Quick Guide
IBM Multiprotocol Switched Services (MSS) Release 2.0/2.0.1

Document Number TR 29.xxxx

Dee Vig

Multiprotocol Switching Services Development
Networking Division
International Business Machines Corporation
Research Triangle Park, N.C.

http://www.networking.ibm.com/nes/nesswitc.htm

January 22, 1998

ABOUT THE AUTHOR

Dee Vig was the chief architect for MSS Release 2.0 and 2.0.1. Dee is a Senior Engineer at IBM's Networking Hardware Division and has 13+ years of experience in product development, systems engineering and marketing. Dee may be contacted via the Internet at dvig@us.ibm.com.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

REFERENCES

[1] C. Alexander, *et al.*, *Quick Guide to the IBM Multiprotocol Switched Services (MSS) Server Release 1.0*, Tech. Rep. TR 29.2170, IBM, Aug. 1996 (available from IBM Networking Technical Reports web page at *http://www.raleigh.ibm.com/tr2/tr2over.html).*

[2] IBM Corp., *IBM Multiprotocol Switched Services (MSS) Server Command Line Interface Volume 1: User's Guide and Protocol Reference*, SC30–3818, January 1997.

[3] IBM Corp., *IBM Multiprotocol Switched Services (MSS) Server Command Line Interface Volume 2: User's Guide and Protocol Reference*, SC30–3819, January 1997.

[4] C. Alexander, *et al.*, *Quick Guide to the IBM Multiprotocol Switched Services (MSS) Server Release 1.1*, Tech. Rep. TR 29.2260, IBM, May 1997 (available from IBM Networking Technical Reports web page at *http://www.raleigh.ibm.com/tr2/tr2over.html*)

[5] ATM Forum, *ATM User Network Interface (UNI) Specification Version 3.0*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] ATM Forum, *ATM User Network Interface (UNI) Specification Version 3.1*, Prentice Hall, Upper Saddle River, NJ, 1995.

[7] ATM Forum, *LAN Emulation Over ATM: Version 1.0 Specification*, AF–LANE–0021.000, Jan. 1995.

[8] J. Heinanen, *Multiprotocol Encapsulation over ATM Adaptation Layer 5*, RFC 1483, July 1993.

[9] M. Laubach, *Classical IP and ARP over ATM*, RFC 1577, Jan. 1994.

[10] M. Laubach, Joel Halpern, *Classical IP and ARP over ATM*, Internet–Draft <draft–ietf–ion–ipatm–classic2–xx>.

[11] J. Luciani, G. Armitage, Joel Halpern, *Server Cache Synchronization Protocol (SCSP),* Internet–Draft <draft–ietf–ion–scsp–03.txt>.

[12] J. Luciani, D. Katz, D. Piscitello, and B. Cole, *NBMA Next Hop Resolution Protocol (NHRP),* Internet–Draft, <draft–ietf–rolc–nhrp–11.txt>.

[13] T. Bradley, C. Brown, and A. Malis, *Multiprotocol Interconnect over Frame Relay,* RFC 1490, July 1993.

[14] J. Case and A. Rijsinghani, *FDDI Management Information Base,* RFC 1512, Sep. 1993.

[15] C. Hedrick, Routing Information Protocol (RIP), RFC 1058, June 1988

[16] G. Malkin, Routing Information Protocol Version 2 (RIP v2), RFC 1723, November 1994

[17] J. Moy, Open Shortest Path First (OSPF), RFC 1131, October 1989

[18] K. Lougheed, Y. Rekhter, Border Gateway Protocol (BGP), RFC 1105, June 1989

[19] G. Malkin, RIP Version 2, RFC 1723, November 1994

[20] S.Deering, Host Extensions for IP Multicasting, RFC 1112, August 1989

[21] G.Armitage, Support for Multicast over UNI 3.0/3.1 based ATM Networks, RFC 2022, November 1996

[22] D. Waitzman, C. Partridge, S. Deering, Distance Vector Multicast Routing Protocol (DVMRP), RFC 1075, November 1988

[23] J. Moy, Multicast Extensions to OSPF (MOSPF), RFC 1584, January 1997

ACKNOWLEDGMENTS

The following people contributed to this document:

Cathy Cunningham
Rana Dayal
Russell Gardo
Tim Gilbert
Anand Gorti
Amit Kumar
Sanjeev Rampal
Matt Squire
Dee Vig
Colin Verrilli
Rama Yedavalli

Trademarks

The following terms are trademarks of the IBM corporation in the United States or other countries or both:

Advanced Peer–to–Peer Networking

APPN

IBM

NetView

OS/2

The following terms are trademarks of other companies:

AppleTalk	Apple Computer, Inc.
IPX	Novell, Inc.
Microsoft	Microsoft Corporation
Netware	Novell, Inc.
Novell	Novell, Inc.
TME 10	Tivoli Systems Inc.
VINES	Banyan Systems, Inc.
Windows	Microsoft Corporation
Windows 95	Microsoft Corporation
Windows NT	Microsoft Corporation
3Com	3Com Corporation

1. Introduction

This document is a quick guide to the new functions in IBM's Multiprotocol Switched Services (MSS) Release 2.0 and 2.0.1. MSS 2.0 was released in October 1997. MSS 2.0.1 replaced MSS 2.0 in December 1997. This document refers to the combined functions of MSS 2.0 and 2.0.1 as MSS 2.0/2.0.1. MSS 2.0/2.0.1 builds upon the functions provided by MSS Server Release 1.0 [1] [2] [3], released in October 1996 and MSS Server Release 1.1 [4], released in April 1997.

A brief summary of MSS Server Release 1.0 and Release 1.1 is provided in sections 1. 2. and 1. 3. below. It is assumed that the reader of this document is familiar with the following networking technologies: routing, bridging, network management, ATM [5] [6], LAN Emulation [7], RFC 1483 encapsulation [8], Classical IP [9], and the Next Hop Resolution Protocol (NHRP) [12].

1. 1. MSS Architecture

Shared media LANs are rapidly being replaced with switched LANs because switched LANs offer very high performance at a relatively low cost. However, when deployed on a large scale, switched LANs run into difficulties. Foremost, to inter–connect switched LANs, a very high speed backbone network is required. Switched LANs also suffer from some of the same scalability limitations as bridged LANs in that they do not control broadcasts, which can cripple large networks if flooded indiscriminately. Traditional routers can segregate layer–2 broadcasts and can be used to inter–connect small switched LANs. However, due to the throughput mismatch between routers and switches, routers can become severe bottlenecks when forwarding traffic between switched networks. This negates the performance advantage of switching. Even if broadcasts were not a problem, protocols like IP require the use of routers to forward inter–network and inter–subnet traffic.

Multiprotocol Switching Services (MSS) is an IBM architecture for providing services in a switched network. MSS services provide a migration path from non–switched to switched networks, exploit the underlying switch fabric, and allow switched networks to scale up. To inter–connect switched LANs, MSS uses an ATM backbone network. ATM can provide enough bandwidth to create large switched networks today and has sufficient scalability to address bandwidth requirements for future growth. ATM is an industry standard and is widely supported by the computer and the telephone industry. Most LAN switches available today can already switch from LAN to campus ATM networks. While next–generation WAN switches being developed are adopting ATM as the WAN backbone technology of the future, many WAN switches available today can already switch from campus ATM to today's Frame–relay based WANs. Thus ATM meets the bandwidth and inter–operability requirements already. Beyond performance and interoperability, ATM can provide new functions such as guarantees on bandwidth and delay, popularly known as Quality of Service (QOS). ATM QOS makes it possible for advanced applications to reliably transport real–time traffic (like video and audio) across ATM. ATM also provides an opportunity to rethink and restructure the way networks are constructed and the functions they are capable of providing. For example, the physical ATM network can be divided into multiple logical networks and an end station can join multiple logical networks using a single physical ATM interface (see Figure 1. and Figure 2.). This is especially useful for servers in that they can have a direct connection in ***each***

logical network they serve and thus bypass intermediate bridges and routers when sending data to clients. ATM provides an infrastructure in which any device attached to the ATM fabric can be zero hops away even when the device is geographically remote. This presents an opportunity to bypass layer–2 bridges and layer–3 routers during steady state data flows.

MSS provides a smooth migration path to ATM by enabling legacy networking hardware and software to take advantage of high speed ATM backbones. This approach allows users to preserve their current networking infrastructure while taking advantage of ATM's many benefits. Thus, MSS exploits the capabilities of ATM and also extends them beyond ATM to LANs. MSS provides a migration path for LAN based networks by providing LAN emulation on ATM and traditional bridging/routing services in a switched network. LAN Emulation provides familiar ethernet and token–ring LAN interfaces on ATM networks.

The MSS architecture is switch–centric. MSS allows switched networks to scale up by preventing LAN broadcasts from flooding uncontrollably, by distributing services across devices supporting the MSS architecture, and by providing redundancy for critical services. MSS exploits the underlying performance advantage of switches by switching traffic when possible, and routing only if it has to.

