MAGNET II: A Metropolitan Area Network Based on Asynchronous Time Sharing

AUREL A. LAZAR, SENIOR MEMBER, IEEE, ADAM T. TEMPLE, MEMBER, IEEE, AND RAFAEL GIDRON

Abstract—**MAGNET II is a testbed for integrated networks. It was designed and implemented based on requirements of real-time network management and control. Quality of service requirements explicitly appear in the design specifications of the network on the media access level. Switching is based on the concept of Asynchronous Time Sharing. The core of the network distinguishes between three traffic classes. The quality of service of these classes is monitored and controlled by the Traffic Control Architecture of the network. Monitoring is supported by observation units distributed throughout the system. The main network resources: switching bandwidth, communication bandwidth, and buffer space, are observable and controllable.**

I. INTRODUCTION

THE architecture of the next generation of integrated packet switching networks will be based on requirements of network management and control [10]. This architecture will incorporate traffic and fault sensing devices on the media access level that will be used for real-time monitoring. Examples are sensors that monitor: the traffic load on a set of buffers, the number of packets blocked by a buffer management system, or the packet delay at congestion points in the network. This architecture is also expected to provide prediction and control capabilities on the media access level. Examples include the prediction of traffic fluctuations and the dynamic allocation of switching and communication bandwidth by switch and link schedulers to different classes of traffic. There are several driving forces for this generation of networks. The one highlighted here is quality of service guarantees.

To this end, we have designed and implemented a network testbed for Integrated Metropolitan Area Networks, called MAGNET II. MAGNET II supports both isochronous and nonisochronous traffic such as video, voice, data, graphics, and facsimile. It has been developed within the framework of a network architecture model called the Integrated Reference Model [4], [7]. Its system architecture consists of a set of switching nodes that are interconnected via point-to-point links in a mesh topology. The current implementation contains three nodes in a fully connected mesh.

Manuscript received September 22, 1989; revised February 15, 1990. This work was supported by the National Science Foundation under Grant CDR-84-21402. This paper was presented in part at the IEEE International Conference on Communications, Boston, MA, June 11-14, 1989, and at the IEEE Global Telecommunications Conference, Dallas, TX, November 27-30, 1989.

The authors are with the Center for Telecommunications Research, Columbia University, New York, NY 10027-6699.

IEEE Log Number 9036225.

The testbed was designed and implemented according to the concept of Asynchronous Time Sharing (ATS). This concept is based on a multiclass network model and asynchronous algorithms for resource allocation. The class is an abstract concept that is specified through delay and loss characteristics. The requirement made on the resource sharing mechanisms is to guarantee the appropriate quality of service for each traffic class.

The concept of quality of service explicitly appears in the design specifications of the network on the media access level. This represents the main novel feature of the network. The core of the network makes a distinction between three traffic classes. Therefore, one of the main assumptions of ATM networks, that the core of the network does not distinguish between different traffic classes, was dropped. This was necessary in order to meet the quality of service guarantee.

The quality of service is monitored and controlled by the Traffic Control Architecture (TCA) of the network. On the media access level, MAGNET II exhibits constructs that support network control and management functions. Monitoring is supported by observation units distributed throughout the system. Provisions have been made to control the allocation of switching and communication bandwidth to each traffic class. The fairness of resource allocation within the same traffic class can also be controlled. Finally, the size of each buffer associated with a traffic class can be defined.

This paper is organized as follows. In Section II, an overview of the system architecture is given. Section III describes the architecture of the Switching Node. Sections IV, V, and VI present the structure of the Ring Switch Fabric, the Bus Switch Fabric, and the Link Switch Fabric. In Section VII, the network monitoring and control capabilities of MAGNET II are presented. Because of space limitations, certain network details are either briefly summarized or not presented. A complete description of the MAGNET II hardware can be found in [19].

II. THE MAGNET II SYSTEM ARCHITECTURE

MAGNET II is a testbed for Integrated Metropolitan Area Networks (IMAN's). Its architecture consists of a communication system that is controlled and managed by a distributed, knowledge-based traffic control system called WIENER [6], [1], [15], [4], [5]. The network architecture was developed under the framework of the In-

tegrated Reference Model (IRM) [4], [7] to support isochronous and nonisochronous traffic and is based on our previous work on ILAN's [9], [17], [20], [14]. (See [18] for a quick reference to our past work.) It satisfies the requirements of an Integrated Metropolitan Area Network (IMAN) and it is ready for field experiments [8].

As already mentioned, the design of MAGNET II was driven by a performance oriented concept called Asynchronous Time Sharing (ATS). This concept is based on a multiclass network model and asynchronous algorithms for allocating network resources. ATS refers to the manner in which scheduling and buffer management resolves contention between the different traffic classes. *Scheduling* consists of switching and communication bandwidth allocation, while *buffer management* refers to buffer space partitioning. ATS calls for dynamic scheduling among the different traffic classes at each contention point in the network. The essential requirement on these resource sharing mechanisms is to guarantee the appropriate quality of service for each traffic class. For each class, the quality of service is defined by a set of attributes such as maximum end-to-end time delay, percentage of contention packet loss, the average number of consecutively lost packets, the minimum average throughput, and the maximum average time delay.

From a user's point of view, the multiclass network model has been implemented with four traffic classes. There are three classes for transporting user information and a fourth class for network management and control. These are called Class I, II, III, and C, respectively. A complete description of the class attributes is given in [10]. The system architecture explicitly allocates resources for only three traffic classes. Class C and I packets share the same switching bandwidth, communication bandwidth, and buffer space. Thus, from the point of view of the system architecture, the network supports only three classes of traffic. (Class C traffic will not be further discussed in this paper.) Priority mechanisms within the same traffic class (as suggested in [10]) have not yet been implemented.

A. The Communication Architecture

The Communication system provides the packet transport path between network users. The architecture of the Communication system is based on a mesh topology in which a set of switching nodes are interconnected via T3/DS3 links (see Fig. 1). Transferring packets between different switching nodes requires that a route be established through the mesh. The Communication system supports several types of routing methods (see Section VI-C).

The structure of the switching node closely follows the basic paradigm described in [10]. It consists of a set of Input and Output Buffers that are interconnected through a switch fabric. The switch fabric suggested in [10] is based on a toroidal structure generated by two rings. The switch fabric implementation for MAGNET II consists of only one such unidirectional ring, called the Ring Switch Fabric.

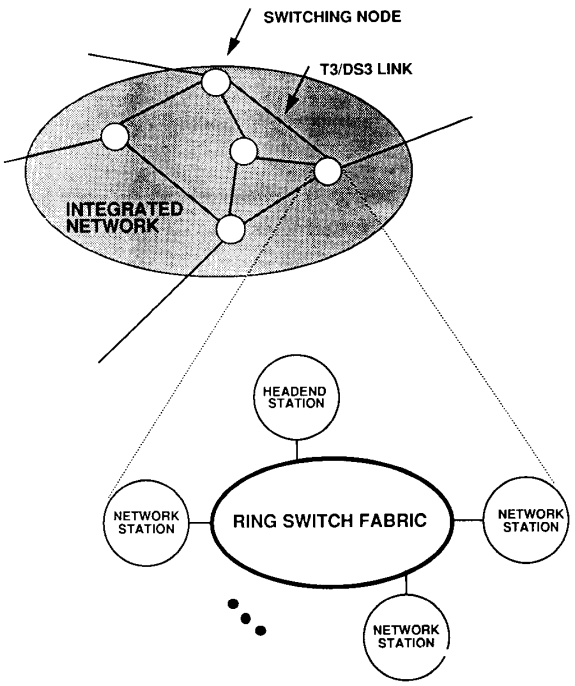


Fig. 1. The system architecture of MAGNET II.

