

1476

# **VIDEO-ON-DEMAND SESSION MANAGEMENT**

*Dissertation Submitted to*  
**JAWAHARLAL NEHRU UNIVERSITY**  
*in partial fulfilment of requirements*  
*for the award of the degree of*  
**Master of Technology**  
*in*  
**Computer Science & Technology**

*by*

**RAM SHAKLA VERMA**

72p + fig + Appendix



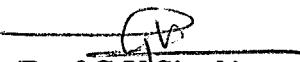
Jawaharlal Nehru University

**SCHOOL OF COMPUTER & SYSTEMS SCIENCES**  
**JAWAHARLAL NEHRU UNIVERSITY**  
**NEW DELHI - 110 067**  
*January 1997*

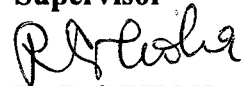
## CERTIFICATE

This is to certify that the dissertation entitled **VIDEO-ON-DEMAND SESSION MANAGEMENT** being submitted by **RAM SHAKLA VERMA** to School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi, in partial fulfilment of the requirements for the award of the degree of Master of Technology in Computer science, is a bonafide work carried by him under the guidance and supervision of **Dr.R.C.PHOHA**. This work has not been submitted elsewhere for any other purpose.

Dean ,SCSS

  
(Prof.G.V.Singh)  
SC&SS, J.N.U.  
New Delhi 110067

Supervisor

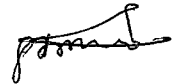
  
(Dr.R.C.PHOHA)  
SC&SS, J.N.U.  
New Delhi 110067

## ACKNOWLEDGEMENT

I wish to express my profound sense of gratitude to Dr. R.C. Phoha, SC&SS, JNU, under whose invaluable guidance and incessant encouragement, my work has taken its present shape.

I wish to express my sincere thanks to the staff members of the computer laboratory, SC&SS for providing me with all the facilities required during the project.

Finally I would like to thank all my friends who cooperated me throughout the work.



( RAM SHAKLA VERMA )

# CONTENTS

1. Introduction	1-3
1.1 Introduction	
1.2 Organization of chapters	
2. VOD Network Architecture	4-16
2.1 System Elements	
2.1.1 Client's set-top terminal	
2.1.2 Access network	
2.1.3 IVN manager	
2.1.4 ATM network	
2.1.5 Video Server	
3. VOD Sessions	17-29
3.1 Session Concept & Resources	
3.2 Basic Operation of VOD Session	
3.3 Hierarchy of Session Management	
3.4 User activity Model	
4. A Multiple-Call Session Model	30-43
4.1 Transition Into Service-2 State	
4.2 Transition Into Service-3 State	
4.3 Call Service versus Call Control Connections	
5. Session Management Protocol	44-71
5.1 Signaling Reference Model	
5.2 Protocol Stack	
5.3 Message Format	
5.4 Session Management Scenarios	
5.4.1. Session Setup	
5.4.2. Resource Request	
5.4.3. Resource Release	
5.4.4. Session Release	
6. Conclusion	72
Appendix	
Bibliography	

## *CHAPTER ONE*

# **INTRODUCTION**

A video-on-demand (VOD) system provides a service which enables the user of the system to request in real time the transmission of a video stream from a collection of the available video material. The most common service for VOD is "interactive VOD," whereby the user gains access to a interactive visual and multimedia application (i.e., digitized video from a storage medium such as a hard disk), via a point-to-point connection. This connection allows the user individual and instantaneous control of the storage medium in terms of start, fast-forward, pause, and rewind etc. actions. Video on demand is one of the most complex services to be provided over B-ISDN, because a VOD system consists of many types of entities that need to communicate with each other before, during and after the actual delivery of a VOD application. Among these entities are user set-top terminal (STT), service gateways, network bandwidth manager, and video servers. A VOD session comprises several distinct services, and each service should be delivered over a set of network connections established between different groups of network entities. During a VOD session, users have to navigate through the network in order to select a service provider and then through the service provider domain in order to select the required video application.

One of the most important aspects of a video on demand system is the exchange of the signaling and control information among the various system components. Different functions related to signaling and control are needed in

order to admit a call or control a video application that delivers some or all of the video stream to the users. Other functions are needed in a switched based video on demand system, in order to setup connections along paths of selected switches across the network. The denominator of these functions is, probably that they are not directly related to the actual delivery of the video signal to the user. Hence they are classified as "signaling and control" or as "management" functions.

A video-on-demand session is defined as an association between Client's and Server, providing the capability to group together the resources needed for an instance of a service. Session are the frame work within which the client interacts with the server to get an instance of service. Here we will develop a user-to-network signaling protocol for communication at session layer, between different entities (client, server and network manager).

This protocol defines the user-network interface protocol for session & resource management . First we will develop a VOD session model that subclassifies the VOD control and management procedures into three management levels : session level , call level and connection level . Next we will use this VOD finite state machine model in order to analyze the relationship between the procedures in the various management levels and to discuss the options and trade-offs involved in the design of VOD session. Next, Based on the VOD model and analysis, we will develop a session layer signaling protocol that will support advanced interactive applications. We will present the session-layer set of messages and procedures that describe the process of establishing and releasing a session and allocating and deallocating session based resources , such as individual transport connections during the life of the session .

# ORGANIZATION OF CHAPTERS -

The dissertation is divided into chapters.

The Chapter 1 is Introduction. It gives the idea about the project and describes the problem.

Chapter 2 concerns the video-on-demand network architecture. The major components of the VOD network architecture are depicted here. Also the impact that the different VOD system elements have on the VOD service from a communications stand point are examined.

Chapter 3 discusses video on demand session. It presents our model for VOD control and management and defines keyterms. A user activity model is developed to describe the usage of system resources.

In Chapter 4 a multiple-call session model is presented. It shows how a multiple-call session can be represented as a FSM with states representing active calls and transitions representing distributed protocols for setting up, taking down or changing the status of calls. Call level management is also discussed here.

Chapter 5 contains the session-layer set of messages and procedures that describe the session setup and releasing of a session and allocating and deallocating session based resources.

Chapter 6 concludes the project.

At the end an appendix and alphabetical bibliography of all references is presented.

## **VOD NETWORK ARCHITECTURE**

In VOD, a Network is a collection of communicating elements that provides connections and provide session control and/or connection control to users. A user is defined as an end system that is connected to a network and can transmit information to, or receive information from, other such system by means the network. Users are categorized as Clients or Servers (or both).

### **. Client-**

Represents the customer premises equipment (CPE) in the interactive video network (IVN) environment, such as: set-top terminals (STTs), video display, TV, etc.. We will use client to reference the collective functionality of the end subscriber and CPE equipment.

### **. Server-**

Represents the collective functionality of the components belonging to a logical information and video service provider, whatever the internal architecture of this provider may be. Each server may physically contain multiple content and signaling nodes, shared or dedicated, in centralized or distributed manner over the network.

The interactive video on demand network consists of numerous video information providers (VIP's) providing high bandwidth information to numerous end users over a transport networks. A vedio on demand system (VOD) comprises many elements that are necessary for the provision of a



complete service. The major components of a VOD network architecture are depicted in fig. Among these are VOD server(s); service operation center; backbone ATM network linking geographically dispersed video servers; network signaling, switching, routing and multiplexing in the head end or central office; access network from the headend or central office to the home or business; and the client,s set-top terminal that receives, demodulates, decodes and converts the video for television playback.

The most likely protocol to be used in implementing a video-on-demand system is ATM. ATM can provide high speed video, voice and data services to consumer in a uniform way. The use of ATM ensures future-proof operation, by allowing various bandwidth and services, such as video, audio, games training courses, etc. to be supported. The VOD architecture is bandwidth transparent, i.e., the server, network, access, and customer premises equipment(CPE) are able to operate with various type of video signal compression schemes and bandwidth requirements. The Federal Communication Commission(FCC) has allocated a local exchange carrier service called video dial tone (VDT), an asymmetric switched video service in which the customer choose among a vide selection of video material and receives on demand, real time response. The service is asymmetric in the sense that the down stream ( to the customer ) channel has much higher bandwidth than the upstream channel. The network providing the delivery of the VDT service is known both as the Level One Network and the VDT network. Services offering by video information provider will be through a regulated, open interface known both as the “open level one gateway” and the “ VDT gateway”.

## **2.1 System Elements-**

Figure illustrates the main element of VOD system. We see the set top terminal in the subscriber premises, by which the user interacts with the services. The communication infrastructure between the customer premises and local switching office, we call the acces network. The term “switching

office” is one we use to characterize the place where services are fed and distributed to individual subscriber. For a designated service community, an IVN manager acting as level 1 gateway will provide the signaling capabilities for the network. Beyond the local switching office, backbone ATM network provide access to servers that do not reside in the local switching office and to regional, national, or other specialized repositories of information. These may or may not allow interactive access to individual subscriber, and the manner in which they are used affects the server design and interface and also depends on several factors, including service characteristics, communication tariffs, and disk storage i/o performance.

A brief overview of the function of these network elements follows.

### **2.1.1 Client’s set-top terminal-**

Set-top terminal ( or personal computer ) is the customer premises equipment that terminates the signaling protocol on behalf of the user to request services from the service provider. It also extracts the video content from the transport stream and directs it to the user’s TV set. The set-top terminal or personal computer along with the television monitor enables viewers to be connected to a video source ( video server ) and browse through a selection of movies or contents such as news stories, software, or games. It is important that the interface specification for the set-top be well designed from the outset because it is a lot more difficult to replace or modify a set-top and its associated user interface than other components in the architecture.

The key components of the set-top device are the line transceiver, demodulator, decompression unit, back-channel interface, and display driver. The line transceiver receives the incoming signals and permits the sending of control information back to the video server. The incoming signal is demodulated to baseband to recover the compressed digital video stream that is sent to the video decompression unit where it is converted to analog form and presented to the video or TV monitor.

## 2.1.2 Access Network-

The access Network comprises the various access arrangements from the ATM networks to the CPE. To use a more general term, we can refer to the network by which this and other broadband services are delivered to the residential users as a community network. This network connects the set-top equipment and ATM network. Numerous networking technologies can and will be employed. Based on broadband switched ATM network, this network architecture uses two access technologies : hybrid fiber-coax (HFC) and switched digital video (SDV). With either access technology this network can provide the entire range of interactive services.

In hybrid fiber-coax (HFC), radio frequency subcarrier modulation is used to transport signals through the access network. These are transported in 6-MHz. channels in National Television System Committee format. Other interval (for example 8-MHz) can also be used, depending on the television standards. A tuner in the STT selects the appropriate channel. The major difference between this and conventional CATV network is the addition of compressed digital video. In the CATV network, one channel of analog video occupies 6-MHz. With digital compression, multiple digital program can be placed in one channel, thus significantly increasing the capacity of the system. Such a system can carry hundreds of digital program, allowing for point to point interactive services such as movies on demand. Because the transmission medium in this access architecture is shared, encryption is used to control access and privacy.

The component of the access network are responsible for:

- Multiplexing of digital programs that occupy a 6-MHz.(or other) channel.
- Modulation and combining of the channels into the RF spectrum for transport on the fiber-coaxial distribution network.
- Routing of signaling messages to-from the STT's, and
- Electrical-to-optical conversion for transport over the optical portion of the distribution network.

The second type of the access network architecture is SDV, in which digital information is transmitted directly as “bits” through a dedicated path to the end user. The access network is configured as a star topology. This approach provides a separate physical drop, or line, for each home, with a dedicated downstream bandwidth on the order of 45 Mb/s.

In the HFC network, video content for a specific home in the neighbourhood is present on the coax feeding all homes. Each STT is responsible for tuning to the appropriate channel and decoding a specific program, while ignoring all other information. In contrast in the SDV network, only the content required by home is actually delivered to it. The available 45-Mb/s bandwidth allows for multiple STT in a home, each receiving a unique program. Switching digital broadcast channel is done outside the home. A channel change request originating at the STT sends an upstream message to the host digital terminal(HDT), where all broadcast channels are available. Similarly point-to-point service like movies on demand, switching must take place in both the HDT and the broadband ATM network.

The HDT is responsible for:

- Initializing/booting each STT as it is powered up,
- Handling upstream/downstream signaling for all interactive services,

- For broadcast services, receiving provisioning information from VIPs that determines which program streams an STT is authorized to decode,
- Handling end user requests for broadcast service by receiving a message from the STT and connecting the STT to the appropriate program stream.
- Routing end user requests for point-to-point interactive services to the video manager, and
- Acting on commands from the interactive video manager to connect an STT to the broadband ATM network for point-to-point interactive services.

### **2.1.3 IVN Manager/Level1 gateway(L1GW) -**

The IVN manager is a processing node that is owned by the network provider to provide level one gateway functions in the distribution network. It provides the interface for the customer for localizing and connecting to a video server, from a selected service provider (session control). Its main function is to provide the end user with an equal access to many video information provider and to establish and tear down session and connection between user and video information provider(s). The level 1 gateway acts as a central intelligence node for signaling and control. The control function provided by the IVN manager and other OSI network layer specific components, such as ATM network, form the control plane of the IVNL1 services. The L2 plane also provides control functions.

The interface for the video information provider has two logical channels carried on a separate physical media: a simplex 155-Mb/s video channel over which run MPEG-2 Transport stream conforming to the MPEG-2 system layer, and a full duplex control channel carried on an IEEE802.3 LAN

with a UDP/IP protocol stack. The interface for the subscriber side of the network has two logical channels: a simplex MPEG-2 digital video channel and a full duplex control channel. Both channels will be carried on a fiber/coax physical layer. The video stream is transported over a standard 6-MHz cable TV channel using 64 QAM modulation. Each 6-MHz TV channel carries several MPEG-2 video streams with an aggregate multiplex rate of about 25 Mb/s and the set-top selects and demultiplexes the desired stream.

The level 1 gateway (or VDT gateway) is the entry point for a video information provider to a carrier's VDT network. It is responsible minimally for the provision of connection management, video stream and signaling message routing, and menu functions. Signaling messages are routed by the message router among the gate controller, the subscriber signaling channel, the video menu server, and the information providers.

The level 1 gateway's (or VDT gateway's) operation can be classified into four phases:

- Establishment of a connection to the subscriber in response to a signal generated by the set-top. This involves allocation of head-end and delivery network resources and, either directly or indirectly, control of the set-top channel tuner and stream demultiplexer.
- Presentation to the subscriber of a menu of available information services. At least two mechanisms are possible here: one is direct display of the video selection menu via connection to the subscriber video channel to a video server; the other is downloading of executable software to the set-top, which then locally generates a service menu. In either case, the subscriber chooses which service to connect to, in a white page fashion.

- Interconnection of the subscriber terminal and a requested information provider. This involves routing an information-provider video channel to the subscriber video channel, as well as the routing of control signals between the subscriber and the information provider. From the information provider, the user receives a rich menu of the actual titles and makes a selection.
- Disconnection of the session between an information provider and a subscriber terminal in response to a request from either party, together with consequent release of head-end and delivery network resources.

## **2.1.4 ATM Network -**

The ATM network provides the interconnection of the various network elements in the VOD architecture. The interconnection includes both signaling and program data transfer, the latter at real time or at network speed where required, semipermanent and on demand. The server information streams such as movies, games, audio, etc., are conveyed from server through the broadband switching network and access arrangement, up to the CPE, using the ATM technology. Both point-to-point and point-to-multipoint connections can be used inside the ATM network.

