_Stream* and *ACE_Message_Queue* facilitate implementing both the pipelined and message parallel architecture. In the pipelined parallel architecture, a service object inheriting class *ACE_Task* overrides the base class function svc() with concrete service logic, among others, including monitoring and synchronized accessing of the associated queue *ACE_Message_Queue*. Upon instantiation and activation with service class member function activate(), a thread pool is created with a specified number of service threads, all executing on service function svc(), which synchronously retrieve messages from the message queue, process them, put process messages to the message queue of the next service object over the *ACE_Stream* and returns. In the message parallel architecture, other than the first service object, no service thread pool and queues are instantiated. The first service object directly calls the main service function of the next service object, and then the next. The control does not go back to the first service object until the message leaves the system.

In the process, the standard interface provided in ACE is used between service objects. Once the function call of put_next() transfers control to the ACE framework, the framework looks for the adjacent service object, and calls its standard interface function put(). Depending on the parallel architecture used, function put() will either call the main service function, or insert the message in the message queue. In the process, the order of service objects is specified in the configuration file rather than hard coded, which facilitates service objects re-configuration. A service object can pass messages to different service objects without changing message passing functions or recompiling modified source codes.

Fig. 24 shows a process design, in which threads are separated into (1) upstream thread, instantiated in the SIP transport service object, which carries messages to the application, and (2) downstream thread, instantiated in the route interface service objects, which carries application data to send out on the network. The rationale of splitting the route feature box into an upstream route interface and a downstream route interface is that route feature box implements a remote lookup of service data and often takes undetermined among of time.

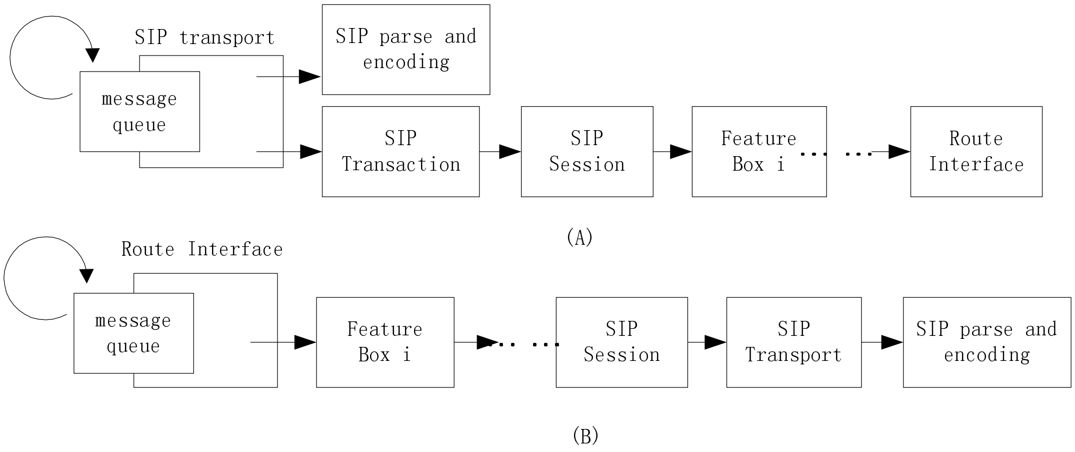

Fig. 24. The process architecture.

In the figure, the SIP transport service function svc() implements a loop that fetches messages from the message queue and then escorts the message upstream through a sequence of service objects via tandem function calls. The thread returns to the thread pool after sending out a data lookup request at the route interface. The message queue in the downstream route interface captures returned responses. In some cases, the thread may not go through the route interfaces. Depending on the message attributes, e.g., while processing a non-local request URI message, a transport thread can escort the message upstream and then downstream, and then generates an egress message.

7. Objects Design

In the object design phase, we summarize all the components described above and form the class diagram of our framework. The framework implements and integrates function modules using the ACE pattern language, which provides a generic architecture skeleton and program control flow. The class diagram is shown in Fig. 25.

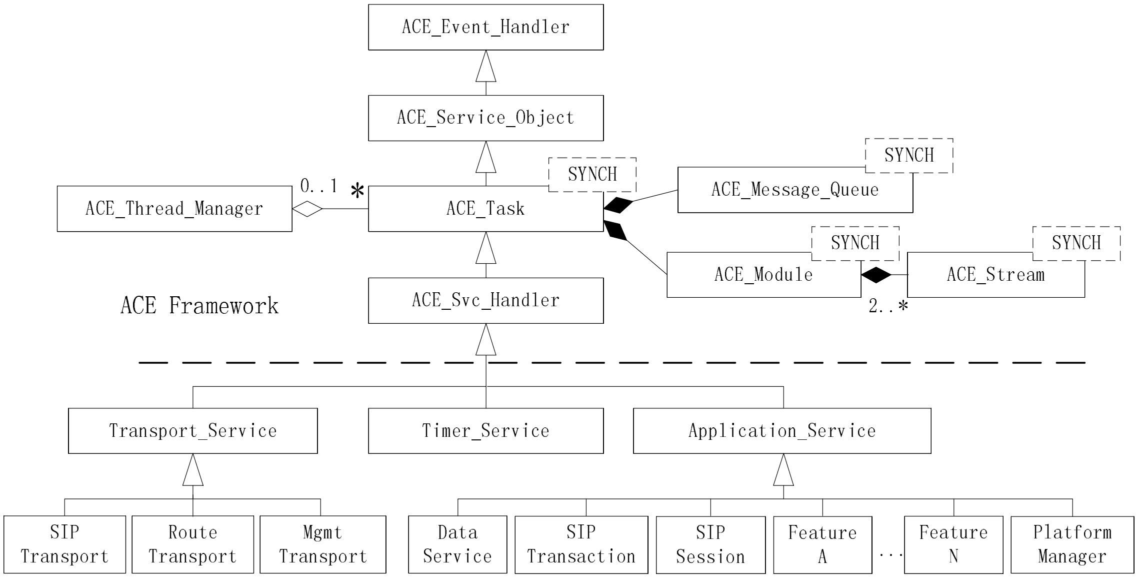

Fig. 25. The class diagram.

8. Related Work

In the literature, little work systematically describes a real system design and implementation from the ground up. In [Bond et al. 2004], BoxOS is proposed as a multimedia telecommunications network and infrastructure for creating services and validating feature interactions in large PSTN switches. The Distributed Feature Composition (DFC) feature components are called boxes, which can be assembled to create voice and signal paths via a virtual switch at the command of routers embedded in each box. Compared with our feature composition model, which concentrates on feature modularity, the DFC concentrates more on the flexible routing algorithm at the cost of complex design and predicted difficult maintenance. Since feature components are distributed in the network, the call setup time could cause performance problems for commercial products.

CHAPTER VII

THE DEVELOPMENT AND DIFFUSION OF VOIP

1. Introduction

In the past decade, VoIP companies have solved most technical problems. VoIP is moving into the mainstream to compete with traditional circuit-switched telecommunications systems [Doherty 2005]. The development and diffusion of technologies, however, can never be isolated from the social sphere. Just as Winston described in his invention model [Winston 2002], the law of suppression of radical potential[6] and the supervening social necessities[7] influence the development and diffusion of VoIP technologies.

Standardization, market forces and government regulation are the three most important factors that influence the development and market diffusion of VoIP technologies.

Standardization enables the emerging VoIP network to interconnect with the existing PSTN networks and ride on top of the Internet, conferring on VoIP users accumulated network externalities from PSTN users and Internet users, and abundant information and applications of the Internet. Standardization also enables worldwide production, which generates low cost and high quality products and services because of large economies of

[6] The pre-existing social formation suppresses the disruptive potential of inventions, e.g., existing market leaders attempt to suppress new market entrance with new technologies.

[7] Social factors that push the invention out into the world, causing its diffusion, e.g., customer needs, cost savings and pressure of revenue growth.

scale. As new service is a major driver for the adoption of next generation technologies, standards help third parties and service providers to develop and carve out new services, and seamlessly integrate with other systems. The intellectual property rights (IPR) problem in the standardization process is also circumvented.

Compared with the rapid development of standard technologies, the market development of VoIP has lagged behind several years because of economic uncertainties after the telecommunication downturn started in 2001. In 2004, along with the awakening of world telecommunications markets, several factors drove the VoIP adoption to a jump start: the widely deployed broadband network, the stabilized capital expenditure and shifted investment structure, carriers' needs of service differentiation and improving profit margins in the competition-intensified wireline market, and the fast market development in business sectors. The market, however, cannot develop at full speed. The existing investment in circuit switches, lack of killer applications, regulation uncertainties, capital barriers and potential virus attacks leads to the assumption that the circuit to packet migration will take a decade.

After the 1970s, the regulation of telecommunications has experienced fundamental change throughout the world. The government discarded the long embraced natural monopoly principle and started to promote competition. The regulation concentrated on such *economic terms* as enforcing interconnection and leasing rights to lower market entry barriers, prohibiting a monopoly from discriminating against an adjacent market to promote competition in the adjacent market for better technologies and services, changing price structure to enable markets to set prices based on demand and supply, and promoting competition and encouraging long term investment.

In the Internet age, the emerging VoIP market lacks dominant players. VoIP vendors and service providers are facing fierce competition from both inside the market and rival telecommunication platforms. The regulation of VoIP therefore concentrates on social factors, such as universal service, emergency service and CALEA[8] that to some extent are unrelated to competition. The competition also evokes regulation disparity for services provided on multiple platforms.

Since next generation wireless (3G) and wireline (VoIP) technologies influence each other and gradually converge into a unified multimedia service platform specification over IP transportation, and vendors and service providers intend to extend products and services to cover both wireless and wireline, we also extend our discussion to the standardization and market development of 3G.

The development and diffusion of the next generation network is still in its early stage and is filled with uncertainties. Due to different cultural, political and regulation backgrounds, different countries have followed different development and diffusion paths. Due to these differences, we concentrate our discussion within the United States. In the section on 3G market development, however, we examine the market of Japan, which distinguishes itself from the world next generation wireless market with a huge success in 2.5G and the first launched 3G service.

[8] Communications Assistance for Law Enforcement Act, which obliges telephone companies to aids law enforcement in its effort to tap phone conversations.

The rest of this chapter is organized as follows. Section 2 discusses the standardization of VoIP and 3G, and its critical role in the development and diffusion of VoIP technologies. Section 3 examines the market development of VoIP technologies and identifies market forces that drive the market adoption. Section 4 describes the current regulatory principles and development trend of VoIP regulation. Section 5 concludes this chapter.

2. *Standardization – Marking the Development Pace of Next Generation Technologies*

Disruptive technologies provide ample opportunities for both incumbent leaders and new entrants to redefine the telecommunications market. Just as Motorola translated itself as a new leader in wireless telecommunications with FM communication in the 1950s, Ericsson with digital switches in the 1970s, and Nokia with GSM technologies in the 1990s, now, companies that lead the development and standardization of VoIP technologies will lead the worldwide next generation telecommunications market.

Due to the interconnection requirement of network devices, telecommunications networks rely on technological standards. Standards mark the evolution path of telecommunications technologies. In the history of telecommunications standards development, standards are usually specified in three ways: (1) proprietary standards set primarily by the market, (2) open standards that are jointly developed by voluntary industry agreements, or (3) standards imposed by national or international standards development organizations [Neil et al. 2003]. In the following, we discuss the standardization of VoIP and 3G, and its critical role in the development and diffusion of next generation technologies.

2.1. VoIP Standardization

For most of the twentieth century, AT&T dominated the development of telecommunications standards. Under the traditional rate-of-return regulation, AT&T had strong incentives to invest in basic research because the cost of research expanded its revenue base. AT&T's legendary Bell Labs has developed technologies for all aspects of telecommunications services. In the 1980s, the International Telecommunications Union (ITU), a branch of the United Nations, adopted the modern "Signaling System 7" (SS7) based on Bell Labs' research. With Bell Systems pushing throughout the United States and the ITU promoting throughout the world, SS7 is used in most of today's circuit-switched networks, supporting call processing and advanced intelligent network features, such as call forwarding and caller ID.

After telecommunications entered the Internet age, Bell Labs, as a research branch of Lucent, was not able to hold the technology leadership. With a smaller basic research budget, the industry struggles to create a new model for standards development. The government only funds some early stage standards development activities, e.g., the Department of Defense funded the initial development of TCP/IP protocols. The standards development work has shifted to formal standardization organizations, such as ITU (International Telecommunication Union), and mostly, to vendor and service consortia, e.g., the Internet Engineering Task Force (IETF), IEEE (Institute of Electrical and Electronics Engineers), and 3GPP (3rd Generation Partnership Project). The IETF consists of large international communities of network designers, operators, vendors, and researchers who oversee the continuing evolution of TCP/IP, the creation and development of next generation Internet addressing scheme IPv6, quality of assurance techniques, and signaling standards for VoIP, paving the road for the next generation telecommunications.

The standards development process of IETF is specified in RFC 2026 [Bradner 1996]. A standard starts as an *Internet Draft* and is published on the IETF's website for public access. Any organization or person can submit an Internet draft for comments. It is a work in process and is subject to update, replacement or obsolescence. If it is not revised or recommended as a *Request for Comments (RFC)*, it will be removed from the IETF website. An RFC has to go through several steps to become a standard. In the first step, an RFC becomes a *proposed standard*. To achieve this, an RFC has to be stable, specific, complete, well understood and has drawn significant interest within the community. It does not have to be implemented and demonstrated, but the Internet Engineering Steering Group (IESG) does require demonstrations of certain mission critical protocols. In the second step, a proposed RFC standard becomes a *draft standard*. A draft standard requires at least two independent successful implementations and demonstrations of its interoperability. Any failed portion has to be removed. A draft standard thus has a high level of confidence on specification details. In the final step, a draft RFC standard becomes a *standard*. A standard is mature, stable and there are significant operation experiences supporting it. It is ready for implementation on a large scale.

2.2. 3G Standardization

Compared with the standardization of wireline communication, the wireless standardization followed a different path. In 1946, AT&T developed the first mobile telephone system. In 1947, Bell Lab scientist, D. H. Ring introduced the cellular concept. However, the U.S. was not able to retain and capitalize on the technological lead [John and Joel 2002]. Japan NTT launched the first commercial cellular mobile phone service in 1979, and the Nordic countries including Denmark, Finland, Iceland, Norway and Sweden did so in 1981. The Unite States lagged behind other developed countries in the 1970s and did not launch commercial service until 1983. The system developed and diffused during this period is called the first generation mobile communication system (1G). The system transmits voice signals through analog channels and is not efficient in spectrum utilization. It does not have roaming functions and communication channels are subject to eavesdropping.

The development and standardization of 2G started in 1980. It took almost eighteen years from its initial conception to its signification penetration in 1998 [Dave et al. 2003]. The system offers secure digital voice and messaging services and makes more efficient use of the available spectrum. GSM (Global System for Mobile communication) is the dominant 2G standard that accounts for more than 60% of all second-generation systems. Other standards such as CDMA and iDEN deployed in the United States, and PDC and CDMA deployed in Japan account for less than 12% of all second-generation systems. The current system is highly optimized for delivering voice service and is difficult to upgrade [Dave et al. 2003].

The definition of 3G system has been continuously evolving since it was first conceived in 1986 by ITU as FLMTS (Future Land Mobile Telecommunication System), which defined a single worldwide system intended to replace all second generation standards, to converge wireless and wireline communication systems and to merge the American, European and Japanese telecommunication standards. Customers can reach the information super highway with a single handset anywhere from Europe to America to Asia. The specification has evolved into IMT-2000 (International Mobile Telecommunications-2000), a framework for worldwide multimedia telecommunication specifications covering air interface, spectrum, bandwidths and services.

In Europe, the standardization of 3G or UMTS (Universal Mobile Telecommunications System) was carried out within various ETSI GSM working groups. Wideband Code Division Multiple Access (W-CDMA) was adopted as the 3G air interface standard, which specifies a pair of 5MHz channels, one in the 1900 MHz range for uplink and one in the 2000 MHz range for the downlink. Although discussions took place between the ETSI and the United States ANSI-41 community with a view to specify a unified standard for all ITU members, in the end, since the UMTS designated spectrums, the 1900 MHz range is used for 2G (PCS) services, and the 2100 MHz range is used for satellite communications in the United States, and due to concerns about technology evaluation for the ANSI-41 system, ITU recognized the CDMA2000 standards including CDMA2000 1x, CDMA2000 1xEVDO under IMT-2000, corresponding to the W-CDMA standard in Europe. [Dave et al. 2003].

Although the air interface standards are all new, the specification of the 3G core network has deviated from its original 3G vision and moved closer to the existing 2G network. Instead of replacing the whole 2G network, the 2G operators and manufacturers desire an evolutionary approach [Dave et al. 2003]. The success of 2G and the recent telecommunication downturn have led the operators and manufacturers to reuse as much existing equipment, development effort, and services as possible. Therefore, the specifications for UMTS are naturally based on the GSM core and CDMA2000, the evolved ANSI-41 core [.

