SESSION MANAGEMENT IN ADVANCED TELECOMMUNICATION SERVICES

Anand Ranganathan
IBM T.J. Watson Research Center, USA

Dipanjan Chakraborty
IBM India Research Lab, India

Keywords: Sessions, PSTN, SS7, H.323, SIP.

Contents

Summary

The end of the 20th century and the beginning of the 21st century has seen incredible

innovations in the telecommunications industry. Increasing numbers of people are getting connected to one another using different kinds of devices and applications that exchange data in different kinds of media and data formats. Most synchronous communication between people occurs in the context of a session, which contains a group of logically related messages that are exchanged between the parties in the session. Hence, an important part of any telecommunication technology is the management of sessions, i.e. how sessions are established, modified and destroyed. The recent explosion of the telecommunications revolution has seen a great increase in the number of operations that can be performed on sessions. In this section, we trace the evolution of session maintenance protocols from circuit switched networks to packet switched networks, and finally to a convergence of telecommunications and the Internet, through SIP.

1. Introduction

Session management technologies for telecommunication networks and services started getting defined quite early on – with the spread of telephones (and the telecommunication network) across the globe. The core backend network infrastructure of a telephone network is primarily referred to as circuit-switched network. The reason for calling it "circuit-switched" is because, in each connection, a dedicated "circuit" or "channel" is reserved from the caller's equipment to the equipment of the person called (callee). This "setup" of the call to reserve a dedicated circuit is done through protocols termed as "signaling" protocols. These protocols (e.g. SS7) actually participate in creating a "session" between the caller and the callee. SS7 (Signaling System No. 7) is currently used as a signaling/session management system in telephone circuit-switched networks.

The Internet on the other hand is a parallel network that saw rapid growth in the 1980s and by 1990s; the Internet had already spread to the households, as the primary network for data communication (email, messaging, and websites). This was initially implemented by DARPA (Defense Advanced Research Project Agency, an organization of Department of Defense, USA) and used to be called the ARPANET (Advanced Research Projects Agency Network). The ARPANET is commonly referred to as the "Father of the Internet". The primary entity being transmitted over this network was data. The primary means of communicating data turned out to be by sending "packets" of data across the physical wires (copper, fiber) and a stack of protocols emerged (TCP/IP) that manage the proper collection of this data and deliver them to the computer applications. This network is primarily referred to as "packet switched" network.

With the rapid spread of the Internet, an amalgamation of the services offered by both the networks (telecom and Internet) started happening. Voice enthusiasts, primarily belonging to the telecom network, realized the need of supporting additional services (like video, data) on the existing telecom network. In 1984, CCITT (International Consultative Committee on Telephony and Telegraphy) set forth to develop a standard architecture to support fully digitized circuit-switched telephone system (known as ISDN) to support both voice and non-voice communication. Similarly, a need to be able to support voice communication over packet-switched network (Internet) arose too.

The need to allow for a “node” residing on a packet switched network to interact with a “node” (e.g. Internet Phone services) residing on a circuit-switched network became more and more apparent. This gave birth to a set of protocols that are tuned to setup paths over this packet switched network, and a set of gateways that interface between the packet-switched and the circuit-switched network. However, the philosophies of the underlying transmission mechanism (one is circuit switched, the other is packet switched) is different in these networks. In short, circuit switched networks are much more reliable, fast and provide much higher clarity of transmission. One of the major challenges in packet switched networks on the other hand has been on how to minimize the delay, how to reduce errors due to lost packets etc. Moreover, since this communication is real-time, it has to be given more priority than other communication happening over the same path. H.323 and SIP are some of the standard session/signal management protocols over these packet switched networks.

SIP has gathered a lot of momentum over the last five years, primarily due to its light-weight nature (compared to H.323) and due to the fact that it’s an application layer protocol, it has the capability of setting up connections virtually across any underlying network.

Since this chapter is dedicated to session maintenance, we first give an overview of signaling and session management protocol in traditional circuit-switched networks (also referred to as Plain Old Telephone Service or Public Switched Telephone Network). In particular, we cover SS7 in more details. We then devote the rest of the chapter to SIP and provide summaries of some cutting edge work that is happening in this area.

