Report Number: WUCS-91-45

1991-07-30

Project Zeus: Design of a Broadband Network and its Application on a University Campus

Jerome R. Cox Jr. and Jonathan S. Turner

This proposal outlines a plan for the design, deployment, and operation of the high speed campus network at Washington University based on fast packet switching technology that has been developed here during the last several years. This new network will support ubiquitous multimedia workstations with high-resolution graphics and video capabilities, opening a wide range of new applications in research and education. It will support aggregate throughputs of hundreds of gigabits per second and will be designed to support port interfaces at up to 2.4 Gb/s. Initial implementations will emphasize 155 Mb/s port rates, with higher rates introduced as the... Read complete abstract on page 2.

Project Zeus: Design of a Broadband Network and its Application on a University Campus

Jerome R. Cox Jr. and Jonathan S. Turner

Complete Abstract:

This proposal outlines a plan for the design, deployment, and operation of the high speed campus network at Washington University based on fast packet switching technology that has been developed here during the last several years. This new network will support ubiquitous multimedia workstations with high-resolution graphics and video capabilities, opening a wide range of new applications in research and education. It will support aggregate throughputs of hundreds of gigabits per second and will be designed to support port interfaces at up to 2.4 Gb/s. Initial implementations will emphasize 155 Mb/s port rates, with higher rates introduced as the demand arises and as economic permits. WE propose to move this technology quickly into an operational setting where the objectives of network use and network research can be pursued concurrently.

Project Zeus: Design of a Broadband Network and its Application on a University Campus

Jerome R. Cox, Jr. and Jonathan S. Turner

WUCS 91-45

July 30, 1991

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130-4899

Abstract

This proposal outlines a plan for the design, deployment and operation of a high speed campus network at Washington University based on fast packet switching technology that has been developed here during the last several years. This new network will support ubiquitous multimedia workstations with high-resolution graphics and video capabilities, opening up a wide range of new applications in research and education. It will support aggregate throughputs of hundreds of gigabits per second and will be designed to support port interfaces at up to 2.4 Gb/s. Initial implementations will emphasize 155 Mb/s port rates, with higher rates introduced as the demand arises and as economics permits. We propose to move this technology quickly into an operational setting where the objectives of network use and network research can be pursued concurrently.

Contents

List of Figures

Project Zeus: Design of a Broadband Network and its Application on a University Campus

Jerome R. Cox, Jr. and Jonathan S. Turner

1. Motivation and Objectives

During the last several years there has been a growing recognition that fast packet switching technology (also known as *Asynchronous Transfer Mode* or ATM) will form the basis of next generation communication networks. One attractive aspect of ATM technology is its inherent scalability, both in the total throughput a network can support and the port data rates. While much of the focus in ATM has been on public network applications, most people now agree that the demand for these new networks will come from computer-based applications needing higher bandwidth than current shared-access LANs are able to deliver. LAN and workstation vendors are recognizing the need to introduce switching within campus networks to expand their capacity and range of applications, and are now moving aggressively to develop products to fill this need.

Washington University has been deeply involved in the development of a switching technology based on small, fixed-length packets called *cells* and the application of this technology to medical imaging. We are currently engaged in a project sponsored by Southwestern Bell and NEC America aimed at demonstrating technical feasibility of the underlying technology and providing some initial applications. We now propose to work with a variety of industrial partners to create commercial implementations of the technology, to apply that technology throughout the university community for the benefit of users and to answer several pressing system questions that can only be addressed in an operational network environment. Thus, our objectives are threefold: 1) to collaborate with industry in the transfer of the technology for all the components needed to construct an ATM network; 2) to make available to users an advanced network with thousands of high performance workstations supported by the necessary hardware and software; and 3) to provide a realistic testbed for communications research on questions concerning network congestion, routing, planning, interoperability, interworking, remote visualization and techniques for operations and management.

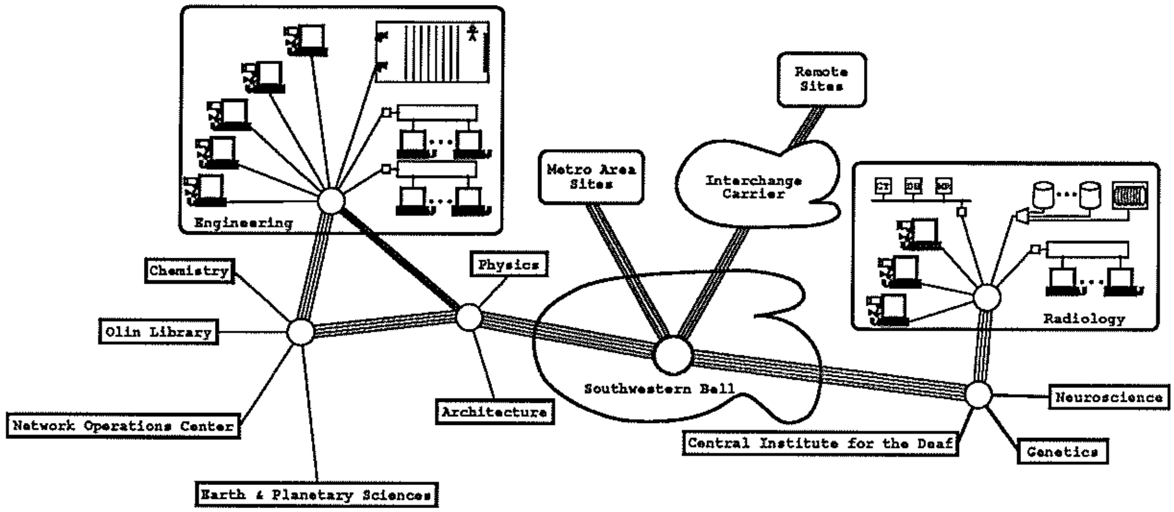


Figure 1: A Fast Packet Campus Network

Figure 1 illustrates the concept behind the proposed ATM network. The system would consist of several switches on each of the university's two campuses. The switches would be connected by transmission links operating at speeds of 155 Mb/s, 620 Mb/s and 2.48 Gb/s. Each switch would support potentially several hundred interfaces, with a variety of port speeds. We expect the majority of ports to be 155 Mb/s but will support higher speed ports as the need for them arises. These interfaces could be connected directly to multimedia workstations and central compute servers or could be connected indirectly through shared access LANs such as Ethernet or FDDI. Video would play a central role in the network, allowing access to centrally stored video information through the network, two-way or multipoint video conferencing and remote classroom instruction using video.

The network will include connections to remote sites using either dedicated or switched channels provided by the local exchange carrier or interexchange carriers. In particular, connection to new broadband services planned by Southwestern Bell would make possible classrooms, medical offices and hospitals, all at locations more convenient to their clientele, and all linked to the university or the medical center by video, high resolution image transmission and shared databases. Connection to interexchange carriers would allow scientists to interact with their colleagues at other institutions and with distant supercomputers via the emerging National Research and Education Network.

2. Applications

A goal of Project Zeus is to explore possibilities which may transform daily practice in a number of application areas and, at the same time to conduct experiments useful in understanding the future bandwidth requirements of these application areas. Those of us familiar with network technology wish to work with scientists and scholars from a broad range of disciplines to increase our ability to generalize the experimental results to new areas.

Many possible applications of the proposed ATM network have been discussed with our colleagues at Washington University. Four seem particularly appropriate for the initial experiments with Zeus network technology. These applications are described in some detail in Section 7, but are described briefly in the following paragraphs.

Medical Imaging and Electronic Radiology. Applications of broadband network technology in medicine are particularly attractive because of the extensive use of medical images in diagnosis and treatment, because of the need for prompt decision making and because of the promising results of early experiments, particularly in the field of radiology.

At Washington University the development of picture archiving and communication systems (PACS) for the acquisition, management and display of radiological images is well underway. A PACS demonstration using prototype ATM technology is scheduled for late 1991. A plan for use of the Zeus network technology is described in Section 7.1.

Optical Sectioning Microscopy. Until now it has not been possible to see inside a biological specimen without slicing it into sections and thereby losing any chance of visualizing a living organism. The optical sectioning microscope analyzes a series of 2D images obtained at different focal planes and displays the 3D structure of the living organism being viewed. This new instrument promises to allow biologists to investigate a variety of fundamental, and previously insoluble, problems.

At Washington University researchers in the departments of Biology, Computer Science and Electrical Engineering and in the Institute for Biomedical Computing are joining forces toward the development of distributed scientific visualization methods applied to the optical sectioning microscope. One of the handful of such instruments in existence, this microscope will be the central focus of this application area. Details of our plans for the use of Zeus technology in this application are presented in Section 7.2.

Earth and Planetary Sciences. The Remote Sensing Laboratory in the Department of Earth and Planetary Sciences at Washington University houses the lead Geosciences Node of the NASA Planetary Data System. The node is responsible for working with the Magellan and Mars Observer missions to ensure that the data acquired from spacecraft exploring the solar system are properly documented and archived. These data are primarily in the form of images that, in most cases, arrive over the Internet. A Seismology Laboratory and a Planetary Geophysics Laboratory will soon be adjacent to the Remote Sensing Laboratory in their new quarters, now under construction. Each of these three laboratories employs high-resolution images in their daily work and to achieve their scientific goals they will require rapid access to large image databases and to a high-performance compute server. Project Zeus will provide the local communications technology required for these three laboratories and the National Research and Education Network will provide the connectivity required for other NASA sites elsewhere in the nation. Our plans for Earth and Planetary Sciences are described in Section 7.3.

Visualization in Art and Architecture. The Urban Research and Design Center in the School of Architecture has identified a research agenda which addresses issues regarding the development of a designer's workstation. The design of buildings, urban places and cities is augmented through multimedia and 3D models that are linked to image collections, graphics and art. Initially, Zeus technology would serve the School of Architecture, but as experience is gained the five divisions of the university associated with the visual arts, Architecture, Fine Arts, Art and Archeology, the Art and Architecture Library and the

Gallery of Art, would participate in Project Zeus. We believe that this new technology will change qualitatively the way the visual arts are practiced and make possible systematic collection and rapid retrieval from a large library of still images, image sequences and supporting information for the use of students and professionals alike. Our plans in this regard are presented in Section 7.4.

Other Applications. Beyond these four applications, we have had experience with collaborators on the existing 10 Mb/s campus network that give us confidence that many medical, scientific and educational applications of Zeus technology can be accomplished in the coming years. In Section 7.6 we review briefly these additional possibilities.

3. Networking Questions

Another goal of Project Zeus is to provide a realistic testbed suitable for communications research on pressing questions of network design and operation. Laboratory experiments, simulations and demonstrations all provide answers to some communications research questions, but until thousands of users with hundreds of applications test a network, questions that depend on real traffic and service patterns remain unanswered.

Network Congestion. Algorithms for the management of network congestion are receiving increased attention within the last year. Although these algorithms require analysis by methods involving both theory and simulation, confidence in them is incomplete until tests in a realistic testbed have been carried out and analyzed. Network instrumentation for this and other related purposes will be a component of Project Zeus network design and at least one approach to congestion management will be investigated.

Efficient Routing. Point-to-point routing has been well studied and efficient algorithms have been incorporated into the design of the connection management software for the prototype ATM demonstration scheduled for 1991. However, an important distinction with respect to other proposed broadband networks for this demonstration network and for Project Zeus is their capability to provide multipoint connections, including multicast video connections with possibly millions of viewers. How robust such connections will be under the stress imposed by the behavior of real viewers must be investigated.

Network Planning and Configuration. Switched networks require more attention to network capacity planning than the shared access LANs currently used in campus networks. We anticipate the need for a general software tool that will support network planning and configuration, allowing detailed consideration of network expansion alternatives, taking into account traffic requirements, physical restrictions on cable and equipment placement as well as installation and maintenance costs. Such a tool must support both analysis of proposed network configurations and synthesis of optimal or near optimal configurations and must take into account the unique traffic characteristics of ATM networks, including their multicast capability.

Interoperability. The fast packet switching technology based on the ATM standard will be deployed both in campus networks (as in the case of Project Zeus) and in public broadband networks. The trade-offs of these two scenarios are different, and thus, can lead to differences in ATM signaling protocols, which are still in their infancy. Moreover, there are a number of proposed ATM switch architectures which may be used, even in the same network. The success of ATM critically depends on the interoperability of different switches and associated signaling protocols. We expect to utilize Project Zeus to investigate the problems associated with the interoperation of several other commercial broadband switches, perhaps including the NEC ATOM switch, the Bellcore Sunshine switch, and the Fujitsu switch.

Internetworking. The existing communication environment is best characterized as an internet consisting of a number of low speed networks interconnected by gateways. We believe that the future communications environment will continue to be an internet because of continued existence of multiple providers, new technological developments, and the different operational trade-offs selected for different network environments. The future internet will include emerging high speed networks, such as ATM and FDDI networks and will need to support a variety of applications, some of these requiring high bandwidth with performance guarantees. This places new demands on internet protocols and gateways. To meet these new demands, we have proposed a novel internet abstraction called the *Very High Speed Internet Abstraction* (VHSI) [31, 32]. The VHSI abstraction is aimed at supporting both connection-oriented applications requiring performance guarantees and classical datagram applications. We plan to use the Project Zeus testbed to address the internetworking issues and demonstrate viability of our VHSI abstraction.

Remote Visualization and Collaboration. Scientific visualization has emerged as a major computer-based field of study. Remote visualization is visualization that utilizes data and computing resources that are physically distributed. We believe a significant fraction of visualization applications will be remote visualization. Efficient remote visualization requires support far beyond what is needed for visualization on a single computer. It requires networks with high bandwidth and low latency, an efficient interprocess communication mechanism among communicating hosts, and proper adaptation and partitioning of the visualization computation. Recent developments in collaboration technologies suggest that computer-based collaboration applications will need to include support for multimedia and visualization. This leads to a new set of interesting issues that have to do with synchronization and concurrency control among various information streams (*e.g.*, data, image sequence, video, voice, and control) and among various end points of a given application. We plan to use the Zeus testbed and a class of remote collaboration applications to research these issues and create a prototype support environment for such state of the art applications.

4. Creating the Network Components

Project Zeus is organized in three phases. Phase 0, now underway, seeks to demonstrate feasibility of the core technology, provide a basis for a more complete design and provide a testbed for application development. The network created in this phase will be primarily an experimental vehicle, rather than an operational network supporting real users. This phase of the project began in 1988 and will continue through 1991.

Phase 1, scheduled to begin at the start of 1992 and run through 1994, will create all the key components needed to establish an ATM campus network and provide extensive support for application development. When complete, the phase 1 network will be an operational system supporting a variety of users in key departments within the university.

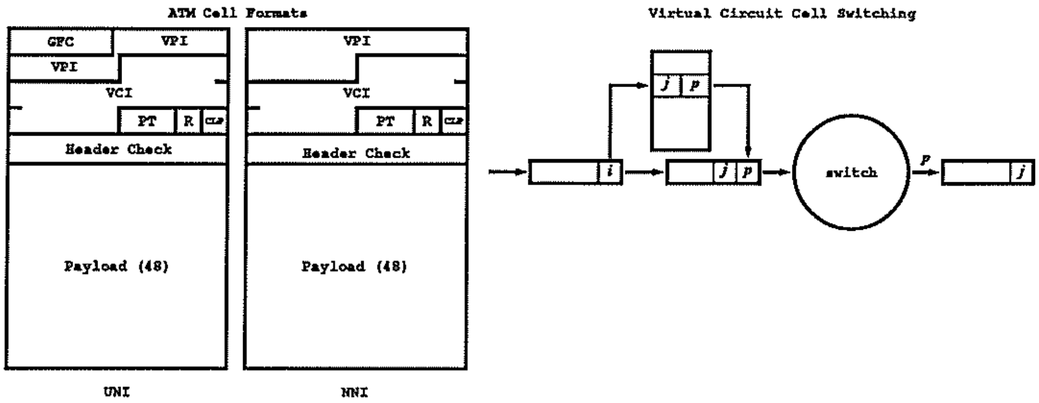


Figure 2: ATM Cell Formats and Cell Switching

During phase 2, which will run from 1994 through 1996, we plan to expand the range of interfaces that can be used to access the network, construct components for larger scale networks and reduce the cost of key network components. The phase 2 network will support users in all departments of the university.

4.1. ATM Network Technology

Project Zeus will take advantage of the emerging technology for *Asynchronous Transfer Mode* (ATM) networks. ATM networks provide virtual circuit oriented cell switching, a form of packet switching in which user data is carried in small fixed-length blocks called cells, where each cell includes a label that identifies the user channel that it belongs to. Communication over an ATM network takes place over virtual circuits which are typically established when some user application is initiated. When a virtual circuit is established, a route is selected, and subsequently all cells transmitted on that virtual circuit are forwarded along the selected route. Since the data objects transmitted by users typically consist of large amounts of data, it is necessary to fragment user data objects into cells and reassemble them at the receiver. This is accomplished using a simple fragmentation and reassembly protocol and supporting hardware in a host's network interface. Host software, is typically free to work with data units of arbitrary size, limited only by internal buffer space. All cell level processing can be left to the interface hardware.

Figure 2 illustrates the essentials of virtual circuit oriented cell switching. As shown on the right side of the figure, the cell's multiplexing label is used to select an entry from a routing table; the selected entry includes the number of the switch output that the cell is to be forwarded to and a new multiplexing label. The figure also shows the ATM cell formats used at the *User Network Interface* (UNI) and the *Network-Network Interface* (NNI). There are two multiplexing options in ATM networks, *Virtual Paths* and *Virtual Circuits*. Cells belonging to different virtual paths are distinguished by their *Virtual Path Identifier* (VPI) and cells belonging to different virtual circuits are distinguished by their *Virtual Circuit Identifier* (VCI). Most often, user connections are implemented using virtual circuits. Virtual paths are essentially bundles of virtual circuits. Two hosts can use a virtual path to multiplex many individual application streams together, using the VCI to distinguish these paths. The network does not interpret or modify the VCI fields of cells on virtual path connections, so the hosts can setup new virtual circuits on an established virtual path without having to request them from the network. In this paper we don't generally distinguish between virtual paths and virtual circuits; the reader may assume that statements made about virtual circuits also apply to virtual paths. Other fields in the five byte header of the ATM cell include a *Generic Flow Control* field (GFC), a *Payload Type* (PT), a *Reserved* bit (R), a *Cell Loss Priority* bit (CLP), and *Header Check* field (HEC). The *Payload* field carries user information and is 48 bytes long.

A key objective of ATM network technology is to ensure consistent performance to users in the presence of stochastically varying traffic. This is necessary to ensure adequate performance for many high speed applications (such as video), which require guaranteed throughput. The objective is attained by selecting virtual circuit paths according to the anticipated traffic and allocating the necessary network resources. This requires that the users specify, at virtual circuit setup, the amount of network resources they require. It also means that if the required resources are not available, a user's request could be refused by the network. The resource specification allows statistical sharing of bandwidth along virtual circuit paths; users specify their peak and average data rates, as well as a maximum burst size. Using this information, the network allocates resources in such a way that almost all information bursts are delivered intact. Soft resource specifications are also possible; users can specify minimum and maximum bandwidth requirements, which the network will accommodate to the best of its ability. In this case, the network can also adjust the resources allocated to established virtual circuits in order to avoid blocking new virtual circuit requests.

Conventional communication networks focus on point-to-point communication. Multicast communication, in which information from a single source is distributed to multiple receivers, is a natural generalization that is essential for efficiently supporting video distribution applications and useful in a variety of other applications as well. The basic concept of multicast virtual circuit switching is illustrated in Figure 3. The figure shows a network consisting of a collection of switches and concentrators, configured to support two multicast virtual circuits, with sources shown at the top of the figure. At each switch involved in a particular multicast virtual circuit, the incoming cells are replicated, assigned new virtual circuit identifiers and forwarded on selected outgoing links. The configuration of a particular virtual circuit is specified in the switching systems' internal control tables and can be modified through control messages.

As in the case of ordinary virtual circuits, the data rate of a multicast virtual circuit is completely flexible, as is the number of endpoints. The number of endpoints can change dynamically during the lifetime of a virtual circuit with new endpoints added and removed over time. The one-to-many virtual circuits is merely a special case of a more general communication model that is described in detail in Section 6. Briefly, a general multicast virtual circuit supports transmission and reception at all the participating endpoints. For each endpoint, transmission and reception can be independently enabled, allowing a very wide range of connection configurations. The bandwidth resource associated with such a multicast virtual circuit is viewed as a common bandwidth pool that is shared by the participating endpoints in whatever fashion they choose. Coordination of transmission by the different sources in a multicast virtual circuit is left to the sources. The network merely monitors the total bandwidth usage and ensures that it does not exceed what has been allocated.

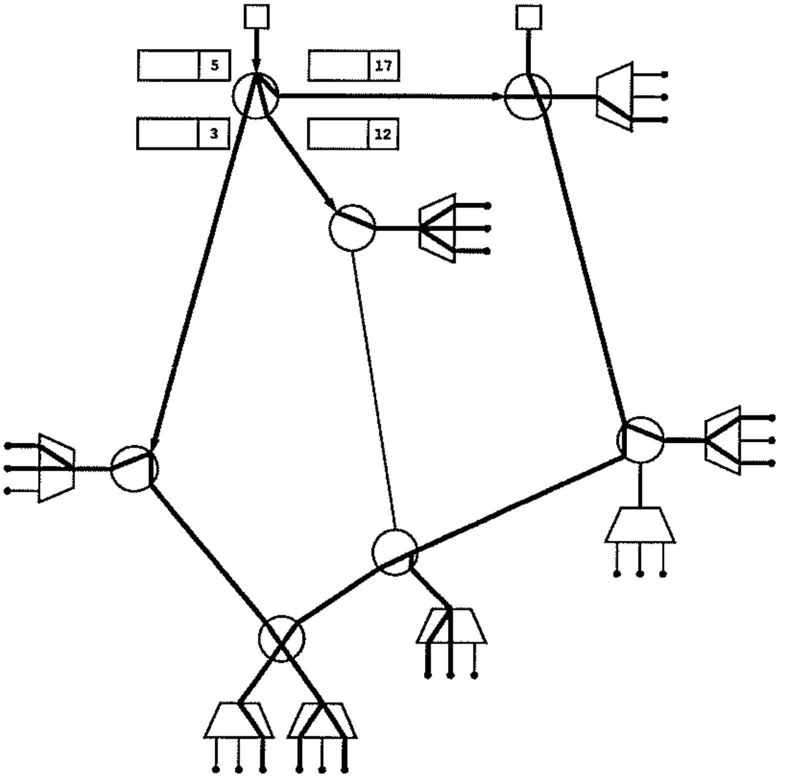


Figure 3: Multicast Virtual Circuit Switching

4.2. Phase Zero

Since mid-1986, the Advanced Networks Group within Washington University's Computer and Communications Research Center has been engaged in an ongoing research project concerned with flexible, high performance communication systems. A key element of this research has been the development of the broadcast packet switching technology, which is compatible with emerging ATM standards, and which supports both point-to-point and broadcast (or multicast) virtual circuits. In late 1988, a project was initiated to demonstrate the broadcast packet technology using a four node network supporting video and medical imaging applications. A new research organization, the Applied Research Lab, was formed to carry out this activity. The project, which has been supported by Southwestern Bell and NEC America, is referred to in this report as phase 0 of Project Zeus.

The overall objectives of phase 0 are to demonstrate the feasibility of the broadcast packet switching technology and its applicability to a variety of high speed applications, provide a basis for a more complete design and a testbed for application development. More specific goals include the design and construction of a network consisting of several 16 port broadcast packet switches, design and implementation of an ATM video interface, an ATM Ethernet interface and a physician's workstation to support medical imaging applications.