Services like Interactive video on demand require error free transmission as well as rapid transfer. In ATM, information flow is organised into fixed size blocks called "cells," each consisting of a header and a information field. Cells are transmitted over a virtual circuit, and routing is performed based on the Virtual Circuit Identifier (VCI) contained in the cell header. The transmission time is equal to a slot length, and slots are allocated to a call on demand. ATM's fundamental difference from STM is that slot

assignments are not fixed: instead, the time slots are assigned in an asynchronous (demand-based) manner. In ATM, therefore no bandwidth is consumed unless information is actually being transported. Between ATM and STM, ATM is considered to be most promising because of its efficiency and flexibility. Because slots are allocated to service on demand, ATM can easily accommodate variable bit rate services. Moreover, in ATM, no bandwidth is consumed unless information is actually being transmitted. ATM can also gain bandwidth efficiency by statistically multiplexing busy traffic sources. Since bursty traffic does not require continuous allocation of the bandwidth at its peak rate, a large number of bursty traffic sources can share the bandwidth.

Discarding cells based on the importance of their contents can be applied to video traffic. If an embedded coding technique is used for the image then coding information is separated into two bits streams: a stream containing essential information and a stream containing picture enhancement information. Cells containing essential information are given higher priority than those containing the picture enhancements. When congestion occurs, only low priority cells are discarded. With this scheme, even when networks become congested, the essential parts of the coded information are transmitted, thus it is expected that cell loss will have only small influence on picture quality.

ATM is by definition, a connection-oriented technique. This connection oriented mode minimizes delay variation since cells belonging to the same call follow the same route. It also minimizes the processing required to make the routing decisions. ATM cells consist of a 5-octet header and a 48-octet information field. The CCITT header format will be used at User-Network Interface (UNI) and Network-Node Interface (NNI). For UNI, the header contains a 4-bit "generic flow control" (GFC) field, a 24-bit label field containing Virtual Path Identifier (VPI) and Virtual Circuit Identifier (VCI) subfields (8 bits for the VPI and 16 bits for the VCI), a 2-bit payload type (PT) field, a 1-bit reserved field, a 1-bit priority (PR) field, and an 8-bit



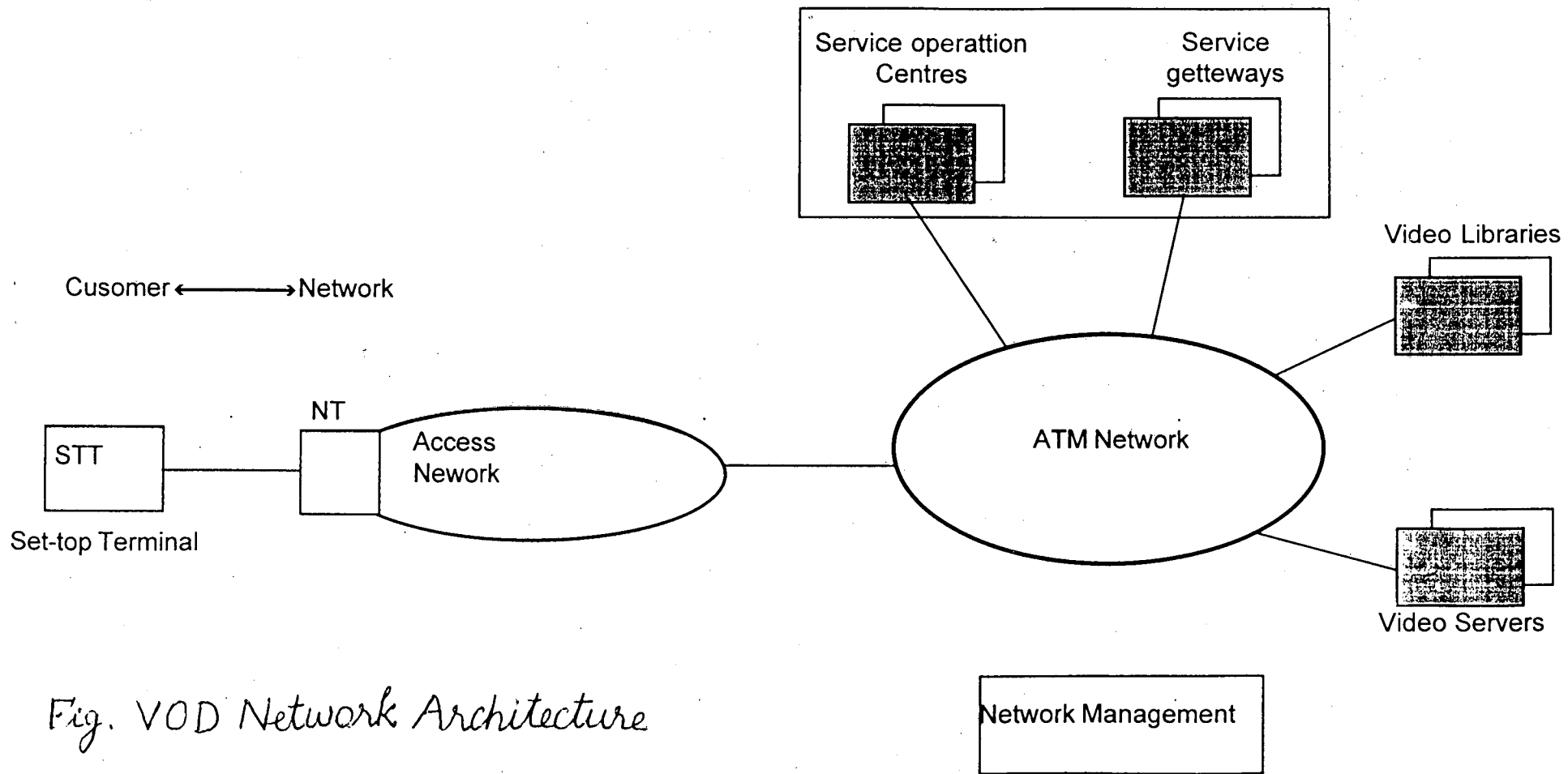


Fig. VOD Network Architecture

header error check (HEC) field. For NNI, the header does not contain a GFC field, and the extra 4 bits are used for a VPI field.

The GFC field is used to assist the customer premises in controlling the flow of traffic for different qualities of service. One candidate for the use of this field is a multiple priority level indicator to control the flow of information in a service dependent manner. The GFC field appears only at the UNI. The “virtual path” concept is adopted in a label field. The VPI provides an explicit path identification for a cell, while VCI provides an explicit circuit identification for a cell. Basically, a virtual path is a bundle of virtual circuits which is switched as a unit by defining one additional layer of multiplexing on a per-cell basis underneath the VCI. A predefined route is provided with each virtual path; thus it is not necessary to rewrite the routing table at call setup. Therefore, call-by-call processing at switched nodes is reduced and call setup delay is decreased.

## **2.1.5 Video Server-**

The video server is the network element providing the source of the video (and other) program material, which can be requested by the customers. The Video server provides an on demand copy of the requested video signal via the ATM network to the user.

The video server consists of the storage and control required to store movies in the compressed format and play them back on the request. It differs from the traditional database server in various ways. It has to perform a number of functions such as admission control, request handling, data retrieval, guaranteed stream transmission, stream encryption, and support of

functions found in VCRs, including pause, resume, rewind and fast forward. Admission control is done for each request by determining if the request can be serviced by the available resources in the system. Because the transmission of the video data is stream oriented, it needs to be delivered to the end viewer without any glitches. The system can service the request only if continuous delivery of the video stream can be guaranteed, once the stream has been started. The non deterministic nature of the disk accesses mandates that intermediate buffer memory be used to transform the bursty disk accesses into a continuous stream that is guaranteed to be glitch-free. Storage constitutes the bulk of the server cost for nearly all systems, whether they store 100 movies or 10000 movies. Storage media would be a combination of magnetic, magneto-optic, and tape devices.

For video on demand, each data stream requires a data rate of at least 1.5 Mb/s. This by itself can be easily supported by a single disk. However, to support multiple accesses to the same movie, we need to interleave the data. The interleaving techniques and granularity of the interleaving are important issues, but in many cases movies will be striped across a number of disks. Redundant arrays of inexpensive disk (RAID) that add additional disks are applicable but costly, and their necessity has not been shown. The RAID levels that are most useful for video are RAID1 and RAID3. In RAID1 data is mirrored, or replicated, usually on two disks, which is very costly. In RAID3 only one additional disk is required for each array to store parity.

Stream management will be explained by stepping through the operation needed to get a stream playback to the home started. Initially a request will come in over the network for a particular movie. The video server may already have cached this movie on disk. If not, the stream manager must initiate the reading of the movie from local or remote archival storage and must invoke disk layout algorithms to determine the most effective striping and placement of the movie within the disk cache. Given the arrangement of the movie on disk a scheduler can generate the disk read request for the stream and insert these into the overall disk request sequence. Finally the stream

manager needs to communicate to the VDT gateway the stream ID (which for an ATM link will be the VCI of the stream) to enable the gateway to setup the connection out to the set-top.

## ***CHAPTER THREE***

# **VIDEO-ON-DEMAND SESSION**

### **3.1 Session Concept and Resources**

The STT and server are isolated from the details of the network control signaling by introducing a higher layer protocol through which the STT and server communicate with the network manager. This is session layer protocol, as opposed to the network layer protocol, which directly controls the switched network. A Session is an association between multiple parties (servers and clients), providing the capability to group together the resources needed for an instance of a service. A resource is a trackable "object" or "element" allocated by the network manager to a video session and is retrieved by the network manager when the resource is no longer needed by the session. This distinction between session and resources allows for the development and offering of advanced applications to the end user. A Session implies relationship between an end user and a VIP for a period of time during which the end user participates in interactive activity. Associated with the session are the series of related connection through the network between server(s) and user's STT.

For the IVN network provider, a session is used as the point of entry to implement higher level policy decisions and enhanced IVN services (such as "Session Transfer"), and to monitor all resources related to a current session between a client and a server. A Network-wide unique sessionID identifies a session in the network. All resources of an instance of service are tagged with the sessionID and are disposed of when the session is torn down. A session will usually comprise more than one connection resources. Session are

depicted as having a control stream and an MPEG-2 stream, but sophisticated application may use more connections (e.g., a connection to carry to movie that is shown in a window whose surrounding background is downloaded with another high speed connection). It should also be noted that the server end of connections within a session do not need to all terminate at the same network access point (NASP): Servers may have multiple ATM attachments to the network, and a Client may be receiving information from more than one access point within a given session with potentially multiple resources making up a connection and potentially multiple connection within a session, there is need to identify the resource connection setup for that session. To do this each resource is assigned a resourceNum unique within a session. Resources that make up the same connection are assigned the same associationTag.

### **3.2 Basic Operation of VOD Sessions-**

Figure shows in the VOD architecture the major communication paths for signaling ("s") and for program information transfer ("p").

The signaling paths from each CPE towards the service gateway (S3), are "nailed-up," i.e., preset by the network management. Also the signaling path between service gateways, video server, video libraries, and the service operation centers are preset (i.e., S5, S6, and S7). Only the signaling path of the CPE towards the selected server (S2/S4) is established on demand under control of the CPE and the service gateway.

a) The set-top terminal is connected with the service gateway (preset signaling connection). From the service gateway, the user receives a list of accessible service operators.

b) On the customer's request for a service operator, the service gateway selects the corresponding video server. A single service operator may own multiple video servers, or may rent capacity on a video server owned by the network operator. This is transparent to the subscriber. As an option, the service gateway could also offer an index service that provides a directory of the contents of the different servers and allows searching over multiple servers and multiple service providers.

c) A path is set up between the set-top terminal and the video server; this path setup can be initiated by the set-top terminal or the video server, using the routing information provided by the service gateway.

d) Once the video server is connected to the set-top terminal, a "program book" is provided, allowing the customer to browse through the program offering. This browsing is done locally at the set-top terminal, and requires basically no interactivity with the server.

e) The customer selects a program out of the program book, and communicates his/her request via signaling to the video server.

f) The program is sent to the set-top terminal, through appropriate actions in the network (generation of a copy of an instant, fast load;...).

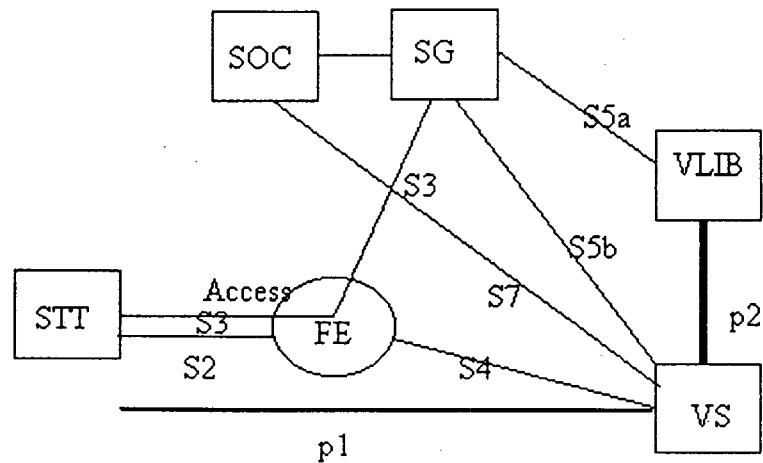


Fig. Signaling and program communication path

A major advantage of the use of the ATM technique is the fact that the signaling paths can be "nailed-up" in the network, avoiding the delay of call setups. The average bandwidth consumed by the signaling paths is negligible.

At the start of a session, a fast load is performed of the first few minutes of the movie from the selected VOD server into the video buffer. This action requires a large bandwidth, e.g., 150 Mb/s, to be allocated in the network during a few seconds. It provides the user with quasi-immediate access to the service. At the same time, an instant of the movie, attributed to a particular ATM VC, is created to continuously feed, at video signal rate (e.g., 2Mb/s), the video buffer with the remaining part of the movie. This instant may already be available in the network, and can be reused by other front-end switches.

If a user pauses and resumes, or shuttles to another part of the movie, he can be connected to another instant, and, a new fast load is performed. The moment at which the movie instant (VC) is created (if it did not exist yet) largely depends on:

- ◇ The amount of storage capacity in the pooled video buffers.
- ◇ The demand requests at that moment for a set of movies.
- ◇ The available network resources.

The time difference between the created instants (also called instant granularity) of a given movie is not constant since instant creation is dynamically controlled in order to improve network resource efficiency.



### (3.3) Hierarchy Of Session Management-

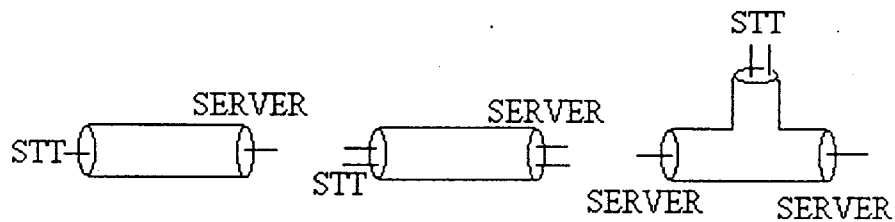
#### (1) Sessions, Calls and Connections-

A Session implies a relationship between the end user and a video information provider (or server) for a period of time during which the end user participates in interactive activity. Associated with the session are series of related connections through the network between server(s) and the user's STT. Suppose that at some time  $t'$ , a user tunes his STT to the VOD channel. Let  $t''$  be the first time after  $t'$  when the STT is tuned to another channel or is tuned off. We say that during time period  $[t', t'']$  the STT was involved in a single VOD session. We represent each session by means of a single contiguous finite state machine (FSM). Most of the states of the session FSM represent active calls. A call is a logical association between the STT and a service provider which is maintained by the network. It serves as an information communication pipe that delivers an instance of service to the user. A call is established by means of the distributed control procedure: the call set-up protocol. It is canceled by means of another distributed control procedure: the call take-down protocol.