After 1998, two bodies – 3GPP and 3GPP2 take over the standardization work of ITU, while ITU shifts its work to harmonize the 3GPP and 3GPP2 concepts. 3GPP (Third Generation Partnership Project) includes five SDOs (standards development organizations) - ETSI (EU), ATIS (US), ARIB and TTC (Japan), TTA (Korea) and CCSA (China). The group produced UMTS specifications and reports based on evolved GSM core networks and the radio access technologies based on W-CDMA [3GPP 2007]. 3GPP2 is the parallel partnership project of 3GPP. It also includes five SDOs – CCSA (China), TIA (North America), TTA (Korea), ARIB and TTC (Japan). The group specified CDMA2000 standards based on an evolved cdmaOne system and using an evolved ANSI-41 network core [3GPP2 2007].

In March 2003 and March 2005, 3GPP issued the UMTS Release 5 and the UMTS Release 6 3G standards, which advance the wireless network toward a full implementation of the 3G vision. The next generation wireless network solution is divided into three parts: (1) terminals, (2) the access networks, including both radio and non-radio wireline broadband access, and (3) the core network serving both wireless and wireline traffic. Among them, the core network, called the IP Multimedia Subsystem (IMS), defines standard functions and interfaces (see Appendix A) based on the IETF Session Initiation Protocol (SIP). It addresses such networks and user requirements as (1) real-time IP-based multimedia person-to-person communications (e.g., voice or video telephony) and person-to-machine communications (e.g., gaming service), (2) integrated real-time and non real-time multimedia communications (e.g., live streaming and chat), (3) high interactivity (e.g., combined use of presence and instant messaging), and (4) simple user setups of multiple services in a single session or multiple simultaneous synchronized sessions [Sonus Networks 2006].

As fixed broadband network and services such as transactions, content distribution, and VoIP over all-IP networks continue to spread, the IMS is extended to cover the next-generation wireline networks to provide a unified architecture converging both wireless and wireline communications. The IMS thus adopts a range of wireless and fixed access technologies, which enable it to support IP-based services over both packet-switched and circuit-switched networks, and both wireless and wireline networks.

2.3. The Critical Role of Standardization

The standardization is critical to the development and diffusion of the next generation network. First, since standards enable interconnection and convergence of wireless networks, wireline networks and the Internet, the next generation users can acquire the accumulated network value from each network and benefit from the integrated multimedia communications.

Although the Internet is endowed with abundant information and exciting applications, since the traditional public switched telephone network was created and optimized to support voice, it provides limited signal mechanism and media processing capability to support multimedia communications such as e-mail, instant message and video. Standard-based next generation technologies support the integration of these multimedia communication methods, e.g., a video conference application integrating voice, video, and shared whiteboard sessions, and enable interconnection with different communication platforms, conferring VoIP users the same network externalities as the PSTN and Internet users. As the wireless network, wireline network and Internet converge into a single network, the number of network components is also reduced, leading to a low-cost network infrastructure.

Second, as price is the key criterion for residential customers to adopt next generation services, standardization, an important factor that leads to low cost and high quality products and services, makes wide adoption of the next generation network possible.

In the history of wireless networks, when a variety of the first generation wireless systems were deployed around Europe and the rest of the world, the network equipment vendors developed products following their own favor. Switch suppliers offered everything from processors, memories, cables, to application software and even the racks. The interoperability between systems of different vendors was poor and plugging in new applications was hard for third parties. As replacing existing equipment involved huge cost and effort, a service provider was often forced to stick with a given vendor from the beginning. The network infrastructure and handsets were expensive. Wireless communications were regarded as the prestige product of the upper classes. The creation of the GSM standard changed this situation [Alan 2001]. The universal standard enabled large production around the world, which reduced the development and manufacturing cost because of economies of scale. Operators can choose from different network infrastructure vendors and users can switch to different service providers while retaining similar service. The competition among infrastructure vendors and service providers further stimulates higher quality and lower price products and services. The standards make roaming possible and ease the regulators task in specifying spectrum use, giving a controlled, yet open and competitive basis for licensing. The success of the GSM network and GSM manufacturers [9] are in large part due to the creation of international standards.

Third, standardization facilitates fast development of new services - the key driver for the adoption of next generation networks.

As competition intensifies, new services will help service providers both differentiate themselves and bring in new revenue. Other than the leading applications such as audio and video conferencing in the wireline network, and instant messaging in the wireless network, which help in bringing in additional revenue from use of these services, service providers often need to develop or customize applications to meet specific service requirements. On the development of new next generation services, most service

[9] The GSM network accounts for 60% of the total mobile communication market and 80% of the European market. GSM manufactures such as Ericsson, Nokia, Motorola and GSM operators such as T-mobile, China Telecom and British Telecom are top companies in financial market capitalization.

providers use features integrated with the equipment, some use service creation environments[10], some use internal programmers, and some use independent software vendors who deliver services independent of the traffic carrier. To achieve seamless system integration, the vendor needs to process open application programming interfaces (APIs) and follow universal standards.

Fourth, standardization enables carriers, especially incumbents to build the next generation infrastructure with multiple vendors for such concerns as reliability, less dependence on a specific vendor, long term availability of support and promising future product offerings. The market is young and no vendor can purport to have it all. Carriers are looking to mix and match suppliers to obtain best of breed elements or use specific vendors for particular applications. The product selection will accelerate the product evolution and the high-quality service resulted will accelerate the market adoption of next generation networks.

Fifth, there is no evidence that patent issues in the standardization hinder the development of next generation technologies.

The patent for the telephone issued in 1876 broke the rapid development of the telecommunications industry for almost twenty years. During the ten years after Bell's patent expired in 1894, more than six thousand independent telephone companies went into business in the United States, and the number of telephones boomed from 285,000 to 3,317,000 [Winston 2002]. Today, to reward and keep firms that developed priority standards in the standards-setting fold, standardization organizations commonly endorse

[10] A software platform with drag and drop interface of basic service building blocks for new service development.

standards containing patented, priority technologies, if any such technology is made available on “reasonable and non-discriminatory” terms. This approach helps to keep the industry leaders on board and increases the likelihood that the standard will be adopted widely. While designing GSM standards, although a large amount of ‘essential Intellectual property rights were inevitable, standard bodies tried to avoid the situation in which a single IPR holder impeded or even blocked the standards development [Rudi et al. 2002]. IPR problems have also been circumvented by negotiations. Companies that took part in cross-license agreements dominate the market for GSM infrastructures and terminals. The largest GSM IPR holders, including Motorola, Ericsson and Siemens also work with other parties either because their IPRs are valuable, or because their product lines are complementary.

Although standardization facilitates the development and diffusion of the next generation network, problems and difficulties exist. As we have seen from the standardization process of 3G – whereas in the transition from first to second generation starting with a clean sheet of paper had been possible, for 3G, there is a stronger commercial drive to reuse the existing infrastructure and take an evolutionary approach [Park and Chang 2004]. Europe and North America have already taken divergent approaches toward standards, which spoiled the original goal of a worldwide 3G standard. In Europe, the EC has mandated W-CDMA in the 3G band while in North America and Asia, the approach has been to allow the market to choose between W-CDMA and CDMA2000 [Neil et al. 2003].

While GSM is a largely self-contained standard, the next generation technologies draw various component standards from other standards bodies, e.g., the IMS adopts IETF protocols and standards to embrace data and the Internet; European Telecommunication Standards Institute's (ETSI) Telecom and Internet Converged Services and Protocols for Advances Networks (TISPAN) proposed to use IMS functions as the "core" of the next generation wireline network. Special coordination and collaboration are needed across different standards forums. The work on a world scale has also introduced wider political and culture influences on the standardization work.

3. The VoIP Market – a Star Market in the Ascendant

Along with worldwide liberalization of telecommunications markets and the development of new communications technologies, new services are being offered to potential customers at an increasing rate. Customer needs, revenue pressure on carriers and vendors, competitive and economic environment, and social culture serve as market forces that create uncertainty, threats and opportunities, influencing the development and diffusion of the next generation network. In this section, with the help of several market frameworks including the product life cycle model, product features model and market segmentation model, we explain the evolution of the VoIP market, examine market drivers and obstacles in the current VoIP adoption stage, survey the current VoIP marketplace, and predict the direction and trend of the VoIP business. In the last section, we briefly analyze the 3G value chain and the worldwide 3G market development.

3.1. The Evolution of the VoIP Market

Most marketers believe that all products are subject to life cycles, just as all creatures have biological cycles. Similar to a creature that progresses from birth to growth and to decline, a product life cycle has four major stages: introduction, growth, maturity, and decline. As a product moves through this cycle, basic market characteristics, e.g., competitions, market segmentations, regulations and technologies also evolve with certain patterns. Market players adjust marketing strategies on product, promotion, distribution and pricing[11] with the product life cycle. In this section, we use the product life cycle model to analyze the evolution of the VoIP market.

In 1994, VocalTec pioneered the first voice communication between PCs executing the same software over the Internet. The development and diffusion of VoIP entered the introduction stage. In the following decade, the VoIP market development has experienced the acceleration in the Millennium Internet boom and the vacillation suffered from economic uncertainties after the telecommunications downturn. Although most carriers have projected significant steps toward VoIP, there were no significant actions taken until 2004. "The technology is growing to maturity, vision is adopted and service providers proceed with caution" were the themes of VoIP during this period [Mitchell 2004].

The earliest service providers including Dialpad and Net2phone offered free PC-to-PC services and low-price national and international PC-to-PSTN services. Due to low

[11] The so-called marketing mix, factors that a company can typically influence when marketers develop marketing strategies.

Internet speed, limited voice processing and transmission technologies over the Internet, e.g., QoS, and lack of universal standards to support intercommunications between the PSTN and the IP network, these VoIP products could not deliver satisfactory voice communications and divorce from the PC platform. The business value created by VoIP was limited. The market development concentrated on product awareness and education. Customers were limited to specific groups such as students, foreign workers, and Internet enthusiasts.

Since the late 1990s, the rapid development of the Internet infrastructure, technologies and services such as expanded fiber backbones, various high-speed access mediums including Cable Modem, DSL and wireless, and fast development of VoIP standards under the ITU and IETF, have paved the road for the accelerated adoption of VoIP.

In the telecommunication market, equipment providers such as Cisco, Nortel and Lucent each develop one or several VoIP product lines, covering from media processing servers, call control servers to IP phones and upgraded network infrastructures. Many small firms concentrate on certain special components, such as various application servers. In the service market, the PC software market took the initial lead in building a mass customer base. Yahoo, Microsoft MSN and AOL messenger provide online conversation functions. Skype, a free peer-to-peer based VoIP software has recorded more than 100 million downloads of its latest release. After years of staggering from a distance, the carrier VoIP market finally made a big breakthrough in 2004. The worldwide carrier VoIP investment surged to $1.73 billion, a strong 37% increase over

2003 [Infonetics Research 2005a]. Carriers were beginning to treat VoIP as a serious strategic element in their long-range plans.

In the United States, Vonage Holdings Corp. announced that it had reached more than 500,000 subscriber lines in March 2005. The corporation had only 75,000 subscribers at the end of 2003. In March 2005, the new subscriber sign up rate had risen to 15,000 per week comparing with 10,000 per week in the fourth quarter of 2004 [Reuters 2005].

After Vonage, major cable companies such as Time Warner, Comcast and Cablevision market VoIP as a part of a voice, data and cable TV service bundle. Cablevision alone had more than 160,000 customers in the New York area within several months, and new installations have reached more than 3,400 per week [Harris 2004].

The active involvement of multiple forces of technology and service development, enhanced competition and sharp increase of investment and revenue indicated that VoIP had entered the rapid growth stage in 2004. The accelerated adoption of VoIP technologies and enhanced competition among service providers and vendors are expected to continue throughout the rapid growth stage, as customers migrate from circuit-switched technologies to VoIP technologies, which is expected to peak in the period from 2010 to 2014 [In Stat Research 2005a]. At the beginning of this period, most IP telephony projects will revolve around cost savings. For example, with VoIP technologies, global enterprises with extensive private voice networks can avoid international tariffs and realize greater savings on global destinations. However, VoIP has far greater capabilities. Applications such as multimedia conferencing and unified messaging built in the VoIP solution will enhance internal communications, business processes and customer relationship management. The business value created by VoIP will shift from the initial cost saving to the value of upper level applications, as it more securely and more maturely integrates with data services.

As the market develops, more service providers will enter the market to provide different VoIP services based on different business models. Intensified competition will further drive price down as participants focus on pursuing mass markets to increase market share and become low-cost producers, which will further accelerate the adoption of VoIP. Market strategies will move from education and awareness of the introduction stage to product differentiation. In the end, as the low profit margin cannot sustain providers and vendors with small economies of scale, both the VoIP service provider market and VoIP vendors market will consolidate with a few dominant players dividing the market, and the rest will be either acquired or driven out of business. The estimated product life cycle of VoIP is shown in Fig. 26.

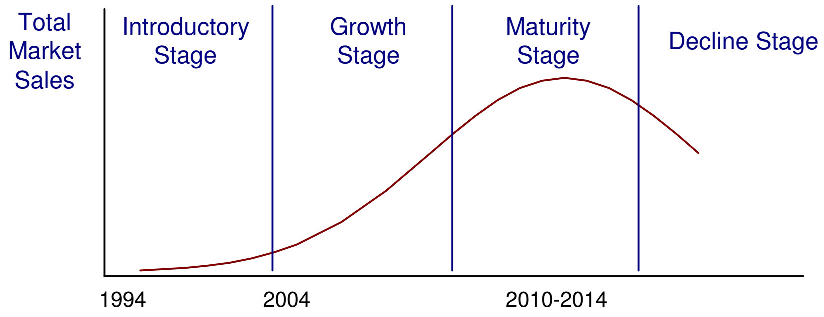

Fig. 26. The four stages of the VoIP life cycle.

3.2. VoIP Market Drivers and Obstacles

In the VoIP adoption process, multiple market forces such as the customer need, revenue pressure on carriers and vendors, competitive and economic environment, and maturing technology and standards act as market drivers that influence the adoption of VoIP. At different adoption stages, each driver has different roles and influences the market with different strength. After the product cycle analysis of VoIP, we now narrow our discussion scope by concentrating on the market environment in the current adoption stage. In 2004, VoIP experienced the first rapid growth in the worldwide market development. Some drivers, while existing ever since the introduction stage, act much more actively in this period, changing the landscape of VoIP. Next, we examine six major market drivers and also pinpoint the obstacles that hinder the adoption of VoIP. Note that most of them will continue to influence the VoIP market development in the future.

First, as 2004 approached, VoIP technologies and products became mature, pushing the VoIP adoption to the next level.

Standards were set in place, enabling products of different vendors, and products of different network domains including the PSTN to interoperate. Fig. 27 shows the accumulated number of pages of RFCs related to SIP or VoIP published by working groups including Audio/Video Transport (avt), sipping (Session Initiation Proposal Investigation), IP Telephony (iptel), SIP for Instant Messaging and Presence Leveraging Extensions (simple), Session Initiation Protocol (sip) and Telephone Number Mapping (enum). By the end of 2001, 974 pages of VoIP related RFCs had been published.

Almost half of them (464 pages) were developed in the avt group that related to real-time transmission of audio and video over UDP/IP. By the end of 2005, the accumulated number of pages had reached 3988 pages. The number of pages of VoIP related RFCs published from 2002 to 2005 is triple that published in the previous six years. The curve is similar to the product life cycle curve with a little advance in time, which partially points out that while market trend leads the direction of technology development, the technology drives the market adoption. Given that the number of pages of RFCs largely represents the technological maturity of VoIP, the figure partially explains why VoIP cannot take off earlier in the Millennium Internet boom.

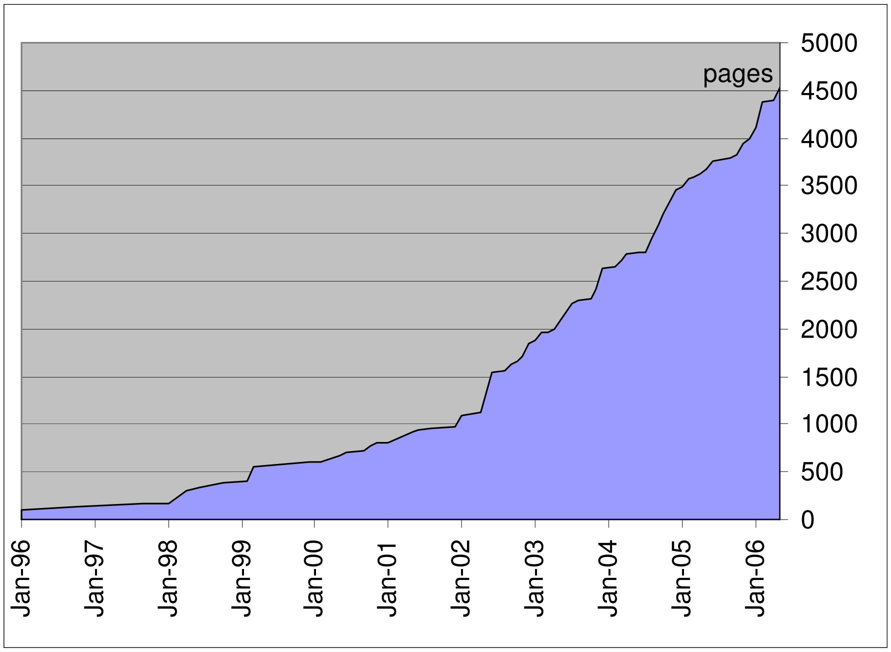

Fig. 27. The number of pages VoIP related RFCs.