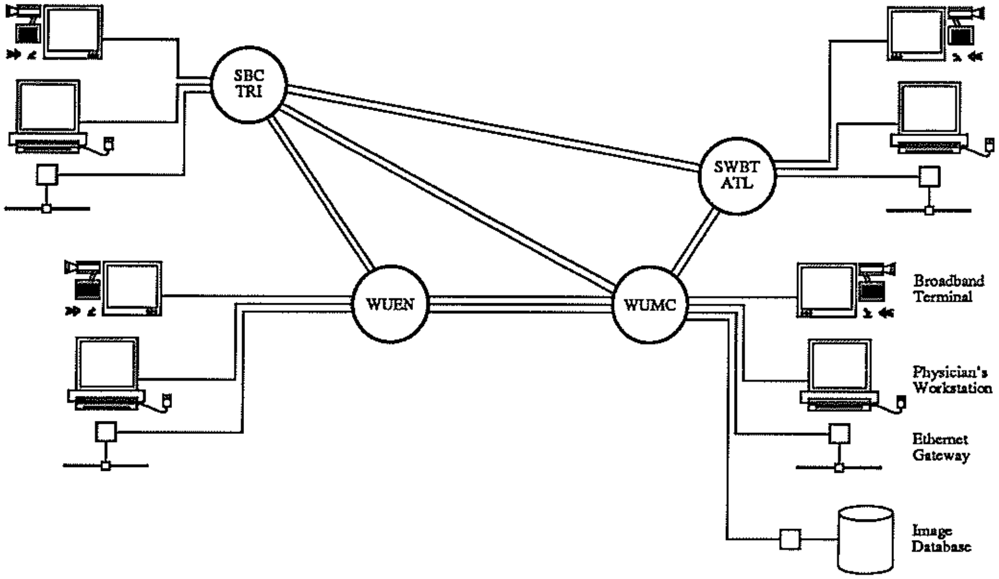


Figure 4: Phase 0 Network Configuration

Figure 4 shows the configuration of the network that is being constructed for phase 0. It consists of four switches located throughout the St. Louis area, each supporting 15 external interfaces operating at 100 Mb/s each. One switch is to be located in the Applied Research Laboratory of the Washington University School of Engineering and Applied Science (WUEN), one will be located at the Electronic Radiology Laboratory of the Mallinkrodt Institute of Radiology in the Washington University Medical Center (WUMC), one will be located at the Advanced Technology Laboratory at Southwestern Bell Telephone's downtown headquarters (SWBT/ATL) and the fourth at Southwestern Bell Corporation Technology Resources, Inc. (SBC/TRI) in St. Louis County. These sites will be connected using single mode fiber links provided by Southwestern Bell. The network will support interfaces to a *broadband terminal* supporting video, a *physician's workstation* for medical imaging applications, and interfaces to Ethernet and FDDI LANs.

4.2.1. Phase 0 Switch Architecture

Reference [39] describes a broadcast packet switching system that can support a wide variety of different applications, including video distribution, LAN interconnection and voice/video teleconferencing, all of which require multicast connections. The overall structure of the system is shown in Figure 5. Data is carried between switches in the form of ATM cells over fiber optic transmission links. The *Port Processors* (PP) perform link level protocol functions, including the determination of how each cell is routed and provide cell buffering. The core of the system is a switching network comprising a *Copy Network*, a *Routing Network*, and a set of *Broadcast Translation Circuits*; these are described below. The *Connection Processor* (CP), is responsible for establishing connections, including both point-to-point and multipoint connections, as well as overall system control. The *Switch Module Interface* (SMI) provides an interface between the CP and core of the switch.

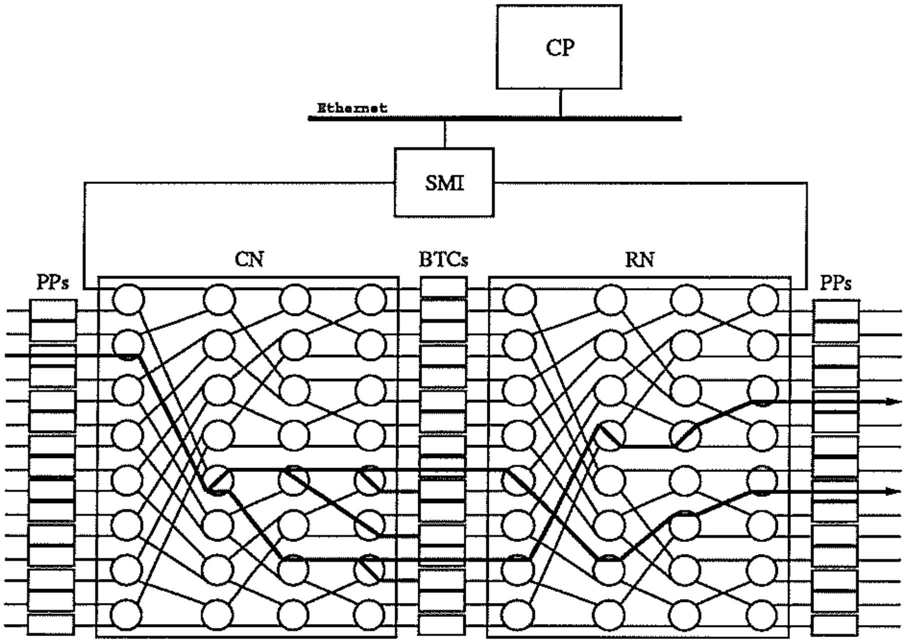


Figure 5: Phase 0 Switch Architecture

When a cell enters the system, it is reformatted by the addition of several new fields containing information needed to process a cell within the switching system. In the case of point-to-point cells, the added fields include an outgoing link number which is used to route the cell through the switch and an outgoing logical channel number. In the case of multipoint cells, they include a *Fanout* field (FAN) which specifies the number of outgoing links that must receive copies of the cell and a *Broadcast Channel Number* (BCN), which is used in a second stage address translation.

The switching network contains three major components, a *Copy Network*, a set of *Broadcast Translation Circuits* and a *Routing Network.* When a multipoint cell having *k* destinations passes through the Copy Network (CN), it is replicated so that *k* copies of that cell emerge from the CN. Point-to-point cells pass through the CN without modification. The function of the Broadcast Translation Circuits (BTC) is to assign outgoing link numbers to the copies of multipoint cells. The Routing Network (RN) delivers cells to the proper outgoing PP, based on the address information given in routing field. The topology shown in the example is a delta network. However, other topologies such as a Beneš topology may also be used.

The Copy and Routing Networks are made up of Packet Switch Elements (PSEs) that contain internal buffers capable of storing several cells. A cell may pass through a node without being buffered at all if the desired output port is available when the cell first arrives. Indeed, in a lightly loaded network, a cell can pass through the CN and RN without ever being buffered. In addition to the data path between switch elements, there is a grant signal used to implement a simple flow control mechanism. This prevents loss of cells due to buffer overflows within the fabric. The entire network is operated synchronously, both on a bit basis and a cell basis--that is, all cells entering a given stage do so during the same clock cycle.

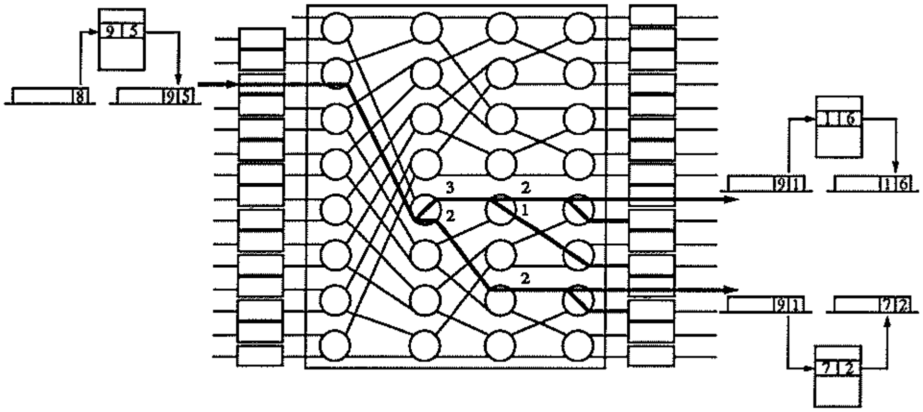


Figure 6: Copy Network Operation

The structure of the Copy Network (CN) is the same as that of the RN. The CN's function is to make copies of multipoint cells as they pass through, as illustrated in Figure 6. When a cell passes through the incoming port processor, the virtual circuit identifier in the cell's header (8 in the example) is extracted and used to perform a table lookup as illustrated in the figure. In the example, this yields a FAN field of 5 and a BCN of 9. At the first stage, the cell is routed out the lower port. This is an arbitrary decision--the upper link could have been used at this point. At the second stage, the cell is sent out on both outgoing links and the FAN fields in the outgoing cells are modified. The upper cell generates three copies and the lower one two. In general, a node in the copy network will replicate a cell if its current FAN value exceeds one-half the number of CN output ports reachable from that node. The FAN values are split as evenly as possible, with an arbitrary decision being made as to which port gets the "bigger half" in the case of an odd FAN value. Point-to-point cells are routed through the CN arbitrarily, taking the "path of least resistance."

When a broadcast cell reaches a Broadcast Translation Circuit, the BCN is used to select a new routing field from the BTC's internal table. This information is then used by the RN to guide the cell to its final destination. In the example, this routing translation is shown for two of the five copies created by the copy network. The first copy is sent to output 6 and will be assigned a logical channel number of 1 on the outgoing link. The second copy is sent to output 2 and will be assigned a logical channel number of 7.

The phase 0 switch supports external links operating at 100 Mb/s and supporting the ATM cell formats. The internal data paths are eight bits wide and the system operates with a clock rate of 25 MHz. All the crucial components are custom integrated circuits fabricated in 2 μm CMOS. There is a total of five distinct chip types, one for the PSEs, one for the BTCs and three for the PPs. The phase 0 switch has a 16 port switching network with 15 external ports. The copy and routing networks are each implemented on a single circuit board, the BTCs are packaged on another and the SMI on another. The PPs are packed four to a board, so the switch as a whole requires eight printed circuit boards.

Figure 7 shows a single switch element. Data enters on the upstream data lines (ud_A,ud_B) and leaves on the downstream data lines (dd_0,dd_1) The *Output Control Circuit* (OCC), at the center of the figure monitors the availability of the downstream neighbors via the downstream grant lines (dg_0,dg_1) and arbitrates access to the outgoing links. The *Input Control Circuits* (ICC), within each of the two *Input Port Controllers* (IPC), receive cells from their upstream neighbors, request the appropriate output link (or links) from the node con-

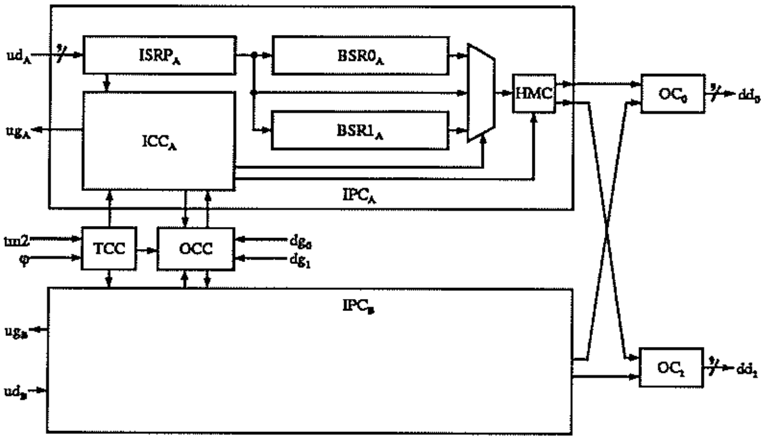


Figure 7: Switch Element

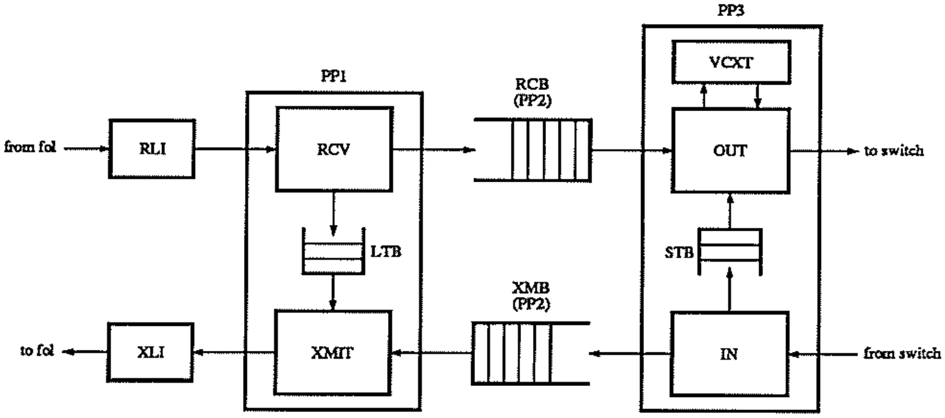


Figure 8: Port Processor Block Diagram

trol, buffer cells if necessary and apply upstream flow control using the upstream grant lines (ug_A,ug_B) as needed to prevent buffer overflow. The chip implementing the switch element has a complexity of about 45,000 transistors.

The Port Processors (PP) form the interface between the external fiber optic links and the switch module's internal data paths. They perform all the link level protocol functions, including the determination of how cells are routed. A block diagram of the PP appears in Figure 8. The major components are described briefly below.

- *Buffers.* The PP contains four cell buffers. The *Receive Buffer* (RCB) is used for cells arriving from the 2 transmission link and waiting to pass through the switch. The *Transmit Buffer* (XMB) is used for cells arriving from the switch that are to be sent to the transmission link. The *Link Test Buffer* (LTB) and *Switch Test Buffer* (STB) provide paths for test cells used to verify the operation of the link and switch respectively. The RCB and XMB each have a capacity of 32 cells. The LTB and STB can each hold a single cell. Together, the four buffers require a total of about 44 K bits of memory.

- *Receive Link Interface* (RLI). Converts the incoming transmission link signal to an eight bit electrical format, and provides a clock recovered from the incoming data stream.

- *Receive Circuit* (RCV). Synchronizes the incoming data stream to the switch's clock, checks the headers of incoming cells for errors, converts the cell to the intermediate format, routes test cells to the LTB and other cells to the RCB, reformatting these into intermediate format and adding parity.

- *Output Circuit* (OUT). Adds five bytes of header information to the front of each cell received from the RCB. Performs logical channel translation and sends cells to the switch. Also reads switch test cells and VCXT read/write cells from the STB and processes them appropriately.

- *Virtual Circuit Translation Table* (VCXT). Lookup table used to translate an incoming virtual circuit number to the routing information needed by the switch. Implemented as a random access memory.

- *Input Circuit* (IN). Routes internal data cells to the XMB, removing the first five bytes of header information and routes all other cells to the STB.

- *Transmit Circuit* (XMIT). Takes cells from the XMB, converts from intermediate to external format, strips parity, computes the header check and synchronizes the outgoing stream to the clock for the outgoing transmission system. Also processes test cells from the LTB.

- *Transmit Link Interface* (XLI). Converts from eight bit electrical format to the transmission link format.

Note that Figure 8 also shows how the components are divided among the different integrated circuits. The RCV and XMIT circuits, together with the LTB are on the PP1 chip, the RCB and XMB each consist of a PP2 chip and the IN and OUT circuits together with the STB and VCXT are placed on the PP3 chip. The RLI and XLI are implemented using commercial components.

4.2.2. Application Interfaces

In the phase zero network, there will be several application interfaces. The first is an audio/video interface for a *broadband terminal.* The broadband terminal consists of a commercial workstation equipped with a video-in-a-window card, together with an ATM audio/video interface as shown in Figure 9. The interface card takes analog audio video from any source, modulates the audio into a band above the video and mixes the two, then digitizes the resulting signal using 8 bit samples at a sampling rate of approximately 10.74 MHz, packetizes the resulting digital stream and transmits the cells on the outgoing ATM link.

On the receiving side, data is extracted from cells and placed in a synchronization memory. At this stage cells are resequenced using a sequence number inserted in the packet by the sending interface. Data is read from the synchronization memory and converted to analog. The audio and video portions are then filtered and the audio is demodulated to baseband. The audio signal is then sent to the audio input on the workstation and played out on the workstation's built-in speaker. The video is connected to the video input on the workstation's video-in-a-window card, allowing it to be displayed on the workstation screen. Data and

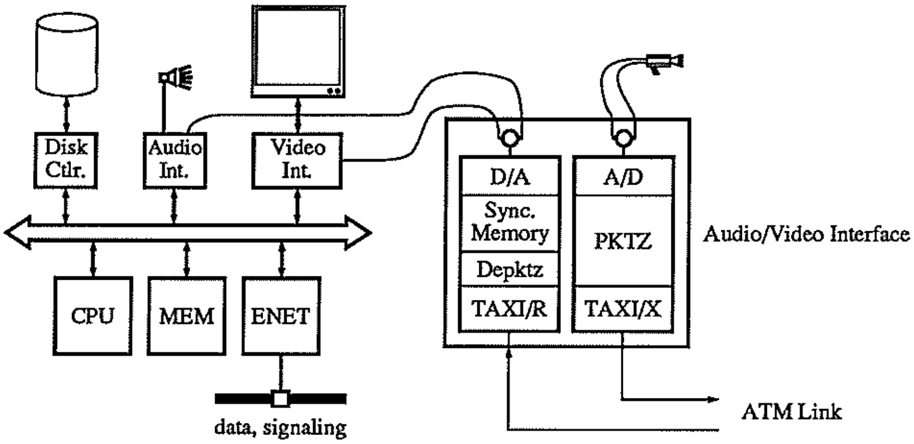


Figure 9: Broadband Terminal

signaling information will not be carried directly on the ATM link. These will pass through an Ethernet interface and be forwarded to the ATM network by an Ethernet portal.

Two versions of the physician's workstation are being supported in the phase 0 network. The first version, based on a DEC workstation and an auxiliary high resolution display, connects to the ATM network through an Ethernet interface, allowing high resolution gray-scale images to be transferred from an image server on a remote Ethernet segment, across the ATM network to the local Ethernet and from there to the high resolution display. A second version is based on a NeXT workstation with a similar supplementary display and a direct interface to the ATM network. The image server used with this workstation connects directly to the ATM network, allowing images to be sent at the full ATM rate. This configuration supports compressed video in addition to still images.

4.2.3. Internetworking

Internetworking will be supported in the phase 0 network using a specialized interface device called an *Ethernet Portal* which is illustrated in Figure 10. The portal will transfer Ethernet packets from the local Ethernet, fragment them into a series of ATM cells and reassemble the received ATM cells at a remote Ethernet. The phase 0 implementation will not include any filtering capabilities, but will simply transfer the Ethernet packets onto a preconfigured multicast virtual circuit. Received ATM cells will be reassembled and repeated onto local Ethernets. Separate, commercially available bridges or routers can be placed between the portal and the actual Ethernet if filtering is required.

As shown in Figure 10, the portal will consist of a commercial EISA bus PC class computer with an ATM interface board including a dual ported memory through which packets are transferred. Fragmentation and reassembly takes place in the memory through a cooperative process involving the Ethernet Controller, the microcontroller on the interface board and the PC's processor, which handles buffer management and makes minor header modifications to ATM cells.

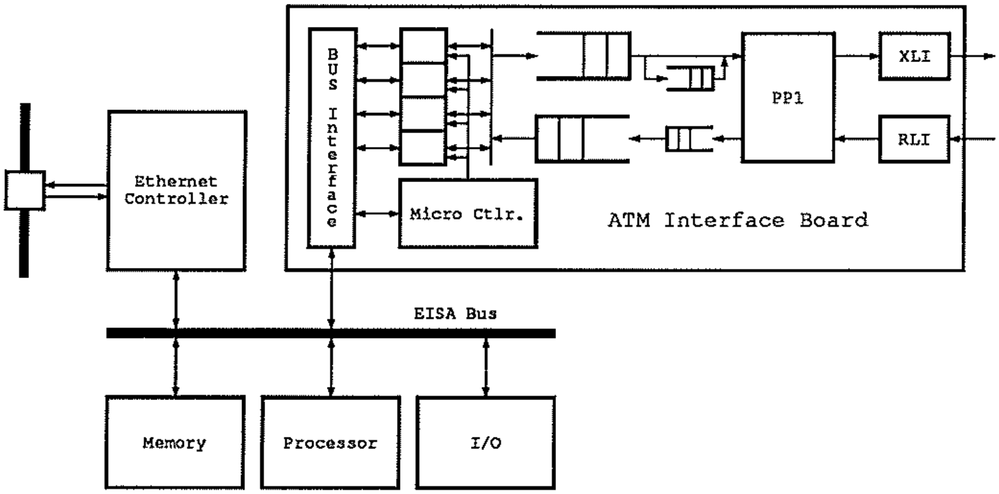


Figure 10: Ethernet Portal

FDDI access will be provided in phase 0, not by a direct interface, but using an Ethernet portal and a commercially available Ethernet to FDDI gateway.

4.3. Phase One

Phase 1 of Project Zeus is to begin at the start of 1992 and run through 1994. Its primary objective is to create the key components needed to establish an operational ATM campus network, develop a set of core applications that use the network and develop support tools to facilitate future application development. The specific goals for phase 1 are listed below.

- Develop an economical switch configuration supporting up to 128 ports at 155 Mb/s each, and an inexpensive concentrator that can be located in a wiring closet within close proximity of desktop workstations.
- Complete the multicast connection management software and add signaling interfaces to workstations. Design and implement basic network management software.
- Design and implement a workstation interface capable of using a full 155 Mb/s link and equipped with cell pacing circuitry. Include support for both coded and uncoded video.
- Design and implement a multiport ATM router that forwards IP datagrams across multiple fixed virtual circuits and provides Ethernet and FDDI connectivity.
- Establish basic interoperability between the campus network and a public network ATM switch in order to support connections to off-campus sites. This will require development of an ATM/SONET interface to the public network.

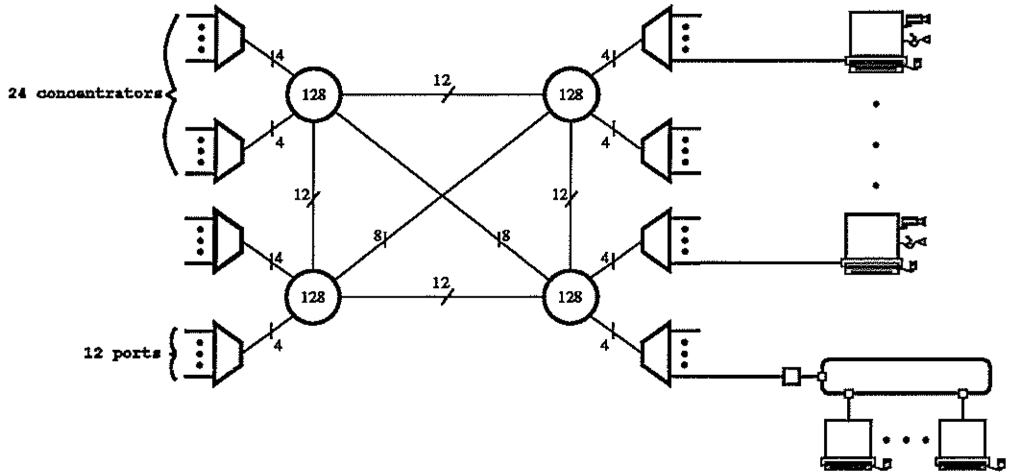


Figure 11: Phase 1 Benchmark Network

Figure 11 shows a network configuration that is intended to be typical of configurations that the phase 1 network components should be able to support. The benchmark network comprises four switching systems, each with 128 ports operating at speeds of 155 Mb/s each. The switching systems connect to users through *concentrators*, where each concentrator has 12 ports and four links to its host system, giving a concentration ratio of 3:1 (that is, each user port has access to 50 Mb/s of switch bandwidth on average). The configuration shown can support over 1000 user ports.

The connections between switches and between switches and concentrators use optical fibers, where four channels are multiplexed onto a single fiber, so just one fiber connects each concentrator to its switch. The connections between concentrators and user devices will use shielded twisted pair wiring. These two steps together drastically reduce the number of optical devices needed to construct the system, thereby reducing the system's cost and improving its reliability.

The system is controlled through a *Control Processor* (CP) associated with every switch. The CPs can exchange signaling messages with user devices and with one another in order to establish or modify virtual circuits. The CPs can also access control tables within their switches and the associated concentrators to configure user connections. The ability to control the concentrators remotely again helps to reduce costs, allowing a single CP to be shared by up 288 users.

Connections to workstations will be provided either through LANs or through direct ATM interfaces. The ATM interface will include the ability to carry a single video channel, as well as high speed data.

The system will provide interfaces to local area networks including Ethernet and FDDI. These can be implemented by adding an ATM interface card to an existing commercial router, that supports Ethernet and FDDI. These routers can be connected via one or more multicast virtual circuits. Workstations that are directly connected to the ATM network can be connected to a router via a virtual circuit, in order to provide datagram forwarding. The standard software will be augmented with a connection-oriented internet protocol.