Every call is associated with one or more connection that usually have a common time base. A connection is an end-to-end bit pipe, associated with a set of quality of service (QoS) parameters, over which call information is exchanged among the call end entities. In ATM networks, for instance, a connection is a directed switched path formed by the hardware of the lower layer protocols. In a TCP/IP network, on the contrary, a connection is a

directed and formed by the software of the upper layer protocols. In order to realize a call one or more connections are needed. Call, like digitized video, may need several connections: one for video information, another one for audio information, and a third one for application control information. There are two type of connections- service connections and control connections. Those connections needed in the various service states are referred to as call service connections, which are used by the application layer for delivering data or application control information. Those needed for the exchange of the call control information (i.e., the message needed for the set-up and take-down of the calls) are referred to as call control connections. This distinction is needed in the circuit-switched networks or in virtual circuit networks like ATM, where control messages must also be sent over pre-established connections.



- Call Example: (a) simple call of one connection  
 (b) simple call of two connection  
 (c) simple call with two server

We have shown a simple (two-way) call consisting of a single connection. In a TCP/IP network, the bullets along the connection represent IP gateways, where in an ATM network they represent VC switches. Figure shows a simple call consisting of two connections. Some of the calls can be multiple-point calls, where more than two entities are involved. A three way

call may consist of one STT connected to two servers or two STT's connected to one server by means of the multicast connections. In the latter case, however, each STT is associated with a different session, and the three-way call can be considered as two simple (two-way) calls. Thus, the one-to-one relationship between calls and STT's is retained.

We observe a hierarchy of sessions over calls over connections, with a 1-to-N relationship between sessions and calls (i.e., one session consists of N calls) and a 1-to-N relationship between calls and connections. Since we assume that all the services, preliminary as well as ultimate, provided during a single session are destined for a single STT, the STT is the end-point of every call. Thus only one call can be active at a time, though more can be nonactive but alive. A call is said to be active when its connections transfer data. A call is said to be alive if it has not been taken down since the last time it was set-up. Figure. depicts two sessions. The session in Fig. is a relatively simple one. At the beginning of this session, call-1 is set up in order to deliver some service to the STT. Then call-1 is taken down and another call, call-2, is set up. Finally, call-2 is taken down and third call, call-3, is set up. When call-3 is taken down, the entire session comes to an end. The session in Fig. is more complex. At the beginning of this session, call-1 is set up, but before it is taken down call-2 is set up and the session has two alive calls. However, only one of them can be active. Then, both calls are taken down and call-3 is set-up. After call-3 is taken down, another instance of call-1 is invoked. Finally call-1 is taken down and the entire session comes to an end.

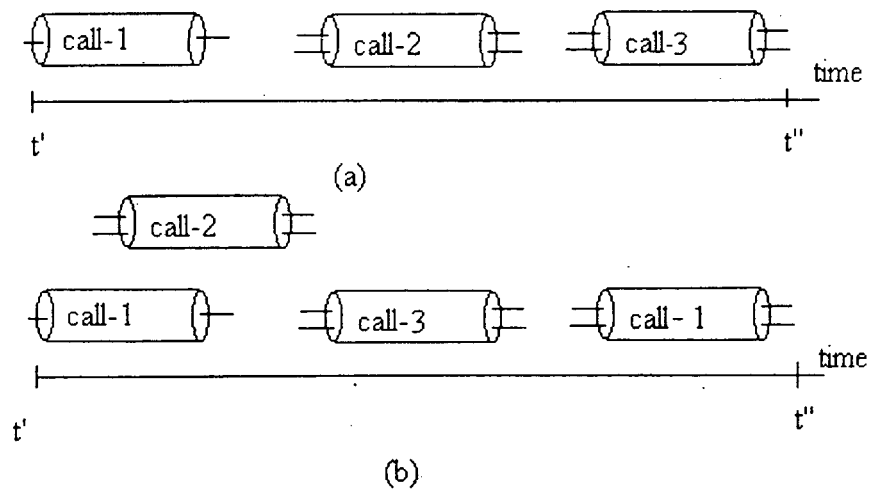


Fig. Session Examples

Based on the hierarchy of connections, calls and sessions, we suggest a generic model for VOD session management.

a) The session management level is where user commands that require changes in the call or in the connection level are translated into network signaling commands. The following are the common commands:

(i) A request for starting a new session or the ending of the current one. In the former case, a new call and/or some control connections are required. In the latter case, all alive calls must be taken down.

(ii) A request for starting the delivery of the new service or ending the delivery of the current service. In the former case, a new call might be set up or a passive call might be activated. In the latter case, an active call might be taken down or become passive.

(iii) VCR-like control commands as fast-forward, pause, play etc. Usually such command donot require change in the call or the control levels, in which case they donot involves session control commands. There might be cases ,however, where such command require changes in the call or the connection levels. For instance, if the STT wants to switch from "play" to "fast-forward," but the fast-forward version of the video is not available at the server that holds the regular version, a new call might be needed.

(b) The call management level is where call signaling and control is performed, in order top set-up new calls, take down an existing call, reactivate passive calls, modify existing calls, etc. Most of the procedure in this level are distributed among several entities. As explained above, application are usually translated by the session management module into call level commands.

(c) The connection management level is where connection signaling and control is performed, in order to set-up, take down, or change the attributes (QoS parameters, bandwidth, etc.) of connections.

## (2) Dependencies of Calls-

Every two calls invoked during a single a session are either dependent or independent of each other. We say that call-2 depends on call-1 if call-2 can become active only after the call-1 becomes active during the life time of the same session.

### (3) Dependencies of entities with Respect to calls-

We consider a call, call-i say, for which entity E is an end point. Then for every entity  $E \neq E$  we have the following:

(1) E is said to be call-i independent of E' if E' is not involved in the process of admitting call-i.

(2) E is said to be call-i semi-dependent of E' if both E and E' are involved in the process of admitting call-i.

(3) E is said to be call-i dependent of E' if E' is involved and E is not involved in the process of admitting call-i.

The admission of a call is performed by means of the call set-up protocol, executed among several entities in the network. Every entity in the network, including the call end-points, may participate or not participate in the call set-up protocol. An entity is said to be involved in the process of admitting call-i if it participates in the call set-up protocol.

## 3.4 User Activity Model -

We are going to develop a user activity model to study the impact of user behaviour on the system design. The user activity model describes the usage of system resources, i.e., network bandwidth and video server usages, by a user as it interacts with the service. Once connected, we assume the user is in one of the two states, the normal and the interaction states. He starts in the normal state, i.e., the video is being played at the normal speed. User stays in

this state for a period of the time which is exponential with parameter a. Then user issues an interactive operation, such as stop, speed-up, etc. User stays in this interaction state for another period of time which is exponential with parameter b. Then he goes back to the normal state, from where he may again go to the interaction state. This may be repeated multiple times until he/she disconnects. Different types of interaction operation will affect the system in different ways. The following are the types of the interactive operations:

(1) *Play/Resume*: The start of the presentation from the beginning or the middle.

(2) *Stop*: Stopping of the presentation, without picture and sound.

(3) *Pause*: Temporarily stopping the presentation, with picture.

(4) *Jump Forward*: Jumping to a target time of the presentation in the forward direction, without picture and sound.

(5) *Jump Backward*: Jumping to target time of the presentation in the backward direction, without picture and sound.

(6) *Speed Up*: Quickly moving presentation forward, with picture and sound (fast-forward).

(7) *Slow Down*: Slowly moving presentation forward, with picture and sound.

(8) *Reverse*: Playing a presentation in the reversed direction, with picture and sound.

(9) *Fast Reverse*: Quickly moving presentation backward, with picture and sound.

(10) *Slow Reverse*: Slowly moving presentation backward, with picture and sound.

Suppose the proportions are as follows: stop/pause,  $q_1$ ; jump forward,  $q_2$ ; jump backward,  $q_3$ ; speed up,  $q_4$ ; slow down,  $q_5$ ; reverse,  $q_6$ ; fast-reverse,  $q_7$ ; and slow reverse,  $q_8$ , where

$$\sum_{i=1}^8 q_i = 1$$

Obviously, "stop/pause" will not require data delivery from the video server. In fact depending on the size of the video buffer, and the size of the jump, even "jump forward" and "jump backward" may not require data delivery either. Basically, if the jump is to a portion of the video already in the video buffer, no data delivery is necessary. If the jump takes us to a portion outside the buffer, then data delivery is necessary. We denote the probabilities that we are within the video buffer in the "jump forward" and "jump backward" operations by  $p_F$  and  $p_B$ , respectively. Both the "speed up" and the "slow down" operation will require data delivery, but the rates may be different from that of normal playback. Let  $K_1$  be the speed up factor, and  $K_2$  the slow down factor. Speed up may be implemented by retrieving selected frames (say every other frame for a two time speed up) from the video server and delivering only these to the user. In this case, the required data delivery rate  $K_1 C$  may actually be the same or even smaller than those for normal playback. Slow down may be implemented by sending data at a reduced rate, resulting in a data delivery rate  $K_2 C$ . These reverse operations are similar to the play, speed up, and slow down operations. Suppose the fast reverse and slow reverse operations also have data delivery rates of  $K_1 C$  and  $K_2 C$ , respectively.

We can now calculate the average data delivery rate to a user. This is the rate at which the data is delivered from the video server to the video buffer, and from the video buffer to the user. The probability a user is in the normal state is  $\beta/(\alpha+\beta)$ , while the probability of being in the interaction state is



$\alpha/(\alpha+\beta)$ . while in the interaction state, the user will be performing various operation with the probabilities listed above.

Thus, the average data delivery reate R is given by

$$R = [\alpha/(\alpha+\beta)*(q_1 + q_2p_F + q_3p_B)] + [\beta/(\alpha+\beta) + (\alpha/(\alpha+\beta))*(q_2(1-p_F)+q_3(1-p_B)+q_6)] * C + [(\alpha)/(\alpha+\beta)(q_4+q_7)]*K_1X + [(\alpha)/(\alpha+\beta)(q_5+q_8)]*K_2C \quad \text{----(1)}$$

Now we calculate the average connection time T of a user, defined as the time from when he/she is connected to a video server to when he disconnects. Suppose the normal playback time, without user interaction, of a video program is T', then T may be longer or shorter than T' depending on the user interactions. For example if we stop the video for t<sub>1</sub> time units, T will be increased by t<sub>1</sub>. Jumping forward by t<sub>2</sub> and jumping backward by t<sub>3</sub> will decrease and increase T bt t<sub>2</sub> and t<sub>3</sub>, respectively. Speeding up for t<sub>4</sub> time units by a factor of K<sub>1</sub> will decrease T by t<sub>4</sub>(K<sub>1</sub>-1), while slowing down for t<sub>5</sub> time units by a factor of K<sub>2</sub> will increase T by t<sub>5</sub>(1-K<sub>2</sub>). Reversing for t<sub>6</sub> time units will increase T by 2t<sub>6</sub> since in this case, the reversing itself case t<sub>6</sub> time units, and then one ends up at a point of the video which is t<sub>6</sub> time units before the reverse operation. Fast reversing for t<sub>7</sub> time units will increase T by t<sub>8</sub>(K<sub>2</sub>+1).

Therefore

$$T=T'+q_1E(t_1)-q_2E(t_2)+q_3E(t_3)-q_4E(t_4)(K_1-1)+q_5E(t_5)(1-K_2)+2q_6E(t_6) +q_7E(t_7)(K_1+1)+q_8E(t_8)(K_2+1) \quad \text{-----(1)}$$

• Where E(t<sub>i</sub>), i=1...8, may be found empirically.

## A MULTIPLE-CALL SESSION MODEL

We model the session management level as a finite state machine (FSM) that translates application layer commands into call management commands, such as call set-up, call take-down, and call change. Such a FSM is presented in the context of a VOD system where two preliminary services must be delivered to the user before a video application can be selected. After a user tunes his STTT to the VOD channel, he may either quit or ask to view the list of service providers. In the latter case, a call is established between the STT and the L1GW or video network manager. This call, during which white page menu is delivered to the STT, we will refer it as call-1. The user may now select a service provider, in which case another call, referred to as call-2 is established between the STT and the L2GW associated with the selected service provider. The L2GW's are the front-end provided by the service provider (video server) for the consumer to select from multiple interactive services. The first call becomes nonactive but is not taken down in order to allow the user fast switching back to call-1 when needed. During call-2, the L2GW provides the user with a selection of available video application, to which preview information can be attached. The user may now quit or select some video application. In the latter case, a call which delivers the requested application is established between the user STT and the video server selected by the L2GW. This call is referred to as call-3.

The VOD session described is represented by the finite state machine in Figure. The session FSM has five states: OFF, NO-SERVICE, SERVICE-1, SERVICE-2, and SERVICE-3. The first state, OFF, indicates that the last session has already ended and a new one has not yet been initiated. The NO-SERVICE state is entered when the STT is tuned to the VOD channel, and the session starts. SERVICE-1 is the state of the session when call-1 is active.

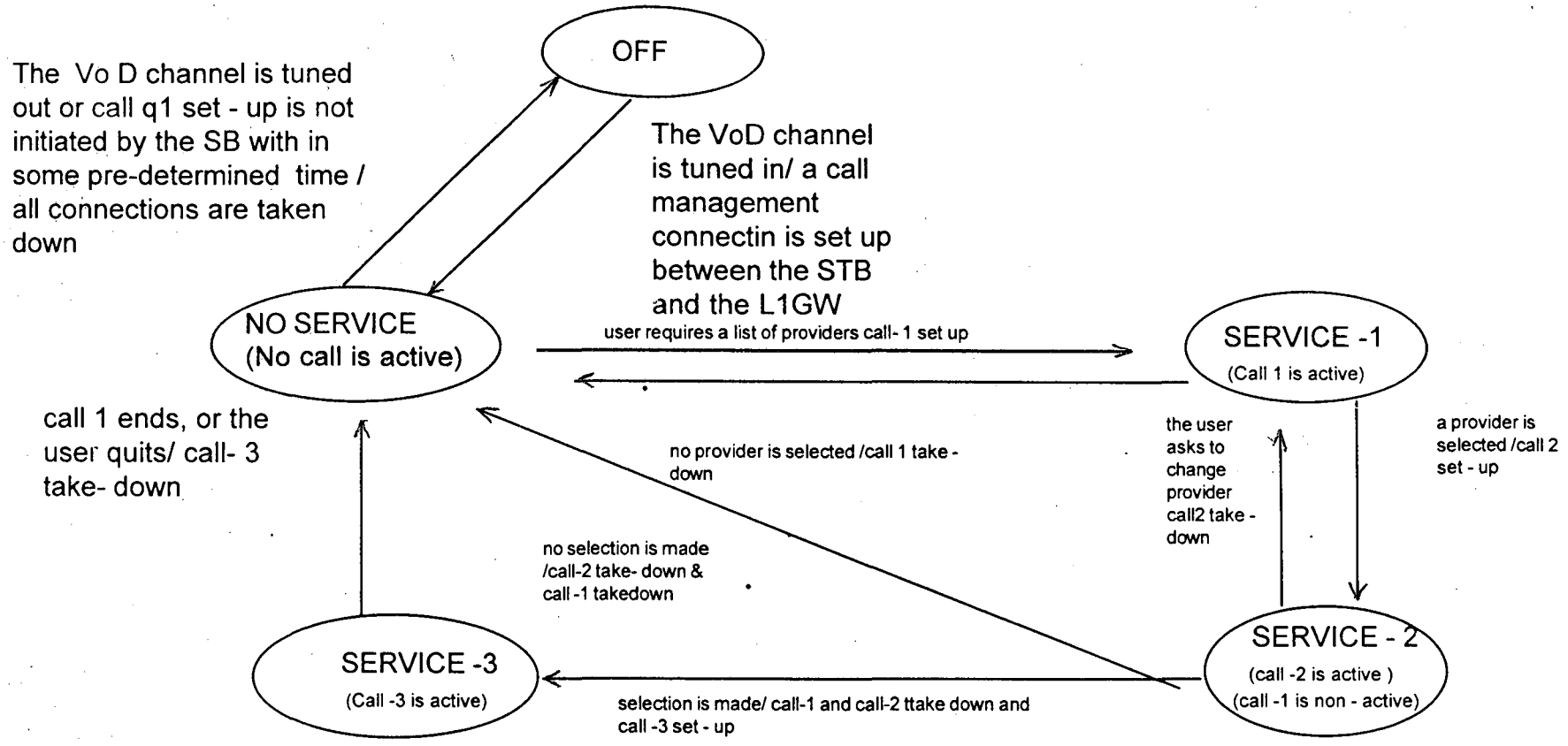


Fig. VOD FINITE STATE MACHINE MODEL

SERVICE-2 is the state when call-2 is active, and SERVICE-3 is the state when call-3 is active. The session management FSM shows only key transition. In particular, those transitions labeled as "call set-up" reflects the case where the set-up is successful, but not the case where it fails due to lack of the network resources, unauthorized access, or for any other reason. When a call set-up fails, the session may proceed in several ways depending on the type of the requested call and the failure reason. The FSM shown here depicts only one possible implementation for management of the particular session. Another possible implementation would be to take call-1 down upon entering SERVICE-2 from SERVICE-1 or to allow only an indirect transition from SERVICE-1 into SERVICE-2 through the NOSERVICE state.