As shown in Fig. 1, a switching node consists of a Ring Switch Fabric that interconnects a set of Network Stations. Both users and the T3 links connect to a switching node through the Network (or Headend) Stations. The Input and Output Buffers are located in the Headend and Network Stations. To support the dynamic scheduling requirement of ATS, each access point to the Ring contains separate Input Buffers for each traffic class.

The main network resources: switching bandwidth, communication bandwidth, and buffer space, are observable and controllable. To clarify our terminology, *switching bandwidth* refers to the bandwidth of the Ring Switch Fabrics, while *communication bandwidth* refers to the bandwidth of the T3 links. Access to switching and communication resources is resolved through a scheduling algorithm based upon time sharing. At each switching node or communication link, the four traffic classes share these resources sequentially in time. Buffer management is achieved via space partitioning. The average throughput and the probability of blocking at the class level are defined as the measure of efficiency of the network. Information regarding the state of the network is furnished on a real-time basis to the traffic control architecture.

B. The Traffic Control Architecture

A general overview of the system architecture of WIENER [6] is shown in Fig. 2. This architecture is embedded into the M-Plane of the Integrated Reference Model [4]. WIENER's system architecture consists of a collection of cooperating expert systems that manage the switching, communication, buffer, and computation re-

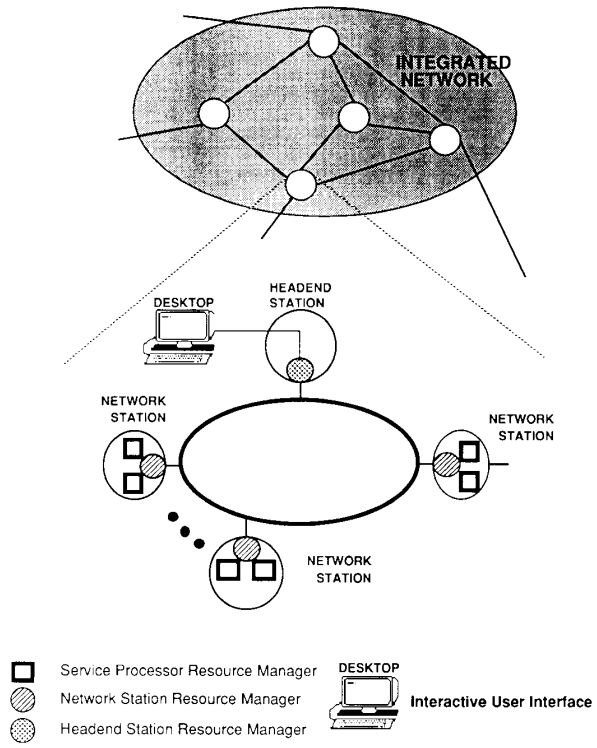


Fig. 2. The system architecture of WIENER.

sources on a networkwide basis. A switching node contains a number of these expert systems, which reside in the Headend and Network Stations (and Gateways [12]). Each of these expert systems is called a resource manager (RM).

The resource managers are responsible for the collection and aggregation of data and statistics generated by the observation units, exchange of information among themselves, and resource allocation and control. They represent an integrated collection of a set of knowledge specialists that are distributed according to their functionality throughout the network [1]. The generic knowledge specialists are: *admission control specialist*, *flow control specialist*, *routing specialist*, and *scheduling and buffer management specialist*. In this paper, only the hardware support for the scheduling and buffer management specialists is described (see Section VII).

A distributed blackboard architecture has been chosen as the basic structure of the statistical database. The statistical database represents a subset of the knowledge database containing all the network objects [15], [16]. Each RM has its own blackboard in the statistical database with possible duplication of information and inconsistencies. The distributed blackboard architecture explicitly represents abstractions of network states and events, knowledge about network behavior, various control problems, and their solutions.

The resource managers comprising the management application process (MAP) coordinate the various knowledge specialists of the network management system by changing the specialist's rules of operation and the basic assumptions about the available information. The four knowledge specialists interact and share information with each other by *writing on* and *reading from* the statistical database. They are not hierarchically structured. A knowledge specialist of a specific type interacts with the corresponding knowledge specialist residing in a different node via *message passing*. This peer-to-peer communication allows independent decision making among different classes of knowledge specialists.

III. The Architecture of a Switching Node

The basic structure of a MAGNET II switching node is shown in Fig. 3. It contains a Ring Switch Fabric (or Ring) that interconnects a set of Network Stations (NS). One station has additional hardware for Ring timing and control and is called the Headend Station (HS). The Ring is implemented with unidirectional 120 Mb/s optical fiber links. Due to transmission encoding, the actual throughput of the fiber links is 100 Mb/s.

The Network Station performs three basic functions. First, it provides an access point to the Ring Switch Fabric for local users. Second, it serves as a local switch that allows local users to communicate with each other without accessing the Ring. Third, it interfaces both local users and the Ring with a T3 link.

A top level diagram of the NS is also depicted in Fig. 3. A Bus Switch Fabric (or Bus), implemented with an industry standard VMEbus, is used to interconnect local users, the Ring, and the T3 link. Local users interface with the Bus either directly through the VME backplane or via an Integrated Network Access (INA). The INA functions as an intelligent multiplexer/demultiplexer that connects multiple users with each other and the Bus Switch Fabric. The Bus communicates with the Ring via the Ring Interface (RI) function. Both the INA and the Link Interface (LI) functions are optional and may not be present in a given Network Station. The NS also contains TCA functions that are not shown in Fig. 3. The Headend Station has the same architecture, although its Ring Interface contains additional hardware.

The Link Interface interconnects the VMEbus of a Network Station with a 45 Mb/s T3 link. The other end of the T3 link connects to the Link Interface of a Network Station which is usually in a different switching node. Thus, using this capability, a set of switching nodes can be interconnected in a mesh topology to form a network.

The Network/Headend Stations are packaged in a standard doublewidth (6U) Euro-card cage with a VME backplane. Thus, the station forms a basic building block that can route and switch among 100 Mb/s, 45 Mb/s, and various user rate input and output streams. These basic blocks can be configured in a variety of network topologies. The choice of the VMEbus provides a wide range of vendor supplied equipment that can be easily integrated into the system architecture.

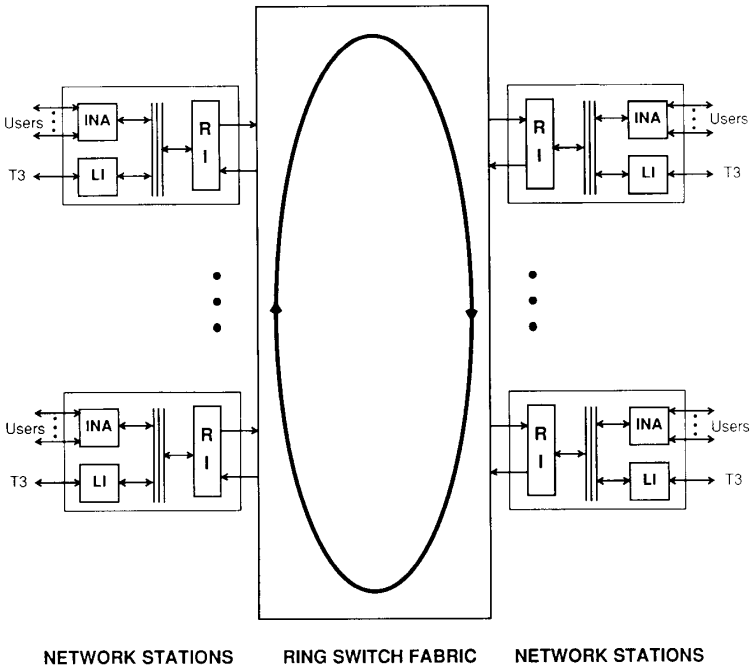


Fig. 3. MAGNET II switching node.

