MULTIMEDIA COMMUNICATIONS

Taught at MIT Fall 1989, Revised 2021

TERRENCE MCGARTY

Multimedia Communications

Terrence P. McGarty Ph.D.
Massachusetts Institute of Technology

Fall 1989

Table of Contents

PREFACE (2021)

This is an edited version of the 1989 class notes on Multimedia Communications which I taught at MIT. I have tried to keep all the original material and eliminated some material which is so out of date that it has little use. However, in view of the slow growth of multimedia communications over this some thirty plus years, I felt an edited and useful version may be of interest. This is NOT a view from the current time but an edit of what was known in 1989. It is worth looking at this and comparing what has been accomplished and what is left undone. The basic and fundamental issues and ideas have not changed. The technology has advanced but the integration of that technology into a true and working multimedia communication environment has not.

The current virus pandemic has placed an emphasis on the need for true multimedia communications as presented in this old notes. Two areas are of most import are education and medicine. Students are left with useless laptops trying to participate in classes using the limited abilities of a Zoom or some other primitive form of communications. In medicine we have telemedicine that uses the same limited technology and, in my opinion, does less than what I was doing at Harvard in 1986-1990. Instead, the current technology implementation has, in my opinion, wasted decades on such things as Apps while leaving the two areas just mentioned a wasteland.

There is a philosophy of multimedia communications, a philosophy as to what we are trying to do as humans in interacting with one another human or group of humans and even our environment. Multimedia communications is ***displaced human interactions***. It tries to erase the sense of displacement; it tries to bridge the gap of technology qua technology. None of the systems currently available have even tried to understand this human interaction challenge. One cannot accomplish this with a screen on a smart phone. A laptop shining in one's eyeglasses creates an ethereal image of unreality. Poor speakers and microphones present an almost comedic interaction, and of course the background is more like a reality home show than a true normal interaction. Finally, the technology itself gets in the way of interacting.

In 1989 I tried to understand and remedy such issues with a philosophy of multimedia communications. Unfortunately, most technologists who design and deploy today's systems are devoid of any understanding of human interactions.

Multimedia Communications must be "displaced human interactions". It is a simple concept but complex to execute. It must be a seamless and as Heidegger noted "ready at hand" and a simple "thing" that disappears when used, making the users act and communicate as if the distance was not there. This is a highly complex challenge. It requires understanding human interaction, human linguistics and the very nature of humanity at the highest level.

This current viral pandemic has demonstrated how poorly we have achieved this goal. Just the thought of using a smart phone to take university classes, or even primary school classes is an insane idea. This is the App world of Silicon Valley, a short term monetizable artifact that has set humanity back, and not allowed us to move forward. How can a Second-Grade teacher interact with her students? Clearly not with what we have now. How can a father read to his daughter, not as we have now? How can a Physician interact with a cancer patient unable to come to the clinic, not as we have now?

Thus, the issuance of this slightly modified revision I believe has value and merit.

I want to also provide a belated thanks to Professor Muriel Medard (MIT) who was my Teaching Assistant during this first course. She did a magnificent effort in assembling key references and sources and in managing the class structure. It is to her and my many other students to whom I dedicate this revision.

Terrence P. McGarty
Thornton, NH January 2021

PREFACE (1989)

This text deals with the ~~ma~~ny ~~i~~ssues that relate to the field of multimedia communications. Multimedia communications is a new field that concerns the creation, manipulation, transmission and processing of complex images, text, graphics, voice and other multimedia entities. Unlike the fields of computer communications or even voice communications, the multimedia environment presents a wide variety of new and complex challenges. The very nature of the message, complex images as well as voice, and the more complex issues of human interfaces, make this a challenging field.

This text develops the field by first providing a structure to the multimedia environment and then developing a theory for its processing, transmission and manipulation. The field of multimedia communications includes the issues of human interfaces, characterizations of complex images, the abstractions of distributed data base and distributed operating systems as well the interaction of multi users in a multimedia environment.

This text grew out of the work that the author has been involved in over a ten year period. This time period included involvement in such diverse areas as CATV, personal computers, image processing, broadband communications and ultimately multimedia systems. It had become quite clear over this period that multimedia communications was becoming a technology unto itself, addressing many issues that combined the interests of many existing fields but also asking questions that had not yet been posed. The key to understanding multimedia communications is not to understand how computers communicate but how humans communicate. Thus the experience that the author has developed in the context of a multitude of end user applications has played a critical role in addressing the issues that led to the development of this book.

This book has been written for the use of students who are studying communications, computer systems or media applications. It is also useful for those practicing in the field and what are developing new and innovative multimedia systems. It is also hoped that it can be used to inspire many of those who have entrepreneurial tendencies to implement many of the theories discussed in this book.

The author would like to thank many of the people who have helped him in the development of this text at NYNEX and MIT.

Terrence P. McGarty Waterville Valley, NH September, 1989

1 INTRODUCTION

Communications has evolved from the simplest form of telegraphy to voice and video and into the efforts associated with computer communications. The challenge of communications is now to deal with the ability to use multiple media simultaneously. For example, if we look at the way creative graphics designer's function, we see that there is high interaction between the images, the spoken word, the use of facial remarks and most importantly the movement of the hands. The ability to fully communicate the creative process is the essential element of the multimedia communication design.

1.1 DEFINITION OF MULTIMEDIA

Media, for the purpose of this book, represents any form of storage or presentation means that allows for the transmission of information from one user to another. Information in its more classic sense is basically the reduction of ambiguity or entropy in the understanding of some concept. Thus the use of a picture, a way of expressing a thousand words, is an alternative means of providing information through a specific form of media. Images, speech, text, graphics are all forms of media. The simultaneous use of all of these forms of media in a session dedicated to information transfer is the essence of multimedia communications.

The history of computer communications is based upon the less than complex transmission of-information by means of simple binary messages and increasing the complexity to that of text and simple graphics. The introduction of more complex image formats is increasing and it is the use of these formats that dramatically changes the way we communicate in a multimedia fashion. In a simple text communications format, whether it be telex of even high speed computer networking, the user is forced to follow the standards of the computer network. The messages are packetized, sending a single message, a single letter or at most a single set of words at a time. The introductions of graphics packages were really a way to attempt to provide images within the constraints of the computer network. True image communications, encompassing the full interactivity of a multimedia information exchange has been severely limited.

A true multimedia multiuser information exchange occurs in a typical ad copy approval session between the client, the ad agency, the publisher and even the pre-press house. The communications involves the use of the images, the use of the hands and voice and the interaction of eyes, body language, charts, graphs, numbers, and finally even the printed words. Multimedia communications attempts to develop the theory and structure associated with such information transfer transactions between sets of individuals. Computer communications has

typically focused on information transmission and transaction between computers and infrequently tolerated the interaction with humans, but always on their terms.

At the heart of multimedia communications is the interpretation that this new media-can bring to bear on our very relationship to what we perceive as knowledge. Marshall McLuhan, the academic who was made famous by his book "The Medium is the Message" is quoted by Peter Drucker at the time he was defending his own doctoral thesis. At that time McLuhan was discussing how Gutenberg had dramatically changed the character of the middle ages with the introduction of the printing press. As he developed the theses to his academic reviewers, they all shook their heads in agreement, that is until he reached his final conclusion. That conclusion, McLuhan stated was that the change in presentation of information changed not only how knowledge was transferred, but more importantly. WHAT WAS KNOWLEDGE. The medium of transfer of information actively altered what was information.

We frequently do not readily understand what McLuhan was saying and look at knowledge on a human time scale. If we were to go back 2,000 years we would see that knowledge was limited to what could be memorized or at most what was kept on the fragile documents in the few libraries such as the one at Alexandria. If we look at the knowledge of those times we find a strong oral tradition and the result was the limitation of ideas to those that had a simple oral rendering. If we look at Newton's works on gravity we find the vestiges of that oral tradition. The presentations in written form still follow the oral form with limited equations and even fewer limited diagrams. In current books on gravitation, there is a plethora of equations and the diagrams are complex and much more enlightening. As we step even further, the use of super computers can now allow these theories to be displayed on high-resolution displays in dynamic form. The knowledge of such thing as fluid flow now is viewed not just as equations, graphs and figures, but as the flow of simulated fluids through simulate boundaries.

As we move towards the current days with the use of film and video, we find that knowledge is viewed as what is on film. The anecdotes, whether they are true or not, of recent presidents believing reports only if they are on film rather than written, show again how the change in available media change what is knowledge. Knowledge is the image or vision of the camera and its ability to position certain ideas in the eyes of the viewer.

As we move forward in the use of new media, especially those that allow for multimedia and multiuser interactions, we further build upon McLuhan's concept that changing media changes what is knowledge. The book as a learning medium will change. The book generally is a static medium that is linear in form, progressing from chapter to chapter. The book is built up by the author to present a sequence of concepts to the user. The book is non interactive, it does not allow for the asking of questions and in response provide a new reordering.

As Nathan Felde has stated, education is a low bandwidth process, entertainment is a high bandwidth process. Education is a questioning nonlinear an iterative process that frequently uses many media. Entertainment is a nonquestioning linear process that lacks significant interactivity. We must recognize that the multimedia user frequently is trying to be educated, that is transfer and absorb information, not be passively a recipient of high bandwidth entertainment.

1.1.1 The Image

We shall expand the concept of an image to mean the embodiment of any media based information bearing element. For example, an image may be a picture, a segment of a voice conversation, a piece of graphics display, or even a segment of binary data. In the general sense, if we let M represent any such image, then we define the image in terms of is extent, E(x,y,z,t), its information content, I, and its location L(x,y,z,t).

The image will be at the heart of the multimedia communications environment. It is the generalization of the basic element that we shall consider at part of the transfer of information. In a voice only environment, the corresponding element is the conversation, the concatenation of the words, inflections, pauses that make up the communications from one individual to another. In a computer communications world the corresponding element is the data file and the transactions that relate to that file. As we expand to the multimedia environment we extend beyond that to the more complex interaction of user with information in many forms.

1.1.2 The User

In a multimedia communications system, the user is generally not the large computer or other information processor but is the human. This dramatically reorients the focus on what is to be communicated and how it is to be communicated. The users objective in a multimedia communications environment is to use the information and to develop and understanding on how the many parts of the information can be used—in achieving "the desired" goal.

1.1.3 Interaction

The interaction allowed in a multimedia environment is dramatically different than that in most other communications systems. As we noted before, the standard form of communications is highly structured and is linear in fashion. Computer communications is based upon the need for well-established and agreed to protocols that permit the users to interact in a controlled fashion. The users in the computer world are generally other computers or in some cases humans who must adhere to the computers well structured dialogue. In contrast, the interactions in a multimedia environment are highly unstructured, significantly nonlinear in their form and are

driven by the more creative side of the human user. Images are complex representations of knowledge elements and the ultimate information is a construct of the combination of the images, the user and the interactions that are developed.

1.2 CURRENT SYSTEMS

There are several simple but representative multimedia system that have been developed in the past few years. The first contain work stations or computers that allow for the use of many types of media, integrating graphics, still images and other types of images—into the screen. The introduction of a
window environment allows for the use of many of these media on a standalone work station. The inclusion of a networking capability further enhances the ability of these work stations to share the resources. For example, a file server is a memory storage device that allows for shared memory to be used by many users on a local area network. The server may have any image type stored on the system and one or several users may window the image onto their customized screen format.

Figure: Sample Current Systems

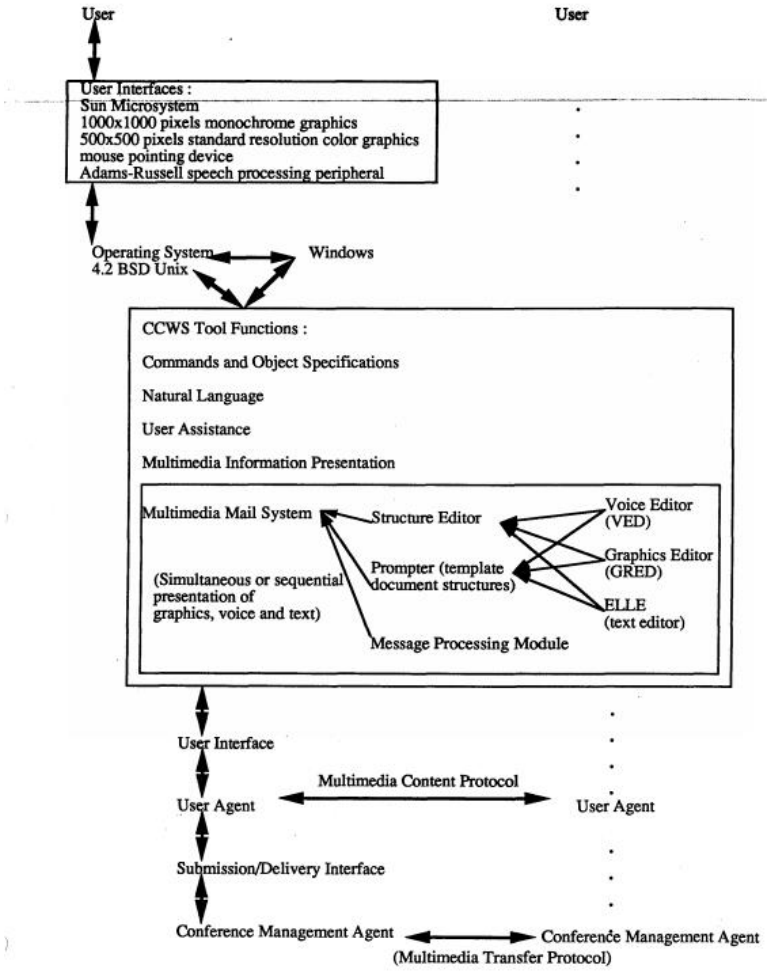


Figure 1.2.a : Sample Current Multimedia Systems - the Command and Control CCWS System.

Figure: Sample Current Multimedia Systems - the DARPA Experimental Multimedia System.

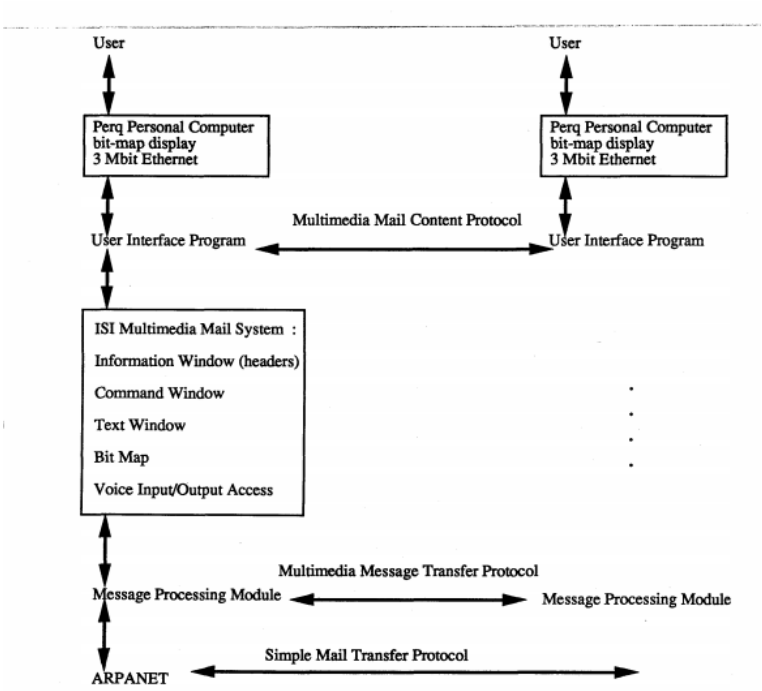


Figure 1.2.b : Sample Current Multimedia Systems - the DARPA Experimental Multimedia System.

1.2.1 Computers

The multimedia work station or computer has been developing over the past few years and a typical design is that of the Apple MAC II system. The MAC II uses a simple window structure with a high resolution bit mapped scree for display. We shall see that this type of work station architecture is typical of many of the first generation multimedia work stations.

Figure: Sample Multimedia Computers

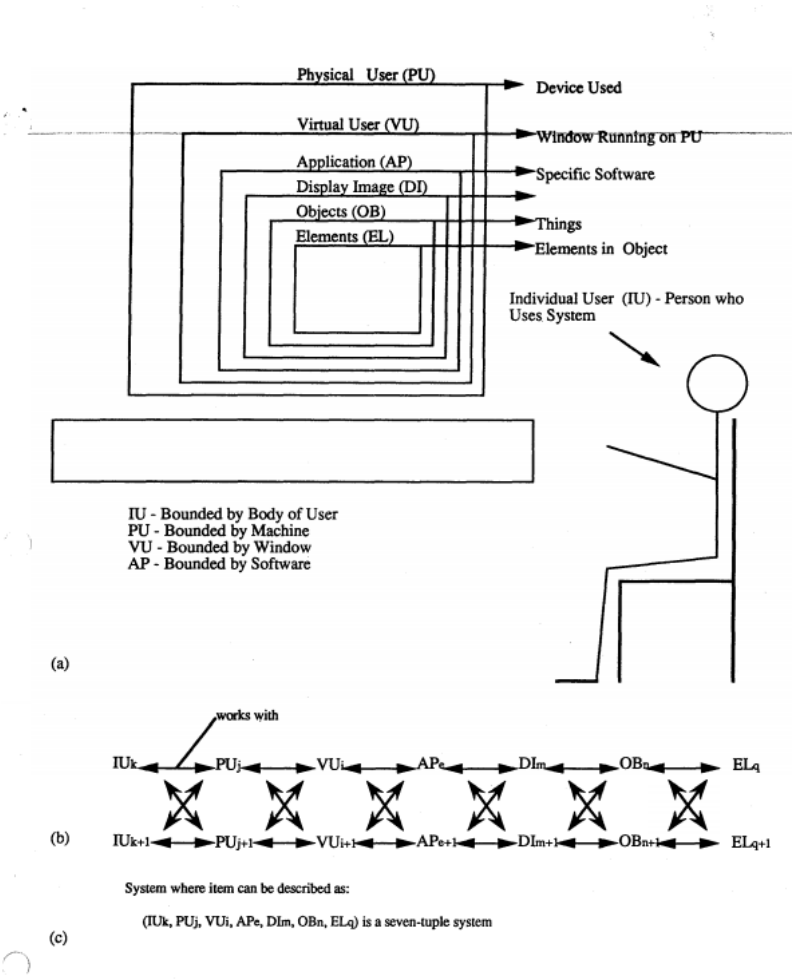
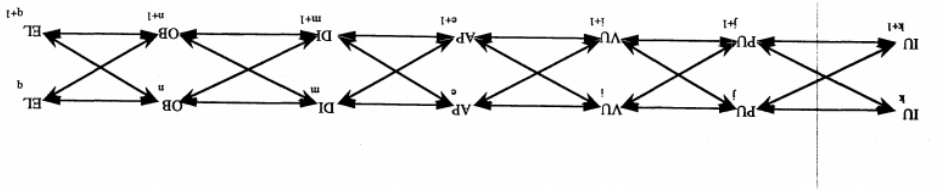


Figure 1.3

Figure: Action Flow



1.2.2 Presentation Formats

The first generation multimedia systems allow for a wide set of presentation formats. The systems will support graphics, text, still images, video of various forms, digitized displays, and integrated voice with the display. Also some systems under development allow for the integration of voice and image as an integrated whole. Some systems allow for the integration and manipulation of various sound elements.

Many of the efforts in the MIT Visual Language Workshop, under Prof. Muriel Cooper, show the capability of combining the visual with sound of many forms.

In a multimedia environment, it is possible to interact with all of the human senses and more importantly to play one off against another. We can exchange sound for intensity and color for the pitch of speech. In the current first generation systems, there is considerable effort using the concept of false colors to bring out new information in visual data that was no there in a simple viewing. A typical example is what can be done using extensive color imaging in the medical area. Currently all X-rays, CAT scans and MRI scans are in gray scale images. A gray scale image quite simply is a system that put at each picture element (pixel) a dot whose intensity can vary from pure black to pure white and has many shades in between. For example the display at each pixel may have 10 bits and thus there could be 1024 different shades of gray. Depending on the human, the eye could resolve anywhere between 256 to- 4096 shades of gray.

Now it is possible to expand the gray scale imaging of human tissue by adding color, wherein the color corresponds to tissue type and its intensity to type density. Thus it is conceivable that an MRI scan could look at the paranasal sinuses and determine not only the different masses and size of tissues but could also identify particular tissue types. The discriminants to do that are available today in the magnetic resonance imaging data.

Figure: Sample Integrated Presentation Formats (Muriel Cooper's)

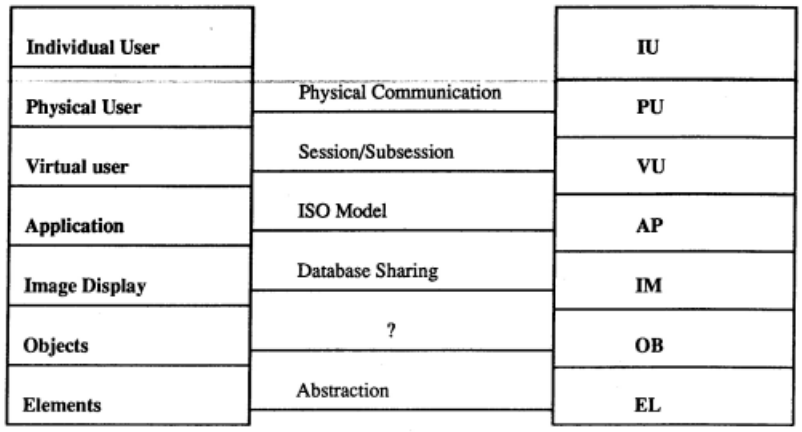


Figure 1.5: The Multimedia Paradigm

1.2.3 Interaction Formats

The end user can interact with the images in many forms. The standard first generation interaction mechanisms include the keyboard, the pen and pallet, the touch screen, and even a voice activated display. Some extreme forms of interaction have included the are/hand point and speak methods developed in the MIT media Lab. The latter types represents some of the first attempts to make the interaction more in line with the activities of the human user. Remembering that multimedia communications attempts to match the end users flexible forms of inputs, the need exists for many types of advanced interaction formats. Attempts have been made to provide for speech response formats that allow for hands off interaction. The problem generally with these types of devices is that they displace the problem of input to a higher human level of making the user coordinate two separate reflexes, sight and sound.

Considerable effort is still required to develop interaction devices and systems to more effectively match the human user. The classic anecdote is the scene in the movie Star Trek, where the chief engineer Scotty, sent back sever hundred years in time, asks for a computer and is shown an Apple MAC. He picks up the mouse and speaks into it and says, "Computer". He is then told that that is a mouse and he must use it to point and the key board to enter the commands. He comments, "How quaint". Yet it is this interaction element that will be the main diver of the McLuhan change of media and knowledge.

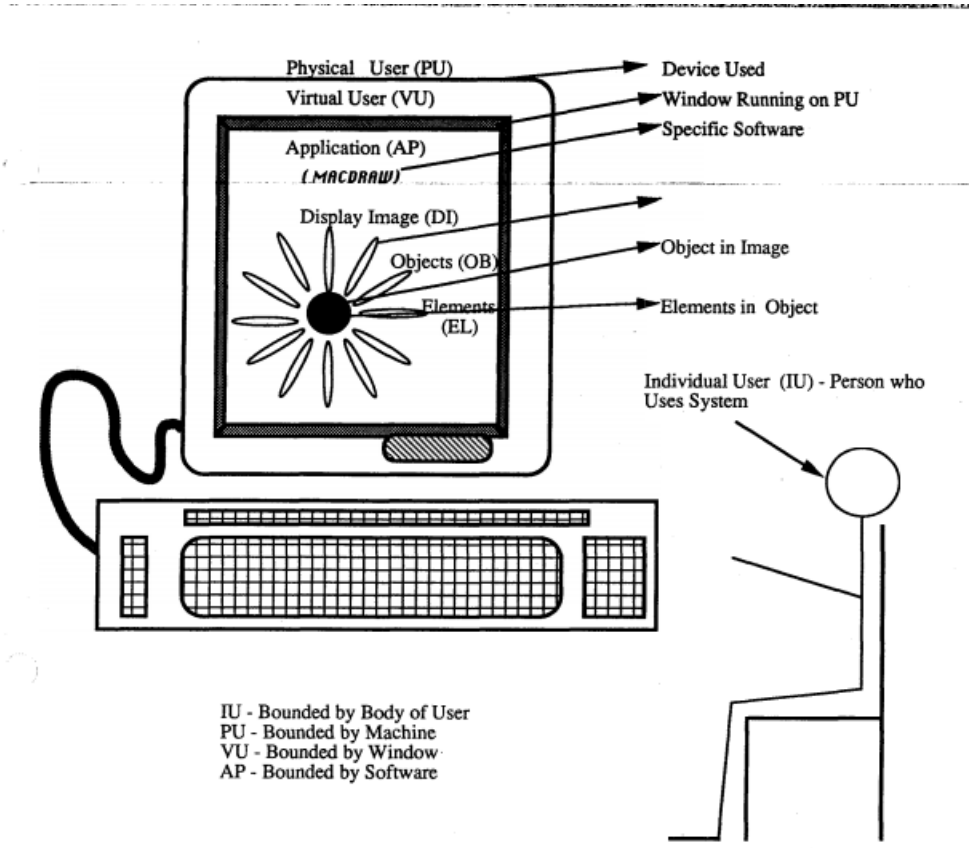
The interaction with the current first generation systems focuses on the interaction of the human user with" the stand alone work station. We can envision extending this by having a network of resources and other users and consider the interaction problems in this multimedia multiuser communications environment. We shall, in the remainder of this book, develop the concepts necessary to show how this more complex system of interaction may evolve.

1.3 THE PARADIGM

The development of a theory for multimedia communications requires the development of a paradigm or world view of the multimedia environment. We have defined the concept of an image and we must now combine together several key concepts that will be sued over again in the book as we develop the multimedia communications environment. All multimedia communications relate to images. Yet those images ultimately relate to the interaction of human users and their mutual manipulation of those images.

We shall now define the key elements of the multimedia paradigm and show how those elements add to a working model for multimedia multiuser communications.

Figure: The Multimedia Paradigm



(a) Structure of Data Management System

System where item can be described as:

(IUk, PUj, VUi, APe, DIm, OBn, ELq) is a seven-tuple system

(c) Specifics of tuple system

Figure 1.6: Multimedia Elementals

1.3.1 Individual User

The human that interacts with information is termed the individual user, IU. In any communications interaction we can envision one or more individual users communicating. The individual user in multimedia communications becomes a key element in the ' communications channel. The individual has an identity that is separate and apart from the other elements that are frequently the key elements in a computer based communications example. Even in the standard world of voice communications there is no recognition of the individual directly. The telephone directory gives only the name of the resident responsible for the payment of the bill. In

multimedia communications, we recognize individuals separately and apart from the other elements in the communications system.

1.3.2 Physical User

The first element in the model is that of the physical user. The physical user is a physical device that is used by an individual for the purpose of communicating with others. Clearly, we can envision that the individual may have access to several physical devices and in turn a single physical device may relate to several individual users.

The physical user, P, may be one of several physical devices that are attached to the network. These devices may be terminals, file servers, imaging devices or there elements. They are unique and are specifically located at a specific location.

Virtual User

The virtual user, V, is a concept that becomes an important element in the multimedia environment. Any physical user may open one or more virtual users on the network. To some degree, the concept of a window is analogous to that of a virtual user. Any physical user may have several virtual users connected to the single physical device.

Application

An application is a specific program that is operative within a physical user environment. For example, in a CAT scan environment, the physical user is the CAT scan system, the virtual user is the specific allocation of resources to a set of scans and the applications may be a combination of both two and three dimensional CAT scan analyses. The virtual user window can open up for the analysis of a specific patient. Thus the identification of a virtual user in this case is with the patient, and within that virtual user is allocated two applications programs that process the CAT scan data and another application program that is used for the maintenance of the hospital records.

Image

The image is the physical embodiment of the multimedia subject to be used in the discussions between users in the session. Tor example the image could be a picture scanned from a video data base, a graphics image generated from a raster or geometric data format, or even a voice message that is interactively generated between the two users.

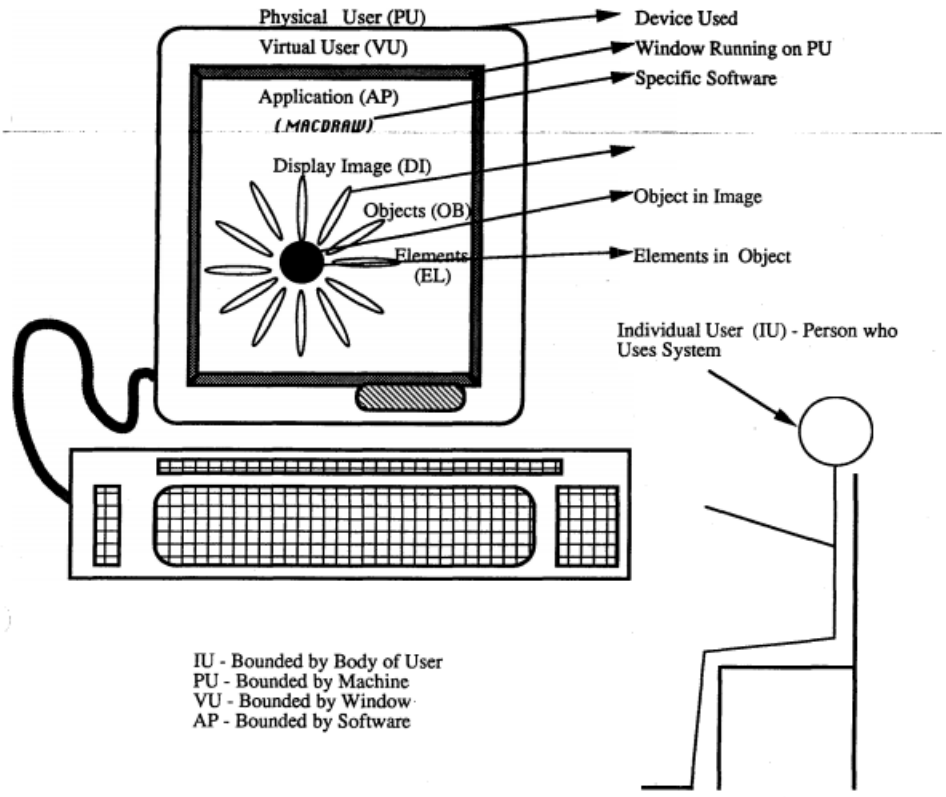
Object

The object is the first step towards abstraction of the image. Whereas the image is the totality of the multimedia message, the object is a bounded portion of the image that can be user definable. Thus if the image is considered to be the full landscape of a photograph, the object may be the part of the landscape that is near a stream or mountain. Objects present localization of images. A further example is that in the general sense the image may be a total conversation, whereas the object may be a portion of that conversation.

Element

The element is the part of an object that can be abstracted to a specific definable abstraction. For example in an image of an X-Ray, there can be a specific object called the spine, and a specific element called the fifth cervical vertebra. The ability to now abstract the fifth cervical vertebra and use this in comparison in MRI, CAT and radiographic studies is the key issue associated with elements.

Figure : Multimedia Elementals



(a) Structure of Data Management System

System where item can be described as:

(IUk, PUj, VUi, APe, DIm, OBn, ELq) is a seven-tuple system

(c) Specifics of tuple system

Figure 1.6: Multimedia Elementals

Figure: Multimedia Elementals

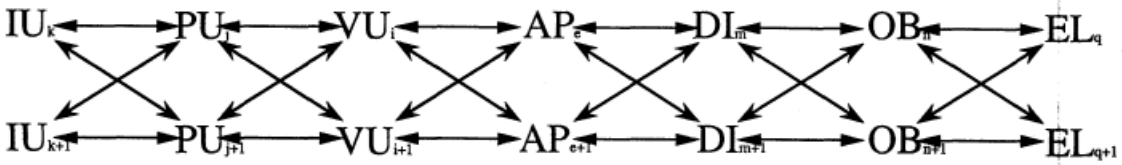


Figure 1.6b: Communication Between Units of System

1.3.3 SubSession

A sub-session is a connection between two virtual users. It is the fundamental building block for multimedia multi user communications. What is important

1.3.4 Session

The session is the key concept in the multimedia communications environment. A session is quite simple a collection of one or more sessions. In a session, a set of virtual users, and their associated individuals and physical users, are linked together into a common communications process. In this book we shall be developing all of the structure necessary to develop the session and to support the session as it evolve in time. A session may last an indefinite period of time and it can survive the coming and going of any of its constituent sub-sessions.

All interaction occurs within the context of a session. In addition the subsessions allow for the adding of any set of virtual users, each of whom may have their own objects, elements or even applications.

Figure: Development of Sessions and Subsessions

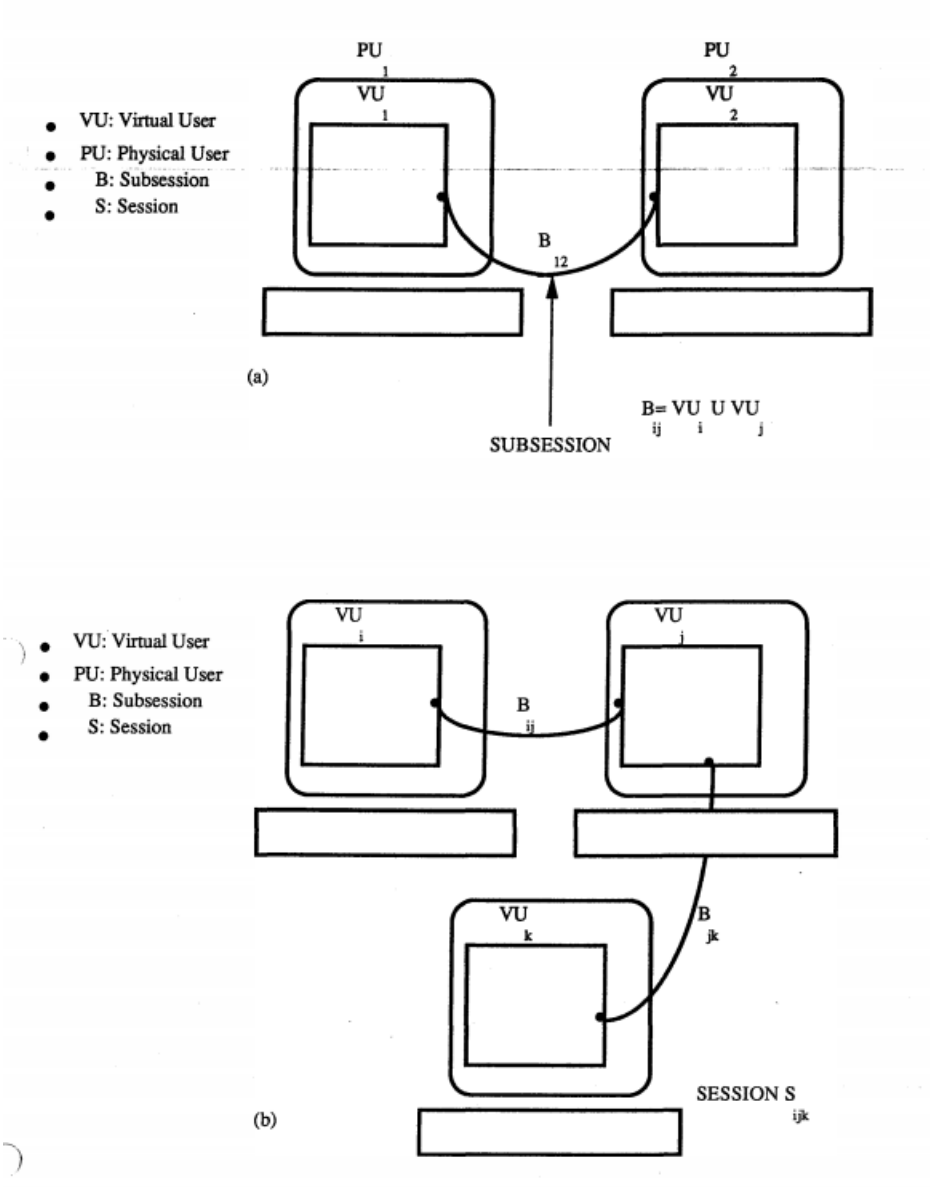


Figure 1.4 Development of Subsession and Session

1.4 ELEMENTS OF MULTIMEDIA SYSTEMS

We have discussed the structure of the current first generation multimedia systems and have discussed the way in which they process data on the work station, display the images and integrate them and provide for interaction with the user. The first generation systems are generally limited in what they can do in these three areas. In this section we shall develop the basic elements of the multimedia communications systems and attempt to anticipate what is possible in the second generation of such systems.

Figure: Multimedia Layering and Interfaces

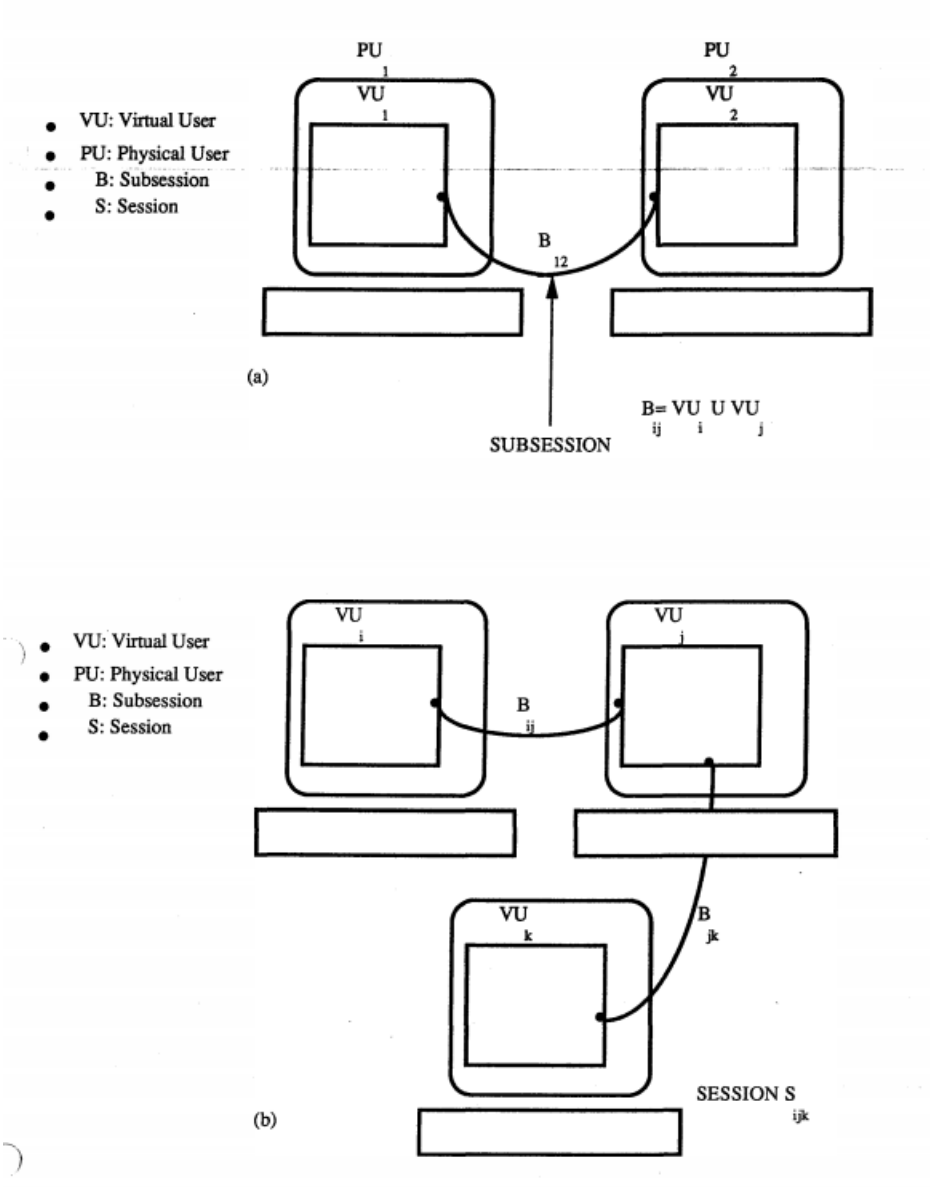


Figure 1.4 Development of Subsession and Session

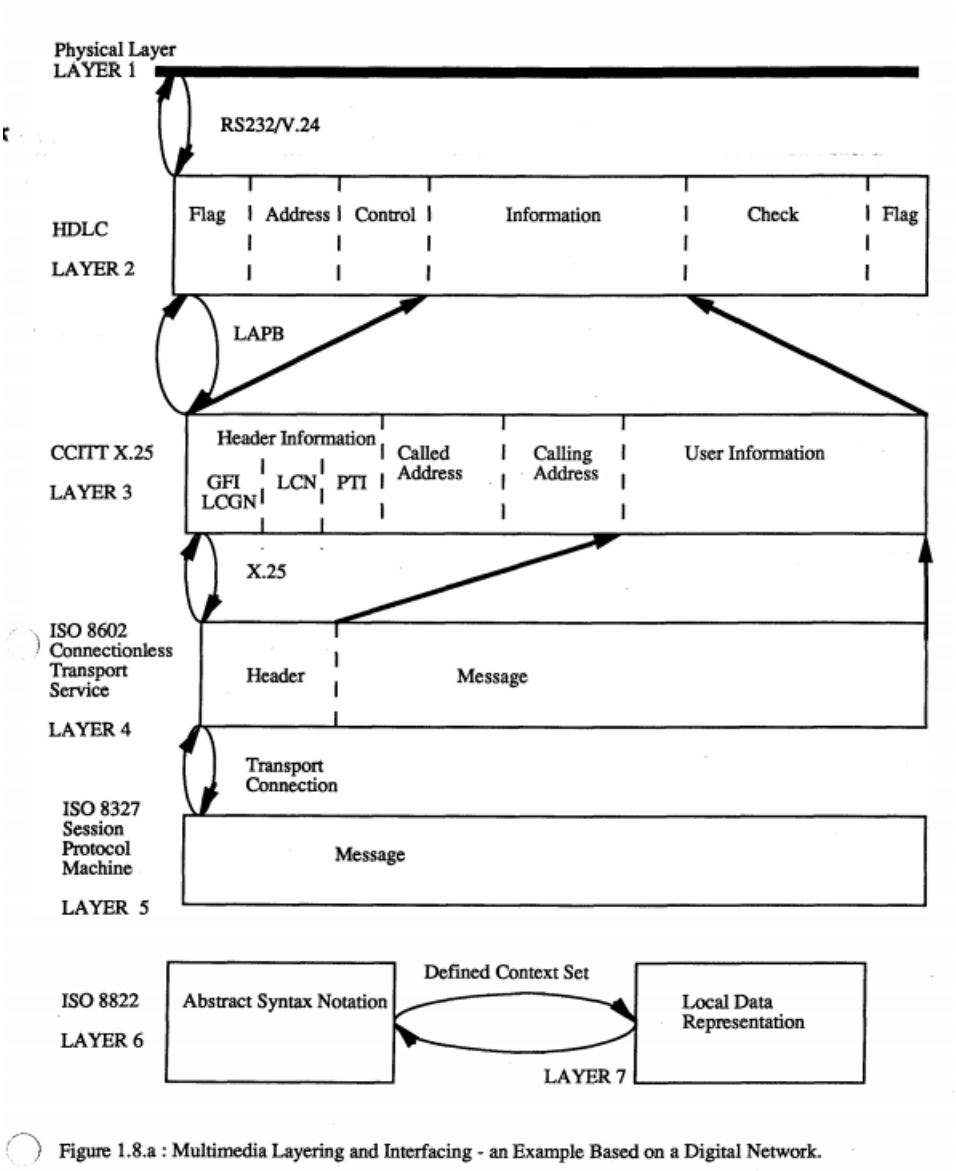


Figure 1.8.a : Multimedia Layering and Interfacing - an Example Based on a Digital Network.

Figure: Multimedia Layering and Interfaces - In*terface between two ISO Layers.*

Link	Network	Transport	Session		Link
Protocol Control Information	Protocol Control Information	Protocol Control Information	Protocol Control Information	Protocol Control Information and Data in Transfer Syntax	Check Sequence

Figure: Multimedia Layering and Interfaces - Layering in the Data Packets.

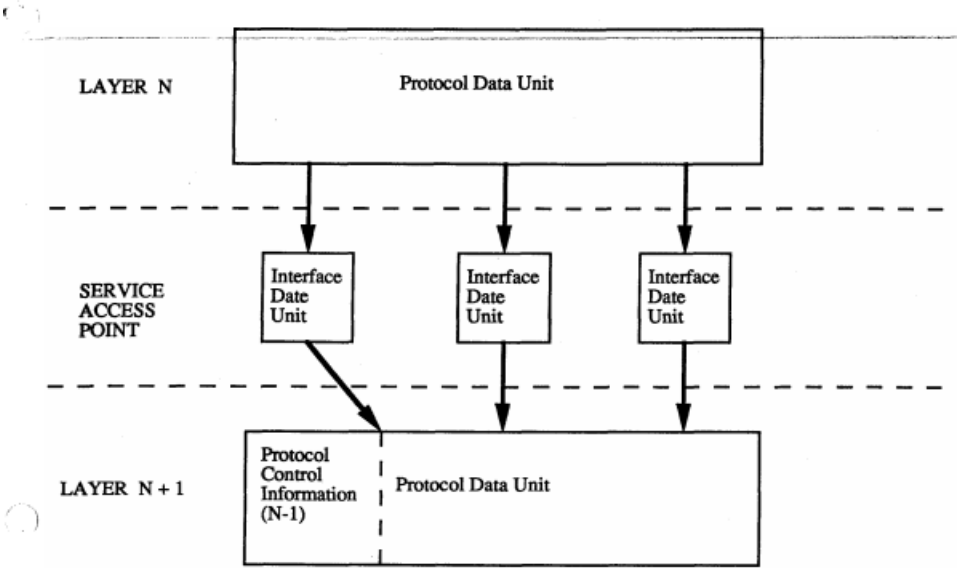


Figure 1.8.b : Multimedia Layering and Interfaces - Interface between two ISO Layers.

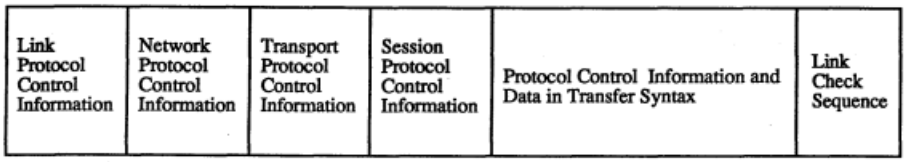


Figure 1.8.c : Multimedia Layering and Interfaces - Layering in the Data Packets.

1.4.1 Display

The display for the first generation systems has generally been a bit mapped display whose size has been limited to about 1000 by 1000 dots. This yields about a million pixels in a display. The pixel may be up to 12 bits deep, so that a single display has 12 million bits. In current production as the first steps towards a second generation display are 2000 by 2000 units of 4 million pixels, each of 24 bits per pixel. This yields 100 million bits per screen. This approaches the display capability of a 35 mm slide.

These second generation displays now must have significant buffer storage to keep the image and must have the bandwidth at the display driver to keep the image refreshed at the rate on 30 times a second. This implies a transfer rate of 3 Gbps.

Figure: Display Architecture - A Simplified View of the Vector General 3400 Display Architecture.

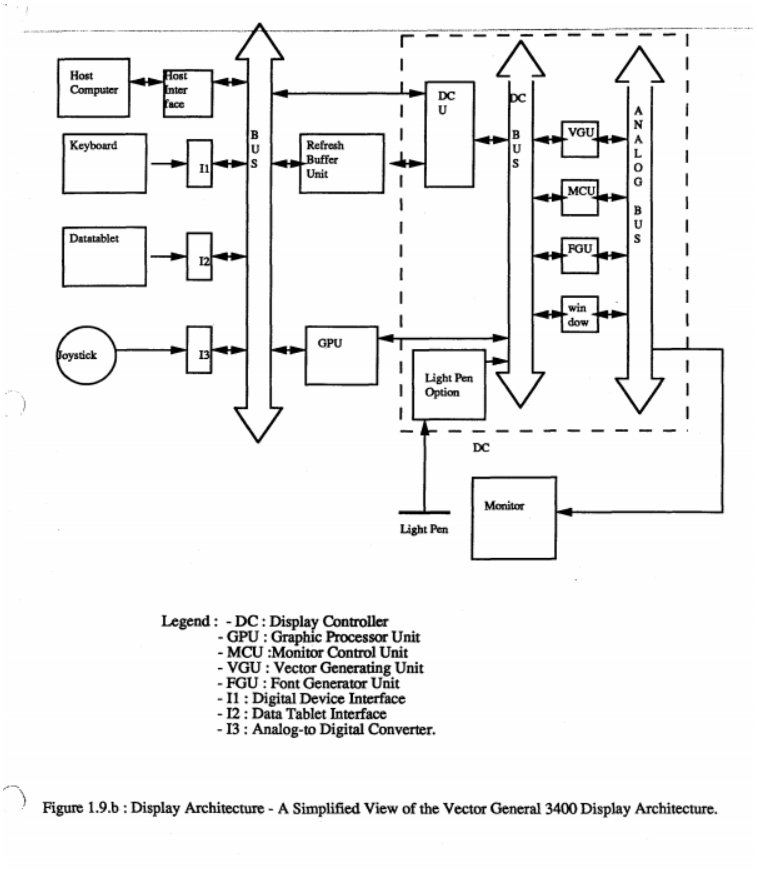


Figure 1.9.b : Display Architecture - A Simplified View of the Vector General 3400 Display Architecture.

1.4.2 Applications Software

The applications software must handle the processing necessary for the display processing as well as for the support of the end user interaction. It is a layered approach to the development of the application software, the lowest layer interfacing with the operating system of the multimedia environment and at the highest layer, and end user interaction capability. The layers include the following:

User Interface Layer: This outer layer allows for support and integration of the end user into the system.

Applications Support Layer: This layer provides for the interface to the User layer and utilization of the local work station for support of local processing.

Data Management Layer: This layer provides for the support of the local data files and management of the images into the display. Such capability now in the X Window format can be found resident in this layer.

1. Operating System
2. Control
3. Supervisor Programs
4. I/O Programs (IGES, COM)
5. Communications Programs
6. Processing
7. Utility Programs
8. Language Translator
9. Information Base
10. Management
11. Application
12. Pagemaker, Videoworks, MacDraw n, Quark Express, (Core, GKS or PHIGS standards for graphics) Legend: - GKS : Graphical Kernel System PHIGS : Programmer's Hierarchical Interactive Graphics Standards.

Figure: Software Architecture - Software Architecture for a Multimedia Messaging System.

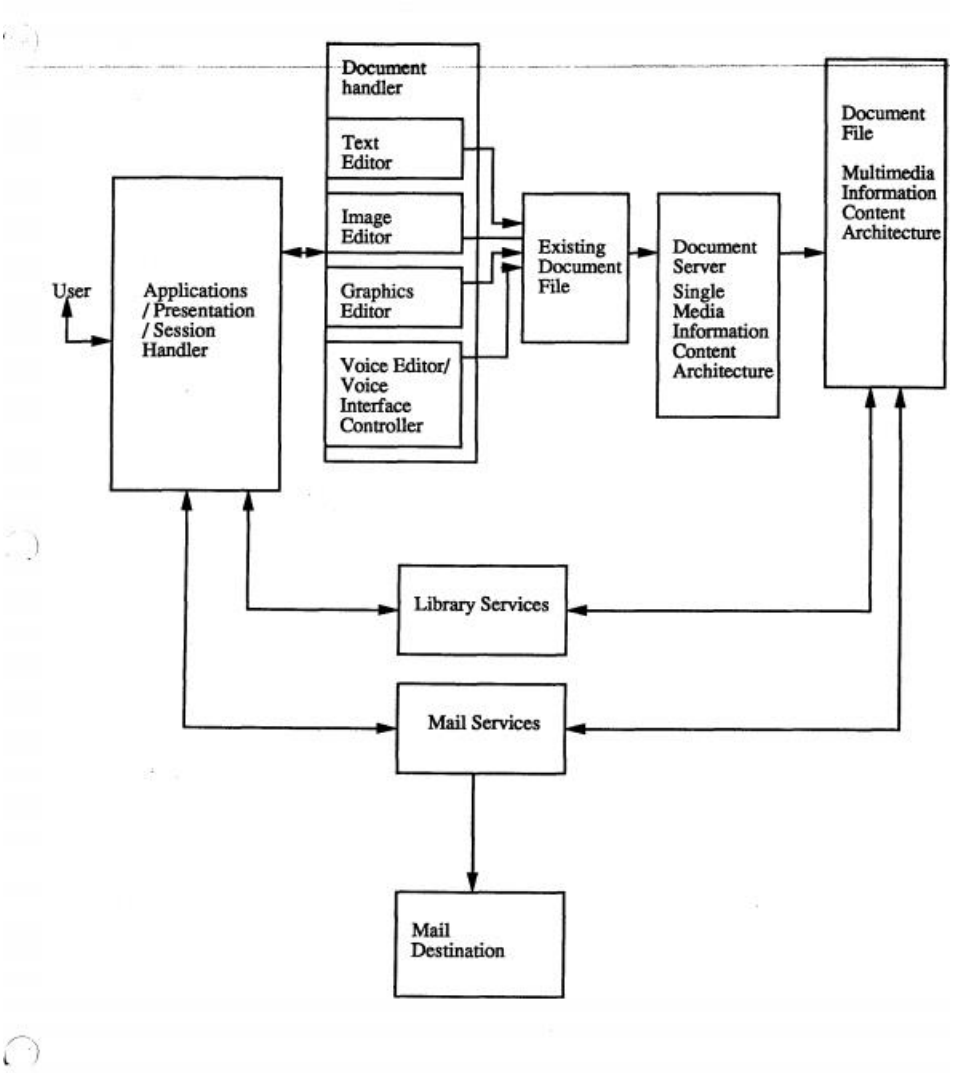


Figure 1.10.b : Software Architecture - Software Architecture for a Multimedia Messaging System.

1.4.3 1.4.3 Services

The services portion of the canonical design allow for the support of the multiuser environment and are the key to enabling the establishment of the session. Typical services may include:

Session Establishment: This service allows for the generation and support of a session between any of the users. A single user may establish a session between themselves and any other virtual user available on the network.

Mail

File

Directory

The services can be implemented in a layered architecture just as we had done with the applications layer. It consists of two layers, the shell and the kernel. They shall allow for the interaction between the individual terminals and the services and the kernel supports the interaction between the individual terminals and the network. The key to successful multimedia services is the ability to work in a fully distributed environment. We shall develop the distributed environment in latter chapters and shall also present a more detailed structure to these two layers of the services.

1.4.4 1.4.4 Network Software

The network that interconnects the end users must do so in support of all of the layered areas discussed above. The network must support the applications areas and furthermore support the services layers. The major difference between multimedia communications and computer communications is that often the computer needs to send the messages in what is called a connectionless format. Packet or datagram are sent as little letters with all the information to get from point A to B. In the multimedia environment, the need for the use of sessions requires that the sessions be supported by a connection based network service. In addition the network must be capable of providing large amount of bandwidth on demand to the user. Consider the case of the second generation high resolution display with 100 M bits of data. To transmit one image requires a 100 Mbps channel from any point to any one of a set of other point. The response time of less than one second is critical to the success of any multimedia system.

This architecture can be either distributed or centralized in nature and the network software must be intelligent enough to support the session based connection path. We shall be developing this concept in further detail in the remainder of the book.

1.4.5 1.4.5 Storage

Multimedia communications requires large amounts of complex image information for use by many users all at the same time. No matter how the cost of memory decreases, the key bit of information will not be sent everywhere at the time it is generated. Thus the network must provide access to memory that is distributed and that also can be accessed in extremely short periods of time. Consider again the example of the high resolution display that requires 100 M bits of image. The transport speed may be 100Mbps to lGbps but the bottleneck may be in the finding of the imaging, accessing it and loading it onto the network.

1.5 KEY PROBLEMS

One of the world's greatest philosophers, Ludwig Wittgenstein, once stated that the essence of true knowledge of any field was the ability to pose the correct set of questions, whose answers are simply stated. Thus with multimedia communications, the development of a new body of understanding is best based upon the positing of the proper set of questions and these questions are in turn based upon a set of key problems that we face in attempting to communicate in this new environment. The following problems and the ensuing questions represent the body of the text.

1.5.1 Problem 1: Characterization

As we have developed in the early part of this chapter, the essence of multimedia communications depends on the concept of the image. The image may vary in form from a high resolution picture to a speech conversation. The first problem is the characterization of the physical image into an electrical form. For example, a speech conversation that varies over a finite period of time may be digitized in a standard form by sampling at the rate of once every 125 msec and using eight bits per sample. This will yield the standard 64 Kbps voice sample of speech.

As a second example of the characterizations problem is that applied to high resolution pictures. Again we can consider sampling the image in two dimensions and in storing the image in this highly sample digital form. The problem is again how well to - sample to retain-the integrity of the basic imager

Another element of the characterization problem relates to how to characterize an image so that it may be manipulated and not just stored. Consider the following example.

Example: In the medical field, there is significant use of both CAT (Computer Aided Tomography) and MRI (Magnetic Resonance Imaging) imaging. CAT scans use X rays to resolve the different parts of the body tissues based upon the density of the tissue and its ability to absorb the X radiation. By passing many different X rays through the body the image can be reconstructed to reproduce the innermost portions of the body. In MRI, the technique uses the ability of different tissues to resonate when under strong magnetic fields, the resonant frequencies dependent upon the tissue type. In the MRI case a multi-dimensional Fourier transform creates images of the human body. Now the physician desires to compare the results of the two scans to determine the status of the cerebellum, the posterior portion of the brain. The characterization problem in this case relates to how do we define the cerebellum as an abstraction in both imaging schemes. We wish to do so not by just comparing bit mapped displays of the same tissue but to do so by a full element abstraction.

1.5.2 Problem 2: Utilization

The utilization problem takes the image characterization and addresses the issue of the human interface. How does the human specify the specific image, object or element, and in turn how does the human user interact with the system to allow for the manipulation of the images. The concepts of the touch screen and the mouse are but simple starts to the development of agents in this area. The general concept of an agent, that is an entity that can abstract both and element and an action are key to understanding the issue of utilization.

Transmission of the information relates to the problem of taking the users view of the world and allow for it to be effective in several places at the same time. In contrast to sharing, which we shall describe in the next section, transmission relates only to the transport for the images in a session context. The present means of transmission are generally quite limited. They use existing telecommunications services that in most cases are not designed for high resolution image transport or for the use of extensive session sharing. The telecommunications transmission channels allow for rates up to 45 Mbps which limit the use to small image sizes or lower refresh rates.

For example, consider a 2,000 by 2,000 pixel array using 24 bits per pixel. This amounts to about 100M bits per image. If such an image were to be used in a full motion video context at 30 times a sec refresh rate, then we would need 3 Gbps transmit capacity. That is almost 100 times greater than the current offered tariffs for transmission services. Thus the transmission problem focuses on two issues, the first is developing higher data rate transmission channels and the second is the developments of compression techniques for the reduction of the data required to transport the true information content.

The true transmission dilemma that faces many users is to wait for the greater data rates or to incorporate the use of often high cost compression technology. The date rates that are now being implemented are moving to the 500Mbps range for local access and to 1.5 6bps for long distance transports—The -problems still exist that one cannot share the data links and thus one is compelled to use the data channels as if they were fully loaded.

1.6 PROCESSOR/ PACKETIZER

The sharing problem relates to how best to take multiple users in different modalities of use and to allow them to share the communications and end users services in a user friendly fashion. The essence of sharing is essentially the development of the session concept that we have developed. As we have stated, the session concept was developed to support the virtual user layer of our multimedia architecture and allows for the interface and support at that level. The applications layer interface will be shown to have the session capability already present in what is called the OSI Layer 5 Session protocols.

The complexity of the sharing problem or that of VU Sessioning relates to the need for connecting diverse end user devices together with diverse end user interfaces. If there were standards that existed for doing so the problem would be generally easy. Such standards exist for the applications to applications interface but fail to exist for the VU to VU interfaces. We spend a great deal of time in this book focusing on the issue of sharing.

Example: Consider the medical application again wherein the user has to work in a complex environment of CAT, MRI and Nuclear medicine scans.

The applications to multi user systems is characterized in the sharing problem but the multimedia multi user system problem is best characterized by the mixing problem. The mixing problem is the one that occurs when different forms of images and media are to be shared and manipulated by multiple users at the same time. These users may have different types of equipment, have their images on different types of databases and different interfaces.

Figure: Mixing Concept in Printing Environment

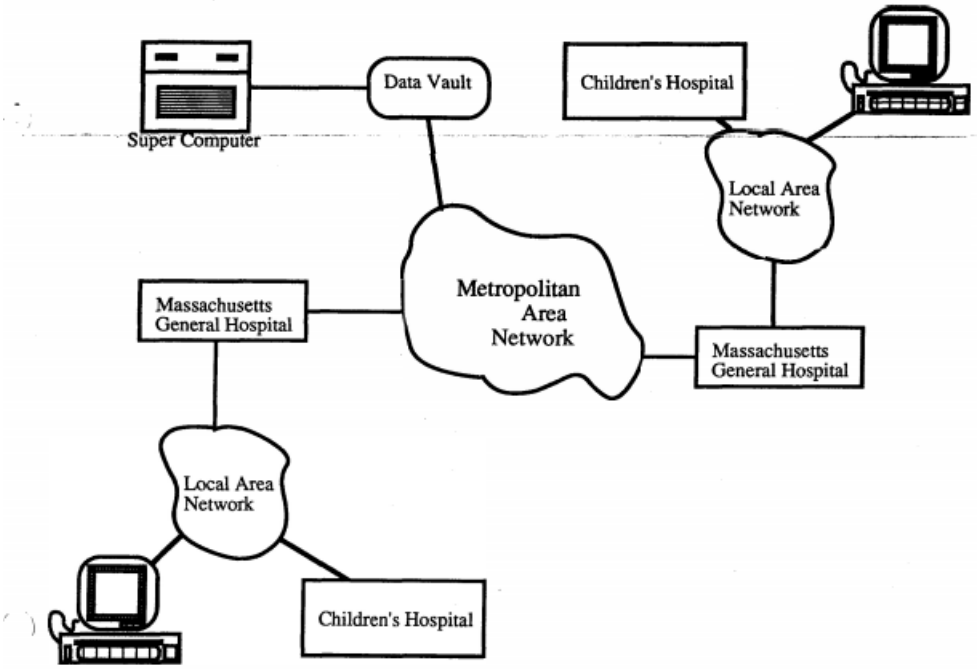


Figure 1.17:Sharing Concept in Medical Enviornment

Figure : Mixing Concept in Printing Environment -Mixing in Multimedia Documents Using an Area Control Mixing Type.

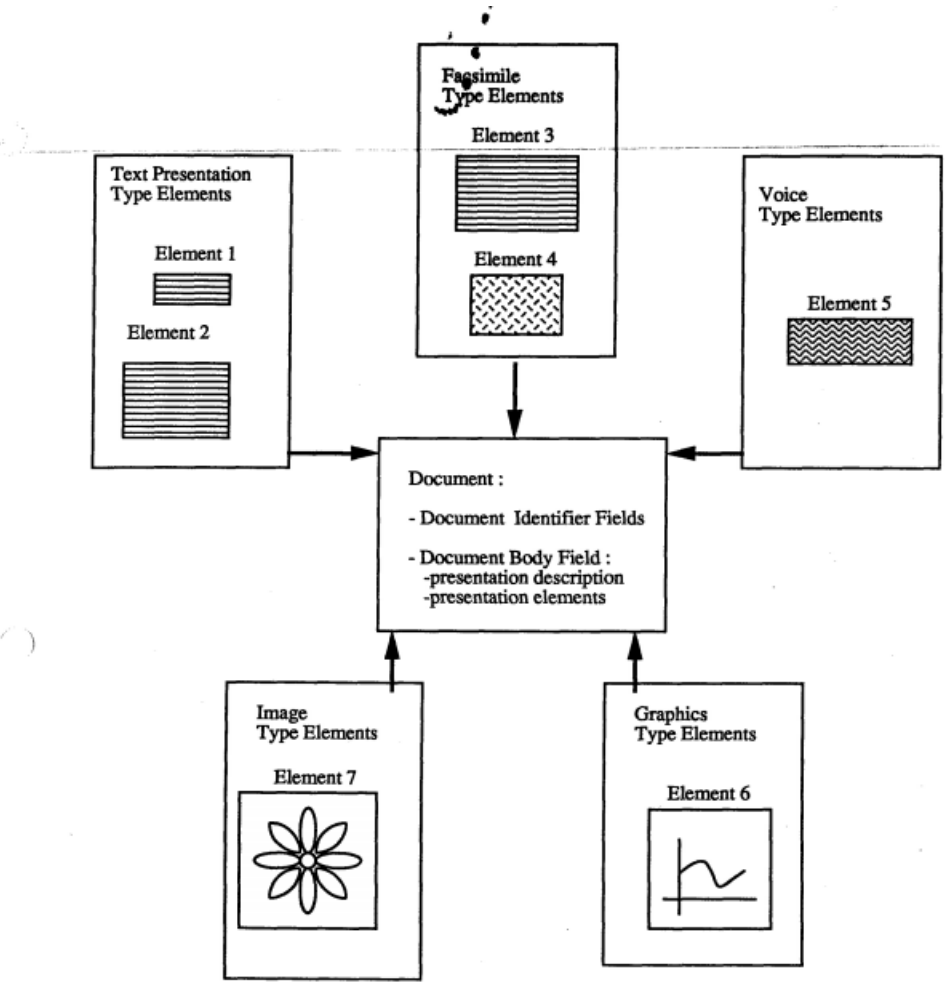


Figure 1.18.c: Mixing Concept in Printing Environment - Mixing in Multimedia Documents Using the DARPA Multimedia Mail Content Protocol.

1.7 OUTLINE OF BOOK

The book follows the discussion on the key problems that frame the multimedia multi user communications area. In this book we spend a great deal of time discussing the characteristics of many forms of multimedia images. Specifically we discuss the way in which users of still images, graphics and video, for example characterize their media. This is important since when we take these existing media that are in a physical and non-electronic form and turn them into electronic, we must do so non-electronic as to preserve their original frame of reference. For example, if we develop a system to display radiological studies from file to a digitized form, then we must preserve the same image density, intensity and responsivity as does the film radiograph.

discusses the many types of multimedia interfaces. These interfaces range from the ways in which we transform the physical media of the image into an electronic image to the ways in which we characterize images in abstraction forms. As we have developed in this chapter, abstraction of images is one of our major concerns. The approach to abstraction may be structural or connectionist. Our major focus in this text is that of a structural approach, leaving the connectionist to those other works that cover them more fully.

The issue of storage of multimedia information is contained in . In this case we are concentrating on multimedia databases that are significantly more complex than those found in normal data processing environments. The multimedia database problem revolves around the issue of the scale of the image and its need for integrity and simultaneity. Another element of the multimedia communications that this chapter considers is that of performance and sizing. These latter two items are of critical importance in designing and developing any new multimedia system.

addresses the issues of communications. We focus on the standard approach to communications by working with the OSI layers. Our approach is different than most others by starting with the higher layers, viz. applications and presentation and session, and working downward. The issue here is that instead of looking at the physical communications link as transport and then finding what is missing, we focus on the session and the essence of the multimedia environment and works downwards. This approach provides new insight into the communications link. We also develop an understanding of the systems and issues associated with broadband communications systems.

The essence of any multimedia system is the development of an overall architecture. The first few chapters focus on the major problems that we have developed earlier in this chapter. develops the overall architecture for such systems. In particular we focus on developing a methodology for architecture development, including the issues we have developed on characterization, processing, storage and communications.

The book develops the theory of multi user systems, showing how to extend the elements of the multimedia environment to that of the many user system environment. We spend a great deal of time discussing the performance and sizing efforts of such system designs. It develops in details many of the issues associated with the implementation of distributed systems. In the multi user multimedia environment, we are dealing with a fully distributed environment. In this environment we must focus on distributed databases and distributed operating system. As we have discussed the multimedia database we must now focus on how it can be expanded to a distributed environment. The issues of distributed operating systems allows for a better understanding on how multi users may now deal with the issues of sessions in a shared

synchronous environment. One of the major issues is that of understanding the overall synergies of a shared network.

Finally it presents the overall conclusions and the directions some of the multimedia system are taking in today's environment. In particular, we discuss many of the issues that will become research issues in the years to come.

2 MULTIMEDIA COMMUNICATIONS

From time to time, a change occurs in the way people communicate, to cause a shift in thought, perception, awareness, and in a sense of what is truth and knowledge. It has been argued that multimedia and multimedia communications is such a shift. This paper addresses that perception from a perspective that is dramatically different than any other taken. The perspective developed in this paper is one that is philosophical. It does not talk to technology or business. It does not refer to companies and equations. The approach is that taken by an engineer who was first trained as a philosopher. An approach that tries to address first principles from the context of the perceptions and understandings of the many other thinkers who have addressed the issues of human knowledge and communications before. We consider the process by starting with Plato and end with the most current thinking in the areas of epistemology and human understanding. This is an evolving approach. The schema presented and developed herein are still formative. They do, however, set a direction and a discipline.

2.1 CURRENT DEFINITIONS OF MULTIMEDIA

Multimedia concepts encompass a wide variety of interpretations. To some, the multimedia environment consists of one that has many elements of storage, to others the environment entails the interfacing of various storage devices. In another extreme there is the concept of a multimedia environment being nothing more than that of a windowed display environment that uses some limited form of hyper media, that is windows in depth besides just length and width.

We attempt to develop a theory of multimedia communications, that encompasses the elements of both the multimedia environment as well as the effective utilization of the rich nature of that environment. Let us begin by understanding the nature first of the human using information in the context of their own limited environment and then expand it to the case where the human can share the information with others.

There have been a significant number of definitions of multimedia. Comparing some of these definitions will be a quick way to determine what multimedia is not. All too often the authors are reporters who have not struggled with the issues of multimedia or at the other extreme they are vendors who define multimedia to be what is in their current product line. The first definition is due to Wright, who describes multimedia in the context of the Global Village. To her, multimedia is;

" Computer based presentations combining two or more media, such as text, graphics, writing, and video and audio signals."

If we look at her definition we note that she defines media in the context of the type of message (text, graphics, video, audio) as well as the way it may be stored (video and audio signals). She does not take into account the complexity of storage, the sophistication of interfaces, the interaction of the senses, nor the intelligence of the human processor. The mere fact that multimedia is "Human" directed as compared to "Computer" directed is not an element in her definition. As a point of fact, Wright goes as far as to define the Human Being as:

"An analog processing and storage device with the bandwidth of 50 bits per second. Human beings excel at pattern recognition but are notoriously slow at sequential calculations."

One is amazed that the author fails to recognize that the essence of multimedia is to more effectively match the input/output capability of the human as a species. We humans are amazingly fast at processing visual information, less fast at aural information and slower at the use of our other senses. Ms. Wright in her Scientific American presentation totally misses the essence of this new paradigm.

Probably one of the more insightful presentations of the Human as processor is the book by Max Delbruck. Delbruck, the Nobel prize winner in Biology, presents the argument that the evolution of primates, and humans especially is epitomized in the evolution of vision as the primary sense. Three specific developments allowed for this speciation to create the mind of the human. The events are;

1. A transition from an olfactory and tactile creature to a visual creature.

2. The displacement of the eyes from lateral to frontal allowing for the perception of three dimensions.

3. The differentiation of the cone receptors in the retina allowing for color perception.

Starting with these three events, Delbruck demonstrates that the human mind is intimately connected to these events, and as such is a complex image processor. This is the basis of the multimedia paradigm. This is what Wright has totally missed in her exposition.

A second presentation of multimedia is done by Jane Morrill. Morrill defines multimedia in the following even more interesting form.

"Flip charts? Overheads? Isn't this the computer era? Surely, with all the high-speed machines, CD-ROMs, synthesizers, and image processing capabilities available, there must be something

that will convey your message better than flip charts and overheads. Well, there it is. It's called multimedia, and it marries the best of image, voice, text, and video processing."

Her attempt at a definition is a tautology, in that you know what it is when you see it. Her article then proceeds to develop multimedia in the context of what vendors of workstations have in terms of different storage devices. She again does not even acknowledge the issue that the human is a factor and that multimedia is a blending of the human in the stream of the information flow.

A third paper by Shandle defines multimedia in altogether different terms. Specifically he calls it:

"...-the delivery of video, audio and other heretofore exotic data types to the desk top -...."

Again there is no mention of the human, why the data is to be transported or even what it is to be used for.

2.2 PHILOSOPHICAL PILLARS

To fully understand multimedia it is necessary to explore the work of two sets of major thinkers. The first is Marshall McLuhan and the second is Winograd and Flores. McLuhan, in Understanding Media, and Winograd and Flores in Understanding Computers and Cognition have complemented each other in a way in which their convergence of ideas lays the ground work for "Understanding Multimedia Communications". We shall be relying on these two sets of authors for a guide through the development of the meaning of multimedia. McLuhan has been discredited of late because of his simplistic views. We shall argue and shall attempt to show, that this may be a direct result of the critics, frequently the Pop Press, not understanding how perceptive McLuhan was in his more academic treatises. McLuhan will be the definer the "bright line" that results when a paradigm shift occurs in a new medium.

Winograd and Flores are the other set of lights that the author shall rely heavily upon. Unlike the Pop writers of the above definitions, Winograd and Flores have developed one of the most seminal works in the areas of computers and computation that have ever been done. These authors have used access to the most recent philosophical understandings of knowledge and knowledge processing from a philosophical perspective to develop a philosophy for computing, especially software development. We shall, in this paper, develop and extend these concepts for the multimedia area.

Drucker, in his biographical sketches of his contemporaries, remarks on his first encounter with McLuhan. It was during a presentation that McLuhan was making on the results of his do doctoral studies. His presentation reflected upon the impact that the printing press has had upon

the university system in the late Middle Ages. He contended that the modern university came into being in the sixteenth century because of printing, which changed not only the method of instruction but, more importantly, what the university intended to teach. He further contended that the cultural results of this period had little to do with the Renaissance and was all a direct result of the printing press.

To quote Drucker, who paraphrased McLuhan;

"Did I hear you right," asked one of the professors in the audience, "that you think that printing influenced the courses that the university taught and the role of university all together." "No sir," said McLuhan, "it did not influence; printing determined both, indeed printing determined what henceforth was going to be considered knowledge."

Thus this led to McLuhan's famous phrase that the medium is the message. Specifically, as we developed a new medium for human communications, we dramatically altered the nature of the information that was transferred and the way in which the human perceived what was "truth" and what was not. The television generation of the 1960's was an clear example of the impact of television versus film in portraying the war in Vietnam as compared to the Second World War. The perception of these two events was determined by the difference of the two media that displayed them to the pubic masses. Television allowed for a portrayal that molded more closely to the individual humans impact of the events as compared to films overview of the groups involvement's. Both media deal with the same senses but they are different enough to have determined two different outcomes of the wars. This conclusion is a McLuhanesque conclusion but is consistent with the changes that McLuhan was recounting in the 1960's in his publications.

The important observation that McLuhan makes is not often understood. He really means that the medium defines what is knowledge. A new medium, as a general construct, will define a new knowledge base. We all too often define knowledge so obtained with truth. In fact, truth is that relative reality that we find comfortable to our understanding, and all too often ascribe an absolute character to it. The essence of this paper will deal with these two issues; knowledge as defined in the McLuhanesque sense, and truth as a phenomenological expression of that knowledge. Multimedia communications will alter those definitions and will dramatically change the way we see, think, and ultimately act. We argue, for example, as with McLuhan, that television violence, for example, changes what is knowledge, the acceptance of moral norms, and this change in moral knowledge is reflected in the truths of everyday existence. The expansion of multimedia communications will take this minor concern many levels higher. Thus multimedia communications is a technological issue, a philosophical consideration, and ultimately a moral imperative.

1.3 Multimedia Structures

Multimedia consists of three major dimensions; the storage media, the interface media and the senses. A true multimedia environment provides a full mix of all of these. Let us begin by first considering the senses. The five senses are; sight, sound, touch, taste, and smell. As humans, we specialize in the first three senses, and only secondarily allow for use of the last two. Max Delbruck, the noted biologist and Nobel prize winner has discussed the uniqueness of the human as one whose dealings with his environment is more visual than olfactory than any other species.

Most animals have highly developed senses of smell and frequently their total social patterns revolve around them. We as humans revolve around sight first and sound second. Point of fact, as cultures evolve, we find that the use of the other senses are less and less socially acceptable. We no longer sniff at each other, nor do we employ the other two senses as other parts of Animalia do.

The senses element of multimedia provides a basis for the understanding of the limitations and emphasis of the human as a processor of information. The information is contained in a combination of sensory inputs. These inputs are all parts of the communications environment. A discussion of an ad layout for a print publication involves the image, the voice interaction as well as the human interaction of peripheral nature that ensues during the presentation of the ad copy. To develop multimedia effectively, it is necessary to evoke all elements of the human response.

The interface element of multimedia communications describes how the human interfaces with the electronic interlocutor, the electronic medium. Again we all too frequently view the interface in simple terms as a computer display and nothing else. The current view of multimedia communications is that of a display terminal that allows for both a windowing environment and a hyper media environment. It may also extend to including video. As we have just described, the essence of multimedia is to maximize the input from the senses, and also balance the communication of all sensory data on the topic to be discussed.

Thus the interfaces must communicate all five senses both from the electronic medium and to the medium. Visual displays are but a part of that communications. Effective multimedia today must also integrate voice and sound with some form of tactile interface. A now primitive example of this is the first set of electronic games wherein the human had visual, aural and tactile interfaces with the electronic medium.

The final element of a multimedia environment is the storage media. The storage media in some sense dictates what is effectively communicated and the ability of the human to retrieve and process the information. The media today has allowed for the storage and retrieval of information in short times for large amounts.

Here the author defines Multimedia Communications in a broad context. "The term multimedia is used in many different contexts and is greatly misunderstood. In the current popular press it is viewed as merely a set of mechanisms for the storage of different data types on a local basis and their display to a single user. In the extended view of multimedia communications it is understood as an environment for the sharing of information in various forms, video, image, voice, text, records, etc., that are stored in different locations. Diagnostic imaging places significant challenges upon the diagnostician as well as the attending physician and multimedia communications can assist in meeting this challenge. There is an increasing need for the set of physicians to act in concert in combining their talents for the best delivery of service to the patient.

Multimedia communications is characterized by the following factors:

1. Multi-Sensory: It uses several of the human senses in transferring, processing, and creating information.

2. Multi-User: It interconnects several users of the information into a conversational mode and allows a dialog based on a fully interconnected set of media.

3. Displaceable: It allows for the establishment of communications and information transfer that is displaced in both space and time from the source.

4. Interactive: It permits a real time interaction between any of the users of the medium, whether the users be human or databases or applications software.

The multimedia environment is one that is user centered and is designed to meet the users' needs in interfacing with complex images and in conveying information from one location to another. Multimedia is not just a description of how the data is stored, it is, more importantly, the description of a philosophy of human interaction with complex data elements in a multi-sensory fashion."

We must ask the question of what does the human want to do with multimedia that the human does not either want of is able to do with the classic single media information sources. There are several processes that are necessary;

1. Define: The user desires first to define a multimedia object. This object is in sharp contrast to a normal data object that is typically a structured and bounded alphanumeric data element, convertible into a digital representation. A multimedia object is the concatenation of video, voice, text, and other sensory representations of the event at hand. A multimedia event is the analog of a data object. The data object is the representation of a definable and measurable

term used in common communication, such as the word, NAME. In contrast the multimedia event is a collection of multimedia presentations that have a temporal and spatial extent to them. That is a multimedia even is a set of voice segments, a set of video elements, and a set of text frames.

2. To define a multimedia event means to concatenate in a rational form the set of disparate multimedia elements into a connected event. The process of connection is complex but it goes to the heart of human communications and understanding.