The transition between the various states are triggered by application (user) commands and translated into call management commands that set-up, take down, and some times just change the status of calls. The purpose of the call set-up protocol is mainly to request the provision of a specific service and to allocate the resources needed by that service. The purpose of the call take-down protocol is to stop the provision of the service and to release the allocated resources. The entities participating in a call set-up or call take-down protocol are determined according to the resources needed by the call, which depend on the call type. We consider the set-up protocol for the call-1. During the call-1, the LIGW connected to the STT is supposed to deliver the white page menu to the STT. This is a basic service, which does not require much of the network resources. Assuming that such a service can be locally approved by the LIGW, the set-up protocol is essentially a simple two-way handshake protocol between the STT and its LIGW. When the LIGW accepts the call, the transition from the NO-SERVICE into SERVICE-1 is enabled. If the call is not accepted for some reason, the LIGW may respond to the call-1 request message with a call-1 reject message.

## 4.1 The Transition Into SERVICE-2 State-

In service-2 state the STT is connected to a particular service provider or servers. The Level-2 Gateway (L2GW) sends the menu of the server and some film clips. Preview service can also be provided in SERVICE-2 state. Some resources is needed to support low bandwidth navigation need and some film clips. So the L2GW also communicates with Network bandwidth manager in call-2 setup protocol. The transition from the service-1 state into the service-2 is made directly i.e. the call-1 is not taken down when the session enters SERVICE-2, so a direct transition from SERVICE-2 into SERVICE-1 is possible by invoking the call-2 take down protocol. The main advantage of this approach is that it, reduces the time and overhead for reentering SERVICE-1.

The transition of the STT session into the SERVICE-2 state depends mainly on the relationship with regards to call-2 between the L1GW and the L2GW, providing service during call-2. Since the L1GW is owned and managed by the network operator, whereas the L2GW the front end of the video server/service provider, we donot consider a case where the L2GW is not involved in the setup of call-2. This leaves the only two possibilities:

- (1) The L2GW is call-2 independent of the L1GW; i.e, the L2GW is involved in the setup of call-2, whereas the L1GW is not involved.
- (2) The L2GW is call-2 semidependent of the L1GW; i.e., both the L1GW and the L2GW are involved in the settop of the call-2.

The first possibility is more likely when the L2GW has its own mechanism for authentication, authorization, and billing. But even when L2GW is having these mechanism, the L1GW should also support these; because user is directly conneted to the L1GW and it is L1GW which charge

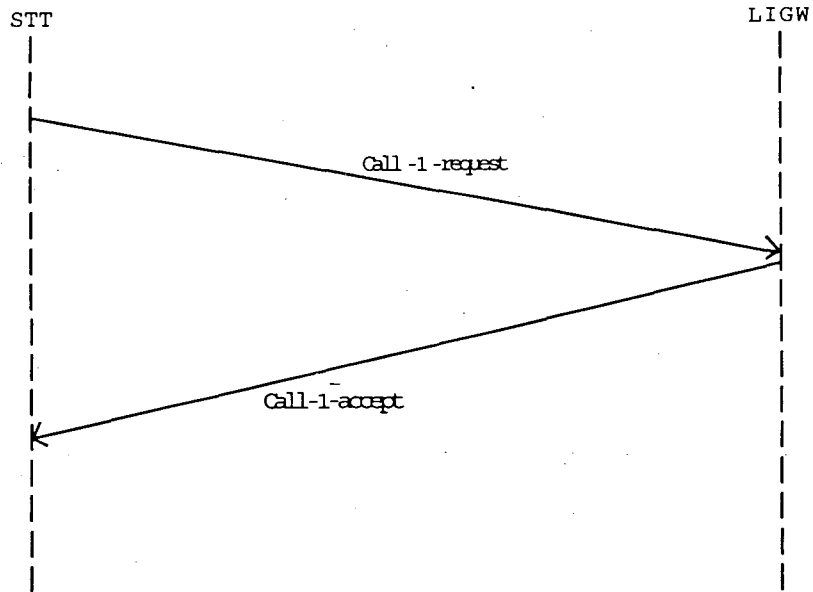


Fig. Call-1 set-up protocol

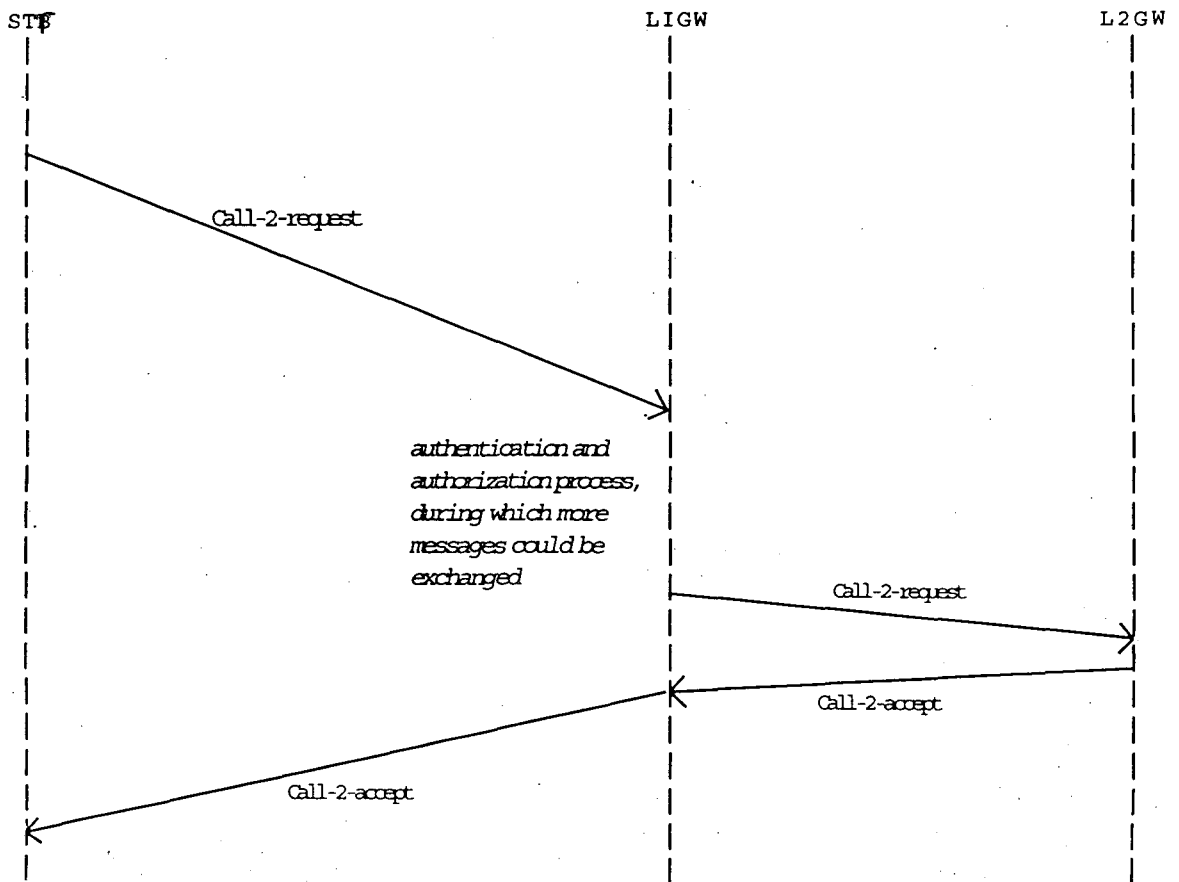


Fig. Call-2 set-up protocol

from the client. So these functionality are provided by the L1GW on behalf of the L2GW. To this end, the L1GW must be involved in the transition of the STT into the service two state.

The structure of the call-2 setup is given. The STT sends a call-2-request message to its L1GW. One of the parameter attached to this message is an explicit or implicit address of the L2GW to which the STT wishes to be connected. This address has been determined by the STT using the white page menu provided by the L1GW, while the STT was in SERVICE-1. When the L1GW receives a call-2 request message, it invokes some authentication and authorization procedure to make sure that this call request should be forwarded, with some possible modifications, to the appropriate L2GW. Now to provide the menu navigation needs and preview service, some resources are needed. So the L2GW sends a bandwidth request message to bandwidth manager. On allocating the bandwidth the network manager sends the bandwidth allocated message to L2GW. After this a call-2 accept message is sent by the L2GW to the L1GW and forwarded to the STT. After receiving this message, the STT enters the SERVICE-2 state.

#### 4.2 The Transition Into the SERVICE-3 State -

The service provided during call-3 demand more network resources and last much longer, than the service provided during call-2. to get high speed channel for a long duration, a network bandwidth manager will have to participate in the setup of call-3. The L2GW must participate in the setup of call-3. We make this assumption because we view the L2GW as the center of the service provider complex, where most of the management and control activities are performed. An important implication of this assumption is that, regardless of the STT intelligence (i.e., regardless of whether or not the STT needs to stay in SERVICE-2 while the user browses the yellow-page menu), the transition into the SERVICE-3 state will be made only from the SERVICE-2 state. The call-2 is not kept active upon entering the SERVICE-3

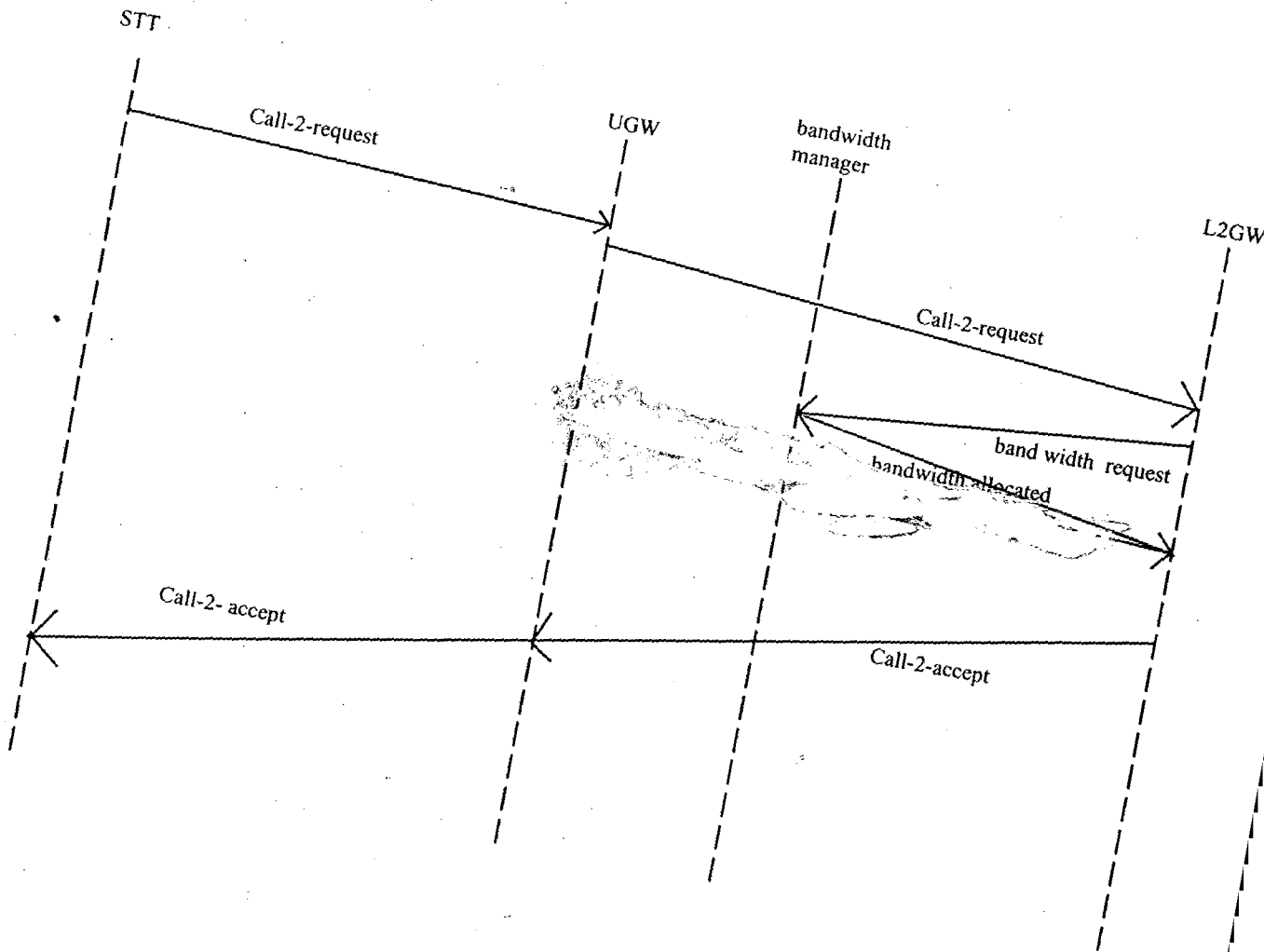


Fig. Improved call-2 setup



state. The reason for this is that call-3 is envisioned as a long call, and because the likelihood that the STT quits in an early stage of this call is relatively low, especially since the preview service is provided in the SERVICE-2 state.

The call-3 setup protocol is presented in Figure. It is assumed that the video-server is call-3 semidependent of its L2GW. The STT sends the L2GW a call-3 request message. The L2GW uses a two way handshake with the video-server in order to verify that the latter can provide the requested service. After the video server confirms, the L2GW ask the bandwidth manager to allocate network resources for the call. When the resources are allocated, the L2GW sends the call-2 and call-1 cancel message to L1GW. When the call-2 and call-1 are canceled, the L2GW informs the two end-points of call-3, the video server and the STT, that call is accepted.

In figure. the video server also call-3 semidependent of its L2GW, but unlike in the previous case the task of admitting the call is distributed between the L2GW and the video server. The L2GW determines whether to accept or to reject the request message sent by the STT. Then it selects the appropriate video server. Now the bandwidth request message is sent to the bandwidth manager. After the allocation of resources, call-3 accept message is sent to the L2GW.