2. Overview of Session Maintenance in Circuit Switched Networks (PSTN)

During the initial days of globalization of telephony, the infrastructure used to be based on circuit switched networks. Several protocols/standards have been used for establishing and controlling sessions on Plain Old Telephony Service (POTS) or Public Switched Telephone Network (PSTN) backbone, each of which probably deserves a whole book by itself. We attempt to summarize the information and present Signaling System No. 7 (SS7) in a bit more details.

Signaling is the basis of starting a session between a caller and a callee in telephone networks. Signaling protocols are used to establish an interconnection of the subscriber to the service (e.g. calling a phone number) through telephone network elements (like the switch). Signaling also gives status information (like dial tone, busy tone, ringing tone) to the subscriber. Signaling is of 3 broad categories. (1) Supervisory signaling (2) Address signaling (3) Audio-visual signaling. Supervisory signaling provides line or circuit related information required for a session (voice call) to be established. It gives such information to both the subscribers as well as the switches. For example, it tells the caller whether the callee phone is “off hook” or “on hook”. Switches that form the “control points” inside a PSTN network also uses supervisory signaling to know such information. Address signaling gives information like whether the end point has rotary dial or whether there are push buttons on the end equipment. Audio-visual signaling informs the caller regarding current progress of a call, like ringing, paging, whether a busy tone is going on. Note the subtle difference in supervisory signaling and audio-visual signaling. Supervisory signaling provides information on "whether a phone is busy or idle". Audio-visual signaling provides information whether a call that has been established till the callee is receiving a busy tone or a ringing tone.

Signaling information is transmitted through several means like pulse duration (each duration carries a different meaning), pulse combination, signal frequency as well as binary codes.

E and M signaling protocols are forms of supervisory signaling used between the trunk-signaling equipment and the switch. The trunk-signaling equipment is used to transmit supervisory signals (user "on-hook" or "off-hook") over the trunk line. E-leads carry signals to the switching equipment and m-leads carry signals back to the trunk-signaling equipment.

Supervisory signaling (or line signaling) has limits on distance that it can travel. Alternating Current (AC) signaling is used to extend these limits. AC signaling is further divided into low-frequency, in-band and out-band signals. We leave out further details of these signals in this chapter.

Moving to digital telephony (voice transmitted using digital encoding techniques); it started with initial development of Pulse Code Modulation (PCM) technology. PCM represents an analog signal using digital sampling. The wave (say a sine wave) is sampled at regular intervals and represented in a binary code, and transmitted over to build the wave at the receiver. PCM serves as the basic technology behind speech telephony. With the evolution of data transmission, telephone network later turned out to be the default network for other services such as facsimile (fax) and data transmission using modems. Integrated Services Digital Network (ISDN) is a system that has been developed to easily integrate various forms of services, allowing, voice, video and data to be transmitted at the same time. This system allows digital transmission over analog copper wires, making transmission more efficient.

2.1. ISDN Session Channels

ISDN primarily consists of two types of channels. (1) the B-Channel (or the bearer channel) (2) the D-Channel. B channels are used for basic user-to-user communication for voice, data, video etc. D channels are used for transmitting control information. Hence D channels carry circuit-switching control information for its related B channels. There are various bundlings of B and D channels available through out the world. However, there are two broad varieties (1) Basic rate interface (BRI) (2) Primary rate interface (PRI).

BRI (also known as Basic Rate Access – BRA) consists of 2B + 1D channel. The D channel has a transmission capacity of 16 kbits/s. The B channels can be used independently and simultaneously. The PRI variety has several B channels (depends on country) and 1 D channel of 64 kbits/s. For example, North America and Japan has 23 B channels with an aggregate bit rate of 1.544 Mbits/s. Whereas Europe and Australia has