3. Query: The query in a data object case is a way to do one of two simple tasks. A data object may be either selected or enumerated. The selection process is based upon the ability to take a data object and recognize that it has a unique representation in some stable denumerable set. For example we can use the alphabet or a binary representation for any data element. Once we select a data object we can then enumerate all of the objects that meet a certain criteria, by again matching and now counting. Thus we can answer the question of how many patients over forty have high blood pressure. We first use the select process on high blood pressure and then the enumeration on patients. In a multimedia environment, we are now posing much more complex queries.

4. Store: This means that we must store complex multimedia objects, composed of video, voice, image, pointer movement, text etc. that may reside on different storage devices at different locations. We must be able to retrieve them in the same order and timing that they were stored and do so in a minimum time.

5. Process: We must be able to process multimedia objects, to alter, enhance, combine them. We must perform the processing in a fully distributed fashion, using the resources from multiple processors.

6. Display: The display of the multimedia objects includes not only the display of the image or video but the "display" of the voice and other sensory elements of the multimedia object.

7. Communicate: Communications means the development of a conversational mode. Conversation is key to communicating in a multimedia environment. Thus, we must not reproduce a communications environment that is attuned for the computer but one that is matched to the human user. The essence is the ability to effectively share the multimedia objects in a dialog fashion, interactive and interpretive.

These multimedia processes are to be done in a fashion that is transparent to the user. They must also be done in a fashion that is resonant with the way the users currently performs the tasks.

There are several dimensions that can be used to characterize the extent of the multimedia environment. These dimensions are;

1. Time and Duration: This dimension shows the amount of simultaneity that the medium allows both for a single user as well as for a collection of users. Further a dimension of durability to the environment is essential as the complexity of a multimedia object requires that time pass until it has its full representation. Thus unlike a mono-media object that can be represented to a single user in a fixed period, the interlining of media and users requires a sustainability of the environment.

2. Communication and Conversationality: This characteristic is one of allowing for a multimedia multi-user environment that permits a full sharing of the environment a an equal basis amongst all of the users. It further allows the users to interact with any other user while at the same time allowing this interaction along any one of the multimedia dimensions.

3. Interactivity and Responsiveness: This dimension relates to the ability of the environment to allow one or several users to utilize all elements of the medium and at the same time to pose questions that are robust in a multimedia sense and to obtain adequate answers.

4. Presentation and Interaction: The interactiveness of the environment is a key element of understanding the

5. Non-Linearity and Hyper-Dimensionality: This dimension of characterization allows for the movement amongst the object in an unbounded fashion. It allows for movement in space, all dimensions, and time, as well as in point of reference. The spatial movement entails the ability to view at different magnifications that is common amongst hypermedia environments. The ability to view at different points of reference allows one users to accept the reference frame of another to view the object.

6. Sense-Complexity and Representation: This dimension allows for the combining of multiple sensory elements into the multimedia objects as well as the presentation of those elements either as direct manifestations or as appropriate analogs.

In the context of this paper there will be three elements that define multimedia and multimedia communications; the message, the medium and the messenger. The ultimate result is the impact of the information created on the environment. Without the result, there has been no transfer of information. The objective of this paper is to show the relationships between all three of these elements and the blend with them a set of philosophical underpinnings that will allow them to be used in analyzing the development of multimedia communications.

We first define the three elements of message, messenger and medium.

1. Medium: The medium is the collections of all physical elements outside of the mind of the creator of the message that facilitates the externalization of the message. Paper, a video screen, a hypermedia environment, a set of signaling flags, a stone tablet are all the elements of the medium.

2. Message: The message is the "idea" to be transferred from one individual to another. It is the information content to be transferred and thus to acted upon. It is an actionable element of internalized conceptualization.

3. Messenger: The messenger generally is thought of as an individual. In our context, the messenger is the collection of any and all entities that move the message from one point to another. Recall that the movement may be i space and time. Recall also that the channel used in the movement may be a "noisy" channel that can introduce errors

There are three schools of philosophical thought that will be used to assist in developing and overall theory of multimedia communications and these schools, complementary in many ways, assist in each of these three elements. They are;

1. Semiotics: This is the study of "signs". In the current context, semiotics helps us with the study of the medium elements of multimedia communications. For indeed, the medium consists of the signs that are used in the conveyance of the idea. To the semiotician, the world is viewed in the sense of a pure sign, an outward display that can be interpreted to mean what the displayer intends. The semiotician reflects on the issues of identifying the sign and then identifying the meaning.

2. Deconstructionism: This is the school of thought that tries to understand the idea or the essential information from the message. What was really meant by the message sent. It takes the combination of the message, the medium and the messenger, in context, and tries to determine the essential message or information content. It does this through the inherent assumption that messages are actionable and it is through the direct of implied actions that the information may be revealed. The deconstructionist relies upon the deeper philosophical underpinnings of the conveyer of the sign. To properly deconstruct a sign, the deconstructionist must understand what world view or philosophical sets of precepts the holder of the sign has. Deconstruction begs the question of a real philosophical ethos, expressed or not expressed.

3. Hermeneutics: This is the study of what the messenger really meant by the words uttered. Hermeneutics focuses on the message via the messenger. It is an extension of biblical study of

trying to reinterpret the message of the gospels and old testament writers from the human perspective of the writer, in our case the messenger.

Our approach is to define and the describe each of the elements in the above and then to apply them to multimedia communications. The application will focus on two issues. First we will develop a set of analytical tools that will allow the user to understand an existing multimedia environment. Second, we will develop a methodology to synthesize the multimedia environment that meets the objectives of the designer. The reader must note that it is the joining of the words, the ideas and the signs that leads to the actionable event.

This paper is structured to address each of these above issues. It first focuses on the message, the medium and the messenger. It then shows how the three philosophical schools ca converge to better understand the issues of multimedia communications.

2.3 MULTIMEDIA; ELEMENTS AND STRUCTURE

Multimedia technology or systems have been viewed as multisensory end user focused technological add-ons to existing or new computer processing environments. Multimedia is a catch phrase for new technology that displays, either orally or visually, information that may be stored in an dense and complex form at the end users site. It has been viewed in many ways as a means to an end, yet viewed as an end in itself. To date, there has been no "philosophy" of multimedia, no Marshall McLuhan of the multimedia world to articulate who wants to use it and for what purpose. The key questions from a business perspective is how does the set of multimedia technologies provide increased value to business customers, and how does the consumer value the multimedia technology in their daily lives. The use of multimedia technology and its acceptance is an economical factor, based on value creation and perception on the part of the user. It is not, as has all too often been suggested, a technologically driven market.

2.3.1 Elements of Multimedia

To begin, the elements of multimedia as currently understood by the hardware and software providers can be presented. There are six major categories of multimedia technology and elements. These are;

1. Displays: The MM display areas focuses on enhancement of images, improved resolution, and display presentation processor enhancement. The display may be generically for any one of the senses; sight, sound, touch, or others that may be appropriate. HDTV is an example of one area of progress. Especially digital TV.

2. Processors: The processors and processes are the special software or hardware elements that locally enhance, manipulate, display, or interpret multimedia elements.

3. Content: This is the more general software that is created and used on a multimedia system. It is the result of authors works in creating information and interactive systems for multimedia.

4. Storage: This is the composite of elements that are used for a multimedia database, storage and retrieval system.

5. Enhancement: These are special purpose non-co-located processors and processes that are used for special purpose enhancements.

Communications: This is the ultimate heart of a multimedia system, namely the ability to communicate in a full and complete multimedia fashion. This is a complex amalgam of protocols, transport infrastructures, servers, and interfaces that allows and support multimedia communications amongst a group of users. All other elements are at best users to interface or utility, and not user to user. Multimedia communications systems support sessioning; namely synchronization and orchestration of multi-media/multi-user conversationality.

These elements must be combined in a fully integrated and networked fashion to deliver a transparent system or service to as wide a variety of end users as is possible.

2.3.2 Value Creation in Multimedia

The use of any new technology must fit within the overall context of value creation to the user. Value, as a concept, may be related to the microeconomic concept of the utility function of demand but it more closely is related to the ability to market the new product or service directly to the customer. Namely, value in the context of multimedia communications is best defined in being able to quantify for the user what savings in expenses or capital shall accrue from the use of the new product or what new revenue stream will result. Ultimately, value in an commercial context is nothing more than the increase in the net present value of the business entity. In the consumer market value is measured by means of its ability to displace other expenditures that the consumer already has made for a perceived greater value from the product offered.

In either case, value creation is essential when determining what multimedia has to offer. The focus in this White Paper is primarily commercial and not consumer. The latter has a greater risk threshold in uncertainty and will be dealt with latter. In the commercial context, therefore, value creation is a definable and measurable result of using the new technology in an existing business context. The business imperative, therefore, is that multimedia services must create value for the firm.

2.3.3 The Multimedia Food Chain and Its Missing Links

The lack of understanding of what multimedia is, combined with the need to meet the measures of the customers value chain, dictate that all elements of the customer's operational environment must be understood and considered to effectively establish and operate an effective multimedia service business. This is called the "food chain" concept. Having even a single element missing will result in starvation no matter how robust the other elements in the chain are. Thus, it is essential to determine if, in the development of a service business, any one company, supplier, customer, or other such entity, has been left from the flow of the service and thus will cause it to fail.

For example, in the healthcare market, imaging and multimedia applications have been developed in significant numbers. However, several key elements have been missing. Specifically; bill transaction processing and record management and keeping. The current approaches to the delivery of medical imaging systems attempt to support applications in radiology by merely replacing the viewing screen. However, this is but one step in the process. AT&T and Philips have learned the hard way that such a point replacement is unacceptable. The system that they have developed, COMVIEW, has had limited acceptance. The system is a typical high end multimedia system. It handles complex images and text, allows for the integration of voice into the overall system, and provides a limited amount of record management. However, it does not readily fit the pattern of the radiological suite in most hospitals. It does not solve the integration of record management and does not solve the issue of patient record keeping and billing.

In the Hospital environment, the use of multimedia will be driven by the need to treat each medical Department as a profit and loss center. Revenue must be ascribed to each procedure and each patient and expense tracked. Quality care will be an equal imperative. Thus a multimedia system must be one that starts with that premise, allows for graceful and incremental migration and addresses the needs of the physician, the technician, the nurse, the administrator and the support staff. The same can be said about all other market applications. Moreover, multimedia systems and services in healthcare and other similar markets, will have a significant impact because these markets are very information rich, and require communications of this information to many people. It is this nature of information richness, in both type and form, combined with the need to transact along with the sharing of the information that establishes a need for multimedia services.

2.3.4 Elements and Structures

The end user interface in a multimedia environment is dramatically different than that in a traditional data processing environment. Specifically, in a multimedia environment, we have

more intimately involved the human as a processor and evaluator of the information flow and we furthermore have complex information sets flowing from one point to another as well as in a conversational fashion. We will develop the end user concept in terms of the paradigm developed by Winograd and Flores.

In the development of a theory for design of computer systems involving the human user, Winograd and Flores invoke the theories of the German Philosopher, Heidegger. Specifically they refer to four key propositions of the philosopher that impact the overall end user interface issue in the multimedia environment. These are:

Our implicit beliefs and assumptions cannot be made explicit.

We all too often may make the statement, "You know what I mean." In so doing we are creating to mistakes. First, the other may never know what we mean just by the nature in which we individually perceive experiences and objects. Second, we may, ourselves, not have the insight to our own true beliefs, because we all too often find ourselves questioning them. Hermeneutics, the study of meaning in documents, has been expanded by Gadamer to investigate human reasoning. Thus, indicates Gadamer, our understandings can change with the time and place. This changing makes the explicit articulation specious at best.

Practical understanding is more fundamental than detached theoretical understanding.

Heidegger has a concept called "thrownness", part of being-in-itself. We know something only by being thrown or involved in it. We know what a radiologist does with an image and how he manipulates it for understanding by doing the process ourselves. We cannot expect the user to detail their beliefs and in fact those understandings are time varying.

We do not relate to things primarily through having representations of them.

We relate to things themselves. We do not relate to a representation. The representation to the "thing itself" is done in the context of the task to be accomplished. For example, teleconferencing is useful is we are not to relate to the person but to a subject whose essences can be presented directly through the medium, rather than just a representation. We find that teleconferencing is inadequate for personal contact since the contact is through a representation.

Meaning is fundamentally social and cannot be reduced to the meaning giving activities of individual subjects.

Meaning is obtained in dialog, in a conversational fashion, with the ability to meet consensus. Gadamer and Heidegger both relate meaning to the social process of communicating. Both also relate the evolution of meaning to the ongoing set of discourses.

Specifically, social or conversational activity is the ultimate foundation of intelligibility. This means that both in the design process as well as in the operations process, the need is critical to have the communications channel be conversational if the intent is to convey intelligibility. If the intent is only to transfer predefined package from one point to the other them the conversationality is not essential. In a multimedia environment, intelligibility in the context of the various media and thus intelligibility demands conversationality.

We can also try to better understand the interface by recognizing that the challenge is matching man to the machine. To do this we are frequently pressed into a metaphorical set of analogies. Typical is that of "man as the human computer". Metaphors of this type are both powerful explanations of new concepts and clear statements of our total lack of understanding of the issue. To quote from Mac Cormac:

"Explanations without metaphor would be difficult if not impossible, for in order to describe the unknown, we must resort to concepts that we know and understand, and that is the essence of the metaphor, a juxtaposition of the familiar and the unfamiliar."

MacCormac further quotes from Arbib, Man a Machine, The Metaphorical Brain;

"We want to understand how people think and behave....In some ways the brain of a man is like the computer of a robot, in others it is akin to the brain of a frog. Our aim here is to convey an understanding of the brain in terms of two main metaphors: The cybernetic metaphor, "Humans are machines," and the evolutionary metaphor, "Humans are animals." We shall downgrade the differences, but we hope to learn much from the similarities."

The harshness of Arbib pronouncements are striking. For indeed he represents the voice of many computer designers who view the human at best as a fellow computer and at worst a level above slime mold. Metaphor is powerful for it is in essence our most gracious way as humans to express our total ignorance of the true essence of the problem at hand. To use the Heidegger view, the use of metaphor is essential if we have not experienced the throwness of the problem, that we have not immersed our total being in the basic uncertainty at hand. We use metaphors as a way of re-expressing what we already know rather than understanding the unknown.

The danger in the use of such metaphors is clearly that we fail to come to deal with the needs of the end user in interfacing with the multimedia world. We view the end user as another peripheral computer system and not as an entity that must be thrown within the environment to

best profit from its performance. Thus, as we saw in the last section, current authors view multimedia as nothing more than another display or another storage device. Their world view, as so aptly developed in Kuhn's thesis, is limited to the existing paradigms. They talk metaphorically as man as computer, or worse as man as frog.

This world view is dramatically different from that of the rational school of thought that focuses on the idea that there exists a perfect truth independent of the individual and that through proper perception as a single individual this truth can be made clear. Heidegger's approach is that we must combine the rational objective world with the totally subjective individualistic world into an environment where the human users becomes part of the environment of the media.

As Gadamer has stated (see Warnke), we understand in a dialog manner. Specifically:

"If one examines Gadamer's analysis ...all knowledge of the natural and social world...is grounded in traditional orientations. We never come upon situations, issues or facts without already placing them within some context...and interpreting them in some fashion."

"In equating the logic of understanding with the structure of dialogue, Gadamer suggests that the proper answer is that..in genuine conversations ...all participants are led beyond their initial positions towards a consensus.."

Thus the process of consensus in a conversational mode is what leads to new understanding. All initial constructs are based upon prior prejudices that can best be formed in the context of metaphors. If our goal in developing new user interfaces is the ability to allow the users to understand, as viewed by Gadamer, then we must do so as to support the conversational modality and to allow the reaching of consensus.

Winograd and Flores have noted six effects of accepting the Heidegger world view. These are;

You cannot avoid actions.

Even inaction is a form of action. Managers, as developed by Simon, interact with their day to day industrial environment, and managers who act by inaction have the corresponding results.

You cannot step back and reflect.

Events exogenous to us are continually occurring and any attempt to stop time to best understand the situation is at best specious. At worst, it becomes inaction. The concept of hermeneutics is one that extended to the environment of the end user say that we make interpretation with what is at hand and what is part of our tradition.

Effects of actions cannot be predicted.

We can anticipate, we can plan and we can strategize, but the world is filled with uncertainty. As such, we act in an environment where the exact outcome is uncertain. The user must anticipate that but not be fearful of it.

You do not have a stable representation of the situation.

Every situation is a representation in flux. When a user accesses a system, there are many factors that impinge on the interaction of the user, their needs and responses. No system interface to a user should assume a stable representation of facts. Designs should be such as to prepare for ambiguity.

Every representation is an interpretation.

X rays are inherently representations of physiological factors. In looking at an x ray a physician is looking at a representation and performing an interpretation. When we design a user interface, we are representing a set of facts to the users. The act of the designer in representing the facts is an act of the designer in interpreting for the user the facts. Thus in designing the interface, the designer must be aware of the fact that they are entering into the interpretation process. Not only is the user interpreting but so too is the designer for the user.

Language is action.

Speech through our language is a spontaneous reaction to a set of situations. In the design of computer interfaces we spend many hours on structuring the presentation of the visual material. Images are carefully scrutinized. Speech, in a multimedia context is fluid and open to instant interpretation that may not be consistent with the other participants in the multimedia session. For example, our tone of voice may make us appear arrogant, our questioning may make us appear petulant and our suggestions may make us appear pedantic. Despite all our structured work on the interface, the instantaneous impacts of the language may override the setting. Thus a multimedia environment must have the flexibility to self-correct.

There are eleven design guidelines that Winograd and Flores have articulated and these play well into the end user interface effort associated with multimedia system. These guidelines are as follows;

1. There are no clear problems to be solved. Action needs to be taken in a situation of irresolution.

2. A business is constituted as a network of recurrent conversations.

3. Conversations are linked in a regular pattern of triggering and breakdown

4. On creating tools were designing new conversations and connections.

5. Design includes the generation of new possibilities.

6. Domains are generated by the space of potential breakdown of action.

7. Breakdown is an interpretation - everything exists as interpretation within a background.

8. Domains of anticipation are incomplete.

9. Computers are tools for conducting the network of conversation.

10. Innovations have their own domains of breakdown.

11. Design is always already happening.

If we follow these design rules in developing the human interface and if we understand the underlying theories of human understanding and intercommunication, this will assist the designer in being flexible to converge to a more stabler interface.

Language is a means to expresses knowledge. Language is also a means to gain and crate knowledge. We can now expand the concept of a language from what we see as words and what we hear as speech to what we see as actions and what we create as situations. The ability to provide for more breath of language as intercourse and interaction, and the ability to extend that intercourse to all of the senses, no just sight and sound, allows for the attainment of the fulfillment of a multimedia communications environment.

The concept of the end user interface that we developed in the last section centered around the need for conversationality. Conversationality is embodied in the concept of a session which is the electronic implementation of communications in a multimedia environment. We shall discuss the concept of a session in more detail in Section 5 on communications. Simply put, however, the session is the underlying communications construct that ensures the binding of users

together. Understanding the multi-user environment will provide the boundary conditions that envelop the communications environment.

The essence of multimedia communications in a multimedia environment is the embodiment and completion of transactions. Transactions are an ordered set of actions taken by the set of participants in the session, whose completion leads to the successful

The users in a multimedia world are categorized into several classes. We can consider three levels for discussion purposes and we may possibly expand these as we proceed. The three are the end user, the physical and the virtual user. The end user is a definable entity that has action responsibility for effecting the transactions that occur in the system.

Data elements in a commonly accepted database are collected according to some schema and labeled accordingly. For example, we may consider a typical database containing the elements of name, address and phone number. In a computer database these elements are then encoded into a digital format and the name, address and phone are stored in some binary form for latter retrieval. The retrieval process can be performed by asking a simple question of the database, specifically, "List all of the name for the case where the zip code is 05XXX." Here we have asked the database to perform certain acts. First to go through all of its records and perform a match on the basis of zip codes. This is a redials performed task since all it entails is matching a bit pattern for the desired code with the bit pattern for each data entry along the zip code field. We know how to do this. The we accumulate all the names and finally print them out.

There are certain inherent structural assumptions that we make in this type of database. First we assume that each data element of a data object has a defined structures. Second we assume that each data element is decomposable in a finite sequence with uniquely defined data objects. For example, we know that address is composed of the objects of street number, street name, city, state, and zip code. We further know that zip code is composed of five digits, not characters. We also know that there is no other representation for address and that further the object address contains no other information. Address, for example, cannot tell me about the type of house, its color or the number of windows. The questions that I ca ask are a priori stated and implicit in the structure of the data object.

Now consider a multimedia data object. The object is an image, specifically an x-ray. Now we can create a patient record which contains the information of the type, name, address, date of admission, attending physician, blood tests, and x-ray. The objects, with exception of the last in this record, are of the type that we discussed above. They are bounded, fixed, and definable. Specifically we can ask a specific question and obtain a quantitative answer. In contrast we cannot ask questions of the x-ray on a specific basis and hope to get an unambiguous answer. We

may ask for the last name and get the answer "Jones". We can ask about the x-ray the question of the disease and get a totally ambiguous answer.

A more common data object may be considered to be a self-correct hyper object if we allow for the depth that may be part of the decomposition of the object. For example, we have the record for "Student" which is composed of name, address, and grades. Name is further decomposed into last, firs, middle, and even parents. Parents is decomposed into mother and father.; And they in turn into last, middle, and first. The common data object is in itself a hyper object having depth and extent. However, the depth and extent has been defined a priori and is part of its very creation. The database designer had defined fields for allow of the elements and for the relationships between the hyper objects depth and co-relations.

If we extend the hyper object case to the multimedia object, we fond dramatic differences. Consider again the case of the x-ray. At best the database designer assigns 100 million bits of space for the data object. There is no structure. For example, the image may be that of a chest x ray and we may be interested in the lungs, the heart, the stomach, or any one of several organs. In the context of that interest, we are further interested in asking a set of important questions as to the nature of the lung's clearness, and if not clear, what is the nature of the perceived shadowing. All of these are useful for a careful diagnosis. Thus with a multimedia object, the object contains information in a complex form, a form that is processed by the cerebral cortex only a posteriori, and not a priori. Further, the information has dimensions that are not fully known before the interaction with the end user. This is the essence of the concept of Heidegger that we had discussed as an integral part of the Winograd and Flores theory of computation. A multimedia object is truly a new paradigm for the representation of data.

The record in a standard database is quasi static in nature and can be changed in a typical transaction processing system only upon the commands of the central data base administrator or upon the allocation of commands in a distributed database environment. Thus, for example, we can envision a database that has the record for the number of seats available on a typical airline flight. The seats available may be changing with time but the change is controllable and is synchronized by some overall database mechanism.

In a multimedia object, there is change in the object that is a natural progression of the object in time. For example, if the multimedia object is a speech signal, we know that if the signal was sampled at 64 Kbps then we must play the record back at the same speed. If we create a compound multimedia record of voice and video, we must now synchronize not only within the objects but also between the objects. Time now plays an integral role in even static records. If we now extend this to the case of a dynamic transaction like multimedia record, we find that the maintenance of the synchronization within and between objects is as critical an element as is all others.

Thus multimedia objects have a complexity and richness of structure and dynamism and interrelationship of form that separates them dramatically from the typical data records that we use as part of the typical data environment.

Multimedia objects have three major dimensions of complexity. They have a temporal, spatial and logical structure. The temporal structure there is tempo to the object as there may be in a speech segment or a video segment. The tempo may also relate to the movement of a cursor and its relationship to other objects. In addition, the temporal object structure may be a segment (namely bounded a priori such as an image) or a stream (uncertain terminus such as a speech or video segment). The temporal structure of a multimedia objects reflects the timing, sampling rate, type of object, appropriate delivery time and the boundedness of the object.

The spatial characteristic reflects the ability to decompose the multimedia object into spatial parts. This is particularly true of images that have some spatial dimension in two or three dimensions. The basic need her is to have a decomposition structure that can be discussed about the object in an in band or out of band fashion. The metaphor of boundedness is a concept that reflects the location of the information on the spatial decomposition of an object. If the object has its rules for decomposition imbedded in the data stream itself, we call it in band. If the decomposition information is located as a separate descriptor, namely a data header, we call it out of band. We must deal with this fact. In images, we restrict the boundedness to a finite set of display elements. Thus we know a priori that an image is 2,000 by 2,000 pixels of 24 bits per pixel, rather than an unknown length voice segment.

In the logical domain, we also are concerned about decomposing images. Logical decomposition may be simple to state but very difficult to implement. For example we may want to decompose objects in the form of type of bones in x-rays or by excited speakers in a voice segment. There is currently limited analytical processing power to extract this type of decomposability.

We can then further combine these simple multimedia objects together into a compound multimedia object. The compound object now has needs for orchestration, that is timing all of the timings of the simple objects. The concept of orchestration is simply just that, being the conductor to assure that all of the instruments in the symphony are not only properly tuned by timed with regard to each other. We also worry about the issue of concatenation, the opposite of decomposition. Concatenation states how we handle the spatial and logical combing of the simple objects.

Communications in a multimedia environment is a process of allowing agents of the communications process at various levels to interexchange information in support of the complete end to end flow of multimedia elements. This definition allows for the existence of

agents which are definable and locatable entities that have specific responsibility for communications flow and management. In the world of multimedia communications, these agents must be empowered with significantly more responsibility than is typically the case in a data communications world.

We will develop the concepts of the communications environment in the multimedia world by first developing the concept of a session. A session is a shared multimedia conversational mode of communications allowing multiple users and devices to share a common working space in such a way as to make the communications interfaces and transport not only transparent but acting as a facilitator of the communications process.

A session provides for several key layers of functionality. These are:

Event Management: Any multimedia activity leads to a transaction. The event manager is the transaction manager. Multimedia events are complex and thus the manager must deal with these distributed complexities.

Dialogue Management: The multimedia conversation must allow sharing of resources and must encourage and facilitate conversationality. Dialogue management is that function.

Activity management: Activities are extended events, they are displaced conversational elements that must be remembered and connected.

Synchronization: The orchestration of the many elements in a multimedia conversation is the role of the synchronization function.

We must define entities that connect together and further define the entities that ensure that this session level connectivity is supported.

2.5 Definition of Multimedia Services

Multimedia Services requires a working definition. This section attempts to provide such a definition. It differs dramatically from many of the definitions offered for multimedia in many of the current trades or even in the business and strategic plans for companies purportedly in this business. The following is the definition:

Multimedia Services is a set of services offered to a community of users that enables and empowers them to perform tasks in a collective fashion that can be accomplished in a significantly more productive fashion by including multisensory information and interexchange in a fully conversational mode, transparent to any and all users, and allowing all of the processes

performed to be monitored, recorded, retrieved, and transacted in a fully electronic fashion. Multimedia Services responds to the needs of the users and can be measured in the context of increasing value to the user base and can be provided in such a fashion that it does not neglect any element of the organizational food chain.

Multimedia Services has certain characteristics that empower it to enable the user to achieve the value that has been determined. Multimedia Services must account for the elements of the human and organizational environment. From the human perspective, several of the common features are as follows:

2.5.1 Human Factors

1. Conversationality: A system, as service, must allow, and more importantly encourage and support, a conversationality amongst users that enables all of the multimedia senses. Understanding and meaning are essential a social act that requires the full sensory interaction amongst individuals and permits them to act together. Multimedia services must not only enable this type of activity but must do so in a fashion that is consistent with existing social paradigms. The essence of multimedia conversationality is its ability to empower the user to be "present-at-a-distance".

2. Throwness: The ability of the user's actions to have some immediate or consequential effect on their environment has been called being thrown into the environment and the term throwness has been used. This capability implies that the user of the service is able to manipulate the object or objects at hand in all of their sensory dimensions. The use of a multimedia service in the design process must be such as to allow the designer to see the impact of the design in terms of its tactile and visual elements, and also the aural elements that may be part of the design. The designer must also have the ability to "use" the designed object in the context of its use to better understand it functionality and acceptability.

3. Breaking Down: The individual does not understand something of an object, especially a multimedia object, unless they have the ability to break it down, namely do something with it to make it function as we know it. Take for example the use of multimedia in healthcare pathology. When a physician views a slide of a tissue sample, they are viewing it after having done so many prior times and having related certain cell boundaries and shapes, blurs to the uninitiated, yet critical to diagnosis of certain conditions of the body organs.

4. Transparency: A multimedia service qua service must be transparent to the users. For example, setting up a simple conference call on a telephone is made so by calling an operator who then calls all of the participants and adds them to the conference. Cumbersome as this may

be it creates a sense of transparency, namely, no one user must know how the call is to be set up. Multimedia services must be operator independent and yet have transparency to operations. This will be one of the major challenges of the service. It will also be a barrier to exit in a competitive market.

5. Action: The service must allow and enable action. This may mean that in a multimedia session, control of the conversation can be taken around the session, allowing any single individual to explain a point through demonstration. It may mean the obtaining of an object or element on demand, and then having the ability to manipulate it at will. It is the ability to demonstrate, contradict, confer, and compromise by example.

The second set of factors to be part of a multimedia service relate to the organizational structure that the service fits into. Specifically:

2.5.2 Organizational Factors

1. Transaction: The service must support the total set of transactions that will occur through its use. At the simplest end, it must create and process bills for services rendered as part of the service. It also must track each interaction of a user with the system in order to support customer service and assist customers in the event of problems.

2. Productivity: The service must address the needs of the customer for productivity improvements. This means that the service must immediately reduce costs through more effective use of existing manpower. For example, in the advertising application, the service must reduce the cost and time associated with the development and production of ad copy. In short, it must be cheaper and faster.

3. Infrastructure: The service must become transparent to the user. It must exhibit all of the qualities of an infrastructure.

4. Value: The service must have a calculable value to the user. The service must address specific processes or operations, with known cost structures, and must definitively show how the costs are reduced and the value of the business unit increased.

5. Holistic: This implies that the service must integrate into the overall way in which the business operates. It should not optimize one part and fail to address inefficiencies in others. It must address the entity as a whole.

2.4 THE MESSENGER; COMMUNICATIONS AND INFRASTRUCTURE

The messenger is the process and processes that link with the medium to effect the transaction of the message. It is of interest to see that the old AT&T used Mercury as the messenger. Hermes, also a messenger, although a mischievous one, is a better choice. Mercury delivered the messages correctly. Hermes would always put a little twist on them. We shall in this section develop the concept of a communications system and then move forward on how that system may be viewed as an infrastructure. A great deal of talk about infrastructure has occurred as of late with little definition. We try herein to define out terms. This will be critical as we apply philosophical principles to the issue of messages, messengers and media.

2.4.1 Communications Systems

There are four architectural elements in the telecommunications network. These elements are the control functions, the transport function, the interconnect function, and the interface function. We now provide further detail on these functions. It should be noted that these functions have evolved over the years in content and complexity. We view these elements in the context of a communications network that must support the most advanced current concepts in communications. Specifically, the world view adopted in this paper that lead to an interpretation of this architecture are:

(i) End users desire to have interactions in a real time fashion with images and other high resolution information that must be provided in a fashion that meet both time and resolution requirements (See Barlow).

(ii) The end user devices are extremely intelligent and complex and can operate in a stand-alone environment.

(iii) The users desire to operate in a totally distributed fashion. Data bases will be a different location, users are at different locations and input output devices are also at different locations (See Dertouzos and Moses, and de Sola Pool pp. 57-59 for details on these directions).

(iv) The network may provide different levels of service to different users. There is no need to provide universal service of full capability to all end users.

This view of the network will significantly influence how extensively we defined the elements and in turn will impact the combination of those elements in an overall architecture. All of these assumptions on the world view are different than before, in an all voice world. In this paper, we define a network as an embodiment of an architecture, in all of its elements.

The architectural elements are control, transport, interconnect and interface.

1. Control: Control elements in an architecture provide for such functions as management, error detection, restoral, billing, inventory management, and diagnostics. Currently, the voice network provides these functions on a centralized basis, although in the last five years there have evolved network management and control schemas and products that allow for the custom control and management of their own network. Companies such as IBM, AT&T and NYNEX have developed network management systems that move the control from the network to the customer (See McGarty and Ball, 87, for a detailed discussion of the different types of control and network management strategies). On the sub-network side, companies such as NET, Timeplex, Novell, 3-COM and others have done similar implementations for local area networks, data multiplexers and other elements. Centralized network control is now longer necessary and in fact it may not be the most efficient way to control the network. What is important, however, is that network control providing the above functions is an essential element for either a public or private network. Thus as we consider network evolution, this element or set of function must be included. Control has now been made to be flexible and movable. The control function is probably the most critical in the changes that have been viewed in the context of an architecture. All buildings need windows, for example, but where one places the windows and what one makes them of can yield a mud adobe or the cathedral at Chartres. The same is true of the control element. In existing networks, the control is centralized, but in newer networks, the control is distributed and empowered to the end users. The users can now reconfigure, add, move, and change their network configuration and capacity

2. Transport: The transport element is provided by the underling transport fabric, whether that be twisted pair of copper, fiber optic cable, radio or other means. Transport should not be mixed or confused with other elements of the network. Transport is merely the provision of physical means to move information, in some form such as digital, from one point to another. At most it is expressed in bits per second and at best it is expressed in bandwidth only. Bandwidth as a transport construct is the most enabling. Transport does not encompass the need to change the information or to do any other enhancement to the information.

3. Interconnect: The interconnect element of the architecture describes how the different users are connected to one another or to any of the resources connected to the network and is synonymous with switching. Interconnection assumes that there is an addressing scheme, a management scheme for the addresses, and a scheme to allow one user to address, locate and connect to any other user. Interconnection has in the past been provided by the Central Office switches. As we shall discuss latter, this implementation of an architectural element was based on certain limitations of the transport element. With the change in the transport element of structures allowing greater bandwidth, the switching needs have changed. Specifically, distributed systems and scale economies of the distributed Architectures allow for interconnectivity controlled by the CPE and not the Central Office. As we shall show later, the advent of Local Area Networks and CATV voice communications are ones using distributed

interconnectivity elements. There are three general views of interconnection that are valid today; the Telcom, the Computer Scientist, and the User. The Telcom view is based on the assumption of voice based transport with universal service and the assumption of the inseparability of interconnect and control. The Computer Scientist view is based upon the assumption that the network, as transport, is totally unreliable, and that computer hardware and software must be used in extremis to handle each data packet. Furthermore the Computer Scientist's view of the network is one where timeliness is secondary to control. The Computer Scientists view has been epitomized in the quote, "Every Packet is an Adventure". This is said with glee, in that each data packet is set out across the network and it is through the best of hacking that the Computer Scientist saves the packet from the perils of Scylla and Charybdis. The third view is that of the user, who is interested in developing an interconnect capability that meets the needs and minimizes cost. This is minimization of both obsolescence and cost strategy.Thus an investment must try to follow the curve. In a hierarchical view of interconnect, such as a large centrally switched network, the changes occur once every few years. Thus the lost cost or performance efficiency can become significant. In contrast, in an end user controlled environment, with a fully distributed architecture, the lost efficiency is minimized as technology advances.

4. Interface: The interfaces are the end users connection to the transport element. The interface element provides for the conversion from the end user information stream and the information streams that are used in the transport form of the network. For example, the telephone interface for voice is the analog conversion device. Moreover, we must also view interfaces as being composed of not only the hardware that provides for the physical interconnection with the end user but the software that assists in the logical interconnection. The telephone hand set is merely the first physical step in interconnection. A second step included the PBX which included software that allowed for additional features. However, these features are frequently not used because of the user lack of acceptance.

We have divided the network elements into these four categories to demonstrate that there are clearly four distinct and separable areas for growth and policy formation. Issues of regulation, due to potential monopolist control are always a concern, but it will be demonstrated that in all four there are economies in market disaggregation.

Understanding that there are several varying architectural designs allows one to better understand that each reflects not only connectivity but also the world view.

2.4.2 Infrastructures

Let us extend the concept of infrastructure. We find that there is a great deal of discussion about infrastructures, in both academic and governmental circles. Unfortunately, none of the participants deem it appropriate to define what they are talking about, namely what do they mean

by infrastructure. We feel that such a definition is critical to developing multimedia communications concepts, since the concept of an infrastructure will be at the heart of the change process. In our context, an infrastructure is a shareable, common, enabling, enduring, resource, that has scale in its design, and is sustainable by an existing market, and is the physical embodiment of and underlying architecture. Specifically;

1. Shareable: The resource must be able to be used by any set of users in any context consistent with its overall goals.

2. Common: The resource must present a common and consistent interface to all users, accessible by a standard set of means. Thus common may be synonymous with the term standard.

3. Enabling: The resource must provide the basis for any user or sets of users to create, develop and implement any and all applications, utilities or services consistent with the underlying set of goals.

4. Enduring: This factor means that for an infrastructure to be such, it must have the capabilities of lasting for an extensive period of time. It must have the capability of changing incrementally and in an economically feasible fashion to meet the slight changes in the environment, but must meet the consistency of the world view. In addition is must change in a fashion that is transparent to the users.

5. Scale: The resource can add any number of users or uses and can by its very nature expand in a structured form to ensures consistent levels of service.

6. Economically Sustainable: The resource must have economic viability. It must meet the needs of the customers and the providers of the information product. It must provide for all of the elements of a distribution channel, bringing the product from the point of creation to the point of consumption. It must have all of the economic elements of a food chain.

7. Physical Embodiment of an Architecture: The infrastructure is the physical expression of an underlying architecture. It expresses a world view. This world view must be balanced with all of the other elements of the infrastructure.

An infrastructure is built around the underlying architecture. An infrastructure is in essence the statement of the architecture which in turn is the conceptual embodiment of the world view.

Infrastructures as physical embodiments of architectures must, to have economic lives that are meaningful, be developed when the world view, technology and user needs are stable. If any of

these three are in states of significant flux, the infrastructure may soon not meet the change in the world view and then become obsolete.

It is important to distinguish between architecture and infrastructure. We have extensively defined architecture in terms of its three parts: elements, world view and technology. Infrastructure unfortunately has been reified in terms of some physical embodiment. The discussion of NREN being an infrastructure is viewed by many as being a determinate thing. Kahin has, however, de-reified the concept in terms of its being an embodiment of a concept or set of common goals. We expand that and state that an infrastructure is an enabling capability built around a common construct.

There are four types of infrastructure views that are pertinent to the current discussions of networks. These are of particular import to such networks as NREN since they will lead to the policy directions that it will take. These four infrastructure types are as follows:

1. Physical: This is the most simplistic view of an infrastructure. It requires a single investment in a single physical embodiment. The old Bell System was such an infrastructure. The National Highway system is such an infrastructure.

2. Logical: This network may have separate physical embodiments, but all users share a common set of standards, protocols and other shared commonalties. All users have access through an accepted standard interface and common higher level transport facility. IBM had attempted in their development of SNA in the mid 1970's to develop a logical infrastructure in data communications. This was expanded upon by the ISO OSI seven layer architecture, selecting a specific set of protocols in each layer.

3. Virtual: This type of infrastructure is built on intermediaries and agreements. It provides shared common access and support interfaces that allow underlying physical networks to interconnect to one another. Separately, the individual networks may use differing protocols and there are no common standards. The standards are at best reflected in the gateways to the interconnection of the network. Thus this infrastructure is a loose binding through gateways. It is in many ways what is the INTERNET today, if we include all of the subnets.

4. Relational: This type is built on relationships between the network parties and the establishment on higher level accessing and admission. Specifically, a relation infrastructure is based on agreements on sharing addresses, not necessarily common addressing, and on the willingness to share data formats and types. It is an infrastructure based on shared common interests but not shared common access. This type of infrastructure is what in essence exists in most cases today. Users can move from network to network through various gateways. The difficulty is the fact that the interfaces are cumbersome and may require sophistication on the

part of the users. However, more intelligent end user terminals and interfaces will reduce this cumbersome interface problem.

Our conclusion is that understanding the type of infrastructure that the coalition of users want, will also impact the architecture, based upon an imputed world view. Arguably, a physical infrastructure leads to maximum hierarchical control and the resulting impacts that such control leads to. This is a critical issue for networks such as NREN, since by choosing infrastructure and architecture may not be as uncoupled as desired. In particular, the selection of Gbps capability may really be GHz capability and is best suited to a Virtual or Relational infrastructure.

2.4.3 Current Infrastructure Options

There is a considerable amount of effort to define and implement an information infrastructure. In this section we describe some of these current proposals, many of which are still quite formative and lack substance. In some case we shall try to place them in the context of the constructs that we have developed in the preceding sections. In order to fully describe these infrastructures, it is also necessary to deconstruct the work of the authors, understanding their meanings in the contexts of what they are saying and taking an approach that blends the hermeneutics of Gadamer with the semiotics of Levi-Strauss. We now address several of the more current views of infrastructures. In each case, we describe as best can be done the concepts of each of the individuals, and then attempt a deconstruction in terms of their underlying architectural assumptions, their view of infrastructure and more importantly their world view of information and information networks.

(i) Dertouzos Infrastructure

This is the most widely discussed of the information infrastructures having been proposed by Professor Dertouzos who is a Computer Scientist and the Head of the Laboratory for Computer Science at MIT. Simply put, he defines the information infrastructure as:

" Common resource of computer-communications services, as easy to use and as important as the telephone.."

Dertouzos states that there are three elements to his vision of an information infrastructure. These are:

Flexible Transport: This includes bandwidth on demand, flexible pricing and security and reliability.

Common Conventions: This includes his concepts of E Forms and Knowbots. The former is a set of standard for formats and the latter are intelligent agents for the movement and processing of data.

Common Servers: This is a set of common file servers or generalized servers to provide directories, text/image translation, data base access and active knowledge.

In the paper, Dertouzos discusses this architecture and he uses as an example a system conceived of and designed by the senior author (McGarty 1990 [1],[2], 1991 [1],[3]). In the author's system, the assumption was to both empower the end user and to do so in an incrementalist fashion. The architecture shown in this second system was based upon:

1. Available Transport: Take what is present and build in an economically viable fashion. Build communications on an incremental and economically effective basis.

2. Open Interfaces: Use standards as appropriate, and allow the users the freedom to meet their economic needs. Recognize the changing needs of the user and buyer and incrementally change to meet the evolving needs.

3. Client Server Architecture: Maximize use of end user terminals and empower end user applications development. Provide tools and not strictures.

The system designed and operated by the author actually connects the MIT campus with hospitals, publishers, and other economic entities in a build-a-little, test-a-little, use-a-little approach that allows for use acceptance and economic justification. The Dertouzos Infrastructure assumes directions that are significantly different and diverge from the end user driven approach of the author but take a more centralist approach. This latter approach has been advocated by Moses in his discussions on the subject, yet are somewhat counter to Moses' layered organizations that maximize flexibility and minimize complexity.

(ii) Kahn Infrastructure:

The vision of Bob Kahn, of CNRI is one of a broad band research backbone, loosely coupled, with dark fiber and as high a bandwidth as possible, read data rate. This proposal, frequently confused with the Gore infrastructure, is generally more open and flexible. However, it too lacks any economic underpinnings.

(iii) Gore Infrastructure:

Gore, the Vice President of the United States in the Clinton Administration, son of the initiator of the Federal Highway system, has argued for a single network, government directed and funded, hierarchical in fashion, that allows everyone to have access to every bit. Consider his comparison of data bits to corn kernels;

" Our current national information policy resembles the worst aspects of our old agricultural policy, which left grain rotting in storage silos while people were starving. We have warehouses of unused information "rotting" while critical questions are left unanswered and critical problems are left unresolved."

He believes that every bit is a good bit. He further has no value concept of information. His definition is clearly the one of quantity and not value. Researchers are not necessarily starving for lack of bits. Quite the contrary, there is a need for coherent data reduction. He further states;

"Without further funding for this national network, we would end up with a Balkanized system, consisting of dozens of incompatible parts. The strength of the national network is that it will not be controlled or run by a single entity. Hundreds of different players will be able to connect their networks to this one"

He is somewhat contradictory. On one hand he states that there should be one network and not many, on the other hand he has all the separate networks connecting to this one. In this case, his world view comes through clearly. He wants a hierarchical or at most centralized architecture as well as a physical architecture. The proposal lacks the flexibility of an economic entity.

(iv) Heilmeier Infrastructure:

Heilmeier, the new President of Bellcore, the R&D arm for the Bell Operating companies on the regulated side, advocated a hierarchical, BOC controlled, network intensive, monolithic network. This is not surprising considering his extensive stay in Washington as a government bureaucrat. He further argues for control of both wire based and wireless networks. He is quoted as saying;

"I'd like to see a bona fide information infrastructure rather than a fragmented world of different systems for everything."

Networks are currently fragmented and as a result of this fragmentation local economic optimization has occurred. In contrast to the hierarchical, centrally controlled view of Heilmeier, also formerly head of DARPA, wherein he views the need for a single point of control and direction, the world of communications networks and information networks have grown through the increased power of the end user interfaces and interconnected distributed throughout the network. In addition growth has resulted from less control in the network and less centralization.

The work of the author (McGarty 1990 [1], [2], 1991 [1],[3]) has shown an architecture for a distributed multimedia environment that has been built and is still in operation that uses a mix of communications channels and thrives on those channels that have the least functionality. Specifically, dark fiber transport is the most enabling and empowering of any communications channel.

(v) Proposed Infrastructure

Infrastructures are enabling entities. As we have discussed, an infrastructure does not have to be a single centrally controlled, managed, and funded entity to be effective. In fact an infrastructure on the loosely constructed basis of a relational infrastructure is just as effective as the extreme of a physical infrastructure. We make the following observations, and based on the prior developments in the paper propose an alternative direction for infrastructure development.

(i) Technology is rapidly changing and will continue to do so. The directions in technology are towards increased processing capability per unit workstation and increased capacity in performing both complex processing tasks while at the same time handling sophisticated protocol procedures.

(ii) User terminals are expanding in a network multimedia environment that is empowering the end users to both use many new media types as well as dialog in a conversational basis with other users in the same network.

(iii) End users are becoming more pervasive and training of users based upon strict confines of computer languages are disappearing. The end user is empowered to act and to use information system with no training or education. Citibank, in its development of the ATM network has ensured that the systems have minimal need for human intervention or training. In addition, the Citibank home banking product, the most widely used of any home banking products on PCs, is almost instruction free. The Apple MAC computer is also another example of enduser empowerment through intimidation free end user interfaces.

(iv) Successful technology development in a productive fashion has best been effected within the constructs of entrepreneurial small companies that allow for the creation of new ideas judged by the dynamics of a free market. Large centralized technology development organizations have time scales that are much longer than the time scales of the underlying technologies. The developments in the computer industry of today are prime examples.

(v) Users are not only empowered to use systems in a variety of ways but they are also able to select from a wide variety of systems, interfaces and data sources. To quote A.G. Fraser of

Bell Laboratories: "Every standards body seems to be churning out protocols left, right and center. We may already have passed the point where we can all come together." (Coy, 1991)

Thus, distributed networks, interfacing with disparate other networks, through gateways is already a reality.

These observations then indicate that with a changing base of customers, a changing set of needs and an already progressing infrastructure that is relational at best, that to continue to maximize our technical creativity it is best to match the information infrastructure to our cultural paradigms. Thus it is argued that the proper evolution of an information infrastructure should be along the relational model. That, in fact, the physical extreme is counter to the trends of user empowerment and economic efficiency. It further could provide a roadblock to technical creativity.

2.5 DECONSTRUCTION

Deconstruction consists of the unraveling of actions and physical realities to determine the underlying sets of truth, such being definable and determinable. In its simplest state, the unraveling is the determining of intentions within the context of how the intended views the world at the time of the intent. Deconstruction from the perspective of multimedia communications is a toll that helps the designer understand the biases, prejudices and limitations of the designer. Deconstruction allows for the ultimate use of the technology in the context of the user. We start the development of the deconstructionist approach with the development of the concepts of paradigms and world views. We have already discussed them in some details but it will be critical to the deconstructionistic approach to have them developed in full detail. Then we shall move to the analysis of a set of key questions that appear when we begin the development of a multimedia communications theory.

2.6 PARADIGMS AND WORLD VIEWS

The concept of an information system or communications architecture has been a cornerstone in the development of new information and communications systems. However, the structural elements of these architectures have not played a role in the development of policies. In this section we will develop the concept of an architecture as a means to understand the network as both a market and regulatory entity, and will provide a new set of perspectives for viewing the network in terms of a new paradigms and world views.

The concept of a telecommunications architecture has been a cornerstone in the development of new telecommunications systems. However, the structural elements of these architectures have not played a role in the development of policies. In this section we will develop the concept of an architecture as a means to understand the network as both a market and regulatory entity, and

will provide a new set of perspectives for viewing the network in terms of new paradigms and world views.

An architecture, first, requires that the underlying system be treated in terms of a set of commonly understood elements and that these elements have a clearly demarcated set of functions and interfaces that allow for the combining of the basic set of elements. The way the elements then can be combined, reflected against the ultimate types of services provided, determine the architecture.

An architecture, secondly, is driven by two factors; technology and world view. Technology places bounds on what is achievable, however those bounds are typically well beyond the limits that are self-imposed by the designer or architect in their view of the user in their world. This concept of architecture and the use of design elements is critical in understanding the paradigms used in the structure of information systems (See Winograd and Flores, pp. 34-50, especially their discussion of Heidegger and Throwness in terms of design). World view is the more powerful driver in architecture (See Kuhn, pp. 72-85). We argue in this paper that it is essential to develop a philosophical perspective and understanding of how to view networks. We argue with Winograd and Flores, and in turn with Heidegger, that we must be thrown into the network, to understand the needs of the users, and to understand the structure of the paradigms that are used to construct the world view.

The concept of a paradigm is in essence the collection of current technologies that we have at hand for the network and the ways we put these elements together. However, the true meaning of a paradigm is in the context of the examples or experiments that we all relate to with that technology. Paradigms are not technology in and of itself, but technology as example. New paradigms result from new technologies. New technologies allow for the placing of the elements together in new ways. Kuhn, then goes on to demonstrate that the world view, that is how we view ourselves and our environment is based upon the our acceptance of these paradigms, as either collections of techniques and technologies or as collections of embodiments of these techniques and technologies in "examples". We then tend to accept this as the way things are and should be. Then Kuhn argues, as the technologies change, changes in the paradigms do not occur in a continuous fashion but almost in quantum leaps. The new paradigms build and congeal until they burst forth with new world views. It is this model that we ague applies to the evolution of broadband.

Thus, architecture is the combination of three parts: the common elements, the underlying technology and the world view. We depict the conceptualization of architecture as the amalgam of these three elements. We shall develop this construct more fully as we proceed.

The concept of a world view is an overlying concept that goes to the heart of the arguments made in this paper. To better understand what it implies, we further examine several common views and analyze the implications of each. If we view our world as hierarchical, then the network may very well reflect that view. If we further add to that view a bias towards voice communications, these two elements will be reflected in all that we do. The very observations that we make about our environment and the needs of the users will be reflected against that view. As an external observer, we at best can deconstruct the view and using the abilities of the hermeneutic observer, determine the intent of the builder of the networks.

To better understand the importance of an architecture we develop the concept of the historicity of architectures based upon the work of Kuhn and ten that of McLuhan. Kuhn begins his thesis of how scientific revolutions occur by the introduction of the concept of paradigms. He defines these as (see Kuhn p. 175);

"...the term paradigm is used in two different senses. On the one hand, it stands for the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community. On the other, it denotes one sort of element in that constellation, the concrete puzzle-solutions which, employed as models or examples, can replace explicit rules as a basis for the remaining puzzles of normal science, the first sense of the term, call it sociological."

The concept of a paradigm is in essence the collection of current technologies that we have at hand for the network and the ways we put these elements together. New paradigms result from new technologies. New technologies allow for the placing of the elements together in new ways. Kuhn, then goes on to demonstrate that the world view, that is how we view ourselves and our environment is based upon the our acceptance of these paradigms, as either collections of techniques and technologies or as collections of embodiments of these techniques and technologies in "examples". We then end to accept this as the way things are and should be. Then Kuhn argues, as the technologies change, changes in the paradigms do not occur in a continuous fashion but almost in quantum leaps. The new paradigms build and congeal until they burst forth with new world views. It is this model that we ague applies to the evolution of broadband. It is this philosophical view, almost Hegelian in form, that is essential in understanding the underlying and formative changes in paradigms that will change our world view.

As a second perspective of the impact of technology as a dominant driver, we can refer to McLuhan and his development of the concept of media. Drucker has referred to the presentation of McLuhan's doctoral thesis and McLuhan is quoted as follows (See Drucker, p. 250):

"Movable type, rather than Petrarch, Copernicus, or Columbus was the creator of the modern world view.. "Did I hear you right," asked one of the professors as McLuhan had finished reading, "that you think printing influenced the courses the universities taught and the role of the

university, altogether?" "No, sir, " said McLuhan, "it did not influence; printing determined both, indeed, printing determined henceforth what was going to be considered knowledge."

This concept later evolved into the medium being the message. In our context it is the fact that both Kuhn and McLuhan recognized, albeit in differing fields and in differing ways, that fundamental changes in technology and technique, call it paradigm or the medium, will change the world view, also the message. It is the importance of understanding the change in the technology, its function and evaluate the possible change that this will have in the world view. It will be argued, that much of the thinking in the current broadband areas, NREN in particular, is based upon outmoded techniques and structures, and that a differing world view will evolve.

Take, for example, the use of twisted pair, pairs of copper wire, to transport telephone traffic. For years it was implicitly assumed that this transport medium was limited to 4,000 Hz of bandwidth, that necessary for an adequate quality voice signal. Specifically the world view was that of a voice network that was to be used for voice traffic only. Ten years ago, this was a true limitation, since the transmission was forcefully limited to 4,000 Hz by inductive loads or coils on the telephone lines, assuring that you could do no more than the 4,000 Hz of bandwidth. Then, there was a short period in the mid-1980s, when Local Area network manufacturers found that you could transmit 1.544 Mbps over the common twisted pair, and that data was viable in what was assumed to be a voice only medium. What had been almost religiously believed to be a limit was found to be untrue. Then with the introduction of digital switches, the old "inductive loads" were returned with the switch now limiting the data to 4 KHz or 64 K samples per second. The world view of a voice only network took hold again, but this time in the context of a data rate limitation, rather than a bandwidth limitation. In the early 90's there is another attempted break out of the world view and to put 100 Mbps on twisted pair, so called FDDI circuits. Again, due to the limitations on the part of the network as a voice dominated system, the world view keeps this high data rate capability on the customer's premise only, and not the network.

We describe this transport world view evolution. Here we indicate the two dimensions of information transport, bandwidth and data rate. The designer of the transport facility may limit the data rate by selection of signaling format or delimit bandwidth by filtering. Twisted pair actually has a bandwidth-data rate profile. It encompasses a large capability of either providing bandwidth or data rates to the user. The two limiting world views are indicated as two solid lines, one at 4,000 Hz and one at 64 Kbps. Both are voice only world views. We can readily see, that with optical fiber superimposed the same issue of architecture dominated by world view may result. In the fiber case, the result may be a segmenting of the architecture along selected data rate lines, again formed by the voice world view.

Thus, architecture can be defined as the conceptual embodiment of a world view, using the commonly understood set of nonstructural elements, based upon the available set of technologies.

For example, Gothic architecture was a reflection of the ultimate salvation in God in the afterlife, in a building having a roof, walls, floors, and windows, and made of stone and glass. Romantic architecture was, in contrast, a celebration of man, using the same elements, but some employing a few more building materials. The impact of the differences in world view are self-evident in the embodiments of the architecture. (See the discussions on the impact of world view on architecture in Wolfe. In addition see the cultural or world view impact on the Gothic architectures in Jantzen and in Toy.)

Let us consider a second example of the impact of world view on architecture, specifically the difference between the ISDN architecture and the architecture embodied in Local Area Networks, LANs. ISDN is an architecture consistent with a voice dominated, hierarchical world view of single points of control. LANs are architectures of world views that reflect both end user self-empowerment and the environment of a data driven utility. This evolution in thought is critical to understand the impact of world view. The LAN is an embodiment of empowerment of the individual view, developed in the context of the 1960's and 1970's. The LAN concept, originating at such locations as XEROX PARC, was driven by the developers needs to enable and empower the end user with computing capabilities heretofore unavailable.

Out of this view came the LAN architecture of a fully distributed system, using a coaxial transport mechanism to do nothing more than provide bandwidth. The transport mechanism is a broad enabler. The actual implementation of the details is done at the users terminal in hardware and software. This is in sharp contrast to ISDN, where the ISDN central switch does the enabling. In ISDN, bandwidth is not provided, rather it is a voice based data rate, 64 Kbps or multiple thereof. Consider this contrast in terms of how cable TV companies provided voice communications in the early 1980's. Both Cox and Warner, using variations on LAN technology, delivered a voice, video, and data service over the coaxial transport medium, by empowering the end users terminal, not by regimenting the transport network.

Technology also plays a very pivotal role in telecommunications. Alfred Kahn (1971, p 300), indicates that in the pre-divestiture period of the Bell System, the arguments for the needs of both vertical integration and need for monopoly control were based on technology. Specifically, there was a contention made by the Bell System that a single point of control to the network was essential. Also, it was argued that an adequate scale economy was attained only through a single monopoly. Indeed, given the state of technology of that time, the argument may have held. For in point, the loaded copper transmission capabilities allowed only limited transport, namely one voice channel per twisted pair. However, as we shall demonstrate, the underlying technology has

provided a dramatic change in the underlying system. Functions now provided by the network, may be more efficiently provided by intelligent Customer Premise Equipment (CPE). The question to be posed is; what is the role of the network, and how do we provide the dimensions of creative freedom to allow these new roles to evolve? To effectively approach this problem, we must first develop a canonical structure of a network.

Before continuing, we will define in a more structure fashion the paradigm, the world view, the architecture and the technology base. First, Kuhn defines world view as:

" An entire constellation of beliefs, techniques, and so shared by the members of the community."

And he further defines the paradigm as, first:

"One sort of element in that constellation, the concrete puzzle-solutions which employed as models or examples can replace explicit rules as a basis for the solution of the remaining puzzles of normal science."

and then second as:

" Paradigms are what the members of the scientific community share and conversely a scientific community consists of men who share a paradigm."

We however, take from these definitions and place them into the context that we have developed in this section.

Definition: A paradigm is a unique and defining experiment or demonstration that in and of itself crystallizes a concept as representative and descriptive of a broader class of similar ideas.

The classic paradigm was that of Watson and Crick in developing the structure of the DNA molecule. The defining moment was best described in their paper in Nature that off-handedly alludes to the DNA molecule having the properties of self-replication and thus containing the genetic information for all life. This defining experiment then led to the massive changes in Botany, Biology, Medicine and even computer science. The paradigm is the rite of initiation for any believer in the new religion. The paradigm for Christianity was the resurrection of Jesus, and that of Judaism was the pact of Abraham with the Lord.

To understand all other elements of deconstructionist thought, it is first necessary to find and identify the defining moment, or the paradigm.

Definition: The world view is the collection of all beliefs that follow from a consistent application of the paradigm in reflecting life and its existence.

The world view of the DNA paradigm is genetic engineering, of the Univac paradigm, is the central processor, of the Church in the Middle Ages is salvation after death. The world view is more than a single statement, it is a collection of beliefs. All the beliefs are predicated on the unique and singular interpretation of the paradigm. Until Martin Luther, the Church in Rome was in control. Luther, through the printing press and the availability of the bible to all challenged and changed this. The new paradigm in this case was the printed bible in vernacular. The world view was that of Protestantism.

All too often we see the world view in parts. Even if we see the world view in total we may fail to see the paradigm. The religious organizations typically define and deify their paradigms. Jesus at the mount, Mohammed and the angel, Joe Smith and the angel, May Baker Eddy and divine revelation, and even the Buddha and his revelations. They are the paradigms. In religion it is critical to have the paradigms unchallenged. They are thus constructed.

Definition: The technology base is defined as the set of all technologies available to demonstrate the paradigm and to implement the representation of the world view.

We define technology in its broadest sense. We may have a paradigm of a god who has shown themselves through the visions of a mushroom. The world view is that we must perform all of our important acts under this influence. The technology allows us to analyze the mushroom and determine based on our neuroscience knowledge that the effects are those of psilocybin, not of our god. That will challenge the paradigm. Another is the ability to measure the age of the earth and to carefully, through genetic analysis, determine the evolutionary pathways of all creatures. This may challenge the paradigm of the creationists. A third, and current one is the technology to determine that homosexuality is genetically based. This challenges the paradigm of choice in the homosexual community and thus attacks a power base. Technology can thus either empower the world view and express it or it can challenge and destroy the paradigm.

Definition: Architecture is defined as the conceptual embodiment of a world view, using the commonly understood set of constructural elements, based upon the available set of technologies.

Definition: The design is the current implementation of an architecture using the current technology base.

Definition: The process of deconstruction is the process whereby the current observer, by using the understanding of the technology base evolution, and by understanding and structuring the

current and prior architectures, can determine the base paradigm and thus infer all elements of the world view of the implementor of the architecture.

We shall use this process in the development of many multimedia systems. The design of a multimedia system must understand the paradigms of the users. This is inclusive of all of the user sets, namely the food chain elements that we discussed before. To deconstruct, we must understand and identify the paradigm. Without that we cannot hope to achieve the converasationality that we seek.

Finally, we add the definition of an infrastructure, recapturing what we developed in the last section.

Definition: An infrastructure is a shareable, common, enabling, enduring, resource, that has scale in its design, and is sustainable by an existing market, and is the physical embodiment of and underlying architecture.

2.7 QUESTIONS

We proceed with the effort of developing a deconstruction theory by developing a set of key questions that ultimately relate to our understanding of object or entities. It will be through this understanding that we come to grasp the underlying structures of multimedia.

1. What is an object?

An object is and entity in a multimedia environment that is actionable. It is a single or compound collection of multimedia elements, combined from one or several users, that creates a transaction. Actionable implies that the objects can be used by others or shared amongst others and the sharing or use results in human actions or responses. The concept of actionable objects is at the heart of multimedia communications.

2. Do Objects exits?

Objects exists through the transactions that they create. Does a conversation exist? Words are created, sounds are made, and a consensus is achieved. Possibly the objects used in the conversations are perceived differently by each of the participants, and possible the consensus is perceived differently, even if it is placed in writing and agreed to by all. The true sign of the objects existing is their ability to create a transaction at some point, namely a change element.

3. What role does the human play in the expression of the characteristics of an object?

Does a tumor such as a lymphoma exist? The patient presents with a symptom of a swelling in the neck. The physician palpates and remembers the past 99% of the cases were lymphomas, by definition, the pathologist reads the slide and sees the telltale cell pathologies. The object of the lymphoma exists. The existence is in the consensus of the individuals. The ultimate existence is in the life or death of the patient

4. How does a human deal with objects?

The human deals with the set of objects through their totality of senses. The object may be visual, all too frequently what we are trained to respond to in today's culture, it may be a sound, a touch, a smell, or any of the other sensual combinations. The human rationalizes the sense objects and creates a communal sense of agreement with others with regard to a collection of such objects. The human deals with multimedia communications objects in a conversational and communal fashion. Unlike a book, or a video game, which is an internalized experience, a multimedia communications based object is externalizable, and allows for the displaced conversationality described earlier.

5. Why does a human deal with an object?

The human deals with the objects through their senses and with the interaction of other humans. It is a consensus-conflict resolution process.

6. Why does a human react to the presence of an object?

The human reacts to the object by transacting an event. The multimedia objects are essentially demands on the individuals senses that will not go unanswered. It is an active medium that demands a reaction and response. Unlike television which may be internalized and has responses that are displaced in their cultural context, multimedia communications is in the context of a real time communal dialog.

7. "What is communications?"

Communications in the context of multimedia is a conversational transaction. It allows multiple individuals, humans or otherwise to interact, converse, and through that conversation to transact. Communications is the movement of information between individuals, movement that entails the interaction of as many of the sense as is possible.

8. "What is conversation?"

Conversation is the sharing of senses, the establishment of consensus or the resolution of conflict, with the end result be the transaction. A conversation requires two or more humans, a common shared medium, co-temporaneous or otherwise, and a shared set of beliefs, that allow for a common understanding of the signs used on the conversation. The semiotician would look at the conversation in terms of the external signs. The deconstructionist would look at the underlying pre-beliefs of the communications. The hermeneuticist would try to understand the motives of the messengers in the conversation. In the case of multimedia communications we are using signs that are externalities of ideas that we wish to transmit from one individual to another. The end result of this transmission of ideas is a transaction, a change in the state of the receiver or even the sender of the ideas.

Multimedia communications is about conversations. It is about using all of the constructs available to communicate information that results in that desired or otherwise transaction. The challenge is to determine what elements are necessary for the state change. The further challenge is to determine if that change really occurred.

9. "How do humans interact with information?"

Humans interact with information by responding and creating a transaction. Information allows the change of knowledge state in the individual. If I am provided with a new piece of information, I change my understanding of something that I had not fully understood before. It may clarify it or it may dramatically alter it. If there is no change, then we argue that there was no information.

10. "What is a representation?"

A representation is an embodiment of a collection of information elements. In a multimedia communications environment we are almost always dealing with representations and never the actual entity. For example, in a radiological environment, we look at the MRI of the brain of a patient with multiple sclerosis. We see in the brain the white spots that represent the demyelination of the brain cells, the evidence that there is a sclerotic event. If we were the surgeon, we would most likely see nothing, if the pathologist, we would see the demyelination on a different scale by staining the cells and see the loss of myelin. All are representations of the underlying disease process. The reality is as best defined a genetic and viral induced process that is the humans immune system rejecting the self. This is the long winded way of defining the entity via the process as compared to defining the entity via the representation. The representation defines an entity that may be defined in no other way.

11. "Does there exist an abstraction that can be shared by two people that allows common understanding?"

The essence of multimedia communications is the delivery of abstractions. An abstraction is a representation of an entity delimited by the technological resources available in the multimedia communications environment. A full representation may require all the senses, or a breath of the senses that exceeds what is generally available to the current technology. The abstraction is the mapping of the representation onto the technology base. Sharing abstractions may allow two or more people to attain a common understanding if there is adequate information conveyed in the abstraction. For example, two physicians may see a chest x ray and determine that the patient has a carcinoma of the left lung. The spot is clear and non-disseminated, and based upon mutual shared prior experience it is clear what the abstraction is saying. If one of the viewers is the patient who is not medically trained, the abstraction would not convey such information. Thus information exchange requires a shared abstraction as well as a shared experience base with that set of abstractions. Expecting that abstractions per se convey information is not viable.

12. "Does a representation exist independent of the observer."

This is the dualism between forma and substance, between essence and existence. Is there such a thing as a rose, or is there a set of observable that when clustered together reach the consensus in most people that this is a rose. The body of plant Systematics faces this problem at all times. The answer is that there is no such thing as an abstract representation. Information and determination, the transaction, take place with a set of noisy observations that have been characterized and categorized.

13. "Does a representation change in its meaning when observed by more than one observer."

This is the same as asking if consensus changes if there are more or less participants in the process. Will a jury meet a different decision if there are different jurors. The answer is possibly. Consensus is convergent but there is no abstract convergent point.

Why are these questions important and how do they relate to multimedia communications. Multimedia communications consists of communicating objects that relate to the human in only one fashion; through a transaction or through the empowerment to act. When designing a system, we empower the users to act. They must do so through the use of objects, the manipulation of the objects and the gaining of a consensus.

4.3 Answers?

The questions posed in the last section and the discussion concerning them clearly indicate that a certain set of long standing issues are still at the heart at how we view ourselves as creatures. The

main driving issue has been the mid and body dualism. As we creatures that think with and act with our minds, or are our bodies the totality of ourselves. To quote from Winograd and Flores:

"...mind-body dualism...rests on several taken-for-granted assumptions:

1. We are inhabitants of a real world made up of objects bearing properties. Our actions take place in that world.

2. There are objective facts about the world that do not depend on the interpretation or even presence of any person.

3. Perception is a process by which facts about the world are registered in our thoughts and feelings.

4. Thoughts and intentions about action can somehow cause physical motion of our bodies."

This concept of the ideal form and the ideal an achievable entity is as old as Plato and Aristotle. The concept of the ideal form, as a Platonist would state, is that there is a true idea of a daylily. It is an abstraction that is the daylily, and what we see as humans is a mere shadow of its true form. To the multimedia communications, we then ask how does a Platonist communicate, namely, does he try to use the abstraction that closely matches the form? Copleston speaks on this with regard to Plato:

" I would point out that the essence of Plato's doctrine of Forms and Ideas is simply this: that the universal concept is not an abstract form devoid of objective content or references, but that to each true universal concept there corresponds an objective reality."

Continuing he states further:

"In the Republic it is assumed that whatever a plurality of individuals have a common name, they have also a corresponding idea of form. This is the universal, the common nature or quality which is grasped in the concept."

It is the attempt to describe the "nature" or essence of things and to use this as a means to communicate that is the basis of many of our problems in design. An example is the compression of speech or video. We compress to avoid the need for more bandwidth. We compress also because we believe that by doing so we get to the essence of it. We do so in a Shannoesque fashion, assuming that there is an essence of bits, minimal as they may be. This extension is best described by Popper:

" I use the name methodological essentialism to characterize the view, held by Plato and many of his followers, that the task of pure knowledge or "science" to discover and to describe the true nature of things; their hidden reality or essence. ...All these methodological essentialists also agreed with Plato in holding that these essences may be discovered and discerned with the help of intellectual intuition. A description of the essence of the thing they called the "essence"."

An extreme position to this essence approach is the positivist approach expressed by Ayer when describing the early work of Wittgenstein.

"..the main theses of the Tractatus can be easily summarized. The world is said to be totally of facts which themselves consist in the existence of what are called.. atomic facts.. or states of affairs. The states of affairs consist of simple objects, each of which can be named. The names can be significantly combined in ways that express elementary propositions. Each proposition is logically independent of all its fellows. They are all positive and each of them depicts a possible state of affair which constitutes its sense....The fact that they are logically independent means that in order to give a complete account of reality one has to say which of them is true or false."

The development of multimedia is the development of new metaphors. MacCormac best describes this change that metaphor can take:

" Metaphor can be described as a process in two senses: (1) as a cognitive process by which new concepts are expressed and suggested, and (2) as a cultural process by which language itself changes...epiphors are metaphors that express more than they suggest..diaphors suggest more than they express."

He goes on to state:

"Generations of students who have passed through introductory philosophy courses in colleges and universities have come to believe in the division between the mind and nature. The rise of cognitive psychology in opposition to behaviorism, which denied the existence of the mind, fins comfort in the philosophical efforts to build a foundation for knowledge. The account that I have presented of metaphor as a cognitive process presumes the existence of the mind existing as a deeper level of explanation that of semantics and surface language."

The essence of the Heidegger philosophy as relates to multimedia design has been best described by Winograd and Flores:

"We...present...a...discussion of Heidegger's philosophy,...

(1) Our implicit beliefs and assumptions cannot be all made explicit.

(2) Practical understanding is more fundamental than detached theoretical understanding.

(3) We do not relate to things primarily through having representations of them.

(4) Meaning is fundamentally social and cannot be reduced to the meaning-giving activity of individual subjects."

The final element of Heidegger's approach is the breaking down effort of providing information in a way in which it is broken down or handled by the user.

"... Heidegger's ...insistence that objects and properties are not inherent in the world, but arise only in an event of breaking-down in which they become present-at-hand...In sum, Heidegger insists that it is meaningless to talk about the existence of objects and their properties in the absence of concernful activity with its potential for breaking-down."

The latter comment on Heidegger is the essence of multimedia communications. The breaking down if the basis of a transaction, of a change in state of the human or humans in the conversation. We must then take into account the impact of this new medium. It is an impact with many dimensions of consequences. To quote McLuhan:

"The personal and social consequences of any new medium result from the new scale that is introduced into our affairs by each extension of ourselves or by any new technology."

And to further the quote:

"The message of any medium or technology is the change of scale or pace or pattern that it introduces with human affairs."

The multimedia revolution is on the scale of all other revolutions. It will be a revolution if and only if it does to the multimedia word what the printing press did to the written word. To quote McLuhan:

"The French Revolution, as per de Tocqueville, was a result of the homogenizing nature of the printed word."

Homogenizing means making it accessible to all. Making it accessible means making it actionable, and actionable at a distance. The actionable at a distance, and the ability to have the

throwness in the medium and the integration with the message is what will make for the revolution.

2.8 HERMENEUTICS

Deconstruction is the process whereby the current reader attempts to place themselves in the context of the writer, both used in generic terms, and determine what the message was that was meant to be sent, relative to the context of it being sent. It in many ways is a Bayesian analysis of a human communications process. Hermeneutics is complementary to this effort. Hermeneutics, named after the god Hermes, is in essence the attempt to understand the environment of the singer and writer of the songs, but to sing them as clearly and faithfully as one can.

2.8.1 Hermeneutic Principles

Hermeneutics is originally imbedded in the interpretation of texts. In our analysis the "text" is the broadened entity of the multimedia environment. The "text" must be expended into the context of the Messenger as carrier of the Message. Thus the need for a hermeneutic understanding is to focus on the messenger and what does the messenger bring to us about the message. In very practical terms, therefore, the same hermeneutic arguments that allowed us to address how to interpret the bible, are and will be at the heart of how do we interpret an x-ray and blood smear in a multimedia environment. The question posed is that does the multimedia environment land through the changed medium ad different message since it is communicated by a different messenger. Or, do we try to ensure that he messenger is kept intact. Another analogy is does the bible change when one understands it from televangelists rather than reading it. Or historically, does the bible change when each person reads it as compared to having it preached from the pulpit. The latter change led to the reformation and the end of the hegemony of Catholicism in Europe.

The hermeneutical school is a contrast to the positivists who argue that one can obtain objective knowledge. To the hermeneutic student, all knowledge is "interpretation". It is in this noisy channel that we try to obtain information. It is thus upon this noisy information that we ultimately act. The multimedia architect must take this into account in the designs of their systems. The channel is inaccurate and furthermore the messenger may actually be devious. The try of hermeneutics attempts to address the issue of devious messengers in the context of texts. We argue that the same questions and approaches are essential for the development of "multimedia text" messages.

The issue relating to texts is developed in Warnke. She states the problem in a historical context:

"Questions of interpretation had been raised earlier, in particular in the Reformations Challenge to the catholic reign of the bible. Did an understanding of Scripture require a prior acceptance of

the precepts of the Catholic faith or could it be understood on its own?...Schleiermacher significantly expanded the scope(asking) how many could be comprehended, what methods would permit an objective understandings of texts and utterances of any kind...Dilthey even asked broader questions: what were the methods that would permit an objective reading of any kind, including actions, social practices, norms and values? How could the understanding of meaning be raised to the same methodological clarity characterized in the natural sciences? How could it find as solid a basis for methodological progress?"

Gadamer is the most recent and articulate espouser of the hermeneutic school. He has evolved and matured the hermeneutic approach from one of literal translation to exposition. To quote Warnke describing the evolution in hermeneutics in Gadamer;

"The Bible is assumed to have a normative authority for everyone and the task of the hermeneutic understanding is therefore simply to help transmit the content of its normative claims."

Simply put, Hermeneutic in this context states that the deconstuctionist approach may be use in a relative setting, but in the context of a normative setting such as the law of God, we us either hermeneutic approach to seek the "truth" or normative facts. The difference is best stated in terms of selecting a justice of the Supreme Court. Judge Bork is the classic hermeneutic seeking the letter of the constitution. Justice Douglas is the decosntructioninst trying to take Madison's words and placing them in a current time frame.

Gadamer takes the hermeneutic goal of positivism and objective answers and introduces the subjective. This is characterized again by Warnke:

"Hermeneutics, as Gadamer conceives of it, then, is no longer to be seen as a discourse on methods of the "objective" understanding as it was for the hermeneutic tradition of Schleirmacher and Dilthey. It no longer seeks to formulate a set of interpretative rules; rather, in referring to his analysis as "philosophical hermeneutics", Gadamer turns to an account of the possibility of understanding in general, conditions that in his view undermine faith in the ideas of both method and objectivity. Understanding is therefore rooted in prejudice and the way in which we understand it is thoroughly conditioned by the past of by what Gadamer calls "effective history"."

From our perspective, as designers of multimedia systems, we have a cultural environment that we are working in. We have an environment with a history, a past, a culture, and a noisiness that makes event objective transmission of information a transformation of information. The hermeneutic channel challenge is to model the channel, to deconstruct its structure, to generate the optimal processing filters for the complex messages that are to be transmitted.

From a historical perspective, the hermeneutic problem for the multimedia designer is the same as the random noise problem was for the designer of signal detection systems of the 1940's and 1950's. As we had indicated before, it was the work of Shannon who determined that information was essential the elimination of uncertainty. Shannon's teacher, and in many ways mentor, was Norbert Wiener. It was to Wiener, who had both mathematical and philosophical raining, that the development of the concepts of the detection of signals in noise is credited. Wiener conceived of the use of the correlation and auto correlation functions. He introduced the history of an random process in a structured for that has led to information theory and the processing of signals in computers and communications. It is the same construct that we are trying to develop that is of utmost importance. The mathematical theory may not be in place, but the philosophical constructs to develop them must first be worked through.

Warnke goes on in terms of hermeneutic development:

"Hermeneutics thus has a largely pedagogical task: it is supposed to exhibit the truth that inheres in a given claim so that its audience can understand and learn from it. As hermeneutics develops, however, attention is redirected from the understanding of truth context of a text and towards the understanding of the intentions. The aim of understanding is no longer seen as knowledge of die Sache- a substantive knowledge of claims to truth or normative authority. It is seen rather as insight into the historical and biographical circumstances behind their expression. Understanding becomes genetic: what were the conditions under which the agents acted, spoke or wrote as they did?"

This leads to a focus on the circumstances qua written word. Deconstruction focuses on the meaning qua circumstances. We must enter into the hermeneutic thought process from the perspective of multimedia communications because the displacements of Gadamer and Haberamus are physical and temporal, but the displacements of multimedia communications are electronic and cultural. The nature of displaced understanding is the same in both cases. The issues, we argue, are also isomorphic.

The complementarity of approaches can be related to the complementarity of ideas in cultural context. For example, Bloom recounts the issue in the context of the French;

" Descartes and Pascal are national authors, and they tell the French people what their alternatives are, and afford a peculiar and powerful perspective on life's perennial problems...On my last trip to France I heard a waiter call one of his fellow waiters a "Cartesian"...Descartes and Pascal represent a choice between reason and revelation, science and piety."

These two authors are also choices between two types of certainty; the divine and the mind. In reality, as we have seen with Gadamer, certainty of any kind has its limits. The positivist school still argues for the existence of absolute certainty. The hermeneutic school of Gadamer eschews such certainty. At best we can interpret. At worst, the interpretation is a reflection of our own past, history, biases and intents. Even in medicine, the process of diagnosis is one of noisy interpretation. Certainty may exist in a pathology slide that portrays without doubt a malignancy. The outcome of that diagnosis may still have some uncertainty.

Namely, there is a philosophical underpinning in our general line of communications. It is this basis that will lead to understanding or cacophony. Gadamer goes on to define his focus in hermeneutics as follows;

" The task of philosophical hermeneutics, therefore, is ontological rather than methodological. It seeks to throw light on the fundamental conditions that underlie the phenomenon of understanding in all of its modes."

Gadamer further stresses the importance of language in this process:

" Language is the fundamental mode of operation of our being-in-the-world and the all-embracing form of the constitution of the world."

The issue of language being the form of history, the carrier of the formation, and the medium of the message, and thus being the message itself, has implications to the hermenuticist. One of these is the change in the nature of the media for the transmission of the e message. Weizenbaum notes:

" The computer has thus begun to be an instrument for the destruction of history. For when a society "legitimates" only those data that are "in the standard format" and "that can easily told to the machine" then history, memory itself, is annihilated. The New York Times has already begun to build a database of current events...from which historians will make inferences as to what really happened."

The conversion of words in the twelfth century, with the rediscover of the Greeks and their thoughts, which arguable led to the enlightenment of the twelfth century itself, was a result of the translation, and not simple transliteration of the Greek texts. As Illiach and Sanders remark:

" The Greek work was not to be turned into Latin verbum pro verbo. Instead, the meaning was to be detached from the words of one language and made to reappear in another; content, stripped of its form, was to be preserved. Theories about translation changed very little - translation was described as an attempt to divulge the secrets of one language into another- until the

hermeneutics of the 1950s. Only then did the study of translation as applied linguistic theory become separated from literary theory."