The next generation products evolve into business class products that support more business activities. In the past, a service provider often chose proven products; first available untried solutions, even with cutting edge technologies, are not desirable. The service provider cannot afford downtime and lack of fault resilience. After continuous development and testing in the field, though still new to the circuit-switched equipment, several product lines have passed the learning and development curve and become reliable and feature abundant, qualified to be put into live networks to replace the traditional switch or deploy at new locations.

Second, due to increased adoption of broadband service, potential needs of promising bandwidth intensive services, emerging access methods and less regulation uncertainty, broadband network infrastructure growth is accelerated, laying the fundamental ground for the VoIP diffusion. By the end of 2004, the worldwide DSL subscriptions had increased 70% from 2003 to 99 million, and the Cable subscriptions had increased 22% to 41 million [Infonetics Research 2005b]. Besides the rapidly growing customer base of broadband service, promising services, e.g., IPTV service, a potential portfolio of IP-based, TV centric service that intends to deliver TV, video on demand, HDTV and videophone as a service bundle, became major drivers for massive broadband investment. On the regulation side, the FCC exempted Regional Bell Operating Companies (RBOCs) from obligations to unbundle the fiber network infrastructure and share with rivals. The decision eliminated regulation uncertainties and spurred investments in new fiber optic networks capable of providing data, video and voice services to consumers, paving the road for RBOCs to more vigorously compete with Cable modem services. In 2004, Verizon delivered fiber to 1.5 million homes. In 2005, this number was doubled to 3 million homes [Reardon 2005]. Incumbent Local Exchange Carriers (ILECs) and Competing Local Exchange Carriers (CLECs) together now have delivered Fiber To The Home (FTTH) to more than 398 communities in 43 states in the United States [FTTH Council 2005].

In the aspect of new access methods, new broadband wireless technologies such as WiMAX (IEEE 802.16) and WiFi (IEEE 802.11) create new service possibilities and strategic opportunities for service providers. Competition compels wireless service providers to plan on intensive investments in wireless broadband network expansion.

Third, as competition intensified in the current wireline market, the next generation voice infrastructure facilitates carriers to achieve future service differentiation.

After decades of market development, the wireline market has entered the maturity stage. The market is saturated, and the sales growth rate and profit margin have decreased. For most service providers, as local service providers, long-distance carriers, and cable companies each watch others' businesses and customers, the top line revenue growth and profit margin enhancement will be difficult to achieve in the foreseeable future, even for well-positioned ILECs, whose core businesses are eroding due to competition and technology substitution [Mitchell 2004]. To stem losses, service providers are compelled to provide differentiated service in hopes that it would leverage the market share in the shrinking market. The open and standards-based architecture of next generation voice enable service providers and Impendent Software Vendors (ISVs) to design services to achieve competitive service differentiation. In the current market, most service providers see the promise of new services as a major driver and reason to deploy VoIP equipment.

Fourth, the stabilized capital expenditure and shifted investment structure enable carriers to invest in the next generation infrastructure. In 2002, when large cutbacks in capital expenditures and sustaining operations were the norms, capital expenditures were tied to getting new customers and limited to small network rollouts. Although most carriers have projected significant steps toward the next generation voice network, the actual execution of the plan suffered from economic uncertainties. In 2004, capital expenditure budgets were not slashed further and showed the first return to growth since the 2001 telecommunication downturn. In the worldwide market, capital expenditures increased at 9% in 2004 to $161 billion. In 2005, capital expenditures continued to grow at 6% to $174 billion. Capital expenditures are expected to continue to increase with similar trends in 2006. In North America, capital expenditure growth is slower, 1% in 2004 and 6% in 2005, reaching $61 billion [Infonetics Research 2006]. Capital expenditure categories, however, are shifting from traditional time division multiplexed (TDM) products to IP-based products, focusing on areas like mobile wireless, DSL, VoIP, IPTV and enterprise data services. The worldwide service provider VoIP equipment revenue totaled $1.73 billion in 2004, a strong 37% gain over 2003, with half the market in North America and over a quarter from Asia Pacific [Infonetics Research 2005a]. Carriers were beginning to treat VoIP as a serious strategic element in their long-range plans.

Fifth, the next generation network meets carriers' needs of improving profit margins. As competition intensifies and margin slips, and savings are not coming from further capital expenditure cuts, carriers need to lower their total cost of network ownership. The packet voice promises lower operation costs. Carriers have shifted spending from circuit-switched networks to next generation IP-based networks. Such new equipment enables service consolidation, reduces the number of networks and creates the foundation for additional and margin-rich services. The next generation positive capital spending combined with the need for operation cost reduction drives the adoption of the next generation network.

Finally, that business customers are quickly adopting the next generation voice network, pushing carriers to capture the market that they would otherwise lose [Mitchell 2004]. In North America, 29% of large, 16% of medium, and 4% of small organizations had adopted VoIP by the end of 2005 [Infonetics Research 2005c]. The business market is still at the beginning of the high growth stage and is expected to grow at a CAGR (Compound Annual Growth Rate) of 21% until 2010 [Juniper Research 2006]. Businesses of all types and sizes are starting to evaluate the merits of the VoIP platform. If a carrier does not act to catch the market, it may lose it, as these businesses will either roll it out themselves or go to a competitor that would roll it out for them.

Despite the accelerated adoption of next generation networks, it is not a revolution. Obstacles exist and hinder the development and diffusion of VoIP.

On the supplier side, although most service providers acknowledge that the voice network is moving to packet technologies, the existing investment in circuit-switched equipments dampens the urge to migrate. For most existing voice carriers, especially for tier one and tier two service providers, such as RBOCs, and Incumbent Exchange Carriers (IXCs), VoIP does not represent a new revenue stream, but a replacement for the circuit-switched network. As a result, most VoIP equipment is adopted when carriers need more capacity or expands to a new location. The complete circuit to packet migration for Class 5 switches is expected to take at least ten to fifteen years.

Second, although VoIP promises new services, many of them are obscure and rely on yet unproven business cases. The current deployment of VoIP still performs the one-for-one replacement for the circuit-switched network. Is there a killer application other than voice? Vendors look for service providers to tell them which new service to sell, and service providers look to vendors for the answer [Mitchell 2004]. It challenges the industry to define and successfully market value-added VoIP applications for which customers are willing to pay. These applications might include worker mobility, wireless/wireline integration, unified messaging, number portability, and conferencing. However, the strong market need for these services similar to what the market experienced with cell phones or WiFi has not appeared yet.

Third, capital requirements are hard barriers for vendors to overcome. Despite the market want and need, carriers cannot secure budgets for new technologies. This barrier will diminish as the shift from TDM to packet technologies continues, freeing up billions of dollars of investment currently spent annually on legacy voice networks.

Fourth, due to inherent difficulties in transmitting a voice stream over a packet network, the voice quality problem has challenged VoIP ever since it was born.

Although improved in recent years, the voice quality of VoIP is still inferior to that of the traditional PSTN. The service provider cannot guarantee the bandwidth required to handle calls and associated services effectively due to limited bandwidth and lack of QoS controls. Optional solutions will be to use protocols such as multiprotocol label switching running on network routers, which provides switching capability and QoS by giving priority to certain IP packets. A similar technology called PacketCable, promoted among cable operators, enables QoS to IP-based services including VoIP, interactive gaming and video programming.

On the demand side, price, technological comprehension and customer perceived value of the service affect VoIP adoption. Large organizations with strong technical support and high cost and saving ratios are adopting VoIP faster than small and medium-size organizations, and organizations adoption as a whole is faster than residential/SOHO (small office home office). In the current stage, VoIP customers benefit from cost savings, e.g., lower long distance charges due to the pricing model of the Internet, where there are no time-sensitive or distance-sensitive charges, and less regulation enforced charges, e.g., high above cost access charges. The voice quality, reliability, availability of emergency services and regulation uncertainty hinder VoIP adoption. In North America, the business service market is expected to double by 2010, reaching $18 billion, and the residential service market is expected to increase from $295 million to $4076 million in 2008 [Frost & Sullivan Research 2005].

3.3. VoIP Product Features

A product can be viewed as a bundle of features or attributes, which together form the basis of customer preferences for the product [Lancaster 1996]. A VoIP product includes (1) basic technical characteristics such as reliability, fault tolerant capability, voice quality and data communications capacity, (2) value-added or enhanced features such as voicemail, call diversion capability, conference call, and (3) commercial elements such as "purchase and activation arrangements, pricing structures and levels, billing and payment arrangements, and after-sales customer service" [McBurney and Parsons 2002]. The demand will crucially depend on the particular set of features that a VoIP product offers, or the utility that each customer derives from VoIP, a function of its specific attributes. At the current stage, for carriers, lower equipment cost and lower operating cost, and for residential customers, prices are the most compelling attributes for VoIP.

VoIP uses an open architecture and universal standards, which lead to increased competition and mass production, lowering the equipment cost. Telecommunications systems are typically sold on a "price per port" basis. For a typical 100,000-port TDM switch, with per port costs ranging from $65 to $150 for basic functionality, the switch would cost about $6.5 to $15 million. The switch typically supports access ports and Inter Machine Trunk (IMT) ports in a 60/40 configuration, or 60,000 access ports and 40,000 IMT ports, as more traffic is bound to remain local and not require switching out to IMT. For a typical VoIP switch with similar capacity, or 60,000 access ports and virtually unlimited IMT ports, the cost is roughly about $50 to $75 per port or $3 to $4.5

million per switch, saving the carrier approximately 60 percent per switch [Sonus Networks 2006].

Softswitches are cheaper than circuit switches in terms of operations, administration, maintenance and provisioning (OAM&P). The OAM&P activities incur a large portion of the ongoing expense for a service provider to run the network and switches, which accounts for 60 to 70 percent of the overall expense. The OAM&P saving enables a VoIP service provider to offer voice service at 20 percent off the traditional rate [Ohrtman 2002].

The OAM&P saving first comes from less power and space cost. A softswitch can take as little as one-thirteenth of the space that a traditional circuit switch requires. For example[12], a softswitch with 36,000 Digital Signal Level 0[13] (DS0) can be placed in one 7-foot rack, or 12 square feet of space. On the other hand, a Class 4 switch with the same capacity needs 13 racks, or 156 square feet of space. The softswitch offers a 92 percent real estate cost saving over the Class 4 switch. Softswitches use less power. In the above example, if we assume that each rack consumes the same amount of power, the softswitch uses 8 percent of the power of the Class 4 switch.

Second, a packet-switched network is more efficient than a circuit-switched network on bandwidth utilization, which further reduces the OAM&P cost.

[12] The example is taken from [Ohrtman 2002].

[13] Digital signal 0 (DS0): a basic digital signaling rate of 64 kb/s, corresponding to the capacity of one voice-frequency-equivalent channel. The DS0 rate may support twenty 2.4-kb/s channels, or ten 4.8-kb/s channels, or five 9.67-kb/s channels, or one 56-kb/s channel, or one 64-kb/s clear channel. Multiple DS0s are multiplexed together on higher capacity circuits. 24 DS0s make a DS1 signal. When carried over copper wire, this is the well-known T-carrier system, T1 (the European equivalent is an E1, containing thirty-two 64- kb/s channels).

Since the 1950s, digital transmission has gradually replaced analog transmission due to its noise resistance and transmission capability. Compared with a single-channel copper wire, a T1 line can carry twenty-four 64-Kbps channels through time division multiplexing (TDM). Voice is encoded on each channel according to International Telecommunication Union (ITU) Recommendation G.711, finalized in 1972.

After G.711, other coding standards such as G.723 and G.729 were finalized to use more sophisticated coding algorithms and transmit speech at lower rates of 5.3/6.4 Kbps and 8 Kbps each respectively. Some coding techniques use silence suppression such that no traffic is generated when silence is detected in the traffic. Given that most conversations involve silence in one direction more than 60% of the time (that one person speaks and the other listens provides 50% savings; the pauses between words and sentences add another 10%), silence suppression saves significant bandwidth. These coding schemes, however, are difficult to implement in circuit-switched networks due to enormous investment overhead. As the circuit-switched network signaling protocol Signal System No. 7 (SS7) lacks support for coding schema negotiation, each network device along the voice path has to implement the same coding schema. This seems impossible especially for international long distance calls that traverse different network domains. In addition, in the circuit-switched network, route and bandwidth are reserved throughout the call, so bandwidth saving due to silent suppression does not apply.

The VoIP solution allows two ends of a call to negotiate the session such as coding schema, transmission rate and whether to use silent suppression. VoIP has great potential to reduce bandwidth requirements without compromising customer needs significantly.

As the bandwidth accounts for a large portion of a carrier's operating cost and initial investment, VoIP means much lower capital cost.

Finally, as VoIP enables integrated services over a unified IP infrastructure, business customers can save from maintaining a single IP network. Multiple applications including VoIP could share a single IP infrastructure cost. VoIP service providers can also reduce the operating cost and provide residential customers bundled services with lower incremental cost than subscribing to each service alone.

3.4. VoIP Service and Vendor Market Analysis

The wireline market in the United States is a $200 billion market [TIA 2006]. While fighting against revenue decline due to continuously losing customers to wireless service, wireline carriers treat VoIP as a serious strategic element, hoping that the introduction of bundled services at a flat rate would neutralize the advantage of wireless rivals. The TV program distribution combined with broadband Internet access would also give current subscribers incentives to retain their wireline service.

On the other hand, VoIP technologies enable diversified operators to enter the market, making competition even fiercer. The VoIP adoption process is a market share reshuffle process for multiple operators. In this section, we examine major players in the VoIP service market and vendor market, their positions in the competitive landscape, and their roles in the adoption of VoIP.

3.4.1. The VoIP Service Market

In North America, VoIP service providers include (1) non-facilities-based virtual network operators or VNOs, (2) long distance telephone operators or IXCs, (3) cable companies or MSOs, (4) local telephone operators including ILECs, (5) Other operators such as Internet service providers and wireless operators [Lam 2004].

The VNOs, such as Vonage, Primus and 8x8 offer the VoIP service independent of the network access service. The minimum upfront capital investment for network infrastructure and new subscriber acquisition cost enable VNOs to offer telephone services at low cost and compete with traditional circuit-switched voice. At the current stage, VNOs are the main drivers of the residential VoIP market. In the long run, however, with price being the major decision criteria for residential VoIP, many VNOs are difficult to sustain at existing price levels. After RBOCs and MSOs enter and compete in the VoIP market, VNOs, as low-priced providers, will be forced to drop prices. If VNOs cannot find the specific market niche to address and the compelling factor is limited to price, given the long-term market cannot sustain a large number of service providers, many VNOs will either be acquired or be run out of business.

The most significant market development will come from ILECs, IXCs, and MSOs. Although moving at slow pace, ILECs, IXCs, and MSOs will change the landscape of future telecommunications.

ILECs, or Incumbent Local Exchange Carriers, including four RBOCs: Verizon, SBC (renamed AT&T after acquiring AT&T), Quest and BellSouth, are telephone companies that provided local services when the Telecommunications Act of 1996 was enacted. The traditional telephone service, which is still a cash cow for RBOCs, is facing fast decline with the rapid market development of VoIP services. RBOCs are positioned to lose big in the VoIP adoption due to customer and revenue loss from margin-rich calling features, such as caller ID, call waiting, call forwarding and voice mail built into most VoIP offerings, which alone representing 15 to 30 percent of revenue source or about $30 billion per year for major carriers [TIA 2006]. While facing great pressure on losing the existing circuit-switched customer base, RBOCs have started to develop differentiated next generation services, including IPTV. Although starting late, the infrastructure investment from RBOCs will represent the single largest investment, covering new residential and business services development, broadband infrastructure buildups on fiber optic networks and packet updates of both local and tandem switches.

MSOs, or Multiple Services Operators, refer to cable TV companies that also provide Internet access. While owning access to end customers' homes though cable TV networks, MSOs have a lower cost structure to provide VoIP as a value-added service over the existing service bundle including Internet access and TV programming. The bundled service increases customer switching costs and ARPU (Average Revenue Per User) and positions MSOs to be the biggest winners in the VoIP diffusion. Supported by service providers such as Global Crossing, Level 3, Sprint and MCI for hosted VoIP platform, PSTN interconnection, local access number and regulation compliance, MSOs entered the voice market to compete with RBOCs. All major North American MSOs now have VoIP offerings [Lambert 2005].

IXCs or Inter-eXchange Carriers, often referred to as "long-distance carriers," provide connections between local exchanges in different geographic areas, or the interLATA service as described in the Telecommunications Act of 1996. The biggest IXCs include AT&T (now part of the former SBC), MCI (now part of Verizon) and Sprint. Given their large share of the business market, IXCs will concentrate on developing wholesale VoIP service and business VoIP service and using VoIP packet tandem applications to increase operational efficiencies, defending against RBOCs' encroachment on the business market. In the current local service market, IXC services are typically based on UNE-P[14], which adds to the cost of doing business, and most likely will not be around much longer. If they want to continue to compete in the local market, to move to VoIP is a better choice to avoid the UNE-P dead end, and to offer local services quickly, cost-effectively, and profitably.

Other service providers, including ASPs and ISPs, have a much smaller impact on the overall VoIP market over time due to their small capital budgets and customer base.