30 B channels bundled with 1 D channel, giving an aggregate bit rate of 2.048 Mbits/s. So, for both the BRI and the PRI varieties, the D channel is the primary control/session establishment channel. D channel information initiated from the user's terminal is later converted to CCITT Signaling System No 7 (SS 7) at the first available Signal Transfer Point (STP). There after SS7 is used to transfer the control information through multiple STPs till it reaches the end terminal (receiver). SS7 is a standard that we cover in bit more details later in this chapter. Thus D channel is responsible for call setup and termination, supervision and all functions dealing with session control. ISDN is conceptually related to the OSI (Open Systems Interconnect – a model on which Internet has been developed) model too. D channels conceptually map to the first 3 layers of the OSI stack (Physical, Data Link, Network layer). You might also hear about B-ISDN. B-ISDN or Broadband ISDN is also another network that was designed to support end-to-end voice and asynchronous data services using Asynchronous Transfer Mode (ATM) as the underlying technology. It has been superseded by Internet.

2.2. Common Channel Signaling System Number 7 (SS7)

In this section, we go into SS7, a predominant standard SS7 used for session control over digital (and analog) PSTN networks. SS7 is a global standard for telecommunications defined by International Telecommunications Union's Telecommunication Standardization Sector (ITU-T). This is the primary protocol that is used by the PSTN network elements to exchange control information over a *digital* signaling network for call setup, call routing etc. Hence, the signaling part is *digital* even though the actual call may use analog voice transmission. Apart from basic call setup and tear down, it is used as a control protocol for other wireless services like roaming, mobile subscriber authentication, local number portability (LNP), toll-free and wire line services, as well as enhanced call features like call forwarding, three-way calling etc.

SS7 control messages are sent over "signaling links". Signaling occurs out-of-band meaning, there is a separate channel for control messages and it's not mixed with voice communication channels. It should be noted that even though SS7 forms the backbone for transmitting ISDN session information, SS7 is by itself a standalone entity in itself that currently dominates signaling in PSTN networks. As such, the whole architecture of SS7 is totally dedicated to signaling. SS7 has been designed to operate on digital networks. The signaling protocol can meet the present and future requirements of information transfer and is geared towards providing loss-less and duplication-free transmission of information.

2.2.1. SS7 Architecture Overview

An SS7 network consists of *signaling points (SP)*. Each signaling point is identified by a numeric point code. SPs are primarily of three categories: (1) Service Switching Point (SSP) (2) Signal Transfer Point (STP) (3) Service Control point (SCP). SSPs are the points that are directly connected to your phone through voice trunks. They are responsible for originating and terminating calls. SSPs interact with other SSPs for call setup and tear down. STPs sit in between SSPs (unless two SSPs are directly connected) to route network traffic. Hence STPs interact with SSPs (as well as SCPs) and help in message forwarding. STPs interact with other components of the architecture using packet switching. This eliminates the need for direct lines between SSPs. SCPs are essentially data bases that contain information on how to route a call, i.e. information to determine the path of a call. SSPs interact with SCPs to determine the message route for establishing a voice call.

2.2.2. Relationship to OSI Stack and Layer Overview

SS7 stack resembles OSI stack up to layer 3 (network layer). CCITT working groups who were standardizing SS7 were mainly concerned with delay in setting up the call – one of the primary reasons for SS7 to be there. CCITT Rec.Q.709 specifies that 95% of all calls should not have a post dial delay of more than 2.2s. Hence a limit was placed on the number of STPs that a particular call setup signal can traverse.

All these requirements called for "minimal" needed functionality in SS7 for achieving its purpose. Figure 1 shows the relationship of the SS7 stack to the OSI model. Note that this is a loose mapping. Conceptually, SS7 system consists of (1) Message Transfer Part and (2) User Part. Message Transfer Part (MTP) deals with hardware and software components needed to build the core signaling substrate. User Part conceptually represents different types of end-users (e.g. Telephone User, ISDN user etc) that can use the system. As such, SS7 defines three user parts: (1) Telephone User (TUP) (2) Data User (DUP) (3) ISDN User (ISUP). Apart from these, there is another subcomponent called the Signaling Connection Control Part (SCCP). It sits on top of the core MTP layers. It provides connection-less and connection-oriented services to users and applications. We briefly describe the functions of each layer in the SS7 stack.