Ironically, the twelfth century Latin of the early universities, such as those in Paris, was a Latin ready for the expansion of the new technology of Guttenberg. It was now a language whose form was prepared for text. Again Illich and Sanders remark;

" Division into words first came into common use in the seventh century. It happened at the northern frontiers of the known world, where Celtic "ignoramuses" had to prepare for the priesthood and needed to be taught Latin. Division of words was thus introduced as a means of teaching Latin to barbarians as a foreign language."

The very structure of language had made a transition into a form that would allow it to be further tempered by a technology and thus be transformed into the new medium. The bilateral change of the understanding of the environment of the author and the tempering of the environment of the reader then leads to the full hermeneutic context. As Winograd and Flores remark:

" .. Gadamer takes the act of interpretation as primary, understanding it as an interaction between horizon provided by the text and the horizon that the interpreter brings to it. Gadamer insists that every reading or hearing of a text constitutes an act of giving meaning to it through interpretation."

Thus the hermeneutics of Gadamer is evolutionary from the revelation of the underlying eternal truth, to the evolving interpretation.

2.8.2 Hermeneutic Methodology Applied

In this section we take the philosophical theory developed in the prior section and address it to the problems of the multimedia methodology. We develop the design principles in hermeneutic observations. The process is to state the observation and then to reflect on the appropriate design principle.

1. What we observe is a reflection of what is there filtered through the understanding of what we think should be there.

2. Conversations have history, and the history is often unknown, and if known may not be aware to the conversant.

3. Absolute "truth" does not necessarily exist in a conversation. Consensus may converge but convergence is not to truth. The convergence of consensus is not consistent.

4. Tradition, authority, and history are integral elements in filtering understanding. The issue of authority is an integration of visions of the presenter and the information presented. Choice of words and language to create a "text" are the basis of power in the relationship.

5. Communications is a sociological interpretive process that seeks to attain consensus in resolving an ambiguity. Actions are the result of that consensus becoming an agreed to common state amongst the community in the process.

2.8.3 Communications at the Conversational Layer: Hermeneutics Applied

Having established the design rules and provided the understanding of the hermeneutic elements, we can now take these and place them in the context of multimedia communications. The issue is one of establishing conversationality. Coversationality is as we have defined the ability of one or more humans to enter into a dialog with a system and themselves relating to elements of information in the system adequate enough to eventually create a transaction.

The OSI layered communications architecture has evolved to manage and support the distributed communications environment across error prone communications channels. It is presented in detail in either Tannenbaum or Stallings. A great deal of effort has been spent on developing and implementing protocols to support these channel requirements. Layer 7 provides for the applications interface and generally support such applications as file, mail and directory. The requirements of a multimedia environment are best met by focusing on layer 5, the session layer whose overall function is to ensure the end to end integrity of the applications that are being supported.

Some authors (See Couloris and Dollimore or Mullender) indicate that the session function is merely to support virtual connections between pairs of processes. Mullender specifically deals with the session function in the context of the inter-process communications (IPC). In the context of the multimedia object requirements of the previous section, we can further extend the concept of the session service to provide for IPC functionality at the applications layer and specifically with regards to multimedia applications and their imbedded objects.

The services provided by the session layer fall into four categories:

1. Dialog Management: This function provides all of the users with the ability to control, on a local basis as well as global basis, the overall interaction in the session. Specifically, dialog management determines the protocol of who talks when and how this control of talking is passed from one user to another.

2. Activity Management: An activity can be defined as the totality of sequences of events that may be within a session or may encompass several sessions. From the applications

perspective, the application can define a sequence of events called an activity and the session service will ensure that it will monitor and report back if the activity is completed or if it has been aborted that such is the fact.

3. For example, in a medical application, we can define an activity called "diagnosis" and it may consist of a multiple set of session between several consulting physicians. We define a beginning of the activity when the patient arrives for the first visit and the end when the primary physician writes the diagnosis. The session service will be responsible for ensuring that all patients have a "diagnosis".

4. Synchronization: We have seen that at the heart of a multimedia system is a multimedia data object. Each of the objects has its own synchronization or timing requirements and more importantly, a compound object has the orchestration requirement. The session service of synchronization must then ensure that the end to end timing between users and objects is maintained throughout.

5. Event Management: The monitoring of performance, isolation of problems, and restoration of service is a key element of the session service. Full end to end network management requires not only the management of transport and subnetwork, but requires that across all seven OSI layers, that overall end to and management be maintained (See McGarty and Ball).

The servers are conceptually at a level above the transport level. We typically view the transport servers as communicating distributed processes that are locally resident in each of the transmitting entities. This then begs the question as to where does one place the session servers. Are they local and fully distributed, can they be centralized, and if so what is their relationship to the Transport servers. Before answering these questions, let us first review how the session services are accessed and how they are communicated.

Session services are accessed by the higher layer protocols by invoking session service primitives. These primitives can invoke a dialog function such as Token_Give. The application may make the call to the S_SAP and this request may be answered. There are typically four steps in such a request, and these are listed in Stallings who shows that the requests are made of the session server by entity one and are responded to by entity two. The model does no however say where the session server is nor even if it is a single centralized server, a shared distributed server, or a fully distributed server per entity design. We shall discuss some of the advantages of these architectural advantages as we develop the synchronization service.

5.3.1 Dialogue Management

Dialog management concerns the control of the end user session interaction. Specifically, who has permission to speak and when, who can pass the control and how is that implemented. In this section we shall describe the environment for the dialogue management function and develop several possible options for implementing this function.

Dialog management requires that each of the virtual users have a token or have access to a baton in order to seize control of the session. In the course of a typical session, the two virtual users fist establish the initial subsession that becomes the first part of the session. The addition or binding of other virtual users through subsessions to the session allows for the growth of the session. The baton or token may be a visible entity that is handed from one to the other or it may be hidden in the construct of the applications.

Consider the session level service called dialogue. The service can be implemented in four possible schemes. These schemes are:

(1) Hierarchical: In this scheme there is a single leader to the session and the leader starts as the creator of the session. The baton to control the session can be passed upon request from one user to another, while full control remains with the session leader. The session leader may relinquish control to another user upon request and only after the leader decides to do so. The leader passes the baton from users to user based upon a first come first serve basis. It is assumed that each users may issue a request to receive the baton, and that any requests that clash in time are rejected and the user must retransmit. There transmit protocol uses a random delay to reduce the probability of repeated clashing. The leader always acknowledges the receipt of the request as well as a measure of the delay expected until the baton is passed.

(2) Round Robin: In this scheme, the baton is passed sequentially from one user to another. Each user may hold the baton for up to Tbat sec and then must pass the baton. When the baton is held, this user controls the dialogue in the session.

(3) Priority: In this case, all of the users have a priority level defined as Pk(t), where k is the user number and t is the time. We let the priority be;

$$Pk(t) = Rk(t) + Tk(t) + Dk(t)$$

Here R is the rank of the k th user, T is the time since the last transmission and D is the data in the buffer. We assume that some appropriate normalization has occurred with this measure.

Every Tcheck seconds, each users, in sequence sends out a small pulse to all other users, on a broadcast basis, and tells them their current priority. Each user calculates the difference between theirs and all the others. User k calculates a threshold number, TRk, which is;

$$TRk = \max |Pk(T)-Pj(T)|$$

If TRk > 0, then user k transmits its packets for Tsend seconds.

(4) Random Access: Each user has a control buffer that indicates who has control of the session, namely who has the baton. The session is broken up into segment Tsess in length, with Treq seconds being relegated to a baton ownership selection period and Tsess-Treq being left for the session operation. During Treq , all of the users transmit a request packet that is captured by all of the other users buffers. Treq is broken into two parts, Tsend and Teval. These requests are broadcast in Tsend.

Now after the sent messages are received, one of two things can happen, the message is received or it collides with another message and is garbled. If the message is garbled, the buffer is not loaded and is left empty. If it is filled, then each buffer during Teval sequentially broadcasts its contents and all of the users listen to the broadcast and record the counts, Nk where Nk is the number of votes for user k in that call period.

The choice of baton control is then;

$$\text{Choose user k if } Nk = \max j \mid Nj \mid$$

else restart Treq again.

5.3.2 Activity Management

Activity management looks at the session as an ongoing activity that users may come and go to. This services provides an ability to easily add, delete and terminate the entire session.

An activity in the terms of the session is a total bounded event that can be compartmentalized in such a way that other events may be locked in suspension until that event is complete. Activity management is in the session layer a function similar to transaction management in a transaction processing system. It allows for the definition of demarcation points that permit suspension of activities in other areas until the activity managed transaction is complete. Activity management can also be developed to manage a set of events that can be combined into a single compound event.

There are several characteristics that are part of activity management:

1. Activity Definition: This allows for the defining of an activity as composed of several dialogue. It allows for the defining of the activity as a key element of a single session or even to expand over several sessions. Activity definition is the process of informing the session server of the beginning and end parts of an activity and in providing the session server with an identifiable name for the activity.

2. Activity Integrity Management: Activities are integral elements of action that cannot be segmented. The activity management system must ensure that once an activity is defined and initiated, hat no other activity that could interfere with the existing one is allowed to function.

3. Activity Isolation: The ability to provide integrity is one part of managing the activity. Another is the ability to isolate the activity from all others in the session. An activity must be uniquely separable from all other activities, and this separation in terms of all of its elements must be maintained throughout its process.

4. Activity Destruction: All activities must be destroyed at some point. This is a standard characterization.

There are several sets of activities that are definable in a multimedia environment. These are as follows:

1. Compound Multimedia Object Transfer

2. SubSession Management

3. Dialog Control

The algorithms to perform the activity management functions are developable consistent with the OSI standards. There are no significant special development necessary.

5.3.3 Synchronization Management

Synchronization is a session service that ensures that the overall temporal, spatial and logical structure of multimedia objects are retained. In this case we have a source generating a set of Voice (VO), video (VI), and Image (IM) data objects that are part of a session. These objects are simple objects that combined together form a compound multimedia object. The object is part of an overall application process that is communicating with other processes at other locations.

These locations are now to receive this compound object as show with the internal timing retained intact and the absolute offset timing as shown for each of the other two users.

In this example, the synchronization function provided by the session server to the applications processes at the separate locations is to ensure both the relative and absolute timing of the objects. The location of the functionality can be centralized or distributed. Let us first see what the overall timing problem is. Consider a simple SMO synchronization problem. The network than transmits the packets and they arrive either in order or out of order at the second point. The session server must then ensure that there is a mechanism for the proper reordering of the packets at the receiving end of the transmission.

Let us consider what can happen in this simple example.

First, if the BMO of the SMO is very lengthy, then as we packetize the message, we must reassemble it in sequence for presentation. Let us assume that the BMO is an image of 100 Mbits. Then let us assume that the packet network has a packet delay that will be low if there is no traffic and grows as traffic increases. Now let us assume that the network provides 500 bit packets transmitting at 50 Mbps.

Second, let us note that there are 200,000 packets necessary to transmit the data. Each packet takes 10 microseconds to transmit. If we assume that there is a load delay of 5 microseconds per packet, then the total transmit time goes from 2 to 3 seconds.

We can also do the same with a compound object. In this case, we take the CMO and note that it is composed of SMOs. The SMOs must then be time interleaved over the transmission path to ensure their relative timing. It is the function of the session service to do this. The application merely passes the CMO and its header information as a request to the session server to ensure the relative timing is maintained.

Here we have a CMO entering the network, knowing that the session server at Server 1 must not only do the appropriate interleaving but it must also communicate with the other servers (in this case K and N) to ensure that de-interleaving is accomplished. We show the session servers communicating with the network through the T_SAP and that in turn takes care of the packetizing. However, we also show that the session server, 1 and N, communicate in an out of band fashion, using some inter process communications (IPC) scheme, to ensure that the relative actions are all synchronized amongst each other. We can now envision how the architecture for this can be accomplished. There are two schemes:

Centralized: It assumes that each application (A) has a local client (CL). The application communicates with the local client (CL) to request the session service. The session server is

centrally located and communicates with the application locally by means of a client at each location. This is a fully configured client server architecture and can employ many existing techniques for distributed processing (See Mullender or Coulouris et al).

Distributed: In this case we have a set of applications, and cluster several applications per session server. We again user local clients to communicate between the session server and the applications. The clients then provide local clusters of communications and the session servers allow for faster response and better cost efficiency. However, we have introduced a demand for a fully distributed environment for the session managers to work in a distributed operating system environment. As a further extreme, we could eliminate the clients altogether by attaching a session server per applications and allow for the distributed processing on a full scale.

The major functions of the session server in its synch mode are:

To bind together simple objects into compound objects as requested by the application.

To provide intra object synchronization to ensure that all timing within each object is met.

To orchestrate amongst objects to provide inter object timing.

To minimize delay, slippage, between simple objects.

To minimize delay, latency, between different users.

To effect these requirements, we have developed and implemented a scheme that is based on a paradigm of the phased locked loop found in communications (See McGarty and Treves, McGarty). Here we have a distributed session server architecture receiving a CMO from an application. The session server passes the message over several paths to multiple users. On a reverse path, each server passes information on the relative and absolute timing of the CMO as it is received using the session services primitives found in the OSI model. Generally for segmented BMOs this is a simple problem but with streamed BMOs this becomes a real time synchronization problem.

5.3.4 Event Management

Event Management deals with the overall end to end management of the session. It is more typically viewed as a higher level network management tool for multimedia communications. In the current sate this service is merely a reporting mechanism. Although ISO has expanded the network management functionality of the seven layers, most of the functionality is still that of

event reporting. In this section we discuss how that can be expended for the multimedia environment.

Event management at the session layer provides for the in band signaling of the performance of the various elements along the route in the session path as well as reporting on the status of the session server and the session clients. We note that each entity in the session path, which is limited to all involved clients and all involved servers provide in band information on the status of the session. In particular the in band elements report on the following:

Queue size at each client and server. The queue size can be determined on an element by element basis.

Element transit and waiting time. For each element involved in a session, the time it takes to transit the entire block as well as the time that the block has been in transit can be provided.

Session synchronization errors can be reported in this data slot. These errors can be compared to lower level errors and thus can be used as part of the overall network management schema.

The structure of the event management system has been effectively demonstrated. It is represented as a header imbedded in the transit of every data block. We can generate specific event management blocks that are also event driven and not data transit driven. These are generated by direct transmission of such blocks as overhead devoid of data content.

2.9 SEMIOTICS

The theory of semiotics is the theory of signs. Umberto Eco, academician, leader in the field of semiotics, and author, has written several novels, one being the "Sign of the Rose". The novel is about the fourteenth century and about a murder in a monastery. The hero uses the signs that are left behind, the clues in common parlance, to play two roles. The first is the common role of the clues to the murder mystery and the second as the signs to the change of the old guard to the new. It is at the steps of the Renaissance and the signs of most importance are the books of old knowledge. Thus Eco combines all elements of the deconstructionist, with those of the hermenuticist and those of the semioticist.

2.9.1 A Definition

Eco defines semiosis as;

" Semiosis is the process by which empirical subjects communicate, communications processes being made possible by the organization of significant systems. Empirical subjects, from a semiotic point of view, can only be defined and isolated as manifestations of this double

(systematic and processual) aspect of semiosis. ...Semiotics treats subjects of semiosic acts in the same way: either they can be defined in terms of semiotic structures or-from this point of view-they do not exist at all."

The elements of semiosis comprises signals and signs. These Eco defines as:

" We are now in a position to recognize the difference between a signal and a sign...as sign may be an expression system ordered to a content, but could also be a physical system without any...purpose. A signal can be a stimulus that does not mean anything...a sign is always an element of an expression plane... correlated to elements of a content plane."

Specifically, Eco combines these as:

"Semiotics is mainly concerned with signs as social forces."

Semiosis is based upon the extension of Pierces concept of pragmatism. The Pierce view of pragmatism is presented by Copleston;

" Pragmatism, as Pierce conceives it,...is a method of reflection having for its purpose to render ideas clear. ... Pierce divides logic into three main parts, the first of which is speculative grammar...concerned with the ...meaningfulness of signs. A sign...stands for an object to someone whom it arouses a more developed sign..the relation of significance, or the semiotic function of signs, is for Pierce...a relation between sign, object and interpretant."

The issue of semiosis as a carrier of information is also embedded in the issue of a code. Eco defines this as follows:

"..a code establishes the correlation of an expression plane with a content plane...a sign-function establishes the correlation of an abstract element of the expression system with an abstract element of the content system...a code establishes general types.. producing the rule which generates concrete tokens...both ...represent ...the semiotic correlation and with which semiotics is not concerned.."

The most descriptive distinction of what semiosis does in comparison to hermeneutics is best described by Eco:

"In order to understand the history of Christian theology, it is not necessary to know whether a specific actual phenomenon corresponds to the word, transubstantiation, ... it is necessary to know what cultural unit ... corresponds to the content of that word.

The semiotic object of a semantics is the content, not the referent, and the content has to be defined as a cultural unit.."

The totality of semiosis is its ability to combine its constructs with all elements of signals and signs.

"...semiotics has been provided with a paramount subject matter, semiosis. Semiosis is the process by which empirical subjects communicate, communications processes being made possible by the organization of significant systems. Empirical subjects, from a semiotic point of view, can only be defined and isolated as manifestations of this double... aspect of semiosis. ... Semiotics treats subjects of semiosic acts...either they can be defined in terms of semiotic structures of-- from this point of view -- they do not exist at all."

We can now develop some of the basis theory of semiotics as presented by Eco. We start with a set of definitions and then discuss some of the implications.

Expression Plane: This is the plane of understanding in which one creates meaning through and abstract expression. In physics, for example the expression plane may state the following:

This expression is the expression for the propagate of light or any other electromagnetic wave in free space. Another example is the composition of a gene;

ATTGTAAGCCGGATTTTC

This is a set of nucleotides that imply their complements on the opposite side of the DNA pair. This sequence means a great deal to a molecular biologist. They may imply blue eyes, green wings, or red flowers. Both of these items are in the expression plane.

Content Plane: The plane in which one creates meaning through concrete expressions expressed in a continuum of actions in the physical world. Simply put, it is everything that you can think about it..

Let us take the propagation equation. This may be the expression of a rainbow. The content plane is the rainbow and the full panoply of what it evokes. It is macro and micro in its expression. It is the emotion that it may evoke leading to poetry, the grandeur it mat evoke in nature and whatever else may do.

In the genetic case it is the fullness of the expression of the gene and its characteristics. It is the expression of what green eyes may have and the human emotions that they may evoke.

In the case of multimedia, we design and implement our systems in the expression plane. We express the in the content plane to ourselves as persons. The design problem will then be to connect the expression and content planes effectively.

Sign: An element of the expression plane correlated to an element of the content plane. It is a correlation recognized by the human.

Signal: A pertinent unit of a system that may be an expression system ordered to a content, or could be a physical system with no semiotic purpose. Units of transmission which can be computed quantitatively irrespective of their respective meaning.

Sign-Function: A relationship between system on E/C Planes, "Expression/Content".

Token: Types are generated by a code.

Token-Sign: Relation between units on E/C planes.

Type: Code generates types producing a rule which generates concrete tokens.

Code: Establishes general types, therefore producing the rule that generates tokens.

Connotative Relationships: A concatenation of the E/C plane such as shown below;

This example demonstrates the concatenation possible between the Expression and the Content and how content may be concatenated upward itself. Eco makes this example one step further when he states:

"Semiotics is mainly concerned with signs as social forces"

2.9.2 Semiotic Application

As we have defined the semiotic approach it deals with the signs or externalities of the process of communications. In the multimedia world these externalities are the elements themselves, the data objects, the storage objects and the actual human interaction with devices. All are signs.

Signs are means by which we relate from one plane upward to another. Signs in the multimedia world focus on relating from the image and the test and the voice segment to the determination of the diseased state. The design problem from the multimedia semiotician then is to go from expression to content. The design problem for the seomiotic architect is the opposite; from the content to the expression plane. How, does the semiotic multimedia architect ask, does one

ensure that the expression of the multimedia system contain adequate elements to reflect the process in reverse and ensure the commonalty of expression by the richness of the expressionist?

2.9.3 The Multimedia Database; A Semiotic Example

We shall use the multimedia design problem of the database as a starting point in the application of semiotic theory to the design of the system from an architectural perspective.

In a more standard computer communications environment, the data objects have significant structure and they are frequently integrated into a system wide data base management system that ensures the overall integrity of the data structures. In a multimedia environment, the data elements are more complex, taking the form of video, voice, text, images and may be real time in nature or can be gathered from a stored environment. More importantly, the separate data objects may combined into more complex forms so that the users may want to create new objects by concatenating several simpler objects into a complex whole. Thus we can conceive of a set of three objects composed of an image, a voice annotation and a pointer motion annotating the voice annotation. The combination of all three of these can also be viewed as a single identifiable multimedia object.

Before commencing on the issues of communications, it is necessary to understand the data objects that are to be communicated. We can consider a multimedia data object to be composed of several related multimedia data objects which are a voice segment, an image and a pointer movement (e.g. mouse movement). As we have just described, these can be combined into a more complex object. We call the initial objects Simple Multimedia Objects (SMOs) and the combination of several a Compound Multimedia Object (CMO). In general a multimedia communications process involves one or multiple SMOs and possibly several CMOs.

The SMO contains two headers that are to be defined and a long data sting. The data string we call a Basic Multimedia Object (BMO). There may be two types of BMOs. The first type we call a segmented BMO or SG:BMO. It has a definite length in data bits and may result from either a stored data record or from a generated record that has a natural data length such as a single image screen or text record.

The second type of BMO is a streamed BMO, ST:BMO. This BMO has an a priori undetermined duration. Thus it may be a real time voice or video segment.

A simple multimedia object, SMO, is a BMO with two additional fields; a Synchronization field (Synch) and a Decomposition field (Decomp). The Synch field details the inherent internal timing information relative to the BMO. For example it may contain the information on the sample rate, the sample density and the other internal temporal structure of the object. It will be a useful field in the overall end to end timing in the network.

The second field is called the Decomp field and it is used to characterize the logical and spatial structure of the data object. Thus it may contain the information on a text object as to where the paragraphs, sentences, or words are, or in an image object, where the parts of the image are located in the data field.

These fields are part of an overall architecture requirement finds it necessary to provide an "out-of-band" signaling scheme for the identification of object structure. The object structure is abstracted from the object itself and becomes an input element to the overall communications environment. Other schemes use in-band signaling which imbeds the signal information with the object in the data stream. This is generally an unacceptable approach for this type of environment.

When we combine these objects together we can create a compound multimedia object. A CMO has two headers, the Orchestration header and the Concatenation header. The Orchestration header describes the temporal relationship between the SMOs and ensures that they are not only individually synchronized but also they are jointly orchestrated. The orchestration concept has also been introduced by Nicolaou. In this paper we further extend the orchestration function beyond that of Nicolaou. The concatenation function provides a description of the logical and spatial relationships amongst the SMOs.

We can also expand the concept of a CMO as a data construct that is created and managed by multiple users at multiple locations. In this construct we have demonstrated that N users can create a CMO by entering multiple SMOs into the overall CMO structure.

The objectives of the communications system are thus focused on meeting the interaction between users who are communicating with CMOs. Specifically we must be able to perform the following tasks:

Allow any user to create an SMO and a CMO.

Allow any user or set of users to share, store, or modify a CMO.

Ensure that the user to user communications preserves the temporal, logical and spatial relationships between all CMOs at all users at all times.

Provide an environment to define, manage and monitor the overall activity.

Provide for an environment to monitor, manage and restore all services in the event of system failures or degradation.

We shall see in the next section that the session layer service address all of these requirements.

2.10 CONCLUSIONS AND OBSERVATIONS

We have introduced a new paradigm of viewing information and have developed a world view to place that construct in. The construct tries to relate the human interface problem and the structure of both information and knowledge of events into a single schema. We have addressed the issue of what is multimedia communications, what are the elements of the design process in this new medium, what are the design dictates of the total environment, and what rules for design can be abstracted from the change in the experience in interacting with media.

Whatever the ultimate design factors that may evolve, there are certain obvious design observations that can be made based upon the current understanding of the multimedia area. We must consider that in designing in a multimedia world we are doing so in a distributed service environment. It is not a manufacturing business where the product is assembled in a factory under the control of the owner and where quality can be carefully maintained and managed. It is a service assembled in the hands of a user in an environment that is neither predictable nor constant. Thus the design challenge is to incorporate the stochastic nature of the shared user environment.

To this end there are several design rule that have evolved in the process of understanding such systems. Specifically:

The next step should always be obvious.

Whenever designing a multimedia system, the presentation to the user, through whatever senses, should clearly indicate what the next step should be. Where should I type next, where can I place the pointer, which device do I use, what should I say? All too often designers let the user have the freedom to create. This results in ambiguity, frustration and visual, aural and tactile dissonance.

Form matches function.

What are we using the system for and why. What is the function and the form of the system should match the function that it has been designed for.

There should be consistent paradigms.

When the system is designed to edit images, the editing tools should be the same in all configurations. The access mechanisms, if they are on the left should always be on the left.

Execution should be smooth.

Tactile and visual dissonance are common factors in poor design. A smooth design should be such as to enhance the conversationality mode of the session.

The question should always be obvious.

State what you want. The statement should be clear and not allow for any secondary interpretation. If the question is complex, then it should be broken down into smaller segments and simpler questions. To paraphrase Wittgenstein, the essence of true understanding is the ability to pose the question in such a way that the answer is clearly yes or no.

The answer should always be obvious.

When answering a question, the answer should always be clear and obvious. Again if a complex answer is to be presented it too should be segmented.

There should be no ambiguity of expectations.

The users and the designer should have the same set of expectations for the deployment of the system. "I never thought they would do that with it!" is a common complaint. If all else fails, listen to the customer, user, etc.

If we can develop systems that follow these guidelines in all of their dimensions, then there should be a smooth road to transition in the multimedia area.

3 MULTIMEDIA ELEMENTS

This chapter discusses the different media elements that encompass the multimedia domain. Specifically we develop the detailed characterizations of the various media types, developing an understanding of the issues of how they map into communications and reproduction formats. One of the issues discussed is the manner in which the media are presently presented and the transition from what is frequently a mechanical process to and electronic process.

The two issues that we focus on in this chapter are those relating to the characterization of the overall source in a multimedia environment. Those two issues are the characterization and the specific sizing of the multimedia image in terms of bits. We know that the bit sizing is essential for the source characterization and the we shall latter be using it as an integral element of the system sizing and performance.

3.1 THE MEDIA CONSTRUCT

The main issue discussed in this chapter is how we represent the reality of the observed world in a form that can be stored permanently and processed at latter times. A simple example is how we take a picture of a particular person or place. The picture is merely a representation of the person or place and in many ways does not duplicate all the features of the person or place. We may then take that film medium and digitize it for storage in a random access memory to be post processed to either extract features or enhance the image. We can then readily ask the question of, "How well does the digitized image reflect reality?".

Thus the key element in understanding the use of media for the processing of information relates first to how we store the information on the different media and second how well the stored representation relates to the original reality. Media characterization starts with the following paradigm:

Reality exists and does so independently of the observer. The reality has an extent that can be projected onto an observer through many means. The representation of the abstract reality is unique at any one instant and this uniqueness can be represented onto each of the observers senses in a definable and measurable fashion.

Senses are the major inputs to the human for the purpose of transferring information about reality. The senses common to the human are sight, sound, touch, smell, and taste.

Observations are the results of reality projecting itself onto an observer through their senses. For example, there may be a representation of a reality in terms of the electromagnetic waves that is generates and there may be a model for the optical response of the human. The combination of this quantifiable reality, with this observation system can be further mixed with

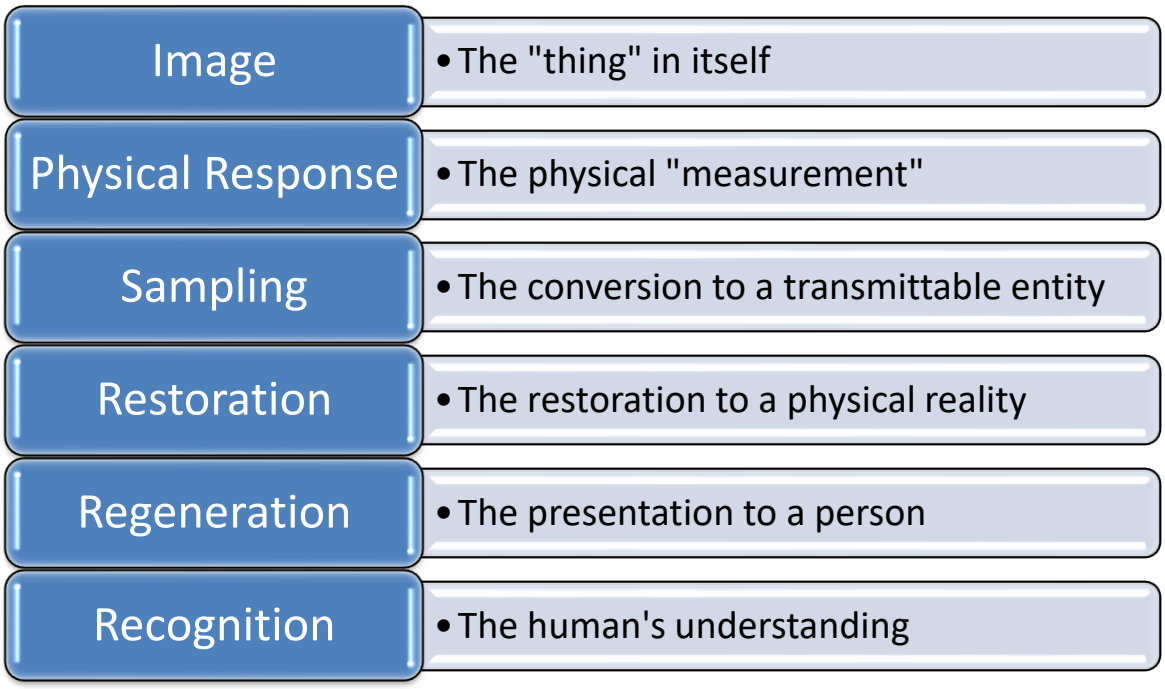
We depict this paradigm of the world that we have viewed with an exposition of the example that we have just discussed. In this chapter we shall assume that we deal only with the senses of sight and sound and that we shall discuss the different media that can be used to store the information in a specific form of representation. We further are concerned with the accuracy of that representation adequately matching the information presented from the reality. It should be noted that we nowhere deal with relating ourselves to the ultimate reality. This is a very Platonic approach to understanding our existence but the reader should not read any significant philosophical positions in this paradigm of a world view.

Medivim Characterization Effectiveness is a measure of the accuracy in which the representation matches the information. Since both of these are quantifiable we can analytically measure the effectiveness of the medium in term of the ability to reconstitute the original information bearing signals.

Thus the two major efforts in this chapter are the ability to understand how information is reduced to storage on different media and in turn to quantitatively determine the effectiveness of that medium in retaining a close proximity to the original information.

We define the information of this reality as the electromagnetic wavefront generated by the reality, projected along the visual senses. Let this information quantity be given by:

(2.1) E(x,y,z,t)



We now take a picture of this reality and generate a photoemulsion that can also generate an observation or an information based image of its own. This information image can also be represented by an electromagnetic field:

(2.2) El(x,y,z,t)

Now it is possible to measure the error induced between the emulsive image capture medium and the original electromagnetic information image. We define a measure ||E-El|| as some metric on a three dimensional space (see McGarty) that allows for the specification of errors. Thus it is quantitatively possible to measure the effectiveness of medium characterization.

Now we can further take the photographic image an digitize it. To digitize an image we must perform several tasks. These are typically:

Time sample the wavefront at t(k) seconds, where k+l to N.

For a single sampled wavefront, sample the wavefront in space at point xi and yj. Call the samples E(i,j,k).

Now quantize the samples of E(i,j,k) into quantized values called S(i,j,k), such that the quantization is close to the original signal.

Now let us assume that we are sampling the picture at the rate of thirty times a second, and that we have 1 thousand by one thousand element display, that is 1 million pixels, and that we quantize in one of 1024 levels (e.g. 10 bits) so that we are generating 300 million bits per second of data on this storage method.

We have now stored the data digitally and we wish to retrieve it. To do so we need a display that reproduces the image using the stored data files. Let us assume that the display is 512 by 512 in dimension and can respond to only 8 bits of phosphor response. The image generated and transmitted towards the human eye is the electromagnetic field E2(x,y,t). We can now describe how the system can effectively reproduce the image by the error:

(2.3) $||E-E2|| = ||E-G(E1)|| - ||E-G(F(E))||$

where G and F represent the transformations of the two media storage methods.

This simple example of using three different media raises the two points that we wish to present in this chapter. First, we must be able to characterize how primary reality is stored and transformed by the storage media, and second what the effect of this transformation is on the overall restoration of reality.

Media characterization starts with a specific reality. That reality, for our purposes, is delimited to the visual and oral areas. Specifically it is related to what we see and hear. Expanding the reality to the other three senses (smell, touch, and taste) have not been focused on in the present are of media communications, however, they surely will evolve as we come to better understand how to transfer information through these senses.

There are multiple types of information bearing media that are used. Several of the more common forms are still images, full motion video, and voice. In this section, we develop models for the characterization of these three types of media and provide a basis for the development of generalizations for many other types of media.

3.2 IMAGES

The still image is the most common form of display format for Pictorial information. The still image generally tires to capture some form of reality that is represented, for example by a scene in nature. That scene in nature may in this simplest form be a view of a tree on a hill with the clouds in the background.

The first question is how do we take this complex still image and store it in a finite amount of memory for latter retrieval and display. That problem had been answered over a hundred years ago with the development of the photograph, a physical/chemical storage medium. The

photograph is only an approximation of the true reality however. It uses a form of silver salts that are responsive to light falling on them and darken to different)degrees. The resolution of such images is that of the underlying silver nitrate solution and it generally is better than that of the human eye. Thus the viewer of one hundred year old photographs will not see any degradation of the image but will perceive a clear and well preserved representation of the scene. The key observation of this fact is that the eye has a key role in determining an acceptable degree of image resolution and with such forms of mechanical/chemical memories as 35 mm film, there is more than adequate resolution.

With the use of computers however and the needs of other media such as the newspapers, there are lesser degrees of resolution required in image display. This type of resolution is what many of us see in the print out of our computer terminals or in the daily newspaper. In the case of the computer printer the image is a composition of small black dots, the density of the dot proportional to the intensity of the image.

3.3 MEANS OF CHARACTERIZATION

We can now proceed to discuss the details of the characterization of the different forms of images. We shall take as the starting point, a reality of a fixed scene and from that scene develop the ways in which we first characterize achromatic (eg black and white) and then chromatic (eg color) still images.

Before beginning with the detailed discussions of the two major types of images; achromatic and chromatic, let us first present a brief discussion of the way the human eye responds to images.

The eye is composed of the following elements:

Cornea: This is the outer surface of the eye and generally encompasses the overall outer surface of the organ. or Iris: This is a circular and pigmented membrane behind the cornea that allows for the entry of light. It opens and closes in one of its parts, the pupil, upon response to the total intensity of the light incident on the eye.
The retina is the neurosensorial layer of the eye and its area is about 12J cm^2. It transforms the incoming light into electrical signals that are transmitted to the visual cortex

The anatomy of the retina shows five types of cells organized in layers. The furthest layer from the incoming light is that of photocells. There are two types of photocells: rods and cones. A normal eye contains about 130 million rods and 6.5 million cones. Rods and cones are different enough to be examined separately. Rods are sensitive to shapes and need low luminance (scotopic vision). In contrast, cones need daylight (photopic vision). They detect color and distinguish details.

Their distribution in the retina is highest in the vicinity of the optical axis of the eye. That is why a precise detail vision is obtained only when the eye "fix" them, in other words, when their image is formed at the fovea. In this region there are about 120 cones per degree which fix the visual resolution to 1 min of arc Photoreceptors are responsible for transforming the in-

The eye is composed of the following elements:

Cornea: This is the outer surface of the eye and generally encompasses the overall outer surface of the organ.

Iris: This is a circular and pigmented membrane behind the cornea that allows for the entry of light. It opens and closes in one of its parts, the pupil, upon response to the total intensity of the light incident on the eye.

Pupil: This is the specific opening at the center of the iris that responds to the light level.

Lens: This is element of the eye and is adjusted by muscle movement to allow for the positioning of images on the rear part of the eye.

Retina: This is the most sensitive part of the eye. It contains the sets of nerves that allow for the reception of the light and its transmittal to the brain for processing.

Optic Nerve: This is the connection point between the eye and the never pathway to the brain. within the retina are two major types of nerves cells; rods and cones. They are called such by their general nature since cones have a cone shaped head and the rods are generally round with no tapering toward their heads..

There are 130 million rod type cells in the typical human retina and there are about 6.5 million cone cells. The cone cells are distributed at about 120 cones per degree of arc which allows for a resolving capability of 1 min of arc (see Kunt at. al. p 175). Rods are the sensors for the shapes of objects whereas the cones are the sensors for color and light intensity. The eye responds to light over the range from 350 nm to 750 nm. The mechanism for this response is the photochemical responses of the molecules called carotenoids. These molecules are coiled tightly until impacted upon by photons. Then they rapidly uncoil and when recoiled release electrical energy that triggers the nerve cells to transmit their messages to the visual cortex. Thus vision is a complex optico-chemical-electrical process.

The resolving power of the eye is one of the more critical factors that we shall be focusing on in this chapter. Let us take a closer look at the 1 min of arc number and see how this effects types of images that we view.

Let d equal the distance between two images at the image plane and let A be the angular distance in degrees of arc between the two points.

$$d = r A$$

and we can substitute A of 1 min of arc to this equation. Recall that 1 min equals 1/60 degree, and one degree is (2 pi/360). Thus we have one minute is equal to 0.0003 radians. At one meter viewing distance we have a resolution of 0.3 mm. At 0.3 m or about one foot, we have a resolution of 0.1 mm. We can compare this to a typical television set of 1 foot in height. If there are 525 lines per scan then there are 1mm between the lines. If we sit 1 foot from the set we can resolve 0.1 mm or ten times the resolution of the set. However if we view it as 10 feet then we are at the resolving power of the human eye.

Thus we can see that the issue of better resolution of the human eye to such technologies as HDTV may be fundamentally specious in that unless we view such system at close range, we cannot hope to obtain an fundamental improvement.

In a fashion similar to the spatial resolving power of the eye we can examine the fundamental color perceiving power of the eye also. This has been discussed by Shepherd and others and will be left to the reference materials.

The eye has six major modalities that impact on how it is effective in multimedia communications. These modalities are:

Photosensitivity: This is the sensitivity of the eye to the intensity of the light. In particular it is important in the eyes ability to receive light from diffuse source that reflect light from nondirect elements. A typical example is the reception of light from movies theater screen which have marbled surfaces to create diffuse reflections. Without those reflections, we should have a specular reflection which would be visible to the viewer in the direct reflection direction.

Form Discrimination: This factor relates to the resolving power that we have discussed. The discriminating capability is dominated by the position of the cone neurons in the retina and thus limit the resolving capability of the eye.

Movement Discrimination: This is the ability of the eye to detect change. This factor is important in such areas as film and television. The eye has a time constant on the order on 0.1 second, such that any events that are shorter than that are not discriminated by the eye.

Binocular vision: This factor allows the human to discriminate depth in a spatial context. It is a function of the two eyes in resolving the differential responses on the retina and the ability of the visual cortex to process the images and resolve depth.

Polarized Light Discrimination: There is a limited ability of the eye to discriminate in the polarization of light.

Color Discrimination: The eyes of humanoids are one of few anthropoid eyes that can respond to color. The color response is viewed as an evolutionary response to the limitation of night vision in the species.

Thus all of these elements of the eye should be considered when designing and operating a multimedia system

3.3.1 Achromatic Images

Achromatic images are those that are either black or white, or even mixed with shades of gray. More specifically, achromatic images do not have any color. An achromatic image is generated by the response of the eye to electromagnetic waves that are reflected from a surface and received by the eye. If we have a surface that reflects light at all frequencies equally well, and that total reflection occurs at the surface, then that surface is perceived as white. On the other hand if we illuminate a surface with a source of light that has all frequencies and the surface reflection characteristics are such that it reflects all incident light with zero reflectance then that element appears black.

Let us consider this black/white occurrence i more detail. First we must remember that we see objects only as a result of the light that they reflect, or in some cases emit. Let us focus on reflected light only. Let us assume that the incident light is characterized by a wave of the form E(x,y,z,t) for all points in space. Further let us assume that the wave can be characterized by a Fourier transform called E(x,y,z,f), where:

$$E(x,y,z,t)=\int E(x,y,z,f)\exp(-2\,ft)\,df$$

If we further calculate the amplitude of the wavefront at each frequency, or equivalently look at the spectrum of the signal, we see that for white incident light we have:

$$||E(x,y,z,f)|| = \text{Const}$$

Which is constant for all frequencies or wavelengths.

Now let us assume that the reflected wave is given by ER(x,y,z,t) and that we can show that if the surface has a reflectance at frequency f of, R(f), then;

$ER(x,y,z,t) = \int E(x, y, z, f)\ R(f)\ \exp(-2\ ft)df$

Now if we assume that R(f) is independent of frequency, that is:

R(f) = R We then have;

ER(x,y,z,t) = R E(x,y,z,t)

Then, clearly, if the incident wave is white light, we have for:

ER(x,y,z,t) = E(x,y,z,t) (R=l and is white)

or;

ER(x,y,z,t) = 0 (R=0 and is black)

When we discuss chromatic surface we shall see that this changes since R is now frequency dependent.

When we consider black and white images, we are generally considering that the white light is reflected from the surface, independent of frequency. This means that the surface has no specific reflection characteristic. At most, the surface may absorb the light uniformly, thus resulting in a graying of the light. The grayness in this case is a result of the loss of light intensity. Thus what we see from a surface is the level of intensity of white light. If we let I be the light intensity, then using standard electromagnetic principles, we have:

$I = | |E_R| |^2$ Using 2.8 we can write;

$I - R_2 | |E| |$

Let us now consider the effects of light on film. Photographic film is simply a gelatin base interposed with silver halide crystals. These crystals are composed of silver halides of the form such as AgBr, eg silver bromide. When light shines upon these crystals, the reaction occur:

(2.13) Br> Br + e

)2.14) Ag+ + e > Ag

That is, silver is reduced into the gelatin substance. Upon development (see Jacobson et al) the developer enhances the result of silver generation in the crystals that have been exposed to the light. Thus in a region exposed to light, a large set of silver crystals are produced and in those areas not exposed that are small i number. During development, the remaining silver halides are washed away, leaving only silver in the developed plate, acting as a darkening agent to transmitted light. This result in the standard negative that we have seen in many cases. We should remember that the density of photons is directly proportional to the light intensity so that we can expand the analysis to what we have discussed in the development Of (2.11).

Now the resolution of film depends upon how well we can generate a suspension of silver halide crystals, and also how effectively we can develop the film. Such effects as diffusion of the light and scattering in the film plate may lead to the less than crisp photos we may have often seen. In the most effective film resolution that we may have experienced, we look at the standard x-ray film. Here the same film process is use but the incident radiation is at the x-ray frequencies or wavelengths. The x-ray is used because of its ability to penetrate the hard and soft tissues of the body. The net result is a film plate that is exposed in proportion to the transmitted, rather than reflected energy.

The x-ray is a typical example of the continuous gray scale image that is what we commonly refer to as the black and white photo. Continuous gray scale represents the fines gradation of exposed silver crystals that can be obtain on photographic file. The resolution of the typical x-ray film is determined by the density of the silver crystals on the film plate and the ability of the developer to keep the diffusion of the developer process to a minimum.

If we were to digitize the x-ray film, on the basis of human ability to resolve the resolution of the image in terms of line resolution and to resolve the image in terms of shades of darkness, a display of 20" by 20", having 2,000 by 2,000 pixels, or 4 Million pixels per 400 sg in is needed. This equals a resolution of 10,000 pixels per sq inch. In addition the display per pixel must resolve 4096 levels of gray, which is determined to be the maximum level of resolution of shades of darkness of human resolution. Thus a figure of merit for an x-ray is 48 million bits per x-ray, based upon 4 million pixels, at 12 bits per pixel.

This x-ray example is the simplest example that demonstrates the need for combining the concept of the original image of the human body, the ability to express it in terms of a digitized continuous tone image, and the corresponding use of human responsibility as the measure of comparing the closeness of the digitized image.

This x-ray example has led us into the discussion of gray scale techniques. A gray scale image is typically a digitized version of a black and white image. Remember that the black and white

image is in reality a quasi-continuous display of the intensity either reflected or transmitted by an achromatic sources, from or through a surface. The gray scale representation is the means of taking that quasi-continuous image and reducing it to a bounded digital representation. We say that the original image is quasi-continuous because of the fact that the silver generated on the film developer plate is microscopic in level and for the most part is as continuous as we can get in nature. In its ultimate limit all such figures are truly discrete.

Let us continue with our gray scale representation. There are two elements that are of concern in the digital representation of a gray scale image. They are:

Pixel Resolution: This issue relate to how many pixel per square inch are necessary to be unresolvable by the human eye. Based upon both psychological and psychological studies, we find that (see the reference by Kunt et al) the retina can resolve images set apart by 1 min or arc. Anything smaller than that and the image falls on the same cell. Thus if we have a viewing distance of about three feet from the screen, we can use the following:

d = r ang

where d is the distance of the pixels on the screen, r is the 3 feet, and ang is the resolutions in radians. Substituting the parameters we find that d approximately equals 0.01 in or 100 per inch. This yields 10,000 pixels per square inch. This is exactly what we displayed earlier.

Lightness resolution: In a similar fashion, we have found that the retina can respond to about 4000 shades of light intensity and this yields about 12 bit of shades of gray that can be used.

Thus for gray scale coding, the human eye needs about 10,00 pixels per square inch with 12 bits per pixel. This means that the typical, continuous gray scale digitization image requires 120,000 bits per square inch at a 3 foot observation distance.

Let us examine the gray scale quantization a bit further. Let I represent the light intensity at a particular pixel. Let us start with a single bit gray scale quantization. That is at the pixel we use either black or white. In particular, let 10 be the threshold and say:

P = P0 if I < 10

and

P = P1 if I >= 10

Then code this pixel into 0 or 1 according to this level of quantization. The choice of 10 is not arbitrary, as we shall show shortly. Let us expand this to two bits.

(2.x) P = P0ifI < 10

P = PIif10 <=I <= II

P = P2ifII <=I <= 12

P = P3if12 <=I

We can now code this into two bits. Again the choice of II and 12 are not arbitrary. We can expand this into 12 bits, or 4096 levels by selecting 10 through 14095. Again the choice is not arbitrary. We find this through many measurements of human responses and it is also related to the ability of the retina to respond to stimuli as well as the brains ability to distinguish these elements.

Now we can view gray scale coding in this form as the ultimate in the achromatic coding or images from a continuous form. At the other extreme is the mapping of images onto dot matrix printers that are common in the PC world. A dot matrix printer can produce about 50 dots per inch or 2500 dots per square inch. Again compare this to the eyes ability to del with 10,000 per square inch (see Durrett p. 223). However, the dot matrix printer prints a dot, black, or no color, white. Thus it is only one bit per pixel. This is simply 2500 bits of information per square inch. This compares to the eyes ability to deal with 120,000 bits per square inch. This is a 50 to 1 reduction in information.

We can expand this capability with a technique called digital halftoning. Quite simply, halftoning is a technique that combines the limited capability of having only one bit per pixel into sets of such pixels that, to the eye, look like more than just one bit, they looks like gray scale. It is based upon the fact that collections of bits have reflectivity parameters that appear to the eye as gray scale. The work by Ulichney shows how this gray scale factor can be obtained. In addition, there are many gray scale algorithms that permit the optimization of the visual response. These are presented in both Ulichney and the work of Foley and VanDam. Let us use a simple example to show how halftoning is developed.

We have a set with no dots, one with one, two, three and four. Thus with 2 dots we have 2^2 or 4 possible entries. Each of these entries have less and less reflectance or in turn have a perceived shade of grayness. Let us assume that we are taking a gray scale image with 100 dots per inch and compressing it down to a halftone image with only 50 dots per inch. Let us further assume

that we wish to have the halftone image approximate the gray scale image. We shall proceed as follows:

Use the halftone basis set of the 2x2 matrix and note that since halftone uses only 50 dots per inch, that two dots will equal 4 dots of the gray scale image. Specifically if we define:

D GS = the gray scale image density in dots per inch

D HT = the halftone image density in dots per inch

N HT = the size of the half tone cell in dots

G HT - the number of gray scale levels in a halftone set

then we can show that we can compress the gray scale into the halftone.

The gray scale compression can be given as follows. The number of gray scale dots compressed is N GS which equals:

N GS = N HT * (D GS/ D HT)

But in the N GS cells there are L GS gray scale levels per cell which equals L GSB, gray scale bits. Remember that this is at most 12 bits or 4096 levels. Thus there are a total of:

TB GS = L GSB * (N GS * N GS)

TB GS gray scale bits in the gray scale cells. For example in the case of 4x4 gray scale and 12 bits per cell we have 16x12 or 192 bits per gray scale cell.

However, in the halftone cell, we have 2x2 or 4 cells with only one bit per cell. This yields an equivalent

TB HT = L HTB * (N HT * N HT)

In this case there are 4 bits. Thus the halftone has reduced the bits from 192 to 4, or a bit reduction of 48 times. We obviously loose a significant amount in this process, but the resulting image is often quite acceptable.

To select the proper halftone bit pattern, we must look at the several sets of gray scale dots or cells and average the gray scale level, and then further quantize it to fit the gray scale levels that have been allowed in the halftone quantization.

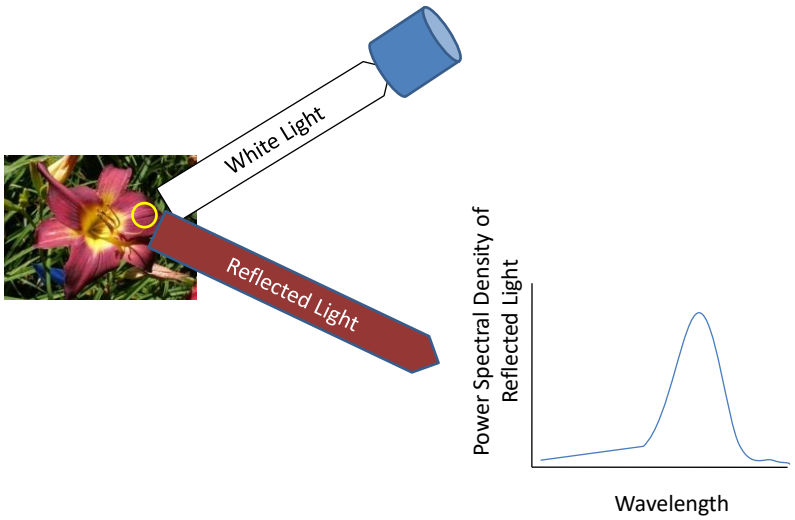
This process of halftone production is a well-accepted practice in the field of newspaper print and in many other printing systems. It is the most obvious choice in use of dot matrix printers.

There are several problems that arise with the halftone imaging however. They typically result from the use of a set of periodic basis functions to generate the pseudo-gray scale images. This can be eliminated by the use of entering randomness in the selection of the dot location and in the assurance of no long run sequences of the same pattern even if the average gray scale density remains relatively constant.

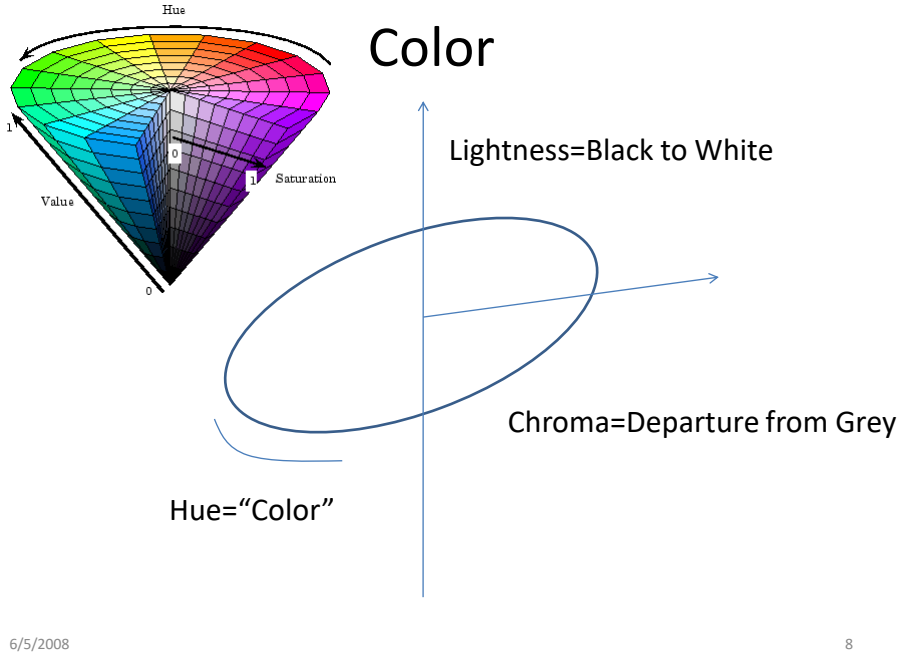
3.4 CLASSIC COLOR THEORY

Color can be viewed from several perspectives and the two focused on herein are the human eye and the measure power spectrum. The human eye views color in a complex manner since the eye receives color stimuli via sets of sensors which are tuned to three possible visible frequencies, the classic red, green and blue.

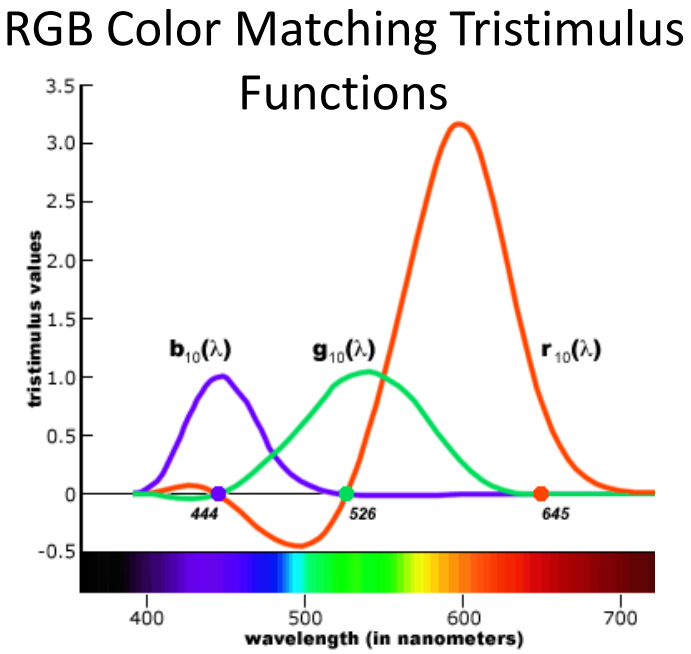
We see a flower as a certain color. There is "white" light shown upon the flower and the light is reflected from the petals and sepals and what we perceive is a red flower. This perception is a combination of two things; what part of the incident white light is reflected, and how our eyes process that reflected light. Thus color has two meanings for us. First color is nothing more than a reflected spectrum of electromagnetic waves in the optical frequency band. Second it is what we perceive as a human observer and in turn name as a color. The latter approach can and often is quite subjective. This latter approach is the basis of print color, paints, dyes, pigments and the like where the end point is the presentation of some artifact of a desired color. It is the former or first approach we seek to use, namely what is there independent of the observer, specifically the spectrum.



In the context of color as perception there is a collection of terms which should be understood. Hue is a synonym for a color. Red is thus a hue as is orange. The hues cover the visible spectrum. Then there is the lightness, ranging from whit to black. The third element is the chroma which is the departure of the hue from gray. We show these elements below.



The color school then takes three more steps. These are the CIE models for color. The first step is the Tristimulus models. The Tristimulus function is shown below. This is NOT a spectrum. In addition negative values mean more positive stimulus. These are also the result of extensive experimental modeling. The red, green and blue Tristimulus model as shown below characterizes three stimuli which affect the three receptors of color in the eye.

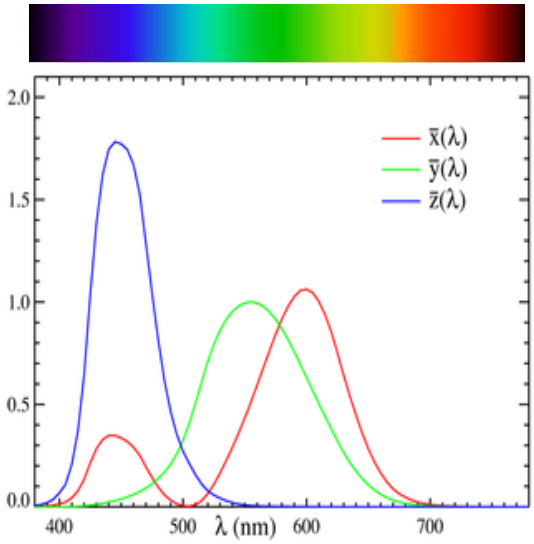


The above Tristimulus curve, called the color matching function, were experimentally determined through an experiment where a person looked at a test color centered at a wavelength as shown on the horizontal axis above and tried to match it by adjusting a red, blue and green lamp in the reference field. This could be accomplished for all regions except between 444 and 526 nm. In that region a red light had to be added to the test field to adjust the color to match. In effect the test color was changed by adding red. This adding of red is accounted for by the negative portion of the above curve[1].

Following this above model based upon experiment is the spectral approach called the standard observer consisting of the X Y Z model as shown below. They do effectively represent quasi spectral responses since they are all positive. It is possible to transform between the RGB and the XYZ formats.

[1] See Berns p. 49.

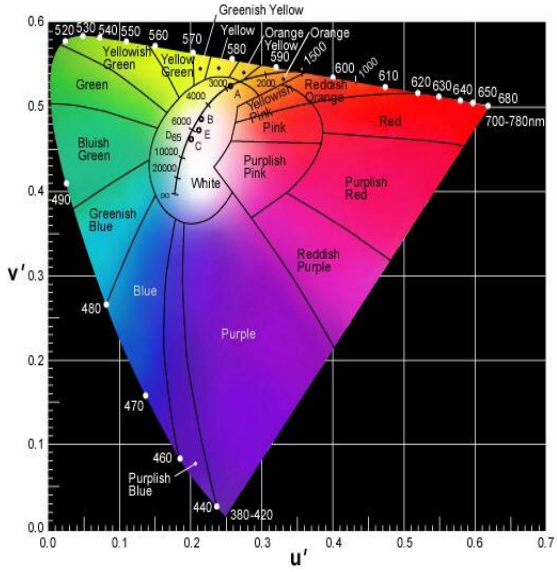
CIE Standard Observer Curves



6/7/2008 12

The CIE Chart is a manifestation of how the three stimuli above can be added together to create a broad set of colors, hues. It is possible to go from red, thru green and then to blue. One need only mix the three stimuli in the proper ratio. Then other hues can also be generated.

CIE Chart



6/5/2008 17

To understand this a bit better we analyze the RGB system first. We start with a source specified by intensity I dependent on wavelength. This allows us to define the following:

$I(\lambda) =$ *the spectrum of a specific sample*

Define

$$R = \int I(\lambda) r(\lambda) d\lambda$$
$$G = \int I(\lambda) \bar{g}(\lambda) d\lambda$$
$$B = \int I(\lambda) b(\lambda) d\lambda$$

Note, as we had stated, the color matching primaries show negative values because the negative was the way the CIE arbitrarily represented an excess positive contribution required to be added to a primary to achieve the desired spectrum response while keeping the elements normalized. Specifically:

$$\int \bar{r}(\lambda) d\lambda = \int \bar{g}(\lambda) d\lambda = \int \bar{b}(\lambda) d\lambda$$

Now this also implies the following are to be true:

$$r = \frac{R}{R+G+B}$$
$$g = \frac{G}{R+G+B}$$
$$b = \frac{B}{R+G+B}$$

and

$$r + g + b = 1$$

In a similar manner we can do the same for the XYZ system. This is done as follows:

$I(\lambda) =$ *the spectrum of a specific sample*

Define

$$X = \int I(\lambda) \bar{x}(\lambda) d\lambda$$
$$Y = \int I(\lambda) \bar{y}(\lambda) d\lambda$$
$$Z = \int I(\lambda) \bar{z}(\lambda) d\lambda$$

And as was the case for RGB we also have the normalizing factor. Not that it is this normalizing factor which assures our ability to deal with the triangular plot of color.

$$x = \frac{X}{X+Y+Z}$$
$$y = \frac{Y}{X+Y+Z}$$
$$z = \frac{Z}{X+Y+Z}$$
and
$$x+y+z=1$$

Finally there exists a set of transforms which allows one to convert from one to the other. This is shown below:

$$\begin{bmatrix} r \\ g \\ b \end{bmatrix} = A \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$
A such that sums of r,g,b and x,y,z are unitary

Therefore for any color C we can write it in one of the following two manners:

$$C = rR + bB + gG$$

$$C = xX + yY + zZ$$

Thus an x,y plane can be constructed such that any color can be characterized as a pair of coordinates (x,y). This is the CIE Chart which we have shown above. It must be noted that all of this analysis is predicated on how "we" see color and not in any context of how it is created or the underlying physics of color.

Now there are two other brief examples worth noting. First is the concept of additive colors, such as those we see when we add lights. This was the basis of what Newton did in his early experiments. By adding lights we can ultimately create white. We show that below.

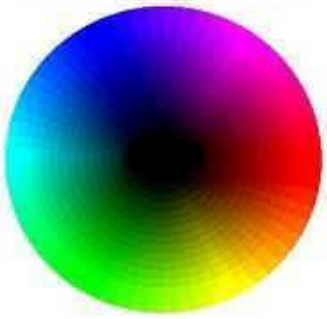
Additive Primary Colors

Additive Colors combine to form white. Traditionally adding lights is additive whereas adding colorants, pigments or dyes, is subtractive.

6/5/2008 19

The opposite is the subtraction of light, and this is the result of adding pigments of different colors together in an oil painting. If we were to add all the colors together then we obtain black and not white. This is subtractive, for we are in reality removing colors by the use of those pigments. In many ways this is the difference between water colors and oil paint. We show this subtractive result below:

Subtractive Primary Colors



Subtractive Colors form black. Subtractive mixing involves the removal, subtraction, of light from the mix. Removing all light ultimately results in black. Absorption only is called simple subtractive mixing whereas combining this with scattering is complex subtractive mixing.

6/5/2008 20

Notwithstanding the above detail and its use in many industrial processes, these methods used in classic colorimetry are methods that rely upon the human by necessity being part of the process. We when looking at plants, shall disregard the human.

3.5 SPECTRA AND MEASUREMENTS

The measurement of absorption spectra can be accomplished by a variety of means. We present here two methods; classic spectrometry and Fourier Transform Spectrometry (FTS) also called Fourier Transform Infrared Spectrometry, FTIS. However FTS can be applied to the optical bands as well.

The goals using these methods are as follows. First to determine the absorbance and extinction coefficients of the secondary products that are colorants. This means that solutions of purified anthocyanins, Peonidin for example, would be used and their absorbance and extinction coefficients determines for all wavelengths over the optical band[2]. This is accomplished for all targeted absorbents. Second, perform the same on all known colorants found in a target plant cell. This could include any secondary product or even proteins which have absorbent properties. Generally the other chemical elements react in an absorbent manner out of the optical band. Third, perform the analysis on target cells. Our approach is to perform this on a cell by cell basis thus requiring focused optical positioning.

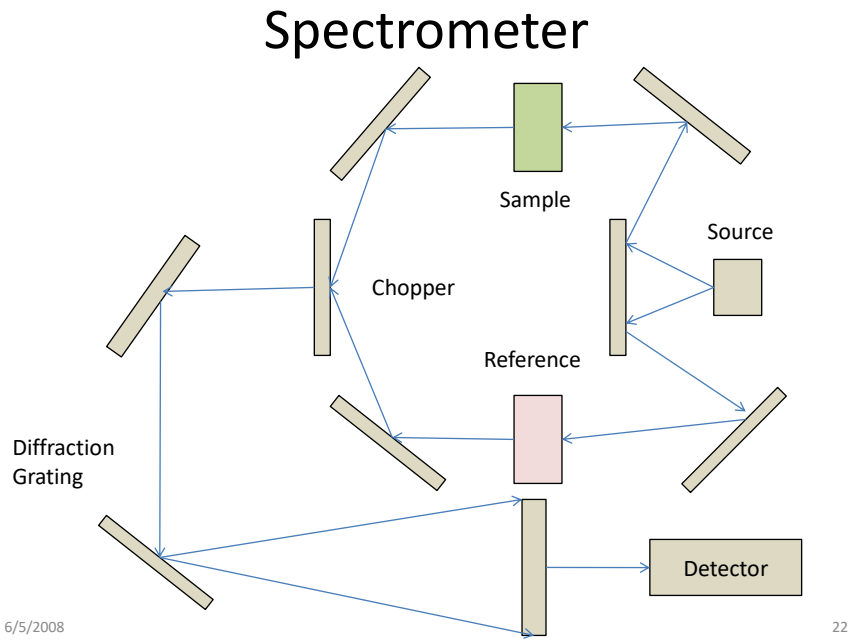
3.5.1 Classic Spectrometer

We first look at the classic spectrometer. It uses two paths for transmission, one through a cell with the target secondary and another cell without any secondaries or colorants. The second cell is a reference cell. The reason for this approach is to calculate the difference in absorption. The spectrometer is shown below. It functions as follows:

- Select a Reference and a Sample. Source Spectrum is to be determined using reference.
- Send light from source through both sample and reference. The source must be broadband wavelength. There will be no need to regulate amplitude across the band since a difference signal is obtained and the result will be expressed as a ratio.
- Chop the signals using electronic chopper so that half interval it is sample and other half it is reference. This can be accomplished with a time controlled electronic device or even a mechanical rotating wheel which can be synchronized to the measurement elements.
- Send to Diffraction Grating to spread out signals over visible spectrum.

[2] See Cantor and Schimmel, Part II, pp 380-388.

- Sample from one end of spectrum to the other by mechanically sampling the diffraction grating spread out. Remember that the diffraction grating act as a prism and spreads out the signal spatially over the optical band.
- Use the reference as the baseline and then measure the ratio or the difference of sample to reference and plot. This generally requires just a difference amplifier at the measuring point and synchronizing it with the chopping signal.
-



The spectrometer, as shown above, functions well for the determination of relative absorption. It is a long and sometimes cumbersome process because the screen in front of the detector is scanned slowly and this provides the signal used to ascertain the difference measurement. There is an issue of accuracy and precision in the collection of data and there is also the issue regarding the amount of light intensity requires. One should remember that as we spread the light out through the grating we see the spectrum now spatially but in so doing reduce the signal strength of each segment. The spectrometer has advantage and disadvantage in this configuration.

3.5.2 Fourier Transform Spectrometer

The FTS is a more recent embodiment of a spectrometer and it eliminates many of the accuracy and precision issue of the classic spectrometer. In many ways it may be viewed as a mini-CAT scanner in that it collects data which is the Fourier Transform of the desire waveform, namely the absorption spectrum.

The FTS works as follows:

1. A target sample is placed in front of a detector. The detector is a broadband detector and it provides at its output the integral of all the power entering across the optical spectrum. The optical spectrum will be the target spectrum of interest so we delimit the detector to that. We also assume we know the detector response and that this can be adjusted for by means of signal pre-emphasis. This means that the detector works as follows:

$$P = \int S(f)df$$

Here P is the total power and S(f) is the power spectral density of the combined signal. We will look at that in some detail in a moment. Now we assume that the detector may itself have a spectral sensitivity given by H(f). Thus what we really receive if we do not pre-process is:

$$P = \int S(f)H(f)df$$

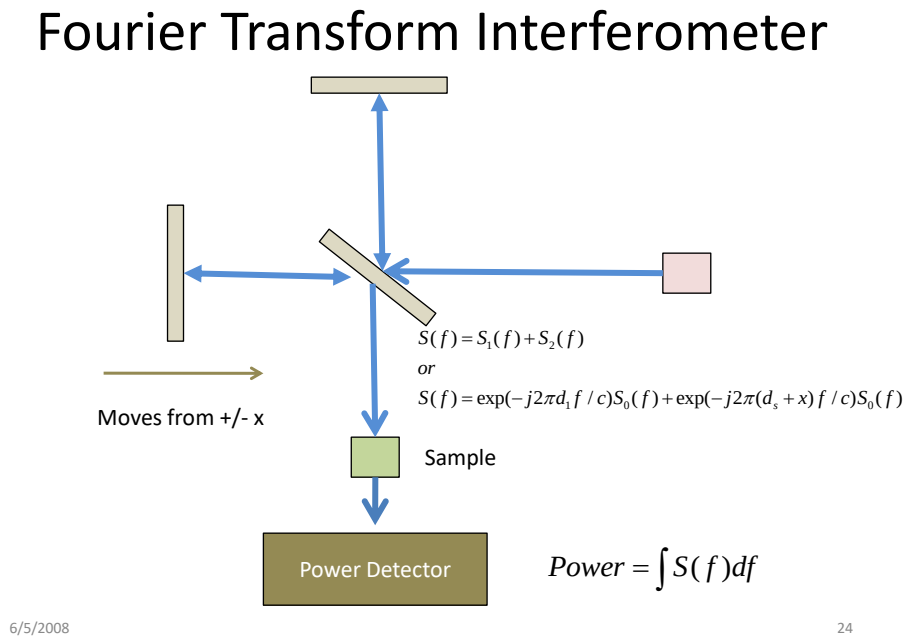
Which may bias out answer? The way to avoid this is to do some pre-emphasis on the front end by using filters which do the following:

$$P = \int H(f)G(f)P(f)df$$
$$= \int P(f)df$$
$$if$$
$$H(f)G(f) = 1$$

This is optical pre-emphasis filtering as one does with FM radio. This is a standard approach.

2. Now let us go back to the input. We assume we have a flat frequency broadband source of radiation emitted from the source. If now we can also pre-emphasize that as well. Then this source follows two paths. Path 1 is a fixed path up and down and through the sample. Path 2 is one that goes to a reflector whose portion is changing uniformly in time and is accurately measures. This second path then send the same signal with the sole exception that it is phase offset from the main path. At times it may be totally in phase and at time totally out of phase. For every position x of the reflecting mirror we measure the combined power spectrum received, the integral of both signals, measure as their amplitude.

3. It can be shown, we do so below, that if one collects the P values and notes them as P(x) then if we sample x properly we obtain samples of the Fourier Transform of the S(f) function. Thus collecting P(x) for the correct values of x and doing so with enough samples we can then perform an inverse Fourier Transform to readily obtain S(f). This is FTS.



The details can be displayed simply as follows. The signal received is the direct and the reflected and they are combined in complex space to account for the phase difference as shown below:

$$S(f) = \exp(-j2\pi d_1 f / c)S_0(f) + \exp(-j2\pi(d_s + x) f / c)S_0(f)$$

$$\approx S_0(f)\cos(2\pi fx / c), \text{ so that the received power is:}$$

$$P(x) = \int S_0(f)\cos(\frac{2\pi}{c} fx)df$$

The result is that P(f) is the real FT of S(f).

FTS shows that we can obtain P(x) and it is the Fourier Transform of S(f) the absorption spectrum of the sample. We take P(x) for many values of x and then inverse FT.

2.2.1.2 Chromatic Images

The ability of the eye to discriminate colors is one of the factors that discriminate the human from the other animal species. The achromatic discrimination abilities relies upon the ability of the eye to be sensitive to the intensity of the radiation of the surface of the retina. The color discrimination capability is the ability of the same neurons to respond in a differential fashion to photons of differing wavelength.

In this section we shall develop the concept of chromatic images, and show how those image are represented and formed and how we can characterize those images for the ultimate purpose of sizing the source and storing the data. The normal visible range of the eye is from 450 nm to 650 nm. Some humans may respond to slightly lower and higher wavelengths. The typical sets of standard colors are :

Pure Blue 470 nm o Pure Green 505 nm o Pure Yellow 575 nm o Red > 610 nm

The concept of pure green is really a psychometric factor. It states that when this color, defined in terms of a narrow linewidth source of radiation, is shown to a large group of people, then they agree in large numbers that the wavelength shown is the pure representation of yellow to them. In contrast, there does not seem to be a uniquely agreeable wavelength for red.

We can define the colors that we have seen in terms of the source of the light if the light is transmitted directly into the eye or of the reflective nature of the surface for light that is reflected into the eye. Let us consider first the issue of transmitted light and then reflective light.

If we use a narrow laser source, we can define the transmitted light in terms of the power density an x,t function of wavelength. Define P(l) as this spectrum and we have:

P( ) = Po f( )

where Po is the peak intensity and f( ) is the spectral characteristic. Light is said to be saturated depending upon how broad a bandwidth the source is. Thus a narrow band light source is said to be more saturated.

We can in a similar fashion look at reflective surfaces and consider the effects of white light being reflected from the surface and entering the eye. The material on the surface will have differing reflective characteristics depending on the wavelength. Let Po be the intensity of the incident radiation, and we shall assume that it is white light. Let R(L,x,y,z) be the reflectivity of the surface as a function of the wavelength and the position. Then we can define the light incident on the eye in the standard fashion. These types are subtractive and additive.

The subtractive techniques takes white light as the source in a transmissive fashion. It then passes the light though filters that take out narrow bands of light and leave a combination of all other colors. Here we have shown white light on a yellow filter that passes right that has had yellow removed and then we pass it through a filter that has a capability to remove cyan. The resulting color is green as perceived by the observer. It should be noted that this perception

depends highly on the narrowness of the filters and the ability of the filter to be centered on the desired wavelength.

Consider now the additive form of color mixing. This is the more common form and the one that we shall be using at length. In this case we take three primary colors, red, green and blue and start with equal intensities of each. We then vary the intensity of the separate colors to create the total color. The resulting total additive color is:

IT = I(R)+I(B)+I(G)

The resulting colors can be a wide mixture of all colors perceived by the human eye. Here we have shown somewhat broadened spectra. We then combine two of the spectra and generate a third combined spectrum that peaks somewhere in between the two. It is this effect that allows for peaking at any point and thus having the psychometric effect of a continuum of colors.

The recognition of this additive effect of emitted spectra and the generation of continuous color has led to various characterization of color mixtures. The simplest is the chromaticity chart that starts with the additive equation of the sources and normalizes the overall intensity. That is we start with the equation:

T = 1 + R+B+G

so that we can define :

B = 1-R-G

Now we can plot the amount of red and green on a two dimensional chart, knowing that the amount of blue is merely subtractive.

over as large a range as possible. Calculation of the luminance of a color would be made much easier if the luminosity coefficients of two of the primaries were equal to zero. The luminance of a color would then be equal to the number of units of the other primary used in the match.

It was these considerations, among others, that led the CIE to propose a new set of primaries. The spectral locus is totally contained within the triangle formed by these new primaries denoted by (X), (Y) and (Z) (as can be seen in Fig. 5) implying that all spectral colors can be matched with a positive quantity of each primary. The use of such "nonphysical" primaries should be no cause for concern. For measurement purposes any real set can be used and the results can be transformed by a 3 X 3 matrix to the nonphysical set.

Luminosity information is obtained from the tristimulus value of the (Y) primary-the luminosity coefficients of the other two primaries are equal to zero. The resulting tristimulus values 7, 7 and T are shown in Fig. 7. Note the all-positive nature of the functions, and that y is identical to the Vx curve of Fig. 2. The xy chromaticity diagram for the 1931 CIE Standard Observer is shown in Fig. 8. The equal-energy white (£) has the coordinates (1/3,1/3) because it is the reference white for this system. All the color mixture properties that we have previously described for the rg diagram are valid for this diagram, however, the equations for color mixture are especially simple for this case. If we are given the chromaticity's and their luminance's Lx and L% the chromaticity of the mixture is simply that is illuminated by a light of a specific spectral distribution. The spectral distribution of the reflected light may, of course, be thought of as being composed of an infinite series of spectral color. To determine how much of each primary is needed in the mixture, a product is formed of each of the tristimulus values with the spectral reflectance of the object as shown in Fig. 6. The areas under each curve, as obtained by integration, are the desired tristimulus values *R, G* and *B* for the sample.

In 1931 such calculations were commonly performed on desk calculators, and the negative lobes of the functions of Fig. 4 introduce negative product terms in which the negative sign is error prone with repetitive summing and differencing operations. It would be much better if there were no negative lobes, and it would be convenient if the quantities were zero

We can then generate any desired color by combining the amount of red or green on this normalized basis. Note that at R=G=0 we have pure blue. In addition we can note the extension of this chart beyond the region for the sets of colors generating them. This is a measure of the psychometric factors generating these colors.

The implications for the use of this psychometric definition of color and the human response falls into many categories. Two of them are most important in the areas of multimedia communications. The first is the use of color in the RGB combinations for the generation of multiple colors on a CRT screen. The basic CRT screen is in actuality a set of overlays of Red, Green and Blue phosphors. Using these combinations we can generate any set of colors on a CRT screen. Consider the types of CRT displays (Merrifiedl p.70). The RGB phosphors are located in a spaced form on the surface of the screen. Three different electron guns are displayed behind the screen and are focused through a set of masks. These masks are called shadow masks and ensure that the guns of the different colors do not impinge on the wrong phosphor.

The resulting image is a collection of the RGB phosphors that are illuminated by an intensity in proportion to the color that is desired at the specific location. The granularity of the shadow mask is a measure of the pixels per inch of the display. For example, the distance between the Red phosphors on the horizontal or vertical axis may vary from 0.6 to 0.2 mm. The discrete pixel

$y^* =$

is the combination of a set of pixels. If we connect an RGB combination in a triangular fashion, then we can see that with 0.3 mm spacing, we see that the pixel of color, an RGB triangular combination is on the order of 0.2 mm. As we discussed in the resolving power of the eye, this gives significant resolution at close distances.

Specifically, with 2,000 by 2,000 monitors, we can have 6,000 by 6,000 individual elements of color at the RGB level yielding the overall combination of 4 million pixels. With a 60 cm by 60 cm display, we have 3000 RGB cells at 0.2 mm spacing. In the current state of the art we have 0.1 mm spacing of these cells.

The shadow-mask color CRT assembly consists of three closely spaced electron guns, a shadow-mask, and a three color phosphor screen. Focused electron beans emitted from each primary gun pass through apertures in the metal shadow-mask and impinge upon phosphor dots for each corresponding color. The three electron guns are arranged in an equilateral triangle, or delta. Each shadow-mask aperture allows the three electron gun beams to project onto an inverted delta or triad of phosphor dots. The angle of incidence of an electron beam as it passes through a shadow-mask aperture determines the color of phosphor dot it excites. Electron beams of a particular gun are blocked by the shadowing of the mask from impinging upon the other two color phosphor dots of each triad. A shadow-mask CRT has a very simple mechanism for selecting color. The three independent guns in the shadow-mask design provide independent control of the luminance of the red, green, and blue phosphors. In this manner, it is possible to reproduce any color within the chromaticity triangle formed by the primary colors.

Several configurations of gun alignments, mask structures, and phosphor arrangements are available. The in-line gun shadow-mask

CRT has the mask and pi the in-line gi through a sh cause of the Resolution c to their small beam at the also available vertical pho a higher lurr The granular distance best ranging from are considered to be confused several mass electron optics.

The ways of storing still images has evolved significantly over the past decade. The method most used ten years ago was that of paper tape and today there is extensive use of laser disk storage technology. In this section, we will develop some of the current and some of the potentially significant means of storing and retreading still images.

As part of the storage process, we first want to consider the means of compressing the image data and as part of that understanding the ultimate information in the image itself. This latter issue relates to the minimum number of bits that are necessary for representation and storage of the image.