3.4.2. The VoIP Vendor Market

The diffusion of VoIP technologies has cultivated many opportunities for diversified vendors. As new startups emerge, grow up, consolidate with existing vendors or go out of business, the market is crowded with vendors specializing in small parts of the overall solution. Although product specialization is essential for the VoIP equipment market to develop, as carriers select vendors for long-term investments and solutions instead of a

[14] A regulation term that allows a CLEC to lease a combination of UNEs (Unbundled Network Elements) including local loops and switches. UNE-P enable the CLEC to deliver end to end service without any of its own facilities.

single product, a reliable, scalable, and bright future products offering and stable financial landscape will be rationalized, as carriers' next generation service plans evolve over time. Acquisitions, mergers and strategic partnerships will strengthen product offerings and provider a better prospect for service evolution path from VoIP to multimedia service offerings. With active RFPs[15] and purchase decisions taking place, the established telecom vendors will begin to acquire smaller and focused startups. Further consolidation is inevitable.

In 2005, the VoIP products mainly consisted of media gateways, session border controllers, softswitches, media servers and application servers.

The media gateway is a hardware platform with DSP cards that provides media and signal transformations between the PSTN and the packet network. It has two primary carrier applications: trunking and access. Trunk gateways serve IP transit purposes, interconnecting local circuit switches or tandems. Typically, high-density trunk gateways support more than 10,000 DS0s per chassis or unit, and mid-low density trunk gateways support less than 10,000 DS0s per chassis or unit. Access gateways convert circuit to packet calls in the local loop. In the market, some media gateways also incorporate certain functionalities of the session border controller to manage interaction of various networks.

The media gateway leads the VoIP market development in the rapid growth stage as increasing needs of circuit-to-packet conversion between circuit-switched local networks

[15] Request for Proposal, a document that an enterprise sends to a vendor inviting the vendor to submit a bid for hardware, software, services or any combination of the three. An organization typically issues the RFP to assess competing bids.

and packet based domestic and international long distance networks. The media gateway market will grow up with the VoIP diffusion and diminish with the circuit-switched network. Since the circuit-to-packet migration will take a decade, the media gateway market is expected to grow at a high CAGR of 31 percent to 2011, with revenue reaching $6.5 billion, or half the total VoIP equipment market revenue [In Stat Research 2005b]. The leading vendors in the low and mid-density media gateways market segment include Cisco, Siemens, Lucent, Huawei and Nortel. Among them, Cisco has about a quarter of the total market share, and the others each has about 10% of the market. In the high-density media gateway market, Nortel, Sonus and Tekelec share about 80% of the market, and Huawei, Cisco, Metaswitch and other small vendors share the rest of the market[16].

Session Border Controllers (SBC) control and manage real time multimedia traffic flows between IP networks, and handle IP interconnection functions required for real-time communications, e.g., access control, NAT/firewall traversal, bandwidth policing, accounting, signaling exchange, legal intercept and packet processing for QoS. The borders between IP networks include both inter-service provider borders (peering borders) and service provider-customer borders (access borders). The SBC can be either standalone or integrated with other network devices. Vendors with standalone products value the need of protecting the softswitch via topology hiding (IP masking) and preventing softswitches from denial of service attacks; vendors with an integrated product believe that the SBC functionality should be distributed to other network devices.

[16] The market share information is from Sonus internal documentation, with detailed numbers omitted.

The standalone SBC is expected to grow and then decline as the SBC function is integrated with other platforms. The leading SBC vendors include Acme Packet (about 50% market share), Netrake, Jasomi and Juniper.

A softswitch resides on a server or a dedicated hardware platform and provides call control, signaling, and intelligence for Class 4 transit service, or Class 5 local exchange service, or both. The real softswitch implementation may include signaling gateways or application server functions, and may be either standalone, integrated with media gateways or integrated with traditional circuit switches that support both TDM traffic and packet traffic.

Softswitches share one third of the VoIP market. It is projected to grow at a CAGR of 42 percent to 2011, with revenues growing from $560 million in 2004 to $4.8 billion [Infonetics Research 2005d]. The market leaders include Nortel, maintaining one fourth of the total market share, and Ericsson, Siemens, and Italtel, each maintaining about 10%.

Softswitches are less expensive to install and maintain, and use less space than traditional circuit switches. Softswitches will gradually replace class-4 tandems and class-5 switches, realizing circuit to packet conversion on both the trunking and local loop level. As we have discussed in the previous section, the circuit-to-packet migration is expected to take a decade, depending on the broadband deployment, revenue opportunity and competition. In the future, softswitches will be designed on the IMS architecture and able to realize fixed and mobile convergence.

The application server host applications unrelated to fundamental call controls. These applications include (1) local exchange applications, e.g., IP Centrex, (2) voice processing with media servers, e.g., conference bridge, instant messaging, presence, unified messaging and Interactive Voice Response (IVR), and (3) Intelligent Network SS7 applications, e.g., 800 calling.

Media servers process, manage, and deliver media requests made by voice application servers or softswitches in a packet network. It provides basic features such as audio play and record, and advanced features such as media stream blending and text to speech transformation. The softswitches may control the media server for simpler functions, such as network announcements, and application servers may control it for more complex applications, such as voice messaging, conference bridges and IVR.

Since application servers and media servers are deployed as centralized resources and support up to 20,000 concurrent calls per unit, a relatively small number of them will be deployed to serve the whole network. The current application server and media server market is about $246 million, and is expected to reach $1.76 billion in 2012 [Frost & Sullivan 2006]. In the coming years, with service providers becoming increasingly interested in offering enhanced features such as multimedia conferencing, unified messaging, video and multimedia, the media server and application server will become increasingly important, even though the market will remain small compared with the other segments.

In 2005, the vast majority of sales were still core infrastructure elements: media gateways and softswitches. This will change as new services are rolled out and other emerging product categories are adopted, such as voice application servers, session border controllers, and media servers.

3.5. The 3G Market

In the last section of VoIP market discussion, we briefly discuss the 3G market development, an inseparable component of the next generation network, as fixed and mobile converge. Compared with VoIP, 3G is more concentrated on new services other than voice that are enabled by high-speed wireless data transmission. The new services generate revenue opportunities for parties other than network operators and traditional wireless service providers. The 3G value chain captures this change.

A value chain describes a string of collaborating companies involved in delivering products or services to the markets to maximize value generation. In the 2G system, the value chain is simple – users purchase handsets and billing packets from operators through retail outlets. The users generate the network content – voice and short messages. The operators control the value chain and services. In contrast, since the 3G open network architecture allows third parties to access the network and build services, the 3G contains a much complicated value chain. A possible value chain of 3G is given in [Dave et al. 2003] (Fig. 28).

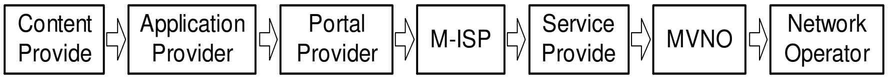

Fig. 28. The 3G value chain.

Other than the network operator and service provider in the 2G values chain, the 3G value chain also contains:

- Mobile virtual network operator (MVNO) – owns more network infrastructure than service providers.
- Mobile Internet service provider (M-SIP) – terminates data calls on an IP network and provides users with IP address and authentication.
- Portal provider – provides a mobile homepage and set of services associated with the portal provider, such as advertising and referrals.
- Application provider – supplies products purchased by users either by pay-per-use or by subscription.
- Content provider – powerful players in the Internet and mobile Internet application space. They sell or license their content such as music or web pages to portal providers

The value is shifting from network operators to content providers. According to KPMG's estimates, 25% of the total revenue will reside in the transmission and the remaining 75% will be divided among content creation, aggregation, service provision, and advertising [Dave et al. 2003]. In some countries, the number of 3G licenses awarded is greater than the number of incumbent 2G operators, creating new entrants. Coupling the opportunities for service providers and MVNOs, the challenges for licensed 3G operators are therefore even greater. The suggested 3G services include multimedia messaging, location based service, mobile commerce and business-to-business m-commerce [Dave et al. 2003].

As we have noticed, the 3G value chain parallels the Internet service. Will the 3G follow the same development path as the Internet? 3G is facing the same situation as VoIP. While the Internet service exhibited a unique and irreplaceable initial demand, the role of 3G represents a replacement demand that is subject to replacing the existing 2G for which substantial investments have already made. The market has evolved from original demand to replacement demand. The global wireless handset market - a useful indicator of this shift – indicates that in the early 1990s, most people bought their first handset, whereas in 1999, 40% of unit sales account for replacement. It was estimated that this number will continue to increase to 70%-80% in the next few years [Dan 2003].

The question therefore becomes: Is the unique feature of 3G compared to 2G strong enough to be the driving force for its diffusion? Or does the need of high-speed data service via the wireless link become the supervening social necessity? Will this necessity be strong enough? For VoIP, the initial cost saving is justified for the circuit-to-packet migration. For 3G, it purely depends on the customer needs for new services, which is still obscure within the VoIP world. The current research is focused on producing forecasts for the growth of wireless data service and the answer is not known. However, we are going to examine this potential based on Japan, which has a huge success in 2.5G and launched the first 3G service.

On the road to providing high speed wireless data services, the 2.5G system serves as an interim product, which is essentially an improved 2G system that provides higher-speed data service, up to 144kbit/s. The two 2.5 systems are the GSM Packet Radio System (GPRS) based on the GSM network core, and the CDMA2000 1x system based on the CDMA network core. The most successful 2.5G service in the world telecommunication market so far has been the i-mode service in Japan. The i-mode service allows users to access their e-mail and text messaging through mobile phones. Other services include viewing news and downloading ring tones and cartoon characters. NTT DoCoMo has adopted Compact HTML (C-HTML) as the script writing language for i-mode websites. C-HTML allows the web pages to be quickly displayed on the small screen of i-mode enabled terminals. According to NTT DoCoMo, more than 50,000 websites are available for i-mode terminals. The basic charge for i-mode is about 300 Yen ($2.50) per month plus 2.4 Yen (2 cents) per kilobyte downloaded [Dave et al. 2003]. The service has attracted more than 33 million users three years after its initial launch in February 1999. The other operator in Japan, J-Phone, launched sha-mail (picture mail), a service that enables a user to send photos taken from the phone's build-in digital camera to other users. The number of picture enabled handsets in use exceeded 20 million in 2003 [Kenichi 2004].

In 2001, NTT DoCoMo started 3G mobile service Freedom of Mobile Access (FOMA) that allowed users to access the Internet at up to 384 kbps speeds using packet transmission. NTT DoCoMo failed to achieve its targeted 1.46 million subscriptions at the end of fiscal year 2002 due to the limited service area of FOMA. In contrast, KDDI launched its CDMA2000 1X 144 kbps download service in April 2002 [Kenichi 2004]. Different from FOMA, CDMA2000 1X does not require substantial change to existing equipment, which in turn allows the service to be easily expanded to regions where mobile phone services are already offered. The 3G enabled phones from KDDI are compatible with the existing service, not with the more advanced communication service. In April 2003, KDDI had more than 6 million 3G subscribers and accounted for 10% of the mobile phone subscribers, and FOMA, 0.5% [Kenichi 2004].

According to the WIP Japan Survey in 2002, about 40% of the total population accessed the Internet via mobile phones. Mobile Internet was especially popular among young people with an average age of 32.2. The penetration rate is also related to education. 71.3% of the wireless Internet users possessed college degree. The main usage of mobile Internet was e-mail, which accounted for half of all mobile Internet users [Kenichi 2004].

In the literature, several articles have concentrated on the key factors that enabled Japan 2.5G to success [Kenichi 2004; Dave et al. 2003]. First, Japan companies focus more on the user's need than on the pure 'cool' technology. The i-mode technology is not advanced. Comparing with the far less successful Wireless Application Protocol (WAP), which is roughly equivalent to i-mode, i-mode was sold to users as a special service (with application and content useful for people 'on the move'), whereas WAP was hyped as 'just like the internet'. Second, because of Japan's national culture, people emphasize the importance of groups and organizations rather than individuals. Contrary to studies in the United States and other countries, where greater use of the Internet was associated with a decline of users' communication with family members and their social circle, the Internet has a positive effect on sociability in Japan. According to the result of the WIP Japan Survey, mobile phone users are more sociable and more interested in getting the newest fashion handset. Similar studies conducted in Korea, which has a relatively similar cultural background to Japan in group-oriented nationality, may suggest that culture character is related to the high penetration rate of mobile data communication.

Although 3G is still not prevalent in Japan, we still perceive the potential needs of wireless data service through the 2.5G service. Will the same situation happen in the rest of the world? The companies can reuse Japan companies' success factors to promote 2.5G or 3G, but the culture is still local. The telecommunication service perhaps can be designed to satisfy special customer needs and takes the culture differences into consideration.

4. The Development of Telecommunications Regulations - from Monopoly to Competition

In most industries, a product's average cost[17] has a U-shaped curve: it decreases with *economies of scale*[18], and then increases as marginal cost[19] increases because of the law

[17] Average cost is equal to total cost divided by the number of goods produced.

[18] Economies of scale refer to the decreasing per unit cost as output increases. It tends to occur in industries with high capital cost in which those costs can be distributed across a large number of units of production [Wikipedia].

[19] Marginal cost is the change in total cost that arises when the quantity produced changes by one unit. The marginal cost also has a U-shaped curve. Marginal costs "decrease as the volume of output increases due to economies of scale, which include factors such as bulk discounts on raw materials, specialization of labor, and more efficient use of machinery. At some points, however, diseconomies of scale enter in and marginal costs begin to rise; diseconomies include factors like more intense managerial supervision to control a larger work force, higher raw materials costs because local supplies have been exhausted, and generally less efficient input [Answers]." The marginal cost curve intersects with the average cost curve at minimum average cost when marginal cost increases.

of *diminishing return*[20]. Manufacturing firms would like to increase input until the average cost curve no longer decreases. The firm is operating at optimum output with minimum average cost and the size of production reaches an "ideal" size, or the *minimum efficient scale* [McConnell and Brue 2001].

In the telecommunications market, given enormous fixed costs and negligible constant marginal costs, a carrier's average cost continues to decrease as customers increase in a certain geographic market. The minimum efficient scale is so large that a single carrier could serve the whole market. Duplicate facilities of multiple carriers in the same geographic market would often dilute the incumbent's economies of scale and cause a carrier to produce less than the minimum efficient scale. In economics, when the minimum efficient scale is large enough so that there is no room for two or more firms to produce at minimum efficient scale, a natural monopoly results [McConnell and Brue 2001]. Given the premise that a telecommunication market is a natural monopoly market and multiple carries would only cause higher costs and higher prices for end consumers, the government had traditionally awarded the whole telecommunications market to a single firm in exchange for a commitment from the firm to provide reasonable services at reasonable rates. The regulator prohibited new entries by granting exclusive franchises to monopolists, covering from telecommunications equipment manufacture to telecommunications services.

[20] "In a production system, having fixed and variable inputs, keeping the fixed inputs constant, as more of a variable input is applied, each additional unit of input yields less and less additional output." [Wikipedia]

Starting in the 1970s, regulators began questioning the above assumptions. In economic paradise, the regulator ensures that a firm produces at an ideal size with the lowest average cost and sets welfare-maximizing prices. In the real world, however, regulators often lack resources, commitment, expertise and important information about the overseen market. The task of overseeing and directing market activities are therefore difficult to achieve [Armstrong and Sappington 2005]. The regulated firm lacks motivations to reduce production costs and to maximize consumer welfare. In the meantime, with the rapid development of telecommunications markets, the minimum efficient scale becomes relative smaller to the huge and still fast growing customer demand. The continued development and implementation of new public policies, which promote competition and demand privatization, change the telecommunications market structure. The old national dominant network operators have been broken into multiple competitive entities[21]. The market share has been divided among breakups and new IPOs. The privatization of state-owned enterprises and the associated deregulation/liberation become the driving force of signification improvements in the financial and operating performance in both developed and developing countries, as Bernardo et al. [2002] found when he investigated the financial and operating performance of 31 national telecommunication companies that were fully or partially privatized through public share offering.

[21] After the U.S. Department of Justice broke up the Bell system into a new AT&T and seven Regional Bell Operating Companies (RBOCs) in 1984, British Telecommunications was privatized in 1985, Japan NTT in 1980, and Korea Telecom in 1997 [Dan 2003].

In the telecommunications equipment market, for example, the Federal Communications Commission (FCC) in the United States created *Part 68 rules* [FCC 2007c] to order telephone companies to (1) unbundle equipment sale and service offerings, and (2) cooperatively use of devices from unaffiliated equipment manufacturers that meet established standards [Cannon 2001; Cannon 2003]. The open competition triggered enormous price decline of telecommunications devices and explosive of new end user devices.

In the Telecommunications Act of 1996, Congress totally dispensed with the natural monopoly premise and declared both the local and long distance market open for competition. The main direction that telecommunications regulations are heading towards is to provide a fair competition market environment and promote long-term investments.

In the rest of the chapter, we scope our discussion of telecommunications regulations within the United States. In section 1, we first discuss pro competition regulations including setting up interconnection obligations and leasing rights, prohibiting anti-competitive behavior, and adjustment of rate and universal service fees. The principles behind these regulation activities reflect native incentives of the FCC's decisions on most current issues. The principles are further exemplified in section 2, which discusses pro competition regulations in Telecommunications Act of 1996, and in section 3, which discusses pro competition regulations for broadband and VoIP. The last section briefly covers spectrum and wireless service regulations.