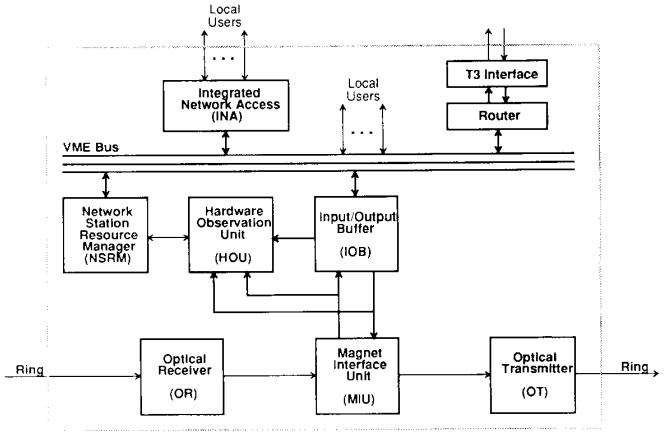


Fig. 4. MAGNET II network station.

A. Network Station

A block diagram of the Network Station architecture is shown in Fig. 4. A brief description of each block is given below.

OR: The Optical Receiver (OR) receives a 120 Mb/s optical fiber input link, performs optical-to-electrical conversion, clock recovery and transmission decoding, and outputs a 100 Mb/s serial data stream and 100 MHz clock to the MIU.

MIU: The Magnet Interface Unit (MIU) converts the 100 Mb/s serial input data into 16 bit parallel words, transmits these words to the IOB, receives 16 bit output words from the IOB, performs parallel-to-serial conversion, and outputs a 100 Mb/s data stream and clock to the OT. The MIU also generates timing and control signals for the IOB.

OT: The Optical Transmitter (OT) receives the 100 Mb/s serial output data and clock from the MIU, performs transmission encoding and electrical-to-optical conversion, and drives the optical fiber output link at 120 Mb/s.

IOB: The Input/Output Buffer (IOB) interconnects the VMEbus with the Ring Switch Fabric. It contains buffers for packets that are being transferred between the Ring and the Bus, controls local access to the Ring, and supports various traffic control functions.

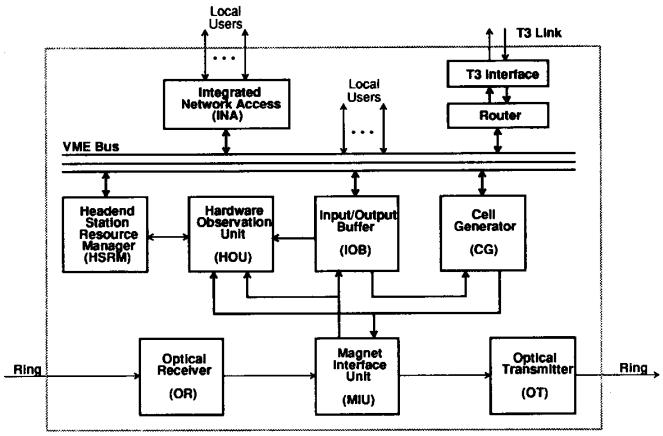


Fig. 5. MAGNET II headend station.

HOU: The Hardware Observation Unit (HOU) monitors the MIU 16 bit input and output interfaces and the IOB buffer status signals, collects traffic statistics, processes the raw data, and delivers the results to the NSRM. The HOU contains an Inmos Transputer processor and communicates with the NSRM via a Transputer serial link.

NSRM: The Network Station Resource Manager (NSRM) is a single board computer, based on the Transputer, with a VME interface. The NSRM is part of the distributed WIENER Traffic Control Architecture and is responsible for controlling the Network Station.

INA: The Integrated Network Access (INA) functions as an intelligent mux/demux that interconnects multiple users with each other and the VMEbus. The INA has a parallel architecture based on the Transputer. Users access the INA via 10 or 20 Mb/s Transputer serial links.

ROUTER: The Router interconnects the VMEbus with a T3 Interface. It contains buffers for packets that are being transferred between the T3 link and the Bus, performs address translation for packets arriving from the link (input packets), and modifies the packet header.

T3I: The T3 Interface (T3I) connects the Router with a 45 Mb/s T3 bidirectional link. It receives MAGNET II packets from the Router, envelopes the packets into the T3/DS3 frame format, and drives the T3 output line. It also receives the T3 input line, deenvelopes MAGNET II packets from the received data, and sends the packets to the Router.

B. Headend Station

A block diagram of the Headend Station (HS) is shown in Fig. 5. It contains the same blocks as a Network Station with the following changes/additions.

- The MIU in the HS contains a 100 MHz oscillator that is the Ring master timing clock.
- The Headend Station Resource Manager (HSRM) replaces the NSRM. The HSRM is implemented with the same single board computer as the NSRM. It is responsible for controlling the Headend Station and is a key part of the WIENER Traffic Control Architecture.
- The HS contains a Cell Generator (CG) module. The CG creates cells on the Ring, synchronizes packets passing through the HS with the output cell timing, and implements the Ring scheduling policy.

IV. Ring Switch Fabric

The Ring Switch Fabric (or Ring) consists of unidirectional optical links connecting a set of stations in a ring topology. Conceptually, the Ring can be viewed as a 100 Mb/s serial data link that starts in the HS, passes sequentially through each NS, and terminates back at the HS. Although it is not explicitly shown in the block diagrams of Figs. 4 and 5, the Ring data path does not pass directly through the MIU. The data path through an NS is OR-MIU-IOB-MIU-OT. For the HS, the path is OR-MIU-IOB-CG-MIU-OT.

Fig. 6 shows a top level view that highlights the packet buffers in the IOB that interface the Ring Switch Fabric with the Bus Switch Fabric in each Network Station. Each IOB contains four separate buffers. The Input Buffer stores packets going from the Ring to the Bus. Three Output Buffers (one for each traffic class) store packets going from the Bus to the Ring.

A. The Basic Transport Mechanism

The transmission time on the Ring is divided into contiguous *cells*. Each cell has a fixed length of 1024 bits. For a 100 Mb/s data rate, the bit interval is 10 ns and a cell has a time duration of 10.24 μs. The basic unit of information transfer on the Ring is called a *packet*. Cells are either *empty* if they are unoccupied, or *busy* if they contain a packet. Stations on the Ring place packets in empty cells and remove packets from busy cells. Cells are marked with an access code that indicates which traffic classes are allowed to use the cell. Packet destinations are determined by an 8-bit address field, allowing a maximum of 256 stations on a Ring. Normally, packets are removed from the Ring by the destination station. A busy cell that becomes empty because a station removed the packet can be immediately reused by the same station.