The issue of reconstituting still images is one of obtaining either the best video display techniques or in using many of the existing and improving methods of hard copy display. We often think of reinstating the image on a video display and do not recognize that the use of hard copy display terminals is also a significant means of reconstituting the image. There are dramatic advances being made in the areas of both displays and hard copy devices and these will be reflected in the way the systems will evolve with the capability to provide significant editing and composition interactivity on displays.

As we have discussed with the set of still images, the key factor in developing a means to display images is finding an acceptable tradeoff between the ability of the human eyes to respond and the limited processing and storage capability of the computer. In the world of video images, there again is the resolving capability of the human eye, combined with the issue of the field of view of the image and the update rate of the images to meet the eyes ability to comprehend motion. Thus the selection of the size of the standard television set in terms of its aspect ratio was based upon significant testing of the viewers ability to view information in that form. The update rate of Television is 30 frames a second, a rate that matches the lower limit of the human response curve for recognizing image motion.

3.5.3 Video Means of Characterization

The video image is more than just the sequencing of a set of still images. There are several psychological and physiological factors that relate to the scan rates and interlacing of images to provide an adequate video image. In this section, we will focus on the scanning input to the video display and shall also discuss the scanned output of the display.

The video displays follow techniques that extend those of the CRT displays that we developed in Section 2.2. Specifically, we scan the screen in frames, in the US and other countries using NTSC video, this is 525 lines of scans per frame and 30 frames per

second. A frame is composed of two fields, the first field being 262.5 lines and the second field being a same number of lines but these lines spaced between the first field lines. The two fields compose a single frame.

The signal that is generated at the transmitter is a combination of the red, blue and green values of the image that is scanned at the original location. This is not the signal that is transmitted. What is transmitted is a set of three video signals that is a linear transformation of the RGB signal set. This composite transformation set is generated so as to make the received signal backward compatible with the standard black and white sets that were in existence in the United States at the time of the introduction of color TV. In addition the signals are required to fit into the same bandwidth as the black and white signals.

If we let R, G, and B be the red green and blue signals, then in NTSC video we generate a signal set that is composed of the following three signals;

$$Y=0.3R+0.59G+0.11B$$

$$I=0.6R-0.28G-0.32B$$

$$Q=0.21R-0.52G+0.31B$$

These signals are called the brightness or luminance signal, the inphase signal and the quadrature signal. It is found that with this transformation and the scan rates selected for capture and ensuing reproduction of the signal, that Y occupies 4.2 MHz of bandwidth and if processed by itself provides an adequate black and white signal. I occupies 1.5 MHz of bandwidth and carries the orange/cyan mix of colors in its signal. The Q signal is 0.6 MHz in bandwidth and carries the green/purple colors of the signal.

It is actually possible to display these signals by themselves, that is the I or Q signals and see what the resulting image appears like.

The composite video signal of Y,I and Q are carried in a complex fashion. The Y signal occupies the lower end of the 6 MHz bandwidth spectrum and is not encumbered by any other signal. The I and Q signals are carried at a frequency that is 3.58MHz up from the carrier. The I signal is sent using a sine modulation and the Q signal using a cosine modulation. Coherent demodulation is used at the receiver so that the two signals may be separated. At the very upper limit of the band, or the carrier plus 4.5 MHz is located a small FM voice carrier that is used to include the voice signal. All of the video signals are carried in an AM format.

At the receiver, the signals are recovered and the Y, I and Q signals are placed through a transformation to recover the R,G and B signals again. These are then used to excite the three phosphors on the television screen.

Storage of video information is a significantly greater challenge than that of the still image. The still image storage is for a single or set of single images, whereas the storage requirements for video are based upon a continuum of many such frames at the scan rate of the device. In this section, we shall first characterize and size the storage requirement and then provide a set of alternative storage systems.

The means of reconstituting a video image is usually performed on a video display. In this section we hall concentrate on the display architecture and show how we can implement many of the schemes that we have discussed.

The human voice is a complex information transmission medium that is created with the interworkings of the human brain, overall nervous system and the vocal chords of the human body. Combined with the creation of speech is its reception by the human ear which has its own characteristics.

3.5.4 Voice Means of Characterization

The voice is the second of the sense that we find in a multimedia system and its characterization can take on several dimensions. First, we can look at the signal itself, that is the conversion of the sound pressure wave into and electrical wave, and develop the digitization of that signal. From that digitization we can determine the bandwidth necessary for adequate transmission and size the storage requirements. A second approach is to recognize the speech using many of the speech recognition systems and then convert the speech into a new form, the written word. With the developments in speech recognition it is now possible to do this type of characterization for many applications.

Let us first develop an understanding of the generation and reception of speech in the human. The basis elements of the speech generation path are:

Oral cavity including the tongue and mouth, within the confines of the lips and down to the velum at the rear of the mouth.

Nasal Cavity from the nostrils at the front to the pharynx at the ear.

Oral pharynx from the velum down to the epiglottis,

Nasal Pharynx at the rear of the nasal cavity and above the oral pharynx.

Laryngeal pharynx region from the velum down to the vocal chords.

Each of these areas are resonant regions that are used for the generation of particular types of sounds. The vocal chords are used in particular in the generation of speech by passing the air from the lungs through the chords and by exerting various modulations o the chord openings by the local chord muscles. This air can then be made to resonate at different modes in the various cavities and the resulting sounds are speech. If we were to block any one of the resonant cavities, all other elements working properly, we would seriously affect the sound. This is a common effect when a person has a cold.

The roof of the mouth can be divided into two principal regions. In front, the roof is formed by a bone called the palate which separates the mouth from the nasal cavities and supports the upper

teeth. At the back of the palate, the roof is formed of muscle and connective tissue; this structure is called the velum, or soft palate. (Writers who call the velum the soft palate refer to the palate as the hard palate to distinguish the two.) The uvula is a small fleshy appendage at the rear of the velum. The velum can be lifted by a muscle and pressed against the back wall of the pharynx to seal the nasal passages off from the rest of the vocal tract. At the front of the palate there is a ridge, formed by the thickening of the bone where the front teeth are inserted; this is called the alveolar ridge.

The tongue is a large system of muscles connected in front to the lower jaw and in back to bones in the throat and head. Its biological functions include tasting and manipulation of food during mastication. For our purpose it is convenient to divide the tongue into regions. In the absence of distinct landmarks, it is difficult to define these regions precisely, but they are approximately

The ear has three major parts; the outer ear, the middle ear and the inner ear.

The outer ear extends from the pinna or outer ear surface to the eardrum through the meatus. This is itself a resonant cavity for the propagation of the sound entering the human ear. The middle ear is compose of the ear drum, the bone structure called the ossicles and the entry into the Eustachian tube. The ossicles contains three bones called the hammer, anvil and stirrup. These bones are used for the transmission and conduction of the sound responses into the nerve endings that are key to the transduction of sound into the overall nervous system.

The inner ear contains the cochlea which converts the mechanical vibrations into electrical impulses in the hearing part of the central nervous system.

The human ear can hear sounds from the range of 16 Hz to 16 KHz with some exceptions up to 20KHz. The sound pressure limits range from 0 dB or 0.0002 dynes/cm2 to 130 dB. At the high level the human ear generates pain and is damaged. At the lower level there is no perceptible response.

3.6 HEARING AND PERCEPTION

By hearing we mean the process by which sound is received and converted into nerve impulses; by perception we mean, approximately, the postprocessing within the brain by which the sounds heard are interpreted and given meaning.

3.6.1 Hearing

We will start with the anatomy of the ear. The ear is divided into three parts: the outer ear, the middle ear, and the inner ear. The outer ear consists of the *pinna* (the visible, convoluted cartilage), the external canal *(external auditory meatus),* and the eardrum *(tympanic membrane).*

The pinna protects the opening; its convoluted shape is thought to provide some directional cues (Schroeder, 1975). The external auditory meatus is a nearly uniform tube approximately 2.7 cm long by 0.7 cm across through which the sound passes to reach the eardrum. Like all tubes, it has a number of resonant frequencies, of which only one, at approximately 3 kHz, falls in the frequency range of speech. The tympanic membrane is a stiff, conical structure at the end of the meatus. It vibrates in response to the sound and is the first link in a chain of structures which transmit the sound to the neural transducers in the inner ear.

The middle ear is an air-filled cavity separated from the outer ear by the tympanic membrane and connected to the inner ear by two apertures called the *oval* and *round windows.* The middle ear is also connected to the outside world by way of the *eustachian tube,* which permits equalization of air pressure between the middle ear and the surrounding atmosphere.

The middle ear contains three tiny bones or *ossicles* which provide the acoustical coupling between the tympanic membrane and the oval window. These bones are called the *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup). The malleus is attached to the tympanic membrane, the stapes to the oval window, and the incus connects the two. The function of the ossicles is twofold: (1) impedance transformation and (2) amplitude limiting.

The signal generated by the human voice system generates a pressure wave of the form:

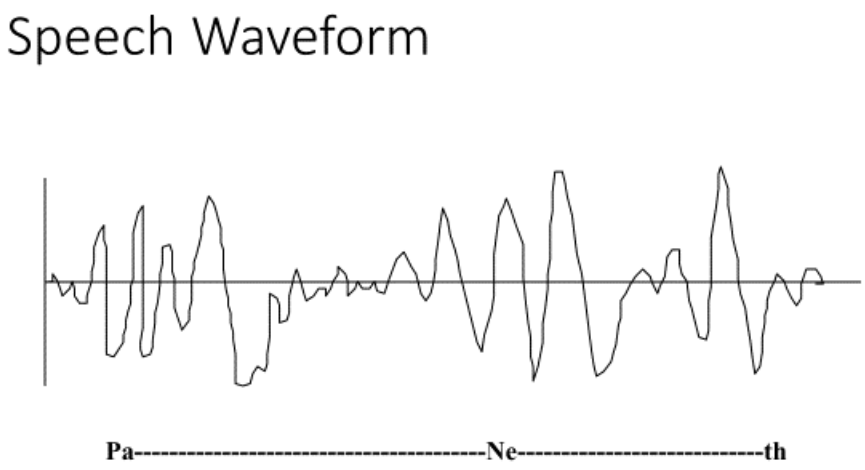
$p(x,y,z,t)$

which is spatially and temporally dependent. The wave is received by the ear and is processed into recognizable speech. We frequently can take this pressure wave and receive it by an electrical transducer like a microphone and convert it into an electrical signal. The resulting signal is a representation of the temporal characteristics of the speech. It does not provide full spatial characteristics and in fact involves the impact of the directionality of the receiving transducer.

As we have already noted, speech is constrained in the 16 Hz to 16Khz range so that the speech signal:

$s(t)$

is composed of only those specific frequencies.



If we define s(t) as the speech signal, we say that this signal is a random process since it changes in time in a random fashion. We can determine the correlation function of this process by talking its average as it is mixed against a shifted version of the signal. We define R(T) as the correlation function of the signal. Specifically if we let E[ ] be the averaging operator on the random process, then we have:

$$R(T) = E[s(t)\, s(t+T)]$$

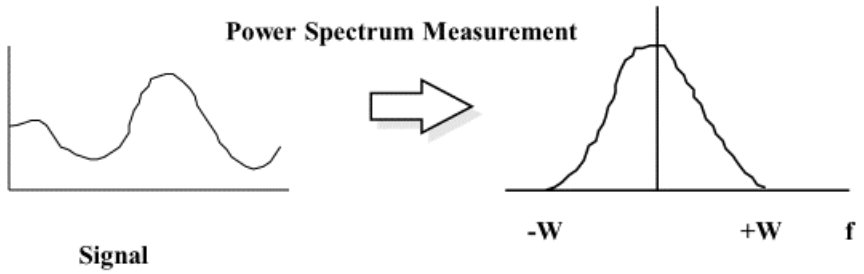
where we have assumed that the process is stationary, that is R is independent of t.

The spectrum of the signal is defined as the following:

$$S(f) = \int R(T) \exp(-j2 \quad fT)\, dT$$

The spectrum for a typical voice signal is shown below

The Nyquist Rate



Sample at T = 1/2W; the Nyquist Rate

We can now use the Nyquist sampling theorem that states that if we sample a signal at twice the maximum frequency rate, then we can totally reconstruct the signal. Thus if a signal is from 0 to l0 KHz, we must sample at 20,000 per second. For voice this means for full fidelity we must sample at 32,000 per second. Generally this is not the case and we frequently limit the upper voice signal in telephony rates to 4 KHz and the sample rate is 8,000 per second.

We can now quantize the samples to create a digital sample. If we quantize s into two levels all we need is on bit (eg 0 or 1). If we quantize into four levels we can do that with two bits. If we continue, we can quantize into 256 levels and do this with 8 bits. This 8 bit quantization has been shown to give minimum human distortion in terms of the human response.

Thus typical speech is sampled at 8,000 times per second and at 8 bits per sample. This means that voice is captured at the rate of 64,000 bits per second. High fidelity speech may be at the 32,000 sample rate and may even use 10 bit quantization. Thus for high fidelity speech we have 320,000 bits per second. Let us review this in further detail.

First we have the general signal s(t) which is the voice signal.

Second, we take and sample the signal at the sample points. The sample pints are twice the maximum frequency of the spectral density. The corresponding sample time intervals are defined as T sec. For 4 KHz bandwidth this is 1/8,000 sec or 125 micro second sampling.

Third, enumerate the sequence of samples;

(s(kT) } : k = 1, , N

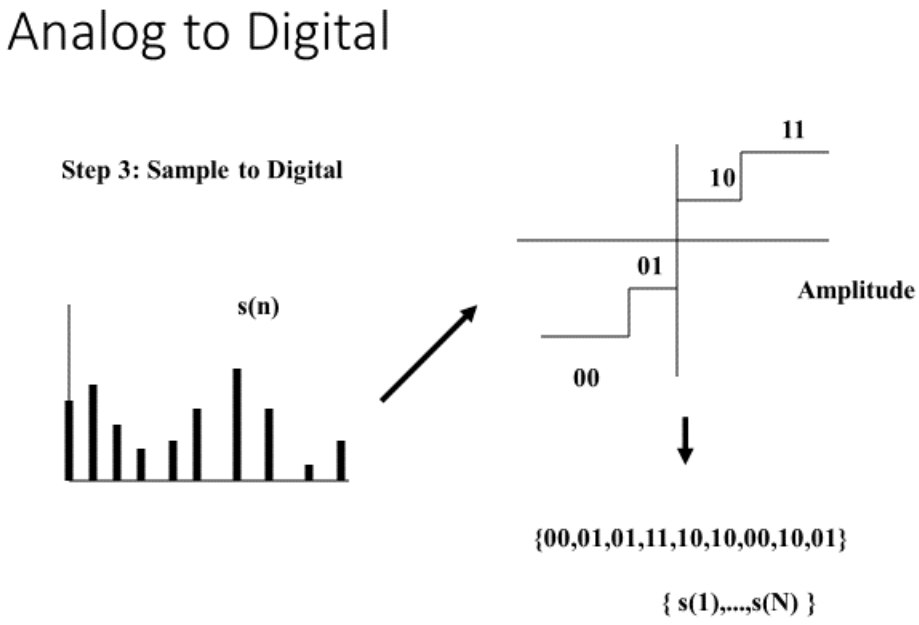
as the sampled signal.

Fourth, using a quantization scheme, select the number of quantization levels, using an a priori choice of through an optimizing process, but choose the level to be consisted with a binary format. That is the quantizing levels should be some factor of 2 to a power. Let us assume that there are 256 levels or 8 bits of quantization.

Fifth, choose the quantizing levels either a priori or through an optimizing scheme. The levels are defined as follows. Let Lk be the kth level and let the total level be;

{Lk} : k=l,...,2N

where N is the quantization bits.



Then we have the quantization mapping of :

s(kT) is in Ln if Ln< s(kT) < Ln+i

Sixth, then map s(kT) into bit pattern Bn which is the binary level of Ln.

Seventh, create a data stream of the form:

[B1,B2,,Bn,Bn+l,]

where Bi is the specific bit stream:

Bi = [0,1,0,1,1,...,0]

3.6.2 Voice Sampling.

Voice storage systems are generally well developed and the major factors in such systems are the ease of access and the ability to store large volumes of speech while still retaining the quality of the original.

3.6.3 Means of Reconstitution

Speech regeneration can be done in one of two ways. As we discussed earlier in the section, we can take the digitized version of the speech and regenerate it directly by reversing the process. A second approach of speech generation from the written word, this is called speech synthesis. The development of speech synthesis systems has been significant over the past several years and these now provide interesting alternatives for use in multimedia systems.

Voice Coders Alternatives

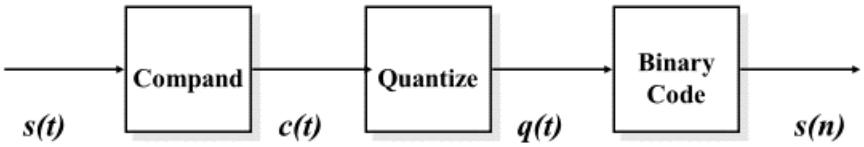
Open Loop Designs

- **PCM**
- **D*PCM**
- **DPCM**
- **Delta Mod**
- **Adaptive Delta Mod**
- **CVSD**
- **Embedded Delta Mod**
- **RELP**
- **Sub Band Coding**
- **Transform Coding**

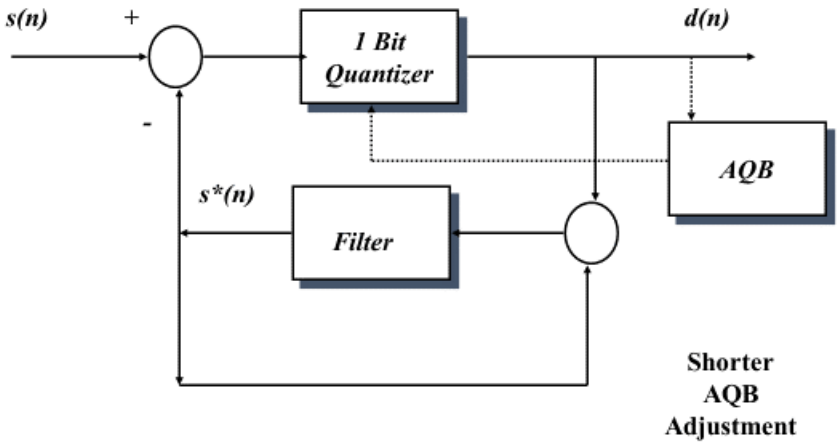
Analysis by Synthesis Coders

- **VSELP**
- **CELP**
- **BPE**
- **MPE**

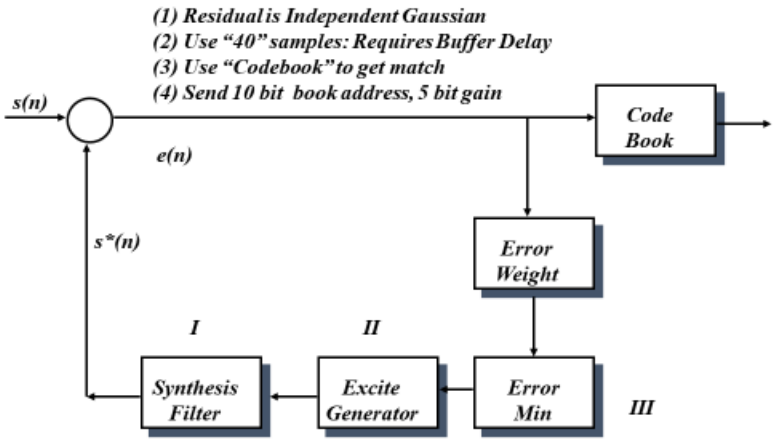
PCM


Compand
Quantize
Binary Code
s(t)
c(t)
q(t)
s(n)

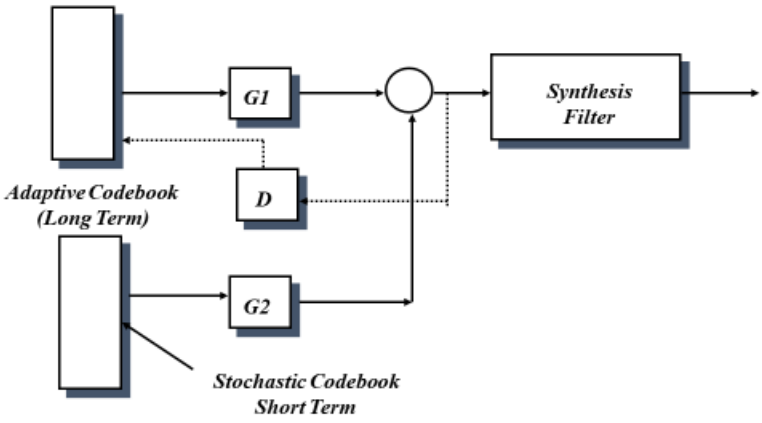
CVSD


s(n)
+
-
1 Bit Quantizer
d(n)
AQB
s*(n)
Filter
Shorter AQB Adjustment

CELP (Encoder)



CELP (Decoder)



3.7 GRAPHICS

Graphics is distinct from the other three representations. In still image, video and voice the information medium is naturally created and is generated from some form of human activity. In

the graphics world, the information is created by a computer to represent some form of reality but is not a priori created by the natural world. Graphics development has progressed significantly over the past few years. The original systems were those that were developed to use a minimum of storage and a minimum of processing. In addition they were designed for vector oriented displays and not for fully bit mapped displays. In this section we shall discuss three of the graphics standards and show how they have evolved through what is now their present state.

The first standard that we shall discuss is the GKS, Graphics Kernel System, standard that was originated in the 1970s. It is the oldest standard and many of its characteristics reflect the limitations on the technology at the time it was developed. The standard uses the basic vector graphics approach and although it is designed for use on many machines, it does not press the capabilities of the processors to allow for maximum flexibility. Thus it has only two dimensional characteristics and requires significant detailed definition for a fully flexible design.

The second standard is the PHIGS, Programmer Hierarchical Interactive Graphics Standard, that is much more flexible and assumes an generally more capable processor. It however still relies on the vector graphics approach to display.

The third standard, although a defector standard, is that of PostScript. This system was developed in conjunction with the work on the Apple Mackintosh system and takes full advantage of the bit mapped displays. It functions on smaller machines, utilizing the power that is available in the smaller processors, smaller in size but significantly greater in processing capability. In addition is recognizes the availability of the bit mapped display and the inherent advantages of the existing MAC display software and the introduction of laser printers.

3.7.1 Means of Characterization

Graphics characterization has evolved over the past ten years with significant growth in capability. The graphics software packages are now extensively used in engineering, architecture, and in many artistic applications to design both still and full motion depictions of real and imagined events. Unlike stills, video and voice, the relationship to the external reality for this media characterization is less strong. For example, we can think of a designer who is developing a new auto body and in so doing is using a graphics design package to help do so. The graphics package starts with the ideas that the designer has and follows the designers lead in developing the body design of the auto. In this case, the output is the reality and not the input.

Many standards have been developed for the development of graphics displays. These standards rang from such as GKS, the Graphics Kernel System, to PHIGS (the Programmer's Hierarchical Interactive Graphics Standard), the CGM (Computer Graphics Metafile).

Graphics systems take a set of from desired images and reduce them to a defined set of primitive elements that in sum make up the resulting image. Let us consider a simple example. We may desire to develop the graphical representation of a computer terminal. We can first start with a set of rectangles that can represent the overall structure of the terminal, the key board and the display device. We can then add to these basis elements sets of rectangles, circles or ellipses that make up the keys, displays, logos etc that make up the final terminal. We can then apply further and further detail to the device to make up a more complex and detailed display. We may even use the capability of creating complex curved surface by using spline functions that are basically higher order polynomials defined on a limited set of R N.

The net result of a graphics system is a picture representation that can fall into one of two major categories, vector graphics or bit mapped. In the bit mapped systems of generation, the images are created by a fully random selection of bits and color associated with the pixels. That is we can imagine that the end user may be capable of entering in to the system the image on a bit by bit basis and doing so by selecting a pixel and then choosing the colors or intensities for the pixel. This would mean that for a fully bit mapped system, we would have to have N by N entries with the B bits chosen per entry. Thus there would be N*N*B steps involved. For the 2K monitors, this would entail almost 100 million steps per screen. Such an endeavor is beyond the capabilities of most humans.

The alternative is to select a system that has a set of basis functions such as circles, rectangles, ellipses, etc and that in addition is aided by a set of primitive commands that allows for the placing, orientation and coloring of the basis functions. The net result is a vector oriented system that does not require a bit by bit painting of the display device. In addition this system of basis functions and primitives allow more readily for the animation of the graphics images.

Graphics standards have several objectives in the development of their design. Typical sets of objectives that will be found in the three systems that we are to discuss are as follows:

1. Machine Portability

2. Application Program Independence

3. Display Independence

4. Hierarchical Designs

5. Ease of Compilation

We shall see that some of these goals are not fully achieved in all of the approaches but that they are generally thought of in the implementations.

3.7.2 GKS

GKS, the Graphics Kernel System, is a graphics design tool that has the status of being the first graphics standard. It is composed of six general elements:

1. Primitives
2. Attributes
3. Attribute Bundles
4. Segments
5. Viewing
6. Input

Let us describe each of these in turn.

Primitives: The primitives or output primitives, or output functions are those functions that are defined in GKS that are used for the generic characterization of the basic sets of elements that are to be drawn from in GKS. The primitives consist of the following set:

Polyline: This allows for the drawing of a set of connected lines from any point to any other set of points.

Polymarker: These are markers that are placed on the figures to specify particular locations on two dimensional surfaces.

Text: This allows for the display of arbitrary strings of characters.

Fill Area: This allows for the defining and filling of arbitrary boundaries of polygonal figures.

Cell Array: This displays a grid of rectangular elements.

Generalized Drawing Primitive: This primitive allows for the generation of arbitrary shapes of figures.

Attributes: Attributes are specific characteristics than can be attributed to any of the primitives. The attributes are specified in the form:

SET FILL AREA COLOR INDEX

SET POLYMARKER COLORINDEX

SET TEXT ALLIGNMENT

For example the attribute SET LINEWIDH SCALE FACTOR affects the polyline line width and allows for arbitrary scaling.

Attribute Bundles: These are collections of attributes that allow for the defining of specific primitive attributes out of the main source code. For example we can use the expression:

SET POLYLINE INDEX

to set the polyline to the characteristics of index 1. Index one may be machine dependent and the applications programmer may not have to worry about the portability of the code on a machine dependent basis. Thus GKS allows for the assignment of attribute and primitive bundling through the index function.

Segments: Collections of primitives and attributes, and even attribute bundles create segments, which are separate descriptive parts of the total graphics picture. For example if we are drawing a house, we can create a segment that is the roof of the house, composed of many polylines, and their attributes and many of the other primitives as necessary. The net result is the segment of the graphics representation.

Viewing or Transformation Functions: This element of GKS allows for the definition and description of the viewing space of the GKS display. Typical viewing functions are:

SET VIEWPORT SET WINDOW

SET WORKSTATION WINDOW

These functions are used to specify such functions as the location of the window coordinates on the screen.

Input: This function provides for the operator input to the program and allows for the entry of such elements that can provide location, select numbers and text for input, and select or pick

specific segments of graphics for display or inclusion. The input functions are in effect a command language for the operation of the GKS kernel.

It has three main layers. They are:

1. Primitive Attributes that are used to characterize the primitives and create the individual segments of the GKS image.

2. Segment Attributes that are used for the detail specification of the segments.

3. Workstation Attributes that provide the details associated with the specific display device. These are typically given in terms of specific device drivers.

A typical GKS image generation shows how the sets of primitives scan be combined to generate a set of segments. These segment in turn are used to generate the overall image. GKS descriptions are generally machine dependent and are structured by the collection of segments of the image. GKS allows for device independence by allocating the device specific portions the device drivers that are separate from the source code.

We can now take any GKS image and address the issue as to how we characterize the image in terms of data storage. Specifically, let us consider the image of a house. The hose is composed of the following segments:

1. Roof

2. Chimney

3. Body of House

4. 10 Windows

5. 10 Window Shutters

6. Door

7. Light

The attributes associate with the roof are those of color, texture, scale and several other factors, similar attributes are attributable to the other primitive in the segments.

Now we can look at the image from two points of view. First is we look at the 1,000 by 1,000 monitor at 12 bits per pixel, we have

12 million bits for the display. If on the other hand we look at the segment characterization, we can say the following:

1. 7 Separate Segments
2. 10 primitives per segment
3. 20 Attributes per segment
4. 20 Characters per segment
5. 8 bits per character.

This yields 244K bits for the storage of this image using GKS and no coding for either primitive or attribute expressions. This is a 50:1 image compression based just upon the fact that we know how the image is constructed. This is typical of many of the image systems that are found in the graphics area.

3.7.3 PHIGS

PHIGS is a second generation graphics standard. PHIGS is dramatically different from GKS in many dimensions, the two most important being the hierarchical nature of PHIGS and the second being the ability to dynamically center attributes as the application program is being run.

Here we first show the basic application program which the PHIGS terms are imbedded. This interfaces with the PHIGS package that includes the graphics system to actually generate the graphic images and its associated graphics data elements.. These are then interfaced with the Input device drivers and the graphics display drivers that finally interface with the operator. As with the GKS approach, there is a segmentation of the device dependent elements and the imbedded image dependent elements.

PHIGS has the two basic elements of the primitive and the attribute. The primitive carries with it the same function of form and shape that we had in GKS. Thus the primitive of a line or circle can be found in the PHIGS format. The attribute is similar to that of GKS in that is associates with a primitive a specific set of characteristics such as shading and width.

The difference between PHIGS and GKS is that in GKS we bind the attribute and the primitive together once and for all. In PHIGS, we can associate a set of attributes and primitives together

as needed and this association can change with time. The binding of the two elements is an output process not a definition process.

This allows for a significantly more flexible design and it allows for the ability to perform animation in a three dimensional context which is not possible with the GKS type of design. Like GKS however it is a structural language and builds from the vector graphics paradigm.

PHIGS is a hierarchical structure. Specifically it allows for the development of STRUCTURES that are bindable primitives and attributes and these structures can be composed upward or decomposed downward. For example in GKS the HOUSE was defined in terms of the roof, the chimney the windows, etc. In PHIGS, we can take the house and take any one elements and further decompose it. Thus the roof can be decomposed into the shingles, the nails, the edges etc. The shingles can be further decomposed into the tar, the paper, and the gravel. We could continue this process

down to whatever detail is necessary. This approach in PHIGS allows the graphics designer to provides greater detail in the specific image.

3.7.4 PostScript

PostScript was the development of John Warnock while at Xerox PARC and is focused on the truly bitmapped display devices such as high resolution terminals displays and laser printers. The PostScript approach is one that mimics the paradigm of the printer placing ink on the page to obtain the effect that is sought by the graphic artist. It is a much more flexible development language and implementation system than many of the other alternatives. It, in addition, is directed at the transfer of the developed graphics to paper, whereas the PHIGS approach is optimized for the continuing display on electronic media.

The basis design object in PostScript are the:

1 o Text Elements that provide the large selection of standard and customizable text fonts.

Geometric elements that provide for the definition of circles, lines and rectangles.

Sampled Image elements that permit the direct importation ad manipulation of external images.

These are all combined in the overall construct of the imaging model. This is the equivalent to the segment of GKS or the structure of PHIGS. The imaging model is composed of three elements;

Current Page: This is the abstraction of the working space on which the graphics layout is made. The current page is the blank space that is filled out by the designer.

Current Path: This is an abstraction that may be independent of the current page and it is the collection of the basic graphics objects.

3.8 CONCLUSIONS

This chapter provides the reader with a method for the characterization and modeling of various multimedia information sources. The source characterization problem is one of the key problems in the development and analysis of multimedia systems. The source of a single user must be composed of a set of the fundamental sources that that user may interact with. Those fundamental sources are those that we have developed in this section, and possible others that may be developed as new technologies are developed.

The most significant fact that the reader should obtain from this chapter is the recognition that in a multimedia environment, we are always striving to duplicate or record a reality that exists external to the machine. The techniques for doing so have been developed both within the context of the computer world and within the world of print medium. We too often forget the importance of the print medium as a means of transferring information from one location to another and from one person to another. We also are seeing that with the introduction of such applications as digital printing systems, we are seeing the melding together of the print and electronic industries from both the operations and applications viewpoint.

4 MULTIMEDIA INTERFACES

There is presently an extensive body of techniques that allows for the interfacing with multimedia systems. The development of the use of windows, the application of high resolution screens and the capability to transmit at Gbps rates allows for the implementation of much more sophisticated technology. In this chapter, we focus on all the interfaces to the multimedia elements, those that are direct to the end user and those that are more closely linked to the processors.

In the last chapter we focused on the different types of presentation alternative afforded to the end user and in the process developed sizing models for the different types of media objects. In this chapter we focus on the human interaction elements with a multimedia terminal. This interaction has two basic bounding elements. The first is the understanding and sophistication of the end user and the second is the complexity of the application that is being used in this part of the overall environment. Typically we desire that the end user have a minimal amount of training for the system, however, this often goes contrary to the desire to have a fully flexible design for the system. The second desire is to have all applications share a common context of use rather than developing a visual and tactile dissonance with the user. Such a goal may be desirable but doubtful ever reachable.

This in this chapter we develop and overall architectural paradigm for the interface to the multimedia system and then give some examples of interfaces that exist today as examples. We then extend the effort into the modeling of the interfaces and the modeling of the total end user source, including the effects of the image characterization. This modeling effort is a critical element in the analysis of the overall design of multimedia system since it helps answer the many system and human factors questions that have been developed over time.

In the paper by Hartson and Hix, they clearly state that the objective of interface design is not just to construct good interface but to develop environments in which good interfaces can be generated. In this chapter, we focus not on the interface as an end in itself but as a means to develop better interfaces. Thus our first efforts are directed at analyzing interfaces and their performance. We specifically develop models for interfaces so that we may determine their performance relative to a set of predefined performance criteria.

We then use that performance modeling as a tool to develop a synthesis of the Harston interface design, knowing the methodology of analysis. Interface design is almost a Marxian process or Hegelian dialectic, thesis or first attempts, antithesis or responses, and hopefully a synthesis of

the key elements of a good design. To reach that however, we need to understand what constitutes a good design. Our approach is to focus on as many tangible issues that are possible.

4.1 INTERFACE ARCHITECTURES

Many works have been written on the human factors issues of man machine interfaces. The simplest rule however is the one that states that the interface should be intuitive and lead to a minimal amount of sensory dissonance. Sensory dissonance is the phenomenon that results when the human user is asked to do something in a way that is radically different that he has been accustomed to. A typical example of tactile dissonance is found in the area of operator service stations in the telephone companies where calls are made to obtain telephone numbers. The key entry devices are not in the standard QWERTY format but in an ABCDEF format. This was done after it was found that the incoming operators could not be required to have typing skills. The net result is that when an experienced person tries to use the system there is sever dissonance and the response time is dramatically reduced.

The development of interface architectures may proceed along many lines. Sutcliffe has described the Command Language Generation methodology with its four levels. These levels are:

Task Level:The level at which we have input and output entries and we focus on what is being done. Specifically this level focuses on the need that are satisfied by the user in applying the interface mechanism.

Semantic Level: At this level we deal with the meaning of the processes that we are performing and we deal with the underlying semantics of the tasks. Many elements of this approach deal almost with the Chomsky like analysis of language and the ability of the language to convey meaning.

Syntactic Level: This relates to the formal relationship of the formal interrelationships between the symbols or words in the interface language or grammar. The syntactic level deals with the semiotic elements that focus on the general philosophical theory of signs and symbols and their use in natural languages.

Interaction Level: This is the simplest level in that is focuses on the simplest manual entries of the end user. What key movements are made and how is the mouse moved to enter the data.

Thus we must consider developing an architectural construct for the multimedia interface. This construct should first start with a clear understanding of the requirements of the processes being executed and the sophistication of the user executing them.

4.1.1 Requirements and Specifications

The design of interfaces requires and understanding of the application and the level of expectation son the part of the end user. There are however several factors that are general in nature and are common in all user interface requirements. These elements are as follows:

Stability: The interface operations should be stable in that actions on the part of the user do not lead to fault states that are unrecoverable. A typical stable interface requirement is one that sates that there shall be no set of inputs that can occur that would lead to a locking up of the system.

Recoverable: This system should have the capability of providing a resolvable mode of operation that allows the end user the ability to return to where they came from without excess operations. Typical in this area is the use of the ESC key in PC applications.

Consistency: This implies that the same result will occur if a transaction is entered in the same way no matter what state the system is in at the time of the transaction entry. It means that there is no level of ambiguity of response that should be anticipated by the end user.

Non-Ambiguous: The interfaces requests for responses should be clear and unambiguous and the level of semantic complexity should be kept at a minimum. If two alternatives are presented to the end user then both alternatives responses should be equally stated in the presentation.

Minimizable: This implies that the transitions needed to effect certain actions should be of minimum duration. The steps necessary to effect any transaction should not be excessively long.

Extensible: The system should be such that all states are reachable\e from any other state,.

4.1.2 Elements and Alternatives

The architectural models of the interface that we are developing in this sections are composed of several layers of elements. We can consider these to fall into two general categories. They are the static and dynamic characteristics.

The static characteristic of the interface describes how the interface is viewed in a single interaction. We can look at the static user interface to be composed of the following elements:

Input: This represents the method, means and technique to enter information into the system. It can range for a mouse entry to a keyboard entry and include touch screen or even digital scanning devices. The entry technique depends upon the specific application, the nature of the end user and the environment.

A classic example of a total mismatch of input mechanisms is the attempt to put a mouse on the floor of an investment baking house trading station. In that environment, and device that is not tied down to withstand hurricane forces can and will become a projectile. Thus the first sets of mouse type entry devices found their way flying across many a trading floor. It sounded like a great idea for the UNIX designer but for the actual user was appropriate for other applications.

Output: This represents the means, methods and techniques to retrieve information from the system. The output in the static environment represents the one time application of the information transferred to the end user.

Presentation: The presentation element of the static design represents how the information is presented in form and design to the user for the purpose of inputting data obtaining output or moving into a more dynamic mode of operation.

The dynamic elements are as follows: States, Tasks, Dialogs, Transitions, Interactions

4.1.3 Performance Issues

In evaluating interface systems, as we stated earlier in this chapter, we are basically developing models for the interface and showing how the interfaces relates to and performs within a larger and overall multimedia communications environment. Specifically, the interface system should ensure that a rapid and error free level of performance is attainable by the user.

Thus as part of the development of a multimedia source interface description, we need to determine the following types of performance measures:

Source Rate: The source rate of the total interface can be generated from the combination of source characterization of the complex images as well as the models of the source within the context of the human interface.

Response Time: The time between requests and responses is termed the response time. Thus the time that the user requests an image and the time that the image is delivered is one measure of

the response time. Another, more general abstraction, is the response time that is from the time the user begins the inquiry to the time the inquiry is complete. This is more than the time from the request of image to response of image. The image is only part of the inquiry process. We must be capable of modeling and accounting for the entire inquiry process, including but not limited to that of a specific multimedia image.

4.2 PRESENTATION INTERFACES

There has been a significant development of various presentation interfaces over the past few years. This development has focused on the need for a set of standard and portable interfaces and development environments that can be used for a wide variety of applications. The driver for this has been in many cases the world of UNIX and C. UNIX is fundamentally an operating system that was developed in response to the need for a multiuser/multitasking environment that could function on a smaller size machine (PDP 7). C was the language that was developed to implement C and satisfied the needs of the developers as an elegant character and string manipulation medium. Neither was developed in the context of dealing with less sophisticated end users and moreover neither envisioned the growth of sophisticated display and I/O devices. Thus it has become necessary to develop overlays to these basic elements of the computing world.

This section details several of the more common examples and presents them in the layered architectural context that we had developed in the previous section.

4.2.1 X Windows

The X Window interface is a current example of the end user interface development capability that is available for use on many platforms. X Windows was a joint development effort that was done at MIT for the purpose of running on top of UNIX type operating system and affording the C programmer a more effective end user interface capability. X Windows provides and interface to the applications program and in connected by calls to the X Library of calls. The text by Schleiffer, Geddys and x provides significant detail of the operations of the X interface.

X is en environment, an environment to develop the user interface. It is not itself a user interface. It is a fairly complex environment that allows for the interfacing not only with the user but with the applications program and the network of other users in the system as well as complex form of multimedia information.

X is composed of the following sets of elements. They are:

Application: The application is the end user program that is intended for use by the user for effecting several transactions on the system.

X Server: The X server is a program that is run on the users machine or another machine but it acts as the intermediary between the applications program and the handling of input and output to the display. The X Server is the key X ingredient as an operative agent in the communications network.

Workstation: This is the physical device that an application is run on. Frequently in the X environment, there is one X Server per workstation.

Clients: The clients are all the users on the X environment and may usually be represented by other applications programs.

X Library: This is a set of functions presented in the form a X primitives that support the X windows applications. The X Library)(X lib) is structured to support a specific C language environment.

X protocol: This is a set of tools that allow a single user to provide other users interfaces to be independent from other clients.

X Window Manager: This provides for the management facility for the window layout and interaction and assists when many windows are in action.

X tool Kit: This is a user interface subroutine library that does certain complex tasks using already written X code that employs the primitives from the X lib.

Events: These are time stamped results of users actions such as the entry of a key stroke or movement of a mouse.

The overall resource architecture available in the X environment. This resource architecture allows for the access of such elements as: Windows, Graphics Contexts, Fonts, Color Maps, Pixel maps, Cursors.

A typical X systems architecture shows the relationship of the workstations and their related X servers and shows how this flow into the network through the X protocol. Using the X lib functionality, we see how X provide access to the other elements of the X architecture. The X protocol provides for the control over the flow of 11 of the multimedia elements in the system. There is a significant disadvantage of this layered architecture. It is that the X protocol can often cause thrashing of the data elements as the system tries to display complex multimedia objects. In the present configuration, it is not common for the X protocol calls to file servers to take 15 to 45 seconds to display a high resolution image.

There are many types of servers in an X environment. They range from printer servers, database servers, name servers and communications servers. Each can be addressable within the X context.

the X window environment has a hierarchy of the windows themselves. The hierarchy allows for the definition of a parent window, shown as WO in this example. From this parent we can define a course of other windows that can be generated and manipulated.

X lib is structured into the following major functions:

1. Display Functions
2. Opens displays
3. Places information onto the display
4. Closes the display
5. Window Functions
6. Create
7. Destroy
8. Map
9. Unmap
10. Attribute assign
11. Configure
12. Translate
13. Window Information Function

4.3 GRAPHICS USER INTERFACE (GUI)

We have seen that X windows and the more applications oriented support of DEC windows or UIL allows for the development and support of the end user interface. This has been formalized in the structure of the Graphics User Interface architecture (GUI) that shows how these elements relate to the hardware, operating system and ultimate the applications and the end user.

The general architecture of the GUI contains the following elements:

Hardware: This is the standard platform on which the system will operate.

Operating System: The standard operating system of choice. Frequently this will be UNIX, PS/2, VMS or DOS.

Graphics System: This element is the actual graphic system used for the generation of the display graphics. X in X lib has a certain graphics system capability but is very limited as a user friendly graphics generator. PHIGS, GKS or PostScript has the better graphics capability.

Window Environment: X is a typical window environment. The wind approach allows for the opening and closing of specific applications oriented windows that allow for the focusing of the users interest and application.

Applications Manager and File Manager: This is the applications and often platform specific display interface that allows for the direct interaction between the screen and the window manager.

User Interface: This is the direct end user interface. OSF Motif is a typical example.

Applications: This is the end user applications.

There are advantages and disadvantages with the types of use interface development tools that we have just been describing. They are:

Advantages

Provides a modular and transportable development environment.

Disadvantages

Adds significant operations overhead on the system thus introducing delays and other operational factors.

Does not allow for customization and real time optimization of the applications.

Requires significant memory allocations to store the library functional and other parts of the GUI elements.

4.4 MM SERVICES

Having developed some of the issues related to the development environment of the end user interface, we can now focus on several specific end user interfaces and review how they meet the requirements and specifications that we have discussed. All of these interfaces have been

developed in an X windows context and thus have been developed within the development context discussed in this section.

The session screens allow the end user to enter into sessions with other users, applications programs, or data bases. What we have shown is a simple set of directions for accessing and determining the specifics of the session operation.

4.4.1 State Machine Analysis and Petri Nets

We shall now take the state diagram construct developed in the last section and apply it to several specific applications. Specifically we shall apply it directly to the four services that we have discussed in the last major section; mail session, file and directory. These services are the type that would be of use to the typical end user of the communications service.

Presentation: This represents one screen that presents the information to the end user. The presentation element is the heart of the state dynamics model.

Input Action: This is the set of input actions that are made in response to the elements that are presented on the presentation screen.

Read Data Base: This action is one half of what is called a database transaction. It reads a database element that is used as an integral part of the overall end user effort.

Write Data Base; This is the second part of the total database transaction and represents the second possible response to an end user input action.

Process: This is the actions that are taken internal to the system that combines the result from the presentation, that actions of the database transaction and the internal mechanism that are part of the user interface.

Output Action: This is the action that results from the process and the database transaction elements or it may even result directly from the input action and the presentation. The output action then leads to movement to another presentation.

Using the state dynamics model, we can see that the states are effectively the presentation elements, and that the transitions between are comprised of all, of the other elements. This transition path is a block representing the transition process that we have developed in the state dynamics model.

Consider the example of a simple user interface that is part of the withdrawal of funds from an ATM. We have four presentation elements or screen that are presented to the end user for s decision to be made. The screens are:

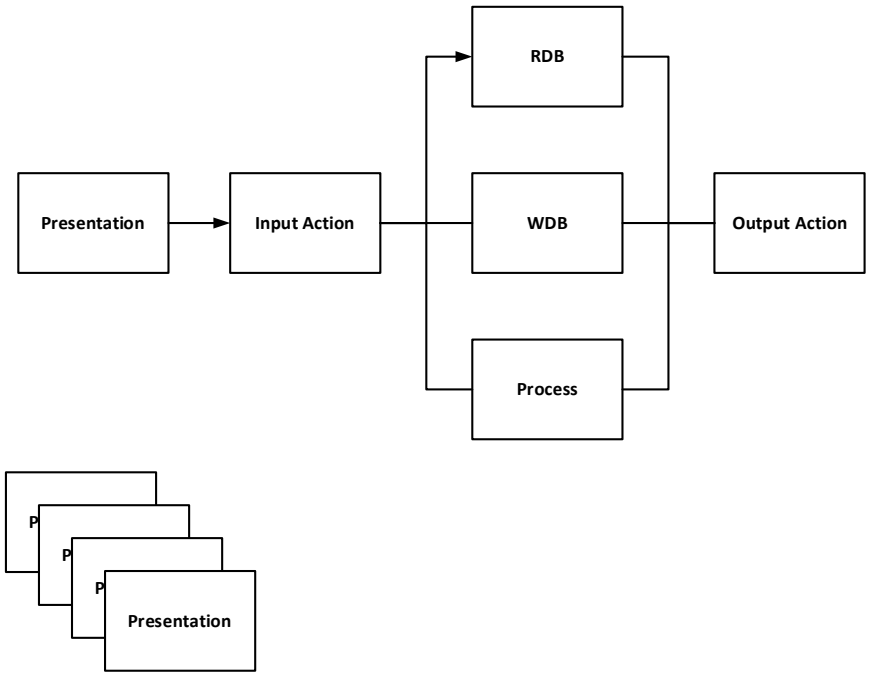
Screen 1: Indicates that the user must log onto the system with some PIN number or other security code mechanism.

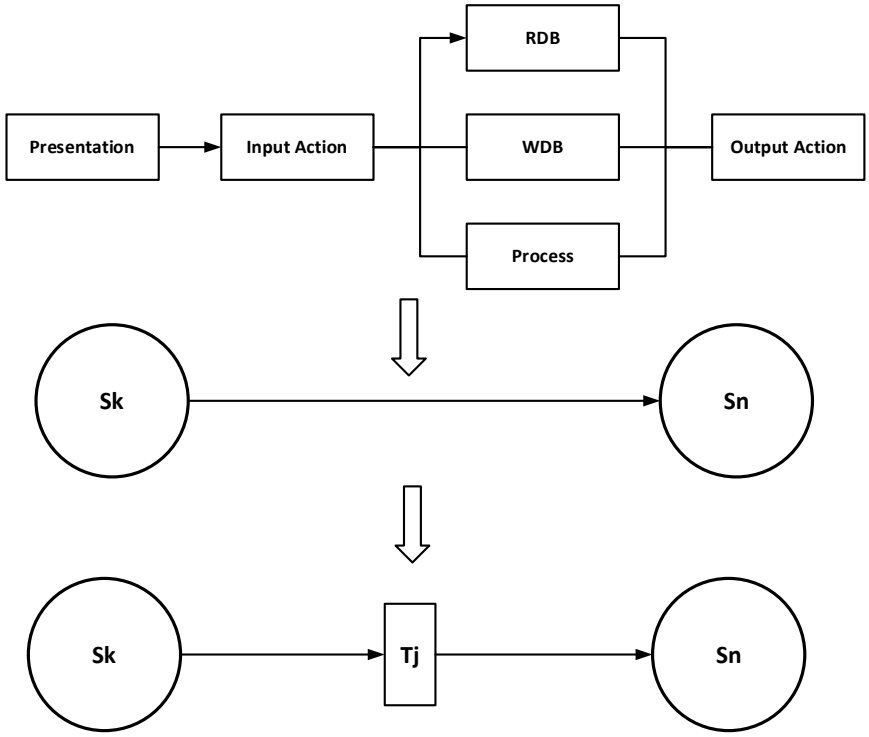
Screen 2: Asks the user if they desire to deposit money or withdraw funds.

Screen 3: This is the withdrawal screen and has all the information necessary to withdraw funds.

Screen 4: This is the deposit screen and it also contains all the necessary information.

Movement occurs between the stets. The movement requires that the other elements of the model be evoked, namely the database elements, and the input and output functions, each associated with a particular screen. For example, let us focus on screen 3 which is the withdrawal screen. It has the following elements:





In this figure, we show that the basis element are the states and the transitions. These two elements can be combined into a general theoretical structure called the Petri Net. We shall rely upon the work of Marsan et al to develop this theory. This we shall also extend with the work of Stotts and Furata which combines this with the Hypertext model. We can now begin to define the Petri Net.

Definition: The Petri Net, PN, is a tuple consisting of the following elements:

P = A set of places = {Pl,....,Pn}

T = A set of Transitions = {Tl,....,Tm}

A = A set of directed arcs; (PXT)X(TXP)

and :
PN = {P,T,A}

In a PN there are two sub sets call pre sets and post set. These are the elements that make up the arcs. We define these as follows:

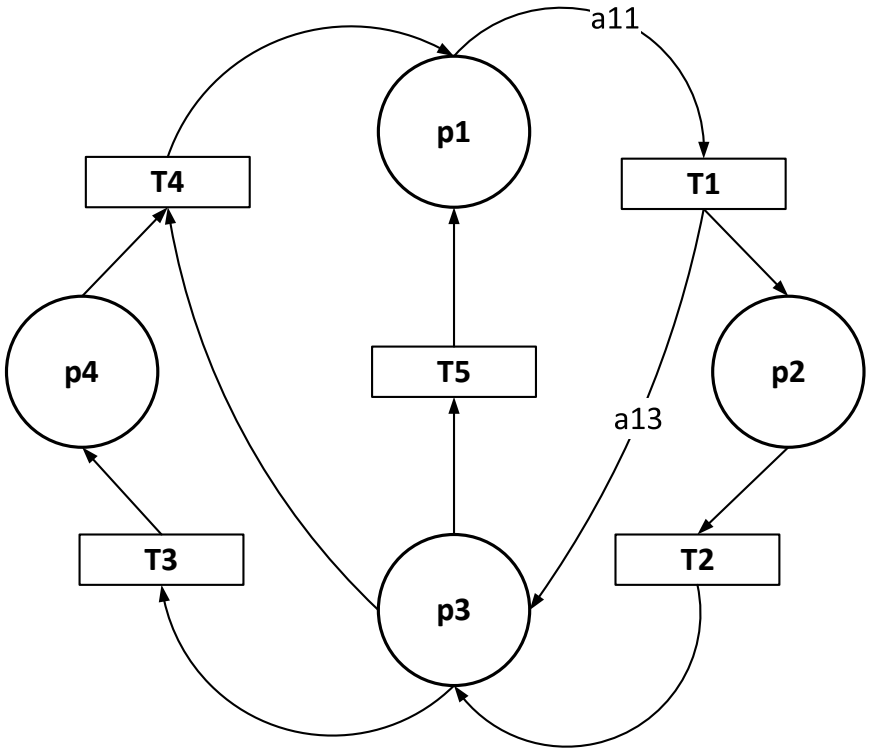
preT = {P: (P,T) F}

and

postT = {P: (P,T) F}

where F is the set of all acceptable P,T pairs.

Consider a sample Petri Net as shown.



Using the state dynamics model, we can see that the states are effectively the presentation elements, and that the transitions between are comprised of all, of the other elements as shown. In this figure we show the relationship to the state and the transition path. This transition path is a block representing the transition process that we have developed in the state dynamics model.

In this figure, we show that the basis element are the states and the transitions. These two elements can be combined into a general theoretical structure called the Petri Net. We shall rely

upon the work of Marsan et al to develop this theory. This we shall also extend with the work of Stotts and Furata which combines this with the Hypertext model.

We can now begin to define the Petri Net.

Definition: The Petri Net, PN, is a tuple consisting of the following elements:

P = A set of places = {Pl,....,Pn}

T = A set of Transitions = {Tl,....,Tra}

A = A set of directed arcs; (PXT)X(TXP)

and :

PN = {P,T,A}

In a PN there are two sub sets call pre sets and post set. These are the elements that make up the arcs. We define these as follows:

preT = {P: (P,T) F}

and

postT = {P: (P,T) F} where F is the set of all acceptable P,T pairs.

Consider a sample Petri Net. In this example there are four places, similar to states or presentations in the original model description. There are five transitions. These transitions represent the accumulation of all the steps that are required to be accomplished to get from one place to another. The transitions control the movement from one place to another, whereas the places are less active players in the PN model.

We have also identified the pre and post sets by the use of the terms aij, where we use a pre and post notation attached.

We can now introduce the concept of a marking, which is a way in which movement can bet determined around the PN> A marking is simple a set of integers that are assigned to a set of places, and an algorithm that details how those integers are propagated around the network. Let us define a marking as follows;

Definition: A marking is an n tuple, {bl,/bn}, where bk is a binary integer, 0,1, and;

M = {b1….bn}

where;

M: P-> {b1….bn}

Further we define MO as the initial marking and Mf the final marking.

Firing of a PN consists of taking one token in each preT and adding it to the token in postT.

we can now further define the actual execution algorithm for a PN.

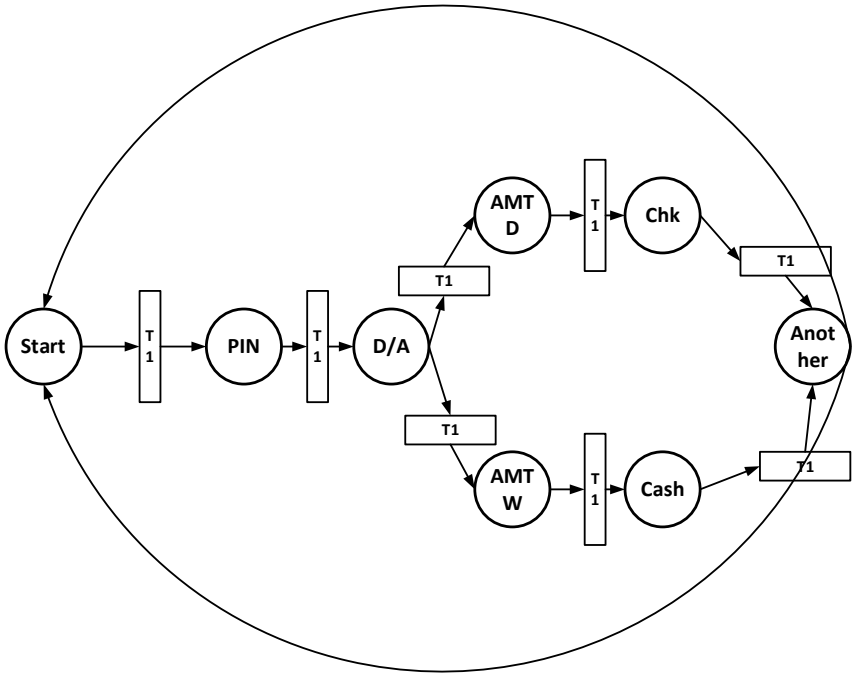
Definition: The execution algorithm for a PN consists of the following steps:

(i) Any Ti is enabled when ALL of its input places contain one or more tokens.

(ii) A transition Ti, that is enabled; can fire and when it fires it removes one token from each input Pk and places a single token in each output Pj.

(iii) Once the firing occurs, the tokens are repositioned and a new marking occurs. The marking sequence may be given by M(0), M(l), , M(n),...

It should be noted that there are two non-determinitsic token transitions that can occur. These are at the deposit/withdraw place and at the Another Transaction? place. In this case we shall leave them deterministic. In latter parts of this chapter we shall sow how we handle this probabilistically.



m1	m2	m3	m4	m5	m6	m7	m8
1	0	0	1	1	0	1	1
1							
0							
0							
1							
0							

In the transition of these states, we see that M(0) is the initial state and that initial state may be anything that we wish to start with. From an initial M(0), we generate the sequence of markings, M(l),....,M(n),... These are generated from a generic state transition function;

M(l) = G(1,0) M(0)

or in general we have;

M(k+1) = G(k+l,k) M(k)

The transition function can be calculated based on the transition elements and the arcs. We leave this detail to the problems at the end of the chapter.

Having defined the concept of the PN we can now extend this to the concept of a Hypertext. This is developed in the context of the work of Stotts and Furuta. It is as follows;

Definition: A hypertext is an n-tuple that is as follows;

H = {PN,C,W,B,Pl,Pd} where;

1. PN = any Petri Net

2. C = a set of document contents. This may include any set of text, graphics etc that may make up the hypertext document.

3. W = a set of windows. These windows may be ordered or otherwise.

4. B = a set of buttons. These are any actions that may create a response on the system.

5. PI = a logical projection of the document.

6. Pd = a display projection of the document.

Using the concept of the PN Hypertext we can determine several important solutions to problems concerning Hypertext. These problems are:

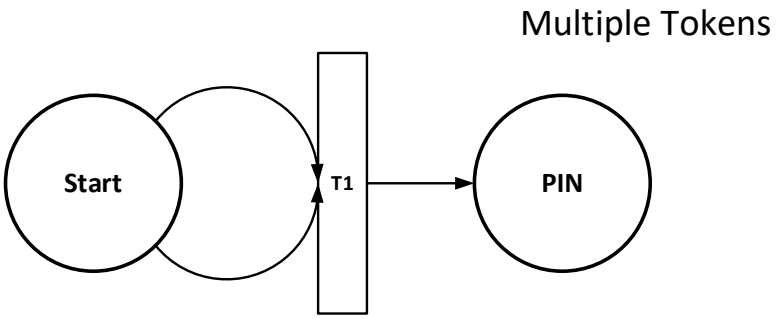
1. Display Complexity: Using the PN graph and the Hypertext adjunct, it is possible to determine the number of hypertext windows that are simultaneously need to display the information. This is the number of marked places in the PN associated with the Hypertext.

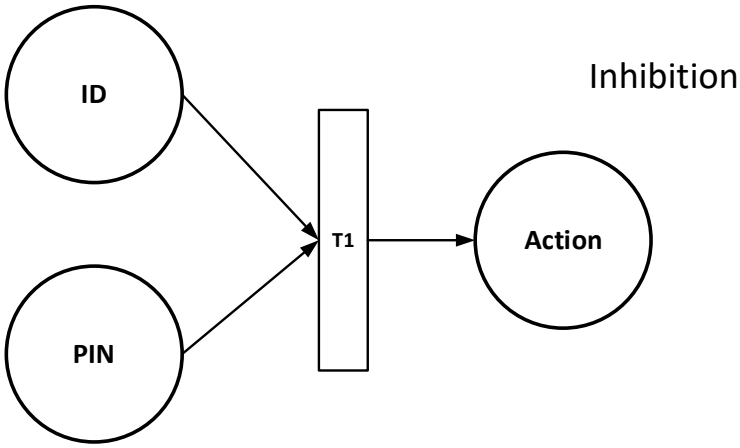
2. Path Size and Synchronization: Using the PN formulation, we can determine the length of the path needed and the level of concurrent path synchronization.

3. Reachability of Places: This is performed in developing the set R(M(0)), the reachability states of the initial state.

There are several extensions to the Petri Net concepts that we shall develop. These will become essential as we develop the model for the sizing of the source characteristics.

Definition: A PN place Pn has multiple arcs. The threshold Tj associated with Pk fires if and only if there are two tokens in Pk. In addition, only one token is passed onto Pn.





Definition: An arc is called an inhibited arc, if and only if all the arcs contain markings, except for the inhibitor arc.

Definition: If we let R(MO) be the reachability set of PN, with initial state MO, then the dead markings are those sets in R(MO) that go nowhere. Namely;

M(DS) is a dead set iff;

MDS(k+l) = G(k+l,k) MDS(k)

Definition: PN is called a safe net if the number of tokens <= 1 for all Pj and for all R(MO). We further define PN as STRICTLY CONSERVATIVE (SC) if for all M in R(M(0));

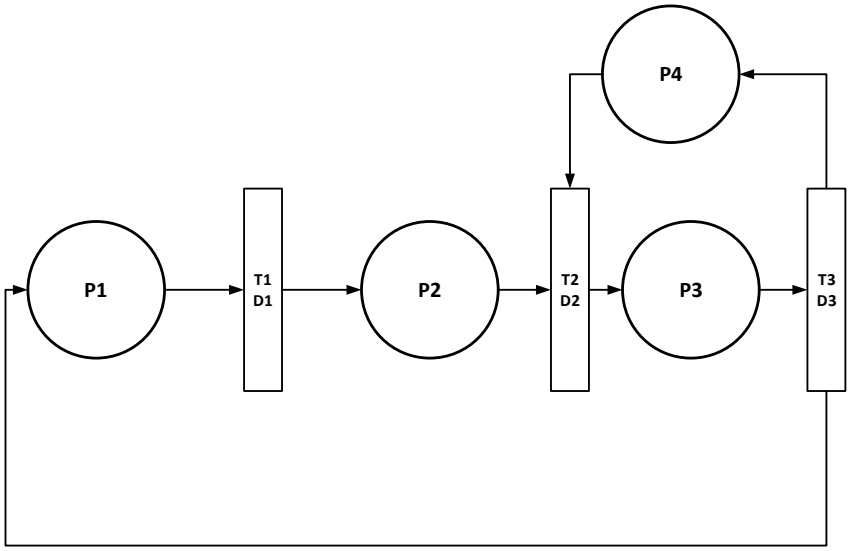
sum of tokens is constant

If PN is SC, then if C(x) is the cardinality of the set x, we have:

$\sum C(preA)\ C(preA) = \sum C(postA)$ i=l i=n

We now introduce two new concepts. The first is the timed PN and the second is the stochastic PN. The timed PN introduces the concept of timing for the PN concept. The stochastic PN allows for the generation of PN transitions that are stochastic in nature. Specifically we have introduced the concept of random timing that is the basic element in determining the ultimate source characteristics.

Definition: A timed Petri Net, TPN, is an n-tuple where D is a set of delays at each of the transitions before firing. Specifically, if we look at the example, we see that at transition Tk, we have a delay Dk that is in place before the transition fires.



Let us expand on the concept of the stochastic PN. In this case we expand on the TPN by applying the random variable to the timing at the transitions.

Definition: A stochastic Petri Net, SPN, is an n tuple that is:

SPN = {P,T,A,M(0),L}

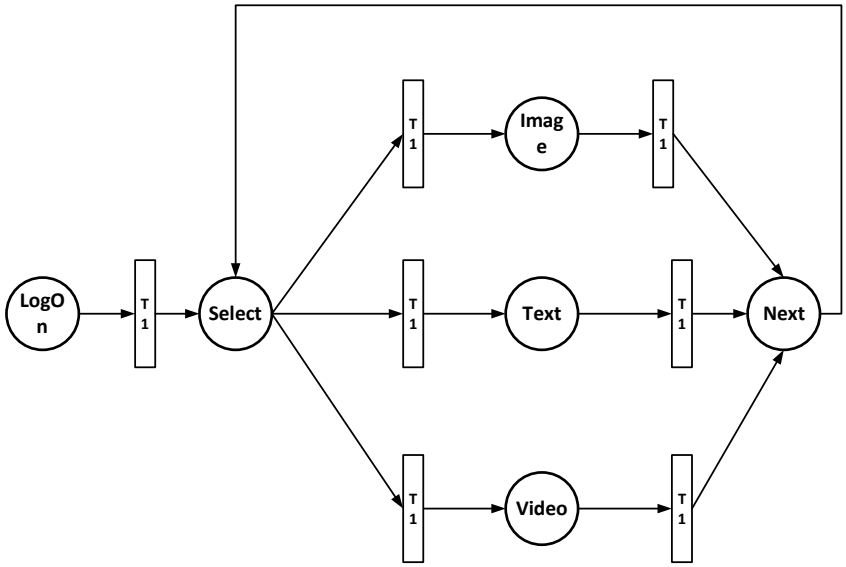
where L is a set of random delays at the individual transitions.

We can then define the probabilities associated with the system going from one state to another, or even the probability of being in any state. In particular it can be shown that the SPN is isomorphic to a Markov chain.

Finally, we introduce the structure that we begin this section with. Specifically, we introduce the concept of random transition path selection, as well as random delays at transitions. We also

introduce the input and output of data at the time of each transition. We first introduce a simple example to develop this concept.

Example: Consider the sample system that depicts pictures, text, or video. At two nodes we have options that could occur, one allow for the choice of the three images, the second is to decide if the next step is logging off or of starting over again.



What we see in this SPN is that there are non-deterministic choices for these two transition nodes. To alleviate these non-deterministic problems, we introduce an random selection choice at each node. Thus we can define the probability of choosing video as ql and that of text as q2 and that of video as q3. We note that the sum of the q's is unity as expected.

In addition, when we decide to see a picture, when the transition is activated, it generates an output that is sent to another similar machine. The output is a packet of information of certain length. In a similar fashion, there is an input packet that contains the information necessary to regenerate the picture requested.

Definition: A Source Dynamic Model (SDM) is an n-tuple that is characterized as follows:

SDM = {P,T,A,M(0),L,Q,I,0}

where the first part of the SDM is a SPN, and;

Q are the transition node probabilities

I are the inputs that occur at each transition when activated

0 are the outputs that occur at each transition when activated.

We shall use this model for the development of the source characterization.

4.4.2 Performance Analysis

The modeling of the multimedia source will require the use of probabilistic techniques that are found in many other areas. In particular we shall show that the end user interface state diagram ca n be modeled effectively using the finite state Markov state machine model and that statistics on the overall state averages can be readily drawn from the technique.

We shall be focusing on the development of performance issues as well with the development of the source model. These typical performance issues are:

1. Response Time
2. Average Holding Time
3. Number of Transitions per Unit time
4. Stability of the state protocol
5. Source generation rate

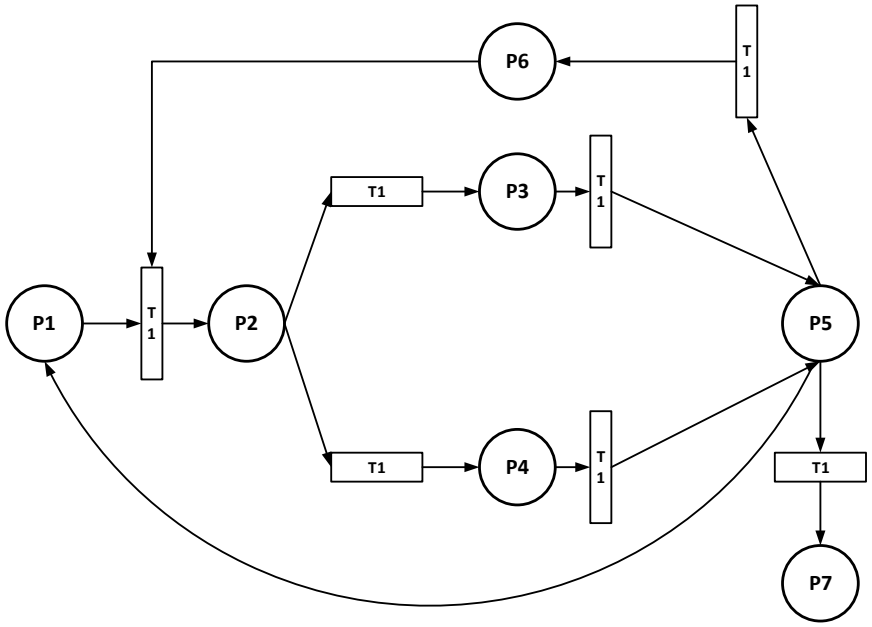
4.5 MM SOURCE MODELING

Multimedia source modeling combines the factors developed in the interface model and combines them with the model development that we performed in the last chapter. We can therefore takes a session interaction at one user, and develop the types of input to a communications network that would be anticipated from such as system source. We can then further expand this source to include a collection of sources to encompass a complete array of users that can provide the cumulative load on the total multimedia communications network.

The source modeling will include the combining of the elements developed in the system state diagrams with the additional factor of image sizing and packaging within the dialogue of the end user session.

4.5.1 Model Elements

It contains the elements that we have developed in the SDM system. It is based on a standard Petri Net models and we have introduced the timed transitions and the stochastic selection of transitions at the places indicated.



To develop the source model, we need to do the following steps:

1. Define and develop the PN for the characterization of the Hypertext interaction with the end user.

2. Assign probability distributions to the firings of the transitions. Typically, exponentially distributed random variable are adequate for this assignment.

3. Assign transition probabilities to each set of non-deterministic transition. A non-deterministic transition is one in which an arc from a place to a transition occurs several times from one place. The transition probabilities are for each non-deterministic place, Pj, and we have the transition from Pj to Pk or Pj to Pn. Define the transitions:

q jk = P[to Pk | from Pj]

and note that for each j:

qjk = 1

k Hj

where Hj is the set of all transition places from Pj. 3.4.2 Model Structures

The source model characterized in the previous section combines the results that we had developed in and those that we have developed in this chapter. In particular, we can see that

the source can be given in terms of a fully stochastic model. Let us begin by modeling the input and output responses for the SDM. These responses are the corresponding input and output traffic on the communications network.

The input and output messages consist of the following possible elements:

1. Images
2. Stills
3. Video
4. Voice
5. Text
6. Graphics

We can now characterize each of these in terms of bits of information, and furthermore characterize them in terms of data rates, namely bits per second.

Since the SDM is a stochastic Petri net, which is in turn a timed Petri net, we know that responses occur a set of define, albeit random, times. We call these times {LI,...,Ln,..}. Thus we have a sequence of input and output messages. Let us define Mi(t-Lk) as the input message at time Lk and Mo(t-Lj) as the output message at time Lj.

There are two classes of sources that we have developed. They are continuous and transient. A continuous source is one that never stops and continuously generates input and output messages. A transient source is one that does not last indefinitely and has a continuously decreasing probability of Lm continuing. We can provide a more definitive definition as follows.

Definition: A place Pterm is called a terminal piece if no transitions emanate from Pterm and there are no input or output messages associated with Pterm.

Definition: A SPM is said to be continuous if there is no Pterm.

Definition: A SPM is said to be transient if there exists a Pterm and if for all M(0) Pterm belongs to R(M(0)), and M(k) converges to Pterm as k increases. Specifically;

Let M( ) = {x,x,, z}

where z marks Pterm and can take 0,1. Let:

M(oo) = {0,0,,0,1}

Then M(k) approaches M(oo) with probability approaching 1 as k approaches oo.

We can define the input and output message streams as follows: mout(t)

= mo,1(t-Ll)+mo,2(t-L2) ++mo,k(t-Lk) +..+mo,n(t-Ln)

min(t)

= mi,1(t-Ll)+mi,2(t-L2)+...+

where:

mo,k(t-Lk) = response, request, image etc.

which depends on the probability assigned or directly upon the place transition characterization.

3.4.3 Model Performance Analysis

We can now take the models that we have developed and convert them onto source models for the purpose of systems performance analysis. PI is the initial state and P4 is the final state. We associate the time delays Lk with each transition.

We start by enumerating all of the paths through the network. A path is an ordered sequence of places that define the network. Associate with each path is a probability of occurrence. We assign to the nondeterministic choice at P3 by assigning the probability of p to going from P3 to P4 and q or (1-p) to going again to PI. With this we can assign the following:

Path Probability

P1,P2,P3,P4 p

P1,P2,P3,P1,P2/P3,P4 p*q

PI,P2,P3,PI,P2,P3,PI,P2,P3,P4 p*q*q

etc. We can readily add the probabilities and see that it adds to unity. Associated with each path is a duration Dk, where:

Duration Probability

L1+L2+L3 p

L1+L2+L3+L4+L1+L2+L3 pq

etc. From this we can determine the average duration of a transient SDM.

We define E[L] as the average path length in seconds, where it is the length of each path times the probability of that path.

In a similar fashion, we can generate a model for the total number of input and output bits generated in this network. We do this again by following the progress through each of the paths in

the network. We can then define the average number of input and output bits as;

E[I] = average input bits

E[0] = average output bits

Using this and the source duration we have the average input and output rates of the source;

Ri = E[I]/E[L]

Ro = E[0]/E[L]

If we have many such sources, then we can generate an average source rate by knowing the total number of sources, determining how frequently they are activated, and then averaging this over the individual sources own statistic. We expand on this in the problems.

4.6 CONCLUSIONS

This chapter provide the designer with the second element that is necessary for the development of the source model in a multimedia context. It allows for the characterization and development of a multimedia stochastic source and the specification of that source in terms its average characteristics as well as a full probabilities characterization. We shall see how this concept will be latter integrated into the overall system design and evaluation.

The key issues to be obtained from this discussion are those that relate to both the design and evaluation of the end user interface. Many works have been written on the end user design factors but as we continually see there are all too often designs that do not provide for ease of use and simplicity of access. There is the continual battle to provide the end user with all the flexibility that could be desired while at the same time increasing the complexity of the interface.

5 MULTIMEDIA STORAGE

The storage of multimedia objects is a major issue not only to the basic archiving of the data but goes to the heart of the issue of multimedia characterization. As we have discussed before, the ability to do more with complex multimedia objects than just raw digitization will represent the major capability to deal with complex abstractions. Thee complex image abstraction concept is the heart of multimedia storage.

This chapter discusses the ways in which complex multimedia elements can be compressed and reduced to simple data elements. It also addresses the more complex problem of how best to abstract elements of images and to combine them in a fully interactive multimedia session.

Storage is one of the elements that are key to the overall design and performance of the multimedia communications environment. Often the concern is the availability of communications channels that are fast enough to bring the images to the end user. The problem in reality is not the communications channel but the storage element. In this chapter, we focus on the individual storage elements and blend them together into a full multimedia storage capability. In addition we focus on the system issue associated with the system performance and sizing that is the driving force in all of the design tradeoffs in this text.

5.1 STORAGE ARCHITECTURES

In this section we present an overview of the key characteristics of storage devices and in addition provide the key factors that go into the performance analysis of the storage systems.

Storage Elements

Storage Access Characteristics

Storage Performance Parameters

The overall performance factor for the operations of a storage system is the access time for a specific file element. In a multimedia system, the performance factor is the access time for the compound multimedia file element. There are four major elements of the access time. These are:

Bus access speed and time

Memory unit latency access time

Data element seek time to find the data.

The data transfer time from the medium to the memory display device.

Multiple Storage Techniques

We have just completed the analysis of system that have single storage media and the performance that results. When dealing with

multiple media we are dealing in two dimensional of further integration. These dimensions are:

Image context integration, integrating voice, video, text, images etc from differing multimedia sources.

Storage device type integration, taking multimedia storage from multiple devices at the same time and creating a compound multimedia storage construct.

5.2 STORAGE ALTERNATIVES

There are many storage media that are available for the temporary storage of the multimedia elements that are used in an operational context. These media are in range for very fast but not grate in extent, to fairly slow but of high density. One of the driving factors for the choice of the specific storage medium is the cost per bit of storage and its accompanying access time. In this section, we shall not focus on the cost issues since they are generally so volatile but urge the reader to review the issue whenever developing a system design.

High Speed Memory (RAM)

Disk Storage

CD ROM Storage

Tape Storage

5.3 FILE FORMATS

Having the specific storage device is one of the elements in the overall storage design problem. The second, and equally important, is having the proper file formats. There are many types of file formats for the efficient storage and retrieval of multimedia data elements. Clearly, it should

be obvious that the storage of voice, video and images may require different file formats from both the perspective of placing the information in storage and in ultimately retrieving it.

5.4 MULTIMEDIA FILE RETRIEVAL ALTERNATIVES

Having discussed the overall issues, technology and file structure, we now will develop the overall system view of a multimedia file receival system. In this section we shall focus on the overall implementation of the architectural alternatives and develop methods and models for the analysis of performance and sizing of the multimedia file structures.

Storage Devices

Access Techniques

5.4.1 Performance Measures

5.4.2 Access Optimization

Having developed the measures of overall, system performance, there are several issues that relate to the optimization of the file system architectures. We have noted that several issues aware key to the design and performance of a multimedia system. These issues are:

Access time per element

Multimedia access time and compound access integration time

Consistent file naming and formatting of data.

5.5 CONCLUSIONS

In this chapter we have focused on the element of the multimedia communications system that stores and retrieves the data elements. The key issues are those that relate to the technological alternatives and the concerns relating to the performance of the system as an overall communications system. The issues of storage capacity, the layout of the specific files and the analysis of the access time to retrieve files are key to the development of an efficient overall system. We shall use these result as an element of the overall system design latter in the text.

6 NETWORKS AND MULTIMEDIA

The telecommunications network has evolved from a structure that was initially imbedded in a regulated monopoly through one that has been dictated by judicial mandate. The natural economic forces that tend to work in other market segments have been deliberately left out of this market. The result is a fragmented and less than efficient market for the delivery, development and expansion of services to the end users. There are however, alternatives that may allow for a repositioning of these current structures and permit a restructuring of the current communications market infrastructure. This paper develops a set of alternative architectural constructs, integrates several technological trends, and describes multiple evolutionary paths. The structure and viability of the paths are discussed in terms of their economic viability. This paper also discusses the potential and evolution of broadband networks and their role as infrastructure elements in a national network.

6.1 INTRODUCTION

Networks have been developing in various forms since the time of divestiture, in response both to the opportunities afforded by deregulation as well as by the needs of the users themselves. There has evolved a clear lack of cohesiveness to the design and application of the network schemes and the evolution seems to presage a movement towards networks optimized for usage by a defined and bonded collection of shared users. The National Research and Education Network (NREN) concept clearly falls into this category. To better understand the implications of such evolutionary trends and to better develop a base of knowledge for the development of effective policy in this area, it is necessary to have a construct or model for these evolving networks.

The NREN concept is an evolutionary progression of networking capabilities that starts with the introduction of a shared data network at the 1.5 Mbps rate and then transitions to a 45 Mbps set of rates and then in its third stage it acquires a Gbps capability. The major focus is on that later stage capability.

It is the development of this area that has been considered as the infrastructure portion of the effort. To quote from an OSTP report, (See Kahin, p.4);

" The NREN should be the prototype of a new national information infrastructure which could be available to every home, office, and factory."

This raises the expectations from that of a network to that of an infrastructure. In this paper, we shall develop the concept of infrastructure against the context of communications networks. We

shall address the issue of NREN and its counterparts being either infrastructure or merely another network. For in the same paper, Kahin goes on to define the NREN as;

" Despite the name, NREN is not conceived as a centralized national network...The NREN is conceived as more coordinated than the present Internet...the vision of the NREN generally includes eventual transition to commercial users."

Thus there are two views of the NREN; that of a prototype network and that of a fully operational entity.

There are similarly multiple views of many of the types of networks that may be developed for the purpose of developing national expertise in the areas of broadband communications. The views, as indicated, flow from those that view the need for a purely research oriented network to those who see the current need for a fully operational physical national infrastructure.