4.1. The Principles of Pro Competition Regulations

The deregulation process is far more complex than removing all legal restrictions. The regulator has to design detailed policies to foster vigorous long-term competition, such as enforcing interconnection and leasing obligations, supervising anti trust behaviors, changing price structure to better reflect operating costs, reducing costs while customers switch suppliers and removing barriers for new entrants. The regulator has to inquire about benefits and costs of specific policies or decide not to regulate at all. Since comprehensive directions are not available, in this chapter, we investigate three most important aspects of FCC regulations, including interconnection obligations and leasing rights, prohibiting anticompetitive business conducts, and adjustment of rate structure and universal service fees. The purpose is to draw some broad conclusions on the principles behind these regulation activities based on which the FCC make decisions.

4.1.1. Interconnection Obligation and Leasing Rights

It is a broad consensus that the government should impose interconnection obligations on telephone networks because of the network effect. The network effect causes a good or service to have a value to a customer dependent on the number of customers owning that good or using that service. Network effects become significant after a certain subscription percentage has been achieved, since after this point, additional people will subscribe to the service or purchase the good because of the accumulated value from other customers. In the telecommunications market, the value of a network is proportional to the number of customers that can be reached, as stated in Metcalfe's law[22].

In the absence of interconnection obligations, since large telephone companies can reach many more people and has better offerings because economies of scale produce lower average costs, large telephone companies often tend to refuse to interconnect with small ones and squeeze them out of the market. The phenomenon can be seen from the telecommunications market in the 1900s, when telephone networks were first developed in the United States [Winston 2002]. After American Telephone & Telegraph (AT&T) owned a collection of large local exchange operating companies - Bell System and bound them together through the single long distance network of that time, it compelled independent companies into joining the Bell System by refusing to interconnect with them. Since independent companies could not serve customers for calls going out of their service area, unless they somehow duplicated the national infrastructure, the market became a monopoly. The interconnection obligation is essential for a competitive market to invite new entrants.

The leasing rights refer to obtaining access to certain rivals' network elements on an unbundled basis. In dealing with leasing issues, the regulator has to solve such questions as to what extent (all network elements or just local loops) to allow leasing and at what rates. The regulator needs to balance three concerns (1) whether the leasing obligation

[22] "Metcalfe's law states that the value of a telecommunications network is proportional to the square of the number of users of the system (n^2). First formulated by Robert Metcalfe in regard to Ethernet, Metcalfe's law explains many of the network effects of communication technologies and networks such as the Internet and World Wide Web [Wikipedia]."

would undermine the incentive of carriers to invest in new facilities. Since short-term leases at guaranteed low rates are much less riskier than huge up-front capital expenditures, leasing rights mean having to share the fruits of capital expenditures that succeed while still bearing the full loss of ones that fail; (2) without regulated access rights to certain network elements, would telecommunications carriers be able to provide the service that they seek to offer; (3) whether competition exists, either from the same platform or from other platforms [Nuechterlein and Weiser 2004].

In 1996 and afterwards, the FCC required broad interconnections for local incumbents and imposed little on cell phone carriers and broadband carriers. We will further discuss the interconnection obligations in the Telecommunication Act 1996 in section 2 and interconnection obligations for the next generation telecommunications in section 3.

4.1.2. Prohibition of Anti-Competitive Behavior

The relationship between service providers can be either horizontal or vertical. The horizontal relationships are between competing providers of substitutable services, e.g., broadband access services from cable companies and telephone companies. The vertical relationships are between service providers in adjacent markets providing complimentary services, e.g., the broadband service from Verizon and the VoIP service from Vonage. Since vertical integration of adjacent services is often cost efficient and produces significant economics of scope, a dominant provider of one service would like to explore adjacent service markets. The government often oversees such business conduct and gets involved if the firm abuses its dominant power in adjacent service markets.

In the telecommunications market, last mile transmission service providers such as telecommunication companies and cable companies often integrate adjacent services, e.g., the Internet access service and VoIP service, and could therefore have incentives to abuse their monopoly power on transmission services to discriminate against firms providing similar vertically integrated services on access terms and access qualities.

According to the *one monopoly profit* theory, the total profit a monopoly could earn from adjacent markets through leveraging its monopoly power in its own market is equivalent to the extra profit it could earn anyway simply by charging more for the monopoly product itself. An exception to this principle is that when the dominant product is subject to price regulations, the firm could thus have incentives to extract profits from adjacent markets [Nuechterlein and Weiser 2004], as exemplified in the case against AT&T.

Under the Kingsbury Commitment of 1913, AT&T and its subsidiaries dominated each segment of the telecommunications market for most of the twentieth century: the Bell Operating Companies (BOCs) controlled most major local exchange markets; Western Electric, a part of the Bell System, was the exclusive equipment provider for BOCs and controlled the telephone equipment market; and AT&T Long Lines controlled the long distance market.

In the 1970s, the Department of Justice antitrust division alleged that (1) AT&T’s relationship with Western Electric was illegal, and (2) AT&T monopolized the long distance market [Economides 1999a]. In 1984, the Modified Final Judgment ("MFJ") settled the case and divested both manufacturing and long distance from local service, breaking AT&T into separate marketing companies: AT&T Long Lines, the research and equipment manufacturing unit – Lucent Technologies, and Regional Bell Operational Companies (RBOCs). The court further subjected RBOCs to various restrictions including a ban on the long distance market and telecommunications equipment market.

There are two rationales for splitting AT&T and confining ROBCs to local telecommunications markets [Economides 1999b]. The first was to prevent monopoly leveraging in an environment lacking interconnection obligations. Being subject to price caps on local rates, the pre-divestiture AT&T leveraged its monopoly power in the local exchange market to discriminate in the adjacent long distance market by charging high above cost rates for long distance services and suppressing other long distance carriers by refusing to interconnect with them. If the RBOCs were permitted to enter the long distance market, they might have incentives to discriminate against other rival long distance providers, e.g., reserving insufficient capacity and degrading rivals' quality of service, in favor of their own long distance operations. For this, the antitrust decree subjects the Bell companies to the affirmative equal access obligation, which offers all long distance rivals the same access terms as AT&T.

The second rationale was related to predatory cross-subsidization. While cross-subsidization may not be a problem in effectively competitive markets, its presence in monopoly and near-monopoly markets has historically concerned regulators for its harm to consumers and potential competitors. The pre-divestiture AT&T provided many services. Whenever competition arose, AT&T assigned costs away from the competitive market to the uncontested markets and undersold rivals with lower prices. After rivals were bankrupted, AT&T could charge consumers an even higher price.

The MFJ caused the long distance market to prosper. At the end of 1996, fifteen years after the MFJ settlement, the long distance market share of AT&T fell from 85% to 53%, and five large facilities-based competitors, including AT&T, MCI, Sprint, WorldCom and Frontier, and numerous wholesale service providers, effectively competed in the long distance market [Economides 1999a].

To prevent monopoly leveraging behavior and maintain a free competition environment in the adjacent market, the regulator decides whether to intervene based on such aspects as whether a monopoly (usually transmission service providers) exists in a given market, whether the monopoly has the incentive to leverage its monopoly power to suppress competition in the adjacent market, whether the threat of monopoly abuse is outweighed by efficiencies of vertical integration, and the regulation cost, administrative burdens, unintended and unpredictable effects of regulations on the market.

4.1.3. Adjustment of the Rate Structure and Universal Service

In the old telecommunications market that lacked competition and price adjustment mechanism, the regulator relied on complicated regulations to set “reasonable” service rates.

To compensate for telecommunications companies’ effort to serve the public, telecommunication companies are entitled to earn a “reasonable rate of return” on their overall cost. The local telecommunication companies file tariffs to recover the overall cost based on the reasonable rate of return.

The tariff, however, does not reflex the actual cost of each specific service. To maximize the subscribership of the telephone network, the government has long adopted a program called "universal service." The program requires that some consumers, e.g., rural telephone users, be provided with basic telephone services well below their actual cost. To achieve this, the deficit from rural area residents is subsidized from resources such as high above cost access charges from long distance companies, high above cost business rates, and extra revenue from urban area residents.

On the one hand, the universal service facilitates ubiquitous telephone subscribership that benefits not just an individual consumer, but also the whole society in enhancing economic development, democratic participation and public safety. The telephone service is just like postal delivery, so fundamental to modern life and it should be provided to all Americans as a civil right. On the other hand, the universal service distorts the price structure of the telecommunication service. The historical method of raising and distributing subsidies was also implicit, hidden and inefficient. It is often unclear who is subsidizing whom [Economides 1999a].

The old price structure prohibits incumbents from adjusting service prices according to market competitive pressure. The new entrants can often cream-skim the most profitable market segment, e.g., urban business customers, and leaving the incumbent to serve the less profitable customers, e.g., rural residential customers [Armstrong and Sappington 2005]. The free competition market requires prices to align to the marginal costs more closely. The traditional price structure and universal service regulation cannot be sustained in the new competitive era. Over time, in order to provide low rate access service for both high cost areas and public institutions, an explicit tax like competitively neutral mechanism is needed to raise subsidies from broad services for all qualified providers.

4.2. Pro Competition Regulations in the Telecommunications Act 1996

As 1996 approached, at least three basic opportunities appeared on the horizon of a free telecommunications market. First, the telecommunications sector had witnessed dramatic cost reduction in transmission, switching and information processing because of fiber and integrated circuits technologies [Economides 1999a]. Second, new entrants could enjoy significant economies of scale while serving just a fraction of the total customer base. Among others, MCI showed that multiple carriers could profitably compete for a share of the long distance market. The telecommunications markets were not or at least no long natural monopolies. Third, new entrants were eager to enter the local market segment that paid the most implicit cross subsidies [Hazlett 1999]. In the 1980s and 1990s, the emergence of the competitive access market enabled long distance companies to bypass some or all of the incumbent local networks and directly connect with the largest business customers for voice and data transmission.

Twelve years after the breakup of AT&T, the long distance market was successfully transformed from a monopoly to an effectively competitive market. Building on such experience, the 1996 Telecommunications Act attempts to formulate a competitive local exchange market, facilitating customers to acquire the benefits of technological advances.

The Act targets to eliminate both economic, regulation and business barriers for new rivals, and maintain low telephone service rates for all residential customers.

The task of eliminating economic barriers concentrates on enforcing interconnection obligations and leasing rights. As we discussed in section 1, investing enormous fixed costs to build a ubiquitous network to reach their prospective subscribers might be infeasible for new entrants with a small customer base. The interconnection obligations and leasing rights address the economies of scale issue and allow new entrants to build customer base first and then facilities.

In the local exchange market, carriers are divided into two warring sides on the debate of interconnection obligations and leasing rights: the ILEC and the CLEC. ILECs, or incumbent local exchange carriers, including the four regional Bell companies: Verizon, SBC, BellSouth and Qwest, possess market powers in the local exchange market. CLECs, or competing local exchange carriers, including traditional long distance companies AT&T (now part of SBC, renamed AT&T) and MCI (now part of Verizon), whose customers mainly consist of enterprise customers and a small fraction of small business and residential customers, compete to enter the local market.

In eliminating the economic barrier and promoting greater competition in the local exchange market, sections 251 and 252 of the Act mandate interconnections between rival carriers and receiving reciprocal compensation based on cost for calls traversing multiple carrier networks. Although both incumbents and new entrants are subject to interconnection obligations and leasing rights, since incumbents control most last mile access, the Act subjects incumbents to more obligations and new entrants expansive rights. Competing rivals can demand interconnection with incumbents at “any technically feasible point,” e.g., collocating facilities on incumbents’ properties, with just, reasonable and nondiscriminatory rates, terms, and conditions. The Act also permits new entrants to lease unbundled incumbents’ network facilities (UNEs), or capacity on those facilities. The lease is compensated based on the long run forward-looking cost incurred at the current most efficient carriers, or Total Element Long Run Incremental Costs (TELRIC). The actual rate that state commissions set on a particular network element is a tool that, rather than reflecting its actual cost, creates a margin between the wholesale and retail rates to invite new entries, and at the same time, does not dampen all carriers’ investment incentives in new facilities [Nuechterlein and Weiser 2004].

In the elimination of regulation and business barriers, sections 253 and 271 of the Act excludes state and local authorities from prohibiting new entries into the interstate and intrastate market. The Act also creates procedures to allow local incumbents to enter the long distance market, cable companies to enter the telephone market, and telephone companies to enter the video program market.

The Act allows the Bell companies to enter the long distance market upon showing evidence that they have loosened their monopoly grip on the local market and have set up separate long distance affiliates and have precluded anticompetitive practice against unaffiliated long distance companies.

The Act triggered a race between local companies and long distance companies on providing a complete “bundle” of local and long distance services to consumers. By the end of 2003, the four Bell companies – Verizon, BellSouth, SBC and Quest had all begun to offer long distance services to their existing local customers at a low incremental cost, and data and voice services for the more profitable enterprise customers with far-flung branch offices. On the road of entering adjacent markets, however, it was not until a decade later that Cable companies expressed serious interest in providing telephone services through VoIP, and did telephone companies plan to compete in the video programming market through fiber optic technologies.

To maintain low telephone service rates for all residential customers, considering that greater competition will erode the traditional implicit price subsidies, Congress ordered the FCC to set up a “universal service fund” as explicit subsidies from all companies that provide interstate and international services, including long distance companies, local telephone companies, wireless telephone companies and paging companies. The broad base of universal service funds reduces the size of distortions in the prices of other services, and therefore, facilitates fair competition. In 2004, the proposed contribution factor was about 9% for a total about $6 billion [UniversalFund.org 2007]. The federal administrator doles out the fund to (1) provide need-based subsidies for low-income households; (2) provide the non-need-based government mechanisms to keep telephone rates for high cost customers affordable; (3) fund broadband connections to the nation’s schools and libraries; and (4) fund connections to rural health care facilities [UniversalFund.org 2007].

The fund contribution based on interstate revenue cannot be sustained in the new competition environment when traditional long distance companies become less and less profitable, as customers began to use e-mails, messaging and VoIP phones as substitutes for the traditional long distance service.

In spite of the wide disagreement over the nature and administration of the Universal Service Fund, there is no consensus on how to fix it. In 2005, Senator Ted Stevens (R-AK) sponsored a bill called the Universal Service for Americans Act that would increase the universal service tax base to include broadband ISPs and VoIP providers, to fund broadband deployment in rural and low-income regions of the country. Senator John Sununu (R-NH) argued that such subsidies distort competition and thwart progress in the arena of broadband access. In 2006, the FCC required all VoIP services that connect to the PSTN network to contribute to the Universal Service Fund.

4.3. Pro Competition Regulations for Next Generation Telecommunications

The emergence of IP-based next generation telecommunications networks raises new challenges for the old regulation environment. In the old environment, massive investment in capital was required to set the network infrastructure. The telephone networks were first built at a local level, and then across states - a strong state role based character. In the new era, the packet-based network has much lower market entrance barrier and is borderless in nature. It is impractical to divide the service into 'interstate' and 'intrastate' like the traditional telephone service. As companies under different regulated titles began to compete for the same market, the current regulatory structure contains disparity elements that could potentially aid some competitors and handicap others, hindering the development of long-term competition. A consensus has been growing that even the 1996 Act are not sufficient to address the changing telecommunications environment. The current regulatory debate focuses on such issues as to what extent should regulations should be applied to traditional providers that enter new markets where they do not hold market power, and to what extent should existing regulations impose on new entrants of the traditional market, and the appropriate regulatory framework to be imposed on new technologies that are not easily classified under the present framework [Gilroy and Kruger 2006]. A more antitrust-oriented 'horizontal' approach that focuses on the layering architecture of the Internet is proposed to replace the current vertical model [Mindel and Sicker 2006; Frieden 2004]. The model identifies those layers that are subject to continuous regulation because there is insufficient competition within them and those layers that should not be subject to such regulation because they are presumptively competitive.

In the discussion of the next generation telecommunications regulation, we concentrate on the regulation of broadband and VoIP.

4.3.1. Regulation in a Broadband World

The broadband infrastructure bestows a comparative advantage for "knowledge based" industries such as data processing, banking, insurance, consulting, customer relationship management and government stewardship [Frieden 2005]. Poor new telecommunications infrastructure development would spoil the overall development of information and communications technologies and global marketplace success. Broadband regulation is critical to the broadband infrastructure development and affects services such as voice, data and TV programming that build on top of it.

The cross platform competition from cable, DSL and wireless protects consumer welfare and keeps service rates at a reasonable level, avoiding unintended consequences of direct regulation. Since the regulator has not limited the rate on broadband Internet access, the broadband providers lack strong incentives to discriminate against unaffiliated applications and content providers [Nuechterlein and Weiser 2004]. Such discrimination could also drive customers to the rival platform and decrease the value of the platform itself. The basic challenge of current broadband regulation is to remove roadblocks to investment and promote long-term investment in expensive broadband facilities.

(1) *Interconnection obligation and leasing rights*

The long distance telephone network has enormous overlap with the Internet backbone. The largest long distance carriers, such as AT&T, MCI, Sprint, and Quest, own the largest Internet backbones. The backbone providers sell transport services to the Internet Service Providers (ISPs), who in turn provide access facilities to connect both end users and content providers to the broader Internet.