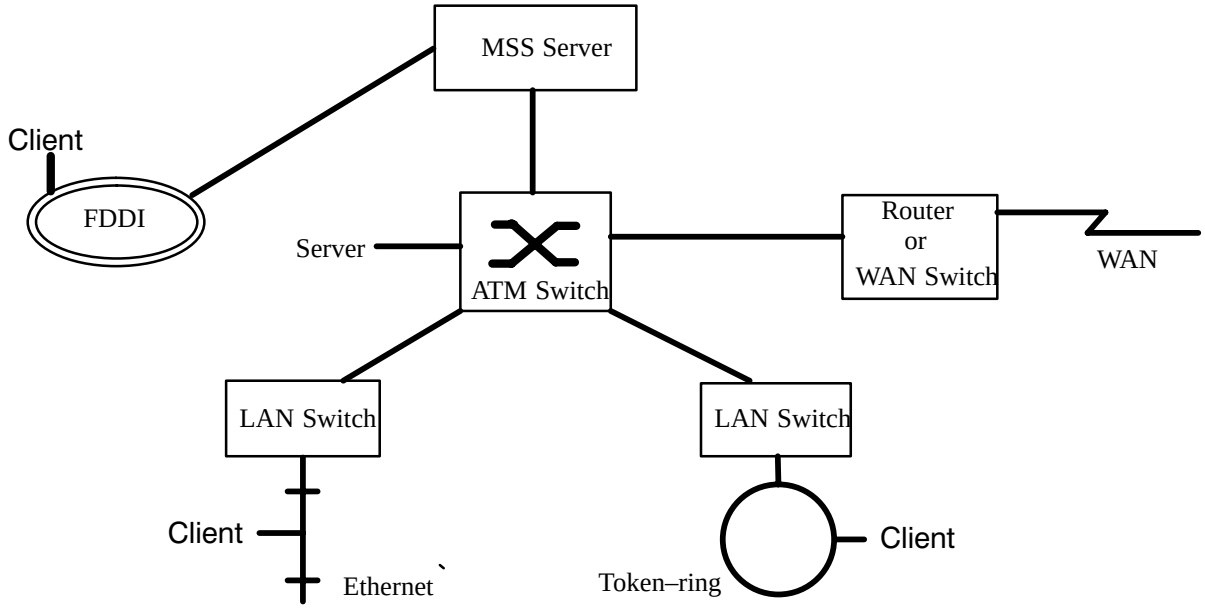

Figure 1. Physical view of a MSS network

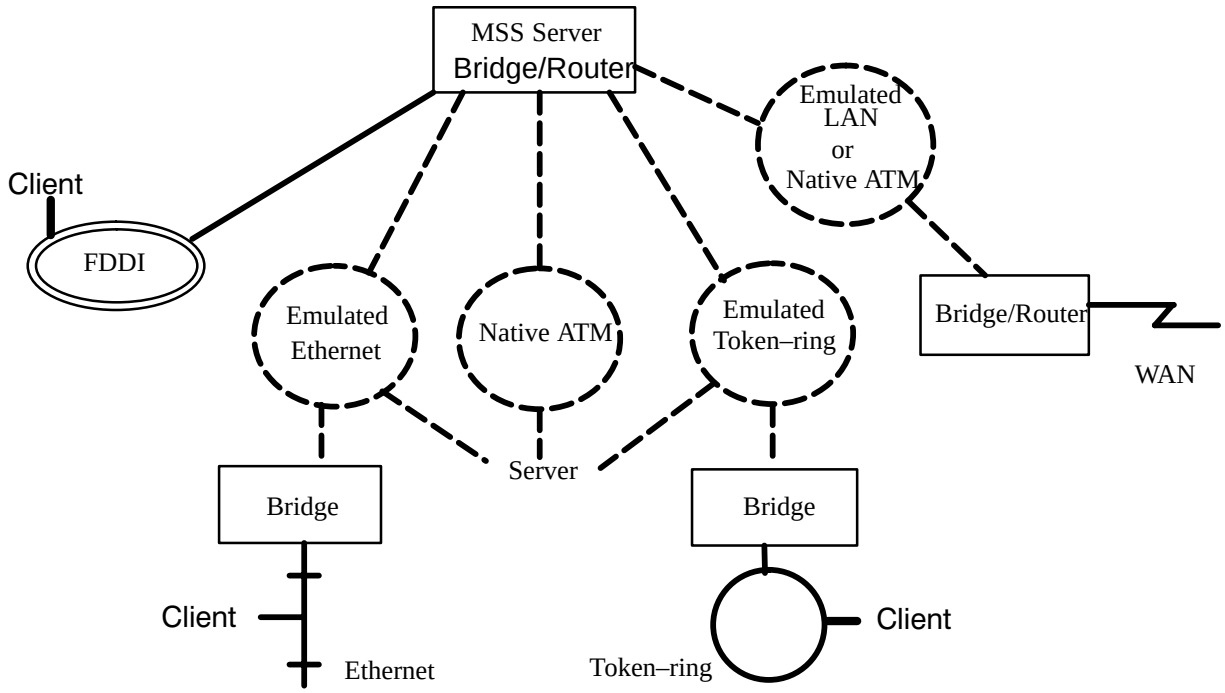

Figure 2. Logical view of a MSS network

Figure 1. shows the physical view of a typical MSS network consisting of ethernet and token–ring LANs connected to an ATM backbone network using LAN Switches. Physical connectivity to FDDI LANs is provided by the MSS Server but it could also be done using a FDDI to ATM switch or bridge. Figure 2. shows a logical view of the same MSS network. The MSS Server provides traditional bridging and routing services on a switched network but with additional enhancements to exploit the underlying switch fabric. Enhancements like SuperElan bridging provide traditional bridging service during the initial flow but the bridge is bypassed during steady state forwarding. End stations then communicate ***directly*** using a switched connection. Similarly, NHRP and Route Switching are used to bypass intermediate IP routers and provide zero–hop IP routing. To control broadcasts in a switched network, MSS monitors and learns protocol–specific broadcasts. It can then change subsequent broadcasts into unicasts based on what has been learned. This not only reduces network traffic, it also reduces unnecessary interruptions to end–stations. Another form of broadcast control is provided by Virtual LANs (VLANs). VLANs dynamically partition a switched network into multiple protocol and subnet–specific broadcast domains. This partitioning is similar to what is done by traditional routers, however VLAN partitions are created dynamically and they are not restricted by physical location as is the case with traditional routers. This makes it possible to move end stations within the switched network without reconfiguration.

In addition to LAN Emulation, native ATM support is provided for layer–3 protocols like IP (Classical IP), IPX and APPN. This means that these layer–3 protocols are transmitted directly over ATM (layer–1) without using LAN Emulation (layer–2). Native ATM protocols can take advantage of unique ATM features like Quality of Service (QOS), which is not exploited by standard LAN Emulation. MSS can also encapsulate and bridge traffic natively on ATM without using LAN

emulation. This bridging is done using RFC 1483 encapsulation [8], which is widely supported by LAN and WAN switch vendors. RFC1483 based bridging allows interoperation with service provider equipment like ATM to Frame–Relay switches, which do not support LAN Emulation. An example of ATM to Frame–Relay inter–networking using 1483 bridging support in the MSS Server is shown in Figure 3..

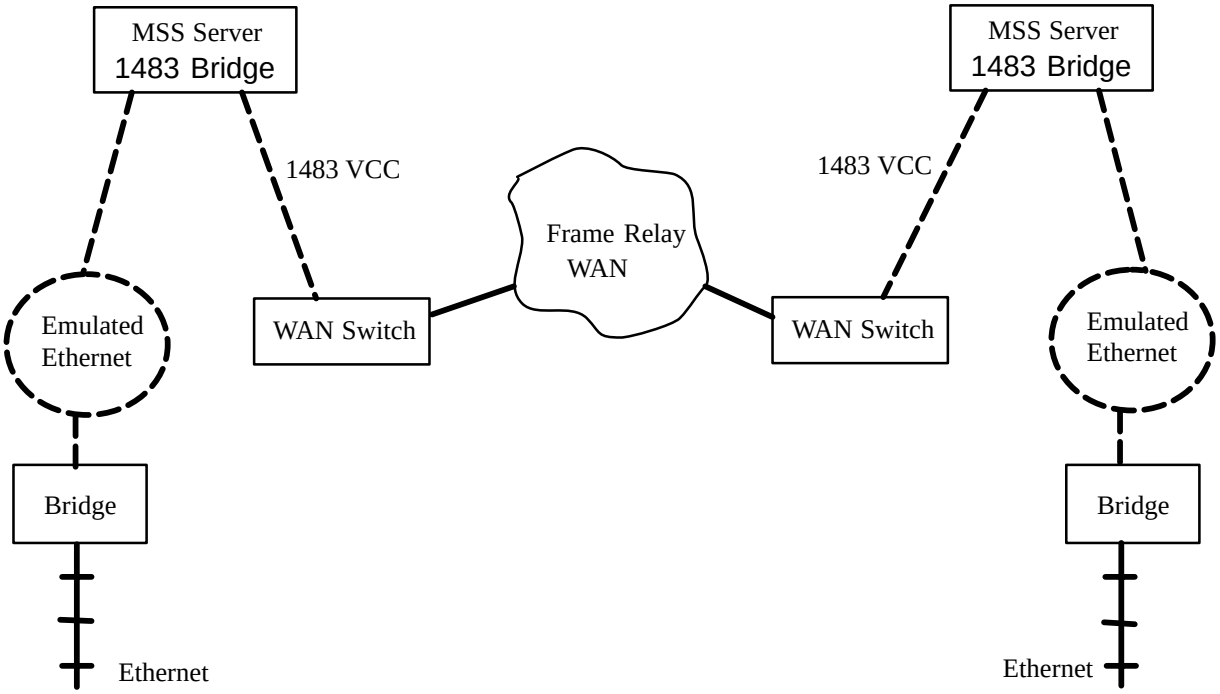

Figure 3. ATM to Frame–Relay Internetworking

MSS benefits are extended to LANs by enabling LAN attached stations with MSS services such as NHRP. By using NHRP in the end stations, layer–3 forwarding is distributed and scalability increases. With NHRP, end–stations get better network throughput by bypassing intermediate routers. This is especially beneficial for servers because they transmit large amounts of information.