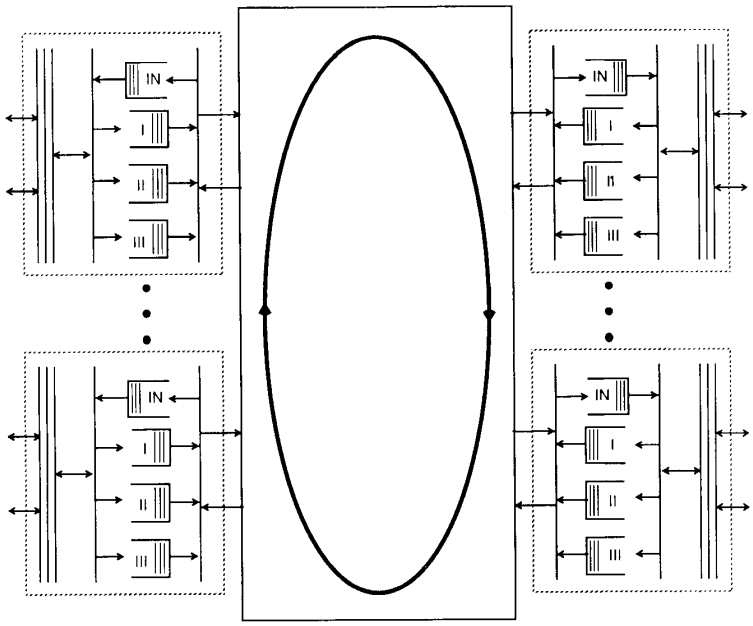


Fig. 6. Ring switch fabric.

Destination removal improves utilization of the switching bandwidth. The increase in the effective capacity depends on the actual connection pattern. For a uniform distribution of the load pattern, the effective throughput can reach up to twice the nominal Ring capacity [20].

The Ring Switch Fabric supports two basic packet types: normal and multicast. A normal packet has a single destination. A multicast packet can have multiple destinations. Instead of a destination field, the header of a multicast packet contains an 8-bit multicast number. Each station contains a multicast address table, located in the IOB, that determines which multicast numbers will be received by the station. Multicast packets are removed by the source station instead of the destination station. Thus, a multicast packet will make one complete trip around the Ring, giving each station an opportunity to receive it. The multicast table in the IOB is controlled by the WIENER system. The NSRM (or HSRM) in each station, via the VMEbus, updates the table as multicast calls are set up/taken down.

The CG module in the HS generates a continuous stream of cells on the Ring. The beginning of each cell is indicated by a 4-bit SYNC pattern, which is followed by an 8-bit Media Access Control (MAC) field. These 12 bits are called the cell header. The remaining 1012 bits can be used to carry packet data. Although not strictly correct, the cell header is generally considered to be part of the packet header. Thus, a full-sized packet is considered to contain 1024 bits.

A cell travels around the Ring and returns to the CG in the HS where it is terminated. If the cell is currently busy, then the packet is removed from the cell, stored in a Transfer Buffer, and inserted into the next new cell generated by the CG. The MAC field contains five parts which are defined below.

- AC: Access Code. The Access Code indicates the current media access procedure (scheduling policy) and which traffic classes are eligible to use the cell.
- CS: Cycle Start. CS = 1 indicates the first cell of a new subcycle.
- BC: Busy Cell. Indicates whether the cell is empty (BC = 0) or contains a packet (BC = 1).
- BR: Busy Record. Indicates whether the cell has been used (BR = 1) or has not been used (BR = 0) during its trip around a Ring.
- T: Transfer. T = 1 indicates that the packet in the cell has passed through the HS.

The MAC field is initialized by the CG when it generates the cell. The AC and CS bits are determined by the Ring Scheduler. When the CG generates an empty cell, the BC, BR, and T bits are initialized to 0. When the CG inserts a packet from the Transfer Buffer into new cell, the BC, BR, and T bits are initialized to 1. If a station on the Ring puts a packet into an empty cell, it sets BC = 1, BR = 1, and T = 0. If a station removes a packet from a busy cell, it sets BC = 0.

When a busy cell is received by the CG, the T bit is checked. If T = 0, the packet in the cell is transferred as described above. However, if T = 1, the packet is discarded. Thus, a packet is allowed to pass through the HS only once. This restriction prevents packets that were not correctly removed from endlessly circulating around the Ring.

MAGNET II supports four different packet sizes: 128, 256, 512, and 1024 bits. As described above, the first 12 bits represent the cell header and cannot be used to trans-

port user data. Regardless of the packet size, only one packet is allowed to occupy a cell. Thus, unless it is the largest size (1024 bits), a packet will not completely fill a cell. The remaining space in a cell is unused. The size of a packet is indicated by the packet size field in the packet header.

When a station receives a packet, the IOB hardware transfers the packet data from the cell to the Input Buffer. When a station transmits a packet, the IOB hardware transfers the packet data from one of the three Output Buffers to the cell. During either operation, the hardware examines the packet size field to determine the actual size of the packet. After the correct amount of data has been transferred, the hardware terminates the buffer read/write signals for the remainder of the cell.

B. Ring Scheduling

The basic scheduling policy implemented by the Ring Switch Fabric is very similar to the general concept described in [10]. The continuous stream of cells generated by the CG is divided into cycles. Each cycle is divided into three subcycles (I, II, III). Thus, when a cell is generated, it is assigned to one of the subcycles. The cell's access code is initialized with the appropriate subcycle indication. This assignment determines which traffic class is allowed to use the cell. Cells belonging to subcycle I can only be used by Class I packets. Subcyle II cells can only be used by Class II packets. Subcycle III cells can only be used by the Class III packets. This restriction does not apply to packets being transferred through the HS. A packet in the Transfer Buffer is placed in the next new cell even if the packet's traffic class does not match the cell's subcycle assignment. However, once this packet is removed, the cell can only be reused by the appropriate traffic class.

The boundaries between subcycles are controlled by the CG using the maximum length moveable boundary scheme described in [10]. The length of a subcycle is measured in cells. The CG contains three variables (MAX I, MAX II, and MAX III) that determine the maximum length of subcycle I, the maximum length of subcycles I and II combined, and the maximum length of the entire cycle. For example, if MAX I = 5, MAX II = 9, and MAX III = 15, then the CG will start a cycle by generating 5 subcycle I cells, followed by 4 subcycle II cells (MAX II-MAX I), followed by 6 subcycle III cells (MAX III-MAX II). After generating the last subcycle III cell, the CG starts a new cycle and repeats this procedure. The values of MAX I, MAX II, and MAX III are determined by the WIENER traffic control system. The HSRM, via the VMEbus, program hardware locations in the CG with the appropriate values.

In addition to the maximum length constraint, the CG uses a moveable boundary procedure. When a cell returns to the CG, the BR bit in the cell header is checked. A BR = 0 indicates that the cell traveled around the entire Ring without being used. If the CG is still in the same subcycle, then it automatically switches to the next subcycle (even though the maximum length constraint has not been reached.) For flexibility, this procedure can be disabled in the CG by the HSRM. If the moveable boundary procedure is disabled, subcycle boundaries are determined only by the maximum length constraint.

As described in [10], the scheduling policy uses an access limit method to fairly allocate the available bandwidth among multiple users. The IOB module in each station on the Ring contains three LIMIT variables (L^I, L^{II}, L^{III}). A LIMIT variable determines the maximum number of packets of the corresponding traffic class that the station can transmit during one cycle. LIMIT variables can be assigned the values 0–255 or NOLIMIT. The NOLIMIT assignment means that the station can transmit an unlimited number of packets of that traffic class. The values of the LIMIT variables in each station are determined by the WIENER system. The NSRM (or HSRM) in each station, via the VMEbus, programs hardware locations in the associated IOB with the appropriate values.