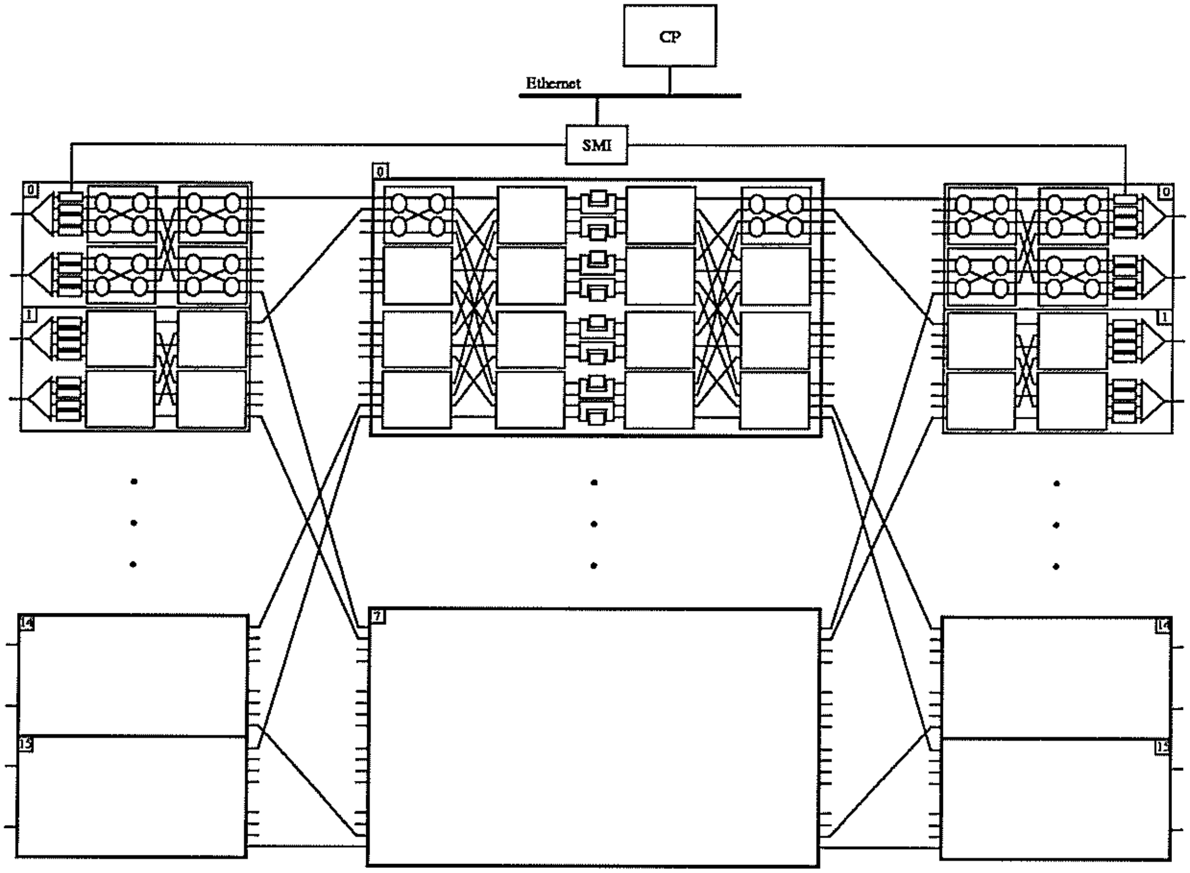


Figure 12: Phase 1 Switching System

Bandwidth management and congestion control in the phase 1 network will be based on a dynamic buffer allocation mechanism to be implemented in the phase 1 switches. This mechanisms allocates buffers to contending bursts on established virtual circuits, ensuring that during congestion, the network delivers complete bursts rather than dropping cells arbitrarily from different bursts.

4.3.1. Phase 1 Switching System

The switching system for the phase 1 network is shown in Figure 12. The configuration shown would support 128 links. As in the phase 0 switch, the system consists of a set of *Packet Processors* (PPs), a *Copy Network* (CN), set of *Broadcast Translation Circuits* (BTC) and a *Routing Network* (RN).

The copy and routing networks are each constructed from binary switch elements similar to those used in the phase 0 system, but these are packaged four per integrated circuit. The networks are also split between two sets of circuit boards. The PP cards (shown on both the left and right) each contain an eight port subsection of the copy and routing networks, as well as eight PPs and two multiplexor/demultiplexor pairs.

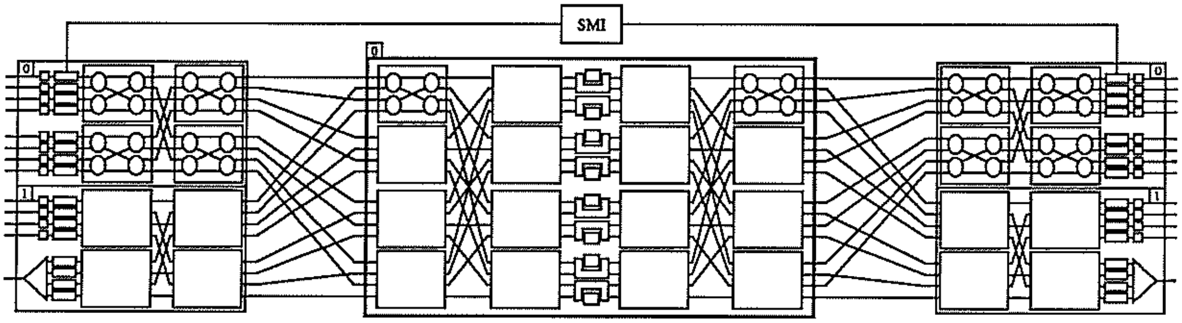


Figure 13: Phase 1 Concentrator

The network cards (in the center) each contain a 16 port subsection of the copy and routing networks plus 16 BTCs.

The PPs will be constructed as a chip set including two or three commercial memory chips and one or two application-specific integrated circuits implementing a custom controller. The PP controller will implement a resequencing buffer which will reorder cells received from the switch so that they are transmitted on the downstream link in the proper order and a fast buffer reservation mechanism to allocate buffers on a per burst basis. The BTCs will be constructed from a two chip set, one a commercial memory chip, and the other a custom BTC controller chip supporting either two or four channels.

The primary transmission interface to the system will be a 620 Mb/s optical interface consisting of four multiplexed streams at 155 Mb/s each. While the rates are chosen to be compatible with the SONET standard for optical transmission in public networks, the coding scheme used within the campus will be a variation of the 4/5 coding scheme used in the physical layer of FDDI. This approach leads to a much simpler interface design and allows the use of relatively inexpensive commercial transmission chips. Nonmultiplexed interfaces to PPs are also possible of course, although we would expect such interfaces to be relatively rare.

With the packaging described, the system will support an n port configuration with $1+3n/16$ circuit cards. For example, the 128 port system shown in Figure 12 would require 25 circuit cards, while a 64 port system would require 13.

Figure 13 shows the configuration of the phase 1 concentrator. Notice that the concentrator uses the same components as the switching system, but is configured with just one network board and two PP boards. Access to the associated switch is through the multiplexed interface shown at the bottom left and right, while connections to user devices are provided by the remaining 12 nonmultiplexed interfaces. The concentrator includes complete copy and routing networks allowing local switching functions to be performed. So in particular, multicast connections can branch at the concentrator, fanning out from a single input to multiple outputs.

The concentrator contains no local CP. Control is provided through signaling messages received from the associated switch on the multiplexed interface. These messages are forwarded to the SMI, where the local microprocessor interprets them and carries out the requested actions, typically sending a message to one of the PPs or the BTCs in order to establish or modify a user connection.

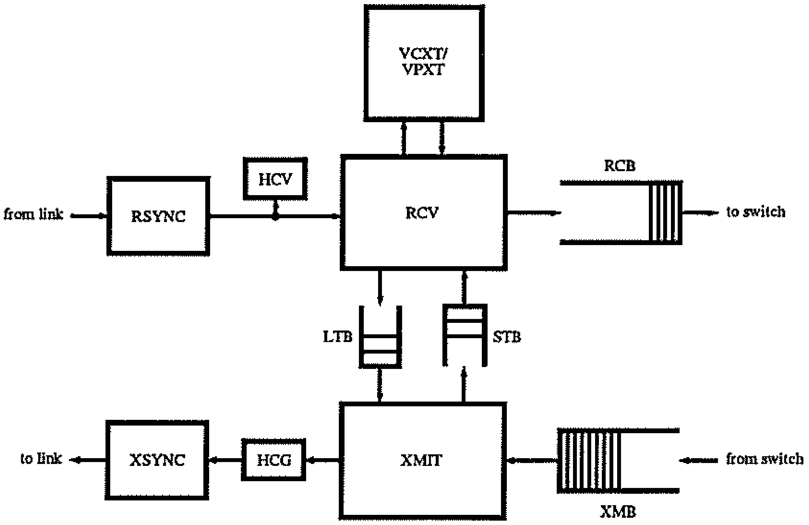


Figure 14: Phase 1 Port Processor

Figure 14 shows the structure of the port processor for the phase 1 switch. Cells coming from the link enter at the top left and pass through a *Receive Synchronizer* (RSYNC), are checked by a *Header Check Verification circuit* (HCV), are reformatted and routed by a *Receive circuit* (RCV) which performs virtual circuit and virtual path translations, and then pass to a *Receive Buffer* (RCV) before going on to the switch. Cells coming from the switch enter a *Transmit Buffer* (XMB), a *Transmit circuit* (XMIT) that does some simple reformatting, a *Header Check Generator* (HCG) and a *Transmit Synchronizer* (XSYNC). The synchronizers allow the switch to support any link speed up to 155 Mb/s. In particular, direct 45 Mb/s interfaces are possible.

Cells can be looped back to the link through the *Link Test Buffer* (LTB) or the switch through the *Switch Test Buffer* (STB) to verify system operation. The STB is also used to provide a control path from the CP to the *Virtual Circuit/Path Translation Table* (VCXT/VPXT).

The (VCXT/VPXT) tables will be implemented using an external static RAM. These will be divided between a VCXT table and a VPXT table. When an incoming cell is received, its VPI is used to extract an entry from the VPXT. A bit in the VPXT entry determines whether or not VCI translation is also required. If VCI translation is not required, the outgoing link and outgoing VPI are obtained from the VPXT entry and the VCI is passed through without modification.

The RCB and XMB will be implemented using random access memory and a buffer controller. Since cells are routed independently through the switch, it is possible for them to exit in a different order than that in which they entered. Consequently, the buffer controller for the XMB includes circuitry to resequence cells following reception from the switch. This resequencing mechanism is described in detail in reference [42].

The XMB controller also includes circuitry to allocate buffer space in the XMB to virtual circuits when the start of a burst is detected. Moreover, it can distinguish high and low priority cells, allowing the buffer space occupied by low priority cells to be preempted if necessary to accommodate an arriving high priority

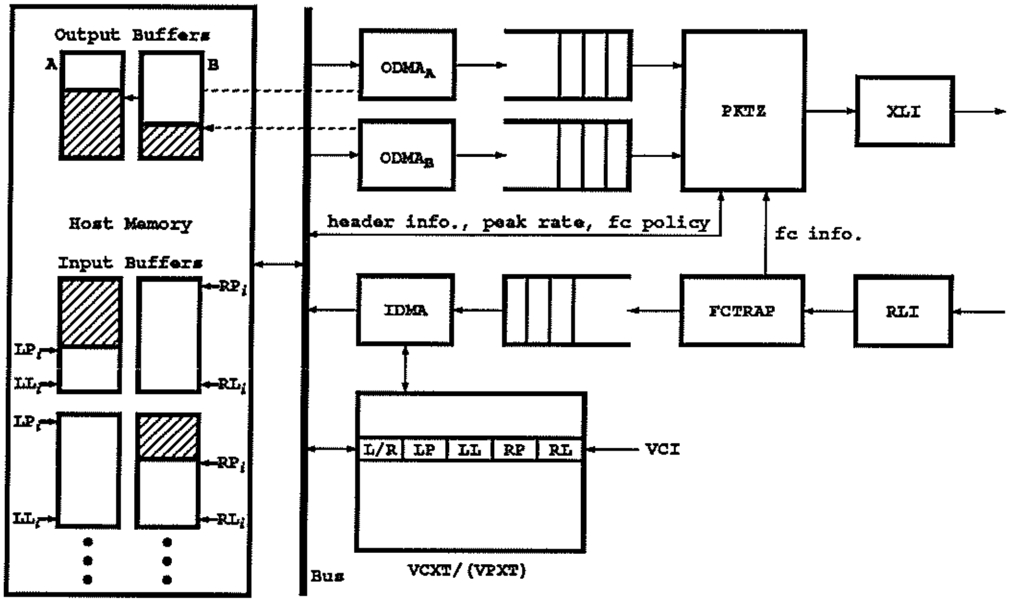


Figure 15: ATMizer for Phase 1 Network

cell. The receive circuit in the port processors at the user-network interface include *Access Resource Management* circuitry to monitor the peak and average bandwidth consumed by each active virtual circuit. More details on the resource management mechanisms are given in reference [43].

4.3.2. Application Interfaces

We plan to implement a rich set of application interfaces in the phase 1 network, allowing for a wide range of user applications. First, the phase 0 broadband terminal will be upgraded to support the higher link speeds of the phase 1 network. The NeXT-based physician's workstation and the medical image server can also be used in the phase 1 network, as they have been designed to operate at up to 155 Mb/s.

We plan to design an ATM interface to a commercial workstation that would directly support the bandwidth management and congestion control mechanisms provided by the phase 1 network. A possible design for this interface, called an ATMizer, is shown in Figure 15. The ATMizer can be programmed by the host software to concurrently transfer data from two memory-resident buffers to two different virtual circuits. The ATMizer transfers data from the buffers to a pair of on-board fifos, adds the ATM cell header information and paces the transmission of cells to the network, so as not to exceed the peak transmission rate specified by the host software. On reception, ATM cells are transferred into one of many memory-resident buffers, using an on-board virtual circuit translation table to identify the appropriate buffer and the current position in the buffer. Flow control cells received from the network are passed to the output channel, which typically suspends transmission of the appropriate virtual circuit and interrupts the workstation's CPU. At this point, the workstation would typically reprogram the output channel to transmit data belonging to a different virtual circuit.

We are planning to support medical imaging in the phase 1 network using a high resolution film scanner and laser printer combination, to allow x-rays acquired at a remote site to be transferred at high speed and printed at a central medical center. This will require ATMizers for both the scanner and the printer.

4.3.3. Internetworking

In phase 1, we plan to expand the internetworking capabilities of the Project Zeus network. First, the Ethernet portal developed in phase 0, will be modified to support the higher link speeds required for the phase 1 network and the software will be augmented so that it can function as an IP router, with the capability of sending and receiving fragmented IP datagrams across multiple virtual circuits and forwarding them to the appropriate destination. We also plan to implement a connection-oriented internet protocol, known as MCHIP, in the Ethernet router [23, 24]. MCHIP can provide consistent performance to connection-oriented applications using statistical resource allocation and supports multipoint connections at the internet level.

In addition, we plan to work with a commercial router company to design and implement an ATM interface to an existing router product with Ethernet and FDDI interfaces, allowing the router to forward datagrams over ATM virtual circuits. In addition, we plan to port the MCHIP software to the router, providing virtual circuit connectivity to hosts resident on Ethernet of FDDI.

Another crucial aspect of internetworking in the phase 1 network will be the connection of the phase 1 switch to a public ATM network. We plan to implement an SONET interface to the phase 1 network to allow virtual circuits to be established across the public network.

4.4. Phase Two

In phase 2 of Project Zeus, we plan to expand the range of interfaces that can be used to access the network, construct components for larger scale and higher speed networks and reduce the cost of key network components. In addition, we plan to extend access to the networks to departments all across the university. The specific goals of phase 2 are listed below.

- Develop a core switch supporting 256 ports at 620 Mb/s each, together with an inexpensive concentrator with 16-48 user ports. Improve the functionality and performance of the connection management software.
- Design and implement compatible ATM interfaces for multiple workstations, including support for HDTV video and multiple coded video channels.
- Extend interoperability with public network switches to include signaling, as well as basic cell transfer. Add 620 Mb/s SONET interfaces to the campus network.
- Design and implement internet processors capable of processing fragmented packets without reassembly. These would support both IP datagrams and MCHIP packets.

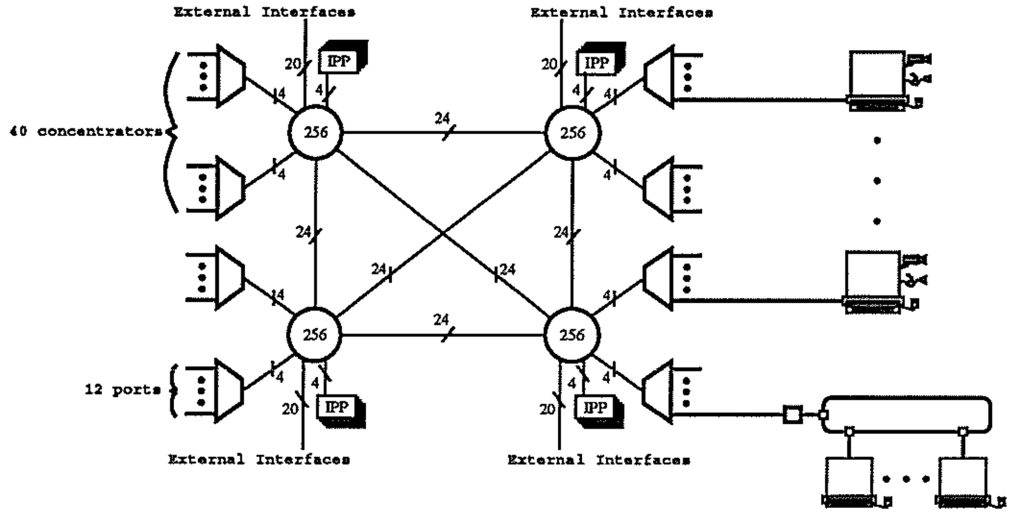


Figure 16: Phase 2 Benchmark Network

Figure 16 shows a typical network configuration that the phase 2 components should be able to support. The benchmark network comprises four switching systems, each with 256 ports operating at speeds of 620 Mb/s each. The switching systems connect to users through *concentrators*, where each concentrator has up to 48 ports at 155 Mb/s each plus four links to its host system, giving a concentration ratio of 3:1 (that is, each user port has access to 50 Mb/s of switch bandwidth on average). The configuration shown can support over 7600 user ports. The system is controlled through a Control Processor (CP) associated with every switch as in the phase 1 network.

The system will provide interfaces to multimedia workstations supporting multiple channels of standard video and in some cases high definition video. These will be carried in coded form with standard video probably coded at a rate of about 16 Mb/s, and high definition video at about 100 Mb/s. *Internet Packet Processors* (IPP) will be provided at each switch to support processing of fragmented datagrams and MCHIP packets. The phase 2 network will also provide a wider range of back-end interface options, including SONET interfaces at 620 Mb/s.

4.4.1. Phase 2 Switch

The overall organization of the phase 2 switch is shown in Figure 17. As in the original architecture there is a copy network and a routing network, and a set of intermediate broadcast translation circuits. The networks are constructed using a three-stage Beneš topology and 16-port switch elements. One stage provides a distribution function to balance traffic loads within the system and can be dropped if the performance improvements don't merit the added cost.

The switch elements support byte wide data paths and if constructed using 1.2 μm CMOS integrated circuits, can be operated with a clock speed of 100 MHz, giving a data path speed of 800 Mb/s, which due to the shared buffering, is fast enough to support external link speeds of 620 Mb/s. The switching elements

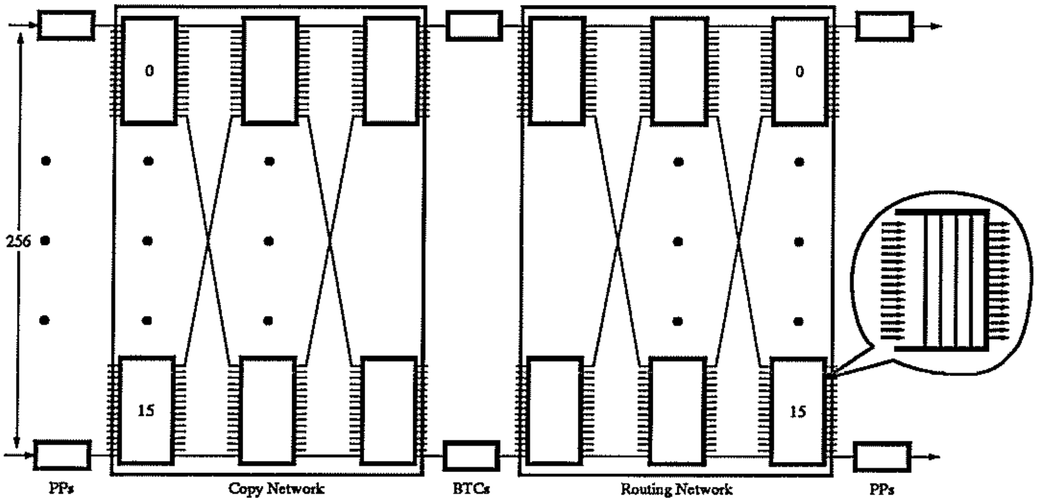


Figure 17: Phase 2 Switch Organization

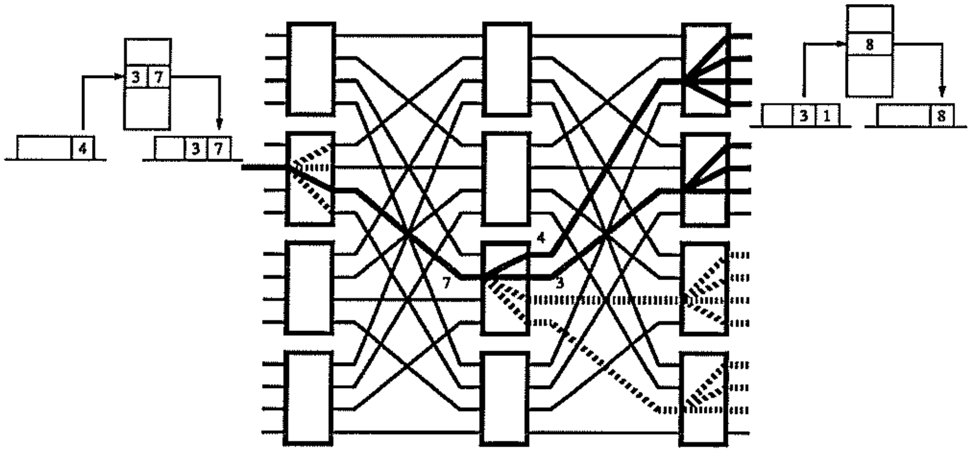


Figure 18: Broadcast in Phase 2 Switch

implement a shared buffer structure, where the buffers within a switch can be used for cells coming from any input link. As in the phase 1 switch, a simple hardware flow control mechanism stops the flow of cells between adjacent switches when there is no room in the buffer.

Figure 18 shows how broadcast is handled in the copy network. When the cell is received at the incoming port processor, a virtual circuit translation is performed, yielding a fanout and broadcast channel number, as in the phase 1 network. In the first stage of the network, the cell is routed to a port selected to distribute the load most evenly. In the second and third stages, copying takes place and the fanouts of the copies are modified. Notice that in the second stage, only two copies are made; in general copying is delayed as long as possible and the smallest possible number of copies is produced. The lower two ports could have been selected to receive the two copies. The choice is made based on load considerations. Broadcast translation is done in much the same way as in the phase 1 network. The one exception is that the Broadcast

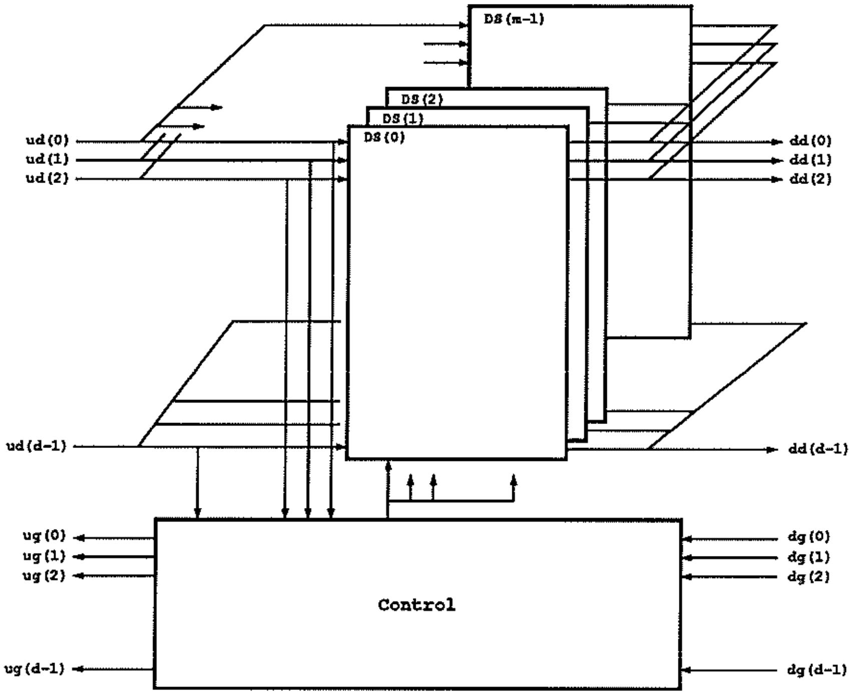


Figure 19: Bit-Sliced Switch

Channel Numbers (BCN) are divided into two types, those used for *small fanout connections* (fanout ≤ 16) and those used for *large fanout connections*. The small fanout BCNs can be used for up to 16 different user connections, making it possible to support over 32K multicast connections using BTCs with 4096 entry tables.

The switches used to construct the copy and routing networks use a bit-sliced structure to achieve the lowest possible complexity. Figure 19 shows the design of a bit-sliced switch with *d* input and output ports and supporting *m* bit wide data paths. In the phase 2 switch *d*=16 and *m*=9. Cells enter on one of the *d* *upstream data* lines (ud_i) at left, and the *m* bits of each cell are distributed across *m* separate *data slices* (DS). The cells exit from the switch element on the *downstream data* lines (dd_i) at the right. The switch element contains sufficient internal buffering to store several cells for each port and implements a simple hardware flow control mechanism to prevent cells from overflowing these buffers.