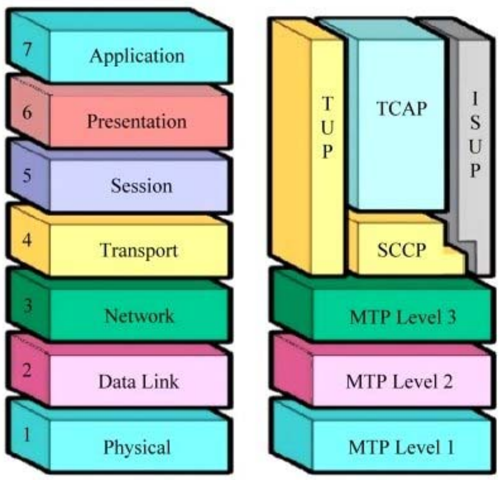

Figure 1: Relationship of SS7 to the OSI model

Message Transfer Part Level 1 (MTP 1) defines the physical, electrical and functional characteristics of the signaling data link. It also specifies means to access it. The connectivity is of 64 kbits/s. Each data link consists of a bi-directional transmission path for signaling consisting of 2 data channels in opposite directions. They connect digital switches (SSPs) and form the basic physical channel of communication between SS7 architecture components.

Message Transfer Part Level 2 (MTP 2) deals with end-to-end transfer of signaling messages from one signaling point to another. It ensures accurate transmission of a message across a signaling link. MTP 2 implements several functions like signaling unit delimitation, signal unit alignment, error detection, error correction, signal link error monitoring and flow control. It is equivalent to the OSI data link layer. Each signaling message consists of variable-length signal units. The signal unit delimitation and alignment part takes care of identifying the signal units through unique 8 bit patterns (referred as flag). A Cyclic Redundancy Check (CRC) of 16 bits is appended to each signal unit. This is used to detect errors. Error correction protocols in SS7 MTP 2 consider various kinds of signaling links (intercontinental links or local links) to decide the appropriate means of correction. Flow control deals with handling congestion over the established channel between two signaling points.

Message Transfer Part Level 3 (MTP 3) is equivalent to OSI network layer. It ensures that a signaling message originated from a particular user part is delivered to the corresponding user part at the destination point. The signaling could be done directly (SSP to SSP) or via an STP. This layer takes care of signal message routing. This layer also decides on the particular signaling link to be used and hence does load sharing. MTP 3 also does network management where it handles loss of signaling link, loss of signaling points and operation degradation due to congestion.

Hence, these core set of STPs are used to set up domestic as well as international calls. International calls go through International Signaling Points (ISP) that connects to National Signaling Points (NSP). Conceptually, both ISP and NSP can be co-located.

2.2.3. Signaling Connection Control Part (SCCP) and User Parts

The SCCP sit on Layer 4 above MTP Level 3 and provides the SS7 end points with additional services such as support for connection-less and connection-oriented network services. It supports transfer of both circuit-related as well as non-circuit-related signal information. It thus essentially supports logical signaling over the SS7 network. This is used by some user parts like the ISDN user part for some of the services offered by ISDN. Four classes of services provided by SCCP are (1) basic connectionless service (2) sequenced connectionless service (3) basic connection-oriented service (4) flow control connection-oriented service. Connection oriented services further are classified into (1) Temporary signaling connection (2) Permanent signaling connection. Temporary signal connections are analogous to regular dialed telephone connections made every day. Permanent signaling connections are different and are used in leased telephone lines. Connection-oriented services offered by SCCP include setting up the connection, data transfer and tearing down the connection. The connection establishment phase decides on the type of operating functions to be used during data transfer. Examples of such data transfer functions are flow control, connection identification, segmenting/reassembly (at the logical level) option etc.

Coming to the other User Parts, the Telephone User Part (TUP) is used to manage analog circuit call setup and tear down. The ISDN User Part (ISUP) on the other hand defines protocols to set up, manage and tear down circuits that carry voice and data between two end points. Note that ISUP is used for ISDN as well as non-ISDN calls. Essentially, whenever calls go out-of-switch, then ISUP is used to establish a trunk connection between the caller's SSP and callee's SSP. It encompasses standards that are used to send back an acknowledgement when an user picks up a ringing phone as well as procedures for hanging up (tearing down) a connection.