V. Bus Switch Fabric

The Bus Switch Fabric (or Bus) is the backbone of the Network Station architecture. As described in Section III, it interconnects local users, the Ring Interface, and the Link Interface. The Bus Switch Fabric is implemented with an industry standard VMEbus. A detailed diagram of the station architecture was given in Fig. 4.

The local users shown in Fig. 4 are divided into two categories. Those that interface with the Bus directly through the VME backplane are called internal users. Those that interface with the Bus via the INA are called external users. Conceptually, the NSRM/HSRM and the INA are also considered internal users.

The Bus provides a communication path that is used in three ways. First, to transfer MAGNET II packets between internal users, the packet buffers in the IOB, and the packet buffers in the Router. Second, to transfer information that is not in packet format between internal users. Third, to transfer control information between the NSRM/HSRM and the IOB, CG, and Router modules. Normally, most of the Bus traffic falls into the first category.

The goal of the Bus Switch Fabric implementation is to provide a transfer path that is basically transparent. That is, provide a Bus whose bandwidth is sufficiently large so that access queueing delays will be negligible. The VMEbus provides a 32-bit data bus that is capable of transfer rates up to 320 Mb/s. This bandwidth will support the maximum possible combined input/output load to/from the IOB (200 Mb/s) and the Router (90 Mb/s).

A diagram that explicitly shows the packet buffers in the IOB and Router modules is given in Fig. 7. Each buffer can currently hold a maximum of 16 packets. Every buffer has an associated THRESHOLD variable, which can be assigned a value of 2, 4, 8, or 16. The THRESHOLD variable determines the maximum number of packets that are allowed in the buffer. Once the THRESHOLD value

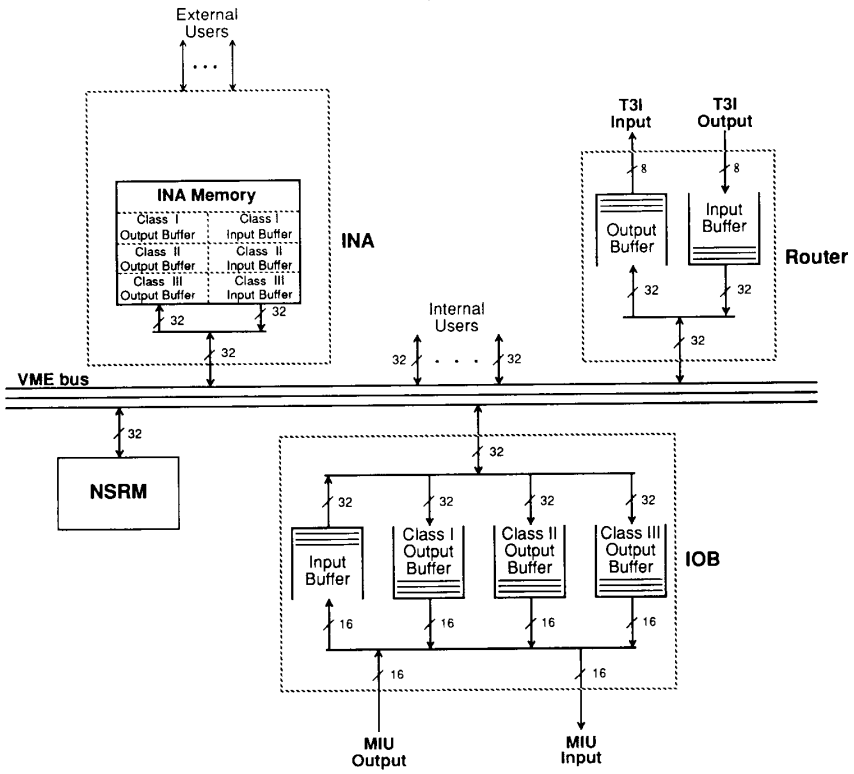


Fig. 7. Bus switch fabric.

is reached, the buffer is considered full and no additional packets will be accepted. The IOB contains four THRESHOLD variables (B^{I}, B^{II}, B^{III}, B^{IN}) and the Router contains two (B^{IN}, B^{OUT}). The values of the THRESHOLD variables for each buffer in a station are determined by the WIENER system. The NSRM (or HSRM) in each station, via the VMEbus, programs hardware locations in the associated IOB and Router with the appropriate values.

Since the Bus Switch Fabric appears as a transparent transfer path, there is no need for scheduling in the manner employed by the Ring Switch Fabric. However, to prevent simultaneous transfers by more than one user, the VMEbus contains an Arbitration subsystem. The Arbiter allocates the Bus to one user at a time using a prioritized or round-robin algorithm. The hardware design of each Output Buffer in the IOB and Router modules also includes a destination scheduler that allocates access to the buffer on a packet basis. Before a user can transfer a packet to an Output Buffer, it makes a request to access the buffer. The associated destination scheduler places the request on an FIFO queue. When a request reaches the head of the queue, the appropriate user is notified via a VME interrupt. The interrupted user now has control of the buffer for one packet transfer. When the packet transfer is completed, the destination scheduler will serve the next request in the queue.

VI. Link Switch Fabric

The Link Switch Fabric consists of a full duplex 45 Mb/s T3 communication link and associated hardware that interconnects the Bus Switch Fabrics of two Network Stations. In general, the Network Stations are considered to be in separate Switching Nodes. The Link Switch Fabric includes the T3 Interface (T3I) and part of the Router (RTR) module in each Network Station, as well as the actual T3 link. A block diagram is given in Fig. 8.

As described in Section III-A, the Router interconnects a VMEbus with a T3 Interface. It contains two packet buffers. The Input Buffer stores packets going from the T3 Interface to the Bus. The Output Buffer stores packets going from the Bus to the T3 Interface. Before placing incoming packets in the Input Buffer, the Router performs address translation and modifies the packet header. The T3 Interface connects the Router with a T3 link. It receives MAGNET II packets from the Router Output Buffer, envelopes the packets into the T3/DS3 frame format, and drives the T3 output line. It also receives the T3 input line, deenvelopes MAGNET II packets from the received data stream, and sends the packets to the Router Input Buffer.

Although MAGNET II defines four different packet sizes (see Section IV-A), the current implementation of the Link Switch Fabric only supports the largest size

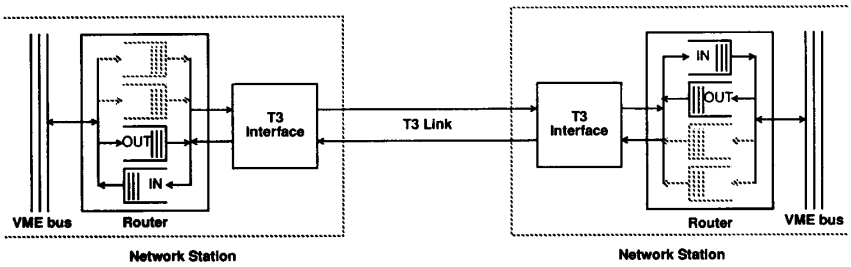


Fig. 8. MAGNET II link switch fabric.