Internet backbones interconnect with each other based on two arrangements: peering or transit [Oxman, 1999]. The peering arrangement usually involves backbone providers of equivalent market presence. The traffic is handed off at the closest point to the origination at no cost. The transit arrangement involves backbone providers of unequal size, under which smaller ones compensate bigger backbone providers for interconnection.

Currently, the interconnection arrangements are completely unregulated. “Each backbone provider bases its decisions on whether, how, and where to interconnect by weighing the benefits and costs of each interconnection [Oxman 1999].” So far, no backbone provider has grown large enough to dominate the market. When large backbones compete for the transit business of smaller backbones to increase their revenue, the competition keeps rates at a reasonable low level.

In the meanwhile, antitrust authorities closely watch the backbone market to ensure that no provider grows big enough to dominant the market. In 2000, the Department of Justice blocked the proposed merge of MCI-WorldCom and Sprint for fear that the excessive market share controlled by the combined company would undermine the competitive market formed since the Department of Justice challenged AT&T's monopoly of the telecommunications industry 25 years ago [Borland 2000].

On the debate of leasing obligations, the regulator is facing the same challenge in the circuit-switched network, that is, on the one hand, to promote the incentive of incumbents and new entrants to make long term investment in network facilities to challenge cable companies, who have already taken a substantial lead over telephone companies, and on the other hand, to empower new entrants to compete in the broadband access market.

In the modern perspective of Schumpeter’s “creative destruction” theory, the emergence of new insurgents because of new technological revolutions, e.g., broadband access via power line, satellite, and various fixed and mobile wireless services, while being free from regulations, will undermine any market dominance of established monopolies, e.g., telephone and cable companies [Nuechterlein and Weiser 2004]. Subjecting the second place telephone company to invasive leasing obligations would be perverse.

The Triennial Review Order distinguished leasing obligations of circuit-switched facilities and the next generation fiber-oriented broadband facilities [FCC 2003b]. To build a fair competition market and diminish regulatory disparity for different platform providers to compete in the same market, exempting leasing obligations of telecommunication companies would help them to catch up with cable companies, who are subject to no such requirements. Although in the face of some CLEC impairment, the FCC exempted these fiber-oriented facilities from unbundling obligations in fear that it would deter both ILECs and CLECs alike from building broadband networks, for which ILECs have no special competitive advantage and the competition from cable companies is already persistent and substantial.

(2) *Prohibition of Anti-Competitive Behavior in the Broadband World*

In the 1970s and 1980s, when circuit-switched networks were served as the single platform to access the Internet, the FCC issued a series of orders called *Computer Inquiries* to control the anticompetitive behavior of local telephone companies on narrowband Internet access. The order first differentiates the information service from the traditional telecommunications service. According to the definition further refined in the Telecommunications Act of 1996, the traditional telecommunications service, or common carriers service, is a "pure transmission service over a communications path" [FCC 1996]. The emerging information service, or enhanced service, is offered over common carrier transmission facilities for interstate communications and uses computers to process the transmitted format and content, and transmission protocols. The Internet Service Providers (ISP) such as AOL or EarthLink provide an information service, not telecommunications service.

The Computer Inquiries further require each telecommunications service provider to separate the transmission service from the information service, tariff the transmission service and sell it on a nondiscriminatory basis to all information service providers including its own affiliated business. Since ISPs have strong incentives to allow customers to access all applications and content providers on the Internet to maximize their service value, the ISP creates an insulation buffer that prevents telephone companies from abusing their dominant power on the last mile transmission to stifle competition in the adjacent application and content market [Economides 1999b]. Most importantly, the Inquires classify information service into Title I of the Communications Act, which contains few rules of any kind and preempts any state regulation on such services.

As the Internet access technique developed into the broadband age, and half the United States residential customers began to use broadband to access the Internet, telephone companies are facing cross-platform competition from cable companies and wireless companies. The competition to some extent safeguards the consumer's interest, avoiding unintended consequences that regulation could produce. So far, there is no strong evidence that major broadband providers leverage their market power at the physical layer to crush competition at higher layers, e.g., using QoS techniques to slow competitors' packets while speeding up their own, because broadband providers lack of incentives to discriminate against unaffiliated applications and content providers as regulators have not capped broadband access rates [Nuechterlein and Weiser 2004]. The malicious behavior will subject the provider at least to the risk of antitrust sanctions. Furthermore, the provider is already facing fierce competition from other platforms; limiting services might simply degrade the value of the platform and reduce the rate that consumers would like to pay for it.

Given the above facts and analysis, regulators hesitate to impose non discrimination rules on broadband provides, for (1) such non discrimination rules are inherently subjective, e.g., for VoIP applications, it is good for a network to differentiate and give higher priority to latency-sensitive voice packets over data packets; (2) The detailed regulation task of enormous particular network usages in such a technologically uncertain environment involves substantial cost and risk; and (3) The regulation could generate unintended effects and make a high stakes bet on how the market will evolve, e.g., improper regulation would risk discouraging companies long term investment in new technologies as such investment is not sure to be recovered if short term monopoly are not allowed [Nuechterlein and Weiser 2004].

The multiple independent ISP model in the narrowband world is not a straightforward way to safeguard the nondiscriminatory access of applications and content in the broadband context. It involves significant technical and regulatory costs much like requiring incumbent telephone companies to unbundle their network elements to CLECs.

4.3.2. VoIP Regulation

The VoIP technology turns voice calls and related features into another set of applications running on top of the broadband connection. In the enterprise market, VoIP is becoming the industry norm and gradually phasing out the PBX solution. Among other advantages, VoIP permits a range of new attractive features and allows companies to cut costs by integrating voice and data networks. Although the transition to VoIP is going more slowly in the residential market, it will pick up the pace with the proliferation of broadband.

The remarkable growth of VoIP, together with the rapid expansion of broadband and wireless service, is causing permanent structural change in the telecommunications markets. The business model of traditional telecommunications services built on centralized switches cannot last, as applications are moved toward the user side and become much less expensive. It has seen accelerated revenue loss from landline voice services as customers switch to VoIP technologies that provide cheaper offerings of local, long distance and enhanced services, and can often exclude access charges while placing long distance calls. Between 2000 and 2004, landline voice revenue declined from $229 billion to about $196 billion, a decline larger than the combined 2004 revenues for the movie and radio industries' main source of income [Bianco 2006]. Some local companies first refused to provide "naked DSL" service to customers who did not also subscribe to their conventional circuit-switched telephone service over the same loop, but as VoIP has become widespread, given the choice between keeping their customers on less favorite terms or losing them outright, either have started to offer “naked DSL” or simply provide their own VoIP service in the hope of winning customers back [Lake 2005].

VoIP regulation highly depends on the classification of VoIP services. Typically, VoIP can be classified into three categories based on the customer premises equipment at the end points of the transmission: computer to computer service (IP-to-IP), phone to phone service (PSTN-to-PSTN) and computer to phone service (IP-to-PSTN). The regulatory treatment of IP-to-IP service and PSTN-to-PSTN service are clear, as seen from the FCC ruling on the Pulver case and AT&T case.

Overshadowed today by Skype, Pulver.com offers similar free voice calls and instant message services called Free World Dialup (FWD) over the Internet. In early 2004, while reasoning that Pulver provides no actual transmission function, the FCC ruled that Pulver provides a pure Title I information service, not a Title II telecommunications service [FCC 2004a]. The ruling exempts Pulver from virtually any federal and state-level regulation, and a long list of Title II obligations, among others, including access charges, leasing unbundled network elements and the universal service fund. In the future, as there will be no entity or infrastructure over which national power can be exercised, and the true IP-to-IP providers often have no revenue stream to tax or divert, the IP-to-IP service will not be subject to economic regulation and escape most social regulations [Bianoco 2006]. Although government agencies wish to impose CALEA-like wiretapping obligations on IP-to-IP services, there is no obvious way to enforce it because the software provider can operate the service in any country that imposes no such obligations.

In contrast to the Pulver case, the FCC ruled that AT&T's "phone-to-phone" long distance services were telecommunication services, not information services, although the service uses IP backbone in the middle. The FCC stated that "users of AT&T specific service obtain only voice transmission with no net protocol conversion" and "do not offer a different service, pay different rates, or place and receive calls any different than they do through AT&T's traditional circuit-switched long distance service [FCC 1997]." The ruling subjected these calls to Title II obligations.

Comparing with the clear cut regulation of the IP-to-IP and PSTN-to-PSTN service, the regulation of the IP-to-PSTN service is much more ambiguous, as to (1) whether to treat the service as a traditional telephone service and divide it into intrastate and interstate components and let the states regulate the latter, and (2) whether to treat the service as a telecommunications service or an information service and exempt it from the common carrier obligations under Title II. In the following, we concentrate our discussions on the regulation of IP-to-PSTN service.

After Minnesota Public Utilities Commission ordered Vonage to obtain state certification as a telephone company, file a tariff, and provide the same 911 service as regular telephone companies, a federal district court justified that Vonage was providing an "information service" rather than a "telecommunications service." When the Minnesota PUC appealed that decision to the Eighth Circuit, the FCC jumped in and ruled that Vonage was providing an interstate service instead of an intrastate service and exempted Minnesota PUC from regulation on such issues.

The FCC reasoned that, given current technical difficulties, it is hard to divide the service, like conventional telephone services, into distinct "interstate" and "intrastate." Attempting to locate subscribers for the sake of regulation itself would force changes on the service and would be negative for the development of innovative services and applications. The commission, although deferred the resolution of whether to put the service into Title II "telecommunications service" or Title I "information service," asserted exclusive federal jurisdiction on the service and preempted most state regulation, a character typically defined in Title I [FCC 2004b].

The FCC generally avoided adopting a broad policy related to the regulation of VoIP, indicating that it wanted to avoid marking decisions that might causes problems for the nascent VoIP industry. On the one hand, the FCC appreciates that a well functioning market would protect consumer welfare more than government agencies. When the FCC concludes that a relevant market is competitive, it would classify the service as an "information service" under the Title I jurisdiction to preempt the state regulation, and exempt some or all Title II requirements when appropriate. On the other hand, the FCC can also subject a service classified under Title I to obligations imposed by Title II on conventional "telecommunications services" using the ancillary authority [FCC 1996]. The FCC's ancillary authority becomes increasingly important as a growing number of IP products resemble the service that has traditionally regulated under Title II.

The purpose of placing a service under Title I is to free it from legacy regulations, e.g., quality of service and tariff filing requirements, and rate regulations designed to protect consumers in a monopoly market with formidable barriers to entry. In a competitive market with relative low entry barriers, consumer protections are naturally built in. Therefore, the question left is how to impose obligations that are to some extent unrelated to the level of competition in a particular market. In the VoIP context, these include, among other things, interception of communications by law, the E911 service, accommodation of the needs of the disabled, and consistency with universal service goals.

Different from the circuit-switched network, the VoIP network does not naturally possess a caller's phone location. Some service providers, like Vonage use intermediaries to connect emergency calls to the traditional 911 calling network. The E911 service obligation subjects VoIP providers to enormous burdens. The FCC initially chose to rely on industry initiatives to address E911 and would like the providers to follow a staged approach similar to the path of cellular telephone providers who are also facing the E911 challenge of their own. Whether to require VoIP and wireless providers to implement the same E911 service as wireline providers no matter at what cost, depends on complex and ultimately subjective judgment about the importance of E911, and considerations of regulatory parity of multiple platforms including wireless, wireline and VoIP. As consumers began to switch to and increasingly rely on wireless and VoIP telephony, the regulator inclines to subject wireless and VoIP providers to "non-economic," or social regulations to which wireline ILECs have long been subject. In

June 2005, the FCC required interconnected VoIP service[23] providers to provide E911 service.

Although the FCC bears the perspective that non regulation, or least regulation, of the Internet and Internet application should be a guiding principle, it is expected that VoIP services will be subjected to most of the social obligations imposed on PSTN in the future. Other than the reason of removing regulation disparity for all voice service providers, another reason is just as [Bianco 2006] pointed out: as incumbents struggle to slow their loss of billions of dollars revenue in voice services while transiting to new broadband service, it is true and often cost effective that incumbents continue to lobby legislators and regulatory agencies to impose regulatory obligations on new VoIP competitors to raise their cost and slow down their growth. In September 2005, the FCC mandated interconnected VoIP providers to compliance with the security requirements needed to implement the Communications Assistance for Law Enforcement Act (CALEA). The order extended to include all facilities-based broadband Internet access providers in May 2006. In June 2006, the FCC required all voice over Internet Protocol services that connect to the public-switched telephone network to contribute to the Universal Service Fund.

On economic terms, interconnection and inter-carrier compensation are two important issues that will have broad implications on the development of competitive VoIP services. Under the Telecommunications Act of 1996, all "telecommunications services" providers are required to interconnect with each other. Since the IP-to-IP

[23] Interconnected VoIP service allows customers to make and receive calls to and from traditional phone numbers, usually using an Internet connection.

service is classified as an “information service” and it is still unclear whether the IP-to-PSTN service should be classified as a “telecommunications service,” providers of IP-to-IP and IP-to-PSTN services have no automatic right under the law to interconnect their networks and interconnect with the PSTN.

The inter-carrier compensation is one of the most important economic terms that define the interconnection obligation. In the PSTN world, under current regulations, a long distance carrier pays access charges to local carriers on each end of the long distance call at a rate high above its actual cost [Klein 1997]. At the local level, a calling party’s carrier, whose customers originate calls that “cause” the extra cost for the terminating carrier, pays at a lower reciprocal compensation rate to the terminating carriers to which local calls are handed off. The latter inter carrier compensation strategy is termed the calling-party’s-network-pays (CPNP) rule [FCC 2001].

In the VoIP world, VoIP providers have never owed access charges to any local provider for calls originating on the IP side, simply because the call never passes through the public switched network. On the terminating PSTN side, ILECs claim that VoIP providers owe them access charges, whereas VoIP providers argue that ILECs should be compensated at lower reciprocal compensation rates as the call is handed off to a point local to the ILEC, and in cases when a call initiated in the PSTN network is handed off locally to VoIP providers, the ILEC should compensate VoIP providers based on the CPNP rule.

In each scenario, the stakes are enormous, disadvantaging some consumers and companies that benefit from cross-subsidies. The FCC has proposed to formulate a unified regime for intercarrier compensation based on costs that contains no implicit subsidies [FCC 2001].

In either case, the IP side will be exempted from access charges, and no access charges for IP-to-IP calls. These savings will allow VoIP providers to differentiate themselves from circuit-switched telephony providers. The differentiation accelerates the migration from circuit-switched services to VoIP services throughout the market.

4.3.3. The Spectrum and Wireless Regulation

The U.S. government claimed that the spectrum is a public resource in the 1920s. Under the premise that the spectrum is scare, and unregulated use would lead to its despoliation and cause widespread interference, the Supreme Court designated the FCC to manage the spectrum use. The two most critical aspects of the traditional spectrum management regime include (1) allocation of particular blocks of frequencies – bands for specific services, and (2) assignment of those bands to particular licensees [FCC 2007a].

Designed to assign spectrums to the "most qualified" users, the first FCC regime for licensing the bulk use of prime spectrum was through "comparative hearings." In practice, however, the regime tended to favor incumbents and those with strong political ties [Faulhaber and Farber 2003]. The process was expensive and time-consuming as the stakes involved were huge and the loser was entitled to appeal. The FCC nonetheless used comparative hearings until the 1980s confined under the Communication Act of 1934, which guaranteed license applicants a right to a hearing and did not provide an alternative assignment mechanism [Nuechterlein and Weiser 2004].

In 1984, Congress authorized lotteries as a replacement for comparative hearings for cellular telephone licenses. After a few years of practice, lotteries proved to be a poor mechanism as well. First, the huge economic prospect of obtaining a free license generated hundreds of thousands of applicants that almost collapsed the FCC's prescreening process. Second, since the winner has no obligation to keep the licenses themselves, the winner could sell the license on the secondary market and acquire enormous sums, which should be obtained by the public treasury. It is therefore important to assign the license right in the first place [Cramton 1998].

In the early 1990s, Congress authorized the FCC to use auctions to assign spectrum licenses to cut down the front-end delays and raise revenue in the process. Auctions are an effective means of assigning spectrum licenses to those who would make the most efficient use of them and increasing opportunities for competition in the telecommunications services market. The auction revenue is transferred to the government and available to the public, compared with comparative hearings, in which most expenditures went to resources to increase the chances of winning a license, such as the time of lawyers and engineers in preparing applications, litigating, and lobbying. Although privately valuable, such expenditures are largely socially unproductive [Kwerel 2001]. The 1996 Act requires auctions for most initial spectrum licenses.

From comparative hearing to auctions, the FCC has made an important step toward market-oriented mechanisms for assigning licenses. The FCC's spectrum management regime, however, remains an exemplar of command regulation. One of the most influential proposals for licensing reform is the property rights model, proposed by

Nobel Laureate Ronald Coase, which uses free markets to allocate the spectrum and assign licenses.

The property rights model proposed similar functions as the law governing private transactions for the purchase and sale of land. The spectrum owners would freely trade or lease segments of spectrum in a robustly competitive secondary market. The role of the government would be to define the relevant property rights and enforce contractual agreements. With well-defined property rights, the free market will generally allocate resources to their most efficient use so long as transaction costs are low [Faulhaber and Farber 2003]. For example, a firm using assigned spectrum for its own internal communications between its local offices would be free to sell its licenses (relatively expensive) to a wireless telephone carrier and purchase capacity on a fiber optic network (relatively cheap) instead. Both parties would benefit from the trade.