Since from our assumptions that the call-3 setup can be invoked only when the STT session is in the SERVICE-2 state and the call-2 is not kept alive. While the SERVICE-3 state is entered follows that the call-2 take down protocol should be performed along with call-3 setup. Call-1 is kept alive in the SERVICE-2 state, so this call should also be taken down during the transition into SERVICE-3. Thus if call-1 is kept alive in the SERVICE-2 state, then regardless of the dependency of the L2GW in the L1GW with regard to call-1:(i) both call-1 and call-2 need to be taken down during the setup of call-3, and (ii) the L1GW has to participate in the setup of the call-3. So two way handshake between the L2GW and the L1GW is done for taking

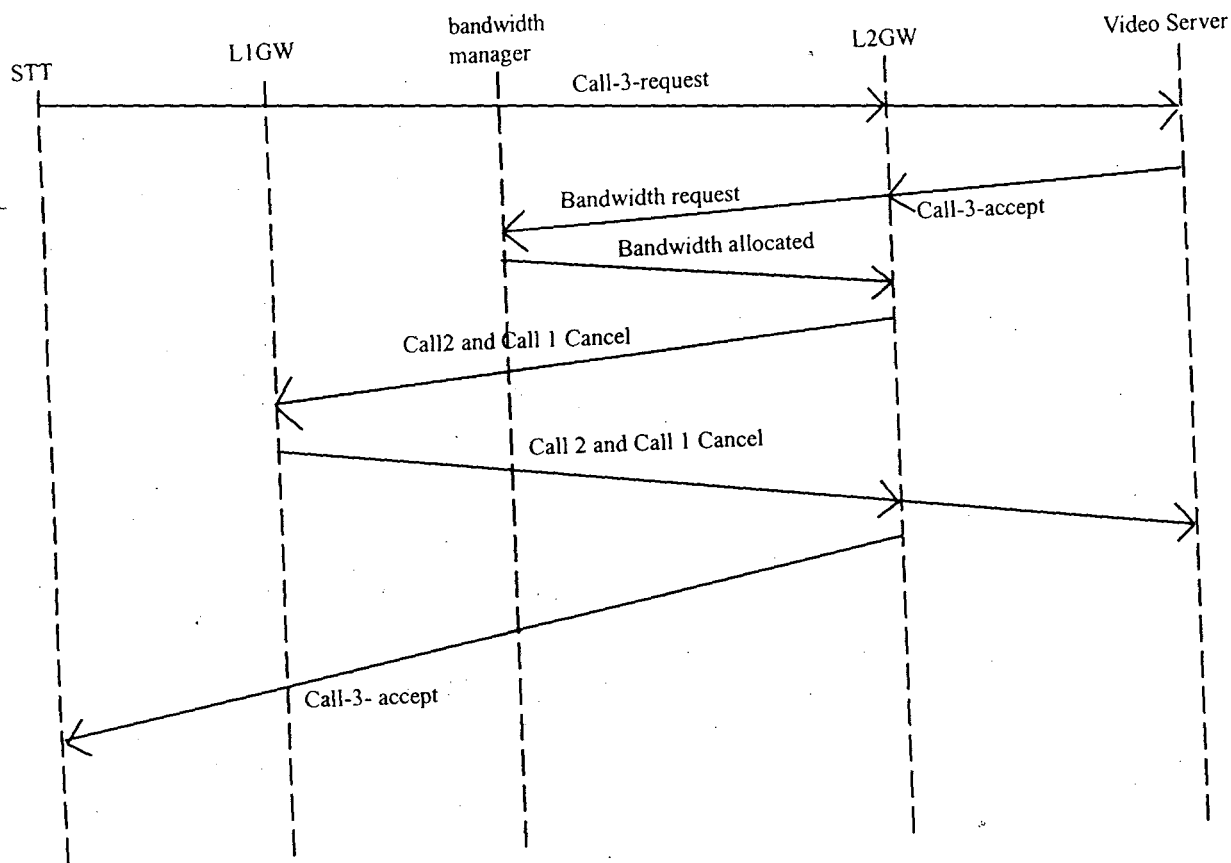


Fig. Call-3 Setup @ Centralized Approval

call-2 and call-1 down. Another issue related to the setup of call-3 is the state that the session enters upon setup failure. Since call-2 is taken down only after is setup, the best option is staying in the SERVICE-2 state. This gives the STT an opportunity to select another service or to quit by invoking the call-2 take down protocol.

#### 4.3 Call Service versus Call Control Connections -

The information exchanges during each call in a switched ATM/STM network require one or more connections. In some cases, one connection that carries all type of information is sufficient. In other cases, where different type of the data (video, voice, timing etc.) require connection with different QoS parameters, more connections from the video server to the STT are needed. To allow the STT to perform special application control operations like fast-forward or pause, a service connection in the reverse direction is needed as well. Connections are established in the network by means of the connection signaling protocols. Different protocol are defined for the user-network interface (e.g., Q.2931 for ATM) and for the network-network interface. In the network-network interface protocols, connection setup message are sent in a tandem along the route over which the connection is to be established, and the routing table of the switches along this route are updated to enable the switched connection.

In a switched networks a connection is needed before any message is sent, so connection should be established not only before the delivery of the call information in the various session service states (SERVICE-1, SERVICE-2 & SERVICE-3), but also before the exchange of call control information during the transition among the states. Those connection needed in the various SERVICE states are referred to as call service connections, whereas those needed for the exchange of the call control information are referred to as call

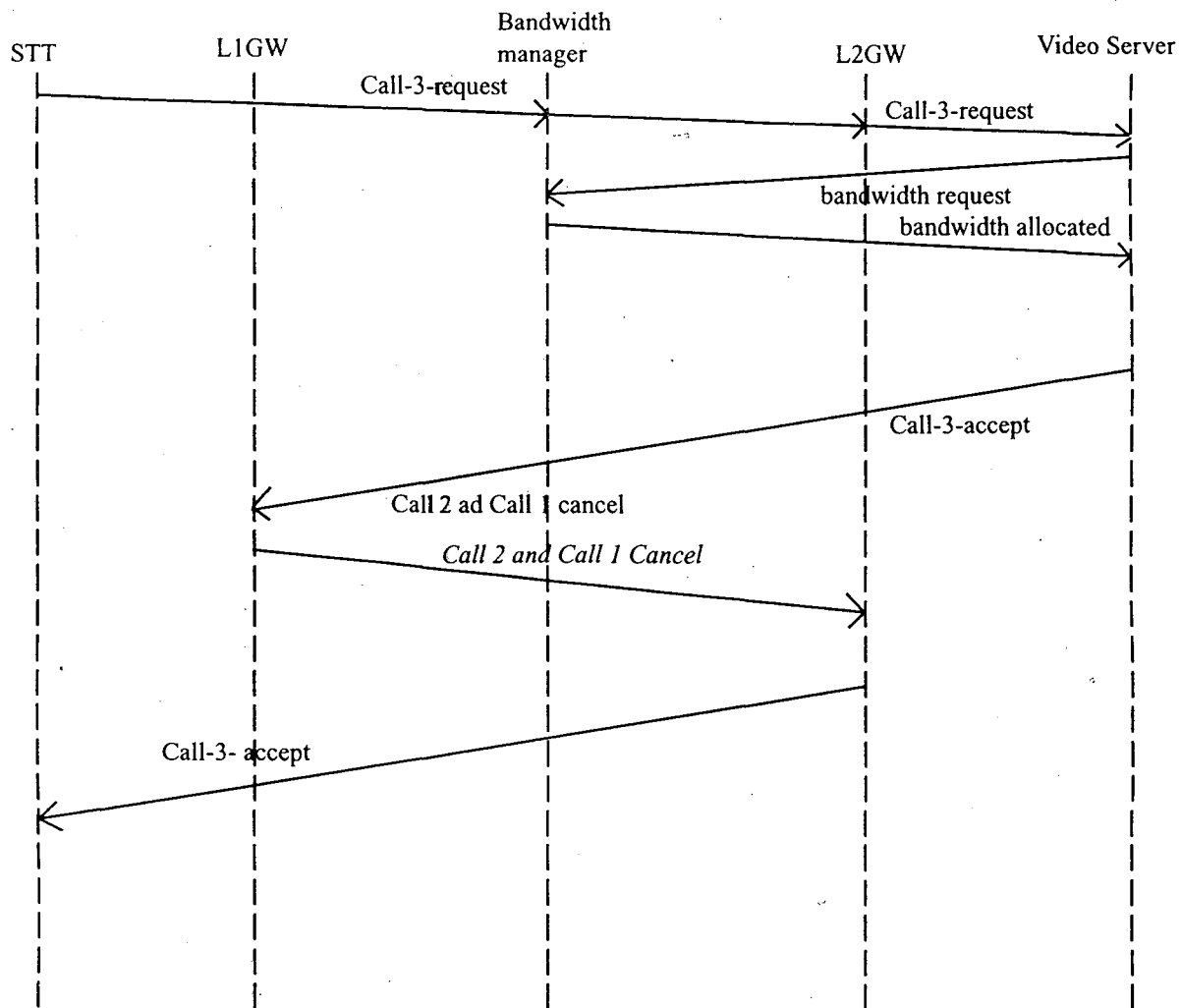


Fig. Call-3 setup distributed Approach

control connections. An important video-on-demand session management decision is when to setup all these connections.

The call service connections can be established during or after the execution of the call setup protocol. In figure we have shown the case where the connection in the SERVICE-3 state are established after the call-3 setup protocol is completed; i.e., after the STT and the video server receive a call-3 accept message from the L2GW. However, call-3 cannot be considered active and the STT session does not enter the SERVICE-3 state until the service connections are available.

In figure, we have shown the other option also, where the connections are established during the call setup protocol. When the bandwidth and other resources needed for the connections are granted, the L2GW asks the video server, or the STT, or both to setup the required connections. After the connections are established, the L2GW is informed, and call is considered active. The advantage of the approach, where connections are setup after the call setup protocol finishes, is that the call setup protocol last for a shorter period of time. This approach is suitable if entities participating in call setup protocol, like L2GW, need to complete the protocol before they are able to support other requests for service.

The second approach, where connection are setup during the call setup protocol, results in more compact solution because it couples connection management with call management. This is a more natural approach because the application connections are the most important building blocks of the calls. The call control connection between the two entities can be a semipermanent one, established in advanced for the exchange of call control messages between these two entities.

In ATM-based VoD networks the approach of coupling call and connection is taken one step further, by using the call setup protocol in order

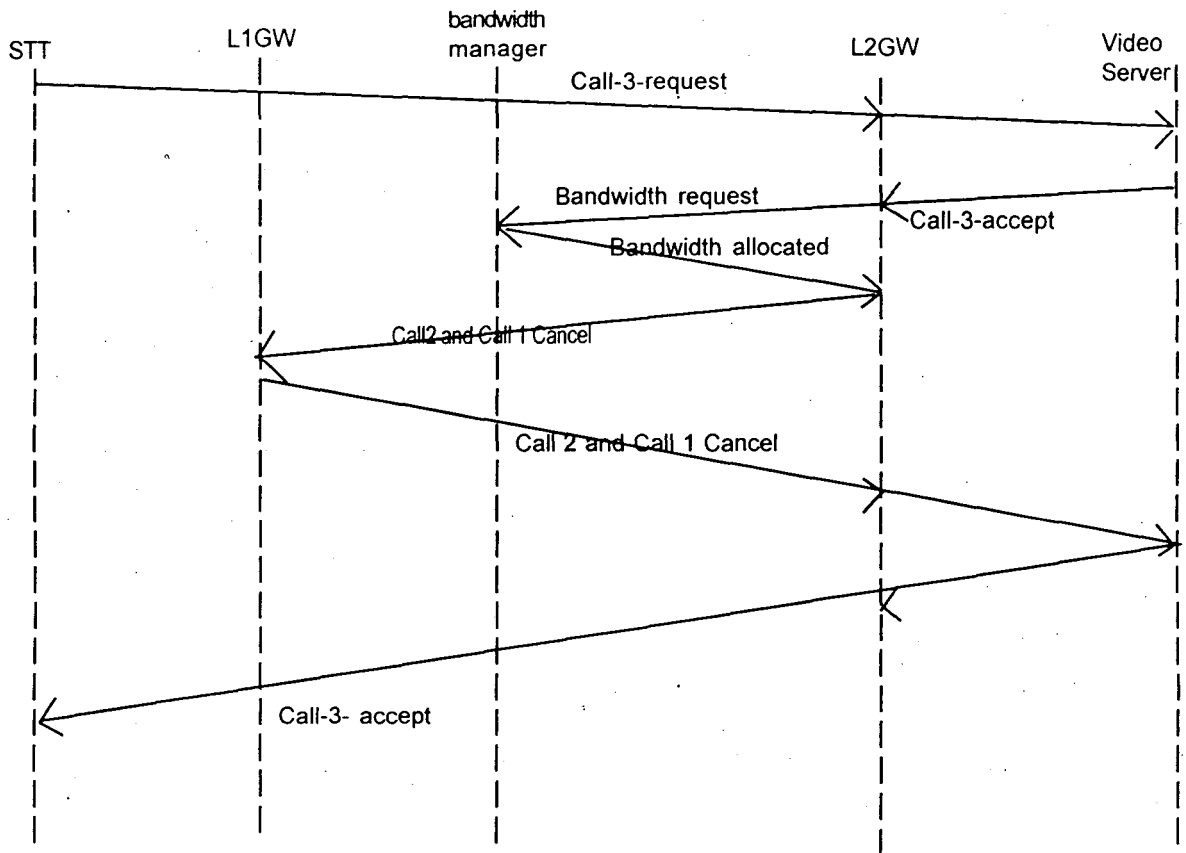


Fig. Service-3 application connections are established during call-3 set-up.

to actually setup the call service connection as well. Such an integration of the call control with connection control may reduce the total time and resources spent by the system for management and control operations.

We know that in ATM network two levels of the connections are defined: virtual channel(VC) connection, over which cells are routed, and virtual path (VP) connections, used as the building blocks of VC connection. Only the end nodes of a connection over which a VC connection has to be established need to participate in the VC setup protocol.

So if an appropriate lay out of the VP connection is setup in advance between the various component of the VoD system, the setup of the most of the VC connection can be performed by the call setup protocols.

Suppose that each L1GW has a VP with every L2GW and that every L2GW has a VP with each of its video servers. We also suppose that each STT that enters the NO-SERVICE state is connected by means of a VP to its L1GW.

This arrangement is shown in the fig. In such a case when the session moves from the NO-SERVICE into SERVICE-1 state, the call service VC connection needed between the STT and L1GW can be established over the existing VP. This information needed to be exchanged between the STT and the L1GW to this end, like VCI values or QoS parameters, can be attached to

the call-1 setup protocol messages. If the session moves from the SERVICE-1 state into the SERVICE-2 state, a VC between the STT and the L2GW can be established over two VP connections the L1GW.