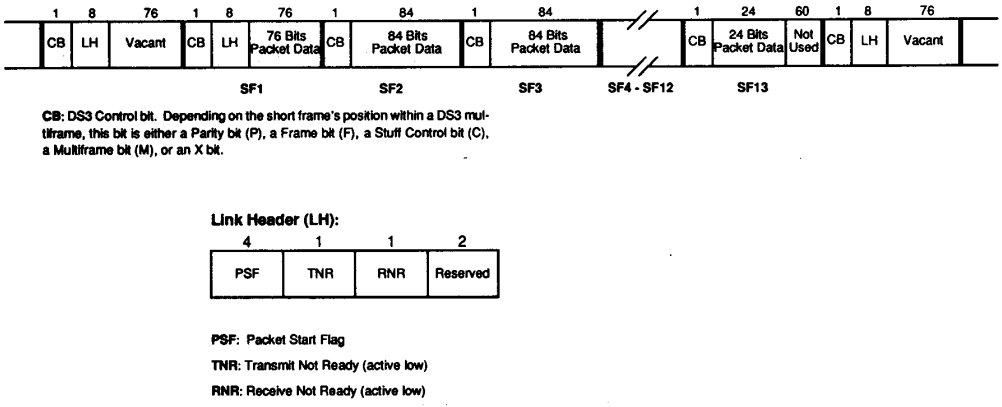


Fig. 9. Mapping MAGNET II packets into the DS3 format.

(1024 bits). Thus, all packets transferred across the Bus Switch Fabric to/from the Router must be 1024 bits.

A. Transmission Format

The DS3 transmission format partitions the bit stream into 85 bit Short Frames. The first bit of each Short Frame is a DS3 control bit. A DS3 Multiframe consists of 56 Short Frames. The definition of the control bit depends on the Short Frame's position within the Multiframe. The remaining 84 bits of each Short Frame can be used for data.

The T3 Interface envelopes MAGNET II packets into the DS3 frame format. The enveloping procedure is based on the Short Frame. A single MAGNET II packet is placed in an integer number of consecutive Short Frames. It requires 13 Short Frames to carry a 1024 bit MAGNET II packet. This procedure is independent of the Multiframe format. That is, the sequence of 13 consecutive Short Frames can begin anywhere within the Multiframe and may even cross the boundary between Multiframes.

When a packet is not available for transmission, the T3I inserts an 8-bit Link Header (LH) into each outgoing Short Frame. The remaining 76 data bits are not used. The Link Header contains a 4-bit Packet Start Flag, 2 Router Status bits, and 2 bits are reserved. In this situation, the Packet Start Flag is XX11, indicating that this is not the start of a packet.

When a packet is ready for transmission, the T3I inserts an LH with a Packet Start Flag of 0000 into the next Short Frame. This indicates that this is the first Short Frame (SF1) of a 13 Short Frame packet envelope. The T3I puts 76 bits of packet data in the 76 remaining bits of SF1. For the next 11 Short Frames (SF2-SF12), the T3I does not insert an LH and uses the entire 84 bits for packet data. For the last Short Frame (SF13), the T3I does not insert an LH, places the last 24 bits of packet data in the SF, and does not use the remaining 60 bits. This procedure is illustrated in Fig. 9.

The actual transmission rate of a T3 link is 44.736 Mb/s. It requires $13 \times 85 = 1105$ bit times to transmit a 1024 bit MAGNET II packet. The efficiency is $1024/1105 = 92.67\%$, which yields an effective data throughput of 41.46 Mb/s.

B. Link Scheduling

The concept for the Link Interface design contained separate Output Buffers for each traffic class and a scheduling scheme similar to the Ring Switch Fabric. However, at the time the Router module was developed, packaging constraints limited the implementation to a single Output Buffer. Thus, in the current design, all packets are served on an FIFO basis and there is no scheduling among the three traffic classes. Due to advances in device technology, a Router design that fully implements the Link Interface concept is now feasible. We plan to upgrade the

Router design in the near future. The new design will include three Output Buffers and a Link scheduler. The additional Output Buffers are shown in Fig. 8 with dashed lines. Each Output Buffer will have an associated THRESHOLD variable (B^I, B^{II}, B^{III}). The Link scheduler allocates bandwidth in the same manner as the Ring scheduler described in Section IV-B. It will divide the link bandwidth into cycles, and each cycle will be divided into three subcycles (one for each traffic class). The scheduler will contain three variables (MAX I, MAX II, MAX III) that determine the maximum subcycle boundary positions. Since a Link has a single access point, there is no need for LIMIT variables. Note that the operation of the Ring scheduler and the Link scheduler are, in general, asynchronous to each other. Thus, in a switching node with multiple T3 links, the Ring scheduler and the various Link schedulers operate independently.

C. Routing

As already mentioned, MAGNET II consists of a set of switching nodes that are interconnected via T3 links in a mesh topology. The path a packet travels through the mesh is called a route. The network supports three methods for determining a packet's route: Datagram (DG), Virtual Circuit (VC), and Multicast (MC). As a packet travels across the mesh, the address fields in the packet header are modified by address translation in the Router to represent the current source station and the next destination station.

For Datagram routing, each packet is treated as a separate entity and may be routed differently from other packets belonging to the same source/destination pair. When a Datagram packet enters a Ring through a Router, address translation uses the final destination address in the packet header to determine the next destination address.

For Virtual Circuit routing, a call set-up procedure initiates the establishment of a virtual circuit between source and destination. The call is assigned a virtual circuit number, which is included in the header of each packet. All packets belonging to the same call (virtual circuit) follow the same route through the network for the duration of the call. When a Virtual Circuit packet enters a Ring through a Router, address translation uses the virtual circuit number in the packet header to determine the next destination address.

For Multicast routing, virtual circuit paths are set up from the source to multiple destinations. Each multicast call is assigned a global multicast number, which is included in the packet header. For each Ring on the multicast route, the global multicast call is assigned a local multicast number. When a Multicast packet enters a Ring through a Router, address translation uses the global multicast number in the packet header to determine the local multicast number.

Address translation is accomplished by a table look-up technique. Each Router contains three look-up tables, one for each routing mode. The translation tables are controlled by the WIENER system. The VC and MC tables are updated as virtual circuit or multicast routes are set up/taken down. The DG tables are changed when network loading conditions require usage of alternate paths. In each station with a Router, the NSRM (or HSRM), via the VMEbus, determines the contents of the translation tables.

VII. Network Monitoring and Control

As already described in Section II-B, the WIENER Traffic Control Architecture follows the structure of a distributed feedback control system. It measures the traffic characteristics and network performance, and dynamically allocates network resources according to the traffic load and the quality-of-service requirements negotiated during call/session set up. Thus, the network resources are controlled by a set of algorithms that use real-time network observations. In what follows, the hardware support for implementing the control policies of the scheduling and buffer management specialists mentioned in Section II-B is presented.

In Fig. 10, a conceptual diagram of the process of monitoring and control of a switching node on the media access level is depicted. As already discussed in Section III, the switching node consists of a set of Network Stations that are connected to the common bandwidth resource of the Ring Switch Fabric. Monitoring is performed by the Hardware Observation Unit (HOU) in the Network and Headend Stations (see Figs. 4 and 5). Traffic control is exerted by the NSRM's and the HSRM through the set of parameters shown in Fig. 10.

The quality of service is evaluated by monitoring the buffer occupancy distribution, the packet time delay distribution, the packet loss, and the gap distribution of the consecutively lost packets [8]. This information is obtained by first recording and then evaluating the state and event variables of the IOB buffers. The processed observations are presented by the HOU to the NSRM's and the HSRM as part of the statistical database. These data are subsequently used by the scheduling and buffer management knowledge specialists to determine the current control policy. The control policy is exerted through a set of parameters that determine the amount of the switching bandwidth, communication bandwidth, and buffer space allocated on each traffic class.