While moving slowly toward a free market spectrum management regime, starting in the mid-1990s, the FCC auctioned the Personal Communication Service (PCS) licenses for generic mobile communications service uses. Unlike most other licenses, PCS licenses neither restrict the spectrum for particular uses nor are dedicate to a particular technology. In 2003, the FCC reformed its Secondary Market Order, which relaxed the standard of certain spectrum transfer [FCC 2003a].

In 2006, the U.S. market had 160 million cell phone users and a national penetration rate of 54% [Horrigan 2003]. Competition in this market is fierce: at least four alternative wireless service providers exist in most living areas throughout the country;

customers can transfer from one carrier to another with their numbers; and prices have continued to fall since 1990.

Unlike the wireline network buildup process, which incurs enormous expenses in installing copper or fiber from the central office to each home and business in a given service area, the wireless network buildup process has different cost patterns. The last mile connection between a wireless carrier and its subscribers consists of signals on different bands. A carrier can choose to build large cells in less populated area to reduce the effect of economies of scale. Therefore, the main obstacles are regulations on spectrum use, not economic. Since the 1980s, the FCC has allocated and assigned abundant spectrum to multiple wireless service providers, among others, including the two 25 MHz of spectrum in the 800 MHz band assigned to an incumbent LEC and an independent provider for each service area via lotteries, and a total of 120 MHz of PCS licenses that were divided into 10, 20 or 30 MHz bands and auctioned on a local or regional basis [FCC 2007b].

The spectrum abundance produced the intensely competitive wireless service market, which made pervasive regulation of the wireless market unnecessary. While lacking of a natural monopoly, the current regulation work concentrates on interconnection and standard setting challenges [Nuechterlein and Weiser 2004]. Observing that, because no carrier dominates the wireless industry and most carriers have much to gain and little to lose from negotiating efficient roaming arrangements for their subscribers, the FCC has generally declined to regulate roaming agreements between cellular network providers. For interconnections between wireless and wireline networks, the FCC has required

ILECs to interconnect with unaffiliated cellular operators and conditioned their own cellular licenses based on such obligations.

5. Conclusions

This chapter investigates the development and diffusion of next generation network technologies from the standardization, regulation and market points of view. The worldwide standardization of 2G and the Internet has contributed to their huge success around the world. The comparison of the standardization process of 2G, the Internet and the next generation network predicts that standardization can also boost the development and diffusion of the next generation network. During the development of next generation networks, most countries in the world are experiencing privatization and deregulation of telecommunication policies. Governments have created more competition through privatization and deregulation, and tried to promote long term investment, although sometimes overestimated or underestimated the growth potential of certain technology. Also, as the IP-based telecommunications network is changing the regulatory environment in which the old telephony market was developed, the roles of organizational and legal forces are enduring a fundamental change. The market can fail unless supported by sensible regulation facilitating competition. VoIP technology provides both dazzling opportunities and traps for both the old and new business models, as the market analysis of VoIP has indicated. Although promising new services, the next generation network is different from the Internet as it is subject to replacing existing telecommunication systems verses the original demand of the Internet. Does the wireless data and VoIP generate enough supervening social necessity for their diffusion?

Although representing only a small portion of the whole telecommunications market, the accelerated adoption of VoIP and the success of 2.5G in Japan illustrate the growing potential for next generation networks. It is not a question of whether or not, but a question of how fast the next generation network will prevail.

CHAPTER VIII

CONCLUSIONS

In the spirit of the Doctor of Engineering Program, the internship objectives were set to develop understanding and experiences in various areas that affect engineering practice. In this Record of Study, I have surveyed and analyzed the current standardization status on VoIP security and proposed an Internet draft on secure retargeting and response identity. To support product line development and enable product evolution in the current fast rising VoIP market, I have proposed a generic development framework for the SIP stack and SIP applications. I have also investigated the current status and influence of standardization, market forces and government regulation on the development and market diffusion of next generation telecommunication technologies.

REFERENCES

3GPP. 2007. About 3GPP. Available online at http://www.3gpp.org/About/about.htm.

3GPP2. 2007. About 3GPP2. Available online at http://www.3gpp2.org/Public_html/Misc/AboutHome.cfm.

ALAN, C. 2001. *Future Mobile Networks: 3G and Beyond*. Institution of Engineering and Technology, Herts, United Kingdom.

ARANGO, M., DUGAN, A., ELLIOTT, I., HUITEMA, C., PICKETTPP, S. 1999. Media Gateway Control Protocol (MGCP). RFC 2705, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc2705.html.

ARKKO, J., CARRARA, E., LINDHOLM, F., NASLUND, M., NORRMAN, K. 2004. MIKEY: Multimedia Internet KEYing. RFC 3830, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3830.html.

ARKKO, J.,CARRARA, E., LINDHOLM, F., NASLUND, M., NORRMAN, K. 2005. Key management extensions for Session Description Protocol (SDP) and Real Time Streaming Protocol (RTSP). Internet Draft, Internet Engineering Task Force. Available online at http://www3.ietf.org/proceedings/06mar/IDs/draft-ietf-mmusic-kmgmt-ext-15.txt.

ARMSTRONG, M., SAPPINGTON, D. 2005. Regulation, competition and liberalization. Available online at http://www.econ.ucl.ac.uk/downloads/armstrong/reg2.pdf.

BERNARDO, B., JULIET, D., MARCELLA, F., WILLIAM, L. M. 2002. Privatization and the source of performance improvement in the global telecommunications industry. *Telecommunications Policy 26*, 177-196.

BIANCO, M. C. D. 2006. Voice past: the present and future of VoIP regulation. Available online at http://www.thevpf.com/txt/Del_Bianco.pdf.

BOND, G. W., CHEUNG, E., PURDY, K. H., ZAVE, P., RAMMING J. C. 2004. An open architecture for next-generation telecommunication services. *ACM Transactions on Internet Technology 4*, 1, 83-123.

BORLAND, J. 2000. DOJ files to block WorldCom-Sprint merger. Available online at http://www.news.com/2100-1033-242457.html.

BRADNER, S. 1996. The Internet standards process. RFC 2026, Internet Engineering Task Force. Available online at http://www.ietf.org/rfc/rfc2026.txt.

CAMPBELL, B., ROSENBERG, J., SCHULZRINNE, H., HUITEMA, C., GURLE, D. 2002. Session Initiation Protocol (SIP) extension for instant messaging. RFC 3428, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3248.html.

CANNON, R. 2001. Where Internet service providers and telephone companies compete: a guide to the Computer Inquiries, enhanced service providers and information service providers. In *Communications Policy in Transition: The Internet and Beyond*. MIT Press Cambridge, MA, 3-34.

CANNON, R. 2003. The legacy of the Federal Communication Commission's Computer Inquiries. *Federal Communications Law Journal 55*, 2, 167-206.

CHEN, E. Y. 2006. Detecting DoS attacks on SIP systems. In *Proceedings of the 1st IEEE Workshop on VoIP Management and Security*, Tokyo, Japan, 53-58.

CRAMTON, P. 1998. The efficiency of the FCC spectrum auctions. *Journal of Law and Economics 41*, 727-736.

DAN, S. 2003. Globalization of wireless value system: from geographic to strategic advantages. *Telecommunications Policy 27*, 207-235.

DAVE, W., PHILIP, E., LOUISE, B. 2003. *IP for 3G*. John Wiley & Son Ltd, England.

DOHERTY, S. 2005. VoIP services beat regulation for now. *Network Computing 16*, 26-27.

DRYBURGH, L., HEWETT, J. 2005. *Signaling System No. 7 (SS7/C7): Protocol, Architecture, and Services*. Cisco Press, Indianapolis, IN.

ECONOMIDES, N. 1999a. The Telecommunications Act of 1996 and its impact. *Japan and the World Economy 11*, 4, 455-483.

ECONOMIDES, N. 1999b. Competition and vertical integration in the computing industry. In *Competition, Innovation, and the Role of Antitrust in the Digital Marketplace*. Kluwer Academic Publishers, Norwell, MA.

Elwell, J. 2007. Connected identity in the Session Initiation Protocol (SIP). Internet Draft, Internet Engineering Task Force. Available online at http://quimby.gnus.org/internet-drafts/draft-ietf-sip-connected-identity-03.txt.

FAULHABER, G. R., FARBER, D. J. 2003. Spectrum management: property rights, markets, and the commons. In *Rethinking Rights and Regulations: Institutional Responses to New Communication Technologies*. MIT Press, Cambridge, MA, 193–226.

FCC FEDERAL COMMUNICATIONS COMMISSION. 1996. The Telecommunications Act of 1996. Available online at http://www.fcc.gov/Reports/tcom1996.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 1997. In the matter of petition for declaratory ruling that AT&T's Phone-to-Phone IP telephony services are exempt from access charges. Available online at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-04-97A1.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2001. In the matter of developing a unified intercarrier compensation regime. Available online at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-01-132A1.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2003a. Promoting efficient use of spectrum through elimination of barriers to the development of secondary markets. Available online at http://www.ntca.org/content_documents/FCC_03113PFR_122903.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2003b. Triennial Review Order. Available online at http://www.fcc.gov/wcb/cpd/triennial_review.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2004a. In the matter of petition for declaratory ruling that pulver.com's Free World Dialup is neither telecommunications nor a telecommunications service. Available online at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-04-27A1.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2004b. In the matter of Vonage Holdings Corporation petition for declaratory ruling concerning an order of the Minnesota Public Utilities Commission. Available online at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-04-267A1.pdf.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2007a. Wireless Telecommunications Bureau. Available online at http://wireless.fcc.gov.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2007b. Broadband PCS Home. Available online at ttp://wireless.fcc.gov/services/index.htm?job=service_home&id =broadband_pcs.

FCC FEDERAL COMMUNICATIONS COMMISSION. 2007c. Part 68 Home. Available online at http://www.fcc.gov/wcb/iatd/part_68.html.

FOWLER, M. 2002. *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional, Boston, MA.

FRIEDEN, R. 2004. Adjusting the horizontal and vertical in telecommunications Regulation: a comparison of the traditional and a new layered approach. *Federal Communications Law Journal 55*, 207.

FRIEDEN, R. 2005. Lessons from broadband development in Canada, Japan, Korea and the United States. *Telecommunications Policy 29*, 595-613.

FROST & SULLIVAN RESEARCH. 2005. North American residential VoIP market poised to increase growth rates by entering the mass market. Available online at http://www.przoom.com/news/1781.

FROST & SULLIVAN RESEARCH. 2006. Media server equipment demand rises. Available online at http://press.xtvworld.com/article11474.html.

FTTH COUNCIL. 2005. More than two million homes now connected to next-generation broadband. Available online at http://www.ftthcouncil.org/?t=242.

GAMMA, E., HELM, R., JOHNSON, R., VLISSIDES, J. 1995. *Design Patterns*. Addison Wesley Professional, Upper Saddle River, NJ.

GANNOD, G. C., LUTZ, R. R. 2000. An approach to architectural analysis of product lines. In *ACM/IEEE International Conference on Software Engineering*, Limerick, Ireland, 548-557.

GENEIATAKIS, D., KAMBOURAKIS, G., DAGIUKLAS, T., LAMBRINOUDAKIS, C., GRITZALIS S. 2005. A framework for detecting malformed messages in SIP Networks. In *Proc. of the 14th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN),* Greece.

GILROY, A. A., KRUGER, L. G. 2005. Broadband Internet access: background and issues. Congressional Research Service. Available online at http://fpc.state.gov/documents/organization/44913.pdf.

HANDLEY, M., SCHULZRINNE, H., SCHOOLER, E., ROSENBERG, J. 1999. SIP: Session Initiation Protocol. RFC 2543, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc2543.html.

HARRIS, M. 2004. Cable finally finds its voice: PacketCable technology provides MSOs a competitive edge over ILECs. *America's Network 108*, 17, 14.

HAZLETT, T. W. 1999. Economic and political consequences of the 1996 Telecommunications Act. Available online at http://www.bandwidthmarket.com/news/consequencesTA.pdf.

HORRIGAN J. B. 2003. Information goods and services in the United States. Available online at http://www.pewinternet.org/pdfs/PIP_Info_Consumption.pdf.

IN STAT RESEARCH. 2005a. Mass migration to VoIP expected within a decade. In Stat Press Releases. Available online at http://www.instat.com.

IN STAT RESEARCH. 2005b. Media gateway sales boosted by VoIP services. In Stat Press Releases. Available online at http://www.instat.com.

INFONETICS RESEARCH. 2005a. Worldwide carrier VoIP equipment up 36% to $1.7B in 2004. Infonetics Press Releases. Available online at http://www.infonetics.com.

INFONETICS RESEARCH. 2005b. DSL port shipment up 2% in Q1 2005. Infonetics Press Releases. Available online at http://www.infonetics.com.

INFONETICS RESEARCH. 2005c. Enterprise VoIP adoption in north america starting to go mainstream. Infonetics Press Releases. Available online at http://www.infonetics.com.

INFONETICS RESEARCH. 2005d. Global next generation voice products market up 40% in Q1 2005. Infonetics Press Releases. Available online at http://www.infonetics.com.

INFONETICS RESEARCH. 2006. Capex spending by public telecom carriers up 6% in 2005. Infonetics Press Releases. Available online at http://www.infonetics.com.

JAIN, D., AND SCHMIDT, D. C. 1997. Service configurator: a pattern for dynamic configuration of services. In *Proc. of the 3rd Conference on Object-Oriented Technologies and Systems,* Portland, OR, 16-28.

JAZAYERI, M., RAN, A., LINDEN, F. 2000. *Software Architecture for Product Families*. Addison-Wesley Professional, Upper Saddle River, NJ.

JENNINGS, C., PETERSON, J., WATSON, M. 2002. Private extensions to the Session Initiation Protocol (SIP) for asserted identity within trusted networks. RFC 3325, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3325.html.

JOHN, L. K., JOEL, W. 2002. Ma Bell's orphan: US cellular telephony, 1947–1996. *Telecommunications Policy 26*, 189-203.

JOHNSTON, A. 2003. *SIP: Understanding the Session Initiation Protocol*. Artech House Publishers, Norwood, MA.

JUNIPER RESEARCH. 2006. VoIP service revenues to reach $18B in business sector by 2010. Juniper Research Press Releases. Available online at http://www.juniperresearch.com.

KENICHI, I. 2004. Internet use via mobile phone in Japan. *Telecommunications Policy 28*, 43-58.

KLEIN, J. I. 1997. Preparing for competition in a deregulated telecommunications market. In *Glassier Legal Works Seminar*, Washington, D.C.

KRUCHTEN, P. 1995. The 4+1 view model of architecture. *IEEE Software 12*, 6, 42-50.

KWEREL, E. 2001. Auctioning spectrum rights. FCC Auctions Home. Available online at http://wireless.fcc.gov/auctions/default.htm?job=auctions_home.

LAKE, M. 2005. Naked DSL: no shoes, no shirt, no service. Cnet Reviews. Available online at http://reviews.cnet.com/4520-6028_7-6215358-1.html.

LAM, K. 2004. Voice over IP industry analysis. Available online at http://www.met.utoronto.ca/met/pdffiles/K.%20Lam%20-%20040923_VoIP_ Presentation.ppt.

LAMBERT, P. 2005. Cable triple play: the VOIP card. Heaving Reading. Available online at http://www.heavyreading.com.

LANCASTER, K. J. 1996. A new approach to consumer theory. *Journal of Political Economy 74*,132-157.

MARTIN, C. R. 1998. UML tutorial: finite state machines. In *Engineering Notebook Column C++ Report*. Available online at http://www.objectmentor.com/resources/ articles/umlfsm.pdf.

MCBURNEY, P., PARSONS, S. 2002. Forecasting market demand for new telecommunications services: an introduction. *Telematics and Informatics 19*, 3.

MCCONNELL, C. R., BRUE, S. L. 2001. *Economics*. McGraw-Hill Companies, New York, NY.

MCGREW, D., RESCORLA, E. 2006. Datagram Transport Layer Security (DTLS) extension to establish keys for Secure Real-time Transport Protocol (SRTP). Internet Draft, Internet Engineering Task Force. Available online at http://tools.ietf.org/wg/tls/draft-mcgrew-tls-srtp-00.txt.

MINDEL, J. L., SICKER, D. C. 2006. Leveraging the EU regulatory framework to improve a layered policy model for US telecommunications markets. *Telecommunications Policy 30*, 136-148.

MITCHELL, K. 2004. VoIP: Vision becomes reality. Telephone Online. Available online at http://telephonyonline.com/access/analysts/telecom_voip_vision_becomes.

NEIL, G., DAVID, S., LEONARD, W. 2003. Standards in wireless telephone networks. *Telecommunications Policy 27*, 325-332.

NILS OHLMEIER. 2006. VoIP RFC watch. Available online at http://www.rfc3261.net.

NUECHTERLEIN, E. J., WEISER, J. P. 2004. *Digital crossroads: American telecommunications policy in the Internet Age*. MIT Press, Cambridge, MA.

OHRTMAN, D. F. 2002. *Softswitch: Architecture for VoIP*. McGraw-Hill Professional, New York, NY.