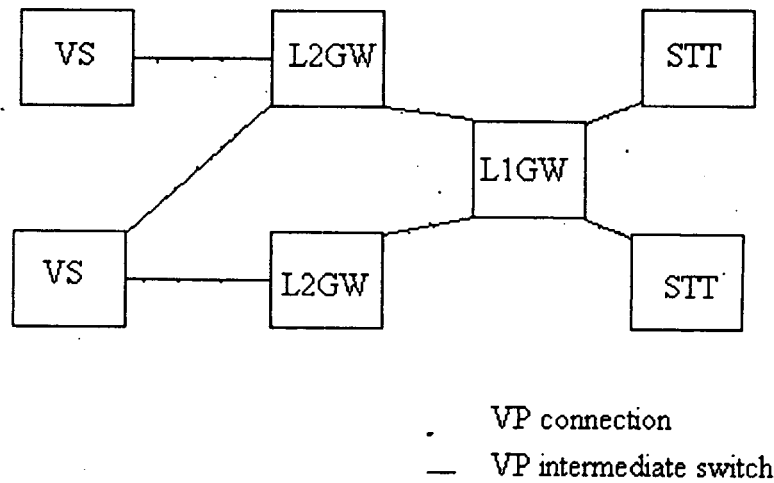


Fig. VP Layout in an ATM based VOD network



## *CHAPTER FIVE*

# **SESSION MANAGEMENT PROTOCOL**

The STT and Server are kept isolated from the details of the network control signaling by introducing a higher layer protocol through which the STT and server communicate with the Interactive Video Network (IVN) manager. This is a session layer protocol, as opposed to the network layer protocol, which directly controls the switched network. A session implies a relationship between an end user and a VIP for a period of the time during which the end user participates in interactive activity. Associated with the session are series of related connection through the network between server(s) and the user's STT. To allow the STT and the server to control the network at the session level, the IVN manager assumes the role of the session manager and proxy signaling agent. STTs and server interact with the IVN manager through session level protocol to establish sessions and add resources to a session. The IVN manager also provides the level-1 gateway functionality necessary to satisfy the FCC's video dial tone requirement for equal access to any VIPs.

The software architecture of the IVN manager is modular, allowing multiple network architecture to be built on a common software platform. To provide modularity the IVN manager software is divided into four major subsystems: communication protocols (specific to STTs, servers, and network hardware); session and connection management; subscriber services; and billing operations, and network management interfaces. Each network element that the IVN manager communicates with has individual modules in the

communication protocols subsystem. The module that comprise the session and connection management subsystem perform the session, connection, and resource allocation functions. The IVN manager provides:

- Network communications,
- Session and network resource management,
- Subscriber services, and
- Network provider support.

The session management function includes session set-up and tear-down. The network resource management function performs resource allocation for network resources such as bandwidth. The IVN manager is the end user's initial entry into the network and guides the end user in service selection. This provides the Level-1 gateway role for the local exchange carriers (LEC) networks. So the IVN manager performs the function of the session management, network resource management and allocation, and bandwidth management for the network. It correlates and tracks the resources used by the session across the network, including session establishment, session ongoing services, normal session release, abnormal session termination, and session auditing. During a session, it can add and drop resources, and even cancel the session. To setup the various paths for the session, the IVN manager, STT, and network access equipment use the network-level signaling protocol to established a session based connection between the IVN manager and the STT.

## 5.1 Signaling Reference Model -

Figure shows a signaling reference model for the interactive video network. We will give the session layer set of messages and procedure that is used to establish, manage, and tear down session between clients and servers

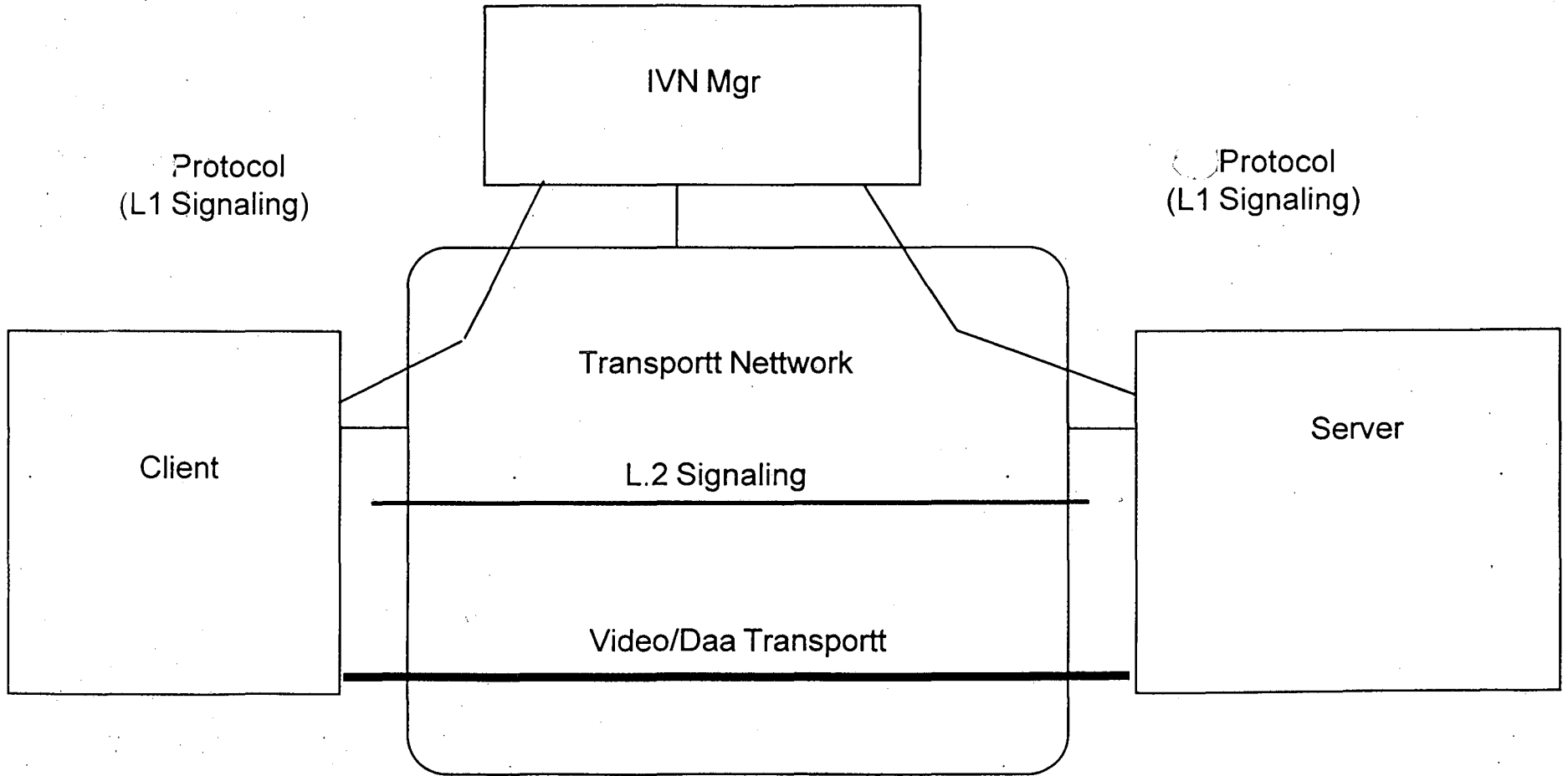


Fig. SIGNALING REFERENCE MODEL

over the network. The protocol defines the messages, scenarios, and resource structures for supporting session level management functions, such as:

- Establishing and terminating sessions, as requested by the client or server;
- Dynamic resource allocation/delocation during the life time of a session, as requested by the client or server;
- Forwarding and transfer of sessions and resources between servers and applications; including support of a distributed application environment; and
- Status and other OA&M activities for monitoring the state of a session.

As shown in the figure, clients interact with the IVN manager over a transport network to establish a session with the client's desired server. For ATM network that support switched services, Q.2931 protocol at the network layer is used to establish specific virtual circuits. Here we will give the signaling protocol between the network manager and the client as well as the signaling protocol between the network manager and the server. After a session is established, direct signaling between the client and server occurs via the user-to-user signaling protocol, without involving the IVN manager. Video and data are transferred over a unidirectional pipe from the server to the client. The control paths are typically bidirectional.

The signaling given here is non-associated. It allows Q.2931 signaling to be on different VC and different facility than the media stream which are established as a result of the signaling. The network signaling takes place over an ATM link connecting the IVN manager to the ATM switched network while the resulting switched connection are established between the server and client. Here the IVN manager acts as a proxy signaling agent. This relieves and even shields the client and servers from complex and detailed interactive video network signaling and management issues while giving them the control of the

network environment suitable for their applications. Non associated signaling allows for a single signaling link to a media server which may have multiple facilities for media output. There is no need for a signaling channel on each facility. This optimizes the network by simplifying provisioning and maximizing bandwidth available for media.

## 5.2 Protocol Stack -

In the context of the Open System Interconnection (OSI) protocol stack, the session management protocol resides in the session layer. By communicating at the session layer, the client is independent of lower layer details.

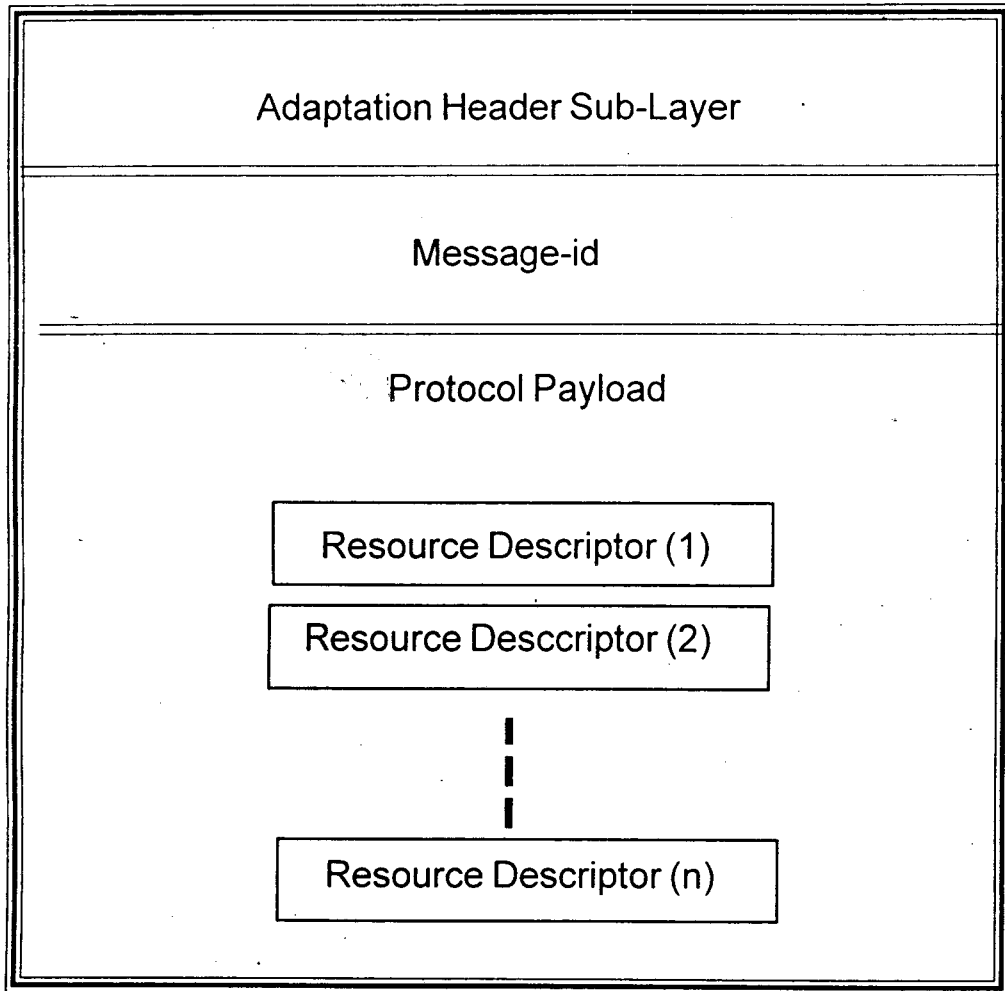
The protocol defines the procedure and the dialog between the client, server, transport network, and the IVN manager for the purpose of establishing and tearing down interactive sessions. To achieve this, the protocol provides mechanism to utilize various lower layer mechanism, such as Q.2931 or TCP/IP capabilities. An adaptation header layer is provided to allow a common interface between protocol messages and any other messaging protocols that may be transported over the lower-layer transport network.

## 5.3 General Message Format -

The format of protocol messages are shown in figure. The messages vary in the size depending on the number and the type of the resources being requested. Multiple resources of different or similar types could be assigned in one message.

*Adaptation Header -*

# Protocol General Message Format



resource -request-id
resource-id
resource-type
request type
resource-length
Resource-Data(1:m)
_____ byte
_____ byte2
_____ byten

An adaptation header layer is proposed to isolate the protocol from lower layer dependencies. It contains transport-dependent information, such as UDP or AAL information, in addition to any encryption keys.

. *Message-id* -

Message-id is a two-byte operation code for the message.

. *Payload* -

For each messages, there is a variable-size payload. It contains fields that identify a unique session and the resources associated with it. Resources are represented in terms of objects referred to as resources descriptors.

## 5.4 Resources -

A resource is a trackable object or element allocated by the network or network manager to a video session as required, and is retrieved by the network manager when the resource is no longer needed by the session.

### Resource Abstraction -

From a view of software technology, a resource can be viewed as an object with a set of characteristics. This abstraction of resources allows for the introduction of new network architectures and for the integration of new network elements into existing networks without the need for major changes to the protocol itself.

### Resources Descriptors -

Resources are requested in the form of resource descriptors and are allocated to the clients, servers also in the form of resource descriptors. A descriptor is a self-contained data structure that contains its identity and its resources. The resource descriptors is shown in figure.

There are two main types of resource descriptors :

- . resource descriptors for interactive video network users for requesting resources from the network manager.

- . resource descriptors for the interactive video network manager for allocating resources to network users.

The content of the resource descriptor is network specific (passband or baseband) and is related to the specific transport layer for the network. A passband system uses RF carrier modulation to transport signals through a shared access transport network, and a baseband system transmits the digital voice, video, and data signal directly as bits through a dedicated transmission path to the subscriber, different resource descriptors may be used for each.

## 5.5 Session Management Scenarios -

The Session management sequence of commands can be described in terms of scenarios. The scenarios consists of parametrized messages exchanged between the client and the interactive video network manager. When a protocol message is described, only parameters relevant to the script will be mentioned. Whoever sends the first message of a scenario, this scenario is said to be initiated from that side.