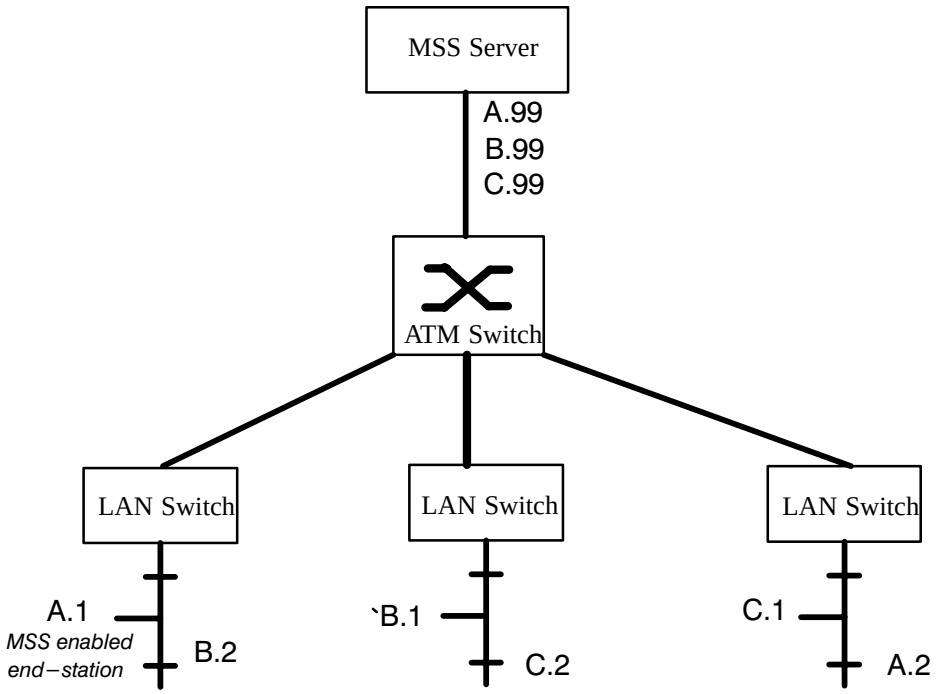

Figure 4. Physical view of a MSS enabled IP network

Figure 4. shows three LAN switches connected to an ATM switched backbone served by an MSS Server. IP stations on three different IP subnets A, B and C are distributed across the three LAN switches. The MSS Server is configured as an IP router to route between the three subnets. IP station A.1 is MSS enabled to do zero–hop routing using NHRP.

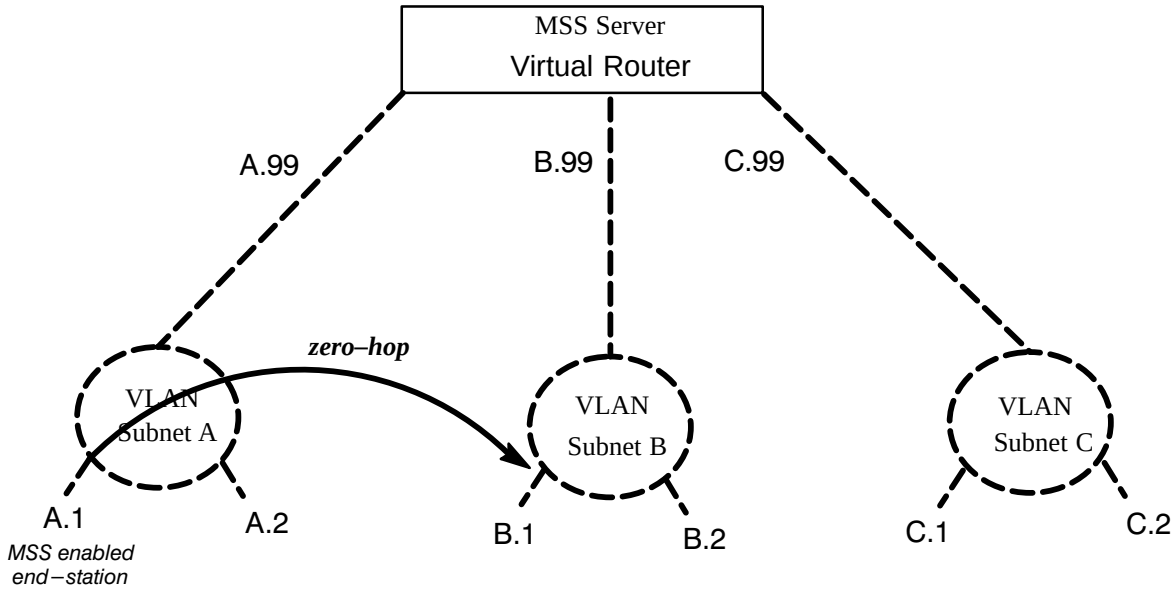

Figure 5. VLAN view of a MSS enabled IP network with zero–hop routing

Figure 5. shows the logical (VLAN) view of the switched network shown in Figure 4.. The network is logically divided into three IP subnet VLANs and the MSS Server is the virtual router connecting the VLAN subnets. Normal, inter–subnet traffic passes through the virtual router in the MSS Server (one hop). However, traffic from MSS–enabled stations like A.1 bypass the MSS Server (zero–hop). This is possible because A.1 receives shortcut information from the MSS Server to send data directly to a destination on another subnet.

1. 2. Summary of MSS Server Release 1.0

MSS Server R1.0 was released in October 1996. A brief summary of the MSS Server R1.0 hardware and software is provided in this section. For details, refer to MSS Server R1.0 documents [1] [2] [3].

1. 2. 1. Overview of MSS Server Release 1.0 Hardware

MSS Server hardware is available in two forms, the IBM 8210 MSS Server (see Figure 6. on page 8), which is a standalone product, and the IBM ATM MSS Server Module (see Figure 7. on page 8), which is a double–wide module installed for the IBM 8260 hub. The 8210 has two slots for network adapters. Two versions of the 155 Mbps ATM adapter are available for the 8210 network adapter slots: a single–mode fiber and a multimode fiber. The MSS Server Module has a single internal 155Mbps ATM interface and connects to the 8260 backplane. With the exception of the ATM interface differences explained above, both products provide equivalent functionality, namely:

- 155 Mbps ATM interface
- 100 MHz PowerPC 603e processor
- 512 KBytes of L2 cache
- 32 MBytes of processor DRAM
- 8 MBytes of buffer memory per ATM interface
- 12 MBytes of flash memory

  512 KBytes of flash memory is reserved for the firmware and another 512 KBytes of flash memory is reserved for up to 4 configuration files. The remaining 11 MBytes of flash memory can be used to store 1 operational code image.

- Serial service port (EIA 232) provides access for local and remote management of the MSS Server (described in 1. 2. 2. 6. on page 19). The MSS Server software uses this interface at 19,200 bps.
- 2 PCMCIA slots containing
  - a 260 MB PCMCIA hard disk which can be used to store multiple operational code and configuration images as well as error logs and dumps. Two banks are provided on the hard disk to store operational code and configuration images. Each of the two banks can hold 1 operational code image and up to 4 configuration files.
  - a PCMCIA Data/FAX modem or in the US and Canada, a Voice/Data/FAX modem.

    The modem provides access for local and remote management of the MSS Server (described in 1. 2. 2. 6. on page 19). The MSS Server software use the modem interface at 19,200 bps.