The *control slice* shown at the bottom of the figure contains the circuitry used to control the operation of the switch. It receives a set of *downstream grant* signals (dg_i) from the downstream neighbors and generates a corresponding set of *upstream grant* signals (ug_i) which are sent to the upstream neighbors. In general, a switch element asserts an upstream grant signal ug_i if it is prepared to receive a cell on the upstream data lines ud_i. The cells flowing through the switch element are organized so that all the control information (in particular, the addressing information) passes through the first data slice DS_0. This allows the control circuit to easily monitor the control information for all cells entering the data slice. Using this information, together with the downstream grants and the internal status of the switch elements, it makes control decisions and broadcasts those decisions to the data slices. In addition, the first bit of the cell in *every data slice* is a control bit indicating the presence or absence of a cell.

Preliminary studies of the bit-sliced switch element indicate that two data slices can be packaged in a single integrated circuit (using 1.2 μm CMOS) as can a single control slice. This allows the copy and routing networks to be implemented using just 440 chips. We anticipate that an n port phase 2 switch can be implemented with $1+7n/32$ boards, slightly more than for the phase 1 switch, but providing four times the throughput per port. So, for example, 57 boards are needed for a 256 port system and 15 for a 64 port system.

One or more concentrators will also be needed for the phase two network. We anticipate a concentrator built around a core that supports 16 ports at 620 Mb/s. While some of these ports would be used directly for external connections, we expect that most of the 620 Mb/s ports will be subdivided into four 155 Mb/s ports. A likely configuration for the concentrator is 32 ports at 155 Mb/s each and eight ports at 620 Mb/s, some of which would be used to connect to a central switch and others for providing high speed access to users.

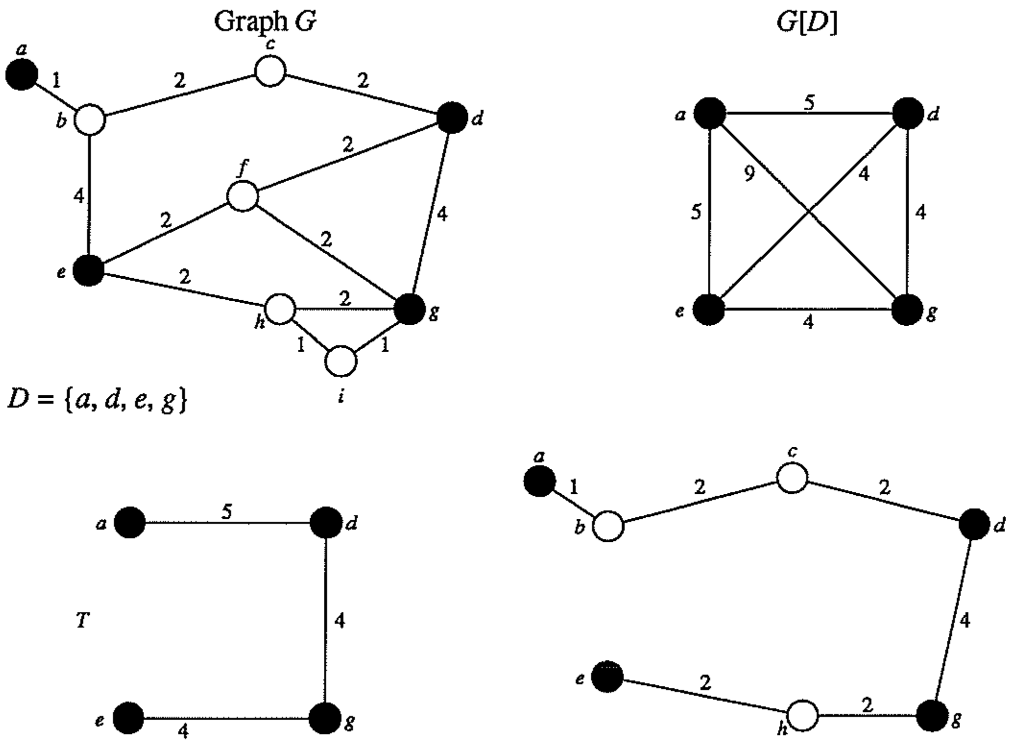
5. Routing and Resource Management

5.1. Multicast Routing

In a packet-switched network which uses virtual circuits, the primary goal in routing connections is to make efficient use of the network resources. For example, we favor an algorithm which can handle the largest number of connections for a given set of network resources. In a point-to-point network, routing is often treated as a shortest path problem in a graph. Here the network is modeled as a graph $G = (V,E)$, where the nodes of a graph represent switches and the edges represent links. In addition, we have two functions CAP: $E \rightarrow \Re^+$ and COST: $E \rightarrow \Re^+$ which give us the bandwidth and cost of each edge (link). In this model we equate cost and edge length. At the time a connection is established, a shortest path with sufficient available bandwidth connecting the pair of endpoints is selected.

Routing of multipoint connections may be modeled in a similar way. In the multipoint problem we wish to connect a set $D \subset V$. Instead of the shortest path, one is interested in the shortest subtree which contains the set D. Finding the shortest subtree connecting a set of points is a classical problem in graph theory known as the Steiner tree problem in graphs. This problem has been shown to be NP-complete by Karp. Consequently, one is forced to consider approximation algorithms which are not guaranteed to produce optimal solutions.

There are several polynomial-time approximation algorithms for solving the Steiner tree problem which we have used as a starting point for work on multipoint routing. The *minimum spanning tree heuristic* (MST) produces solutions whose costs are never worse than twice that of an optimal solution. Our experimental evaluations of MST indicate that it typically yields solutions that are within five percent of optimal. Figure 20 illustrates an example of the application of MST. Here we are asked to connect the set of four nodes $D = \{a, d, e, g\}$. The first step of the algorithm involves constructing a derived graph $G[D]$. This graph is a complete graph on the four nodes in D, where the length of each edge corresponds to the length of the shortest path in the original graph G. The second step involves finding a minimum spanning tree for $G[D]$. This can be done using one of several polynomial time algorithms. Finally the edges of the minimum spanning tree for $G[D]$ are mapped back to paths in the original graph, taking advantage of path overlap. Note that the solution here has cost two units more than optimal.



A minimum spanning tree for $G[D]$ Basic MST solution, cost = 13

Figure 20: An Example of the Application of MST

We have studied a more sophisticated algorithm for the Steiner tree problem known as Rayward-Smith's algorithm. We have shown that it produces solutions that are no worse than twice optimal, and surprisingly, that it also can produce solutions that are that bad. We have devised a generalization of Rayward-Smith's algorithm which we conjecture produces solutions that can be arbitrarily close to optimal, at the cost of increased, but still polynomial running time. We have also developed iterative versions of MST and Rayward-Smith's algorithm which have the same worst-case performance as the ordinary versions but perform slightly better in practice.

MST and Rayward-Smith's algorithms are mainly useful as a standard of comparison against which to measure other routing algorithms, since they require global information about the network state that is not generally available. Our plan in Project Zeus is to use a much simpler incremental algorithm that works in the following way. When adding a new endpoint to an existing multicast connection, we start from the new endpoint and select the shortest available route that leads to the owner of the target connection. Since the owner's address is embedded in the connection identifier, this information is always available. The search stops at the first switch along the path through which the target connection passes. This approach then reduces the multicast routing problem to a point-to-point routing problem, allowing us to take full advantage of prior work on point-to-point routing. Simulation studies of this and similar algorithms have shown that the resulting connections are generally within about 20-30% of the optimal connection. This appears to be adequate in the campus network environment and its essential simplicity makes up for its lack of optimality.

See references [20, 21, 45, 46, 47, 48, 49] for further details.

5.2. Bandwidth Management and Congestion Control

A central objective in ATM networks is to provide virtual circuits that offer consistent performance in the presence of stochastically varying loads on the network. This objective can be achieved in principle, by requiring that users specify their expected traffic characteristics when a virtual circuit is established, so that the network can select a route that is compatible with the specified traffic and allocate resources as needed. While this does introduce the possibility that a particular virtual circuit will be blocked or delayed, it allows established virtual circuits to receive consistent performance as long as they remain active.

Ideally, a bandwidth management and congestion control mechanism should satisfy several competing objectives. First, it should provide consistent performance to those applications that require it, regardless of the other virtual circuits with which a given virtual circuit may be multiplexed. Second, it should allow high network throughputs even in the presence of bursty traffic streams. Third, the specification of traffic characteristics should be simple enough that users can develop an intuitive understanding of the specifications and flexible enough that inaccurate specifications don't have seriously negative effects on the user. Fourth, it should not artificially constrain the characteristics of user traffic streams; the need for flexibility in ATM networks makes it highly desirable that traffic streams be characterized parametrically, rather than by attempting to fit them into a pre-defined set of traffic classes. Fifth, it must admit a simple realization for reasons of economy and reliability. Less crucial, but in our view, also important, is the requirement that the bandwidth management mechanism accommodate multicast virtual circuits with multiple transmitters. All proposals we have seen for connection management in ATM networks have serious deficiencies with respect to at least one of these objectives. We describe here an approach using fast buffer reservation which appears to satisfy them all.

To preserve the integrity of user information bursts, the network must detect and track activity on different virtual circuits. This is accomplished by associating a state machine with two states with each virtual circuit passing through a given link buffer. The two states are *idle* and *active*. When a given virtual circuit is active, it is allocated a prespecified number of buffer slots in the link buffer and it is guaranteed access to those buffer slots until it becomes inactive, which is signaled by a transition to the idle state. While in the active state, the virtual circuit is guaranteed access to its specified number of buffer slots. If it attempts to use more than its allocated number, the extra cells are marked as excess and placed in the buffer. The excess cells are treated as lower priority and may be discarded in the presence of congestion. Transitions between the active and idle states occur upon reception of user cells marked as either *start-of-burst* or *end-of-burst*. Other cell types include *middle-of-burst* and *loner*, the latter designating a low-priority cell that is to be passed if there are unused buffer slots available, but which can be discarded if necessary. A forced transition from active to idle is also made if no cell is received on the virtual circuit within a fixed timeout period.

Figure 21 illustrates the buffer reservation mechanism. For virtual circuit i, the mechanism stores the number of buffer slots needed when the virtual circuit is active (B_i), the number of buffer slots used by unmarked (non-excess) cells (b_i) and a state variable (s_i: idle, active). The mechanism also keeps track of the number of unallocated slots in the buffer (B). The detailed operation of the state machine for virtual circuit i is outlined below.

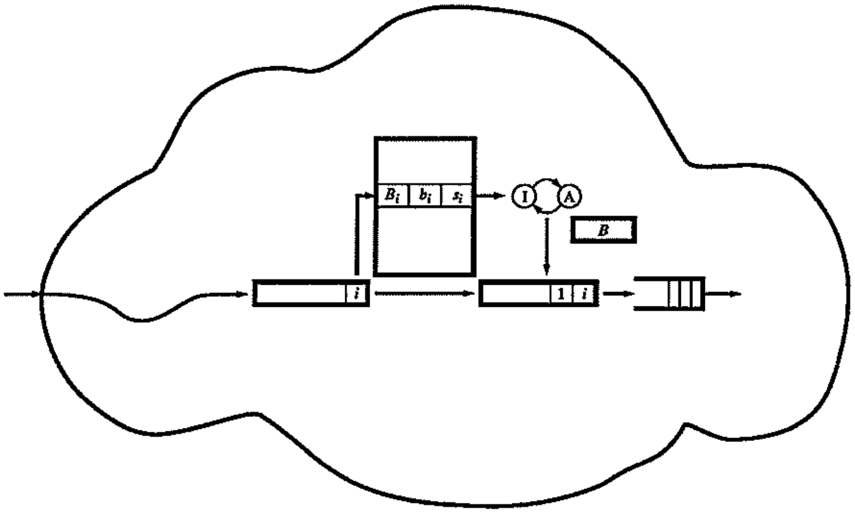


Figure 21: Fast Buffer Reservation

When a start cell is received:

- If the virtual circuit is in the idle state and $B - B_i < 0$, the cell is discarded.
- If the virtual circuit is in the idle state and $B - B_i \geq 0$, s_i is changed to active, a timer for that virtual circuit is set and B_i is subtracted from B. If $b_i < B_i$, b_i is incremented and the cell is placed (unmarked) in the buffer. If $b_i = B_i$, the cell is marked and placed in the buffer.

If a start or middle cell is received while the virtual circuit is in the active state, it is queued and the timer is reset. If upon reception, $b_i = B_i$, the cell is marked; otherwise, it is left unmarked and b_i is incremented.

If a middle or end cell is received while the virtual circuit is in the idle state, it is discarded.

If an end cell is received while the virtual circuit is active or if the timer expires, s_i is changed from active to idle and B_i is subtracted from B.

If a loner is received, it is marked and placed in the buffer.

Whenever a cell is sent from the buffer, the appropriate B_i is decremented (assuming the transmitted cell was unmarked).

When a virtual circuit is routed, the software that makes the routing decisions attempts to ensure that there is only a small probability that the instantaneous demand for buffer slots exceeds the buffer's capacity. This probability is called the *excess buffer* demand probability and might typically be limited to say, 1%.

To make this precise, let λ_i denote the peak data rate of a given virtual circuit and let μ_i denote the average rate. If the link rate is R and the buffer has L buffer slots, the number of slots needed by an *active source* with peak rate λ_i is defined to be $B_i = \lceil L\lambda_i/R \rceil$.

Since B_i buffers are allocated to a virtual circuit when it is active, the virtual circuit's instantaneous buffer requirement is either 0 or B_i. If we let x_i be a random variable representing the number of buffer slots needed by virtual circuit i at a random instant, then $\Pr(x_i = B_i) = \mu_i/\lambda_i$ and $\Pr(x_i = 0) = 1 - \mu_i/\lambda_i$.

Consider then a link carrying n virtual circuits with instantaneous buffer demands $x_1,...,x_n$. Define $X = \sum_{i=1}^{n} x_i$. Note that X represents the total buffer demand by all the virtual circuits. Suppose we have a new virtual circuit with buffer demand x_{n+1} and we want to decide if it can be safely added to the link. We first must compute, the probability distribution of the random variable $X = X + x_{n+1}$. This can be obtained by numerical convolution of the distribution of X with the distribution of x_{n+1}, assuming that the idle, active behavior of the new virtual circuit is independent of the existing virtual circuits. To decide if the virtual circuit can be accepted we then simply verify that $\Pr\{X' > L\}$ is small. For a link with a 256 slot buffer, we estimate that about 2000 multiplications and 1250 additions are required to compute the distribution of X' and $\Pr\{X' > L]$. Using fixed point arithmetic, this can be done in less than half a millisecond on a 10 MIPS processor.

To complete the bandwidth management scheme, a traffic monitoring mechanism is required at the user-network interface to ensure that the virtual circuit's long term average data rate does not exceed the value specified at virtual circuit establishment. We have designed an appropriate mechanism of this sort, and generalized it to support multicast virtual circuits with multiple sources. We have studied the implementation complexity of this approach and estimate that the incremental cost of adding bandwidth management hardware to a port controller of an ATM switch to be no more than 10%.

See [42] for further details.

6. Connections, Calls and Protocols

The Project Zeus network is a connection-oriented communication network supporting both point-to-point connections, linking two clients, and multipoint connections, linking three or more clients. Connections can be unidirectional, where data flows from one client to one or more other clients, or bidirectional, where all data transmitted by each client is received by all others. Connections are created and destroyed in response to requests issued by clients.

To help motivate the detailed description of the network call model presented in Section 6.1, an example bidirectional point-to-point connection linking clients A and D is presented here (see Figure 22). The circles labeled $N1$, $N2$,... represent switching systems, which we refer to as nodes. In this example, the connection flows through nodes $N1$, $N3$ and $N4$. The connection is bidirectional, meaning that both A and D can transmit and receive on the same connection. Certain attributes are associated with each connection. The identifier <A,1> in the example distinguishes the connection from all others in the network and provides a means whereby other clients can request to be added to this connection. The bandwidth attribute governs the maximum rate at which clients can transmit data and is used by the network to restrict the maximum data flow over internal links so that overflows can be avoided. In this example, the bandwidth is limited to 10 Mb/s (the actual bandwidth specification used in the network is more complicated than this, as we will elaborate on in Section 6.1). The permission attributes specify the default receive/transmit rights given to

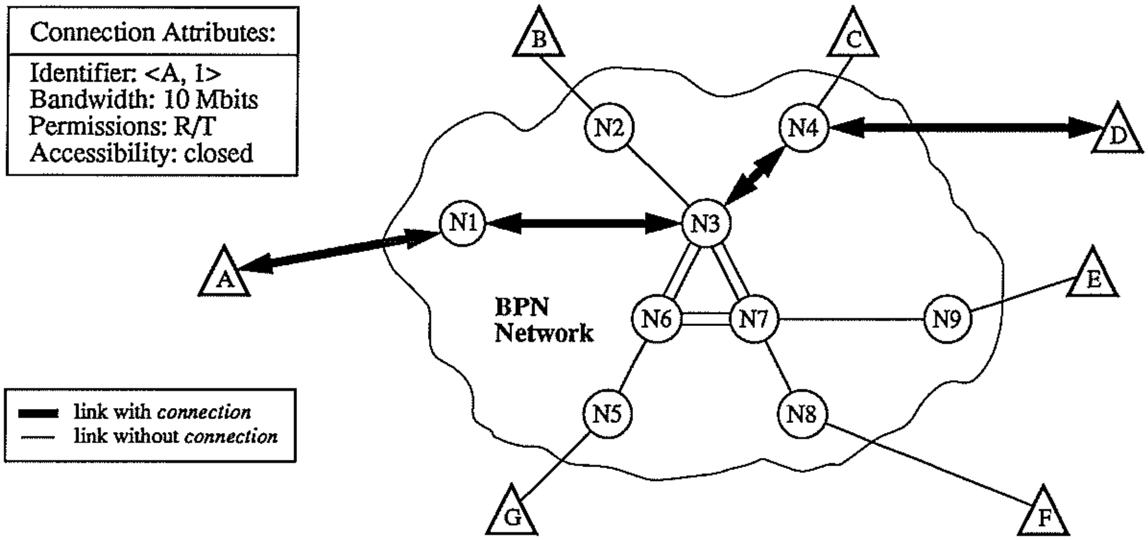


Figure 22: A Point-to-point Connection

other clients who join the connection. Here, the permissions indicate both receive and transmit, meaning that if a new endpoint is added to the connection, it will be allowed to transmit and receive, unless this default permission is overridden. The last attribute, accessibility, governs whether other clients are allowed to freely join the connection. In this case, the accessibility is closed, meaning that other clients are only allowed to join if invited by the owner. A point-to-point connection with these characteristics would be appropriate for data exchange or simple telephone calls.

A second example showing a unidirectional multipoint connection, linking client *A* to clients *B*, *D*, *E* and *F*, is diagramed in Figure 23. All data sent from *A* is replicated at nodes *N*3 and *N*7 and forwarded to the other clients. A connection such as this might be used for broadcast video distribution, where *A* is the supplier of the video feed and the other clients are customers of *A*. The attributes shown reflect the different intended use of this connection from that of Figure 22. The bandwidth is set to 50 Mb/s, suitable for higher rate video transmissions. The default permission is receive only so that other clients cannot interrupt broadcast source *A* with their own transmissions. Finally, the accessibility is open so that any client can freely join the connection. A connection such as that shown in Figure 23 could be used to support a private video conference call by changing the permission to receive/transmit and changing the accessibility to closed.

6.1. Connection Management and the Network Call Model

With the above two examples indicating the general idea of a connection, let us proceed to more careful description of connection management, a collection of algorithms, data structures and protocols used to create, maintain and destroy connections among network clients. An integral part of connection management is the call model, which describes the network view presented to clients of the network, that is, the network's outward functionality and expected behavior in response to client requests. In our model,

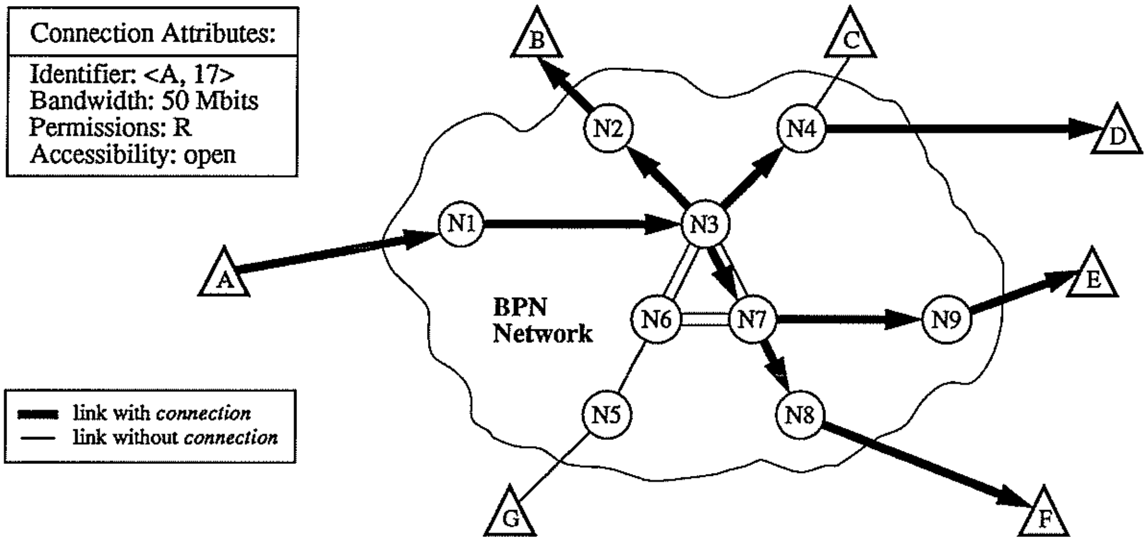


Figure 23: A Multipoint Connection

a *call* is a collection of one or more connections between two or more clients of the network. A *multipoint call* is a call involving two or more clients; a *point-to-point call* is a special case of a multipoint call involving only two clients. Data sent over a connection by one participant in a call is received by all other participants electing to receive on that connection.

When a call is created, a single connection is created between the network and the client who created the call. This client is designated the owner of the call. Additional clients, or endpoints, are added to the call by:

- Invitation from the owner, where the invited party has the option of refusing the invitation (used, for example, to initiate a data connection or a telephone call).
- Request from a client not currently in the call to be added, where the owner has the option of denying the request (used, for example, to connect to a broadcast video source).
- Request from a third party, not necessarily in the call, to add a client, where both the owner and the client being added have the option to refuse (used, for example, when a member of a conference call other than the owner wants to add another party).

In these ways, a call is allowed to grow to include any number of participants.

Once a call has been created, additional connections can be added to the call as well. Multiple connections might be grouped into a single call, for example, to separate the video and audio feeds in a conference call. Each connection of a call has a bandwidth specification and a receive/transmit permission. The connection bandwidth is described by three parameters: peak, average and burst length. The peak bandwidth is the maximum transmission rate that any client can use when transmitting. The average bandwidth is the maximum long term average transmission rate of all clients' combined transmissions. The burst length is the maximum period of time that a given client can transmit continuously. If the bandwidth of a single client transmission or of all client transmissions combined exceeds the connection's bandwidth specification, client transmissions may be discarded by the network. The bandwidth specification is given when connections are created and is used by the network to decide whether to accept the connection or not (that is, whether the requested bandwidth can be supported using the available network resources). This helps the network avoid internal overloads by restricting the total expected bandwidth allocated on each link to less than the link's capacity. Requests that would exceed the capacity of any link traversed by the connection are denied.

As indicated above, each connection has a default receive/transmit permission that governs whether clients are allowed to receive and/or transmit on that connection. The default permission can be overridden by the owner to establish different individual permissions for some clients. For example, if a video conference call consisted of two connections, one for video and the other for audio, where the default permission was set to receive/transmit on both, the owner would need to override this default for a client lacking video equipment who wanted to participate in the call on the audio connection only. In this case, the client's permission on the video connection would be set to neither receive nor transmit, while it would remain the default receive/transmit on the audio connection.

The attributes of calls and connections are allowed to change during the course of a call, making calls very dynamic. That is, clients can be added to and deleted from a call while other clients are actively communicating. The bandwidth of connections can be increased or decreased as clients' demands change, and client receive/transmit permissions can be updated to allow new patterns of communication. The dynamic nature of calls increases the set of applications that can be supported by our network and makes the applications more flexible in meeting client needs.

6.2. Connection Management Access Protocol

Clients of the network interact with the network using our Connection Management Access Protocol (CMAP). CMAP defines the interface between clients and the network used to create, manipulate, and destroy calls. CMAP does not address the procedures for data exchange; we leave this to other protocols which are specified separately. CMAP is layered on top of ATM User-to-Network Interface (UNI) protocol [3, 9, 26].