A policy is implemented through programming the corresponding hardware locations. Switching bandwidth allocation is controlled by a knowledge specialist residing in the HSRM. Communication bandwidth allocation is exerted by a knowledge specialist residing in the NSRM associated with the corresponding link scheduler. A knowledge specialist on the same NSRM also sets up the routing tables as part of a distributed routing algorithm.

The WIENER Traffic Control Architecture contains monitoring and control agents. Section VII-A describes

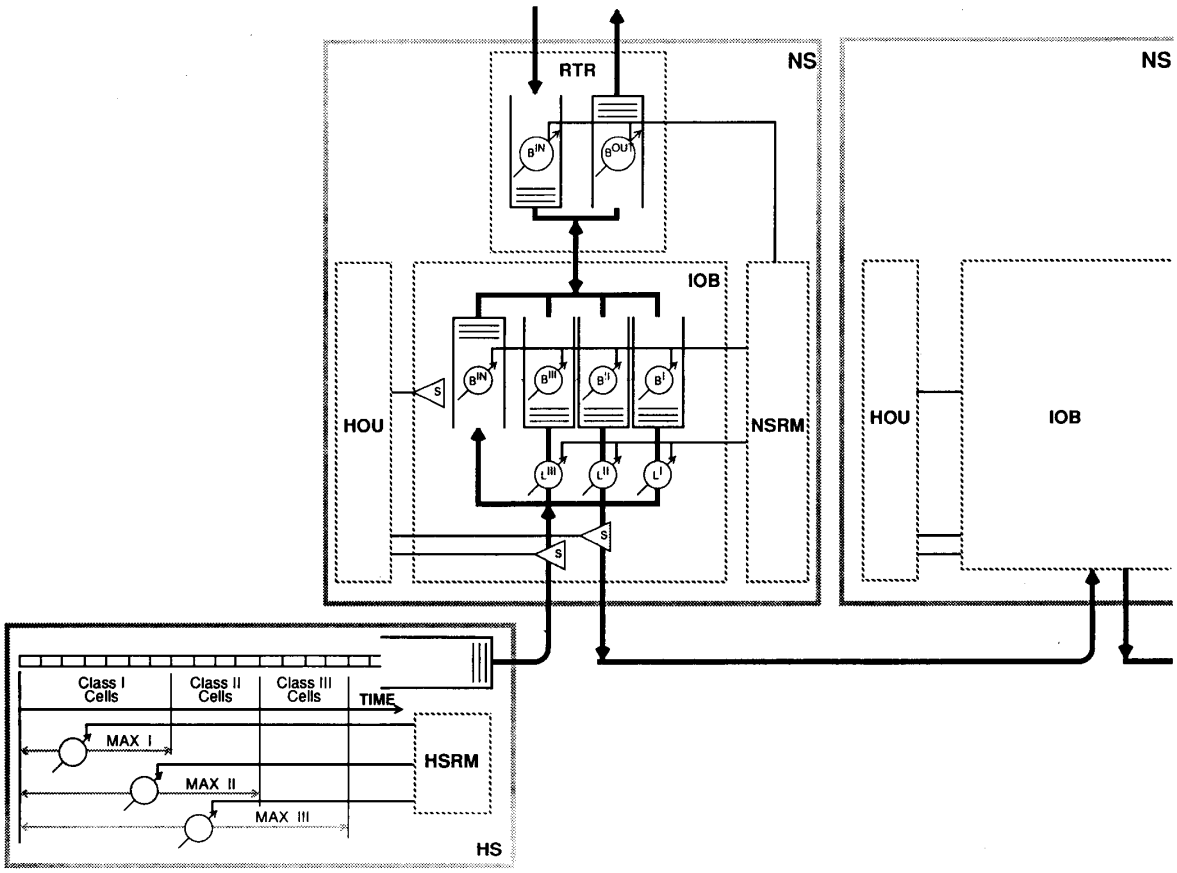


Fig. 10. Monitoring and control of a switching node.

the Hardware Observation Unit that is used for monitoring, while Section VII-B contains a description of the control parameters and their operation.

A. Monitoring

Network activity in a switching node is monitored by the HOU of each station. The HOU has sensors that are shown in Fig. 10 as triangles marked with an S. The HOU can observe the Ring data path as it enters and leaves the station and the state of the packet buffers in the IOB. The data path sensors allow the HOU to extract cell and packet information as cells pass through the NS. Cell headers contain the Media Access Control (MAC) field described in Section IV-A. Busy cells contain packet headers which provide source, destination, type, size, and class information. The HOU uses these data to derive the status of Ring activity and traffic profile. The IOB buffer sensors monitor packet arrivals and departures at each buffer. These data allow the HOU to extract buffer status information such as arrival and departure rates, average occupancy, and waiting times, etc.

The HOU contains three functional blocks: Acquisition Unit, Processing Unit, and Storage Unit. The Acquisition Unit captures headers and buffer state transition signals, and packs them into record formats (I-Records). These records are placed in the Storage Unit. The Storage Unit provides RAM for the HOU. Memory is split into two areas: one for the Processing Unit code and data, and the other for the storage of the I-Records. The Storage Unit memory is dual ported. The I-Records are written by the Acquisition Unit and, at the same time, can be read and manipulated by the Processing Unit. The Processing Unit consists of a 32-bit Transputer (T800). The Processor Unit controls the Acquisition Unit, reads the History Buffer, and executes statistical evaluation algorithms.

The HOU is designed to operate in two acquisiton modes.

- *Continuous Mode (CM):* The HOU records an I-Record for every cell that passes through the station. Each I-Record is identified with a cell number that is modulo 64K. Using this mode, all network activity at the local station can be observed. This makes it easy to determine, for example, the bandwidth available on the Ring to a Network Station.
- *Event Mode (EM):* The HOU stores I-Records only when there is a local event. A local event occurs when-

ever a buffer changes status, that is, when any of the buffers in the IOB has an arrival or a departure. In this way, infrequent events can be recorded over long periods of time.

B. Control

The NSRM's and HSRM determine the allocation of switching bandwidth, local access, and buffer space by assigning values to a set of control variables. This is accomplished by programming hardware locations in the IOB, CG, and Router via the VMEbus. There are three types of control variables in the network: subcycle boundaries, LIMITS, and THRESHOLDS.

The subcycle boundary variables define the maximum length of each subcycle as described in Section IV-B. There are three variables (MAX I, MAX II, MAX III) that control the maximum length of subcycle I, the maximum length of subcycles I and II combined, and the maximum length of the entire cycle. The values of these variables determine the total amount of switching bandwidth that is allocated to each traffic class. This allocation can be modified on a cycle basis by the moveable boundary procedure described in Section IV-B. The three MAX control variables are shown in Fig. 10 as part of the Headend Station.

The LIMIT variables determine the access bandwidth allocation to different traffic classes that belong to the same Network Station. As discussed in Section IV-B, each station is assigned a set of three LIMIT variables L^{I}, L^{II}, and L^{III}. These variables determine the maximum number of packets each station can transmit during one cycle and the subdivision of this bandwidth among the traffic classes. The controls are shown in Fig. 10 (as part of the IOB) as circles marked with the name of the corresponding variables.