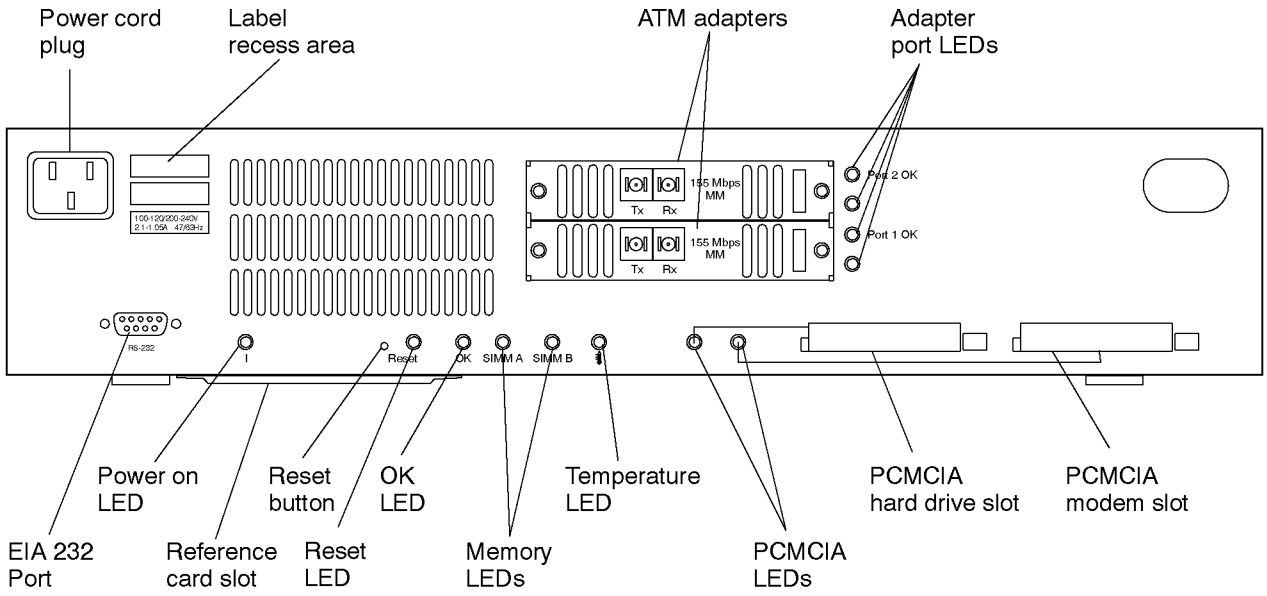

Figure 6. 8210 MSS Server

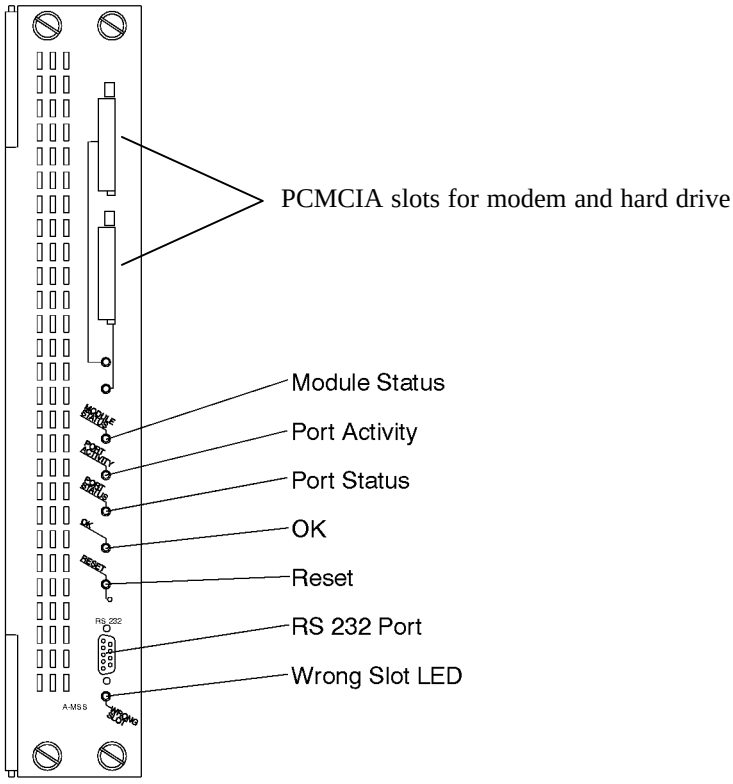

Figure 7. ATM MSS Server Module

1. 2. 2. Overview of MSS Server Release 1.0 Functions

MSS Server Release 1.0 includes the following functions:

- LAN Emulation (LANE)
- Classical IP (CIP)
- Bridging
- Routing
- Filtering
- Configuration, Monitoring and Management

An overview of these functions is provided in the following sections:

1. 2. 2. 1. LAN Emulation (LANE)

The ATM Forum has standardized a technique called LAN Emulation (LANE) which allows ATM networks to provide the appearance of Ethernet and Token–Ring emulated LANs (ELANs). LAN Emulation exploits high–speed ATM links while protecting existing LAN infrastructure since LAN–based applications do not have to change and existing LANs can be connected to ELANs using bridges or LAN switches. LAN Emulation also eases the migration to ATM because it allows incremental installation of ATM adapters in stations with high bandwidth requirements eg. servers. Furthermore, LAN Emulation simplifies network management because ELAN membership is not based on physical location. End stations can move from one ELAN to another without reconfiguration or rewiring. End stations can also join multiple ELANs using a single physical ATM interface. This is especially useful for servers in that they can have a direct connection in each ELAN they serve and thus bypass intermediate bridges and routers when sending data to clients.

The MSS Server LAN Emulation support is ATM Forum compliant and includes LAN Emulation Client (LEC) and LAN Emulation Services (LECS, LES, BUS). IBM value–add extensions include Intelligent BUS, Broadcast Manager, LAN Emulation Redundancy, Security and BUS Monitor. These functions are described below.

- LAN Emulation Client (LEC)

  LECs reside in end systems (ATM–attached hosts, or bridges/LAN switches representing legacy LAN hosts) and join ELANs. LECs provide a MAC–level service (either Ethernet/IEEE 802.3 or IEEE 802.5 Token–Ring) to higher level software. All LECs on an ELAN must be the same type. ELANs can be bridged to other ELANs or to legacy LANs using bridges. A bridge's LEC interface on an ELAN is called a Proxy LEC because it acts as a proxy on behalf of stations and rings on the other side of the bridge. To participate in ELANs, LECs request services from the

LECS, LES, and BUS. Token–ring and ethernet LEC interfaces can be defined and configured for bridging, routing or IP based box services like telnet and SNMP.

- LAN Emulation Configuration Server (LECS)

  The LECS reduces the network management burden by serving as a centralized repository for configuration data, thus minimizing configuration at the clients. LECs can use ILMI signalling to the ATM switch to dynamically discover and connect to the LECS. LECs can also be configured with the ATM address of the LECS or they can establish a VCC to the Well–Known LECS ATM Address defined by the ATM Forum. The MSS Server and the 8260 ATM switch supports all three methods.

  To get their initial configuration, LECs send a LE_CONFIGURE_REQUEST to the LECS. The LECS provides configuration data in a LE_CONFIGURE_RESPONSE to LECs. This information includes the address of the LES/BUS (ELAN) to join based on user defined policies. A policy is a criteria that the LECS uses to assign a LEC to a LES. Policies can be combined and prioritized. Policies can be created based on (1) ATM Address of the LEC, (2) MAC Address of the LEC, (3) Route Descriptor served by a proxy LEC, (4) ELAN Type of LEC, (5) Max Frame Size supported by the LEC, and (6) ELAN Name requested by the LEC.

- LAN Emulation Server (LES)

  LES provides address resolution service on an ELAN. LECs join an ELAN by sending a LE_JOIN_REQUEST to the LES which includes the LEC's MAC and ATM address. The LES establishes a point–point control VCC with the LEC and adds the requesting LEC to the appropriate point–multipoint distribute VCC and returns a LE_JOIN_RESPONSE.

  LECs wishing to find the ATM address associated with a MAC address or a Route Descriptor, send a LE_ARP_REQUEST to the LES. The LES sends this information in a LE_ARP_RESPONSE to the requesting LEC if the target is known. If the target is unknown, the LES forwards the request to all registered proxy LECs which in turn respond to the LES if they are proxying for the target. The requesting LEC uses the destination ATM address specified in the LE_ARP_RESPONSE to setup a data direct VCC to the destination and sends frames to the destination LEC.

- Broadcast and Unknown Server (BUS)

  In shared media LANs, delivering a frame to multiple or unknown destinations is straightforward because all stations see the packet. For LANs emulated on point–point networks like ATM, the BUS provides the services to deliver broadcast/multicast and unknown destination frames. A LEC discovers the BUS by sending a LE_ARP to the LES requesting the ATM address for the broadcast MAC address. The LEC establishes a point–point VCC to the BUS which in turn adds the LEC to a point–multipoint VCC.

  LECs send broadcast and multicast frames on a point–point VCC to the BUS which in turn distributes it on a point–multipoint VCC to all LECs on the ELAN. In addition to forwarding

broadcast and multicast frames, the LECs can also forward initial unicast frames to the BUS until a data direct VCC is setup to the destination LEC. The BUS forwards the unicast frames to the appropriate destination LEC on a point–point VCC if the destination is registered with the LES; otherwise, the BUS forwards the frame to all LECs on a point–multipoint VCC.

The following IBM value–add extensions to ATM Forum LAN Emulation are also included:

- Intelligent BUS (IBUS)

  The IBUS is an enhancement of the standard BUS. It can maintain separate point–multipoint VCCs for proxy and non–proxy LECs. The IBUS forwards broadcast and multicast frames on both point–multipoint VCCs. Unicast frames to registered destinations are forwarded on a point–point VCC to the destination. Unicast frames to unregistered destinations are forwarded only to proxy LECs on the separate point–multipoint VCC. This reduces client perturbation due to "nuisance" unicast frames i.e. non–proxy LECs are not bothered by unicast frames for proxy LECs.

- Broadcast Manager (BCM)

  BCM is an enhancement of the standard BUS. Without BCM, every multicast frame sent to the BUS is forwarded to all LECs on the ELAN. Furthermore, LECs that include the proxy function to provide bridging support then forward the broadcast frame onto other LAN segments. All end stations receive and process every broadcast frame, which can significantly increase end–station processing in large networks . BCM examines specific broadcast and non–broadcast frames received by the BUS and learns the association between source MAC address (layer–2) and source protocol address (layer–3). On subsequent broadcasts to learned protocol addresses, BCM transforms broadcast frames into unicast frames, and sends them only to interested LECs and end stations. By reducing broadcast frames, BCM reduces both network traffic and end–station overhead associated with filtering nuisance broadcast frames. Thus, BCM can improve overall system performance and enable practical deployment of larger ELANs. BCM can transform IP ARPs, IPX RIP and SAP advertisements and NetBIOS broadcasts (Add.Name.Query, Name.Query, Name.In.Conflict, Name.Recognized. Status.Query and Datagram). For NetBIOS, BCM also filters out frames like Add.Name.Query that are broadcast repeatedly[1].

  For token–ring ELANs, BCM includes an optional feature called Source Route Management (SRM) to further manage IP and NetBIOS broadcasts. SRM learns the ring topology behind each LEC by recording the Routing Information Field (RIF) of frames received by the BUS. After BCM has transformed an IP or NetBIOS broadcast into unicast as described above, SRM further transforms All Routes Explorer (ARE) or Spanning Tree Explorer (STE) frames into Specifically Routed Frames (SRF). Such frames would no longer need to be transmitted onto each ring in the source–route bridged network, thus conserving network bandwidth.

1. Historically, broadcast frames were repeated to ensure that all devices receive the frame when the network is heavily congested.

- LAN Emulation Redundancy,

  Without support for redundancy, ATM Forum LAN Emulation can become the single point of failure in a network. The MSS Server can provide redundant LAN Emulation Services (LES, BUS, LECS). A LES/BUS pair can be configured as primary or backup. The primary LES/BUS sets up a "Redundancy VCC" to the backup LES/BUS (see Figure 8.). The presence of this VCC indicates that the primary LES/BUS is operational, and the backup LES will not accept connections from LECs (which forces requesting LECs to repeat the configuration phase). Conversely, if the Redundancy VCC is not present, the backup LES/BUS services ELAN requests in the usual manner.