CMAP uses a concatenation of a unique identifier that distinguishes multiple calls owned by the same client and the call owner's network address as a *call identifier*. Note that the call identifier is globally unique and must be distinguished from the VPI and VCI, local identifiers that pertain to switches and their associated links. The call identifier is not involved in routing individual cells, but is always involved in the creation, modification or destruction of a call.

To illustrate some of the CMAP operations, Figure 24 shows the steps performed when a point-to-point call containing two connections is created between clients *A* and *D*. In step 1, an *open_call* request is issued by *A*. Included in the request is the bandwidth specification for the connection and the default permissions. Upon receiving the request, node *N*1 makes sure that the required resources are available, then responds to *A* with an acknowledgment in step 2, creating a call between *A* and *N*1. In step 3, *A* sends an *add_endpoint* request for node *D* to the network. This causes the network to internally forward the request to *N*4, where the network, in step 4, sends an *invite* prompt to *D*, asking if *D* would like to join the call. This request

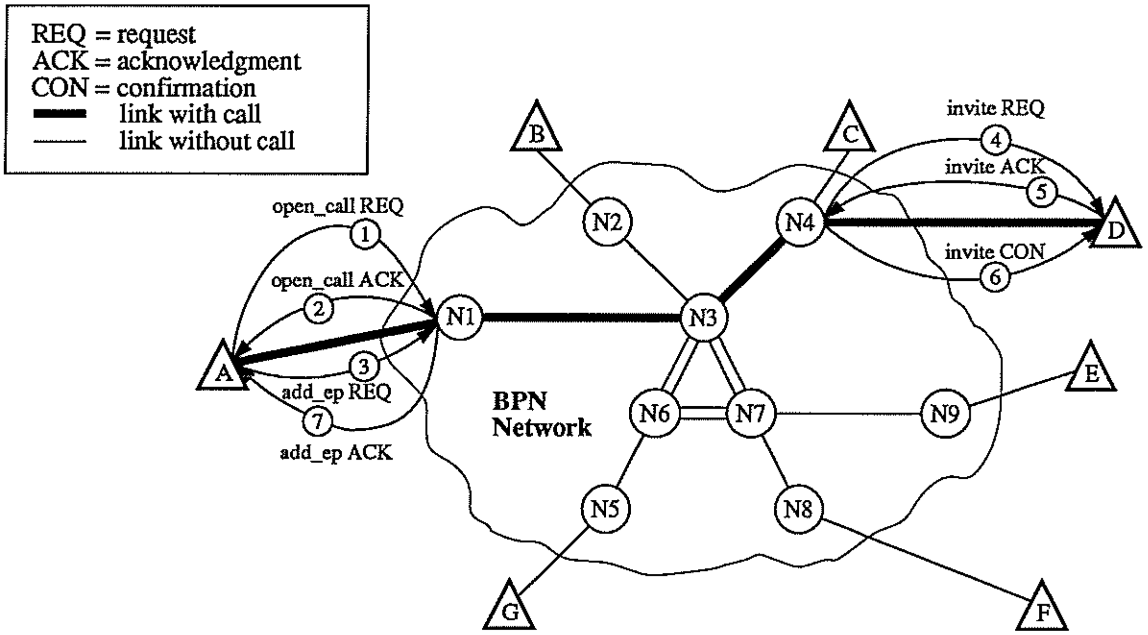


Figure 24: Steps in Creating a Point-to-point Call

contains the bandwidth specification and default permissions of the call so that *D* can decide whether it can accept the call (that is, whether it has the resources to support such a connection and whether it is willing to join using the indicated permissions). If *D* has the required resources and the desire to communicate with *A*, then *D* acknowledges in step 5. Finally, the network builds the connection between *A* and *D*, confirms that *D* has been added in step 6, and acknowledges the *add_endpoint* request by *A* in step 7. If the network does not have the necessary resources to support the connection, negative acknowledgments would be sent instead in steps 6 and 7.

The previous example is extended in Figure 25 to a multipoint call by issuing another *add_endpoint* request for client *F*. Any of the three clients, *A*, *D* or *F*, can issue the request. Figure 25 shows the steps when *F* requests to be added. In step 1, *F* sends the initial *add_endpoint* request. This prompts the network to search for the call, tentatively allocating bandwidth along the links on each hop. Once the call is located, the bandwidth allocations are committed, creating the multipoint communication channel. Finally, in step 2, the network responds to *F* with an *add_endpoint* acknowledgment. Once again, if the necessary resources for adding *F* did not exist in the network, a negative acknowledgment would be sent instead. The dialog box in Figure 25 indicates where data is copied at node *N*3 in the call tree created between the three clients *A*, *D* and *F*. Additional endpoints may be added or dropped from the call while existing endpoints are communicating, as *F* was added in this example.

6.3. Implementation

We refer to the system that implements connection management, our call model and our CMAP protocol as the Connection Management Software System (CMSS). A copy of the CMSS runs at each network node and

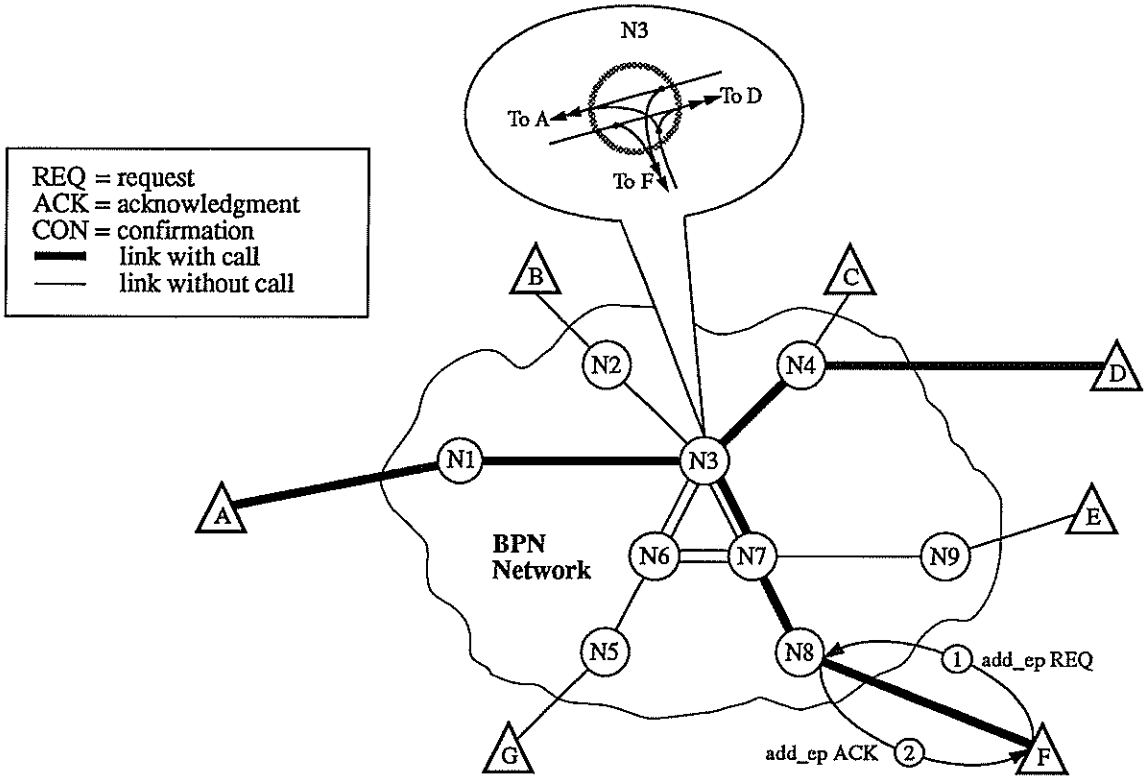


Figure 25: Steps in Extending a Point-to-point Call to a Multipoint Call

is responsible for the following: updating switch tables when creating connections through the switch fabric; bandwidth management for call acceptance decisions; routing of requests destined for remote clients and remote network nodes; and general OA&M (Operations, Administration and Maintenance). Each CMSS runs on a general purpose UNIX computer and services the requests from all clients attached to its node. For those requests that cannot be completely satisfied locally, the CMSS enlists the services of other CMSS systems running on remote nodes. In order to increase the number of clients that can be supported by a single CMSS, we have generalized the notion of a node to include a group of one or more switches all under the control of the same CMSS. Details of CMAP are contained in references [8] and [14]. Other related information can be found in references [7] and [17].

7. Creating the Applications

The reason for incorporating applications in Project Zeus is to explore possibilities which may transform the daily practice in these areas. Demonstrations of serious use of the technology by scientists and scholars can do much to shape the future of both the application area and the network. In this regard, a specific aim of Project Zeus is to facilitate experiments useful in determining the network bandwidth requirements of a variety of application areas. However, it is often the case that the perceived nature of the application depends on the network solution chosen. For example, network bandwidth that is too narrow may cause our collaborators to limit their scientific horizon. On the other hand, network bandwidth that is too wide may not be cost-effective until well beyond the user's planning horizon. We feel that Zeus technology strikes a balance between these extremes. Phase 1 demonstrates a cost-effective network architecture that delivers bandwidth to an application well beyond the bandwidth routinely available today and yet in phase 2 can be scaled to gigabit bandwidths wherever the need is manifest.

Thirteen possible applications are listed in this section and although the list is incomplete, there are too many for all of them to be turned into working applications in the next two years. We wish, therefore, to select a small number of promising applications for our initial experiments in phase 1 of Project Zeus and to expand to include the others in phase 2. Criteria for selection of the phase 1 application areas are as follows:

- Need for high-speed, multimedia communication.
- Evidence of substantial present investment in networking.
- Contribution of diversity to the set of selected application areas.
- Opportunities for communication with other schools and institutions.
- Indications of substantial impact on the future of the application area.

We have selected four application areas that satisfy these criteria and they are discussed in detail in the paragraphs below.

7.1. Medical Imaging and Electronic Radiology

Medical imaging has been steadily moving from film to digital storage and display over the past two decades. Computed tomography led the way when it was introduced in 1972. Today almost all medical images are either captured or can be made available in electronic form. The advantages of storing, managing and delivering images electronically instead of photographically are substantial. Many radiology departments are on the threshold of converting to an electronic system for acquisition, storage and display of all medical images. These electronic radiology systems increase the potential for making the electronic medical record complete by adding all the relevant medical images to the necessary text, laboratory results and graphics. A complete electronic medical record may be the incentive that doctors need to overcome their natural caution regarding the acceptance of informatics assistance in medical decision making. As system costs decrease and the quality of electronic images improve, the transition seems inevitable first to electronic radiology and later to the complete electronic medical record.

If these developments are inevitable, what is impeding the first step, the transition to electronic radiology? Important questions that have yet to be answered are:

- What gray scale and spatial resolution is required for each source of diagnostic medical images, the imaging modality? Does the required resolution change depending on whether the purpose in viewing the image is primary diagnosis or some other secondary use?

- Is lossless compression cost-effective for the transmission or storage of images? Can lossy compression methods be used? If so, what distortion is acceptable for each imaging modality and viewing purpose?

- Does the physician need to control contrast, change magnification or employ image enhancement routines? Would a small number of predetermined image manipulation settings be sufficient and eliminate the need for physician interaction with more flexible procedures?

- What traffic patterns can be expected after electronic images are widely accepted by physicians? Are these traffic patterns predictable from present experience with film or are they qualitatively different?

- What response time from image request to its display will be accepted by physicians? Do the results depend upon the viewing purpose, the system cost or both?

The answer to all of these questions will affect the bandwidth needed by an electronic radiology system. Unfortunately, the answers are difficult to obtain in the laboratory. Many helpful psychophysical studies have been carried out and more are planned, but in the final analysis, acceptance of an electronic radiology system by practicing physicians can only by determined be field trials. Premature bandwidth limitations can seriously bias the results of such a trial.

The Mallinckrodt Institute of Radiology (MIR), Washington University's department of radiology, is a leader in electronic radiology research, development and deployment. A prototype electronic radiology system has been developed and a clinical trial in MIR's chest service is underway. In this trial, phosphor plates exposed in a portable x-ray unit are processed by a computed radiography scanner to produce digital chest images. These images are displayed in soft copy form on high-resolution displays along with any available previous images. A high speed image server has been developed and integrated with patient and image databases in MIR's Radiology Image and Information Manager (RIM) [11]. Images will also be available on intermediate-resolution displays placed in selected intensive care units within the hospital.

Images used for primary diagnosis in this study have spatial resolution of about four megapixels with gray scale resolution of 10 to 12 bits per pixel. MIR believes that a response time under two seconds is desirable. We currently experience a one second time interval for the database retrieval of image information from RIM and about two seconds for the delivery of the image itself over a 40 Mb/s point-to-point fiber link from the image server. Even with this limited prototype, much useful information will be obtained from the present clinical trial before the end of 1991.

The plan indicated in Figure 26 will be refined with the aid of the results from this trial and the improved understanding of the technology developed under Project Zeus. Here, images are acquired from Magnetic Resonance (MR), Computed Tomography (CT), Computed Radiography (CR) and Film Scanner (FS) devices. Images are stored for rapid viewing in an image server and archived for the long term on an optical storage device. A radiology database management system (DBMS) keeps track of both patient and image information that augments the image itself. Ethernet technology can be used for connectivity to the DBMS since no image traffic is involved. Likewise, the various data acquisition devices can use either Ethernet or FDDI networks since, in most cases, the average data rate is low and response time is not an issue.

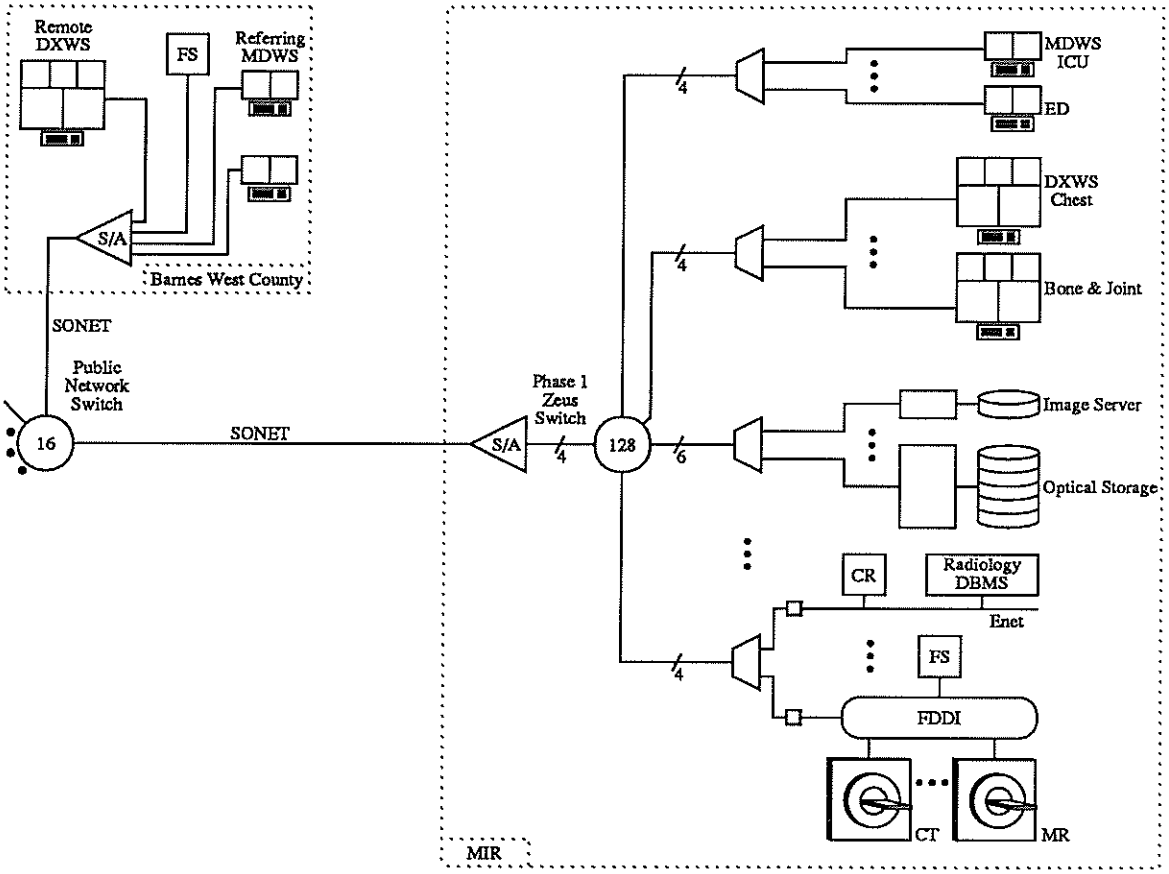


Figure 26: Broadband Electronic Radiology System

There are two kinds of display screens associated with the imaging workstations shown. A diagnostic workstation (DXWS) has multiple display screens (for example, two high-resolution screens and three intermediate-resolution screens). A medical doctor workstation (MDWS) has one or two intermediate-resolution displays. The DXWS is used for primary diagnosis of radiographic images such as those in the chest and the bone and joint services. The MDWS is used by referring and attending physicians to understand and advise on the diagnostic report. These physicians may be at their office, in an Intensive Care Unit (ICU) or in the Emergency Department (ED). Both the DXWS and the MDWS incorporate the capability for multimedia including voice, video, graphics, in addition to high-resolution images so that conferencing between radiologists, specialists and other physicians is facilitated.

The 128 port Zeus switch will provide connectivity between image sources and image storage and between image storage and image display. A SONET/ATM interface and multiplexor/demultiplexor (designated S/A) connects the Zeus switch to a broadband public network switch over a 620 Mb/s SONET channel. The public network switch connects to a satellite hospital location that can benefit by weekend, night and subspecialty coverage from the medical center.

Most of the components of this overall plan will be in place by the end of 1991, at least in prototype form. The prototypes need to be converted into production versions with modifications derived from our current experience. We believe we are ready to embark on an important set of experiments that can help answer many of the pressing questions regarding the feasibility and effectiveness of electronic radiology.

7.2. Optical Sectioning Microscopy

Organisms and their constituent cells are obviously three-dimensional entities. Yet, until recently, it has been impossible to visualize living biological specimens in 3D. In the past, biologists could peer inside a 3D tissue only if the tissue were physically sliced into sections. Now, however, living specimens can be examined by an important new technique known as optical-sectioning microscopy.

In optical-sectioning microscopy, a series of 2D images are acquired at different focal planes throughout a specimen. In any particular focal plane, an image consists of light directly from that focal plane, plus out-of-focus light from surrounding focal planes. It is this contaminating out-of-focus light that has heretofore made 3D imaging problematic, and necessitated physical sectioning of specimens. Optical sectioning has become practical only recently due largely to the development of highly linear and extremely light sensitive charge-coupled device (CCD) cameras. These CCD cameras permit quantitative imaging at light intensities so low that the specimen is unharmed. The resultant images from the CCD camera can then be processed by various computational algorithms designed to diminish the contributions of out-of-focus light. The camera's linearity is critical for these processing algorithms which rely on an imaging model of a linear superposition of light sources in the specimen.

There are only a handful of optical-sectioning microscopes in existence. One such instrument is in the Biology Department at Washington University. Biologists here are using this new tool to tackle a variety of fundamental, and previously insoluble, questions. These studies are of broad importance to biology and include the determination of the 3D architecture of chromosomes; time-lapse visualization of so-called molecular motors which power the transport of particles within cells; inspection of the mechanisms by which cells divide; determination of a cell's response to various environmental stimuli; examination of how cells move and reorganize themselves in a developing embryo; and mapping the pathways of neurons in the brain. Nearly all of the applications generate high-resolution 3D movies as a routine part of data analysis.

To provide a flavor for the significance of the biological questions being addressed, and also to provide a sense for some of the specific methods employed in these studies, one of the preceding applications is considered in more detail below.

A fundamental, unsolved question in biology is how a single-celled embryo gives rise to a complex, multicellular adult organism. During this process of embryonic development, cells of virtually all animal species execute a complicated series of movements known as morphogenesis. These coordinated cell movements are responsible not only for generating the shape transformations required to produce the adult, but also for juxtaposing previously separate cell populations, thereby inducing new pathways of cell differentiation.

Very little is known about how these critical cell movements of morphogenesis are orchestrated during embryonic development. Ignorance of these phenomena has been due in large part to an inability to observe how individual cells behave in the 3D mass of embryonic tissue.

Optical-sectioning microscopy is being used to study 3D morphogenesis in a simple organism, the cellular slime mold *Dictyostelium discoideum.* Dictyostelium is widely studied as a model system for the processes of embryonic development. In Dictyostelium, cells form a multicellular mass that undergoes dramatic shape transformations and ultimately gives rise to a fruiting structure. This fruiting body consists of only two differentiated cell types that are always arranged in a stereotypic pattern. This entire process of development requires only one day. To visualize how cells move during this process, a small percentage of cells is marked with a dye and then these labeled cells are mixed with unmarked cells. The motion of the marked cells can be followed by acquiring 3D images of the multicellular mass at successive time points in development.

Two questions about cell movement are crucial. First, what are the trajectories that cells pursue during development? Preliminary observations demonstrate that there are at least two types of motile behavior. Some cells move along 3D spirals, while others jiggle randomly in place. These observations raise new questions about what signals guide certain cells along spiral trajectories, why some cells do not respond to these signals, and whether the two types of motion observed are correlated with the two types of cells that ultimately differentiate. Such new questions can be addressed by analyzing motion in various mutant lines that, for instance, are known to have impaired responses to putative signaling molecules, or that differentiate altered percentages of the two cell types. A second critical question about cell movement is what is the mechanism by which an individual cell moves? To address this issue requires first an understanding of the 3D deformations that a cell undergoes when it moves. Then an analysis of various mutant lines that lack components believed to be crucial for movement will help elucidate the molecular mechanisms underlying the observed deformations of a moving cell.

The above description of just one scientific problem gives a sense of the specific methods that might be employed by investigators using the microscope in their research. The computational tasks are varied and often complex. We believe that future research will benefit tremendously from the distribution of these computational tasks for at least three reasons.

- Individual biologists will be able to analyze and process data at a workstation in their own lab.

  In addition to the computer interfaced to the microscope, there are already three workstations dedicated to optical-sectioning data analysis located in separate Biology Department laboratories. Several other biologists are planning to purchase workstations of their own in the near future. Project Zeus will tie all of these workstations into a network that will provide individual researchers access to each other and to their archived data, as well as access to any other computer on the network where their data can be processed (see Section 9.6). This will facilitate research enormously since at present only one person can use the image-processing system at a time.

- Collaboration will be facilitated both locally and nationally.

  The optical-sectioning project spans four departments at Washington University: Biology, where the microscope is housed; Computer Science, where networking studies are done; Electrical Engineering and the Institute for Biomedical Computing, where (at both locations) new image-processing algorithms are under development. Project Zeus will interconnect the four different sites and permit researchers at each location to visualize immediately the results of their colleagues at any other site.

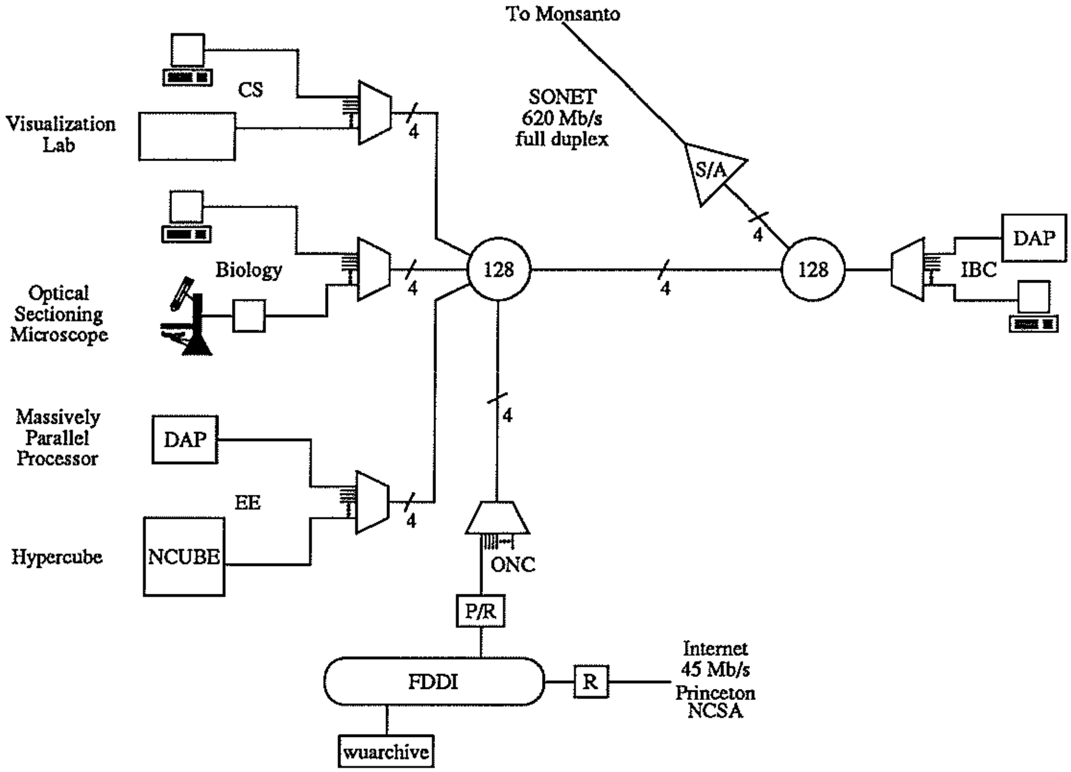


Figure 27: Application of Project Zeus to Optical Sectioning Microscopy

Since the optical-sectioning microscope is a rare commodity, several Washington University biologists are also collaborating with colleagues at other institutions who wish to use this facility. For example, in the slime mold application discussed above, a collaboration is underway with biologists at Princeton University who are developing techniques to visualize gene expression in this organism. Their method requires the ability to visualize very weak fluorescent signals in 3D in a living specimen, and therefore is ideally suited for the optical-sectioning microscope. Presently, research is hampered by slow communication: image sequences from the computer screen are videotaped and then mailed to Princeton where they are analyzed, and modifications to the experimental protocol are made. The ability to relay images to Princeton as they are collected from the microscope coupled with the ability to discuss and analyze these data via teleconferencing will dramatically facilitate this and all other collaborations. Together, Project Zeus and the Internet will provide biologists these opportunities.