ONO, K., TACHIMOTO, S. 2007. End-to-middle security in the Session Initiation Protocol (SIP). Internet draft, Internet Engineering Task Force. Available online at http://www.ietf.org/internet-drafts/draft-ietf-sip-e2m-sec-04.txt.

OULU UNIVERSITY. 2006. Protos Test-Suite: c07-sip. Available online at http://www.ee.oulu.fi/research/ouspg/protos/testing/c07/sip/index.html.

OXMAN, J. 1999. The FCC and the unregulation of the Internet. Office of Plans and Policy, Federal Communications Commission. Available online at http://www.fcc.gov/Bureaus/OPP/working_papers/oppwp31.pdf.

PARK, H. Y., CHANG, S. G. 2004. Mobile network evolution toward IMT-2000 in Korea: a techno-economic analysis. *Telecommunications Policy 28*, 177-196.

REARDON. M. 2005. Verizon plays hardball on pricing. Cnet News. Available online at http://www.news.com/Verizon-plays-hardball-on-pricing/2100-1037_3-5942158.html.

PETERSON, J. 2002. A Privacy mechanism for the Session Initiation Protocol (SIP). RFC 3323, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3323.html.

PETERSON, J. 2004. Session Initiation Protocol (SIP) Authenticated Identity Body (AIB) format. RFC 3893, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3893.html.

PETERSON, J. 2005. Retargeting and security in SIP: A framework and requirements. Internet Draft, Internet Engineering Task Force. Available online at http://www3.tools.ietf.org/html/draft-peterson-sipping-retarget-00.

PETERSON, J., JENNINGS, C. 2006. Enhancements for authenticated identity management in the Session Initiation Protocol (SIP). RFC 4474, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc4474.html.

PROTOCOLS.COM. 2007. H.323 protocol suit. Available online at http://www.protocols.com/ pbook/h323.htm.

REUTERS CORP. 2004. The Reuters abridged business summary. Available online at http://finance.yahoo.com/q/pr?s=SONS.

REUTERS.COM. 2005. Vonage numbers: 1 mln customers by year-end 2005. In *ZDNet Research Blog, July*.

ROACH, A. B. 2002. Session Initiation Protocol (SIP) specific event notification. RFC 3265, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/ rfc3265.html.

ROSENBERG, J., SCHULZRINNE, H. 2002a. Reliability of provisional responses in Session Initiation Protocol (SIP). RFC 3262, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3262.html.

ROSENBERG, J., SCHULZRINNE, H. 2002b. Session Initiation Protocol (SIP): locating SIP Servers. RFC 3263, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3263.html.

ROSENBERG, J., SCHULZRINNE, H. 2002c. An offer/answer model with Session Description Protocol (SDP). RFC 3264, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3264.html.

ROSENBERG, J., SCHULZRINNE, H., CAMARILLO, G., JOHNSTON, A., PETERSON, J., SPARKS, R., HANDLEY, M., SCHOOLER, E. 2002. SIP: Session Initiation Protocol. RFC 3261, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3261.html.

ROSENBERG, J. 2006. Coexistence of P-Asserted-ID and SIP identity. Internet Draft, Internet Engineering Task Force. Available online at http://tools.ietf.org/wg/sip/draft-rosenberg-sip-identity-coexistence-00.txt.

ROSENBERG, J. 2007. Obtaining and using Globally Routable User Agent (UA) URIs (GRUU) in the Session Initiation Protocol (SIP). Internet Draft, Internet Engineering Task Force. Available online at ftp://ftp.rfc-editor.org/in-notes/internet-drafts/draft-ietf-sip-gruu-15.txt.

ROSENBERG, J., CAMARILLO, G., WILLIS, E. D. 2007. A Framework for consent-based communications in the Session Initiation Protocol (SIP). RFC 4453, Internet Engineering Task Force. Available online at http://www.ietf.org/rfc/rfc4453.txt.

ROSENBERG, J., JENNINGS, C. 2008. The Session Initiation Protocol (SIP) and Spam. RFC 5039, Internet Engineering Task Force. Available online at http://tools.ietf.org/html/rfc5039.

RUDI, B. BART, V., JAN, S. 2002. Intellectual property rights and standardization: the case of GSM. *Telecommunications Policy 26*, 171-188.

SCHMIDT, D. C., SUDA, T. 1994. Measuring the impact of alternative parallel process architectures on communication subsystem performance. In *Proceedings of the 4th International Workshop on Protocols forHigh-Speed Networks*, Vancouver, British Columbia, 103–118.

SCHMIDT, D. C., STAL, M., ROHNERT, H., BUSCHMANN, F. 2000. *Pattern-oriented software architecture, Volume 2, Patterns For Concurrent And Networked Objects*. John Wiley & Sons, Inc., New York, NY.

SHAW, M., GARLAN, D. 1996. *Software Architecture: Perspectives on an Emerging Discipline*. Prentice-Hall, Inc., Upper Saddle River, NJ.

SONUS NETWORKS Inc. 2006. Available online at http://www.sonusnet.com.

TELECOMMUNICATIONS INDUSTRY ASSOCIATION (TIA). 2006. Spending in U.S. telecommunications industry rises 8.9% in 2005 reaching $856.9 Billion. Available online at http://www.tiaonline.org.

UNIVERSALFUND.ORG. 2007. Universal Service Fund. Available online at http://www.usac.org/about/universal-service.

VOIPSA. 2005. VoIP security and privacy threat taxonomy. Voice over IP Security Alliance. Available online at http://www.voipsa.org/Activities/VOIPSA_Threat_Taxonomy _0.1.pdf.

VOLTER, M., SCHMID, A., WOLFF, E. 2002. *Server Component Patterns: Component Infrastructures Illustrated with EJB*. John Wiley & Sons, Inc., New York, NY.

WATSON, M. 2002. Short term requirements for network asserted identity. RFC 3324, Internet Engineering Task Force. Available online at http://www.faqs.org/rfcs/rfc3324.html.

WINSTON, B. 2002. *Media Technology and Society*. Routledge, New York, NY.

WU, Y., BAGCHI, S., GARG, S., SINGH, N., TSAI, T. 2004. SCIDIVE: A stateful and cross protocol intrusion detection architecture for voice-over-IP environments. In *International Conference on Dependable Systems and Networks (DSN'04)*, Florence, Italy, 443-442.

ZIMMERMANN, P., JOHNSTON, A., CALLAS, J. 2006. ZRTP: extensions to RTP for Diffie-Hellman key agreement for SRTP. Internet Draft, Internet Engineering Task Force. Available online at http://tools.ietf.org/wg/avt/draft-zimmermann-avt-zrtp-02.txt.

APPENDIX A

CONTENT OF SIP MESSAGES

```
INVITE sip:alice@tamu.edu SIP/2.0
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKnashds8
Max-Forwards: 70
To: Alice <sip:alice@tamu.edu>
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 INVITE
Contact: <sip:bob@server33.tamu.edu >
Content-Type: application/sdp
Content-Length: 142

v=0
o=Bob 53655765 2353687637 IN IP4 server33.tamu.edu
s=Session SDP
t=0 0
c=IN IP4 10.1.3.6
m=audio 3456 RTP/AVP 0 1 3 99
a=rtpmap:0 PCMU/8000
```

The 302 response contains:

```
SIP/2.0 302 Moved Temporarily
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKnashds8
To: Alice <sip:alice@tamu.edu>;tag=10435993456
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 INVITE
Contact: sip:alice@pc33.tamu.edu
Content-Length: 0
```

The ACK request contains:

```
ACK sip:alice@tamu.edu SIP/2.0
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKnashds8
Max-Forwards: 70
To: Alice <sip:alice@tamu.edu>;tag=10435993456
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 ACK
Contact: <sip:bob@server33.tamu.edu >
Content-Length: 0
```

The updated INVITE contains:

```
INVITE sip:alice@pc33.tamu.edu SIP/2.0
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKnashds8
```

```
Max-Forwards: 70
To: Alice <sip:alice@tamu.edu>
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 INVITE
Contact: <sip:bob@server33.tamu.edu >
Content-Type: application/sdp
Content-Length: 142

v=0
o=Bob 53655765 2353687637 IN IP4 server33.tamu.edu
s=Session SDP
t=0 0
c=IN IP4 10.1.3.6
m=audio 3456 RTP/AVP 0 1 3 99
a=rtpmap:0 PCMU/8000
```

The 200 OK contains:

```
SIP/2.0 200 OK
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKnashds8
Max-Forwards: 70
To: Alice <sip:alice@tamu.edu>
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 INVITE
Contact: <sip:bob@server33.tamu.edu >
Content-Type: application/sdp
Content-Length: 142

v=0
o=Alice 53655785 2353687684 IN IP4 pc33.tamu.edu
s=Session SDP
t=0 0
c=IN IP4 106.52.21.5
m=audio 6546 RTP/AVP 0 1 3 99
a=rtpmap:0 PCMU/8000
```

The final ACK request contains:

```
ACK sip:alice@pc33.tamu.edu SIP/2.0
Via: SIP/2.0/UDP server33.tamu.edu:5060;branch=z9hG4bKn009
Max-Forwards: 70
To: Alice <sip:alice@tamu.edu>;tag=10435993456
From: Bob <sip:bob@tamu.edu>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 1 ACK
Contact: <sip:bob@server33.tamu.edu >
Content-Length: 0
```

APPENDIX B

A LONG DISTANCE CALL SETUP EXAMPLE

In this section, we demonstrate the long distance call setup process and examine the detailed roles of and interactions between Sonus network components in a carrier environment. A simple example is call set up between two circuit-based end office switches, one in Boston and the other in Chicago. As a side-to-side comparison, we first illustrate the call setup process in a traditional circuit-switched SS7 network.

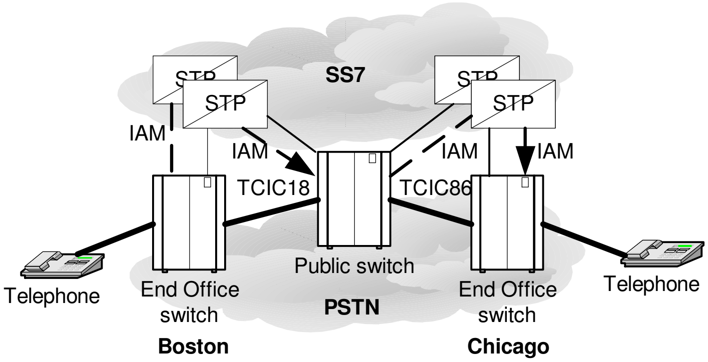

Fig. 29. A traditional circuit-switched network call setup process.

Fig. 29 shows a traditional circuit-switched network arrangement within the public switched telephone network (PSTN). Calls destined for subscribers outside the service area of an end office switch are passed to a tandem switch. The call traffic is carried over the trunks that connect the switches. The switches are also connected to a separate SS7 network using A-links (access links) to Signaling Transfer Points (STPs). The call signaling messages travel through the SS7 network (shown by dotted lines) and the actual call path (voice or data payload, represented by a thick line) is established through the PSTN.

(1) The caller goes off hook and dials the destination phone number. The Boston end office switch detects the seizure when the caller goes off hook and collects the dialed digits.

(2) The Boston end office switch reserves a circuit using TCIC (Trunk Circuit Identification Code) and sends an IAM message (Initial Address Message) to the tandem switch through the STP (Signaling Transfer Point). The IAM message contains the CPN (Called Party Number) as well as the TCIC to be used.

(3) The tandem and the Boston end office switches now have TCIC 18 reserved.

(4) Following the same procedure, the tandem switch and the Chicago end office reserves TCIC 86.

(5) The Chicago end office switch seizes the line of the called party and applies ringing voltage.

(6) An ACM (Address Complete Message) is sent in the backward direction indicating that all the address signals required for routing the call to the called party have been received. The voice path is cut through in a backwards direction so the caller can hear ring back tone provided by the far end office switch.

(7) The call is answered and an ANM message (Answer message) is sent in the backwards direction indicating the call has been answered.

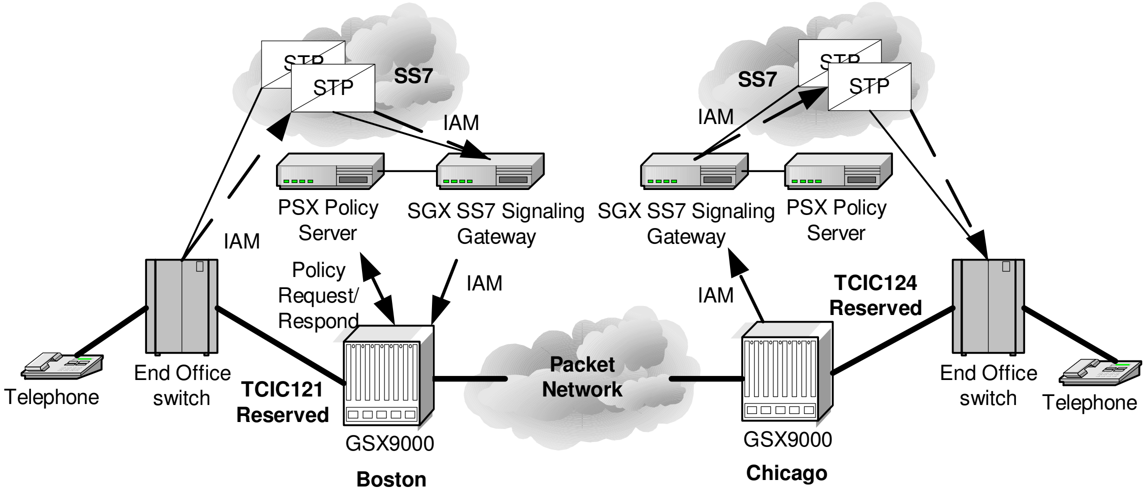

Fig. 30. A VoIP call setup process.

Fig. 30 shows a VoIP call setup process. A call is established from Boston to Chicago. The call is routed over an ISUP trunk group into a GSX located in a Boston POP (point of presence). Instead of handing the call off to a long distance carrier, the Boston GSX routes the call out on a packet port over a packet network to the Chicago POP. The Chicago GSX then routes the call out on an ISUP trunk group back to the PSTN to reach the end destination. Again, the dotted lines represent signaling messages and the thick, solid lines represent the actual call path (voice and data payload).

(1) The Boston end office and the GSX reserves TCIC 121 following the same first three steps as described in the traditional circuit-switched network call set up process, except that the IAM message is passed to the GSX through the SGX over the management network.

(2) The GSX asks the PSX “What do I do with this call?” in a policy request. The GSX provides the contents of the IAM message, including the called party and calling party phone number. The PSX analyzes these digits and makes routing decisions based on the requirements of the call. The PSX also traverses a Service Selection Graph (SSG) to determine if special services apply to the call. In this example, the PSX finds a routing label with one route: a gateway named Chicago and a trunk group named CHGGSX_TG1 on that GSX.

(3) The PSX sends a policy response to the Boston GSX, instructing it to route the call to the Chicago GSX at IP address 10.8.2.100, and then send the call out on trunk group CHGGSX_TG1 off that GSX.

(4) The Boston GSX and Chicago GSX exchange signaling messages over the packet network using a proprietary signaling protocol or SIP.

(5) The Chicago GSX selects TCIC 124 from trunk group CHGGSX_TG1 for the call and sends out an IAM message to the end office through the SGX. Now, TCIC 124 is reserved on both the end office switch and the Chicago GSX.

(6) When the switch serving the destination applies ringing voltage to the destination phone, it sends an ACM (Address Complete Message) back to all the switches involved in the call, telling them to cut through the voice path in a backward direction. This allows the caller to hear ringback tone generated by the last switch involved in the call (the one serving the destination). This signaling is performed point-to-point: between the last switch and the Chicago GSX, between the two GSXs, and between the Boston GSX and the first end office switch. Note

the path and form the signaling takes (the ACM message is passed over the SS7 network between each GSX and an office switch, and proprietary gateway-to-gateway signaling is passed over the packet network between the two GSXs).

(7) The call is answered and an ANM (Answer Message) is sent in the backward direction indicating that the destination has gone off-hook.

(8) During the call, voice packets are exchanged between the two GSXs using the negotiated physical port IP address. The codec used for the voice traffic was negotiated earlier between the two GSXs.

VITA

Xu Yang is currently a senior software engineer at Sonus Networks Inc, where he has worked on the SIP protocol stack, and related application and feature development for VoIP and IMS systems.

Prior to this, Xu Yang was a graduate research assistant at the Real Time System Lab within the Department of Computer Science, Texas A&M University, where he conducted research on secure election algorithms in peer-to-peer networks.

From 1999 to 2001, Xu Yang worked at Alcatel USA intelligent network R&D department, where he worked on database replication and SIP user agent demo, and won the Alcatel Night Out Award. Xu Yang also worked for China Mobile (Shenzhen) from 1997 to 1998.

Xu Yang earned his Master's Degree in Computer Science from the University of Texas at Arlington in 1999 and Bachelor's Degree in Computer Communications from Beijing University of Posts and Telecommunications in 1997.

Xu Yang currently lives in Boston, Massachusetts. In his spare time, he enjoys various outdoor activities such as tennis, soccer and hiking in the beautiful Massachusetts state parks and skiing in many of the gorgeous New England ski resorts.

Xu Yang may be reached at Sonus Networks, Inc., 7 Technology Park Drive, Westford, MA, 01886. His email is clakxuyang@gmail.com.