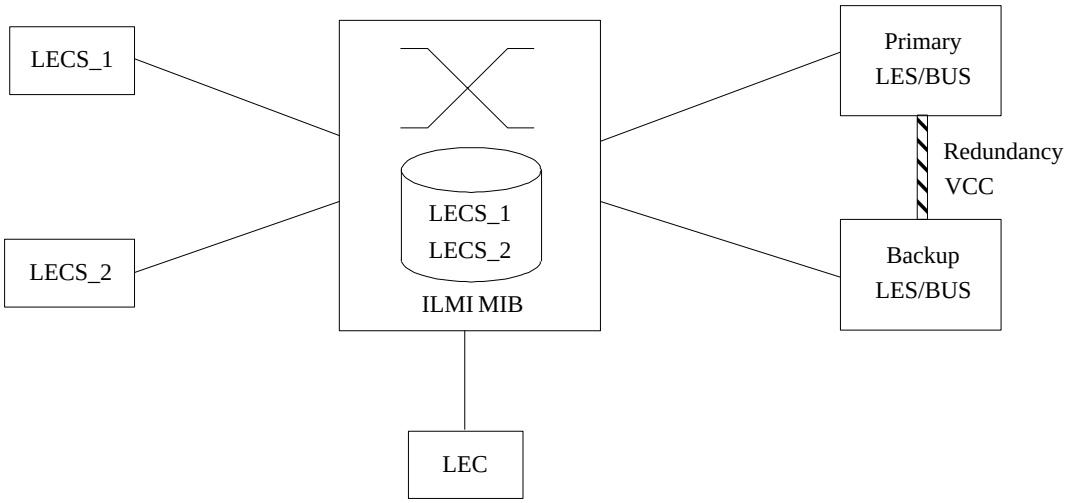

Figure 8. LAN Emulation Redundancy

If the primary LES/BUS fails, the LECs lose their connections to the LES/BUS and attempt to rejoin the ELAN by contacting the LECS. If the ELAN's LES/BUS is co–resident with the LECS, the LECS provides the address of the primary or secondary LES/BUS to the LEC based on the status of the redundancy VCC. If the LES/BUS is not co–resident with the LECS, the LECS cannot check the status of the LES/BUS, so it keeps a short term memory (5 minutes) of which LES/BUS was assigned to each requesting LEC. If a LEC re–requests configuration from the LECS within this time period (an indication that the previous LES/BUS is probably unavailable), the LECS response alternates between the primary and secondary LES/BUS.

When the primary LES/BUS recovers, it restores the redundancy VCC. This causes the backup LES/BUS to release its connections to all the LECs which in turn causes all the LECs to start the process over again and connect to the primary LES/BUS. LECS redundancy can be achieved by identically configuring a LECS on a different MSS Server and registering it with the ATM Switch.

A MSS Server need not be dedicated to backup functions. It can provide primary LES/BUS services for some ELANs and backup LES/BUS services for others.

- Security,

  Since LECs may be configured with LES ATM addresses, they can bypass LECS based access control. To control ELAN membership, an MSS LES can be configured to validate LEC join requests with the LECS. In this mode, the LES forms an LE_CONFIGURE_REQUEST on behalf of the LEC using information from the LE_JOIN_REQUEST. These LE_CONFIGURE_REQUESTs include the Source LAN Destination, Source ATM Address, ELAN Type, Max Frame Size, and ELAN Name from the LE_JOIN_REQUEST. If the LECS policies do not allow the LEC to join this ELAN, then the LECS rejects the LE_CONFIGURE_REQUEST. This causes the LES to reject the LE_JOIN_REQUEST from the LEC.

- BUS Monitor

  The BUS Monitor provides a way to pinpoint end–users that may be over–utilizing the BUS. When enabled, the BUS Monitor periodically samples the traffic sent to the BUS on a particular ELAN. At the end of each sample interval, the BUS Monitor identifies the top users of the BUS by their source MAC addresses, LEC ATM addresses, and the number of sampled frames each of them sent to the BUS. The number of top users to record, number of seconds in each sample interval, sample rate (i.e., sample one out of every “sample rate” frames), and number of minutes between sample intervals are all configurable.

- Default IP Gateway redundancy for ELANs

  The MSS Server provides Default IP Gateway redundancy for ELANs. The Default IP Gateway Redundancy works in conjunction with LAN Emulation Redundancy. One MSS Server can be configured as the primary Default IP Gateway and primary LES/BUS for a subnet/ELAN and another MSS Server can be configured as the backup (as shown in Figure 9.). Both MSS Servers have LEC interfaces on the subnet. However, only one of the LECs is active at any given time. The LECs have the same MAC address, and the same Default Gateway IP address is configured on both LEC interfaces. The Default Gateway function is performed by the MSS Server with the active LEC. The backup LEC continues to periodically join the ELAN. In the event of a failure at the primary MSS Server (or the ATM port associated with the primary LEC), the backup LEC is able to join the ELAN. Since the primary and the backup LECs have the same MAC address, IP ARP entries cached at the end hosts will still be valid and the default IP gateway switch over is transparent to the end hosts.

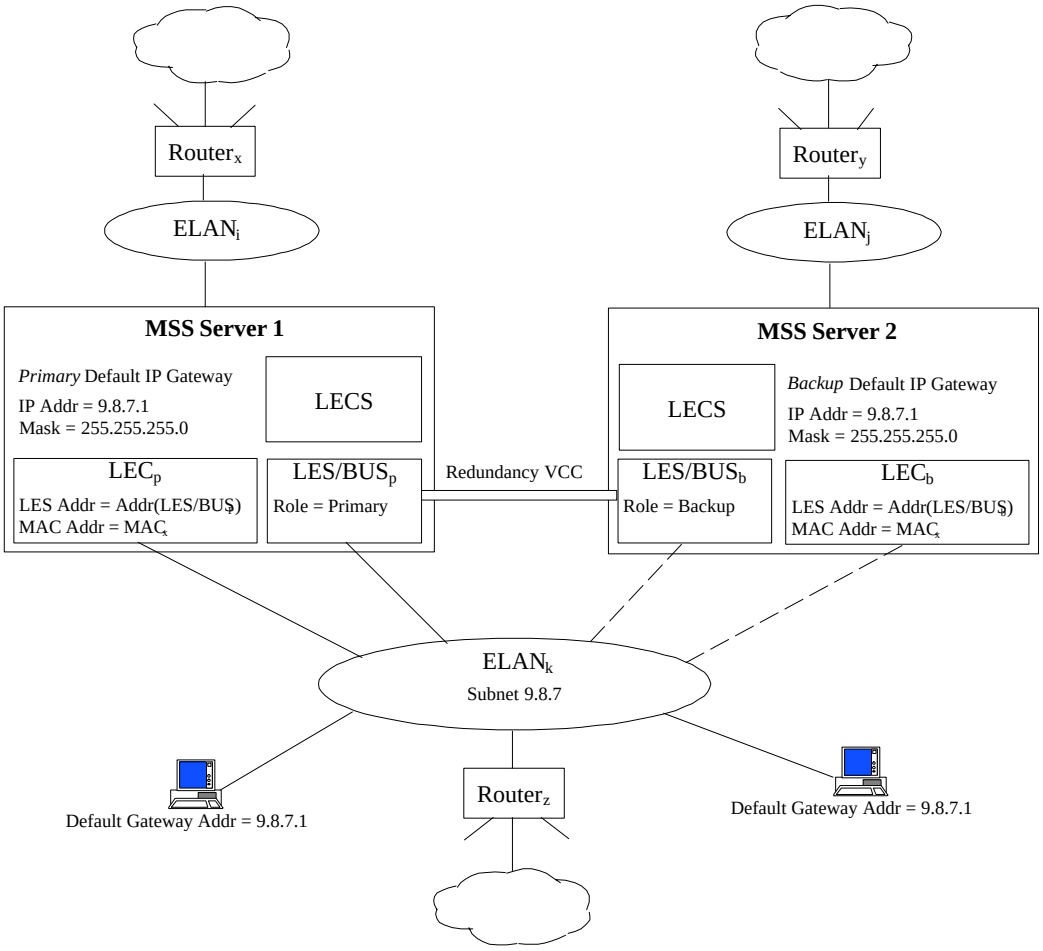

Figure 9. Default IP Gateway Redundancy for ELANs

1. 2. 2. 2. Classical IP (CIP)

The IETF has standardized a solution in RFC 1577 [9] called Classical IP (CIP) for sending IP traffic directly over an ATM interface while keeping the ATM infrastructure transparent to IP. Most IP applications that run today in a LAN or WAN environment will see no difference in functionality; however, their performance and throughput gains may be substantial. In addition to the high link speeds that ATM provides, Classical IP (CIP) requires fewer framing bytes than, for example, LANs which contain source and destination MAC addresses. Less bandwidth is used for overhead bytes, and more is used for data. In addition, no broadcast traffic is required for the resolution of ARP frames. In a broadcast environment, ARP traffic can adversely affect all stations in the subnet. In CIP, the ARP traffic only affects the ARP Server and the client requesting the information. Other stations on the subnet are unaffected by this traffic. The same benefits of moves, adds, deletes, etc., described for ELANs (in 1. 2. 2. 1. on page 9) apply to the CIP Logical IP Subnet (LIS). An end station can join multiple LISs using a single physical ATM interface. This is especially useful for servers in that they can have a direct connection in each LIS they serve and thus bypass intermediate bridges and routers when sending data to clients. Membership is not based on physical location.

Logically related stations are grouped into the same LIS. While all members of a LIS must support the Classical IP model, the MSS Server can route between subnets that are CIP–based and subnets that are LANE–based. The Logical IP Subnet contains all of the properties of a normal IP subnet. However, because ATM is a Non–Broadcast Multiple Access (NBMA) network, the existing broadcast method for resolving addresses cannot be performed. ARP Servers and ARP Clients were developed to solve this problem.

MSS Server CIP support includes ATM ARP Server, ATM ARP Client and SVC/PVC/QOS support. IBM value–add extensions include ATM ARP Server redundancy and Default IP Gateway redundancy for LISs. These functions are described below.