2.2.4. Performance Parameters

Circuit switched networks has to support strict guarantees for performance of the network. SS7 performance is measured by (1) Message Delay (2) Signaling traffic load (3) Error rate. Typical examples of performance guarantees are: The unavailability of a signaling route set should not be more than 10 min/year. Only 1 out of 10^{10} signal units can go undetected. Similarly, 1 out of 10^{7} messages is allowed to be lost in this network. This gives an idea of how strict the enforcement guarantees are on these backbone networks that support our day-to-day voice calls.

2.2.5. SS7 Signals over the Internet

With the outbreak of the Internet, there has been a spurt of activity to support transport of SS7 signals on top of IP. Internet Engineering Taskforce (IETF) set up a working group called Sigtran that develops standards for transport of PSTN signals (read SS7) on top of IP. RFC (Request for Comments) 2719 defines the protocol specifics. Sigtran consists of (1) The standard IP network (2) common channel signaling transport layer (3) an adaptation layer. There has been several standards in the adaptation layer like M2PA (MTP Layer 2 peer-to-peer adaptation), M3UA (MTP Layer 3 User adaptation) etc. Detailed documentation regarding these can be found in the corresponding IETF RFCs, starting with RFC 2719.

-
-
-

Bibliography

A. Roach (2002), Session Initiation Protocol (SIP)-Specific Event Notification, Request for Comments 3265, Internet Engineering Task Force. [IETF request for comments describing an extensible framework by which SIP nodes can request notification from remote nodes indicating that certain events have occurred]

Alan B. Johnston (2004), SIP: Understanding the Session Initiation Protocol, Second Edition (published by Artech House) [Book describing the SIP protocol and its use in IP telephony]

G. Sidebottom, K. Morneault, J. Pastor-Balbas (2002) Signaling System 7 (SS7) Message Transfer Part 3 (MTP3) - User Adaptation Layer (M3UA), Internet proposed standard RFC 3332. [This Request for

Comments defines a protocol for supporting the transport of any SS7 MTP Level 3 to User (ISUP, SCCP) signaling]

Gonzalo Camarillo (2002), SIP Demystified (published by McGraw-Hill) [Book describing the use of SIP, architectures based on SIP and how SIP interacts with other protocols]

H. Lei, and A. Ranganathan (2004) Context-aware unified communication. In Proceedings of 5th International Conference on Mobile Data Management (MDM), Berkeley [Paper describing use of SIP for context-aware unified communication across heterogeneous devices]

H. Schulzrinne et al (2003) Real-time Transport Protocol, RFC 3550, 3551. [Request for Comments that defines the specification of Real-time Transport Protocol, a protocol used for real-time data transfer over IP]

Henning Schulzrinne and Elin Wedlund (2000) Application-Layer Mobility using SIP Mobile Computing and Communications Review (MC2R), Volume 4, Number 3, July 2000. [Paper describing how SIP (can help provide terminal, personal, session and service mobility to applications ranging from Internet telephony to presence and instant messaging]

Henry Sinnreich and Alan B. Johnston (2001) Internet Communications Using SIP: Delivering VoIP and Multimedia Services with Session Initiation Protocol (published by John Wiley & Sons) [Book describing use of SIP for different kinds of multimedia communication sessions and available vendor offerings and services in this area]

Hong Liu and Petros Mouchtaris (200) Voice over IP Signaling: H.323 and Beyond, IEEE Communications, October 2000. [Comparison of H.323 with SIP and looking at the next steps of evolution of session maintenance]

Internet Engineering Task Force (2001), An Application Server Component Architecture for SIP [An IETF Draft describing the structure of an application server for providing advanced call management features]

ITU-T (2003), "One way Transmission Time" ITU-T Rec. G.114, May, 2003, [An ITU-T recommendation that defines performance parameter definitions for quality of Speech and other real-time data applications over IP Networks]