### 5.5.1 Session Set-Up -

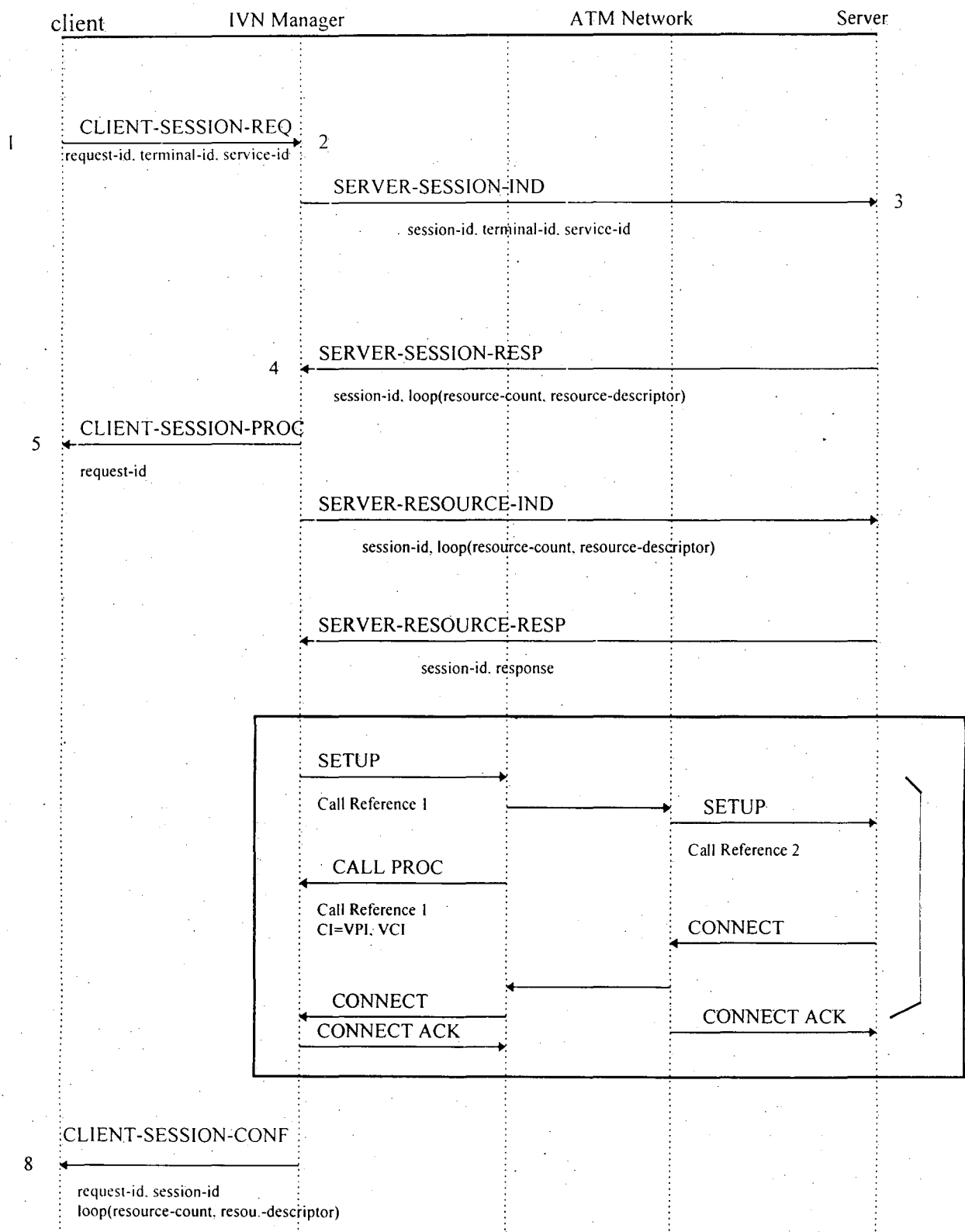
The session set-up request starts with an end user/subscriber selecting a multimedia service from the menu displayed on a television screen. This selection triggers the set-top terminal (STT) connected with the television,



acting as the client, to initiate a session request message. The client sends a Client-Session-Request message to the interactive video network manager. This message contains a terminal-id, which identifies the client and a service-id, which identifies the selected interactive service. Since first the client selects a particular service provider's menu so in this case the service-id will be the server-id of service provider.

The session control entity in the network manager verifies the terminal-id from the network provider's point of view. After the IV network manager has validated both the client and server, it send the Server-Session-Indication message to the indicated server using low-level transport and network address of the server (e.g. IP address or E.164 address) that the IVN manager maintains. This message contains the client-id, service-id, and session-id parameters. The session-id is assigned by the IVN manager to identify the session being established between client and the server, for the purpose of maintenance and billing. The client only knows about the high level service-id.

Upon receipt of the Server-Session-Indication message from IVN manager, the server verifies if the client (identified by client-id) can receive the service (identified by service-id), and if positive, accepts the request. After accepting the request, server starts to identify all network resources that it needs in order to support the selected service. Each connection's required attributes are described in a connection resource descriptor. Each resource descriptor include a resourceNum allocated by server. The Server-Session-Response message contains "resource-count" resource descriptors. If the service needs ATM SVC connections, the server will include the proper number of "ATM SVC" resource descriptors. After determining these resources details the server sends a Server-Session-Response message to IVN manager.



Only Relevant parameter in each message are shown

Figure. Session Setup

Upon receipt of the server session response message, the IVN manager processes the resource descriptors included in the message. It goes through the request list and try to allocate the resource. For an "ATM SVC" resource, the IVN manager verifies the requested properties, such as user cell rate and QoS, against the network provider's, high layer policies, such as rejecting requests of bandwidth greater than 10 Mb/s during evening hours. If the IVN manager fails to allocate a single non negotiable resources, the whole request list fails. The whole request list also fails if the IVN manager fails to allocate all negotiable resources. If some but not all negotiable resources fails, the protocol allows the IVN manager to propose alternate values back to the server. The IVN manager translate parameters in the "ATM SVC" resource descriptors into Q.2931 information element, parameters. The IVN manager constructs a resource assignment descriptor for each allocated resources where it includes the resource-id and all the relevant information about the assignment. The resource-id uniquely identifies the allocated resource within a given session. The IVN manager then collects all assigned resource descriptors into a server resource indication message, and sends it to the server, indicating the requested connections are available.

When the server receives the server resource indications message from the IVN manager, it iterates through the resource assignment descriptors. For any negotiable resource that fails, the server has the option to accept the alternate values proposed by the network or not. In any case the server sends a reply in server resource response message indicating its decision.

When the IVN manager receives a server resource response message signal from the server indicating server's acceptance of the resource allocated, it initiates a Q.2931 non associated call/connection procedure.

The generic identifier transport information element (GITIE) has been defined to hold a session identifier and resource number. The GITIE may be

included in the Q.2931 SETUP/CONNECT messages, as required. On the upper layer application, the GITIE at connection setup includes session-id + resource num=resource-id. This permits the applications at the client, network, and server, to associate ATM connection to the session.

## Q.2931 Protocol

The IVN manager initiates a non associated call/connection procedure by sending a Q.2931 setup message by the IWU to its ATM user-network interface(UNI). The call reference used will be selected by the IVN manager. The following information element is sent to the ATM UNI.

. Calling party number = ATM address of the client as derived from the terminal id field .

. Called party number= ATM address of the server as derived from the resource descriptor field within the server session response message. This feature allows a server to redirect the recipient of the ATM call to any distributed end points it manages.

. ATM adaptation layer parameter, ATM user cell rate and QoS parameter values imported from the ATM connection resource descriptor field within the server session response message.

. GITIE = resource-id corresponding to this ATM connection .

When the server is informed of the resource SETUP ON its ATM UNI, it obtains the resource-id from the generic identifier transport IE in the SETUP message, and associates the connection to the session.

The server maintains this association until the ATM SVC connection is released. The IVN manager maintains the relationship between the call reference at its UNI and the session-id / resource-id pair to maintain its ability to retrieve either one from the other.

The ATM network acknowledges the SETUP with a CALL PROCEEDING message . From that message, the assigned VPI and VCI values and the resource-id will be used by the IWU to connect the client to the ATM SVC connection being setup.

After connections are established, the IVN manager informs the client through the client session confirm message. The list of resource descriptor in this message informs the client of the connection resources it will use. Both client and server are now ready to exchange user to user message over the acquired connection(s). Now the client can tune to those resources and starts to receive the service from the server. The subscriber now sees the welcome screen of the service he/she has selected a few second earlier . So client receives the list of movies for the service movie-on -demand.

Based on the above analysis the algorithm for Session Set-up is designed. It shows how the client, Server, and IVN manager communicates with each other for Session Set-up.

## Algorithm for Session-Setup -

### Client -

At the time of starting Session the client uses following algorithm for Session Setup.

Send CLIENT-SESSION-REQ(request-id, terminal-id, service-id) to IVN manager

while (!receive CLIENT-SESSION-CONF from IVN manager)

{

    receive message from IVN manager;

    if (message got == session not possible)

        exit();

}

start communication with server

### IVN Manager -

The interactive video network manager uses following algorithm at the time of session setup.

On receiving CLIENT-SESSION-REQ it processes as below-

verify terminal-id, service-id in CLIENT-SESSION-REQ

if (not valid) then

{

    send message to client - session not possible

    return

```

    }
    send SERVER-SESSION-IND(session-id, terminal-id, service-
id) to server

    on receiving SERVER-SESSION-RESP(session-id,resource-
list) from server

    if (server do not accept session)
    {
        send message to client - session not possible
        return
    }
    check resources needed for session and allocate them

    if all necessary and some negotiable resources are free then
    {
        send SERVER-RESOURCE-IND(session-id, list of
allocated resources) to server
        receive SERVER-RESOURCE-RESP(session-id, response)
from server

        if (responce is affirmative)
        {
            SETUP-CONNECTION

            send CLIENT-SESSION-CONF(session-id, resource-list)
        }
        else
        {
            request has failed
            send message to server and client
        }
    }

```

Server-

While accepting the Session Setup request the server uses the following algorithm-

confirm/verify if the client can receive the service request

if (not accepted the request)

{

send message to IVN manager - service not accepted

return

}

Send SERVER-SESSION-RESP(session-id, resource-list) to

IVN manager

if (receive a message that session request failed)

return

if (receive SERVER-RESOURCE-IND(session-id, allocated resource-list))

confirm if Session can be Setup with allocated resources

if (resource not sufficient)

{

send non affirmative decision to IVN manager

return

}

else send the affirmative decision to server



After all connection are setup

Start Communication with client

### 5.5.2 Resource Request-

After a session has been setup between the client and the server, what the client receive may be just a list of movies. The initial resource that the server has requested may be enough to support a low bandwidth menu navigation need and some film clips. The subscriber now selects a movie from the list. This interaction is done between the client and the server, without going through the IVN manager, since the resources to conduct such interaction has been established earlier in the session request. Now to support a full length movie the server needs new resources. So the server sends to the IVN manager a server - add resource request message which contains the session-id parameter and the resource descriptor of the new resources requested by server.

The IVN manager processes the resource descriptor as before. For each resource allocated, it creates a resource-id to represent the allocation and constructs a resource descriptor which includes the resource-id and the associated resource information. Descriptors of all the resources are allocated into one message. Now the IVN manager returns the result to the server using the server resource indication message.

Now the server receives the server resource indication message. It processes the resources as assigned by the IVN manager and acknowledges

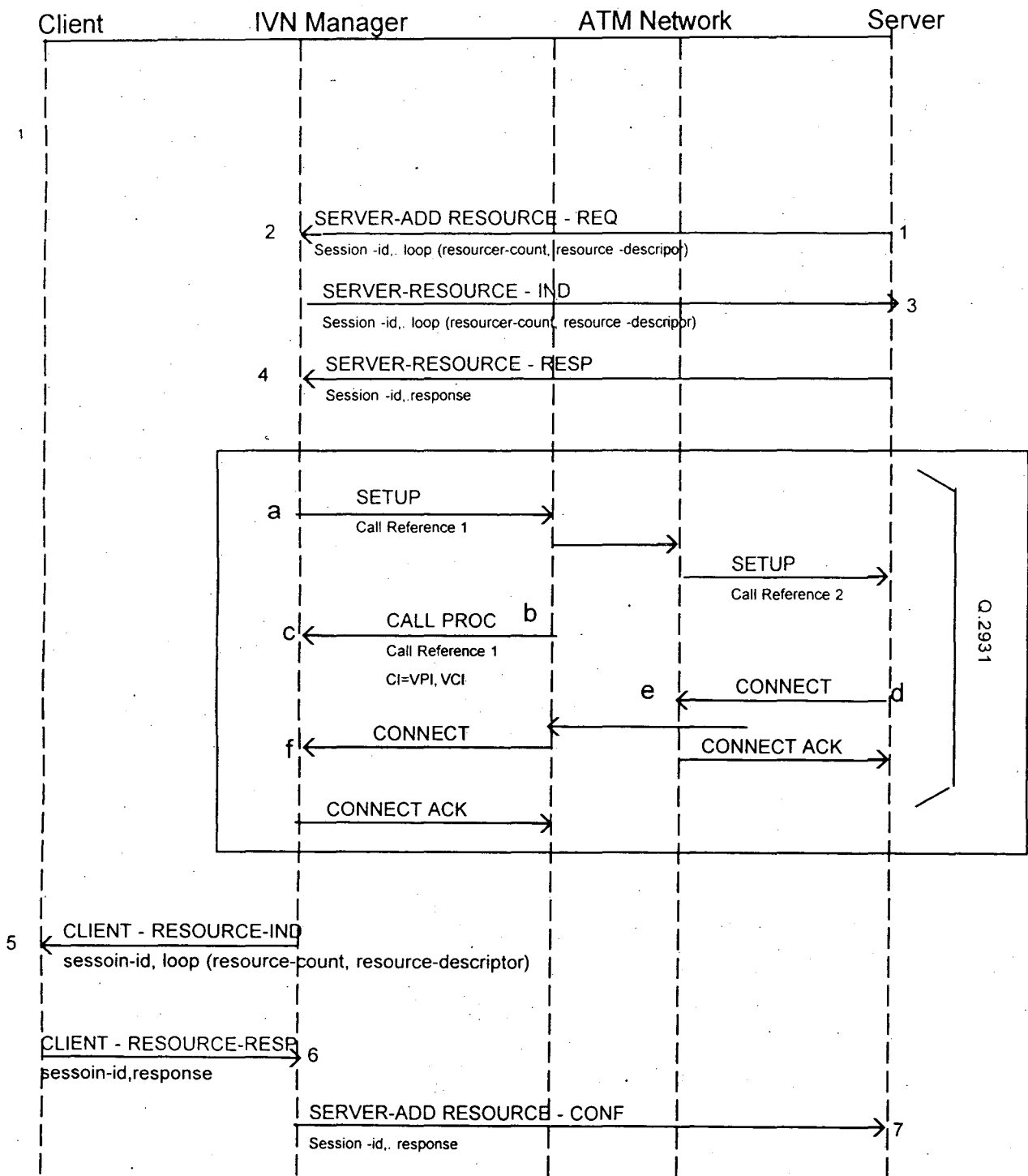


Figure - Resource Request

this message using the server resource message, indicating its acceptance or rejection of the resources.

When the IVN manager receives the response message with an acceptance, the IVN manager initiates a Q.2931 non associated call/connection procedure at its ATM UNI to add more ATM paths over the transport network.

Now the IVN manager constructs resource descriptors corresponding to the resources requested by the server. After the connection are established the IVN manager bundles these resource descriptors in one message and sends to the client as client resource indication message.

After receiving the client resource indication message the client responds this using client resource response message. On receiving this the IVN manager send a server-add resource confirmation to server. Now the client receive data from the server via the newly allocated resources. The subscriber now sees the movie on television screen.

The algorithm showing how the client, IVN manager, and server are involved in the resource request is presented below.

### Resource Request Algorithm-

Server-

When server needs some more resources-

Send SERVER-ADD-RESOURCE-REQ(session-id, resource-descrip) to IVN manager

Gets the SERVER-RESOURCE-IND from IVN manager

if the reponse is affirmative

{

go through all the resources allocated

if it suits the need then

Send positive SERVER-RESOURCE-RESP to IVN manager

else

{Send message to the client and

negative SERVER-RESOURCE-RESP to IVN manager

}

}

else

{ send resource not available message to client

return

}

wait till confirmation of addition of resources arrive

start communication with client

### IVN Manager -

On receiving a request for additional resource

Check if the request resources can be allocated

if all necessary and some negotiable resources are free

```

    {
        allocate resource
            Send SERVER-RESOURCE-IND(session-id,allocated
resources-list)
    }
else
    {
        Send SERVER-RESOURCE-IND(resource not available)
        return
    }
On receiving SERVER-RESOURCE-RESP(session-
id,response) from server
    if (response is affirmative)
        SETUP CONNECTION
        Send CLIENT-RESOURCE-IND(session-id,resource-list)

    On receiving the response/ack. from client
        Send SERVER-ADD-RESOURCE-CONF(session-id,
response) to server

```

### Client-

```

    On receiving CLIENT-RESOURCE-IND(session-id,list of
resources used)

```

Add the resources to available resource table

Send the acknowledgement to the IVN Manager

### 5.5.3 Resource Release-

After a time the requested video (e.g. movie) ends . The server does not want to terminate the session since it is possible that the client's may want to select another movie .So the session returns to the menu of available video.