- Computational performance can be dramatically improved.

All future work on the optical-sectioning microscope will profit from access to a wide variety of computing resources. The optical-sectioning methods in use, not to mention new methods under development, are computationally demanding. A typical time-lapse 3D data set ranges from 150-300 Mbytes. Computational power is required for image acquisition (for maximal temporal resolution), image processing (for optimal removal of out-of-focus light and eventually image segmentation and analysis) and image display (stereo movies of cell motion are routinely constructed). All of these steps are presently accomplished on one computer.

Performance can be optimized at each of the steps outlined above. As suggested in Figure 27, Project Zeus will permit image acquisition on a computer linked to the CCD camera and microscope in the Biology Department, as well as rapid retrieval of images from a 5 Gbyte archive. At present, image processing to remove out-of-focus light is limited to a simple, but fast linear method. Non-linear, iterative methods have been developed and even more complex methods are under development. These latter methods are more computationally demanding, but they provide far superior 3D resolution. High-bandwidth access to a supercomputer and other parallel machines will provide biologists an unparalleled view of 3D structures.

For instance, a new algorithm is under development in Electrical Engineering to visualize the 3D cell-surface contours of the migrating slime mold cells discussed above. The image-processing of these data is now done on a Hypercube and on a massively parallel DAP in Electrical Engineering. Project Zeus will link these machines to the microscope computer so that biologists can visualize the processed images immediately. Rapid processing of the 3D data is critical because it enables the biologist to determine whether to continue examination of a particular specimen. This evaluation can only be made by visualizing the processed data. If processing is not in real time and a particular specimen is unsatisfactory (which is often true), then considerable effort is wasted by collecting data for a specimen that will ultimately prove worthless.

A further advantage of the Zeus network is that it will link the Biology computers to the visualization lab in Computer Science. Sophisticated visualization and recording of time-lapse 3D data is critical both for data analysis as well as data presentation. The visualization lab will assist biologists in both of these capacities.

Motivated by the considerable advantages cited above, the feasibility of networking has been tested using existing campus network protocols for the slime mold application discussed above. Further details of this test are presented in Section 9.6.

7.3. Earth and Planetary Sciences

The Earth and Planetary Sciences Remote Sensing Laboratory is one of six Geosciences Nodes of the NASA Planetary Data System. Derived image data from the surfaces and interiors of all the planets (except earth), along with associated geochemical and geophysical data are maintained by the Geosciences Nodes. As the lead of the Geosciences Node, Washington University coordinates activities among the four subnodes at MIT, Brown, Arizona State, and the Johnson Space Center. The Washington University node in the Remote Sensing Laboratory is responsible for working with data sets acquired during the Magellan and Mars Observer missions to ensure that they are properly documented and archived. The Laboratory receives data sets, analyzes these data [4], publishes selected data on media such as CD-ROM for delivery to the planetary science community [27, 28], provides information to this community and gives them expert assistance. Currently the Laboratory is producing a set of 30 or more CD-ROMs containing the entire Viking Orbiter collection of about 50,000 images.

The Laboratory also maintains one of the fourteen NASA Regional Planetary Image Facilities housed at universities and government institutions throughout the United States and Europe. The Washington University facility serves midwest scientists, science teachers and students by providing access to images and maps produced by planetary missions since the beginning of the U.S. planetary exploration program. Several hundred thousand images, along with ancillary products such as topographic and geologic maps, are available to users. The volume of data is expected to grow exponentially from missions scheduled in the near future, and has already been doubled by the data products from the Magellan mission to Venus.

The Remote Sensing Laboratory, the existing Seismology Laboratory and a new Planetary Geophysics Laboratory are preparing to move into new quarters where they will be located together. These three laboratories will carry out the primary computation and data management work of the Department of Earth and Planetary Sciences. Presently there are a dozen computers linked by a departmental subnet and connected to both the Internet and the NASA Science Internet. This network of computers devoted to the work of the three laboratories is expected to grow substantially in the new, more spacious quarters and include one or more powerful compute servers. Possibilities under study include a Cray or a cluster of several powerful numerics processors.

An important use of Project Zeus by the Remote Sensing Laboratory, is in the analysis of airborne data sets which include visible and reflected infrared imaging spectrometer (AVIRIS) measurements. Each of the AVIRIS data sets contains over 150 Mbytes, consisting of 224 image planes covering the 0.4 to 2.5 µm spectral region. Using a library of the spectra of rocks and minerals it is possible to search a region for known spectral patterns allowing the identification and mapping of the mineral composition of the surface of the planets. In some regions it will be necessary to estimate the mixture of minerals present from the know spectra of the component ingredients. We plan that AVIRIS data will be sent to a server for such computationally-intensive jobs and then to a multimedia workstation where an investigator examines the results. A network organization that can support research in Earth and Planetary Science is sketched in Figure 28. This approach would be severely limited by the Ethernet transmission speeds today, but would work well with Zeus technology and one or more high performance compute servers.

An additional use of the computational equipment shown in Figure 28 is in the development of visualization tools. The Earth and Planetary Sciences are graphics- and image-intensive disciplines. Plans are underway for the development of three-dimensional models of geologic processes and associated 3D computer-generated images. This development will require extremely fast communication between workstations and a fast compute server. In seismology and structural geology, software is being produced for generating three-dimensional images of the upper portions of the Earth. Time sequences, composed of a large number of such images, are to be used not only for research, but for teaching as well. Some classrooms in the new natural sciences building will have the capability of projecting computer-generated images. In order to display a large number of these complex images with low latency, the speed of phase 2 Zeus technology will eventually required.

The high speed communication of the interim Internet and the evolving National Research and Education Network will be available to the investigators in these three laboratories via the ONC connection. Exchange of images with the other five NASA Geosciences Nodes and the Jet Propulsion Laboratory will be an essential aspect of this connection. The massive data storage capabilities of *wuarchive* (see Section 8) will allow backup and national distribution of planetary images set at a site outside the department and directly accessible over the Internet.

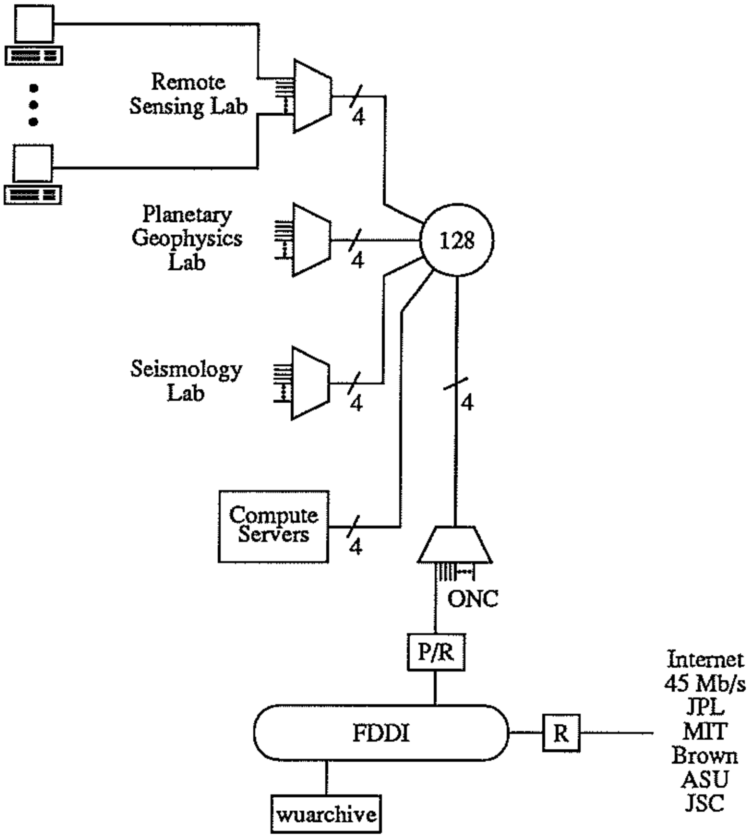


Figure 28: Earth and Planetary Sciences Use of Zeus Technology

7.4. Visualization in Art and Architecture

The Urban Research and Design Center (URDC) in the School of Architecture has identified a research agenda which addresses issues regarding the development of a designer's workstation. The workstation would be part of an envisioned collaborative workplace in which a small group of designers and consultants would contribute data, information and knowledge regarding an architectural or urban design problem. This multimedia workstation would be the window through which a designer would enter and retrieve information in the form of text, graphics and images. Clearly, images play a significant role in design activities in the visual arts. No matter what the medium, architects and urban designers are constantly using images in all aspects of design thinking. Whether it is the perception of the environment, an image in the mind's eye, an abstract drawing or a photographic record, designers use images to conceive of, and manipulate their design ideas. Managing these image collections must occur at a variety of levels in the creative process and is dependent on the type of image that is called upon for reference. Now the capacity to use electronic means of creating, altering, storing, and retrieving images will enable designers to effectively use image collections in ways that have not been possible before.

The designer's workstation including suitable visualization tools will be the primary means by which architects and urban designers could access and use images in their work. A full function CAD program with

3D modeling and photorealistic rendering capabilities is the core on which the proposed system would be built. The workstation would include the capability to display still images and motion video at the desktop. The basic purpose is to enable designers to access a shared visual database of images and graphic information about art, architecture and urban design. Browsing the collection, queries to the database and entering new visual material need to be accomplished as easily and intuitively as possible. Current projects in the URDC which access video stills and motion sequences have used a two screen configuration (the computer monitor and an NTSC video monitor). The objective in Project Zeus is to combine the display of CAD graphics, images and video in a multimedia workstation with access to a broadband network.

In addition to the conventional CAD functions, visualization tools would include computer generated animations, processing for ray tracing, photo and image manipulation, and access to image collections. These various tools and capabilities might not necessarily be on a single platform, but they could be made available through the network and its associated servers.

The URDC has begun to create a database of urban images, data and information for use in the urban design studio. Information regarding well known places including photographic images stored on videodisc and 3D CAD models in the symbol library of the CAD system have been collected. Access to the photographs can be managed through the 3D model by linking the graphic objects in the model to the image collection. Comparative studies and the use of essential elements in the design of urban places can be accomplished by this interconnected resource.

Such a networked environment would also permit architectural students and their instructors to access a visual and text database of information and design knowledge that could be shared for a specific project. The collaborative aspects of this type of working environment are likely to become more commonplace so that urban design and architectural projects effectively utilize interdisciplinary contributions. We envision information being shared over wide area networks in which design consultants could interact and make use of shared visualization techniques. Not only would it be possible to involve other faculty and graduate students on the campus but consultants, clients, and others off campus and in other cities could be included. The URDC has developed a working relationship with the architectural firm of Hellmuth, Obata and Kassabaum (HOK) in St. Louis and opportunities exist for undertaking applied research projects which would link HOK downtown office to the School of Architecture. The project could utilize the SONET loop that Southwestern Bell will install linking downtown with the campus. Moreover, preliminary discussions have occurred with the Graduate School of Design at Harvard University regarding the possibility of a joint design studio which would utilize shared databases and networked communications over the Internet. Both of these types of projects would be enabled by Project Zeus.

Initially the network would serve URDC and the School of Architecture's design studios. Somewhat later, all five divisions of the University that are associated with the visual arts would be served. In addition to the School of Architecture these are the School of Fine Arts, the Department of Art and Archeology, the Art and Architecture Library and the Gallery of Art. Project Zeus would enable the transmission of video and digital images over the network and allow users to share the common visual databases. For example, the Department of Art and Archeology has over 100,000 slides of art and architecture and if they were transferred to a server on the network, faculty and students could access the image collection from their workstations. Moreover, we envision projects in which temporary and personal image collections are a vital part of the communications that take place in collaborative efforts in the design disciplines. A small group of designers focussing on a specific problem could share their observations and schematic sketches over high speed networks such as Project Zeus. Often, in such endeavors browsing the image collection is a powerful technique in seeking new insights to the problem at hand. Full motion video or high speed display of digital images is necessary for effective browsing of image collections. On the other hand, systems which search for a single image as the result of some type of text query may not necessarily require high speed image transmission. However, browsing is still likely to be required when the query does not result in just one appropriate image. Thus, both types of access to the collections are necessary for the activities expected on the designer's workstation.

The plan for URDC to develop a designer's workstation and investigate the collaborative workspace raises several significant questions:

- What image transmission data rates are necessary to enable the browsing of the collection? In addition to displaying a single image as a response to a query, the capability of browsing specific classes of images is a useful and powerful tool for designers. Browsing requires image display rates to be as high as ten frames per second. Image size, compression and storage medium are issues that need to be considered.

- What techniques might be used to enter new images and classify them as a part of the collection? In a collaborative working environment contributors will be bringing photos, video, text and graphics to the group for consideration and easier access and input is imperative for effective use of the system.

- Are new developments in virtual realities and space applicable to the development of the collaborative workspace? Is it possible that the visual collections might be shared in a virtual conference room?

Through Project Zeus we plan to address these questions. Figure 29 shows a possible organization of the network for this application. Each of the five divisions associated with the visual arts would eventually be involved, but the URDC and the School of Architecture would employ the designer's workstation first with migration to the other three divisions as experience is gained and purchase prices decrease. Image databases would be archived on *wuarchive* and available over the Internet. Access to collaborators at HOK and Harvard would be available over the Southwestern Bell SONET testbed and the Internet, respectively. Such a network organization could be a major step in the introduction imaging technology into the visual arts.

7.5. Applications and Their Bandwidth Requirements

The purpose of this section is to help characterize traffic output of various applications in simple terms and argue that these applications require high data rates and can drive the high speed networks, such as Zeus to their limits. We characterize the bandwidth requirements of the candidate applications using four parameters: peak bandwidth, average bandwidth, peak burst duration and the averaging interval. Note that these parameters are closely to the parameters used for resource allocation at the time of connection set up. Because we must speculate about the technology that will be available to us, our characterization of bandwidth requirements of applications is simple and approximate. Bandwidth estimates for the candidate applications are presented in the following paragraphs, and are also summarized in Table 1.

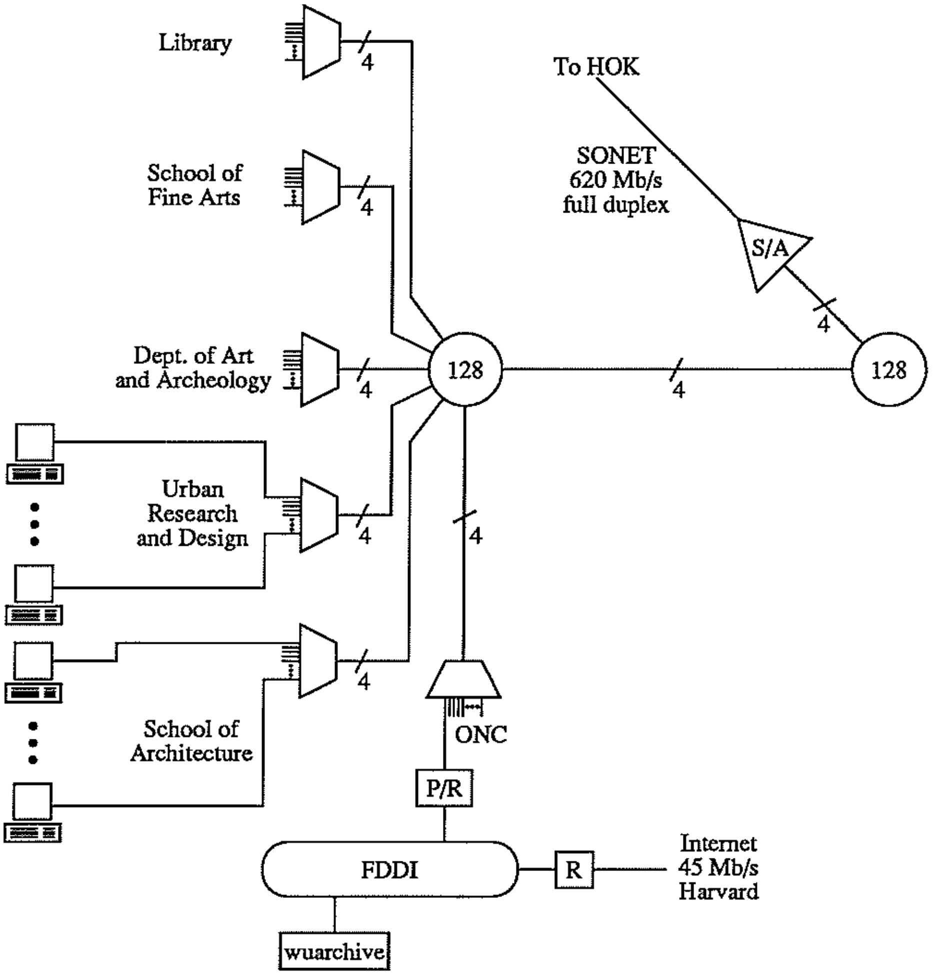


Figure 29: Application of Project Zeus in Art and Architecture

In the case of electronic radiology, current image and display resolutions are $2K \times 2K \times 2$ bytes in size, and a radiologist should be able to retrieve an image in about one second. To the approximations used in the table, this gives a peak bandwidth for phase 1 technology of 60 Mb/s and a peak burst duration of one second. Once the image is on the display, a radiologist typically looks at it for at least ten seconds before requesting the next one. This leads to an average rate of 60 Mb every 10 seconds which is 6 Mb/s. For the phase 2, we expect to provide two images simultaneously of $4K \times 5K \times 2$ bytes each to the radiologist for diagnosis. The retrieval time is one second and interimage time of 10 seconds remain the same, giving peak and average bandwidth of 600 Mb/s and 60 Mb/s, respectively.

It is important to note that our estimates for bandwidth requirements are conservative for the following reason. An estimate includes bandwidth requirement of only the primary data stream in the application. For example, in the case of electronic radiology, the numbers include the bandwidth requirement of just the image stream. If the application is used in a collaboration setting, as described in, it will also include other information streams such as, video, voice, and data.

	Electronic Radiology (Hi-res Images)	Optical Sectioning Microscopy (4D Images)	Earth and Planetary Science (3D Images)	Art and Architecture (Video)
Phase 1				
Peak bandwidth	60 Mb/s	120 Mb/s	120 Mb/s	30 Mb/s
Average bandwidth	6 Mb/s	1.2 Mb/s	12 Mb/s	10 Mb/s
Peak burst duration	1 sec	10 sec	10 sec	---
Averaging interval	10 sec	100 sec	10 sec	10 sec
Phase 2				
Peak bandwidth	600 Mb/s	600 Mb/s	1.2 Gb/s	180 Mb/s
Average bandwidth	60 Mb/s	30 Mb/s	120 Mb/s	60 Mb/s
Peak burst duration	1 sec	2 sec	1 sec	---
Averaging interval	10 sec	20 sec	10 sec	10 sec

Table 1: Application Bandwidth Requirements

Optical sectioning microscopy requires the collection of 1.2 Gb data set having 3D images with 20 slices obtained at each of 60 instants in time; this constitutes a 4D image. These data are stored for processing by a variety of algorithms. An investigator might wish to view the results of several algorithms in sequence while algorithm parameters are adjusted. We assume the compute server takes 10 seconds to receive the data set and 100 seconds to process it, whereupon the investigator will move on to the next algorithm or set of parameters. Perhaps by the time we have phase 2 technology, the compute server might be five times as fast.

In applications in earth and planetary sciences it is necessary to process 3 dimensional images; two space dimensions and one spectral dimension that can have as many as several hundred spectral bands. For example, an AVIRIS data set is 512 × 640 × 224 × 2 bytes. Again, a single data set is 1.2 Gb in size, but the investigator wishes to browse through various visualizations of the data. The environment is not as production oriented as that for electronic radiology and so we assume new images are requested at most every 10 seconds. Display and storage hardware limitations in phase 1 cause us to estimate that the full 10 seconds is available to transfer the image to a double frame buffer while the investigator examines the previously transferred image. In phase 2 this hardware limitation is assumed to be no longer applicable.

In the final application area, art and architecture, the most demanding circumstances occur for the presentation of video clips. Lossy compression is possible since no quantitative use of the images is contemplated. In phase 1 we assume NTSC video streams with JPEG compression (see Section 9.6). In phase 2 we assume HDTV video streams which require about six times the bandwidth of NTSC video. The data rate has considerable variation about the average, but averages over intervals of 10 seconds or longer will smooth most of these fluctuations.

This analysis of bandwidth requirements is preliminary, but it does indicate that the four application areas provide challenging communications problems for Zeus technology. We also see that the four areas provide a diverse set of scientific visualization problems that should benefit by the availability of Project Zeus.

7.6. Other Applications

Medical Informatics. The rapid and useful retrieval of relevant information from the medical record, medical texts and the medical literature is the objective of research by the Washington University Medical Informatics Laboratory in the Department of Medicine. Incorporation of such research results into clinical practice is this group's urgent, but elusive goal. The broadband network provided by Project Zeus will be an important facility in their attainment of this goal and the resultant improvement in medical care.

Radiation Treatment Planning. Substantially improved accuracy in radiation therapy is now possible because of the combination of individual patient imaging technology, on-line dose monitoring instrumentation and special algorithms for on-line dose estimation. At Washington University progress in this area is rapid as a result of a collaboration between the Division of Radiation Therapy, the Department of Physics and the Institute of Biomedical Computing. Through the use of broadband technology, centralization of planning personnel and equipment can be achieved for therapy machines distributed throughout a metropolitan area.

Outpatient Services. A new Ambulatory Care Center is planned for the Washington University Medical Center to be completed in 1995. Medical images are expected to be available as an electronic utility for each of the physician's offices in this new building. The interworking of Zeus technology and other broadband services is anticipated.

Electronic Communication for the Deaf. Central Institute for the Deaf is a world renowned center for education and research in hearing and deafness. A program to enhance the ability of the deaf to use electronic communication through lip reading is now underway at CID. The cell-based video service of Project Zeus is expected to be an important component of this exciting program.

Neuroscience. Washington University is one of the leading centers of neuroscience research. Positron Emission Tomography (PET), developed at Washington University, is a primary tool for mapping the function of the brain. Three-dimensional maps of an individual's metabolic response to cognitive stimuli are already being shared among laboratories that utilize PET technology. The Zeus Project could assist the PET group and other scientists at Washington University to play an important role in the proposed and very ambitious national initiative for mapping the brain and its functions.

Genetics. Geneticists worldwide are engaged in the challenging task of mapping and sequencing the human genome. Washington University has established one of the four national genome centers sponsored by NIH. Electronic communication will be essential for the scientists engaged in this task since the information to be shared far exceeds the identification of the sequence of three billion nucleic acids that constitute a single human genome. The Internet is now needed for this task and its bandwidth is adequate. Later, however, the higher capacity of the National Research and Education Network will be needed.

Molecular Design. A grant to establish a unique Center for Molecular Design has been awarded to Washington University. This Center will use computational and visualization techniques to explore the conformational space associated with experimental or hypothesized drugs. It is becoming increasingly possible to design new, practical pharmaceutical agents and then predict their biological activity even before the first laboratory experiments are begun. These new achievements in computational pharmacology require collaboration between the university's medical campus and the hilltop campus, between the university and industrial partners and between scientific colleagues, nationally. Both Project Zeus and National Research and Education Network will play important roles in this exciting enterprise.

Video Classrooms. The School of Technology and Information Management (STIM) has been charged with the task of bringing continuing education in technology to those working in engineering and informatics. This often produces a conflict for employers who are reluctant to provide the time for travel to the campus in addition to time for classes. "Distance learning" through video classrooms can be an answer. Furthermore, the possibility of two-way video through Project Zeus technology opens new avenues for this segment of STIM's activities.