J. Glassmann et al. Service Architectures in H.323 and SIP – A Comparison,. IEEE Communication Surveys and Tutorials [A document comparing architectures of H.323 and SIP]

J. Rosenberg et al. (2000) SIP Extensions for Instant Messaging. Internet Draft, Internet Engineering Task Force, June 2000. [IETF Draft describing a set of extensions to SIP which allow instant messaging]

J. Rosenberg et al. (2000) SIP Extensions for Presence. Internet Draft, Internet Engineering Task Force, June 2000. [IETF Draft describing SIP extensions for providing user presence information]

J. Rosenberg, H. Schulzrinne (2000) A Framework for Telephony Routing over IP, Internet Informational RFC 2871 [A Request for Proposal that outlines the specification of Session Initiation Protocol]

J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler (2002), SIP: Session Initiation Protocol. Request for Comments 3261, Internet Engineering Task Force. [The next version of IETF Request for comments describing SIP]

K. Morneault, R. Dantu, G. Sidebottom, B. Bidulock, J. Heitz, (2002) Signaling System 7 (SS7) Message Transfer Part 2 (MTP2) - User Adaptation Layer, Internet proposed standard RFC 3331 [A Request for Comments that defines elements of the session control protocol called SS7]

L. Ong et al (1999), Framework Architecture for Signaling Support, RFC 2719 [Comparison of H.323 with SIP and looking at the next steps of evolution of session maintenance]

M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg (1999). SIP: Session Initiation Protocol. Request for Comments 2543, Internet Engineering Task Force. [The original IETF Request for comments describing SIP]

Nortel (2001) SIP and the new network communications model [White paper describing use of SIP in different kinds of applications]

Open Systems Interconnect Architecture, International Organization for Standardization, www.iso.org [A standardization body that specified the OSI stack]

R. Sparks. (2000) SIP Call Control. Internet Draft, Internet Engineering Task Force, July 2000. [IETF Draft describing a set of extensions to SIP which allow for various call control services]

Roger L. Freeman Telecommunication System Engineering, Wiley publishers, 4th Edition [This is an extensive study of the whole gamut of telecommunication systems, starting from its evolution, but focused more on pre-SIP state-of-the-art in telecommunication session maintenance]

T. George, B. Bidulock, R. Dantu, H. Schwarzbauer, K. Morneault, (2005) Signaling System 7 (SS7) Message Transfer Part 2 (MTP2) - User Peer-to-Peer Adaptation Layer (M2PA), Internet proposed standard RFC 4165 [A Request for Comments document that defines the peer-to-peer components (MTP Level 3) of SS7]

W. Marshall et al. (2000) SIP Extensions for Caller Identity and Privacy [IETF Draft describing how SIP can help preserver privacy]

Biographical Sketches

Anand Ranganathan is a Research Staff Member at IBM TJ Watson Research Center. He finished his PhD at the Department of Computer Science in the University of Illinois at Urbana-Champaign in 2005, under the supervision of Prof. Roy H. Campbell. He received his B.Tech in Computer Science and Engineering from the Indian Institute of Technology in Madras in 2000.

He is currently working in the Next Generation Distributed Systems group at the IBM TJ Watson Research Center in Hawthorne NY. As part of his research, he is exploring various aspects of the Semantic Web, mobile and distributed systems and information retrieval. His broad research interests include pervasive computing, mobile computing, context-aware computing, autonomic computing, middleware, artificial intelligence and social networks. His thesis was on the confluence of the areas of pervasive and autonomic computing.

Dipanjan Chakraborty is a Research Staff Member at IBM India Research Lab. He received his Ph.D. in Computer Science from University of Maryland, Baltimore County (UMBC) in 2004.

His research is in the areas of mobile and pervasive computing environments, next generation network protocols and management, peer-to-peer systems with special interests in the fields of service discovery, information aggregation and composition, ad-hoc/sensor networks and application-centric routing. He is also working in the area of business process management. His thesis is in the area of service discovery and composition for pervasive environments. He received a fellowship grant from IBM during the 3 years of his Ph.D candidacy.