This paper addresses several of the fundamental issues that may assist in resolving the issues presented by these network alternatives. Specifically, we address the concept of infrastructure, and describe its multiple embodiments. We then develop the concept of architecture and how that it includes several elements and that a network can be viewed only in the context of its architectural embodiment. We then develop the major exogenous drivers for any network, the end users. Finally, we consider the elements of policy and how policy may be developed in the context of the evolving network world views.

There are several questions that we investigate in this paper. First, what is a network and what is a network infrastructure, and within that context, how does one create a broadband, high data rate, network infrastructure. Second, does such a network capability already exist and if so how does one assess and use it as is. Third, what are the goals that are achievable with such an infrastructure network, in terms of international competitiveness, establishing an information infrastructure, or in terms of a national asset.

The main theme of this paper builds on these questions. Specifically, it will be argued that there are basic architectural alternatives for network designs. The existence of these different architectures is based upon a world view and a set of technologies that enable its embodiment. We argue that there is a fundamental change in the world view, one from a truly hierarchical environment to one that now empowers the end user. The result of this change is a fundamental change in the operative network architectures, moving from the world of large scale infrastructures to those of multiple overlay networks.

The current dominant carriers in the telecommunications market are generally regulated on the theory that they have monopolistic power in their respective markets by either a direct or implied

exclusive franchise to provide services to the customers in those markets. Furthermore, market control allows the dominant carrier to exert price control as either a monopolist in the pure sense or at least as a simple oligopolist. It is the concept of the "Bottle Neck" that has dominated not only policy but also the flow of technology and services to the market. In this paper, we demonstrate that the dominant carrier concept is degenerating in many areas. It is becoming an environment of multiple network providers, where the functions of the classical carriers are now becoming distributed directly to the end users. Furthermore, we shall argue that as a result of the technological changes allowing for multiple carriers, that the current regulatory strictures on the telecommunications market are not only inhibiting to development but are the fundamental essence of the loss of international competitiveness.

The policy issues associated with these classical systems are based upon the fact that, in the past, there has been a perceived lack of growth and innovation, a set of barriers to entry on the part of new innovators and price competitors, and a set of price structures that inhibited the entry of competitors. We demonstrate in this paper that there are now a multiple set of alternatives for the network user. The alternatives are based upon a revised view of what a network is as well as major changes in the technology of networking.

Thus, these changes require an expanded development of new policies for these network infrastructures.

Networks have been viewed by some as an infrastructure such as highways, educational systems, and the military. Typically they have been thought of that way because they serve the general good, they require significant capital investments that are frequently beyond the bounds of most single users, even corporations, and because they must deal with both national and international elements of our society.

Thus, the three reasons for the infrastructure view are the general good, capital intensiveness and global in scope. In this paper, we argue that the concept of networks and infrastructure must be revised and expanded. That networks such as the NREN, are only one of several embodiments of a network, and that infrastructure must be understood in the context of the new world views of networks. Further, we argue that technology is changing what we can do with the network architectural elements and that this added use flexibility to create subnetworks that are optimized to meet their specific economic driving forces, makes the very concept of a single physical network architecture obsolete.

Specifically, market demands require that the networks have more customized structures and that in fact the general network that meets everyone's needs is both counterproductive and adds additional costs.

Second, technology allows for segmentation in the network, using a commodity infrastructure of fiber to allow the end users to have the use specific designs that they need. For example, if we allow fiber to be used at its full potential to the end user, rather than as a segmented and compartmentalized structure matching a voice only world, this may free both market and technical forces to attain the segmentation.

Third, globalization has already occurred with the globalization of the business market.

There are currently five players in the networking area. They are the Government and its own networks, Public Switched networks provided by the common carriers, Private Networks, frequently called Bypass carriers, CATV networks, primarily carrying entertainment and not information, and customer specific networks, such as those of IBM, DEC and other large information intensive companies, as well as the regional educational networks, such as NEARNet and NYSERNet. The five network providers and generally depicts the gross differences between each of them. It is essential to understand not only the microstructure of each of the players in this map, but to also understand that each of the players has a dramatically different world view of what they are to achieve in their networks. In this paper, we shall examine each of these players in the context of the new world view of networks In this paper, we will prove the following thesis:

The use of a centrally conceived, non-market generated and driven, means to develop an economically productive implementation of a new electronic broadband communications network is not only an ineffective act but also is counterproductive to the national economy.

We prove both sides of this thesis in the body of this paper. Moreover, we demonstrate that the broadband network is and must continue to be driven by end users motivated by such basic economic drivers as value creation. The evolution includes not only the basic transport features but the entire complement of elements necessary for a viable communications and business entity. The current environment is such that competitive market forces exist, inhibited only by regulation mired in an obsolete world view that must be adopted to the new paradigms of implementation.

The process that we take to prove this thesis is as follows:

First, we develop the concept of architecture and show that there is both a philosophical and material structure to the concept of architecture. The key observation of this first step will be to note that our understanding of communications architectures is dominated by a world view created by the available paradigms or technologies.

Second, we demonstrate that the underlying technologies have changed dramatically in the area of broadband. The change is dominated by the ability of the end user terminal to play a dramatic role in the network operations and that ability to have a fully distributed design. This technology change is reflected in new paradigms and thus agues for the development of a new world view.

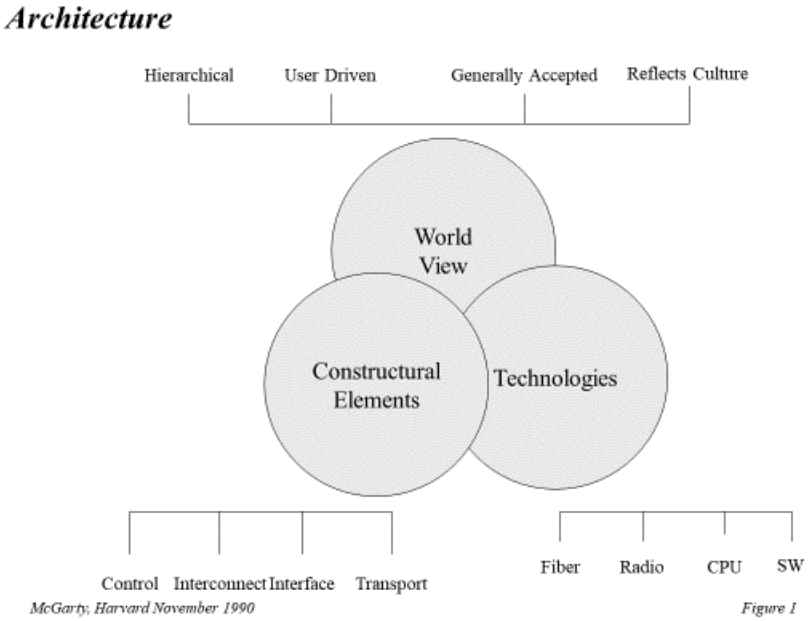
Third, we demonstrate that the market is the ultimate arbitrator. That value creation as a measurable and definable entity is critical to the success of any broadband implementation and that this value creation must be perceived by and understood by the customer. Without the customer as user, buyer and decision maker being the focal element of the process, the effort is doomed to failure.

Fourth, there are several players all trying to evolve in the direction of developing the basic elements necessary for a broadband environment. We review them in some detail and show that there already exists many of the elements in place for such a system and that another scheme will result in reducing the current competitiveness of these players and impact negatively on the economy as a whole These four steps build to the proof of the thesis.

6.2 ARCHITECTURES

The concept of a telecommunications architecture has been a cornerstone in the development of new telecommunications systems. However, the structural elements of these architectures have not played a role in the development of policies. In this section we will develop the concept of an architecture as a means to understand the network as both a market and regulatory entity, and

will provide a new set of perspectives for viewing the network in terms of a new paradigms and world views.



An architecture, first, requires that the underlying system be treated in terms of a set of commonly understood elements and that these elements have a clearly demarcated set of functions and interfaces that allow for the combining of the basic set of elements. The way the elements then can be combined, reflected against the ultimate types of services provided, determine the architecture.

An architecture, secondly, is driven by two factors; technology and world view. Technology places bounds on what is achievable, however those bounds are typically well beyond the limits that are self-imposed by the designer or architect in their view of the user in their world. This concept of architecture and the use of design elements is critical in understanding the paradigms used in the structure of information systems (See Winograd and Flores, pp 34-50, especially their discussion of Heidegger and Throwness in terms of design). World view is the more powerful driver in architecture (See Kuhn, pp 72-85). We argue in this paper that it is essential to develop a philosophical perspective and understanding of how to view networks. We argue with Winograd and Flores, and in turn with Heidegger, that we must be thrown into the network, to understand the needs of the users, and to understand the structure of the paradigms that are used to construct the world view.

To better understand the importance of an architecture we develop the concept of the historicity of architectures based upon the work of Kuhn and ten that of McLuhan. Kuhn begins his thesis of how scientific revolutions occur by the introduction of the concept of paradigms. He defines

these as (see Kuhn p. 175);"...the term paradigm is used in two different senses. On the one hand, it stands for the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community. On the other, it denotes one sort of element in that constellation, the concrete puzzle-solutions which, employed as models or examples, can replace explicit rules as a basis for the remaining puzzles of normal science, The first sense of the term, call it sociological, ..., “

The concept of a paradigm is in essence the collection of current technologies that we have at hand for the network and the ways we put these elements together. New paradigms result from new technologies. New technologies allow for the placing of the elements together in new ways. Kuhn, then goes on to demonstrate that the world view, that is how we view ourselves and our environment is based upon the our acceptance of these paradigms, as either collections of techniques and technologies or as collections of embodiments of these techniques and technologies in "examples". We then end to accept this as the way things are and should be. Then Kuhn argues, as the technologies change, changes in the paradigms do not occur in a continuous fashion but almost in quantum leaps. The new paradigms build and congeal until they burst forth with new world views. It is this model that we ague applies to the evolution of broadband.

It is this philosophical view, almost Hegelian in form, that is essential in understanding the underlying and formative changes in paradigms that will change our world view.

As a second perspective of the impact of technology as a dominant driver, we can refer to McLuhan and his development of the concept of media. Drucker has referred to the presentation of McLuhan's doctoral thesis and McLuhan is quoted as follows (See Drucker, p. 250):

"Movable type, rather than Petrarch, Copernicus, or Columbus was the creator of the modern world view.. "Did I hear you right," asked one of the professors as McLuhan had finished reading, "that you think printing influenced the course the universities taught and the role of the university, altogether?”

"No, sir, " said McLuhan, "it did not influence; printing determined both, indeed, printing determined henceforth what was going to be considered knowledge.”

This concept later evolved into the medium being the message. In our context it is the fact that both Kuhn and McLuhan recognized, albeit in differing fields and in differing ways, that fundamental changes in technology and technique, call it paradigm or the medium, will change the world view, also the message.

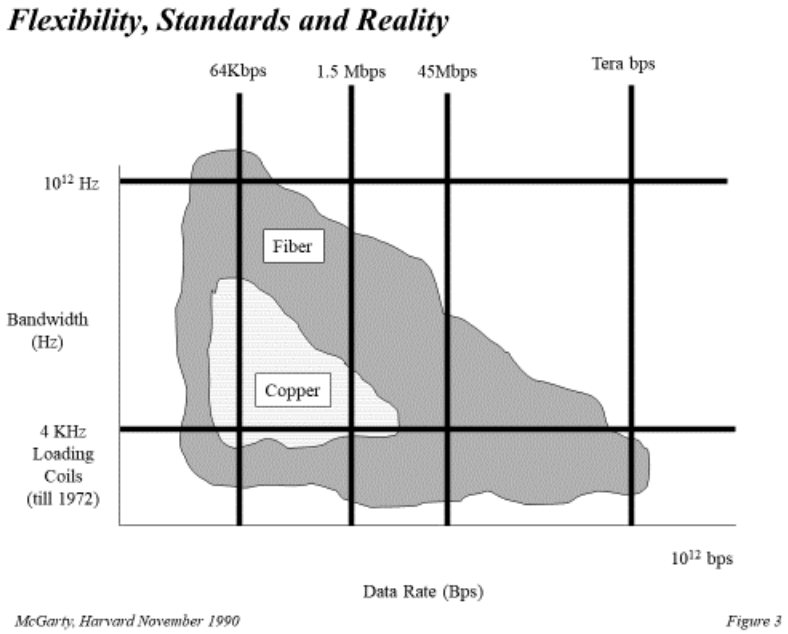
It is the importance of understanding the change in the technology, its function and evaluate the possible change that this will have in the world view. It will be argued, that much of the thinking

in the current broadband areas, NREN in particular, is based upon outmoded techniques and structures, and that a differing world view will evolve.

Thus, architecture is the combination of three parts; the common elements, the underlying technology and the world view. The conceptualization of architecture as the amalgam of these three elements. We shall develop this construct more fully as we proceed.

The concept of a world view is an overlying concept that goes to the heart of the arguments made in this paper. To better understand what it implies, we further examine several common views and analyze the implications of each. If we view our world as hierarchical, then the network may very well reflect that view. If we further add to that view a bias towards voice communications, these two element will be reflected in all that we do. The very observations that we make about our environment and the needs of the users will be reflected against that view. As an external observer, we at best can deconstruct the view and using the abilities of the hermeneutic observer, determine the intent of the builder of the networks. (See Gadamer's interpretation as discussed by Winograd and Flores, pp 27-30. Also see the historical context of the hermeneutic approach in the sciences as discussed by Greene in Depew and Weber, pp 9-10).

Take, for example, the use of twisted pair, pairs of copper wire, to transport telephone traffic. For years it was implicitly assumed that this transport medium was limited to 4,000 Hz of bandwidth, that necessary for an adequate quality voice signal. Specifically the world view was that of a voice network that was to be used for voice traffic only. Ten years ago, this was a true limitation, since the transmission was forcefully limited to 4,000 Hz by inductive loads or coils on the telephone lines, assuring that you could do no more than the 4,000 Hz of bandwidth. Then, there was a short period in the mid-1980s, when Local Area network manufacturers found that you could transmit 1.544 Mbps over the common twisted pair, and that data was viable in what was assumed to be a voice only medium. What had been almost religiously believed to be a limit was found to be untrue. Then with the introduction of digital switches, the old "inductive loads" were returned with the switch now limiting the data to 4 KHz or 64 K samples per second. The world view of a voice only network took hold again, but this time in the context of a data rate limitation, rather than a bandwidth limitation. In the early 90's there is another attempted break out of the world view and to put 100 Mbps on twisted pair, so called FDDI circuits. Again, due to the limitations on the part of the network as a voice dominated system, the world view keeps this high data rate capability on the customer's premise only, and not the network.



Here we indicate the two dimensions of information transport, bandwidth and data rate. The designer of the transport facility may limit the data rate by selection of signaling format or delimit bandwidth by filtering. It encompasses a large capability of either providing bandwidth or data rates to the user. The two limiting world views are indicated as two solid lines, one at 4,000 Hz and one at 64 Kbps. Both are voice only world views. We can readily see, that with optical fiber superimposed the same issue of architecture dominated by world view may result. In the fiber case, the result may be a segmenting of the architecture along selected data rate lines, again formed by the voice world view.

Thus, architecture can be defined as the conceptual embodiment of a world view, using the commonly understood set of constructural elements, based upon the available set of technologies. For example, Gothic architecture was a reflection of the ultimate salvation in God in the afterlife, in a building having a roof, walls, floors, and windows, and made of stone and glass. Romantic architecture was, in contrast, a celebration of man, using the same elements, but some employing a few more building materials. The impact of the differences in world view are self-evident in the embodiments of the architecture. (See the discussions on the impact of world view on architecture in Wolfe. In addition see the cultural or world view impact on the Gothic architectures in Jantzen and in Toy.) Let us consider a second example of the impact of world view on architecture, specifically the difference between the ISDN architecture and the architecture embodied in Local Area Networks, LANs. ISDN is an architecture consistent with a voice dominated, hierarchical world view of single points of control. LANs are architectures of world views that reflect both end user self-empowerment and the environment of a data driven utility. The LAN embodiment as well as its extension in the CATV architecture of voice

communications using a LAN world view. This evolution in thought is critical to understand the impact of world view. The LAN is an embodiment of empowerment of the individual view, developed in the context of the 1960's and 1970's. The LAN concept, originating at such locations as XEROX PARC, was driven by the developers needs to enable and empower the end user with computing capabilities heretofore unavailable. Out of this view came the LAN architecture of a fully distributed system, using a coaxial transport mechanism to do nothing more than provide bandwidth. The transport mechanism is a broad enabler. The actual implementation of the details is done at the users terminal in hardware and software. This is in sharp contrast to ISDN, where the ISDN central switch does the enabling. In ISDN, bandwidth is not provided, rather it is a voice based data rate, 64 Kbps or multiple thereof. Consider this contrast in terms of how cable TV companies provided voice communications in the early 1980's. Both Cox and Warner, using variations on LAN technology, delivered a voice, video, and data service over the coaxial transport medium, by empowering the end users terminal, not by regimenting the transport network.

Technology also plays a very pivotal role in telecommunications. Alfred Kahn (1971, p 300), indicates that in the pre-divestiture period of the Bell System, the arguments for the needs of both vertical integration and need for monopoly control were based on technology. Specifically, there was a contention made by the Bell System that a single point of control to the network was essential. Also, it was argued that an adequate scale economy was attained only through a single monopoly. Indeed, given the state of technology of that time, the argument may have held. For in point, the loaded copper transmission capabilities allowed only limited transport, namely one voice channel per twisted pair. However, as we shall demonstrate, the underlying technology has provided a dramatic change in the underlying system.

Functions now provided by the network, may be more efficiently provided by intelligent Customer Premise Equipment (CPE). The question to be posed is; what is the role of the network, and how do we provide the dimensions of creative freedom to allow these new roles to evolve? To effectively approach this problem, we must first develop a canonical structure of a network.

6.2.1 Elements

There are four architectural elements in the telecommunications network. These elements are the control functions, the transport function, the interconnect function, and the interface function. We now provide further detail on these functions. It should be noted that these functions have evolved over the years in content and complexity. We view these elements in the context of a communications network that must support the most advanced current concepts in communications. Specifically, the world view adopted in this paper that lead to an interpretation of this architecture are:

(i) End users desire to have interactions in a real time fashion with images and other high resolution information that must be provided in a fashion that meet both time and resolution requirements (See Barlow).

(ii) The end user devices are extremely intelligent and complex and can operate in a standalone environment.

(iii) The users desire to operate in a totally distributed fashion. Data bases will be a different locations, users are at different locations and input output devices are also at different locations (See Dertouzos and Moses, and de Sola Pool pp 57-59 for details on these directions).

(iv) The network may provide different levels of service to different users. There is no need to provide universal service of full capability to all end users.

This view of the network will significantly influence how extensively we defined the elements and in turn will impact the combination of those elements in an overall architecture. All of these assumptions on the world view are different than before, in an all voice world. In this paper, we define a network as an embodiment of an architecture, in all of its elements.

The architectural elements are control, transport, interconnect and interface. The overall architecture of the element interrelationship and the elements of the functions of the separate elements. The details on each are described below:

Control: Control elements in an architecture provide for such functions as management, error detection, restoral, billing, inventory management, and diagnostics. Currently, the voice network provides these functions on a centralized basis, although in the last five years there have evolved network management and control schemas and products that allow for the custom control and management of their own network. Companies such as IBM, AT&T and NYNEX have developed network management systems that move the control from the network to the customer (McGarty, 87). On the sub-network side, companies such as NET, Timeplex, Novell, 3-COM and other have done similar implementations for local area networks, data multiplexers and other elements. Centralized network control is now longer necessary and in fact it may not be the most efficient way to control the network.

What is important, however, is that network control providing the above functions is an essential element for either a public or private network. Thus as we consider network evolution, this element or set of function must be included.

Control has now been made to be flexible and movable. The control function is probably the most critical in the changes that have been viewed in the context of an architecture. All buildings

need windows, for example, but where one places the windows and what one makes them of can yield a mud adobe or the cathedral at Chartres. The same is true of the control element. In existing networks, the control is centralized, but in newer networks, the control is distributed and empowered to the end users. The users can now reconfigure, add, move, and change their network configuration and capacity.

Let us briefly describe how the control function can now be distributed. Consider a large corporate network consisting of computers, LANs, PBXs and smart multiplexers, as well as a backbone fiber transport function. Each of these elements has its own control facility for management and restoral. Each has the capability to reroute traffic from one location to another, and the routing systems are programmed into the system as a whole. On top of these sub element control functions is built another layer of control that views the network as a holistic entity. This form of control has been termed a manager of managers.

It monitors all of the sub net elements and takes control if necessary. It is embodied in several independent controllers, each having the capability of taking control from a remote network. This form of organic network control has evolved in recent years and is now common in many corporate networks. In addition, this concept of the organic network was described in detail by Huber in the DOJ report to the U.S. Justice Department during the first Triennial Review of the MFJ (See Huber).

Transport: The transport element is provided by the underling transport fabric, whether that be twisted pair of copper, fiber optic cable, radio or other means. Transport should not be mixed or confused with other elements of the network. Transport is merely the provision of physical means to move information, in some form such as digital, from one point to another. At most it is expressed in bits per second and at best it is expressed in bandwidth only. Bandwidth as a transport construct is the most enabling. Transport does not encompass the need to change the information or to do any other enhancement to the information.

In the early regulatory cases such as the Above 890 Decisions in the microwave systems that were the precursors to MCI (See Kahn (II p12)), the Bell System argued that the technology of transmission limited the transport to only those companies that had the transport, interconnect and control. MCI on the other hand recognized that the customer was able and willing to differentiate these elements of the architecture and would segment them in a more economically efficient fashion. Specifically, in the early days of MCI, customers in the mid-west would select multiple transport paths and would do the control function on their own premises. In addition, the customers were willing to accept lower quality of service for a lower cost of service. The lower quality was reflected by possibly a higher outage time.

It could then be recognized that the horizontal scale economies of all of the network elements, including but not limited to transport, were actually diseconomies of scale in the market. (See Fulhaber for a discussion of a more detailed view of scale diseconomies in terms of the new architectural elements) Fragmentation and segmentation along architecture elements allowed for the growth and efficiency of MCI. The emphasis should also be made on the statement of the FCC Examiner in the MCI case who stated (Kahn II p 134), "MCI is a shoestring operation ... the sites are small and the architecture of the huts is late Sears Roebuck toolshed." It is prescient to note that the examiner used the term architecture for the microwave repeater sites when indeed MCI was changing the architecture of the network. This remark is more than just an embodiment of a metaphor.

In the current network environment, the issue of transport and its enabling capacity has again arose. This has been the case with the introduction of fiber. Fiber may be segmented for the user in terms of data rates or in terms of bandwidth. In the NREN, the three steps are all focused along the lines of increasing data rates, from 1.5 Mbps to 45 Mbps to Gbps. As we have discussed, bandwidth is the more enabling dimension, leaving the choice of data rate and data structure to the end user. This capability is best deployed by using a dark fiber network. The top network is a standard fiber network with repeater at periodic intervals. In current technology limitations these are necessary because of the losses in fiber transport. However, with the current state of the art technology, fiber can be strung for many tens of miles without such repeaters and still maintain adequate transmission capacity.

Thus the repeaters are not there solely as a result of fiber constraints on transport. They are also there because they enforce the voice regime of the voice based world view. Namely, the repeaters do not repeat data rates, they also repeat framing sequences based on 64 Kbps voice frames. Thus any work station must use 64 Kbps as the underlying data fabric. As an extreme example, NREN in its Phase 2 will provide 45 Mbps to the users. Regrettably, there is no 45 Mbps modem. That is, direct access to 45 Mbps is not achievable. It must be sub multiplexed to the equivalent of voice grade digital circuits. Thus the world view is pervasive in this design. The same is true as SONET protocols are used in upgrades to broadband ISDN, especially over an ATM switch (See Fleming for a discussion of broadband switching and the voice paradigm).

In contrast, dark fiber is the provisioning of an optical fiber to be used as the end user sees fit. It is the world view analog of the LAN. The LAN provides co-axial bandwidth of several hundred MHz whereas the fiber provides the bandwidth of GHz to TeraHz.

Interconnect: The interconnect element of the architecture describes how the different users are connected to one another or to any of the resources connected to the network and is synonymous with switching. Interconnection assumes that there is an addressing scheme, a management

scheme for the addresses, and a scheme to allow one user to address, locate and connect to any other user.

Interconnection has in the past been provided by the Central Office switches. As we shall discuss latter, this implementation of an architectural element was based on certain limitations of the transport element.

With the change in the transport element of structures allowing greater bandwidth, the switching needs have changed. Specifically, distributed systems and scale economies of the distributed architectures allow for interconnectivity controlled by the CPE and not the Central Office. As we shall show later, the advent of Local Area Networks and CATV voice communications are ones using distributed interconnectivity elements.

Again, Alfred Kahn noted (II, p 127),

"We have already alluded to the technological explosion in communications after World War II,...The case for a national telecommunications network monopoly has the following aspects. Aggregate investment costs can be minimized.. if the planning for the installation and expansion is done with an eye for the total system....Since any one of the 5 million billion possible connections that the system must stand ready to make at any point in time may be performed over a variety of routes....justifies the interconnection...completely dependent on its own resources alone."

This argument for interconnection, combined with transport and control (namely horizontal integration) was valid in 1970. It however is not valid today. They are separable functions and scale economies are in the hands of the CPE manufacturers not the network providers. In effect, there exists no monopoly in interconnect as a result of these technology changes. This is a dramatic change from 1971 and Kahn's analysis.

There are three general views of interconnection that are valid today; the Telcom, the Computer Scientist, and the User. The Telcom view is based on the assumption of voice based transport with universal service and the assumption of the inseparability of interconnect and control. The Computer Scientist view is based upon the assumption that the network, as transport, is totally unreliable, and that computer hardware and software must be used in extremis to handle each data packet. Furthermore the Computer Scientist's view of the network is one where timeliness is secondary to control. The Computer Scientists view has been epitomized in the quote, "Every Packet is an Adventure". This is said with glee, in that each data packet is set out across the network and it is through the best of hacking that the Computer Scientist saves the packet from the perils of Scylla and Charybdis. The third view is that of the user, who is interested in

developing an interconnect capability that meets the needs and minimizes cost. This is minimization of both obsolescence and cost strategy.

Processing cost or capacity is declining every year. Thus an investment must try to follow the curve. In a hierarchical view of interconnect, such as a large centrally switched network, the changes occur once every few years. Thus the lost cost or performance efficiency can become significant. In contrast, in an end user controlled environment, with a fully distributed architecture, the lost efficiency is minimized as technology advances.

Interface: The interfaces are the end users connection to the transport element. The interface element provides for the conversion from the end user information stream and the information streams that are used in the transport form of the network. For example, the telephone interface for voice is the analog conversion device.

Interfaces were originally called "Alien Attachments". In Kahn (II p. 140-145,) he discusses the history of the interface leading up to the Carterfone decision. The most significant position in CPE control was the Hush-A-Phone debate from 1921 to 1946. The Bell System at that time took total and full control over the quality of the delivery of the service of voice. The Hush-A-Phone company provided a mechanical cup device that could be placed over the mouthpiece of the telephone to assist in making the conversation more private. AT&T took the position that it interfered with the network and the quality of service and battled this for 25 years. Such is not the case today. CPE computer equipment has proliferated and the current costs for 9,600 bit per second modems are comparable to high end voice telephone devices.

Clearly, this fourth architectural element is separate and apart.

We have divided the network elements into these four categories to demonstrate that there are clearly four distinct and separable areas for growth and policy formation. Issues of regulation, due to potential monopolist control are always a concern, but it will be demonstrated that in all four there are economies in market disaggregation.

Network Providers

Type	*Customers*	*Architecture*
Public Switched	RBOCs	Hierarchical Centrally Controlled
CATV	Cable Companies	Local Branched
Customer	Local Campus	Point to point Point to Multipoint Branched
Government	Federal, State, Local	Point to point Point to Multipoint Branched
Private	Corporations	Point to point Point to Multipoint Branched

McGarty, Harvard November 1990 *Figure 2*

Natural monopolies have been studied by many, and in the context of utility regulation there are many key studies. In this context, Spulber has defined a natural monopoly as:

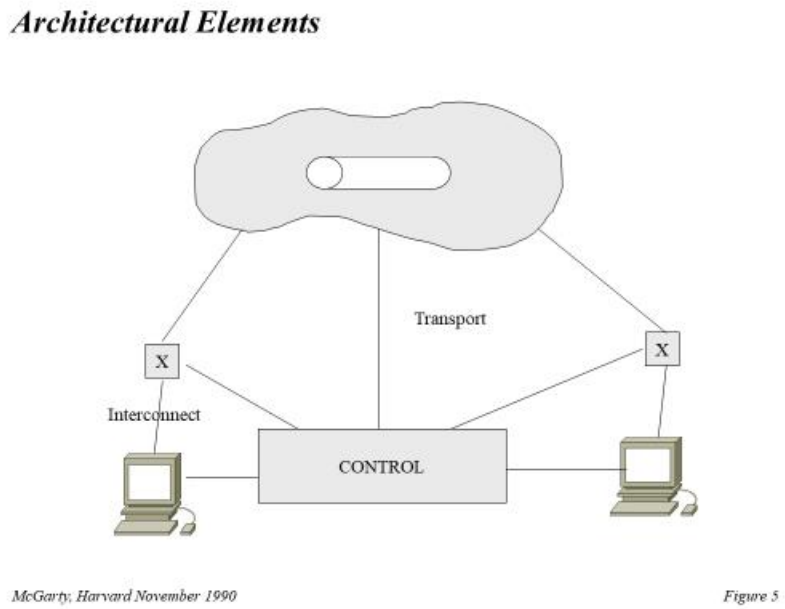
".. a property of productive technology, often in conjunction with market demand, such that a single firm is able to serve the market at less cost than two or more firms."

Natural monopoly is due to economy of scale and in the current architecture, elements of the monopoly concept no longer applies. Potential monopolist tendencies in all of these elements, separately, have been reduced by the ability of the end user to fabricate the elements of the network in a set of separable fashions. Together these elements clearly demonstrate no monopolistic power. The traditional theory on regulation has focused on the control over the transport facilities. As we enhance the network structures and provide differing forms of information transported, concern should also be focused on the other three elements (see Kahn p II 127). In addition, monopoly power even over transport was based on the users inability to individually justify the capital costs on the transport infrastructure. In certain cases, this is no longer the case, finding many users easily justifying payback in less than one year on new transport infrastructure

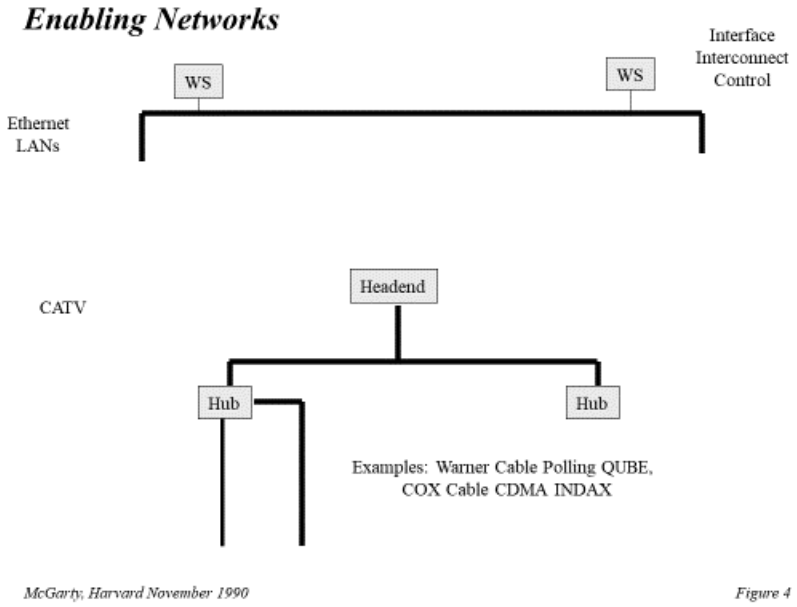
6.2.2 Architectural Alternatives

Is there a natural taxonomy for the set of network architecture alternatives? Do these present limitations on what can be done or are they extensive? Is there a natural limitation in the existing

architectures that prevent the new technologies from introducing the new paradigms to the communications world?



We address these issues in the context of several existing network hierarchies.

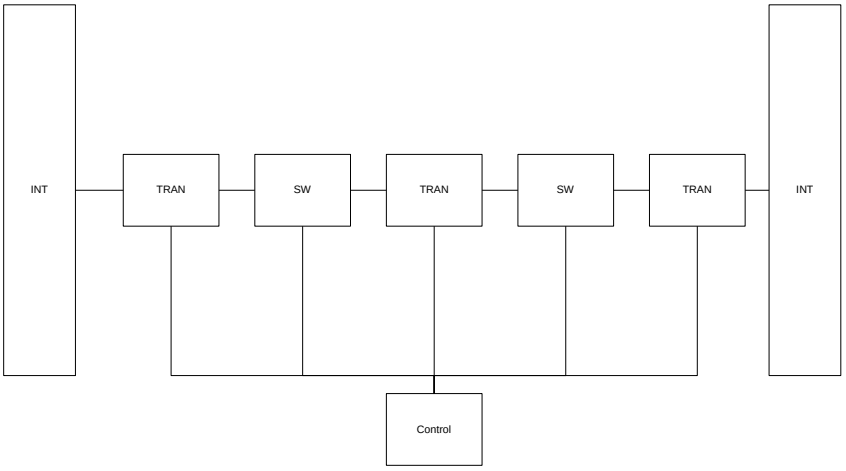


Hierarchical: The current network architectures are structured in a hierarchical fashion. As we have already indicated, there are historical and technical reason for this architecture. We show in

Specifically, we see the set of transmission schemes connecting from a lower level to higher ones. A path may or may not go horizontally. It may go vertically, all controlled by a single control at the highest level.

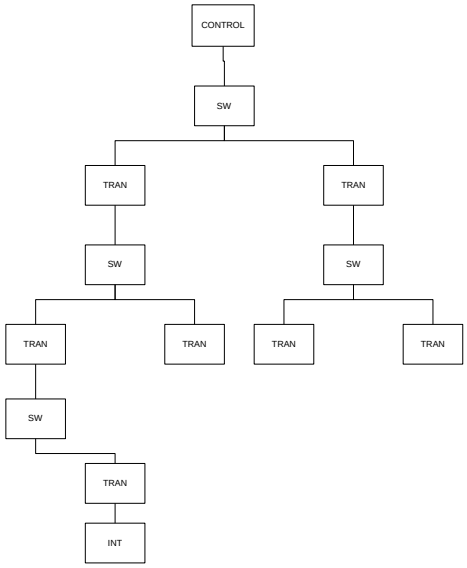
Centralized: A centralized architecture is similar to a hierarchical system in that the control function is centralized. However, the transport elements are not in a hierarchical format.

Centralized



McGarty, Harvard November 1990 *Figure 9*

Hierarchical

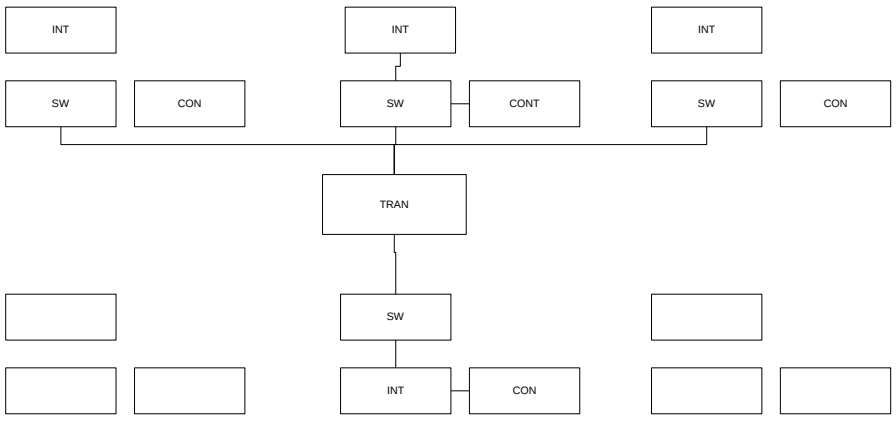


McGarty, Harvard November 1990 *Figure 8*

The hierarchical structure is no longer present, but there is a single point of control. The control element covers all other elements in the system. A typical example of this type of network is that of a large bank in a metropolitan area. Part of the network is the local ATM (Automated Teller Machine) network and the voice network for the bank. Each are separate but the bank controls both from a single point of control.

Distributed: The distributed system has distributed control, distributed interconnection and flat transport alternatives. Here we first note the reduction in concatenated switch and transmission elements. The network is much less dense and the switch is actually co-located with the interface. The LAN networks are typical example of distributed designs.

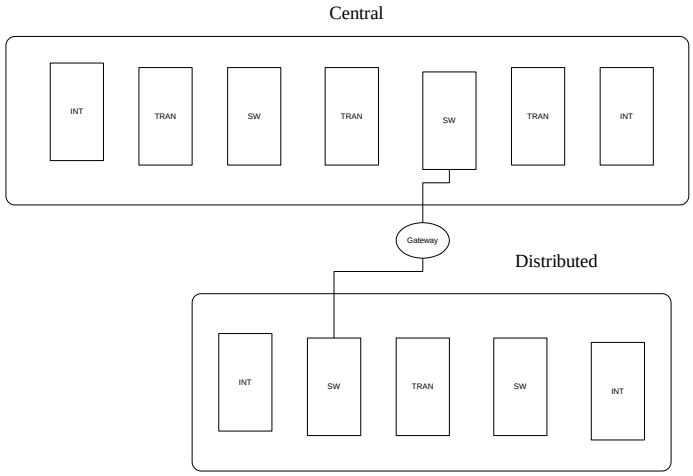
Distributed



McGarty, Harvard November 1990 *Figure 10*

Segmented: A segmented network is really a hybrid. Each segment uses a sub architectures that meets the requirements of the existing system but the networks are interconnected through standard interfaces.

Segmented



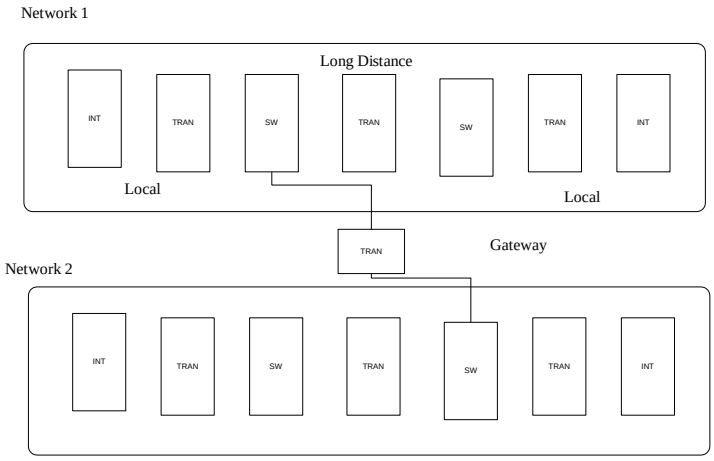
McGarty, Harvard November 1990 *Figure 11*

In this case we show that this network architecture is an amalgam of the first three. What is still common, however, is the partitioning into local and long distance nets. A typical example of this network is that of a large corporate network. Part of the network can be for the voice circuits, controlled at a single point and based upon use of both local and inter-exchange carrier circuits.

The second part of the network is the data network, again using both local and long distance carriers, and control from a separate location.

Partitioned (Local and Long Distance combined in a community of interest): In all of the above, we have assumed that local and long distance transport are separate. This is a world view dominated by the regulatory environment. We can see the segmentation along community of interest lines rather than along these more traditional lines. Thus one community of interest is a network for financial service companies and a second for a network providing service to the residential user. These each have all of the local and long distance services, but are now segmented by the user market or the community of interest. The sub architecture may be any of the above. This architecture allows for local and long distance in separate partitions. It says that you can segment the network by users not just by function. Had the MFJ understood users rather than functions, the results could have been dramatically different. An example of a Partitioned network would be that for American Express or Sears. It contains the set of local and long distance networks as well as subnets for specific distributed applications. However, each of these companies may have access to a separate public switched environment.

Partitioned



McGarty, Harvard November 1990 *Figure 12*

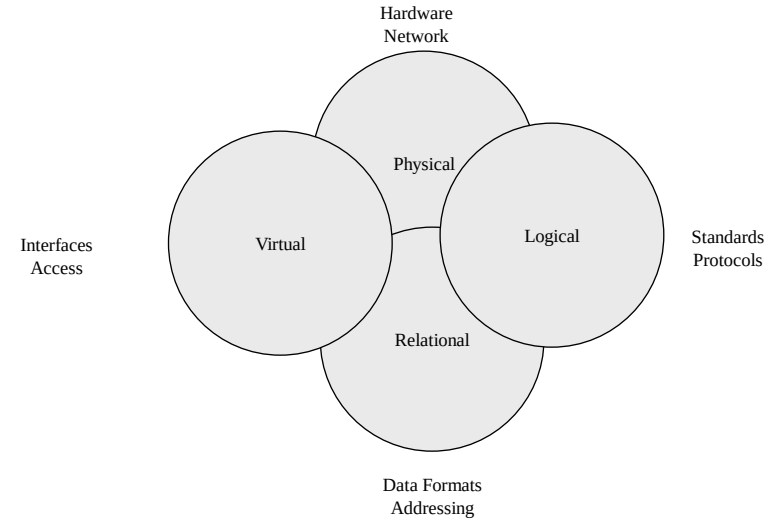
Understanding that there are several varying architectural designs allows one to better understand that each reflects not only connectivity but also the world view.

6.2.3 Impact of Technology on Architecture

We have just discussed the elements of the architecture and the embodiments of design that these elements may lead to. We shall later discuss the details of the technology evolution but it is appropriate at this stage to make several observations about the current impact of technology on architecture.

In the current telephone system, the interconnect element of the architecture is provided by the Central Office Switch and the physical interconnection of the wires from the street to that switch. The point at which the many wires from the street meet the switch are at a device called the Main Distribution Frame (MDF). The Frame must be able to connect any incoming wire to any outgoing wire. The MDF, as it is called, has been the same for over fifty years. It is a manually connected system , where the craft person must connect each incoming telephone wire to a corresponding location on the switch, each time a customer moves or changes their phone number. In computer systems, this is all done in an electronic fashion.

Infrastructure

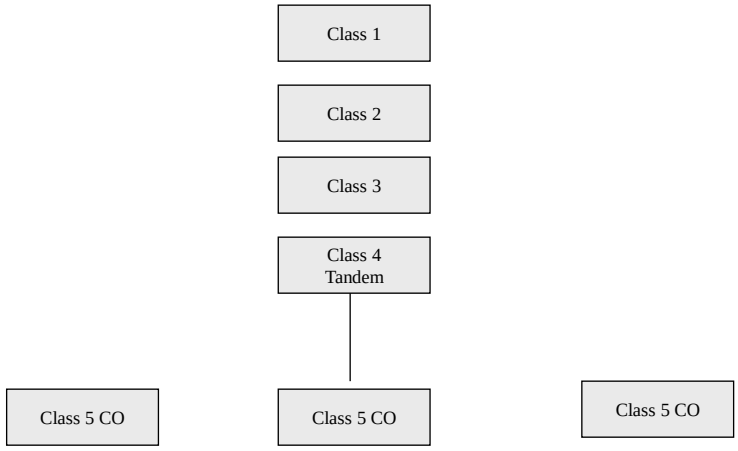


McGarty, Harvard November 1990 *Figure 13*

In contrast, the central processing unit in computers goes through changes once every two years. The standard processing capacity curves show a doubling of processing capability in the same two year period. Computer users have a more rapid turnover of technology because they generally work in an environment with no regulation, shorter depreciation schedules and a focus on meeting specific business needs.

In contrast, the centrally based network must meet a collection of common needs and serve them in a least common denominator basis. The conclusions from these observations is clear. If change is at the heart of the services and technology is driving them, then migrating the elements to the customer of control, interconnect and interface maximize the change and innovativeness of the network.

Current Switched Network



McGarty, Harvard November 1990 *Figure 14*

In terms of a national network, this then begs the question, should not the network, as infrastructure, be nothing more than a broadband transport of open single mode fiber and let all other functional elements be provided by the end user.

Consider what was written by a Bell System polemicist in 1977 at the 100th anniversary of the Bell System at MIT. The author was ***John R. Pierce, Executive Director at Bell Labs***, who stated:

" Why shouldn't anyone connect any old thing to the telephone network? Careless interconnection can have several bothersome consequences. Accidental connection of electric power to telephone lines can certainly startle and might conceivable injure and kill telephone maintenance men and can wreak havoc with telephone equipment. Milder problems include electrically imbalanced telephone lines and dialing wrong and false numbers, which ties up telephone equipment. An acute Soviet observer remarked: "In the United States, man is exploited by man. With us it is just the other way around." Exploitation is a universal feature of society, but universals have their particulars. The exploitation of the telephone service and companies is little different from the exploitation of the mineral resources, gullible investors, or slaves." (de Sola Pool Ed, Pierce, pp 192-194).

The readers should note that this was written nine years after the Carterfone decision and five years before the announced divestiture. Pierce had a world view of an unsegmentable telephone network. This paper has the view of a highly segmentable communications system. The world view of the architecture has taken us from "slavery" of Pierce to the freedom of the distributed

computer networks of today. Kuhn has described technologists as Pierce as the "Old Guard", defenders of the status quo. They defend the old paradigms and are generally in controlling positions for long periods of time.

6.2.4 Architecture versus Infrastructure

It is important to distinguish between architecture and infrastructure. We have extensively defined architecture in terms of its three parts; elements, world view and technology. Infrastructure unfortunately has been reified in terms of some physical embodiment. The discussion of NREN being an infrastructure is viewed by many as being a determinate thing. Kahin has, however, de-reified the concept in terms of it being an embodiment of a concept or set of common goals. We expand that and state that an infrastructure is an enabling capability built around a common construct.

Fiber Distributed Network

McGarty, Harvard November 1990 *Figure 15*

There are four types of infrastructure views that are pertinent to the current discussions of networks. These are of particular import to such networks as NREN since they will lead to the policy directions that it will take. These four infrastructure types are as follows:

Physical: This is the most simplistic view of an infrastructure. It requires a single investment in a single physical embodiment. The old Bell System was such an infrastructure. The National Highway system is such an infrastructure.

Logical: This network may have separate physical embodiments, but all users share a common set of standards, protocols and other shared commonalities. All users have access through an accepted standard interfaces and common higher level transport facility. IBM had attempted in their development of SNA in the mid 1970's to develop a logical infrastructure in data communications. This was expanded upon by the ISO OSI seven layer architecture, selecting a specific set of protocols in each layer.

Virtual: This type of infrastructure is built on intermediaries and agreements. It provides shared common access and support interfaces that allow underlying physical networks to interconnect to one another. Separately, the individual networks may use differing protocols and there are no common standards. The standards are at best reflected in the gateways to the interconnection of the network. Thus this infrastructure is a loose binding through gateways. It is in many ways what is the INTERNET today, if we include all of the subnets.

Relational: This type is built on relationships between the network parties and the establishment on higher level accessing and admission. Specifically, a relation infrastructure is based on agreements on sharing addresses, not necessarily common addressing, and on the willingness to share data formats and types. It is an infrastructure based on shared common interests but not shared common access. This type of infrastructure is what in essence exists in most cases today. Users can move from network to network through various gateways. The difficulty is the fact that the interfaces are cumbersome and may requires sophistication on the part of the users. However, more intelligent end user terminals and interfaces will reduce this cumbersome interface problem.

We show the relationship of these four infrastructures in a diagrammatic fashion. Our conclusion is that understanding the type of infrastructure that the coalition of users want, will also impact the architecture, based upon an imputed world view. Arguably, a physical infrastructure leads to maximum hierarchical control and the resulting impacts that such control leads to. This is a critical issue for networks such as NREN, since by choosing infrastructure and architecture may not be as uncoupled as desired. In particular, the selection of Gbps capability may really be GHz capability and is best suited to a Virtual or Relational infrastructure.

6.3 TECHNOLOGICAL FACTORS

In the previous discussions, we have assumed that there are certain underlying stabilities in the transport structures that enable the separate network providers to perform their tasks. There is a growth in technological capabilities that may cause dramatic changes in structures that we have discussed in the previous sections. In this section, we will focus on some of the dramatic technical changes and discuss the impact that they may have on the market equilibria established.

6.3.1 Transport Capabilities

Transport is the raw power to move information from one place to another. Transport is also viewed in its most primitive form, specifically bandwidth, rather than data rate. In current systems, transport has significant capital in the twisted pair plant as well as the fiber backbone. There are several alternatives to twisted pair that are evolving and we shall discuss their directions briefly. These directions may significantly change the view of capital allocation to the transport portion of the network.

Specifically the raw transport capabilities have been dominated by the capital cost of a twisted pair, with that single pair limited to a single voice channel. There are at least three technology areas of change that may impact the capital asset allocation equation. These areas are developments in fiber, radio and surprisingly in twisted pair itself. One of the goals of transport development is to have "Free" bandwidth access, or as close to it as possible, in relation to the other three network elements.

Under the current structures, this bandwidth is so segmented that it is impossible to foresee a free state.

Another factor in increasing the capital costs of the existing network is the need to add capital in large amounts and not in incremental amounts. Thus, a central office is added with the capability to handle several hundred thousand users. In contrast, with a fully distributed network, it is possible to add user capital as each user is added. We have discussed this issue in the last section and demonstrated that in an environment of rapidly changing technology, the cost-performance curve versus time is rapidly improving, so that small incremental changes allowed by fully distributed system optimize economic performance.

The following are three technological factors that will lead to the goal of freer bandwidth.

Fiber; Uses, Users, and Costs

Fiber has revolutionized the data networks in the United States. A single strand of fiber can transmit 10 12 bits per second of data. If we allocate each home, 100 million residences, with 100 Kbps of full time data, that is 10 13 bits per second if everyone in the US is talking simultaneously in this high speed data fashion. That is the capacity of just a single strand of fiber. A typical bundle of fiber has 25 to 50 strands and these are connected to other such bundles. The current fiber network is structured like past voice networks, and generally does not take advantage of the bandwidth of the fiber. Albeit the technology is not yet totally operationally capable, the world view of the system designers is one that is to use fiber as copper. Use it for one voice circuit after another.

As we have discussed, dark fiber is the most flexible form of access for the more sophisticated user. It enable the user to maximize the impact of processing power and it optimizes the tracking of the cost performance curves. However, it is a fully distributed system and thus requires the careful control of the infrastructure relationships established. It should also be noted, that although the fiber loss characteristics necessary to support a wide dissemination of dark fiber are theoretically available, this is only in a laboratory setting. There are no commercially available fiber systems to allow this in today's network.

Radio Spectra; Access, Capacity and Cost

The current cellular system is being supplanted by new digital cellular technology that can support in excess of 100 times the number of voice channels in the same spectrum. It will be able to support data rates of from 1 Mbps to 100 Mbps. There is pressure to allow the cellular companies to widely deploy this technology and it has been estimated that the capital costs per cellular phone will become less than the capital costs of a wireline phone. The change in technology may make the replacement or build decision not one between fiber or copper, but between copper or cellular. The policy issue is the effective use of bandwidth. If the FCC, in the Common Carrier Bureau, and the FCC, in the Radio Carrier Bureau, can arrive at two different decisions, then there will clearly be a significant policy debate.

The two positions are as follows. First, assuming that the radio bandwidth is free, then if the new cellular technology is less capital and operational cost intensive then copper or fiber, should the FCC allocate the bandwidth in the public interest. If the answer to this is yes, then how should the FCC allocate this resource in the most competitive fashion. Does the current Cellular policy of two operators per system still hold, or should there be an imputed cost to the bandwidth in a bid process. If there is an imputed cost, who is to receive the economic benefit.

The radio technology will allow many others to enter the market, restrained only by the bandwidth limitations on available spectra. This represents a potential destabilizing technology, especially for residential and rural networks.

Twisted Pair; Utilization, Capability and Sunk Cost

As we mentioned before, twisted pair has been limited by culture to single voice channels or 64 Kbps at a maximum rate for data. Current advances allow for transmission at rates of 100Mps on unshielded twisted pair. Thus it is possible to continue to utilize much of the existing copper plant for the types of data services that are required by many customers. Therefore twisted pair should not be considered defunct.

6.3.2 Interconnection

Interconnection is the architectural element that provides for the function of allowing each user to interface with other that are connected to the network. In the existing telephone network, switching as we know it was introduced since bandwidth was very expensive. It was not primarily introduced for the purpose of interconnection. Thus the central offices of today are not there for the sole purpose of connecting one user to another. They are there, primarily, for the purpose of concentration on trunk circuits, and thus preserving bandwidth. If one looks at the development of the communications technology through the Electronic Switching Systems, the intention was always to utilize the voice channel trunks at as great a level as possible. It was not just to interconnect one user to another.

Past interconnect technology began first recognizing that bandwidth was expensive and that concentration on the trunks was also essential. In the past, a twisted pair could support only one voice channel. A central office could readily connect one local user to another local user through the switch by merely closing a crossbar relay or a fereed switch. For connections to other central offices, trunk networks were needed.

Since copper was expensive but of very limited carrying capacity, it was necessary to design the switching or interconnection network to first minimize the need for copper trunks and then to assure that full interconnection could be achieved. The view of the network as a hierarchical design was a world view based upon the paradigm or technology of copper twisted pair, of copper as a limited bandwidth transport vehicle.

Changes in technology show that interconnection can be migrated to the customer premise. As we have indicated, the capability of a single fiber is adequate to handle all of the telephone users in the United States. The interconnection in this system may be done by assigning each user a separate frequency and using a laser tuned circuit to perform the switching function. Thus the switch is at each customer telephone, and all that is necessary for the transmission function is a single strand of fiber. Point of fact, there is research under way at MIT and other institutions to develop just such systems. In extremis, this approach reduces the public switched network to a commodity based transport only facility. If we accept the validity of this alternative world view, then we can dramatically see the changes that may occur in the national network.

Central Switched Networks; Monopolistic Necessity

The central switched network was the result of the bandwidth preserving approach of the nineteenth century. It was further reinforced by the regulatory emphasis on rate of return regulation. In addition, this approach assured a barrier to entry to any form of competition.

The architecture of the current centrally switched networks starts with the Class 5 central office switches, moves up to Class 4 trunk or tandem switches and migrates even up to a class 1 switch, which handle only excessive overflow traffic.

Fully Distributed CPE Based Interconnect; Free Bandwidth

The first entry into a distributed interconnection capability was the local area network technology, LANs, particularly Ethernet connections. The Ethernet connection provides a fully distributed interconnect capability. In the Ethernet configuration, each user has an Ethernet card in their machines and it is in this card that the signaling on the Ethernet channel carries the signaling and interconnection information. The signals have information of the source and destination of the packet message. This is read and decode by each of the terminals on the network.

Ethernet was the first architecture to challenge the existing hierarchical architecture in the interconnect element. It empowered the end user to control the interconnect function directly and enhance it as the needs required. With the introduction of routers, bridges, and gateways, more complex inter-network connections can now be achieved. Thus LAN intensive companies can inter-network all of their locations on a fully switched basis by use of CPE based systems, and use only the direct point to point transport of the common carrier.

As we see the expansion of the bandwidth on a fiber loop, the capital investment of fiber per unit data is de minimus. For example, if the fiber costs are $50,000 per mile, and one can carry even one Terabit per second, or 10 12 bits, then the cost per bit per mile is $5 10 -8 per mile. That is it is one the order of a millionth of a cent per mile. If we use a million miles and the capital costs are a cent per bit per second. In effect, technology is driving the bandwidth costs to zero, as compared to other costs, We depict a simplistic view of a fiber network. In this network, we envision that all of the transport is in the network and that the interconnect, control, and interface are part of the CPE. This is a Terabit per second network that allows the end user the maximum in interface capability. The end user may access any and all of the data capabilities of the transport network, not limited by switching or other elements.

This technological change will undoubtedly change the world view of network providers. If bandwidth is relatively free then the use of processing at the CPE becomes a vital ingredient in the network design. The hierarchical network is no longer the only choice, in fact its viability is called into question. The move towards that end, as we have already shown in the LAN area, is already under way. The development of new means of interconnect will result in significant changes to the network. Specifically, we see;

The network will become more organic. The end user will have direct control and access to the interconnect, interface and control functions.

There will be less materiality of scale. In the current architectures, there is scale and the performance of interconnect is a result of the scale of the network. Simply stated, we need a big network with a great deal of switches so that we can talk to one another. In the network with user controlled interconnect and "free"

bandwidth, the materiality of this scale is no longer a factor. That is, infrastructure is at best irrelevant and at most counterproductive.

Multiple overlay networks are connectable, from both within and without the core net. Thus, viable overlays can lead to local short term optimizations that meet end user needs.

o Intelligence in the CPE is expansive and reduces the capitalization needs for networks. It also reduces the time scale factor for the introduction of new technologies mapped to the technology change curves.

6.3.3 End User Interfaces

The use by the end user is the driver for new network applications. That end user must have some economic justification for the use for it to be meaningful. It must fit the value chain of the user or of the customers. Unless this is the case the result is a technology in search of a market. There are four elements of the end user interface that must be explored, the first the physical constraints and the later three the application specific elements. Specifically, the user interface must first be a functional input/output device, displaying information in its broadest sense as well as enabling the user to interact with all elements connected to the network.

For the most part, the last ten years has seen the evolution of the end user interface from a dumb terminal to an intelligent, highly interactive, display device. It will be in the end user interfaces that the new drivers for network usage will arise. As we have indicated before, the world view of the telecommunication network was based on a voice modality of instantaneous communications. The new modalities are focused on multimedia and multi user communications systems. There will be an integration of video, voice, text, image and other sensory interfaces. More importantly, there will be more processing power in the terminal to perform the tasks that the network now performs best in its hierarchical way.

Terminals and Displays; Uses, Access and Price

The evolution of the end user terminal has been and will continue to be the major driver in changing the architecture of the network. The use that is made of the conversational nature of

communications, the sources and structure of the dialogue, will combine multiple media modalities. The second trend is that of a more intelligent end user terminal. This intelligence is not only in the physical processors in the terminal but also in the distributed software capabilities that reside on top of the device. A fully distributed operating system environment, as well as a fully distributed database environment, will allow for highly interactive communications environments. Users will be able to share information of significant data content in a fully distributed and real time fashion.

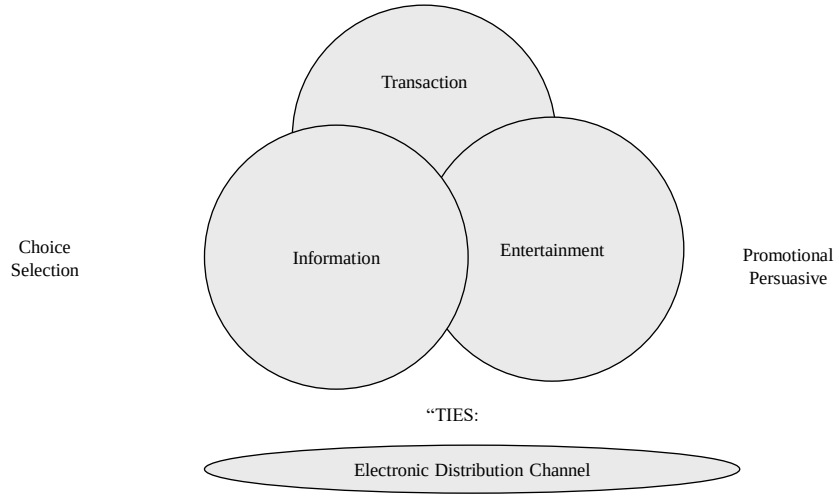
Moreover, the functions of network management, network administration, and interconnectivity will be controllable at the terminal end and not reliant upon a single hierarchical network. In fact, the structure of a regimented network with its hierarchical structure may actually inhibit certain types of applications development. For example, in the areas of multimedia image communications, the need is for as great a bandwidth as possible, with as little network control as possible. The use of the OSI lower layers, such as data link, network and transport, actually add significant delays into the communications link. For example, in the transfer of a compound multimedia image of a 100 Mbit image, a set of four synchronized voice signals, and a video segment, the use of either an ATM switch or SMDS switch, using cell transfer rather than full frame transfer, can result in delays exceeding five seconds. The packetization required seriously effects the overall response time and performance.

Information; Target Markets and Economic Returns

Information networks have evolved over the past fifteen years with the early introduction of the database services of such market leaders as Lockheed, with the Dialog system. This industry has evolved along market niches, providing such markets as the financial services market with a wealth of information sources. Companies such as Reuters, Knight Ridder, Quotron, and others have made this segment a billion dollar industry. However, there has been little or any interest in the consumer side of the information market. Information is a strategic element in a corporation's competitive advantage in the market. It is not, and has not been a major element in the consumer market. The consumer is focused on entertainment and not information.

However, the evolution of the network architectures will depend upon and will influence the information interfaces to the end user. Currently, there are significant delays on the Dialog link due to limited transport and the single location of the database storage. With a fully distributed architecture, the Dialog system can be moved from its current location on the San Andreas Fault and become distributed in multiple locations. Users will be able to obtain real time full text information for off line perusal.

Major Technology/Service Drivers



McGarty, Harvard November 1990 *Figure 16*

Information can also be provided in a distributed form, and no longer in a hierarchical form.

Transactions; The Electronic Distribution Channel

Judge Green has looked to Gateways as a means to enter the information age. As noted, information services are naturally a valued added benefit to commercial users and the consumer is generally interested in entertainment or at most a limited transaction capacity. Transactions are also purchases and to obtain the purchase the seller must not only have access to the buyer but also must promote and persuade. The video medium that currently exists is a promotional medium and creates awareness. It is still questionable whether it is persuasive.

The vision of Videotex has been significantly blemished over the past ten years in the United States. We all too frequently look to the success of Minitel in France, but we all too often have no knowledge of the French Telephone System. Telephone Information Operators in France are almost nonexistent and when available have been known to hang up on customers in less than the most polite fashion. In addition, the French Telephone company, in conjunction with the French Newspapers, developed a way to ensure a noncompetitive advertising market in an electronic fashion. The classic yellow pages were abandoned, the terminals were underwritten by the government as a social action move to develop a technology base, and the public turned to their favorite pastime, now in digital form.

Minitel is not the harbinger of things to come in the area of home information systems. The regional Phone companies have all introduced videotex gateways for use of information services. These have been of little success. It is still to be determined what the consumer wants from electronic information services.

As indicated, the use of the communications channel through the end user interface allows for a distribution or sales channel for products or services. Promotion and persuasion are critical and these require high quality, personalized, self-segmented, full motion video. The impacts on the network are significant and it is possible to perform this task only with the extended network that we have developed.

The architecture of user empowerment is essential and existing networks cannot meet the needs.

Entertainment: Generator of Revenue

The entertainment interface has been the television set and the transport mechanism has been the coaxial cable of the CATV operator. Fiber to the home or other possibilities are dependent upon the entertainment driver. As we shall discuss latter, the CATV provider has a secure monopoly of service. The entertainment interface will be driven by that supplier.

Totality of Interfaces

We have dedicated the totality of interface elements. It is a conjunction of transaction, information and entertainment, overlaid on displays and input/output devices. The focus on the residential side will require an amalgam of these factors. The convergence on the commercial side will require transactions and information. The transaction element is key to the overall success. Thus in evolving networks, we should not disregard the transaction portion.

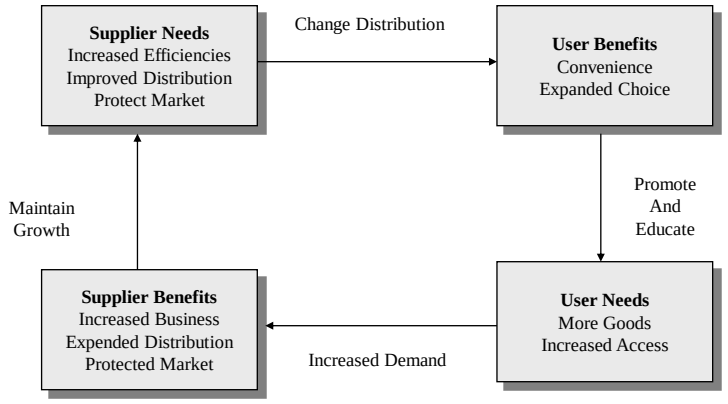
The key conclusion from this section is that technology has changed and with the power to perform most if not all of the network functions placed in the user's hands, at drastically different cost factors, the need for infrastructure, viewed as a single logical reality, is no longer necessary or possible. In the next section, we shall show the need, first, to meet the market demands, and second to create value for all users.

6.4 MARKET ENVIRONMENT

The primary driver for the development, application and usage of any resource or asset is the set of market elements who view this asset as a means to increasing value in some form. We assume that there exists a rational consumer and that in turn the consumer has a set of utility functions that can be maximized or optimized by use of this exogenous asset (See Henderson and Quandt). Understanding the current segmentation of the market by use and user allows for a rational

approach to the focusing of additional assets by the provider of the communications infrastructure (Porter, 1980, 1985). However, one must be careful of the possibility of a circular reasoning that says that the current market is the basis set for extending the future market. The current market is conditioned by the existing trends in prices, accessibility, and performance. If this is changed, we must be careful to assess the change in the future market. Specifically, Mandelbaum has indicated that the users become de facto market makers and that they are empowered by the unbundling of the network elements. The market maker influence is critical in understanding the dynamics of the network architecture revolution.

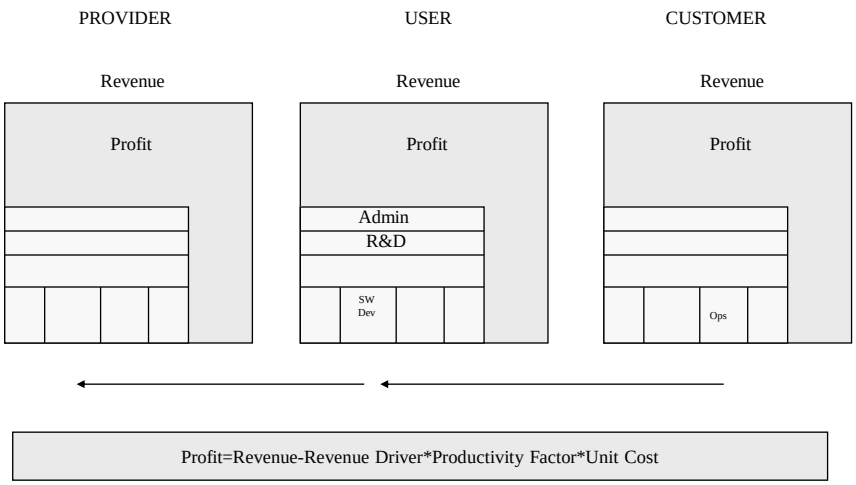
Needs/Benefit Cycle



McGarty, Harvard November 1990 *Figure 17*

Value creation and transfer is the key construct in understanding the market and its relationship to the new paradigms created by broadband networks. Value creation and transfer (See the discussions in Fruhan in terms of the quantitative measures of value for users). Value can be measured if the impact of the new technology or paradigm is understood in terms of its impact upon the user of the new technology. Meeting the needs of the end user is creating this value or the transfer of this value. Understanding the market is truly understanding the needs of the user as the user perceives them and not assuming in a Kantian sense what the user "should need".

User Value Chain



McGarty, Harvard November 1990 *Figure 18*

In the current market understandings of the market even by regulators there is a change in the understanding that value to users must be created. For example, Noam has indicated that:

" Perhaps the greatest common failing of these traditional organizing ways of looking at the telecommunications principles is that they concentrate "supply-side" analysis. That is, they look at the subject from the angle of the production and the producers... Thus, one should not view deregulation as a policy of primarily liberalizing the entry of suppliers. Just as importantly, though much less obviously, it is the liberalization of exit, by some partners, from a previously existing "sharing coalition" of users which has become confining."

Clearly, the need for user empowerment, the formation of new coalitions and the establishment of infrastructures unbound by the physical only are critical. It is essential to look at telecommunications from the users standpoint and not the suppliers. Deregulation is not, as Noam points out, the freeing of the suppliers as much as it should be an empowering of the users. Thus, in this section we discuss the dimensions of the market makers and the market players. We also note that regulation all too often controls the providers and all too often dis-empowers the user.

6.4.1 Market Players

The market players represent the segmentation of the users of the telecommunications markets assets.

They are motivated to do so in a rational fashion and do so as to maximize their utilization criteria. For example, if the set of market players are large corporations, then the use of telecommunications is that of a strategic asset to increase their overall competitive advantage. That advantage may, and often is reduced to a simple set of financial returns. In contrast, the consumer or residential user tends to maximize a utility function that is much less well defined and is of increased complexity. It relates to alternative expenditures for such items as leisure and entertainment.

Commercial User

The commercial user of communications network generally is a corporation who is using this communications capability for either strategic or tactical purposes in their business. Such companies as American Airlines have used the communications networks as a means of capturing market share through the concept of the electronic marketing and distribution channel (See Konsynski and McFarlan for a discussion of how this may expand beyond just a single user into a coalition of users. This builds on the concept of Noam as well as that of de Sola Pool). The tactical application is the use by companies in the operations of point of sale services.

The commercial user's market dynamics are most easily understood. Value creation can be measured because the commercial entity has carefully analyzed and studied the structure of the business and can immediately measure the impact of new means of productivity.

Consumer User

The consumer user of communications has the lowest level of expectations in terms of the delivery of services. This user is expecting at best adequate local and long distance communications. The suppliers of communications view this segment as a significant source of new revenues. The CATV companies use their access channels as means for entertainment distribution, only (See McGarty & McGarty, 1982, for a discussion of the market needs for CATV). They offer a level of service and a level of expectations that is dramatically different than that of the telephone provider.

Consumers have a complex set of benefits. This set is often not readily measurable and there may not be a ready realization of the quantitative benefit. The concept of a utility function or preference function defined for the consumer on an individual is generally quite difficult to measure. In contrast, in consumer markets, market segmentation allows for effective and efficient measures of impact. This will allow for a measurement of the impact of such new products and services that broadband can bring to the consumer.

However, past experience, as we have already discussed, has shown that consumers are quite complex and their behavior in the use of electronic elements has been difficult to predict and variable in its response.

Corporations who have based their future on understanding the consumer have failed in repeated cases in developing electronic services to the consumer. This is a clear warning to academics who attempt to think that their needs are reflective of the masses that make up the consumer market. One can simply imagine an MIT student designing a consumer product, assuming that a UNIX (TM, AT&T) interface and use of LISP is essential because it gives him/her the best response.

Government User

The government user is focused on meeting policy driven directives. The government user is generally slower to act, acts in a highly competitive cost efficient fashion, and is generally risk averse. In addition, the government user must deal with large coalitions of users with often conflicting goals and objectives.

This market represents the most difficult to deal with as a result of the internal stress.

6.4.2 Market Drivers

The market for communications services is driven by the needs of the end users. These clearly are different for the three market segments that we have just described. market drivers are those factors that lead users to demand and use services. The market cycle is shown depicts the need benefit cycle for the user and the provider. This cycle is the same for both the commercial and the consumer user.

For example the cycle has the following dynamics (See McGarty, 1989):

First, The supplier has the need for new revenue. This need is recognized and an investment in a changing distribution capability is created.

Second, the user is provided a certain set of benefits. The user is first to benefit. In some cases, even though the supplier of the service desires to benefit, he must first provide that benefit to the customer.

Third, these benefits convert into user needs. The user takes time to recognize the benefits and then to internalize them into needs. Needs the create demand. A need provides the basis for a sustainable demand.

Fourth, the user needs convert into supplier benefits. This occurs latter in the cycle. It is critical to note the significant time delays in this process.

This cycle takes time for new and innovative products or services. This cycle may also be broken at any single point and result in the dissolution of the specific market opportunity. However, the basic concept of the need-benefit cycle still holds true. The dynamics of the cycle impact the value of innovations to all players. Delays and the impact on imputed interest costs will increase the need for payback to the investor.

Thus the cycle and its dynamics show that value creation results in lower value as the time to complete the cycle increases.

Consider the example of the development of electronic banking, specifically the use of ATMs (See McGarty, 1980, 1981, for a detailed discussion of this technology and its early development). Initially banks wanted or needed to have less costly means of service their customers. In addition, the banks viewed this electronic marketing channel as a means of attaining greater market share based on the concept of location. This service was not needed by the consumer at first but it did provide a set of benefits. These benefits were targeted at the select share of the market, namely the young and well to do segment. By having the ATM network, the bank could provide this segment of their market with the benefits of cash at any time and at any location. This consumer benefit then changed into a consumer need and the broader base of banking consumers requested the service. Then, this resulted in a direct benefit to the banks as the size of the electronic banking led to them (See Zuboff, pp 132-145). The same will apply to accruing the benefits from use of the NREN. The market drivers then are the factors that meet the needs of the supplier and the benefits of the user.

We can expand this into the needs for broadband communications services. As we discussed before, NRI is interested in developing broadband, as defined by Giga bit per second data rates, from a national infrastructure perspective. Their initial focus is on national supercomputing networking. In contrast, there are many applications for imaging that are directly related to solving end user requirements of current interests. For example, the printing, publishing and advertising industry is going through an major infrastructure change. There are Apple Macintosh computers in use for ad composition, editing, and layout. There are digital prepress systems that allow for digital printing. There are not, however, the electronic means to connect these two systems. If there was an effective system, then advertising agencies could cut operating costs an estimated 5%, and press time could be shortened several hours. These are significant impact factors in this business (See McGarty, Nov., 1990).

A second current application is in meeting the needs of the medical imaging community. For example, if a large hospital can take the responsibility of reading the x-rays for a local HMO, the

current costs are on the order of $60 per patient per procedure. It has been shown that just clustering the readings have a scale economy that reduces this cost to $40. If we further include and electronic imaging system with a Hospital Information system, these costs drop to $20 (See McGarty and Sununu, 1991, for a detailed analysis of the impact of the use of imaging technology in a Health Care environment. The authors detail the cost savings directly attributable to this technology and presents results from actual operations. The authors have also developed an extensible methodology for determining cost benefits in this type of broadband environment).

These two simple examples show that there are clear and present opportunities for broadband that are local and that, if properly structured, meet the market needs. These two examples also result in two conclusions. In addition, these examples show that customers can and do, today, create their own infrastructures using the new technologies. In the small, this clearly shows that a global infrastructure is not only not needed but would be redundant. Specifically:

End users drive technology through their value chain productivity gains. Lowered costs or better competitive positioning are essential. Broadband provided scale economies to businesses that heretofore may have had little.

The development of broadband does not require a large infrastructure. It can be built around focused local networks and applications. The needs of the users are not just transport. The interfaces and interconnections are more significant drivers in an information rich environment such as imaging.

Moreover, these examples have shown that today's transport, control, interface and interconnect technologies meet current needs. The migration from current systems to broadband is possible and achievable and are also economically efficient. Speculative applications based upon non-market drive idealizations will clearly not replace currently customer directed and driven applications.

6.4.3 User Value Chain

The generation of value by a user has been discussed in a dynamic sense as the creation of value by increasing productivity on the part of a user or allowing for the development of new revenue sources.

Value was defined in terms of the increase of the flow of funds to a firm by performing a specific task.

The value chain concept, in contrast, is a static view of value projected back onto the operational elements of the firm. It is a key concept in the full grasping of the value flow to a firm with the introduction of a new technology. The value chain analysis, as developed by Porter (Porter,

1980, 1985, 1990), is a construct that overlays the rational utility function maximization process of the commercial user. As we indicated in the preceding paragraphs, we will focus on the commercial user because their utility function is generally more evaluateable and can be readily related to a rational decision process.

We show the provider, the user and the customer. This is the natural food chain of the economic market place. The provider must provide the user with effective supplies necessary for the production and delivery of goods. The user, to attract a customer, must also provide the customer a similar set of benefits. If we view the revenue of the user as the size of the total box, and the expenses as their corresponding areas, then the users profit is the area left over after all expenses are taken care of. The revenue is provided by the customer, the expenses controlled in part by the provider. The company, namely the user, spends money on horizontal elements such as Administrative functions, and vertical elements such as Software development. The allocated profit is represented by revenue less expenses in each segment. The expenses are a product of a factor driven by the customer demand, the revenue factor, the company's productivity, and a unit cost. Thus for a fixed revenue, profit is increased by lowering costs or increasing productivity. The information networks that we have described are productivity enhancers, thus profit enhancers. This is the essence of the Porter theory.

The value chain concept views the user as an operating entity with sequential and simultaneous operations as part of running the business. The sequential operations follow the flow of goods into the establishment, through the processing done to add value and then out of the establishment. The simultaneous operations are followed over all tasks and may include such functions as finance, legal, and marketing. The company can then allocate costs and value to each of these elements, and then can compare them to its competitors.

The costs of each step in the process are the result of three factors; the revenue drivers, the unit productivity, and the unit costs. The use of information or communication networks allow the user to improve the productivity or reduce the costs. This allow for increased competitiveness and thus better margins. If the seller recognizes the value chain of the buyer, then the product that is sold can be positioned in a similar fashion, thus helping the buyer to improve their value chain. This will increase the revenue to the seller.

The value chain analysis provides a methodology to integrate the effects of communications and information services into the evaluation of a business. Porter has done this for many segments of many industries and McGarty (1989) has developed a detailed micro model to use in a detailed competitive analysis. As we look at the market factors, value chain theory states that the use of any new technology must be evaluated in terms of not only the end users value chain but also the value chain of their customers. The chains are linked and the effect is complex

6.5 GOVERNMENT NETWORKS

Recent changes have seen a migration of Government networks from the totally owned and operated FTS to the leased FTS 2000 evolve. It will be argued that the Government, as a customer, is neither a key player in defining architectures nor do they drive the trends in new and innovative systems and services.

In fact, as we shall demonstrate, they are at best the preservers of the existing world view in an attempt to minimize risk and reduce perceived increased costs.

However, before continuing, there are several factors about the Government nets and their past achievements that are appropriate. Dr. Robert Kahn, formerly Deputy Director of DARPA, husbanded the ARPA net through its infancy (See Markoff, 1990). During that phase, he approached AT&T to work with DARPA on the new and emerging packet technology. Kahn indicated that AT&T not only failed to understand the technology at that time but further were very slow in providing the Government with the information to use the existing modems that it provided in its existing network. Thus Kahn, despite the position of AT&T, developed a true infrastructure network. Out of this evolved the current packet nets of today. However, these networks have been supplanted in many ways with other technologies, primarily those on the customer premise, such as LANs, combined with user specific virtual networks. The reason for the change was that the original premise of the packet networks usefulness as an infrastructure which was that of cost reduction for the computer user. With the breakup of AT&T that equation changed and now users have a variety of cost performance tradeoffs that are both available and more attractive.

Currently Kahn is moving in other directions, as head of NRI (The Corporation on National Research Initiatives), a not for profit research consortium, he is pursuing the vision of a broadband network, called Aurora, that will build on a research consortium of various universities. The goal is to interconnect supercomputer over a Giga Bits per second network. In the ARPA Net days, the goal was clear; computer to computer communications in a cost effective fashion. In the current scenario, the goal is more elusive.

However, the world has changed. The world now is one of high competition, intelligent terminals, and a migration of many of the networks old functions into the end users devices. Can even such a government infrastructure be justified in light on the underlying equation? In addition, the drivers of networks today are the applications that the end users are to put on the network, not just needed to transfer large computer data files. Thus one is faced with the quandary of placing a sophisticated network in place to challenge the applications developers, or to let the applications developers stress the network to its limits using the driver of customer or end user demand. A research network must be flexible to stress the technology but not too costly

to result in preordained failure. The major accomplishments of the ARPA Net were electronic mail and layered communications protocols.

The former was what every Net user found of use on the net and the layered architecture was a natural offshoot of all of the interfaces required to meet some form of flexibility.

6.5.1 Structure

The Government networks can be divided into federal and state networks. For the most part we shall focus on the Federal networks. There are two extremes in these networks, DoD based networks and non DoD efforts. We shall not focus on DoD because of their special infrastructure requirements. Non DoD networks have been focused on the meeting of a wide set of standards. They must balance the needs of the voice user and the data user. Further, there is currently a need for both interconnection as well as ensuring security in the network. For the most part, the Government networks have outsourced the management and control. The control functions, such as billing, are eliminated by the bulk buying characteristics.