Archives, Depositories and Libraries. The Washington University campus network operates the largest public electronic archive in the nation. This archive contains electronically accessible programs, bug fixes, news, images, bulletin boards and many other computer related documents. Users retrieve information from the archive over the campus network or the Internet for use at sites throughout the nation and in dozens of countries throughout the world. A tenfold expansion of the archive is planned in the near future to meet soaring demands for this widely accessible electronic depository. We expect that Project Zeus technology will play an important role in the expansion of the archive and its integration into the electronic information retrieval plans being developed by Washington University's Olin Library.

8. Deployment and Operation of the Network

To describe how Zeus technology will be deployed at Washington University it is helpful to understand the capabilities and organization of the present campus network. This network is a rich mixture of technology, services and applications. Backbones on the medical and hilltop campuses provide video and data services for research, educational, administrative and hospital purposes. These backbone networks are managed by the Office of the Network Coordinator (ONC) while the various departments manage the individual subnets and provide end-user support.

Beyond the traditional local network management activities, ONC has the additional responsibility of operating a mixture of secure and open systems in a decentralized network environment, administering a usage-based cost recovery system, providing central software acquisition and distribution, acting as an Internet hub connecting seven industrial and academic sites, and operating the largest and most widely used public computer archive on the Internet.

The challenge of deploying Zeus technology in this diverse and demanding environment offers the opportunity to explore several novel management capabilities inherent in this new approach to networking.

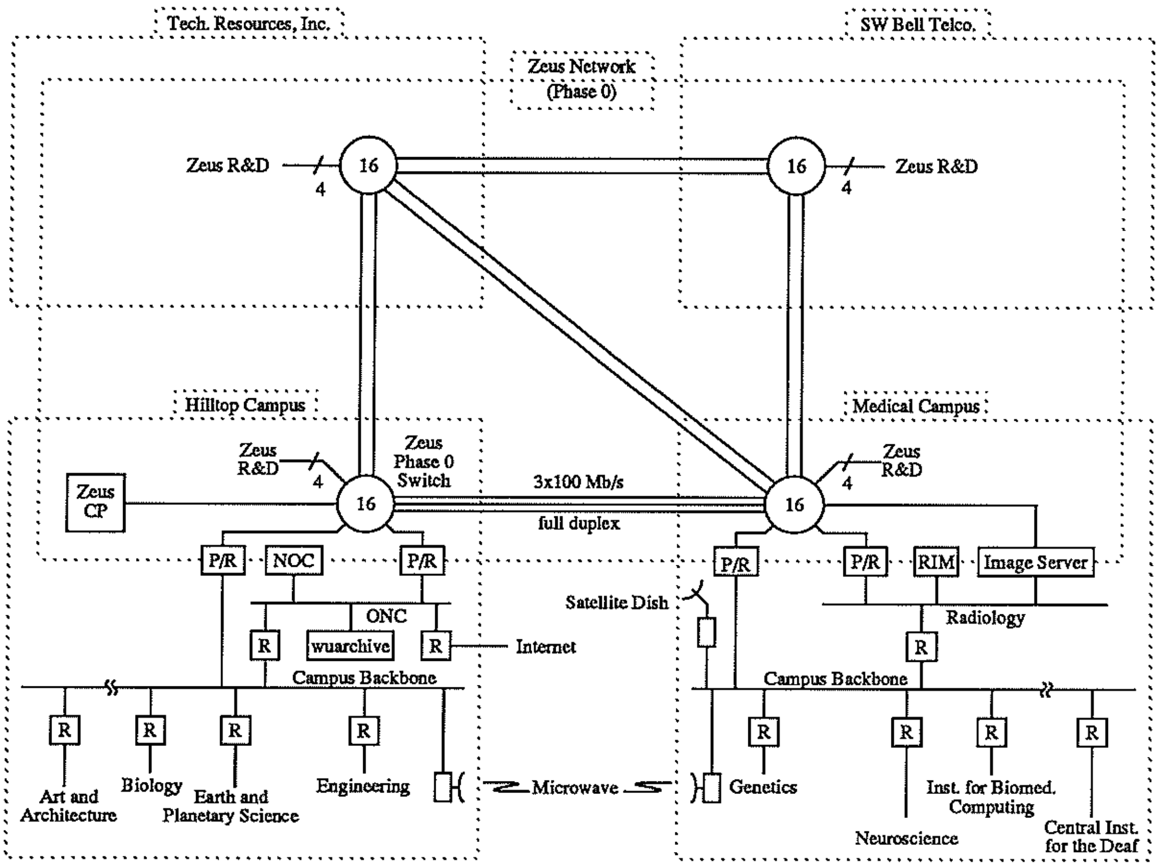


Figure 30: First Stage of Zeus Deployment: Providing Transport

The early involvement by the individuals responsible for managing the present network provides an effective mechanism for insuring that adequate network control and management capabilities will be incorporated into Project Zeus.

8.1. The Campus Network Today

The Washington University Campus Network uses a dual cable, 450 MHz, CATV backbone divided into 66 NTSC channels. The backbone is accessible from most major buildings on both campuses. Modems are used to transport data at full (10 Mb/s) Ethernet data rates that occupy 18 MHz on any set of three contiguous NTSC channels. Careful engineering and deployment of channels has prevented the saturation of any single Ethernet. Four such 18 MHz data channel allocations are currently in operation.

The campus data network is predominantly Ethernet and the two campuses are connected with a pair of 10 Mb/s microwave radio links (see lower portion of Figure 30). Over 24 cisco bridging routers (R) are used to connect 22 subnets containing over 1700 nodes and servicing over 10,000 users. Five of the over 20 active protocols are routed and the remaining protocols are bridged. The use of subnetting and the application of access lists in the routers allow open and secure systems to be intermixed.

Network management of the backbone and training of the subnet managers is provided by ONC. Commercially available management tools such as *Simple Network Management Protocol* (SNMP), LTM and LAN analyzers are widely employed, but substantial custom software has been developed to support the usage-based cost recovery system.

CATV channels 6 through 13 are used for video services. These include satellite downlinks, video conferencing, seminar broadcasts, and several informational channels. Video production services are provided by several groups on campus and most large auditoriums are cabled and equipped with video projectors.

8.2. Deployment of Zeus Technology

We plan a fully integrated, staged deployment of Zeus technology within the evolving campus network. This approach takes advantage of the flexible nature of Zeus technology and recognizes the continuing need to support existing as well as new protocols throughout our network. A smooth transition is anticipated.

The operational components of Project Zeus will be deployed in three stages as the technology and the management tools become available. Operational deployment will lag 6 to 12 months behind the completion of research and development activities to allow for maturation of the components of the technology while still providing opportunities for feedback to the products that arise from Project Zeus.

During the first stage of deployment, as shown in Figure 30, Zeus technology will be used as a transport network for the existing campus traffic through the use of a combination of Ethernet portals and bridging routers (P/R).

The connections between portals are established by default on initialization of the Zeus network. Zeus allows portals to be connected in any manner desired, allowing for the creation of secure, open, private and maintenance paths as required. During this first stage a public, multicast connection consisting of several of the routing portals (P/R) will be used to carry much of the existing backbone traffic. The routers at each portal will also provide most of the required network management functions.

The operational status of the Zeus network will also be monitored by a private connection created between the Network Operations Center (NOC) at ONC and the Zeus control processor (CP). Software compatible with the SNMP will be implemented on the CP to provide status information on the Zeus network. The NOC will monitor and control traffic routed over the Zeus network and the backbone's microwave radio links. New SNMP variables will be developed to monitor the activity and performance of the Zeus network.

During the second stage of deployment of the Zeus network a number of multimedia workstations will be incorporated and their functionality explored. In this stage all devices involved in public connections to the Zeus network are viewed by the present ONC network management tools as participating in a single subnet on the CATV backbone. Primary network management will still be provided by the bridging routers connected to the Ethernet portals as shown in Figure 31. Also at this stage an FDDI ring will be installed in ONC to provide increased bandwidth for *wuarchive* and its connection to the Internet.

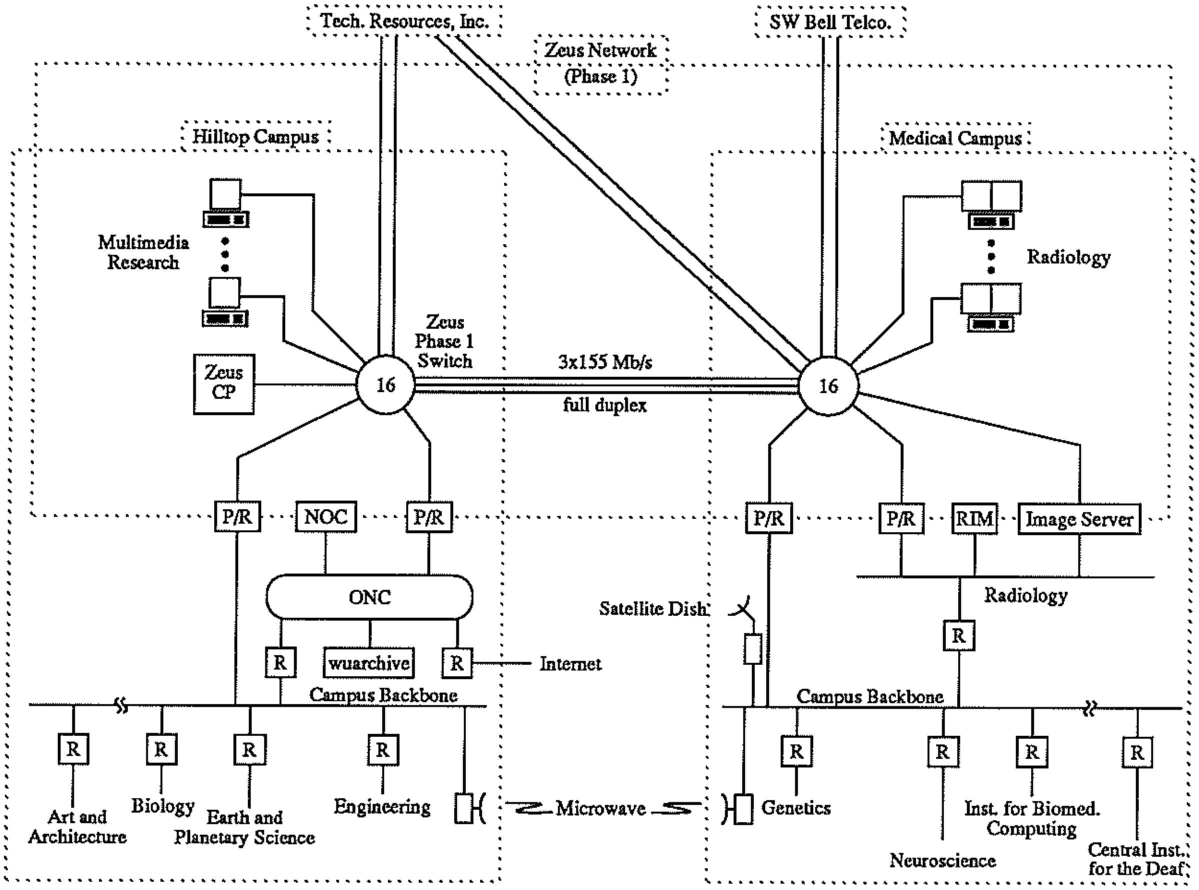


Figure 31: Second Stage of Zeus Development: Multimedia Workstations

8.3. Operation of the Network

Networks operating at very high throughputs require a different approach to network control and management than can be used when the throughputs are lower, as is the case with FDDI and Ethernet. Cells in a fast ATM network move so quickly that there is insufficient time to examine each one. This is in contrast to what is done with the packets that pass through devices like current bridging routers. The smallest transaction on an ATM network that can recorded or controlled is a connection. It appears likely that new protocols will be defined to take advantage of the services offered in ATM networks. There must, therefore, be a set of network management tools developed for a connection-oriented, guaranteed-bandwidth environment.

Many of these developments are one or two years away, and once available will take several years to mature and reach a significant market penetration. The third stage of the deployment of the Zeus network will address the development of prototypes of these network management tools in an environment of demanding applications using both new and existing protocols. Important feedback to the companies preparing these new tools will occur during this final stage of deployment of the Zeus network (see Figure 32).

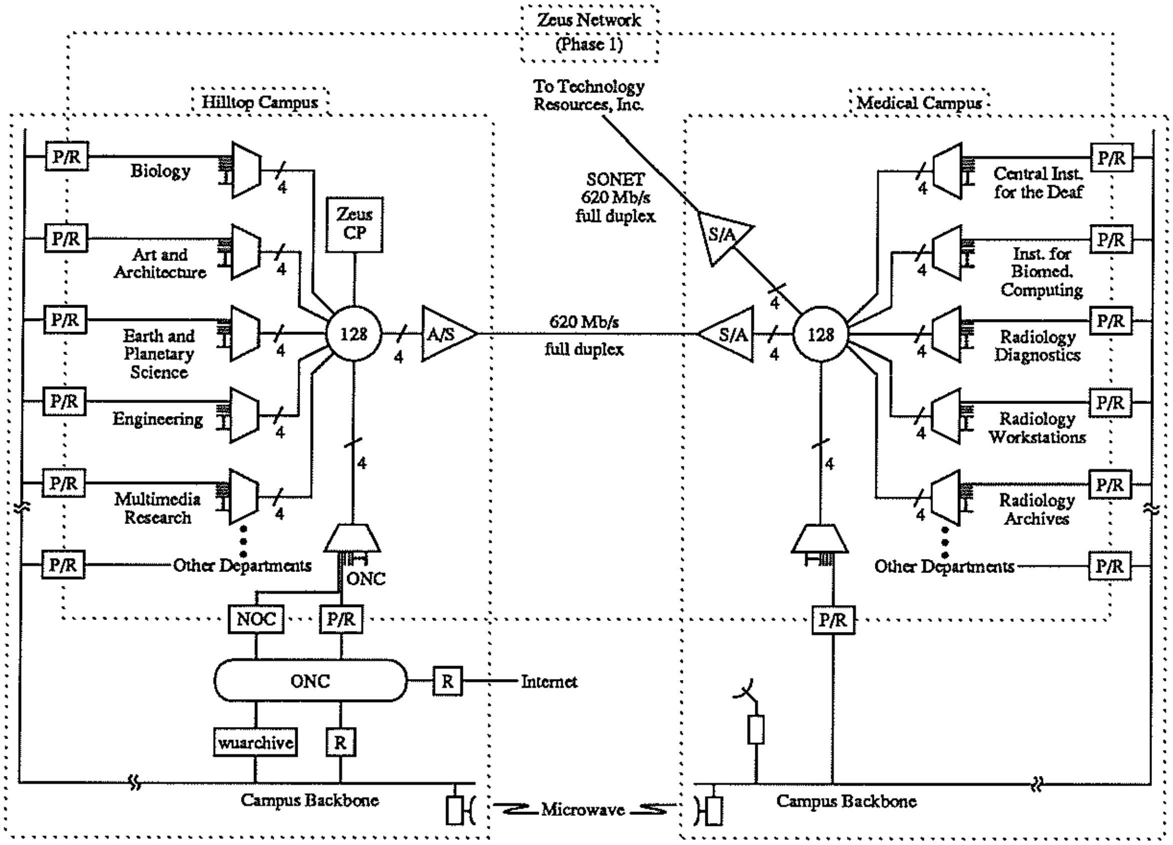


Figure 32: Third Stage of Zeus Development: Network Management

Once mature, the Zeus network will be much simpler to manage than today's networks yet much more effective in providing a variety of private and public services. The three stages of the network deployment provide the opportunity for users to choose to accept increased levels of functionality. An increase in the complexity for a user will only be apparent when that user chooses to move to a new level of functionality.

9. Using the Network to Answer Questions

9.1. Network Congestion

As summarized above, bandwidth management and congestion control in Project Zeus, makes use of a buffer reservation mechanism to handle burst level contention and a call acceptance algorithm that blocks new virtual circuits if the resources they require are not available. The call acceptance decision ensures that the probability is small that the short term load on any given link queue exceeds the link's capacity. The details of our approach are given in [43]. While we are confident that the chosen approach is a sound one, many specific questions concerning its performance can only be answered in the context of an operational network.

ATM networks are intended to serve a wide class of traffic with varying traffic characteristics. Hence, the approach we are taking to bandwidth management and congestion control makes very few assumptions about the nature of user traffic. In particular, we permit the users to specify their virtual circuit requirements in a general way, rather than restricting them to a few discrete circuit speeds. While we restrict the users very little in this regard, nonetheless, the users' actual requirements will affect the efficiency with which the network can be operated and hence its cost. This makes it very important to obtain an understanding of the actual traffic mix that will be present in high speed networks. For our purposes, each virtual circuit can be adequately characterized by its peak and average bandwidth. Specific questions we need to answer are as follows:

- What is the overall distribution of the peak and average data rate among virtual circuits in the campus network?
- What rates are associated with different applications?
- How does an application's rate parameters differ based on the user?
- In situations where network bandwidth is limited, can resource controls on individual workstations (the network equivalent of disk quotas) provide an effective means for controlling virtual circuit level congestion?

One assumption that must be made by any bandwidth management scheme that allows for statistical multiplexing of virtual circuits is that the traffic from different sources in the network is independent. That is, the timing of virtual circuit requests from different workstations or burst arrivals on different virtual circuits are assumed to be mutually independent. Unfortunately, experience in other networks clearly shows that such independence assumptions hold only approximately. External events that are unknown to the network can cause periods of highly correlated traffic, causing the network to perform well below normal expectations. It is important to obtain an understanding of the sources of correlated traffic in the Project Zeus network so that steps can be taken to limit its impact. We suspect that many of these sources will only become evident once the network is operating, and so we plan to include mechanisms in the network to detect the traffic impairments caused by these sources and identify their cause.

While the Project Zeus network controls burst traffic entering the network and provides certain performance guarantees for burst traffic, it also allows individual low priority cells to be sent through the network without constraint and with no guarantee on delivery. An important question we will need to answer in the Project Zeus network concerns how much of this low priority traffic is present, and what cell loss rates are typical for it. We also need to understand if this low priority traffic can impact the burst traffic in any significant way; in particular, we want to know if the presence of uncontrolled low priority cells has a negative effect on the performance within a switching system's network.

One of the important features of the call acceptance algorithm used in our bandwidth management approach is that it can handle a completely general traffic mix, while being fast enough to make rapid decisions. We expect the actual performance to be much better than the estimated upper bound of about one millisecond cited in [43], but are interested in verifying this under actual operational conditions. We wish to determine if there are any effective "shortcuts" to the general algorithm that can be used to cut the decision time down significantly and verify our worst-case assumptions.

The traffic monitoring mechanisms at the access points of the Project Zeus network will measure each virtual circuit's traffic and will flow control the virtual circuit if necessary to prevent the user from exceeding the rate specification given when the virtual circuit was established. We are interested in determining how tightly the rate specification can be matched to the actual traffic and in understanding to what extent the access flow control negatively affects the subjectively perceived performance of an application. These are closely coupled; if a user does not notice the performance effect of the flow control, it may be possible to reduce the rate specification to more closely match the actual traffic. If the user does notice the effect of flow control, additional resources may be needed and we must determine if these should be supplied by augmenting the peak rate, the average rate or perhaps by adding a supplementary virtual circuit to accommodate occasional excess demand.

Multicast virtual circuits with multiple sources provide a novel aspect of the Project Zeus network. They impact network congestion in that traffic streams from different sources in the same virtual circuit share a common bandwidth pool, with the network controlling only the overall bandwidth usage, rather than the traffic from the individual sources. This approach while highly flexible, raises some questions that we feel can only be fully answered in an operational context. We need to know how different sources share the bandwidth pool in typical situations and what end-to-end protocols are needed to facilitate this sharing. We also need to understand how the short term overloads that can be caused by multiple sources in the same virtual circuit might impact traffic on other virtual circuits.

9.2. Routing

As described briefly in Section 5.1, we plan to address multicast routing in the Project Zeus network by reducing the multicast routing problem to a point-to-point problem. While simulations indicate that this approach can be efficient enough for our purposes, experience in an operational network is necessary to really verify its efficacy.

There are many questions concerning multicast for which we lack answers at the moment. While we have designed our routing strategy to adapt to different network conditions, the efficiency of network operation will ultimately depend on the answers to these questions. The most crucial questions concern the nature of the multicast applications in the Project Zeus network. We need to know what applications will make use of multicast and how heavily they will use it. We need to understand how many endpoints are involved in typical multicast connections and whether they are widely distributed across the network or tend to be concentrated in individual departments or other "communities-of-interest." We need to understand how many endpoints are typically required at a single switch; if the fanout at individual switches is always quite small then more economical implementations may be possible.

Another crucial set of issues involves the dynamics of topology changes in multicast connections. It is important to understand how frequently endpoints are added to or removed from multicast connections. Also, we need to understand if there are situations in which endpoint additions and deletions can lead to seriously suboptimal topologies. The dynamics of multicast connections affect not only the quality of the connection topology but also the load on call processing software. To handle highly dynamic situations, we may require optimizations that save time in certain common situations. For example, it may be common for users to switch rapidly back-and-forth between a pair of multicast virtual circuits. If this situation is common, it may be useful to delay the dropping of a branch of a multicast virtual circuit when an endpoint leaves it in anticipation of a future "reconnect."

If endpoint additions and deletions significantly degrade the quality of a multicast connection's topology, it may be necessary to rearrange the connection in order to obtain a more efficient topology. Adoption of such a strategy first requires an understanding of how often it might be needed and the performance improvements it might reasonably be expected to yield. If a rearrangement strategy turns out to be warranted, questions arise concerning under what conditions it should be applied. Since this is likely to be an expensive operation, it should be applied only to those virtual circuits where it can be expected to do the most good. These would typically be long-lived virtual circuits with many endpoints and seriously suboptimal topologies. The next question that arises is how to accomplish the rearrangement. This is a case where a centralized routing computation is probably worthwhile. The new connection topology could be computed based on a snapshot of the network state and the necessary resources set aside in anticipation of the actual reconfiguration. Switching over to the new topology from the old would require coordination of activities among the different switches in the connection. It will be important to develop methods to minimize the temporary disruption that will be caused when the switch is effected.

Because ATM networks support virtual circuits of arbitrary data rate, some new issues arise when selecting a route between two points in the network. Traditionally, networks have sought to spread the traffic among the available routes in order to avoid local congestion. This is exactly the wrong approach for ATM networks however, as it tends to fragment the network bandwidth, leading to unnecessarily high blocking rates for virtual circuits with high data rates. To minimize blocking, we require *packing strategies* that route a virtual circuit along the busiest route that has the capacity to carry it, assuming all other considerations are equal. For example, given a choice of two direct links between the same pair of switches, it makes sense to select the link that yields the tightest packing. The situation is more complicated when the selection is among routes of different length. In this case, there is a trade-off between the route length and the tightness of the packing that must be taken into account. We plan to use the Project Zeus network to obtain an understanding of the bandwidth fragmentation phenomenon and develop packing strategies that reduce blocking due to fragmentation.

9.3. Network Planning and Configuration

Careful capacity planning will be essential to the success of the Project Zeus network. Capacity planning is complicated by the fact that we will not know in advance the typical values for many of the variables that influence the amount of capacity needed in the network. Hence, traffic measurements of the operational network will be crucial to allow us to identify those places where the network capacity does not match the demand and help us understand how to best deploy the available resources to meet the demand.

There is obviously a strong interaction among capacity planning, congestion control and routing. Inefficiencies in congestion control and routing can waste network resources, requiring the installation of additional facilities to achieve satisfactory performance. Conversely, efficient network control mechanisms can allow network planners to delay installation of new facilities, saving money in the process. Hence, all the issues discussed in the last two subsections, concerning the traffic mix and the topology and dynamics of multicast connections affect network planning as well as control. The more knowledge we have about the way the network is used, thc more closely we can tailor the network capacity to the users' needs.

From a cost standpoint, the most crucial part of any network is the part that is closest to the end terminals. Hence, the placement and dimensioning of concentrators requires careful study. Determination of the appropriate concentration ratio is one of the first questions that needs to be addressed. While a 3:1 concentration ratio appears a reasonable starting point, it may in some cases be too large and in other cases too small. The size of the concentrator is one variable that affects the concentration ratio, since the larger the concentrator, the more effective is the statistical sharing among users connected to it. On the other hand, if a concentrator is made too large, it becomes expensive and the distance constraints imposed by the use of twisted pair wiring between the concentrator and the desktop may be difficult to satisfy. Another variable that will affect the concentration ratio is the amount of multicast traffic and in particular whether it is common for users on the same concentrator to be participating in the same multicast. To deal with the uncertain concentration ratio requirements and to satisfy a variety of requirements, the concentrators will be configurable, allowing links to be freely reassigned between user ports and connections to the central switch. This will allow experimentation with different concentration ratios and permit the most economical choice for each situation.