The size of the buffer pools in the IOB and in the Router can be defined. The concept of space partitioning as proposed in [10], however, is only partially implemented through the use of THRESHOLD variables. These variables control the effective size of the associated buffer. The IOB contains four THRESHOLD variables (B^{I}, B^{II}, B^{III}, B^{IN}) and the Router contains two (B^{IN}, B^{OUT}). These controls are shown in Fig. 10 as circles marked with the corresponding variables.

VIII. Conclusions

In this paper, the system architecture of MAGNET II, a testbed for Integrated Networks supporting isochronous and nonisochronous traffic, has been presented. The system architecture of the network has been described with an emphasis on the built-in foundations for a Traffic Control Architecture.

The network has been designed based upon the concept of Asynchronous Time Sharing, and it supports four traffic classes. The core of the network, as currently implemented, distinguishes between three of these classes. In [10], four traffic classes, each with up to four priority levels, have been proposed. In view of the experience we have gained with the exploratory design and implementation of MAGNET II, this choice in modeling future networks appears to represent a good compromise between implementation complexity and a reasonable set of quality of service parameters.

The traffic classes represent an abstract performance oriented concept that can support services such as video, voice, data, graphics, and facsimile. The decision regarding information transport and its associated quality of service remains with the users. These further add in flexibility of network usage should new services arise [11], [2], [3], [13].

A natural generalization of the current implementation that leads to an increase in the total throughput of the switch fabric can be achieved with a topological extension of the ring-type structure to a toroidal or shuffle topology [10]. In addition to the large bandwidth available, the advantage of these architectures is technological. It promises a smooth transition into an *integrated* electrooptical domain.

Acknowledgment

The work presented here involved many researchers at the Center for Telecommunications Research. In particular, we would like to acknowledge the contributions of E. Cadorin, F. Kamikawa, G. Pacifici, and J. S. White.

References

[1] J. T. Amenyo, S. Mazumdar, and A. A. Lazar, "Modeling knowledge-based resource management and control on MAGNET II," in *Proc. GLOBECOM'87*, Tokyo, Japan, Nov. 15–18, 1987, pp. 9.6.1–9.6.5.

[2] T.O. Brunner, and J. S. White, "Implementation of packet telephone and video services on a local area network," Center Telecommun. Res., Columbia Univ., New York, CTR Tech. Rep. 103-88-42, Aug. 1988.

[3] G. Karlsson and M. Vetterli, "Subband coding of video for packet networks," *Opt. Eng.*, vol. 27, no. 7, pp. 574–586, July 1988.

[4] A. A. Lazar, "Object-oriented modeling of the architecture of integrated networks," Center Telecommun. Res., Columbia Univ., New York, CTR Tech. Rep. 167-90-04, Jan., 1990.

[5] ——, "The game of networking," Center Telecommun., Res., Columbia Univ., New York, CTR Tech. Rep. 200-90-37, July 1990.

[6] A. A. Lazar, J. T. Amenyo, and S. Mazumdar, "WIENER: A distributed expert system for dynamic resource allocation in integrated networks," in *Proc. IEEE Symp. Intelligent Contr.*, Philadelphia, PA, Jan. 18–20, 1987, pp. 159–164.

[7] A. A. Lazar, M. A. Mays, and K. Hori, "A reference model for integrated local area networks," in *Proc. IEEE Int. Conf. Commun.*, Toronto, Canada, June 22–25, 1986, pp. 531–536.

[8] A. A. Lazar, G. Pacifici, and J. S. White, "Real-time traffic measurements on MAGNET II," *IEEE J. Select. Areas Commun.*, vol. 8, Apr. 1990.

[9] A. A. Lazar, A. Patir, T. Takahashi, and M. El Zarki, "MAGNET: Columbia's integrated network testbed," *IEEE J. Select. Areas Commun.*, vol. SAC-3, pp. 859–871, 1985.

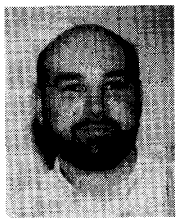
[10] A. A. Lazar, A. Temple, and R. Gidron, "An architecture for integrated networks that guarantees quality of service," *Int. J. Digital and Analog Commun. Syst.*, vol. 3, no. 2, 1990.

[11] A. A. Lazar and J. S. White, "Packetized video on MAGNET," *Opt. Eng.*, vol. 26, no. 7, pp. 596–602, July 1987.

[12] S. Q. Li, M. J. Lee, H. C. Chen, and A. A. Lazar, "An ILAN-ISDN gateway," in *Proc. IEEE Int. Conf. Commun.*, Philadelphia, PA, June 1988, pp. 4.4.1–4.4.6.

[13] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, pp. 834–844, July 1988.
[14] A. M. Mays, K. Hori, and A. A. Lazar, "MAGNET's integrated network architecture," in *Proc. MILCOM'86*, Monterey, CA, Oct. 5-9, 1986, pp. 1.3.1-1.3.6.
[15] S. Mazumdar and A. A. Lazar, "Knowledge-based monitoring of integrated networks," in *Proc. 1st Int. Symp. Integrated Network Manage.*, Boston, MA, May 14-17, 1989, pp. 235-243.
[16] ——, "Monitoring integrated networks for performance management," in *Proc. IEEE Int. Conf. Commun.*, Atlanta, GA, Apr. 15-19, 1990, pp. 289-294.
[17] A. Patir, T. Takahashi, Y. Tamura, M. El Zarki, and A. A. Lazar, "An optical fiber-based integrated LAN for MAGNET's testbed environment," *IEEE J. Select. Areas Commun.*, vol. SAC-3, pp. 872-881, vol. SAC-3, 1985.
[18] S. R. Sachs, "Alternative local area network access protocols," *IEEE Commun. Mag.*, vol. 26, pp. 25-45, Mar. 1988.
[19] A. Temple, R. Gidron, and A. A. Lazar, "The hardware architecture of MAGNET II," Center Telecommun. Res., Columbia Univ., New York, Tech. Rep. 120-88-48, Oct. 1988.
[20] M. El Zarki, A. A. Lazar, A. Patir, and T. Takahashi, "Performance evaluation of MAGNET protocols," in *Local Area & Multiple Access Networks*, R. L. Pickholtz, Ed. Rockville, MD: Computer Science Press, 1986, pp. 137-154.

Aurel A. Lazar (S'77-M'80-SM'90), for a photograph and biography, see p. 483 of the April 1990 issue of this JOURNAL.



Adam T. Temple (S'87-M'87) received the B.S. degree in engineering from Yale University, New Haven, CT, in 1979 and the M.S.E.E. degree from Columbia University, New York, NY, in 1987.

From 1979 to 1985, he designed high-speed signal processing systems for sonar applications at Raytheon's Submarine Signal Division. Since 1985 he has been a member of the Research Staff at the Center for Telecommunications Research at Columbia University. His major technical activities there have been in the areas of broad-band integrated networks and computer communication networks.

Rafael Gidron was born in Tel Aviv, Israel, on March 3, 1958. He received the B.S. degree (magna cum laude) from Tel-Aviv University in 1985, the M.S. degree from Columbia University, New York, NY, in 1987, and is currently in the Ph.D. program at Columbia University.

He has been a Research Staff member at the Center for Telecommunications Research since 1985, working in system and hardware design of broad-band integrated networks. Before joining the CTR, he worked with the Israeli Defence Forces, Tel-Aviv University, and Simteck in Israel and with Renecentrallen in Denmark. His current interests include hardware architecture and resource allocation in broad-band integrated networks, as well as digital signal processing and multiprocessing architecture.