- ATM ARP Client

  The ATM ARP Client, also known as the Classical IP (CIP) Client or the 1577 Client, provides the interface for IP to a Logical IP Subnet (LIS). The CIP client registers its IP address with the ATM ARP Server. When the CIP client has traffic to transmit to other clients on the LIS, the client sends an ARP request to the ARP server with the target IP address. The server sends back a reply with the ATM address of the target IP address (if the target is registered with the server). The client uses this ATM address to setup a connection (VCC) to the target client and send IP datagrams on the connection. CIP interfaces can be used for IP routing or IP based box services like Telnet and SNMP.

- ATM ARP Server

  One ARP Server is defined per LIS. The server maintains the translation of IP addresses to ATM addresses. The server allows clients to register by accepting incoming VCCs and querying the client (with an InATMARP request) for the appropriate mapping information (i.e., the IP and ATM addresses of the client). The ARP Server also responds to ATMARP requests for ATM addresses corresponding to IP addresses specified by the client. Finally, the ARP server updates its tables through aging ARP entries and associated VCCs.

- SVC, PVC and QOS support

  The MSS Server implementation of Classical IP supports both Switched Virtual Circuits (SVCs) and Permanent Virtual Circuits (PVCs). SVCs use ILMI signalling to establish connections. PVCs do not require a signaling protocol, but do require configuration in both the ATM network switches and end systems. SVCs may be generated automatically through the address resolution and call setup procedures of Classical IP, or an SVC may be explicitly configured. Automatic SVCs are brought up and torn down by the ARP subsystem as required for sending IP traffic. Configured SVCs and PVCs are brought up during initialization and kept up indefinitely. PVCs and configured SVCs do not require an ARP Server.

  The attributes of both control channels (connections from a client to a server) and data channels (connections from one client to another) may be tailored to specific user needs. For example, “Quality of Service” characteristics can be specified on a LIS–basis by configuring VCC traffic parameters such as Peak Cell Rate, Sustained Cell Rate, and Maximum Reserved Bandwidth.

The following IBM value–add extensions to Classical IP are also included:

- ATM ARP Server redundancy

  Most ATM ARP Clients are configured with the ATM address of a single ARP Server. Therefore, to be widely applicable, a backup ARP Server must be accessible via the same ATM address as the primary ARP Server. This can be accomplished by connecting two MSS Servers with identically configured ARP Servers to the same 8260 ATM switch. By virtue of being connected to the same 8260 ATM switch, the network prefix of the ARP Servers' ATM addresses will be the same, and since the ARP Servers are configured identically (i.e., with the same locally–administered ESI and selector), their ATM addresses will be identical.

  The first ARP Server that attempts to register the common ATM address with the switch will be successful; this will be the primary ARP Server. The other ARP Server will be the backup because its ATM address registration will fail as a duplicate. MSS ATM ARP Servers (and ATM ARP Clients) retry duplicate ATM address registrations every 15 seconds. Thus, if the primary ARP Server fails, the backup will take over in a timely manner (i.e., when the primary fails, the switch will de–register its ATM address,which allows the next backup registration attempt to succeed). MSS Servers hosting backups need not be idle, since they can provide services for other LISs (and ELANs).

- Default IP Gateway redundancy for CIP LISs

  The ATM ARP Server redundancy mechanism can also provide Default IP Gateway redundancy for the LIS. Default Gateways provide routing services for hosts that do not run routing topology protocols (e.g., RIP or OSPF). These hosts are commonly configured with the IP address of their Default Gateway. Consequently, a backup Default Gateway must be accessible via the same IP address as the primary Default Gateway. An MSS Server can provide the Default Gateway function for a LIS when hosts are configured with the IP address of an MSS ATM ARP Client. In this case, Default Gateway redundancy can be achieved by connecting two MSS Servers with identically configured ARP Clients to the same ATM switch. These ARP Clients have the same ATM address and the same IP address. As previously described for ARP Server redundancy, the duplicate ATM addresses will ensure that only one of the ARP Clients is active at any given time, while the duplicate IP addresses allow the backup ARP Client to takeover as the Default Gateway if the primary fails.

1. 2. 2. 3. Bridging

The MSS Server supports bridging over emulated Ethernet and Token–Ring interfaces. Overall bridge behavior is determined by the configured behavior of individual bridge ports (interfaces). A bridge port can be configured for Transparent Bridging (TB), Source–Route (SR) Bridging, or Source–Route Transparent (SRT) Bridging. Ethernet ports only support TB while Token–Ring ports support all three modes. An overall option is also provided to translate between transparent and source–routed ports. Based on these options, the bridge can have six personalities. Based on the personality, the bridge uses different spanning tree algorithms to prevent loops in the network.

- TB

  Pure TB behavior is activated when all bridge ports are in TB mode. In pure TB mode, the MSS Server only acts as a transparent bridge and the IEEE 802.1d Spanning Tree algorithm is used. The pure TB bridge is sometimes also referred to as a Spanning Tree Bridge (STB). In TB mode, the bridge learns the MAC addresses of end–stations and the associated bridge port by observing the source MAC address of all frames being received. The bridge forwards frames to the appropriate bridge port based on the destination MAC address in the frame.

- SR

  Pure SR behavior is activated when all bridge ports are in SR mode. In Pure SR mode, the MSS Server only acts as a source–route bridge and the IBM Source–Route Bridging Spanning Tree algorithm is used. In SR mode, the bridge is configured with the ring numbers associated with each port. The bridge forwards frames to the appropriate bridge port (ring) based on the next ring in the Routing Information Field (RIF) of the frame.

- SR&TB

  SR&TB behavior is activated when token–ring ports are in SR mode, ethernet ports are in TB mode and SR–TB translation is disabled. In SR&TB mode, the MSS Server acts as both a transparent bridge and a source–route bridge simultaneously, but the two types of bridges do not work together and are thus isolated from each other. The IEEE 802.1d and IBM Source–Route Spanning Tree algorithms are used independently.

- SR–TB

  SR–TB behavior is activated when token–ring ports are in SR mode, ethernet ports are in TB mode and SR–TB translation is enabled. In SR–TB mode, the bridges are no longer independent and frames are “translated” between the two bridged domains. SR–TB translation is supported for LLC based layer–2 protocols like SNA and NetBIOS. SR–TB is typically used to bridge between ethernet and token–ring stations when routing is not possible. The IEEE 802.1d and IBM 8209 Spanning Tree algorithms are used. The SR–TB translational bridge in the MSS can inter–operate with the translational bridges in the IBM 8209/8229, IBM 6611 and the common translational bridges in the IBM 2210 and IBM 2216.

- SRT

  SRT behavior is activated when at least one token–ring port is in SRT mode (which means that the port can support SR, TB, or SR & TB simultaneously) and SR–TB translation is disabled. In SRT mode, the bridges are independent, as in SR&TB mode, but only the IEEE 802.1d Spanning Tree algorithm is used. The SRT bridge performs source routing (SR) when frames with routing information (RIF) are received and performs transparent bridging (TB) when frames are received without routing information. SRT is typically used when some token–ring end stations in a SR network cannot perform source routing as is the case sometimes in Novell Netware based networks.

- Adaptive SRT (ASRT)

  Same as SRT except that SR–TB translation is enabled.

1. 2. 2. 3. 1. Bridge Tunneling

In addition to bridging between emulated LANs, the MSS Server also supports Bridge Tunneling. A bridge tunnel is used to join bridge domains that are interconnected by an IP network. The sending bridge encapsulates frames in an IP packet and addresses it to the bridge on the other end of the tunnel. The encapsulated bridged frame traverses the IP network and reaches the receiving bridge which strips the IP header and handles the frame as if it was received on a locally attached LAN. From the bridge view point, the IP tunnel looks like an ethernet or token–ring LAN so bridging behaviors discussed earlier apply to IP tunnels as well. Bridge Tunneling is typically used to bridge traffic across an IP wide–area network backbone.

1. 2. 2. 4. Routing

Routing support is provided for two widely used protocols, IP and IPX.

- IP Routing

  IP routing is supported between arbitrary combinations of CIP and ELAN–based subnets. Routing Information Protocol (RIP) [15] and Open Shortest Path First (OSPF) [17] are supported as interior gateway protocols. Border Gateway Protocol (BGP v4) [18] is supported as an external gateway protocol. RIP is supported on ELAN–based subnets. OSPF and BGPv4 are supported on both ELANs and LISs.

  IP level filtering is supported to control IP traffic. It is described in 1. 2. 2. 5. on page 18.

- IPX Routing

  IPX routing is supported over emulated token–ring and ethernet interfaces (LECs). It is also supported natively on ATM PVCs and SVCs using RFC 1483 connections to other IPX routers (Note: CIP also uses RFC 1483 encapsulation). One IPX network may be configured per ATM interface. As with Classical IP, Quality of Service characteristics can be specified by configuring VCC traffic parameters such as Peak Cell Rate, Sustained Cell Rate and Maximum Reserved Bandwidth. Netware Routing Information Protocol (RIP) and Service Advertisement Protocol (SAP) are supported as IPX routing protocols.

  IPX level filtering is supported to control IPX traffic. It is described in 1. 2. 2. 5. on page 18.

1. 2. 2. 5. Filtering

The MSS Server provides several types of filters to control network traffic. Filters can be combined together to create complex filters.

- MAC Filters

  MAC level filters can be used to filter packets based on source/destination MAC address or user data and can be specified on a per port basis. User data filters are in sliding window

format and specify an offset from the beginning of the MAC or LLC header, a byte pattern and a mask.

- NetBIOS Filters

  NetBIOS traffic can be filtered by host name or user data and can be specified on a per port basis. Host name filters can contain wildcard characters. User data filters are in sliding window format and specify an offset from the NetBIOS header, a byte pattern and a mask.