6.5.2 Competitive Environment

Clearly, the Government networks are in a noncompetitive environment. At best they are limited market makers as described by Mandelbaum

6.5.3 Optimization Criteria

Performance in the Federal Networks is best exemplified by the FTS 2000 effort. The requirements for the network were developed and the selection of the provider based on cost competitiveness.

6.5.4 Evolutionary Constraints

The major evolutionary directions for the Government networks are based on meeting their user coalitions needs while doing so at a minimal cost. In certain cases, such as in DoD networks, advanced technology will be employed. Other than this specific case, they will generally be followers of other network leaders.

6.6 PUBLIC SWITCHED NETWORK

The Public Switched Network is the result of an evolutionary process that had given rise to the Bell System and has resulted in the current structures under the Modified Final Judgment (MFJ) strictures (Coll; Geller; Kraus and Duerig). The network is based upon a set of technologies that require massive investment in capital assets to allow for the interconnection of many users to

each other. The system requires massive switching infrastructures since the underlying transport facilities, the classic copper cables called "twisted pairs", have very limited information carrying capacity.

6.6.1 Structure

The public switched network is built around a hierarchical architecture that makes several assumptions of its environment and the customer base. These assumptions are (See von Auw for the insider Bell System view, pp 334-398; see Toffler for the external consultants vision of where the Bell System could have gone):

Bandwidth is a costly commodity so that it is necessary to provide for concentration of circuits on trunks and tandem lines. This concentration leads to the need for multi layered switching centers, a Class 5 Central Office being the lowest level.

Voice is the primary means of communications and all circuits are to be considered multiples of voice circuits. In addition the voice is to be sampled at a rate of 8,000 samples per second and the number of bits per sample may be from 7 to 8.

Universal service is necessary in order to meet the needs of the state regulatory bodies, This means that telephone service must be structured to cover the gamut of the rural home to the large corporation.

Quality of service is to be as high as possible with overall system availability to exceed 99.95%. This implies redundancy, disaster recovery systems, and a trained workforce that will permit as near as real time restoral as possible. New York telephone has in twenty years responded to crises that would have sent other US corporations reeling in chaos. Their response allowed restoral of services in extreme conditions of distress (See the recent paper by Bell discussing the fragmentation of the telephone network).

o The focus is on operations, namely keeping the network service up to the performance standards set, and this implies that the work force must have the capability to deal with complex operations requirements in multiple positions. All people should be cross-trained to meet the level of service expected by the consumer of the service.

These assumptions make the local telephone company an infrastructure entity. The telephone company as the local operating company has a highly redundant, high quality of service, with a highly integrated work force.

If we look at this network in terms of our architectural elements we see the following:

Control: The control is highly centralized. The control emanates from a set of methods and procedures, flows through the overall control mechanism for the network and is integrated to the maintenance and restoral efforts.

Interconnect: This is a truly hierarchical interconnect. It is based upon the Central Office Switch, which is designed to conserve bandwidth at the trunk side. It provides a level of common access based on a single voice channel.

Transport: The basic transport is the twisted pair to the end user. There is fiber in the loop and some fiber to the user, specifically those users in the higher data rate category.

Interface: Generally the interface if the telephone handset. More recently, the interface has been expanded to include data sets.

In contrast to the ongoing performance levels of the Regional Bell Company, the interexchange carriers (IECs) do not have to meet the same levels of service. The recent AT&T service outage in New York with their Class 4 and Class 3 switch outages showed how the level of service has dropped for the IEC level (See Bell p. 36). Specifically, the IECs have recognized that a level of service may be priced on a differential basis. This is in contrast to the pricing structures for the local operating companies whose service levels are more closely controlled by the state Public Service commission. It is of question if there is to be a divergence in service levels over time, between these two types of carriers.

6.6.2 Competitive Environment

There has been great discussion on the threat to the local carriers from the bypass carriers. The bypass carriers have had some impact but has not been as serious as expected. The true competition has typically been from the changes in technology. Specifically the user has in the past used a set of single voice circuits not having the capability to multiplex them over several higher data rate lines. With the introduction of T1 or DS1 (1.544 Mbps) circuit, now it is possible to get 24 voice circuits for the cost of 8.

In addition it is also possible for customer to get DS3 or 45 Mbps circuits for 8 times a DS1. Thus the technological innovations in multiplexed circuits are the internal competitors to the existing single voice circuits.

6.6.3 Optimization Criteria

The economic performance of a public switched network has evolved over the past ten years in a new direction. The rational behavior of any firm is to maximize their profits (Pindyck and Rubenfield). This rational economic behavior has been the incentive for any company to invest.

However the PSN is inherently a monopoly in many market segments and thus operated as a monopolist. In 1980, the strategy of the PSN provider was to maximize the rate base subject to the constraints applied by the Public Service Commission (Kahn, Spulber).

We introduce performance elements, P i, which defines the measure that is maximized or minimized in the rational business decision process. We can relate it to the standard utility function of microeconomics, but will not do so in this paper. In the past, the optimization criteria for a telephone operating as a public utility was;

P PSN = max (Rate Base Universal Service, Level of Service)

In contrast as the regulation has been changed some new forms of performance measures have evolved.

Specifically, we have;

P PSN = min(Cost| PSC Price Caps)

From a rational investment point of view, the corporation should be acting in such a fashion so as to maximize the net present value of the company as reflected in its discounted cash flow. This is the strategy that leads to maximum return for the shareholder (See Fruhan for value creation, transfer and destruction, also Lawrence and Dyer for the analysis of the transition of AT&T from strategy 1 to strategy 2). This optimal strategy can be given as:

P PSN = max(NPV(Cash Flow)| Level of Service)

6.6.4 Evolutionary Constraints

The current evolutionary thinking in the public switched network takes it from ISDN to Broadband ISDN with ATM switching. It is an evolutionary path that is built on the world view of the nineteenth century system of hierarchical switches and the need for universal service. It is driven by a need to provide the same data transport, at almost the same time, to business customers as to the residential customer. It does not readily admit the flexibility of multiple overlay networks, allowing for segmentation of service. It is predicated on the assumptions of high per user capital investment and allocation of costs in a rate based world. Much of this thought process is, however, driven by responding to a Public Service Commission demand on a state by state basis for equality of service.

The question may be asked as to why the local operating companies have not taken a more active role in the development of the regional networks. There are several obvious answers and those not so obvious.

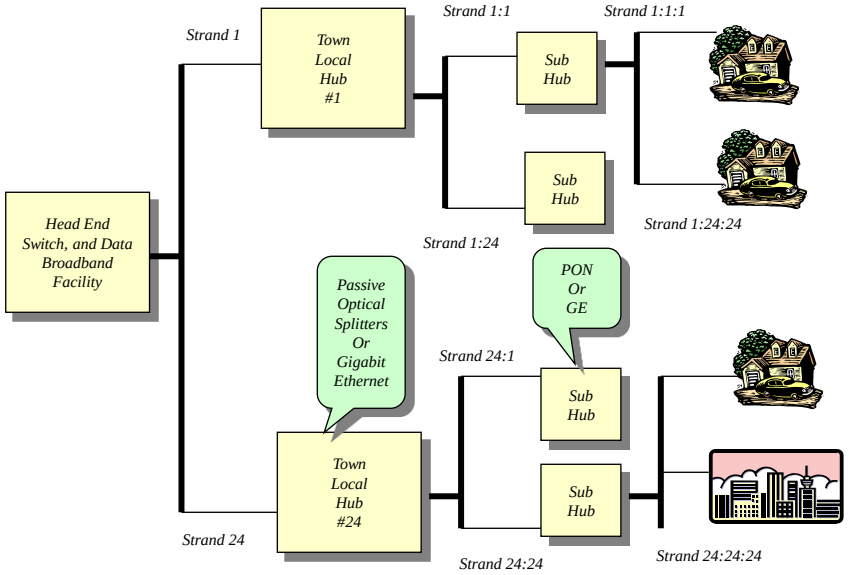
Clearly, in the case of most regional networks, NYSERNet being one, inter LATA transport is necessary.

New York Telephone took a key role in the early stages, however it had to deal through Rochester Telephone to legally work with the NYSERNet efforts. In fact, it is even considered in violation of the MFJ to manage an inter LATA network. Thus, for MFJ and other regulatory reasons, the local operating companies are wary of involvement. A second reason is the general voice focus of the Regional Phone Companies. The emphasis on data in the networks of the type such as NREN go beyond the general network infrastructure supported by the local operating companies. A third reason is the questionable economic viability and extensibility of such academic networks.

6.7 CATV NETWORKS

CATV networks have evolved over the past thirty years from a simple local distribution service for off the air television transmissions to highly complex communications networks with a centralized architecture.

CATV Network Structure



McGarty, Harvard November 1990 *Figure 19*

In the early 1980s the CATV carriers were required to make all of the cable systems have the capacity to carry two way transmissions and it was implemented in 1980 by the Warner Cable company in the now classic QUBE system (See McGarty 1982, 1983 for a technical description of the system and Couch, McGarty and Kahan for the market structures). Two way cable was further expanded to provide for both data and voice transmission. The first data circuits were

provided in 1982 by Warner in Pittsburgh, providing 1.5 Mbps data communications to Westinghouse in a metropolitan area network. In 1983, Cox Cable (see Tjaden) implemented, with the assistance of MCI, a fully switched voice communications system over the cable network. In 1983, Warner Cable reached 100% penetration of optical fiber in its backbone trunks, from hub to head ends, in the large Metro systems.

In late 1983, Warner Cable developed, and field tested a full motion video, on-demand, videotex system in a joint venture with Bell Atlantic, Bank of America, Digital Equipment Corporation and GTE. After determining that the time was not appropriate this trial was canceled (See McGarty and McGarty, 1983). However, GTE took the Warner developed technology and further developed for its cable trials in California. Unfortunately the GTE trials lacked a market driven partner and they too failed. Despite the failures, these trials showed that the cable industry, almost a decade ago, had developed, tested and marketed systems that are still to be implemented in the telephone network. This limit is not a technical or market limitation, it is clearly a limit of judicial mandates on market expansion.

Thus in many ways the cable industry was far ahead in meeting the needs of all classes of users as compared to the telephone companies. Much of this was a direct result of the lack of regulation and willingness on the part of the customer to infer the existence of and agree to an underlying price performance curve.

6.7.1 Structure

The CATV networks are structured to provide for the distribution of entertainment to the home. These networks are centralized in form; they have a head end that operates the local area, a set of hubs that are distribution points and possibly sub hubs for local regeneration. The systems are generally one way broadcast but have a two way capability. The one way may be 50 MHz or more and the return channel is 50 MHz or more. The systems are primarily coaxial cable, although they are being expanded to include fiber. Some of the existing systems have a fiber backbone The architectural elements for the CATV networks are as follows:

Control: The control is generally from a central facility in the network. The systems are broadcast only and thus distributed control of the network is limited.

Transport: The transport is very passive. The coaxial cable used in these systems is a limiting factor. It has repeaters along the route and these are limiters to the cable design. With the introduction of fiber in the CATV local loop, however, this will change and the cable transport will have the capacity for any bandwidth. Even in the current systems, there is not the segmentation as in copper. Cable could provide any user in the reverse path with any part or all of the reverse 50 MHz of bandwidth today. In fact, in several of the Warner QUBE systems, this was commonly used for commercial data circuits. Thus the CATV systems are dramatically

different than Telco system in that ability to assign and allocate full capacity on transport to any user.

Interconnect: The interconnect capability of a cable system is fully distributed. In the QUBE system, a polling scheme was used which was centralized. In the Cox INDAX system, the scheme was a CDMA system that was fully distributed. Thus in CATV both extremes have been used. In fact, CATV used the first version of the IEEE 802.6 protocol in the CDMA designs, this being the basis for SMDS systems.

Interface: Cable allowed a wide variety of interfaces, ranging from TV converters, to PCs and video games. This is dramatically different from the limitations on Telco networks.

6.7.2 Competitive Environment

The cable systems have a natural monopoly in their underlying franchise licenses. These licenses are awarded on a municipality basis and generally assure the local governments a certain percent of the gross revenues from the system. Thus it is of immediate gain to the local governing bodies to maximize the revenue from the cable systems to achieve the greatest revenue source from the cable operator. The current law limits this implicit tax to 5 %. We consider the competitive environment of the cable system under three differing scenarios and these are presented in Appendix A. The conclusion of that analysis is that cable is a stable monopoly under current conditions.

6.7.3 Optimization Criteria

The original criteria for economic performance was one that was driven by returns to investors. The CATV business was a means for many investors, using both tax shelters and capital gains rates to obtain significant overall returns from their investment. Thus, negative cash flow in the early years was very important to shelter income and positive cash flow deferred to later years was essential for the payout.

This led to the following criteria for cable systems.

P CATV = max(Cash Flow Service Areas, IRR)

This has been changed to:

P CATV= min(Cost| IRR, Level of Service)

due mainly to the change in the tax laws. In addition, there has been a maturing of the cable systems and they are now run on the basis of commonly accepted operations and business

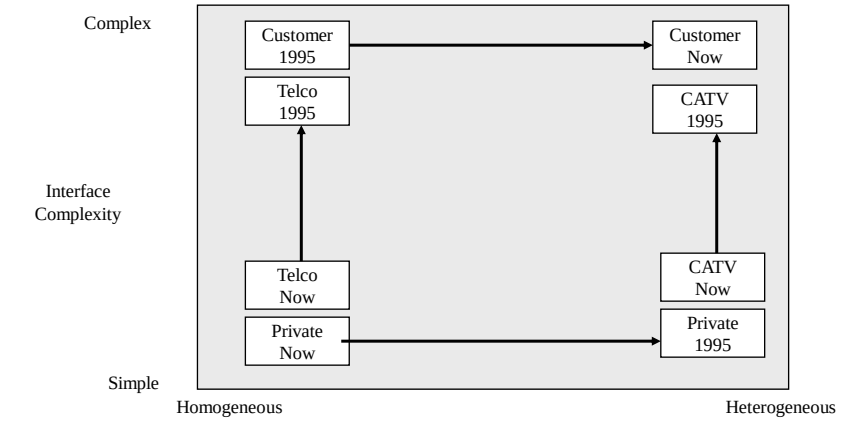
factors. That is, the driving need for capitalization is no longer as critical. It could be questioned, that given the great capital needs for the cable systems, if the tax law had been changed ten years earlier, would there be the cable infrastructure that we see today. Clearly in the current finance markets, the answer would be no. CATV systems were means to ends for many investors in the early 1980's. The success in building a CATV infrastructure was clearly a success of a carefully executed tax policy that allowed for the investment in potentially risky investment. The change in that policy as clearly changed the CATV investment market.

6.7.4 Evolutionary Constraints

The evolution of the CATV networks has been based on the assumption that the CATV operators will assume the massive capital investment risks with the opportunity to attain an adequate return on their investments. That return has for the most part, been focused on the use of the transmission medium for the almost exclusive purpose of entertainment distribution. The fact that the current cable business passes almost 80% of the US homes and has about 50% penetration, thus providing service to 40% of the US households. In contrast however, even at the current rates of service, the revenue for the CATV industry as a whole is only $18 billion, comparable to one RBOC, there being seven, plus GTE. The total RBOCs plus GTE equal the interexchange carriers in revenue, including AT&T, MCI, Sprint and the other IECs.

Thus the total CATV industry is about one sixteenth or 6% of the telecommunications transport business for voice and data. Such a position is not one of any market dominance in the network area. However, as has been noted, this industry has a network in place with both existing capabilities and new transport capacity that greatly expands what the RBOCs have in all their networks. The difference is driven by both the unregulated nature of cable plus the fact that cable is not forced to be the transmission provider of last resort. That is cable is not required to provide "Life Line" services below cost.

Network Evolution Map: Combined



McGarty, Harvard November 1990 *Figure 24*

CATV companies now find themselves in an environment where the RBOCs are potential alternative carriers competing with CATV (See Carnavale). Specifically, the RBOCs have indicated their intent in providing local loop fiber transport. That transport would have the capability to provide transport for CATV systems. The large capital intensiveness of these systems, however, will slow their installation. The inertia of local governments, the true regulators and customers in this market, will probably, for the short term, maintain the status quo. It does however allow for the CATV companies, as they progress to fiber based transports to themselves become the alternative carrier. If the CATV companies, in the current rebuild cycle that will occur in the 1990s, position their rebuilds in fiber, then they are in the position to have the transport base for transmission to the home. It is then possible, in extremis, for the RBOCs to become leasers of transport service from the CATV companies who will have fiber passing the majority of homes, and not vice versa. It will be an issue of who will be at the home first with the transport facility and then it will become an economic decision.

6.8 PRIVATE NETWORKS

Private networks are those offered to select portions of the general market and are owned and operated by network operator other than the common carriers. Typical of such network operators are Teleport in New York and Boston. These private network operators build their own bypass capability, including cable or fiber, along rights away that they may lease from local agencies. They then also provide some limited switching but for the most part connect to the local point of presence of the Inter Exchange Carrier. Their selling proposition is generally lower cost for the

same service. In fact, their actual benefit may lie in the shortened time to installation and the support on moves, adds, and changes.

6.8.1 Structure

Private networks are structured on a local basis, allowing direct access by the end user or customer. They focus on the segment of the market that desires an additional carrier and who are concerned about cost.

The networks generally are built to maximize the local market penetration advantage and are not structured to provide universal service. For example, in Chicago, the Chicago Fiber Network, was built in the coal tunnels as rights of way, directing traffic off the Michigan Avenue business district. Teleport in New York is directed at both the financial district and the mid-town business district.

The networks are, for the most part, fiber point-to-point and have limited any switching. Specifically, they are built around the smart mux technology and are positioned to attract customers on the basis of competing with bulk transport buys. For example, they sell DS0, DS1, and DS3 circuits at significant discounts.

We can now compare the four elements of these networks to those of the others discussed. Specifically;

Control: The control of these networks is centralized and is generally under the control of the organization responsible for their operation. This is an organizational control philosophy and there is generally no reason for not having a distributed control.

Transport: The transport is usually over fiber networks. In contrast to CATV, the transport is controlled, generally by voice mux equipment and is thus no different than the Telco networks. However, some private networks are using special data transport facilities and nonstandard data rates. Thus it is possible to use any amount of the available bandwidth.

Interconnect: The interconnect is highly flexible and allows for all types of device interface. The switching is done by multiplexers that are intelligent and generally are controlled by users segments. Thus distributed control is possible.

Interface: Most interfaces are possible.

6.8.2 Competitive Environment

There are few examples where Private Networks compete with each other, Boston have three smaller versions. Generally their self-competition is along common rights of way and it is yet to be seen that there is sufficient traffic for them to survive. In Boston, as in many other cities, the key to a Private network provider is the Right of Way. Usually the local electric, transportation, sewer, water or gas company has existing rights of way that are usable. The problem however, is the pricing of that right of way as well as the ease of expanding it. For example, in 1984, the New York Metropolitan Transit Authority, MTA, was approached by several private network providers with an offer to lease rights of way. Six years latter there is still no effective policy on how to do it. In Boston the process took only three years. In Washington, D.C. it took eighteen months. In Chicago, only nine months. Typical rights of way rates are $0.50 per foot per month. That is about $2,500 per mile per month. In New York, if rights of way were available, the entire network would need no more than 50 miles of fiber, thus costing $125,000 per month or $1.5 million per year. Such a system could supply service to several hundred customers at the level of a gross revenue in excess of $50 million per year.

Thus the right of way is a fixed fee and if efficiently run, the network may have that fee represent a small percentage of the total revenue base.

The Private carriers receive little if any competition from the common carriers since the common carriers generally have higher rates commensurate with the switching, management, and distribution services that they provide.

6.8.3 Optimization Criteria

The approach of the private network companies is to provides for the least cost design in the early phases subject to minimal market coverage. This financial strategy assures them the ability to compete on a price basis during the startup phase where they capture market share based solely on price competition.

P PVT = min (Cost Service Area) As they have evolved the criteria has changed to;

P PVT = max (Cash Flow| Service Area, Level of Service) which demonstrates the rational approach of focusing on cash flow out of the business. As we indicated in the studies by Fruhan, value creation is based on cash flow subject to market retention.

6.8.4 Evolutionary Constraints

Private Networks have evolved on both the local and the long distance basis. On the local side, there will be a continuing fragmentation of some of these network until the consolidation phase occurs. The consolidation will not occur unless and until it becomes clear what the market is. As we have indicated, the provision of service to the residential user still requires capital per user

that exceeds that for the short term payback. In contrast, the commercial user still has the short term payback and thus will be a buyer for these services. The main evolutionary factors will be:

Market consolidation: Will there be enough clustered revenue potential for the independent private network providers, especially if they are competing on price and in a pure transport commodity market.

Scale consolidation: If there are no other elements other than transport, will there be scale economies in transport on the local level. On the long distance level, such companies as Williams have found scale economies via leveraging their gas transport facilities. Are such consolidations available in local transport? Is there a stable environment in the franchise fee for rights of way. As has been seen in the cable industry this has been the case on the local basis as long as predatory pricing is not perceived? This however will not be the case in the private network area.

Technology consolidation: As we shall note in a latter section, there are certain technologies that may make the capital intensiveness of fiber a non-issue. Specifically radio technologies may be capable of providing some modest bandwidth requirements.

Usage expansion: As the bandwidth is liberated in private networks, as it will be before the public switched network, will such bandwidth drive up the revenue per fiber while at the same time driving down the price per bit per second? It is anticipated that if the private carriers are able to provide dark fiber and that if vendors such a startup as Ultranet and others provide high Mbps customer based switches, the applications will increase.

Clearly, the possibility exists for the Private Network providers to cream skim the market. It has yet to be proven that such has occurred. At present, therefore, there is limited regulation on such networks and it is appropriate to let that continue until they have stabilized with a recognizable character and understood economic effect in the telecommunications community.

The directions of Private networks are in the balance between the Public Switched networks and the Customer Networks. Clearly, the former provide the network of both universality and last resort. The latter provide the network of optimized user customizability. Noam has recognized this dynamic and had indicated a possible direction:

" The breakdown of monopoly is due to the very success of the traditional system in advancing telephone service and making it universal and essential (eg a need). As the system expands, political group dynamics take place, which lead to redistribution and overexpansion. This provides incentives...to exit sharing.. to a system of separate sub-coalitions."

This then begs the question, if this approach is the true market dynamic, what is the stable point of the evolutionary telecommunications architectures. Undoubtedly there is both an economic need and social need for a Public Switched network. User sub-coalitions, namely user empowerment, will lead to effective Customer Networks. What then is the future of the Private Networks? Specifically, if Public networks support the wide base of users and the need for universal service, and if customer networks provide service optimized for the larger user, what does this leave for the private networks. It is argued here that the private networks will become niche market players, providing a boutique service to mid-tier market users of communications service and a broad and common functionality basis. Thus it is concluded that the private networks will have limited broadband functionality, providing at best raw fiber transport, marketable on a limited basis.

6.9 CUSTOMER NETWORKS

Customer networks are the newest of the five networks to have evolved. They are currently the creature of large corporations that have the need to meet specific high density data and voice traffic and academic consortia that are meeting the needs of their research communities. In addition these companies have a wealth of experience in managing data and voice networks and may even be computer companies themselves. Thus IBM and Digital are examples of companies that have their own networks. These companies have had third parties build fiber networks that the individual company then takes over and operates as a part of the corporate infrastructure. These networks cover areas of heavy traffic and are priced in as cost competitive on the basis of being less costly than leased lines. However, in most of the situations discussed the networks also are part of the company's strategy to expand its product offering to include networking products and expertise. Thus IBM and DEC clearly are interested in expanding their product offerings in the network area and the most effective way to do that is through a network of their own.

In contrast, the academic networks, such as NYSERNet and NEARNet, have focused on the needs of their constituent members. Some of the networks use a leased circuit transport but there are many user owned and operated fiber networks extensions. MIT, for example, has an extensive campus wide fiber network that is used to support its local interconnect and transport requirements. A detailed discussion of campus networks is presented in Ames.

On the commercial side, there is also the creation of shared networks for the use of market leverage. The recent study by Kosynski and McFarlan provide several example of corporations and their networks applications and infrastructures. In all cases they lead to a sustainable competitive advantage.

6.9.1 Structure

A customer based network is generally a closed environment that is created solely on the basis on providing better service at lower costs. The Customer Network may be for a single customer or for a coalition of customers. It should be recalled that the network consists of the control, interconnect, interface, and transport functions. The nature of a Customer network is that each of these may be separately or jointly under the control of the customer or the customer coalition. For example, the transport may be leased from a Public Switched Network provider or from a Private Network provider.

The control may be directly under the customer or it may be outsourced. However, the ultimate control is that of the customer.

The network architecture is based on purely economic factors and they are generally established between clusters and high usage and traffic premises of large companies. The networks are built on the following basis:

Transport: The transport is a fiber network that is built between locations. The transport is provided on the basis of both dark fiber and a switched transmission facility common to the switched network fabric.

Interconnect: In the corporate network, the interconnect function is provided by the inherent point to point links and the addressing generally done by the TCP/IP transport layer protocols that are used on the data transport. Voice interconnect is accomplished via the PBX facilities that are on the customer premises.

Control: The control of these customer networks is based on separate network management facilities.

Data networks are controlled by systems that are now a natural part of the computer systems. The control of the Customer networks has taken several operational directions. Some companies have decided that outsourcing of the control directly under their management is the way to proceed. Other companies have maintained control under their direct corporate control. Others such as NYSERNet have found third party companies which that had prior relationships with to take a more active role.

Interfaces: These are provided as a common part of the end user devices.

The corporate networks allow for expansion and customization.

6.9.2 Competitive Environment

The competitive environment is limited in this type of network. Competition is generally from an internal comparison with other carriers. The network is viewed as a strategic corporate asset that is part of the overall company value chain. Thus for example, the network for Citicorp or ADP may be part of the overall financial services system that is provided to their customers. The choice of build versus buy is both a strategic and tactical decision. On a strategic level, the issue is that of establishing a barrier to entry to the competition, allowing the provision of new and innovative services. The tactical issue is that of lowest cost for a required level of performance.

The advantage of a Customer Network is the ability to optimize it for the benefits of the user. Thus NREN, if viewed in the paradigm of the Customer Network solution, will allow for maximization customizability in all four network elements. This is in contrast to the Private Network which will provide a common denominator capability focused on its market niche.

6.9.3 Optimization Criteria

The initial optimization criteria is based on the need to prove the system in on the basis on return on investment. Specifically;

P CUST = max (Reperformance, Coverage)

As the networks have evolved the measure has changed to;

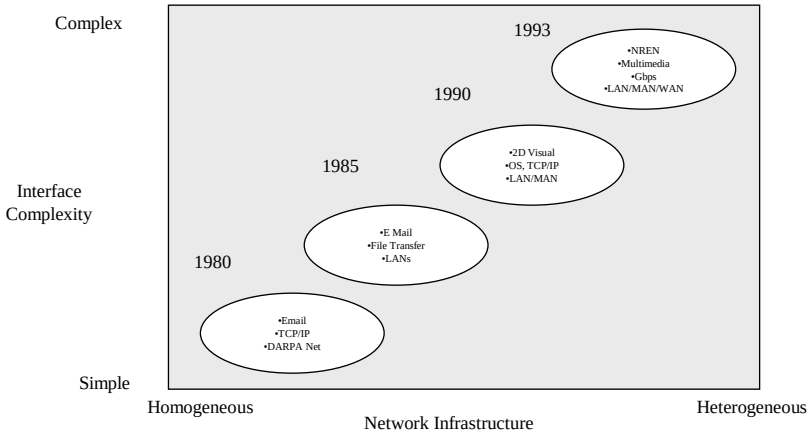
P CUST = min(Op Costs Level of Service, Functionality)

Specifically, the tactical issues of cost containment are key. Levels of service and extent of operational functionality also play roles as constraints.

6.9.4 Evolutionary Constraints

This type of network has the greatest potential for evolving. The key to this expansion is the use of the dark fiber option allowing for the use of data rates that exceed the 45 Mbps maximum provided by the common carriers. Companies such as IBM will undoubtedly use these networks to expand their own products, stretching the abilities of computers to communicate at higher data rates. The challenge is to allow the network to act as a computer backplane and reducing the overall latency of the transport. Such low latency transport will evolve with new applications and new computer architectures. In addition there will be a more fully distributed computer architecture for the use of these transport facilities.

Network Evolution Map: Internet



McGarty, Harvard November 1990 *Figure 25*

In addition to the corporate players in the Customer Networks, the coalition players, be they companies or academic institutions, have a significant opportunity to customize and experiment with these networks. It is possible that it will be with the Customer Network, such as an NREN, that the maximum amount of innovation will occur. This is clearly the case from a commercial perspective with Customer Networks in other areas. The Securities Industry Association Network in New York presents an opportunity for this market segment, in their coalition structure, to establish a highly cost effective transaction processing network. This is leading to changes in database, transaction processing computers and interface workstations.

Network Characteristics

Factor	*Public Switched*	*CATV*	*Private*	*Customer*
Ownership	Regulated Telcos	Unregulated	Unregulated	Customer
Control	Telco	CATV Operator	Third Party	Self Control
Management	Telco	CATV	Third Party	Customer or Third Party
Interface	CCITT, ISO	Some standards	Limited Standards	Customer Specific
Interconnection	Hierarchical	None	Limited	Limited
Extent	Nationwide	Local	Local	As needed
Costs	Tariffs	Negotiated	Negotiated	Cost Based
Evolution	Complex Standard Interfaces	More capacity Complex Interface	More Capacity Limited Service	More complex Increased Interfaces
Examples	RBOCs	Time Warner TCI	Teleport CFO	IBM DEC

McGarty, Harvard November 1990 *Figure 26*

6.10 OBSERVATIONS

Networks represent the asset allocation of capital resources to meet market needs and provide the end users with a means to maximize their utility function. In the area of commercial networks, the network functionality goes to the heart of the value chain of the user and in turn to all those affected by the user.

Networks play an important role in value creation not just value transfer. Networks are no longer just intermediaries that facilitate the work of other information users, they become an integral part of that use.

In this paper we have substantiated several key conclusions.

First: Networks will evolve in a dynamic fashion, not being driven by a single vision but being responsive to a collection of constituencies. This will result in a collection of multiple overlay networks.

The challenge will not be in controlling a single standard, but in developing a flexible interface to these multiple overlay networks. Infrastructures are important, but the infrastructure is not a physical infrastructure, it is rather a logical or relational infrastructure.

Second: The future of networks is driven by uses and users. We are entering a network generation dominated and driven by the end user and their uses. Admittedly, the use of regulation has been argued as necessary for the attainment of limited social goals but for the commercial or government user, networks should be designed and evolved to meet the needs of such business entities. Market makers are the users.

They will define the boundaries of their networks in terms of the uses to which it is applied. Thus, the users will seek networks that are enabling and not delimiting. User communities are not homogeneous as in the classic paradigm for hierarchical networks. It is because of that paradigm shift and the needs to meet niche interests that the change is occurring.

The end user community is not going to wait for an ultimate infrastructure. That is an economic reality.

The end user will be empowered by a new architecture and this architecture will result in sets of quasi-independent multiple overlay networks.

Third: Current regulation is based upon past paradigms. Networks are evolving and creating new paradigms. In this changing world view, regulation serves to buffer obsolescence but should be carefully controlled not to endanger the facilitation of value creation and transfer. Multiple overlay networking is the essence of true competition. If the goal is to maximize the benefit for the consumers and to avoid the monopolistic bottleneck, this is best accomplished by fostering the ability of users to have the networks of their choice.

Fourth: Technology is a fluid change agent in the stability conditions of networking. The free market drivers of minimum cost and maximum utility impact will assist to integrate the new technologies into the alternative network solutions. The technological lifetime of new network architectures may be significantly less that in the past. The capital intensiveness of networks are changing as a result of both existing infrastructure and technological changes. The mindset of the rate based regulation and barriers to entry due to capital investment may no longer be valid. This will force a change in the view of networks as distinct logical or relational infrastructures as distinct from a national highway physical infrastructure.

A Research Network and in some sense LANs are distributed suboptimal consequences of the continuing asymmetrical regulation rather than independently successful technological changes for end user application. If regulatory policy would respond to the changing world view, if it would allow for the implementation of less controlled broadband systems, unfettered all players, optimal solutions are more likely. These more efficient solution would further be in the best economic interests of the user.

This paper has viewed networks not as a national physical infrastructure, as some have argued, a parallel to the national highway system. Rather, it is argued that networks are flexible and less capital intensive logical or relational infrastructure. NREN is a concept that fits that view and could potentially provide the basis for future users and use development. However, NREN as conceived and to be implemented should be nothing more than an academic test bed which may have the potential to evolve into a limited operational network for use limited to researchers. NREN is not and cannot be of any value to a commercial set of entities because of its fundamental lack of value creation end user support.

In addition, technology changes in networks at rates that dwarf the highway model. Roads are built today in a fashion similar to the Romans of two thousand years ago. Networks built today are dramatically different than those of ten years ago. Thus the paradigm that states that networks are the next highway infrastructure may lead to disastrous results if implemented. It will result in the strategy of least common denominator, the most intensive capital costs, and the greatest regulation. In contrast, if we view networks as fluid creativity and value creation for the end user, then it is essential that successful economic development be left in the most open market possible.

6.11 CATV MARKET DYNAMICS

In this appendix we develop the details of the argument that leads to the conclusion that the current CATV market is an established monopoly. We proceed by considering the set of cases that may be considered in the evolution of competition in this market. The demonstration relies on an understanding of Game Theoretic analyses in econometric systems. The presentation is non-mathematical. Consider, now the following cases;

Case 1: Franchise Licenses (Exclusive):

The exclusive franchise license is inherently a stable monopoly (See Shubik, 1987 for a description of the applications towards game theoretic approaches). It allows for a maximization of the cable franchise fee by providing a maximum value to the underlying asset. The only competition that the CATV provider would then have is the alternative displacement competitors such as home video, off the air television, and the movie theaters. Generally, the pay as you buy alternatives such as rentals and movie theaters are incremental and more discretionary, whereas CATV is considered a fixed monthly expense. Thus there is an inherent bias on the part of the consumer, based on the standard consumer inertia factor, to rank order the cable purchase as the first choice. There is also, as has been shown in the past few years, been an inelastic demand for cable services, almost independent of price. The response of the local governments has been de minimus as compared to Federal Congressional bodies. The local governments have seen their local revenue increase almost 350% in the past five years.

Case 2: Non Exclusive Franchise, No Alternative Incremental Transport:

In this environment we assume that each CATV provider must build their own transport network and that the franchises are nonexclusive. We will argue that the result of this is again a stable solution which is again monopolistic, assuming that there is an existing cable provider, established as the dominant. If there are no dominant providers, we will argue that the stable solution is that no CATV provider will either bid or survive.

It should be first noted that the value of the franchise is less in this scenario to the franchise holder.

Non exclusivity has reduced the potential revenue to the local government, so that this and the third scenario have lower marginal returns to the local government.

Let us now consider the two such classes of this case:

Prior Dominant Carrier:

Assume that there is a prior dominant carrier, one who has the existing franchise and that the build of the system has been completed and penetration in excess of a minimum level has been achieved. In this case, the customers have been captured, they are accustomed to the service and have adapted to the monthly fee.

Further assume that the service provided, namely the basic service channels and the pay channels are generally available to any CATV provider on a comparable basis. This is a key assumption, it assumes that there is no differentiator of the service based on the product. Namely, that all CATV products, that is the video content, is comparably the same.

Let us now proceed with the market dynamics under this scenario. It will be noted that this scenarios has the structure of a two person zero sum game. Thus, from an analytical point of view, the dynamics are those that satisfy the Nash criteria (see Shubik, 1984, p. 194 and Luce and Raiffa p. 140; also see the original work by von Neumann and Morgenstern).

We assume that the underlying competitive factor is price, since all providers have equal access to all products.

We assume that the dominant carrier has already established the system at an average cost less than any average cost of a subsequent carrier, and that the marginal cost per new subscriber for the established dominant carrier is smaller than for any new provider.

We assume that the established carrier has raised per monthly rates for several cost increase cycles so that the marginal revenue significantly exceeds the marginal costs and capital requirements. Thus the existing carrier has at least a positive marginal cash flow.

A new competitor enters the market. The new carrier must build a new system and must "buy" the new subscribers from the existing system or from the pool of non-subscribers. The only economic way to do this is with lower prices.

The established carrier then responds with lower prices, still sustaining at least positive marginal cash flow. The new carrier, due to the entry cost structure, and the fact that it has considerable higher marginal cost structures, is at a negative marginal cash flow. This cyclic price war escalates until the newcomer is forced from the market.

The final stable point is again with the single player, who can again raise the rates and wait for another competitor.

The net result for the local municipality is lower franchise fees.

No Prior Dominant Carrier:

In this case, we can readily reach a conclusion based upon the results of the prior case. Since competition is based upon price, there is a stable solution if and only if there is a clear segmentation of the market by geography that allows for the reduction of prices to a level that meets positive marginal cash flow. Again, this solution assures only a minimal franchise fee for the local municipality.

Case 3: Non Exclusive Franchise, Alternative Incremental Transport:

Let us assume that no CATV provider must build their underlying network because it is provided by a third party. Thus there is limited capital investment provided and that transport is provided by a third party, such as the local telephone exchange company. Let us further assume that the transport can be provided at a cost that depends upon only each new subscriber an is available on the market on an equal price basis. Thus there is no competitive advantage on any transport provisioning. Further there is no competitive advantage on the provision of the video product. Thus the only controllable costs of each CATV provider is that of its local sales and operations organizations, all else being fixed. There are then limited scale economies available and thus the marginal costs generally equal the average costs.

Competition in this case is based upon price and the perceived quality of service. It can be shown that this case generates the most cost competitive market and that the franchise fees are

minimal to the municipalities. However, it may be argued that the consumer may get the lowest priced service, the switching costs from one provider to another being minimal.
switching costs from one provider to another being minimal.

7 COMMUNICATIONS ENVIRONMENT

The ability to interconnect many users with multimedia applications has developed over the years into a structured and layered methodology. This chapter presents to the reader an overview of the communications environment, focusing on broadband systems, and expands to coverage of the overall communications lawyering protocols.

The key factor that drives the communications environment for the multimedia user is the need for a connection based service. Connection based services assure the user that there is a synchronous service provided that allows for the end to end integrity of the image.

In this chapter, we discuss the communications element from the aspect of what is needed to effect the overall sessions service and to implement the complete multimedia communications environment. We take an approach that is the inverse from that of the standard communications study, starting with the higher layers of functionality first and moving downward towards the transport function as the latter element.

We view communications as a means to an end, but a means that has a cost in terms of delays, capacity limitations and errors. As indicated, we begin by assuming that the communications medium has all of the elements that are necessary for the errorless interconnectivity of all the users in the session. It is the

assumption of all of the players being in the same room at the same time. As we relax these assumptions, we get farther and farther into the limitations of a real communications environment.

7.1 REQUIREMENTS AND ARCHITECTURE

The issues of the communications environment revolve around the questions of what communications do and how much is needed to effect the necessary task. The multimedia environment places many more requirements on a communications system than is placed by any other usage. In the multimedia environment we are dealing with very high data rate requirements that demand high throughput and this moves quickly into the 100 Mbps to Gbps range of data speeds. In addition we are dealing with an environment that demands that requires that there be minimal delays and that the messages are synchronized. Also we demand that the users be allowed to develop the session structure that we have been developing and that the sessions be transparent to the communications network.

In the multimedia environment, we are dealing with a world of communications that is quite unlike that of the computer to computer communications world. In the computer only world, the end user, namely the computer, can be designed through its software to adjust for the vagaries of

the communications channel. In fact, the development of the layered architecture that we discuss in the next section was a direct result that the communications channel was unreliable from the point of view of

throughput, errors and delay. This then required layer upon layer of protocols to adjust for the channel. As we evolve new channels, we can envision correcting many of the errors of the past and providing a much more enabling environment.

7.2 MM COMMUNICATIONS ENVIRONMENT

The multimedia communications environment is complex in that it demands that any set of users of the system be allowed to be interconnected in any fashion and further that they share the resources of any of the elements of the overall system, either separately or in concert. We have seen that in a multimedia environment that we combine together the forms of communications such as voice, video, image, text, and standard record files, and that we can create many virtual users, such users being real people, applications programs or database and other devices. The ultimate embodiment of a multimedia communications environment is the session which reflects the ability of all of these users to share complex multimedia data objects in a seamless fashion.

When we look at the communications environment of the real world we see that it is often lacking in its ability to provide the resources necessary to effect this capability. The standard communications system is a voiced based network that provides data rates typically of 56 Kbps or less to the end user. Even with the introduction of the proposed ISDN network, this moves up to 64 Kbps, which is still too little too late for the multimedia environment.

There is, however, an available set of higher data rates in the larger network, rates up to 45 Mbps. However, these are not provided on a shared switched basis, only on a dedicated basis. The question may be, why do we need shared and switched. The answer is that we really do not want to pay for dedicated fiber from one location to all other locations on any session connection that we may desire to connect. The shared switched network works well for voice communications, since that is the essence of the current telecommunications network.

The network depicted is that of the shared switched public network which is generally a hierarchical network, consisting of local switches and a higher switch hierarchy.

The environment to support multimedia communications needs is generally quite limited at the present time. For example, at the time of the writing of this book (1989), there are only 45 45Mbps circuits that have been sold under tariff in New York City. In contrast, the city is the media capital of the world, with the highest concentration of media intensives users anywhere. Prime amongst these users is the printing, publishing and advertising industries. Millions of images and hundreds of thousands of hours of video are created and moved from one location to

another around the city on a daily basis. In fact, the most common form of communications is the use of bicycle messenger, each moving at significant speed across the city streets. In fact, these bicycles are moving in the Gbps range!

Thus despite the needs for this multimedia communications market, there is yet an established infrastructure. This is generally because of the lack of the needs of the multimedia communications environment. These requirements are as follows;

1. Effective and Available I/O Devices
2. Shared Switched Networks
3. Layered Architectures
4. Protocols and Standards
5. Connection Based Services
6. Video, Voice, Data Integration
7. Local and Long
8. Distance Interconnect

Based upon the discussion of the multimedia environment it is possible to develop a detailed set of requirements for a general multimedia environment. In this section, we focus on some general requirements issues, but note that special network requirements may be necessary for very specific applications.

7.3 PERFORMANCE ISSUES

As we design a multimedia environment, we are required to develop a set of measures that detail the performance of the network. The performance measures take the loading on the network, based upon the number and types of users and their frequency of use, and determine how well the network can provide the services required. The issue of Grade of Service is a set of benchmark levels that tell the user what to expect from the network. Performance tells the users how the network can vary from the specified grade of service.

The main performance issues to be developed in this chapter are those that have been accepted as the most effective measures of network performance. These general performance measures are;

1. Effective Data Rate
2. Throughput
3. Delay
4. Errors
5. Delay
6. Capacity
7. Variability

7.4 SIZING ISSUES

Sizing is the complement of performance. In the performance model, we assumed that the load on the network was determined and that we then proceeded to develop measures of performance. In the sizing analysis, we assume that the performance level have been given and we then determine how large a user base is supportable. In performing a sizing analysis, we determine the limits of usage of the system.

The major sizing elements for a multimedia system are as follows:

1. Number of Users
2. Number of messages per source per unit time o Message Mix
3. Maximum Data Rate per Channel
4. Maximum Number of Instantaneous Users

7.5 ARCHITECTURES

We have continuously, in this book, addressed the issues of architecture. The architecture issues are the first in any system design and provide the blueprint for the development of any detailed design. The architecture elements are those that must be configured to meet the need that are articulated by the users of the system. In this section, we first detail the architecture elements and present some of the architectural alternatives that we shall be developing in this chapter for the communications element.

There are five architectural element that we consider important for the communications environment. They are;

Network Topology or Physical Layout

Addressing Formalism

Switching Capabilities and Alternatives o Functionality Interfacing o Interconnection

(i) Network Topologies

We depict several network topologies that are in common use. The tree topology is a common topology in such areas as cable television networks. This topology has advantages for networks where the major function is the distribution or broadcasting of high bandwidth data. Thus they find common use in cable networks (see McGarty). The tree network is a hierarchical design that assumes that the system flows both to the headend location of the communications network. The

existing telephone network is of this type in certain locations and is clearly the case in the long distance communications configuration.

The star topologies has been accepted as a viable local topology for fiber networks where the fiber is generally a single mode type and cannot be adequately tapped and bridged at multiple locations. Single mode fiber is not amenable to these types of taps and thus it usually finds itself in point to point interconnects. To expand a point to point connection to a network topology, we find the star architecture as the appropriate one.

The ring topology provides a totally connected structure that allows each users to be interconnected onto the network at their location and further allows for the condition that if the network is severed at any location, there may still be a path to other locations, albeit not functioning under a ring structure. Thus, rings are used for applications where there is a need for some form or redundancy. We can further extend the ring structure to include a dual ring structure, one ring carrying traffic in one direction and the other ring carrying traffic in the opposite direction. We shall see that this becomes a common architecture in the metropolitan area network applications.

The bus network topology is a ring that has been broken. The bus may be a dual bus architecture that is similar to the ring in that one bus can handle traffic in one direction and the second bus handles traffic in the opposite direction.

(ii) Addressing Formalism

Addressing is a fundamental issue in communications networks. We frequently forget how easy it is to communicate in a telephone network because we know how to reach someone by means of their name. However if we look at the process in further detail, we see that the human naming convention is a complex data structure that we have had years of experience with and have accepted almost as part of our cultural history. We know that to call someone, we must know first where they live, we must then find the area code from an independent source, and then and only then can we dial xxx-555-1212. At that point, assuming that there is no telephone strike, we ask information for the number of John Doe. We may further have to delimit this if there are sever entries of that name. This delimitation is the location of Mr. Doe's house. Consider now that we must do this for all of the elements in the multimedia communications network.

Addressing thus applies to users, terminals, layer control elements, data bases and anything else, logical or physical that we wish to address. We shall see in this chapter that the essence of communications is not only getting the bits there but being able to say where there is. That is the addressing issue. We can look at addressing in two dimensions, shared and private networks. In a shared network, we must all agree as to how we are to be addressed. There are limited Vanity

phone numbers, mostly limited to 800-xxx-xxxx. The classic number is the trouble reporting number for N.Y. Telephone for data circuits, 800-1 AM-DEAD!

In private networks, the addressing methodologies are somewhat freer and depend basically upon the limitations of the supplier of the hardware and software. However, as anyone who has run a private network knows, the data and network administration task is significant.

In any shared network, the sharing is done by some form of switching. The function of switching is to allow any user to get to any other user and in addition to allow the user to gain access, on an demand basis, to any set of the network resources. The switching may be done in three different ways.

Centralized: Each user accesses a centralized and hierarchical controlling element that determines the resources necessary and determines the path that is to be taken in providing the connection. The voice telephone network is an example of this type of switching architecture. The signaling to the central switch may be in-band or out of band. An inband signal is one that uses the same channel that the message transport occurs in as is done in the telephone network as today. Out of band signaling is increasing and is typical of the type of signaling in the new Signaling system 7 that will be a part of the ISDN network.

Localized: In this form of switching, the end user equipment takes full responsibility for the selection, management, setting up, and control of the communications channel. Generally this is a complex function and there are very few system that implement this strategy.

Distributed: This is a common switching strategy is its most ^common implementation is in the area of local area networks. It provides for any users to interface with any other user by all users watching message that flow across the network and determining which of the messages are for them and then selecting the message. The advantages of this type of system is that it allows for many users to activity participate in the network but it has the disadvantage of having a higher data overhead.

Interfacing the many elements in a communications network is generally done through a layered architecture. We shall see in this chapter, that the layered approach has been developed in such a fashion that it allows for effective and universal interconnection of all of the elements of a data communications network. The need for multimedia communications has not yet allowed for the development of an equivalent set of layers for multimedia environments.

The interfacing using layered protocols makes several assumptions about the communications environment. First it assumes that there are well defined and structured level that can be segmented into separate control and implementation levels. Thus if the is a level that performs

just electronic functions we can segment all of these function into that level. This is what is done in the OSI physical layer. If there is an agglomeration of functions that handle the switching functions or the graphics interface functions, we could just as easily load all of these elements into that layer.

Thus layering requires a commonality of functions, a collection of those functions into separate layers, and the interfacing of those layers into the overall network. The advantage of the layered architectures is that that they permit the development of software and systems that can be segmented to improvements on a layered by layer basis. The major disadvantage is that with more layers, and more segmentation, we give up speed and performance for the flexibility to develop the individual layers. We shall see in this chapter some of the effects of layering.

(v) Interconnection

Interconnecting networks is a critical design and architectural function that must be considered in any communications network design. The interconnection is driven by the fact that no single network is self-sufficient and that the world is becoming an amalgam of many different types of networks. Overlays of broadband multimedia networks are typical and will become a more common feature.

There are conceptually four generic types of interconnections. They are:

(i)Repeater: This provides for a physical to physical interface connection between to different networks. Thus a repeater may be nothing more than a modulation and demodulation pair of devices in a network.

(ii)Bridge: This type of interconnection takes the first step in interconnection networks where the data format of the bits may be different. Thus if we want to connect one type of LAN to another, or even a LAN to a MAN, we can do so through a bridge, The bridge matches the physical difference in the signal structure as well as the data format.

(iii)Router: This element of the interconnection hierarchies allows for not only the physical and bit format matching but also handles the changes in the addressing between the two networks. Thus we can go from one network where the addresses may be in one format to another network with a different addressing scheme.

(iv) Gateway: When everything is different from one network to another, we need a gateway to match all of the changes from one to the other

7.6 LAYERS AND STANDARDS

There has evolved a standard way to view the communications environment over the past ten years that involves the use of layered protocols. As the reader may recall, we introduced the layered approach first in Chapter one when we introduced the overall multimedia environment. We extended this to software layering in the when we developed the source and in when we integrated the GUI concepts with such elements as X windows. We further extended this in with the data storage and file elements of the design.

In this chapter, the layered constructs have taken even a firmer form with the introduction of international standards. This section reviews the seven layer international standards for communications systems. Layered approaches are useful for both implementation and interfacing purposes. In the implementations context, the layered approach allows for the clear delimitation of functions and shows how they can be partitioned to allow for simpler system design and coding.

Communications networking, like any form of human communications, needs to have a set of protocols or agreed to sets of standards of handling many talkers at the same time. The protocols are ways in which the speakers or participants in any conversation can know how to interrupt and who has precedence in the conversation. A typical example is a meeting in a highly structured company. If the CEO is at the table, all conversation may generally defer to that persons comments and the level of interrupts may depend on

the title and status of the people at the conference table. Inhuman conversations, we frequently have learned these protocols in a cultural environment and they are generally not written down. The need for addressing written protocols occurs when we are dealing in cross cultural environments such as a U.S. business man doing business in Japan, in that case, the difference in and acceptance of protocols is critical.

The same situation is prevalent in communications, and more importantly in multimedia communications. We approach this problem with the development of protocols and these protocols are developed in a highly structured manner. Historically, these protocols were developed in an environment of data communications, where the need was to allow for the interconnection of computers with other computers. The protocols were not envisioned as a way for human interaction with the communications environment. We shall see how this has evolved as we describe the protocols in some detail.

The protocols that we shall describe are in seven layers, each of the layers having the responsibility for a specific set of functions that are necessary for the communications network. Our approach to these protocols is different than most in that we shall begin at the highest layer and work downward. The typical pattern of development and the actual path of development in the standards bodies was from the bottom up since they were interested in protocols that satisfied

the needs of the data communications community. In our analysis, we are interested, first , in the needs of the multimedia communications user.

Thus, we develop the seven layer protocol set from the top layers down. This is a drastically different course in the development and exposition of the OSI protocol set but it reflects the need in the multimedia environment to have the higher layer functions come to the fore.

The seven protocol layers are as follows:

7.6.1 Layer 7: Applications:

The applications layer provides for a wide variety of services that support the overall set of applications that may be running on the end users CPU. These applications can generally run on a standalone basis with no need for the applications layer to even exist. It is necessary, though, when we desire to network the CPUs together in some communicating fashion. A typical example of an applications layer function is the file server function that is an integral part of a local area network environment. In this applications service, a particular applications program, such as a word processor, any desire to obtain a file of letters that have been stored on the file server. The applications program then evokes the file server service and this places a request from the program, through the local area network to the target file server.

Other applications services are such functions as a mail service, a directory service for all the users , a virtual terminal service that allows users to integrate a variety of terminals on the network, and the graphics interface services that we had developed in . Thus the applications layer, is the highest layer that functions to allow for applications to take full advantage of the fact that they can communicate with a variety of other users or applications on the network.

7.6.2 Layer 6 Presentation:

This layer is probably the most improperly named for the purpose of this book. It does not deal with presentation in any way other than possibly how it "presents" data to the applications layer. It has been suggested by Tannenbaum that it be called the representation layer, and by others that it be called the data packing layer. It actually provides a useful function in how the data is to be packaged and repackaged as it comes and goes on the communications network. One service of the presentation layer is that of end to end encryption. This service takes the information that is generated by the end user in an application, and secures it through some encryption scheme that reduces the probability that some nefarious interceptor may capture the data for illegal purposes.

7.6.3 Layer 5 Session:

This layer is the one that the standards bodies have spent the least amount of time on but as we have seen in this book, represents the heart of a multimedia communications environment.

The session is the construct that we have developed that allows many users, in a multimedia environment, to be assured of a seamless, error proof path that ensures the synchronicity of the communications across the path and the graceful interconnection and elimination of any set of users as required. In the multimedia environment, the session is an enduring dialogue between one or many human users and their interactions with applications programs and multimedia data files. Voice and video do not suffer the vagaries of a connectionless network well and require that there be a connection based entity to support their communications requirements, (see Kishino et al, Nomura et al, Verbiest and Pinnoo, and Kishimoto et al) .

In these first three layers we have dealt with concepts that ensure that the higher level communications processes are properly handled. We have implicitly assumed that all other things have been handled in an appropriate fashion. The lower layers will worry about such things as the end to end integrity of the bits, the routing of the bits to the correct places and the physical interfaces being correct.

7.6.4 Layer 4 Transport:

The transport layer is the first "real" communications layer in that it relates to bits and bytes going across the network. It is concerned about the end to end communications of those packets of information, assuring that they get from point A to point B. It assumes that if it send a packet to the lower layers that they, on a one by one basis get where they are supposed to in an errorless fashion, but that this layer must be concerned with the whole set of packets generated by a source.

7.6.5 Layer 3 Network:

The network layer is an integral part of a network that has many parts going in many directions. This may be called a set of subnets. These subnets have to be controlled, and packets that rattle around in this environment must be managed carefully and assured that they get, on a packet to packet basis, from one network node to another.

7.6.6 Layer 2 Data Link:

The data link layer worries about bits getting from one local location to another in an errorless fashion. This layer provides for error detection and management and may also include the capability to correct for errors that occur on the channel. It assumes that the lower layer provides only a transmission path for bits but that the transmission path may have errors.

A significant part of the Data Link layer is the Media Access Control (MAC) sublayer. The MAC layer provides for the local control and access to a distributed and shared communications facility. Typical amongst the MAC capabilities are local and metropolitan area networks.

7.6.7 Layer 1 Physical:

The physical layer includes the physical connection, the modulation scheme and the multiplexing or switching schemes that can be used. Many of the elements of the proposed ISDN (Integrated Services Digital Network) can be found at this layer. The common RS-232 interface is also an element of this layer.

We can explain them in some further detail by considering a specific example and how each of the layers interacts and requires the services of the layers below. The basic transfer element is a data packet and that each of the layers adds a header onto the data packet to effect the services that are provided by that layer. This goes all the way down to the physical layer which is the layer that deals with bits.

Consider the example of a set of multimedia users that desire to communicate in a multimedia mail environment. Let us first assume that thy are using a specific application program to analyze a particular image on the screen. This may be the case of a single user at a single location. That user now desires to share that image with several other users and comment on the results in some multimedia fashion.

We generate a data block that is the multimedia image element. This is the first step in the ultimate process. The data block is in actuality a complex multimedia data object, encompassing video, voice and still image. At the higher layers, the notation of a data block is merely symbolic of a large data element that is to be transmitted across the complex network.

We then invoke the mail service at the applications layer and append to the data block that we wish sent a mail header. This header will contain all the elements that are necessary to evoke the mail function and to ensure that the system sends this total image from one location to all other users in the system. I will further assume that the act of mailing is such that the image being sent is a compound multimedia object, consisting of video, voice and still image, and thus we must have a fully synchronized session based service in place. We evoke the mail service in an applications layer program that provides for electronic mail. It may have also been possible that this mail program may have been resident at another location on another processor. We shall deal with these options at another time.

Since this is a highly proprietary message, we will need to encrypt this message and this service of the presentation layer is invoked and it is represented by the header that is appended on the applications header packet at the presentation layer. This header is a symbolic method of

demonstrating the process of encryption. We evoke the encryption at a point in the system called the data encrypting algorithm, which may be a standalone piece of secure hardware in the communications system.

At the session layer we append a header to the encrypted message to send to a session manager processor. This processor may be a network based element that handles all of the control for the network. This element sends out a set of commands to ensure that a temporary session is established and that all of the session participants are brought onto the session connection. The session manager then continues to clear the channel and to assure that the ultimate block of data associated with this multimedia message is properly handled as a complex multimedia object, and not just a standalone computer record which can be packetized in any random fashion.

The transport layer will now take this packaged and session secure pact and it is required to get it from one point to another, and that is all. It will assume that the data channels are errorless and that they are being routed to the proper locations.

The transport layer takes the session packet and may break it up into smaller packets for transmission and further takes instructions from the session layer so as to ensure the synchronicity of the total packet as it moves across the network. The transport layer control is located in a transport protocol control device that may be part of the computer front end or as a standalone device.

The network layer is needed because we are transmitting this broken up, packets into a network composed of many smaller switching elements. The network layer functions at each of these nodes in such a way to route the packets in the most efficient form from one point to another. This layer focuses on getting the smaller packets across the disparate network.

The data link layer deals with the error control protocol that ensures that the data is kept error free. In this case we assume that the DLC protocol is an HDLC (High Level Data Link Control) protocol that allows for acknowledgments and error detection. Error correction is not provided on transmitted packet and is achieved only by packet retransmission.

Finally, the physical layer performs this task with an RS-232 interface to my computer using a simple on-off 9.6 Kbps modem. This is the easiest layer to understands.

We call view the message at the top most layer as M and then view the actions of each layer as Lk where k represents the layer. At the top most layer we have:

R7 = L7 M

where R7 is the layer 7 representation of the message M that is a result of layer seven operator L7. Thus the mail function may operate by L7 on the message M to generate R7.

In general we can say that;

Rk+1 = Lk+1 Rk

Rk+1 = Lk+l Lk ... L7 M

For example, the operator LI is a modulation operator on R6. Thus R7 is a modulated waveform with specific characteristics.

There are many issues that arise in the discussion of the OSI protocol stacks. Some of them are as follows;

Addressing: How do we know how to get from one point to another in a network. Addresses are necessary at all layers of the protocol.

Each layer has elements, we shall call them entities, that are agents for performing tasks at those layers. The RS-232 interface or the modem are physical entities, and the mail process is an applications entity. We must be able to identify entities at all of the layers.

Active agents are necessary at the layers to effect the operations of the services provided at those layers. These are called service access point (SAP) and these are the points that effect the service operation. We can usually envision these being at the interface between the two layers.

There must be an agreed to format and set of rules for the layers to exchange the data downward. This format is called the interface data unit (IDU).

The IDU consists of a set of control information called the interface control information (Id) and the actual information to be passed along the network, called the service data unit (SDU).

There must be a set of calls or instructions that allow for communications at the same levels in different units in the network. We shall call these primitives and they have their own

syntax. The syntax of the communication language primitives fall into a form :

PRIMITIVE = LAYER.Service_Type.action(parameters)

For example, at the session layer we may be asking the session SAP to create a session and then to add several people to the session. The primitive may take the form:

Session.Create_Session.request(pi,p2pn)

This sends a request to the session processor that a session creation is desired. We can further extend the syntax by including parameter within the primitive to delineate the details of the primitive request. For example:

SESSION.Session_Create.request(source_address,destination_address)

Thus the general form for a primitive will be;

LAYER.Service_Type.action(ul,u2,....,un)

We show the elements of the layers and show how this primitive form of interaction within layers functions

There are four types of actions that can be implemented in the different protocol layers. These actions are;

1. request: asks an entity to do something
2. indication: entity is informed of a request
3. response: entity desires to respond
4. confirm: entity is made aware of request response

The type of Service_Types may vary by layer.

We shall be spending significant effort on the Service Types development of the primitives, especially the Service_Types. These are extremely important at the higher layers for the implementation of the services that can be implemented in these types of networks. We have already discussed several primitive concepts in Chapters 2 and 3 in both the graphics and windowing areas already. The development of these primitives in the communications area will merely be an extension of that effort.

We can envision the two entities in a communications mode, one being the Initiator and the other the Responder. There are two layers of protocol involved. At the Initiator, the user at the higher

layer send a request action to their lower layer. This goes across the network and becomes an indication passed up from the lower layer (Provider) to the comparable layer (User) who then responds. This response is called the response and it goes down and up and it results in a confirm from the Initiators Provider layer to the User layer (see Schwartz p. 88).

7.6.8 Applications Layer

The applications layer is the highest layer and it is structured to support the end user applications as well as provide an environment for the applications to reside in. The applications layer's functions, as we shall describe shortly, focus on direct support of the applications programs. The applications programs may have access to the applications layer services through the primitive functions that we discussed above. It is through these primitives that we obtain access form one layer service to another.

We may view the applications layer services as if they were adjuncts to the operating system, in that we frequently call the operating system services from pure applications program.

7.7 FUNCTIONS

7.7.1 Applications

The applications layer functions both as a service provider as well as a port to the lower layers of the OSI stack. In its function as a service entity, it directly provides services that utilize the network resources that are available to it. Thus, the applications layer can provide such network based services as;

Mail

File Server o Directory

The applications layer may also provide for locally resident services that may be of general use to the end user. These types of services are those of the graphics user interface, terminal emulation services and other specific translations of generic to specific nature. We shall focus on the functionality of the applications layer the relates to its use of the network resources.

7.7.2 Presentation Layer

The presentation layer is in actuality a data re-presentation layer. It takes the data that is provided by the session layer or provided to the session layer and performs certain tasks to represent the data. For example, the presentation layer takes a letter that may be generated in the mail service at the applications layer and encrypts the letter using an encryption service. The service layer

may also take an image that is to be sent at the applications layer and may compress the image using a data compression algorithm of the types that we have discussed in . This layer may also provide a means for changing from

one data format to another, especially in terms of how the data is to be presented to the applications layer.

The presentation layer does not deal with presentation of information to the end user. That, ironically, is dealt with in the applications layer as an applications service. The presentation layer deals with the presentation of data to the applications layer and to the session layer. It focuses on how to change the data from and into forms that meet the overall communications requirements. Also recall that the data or information at the higher layers is not just a block of data it is the entire message.

5.2.2.1 Functions

The presentation layer has several major functions as we have been discussing them. Tannenbaum has presented them as follows;

Provides an access mechanism to the session level services.

Provides a common mechanism for the definition and manipulation of complex data structures.

Provides a management facility for the management of the data structures that are currently in use through the applications or session layer services.

Providing a means to convert a data element from one form to another, on either transmit or receive, for the purposes of more effective or secure communications.

We can see that these types of functions become more complex when we deal with the multimedia environment. Recall, that in the multimedia environment that the session layer tasks on a more significant role and that this role requires that the images be handled in a more robust fashion. The present conceptual structure of the presentation layer deals with data structures that have little time variation. That is they are considered give as a block of data in their entirety and do not have the synchronization problems that we face in video or voice communications. We can envision, possible, the use of presentation layer services that can be used to create the multimedia data elements, and to provide these to the session layer.

Consider, for example, that there is an applications layer program that generates a complex multimedia object. Let us call that object DOk. We let DOk be given by;

DOk = {DOVk,DOIk,DODk,DOTk}

where these sub elements are those for voice, image, video and text respectively. We may envision that the presentation layer has the responsibility to handle the manipulation and integration of this multimedia data object. We have seen in how we can generate this object in a single data file and we shall further see in how we can handle these elements in distributed multimedia database. We see in the presentation layer a functionality that allows us to manipulate these data elements across the network so that we have a mechanism to maintain their total integrity.

7.7.3 Session Layer

The session is the heart of the multimedia communication environment. All that the layers below the session layer do is guarantee that the data packets have gotten from one point to all of the other points that are in the session. They do not guarantee that they got there on time or in most cases even in order. In addition, these lower layers of the protocol sets only view the data transported as a single block, even if it is a created complex multimedia data object. However, the data object may be a sequenced element of an extended session between many users and the context that that block of data finds itself in is as important as the block in and of itself. The session layer is structured to deal with communication "in context". For the most part, it is the only layer that deals with the interactive dialogue nature of conversation environments as a natural part of it functional and service elements.

If we have an error free fiber connection between each and every user, it is conceivable that there is no need for the lower four layers and the only critical layer is that which maintains the session. Ironically, the session layer was not considered as an important layer in the early days of the ISO structures. Tannenbaum discusses the fact that the British had originally proposed that the session layer was not at all needed. As we shall see in this section, the session layer provides some of the most valuable functions from the perspective of a multi user multimedia environment. All the session layer assumes from the services below it is an error free path for the messages that is sends to those layers. All it expects from the layers above is a continual flow of messages that are to be forwarded to those users. The session layer takes the responsibility of managing the interuser conversation, not just the communications.

5.2.3.1 Functions

The functions of the session layer relate to the maintenance and support of the session services that we have just discussed. Some of these functions are as follows:

Dialog Management: This provides for the control over who is intruder control of the session at any one time. We can view the management of a dialogue handled in a variety of forms. The baton approach is based upon a mutual agreement between the session participants that the holder of the baton is in charge and may exercise direct control. The baton may be passed back and forth amongst the session members. Another approach is a priority based session dialogue manager which assigns a priority level to each of the session participants. We shall discuss these and other latter in this section. There are performance issue that relate to deadlock of dialogue management and the delay in session control.

Session Establishment: Session establishment allows for one user to initiate a session and to add other users onto the session in an operational context. The session establishment function allows for the identification and addressing of other virtual users and ensures that they are bound to these session. It also allows for the disestablishment of sessions. We frequently are concerned about the session establishment time as we are concerned about call set up time for a telephone call. Thus in terms of the overall performance of the session establishment function, we look at determining the overall session establishment set up time budget and allocate it to the various system resources used.

Connection Based Service: A session is synonymous with a session. A connection based service is one that establishes a permanent virtual circuit that the session is built upon. The establishment, maintenance and disestablishment of a session is one of the functions of the session layer. The session is built upon the session layer created connection based circuit, that is the permanent virtual circuit that the session layer creates with the transport layer.

Synchronization: When we are dealing in multimedia communications, we are dealing with information that is not only contiguous in terms of spatial relationships but also has a temporal relationship that is critical to its presentation, display and interaction. We may think of what may happen to a video display using an NTSC scan line system if instead of 525 lines being transmitted we get only 524. This would propagate down to the bottom of the system and result in a totally confused interaction. Errors and transport faults do occur and the timing of messages across the network may face significant delays and temporal variances. It is the function of the synchronization service of the session layer to keep all of these complex messages in a truly synchronous mode. The suynchronization service performs this task by inserting various synchronization ticks in the data stream and recovering them at the other locations in the session.

Activity Management: A session is an enduring communications between two or more entities in the communications network. There are moments when the actual communications in the session may be reduced to a low to zero level. We can envision breaking up a session into elements that represent the burst of active interaction between the users and during those times allocate services of the lower layers. We will always want to have the session endure, since the endurance

is a quality of the session. However, it is a function of the session layer to recognize the dormant phases that may occur in a session and to allocate resources accordingly. We do this by creating or defining elements of the total session called activities that are the periods of the active participation of the users of the session and during those periods allocates the resources of the system. This session service allows for the optimum use of the communications resources at the lower layers and assists in the optimization of the shared switched services for transport.

5.2.3.2 Protocols and Architectures

The session layer protocols are similar to those at the layers above but focus on the implementation of the services described in the functions that we presented in the preceding subsection. Recall that the syntax for is as follows:

LAYER.Service_Type.activity

Here we have SESSION for the LAYERS and we still have the four possible activities as: request indication, response, confirm.

These actions relate to the development and specification of the protocol to effect the service. The Service is the indication of which of the services that is being invoked. These Services are: Session, Connection, Dialog, Activity, Synchronization, Report

The Type sub-indicator may tell that it is the beginning, end, discard, resume or interrupt of an activity service. The session Type delimiter provides additional specificity to the command. Recall also that part of the syntax are specifiers to the command syntax the indicate the specifics of the session elements. We have detailed several detailed primitives that are used in an actual multimedia system.

5.2.3.4 Alternatives

Let us consider several implementations of the session layer functionality. The first deal with those of dialogue management and then we consider synchronization and activity management as well as exception reporting.

In dialogue management, the major function is to develop algorithms that allow for the management and control of sessions. This management and control can be handled from the point of view of a token or baton or from a wide variety of other techniques. Let us consider four that are common.

Hierarchical: In this case a session manager in control at all times and all requests for the baton are passed up to the session manager. The requests are handled in some priority fashion based on equal user priority and first come first served to some other set of priorities. It is a non-distributed control algorithm and requires that each session select and adhere to the single point of control.

Baton Passing: This is a round robin approach where the baton is held by a user for a fixed period of time if they need it otherwise it is passed to another users in a round robin fashion, each users is allowed a maximum period to maintain session control and at the end of the period must relinquish it. It may be relinquished earlier.

Priority: In this case, all of the users have a priority, and no two users have the same priority. The priority may be a combination of several factors. Let P(k) be the priority of user k, and we define this as:

$P(k) = R(k) + T(k) + D(k) + I(k)$

where ;

R is the hierarchical level of rank of k

T is the length of time since k last controlled the session

D is the data waiting in k's buffer

I is the image intensity of k's session

This is a totally arbitrary priority scheme but any other priority function may be generated. Th scheme works by all users sending out their priority to the network and the users comparing theirs to others and responding accordingly. The only users to respond is the one with the greatest priority.

Interruption Control: This scheme is a first come first served, shouts the loudest protocol. It is based upon a user's sending a broadcast message out to all users demand access and only when this users messages are received first by all others is this session accepted.

7.7.4 Transport Layer

The transport layer provides for end to end integrity of the communications path. It assumes that there is an errorless and routed infrastructure in the lower layers and ensures that the message that is to be transmitted is done so errorless in toto. For example, the transport layer will receive

a mail message from the higher layers and its takes the entire mail message, partitions it in a fashion to meet the limitations of the transmission paths, and sequences smaller packets across the network to ensure that the total mail piece is sent in an errorless fashion. It is important to note, however, that though this performance may suffice for a computer oriented system, it is not sufficient for the multimedia environment developed in this book. The transport layer allows for the transport on an end to end basis of a message. The session layer ensures the endurance of the conversational mode that is the essence of the multimedia environment. The session layers assumes that the transport layer exists. The transport layer makes similar assumptions of the layers below it.

5.2.4.1 Functions

Transport layer services can be of the connection based type of the connectionless type. The connection based type transport services follow the requirements that we placed on them in the session layer. In this case the transport layer may either support the session layer or may even supplant it in the session

connection service. The connectionless type transport service is merely the ability to provide for a datagram service, ensuring that a single package is transmitted in and error free fashion from one point to another.

Recall that at the transport layer the services are provided to it from the Network SAP and are provided to the session layer from the T SAP. The actual hardware and software that performs the tasks in the transport layer are at the Transport Entity.

The typical functions of the transport layer are:

Connection Establishment: This function provides for the calls down to the lower levels for the establishment of a connection between another user on the network. It is a connection which may also be a permanent virtual circuit and in some ways may emulate the session level connection service. If the transport layer can support a full connection service, then the session service is simply provided by a call to the transport service.

Addressing: Addressing at the transport layer provides for the naming and identification of specific point to be communicated with in the total network. At the transport layer, the addressing function allows for specific identity of the physical and logical location of the virtual user.

Connection Disestablishment: This service is the termination of the connection established in the connection establishment.

Connection Management: The management of a connection is necessary both from the perspective of maintaining the connection as well as providing reports on the status of the connection as it progresses in time. This service allows for interface for the overall network management of the services.

Flow Control and Buffering: The information handed to the transport layer from the higher layers is always the total message and not a fragmented form that is suitable for communications. The transmission paths generally do not support the transport of large blocks of data because the errors inherent in raw transmission would never permit an effective communications system. Thus the communications path uses some form of smaller packets, usually a small percentage of the total information package. Thus the transport layer must accommodate the buffering and segmentation of the total information element and control the flow of the packetized versions across the network.

Recovery: Recovery is a key service at the transport level and it is used as an element of the overall session service. Recovery provides for the restoration and reconstitution of the communications path is there is a fatal failure that causes the collapse of the path. We generally assume that this function is at the transport layer since it has full control over the total network elements.

Level of Service (Quality of Service): The transport level allows for the attachment of level, grades or qualities of

service to the established communications path. These grades of service are related to the performance factors that we shall be discussing in the latter parts of this chapter.

Multiplexing: The transport layer can provide a service that allows for multiple circuits on the same channel by the proper identification of sources and sinks in the network. We shall expand upon this latter.

These services can be created in a fashion similar to those of the higher layers. In the next subsection, we shall focus on some of the specific primitives and how they are architects to provide the transport layer functions or services.

7.7.5 Network Layer

This layer provides for the point to point switching and routing of packets through the network. The services that the data link layer provides is an errorless point to point path. The network layer has to provide the transport layer with a clear path from the beginning to the end of the transmission medium.

5.2.5.1 Functions

The functions of the network layer are as follows: Routing, Congestion and Flow Control, Internetworking, Protocols and Architectures, Performance Issues.

7.7.6 Data Link Layer

The data link layer has traditionally had the responsibility of providing an error free transport on a point to point basis. It assumes that there is an underlying physical connection, but that the connection may be error prone. It further assumes that the connection has been made correctly in some gross sense by commands that have been sent down from possible higher layers in the protocol stack.

There is also a second subfunction that is now incorporated in the DLC layer and it is in the Media Access Control (MAC) sublayer. This layer provides for interconnection of physical devices in an errorless fashion on a distributed network such as a local area network or a metropolitan area network.

7.7.7 Physical Layer

The physical layer describes the interfaces that the computer communications system has with the real world of the communications network. It discusses the physical interfaces, the modulation schemes and signal levels, the data rates and coding schemes, and even the switching and multiplexing schemes. We shall not focus a great deal on this layer in this book since there is a wealth of literature on this layer elsewhere. The reader is referred to such books as Tannenbaum, Schwartz, and others.

There are several functions of the physical layer. They are; Physical Interface, Mechanical to Electrical Conversion, Switching, Transmission, Modulation, Multiplexing, Protocols and Architectures, Performance Issues.

7.8 BROADBAND ALTERNATIVES

The most pressing requirement for the communications system in order to ensure the ubiquity of the service that we are developing in this book are the elements that relate to the transport of data in very high data rates, these rates are in the regions of tens of Mbps to the higher Gbps range. These rates are now achievable with the introduction of fiber technology, fibers that are made of glass filaments. These fibers have been developed and perfected in the past ten years and now are at the stage that they are both ubiquitously deployed and in addition various implementations are available for implementation.

In this section we discuss three of the most common alternatives for the deployment of high Mbps data channels, ones that also have a potential for evolving into Gbps channels. From a technical point of view, a single strand of fiber can theoretically support data rates in the Tbps range, that is 10^{12} bits per second. However, the technology to interface this data rate to the network is not yet available in commercial form. We shall not discuss these factors in this chapter but leave this to the reader. The texts by Personick, Kaiser and others provide significant detail on the fiber technology and the status of the opto electronic interfaces.

The choice of the three techniques discussed in this section is based upon the facts that they are either available of will be currently available in the near future. FDDI is a presently available system that supports 100 Mbps data channels and is being deployed for intra premise network systems. It frequently works on multimode fiber systems and can be interconnected to inter premise systems. SMDS is an embodiment of a metropolitan area network architecture that is being planned for shared switched network deployment. It works at the 45 Mbps rate and can go to the 150 Mbps range. It is a dual bus architecture and can support a shared switched network operation. Finally, ATM is the embodiment of the Broadband ISDN system and is farther off in implementation.

7.8.1 FDDI

We start first with a broadband system that is presently in operation and provides for 100 Mbps capability in internetting local area networks. The FDDI(Fiber Distributed Data Interfaces) systems are typically used in intra facility networks although they are not limited to these areas. The generally are also used in an non switched and non-shared environment, although changes are being developed to permit them further capability.

The FDDI systems were developed to meet the need of the users of LAN systems and to provide the capability to allow these system to be interconnected at higher speeds. Initially, the LAN to LAN interfaces were at considerable lower speeds and thus made for slow file access and transfer time on these types of networks, recognizing the needs for expanded interfaces speeds the use of 100 Mbps was the desired bit rate on the interconnect network. There was, however, no thought originally given to the implementation or operational issue of a multimedia environment. We shall however show how this can be incorporated into this environment.

The architecture of the FDDI system is built upon a dual ring bus structure that uses a fiber optic cable. The fiber cable is frequently a multimode fiber and the bus architecture allows for graceful degradation of the system in the event a single node or a fiber failure occurs. The ring architecture has been chosen in place of the other forms because of the inherent reliability of the ring in the event of a single point failure. Each node has a repeater function as well as the capability to detect the FDDI packets as they are transmitted around the network.

The IEEE 802.5 MAC protocol is used on the ring for the control of the access to multiple users. This protocol is a token passing scheme that passes a token around the network that allows the users to grab the token in the event that they have data to transmit or otherwise to pass the token to other users. As we discussed in the previous section, on the topic of the 802 standards, this particular IEEE standard allows for control on ring structures. The other protocols are generally found on bus structures.

the message an indicate whether it has been received or not properly and retransmit this change message around the circuit.

The message will continue to transmit around the ring, and when received by the original station, it will read the message and will recognize if the message was successfully transmitted or not. If not a retransmission may be performed if it has not been timed out.

Let us first consider the implications of the circulation of this single token. The speed of propagation of a signal in a fiber is approximately 200m/microsec and the free space speed of light is 300 m/microsecond. Thus the speed of propagation of about 0.66 that of free space. Now consider a ring message packet of 100 bits and consider an FDDI ring of 1,000 m. The ring propagation time for a packet to go around the entire ring, assuming that there are no delays anywhere, is 5 microseconds. If the message is 100 bits, then in order for the bits not to overlap the data rate must be greater than 100 bits/5 microseconds or 20 Mbps.

When the token bit is changed, that station now can send data for a period as long as THT (token holding time) seconds, usually 10 msec, and then it must release the token. When a station desires to seize the next token, it does so by looking at the data message being sent around the ring and when it wants to seize the token, it enters a message in the AC byte to indicate its priority. If there is a message there already, it waits. If there is a request there already but it has a higher priority, then it may be able to bump the other request. When the station stops due either to time out of data finished, the token is passed to the requesting station if there is one or a blank token request starts circulating again.

The performance of the FDDI networks has been studied in some detail and we present the result outlined in Bux. Some of the analysis of this protocol for two different channel rates and for the presentation of the mean transfer delay time as a function of the information throughput. Several observations are clear;

The greater the length of the cable the greater the delay.

The greater the length of the packet the greater the delay.

There is an instability that occurs in the delay as the throughput attempts to reach the data rate of the bus. Thus the maximum throughput is less than the maximum bus rate as is expected.

FDDI applications are quite appropriate for intrafacility locations or those involving shorter distances. As the distances increase we have seen that the average delay increases proportionately and that this is an inherent element of the token ring structure.

7.8.2 MDS

This represents a second step in the development of a broadband capability. SMDS, switched multimegabit data services, is a shared switched metropolitan area network (MAN) that is being proposed for implementation in local transport areas. It is built around the 45 Mbps standard.