Design of the backbone network requires an understanding of how information flows among the different parts of the campus. While we expect, significant traffic concentrations within departments and schools, we also expect ongoing collaborations between departments to be exhibited in the traffic patterns present in the Zeus network. Direct links should be supported between switches that have a lot of traffic between them, with tandem connections used for traffic that cannot be accommodated on direct links.

The speed with which network planners can respond to changes in traffic demand is limited by budget constraints and the demands on staff time. Hence, it is prudent to install excess capacity when installing fiber optic cable between switches. This typically involves installing cables with extra fibers, in anticipation of future demand. Similarly, it makes sense to install switches that are expandable to sizes well beyond current needs. To achieve the most economical balance, short-term budget constraints must be balanced against anticipated growth. At the same time, planners should be ready to take advantage of technological improvements that can stretch the capacity of existing plant. For example, wave-length division multiplexors might be used in some situations to allow several 620 Mb/s links to be carried on a single fiber, allowing capacity to be added between switches without installation of new cable.

9.4. Network Interoperability

The following subsection discusses the topic of internetworking among diverse subnets including ATM, Ethernet and FDDI. This subsection discusses the problem of interoperability between public and campus ATM networks. While this is simply a special case of the more general problem discussed later, it is an important enough special case that it merits separate consideration. In this subsection, we briefly outline the areas where differences are expected between public and campus ATM networks and describe how they could be resolved.

At the physical level, the Zeus network will match the public ATM networks in transmission speed but not in transmission format. In particular, the public networks will use SONET transmission systems, requiring that the Zeus switches be augmented by a SONET interface when connecting to a public network. This can be done in one of two ways. One is to design a special port controller board that directly supports

SONET transmission instead of the simpler format planned for Zeus. The second option is a separate interface card that converts between the two transmission formats. An STS-3C framer chip is required for the first approach and desirable for the second. While such devices are not generally available at this writing, they are expected to become available within a year.

At the cell level, the Zeus network will match the public ATM networks very closely. The one exception is the inclusion of a resource management field in the Zeus cell, allowing burst identification by the buffer reservation mechanism described earlier. No similar feature has yet been standardized for public networks. Given current standards, the resource management field cannot be carried transparently through a public network virtual circuit, which may limit the ability to carry bursty traffic through the public network. On the other hand, the resource management field could be carried transparently over a public network virtual path. Detailed consideration of these options will be made as the public network standards develop.

There is currently limited activity in the standards community relating to multicast virtual circuits. The multicast model to be implemented in the Zeus network is far more general than anything currently under discussion in the public networks. To simplify interoperability in the early stages, it may be necessary to use the public network in purely a point-to-point fashion. As more extensive multicast capabilities are added to the public network they can be incorporated into the Zeus network.

The signaling standards for public networks will probably be based on an extension of the current narrowband ISDN protocols. We have chosen a different approach in Project Zeus, as, in our view, it is not possible to implement a general multicast call model by such incremental extensions. On the other hand, if the public network is to be used to interconnect campus networks, the Zeus switch software will require some ability to signal using the public network protocols. We plan to add this capability as the standards for public network signaling progress. We expect that prior to the establishment of a full signaling interface, the public network will provide point-to-point virtual paths. These will be used in the early stages to interconnect Washington University's two campuses and to interconnect the Zeus network to remote sites.

9.5. Internetworking

We believe that the future communications environment will continue to be an internet of heterogeneous subnetworks. It will include high speed networks, such as ATM and FDDI and will need to support a variety of applications, some of these requiring high bandwidth with performance guarantees. One of the roles of the ATM network at Washington University will be to serve as the backbone campus network interconnecting a number of FDDI and Ethernet segments. Thus, the communication environment of Project Zeus can be characterized as an internet of ATM, Ethernet, and FDDI networks. In such an environment, our first objective is to extend support for existing protocols, such as TCP/IP across the ATM network in a transparent way. We plan to achieve this using Ethernet and FDDI portals.

Our next objective is to extend the ATM-like access to hosts and their applications directly connected to Ethernet and FDDI in an internet environment. The ATM access can be characterized as connection-oriented with ability to multiplex constant bit rate (CBR) and variable bit rate (VBR) applications and provide performance guarantees. Having this functionality available end-to-end (for example, from FDDI to ATM to FDDI) is crucial for supporting broadband applications and for the success of ATM technology.

We provide this functionality using a novel internet abstraction called the *Very High Speed Internet Abstraction* (VHSI). The VHSI abstraction is aimed at supporting both connection-oriented applications requiring performance guarantees and classical datagram applications. Important components of the VHSI abstraction include a Multipoint Congram-oriented High-performance Internet Protocol (MCHIP, *i.e.*, equivalent of DoD IP), a high-speed gateway architecture, and resource managers [31, 32].

The important internetworking questions to be addressed within the context of VHSI and Project Zeus include the following:

- In the case of a homogeneous ATM network, resource management on a per connection basis is the key to making performance guarantees to applications. An important question is how we can extend the same performance guarantees to applications running on an internet of FDDI, Ethernet, and ATM. Ethernet and FDDI do no internal resource management, and in such cases we have proposed that the gateways provide this functionality. Gateways can keep track of all active connections and their resource usage and also monitor resource availability in the network. In the case of broadcast LANs, the strategy of gateways acting as resource servers can be conveniently implemented, because a gateway can monitor network traffic while keeping a record of active connections and their resource needs. We will use the Zeus testbed to systematically study the effectiveness of this approach on LANs such as Ethernet and FDDI.

- A gateway is the physical device that interconnects component networks and implements the internet functionality. The important goals of a gateway design in the VHSI include the ability to interface with diverse networks that support data rates up to a few hundred Mb/s, to switch input packets with low latency, and to interface with variable number of input ports (2-64). We proposed a candidate gateway architecture that implements the per-packet processing (critical path) in hardware, and implements connection, resource, and route management (non-critical path) in software. The per-packet processing includes standard functions, such as error detection, address translation, and routing, and also functions such as packet sequencing, fragmentation, and reassembly. We have already done a paper design and detailed simulation of a two-port ATM-FDDI gateway [22]. We plan to use the Zeus testbed to demonstrate the viability of our gateway design, and evaluate a number of associated trade-offs.

- The two most important reasons for using a connection in a high-speed (inter)network access protocol such as MCHIP are assistance in resource management and simplification of per-packet processing. However, introduction of a connection abstraction in a protocol adds the complexity of management of the state and resource binding associated with the connection. There are groups that claim that with minor extensions of datagram IP (*i.e.* without introducing the overhead of a connection abstraction), most of the benefits of the resource management and simplicity of per packet processing can be realized. We plan to use the Zeus testbed to compare these two approaches and examine the claims made about these contrasting approaches.

9.6. Remote Visualization and Collaboration

It is becoming increasingly evident that the visualization is indeed a critical tool for discovery and understanding as well as a tool for communication and teaching. New visualization applications have been rapidly emerging with developments in visualization methodology and underlying technologies. There are four components in the visualization process, namely data, computation, display, and user interaction. As long as the locations of these components are not all the same, the need for remote visualization arises. Furthermore, the scientist may wish to do parts of the visualization computation on separate machines in order to distribute the computation load and achieve better performance. Thus, we believe that a significant fraction of practical visualization will be remote visualization.

In collaboration with scientists in the Department of Biology, we have developed a distributed visualization scheme for the examination of how cells move and reorganize in a developing embryo, an optical sectioning microscopy application explained in Section 7.2. For this specific implementation, 3D data sets consist of twenty 2D images at adjacent focal planes spanning the specimen. Such 3D images are collected every two minutes over a period of two hours generating 60 images. Typical image sizes are 256 × 256 × 20 at 2 bytes per voxel, giving a total data size of about 157 Mbytes. Each 2D slice collected by this method is blurred by out of focus light from the surrounding slices. As noted in Section 7.2, removal of the blurring can be accomplished by a variety of methods. The simplest method (used for this prototype testing) thresholds the data at an appropriate value obtained from the data histogram. Rendering of the cleaned up data is at this time accomplished by using a Simulated Fluorescence Process (SFP) algorithm. This algorithm closely resembles a fluorescence process in that it simulates the flow of excitation light through the 3D data and then creates a projection by gathering the fluorescence from each individual voxel. Finally, the projected images are displayed using the X-window system. Each image is also saved at the display stage, and the researcher can choose to view them individually or as an animation sequence.

The four steps described above are naturally implemented as a pipeline of separate modules. We assign each stage of this pipeline to a different machine so that pipelined parallelism can be achieved. While the raw data is collected and stored on a laboratory computer, the data viewing needs to be on the biologist's workstation remote from the laboratory. Thresholding and SFP are also done on different host computers. Each of the modules is designed such that it can run on different computers to better utilize the computing resources. Communication between modules is through a UNIX IPC socket interface on top of TCP/IP over our 10 Mb/s campus network.

Initial experience with the implementation does show that there is speedup when using the pipeline. The speedup can be attributed to two factors: parallelism resulting from pipeline operation and the reduction in disk swapping due to availability of increased memory at different stages and reduced memory requirement at any given stage of the pipeline.

If the work load is carefully partitioned, communication becomes the bottleneck and it will get worse as the data resolution increases. The communication overhead includes moving the data between user and system spaces, protocol processing, and data transmission, including time to recover from errors. As faster networks become available, actual data transmission time will decrease and the rest of the communication overhead will start to dominate, unless appropriate modifications are incorporated.

The important questions that need further investigation to allow efficient remote visualization on high speed networks include the following:

- A visualization application typically involves a large number of data segments of considerable size. These segments have to be exchanged with minimum communication overhead among different host computers involved. This requires carefully engineered flow control to avoid buffer overflow and data loss. The traditional window based flow control with a tight feedback loop is not appropriate because it is complex to implement and optimize and because high speed networks have large bandwidth-delay product (a measure of how many bits can be in the *transmission pipe*) which makes the feedback less responsive. We are exploring an alternate scheme that uses an open-loop rate control and a simple window control with large data segments [18, 33, 34].

- Error control for visualization applications is challenging because their error control requirements lie on a spectrum between those of data and video applications. A typical data application requires error free data exchange which means every packet loss and corruption has to be recovered. A typical video application on the other hand does not require any error control (as long as the error rate is reasonably low) because the subsequent frames overwrite the previous frame before any correction can be made. In the case of visualization and other imaging applications, the error tolerance varies depending on the application and the processing stage. We claim that an adaptive application-oriented error control scheme is needed. Parametric specification of application's error tolerance and design and implementation of an adaptive application-oriented error control strategy in a protocol are topics of further research.

- An interprocess communication paradigm based on either message passing or remote procedure call is not appropriate for remote visualization applications. A message passing paradigm requires the application to deal with flow control, synchronization, and buffer management. This additional complexity in an application is not justified. In the case of a standard remote procedure call, the calling process gets blocked waiting for the results of a remote procedure. The blocking can lead to poor performance. Thus, both these paradigms are not well suited for an application requiring an exchange of a stream of data segments among processes.

We plan to create a prototype implementation of a novel IPC mechanism, called segment streaming, and a transport protocol based on the two stage flow control and adaptive error control mentioned above. We will use the Zeus testbed and a set of remote visualization applications to evaluate and experiment with our proposed solutions.

Computer-based distributed collaboration applications involve a number (two or more) of participants separated by a network working on a single *problem*. Example applications include a team of software engineers developing and debugging a complex software system; a researcher sharing results of his/her computation with colleagues; and a referring physician getting consulting from a radiologist while the patient is undergoing certain tests. We plan to concentrate on collaboration applications that include remote visualization, imaging, and video conferencing as its components. The collaboration paradigm introduces a number of interesting and challenging research questions, as outlined in the following paragraphs:

- When multiple participants are involved, there are issues concerned with concurrency control among participants and maintaining a consistent state of the data space for each participant. For example, consider an application involving multiple participants controlling a distributed computation and observing its visualization output. The concurrency control has to ensure that only one participant gets to *steer* the computation at a given time. Otherwise, computation can become chaotic and results incomprehensible. Clearly, the exact form of concurrency control depends on the application and collaboration paradigm. Similarly, if *what you see is what I see* (WYSIWIS) paradigm is used, then all participant should see almost identical images or image sequences. This requires application or the underlying protocols to present a consistent view of the data space to all participants.

The problems of concurrency control and data consistency have been addressed in the context of concurrent computing on tightly coupled machines, and recently, proposals have been made to adapt these solutions for distributed collaboration environments. It is our goal to evaluate these proposed solutions, and if found inappropriate, design new ones within the context of Project Zeus.

- When multiple information streams (*e.g.*, data, image sequence, video, voice, and control) are part of a single application, operating system and underlying protocols have to provide mechanisms to ensure that these streams are *synchronized* an each stream receives appropriate quality of service. In other words, the goal is to maintain the temporal relationship of various streams as they are carried over the network. Streams can experience different amount of delay through the network because they are routed via different paths or because they are provided different grade of service. For example, the control stream requires every packet loss and packet error to be recovered while the video stream can tolerate some packet errors and loss. The synchronization problem becomes even more complex if multiple participants are involved, because temporal relationship among streams of different participants have to be also maintained.

  A solution to such synchronization problems has many parts; clock synchronization among participating hosts being one of them, lip-synching for video and voice being another one. Clock synchronization in a network can be achieved using Network Time Protocol [25]. The problem of lip-synching has been studied for video conferencing [29]. Even if clock synchronization and lip-synchronization are assumed, there are issues that have to do with operating system support for real time continuous streams. This requires new scheduling and buffer management algorithms that are not typically part of operating systems such as Unix.

- Compression can improve transmission speeds, but its application is limited to circumstances for which decompression can be accomplished concurrently with transmission. Chip sets that operate in accordance with the recommendations of the OSI/CCITT Joint Photographic Experts Group (JPEG) and code and decode individual images at high speed are appearing today. These inexpensive chip sets will soon be followed by other similar chip sets capable of removing interframe redundancy. These compression methods are lossy and useful for viewing images or image sequences in circumstances where quantitative analysis or medical diagnosis is not required. Compression of from one to two orders of magnitude can be obtained depending on the application. Obviously, such a reduction in the storage required for images is important economically and also to achieve improved disk transfer rates.

  Incorporation of one of these chip sets into the medical doctor workstation (MDWS) will be operational by the end of 1991. Studies of high speed lossless decoding for medical images are underway at MIR, but compression for these methods is typically a factor of two or even less for some images. Even so, the use of reversible compression methods can be important in certain cases where other factors like storage device transfer rates are limiting.

In summary, it is our goal to incorporate appropriate concurrency control, synchronization, and compression mechanisms in operating system and/or underlying protocols to effectively support computer based distributed collaboration applications on Zeus.

10. Summary

We have presented a plan for the design, deployment and operation of a high speed campus network at Washington University: Project Zeus. The network is based on fast packet switching technology, developed here during the last several years. We have shown the details of four applications that can benefit greatly by this technology. Each of these applications utilizes multimedia workstations with high-resolution graphics and video. We believe that these applications can show the way to utilize broadband communications technology in medicine, science, education and also, by extension, in business and industry. The technology itself will support aggregate throughput in switches of hundreds of gigabits per second. In phase 1 of Project Zeus port rates of 155 Mb/s will be emphasized, but the technology can be scaled in phase 2 to 620 Mb/s and thereafter up to 2.4 Gb/s. Higher rates will be introduced as the demand arises and the economics permits. We propose to move this technology quickly into an operational setting where the objectives of network use and network research can be pursued. In this setting the network will be available for use by researchers, instructors and students throughout the university and simultaneously serve as a realistic testbed where pressing system issues can be addressed.

11. Acknowledgments

We had much help with the preparation of this white paper from many of our colleagues and coworkers. Michael E. Gaddis and Richard Bubenik prepared the first draft of Section 6. G. James Blaine and R. Gilbert Jost helped with Section 7.1. James McNally and Guru Parulkar drafted Section 7.2. Raymond Arvidson supplied material for Section 7.3 and Davis van Bakergem prepared the first draft of Section 7.4. Martin Dubetz wrote Section 8 and Guru Parulkar prepared Sections 9.5 and 9.6. Louis Antonopoulos drew or redrew all the figures and helped format the text. We greatly appreciate all their help. We also want to express our thanks to Southwestern Bell and NEC America for their generous support of the Fast Packet Project, which has made it possible for us to take this next step toward high speed multimedia networks.

References

[1] Abdel-Wahab, Hussein M., Sheng-Uei Huan, and Jay Nievergelt, "Shared Workspaces for Group Collaboration: An Experiment Using Internet and UNIX Interprocess Communications," *IEEE Communications Magazine*, November, 1988.

[2] Akhtar, Shahid,"Congestion Control in Fast Packet Networks," Masters thesis, Department of Electrical Engineering, Washington University, St. Louis, Missouri, November, 1987.

[3] ANSI T1S1 Technical Sub-Committee. Broadband Aspects of ISDN Baseline Document.T1S1.5/90 -001, June, 1990.

[4] Arvidson, R.E., V.R. Baker, C. Elachi, R.S. Saunders, J.A. Wood, "Magellan: Initial Analysis of Venus Surface Modification," *Science*, 1991.

[5] Barrett, Neil, "Design of a VLSI Packet Switch Element," Technical Report WUCS-88-32, Department of Computer Science, Washington University, St. Louis, Missouri.

[6] Bubenik, Richard and Jonathan S. Turner, "Performance of a Broadcast Packet Switch," *IEEE Transactions on Computers*, January, 1989.

[7] Bubenik, Richard, Michael Gaddis, and John DeHart, "Virtual Paths and Virtual Channels," Submitted for publication.

[8] Bubenik, Richard, John DeHart and Michael Gaddis, "Multipoint Connection Management in High Speed Networks," *Proceedings of Infocom*, April, 1991.

[9] CCITT. Recommendations Drafted by Working Party XVIII/8 (General B-ISDN Aspects) to be approved in 1990, Study Group XVIII-Report R 34, June, 1990.

[10] Cheriton, D.R. and W. Zwaenepoel, "Distributed Process Groups in the V Kernel," *Transactions on Computer Systems*, 3(2):77-107, May, 1985.

[11] Cox, Jerome R., Stephen M. Moore, Robert A. Whitman, G. James Blaine, R. Gilbert Jost, L. Magnus Karlsson, Thomas L. Monsees, Gregory L. Hassen and Timothy C. David, "Rapid Display of Radiographic Images," *Proceedings of SPIE Medical Imaging V*, March, 1991.

[12] DeHart, John, "CMAP/CMIP Scenarios: A Tutorial," Technical Report ARL-90-03, Applied Research Laboratory, Washington University, St. Louis, Missouri.

[13] Dehart, John, "BPN Connection Management Access Protocol Specification," Technical Report ARL-89-06, Applied Research Laboratory, Washington University, St. Louis, Missouri.

[14] DeHart, John, Michael Gaddis, and Richard Bubenik, "BPN Connection Management Access Protocol Specification," Working Note 89-06, Version 2.0, (August, 1991), Applied Research Laboratory, Washington University, St. Louis, Missouri.

[15] Gaddis, Michael,"Prototype Connection Management: a Progress Report," Technical Report ARL-89-01, Applied Research Laboratory, Washington University, St. Louis, Missouri.

[16] Gaddis, Michael, "FPP Packet Formats and Link Level Protocol Descriptions," Technical Report ARL-89-07, Applied Research Laboratory, Washington University, St. Louis, Missouri.

[17] Gaddis, Michael, Richard Bubenik, and John DeHart, "Connection Management for a Prototype Fast Packet ATM B-ISDN Network," *Proceedings of the National Communications Forum*, Vol. 44, pp. 601-608, October 8-10, 1990.

[18] Gong, Fengmin and Guru M. Parulkar, "A Novel Interprocess Communication Paradigm for Remote Visualization,"Submitted for publication.

[19] Haserodt, Kurt and Jonathan Turner, "An Architecture for Connection Management in a Broadcast Packet Network," Technical Report WUCS-87-3, Department of Computer Science, Washington University, St. Louis, Missouri.

[20] Imase, Makoto and Bernard Waxman, "Worst-case Performance of Rayward-Smith's Steiner Tree Heuristic," *Information Processing Letters*, December 8, 1988.

[21] Imase, Makoto and Bernard Waxman, "The Dynamic Steiner Tree Problem," Technical Report WUCS-89-11, Department of Computer Science, Washington University, St. Louis, Missouri.

[22] Kapoor, Sanjay and Guru M. Parulkar, "Design of an ATM-FDDI Gateway," *Proceedings of the ACM SIGCOMM '91*, September, 1991.

[23] Mazraani, Tony and Guru M. Parulkar, "Specification of a Multipoint Congram-Oriented High Performance Internet Protocol," *Proceedings of Infocom '90*, June, 1990.

[24] Mazraani, Tony, "High Speed Internet Protocols and Resource Management in the Ethernet," Masters thesis, Department of Computer Science, Washington University, St. Louis, Missouri, August, 1990.

[25] Mills, D.L., "Internet Time Synchronization: the Network Time Protocol," DARPA Network Working Group RFC 1129, University of Delaware.

[26] Minzer, S.E., "Broadband ISDN and Asynchronous Transfer Mode (ATM)," *IEEE Communications Magazine*, September, 1989.

[27] National Aeronautics and Space Administration, "Geologic Remote Sensing Field Experiment CD-ROM (9 volumes)," USA NASA PDS GR 0001, 1991.

[28] National Aeronautics and Space Administration, "Pre-Magellan Radar and Gravity Data," USA NASA PDS MG 1001, 1991.

[29] Nicolaou, C., "Architecture for Real-Time Multimedia Communication Systems," *IEEE JSAC*, April, 1990.

[30] Oliver, D. (org.) "Background Materials: Workshop on Computer Graphics in the Network Environment," *Proceedings of SIGGRAPH '91*, July, 1991.

[31] Parulkar, Guru M. and Jonathan S. Turner, "Towards a Framework for High Speed Communication in a Heterogeneous Networking Environment," *Proceedings of IEEE INFOCOMM '89*, 1989. Invited for publication in *IEEE Network, The Magazine of Computer Communications*, March,1990.

[32] Parulkar, Guru M. "The Next Generation of Internetworking," *ACM SIGCOMM Computer Communication Review*, January, 1990. Also, Technical Report WUCS-89-19, Department of Computer Science, Washington University, St. Louis, Missouri.

[33] Sterbenz, James and Guru M. Parulkar, "Axon: A High Speed Communication Architecture for Distributed Applications," *Proceedings of IEEE INFOCOM 90*, June, 1990.

[34] Sterbenz, James and Guru M. Parulkar, "Axon: Application Oriented Lightweight Transport Protocol Design," *Proceedings of ICCC*, November, 1990.

[35] Turner, Jonathan S., "New Directions in Communications," *IEEE Communications Magazine*, October, 1986.

[36] Turner, Jonathan S., "The Challenge of Multipoint Communication," *Proceedings of the ITC Seminar on Traffic Engineering for ISDN Design and Planning*, May, 1987.

[37] Turner, Jonathan S., "Fluid Flow Loading Analysis of Packet Switching Networks," *Proceedings of the International Teletraffic Congress*, June, 1988.

[38] Turner, Jonathan S., "Broadcast Packet Switching Network," United States Patent #4,734,907, March, 1988.

[39] Turner, Jonathan S., "Design of a Broadcast Packet Network," *IEEE Transactions on Communications*, June, 1988.

[40] Turner, Jonathan S., "High Speed Data Link," United States Patent #4,829,227, May, 1989.

[41] Turner, Jonathan S., "Buffer Management System," United States Patent #4,849,968, July, 1989.

[42] Turner, Jonathan S., "Resequencing Cells in an ATM Switch," Technical Report WUCS-91-21, Department of Computer Science, Washington University, St. Louis, Missouri.

[43] Turner, Jonathan S., "A Proposed Bandwidth Management and Congestion Control Scheme for Multicast ATM Networks," Technical Report WUCCRC-91-1, Computer and Communications Research Center, Washington University, St. Louis, Missouri.

[44] Valdimarsson, Einir, "Design of an Eight-Bit VLSI Packet Switch Element," Technical Report WUCS-88-23, Department of Computer Science, Washington University, St. Louis, Missouri.

[45] Waxman, Bernard, "Thesis Proposal: Routing of Multipoint Connections," Technical Report WUCS-87-2, Department of Computer Science, Washington University, St. Louis, Missouri.

[46] Waxman, Bernard, "Probable Performance of Steiner Tree Algorithms," Technical Report WUCS 88-4, Department of Computer Science, Washington University, St. Louis, Missouri.

[47] Waxman, Bernard, "Routing of Multipoint Connections," *IEEE Journal on Selected Areas of Communications*, December, 1988.

[48] Waxman, Bernard, "New Approximation Algorithms for the Steiner Tree Problem," Technical Report WUCS-89-15, Department of Computer Science, Washington University, St. Louis, Missouri.

[49] Waxman, Bernard, "Evaluation of Algorithms for Multipoint Routing," Doctoral thesis, Department of Computer Science, Washington University, St. Louis, Missouri, August, 1989.

[50] Yavatkar, R.S., "Communication Support for Collaborative Multimedia Applications," Technical Report 181-91, Department of Computer Science, University of Kentucky.