- IP Filters

  IP filters can be used to control access based on source/destination IP address, IP protocol number and destination TCP/UDP port number. IP filters can be specified on a per port basis or globally on a router basis. IP filters can also be used to do route filtering by controlling how routes are learned.

- IPX Filters

  IPX filters can be used to control access based on source/destination IPX network, source/destination node address, hop count and IPX protocol type and socket. IPX filters can be specified on a per port basis. Global IPX filters are also supported and apply to all IPX interfaces. Additionally, global RIP and SAP filters are supported to control how routing information and netware services are advertised. All filters can be applied inbound or outbound.

1. 2. 2. 6. Configuration, Monitoring and Management

The MSS Server can be configured from a console interface, a web browser interface or a standalone configuration program. The MSS Server status can be monitored from a console interface, a web browser interface, a telephone/fax interface or a SNMP based network management station. The MSS Server can be managed from a SNMP based network management station. These user interfaces are described below.

- Console Interface

  The command line console provides complete configuration as well as comprehensive monitoring and status reporting facilities. It can be accessed (via the serial port) using a TTY connection from an ASCII terminal or a workstation/PC with ASCII emulation. It can also be accessed using Telnet over a SLIP connection (via serial port or PCMCIA modem) or in–band IP connection (via the ATM interface).

- Web Interface

  The web browser interface enables the functions of the command line console to be accessed via a more user–friendly graphical interface. The MSS Server supports HyperText Transfer Protocol (HTTP) and HyperText Markup Language (HTML). Web browsers may access the MSS Server via SLIP (serial port or modem) or in–band IP connections (via the ATM interface).

- Configuration program

  The Configuration Program is the most user–friendly configuration method. The Configuration Program is a standalone software package for workstations and PCs that provides a graphical user interface for creating complete MSS Server configuration files without connecting to the MSS Server. It also ensures that configured parameters for the different functions do not conflict with each other. Completed configuration files can be transferred to an MSS Server in the following ways:

  i) using the Xmodem protocol to download the configuration file over a TTY connection (serial port or modem)

  ii) using the Trivial File Transfer Protocol (TFTP) to transfer binary configuration files over a SLIP connection (serial port or modem) or in–band IP connection (ATM interface)

  iii) using the Communications Option of the Configuration Program, which utilizes SNMP and requires in–band IP access (ATM interface).

  The Configuration Program runs on the following operating systems: AIX (version 3.2.5 or higher), OS/2 (version 2.1 or higher), and DOS/Windows (Windows 3.1 or higher including Windows NT and Windows 95).

- Data/Fax or Voice/Data/Fax modem

  The MSS Server may contain an integrated Data/FAX modem, or in the US and Canada, a Voice/Data/FAX modem. Both types of modems can FAX status and configuration reports upon request. The modems can also be configured to auto–dial and either page the network administrator or FAX a report when alerts occur. Additionally, the Voice/Data/FAX modem provides a set of interactive configuration, monitoring, and control capabilities using touch tone telephone input with automated voice responses.

- SNMP based Network Manager

  The MSS Server can be managed by an SNMP based Network Manager such as TME 10 NetView. The MSS Server contains an SNMP agent which supports standard and private MIBs and is accessible in–band via IP (ATM interface).

1. 2. 2. 7. Trouble–shooting

Several facilities are included with the MSS Server to aid in network trouble–shooting.

- Event Logging System (ELS)

  ELS is a facility used to monitor events within the MSS Server as it runs. It can be configured to monitor events by subsystems, groups or specific events. For example, ELS can be used to monitor how the bridge is processing broadcasts. ELS events are captured in a circular buffer and can be viewed on the console, captured in a file, or forwarded to a network management workstation as SNMP traps.

- Packet Trace

  A packet trace facility is available for LEC interfaces. The user can set the trace buffer size and the number of bytes to capture for each packet. Traced packets can be viewed from the console or captured in a file.

- Memory Dump

  A facility is provided to dump the contents of the entire MSS memory to the hard disk in the event of a catastrophic failure. This information can be used by IBM service personnel to diagnose problems.

1. 3. Summary of MSS Server Release 1.1

MSS Server R1.1 was released in April 1997. A brief summary of the MSS Server R1.1 hardware and software functions is provided in this section.

1. 3. 1. Overview of MSS Server Release 1.1 Hardware

Four FDDI adapters were introduced for use in the 8210's adapter slot–2 (see page 8):

- Dual Ring optical fiber FDDI adapter
- Single Ring optical fiber FDDI adapter
- Dual Ring copper FDDI adapter
- Single Ring copper FDDI adapter

1. 3. 2. Overview of MSS Server Release 1.1 Functions

The key functions of MSS Server Release 1.1 are support for FDDI, SuperELAN, NHRP, Virtual ATM interfaces, LAN Emulation enhancements, Classical IP enhancements, Routing enhancements, Bridging enhancements, Capacity improvements and Performance improvements. These are described below.

1. 3. 2. 1. Support for FDDI

The addition of the FDDI interface enables the MSS Server to provide a path for incremental migration from existing FDDI backbones to new ATM backbone networks. IP, IPX, and AppleTalk routing protocols are supported over the FDDI interface.

1. 3. 2. 2. SuperELAN

Short–Cut Bridging (SCB), Bridging Broadcast Manager (BBCM), Dynamic Protocol Filtering VLANs and VLAN IP cut–through are new functions in this release that can combine to create a SuperELAN (also called SuperVLAN). A SuperELAN is a collection of ELANs whose LECs can set up data direct VCCs to each other as if they were on the same ELAN. Thus, a SuperELAN behaves like a large ELAN with distributed LAN Emulation services. Furthermore, protocol specific broadcasts between the ELANs are controlled by BBCM & DPF VLANs. Each of the key SuperELAN functions are described below.

- Short–Cut Bridging (SCB)

  Short–cut bridging allows multiple, independent emulated LANs to function as a single SuperELAN. This provides several benefits, including scalability and robustness improvements associated with distributing the LE Service. With short–cut bridging, a client on any of the ELANs in the SuperELAN may setup a Data Direct VCC to any other client on the superELAN.

By bypassing the bridge, unicast data transfer is limited only by the speed of the ATM network. Short–cut bridging support is being provided for transparently bridged ELANs.

- Bridging BroadCast Manager (BBCM)

  BBCM complements short–cut bridging by limiting the scope of broadcast frames within the SuperELAN. BBCM is similar to BCM (described on page 11) except that it manages broadcasts exiting an ELAN via the bridge. BBCM enables hierarchical broadcast management with BCM managing intra–ELAN broadcasts and BBCM managing inter–ELAN broadcasts. BBCM support is provided for IP and NetBIOS.

- Dynamic Protocol Filtering VLANs

  Dynamic Protocol Filtering (DPF) complements short–cut bridging by limiting the scope of broadcast frames within the SuperELAN. DPF limits transmissions of broadcast frames to the appropriate Protocol VLAN (PVLAN). DPF creates broadcast domains by dynamically learning the set of PVLANs active on each ELAN segment. Unmanaged broadcasts are then transmitted only on ELAN segments containing stations that are members of the VLAN. VLANs limit broadcast traffic as well as allow end–users to move around in the switched network without changing their configuration. Routing services are used to inter–connect protocol VLANs. The protocol VLAN is identified by the packet's content: either the protocol type (e.g., NetBIOS) or the combination of protocol type and subnet address (e.g., a particular IP subnet or IPX network). DPF supports multiple IP and IPX protocol VLANs, and a single NetBIOS protocol VLAN. DPF VLANs are supported on SR and TB bridge interfaces.

- VLAN IP cut–through

  DPF also includes support that enables IP stations to establish inter–subnet Data Direct VCCs. This *IP cut–through* support is activated by modifying the subnet mask at one or more endstations so that these stations will *ARP* for destinations that were previously reached through the router. When enabled for IP cut–through, DPF will forward inter–subnet ARPs to the PVLAN associated with the destination subnet. Once the destination MAC address is returned in the ARP response, short–cut bridging enables the establishment of a direct VCC. Thus, IP cut–through enables *zero–hop routing* (by removing the router from the data path), while preserving the original subnet broadcast domains (since DPF uses the *real* subnet masks). IP cut–through can be enabled on a subnet basis or by specific IP hosts (eg. servers).

  An example of SuperELAN short–cut bridging and VLAN IP cut–through is shown in Figure 10.. In this example, IP VLANs keep broadcasts on subnet 1.1.1 separated from subnet 1.1.2.. The server 1.1.2.2 has been configured with a different mask (255.255.0.0) and VLAN IP cut–through has been enabled from subnet 1.1.2 to subnet 1.1.1.. Traffic from the client to the server goes through an IP router (in this example, the IP router is in the MSS Server). Traffic from the server to the client bypasses the IP router (zero–hop) and is forwarded on a short–cut VCC.