The architecture for the SMDS system is comprised of a dual bus structure, not a closed ring, and the forward part of the bus transmits packets with data plus a bit that is called the busy bit. The busy bit indicates if the packet is empty or has data. At each node or station on the SMDS network, the packet is decoded and the busy bit is read. We shall discuss the status of the busy bit shortly.

The SMDS system is a connectionless datagram approach to higher speed communications networks. The network designs are developed for the use in metropolitan area networks (MAN).

There is a second or reverse bus in the dual bus scheme. The reverse bus transmits a packet containing no data but it does contain a request bit (REQ). This bit is generated by any one of the stations upwards of the receiving station. The bit is set to one by the station desiring to transmit.

The stations read both the forward and the backward bits.

The SMDS protocol acts in the following fashion. Recall from the previous paragraph that on the forward channel a station that transmits data transmits not only the data but a bust bit, BUSY. We shall call the complement of this the IDLE bit. If the IDLE bit is l then the packet is empty. Also recall that on the reverse channel we have the request bit, REQ, that if it is 1, it means that some station upstream desires to transmit.

We can now describe the SMDS MAC protocol for transmission purposes. This is called the Distributed Queue Dual Bus (DQDB) algorithm and is the heart of the 802.6 MAC protocol.

Each station has a Request Counter, called RQ. The contents of the RQ at any time k is called;

RQ(k) = contents of the request counter at time k

When a data block on the reverse channel desires to send a message it generates a REQ bit that is loaded into the RQ. Thus when the station sees a reverse packet with a REQ we have;

RQ(k+l) = RQ(k) + REQ(k)

When the forward channel sees and IDLE bit set at time k+2, it subtracts the IDLE bit from the request counter, RQ.

RQ(k+2) = RQ(k+l) - IDLE(k+l)

Thus RQ measures the number of requests left upstream less the number of idle packets that can satisfy those requests. Or, in

effect, RQ is a measure of the status of the upstream demand at any time.

When a node wishes to transmit it transfers the contents of the RQ to another buffer called the countdown buffer, CD.

It then separates the RQ from the forward path and keeps it connected to the reverse. Thus RQ(k) becomes;

RQ(k) = RQ(k-l) + REQ(k)

The countdown buffer is connected to the forward path. Then the CD buffer is as follows:

CD(k) = RQ(k); at the time of transfer

CD(k) = CD(k-l) - IDLE(k); when an empty packet passes

When CD(n) = 0, the station transmits a packet.

This protocol is dramatically different than the FDDI 802.5 protocol since it incorporate a minimum distributed intelligence in each node of the network.

The SMDS protocol has three layers. They are;

Level 3: Provide the SMDS addressing information. Also error detection is handled at this level.

Level 2: Segmentation and reassembly is provided at this level.

Level 1: This is the physical layer functionality.

Some simple performance analysis has been done on the SMDS protocol. To data, however, there is not as detailed a study of the DBDQ protocol as there is of the many other protocols. However, Newman et al have presented an analysis of the DBDQ protocol in terms of the access time delay and compared it to FDDI. As we noted in the FDI case, the delay increase as the network increase in size and also as we increase the traffic. In fact, it is possible to make the FDDI delay dependent upon the distance of the ring alone and have it independent of load. This fact is not at all evident.

SMDS provides an intermediate choice for data communications over a larger area, typically 100km and greater. It allows for many users to access a network and to do so at speeds that are consistent with both the telephone network and he local area network speeds. Thus SMDS presents an ideal candidate for metropolitan area networks (MANs).

7.8.3 ATM/BBISDN

The previous two networks are in various stages of implementation with their standards having been well defined and developed and the technology either available as with FDDI or clearly under development as is the case with SMDS. Broadband ISDN, is in some sense the son or grandson of ISDN which is slowly having limited introduction in some telephone areas. The broadband version of ISDN has two major directions. The first is the development of a synchronous system which has been built around the SONET standard and the second is an asynchronous system called the asynchronous transfer mode (AT) and this provides the most effective transport environment for the full end to end transport.

7.9 NETWORK MANAGEMENT

The issue of network management addresses the need of both the operator of the system to control the resources available and the need of the system user to obtain maximum benefit of the system and all of its resources. The task of network management is best stated by what is has to do for the user of the system. A good network management system must allow the user who is employing a wide set of system resources, and when the system fails in any way to meet the user's needs, such as the return key not work, allowing the user to have a single source of contact who will return all of the resource to the user in a form which allow the user to complete their tasks. This function must be achieved in as transparent and direct fashion as possible.

In this section, we discuss network management from the perspective of the communications system, but we must remember that all of the elements of the multimedia environment must be

controlled in as seamless a fashion. This network management in a multimedia environment must deal with the end user from the interface, through the data and file structures, through the communications network and including all of the network services and applications. This is a goal of such a system and this goal may not be readily achievable in the context of many of today's systems. However, if we begin to recognize the need of this function, we can more readily factor these into the overall design at the earliest possible stages. This is an example of a typical multimedia environment that needs the network management problem. Specifically, the network includes the following elements;

The network management problem can be stated as is shown. We have shown several of the sub network elements and have in addition shown that the elements are comprised not only of their physical interfaces but of all seven OSI layers. The problem then is if there is a network problem in any one layer it must be identified and corrected, and more importantly, if a problem arises as a result of the interaction between layers, then this is the critical issue associated with network management.

The solution of the network management problem is to develop a manager of managers that looks at all of the seven layers of each of the sub network management elements in each. It then can present an integrated system to identify, correct and restore the network functionality.

7.9.1 Functions

The network management system must perform the following functions;

Interfacing: This function allows for the interfacing of the management system to talk to and control other sub network management systems. The interfacing function can be approached in two different ways. The first fashion is the approach of defining a standard interface to the overall management system and have all of the sub network vendors meet this interface. The second approach is to have a flexible and referable interface that does not require vendor modification.

Status Monitoring: This function provides for the monitoring of the status of each of the individual sub network elements.

Performance Monitoring: With the results from the status monitoring the performance monitoring function provides for an evaluation of how each of the sub network elements are functioning with regards to the standard in which they are to perform.

Performance Analysis: The function of analysis determines the details of the system errors and faults and allows for the determination of how and where this can be improved.

Inventory Management: This function is a key ingredient that provides for a ready access to all inventory of other elements of the system as well as the ability to achieve the restoral that is necessary.

Restoral Activation: This function provides for the reconstruction of the assets of the network to perform its overall tasks.

Reconfiguration Management: This function provides for the management of the overall reconfiguration task in contrast to the restoral activation function that actually implements the restoral. The reconfiguration management, looks not only at the resole functions of getting the system back to a prior state after a fault occurrence but also, and more importantly getting the system into a new state, one that is planned and orchestrated.

Report Generation: A necessary function is the preparation of reports that are both event and time driven in nature. An event driven report is one that occurs when a particular event occurs. A time driven report is prepared at specified time intervals. The report generation function of the network manager must be flexible in its ability to meet the changes in the environment as

well as flexible to be presented on various out display devices, print and video.

7.9.2 Architecture

There are many subnetwork elements in the system and there are multiple interfaces in the overall architecture. The overall system architecture shows that the system operates in a multi domain applications. The system must cover two dimensions of operations. The first dimension is the interfaces across multiple sub network management systems, typically integrated at the lower level of the OSI, layers. The second level is within the same network and upwards thorough the OSI layers. A third dimension is desired but highly complex, is that integrating the first two dimensions.

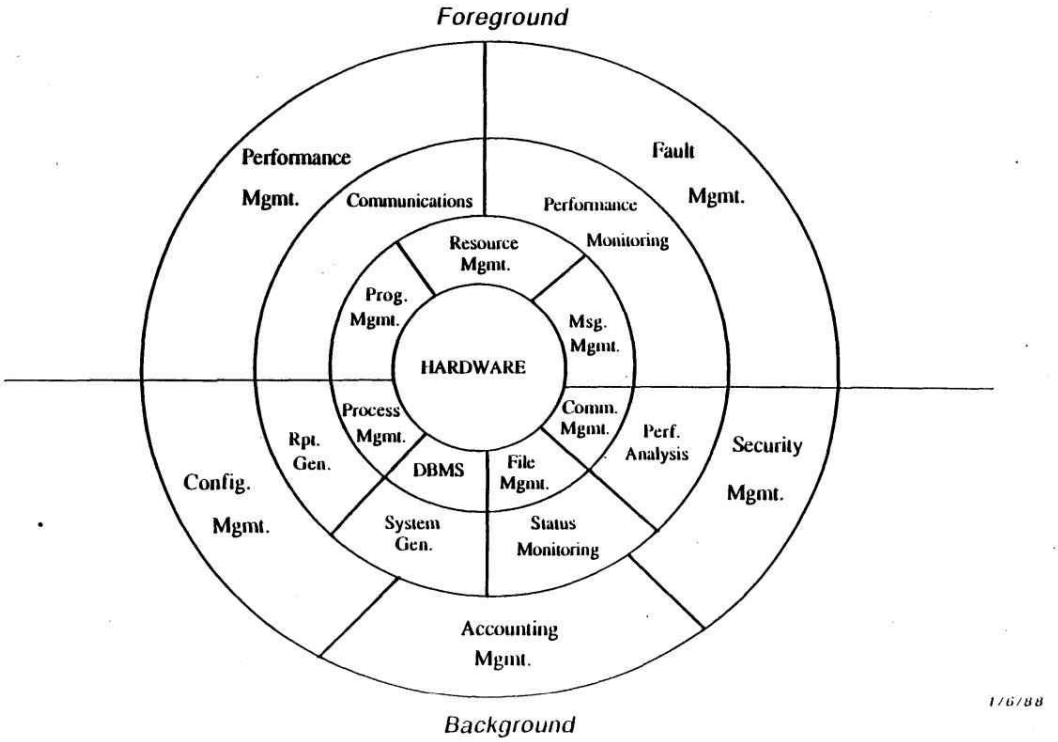
The overall architecture consists of two major elements; the software and the hardware elements. The system architecture shows that there are common interfaces between the overall network manager and the sub network elements, these interfaces may be of a customizable type or even matched directly to each interface element. The software of the system must provide for the flexible interfaces to each of the systems and then must also be able to provide for the functions that we have defined above. The display of the results as also a key element in the design of the network management capability. The display function may have to be flexible to meet the needs of all the end users.

7.9.2.1 Software

The system software architecture is depicted below. It is composed of multiple layers that internally match the interfaces to the local subnetwok managers and externally face the end user. The inner facility of the network manager is the kernel functionality that must meet and interface with the multitude of other subnetwok elements. The outer layer is called the shell and it is through the functionality of this layer that the end users have access to the system.

Figure 2

Network Management Software Architecture



We further take the functionality at each of the two layers and further divide it into foreground or real-time activities and background or nonreal time activities. The need for this partitioning is both for the implementation of the codes as well as for the implementation of the manager on a real time platform or target machine.

The kernel functions include the following:

Communications Interface: This function provides for the subnetwok interface and handles the protocol support to manage the vagaries of the different interfaces. This function allows for the parsing of incoming messages and the same function for outgoing commands. The general

parsing function provides for a flexible command interface and allows for the operation of that interface as a common element in the network manager.

Performance Monitor: This is a time and event driven real time function that monitors all of the incoming status reports and determines if the report indicates an error status. The error status can be divided into several categories;

Error Level Detection: This provides for the determination of the severity of the error and categorizes it instantly for action.

Error Correlation: This is a complex task that takes less severe errors or faults and generates an assessment of the interaction of these faults to generate more complex errors.

1. Performance Analysis
2. Status Monitoring
3. Report Generation
4. System Analysis

In a similar fashion, the shell level functions can be given as:

1. Performance Management
2. Fault management
3. Configuration Management
4. Security Management
5. Accounting Management

7.9.2.2 Hardware

The network management hardware architecture is fairly simple. The network is a complex collection of many elements and the network managers function is to acquire from a wide set of existing interfaces a complete collection of the network faults. These faults are locally generated and they are collected either in band or out of band of the existing signal paths. The in band collection uses the existing network paths during the periods in which they are properly functioning. In the event that they cease functioning a separate set of communications paths must be established. These are the out of band paths. Some architectures have these out of band paths in place for all communications. The latter approach may be costly and also occur the risk that they themselves may be error prone.

7.9.2.3 Implementations

This section describes several of the network management system implementations that are currently available.

7.10 NETWORK PERFORMANCE AND SIZING

The issues of performance and sizing are complimentary issues. Performance addresses the issues of; given a specific load on the system and a specific set of system capabilities, what are the performance factors for the system in such areas as delay, throughput, response time and others. Sizing, on the other hand, addresses the issue of; given the fixed set of performance goals in terms of capacity, delay, throughput etc, how many users can the system support and how large should the system resource should be. Thus it is clear that when we address on of the two issues we address both.

In this section, we address the two issues from the viewpoint of the overall multimedia environment and do not provide a detailed discussion of either of the two elements. We leave these as details to be developed in the references and in the problems the end of the chapter.

The communications element of the multimedia environment is frequently thought of as the most easily understood part of the overall system. All that is needed is faster bandwidth an greater connectivity. It is however the mosey complex because of the time scale of communications network evolution and system design. There are communications network that are intra premises and those that are inter premise. In this chapter we have developed the concepts of the standard architectures that we all assume will take care of the environments and their mixes. However, we see that there is clearly an asset of multiple alternative for the multimedia environment even at the lower layers of the protocol.

In this chapter we have also introduced the concept of network management. This is a critically important element in the overall multimedia system design and must be integrated into all of the elements that we have developed in this book. Unfortunately this is not the case and the user is often let to fend for themselves in the event of errors or faults occurring in some part of the network.

We have also further developed the clear distinction of sizing and performance in this environment and we shall extend it to the overlap environment of the multimedia network as we progress.

8 SESSION MANAGEMENT

Multimedia Communications involves itself with the communications of highly distributed multimedia data objects that require precise timing at and between multiple locations. This paper proposes a way to handle this level of communications through enhancements made at the Session Layer of the OSI protocol standard. The approach taken starts with a definition of multimedia data objects and then develops the required elements for the Session Layer. Detailed implementations are presented and discussions on their performance comparisons are discussed. Multimedia Communications is a discipline that combines the ideas of the human senses, disparate storage and data structures, varying interfaces and complex communications systems. The basic concept of a multimedia environment has evolved from that of the single media data focused world of the computer specialist to the need to provide a fully integrated system for a human user to interact with using information stored on many different storage media. Multimedia consists of a matching of the three elements of the senses, the storage media and the interface devices.

It has been argued elsewhere that multimedia should not be confined to merely the storage of information of multiple storage devices. Rather, multimedia must include the senses and the interfaces as well. In fact, for the purpose of this paper we define multimedia as the confluence of storage, senses and interfaces. Specifically, multimedia relates to constructs of not only information storage but also information processing and communications. It encompasses all of the senses, although we currently only focus on the senses of sight, sound and touch. The definition that we take of multimedia in this paper is an expansive definition. It has been taken to provide a basis for the next step which is multimedia communications, which takes the multimedia paradigm and adds multiple human elements and as such transcends the prototypical computer communications view of the world.

When we introduce the communications concepts, we do so in the context of having multiple users share in the use of the multimedia objects. Thus multimedia communications requires that multiple human users have sensory interfaces to multiple versions of complex objects stored on multiple storage media. In contrast to data communications in the computer domain, where humans are a secondary after thought, and optimization is made in accordance with the machine to machine connection, multimedia a communications is a human to many other human communications process that must fully integrate the end user into the environment. Multimedia communications thus generates a sense of conversationality, it is sustainable over longer periods, and it has an extreme fluidity of interaction.

Various authors have recently addressed the issue of multimedia communications with an architectural approach. (See Little and Ghafoor, Nicolaou, and Steinmetz). The current approaches

focus on one of two extremes, either on broadband communications and the transport mechanism or on the multimedia storage aspects of the system design. Little and Ghafoor have attempted to integrate the presentation and data object side of the problem and have at a higher level, attempted to address the communications issues. Nicolaou has developed a communications architecture that follows the OSI standards but in attempting to introduce the multimedia issues has been forced to introduce several new constructs. Various other researchers in this area have focused on the lower protocol layers and have specifically been concerned with transport layer problems and below.

One of the major challenges to multimedia communications is that today there are broadband architectures that are developed that provide higher speed communications using direct extensions of the techniques developed in the data world of packet communications. Specifically such techniques as ATM and SMDS, as well as FDDI are direct offshoots of local area networks and packet technology. The fail to understand the paradigm that we are developing in this paper that relates to the structure of the multimedia object and the conversationality of multimedia communications.

In this paper, we concentrate on three issues in the area of multimedia communications; the data objects, the conversationality of the interaction and the overall communications architecture. We first note that the data structures in multimedia environments are dramatically different than those in normal computer data communications. Specifically, Mullender has shown that typical data file sizes that are transferred in a UNIX environment are on the order of 2K bits whereas in a multimedia environment the file size may average 100 Mbits. Secondly, a multimedia environment needs to handle real time data interaction such as that in real time voice and video. As is well known, such transport protocols as TCP/IP are not adequate from a delay perspective to support these types of data objects.

The conversationality aspect of the multimedia environment is key to effective communications. In this paper we focus on utilizing the Session layer from the OSI format for the delivery of the multi-user conversationality. Historically, the session layer (See Tannenbaum) has been relegated to a secondary position in the OSI hierarchy. In a multimedia environment, we show that the session functionality, refined and expanded, provides the essential integrating capability for conversationality.

The remaining communications services, at OSI layer 4 and below, become, at best, delimiting factors in the communications environment. In this paper we show that there are certain underlying performance factors of the lower four layers, that when combined control the overall end to end performance as viewed from the users perspective. As a major point in this paper, we argue that the standard approach to communications system design, from the physical layer and up is the wrong way to proceed for multimedia. Specifically, in a multimedia environment, one must, perforce of user acceptance, design the system from the top layers and down.

8.1 MULTIMEDIA DATA OBJECTS

In a more standard computer communications environment, the data objects have significant structure and they are frequently integrated into a system wide data base management system that ensures the overall integrity of the data structures. In a multimedia environment, the data elements are more complex, taking the form of video, voice, text, images and may be real time in nature or can be gathered from a stored environment. More importantly, the separate data objects may combined into more complex forms so that the users may want to create new objects by concatenating several simpler objects into a complex whole. Thus we can conceive of a set of three objects composed of an image, a voice annotation and a pointer motion annotating the voice annotation. The combination of all three of these can also be viewed as a single identifiable multimedia object.

Before commencing on the issues of communications, it is necessary to understand the data objects that are to be communicated. We can consider a multimedia data object to be composed of several related multimedia data objects which are a voice segment, an image and a pointer movement (e.g. mouse movement). As we have just described, these can be combined into a more complex object. We call the initial objects Simple Multimedia Objects (SMOs) and the combination of several a Compound Multimedia Object (CMO). In general a multimedia communications process involves one or multiple SMOs and possibly several CMOs.

The SMO contains two headers that are to be defined and a long data sting. The data string we call a Basic Multimedia Object (BMO). There may be two types of BMOs. The first type we call a segmented BMO or SG:BMO. It has a definite length in data bits and may result from either a stored data record or from a generated record that has a natural data length such as a single image screen or text record. We show the SMO.

Figure: SMO Structure

The second type of BMO is a streamed BMO, ST:BMO. This BMO has an a priori undetermined duration. Thus it may be a real time voice or video segment.

A simple multimedia object, SMO, is a BMO with two additional fields; a Synchronization field (Synch) and a Decomposition field (Decomp). We now depict the SMO structure in detail. The Synch field details the inherent internal timing information relative to the BMO. For example it may contain the information on the sample rate, the sample density and the other internal temporal structure of the object. It will be a useful field in the overall end to end timing in the network.

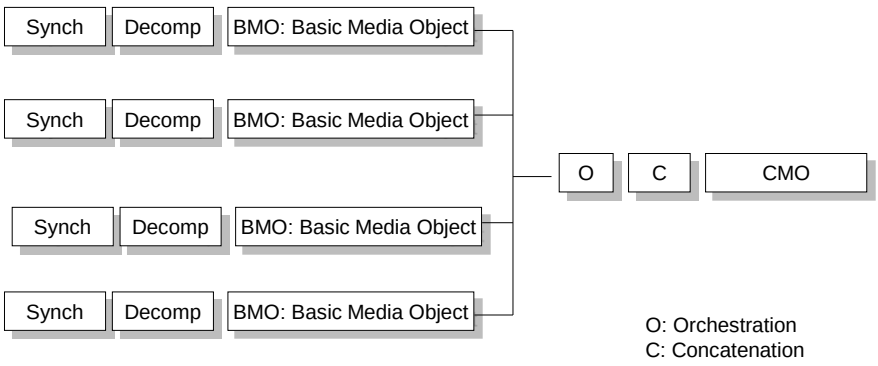
The second field is called the Decomp field and it is used to characterize the logical and spatial structure of the data object. Thus it may contain the information on a text object as to where the paragraphs, sentences, or words are, or in an image object, where the parts of the image are located in the data field.

These fields are part of an overall architecture requirement finds it necessary to provide an "out-of-band" signaling scheme for the identification of object structure. The object structure is abstracted from the object itself and becomes an input element to the overall communications environment. Other schemes use in-band signaling which imbeds the signal information with the object in the data stream. This is generally an unacceptable approach for this type of environment.

When we combine these objects together we can create a compound multimedia object. A CMO has two headers, the Orchestration header and the Concatenation header. The Orchestration header describes the temporal relationship between the SMOs and ensures that they are not only individually synchronized but also they are jointly orchestrated. The orchestration concept has also been introduced by Nicolaou. In this paper we further extend the orchestration function beyond

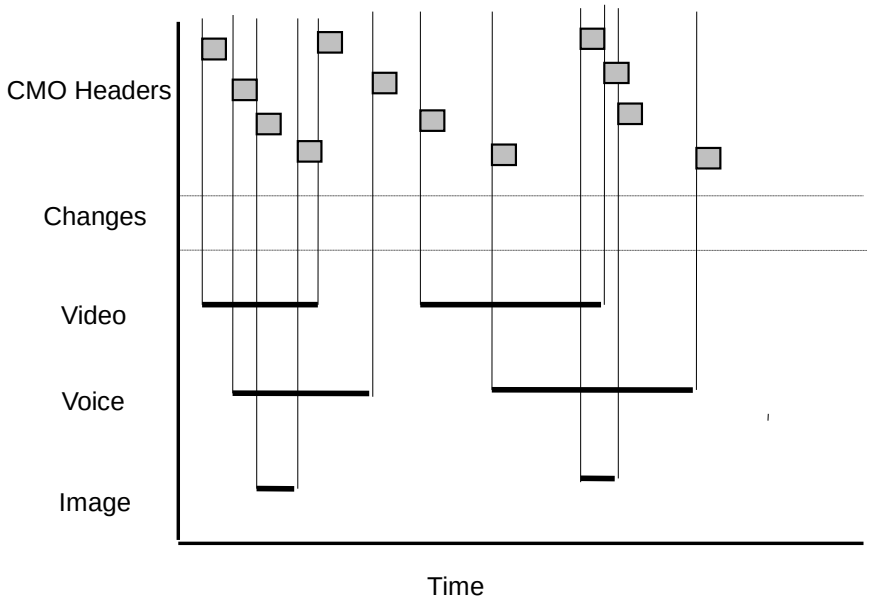
that of Nicolaou. The concatenation function provides a description of the logical and spatial relationships amongst the SMOs.

Figure: CMO Structure



These concepts have been further developed in McGarty[2] and there we have provided more detailed structure to the multimedia data objects. We then also plot the times that the CMO, the concatenation of all simultaneous objects, change in this system. Then we also show the CMO headers that are flowing in the network at each change interval. It is this dynamic process of data elements that must be controlled by the session layer to be discussed in the next session.

Figure: Temporal Interaction of CMOs



We can also expand the concept of a CMO as a data construct that is created and managed by multiple users at multiple locations. In this construct we have demonstrated that N users can create a CMO by entering multiple SMOs into the overall CMO structure.

The objectives of the communications system are thus focused on meeting the interaction between users who are communicating with CMOs. Specifically we must be able to perform the following tasks:

Allow any user to create an SMO and a CMO.

Allow any user or set of users to share, store, or modify a CMO.

Ensure that the user to user communications preserves the temporal, logical and spatial relationships between all CMOs at all users at all times.

Provide an environment to define, manage and monitor the overall activity.

Provide for an environment to monitor, manage and restore all services in the event of system failures or degradation.

We shall see in the next section that the session layer service address all of these requirements.

8.2 SESSION LAYER FUNCTIONS

The OSI layered communications architecture has evolved to manage and support the distributed communications environment across error prone communications channels. It is presented in detail in either Tannenbaum or Stallings. A great deal of effort has been spent on developing and implementing protocols to support these channel requirements. Layer 7 provides for the applications interface and generally support such applications as file, mail and directory. The requirements of a multimedia environment are best met by focusing on layer 5, the session layer whose overall function is to ensure the end to end integrity of the applications that are being supported.

Some authors (See Couloris and Dollimore or Mullender) indicate that the session function is merely to support virtual connections between pairs of processes. Mullender specifically deals with the session function in the context of the inter-process communications (IPC). In the context of the multimedia object requirements of the previous section, we can further extend the concept of the session service to provide for IPC functionality at the applications layer and specifically with regards to multimedia applications and their imbedded objects.

The services provided by the session layer fall into four categories:

Dialog Management: This function provides all of the users with the ability to control, on a local basis as well as global basis, the overall interaction in the session. Specifically, dialog management determines the protocol of who talks when and how this control of talking is passed from one user to another.

Activity Management: An activity can be defined as the totality of sequences of events that may be within a session or may encompass several sessions. From the applications perspective, the application can define a sequence of events called an activity and the session service will ensure that it will monitor and report back if the activity is completed or if it has been aborted that such is the fact.

For example, in a medical application, we can define an activity called "diagnosis" and it may consist of a multiple set of session between several consulting physicians. We define a beginning of the activity when the patient arrives for the first visit and the end when the primary physician writes the diagnosis. The session service will be responsible for ensuring that all patients have a "diagnosis".

Synchronization: We have seen that at the heart of a multimedia system is a multimedia data object. Each of the objects has its own synchronization or timing requirements and more importantly, a compound object has the orchestration requirement. The session

service of synchronization must then ensure that the end to end timing between users and objects is maintained throughout.

Event Management: The monitoring of performance, isolation of problems, and restoration of service is a key element of the session service. Full end to end network management requires not only the management of transport and sub network, but requires that across all seven OSI layers, that overall end to and management be maintained (See McGarty and Ball).

Here we have shown the session entity which is effectively a session service server. The entity is accesses from above by a Session_Service Access Point (S_SAP). The session entities communicate through a Protocol Data Unit (PDU) that is passed along from location to location. Logically the session server sits atop the transport server at each location.

The servers are conceptually at a level above the transport level. We typically view the transport servers as communicating distributed processes that are locally resident in each of the transmitting entities. This then begs the question as to where does one place the session servers. Are they local and fully distributed, can they be centralized, and if so what is their relationship to the Transport servers. Before answering these questions, let us first review how the session services are accessed and how they are communicated.

Session services are accessed by the higher layer protocols by invoking session service primitives. These primitives can invoke a dialog function such as Token_Give. The application may make the call to the S_SAP and this request may be answered. There are typically four steps in such a request, and these are listed in Stallings who shows that the requests are made of the session server by entity one and are responded to by entity two. The model does no however say where the session server is nor even if it is a single centralized server, a shared distributed server, or a fully distributed server per entity design. We shall discuss some of the advantages of these architectural advantages as we develop the synchronization service.

8.3 DIALOGUE MANAGEMENT

Dialog management concerns the control of the end user session interaction. Specifically, who has permission to speak and when, who can pass the control and how is that implemented. In this section we shall describe the environment for the dialogue management function and develop several possible options for implementing this function.

Dialog management requires that each of the virtual users have a token or have access to a baton in order to seize control of the session. In the course of a typical session, the two virtual users fist establish the initial sub session that becomes the first part of the session. The addition or binding of other virtual users through sub sessions to the session allows for the growth of the session. The

baton or token may be a visible entity that is handed from one to the other or it may be hidden in the construct of the applications.

Consider the session level service called dialogue. The service can be implemented in four possible schemes. These schemes are:

(1) Hierarchical: In this scheme there is a single leader to the session and the leader starts as the creator of the session. The baton to control the session can be passed upon request from one user to another, while full control remains with the session leader. The session leader may relinquish control to another user upon request and only after the leader decides to do so. The leader passes the baton from users to user based upon a first come first serve basis. It is assumed that each users may issue a request to receive the baton, and that any requests that clash in time are rejected and the user must retransmit. There transmit protocol uses a random delay to reduce the probability of repeated clashing. The leader always acknowledges the receipt of the request as well as a measure of the delay expected until the baton is passed.

(2) Round Robin: In this scheme, the baton is passed sequentially from one user to another. Each user may hold the baton for up to Tbat sec and then must pass the baton. When the baton is held, this user controls the dialogue in the session.

(3) Priority: In this case, all of the users have a priority level defined as $P_k(t)$, where k is the user number and t is the time. We let the priority be;

$$P_k(t) = R_k(t) + T_k(t) + D_k(t)$$

Here R is the rank of the k the user, T is the time since the last transmission and D is the data in the buffer. We assume that some appropriate normalization has occurred with this measure.

Every T_{check} seconds, each users, in sequence sends out a small pulse to all other users, on a broadcast basis, and tells them their current priority. Each user calculates the difference between theirs and all the others. User k calculates a threshold number, TR_k, which is;

$$TR_k = \max |P_k(T)-P_j(T)|$$

If $TR_k > 0$, then user k transmits its packets for T_{send} seconds.

(4) Random Access: Each user has a control buffer that indicates who has control of the session, namely who has the baton. The session is broken up into segment T_{sess} in length, with T_{req} seconds being relegated to a baton ownership selection period and T_{sess}-T_{req} being left for the session operation. During T_{req} , all of the users transmit a request packet that is captured by all of the other users buffers. T_{req} is broken into two parts, T_{send} and T_{eval}. These requests are broadcast in T_{send}.

Now after the sent messages are received, one of two things can happen, the message is received or it collides with another message and is garbled. If the message is garbled, the buffer is not loaded and is left empty. If it is filled, then each buffer during T_{eval} sequentially broadcasts its contents and all of the users listen to the broadcast and record the counts, N_k where N_k is the number of votes for user k in that call period.

The choice of baton control is then;

Choose user k if $N_k = \max_j | N_j |$

else restart T_{req} again.

For each of the protocols we describe the advantages and disadvantages of each in Table 4.3.

Table 4.3 Dialog Protocol Comparison

Protocol	Advantage	Disadvantage
Hierarchical	Single Point of Control of the Session.	Lacks capability to have open discussion.
Priority	Establishes who is in charge by allocation.	Requires a scheme to give priority that may be open to compromise.
Round Robin	Everyone gets to talk. Egalitarian approach.	May be excessively time consuming.
Random	Strongest player wins.	May not permit dissent.

8.4 ACTIVITY MANAGEMENT

Activity management looks at the session as an ongoing activity that users may come and go to. This services provides an ability to easily add, delete and terminate the entire session.

An activity in the terms of the session is a total bounded event that can be compartmentalized in such a way that other events may be locked in suspension until that event is complete. Activity management is in the session layer a function similar to transaction management in a transaction processing system. It allows for the definition of demarcation points that permit suspension of activities in other areas until the activity managed transaction is complete. Activity management can also be developed to manage a set of events that can be combined into a single compound event.

There are several characteristics that are part of activity management:

Activity Definition: This allows for the defining of an activity as composed of several dialogue. It allows for the defining of the activity as a key element of a single session or even to expand over several sessions.

Activity definition is the process of informing the session server of the beginning and end parts of an activity and in providing the session server with an identifiable name for the activity.

Activity Integrity Management: Activities are integral elements of action that cannot be segmented. The activity management system must ensure that once an activity is defined and initiated, hat no other activity that could interfere with the existing one is allowed to function.

Activity Isolation: The ability to provide integrity is one part of managing the activity. Another is the ability to isolate the activity from all others in the session. An activity must be uniquely separable from all other activities, and this separation in terms of all of its elements must be maintained throughout its process.

Activity Destruction: All activities must be destroyed at some point. This is a standard characterization.

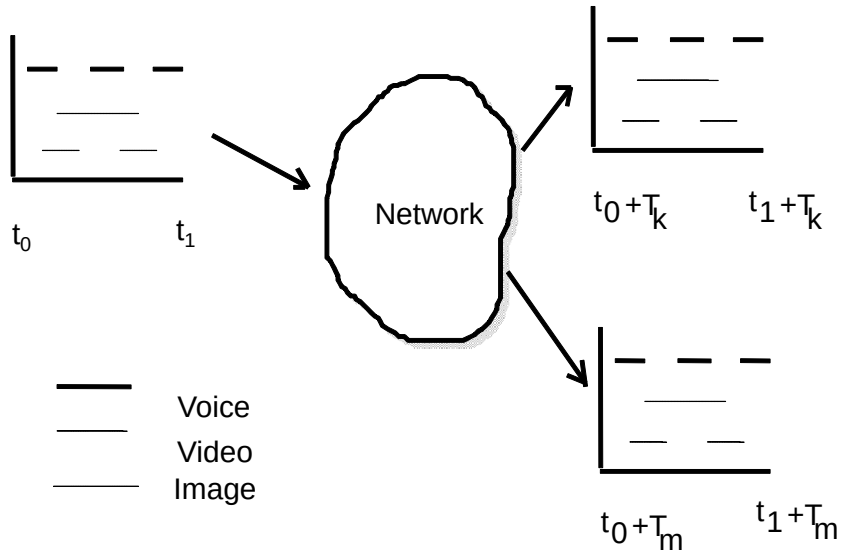
There are several sets of activities that are definable in a multimedia environment. These are as follows: Compound Multimedia Object Transfer, Sub-Session, Management, Dialog Control

The algorithms to perform the activity management functions are developable consistent with the OSI standards. There are no significant special development necessary.

8.5 SYNCHRONIZATION MANAGEMENT

Synchronization is a session service that ensures that the overall temporal, spatial and logical structure of multimedia objects are retained. In this case we have a source generating a set of Voice (VO), video (VI), and Image (IM) data objects that are part of a session. These objects are simple objects that combined together form a compound multimedia object. The object is part of an overall application process that is communicating with other processes at other locations. These locations are now to receive this compound object as show with the internal timing retained intact and the absolute offset timing as shown for each of the other two users.

Figure: Synchronization



In this example, the synchronization function provided by the session server to the applications processes at the separate locations is to ensure both the relative and absolute timing of the objects. The location of the functionality can be centralized or distributed. Let us first see what the overall timing problem is. Consider a simple SMO synchronization problem. The network than transmits the packets and they arrive either in order or out of order at the second point. The session server must then ensure that there is a mechanism for the proper reordering of the packets at the receiving end of the transmission.

Let us consider what can happen in this simple example.

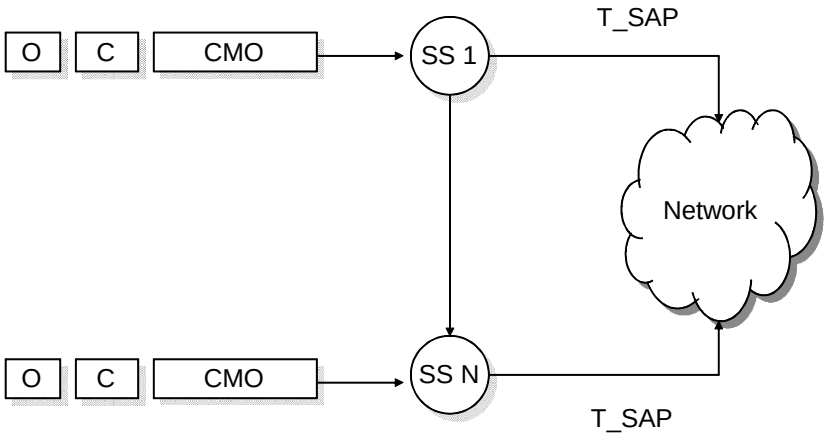
First, if the BMO of the SMO is very lengthy, then as we packetize the message, we must reassemble it in sequence for presentation. Let us assume that the BMO is an image of 100 Mbits. Then let us assume that the packet network has a packet delay that will be low if there is no traffic and grows as traffic increases. Now let us assume that the network provides 500 bit packets transmitting at 50 Mbps.

Second, let us note that there are 200,000 packets necessary to transmit the data. Each packet takes 10 microseconds to transmit. If we assume that there is a load delay of 5 microseconds per packet, then the total transmit time goes from 2 to 3 seconds.

We can also do the same with a compound object. In this case, we take the CMO and note that it is composed of SMOs. The SMOs must then be time interleaved over the transmission path to ensure their relative timing. It is the function of the session service to do this. The application merely passes the CMO and its header information as a request to the session server to ensure the relative timing is maintained.

The architecture for the session synchronization problem is shown. Here we have a CMO entering the network, knowing that the session server at Server 1 must not only do the appropriate interleaving but it must also communicate with the other servers (in this case K and N) to ensure that de-interleaving is accomplished. We show the session servers communicating with the network through the T_SAP and that in turn takes care of the packetizing. However, we also show that the session server, 1 and N, communicate in an out of band fashion, using some inter process communications (IPC) scheme, to ensure that the relative actions are all synchronized amongst each other.

Figure: Synchronization Architecture

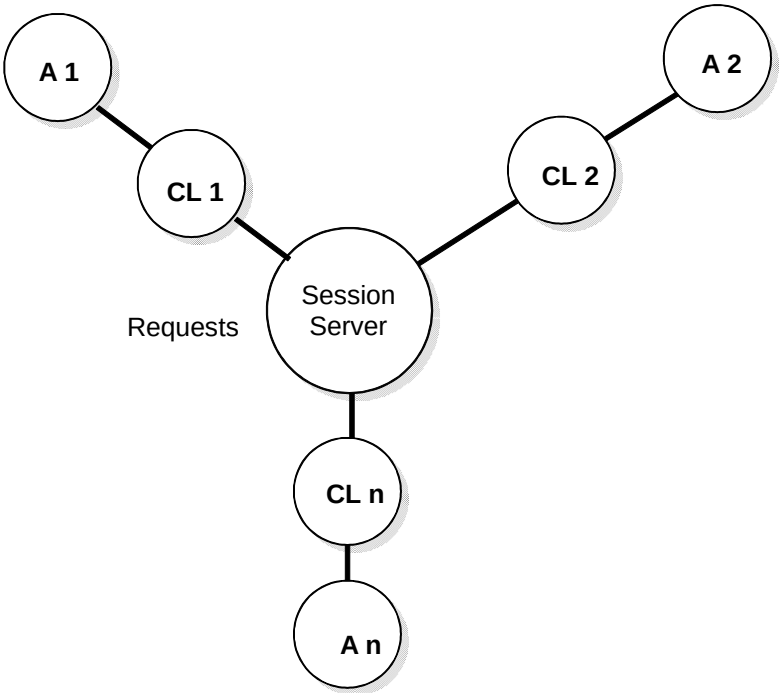


We can now envision how the architecture for this can be accomplished. There are two schemes:

Centralized: It assumes that each application (A) has a local client (CL). The application communicates with the local client (CL) to request the session service. The session server is centrally located and communicates with the application locally by means of a client at each

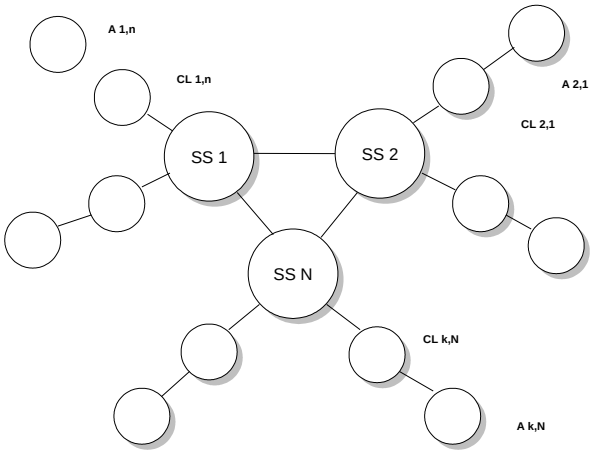
location. This is a fully configured client server architecture and can employ many existing techniques for distributed processing (See Mullender or Couloris et al).

Figure: Centralized Architecture



Distributed: In contrast to the centralized scheme, we can envision a fully distributed session server architecture as shown. In this case we have a set of applications, and cluster several applications per session server. We again user local clients to communicate between the session server and the applications. The clients then provide local clusters of communications and the session servers allow for faster response and better cost efficiency. However, we have introduced a demand for a fully distributed environment for the session managers to work in a distributed operating system environment. As a further extreme, we could eliminate the clients altogether by attaching a session server per applications and allow for the distributed processing on a full scale.

Figure: Distributed Architecture

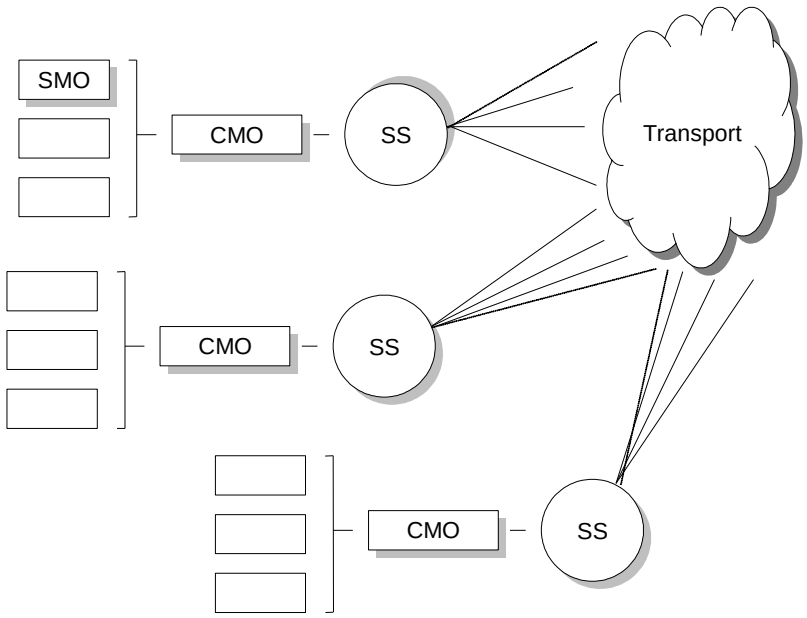


The major functions of the session server in its synch mode are:

1. To bind together simple objects into compound objects as requested by the application.
2. To provide intra object synchronization to ensure that all timing within each object is met.
3. To orchestrate amongst objects to provide inter object timing.
4. To minimize delay, slippage, between simple objects.
5. To minimize delay, latency, between different users.

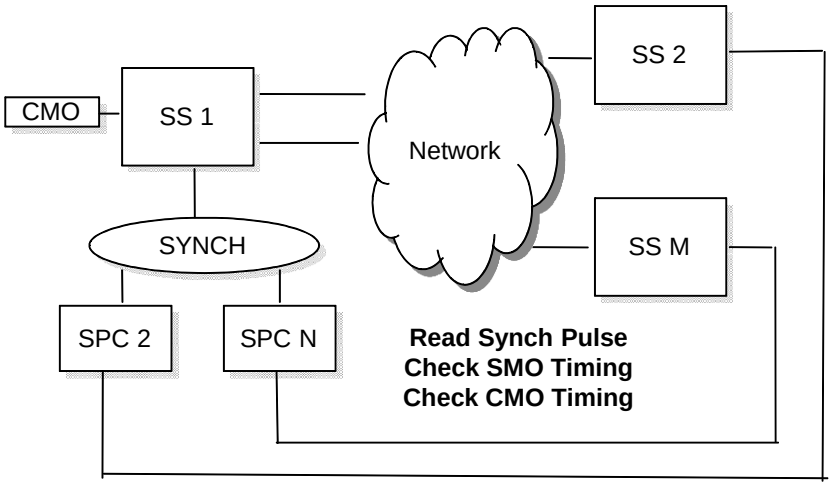
To effect these requirements, we have developed and implemented a scheme that is based on a paradigm of the phased locked loop found in communications (See McGarty and Treves, McGarty). Here we have a distributed session server architecture receiving a CMO from an application. The session server passes the message over several paths to multiple users. On a reverse path, each server passes information on the relative and absolute timing of the CMO as it is received using the session services primitives found in the OSI model. Generally for segmented BMOs this is a simple problem but with streamed BMOs this becomes a real time synchronization problem.

Figure: Synchronization Architecture



Here we show M session servers and at the sending server we do the pacing of the packets to the T_SAP and allow for the interleaving of the SMOs. Based on the commands from the feedback system we provide delay adjustment, through caching and resetting priorities to the T_SAP for quality of service adjustments for the lower layer protocols.

Figure: Detailed Synchronization Implementation



At the receiving session servers, the synch pulses are read by the server, the SMO timing errors are read, knowing the synch header, and an error message is generated. We also do the same for the inter object CMO timing error.

The information is sent back in an out of band fashion to the source session server which in turn controls the synch control pulses for the source session server.

We can provide further detail on the synchronization scheme as follows:

A CMO is generated by the applications program. This may be a totally new CMO or a result of a new SMO addition or deletion.

The Source Session Server (SSS) transmits the header of the CMO to the Receiver Session Servers (RSSs). They then respond with an acknowledgment and in turn set up their internal timing and sequencing tables for local control. They also use the CMO header to adjust their local clock for network timing references.

The SSS commences to interleave, sequence and pace the SMOs of the CMO down to the T_SAP for transport across the network. At this point, the Transport protocol must have certain requirements of either increasing bandwidth (e.g. local data rate requests and also controlling

sequence order. This interaction between the SSS and the T_SAP will define what additional capabilities we will need at the Session layer.

At indicated instances, the SSS inserts local synch pulses in the interleaved CMO. The synch pulses are to be used as local timing reference point for global coordination.

The RSSs read the local synch pulses and relates them to both the SMO and the CMO and obtain offsets from the global system clock that has been updated in the RSS. It then send back the offset of the synch pulses on a periodic basis. The offset is a vector that is given by:

$$E(k,j) = [e(k,j,1),.....,e(k,j,n),e(k,j,M)]$$

where E(k,j) is the offset vector of RSS j at time instant k. The internal values of the vector are the offsets of each of the SMO elements and the last entry is the offset of the CMO.
The SSS uses the set of E(k,j) for j=1,..,N RSSs to calculate an overall error signal to control the SSS. There are two major control features. If the average error is low then the SSS can reduce the insertion of synch pulses and the lower the processing load. If the errors are large, then more synch pulses are inserted to obtain finer loop control. The second element is control over the lower layers. We use the magnitude of the delay offsets to send messages to the T_SAP to change the quality of service parameters for the system.

We have developed several performance models for these protocols and the architecture that has been developed to implement them.

8.6 EVENT MANAGEMENT

Event Management deals with the overall end to end management of the session. It is more typically viewed as a higher level network management tool for multimedia communications. In the current sate this service is merely a reporting mechanism. Although ISO has expanded the network management functionality of the seven layers, most of the functionality is still that of event reporting. In this section we discuss how that can be expended for the multimedia environment.

Event management at the session layer provides for the in band signaling of the performance of the various elements along the route in the session path as well as reporting on the status of the session server and the session clients. We note that each entity in the session path, which is limited to all involved clients and all involved servers provide in band information on the status of the session. In particular the in band elements report on the following:

Queue size at each client and server. The queue size can be determined on an element by element basis.

Element transit and waiting time. For each element involved in a session, the time it takes to transit the entire block as well as the time that the block has been in transit can be provided.

Session synchronization errors can be reported in this data slot. These errors can be compared to lower level errors and thus can be used as part of the overall network management schema.

The structure of the event management system has been effectively demonstrated. It is represented as a header imbedded in the transit of every data block. We can generate specific event management blocks that are also event driven and not data transit driven. These are generated by direct transmission of such blocks as overhead devoid of data content.

8.7 CONCLUSIONS

What we have shown in this paper is that the session layer functions are key to supporting the overall needs of a multimedia communications environment. We have also developed algorithmic approaches for dialog and synchronization services and have shown that these services depend upon the lower layers for support. Specifically, we have shown that if the underlying communications network is jittery in the packet transport provided, the resulting delays associated with the synchronization process can be significant.

Architecturally, we have raised several issues as to how best to provide the session service, specifically where to place and how to communicate with a session server. The session services require considerable entity to entity communications and this may require a distributed environment of session servers all functioning in a fully distributed mode. In the network applications developed to date (See McGarty and Treves), the session server has been centralized and has allowed for communications in a distributed fashion on a UNIX environment using sockets (Berkeley 4.3). However, in future implementations, the session server will be architects in a more distributed fashion.

9 DISTRIBUTED DATABASES

In developing the multimedia communications network, there were many elements that were left defined but not addressed. Specifically, we discussed many of the higher layer services that can be provided in sessioning, presentation and application layers. In a multiuser environment, these services must be implemented to support a fully distributed capability.

Whenever a network is working across several departmental, organizational and geographical boundaries, it becomes inefficient to have a fully centralized system design. In those cases, it is more appropriate to have a fully distributed design.

This chapter discusses the issues and alternatives of distributed elements. In particular we focus on distributed data bases, distributed operating systems and distributed processor configurations. We develop these in the overall context of the architectural requirements of such distributed architectures. In addition, we focus on understanding the performance factors and their tradeoff factors.

9.1 ENVIRONMENT FACTORS

The distributed environment includes the combination of the distributed users and the ability to include many of the systems resources into the ongoing session. The session will have the capability to access the processors and the databases and should

be able to do so in as real time fashion as possible. The evolution of distributed environments has developed over the past twenty years. The earliest system was with the development of the MULTICS operating environment that was developed at MIT for the use in a time sharing computer system. It further evolved with the development of IBM's SNA architecture in the early 1970's. Yet in all of these systems there was still a hierarchical system that controlled all of the systems resources.

A true distributed environment requires that all of the resources and the users can have an arbitrary but definable relationship to all other users in the system. For example, the session concept allows for the development of a complex environment wherein any user in the session can play the lead of follower roles and any resources available in a session can be shared in any fashion with any user or sets of users.

The environment entails all of the resources being at several locations and that there is now single point of management and control. The distributed environment is distributed in five dimensions that we shall discuss in this chapter:

Processors: This is a physical distribution of the systems resources. Typically we envision the location of computing resources to be a one location and the users share that resource in some fashion. Current computer architectures allow for local distribution of assets such as is found in the VAX cluster

concept. Research efforts have allowed for the fuller distribution of computing resources through the use of simple distributed operating systems. However, the fully distributed processor environment is possible with the use of the techniques that are discussed in this section. Thus no single processor is in total control and the processors share resources in an almost parallel fashion. We can envision this as an adaptive parallel processing environment, wherein the connection between the parallel processors is logical and not just physical.

Processes: Distributed processes are already a common ' capability in many systems. The distributed process environment allows for the management of the process execution over several processors at the same time. This may be of use in both real time distributed systems as well as in fault tolerant transaction designs.

Database: The distributed database environment has evolved over the past several years and there is now a body of technology that can support such operations. The distributed database is a key element in a distributed system. It becomes the backbone of operating in an environment where there is data generated by many users and this data has time sensitive nature.

Operating System and Management: The operating system environment is the overall mechanism for controlling the multiple system resources. With the use of distributed processors and the

need for coordinating the set of all system resources, there is a need to do so in a fully distributed environment. The overall management of these resources is done through the operating system.

Communications: The essence of the communications environment is the session. As we have discussed, the session is the access mode for enabling multiple users to share resources, data, and process applications. The distributed communications environment must be built upon the overall structure of the distributed processors and operating system.

We have depicted two possible architectures. The first is the heterogeneous environment, one that is more common in systems operation. The heterogeneous environment shows the need to

share such different resource as memory, CPU, data files, image storage devices and I/O devices. The specific application envisioned is that of the printing and publishing industry.

The second environment is the homogeneous environment. This is the truly distributed processor environment wherein we can envision multiple processor locations of similar devices all sharing in the overall processor loads. An example is the internetting of multiple work stations so as to allow the utilization of all the resources simultaneously.

Recall that a process is a program in execution. A process is typically associated with a single processor and the functioning a process across multiple processors is generally not expected. However, in a fully distributed environment, there is a clear need for such functioning. Consider the following example.

In a medical application are, there is the need for radiological consults between the primary physician, the neurosurgeon, the neuroradiologist and the oncologist for a patient with a primary brain tumor. The consult is conducted in the context of a session and uses the resources of several processors as well as several database. There is an application program that can take the multiple radiologic data, MRI, CAT and nuclear scans, and combine them into a three dimensional image of the patient. The neurosurgeon needs to manipulate this for the selection of laser surgery procedure and the oncologist is interested for the impacts of blood brain barrier effects for post-operative chemotherapy.

The process that is in operation is resident on each of the specialist processors and is shared amounts them.

In this particular case we can envision the application being in the area of financial transactions. The system shown is used for the checking of credit card validation and assuring that the seller can obtain payment from the buyer. The system must track the activity of all the purchasers and seller in the national network. In particular ,care must be taken to track the performance of any single card and to see if it has been used an excessive number of times.

In present day credit card systems, the data base is centralized and allocate checking is done in a hierarchical manner. The data base is accessed via a communications network and the overall cost per transaction can be quite high due to the communications overhead. The distributed database design allows for minimum communications and a distributed set of datafiles with full interconnectivity between them. This system requires that a single card users has their activity tracked as they move geographically through the system. One way to do this is to pass the file off to the local database and then to act as if the system were a fully segmented design. The second alternative is to keep parts of records at a distributed set of locations and to have a set of composite numbers, such a total exposure per card tallied at all or a few locations.

We shall develop this concept in detail in latter sections. However, the overall requirement for a distributed operating system is to allow for the overall management of the systems resources by managing processes, managing data, memory and controlling events. Also the operating system must allow foe control of the overall I/O resources of the system.

The communications issues in a distributed environment revolve around the need to provide a seamless connection between users in such a way as to make the presence of the communications network appear to result in minimal system impact.

Communications in a distributed environment has evolved over the years. As we have shown in , there are standard protocols to handle this communications issue, hate we wish to show in this chapter is that the interaction of database, operating systems , processors and processes may require a modification of such standard communication architectures. In particular the time required of all the protocol overhead may be prohibitive in the operation of a fully distributed system.

We can combine these separate elements and show how they interrelate. The specific example that we have used is that of a Hospital information management system for the care of patient records and one that connects multiple teaching hospitals into a single network.

9.1.1 Network Requirements

The network requirements focus on the needs for the network to, provide more than just a data path. The requirements are driven by the high data rates and bandwidth available to the end user and the ability to configure the channel in a more fluid fashion.

9.1.2 Session Requirements

The session was the key element of the multimedia multiuser communication -paradigm. Sessions have been discussed in detail in their many element. In this section we shall develop the overall requirements on the session concept when it is impeded within a fully distributed system.

Hierarchical schema for part of the COMPANY database have level zero. The level of a nonroot node is one more than the level of its parent node. A descendent D of a node N is a node connected to N via one or more arcs such that the level of D is greater than the level of N. A node N anil all its descendent nodes form a subtree of node N. An occurrence tree can now be defined as the subtree of a record whose type is of the root record type.

The root of an occurrence tree is a single record occurrence of the root record type. There can be a varying number of occurrences of each nonroot record type, and each such occurrence must have a parent record in the occurrence tree; that is, each such occurrence must participate in a PCR occurrence. Notice that each nonroot node, together with all its descendent ra)des, form a subtree, which, taken alone, satisfies the structure that the integrity constraints specified on the relational database schema are not violate)!. In this section we discuss the types of constraints that may be violated by each update operation and the types of actions that may be taken in case an update causes a violation. We use the database shown and discuss only key constraints, integrity constraints, and the referential integrity constraints for each type of update we give some example operations and discuss any constraints that each operation may violate.

For the COMPANY schema if an interface between PASCAL and the networKpBMS were available, it could create the PASCAL program variables. A single record of each record type can be copied from or written into the database usinji the corresponding program variable of the UWA. To read a record from the database, we useJthe GET command to copy a record into the corresponding prograrrm variable. Then we can refer to the field values to print or to use for calculations. To write a record into the database, we first assign its field values to the fields of the program variable and then use the STORE command to actually store the record in the database.

Currency Indicators

In the network DML, retrievals and updates are handled by moving or navigating through the database records, and hence keeping a trace of the search is critical. Currency indicators are a means of keeping track of the most recently accessed records and set occurrences by the DBMS. They play the role ofsposition holders so that we may process new records starting from the ones most recently accessed until we retrieve all the records

9.2 DATABASES

Databases are key elements for the storage of the multimedia data elements. The database issue has received considerable analysis and development for the standard computer files that exist today. These are typically local data bases that are accessed either locally or remotely. The key issue of multimedia communications is the need for both distributed database as well as multimedia databases. The issue of a distributed database is expanded now that we have the capability to transmit data at Giga bit rates. We can now reduce the access time per file for transport to a factor less than that for disk access. This opens new issues for the design of such databases. The multimedia data base represents a new and innovative method of storing multimedia storage. It has two dimensions. The first is the fact that the data may be stored on different media themselves and that may require a nonhomogeneous access and storage protocol. The second is that a multimedia element may no longer be bounded as simply as we have known such computer files as bank records of income tax records. Thus the multimedia database issue is

one of inhomogeneity and unboundedness. Both of these characteristics are typically lacking in standard data bases.

9.2.1 Database Structures

Databases are typically structured collections of data elements that are to be processed on an as needed basis by the end user. Quite simply, we may think of a database as a collection of elements that may be capable of representing the totality of information needed for a specific application. For example, consider the data base that may be needed for the distribution of Christmas cards. Such a data base may need the name, address (street, city, town, and zip code), the relationship (whose side of the family and the specific relationship, such as aunt), the status of sent or received from in the past year, and the gifts that may have been sent or received. What we see in this simple example is that each element of the data base is definable and also is definable in terms of a certain well defined alphanumeric quantity.

In a multimedia database however, we do not have as simple a set of structures. Consider a similar example but for a multimedia case. In this example we are trying to track the progress of a patients recovery from a head trauma sustained in an accident. The characterization is that of patient John Doe. The data associated with this patient includes, X-rays, MRI scans, CAT scans, oral records of the radiologist and neurologist, written records from the primary attending physician, and the cardiologists electrocardiograms. What we immediately see is that the records are not so readily characterized as a finite set of alphanumeric data. The data are video, voice, text, graphics and

image data. The data in addition are of significant and varied length, and the relationship between the different media are suggested but not precisely structured.

9.2.1.1 **Data Base Elements**

A standard structure for database is the entity-relationship (ER) model. An entity is any real object that one desires to discuss and store data on. For example, the patient may be en entity. Associated with any entity is a set of attributes. In the patient example, the attributes may be the name of the patient, the address, age , diseases, and attending physician. We can see that any attribute may be further extended as its own entity with its own attributes. Thus in the present case, we have the disease and it may be characterized by organ, name, level of involvement.

Attribute are generally of bonded form, that is the patients name, has a first, middle and last part, and each part is composed of a set of ASCII characters that do not exceed 50 in length. Thus the name can be at most 150 characters or 1200 bits.

Consider now a multimedia data base in the medical application area. The entity is the patients radiology report. The patient had a set of X-rays and an MRI scan. The "radiology report" entity has the following attributes:

X-rays, including 5 scanned images

MRI images, three sets of 12

Written report of radiologist

Written report of neurologist

Written report of pathologist

Oral report of radiological consult

Oral report of attending physician

In the case of a standard data base, the attributes have a finite set of values that they can take. These finite set of values are called the value sets or domains of the attributes. Names can be up to 50 characters for example. However, the domain of an attribute such as the oral report from a radiologist may be less well-structured and in fact may be unbounded.

We can envision that there can be multiple entities in a large data base. Again consider the hospital data bases that we have been developing. Consider several entities: Patient, Physician, Insurer, Procedures, Prescriptions

The patient entity may contain the attributes: Patient Name, Address, Age, Insurer, Physician , Phone

The physician entity may contain the following attributes: Physician, Name, Address, Phone, Specialty, Backup Physician, Specialty

We can now relate these two entities through a RELATIONSHIP called Patient_Of. Thus we have:

PATIENT Patient_Of PHYSICIAN

This shows that there are two entities that are related through the relationship. This types of relationship is called a binary relationship because it relates one entity type to one other entity

type. There are relationships, that is action statements that relate one entity to another, that ca relate multiple entity types. Those that relate three entity types are called ternary etc.

Consider now a set of three multimedia entities that are prepared for a patient. They are:

RADIOLOGY REPORT

PATHOLOGY REPORT

CARDIOLOGY REPORT

Each of these reports contain attributes the are text, image and voice. We can also have standard entities that can be used in a normal data base. These entities are:

PATIENT

PHYSICIAN

INSURER Consider now two relationships:

Requests_Report

Prepares_Report

We can now have the following statements:

PATIENT Requests_Report RADIOLOGY REPORT

PHYSICIAN Requests_Report RADIOLOGY REPORT

INSURER Requests_Report RADIOLOGY REPORT

Thus we can see that data base are merely ways of storing recorded facts, called attributes, in well-defined bundles called entities, and associating them with each other in relationship. The structure of the language that uses the data base is called the data base language and has its own syntax. However, the syntax is governed by the underling structure of the database.

9.2.1.2 **Relational Databases**

A relational database is a means of characterizing the data and structuring it so that it may be accessed and altered in a fashion that allows for maximum flexibility. In a relational database, we develop a flexible set of relations and build these from the bottom up to have as flexible and accessible format as possible.

The basic building blocks of a relational database are as follows:

Domain of Atomic Values: These are irreducible entries that represent final data values, such as the age of a patient, their first name, or their phone number.

Attribute: This is the name or role played by a particular domain. Thus phone number is the attribute name played by the domain xxx-xxx-xxxx.

Relation Schema: This is a set of attributes, {Al,...,An} that has a name R, which may be the term patient, and has a degree equal to the number of attributes, or irreducible elements.

Relation: A relation is a set of tuples, where a single tuple is the set of attributes of the relation. For example, the relation schema, address may contain, the attributes, number, street, city, state, and zip. This is a 5-tuple. The relation ADDRESS now is the set of all such tuples that are relevant to this application. It could be zero, one or even one million addresses.

Tuple: As we have just noted this is a specific relation entry. In the above case for address, we have:

t = {24, Wood St, Greenwich, CT, 08534} or for the general case:

t = {number, street, city, state, zip} and the relation:

ADDRESS = {tl,t2,,tn}

Key: The key of a relation is an attribute that can be used to uniquely identify the relation. In the case of address, we may use the zip as that attribute if and on if we know that people are in one town at a time. There may be many keys in relationships, and there can always be a primary key.

We can now see how relations can be referred to one another. Let us define three relations through the three schema:

R1(A1,A2,A3,A4,A5)

and rl = {tl,t2,t3,t4,t5,t6,t7,t8,t9}

where ti are all 5-tuples o R2(A6,A7,A1)

and r2={ul,u2,u3,u4}

where ui are all 3-tuples and Al is the same as in rl.

R3(A8,A5)

and r3={vl,v2,v3,v4,v5,... ,vll4}

where vi are 2-tuples and A5 is the same as rl.

We can now point from r3 to rl and from r2 to rl. This pointing is through the common attributes in both cases.

There is an algebra that can be developed using the relations that entails the use and definition of operations on the relational schema. These operations are of the type: Union, Difference, Select, Divide, and many others (see Elmasri and Navathe)

9.2.1.3 Hierarchical Databases

The hierarchical data structure is an older scheme of database and follows the need for a structured environment. In the relational case we define general relations, many of which can change and evolve in time and are related to one another through keys. In a hierarchical scheme, we have a much more structured system that is related a priori.

In the hierarchical scheme we have several elements that make up the data base paradigm. These elements are:

Record: This is the collection of data that are the basis of the database. The record could be the patients account and would be composed of name, address, age, etc.

Parent-Child Relationship: This is the relationship between one record and other records. For example, we can say that Department (eg Cardiology, Radiology, Urology etc) cane be the highest record in the hospital. This may relate down to Patient and Physician. In turn, Patient relates down to Tests' and Insurer records.

Field Type: This is the detailed contents of the Record type. In the patient Example we may have:

Patient = {Name, Address, Age, Disorder, SS Number}

The patient Tests record may include the following Field Types:

Test = {Type, Date, Physician, Result, Next Test Date}

Hierarchical schemes have a fast access time for most applications but have many major drawbacks. It is generally difficult to change data elements or records and we have to be cautious in handling non 1:N relationships in the parent child area. In addition, as we have noted in this example, the patients are grouped by department first and to get statistics on patients, we must sum across departments, not across patients.

The patients, and tests etc are segmented by the hierarchical scheme.

9.2.1.4 **Network Databases**

The network database structure owes its early existence to Bachman and it represents a way to structure the relationships between data records in a logical fashion on the overall data structure. In the hierarchical scheme we had developed the notion of an overlaying hierarchical relationship between all elements of the database. In the relational scheme, we had much greater freedom in defining data elements and then after the fact overlaying relationships in a much freer manner. In the network scheme, we find that there is a middle ground in which we do not have the hierarchical structure but that there is an overriding relational structure in the database that is called the networking of the data elements.

The network data base is composed of the following elemental concepts:

Data Values: These are the actual numerical or letter values that are entered into the data base. Example would be specific names or numbers.

Data Item: This is the name ascribed to the data value. For example, we can use NAME, or ADDRESS.

Record: This is a collection of data values in an organized form. It may represent all of the data values stored on a particular patient.

Record Types: This is the name used to identify a collection of records, such as PATIENT or PHYSICIAN.

We can then find a way to relate these together through the following concepts:

Set Type: A relation between two record types. One of the record types is called the owner and the other is the member. There is only one owner and there may be several members. A set type may be the one Treated_By and it relates PATIENTS to PHYSICIANS.

9.2.2 Data Base Access

The structure of databases is only one of the key factors in understanding the usage and implementation of a database. The second element is the access to the data records themselves. In this area there are three major issues of concern. The first is the issue of the physical access to the data. Specifically what medium is the data stored on and how is the data physically stored on that medium. As with the file structure, we consider the more common forms of file structures in this section and defer the issues of multimedia databases to latter sections.

The primary form of data storage is the magnetic disk for linger term storage and random access memory for shorter-term storage. The issue of medium is significantly impacted upon by the cost of random access memory and its ability to be supported is a nonvolatile environment. With the advent of multi-megabyte chips and the dramatic cost reduction in cost per bit, there are many tradeoffs that can be made in the selection of physical storage medium strategy. We shall not discuss those elements in this text but refer their reader to XXX.

The second issue that we shall discuss in this section is the issue of the logical access to databases. Specifically, what languages do we use to readily access the information. The languages typically used here are called query language and typical example are those of SQL and the other languages that have developed syntaxes for access of structured databases. The query language is the user interface issue that we have discussed at length in other chapters. The query language has develop for structured database from a set of typically arcane commands to a set of commands that today resemble a natural language. It is actually possible to ask such a question as "How many people are there in the Drafting Department". A system may take this sentence and parse it and then know enough to be able to recognize that it must add together all the name files for a specific department.

Physical Access

Logical Access: Access Languages (DBL)

Database Access Interfaces

9.2.3 Distributed Databases

Distributed data bases are essential in environments where there would normally be significant communications from one location to a distant one. A typical example that we have used is the

need for distributed database in the financial transaction processing area. Here we have to keep track of many transactions that occur at various locations.

Let us first consider an example of such a distributed system of databases that is slightly different. This example is that for the storage of 800 numbers in the telephone system. In this example we will first deal with a passive database that is read by the user and there is no transaction made in the database.

The 800 number system in the telephone network works on the principle that local users can dial 800-xxx-xxxx and can get a toll free phone call. The normal billing system will recognize a prefix other than 800 as one to bill and will bill the calling party directly. When the 800 number is referred to in a call, the number is looked at in a database table and it is cross referenced to an actual area code number and that number is fed into the network as the dialed number. At that point the end user does not know that he has skipped around the billing computer and has had his dialed number converted. The conversion table could be placed at multiple locations or it could be centralized. In this case the database is static. The difference between having this data base centralized or distributed is the issue of call set up time, that is the time to place the call.

Let us examine the 800 database tradeoff. Consider the two cases of centralized versus distributed. In the centralized case we have a large data base that is at one location. Let us assume that there are 1 million calls per busy hour and that each 1 thousand calls requires a 1 MIP processor. The call set up time for this system is given by:

T SU = T TR + T I/O + T pro + T Resp + T TR

Here T TR is the transport time, T I/O is the I/O time and T resp is the CPU response time. For a heavily loaded machine we have decreased I/O and response time.

In the 800 distribute case we assume that we can place an 800 processor at each of the 1,000 central office locations and this yields 1,000 MIP machines and reduced I/O per machine and reduced processing time. In addition we have to split the database in such a way that we do not have to multiply it 1,000 times. We do this by recognizing that only x% of the 800 numbers are used by y% of the people in each of the locations. This is often called the 80/20 rule. Thus we can save only a small percent of the conversions for translation and leave the rest for a centralized system or route them to other systems.

Let us now consider a second example that is both read and written into in a distributed fashion. This data base is one used for the buying and selling of commodities such as gold, oil, silver and other such items. This differs from the first case because now there is a dynamic bidding process that must

carefully take into account the timing of the transactions and assure the end user that bid and asked prices for commodities are met.

Let us examine the process in some detail. There are sellers of commodities that present an asking price. For example Seller 1 could say that they are willing to sell gold at $450 per ounce. There are several buyers who are interested in the gold. However they are not willing to pay the asking price. One bids at $448 per ounce as the highest. Then the seller reduces the price to $449. One of the buyers sees this as an opportunity and bids • $446.50. The other buyer seeing that this is a good deal bid $449,125 and the seller closes the deal.

Now there any be several markets for this gold. The markets may be in New York, London, Tokyo, and Hong Kong. One approach is to have the data base used for recording and display the bids to be centrally located in a city such as Chicago. Or alternatively, each city may have its own data base and the system communicates amounts itself in a fully distributed fashion. Again the tradeoff is time to transact and time to respond. In this type of market, the quicker I can respond the better closure on the bid/asked price one can obtain.

A third example of a distributed database is that which can be used for the establishment of a multimedia communications system, where sessions are established and users are sending images and other multimedia objects from one point to another.

In this case, the system is used to set up calls and to maintain the multimedia sessions. The session set up algorithm identifies all of the participants in the session and knows their locations on the network. When a member of the session transmits a packet to a local node, the local node accesses the overall database which knows the status of all participants in the session. It then generates a set of additional packets, customized as is necessary and transmits them to the desired nodes. It must assure that there is certain synchronization in the packet transfer and that all of the packets are synchronously transmitted both interpacket and interuser on a single packet.

In these three example there are certain characteristics that arise that a distributed database must reflect. These characteristics are:

Reduced access time due to the combination of reduced communications time and reduced I/O time due to more efficient loading per processor.

Real time updating of data records and assurance of timing factors so that bid and asked numbers are not reversed in time.

Overall synchronicity within and between packet transmissions to assure end to end synchronicity for display and processing purposes.

9.2.4 Distributed Database Issues

We can now begin to discuss some of the structural elements of a distributed database. We shall use the relational database system for the representation of database in this environment. Approaches using the other schema are also possible and are left to be developed in the problems. The discussion in the section is based upon the work by Ceri and Pelagatti.

Recall that a relation, R, is a table that store data. The data is stored in columns called attributes, Ai. The number of entries in the table of the relation is called the tuples in the data entry.

In the normal non-distributed environment, we have a collection of relations that generate the overall database. The global schema is this collection of data as if we had no distributed environment. Now we ask ourselves, how do we split this up into several pieces so that we may locate the data in multiple locations. We do so in two steps:

Fragment: In this step we take the global schema and generate separate disjunct schemas call fragments and named Rl*i*••••*f*Rri•

Allocate: We assume that the database is to be distributed to k locations, Ll,...,Lk. We then must map the fragments onto these locations. Many fragments may be mapped onto several locations.

We can deal now with the first issue of the allocation of the databases to many locations and focus first on the issue of reading of such databases. Consider the database composed of many sub elements. We desire to distribute this database over many locations. Let us take the example of the 800 number database that is used in the telephone business. This database is a conversion data base that converts the 800 number into a real dialable number of the type xxx-yyy-zzzz. Rather than having this database at one central location and having all customers dial that location, we can break this database apart and distribute it at many locations. The breaking apart function is the fragmentation task, and the placing of these parts at multiple locations is the allocation task.

In fragmenting a database, we are considering the ey need to break it apart into smaller segments that can be accessible in some particular fashion. In the 800 number case we may segment by other areas of the country, by the specific application, by the type of business or other reason. When we fragment, we than may want to assign or allocate the fragments to multiple locations.

Thus if we define D as the total database, we can define D as:

D = {D1,D2,....,Dn}

where Dk is a specific fragment. We can then define Li as the ith location. Let Si be the assignment vector for the ith location;

Si = {Sil,Si2,,Sin}

where :

1 if Dk is at Li

Sik =

0 otherwise

Then we have the database at Li being the following: D(Li) = DT Si

where T is the transpose operator, and we represent it as:

D(Li) = {SilDl,, SinDn}

The optimal segmentation and allocation problem can now be stated as follows. How do we choose the best fragmenting and allocation scheme to either maximize or minimize some specific performance criteria. One approach is to define the cost of a scheme and then to find the fragmentation and allocation that minimizes the cost. Consider the following cost equation. Let CTOT be the total cost of the scheme. The cost equals:

CTOT = CDB + CPRO + CCOM + CLAB

where:

CDB is the cost of the actual storage databases at each location and duplicated as needed at each location.

CPRO is the cost of the processing power needed at each location to manage the database entry and control.

CCOM is the communications cost of the system as needed for the users to access each of the fragmented databases.

CLAB is the labor cost per access, summed over all access. It is a measure of the response time of the system in providing a unit access to a datafield.

Cost for Memory: As we fragment the database, we then allocate it to many locations. The cost for memory is dependent on how much we place at each location. At one extreme, we can fragment and place only parts at one and only one location. At the other extreme we can fragment and place all parts at all locations. The latter will be an upperbound on memory costs and the former a lower bound. Depending on the cost of memory the cost may be significant. The memory cost can be made lower by a slower memory storage system, however, that will increase the labor cost element.

Cost for Processing: As we have done for memory, the processing cost is dependent upon two factors, the number of accesses per location and the amount of memory controlled per location. If we have few locations then we are dominated by access, if we have many locations but significant duplication of the database, then we have significant access processing.

Cost for Communications: This factor depends on how many locations there are and how many locations have the data base elements. If there are many locations and many data base elements then the communications costs are low. If there is only one location then we would expect high communications cost.

Cost of Labor: The labor costs are based on how many people are required for a volume of work. The people are determined by the number of units and the holding time per person. If the overall processing time is dominated by the response time of the human then this is the lower bound. If, however, the response time is dominated by the system delays, then this is a factor. For example, if we have log communications delays, long access times etc, we have large labor costs.

Conceptually, we can conceive of an optimization criteria that says the if this cost is to be minimized, we could then determine the optimum fragmentation and allocation to reach the minimum cost element. Generally this is not readily determined and suboptimum approaches are performed. We demonstrate several of these in the problems.

We can now consider the simple second issue that relates to the real time updating of databases. This is the issue of writing or reading and writing databases. In the previous analysis we have developed an approach to the selection of how to distribute the database based upon the reading of the files. In this area we shall focus on the issues of actively changing the contents of the database. There are two major issues that are key to understanding the active change of a distributed database. The first issue is that of recovery from faults and the second is the concurrency of a database use. Recovery relates to the issue of how to deal with system errors and how to restart the database after such a failure. The issue of concurrency relates to the need

to establish the concept of a transaction, which we shall do shortly, and show that transactions are key elements in read write actions.

The issue of recovery of the database after a failure must be considered. The distributed database is composed of three levels of elements. At the highest level is the data agents that control the flow of the actions of the database. The agents have a define root agent that provides an overall centralized control element. At the second level is the distributed transaction monitors DTM that provide for the communications amounts the different locations. The DTM provide control of the elements in the third level, the Local Transaction managers, LTM. The DTM elements communicate amounts each other and also sent the action signals to the LTM.

The recovery issue relates to a transaction that is occurring that may be halted in mid-stream. If this is the case, it is necessary to remember that it was not completed and possible, but not necessarily, where it left off. It is the function of the root agent, in conjunction with the TMs to keep logs of the status of the transactions. The complete transaction has to be defined and the progress of the transaction is followed through a log process, If a fault occurs, then the log is used as a means to determine if the system is recoverable.

Concurrency is the issue that relates to ensuring that one event does not occur before another that id dependent upon the first. We would not want to post money to an account before we checked the validity of the depositing check, and if posted allow the customer to withdraw the amount not yet checked.

The issue of concurrency control is also critical. As we had done with recovery, concurrency depends on the three layer architecture. Now however, we need to have the agents talk with one another and with the dTMws. We perform the concurrency control through a locking mechanism, combined with a structured and time tagged transaction sequence. The transaction is a complex set of reads and writes, that must be defined a priori. By tagging the transaction sequence as a whole, and providing locking mechanism throughout the database schema, we can assure that improper reads and writes do not occur. The cost may occur in the efficiency of the database, access time.

We can now define a transaction as an atomic unit of access to the database. This transaction is complete by completing all of the elements of the access or otherwise it is not complete. We can define Tj as a transaction. We further define a transaction as a combination of reads and writes to a data base. Let W denote the write function or operation and R the read operation. Let Dk represent the data item on which the operation is to be made. We call Dk a data item and say that Dk is part of relation Rj. A data item is that set of a relation that is materially affected by an operation of a read or write.

We can now define a transaction as an ordered tuple of W and R operations as follows:

Tj = { Rj (Dk) ,Wj (Dm) ,,Rj(Dp)}

That is Tj is a collection of ordered reads and writes, at locations and on data elements k,m,...,p.

Transactions have four basic characteristics that must be observed:

Atomicity: A transaction is the totality of all of its parts and cannot be broken down into sub elements.

Durability: Any system that performs a transaction must assure the operators that the results of the transaction endure.

Serializability: When there are several transactions executing simultaneously, them the result must be the same as if the transactions occurred in a hourly serial manner. That is if there are three transactions TI, T2, and T3, then if we have:

TI = [W1,W1,W1,R1,R1]

T2 = [W2,W2,R2,R2]

T3 = [W3,R3,R3]

We can now define a SCHEDULE, S, as a sequence of W/R commands and one serial schedule is the one;

S = {T1,T2,T3}

Another schedule is a sequence of mixed versions of Wi and Rk. Any admissible sequence form a schedule must be serializable, or equivalent to the first S.

Isolatable: This means that incomplete transactions must not be effected or cannot effect other transactions.

We have seen that we ensure the necessary elements of a transaction are preserved by ensuring that the use of locks, time tags and the other elements that we discussed are used in a distributed database environment.

In the system at any one time there are many transactions that

are occurring. That is there are Ti, Tj, , Tq. This clearly

shows the problem of the distributed database. The issues are as follows:

Concurrency: This relates to changing data elements at different places at different times, when there is a need to assure that a sequence of changes be made in certain order. For example Ti may have a sequence or ordered reads and writes on a set of data elements. Tk may have another set of reads and writes. However there may be an operation in Tj that may proceed an operation in Tk that will alter the way Tk is carried out. A simple example will give the case.

Let us assume that we wish to reserve a seat on an airflight. However another person at another city wants to reserve the seat on the same flight. Let us assume that you get to the counter in New York one minute before the contending person does in San Francisco. Hopefully, the system is first come first serve and you get that last seat.

Let us first examine the transaction that goes into reserving the seat.

A read is made into the database to determine if a seat is available.

A write is made into the database to reserve the seat on the plane.

A read is made into the database to check the credit card, o A write is made into the database to charge the card, o A write is made to a data base to issue the ticket.

Thus the transaction is:

TI = {R1(D1),W1(D1),R2(D2),W2(D2),W1(D3)}

Here we have Wl being a read at location 1 which is New York, and W2 being Phoenix. The second transaction is:

T2 = {R5(D1),W5(D1),R2(D2),W2(D2),W5(D3)}

Now these transactions are occurring almost simultaneously but the reservation system is in New York (1) and San Francisco (5). Let us see what could happen in this case.