As soon as the movie ends the server request the IVN manager to release the resources, in order not to be charged for the high bandwidth resources setup to support movie. First the server uses the server delete resource request message to inform the IVN manager of it request to delete one or more assigned resource. This message includes the session-id parameter and the resource-id of any resources the server wants to release .

The IVN manager then informs the client via client delete resource indication message to stop using the identifies resources. On receipt of the client's -delete-resource response message ,the IVN manager processes the deletion of the identified resources from the established session .

If the IVN manager detects a resource-id identifying an "ATM SVC"resources , it retrieves the corresponding call reference and initiates a Q.2931 Call/Connection clearing procedure at its ATM UNI with a Q.2931 RELEASE message . After all resources have been deleted , the video manager informs the server of the operarion's outcome using the server delete resource confirmation message.

At the time of the resource release the server, IVN manager and client uses following algorithm.

Algorithm -

Server -

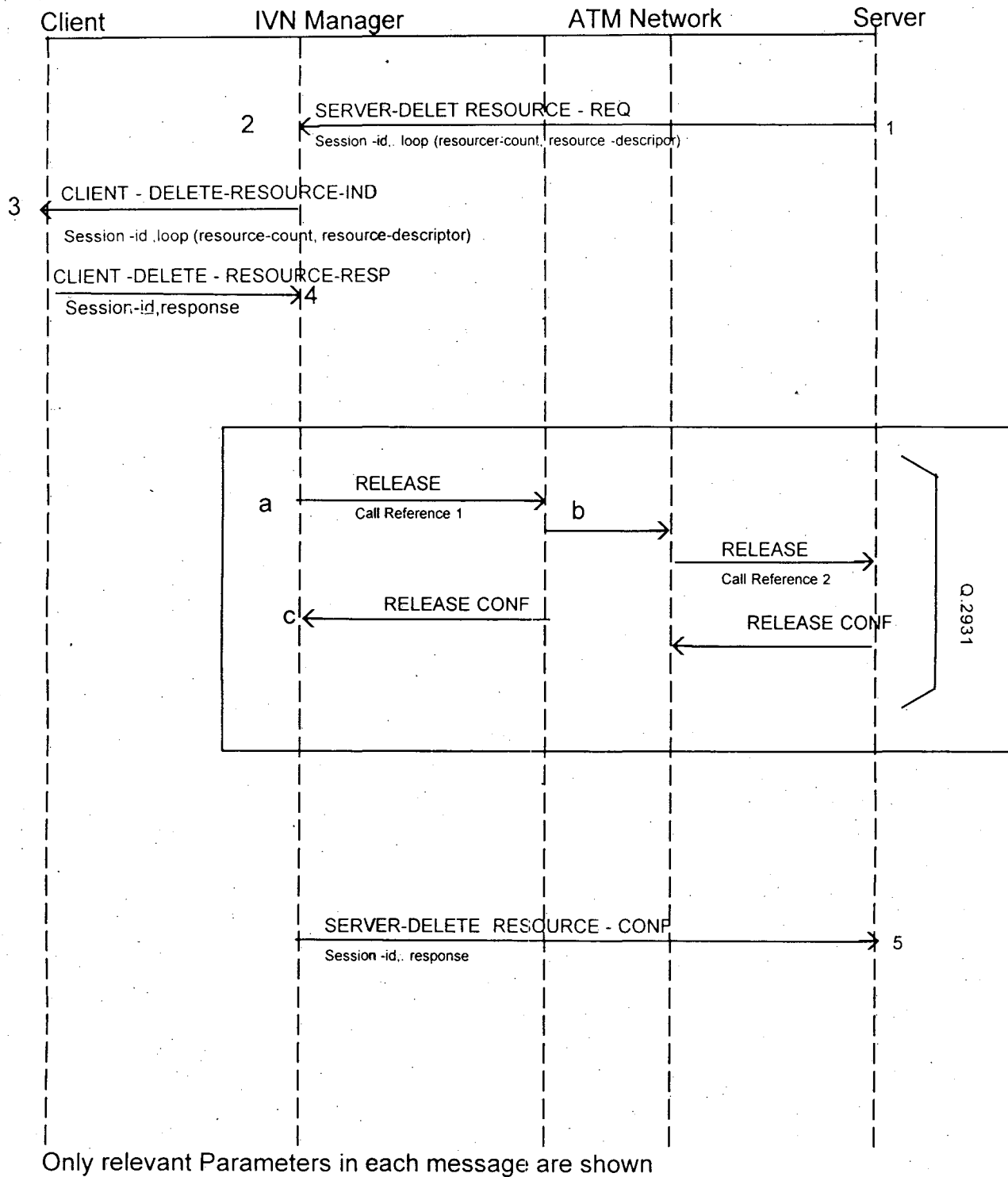


Figure - Resource Release

The server uses the following algorithm to release the resources

Send SERVER-DELETE-RESOURCE-REQ(session-id,resource-list)

Wait for the confirmation from the IVN manager

IVN Manager -

On receiving SERVER-DELETE-RESOURCE-REQ(session-id,resource-list)

Send CLIENT-DELETE-RESOURCE-IND(session-id,resource-list) to Client

On receiving CLIENT-DELETE-RESOURCE-RESP from Client

Delete the allocated resources and move them into resources

Send SERVER-DELETE-RESOURCE-CONF to server

Client -

On receiving CLIENT-DELETE-RESOURCE-IND(session-id,resource-list)

Modify it's table of resources that can be used by it



Send CLIENT-DELETE-RESOURCE-RESP(session-id, response) to IVN manager

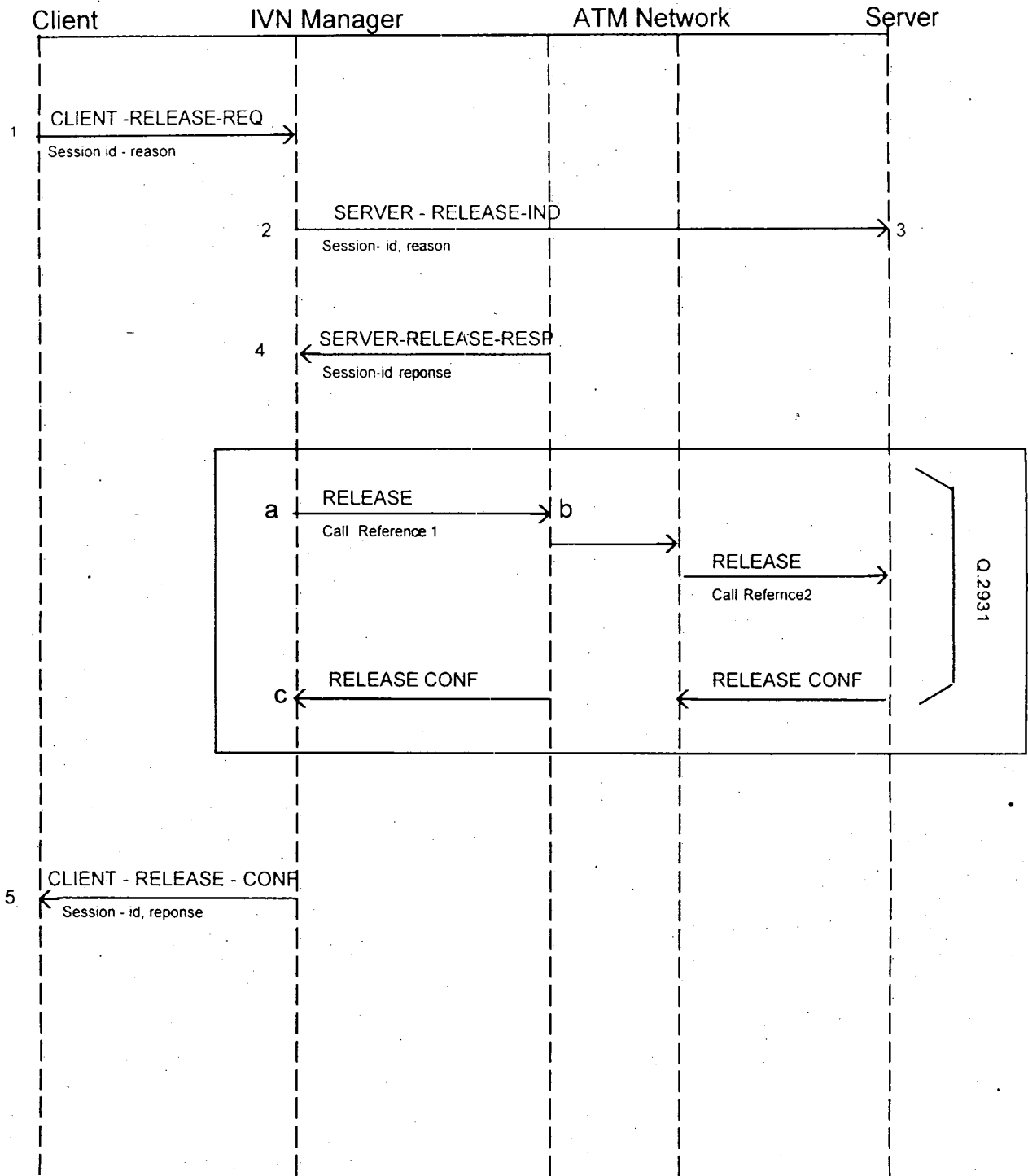
#### 5.5.4 Session Release-

After a session has been established through the session setup, either the client or server can come back to the network to request that the session be torn down. The user wants to tear down the session since user does not want any more movies. He/she then exits the server and then get back to the standard menu of server. This will trigger the STT, acting as a client to initiate a session tear down request.

The client sends a client release request message to the IVN manager containing the session-id. The IVN manager in turn send the server associated with that session, a server release indication message. The server acknowledges the request with a server release response message.

When the server release response message is received, the IVN manager retrieves all resources allocated to the identified session and deletes them. If the IVN manager detects an "ATM SVC" resource belonging to the session, it retrieves the corresponding call reference from the resource-id and initiates a Q.2931 call/connection clearing procedure at its ATM UNI.

After all the resources have been deleted, the IVN manager informs the client of the outcome of the operation using the client release confirmation message.



Only relevant Parameters in each message are shown

Figure - Session Release

## Algorithm-

The Client, IVN manager and Server uses following algorithm at the time of Session Release.

### Client -

For releasing a session the client's follows -

Send CLIENT-RELEASE-REQ(session-id, reason) to IVN manager

loop while (CLIENT-RELEASE-CONF does not arrive)

exit

### IVN Manager -

On receiving CLIENT-RELEASE-REQ(session-id,reason) from client

Send SERVER-RELEASE-IND(session-id,reason) to server

On receiving SERVER-RELEASE-RESP(session-id, reason) from server

RELEASE all the resources of the session

Send CLIENT-RELEASE-CONF(session-id, response)

Server -

If receive SERVER-RELEASE-IND(session-id, reason) from  
IVN manager

Check the message

Send SERVER-RELEASE-RESP(session-id, response) to IVN  
manager

## CONCLUSION-

We have developed a session management protocol that fills in the void of control plane and compliments the enabling technologies and standards nicely. A model that subclassifies the video-on-demand control and management procedures into a three level management hierarchy of session calls and connections is developed. The relationship between procedures in the three management levels and option and tradeoff involved in the design of the VoD session is presented. The protocols level-1 capability is the middleware for the control plane with an API interface for level-2 application. This middle ware is equivalent to the OSI session and presentation layer services.

Though we have presented a simple VoD system, where two calls must precede the selected video application, more complicated systems can be addressed in the same way. We have shown how the session management protocol work on ATM and how connection management can be optimized, taking advantages of the inherent features of ATM. Since a VoD session can potentially comprise more than one ATM connection, so we have shown that using a GIT IE in the Q.2931 SETUP message allows the connections to be identified and utilized at a higher layer. With an open standardization effort, we feel that our work provide benifits to the young video-on-demand industry.

## Appendix

### Abbreviations, and Terms

ATM	-	Asynchronous transfer mode
CATV	-	Cable television
DSM-CC	-	Digital Storage media Command and Control
FCC	-	Federal Communications Commission
GUI	-	Graphical user interface
HDT	-	Host digital terminal
HFC	-	Hybrid fiber-coax
ISDN	-	Integrated services digital network
ISO	--	International Organization for Standardization
IVN	-	Interactive video network
LEC	-	Local exchange carrier
MPEG2	-	Motion picture Experts Group-2
NTSC	-	National Television Systems Committee
OA&M	-	Operations, administration, and maintenance
OSS	-	Operations support system
PIN	-	Personal identification number
PVC	-	Permanent virtual circuit
QOS	-	Quality of service
RF	-	Radio frequency

SONET - Synchronous optical network  
L1GW - Level 1 gateway  
L2GW - Level 2 gateway  
SDV - Switched digital video  
STT - Set-top terminal  
SVC - Switched virtual circuit  
UNI - User-Network interface  
VIP - Video information provider

## BIBLIOGRAPHY

- (1) Change Li Lin, Sheng G., "The Design and Architecture of a Video Library System," IEEE Communication magazine January 1996
- (2) Daniel D. Hilem & Henari V., "Interactive Video on Demand," IEEE Communication Magzin" PP 82-88, May 1994
- (3) Dilip D. Kandlur, Dehanjan S., "Protocol Architecture for multi media Applications our ATM Networks," -- IEEE J SAC, vol.14, no.7, pp.1349-1359, Sept.1996
- (4) Hans R. Appenzllier, "Signaling Sys. no.7 ISDN user part", IEEE SAC, vol. 4, no.3, pp.366-371, May 1986
- (5) Ian F. Akyildiz & Weiyen, "Multimedia Group Synchronization Protocol for Integrated Service Digital Network," IEEE Comm. Mag ,pp.52-59
- (6) Jaime J. Bae & Ttatsuya S., "Survey of Traffic Control schemes & Protocols in ATM networks." Proceeding of IEEE, Vol.79 Number to PP 170-189 Feb.1991.
- (7) J. Richred Janes, "Basebond & pass band system for Inter active video services" IEEE communication mag. PP 90-101
- (8) Jean Paul M, Bawjn V. Ratel, Franks S.R, "Networking Requirements for Interactive video on demands," IEEE comm, mag Aug. 1996 PP. 128-133.
- (9) P. Venjka & Srihari , "Continuity & Synchronization in MPEG," IEEE SAC Vol. No.1 Jan 1996, PP 53-60
- (10) Rofer E Hibman & Irene J. M. & Huns, "The interactive video network," AT&T technical Jouenal Sept. /Oct. 1995, PP 92 -105
- (11) Steven Menger, A Protocol for complex multi media services," IEEE SAC Vol. 9, no 9, Dec. 1991, PP1383-.1394
- (12) Shoshana Loeb, "Delivering Interactive multi media Docoument over ATM networks," PP 52-59 May 1996.
- (13) Tanenbaum, A. S. 'Computer Networks, second Edition," PHI 1996.



- (14) Vahe B., Liam C. and Nancy , "Digital Storage of media Command and control protocol Applied to ATM" PP 1162-1172. IEEE J. Selected Areas communication, Volume 14. No. 6. Aug. 1996.
- (15) Y.H. Chang, David C. "An open system approach to video on demand," IEEE Communication Mag. May 1994 PP. 68-80
- (16) Victor U.K. Performance modoul of Interactive Video on demand system," IEEE J. Selected Areas Communication , Volume 14. Nno. 6 August 1996. PP. 1099-1109.
- (17) W.D. Sincoskie, "System architecture for a large scale video on demand SreVICES," Computer network and ISDN systems, vol. 22, pp 155-162, 1991
- (18) William Sttillings, "Data and Computer Communication" PHI, 1996.
- (19) William & Donglas, "Signaling Challenges in Video dial Tone Networks" IEEE Communication Magazine, Aug. 1996 PP 128-133.