1. New York checks and shows available.

2. San Francisco checks and shows available.

3. San Francisco books seat.
4. New York books seat.
5. New York checks credit
6. San Francisco checks credit
7. New York charges credit
8. San Francisco charges credit
9. New York gets default on ticket print.
10. San Francisco gets ticket
11. You get very mad !!!!!

Even though the data bases update instantly, we still have not kept the integrity of the overall transaction. Admittedly this was a bad way to create a transaction, it however, is typical of what can happen in large scale transaction systems.

The solution of the concurrency problem is time stamping the total transaction and ensuring that all data node controller have the ability to update each other simultaneously.

9.2.5 Multimedia Databases

The multimedia database consist of several media that contain a complex interconnection of video, voice data, text, image and other forms of information. As we discussed in the context of the reactions database, there can be significant connections between many of the different files. For example, consider a voice file that contains certain information on a particular patient record. This file may be associated with a video file of the patients echo cardiogram. It is necessary to ensure that the files are related to one another, that they can be tracked down and also synchronized in time and space. These types of requirements are not typical for alphanumeric records and files.

9.2.6 Physical Access

Let us consider the case of accessing the record of patients in a hospital. There are four types of record that may be of concern for the development of a certain type of procedure. These records are:

Patient records stored in the Relational database disks.

Radiological images stored in a CD juke box storage system and managed by a PACS type data manager.

Text records stored in an image based CD system similar to the image files and front ended by a separate database manager.

A voice storage system that is the storage vehicle for voice messages that annotate the image files.

A video angiography of the patient's hear stored on a CD rom disk.

We can consider the multimedia data object as a meta-object which is the concatenation of the set of four separate data objects. Let us define these objects as follows:

OR is the data object associated with the patients files

01 is the data object associated with the patients radiological records. Note that this may be just x-rays, all x-rays, x-rays and MRI data, or a concatenation of many of these elements.

OT is the collection of text records that relate to this file. It may include the write ups on the current x-rays prepared by the attending radiologist or any other set of elements.

OV is the voice samples of the consulting radiologist or other related physician.

We can now defining the compound data object associate with the patient as OP and definite it as:

OP = {OR,OI,OT,OV>

where {} stands for an appropriate concatenation. We shall

discuss the structure of this concatenation in detail latter in this section.

We can now see how we can relate these compound data objects the physical system design by examining the overall system architecture.

There are separate storage elements and assume for this case that they are generally located in close proximity. There are two extreme strategies for developing physical access to these records. The first assumes that each is a standalone system and the end user must establish all connections to related records.

9.2.6.1 Logical Access

The logical access issues for multimedia database are significantly more complex than for the alphanumeric databases. The SQL type languages for physical database are based upon the simpler structure of the records and files and make the use of simple syntax possible. Queries in a multimedia context are much more complicated. In an alphanumeric system it is possible to ask the question "How many patients in 1989 have had acute endocarditis?". The statement is parsed into the following elements:

How many: Count the number and the answer is an integer.

Patients: Count on the patient records

1989: Condition on the year

endocarditis: Condition on the disease.

Consider now the type of question we may query the data base with in a multimedia format. "Which patients have shown a cholecystitis diagnosis that the physician has expressed concern in the extent of the diagnosis and also the scans are not fully determinate?".

Let us examine this query and compare its elements to the query for an alphanumeric data base query.

Distributed Multimedia Databases

MMDB Design Factors and Issues

9.3 CONCLUSIONS

This chapter developed the concepts of databases and further developed concepts for both distributed database and multimedia databases. The database element is key to the storage and retrieval of multimedia files. Knowing the standard data elements in a typical record file is

necessary but not sufficient for the development of a multimedia database. In we had developed an understanding of the types of files that we could develop in a multimedia environment and studied to types of storage media available. That allowed us to develop the times statistics necessary for the operations of a data file retrieval. In this chapter, we have extended that to the database element, understanding the way the information can be stored and what languages can be used in its retrieval.

10 DISTRIBUTED OPERATING SYSTEMS

The control of the flow of information, requests and the overall management of the session process is generally beyond the control of the typical functions of the communications systems that we had developed in. This management can be viewed as a major function of the operating system of a standard operating system of a stand-alone computer and more importantly the function of a distributed operating system of a fully distributed environment.

There are two extreme approaches that can be taken to viewing the overall operational control and management of a large distributed system. The first approach states that the control can be fully distributed in the layer architectures of the separate elements of the network, and thus follow all of the ISO layered commands. Frequently this works well within a purely record based and generally homogeneous network of multimedia elements. The second approach states that a form of central control, although implemented in a distributed fashion is a means to the same end. This latter approach allows for the integration of multiple system elements, even though they may not be fully ISO compatible. Moreover, it allows for the introduction of new and innovative multimedia interfaces that the ISO level may not anticipate.

In this chapter, we focus on the development of a multimedia distributed operating system that ensures that the sessions are managed properly and that handles all of the other functions of an operating system. These functions include such elements as file and record management, process and session management, device and I/O management and an overall network management function.

As with the other chapters, our focus is to build on existing paradigms and then to expand the concepts to a fully distributed and multimedia environment. Moreover, we shall also develop the sizing and performance analyses that are necessary for the development of an overall system architecture.

10.1 OPERATING SYSTEMS

The operating system is that software that allows all of the different processors to be cholecystitis utilized to their best degree. In particular, the operating system will allow for control of the memory of the system and the files that are to be managed by the processors. It also allow for the overall management of the processes that are running on the system. Finally, the operating system allows for the control of the input and output devices on the system. In a distributed environment, the operating system can be extended to include all of the processes in the system and in turn be the manager of the overall files in the system. As we discussed in the

development of the distributed database environment, the distributed operating environment, builds upon the stand alone system.

In this section, we first develop an understanding of the standalone operating system and then we develop the overall concepts of the distributed operating system. In particular we develop a system called, MEDOS, standing for the Media Distributed Operating System. This system has been developed for use in a fully distributed multimedia environment.

7.1.1 OS Structures

An operating system is simply a collection of software that, at one side controls the hardware and other external assets of the computing environment, and at the other side, provides a ready means for the user of the resources to incorporate them into the overall application. Thus we can view the operating system from two directions, first that towards the resources managed, and second towards the end users need to manage the overall set of resources. There are generally five major functions of an operating system as we understand it today. These are:

1. Process Management
2. Memory Management
3. I/O Management
4. Device management
5. ,

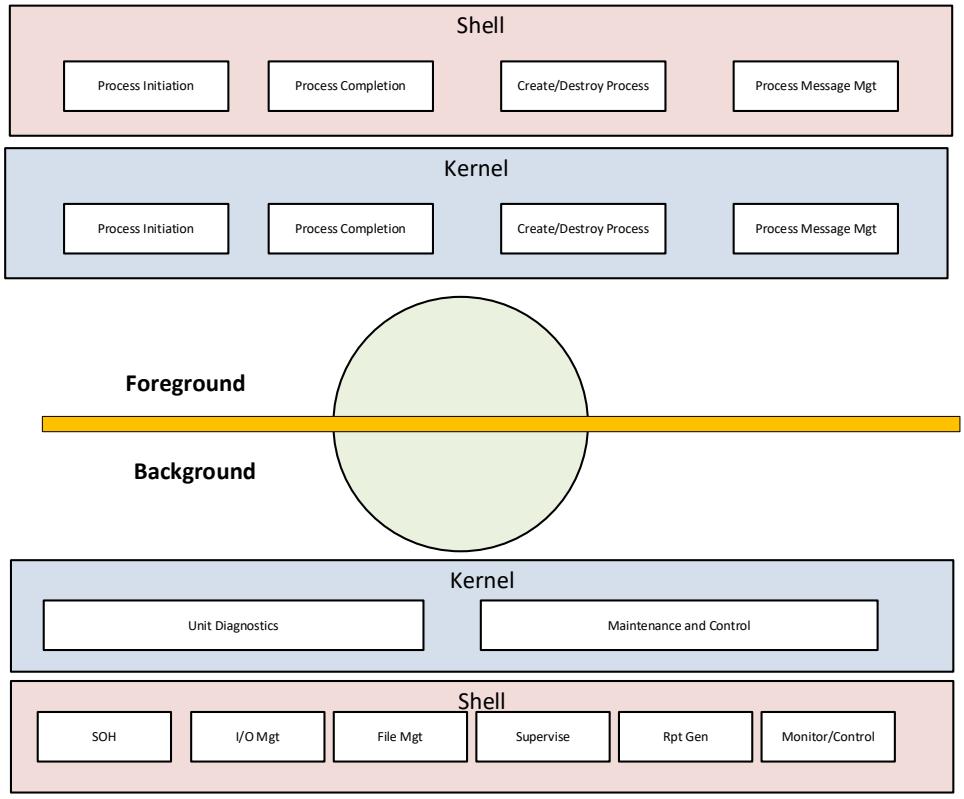
These five major functions relate to typically a single standalone environment that focuses on the need to manage the resources for a single machine. As we have discussed in the development of multimedia communications, the primary element is that of the session. In contrast, in the environment of the single machine, even with multiple users and multiple applications, the primary focus is the process (an program in execution). There are significant differences in the ability to meet the needs of a process as compared to a session. In this section, we shall review the structure of current operating systems, and expand the concept to the area of a fully distributed operating systems focused on multimedia applications.

Operating systems have evolved from the simplest ones that were basically the operator themselves, through batch systems, and upwards to today's systems that have the capability to perform multi-processing and multi-tasking. The development of such common operating system environments such as UNIX and MS-DOS are responses to the end users' needs to manage the

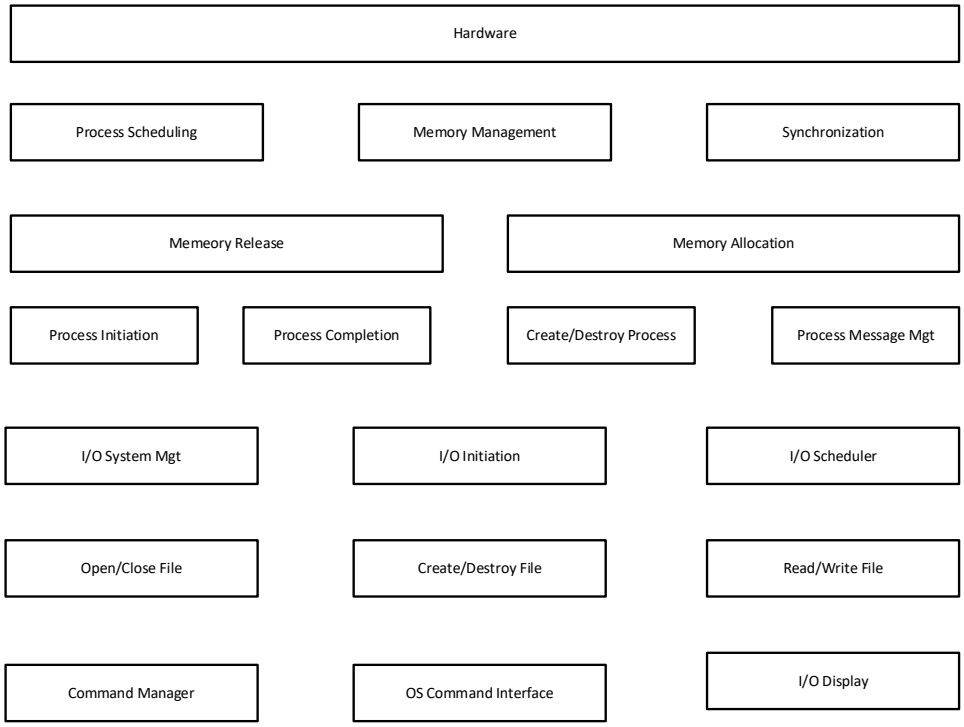
computing resources of the specific location. Expanding the resource management to multiple locations has been achieved through communications channels that we have discussed in the last chapter.

To provide an oversight, we can view operating systems to fall in four general categories. These are a local operating system (LOS), a networked operating system (LOS), a distributed operating system (DOS) and a multi-media distributed operating system (MEDOS). Our objective in this section is to develop the structure of a MEDOS and to do so by showing how it is a natural evolution of the LOS environment. Before continuing, let us first define in some detail each of these separate operating system environments, with emphasis on the requirements of the end user that they satisfy.

Before comparing the different architectures for the operating system environments, let us first consider a canonical environment of an operating system. The Figure below shows the canonical structure for the operating system. It consists of six levels. These levels are:



Another view of such a structure is shown below containing more enhanced elements is as follows.



The specifics as above are defined as follows:

1. Hardware: This is the innermost level and represents the totality of all hardware that the operating system could manage as resources to the end user and the application. Typical of such resources could be the processor or processors involved in the computation, the memory units, printers, displays and entry devices for the end user.

2. Generics: This typically is microcode that is written specifically for each device and may also typically be provided by device vendors for the local access to interface and manage the individual devices. For the most part, the device drivers. And fault monitors are part of the generic elements.

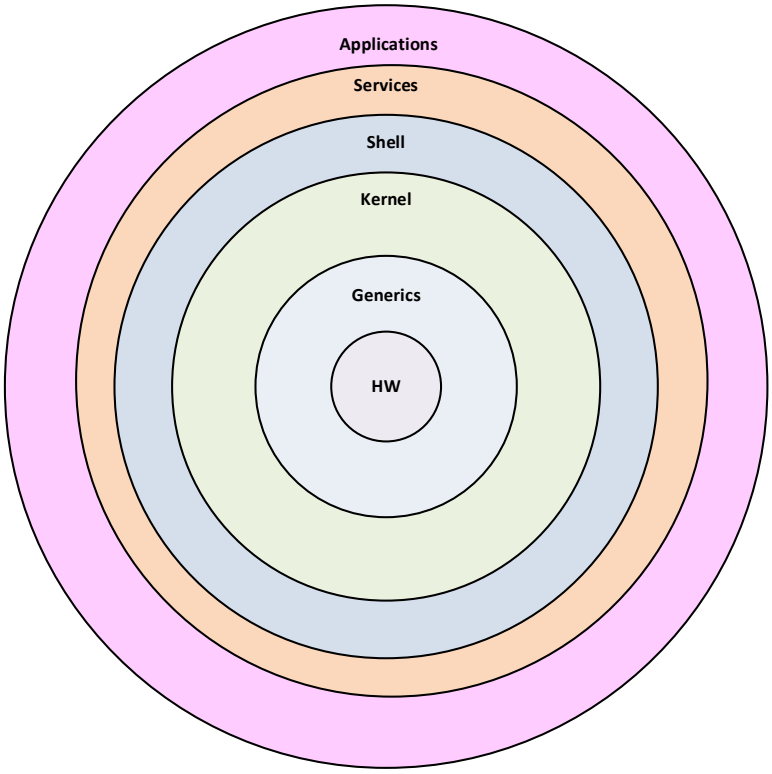
3. Kernel: This is the first part of the operating system. The kernel element, for our purposes, is the layer of operating systems functions that directly interface with the devices and hardware. In this canonical structure, all kernel functions of the operating system are hardware directed, as compared to the shell functions that are applications support directed. Thus memory management is a kernel layer function whereas file management is a shell layer function.

4. Shell: The shell is the second layer of the operating system and is that layer that looks out towards the user. It provides for the user interface and also allows for direct access by the

higher layers to all the service provided by the operating system. Access to the shell layer is through a set of commands that we call primitives. Primitives have a defined syntax and can be used by the end user to not only access all the operating system facilities but to manage them in some detail.

5. Services: The services layer is provides for special support functionality to the end user application. Typical service layer functionality is that of a windowing environment that evokes the functionality of the operating system. X Windows is a typical service layer functionality that supports the capabilities in UNIX base system. Other services could include the ability to perform sessioning, the ability to mail in the network, the ability to perform communications service between various nodes ijn the network and other such services. The service layer can also be thought of as the OSI layer interface at the applications layer.

The layered approach to an OS is shown below:



Typical services that can be supplied by the system are as follows:

1. Communications: This service allows for the connecting of a local environment to any other local environment. This service usually utilized the ISO based layered approach to establish the communications link.

2. File: This service attends to the need of the end user in accessing and maintaining their overall files. In a distributed environment the file service extends to sharing in a common and seamless fashion all of the important file elements.

3. Directory: This service allows access to the location and names of all files on the system. If we consider a file in the extended sense of UNIX then we can see that the directory provides the overall access mechanism for any readable and/or operable entity or agent on the system.

4. Mail: This service allows for the sending of data elements from one user to another on the system.

5. Session: This is and advanced service that allows for the establishment and maintenance of a session based service. We shall be discussing this at length.

6. Gateway: This service allows for the interconnection of multiple users to other networks. It extends the communications service by providing a level of protocol translation and conversion.

7. Presentation: This service consists of means and methods to present information on the end users console. An example, as we have discussed is the X Windows service that is used for bit mapped presentation.

8. Print: This service allows for the output in a hardcopy format of any of the data or processed files in the system.

These services are only examples of the many that can be generated at the services layer. The services are accessed via Service primitives that allow for specification of the details of the service elements. In turn the services are generated by using a collection of the Shell Primitives.

Applications: This is the layer that include the set of programs that make up the end user application. The applications may be one or many, and they may be associated with one or many end users.

We can now use this canonical model and explain each of the four types of operating systems in terms of the canonical form. The reader should note that we have extended the models for the operating system environment beyond what may generally appear in the literature. The classic text by Madnick and Donovan provide an excellent review of operating systems in the mid-

1970's. The more recent texts by Tannenbaum and by Deitel provide excellent updates to the environment.

We consider now the different types of OS such operating systems:

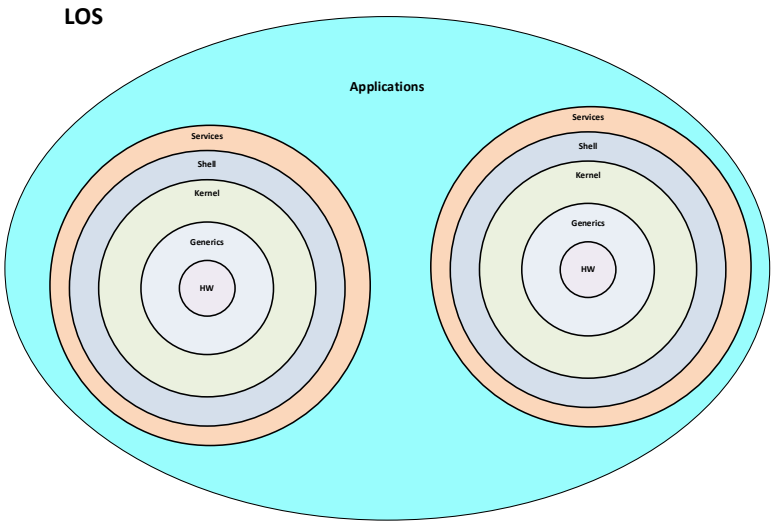
10.1.1 Local Operating System (LOS)

A Local Operating System, LOS, is the result of the evolution of the end users need to readily access many of the system resources and the systems designers needs to allow for maximum flexibility in explaining both the applications run on the machine and the number of users utilizing the machine resources. From the end user's perspective, there is a need to access files that are stored in the system and to manipulate those files for the purpose of a specific application. We allow the access to the LOS through a user interface that is a command interpreter. The commands that allow the user to manipulate files are similar to commands that permits all of the functions in the LOS. We shall call these commands the operating system primitives, or primitives for simplicity.

The LOS focuses on the local environment and its resources. It allows for the input and output of the end users commands and data and also manages all other system resource I/O. In a similar fashion to file management, it manages the system memory, and if necessary can interact with the management of processes to permit certain memory to be brought to faster processing locations for better optimization of applications run time. Typical of this latter approach is a software controlled cache function.

To the user of an MS-DOS system, the commands such as dir (for directory), path, and file, are typical of the file management commands readily available.

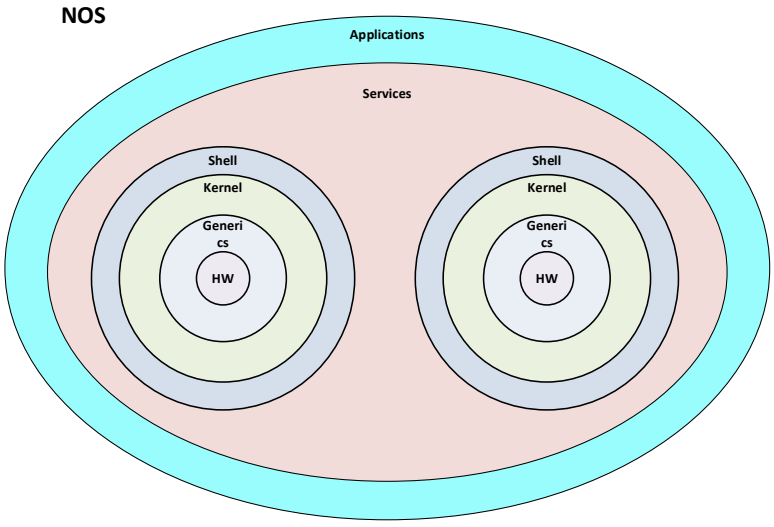
We depict the structure of an LOS operating an application that may share one or more machines. The application is bound between the two machines by the services layer communication service. This means that the machines are totally separate and are self-contained. Only in an abstract sense are the applications shared. In reality they are fully resident in each machine.



10.1.2 Network Operating System (NOS)

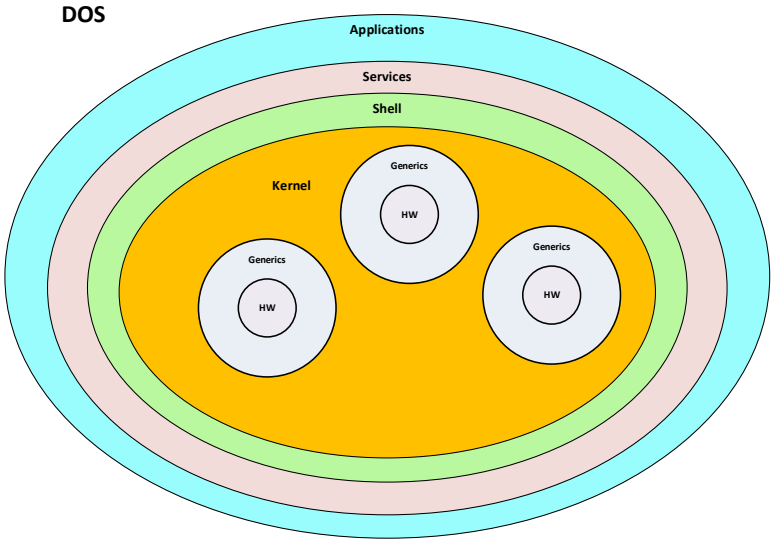
A Network Operating System is the next step up in complexity from the local OS. In the local OS environment, all of the resources are local and are owned and operated by the single OS. In the network OS case, there are multiple locations, but all of the resources are locally owned and controlled. Thus if a user desires to use the resources at another location, that user must log onto those resources through the network.

We can see below in the NOS case, the OS elements are still discrete even though the users may be sharing the applications and now Rtes services layer. The NOS structure allows access to all elements through a shared network communications protocol and also a set of other shares service elements.



10.1.3 Distributed Operating System (DOS)

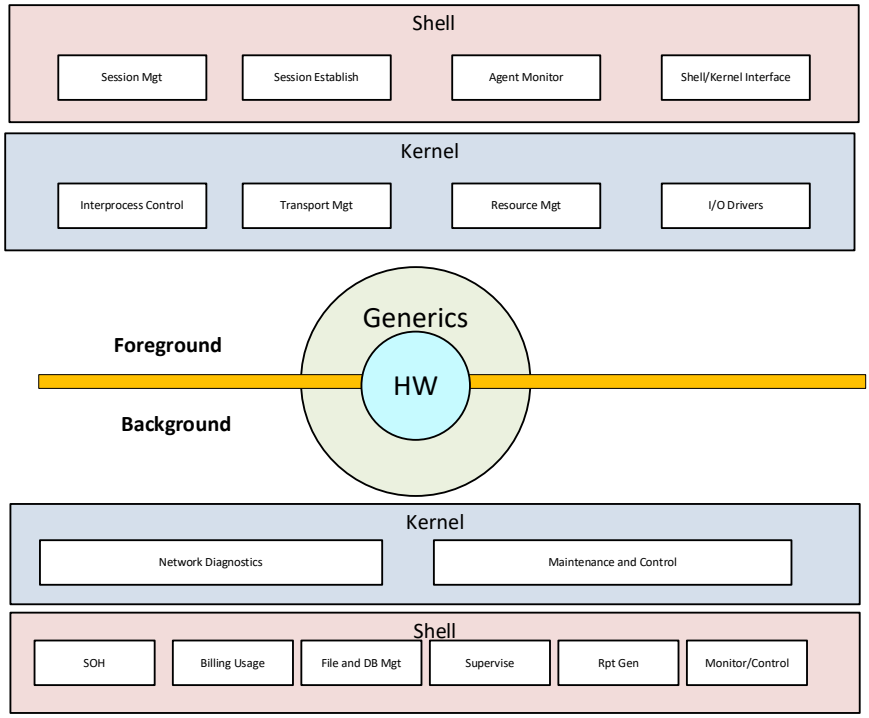
In a Distributed Operating System, only the hardware and its associated generics are still associated on a logical basis with the individual sites. In this case the operating system functions of both the shell and kernel are shared amounts all the users. In a fully implemented DOS environment, no user will be able to identify the specific resource being utilized at any time. The sharing of all of the resources and their access is through the common distributed shall/kernel combination. Below shows how such a DOS environment can be envisioned. One could conceive of the day when a fully distributed system is developed with sharing of even the generics and hardware. Such is not the issue in the present discussion.



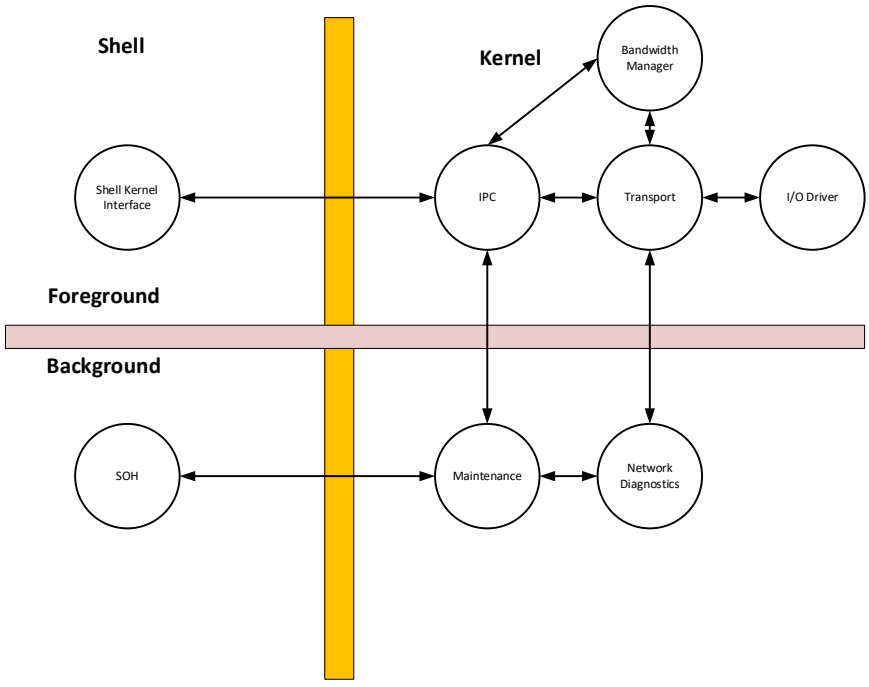
10.1.4 Media Distributed Operating System (MEDOS):

We have discussed the concept of a distribute operating system and the focus has been on the system that has standard data elements and storage. In a multimedia environment, the hardware the lies in between the system is more complex and hardware elements are multimedia in character. Thus the system must combine all of the elements that we have discussed previously.

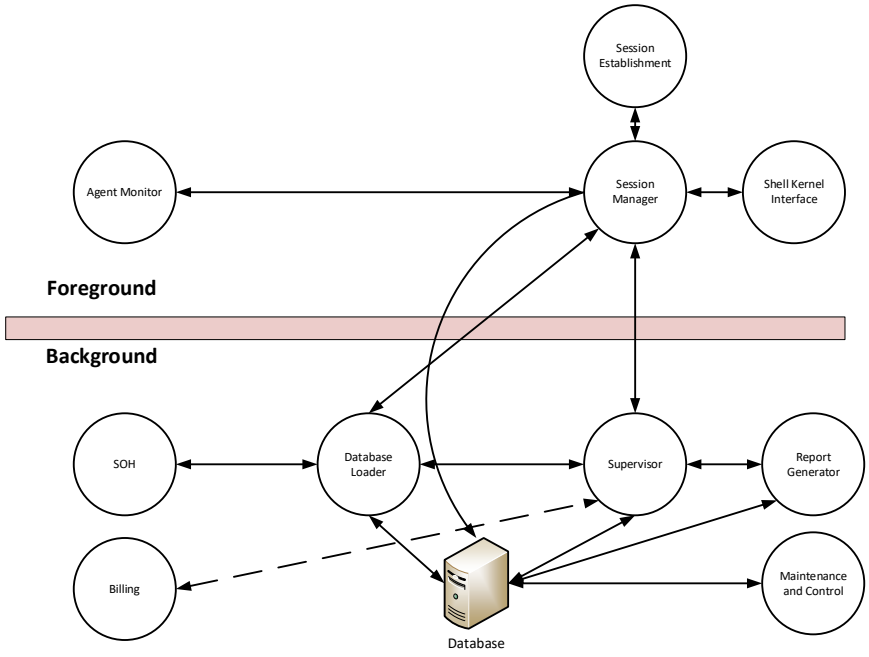
The overall MEDOS architecture is shown below:



The following is the Kernel Architecture:



The following is the Shell Architecture:



The MEDOS system was a multimedia directed distributed system that worked across a multiple set of platforms and data sets using a high speed, 45 Mbps, channel when it was deployed in 1989[3].

10.2 OS DETAILS

Operating Systems have been developed to provide serves to a wide variety of machines and in a large set of environments. Typically, there are four major functions of an operating system (see Madnick and Donovan):

1. Memory Management
2. File management
3. Process management
4. I/O management

We have also included a fifth function (see Tannenbaum) :

[3] Note that the author had architected and named the system in late 1986 and in three years of debelopment it was deployed for test and evaluation.

5. Device management

Very simply, these five functions perform the following tasks.

Memory Management: This function manages the reel and non-real time memory of the system. It allocates space in memory and moves date to and from different parts of the systems memory. As we look towards multimedia applications, memory management becomes even more complex than that for standard digital data. The nature of the storage media changes and the use of real and virtual storage becomes more complex.

File management: A file is defined as a collection of data that has a specified name (see Deitel). In contrast to memory management, which manages the physical resources of the storage devices, file management manages the allocation of data to memory that is used in the context of specific applications. As with memory management, file management for multimedia applications is more complete because of the need to create files of often unbounded and disparate media.

Process management: A process is a program in execution. The management of processes entails the allocation of system resources to ensure the effective execution of all processes. The implementation of interprocess control (IPC) for the coordination and swapping of processes in execution is also a key ingredient of this function.

I/O Management: The management of the various input and output devices is a key element of this function. This includes the management of printers, display devices and other elements.

Device management: Device management typically refers to internal elements of the computation process. It may often be bundled into the management of I/O elements.

These five areas are the classical elements of the operating system structure. However, below, we have displayed the canonical structure of an operating system. This canonical structure shows two major divisions. The first division is that between the shell and the kernel. As we have stated before, the shell is that part of the operating system that relates outwardly to the end user, and specifically to the services layer. The kernel looks inward to the devices. The second division is in terms of foreground and background process. The foreground represents real time or near real time processes that operate in the OS. The background processes are no real-time processes.

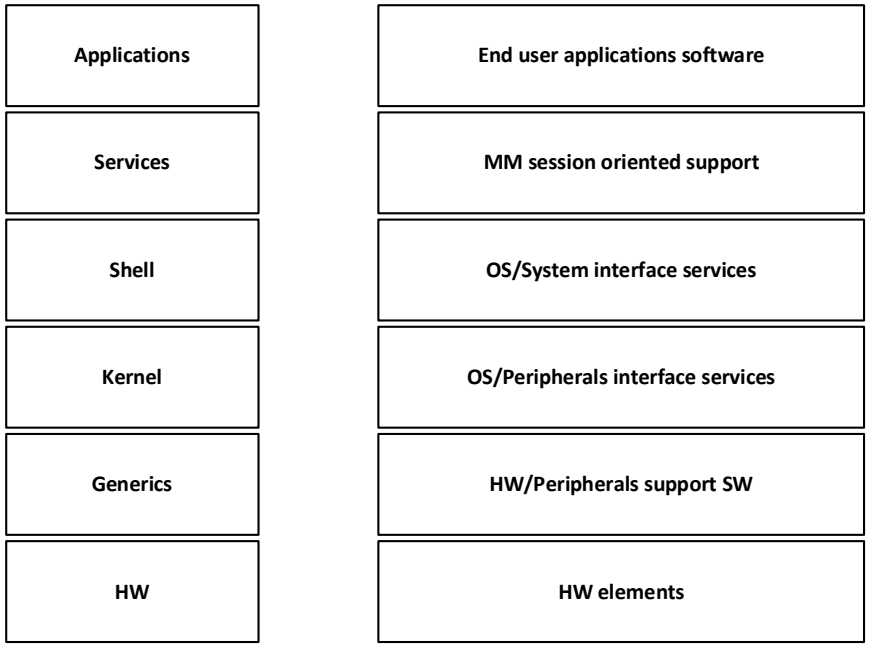
The division of an operating system into these two divisions is an important one from an operational perspective. The shell/kernel division allows the designer to focus on the directional emphasis on interfacing and functionality. The foreground/background partition focuses on the real time nature of the associated processes.

In the canonical structure, we have shown sixteen separate processes that are used to define the canonical operating system. Theses represent those that are normally a part of the standard five elements plus a set that we have added for the purpose of expanded beyond the local OS to a fully distributed environment. We shall now discuss the contents of each of the processes.

1. Shell
    a. Foreground
        i. Process management
        ii. Process establishment
        iii. Agent Monitor
        iv. Shell/Kernel Interface
    b. Background
        i. State of Health
        ii. I/O management
        iii. File Management
        iv. Supervisor
        v. Report Generation
        vi. Maintenance and Control

2. Kernel
    c. Foreground
        i. Media (Memory) Management o
        ii. Communications Interface o
        iii. Resource Management o
        iv. I/O Drivers o
    d. Background
        i. Unit Diagnostics
        ii. Maintenance and Control

Figure Operating System Structure: Canonical Example



We can now use the discussion of the canonical operating system and show how existing operating systems map into the canonical model. In the Figure, we have presented three of the more common operating systems in use, VMS, UNIX and MS-DOS.

10.2.1 Specifics

Now that we have defined the overall canonical model for an operating system, and we have provided a comparison to specific existing operating systems, we can now demonstrate how each of the key processes is developed with module specification diagrams.

The operating system structure that we developed in the last section was a generic structure that can apply to all possible types of operating systems. However, as we have seen in the examples of existing operating systems that are local in nature, many of the elements of the canonical model are not needed in the system. What distinguishes a distributed operating system from a local system is the need to ensure that all of the physical resources are treated as one set of local resources and not just as a collection of separate local sets. Thus, for a distributed operating system, there is a need for significantly increased communications management amounts the physical resources.

Let us begin by considering a simple example in the area of memory management. In a local operating system, we can partition the memory and manage it from a single local point. Let us assume that the system has both RAM memory and Disk memory and that we page memory from disk into RAM for processing purposes. Let us assume that main memory (eg RAM) has capacity of C_H and that there are N processes that are being run simultaneously.

Let Pi represent the ith process. Let Pi take memory space MPi and let the data requested for Pi take space MMi. Thus using an appropriate swapping algorithm, we can use a performance measure of the system and the algorithm that measure the efficiency of the usage of main memory and the efficiency of the time efficiency of use. We call this the machines time-bandwidth product. It is defined as:

$BT = (MEM_{eff} * MIPS)*(Time_{eff}* T_{cycle})$

where:

MEMeff = the percent of main memory fillable on average by the swapping algorithm.

MIPS = the processing rate of the machine in instructions per second.

Time eff = the percent of a cycle time that is used for processing as compared to all other factors such as data transfer etc.

T cycle = the average instruction cycle time

Now the T eff can be calculated as follows. If T load is the total time to load a main memory cell, and T proc is the total processing time for the command of that process in that cell then we define T eff as:

$T_{eff} = T_{proc}$

Now generally we have:

T load « T proc

because the memory access time is significantly short. The access time of the memory is generally short and is dependent mainly upon the I/O characteristics of the memory device. In a distributed environment however, the access time is increased due to the communications associated with the memory call plus the nomad memory I/O access time. Specifically we have:

T load = T I/O + T comm

Now when we design the paging algorithm, or consider the need for virtual memory, even in one machine, we must include the communications effects into the analysis. Frequently:

T comm » T I/O

Let us calculate some of the numbers in these equations to give the student a feeling for the dimensions of the problem. The speed of light is about one foot per nanosecond. If we have a distributed system with separation of 500 miles on average, then this is 2.5 10^{6} feet and equals a delay of 2.5 msec. If the I/O to the memory unit works at microsecond speed, then we have the inequality indicated with three orders of magnitude difference. In addition if we have a 2 MIP machine, and we page in 100 instructions per second, the processing time is 50 microseconds. Clearly T proc is greater than T I/O but much less than T comm. It will be this factor that we shall see dominating the DOS performance.

Table 7.x depicts the mapping of the detailed operating systems functions as described against the canonical model and shows how they relate to the higher level functions commonly used in operating system descriptions. We shall now use the more common structure to discuss the key features that must be considered in a distributed operating system as compared to a local operating system. Each of these issue relate to the problems that we have discussed in the above example. That is, it is now possible to use other memory devices and other processors that are connected more loosely and with greater delays than generally anticipated.

Let us now consider all of the major elements of the DOS and detail their key performance factors and compare these against a standard LOS.

10.2.2 Process Management

The process management function is directed towards the overall management of the processes that are operational in the system. It looks outward towards the processes that are part of the overall applications program and ensures that they are scheduled and interlaced in an effective manner. Process management takes the form of managing the multiple users in the system and matches the users and their needs with the overall set of system resources. A process is taken in and out of the main memory and assigned system resources depending upon the needs of the process, the priority of the end user and the availability of the resources.

10.2.3 Local System Performance

In a local operating system environment, the issues of process management relate to the following items.

1. Process Establishment:
2. Interprocess Communications:
3. Process Monitoring:
4. Process Messaging:

5. Process Scheduling:
6. Process Termination:

10.2.4 Distributed Systems Design Factors

In a distributed environment, we include all of the above items plus several that are in addition to these. Moreover, we need to expand the functionality of the separate elements of the process management function to ensure that processes are properly managed across several resources. Process management in a distributed environment may take on several different aspects. In the local environment, we have a set of processes that can be existing on a single processor. They are managed by the operating system of the local processor. In a distributed system, we can envision several scenarios:

1. Single process per Processor/Multiple Processors: In this case there is a single process that is operational at a single processor and it must communicate with other processes at other processors.

2. Multiple Processes per Processor/Multiple processors: In this case we have multiple processes that are operational at each processor but that they all must communicate between themselves.

3. Single process per Multiple processor/MP: In this case we have a single process that is shared amounts several processors. These processors are separated by many local distances and thus have the dynamics associated with separate locations.

4. Multiple processes per Multiple Processors (MP): In this case we have multiple processes, all communicating amongst themselves, distributed in multiple processors.

Process management in a distributed environment is increased in complexity by the sharing of resources that are at disparate locations but the advantage is the ability to share resources and to dynamically reallocate resources amongst many users.

10.2.5 Device Management

Device management is also known as processor management. It is the physical complement of the process management that we have just discussed in the above are a process management. The processor management function basically relates to the assignment of processes to specific systems resources or processors. The process management function relates to the intra-process management functions, namely how does on manage the process itself and in relation to their processes. The processor management function relates to how we manage the processes in

relation to their outside environment. The processor management function relates to the overall need of the system to perform the following types of task:

1. Job scheduling of the processes to the processors.

2. Traffic control of tasks between and amongst the

3. Synchronization of the processors amongst themselves.

Device management for local systems and distributed systems is very similar. The allocation of resources can be managed on a centralized or a distributed fashion. In either of these cases the specific algorithms may be static or dynamic, and in turn may be time driven or event drive.

Thus, the type of processor manager may be characterized by the tuple:

{Control, Timing, Response}

where the control is central or distributed, the timing is static or dynamic, and the response is event or time driven. Thus there are the following types of processor management systems:

{Static, Dynamic, Event} etc.

This yields eight possible configurations of process management. We develop several algorithms for the management under these domains in both the examples in this section and the problems. In a distributed environment, the same set of variables apply, and moreover the same set of algorithmic devices or artifacts are applicable.

10.2.6 I/O Management

The I/O management entails both the I/O to devices and peripherals as well as I/O to the end user. We shall focus on both of this element. The physical device I/O is controlled by a set of device drivers that talk directly to the device and are controlled by the management portion of the I/O element. The end user I/O are interfaced by what we term primitives, which are command sets that use a defined syntax and protocol and allow the user to interact with the operating system directly.

10.2.7 Memory Management

Memory management provides the management of the physical memory of the system. There are typically many forms of memory that are used in a commuting system. These range from the local RAM memory storage, to local disk and local tape memory. Against these physical storage

media is the allocation of the memory that the process recognizes as important and the memory as allocated and managed by the operating system. The process frequently has need to access many elements of the physical memory and it then needs to changes and redirect that memory as it is processed. The operating system must assist the process in its process management and memory management functions in allocating this memory to be effectively used by the system.

Typically we have a system with a fixed amount of memory that is immediately accessible by the process in operation. Let us defined each process by Pi and let the total memory be given by MT.

Let the set P{I,k} represent the process set at time k, and let this equal:

P{I,k} = { Pi : i I and kT< t < (k+l)T }

10.3 DISTRIBUTED PROCESSORS

In any distributed environment there are not only issues of the distributed data base and the distributed operating system, but there is also the issue of the distributed processors and their interconnection. This section addresses some of the issues of distributed processors.

Processor Types

Processor Performance

Processor Interconnection

10.4 DISTRIBUTED PROCESSES

As we recall from our earlier discussions a process is a program in execution. A program calls upon all of the resources of the operating system, the databases and the processors. Thus there is a significant different in a multimedia distributed environment for the functioning of a set of distributed processes.

10.5 CONCLUSIONS

This chapter developed the overall distributed environment and discussed the key elements of that environment. What should be clear after developing these many issues is that there still are many open questions in distributed environments and these questions are driven by the availability of new technology that allows for interconnection at extremely high data rates. In the past, as we have discussed in the presentation of the ISO model, we have found that the data rates were slow and that it was necessary to provide a significant overlay of protocols to assure end to end performance. A distributed environment in that world was limited to slower acting

systems. In the world of unlimited bandwidth as presented by fiber links, a fully distributed architecture can be developed.

11 CONCLUSIONS

In this book, we have developed a view of multimedia communications that combines the structure of the underlying images and the storage and transport media along with a careful understanding of the end user and their interactions with the different forms of media. Unlike the world of computer communications where the computer can be trained to perform certain tasks that are required for proper performance, the human in multimedia application is operating in a mode of total creativity. This creativity results in complex and often unpredictable interactions with the other human users. This interaction is the essence of the multimedia communications theory.

We have tried in this book to present a canonical structure to view the multimedia multiuser environment, focusing on the concept of the session. We have not focused on the psychological inferences and actions of the end user in great detail. This is the artistic side of multimedia communications and expands beyond just knowledge engineering or even epistemology.

11.1 KEY ISSUES

This book has developed a philosophy of multimedia communications that tries to separate the "artistic" element from the engineering element. It may be criticized that such a separation is difficult if not impossible, but our intentions were to gain a better focus on the system ideas by quantifying the basis elements and not serializing them. Thus the approach has been alone the more common engineering lines, understanding that the system definition and design are the key characteristics. We can now review the overall key issues that we have developed in this book and present what has been developed in context.

Multimedia communications, as we have seen, deals with the handling, manipulation, processing and transmission of complex images that are highly interactive in nature. The insensitivity factor focuses on the "conversational" mode of communications that we have discussed.

1. The Overall System Approach
2. Characterization
3. Performance and Sizing
4. Feel and Form versus Function and Formalism
5. New Paradigms of Presentation
6. Displaced Working Environments
7. End User as Design Element
8. Sessions

9. Distributed Environments

The evolution of multimedia communications will be enhanced by the development and deployment of broadband communications as well as the development of smarter and more sophisticated end user terminals. The most difficult problem is the issue of how do we effectively interface with the human end user. Specifically what common paradigms are most effective in what environments.

The trends that impact this area of multimedia communications are quite simple to articulate at this stage but they may change as we enter into a time of rapid technological change. Specifically, if we look at the model of the multimedia environment, and we map prognosticated changes in each of these areas against the separate elements of the multimedia domain, we can carefully determine the impact of the multimedia environment and determine the returns to that environment.

Let us focus on the specific multimedia areas and discuss the trends in each of those areas that will impact the development of an improved and more advanced multimedia communication system.

These trends provide for a brief introduction of some of the options that are available to the user and developer of multimedia communications networks. They present a view of what

is possible but do not present what will actually occur. We have developed some of these issues based upon what is difficult to do in today's environment but new technology will most certainly alter this view.

The discussion in the past section turns on views of future trends that are natural follow on to what we have discussed in the body of the text. In this final section of the book we wish to share some of the future research directions that are open and whose solutions provide the most fruitful ground for future work.

1. Multimedia Database Structures

2. End User Interface Language Design

3. WAN Communications Architectures

4. Distributed Operating System Optimization

5. Overall System Optimization Performance

6. Multi Sense Interfaces

11.2 SUMMARY

This book presented a vision of multimedia communications as more than just an amalgam of separate parts, disembodied from the ultimate user. It presents a holistic view that recognizes the impact of the user on the services and the underlying technology. Multimedia communications is a concept that is in its infancy but that also offers the user a significant increase in performance, functionality and the ability to disembody themselves from the medium of communications.

As was stated earlier in the book, Marshall McLuhan stated that the introduction of a new medium changed not only how we presented information, but ultimately what was considered knowledge. This is both a challenge and a warning. The technological concepts presented in this book provide for a significantly different view of media or presentation technology. It raises the question of how best to interface the end user to a network that responds in near instantaneous speed at almost all data rates. It, in effect, introduces a plethora of new media, thus challenging and warning us simultaneously.

12 REFERENCES

1. Abbatiello and Sarch, A Typology of Local Area Networks - Confused by the array of LANs, some users fall back on switching, in Telecom m unications and Data Communications Factbook Data C om m unications and C C M I/McGraw -H ill, 1987

2. Aguietal, A Painting Algorithm of Monochromatic Images Using Color-Sequences, translated from Denshi Tsushin Gakkai Ronbushi, vol 67-D, No.3, March 1984, Scripta Publishing Company, 1984

3. Alemany, R. Kasturi , A Computer Vision System for Interpretation of Paper-Based Maps in Applications of Digital Image Processing X - Proceedings, edited by A.G. Tescher SPIE - The International Society for Optical Engineering, 1987

4. Arbib, M.A., A. R. Hanson, Vision, Brain and Cooperative Computation, MIT Press (Cambridge, MA), 1990.

5. Arms, C., Campus Networking Strategies, DIGITAL Press (Maynard, MA), 1988.

6. Atkins, P. Physical Chemistry, Freeman (New York) 1990.

7. Avery, J.E., VAN, The Regulated Regime, Telecom and Posts Div, Dept of Trade and Industry, UK, Conf on Communications, IEE, pp 78-81, May 1986.

8. Ayer, A.J., Philosophy in the Twentieth Century, Vantage (New York), 1984.

9. Ayer, A.J., Russell, The Wodburn Press (London), 1974.

10. Ayer, A.J., Wittgenstein, Random House (New York), 1985.

11. Barlow, W. The Broadband Revolution, Info Tech and Pub Policy, Vl 8, No 1, pp 6-8, 1989.

12. Barlow, W. The Broadband Revolution, Info Tech and Pub Policy, Vl 8, No 1, pp 6-8, 1989.

13. Barrett, W., The Illusion of Technique, Anchor Press (New York), 1978.

14. Bartel, B., S. Matsuda, Seeing Red, Science, Vol 299, 17 Jan 2003, pp 352-353.

15. Bell Labs, Transmission Systems for Communications, fifth edition Bell Telephone Laboratories Incorporated, 1982

16. Bell, D., The Coming of the Industrial Society, Basic Books (New York), 1973.

17. Bell, T., Technical Challenges to a Decentralized Phone System, IEEE Spectrum, pp 32-37, Sept.,

18. Bell, T., Technical Challenges to a Decentralized Phone System, IEEE Spectrum, pp 32-37, Sept., 1990.

19. Bender et al, CRT Typeface Design and Evaluation, MIT Industrial Liaison Report S0288-069,1988

20. Bender, Anti-Aliasing for Broadcast Compatible Television, SID 84 Digest

21. Bender. Anti-Aliasing for Broadcast Compatible Television, SID 84 Digest, 1984CRT Typeface Design and Evaluation, W. Bender, R.A. Crespo, PJ. Kennedy, R. Oakley MIT Industrial Liaison Report S0288-069,1988

22. Berns, R. S., Principles of Color Technology, Wiley (New York) 2000.

23. Bernstein, J., Three Degrees Above Zero, Scribners (New York), 1984.

24. Blackwood, M.A., A. Girschick, Theory of Games and Statistical Decisions, Wiley (New York),

25. Blackwood, M.A., A. Girschick, Theory of Games and Statistical Decisions, Wiley (New York), 1954.

26. Born, M., E. Wolf, Principles of Optics, 4th Ed, Pergamon (New York) 1970.

27. Bostwick, W.E., Program Plan for the National Research and Education Network, Dept. Of Energy,

28. Bostwick, W.E., Program Plan for the National Research and Education Network, Dept. Of Energy, May, 1989.

29. Bradley, Alan, Optical Storage for Computers, Wiley (New York) 1989.

30. Brealey, R., S. Myers, Principles of Corporate Finance, McGraw Hill (New York), 1990.

31. Burns, Alan, Andy, Wellings, Real Time Systems and Their Programming languages, Addison Wesley (Reading, MA), 1989.

32. Business Review, pp. 114-120, September-October, 1990.

33. Bux, Token-Ring Local Area N etw orks and Their Performance Proceedings o f the IEEE, vol.77, no. 2, February 1989

34. Byrne et al, Broadband ISDN T echnology and Architecture IEEE Network, January, 1989

35. Carnevale, M.L., "Untangling the Debate over Cable Television", Wall Street Journal, p. B1, March, 19, 1990.

36. Casavant, Analysis of Three Dynam ic Distributed Load-Balancing Strategies with Varying Global Information Requirements 1987 IEEE 7th International Conference on Distributed Computing System s Computer Society Press of the IEEE, 1987

37. Chang, N.S., K.S. Fu, Picture Query Languages for Pictorial Data Base Systems, IEEE Computer, Nov 1981.

38. Chang, S., T. Kunii, Pictorial Data Base Systems, IEEE Computer, Nov 1981, pp 13-19.

39. Chiariglione et al, An Integrated Approach to Video Standard Conversion and Pre-Processing for Videoconference Codec in Image Coding edited by M. Kunt, T.S. Huang Proceedings of the SPIE The International Society for Optical Engineering, volume 594,1985

40. Chomsky, N., Aspects of the Theory of Syntax, MIT Press (Cambridge, MA), 1965.

41. Christman et al, Which Kinds of OS Mechanism s Should be Provided for Database Management? Experiences w ith Distributed Systems edited by G. G oos and J. Hartmannis Lecture Notes in Computer Science, Springer-Verlag, 1987

42. Christodoulakis, et al, The Multimedia Object Presentation Manager of MINOS, ACM, 1986, pp 295-310.

43. Christodoulakis, S., et al, Design and Performance Considerations for an Optical Disk Based Multimedia Object Server, IEEE Computer, Dec 1989, pp 45-56.

44. Chung and Pereira, Timed Petri Net Representation of SMIL, IEEE Multimedia, 2005

45. Churchland, P.S., Neurophilosophy, MIT Press (Cambridge, MA), 1986.

46. Ciardo et al, A Characterization of the Stochastic Process Underlying a Stochastic Petri Net, IEEE Transactions On Software Engineering, Vol. 20, No. 7, July 1994

47. Clark, A., Microcognition, MIT Press (Cambridge, MA), 1989.

48. Coll, S. The Deal of the Century, Atheneum (New York), 1986.

49. Copeland, T.E., J.F. Weston, Financial Theory and Corporate Policy, Addison Wesley (Reading, MA) 1983.

50. Cortadella et al, Synthesizing Petri Nets from State-BasedModels, IEEE, 1995

51. Couch, S., T.P. McGarty, H. Kahan, QUBE: The Medium of Interactive Response, A Compendium for Direct Marketeers, Direct Mktg Assn., pp 162-165, 1982.

52. de Sola Pool, I., Technologies Without Barriers, Harvard University Press (Cambridge, MA), 1990.

53. de Sola Pool, I., The Social Impact of the Telephone, MIT Press (Cambridge, MA), 1977.

54. Delbruck, Max, Mind From Matter, Blackwell (Palo Alto, CA), 1986.

55. Depew, D.J., B.H. Weber, Evolution at a Crossroads, MIT Press (Cambridge, MA), 1985.

56. Dertouzos, M.L., et al, Made in America, MIT Press (Cambridge, MA), 1989.

57. Dertouzos, M.L., J. Moses, The Computer Age, MIT Press (Cambridge, MA), 1979.

58. Dorfman, R.A., P.A. Samuelson, R.M. Solow, Linear Programming and Economic Analysis, Dover (New York), 1986.

59. Dreizen, Content-Driven Progressive Transmission of Grey-Scale Images IEEE transactions on Communications, vol. COM-35, no.3, March 1987.

60. Dreyfus, Hubert L., Being-in-the-World, The MIT Press (Cambridge, Mass.), 1991.

61. Drucker, Peter F., Adventures of a Bystander, Harper Row (New York), 1979.

62. Dugan, D.J., R. Standard, Barriers to Marginal Cost Pricing in Regulated Telecommunications, Public Utilities Fortn., vol 116, No 11, pp 43-50, Nov 1985.

63. Eco, U., A Theory of Semiotics, Indiana University Press (Bloomfield, IN), 1979.

64. Egan, B.L., Costing and Pricing of the Network of the Future, Proc of International Switching Symposium, pp 483-490, 1987.

65. Egan, B.L., T.C. Halpin, The Welfare Economics of Alternative Access Carriage Rate Structures in the United States, Telecom Journal, Vol 54, No 1, pp 46-56, Jan 1987.

66. Elmarsi, Ramez, Shamkant Navathe, Fundamentals of Database Systems, Benjamam (Redwood City, CA) 1989.

67. Fleming, S., What Users Can Expect From New Virtual Wideband Services, Telecommunications, pp 29-44, October, 1990.

68. Franklin, Problems with Raster Graphics Algorithms in Data Structures for Raster Graphics edited by L.R.A. Kessener, F J. Peters and M.L.P. van Lierop Eurographic Seminars, Springer-Verlag, 1986

69. Freeman, The 35mm Handbook, Ziff David Books, 1980

70. Fruhan, W.E., Financial Strategy, Irwin (Homewood, IL), 1979.

71. Fulhaber, G.R., Pricing Internet: Efficient Subsidy, Information Infrastructures for the 1990s, J.F.

72. Gadamer, Hans Georg, Philosophical Hermeneutics, U. Cal Press (Berkeley), 1976.

73. Gadamer, Hans Georg, Truth and Method, Crossroad (New York), 1990.

74. Gadamer, Hans-Georg, Philosophical Apprenticeships, MIT Press (Cambridge) 1985.

75. Gadamer, Hans-Georg, Reason in the Age of Science, MIT Press (Cambridge), 1981.

76. Gawdun, M., Private-Public Network Internetworking, Telecommunications, Vol 21, No 11, pp 49- 58, Nov 1987.

77. Gechter and OReilly, Conceptual Issues for ATM IEEE N etw ork, January 1989

78. Geller, H., US Domestic Telecommunications Policy in the Next Five Years, IEEE Comm Mag, Vol

79. Geller, H., US Domestic Telecommunications Policy in the Next Five Years, IEEE Comm Mag, Vol 27, No 8, pp 19-23, Aug 1989.

80. Ghafoor et al, An Interconnection T opology for Fault-Tolerant M ultiprocessor System s 1987 IEEE 7th International Conference on Distributed C om puting

81. Goos and Hartmanis, Communication Support for D istributed Processing : D esign and Im plem entation Issues L. Svobodova in Networking in O pen System s Lecture Notes in Computer Science, Springer-Verlag, 1987

82. Grayling, A. C., Wittgenstein, Oxford (Oxford) 1988.

83. Green, Implementation of FDDI : a 100 M bit Token Ring in Localnet 86 Online Publishers, 1986

84. Hall, The Application of Human Visual System Models to Digital Color Image Compression in Image Coding Proceedings of the SPIE - The International Society for Optical Engineering, volume 594,1985

85. Haller, Rudolf, Questions on Wittgenstein, University of Nebraska Press (Great Britian), 1988.

86. Handel, Evolution of ISDN tow ards Broadband ISDN IEEE Netw ork, January 1989

87. Hanson, Owen, Design of Computer Data Files, Computer Science Press (Rockville, MD) 1988.

88. Hasegawa ey al, Com m unication-System Models Embedded in the OSI-Reference Model, a Survey IFIP, North-Holland, 1986

89. Heidegger, M., Basic Writings, Harper & Row (New York), 1977.

90. Heidegger, Martin, An Introduction to Metaphysics, Yale (New Haven) 1959.

91. Heidegger, Martin, Being and Time, Harper & Row (New York) 1962.

92. Heidegger, Martin, Early Greek Thinking, Harper & Row (New York) 1979.

93. Heidegger, Martin, On Time and Being, Harper & Row (New York) 1972.

94. Henderson, J.M., R.E. Quandt, Microeconomic Theory, McGraw Hill (New York), 1980.

95. Hills, J., Issues in Telecommunications Policy- A Review, Oxford Surveys in Information Technology, Vol 4, pp 57-96, 1987.

96. Hoshiko et al, A l00 Mb/s Optical Token Ring N etw ork Suitable for H igh-Speed Inter-Processor C om m unications 1987 IEEE 7*h International Conference on D istributed C om puting System s edited by R. Popescu-Zeletin, G. Le Lann, K.H. Kim Com puter Society Press of the IEEE, 1987

97. Hsing and Liou, The Challenge o f VLSI Technology to Video Communications in Applications of Digital Image Processing X - Proceedings SPIE - The International Society for Optical Engineering, 1987

98. Huber, P.W., The Geodesic Network, U.S. Department of Justice, Washington, DC, January, 1987.

99. Hudson, H.E., Proliferation and Convergence of Electronic Media, First World Electronic Media Symposium, pp 335-339, 1989.

100. Illich, I., B. Sanders, ABC, The Alphabetization of the Popular Mind, Vintage (New York), 1988.

101. Irwin, M.R., M.J. Merenda, Corporate Networks, Privatization ans State Sovereignty, Telecommunications policy, Vol 13, No 4, pp 329-335, Dec 1989.

102. Jackendoff, Ray, Semantics and Cognition, MIT Press (Cambridge) 1988.

103. Jansen. "Data Structures for Ray Tracing Data Structures for Raster Graphics edited by L.R.A. Kessener, F J. Peters and M.L.P. van Lierop Eurographic Seminars, Springer-Verlag, 1986

104. Jantzen, H., High Gothic, Princeton University Press (Princeton, NJ), 1984.

105. Jayant and P. Noll, Digital Coding of Waveforms -, Principles and Applications to Speech and Video, Prentice-Hall Signal Processing Series, 1984

106. Jenkins, F. A., H. E. White, Fundamentals of Optics, McGraw Hill (New York) 1957.

107. Johnson, Design Issuses in Netw ork in g Incompatible O perating System s Localnet 86 Online Publishers, 1986

108. Judd, D. B. et al, Color 2ND Edition, Wiley (New York) 1963.

109. Kaelin, E. F., Heideggers Being and Time, Florida State (Tallahassee) 1988.

110. Kahin, B., The NREN as a Quasi-Public Network: Access, Use, and Pricing, J.F. Kennedy School of Government, Harvard University, 90-01, Feb., 1990.

111. Kahn, A.E., The Economics of Regulation, MIT Press (Cambridge, MA), 1989.

112. Kaiser and Morss, Direct Comparison of 35mm Film and High Definition Television, in High Definition Television Colloquium 85, vol. 2, Minister of Supplies and Services, Canada, 1986

113. Kernel and Cherion, Request-R esponse am d M ulticast Interprocess C om m unication in the in Networking in Open Systems Lecture N otes in Com puter Science, Springer-Verlag, 1987

114. Kim, W., H. Chou, Versions of Schema for Object Oriented Databases, Conf VLDB, 1988.

115. Kirstein, P.T., An Introduction to International Research Networks, IEEE Int Council for Comptr Comm, Ninth International Conf, pp 416-418, 1988.

116. Knightsron The Reference M odel in Standards for O pen System s Interconnections McGraw-Hill, 1988

117. Knoll and Depp, Adapative Gray Scale Mapping to Reduce Registration Noise in Difference Images, in Computer Vision, Graphics and Image Processing 33, 129-137,1986.

118. Kohno, H., H. Mitomo, Optimal Pricing f Telecommunications Service in Advanced Information Oriented Society, Proceedings of International Conf on Info Tech, pp 195-213, 1988.

119. Kolnik and Garodnick, First FDDI Local Area in Proceedings of the 1 2 ^ Conference on Local Computer Networks of the Computer Society of the IEEE, 1987

120. Konsynski, B.R., E.W. McFarlan, Information Partnerships - Shared Data, Shared Scale, Harvard Business Review, pp. 114-120, September-October, 1990.

121. Kraus, C.R., A.W. Duerig, The Rape of Ma Bell, Lyle Stuart (Secaucus, NJ) 1988.

122. Krishnamurthy, E.V., Parallel Processing, Addison Wesley (Reading, MA), 1989.

123. Kuhn, T.S., The Structure of Scientific Revolutions, Univ Chicago Press (Chicago), 1970.

124. Kung, Hans, Theology for the Third Millenium, Doubleday (New York), 1988.

125. Kunii, T. L., Visual Database Systems, North Holland (Amsterdam), 1989.

126. Kurzweil, R., The Age of Intelligent machines, MIT Press (Cambridge, MA), 1990.

127. Larmouth, Com m itm ent, Concurrency and Recovery - ISO-CASE and IBM LU.62 in Networking in Open Systems edited by G. Goos and J. Hartmanis Lecture Notes in Computer Science, Springer-Verlag, 1987

128. Lawrence, P.R., D. Dyer, Renewing American Industry, Free Press (New York), 1983.

129. Leger et al, Distributed Arithmetic Implementation of the D.C.T. for Real Time Photvideotex on ISDN, Advances in Image Processing - Proceedings edited by A. Oosterlinck, A.G. Tescher SPIE - The International Society for Optical Engineering, 1987

130. Levi, L., Applied Optics, Wiley (New York) 1968.

131. Lierop, Intermediate Data Structures for Display Algorithms in Data Structures for Raster Graphics edited by L .R A . Kessener, F.J. Peters and M.L.P. van Lierop, Eurographic Seminars, Springer-Verlag, 1986

132. Liff, Color and Black & White Television Theory and Servicing, Prentice-Hall, Inc., Englewood Cliffs, 1985

133. Linge, David E., Philosophical Hermeneutics, University of California Press (England), 1977.

134. Little, T.D.C., A. Ghafoor, Synchronization and Storage Models for Multimedia Objects, IEEE Journal on Sel Areas in Comm, April, 1990, pp. 413-427.

135. Loomis, Mary, Data Management and File Structures, Prentice Hall (Englewood Cliffs) 1898.

136. Luce, R.D., H. Raiffa, Games and Decisions, Dover (New York), 1985.

137. Lynch, M., S. Woolgar, Representation in Scientific Practice, MIT Press (Cambridge, MA), 1990.

138. Lyons, J., Noam Chomsky, Penguin (New York), 1978.

139. Mac Cormac, Earl R., A Cognitive Theory of Metaphor, MIT Press (Cambridge) 1985.

140. Machizawa, M. Tanak, Image data Compression Based on Interpolative DPCM by Area Decomposition, in, tranalated from Denshi Tsushin Gakkai Ronbushi, vol 69-D, no.3, March 1986, in Sato et al, Systems and Computers in Japan, vol. 17, no.12, 1986

141. Maier, David, The Theory of Relational Databases, Computer Science Press (Rockville, MD) 1983.

142. Management, HIMSS Conference, San Francisco, Feb. 1991.

143. Mandelbaum, R., P.A. Mandelbaum, The Strategic Future of Mid Level Networks, J.F. Kennedy School of Government, Harvard University, Working Paper, October, 1990.

144. Markoff, J., Creating a Giant Computer Highway, New York Times, Sept. 2, 1990.

145. Mayne, Netw ork Services, W ide Area N etw orks in Linked Local Area N etw orks, second edition John W iley & Sons, 1986

146. McCarthy, T., The Critical Theory of Jurgen Habermas, MIT Press (Cambridge, MA), 1978.

147. McCulloch, W., Embodiments of Mind, MIT Press (Cambridge, MA), 1988.

148. McGarty, T.P., R. Veith, Hybrid Cable and Telephone Networks, IEEE CompCon, 1983.

149. McGuiness, Brian, Wittgenstein: A Life, The University of California Press (Great Britian), 1988.

150. McLuhan, M., The Gutenberg Galaxy, Univ Toronto Press (Toronto), 1962.

151. McLuhan, M., Understanding Media, NAL (New York), 1964.

152. Mee, C. Dennis, Eric D. Daniel, Magnetic Recording, McGraw Hill (New York) 1988.

153. Menzes, An Interconnection N etw ork Supporting Ralational Join Operations 1987 IEEE 7th International Conference on D istributed C om puting System s edited by R. Popescu-Zeletin, G. Le Lann, K.H. Kim Computer Society Press of the IEEE, 1987

154. Milgrom, L. R., The Colours of Life, Oxford (New York) 1997.

155. Mill, John S., et al, Utilitarianism and Other Essays, Penguin Classics (England), 1987.

156. Minzer and Spears, New Directions in Signaling for Broadband ISDN S. E. M inzer, D.R. Spears in IEEE Com m unications M agazine, February 1989

157. Monk, Ray, Ludwig Wittgenstein: The Duty of Genius, The Free Press (New York), 1990.

158. Morita and Kataoka, Color Display System for High-Definition Television High Definition Television Colloquium 85, vol. 2, Minister of Supplies and Services, Canada, 1986

159. Morrill, Jane, Multimedia, BYTE, February, 1990, pp. 200- 237.

160. Mounce, H. O., Wittgenstein, U Chicago Press (Chicago) 1981.

161. Muroyama, J.H., H.G. Stever, Globalization of Technology, National Academy Press (Washington, DC), 1988.

162. Nadel, L. et al, Neural Connections and Mental Computation, MIT Press (Cambridge, MA), 1989.

163. Nickerson, R.S., Using Computers, MIT Press (Cambridge, MA), 1986.

164. Nicolau, Cosmos, An Architecture for Real Time Multimedia Communications Systems, IEEE Journal on Sel Areas in Comm, April, 1990, pp. 391-400.

165. Nikolau et al, Processor Performance. Allocation and Relocation of Processes in Distributed Computer System s in Current A dvances in D istributed C om puting and Com m unications Computer Science Press, 1987

166. Noam, E. M., Network Tipping and the Tragedy of the Common Network, J.F. Kennedy School of Government, Harvard University, Working Paper, October, 1990.

167. OHara, S., The Evolution of A Modern Telecommunications Network to the Year 2000 and Beyond, IEE Proc, Vol 132, No 7, pp 467-480, 1985.

168. Osherman, Daniel N., Edward E. Smith, Thinking, MIT Press (Cambridge), 1990.

169. Osherman, Daniel N., Howard Lasnik, Language, MIT Press (Cambridge) 1990.

170. Osherson, Daniel N. et al, Visual Cognition and Action, MIT Press (Cambridge) 1990.

171. Parsaye, Kamran, et al, Intelligent Databases, Wiley (New York) 1989.

172. Pears, David, Wittgenstein, Harvard Press (Cambridge),1969.

173. Peterson, Petri Net Theory and the Modeling of Systems, Prentice-Hall, Englewood Cliffs, 1981.

174. Pierkainen and Harwood, Segmentation of Color Images Using Edge-Preserving Filters, in Advances in Image Processing and Pattern Recognition, edited by V. Cappellini, R. Marconi North-Holland, 1986

175. Pindyck, R.S., D.L. Rubinfield, Microeconomics, McGraw Hill (New York), 1989.

176. Pizano, A. et al, Specification of Spatial Integrity Constraints in Pictorial Databases, IEEE Computer, Dec 1989, pp 59-71.

177. Plaehn, PHIGS : Programmers Hierarchical Interactive Graphics Standards - A Giant Step toward a Universal Graphics Standard, Byte, November 1987

178. Plompen et al, An Image Knowledge Based Video Codec for Low Bitrates, Advances in Image Processing - Proceedings, SPIE - The International Society for Optical Engineering, 1987

179. Poopescu et al, , System s Com puter Society Press of the IEEE, 1987

180. Porter, M., Competitive Advantage, Free Press (New York), 1985.

181. Porter, M., Competitive Strategy, Free Press (New York), 1980.

182. Porter, M., The Competitive Advantage of Nations, Free Press (New York), 1990.

183. Public Utilities Fortn., vol 116, No 11, pp 43-50, Nov 1985.

184. Raisbeck, G., Information Theory, MIT Press (Cambridge, MA), 1963.

185. Rhodes et al, "Locally Optimal Run-Length Compression Applied to CT Images, IEEE Transactions on Medical Imaging, vol. MI-4, no. 2, June 1985

186. Rhodes et al, "Locally Optimal Run-Length Compression Applied to CRT Images in IEEE Transactions on Medical Imaging, vol. MI-4, no.2, June 1985.

187. Rorty, Richard, Philosphy and the Mirror of Nature, Princeton University Press, (Princeton, NJ),

188. Rosenfeld and Kak, Digital Picture Processing, Second Edition, vol. 1, Computer Science and Applied Mathematics, Academic Press, 1982

189. Rossi, B., Optics, Addison Wesley (Reading, MA) 1957.

190. Rousseau, Jean-Jacques, The Social Contract, Penguin Classics (England), 1968.

191. Russell, B., The Problems of Philosophy, Oxford University Press (Oxford), 1959.

192. Rutkowski, A.M., Computer IV: Regulating the Public Information Fabric, Proc of the Regional Conf of the International Council for Computer Communications, pp 131-135, 1987.

193. Sato et al, Data Compression of Gray-Scale Images by Level Plane Coding tranlsted from Denshi Tsushin Gakkai Ronbushi, vol 69-B, no.8, August 1986, Electronics and Communications in Japan, Part 1, Vol. 71, n o.l, 1988

194. Schreiber, A Generalized Image Processing System in "Fundamentals of Electronic Imaging SystemsSome Aspects of Image Processing Springer Series in Information Sciences, Springer-Verlag, 1986

195. Schreiber, A Simple Television System in Fundamentals of Electronic Imaging Systems Some Aspects of Image Processing, Springer Series in Information Sciences, Springer-Verlag, 1986

196. Schrube and Yucesan, Transforming Petri Nets Into Event Graph Models, Proceedings Of the 1994 Winter Simulation Conference

197. Sears, F. W., Optics, Addison Wesley (Reading, MA) 1949.

198. Shakirova, Constructionof Petri Nets Via States of Action Objects And Subjects, Proceedings of the 9th International Conference on Neural Information Processing (ICONIPOZ), Vol. 1

199. Shandle, Who will Dominate the Desktop in the 90s?, Electronics, Feb 1990, pp 48-59.

200. Sharp, Networks and Interconnection Structures in An Introduction to Distributed and Parallel Processing Computer Science Texts, Blackwell Scientific Publications, 1987

201. Sharples, Mike, et al, Computers and Thought, MIT Press (Cambridge) 1990.

202. Shubik, M., A Game Theoretic Approach to Political Economy, MIT Press (Cambridge, MA), 1987.

203. Shubik, M., Game Theory in the Social Sciences, MIT Press (Cambridge, MA), 1984.

204. Silverman, Kaja, The Subject of Semiotics, Oxford University Press (New York), 1983.

205. Simon, H.A., The Sciences of the Artificial, MIT Press (Cambridge, MA), 1969.

206. Sirbu, M.A., D.P. Reed, An Optimal Investment Strategy Model for Fiber to the Home, International Symposium on Subscriber Loops to the Home, pp. 149-155, 1988.

207. Sokal, R., P. Sneath, Principle of Numerical Taxonomy, Freeman (San Francisco) 1963.

208. Spector, Communication Support in Operating Systems for Distributed Transactions Department of Computer Science, Carnegie-Mellon University, 1986

209. Speidel et al, Improved Hybrid Coders with 2-D Signal Processing for M oving Pictures J. in Advances in Image Processing - Proceedings SPIE - The International Society for Optical Engineering, 1987

210. Spulber, D.F., Regulation and Markets, MIT Press (Cambridge, MA), 1990.

211. Steiner, George, Heidegger, U Chicago (Chicago), 1978.

212. Steinmetz, Ralf, Synchronization Properties in Multimedia Systems, IEEE Journal on Sel Areas in Comm, April, 1990, pp 401-412.

213. Strong, J., Concepts of Classical Optics, Freeman (San Francisco) 1958.

214. Sutcliffe, A., Human-Computer Interface Design, Springer Verlag (New York), 1988.

215. Swastek and Verreke, M igrating to FDDI on Your N ext Big LAN Installation in Data C om m unications, June 21, 1989

216. Tannenbaum, A., Computer Communications, Prentice Hall (Englewood Cliffs, NJ), 1989.

217. Taylor, Mark C., Ceconstruction in Context, The university of Chicago Press (Chicago, IL), 1986.

218. Tcha et al, Link-by-Link Bandwidth Allocation in an Integrated Voice/Data Network Using the Fuzzy Set Approach in Computer Networks and ISDN System s, 16, Elsevier Publishers, 1989

219. Temin, P., The Fall of the Bell System, Cambridge Univ Press (Cambridge), 1987.

220. Ten Hagen and Trienkens, Pattern Representation in Data Structures for Raster Graphics edited by L.R.A. Kessener, FJ. Peters and M.L.P. van Lierop Eurographic Seminars, Springer-Verlag, 1986

221. Teorey, Toby, James Fry, Design of Database Structures, Prentice Hall (Englewood Cliffs, NJ) 1982.

222. Terry, D., D. Swinehart, Managing Stored Voice in the Etherphone System, ACM Trans Cptr Sys, Vol 6, No 1, Feb 1988, pp3-27.

223. Teunissen, J. van den Bos, A Model for Raster Graphics Language Primitives edited by L .R A . Kessener, F.J. Peters and M.L.P. van LieropEurographic Seminars, Springer-Verlag, 1986

224. Thomasian, A Perform ance Study of Dynam ic Load Balancing in Distributed System s in 1987 IEEE 7th International Conference on Distributed Com puting System s edited by R. Popescu-Zeletin, G. Le Lann, K.H. Kim Com puter Society Press of the IEEE, 1987

225. Thurber, Getting a Handle on FDDI in Data C om m unications, June 21, 1989

226. Tjaden, G., CATV and Voice Telecommunications, IEEE ICC, 1984. Toffler, A., The Adaptive Corporation, Bantam (New York), 1985. Toy, S., Castles, Heinemann (London), 1939.

227. Tsichritzis, D. et al, A Multimedia Office Filing System, Proc VLDB 1983.

228. Tu et al, A Hybrid Coding Scheme for Still Images :, Adaptive Vector Quantization with Correction Information in DCT Domain, Applications of Digital Image Processing X - Proceedings, edited by A.G. Tescher, SPIE - The International Society for Optical Engineering, 1987

229. Ueltzen, Codex View s on LAN System s Integration Localnet 86, Online Publishers, 1986

230. Ullman, Jeffrey, Database and Knowledge Base Systems, Computer Science Press (Rockville, MD) 1988.

231. van Lierop, Intermediate Data Structures for Display Algorithms, Data Structures for Raster Graphics, edited by L.R.A. Kessener, F J. Peters and M.L.P. van Lierop, Eurographic Seminars, Springer-Verlag, 1986

232. Vandermeulen et al, Surface Reconstruction from Planar Cross Sections for 3-D Integrated Representation of Medical Images, in Advances in Image Processing - Proceedings edited by A. Oosterlinck, A.G. Tescher SPIE - The International Society for Optical Engineering, 1987

233. Venetsanopoulos, Image Acquisition and Presentation, in Advances in Image Processing and Pattern Recognition edited by V. Cappellini, R. Marconi North-Holland, 1986

234. Vickers, R., T. Vilmansen, The Evolution of Telecommunications Technology, Proc IEEE, vol 74, No 9, pp 1231-1245, Sept 1986.

235. Von Auw, A., Heritage and Destiny, Praeger (New York), 1983.

236. Von Hippel, E. The Sources of Innovation, Oxford (New York), 1988.

237. Von Neumann, J., O. Morgenstern, Theory of Games and Economic Behavior, Wiley (New York), 1944.

238. Walach et al, A Modified Block Truncation Coding Technique for Image Compression inAdvances in Image Processing and Pattern Recognition, edited by V. Cappellini, R. Marconi North-Holland, 1986

239. Warnke, Gadamer, Stanford (Stanford, CA), 1987.

240. Weinstein, Getting the Picture -A Guide to CATV and the New Electronic Media, IEEE Press, 1986

241. Weizenbaum, Joseph, Computer Power and Human Reason, Freeman (New York) 1976.

242. Wendland, A Scanning Scheme for a New HDTV Standard, HDTV 85, Minister of Supply and Services Canada, 1986

243. West, E.H., et al, Design, Operation, and Maintenance of a Multi Firm Shared Private Network, IEEE MONECH Conf, pp80-82, 1987.

244. White, Image Transmission over Grade T elephone Lines in Applications o f D igital Image Processing X - Proceedings edited by A.G. Tescher SPIE - The International Society for Optical Engineering, 1987

245. Wiener, N., Cybernetics, MIT Press (Cambridge, MA), 1948.

246. Wiener, N., God and Golem, MIT Press (Cambridge, MA), 1964.

247. Wiener, N., The Human Use of Human Beings, Avon (New York), 1967.

248. Wilkes, M.V., The bandwidth Famine, Comm ACM, pp 19-21, Vol 33, No 8, Aug 1990.

249. Winograd, Terry, Fernando Flores, Understanding Computers and Cognition, Addison Wesley (Reading, MA), 1987.

250. Wirth, T.E., Telecommunications in Transition, U.S. House of Representatives Committee Report, Nov, 1981.

251. Wittgenstein, Ludwig, Philisophical Investigations, MacMillian Publishing Co. (New York), 1958.

252. Wittgenstein, Ludwig, Tractatus, Routledge & Kegan Paul (London) 1922.

253. Woelk, D, W. Kim, Multimedia Information Management, VLDB Conf 1987.

254. Woelk, D. et al, An Object Oriented Approach to Multimedia Databases, ACM, 1986, pp 311-325.

255. Wolfe, T., From Bauhaus to Our House, Simon and Schuster (New York), 1981.

256. Wright, Karen, The Road to the Global Village, Scientific American, March 1990, pp 83-94.

257. Zhgou and Ferrari, A Measurement Study of Load Balancing Performance 1987 IEEE 7th International Conference on Distributed C om puting System s Computer Society Press of the IEEE, 1987

258. Zuboff, S., In the Age of the Smart Machine, Basic Books (New York), 1988.

259. Zurawski and Zhou, Petri Nets and Industrial Applications: A Tutorial, IEEE. Transactions On Industrial Electronics, Vol. 41, No. 6, December 1994

13 INDEX