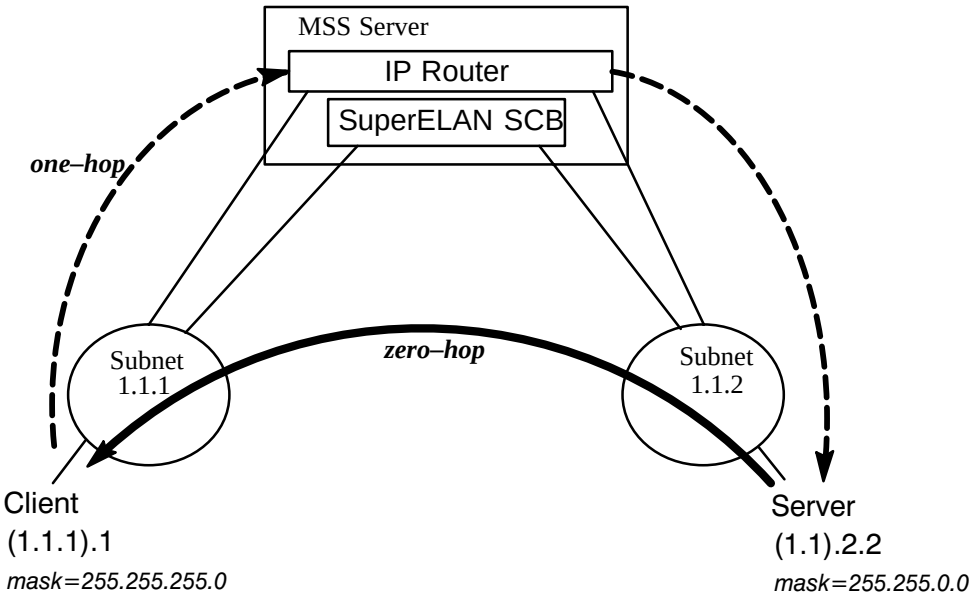

Figure 10. SuperELAN with SCB and VLAN IP cut–through

1. 3. 2. 3. Next Hop Resolution Protocol (NHRP)

NHRP is an IETF protocol (draft specification v11) [12] that can improve network performance by eliminating router hops on Non–Broadcast Multi–Access (NBMA) networks like ATM. When internetworking protocols, such as IP, are run on NBMA networks, the routed path through the network may include multiple hops across the same network. NHRP was designed to better utilize the capabilities of the underlying switched fabric by enabling the establishment of shortcut routes across NBMA networks. A shortcut may be established directly to the destination or, if that is not possible, to the egress router on the NBMA network nearest to the destination. Note, an emulated LAN (ELAN) on ATM is not considered to be part of the NBMA network by the NHRP standard. NHRP is a client–server protocol and the MSS Server implementation supports both, the NHRP Client (NHC) and the NHRP Server (NHS). NHRP support is provided for the IP protocol.

The MSS Server NHRP implementation includes IBM value–add extensions that enable short–cut routes to be established to destinations that do not support NHRP and to destinations that are not on the NBMA network (eg. if destination is accessible via Emulated LAN).

- NHRP Client (NHC)

  NHCs register their IP and ATM address with NHSs. A NHC issues a NHRP Resolution Request to a NHRP Server (NHS) for the destination IP address when the rate of traffic to the destination exceeds a configurable threshold. The request follows the routed path until the egress NHS in the path responds to the requesting NHC with the ATM address of the next hop. The next hop may be the destination itself, or an intermediate router if the routed path exits the NBMA network. The NHRP Resolution Reply which follows the routed path back includes a holding time during which the information is valid. The NHCs revalidate the address mapping prior to

the expiration of the holding time. If the NHS sends a NHRP Purge for a mapping, then the NHC invalidates the mapping regardless of the holding time and sends another NHRP Resolution Request if necessary. If the NHC receives a NHRP Resolution Reply rejecting the NHRP Resolution Request, the NHC will not send another NHRP Resolution Request for a configurable period of time.

- NHRP Server (NHS)

  NHSs are coupled with routers. NHSs maintain layer–3 to ATM address mappings for registered NHRP Clients (NHCs). NHSs receive NHRP Resolution Requests from NHCs and other NHSs requesting shortcuts to a layer–3 destination. If the destination is served by the receiving NHS, then the NHS sends a NHRP Resolution Reply containing the ATM address of the destination to the requesting NHC via the routed path. Otherwise, the NHS forwards the request along the routed path. The request follows the routed path until it reaches the egress NHS in the path. If the destination is registered with the egress NHS, the egress NHS sends a NHRP Resolution Reply containing the ATM address of the destination to the requesting NHC. If the destination is not registered with the egress NHS, the egress NHS sends a NHRP Resolution Reply containing its own ATM address. If a layer–3 to ATM address mapping changes due to a network topology change or some other reason, the NHS sends a NHRP Purge to all NHCs that are using that mapping.

  An example of zero–hop routing with NHRP is shown in Figure 11.. Zero–hop routing is possible when NHC software is installed on attached–ATM hosts. To initiate establishment of a shortcut to *IP Host*$_y$, the NHC in *IP Host*$_x$ issues a NHRP Resolution Request for *Host*$_y$*'s* IP address. As indicated above, NHRP Resolution Requests follow the routed path, so the request flows from *IP Host*$_x$ to *NHS*$_1$, from *NHS*$_1$ to *NHS*$_2$, and then on to *NHS*$_3$. *NHS*$_3$ responds to the request with the ATM address of *IP Host*$_y$. When the NHRP Resolution Reply is returned to *IP Host*$_x$, the ATM address is used to setup a *shortcut VCC*. Subsequent traffic for *IP Host*$_y$ is

transmitted directly over this VCC, bypassing the three intermediate routers, which both lowers latency and increases system throughput.

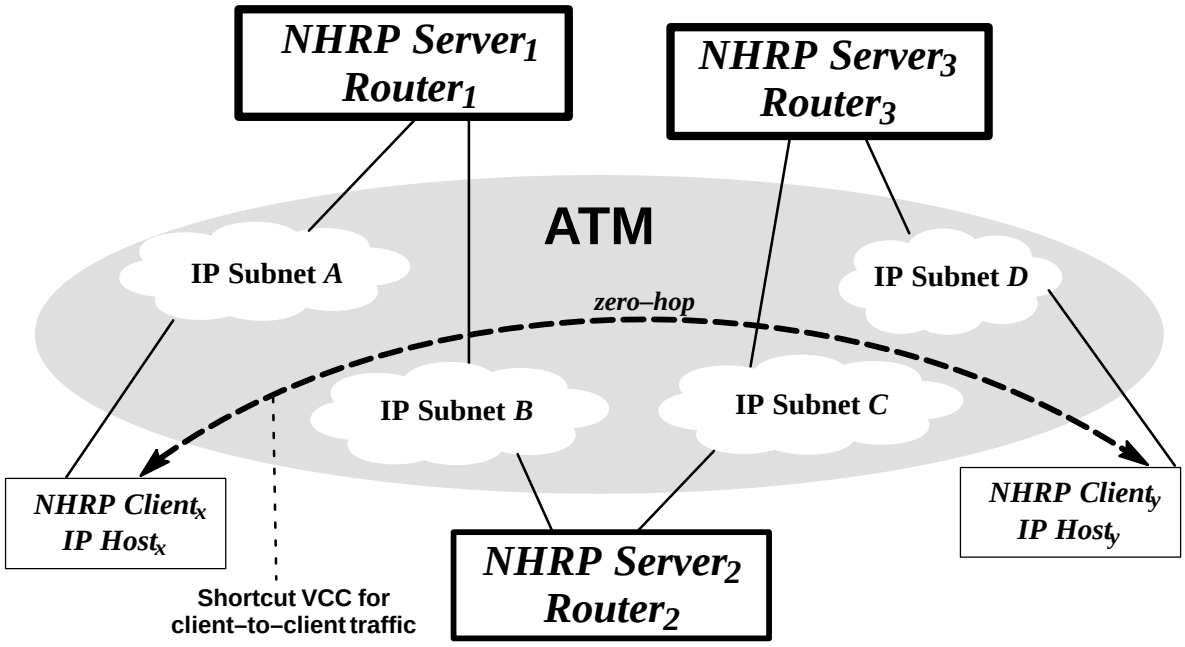

Figure 11. Short–cut VCCs using NHRP

The following IBM value–add extensions to IETF NHRP are also included in the MSS Server:

- Shortcuts to destinations without NHRP clients

  The IETF NHRP standard allows shortcut VCCs to be set up between NHCs only. If the source is a NHC but the destination is not (eg. it is a Classical IP client), then a shortcut cannot be established to the destination. Thus, the benefits of NHRP are limited to environments where both the source and destination support NHRP.

  The MSS NHRP implementation solves this problem. The MSS NHS can send a NHRP Resolution Reply containing the ATM address of the destination even if the destination is not registered as a NHC. The MSS NHS discovers the destination ATM address using other means like querying the router ARP table or querying the ATM ARP Server if the destination is on a CIP LIS. Thus, NHC enabled stations can get zero hop routing while non–NHC stations get one hop routing.

- LAN Emulation short–cuts

  The IETF NHRP standard specifies RFC 1483 encapsulation on ATM for NHRP flows as well as shortcut VCCs. Support for LANE encapsulation is not mentioned. This means that short–cut

VCCs could not be established to stations that are attached to ATM using LANE or attached to ATM via a LAN Switch that supports LANE.

The MSS NHRP implementation supports LANE encapsulation and LANE short–cut VCCs. This is done using vendor private extensions in the NHRP control flows, which are allowed by the IETF NHRP standard. The MSS NHC specifies the LANE shortcut extensions in the NHRP Resolution Request and the MSS NHSs propagate this request along the routed path. If the egress MSS NHS is connected to the destination via LANE, the NHS responds with a NHRP Resolution Reply containing the ATM address of the destination as well as the MAC address of the destination. The MAC address is specified in the vendor private extension field. If the destination is connected to ATM via a LAN switch with a LANE interface, then the ATM address is that of the LAN switch. The requesting NHC sets up a LANE shortcut VCC to this ATM address. Note, the requesting NHC sets up its own shortcut LANE VCC to the destination independent of any LEC, so the NHC can automatically adjust for the destination's LAN type, bridging type or MTU size. Thus, the MSS NHRP implementation extends the benefits of NHRP to the installed base of LANE equipment as well as to the very large base of legacy LAN attached stations connected via LANE bridges and switches.

- Partial shortcuts using Exclude Lists

  The IETF NHRP standard allows a NHC to set up a shortcut across a NBMA network as long as all routers in the routed path on the NBMA network provide NHS services. If the routed path across the NBMA network includes a router that does not provide the NHS function, then the NHRP Resolution Request from the NHC gets lost. Thus, a short–cut cannot be established and packets will follow the routed path.

  The MSS NHS implementation provides a feature called the *Exclude List*, which allows the user to configure the layer–3 address of routers on the NBMA network that do not provide the NHS function. The MSS NHS will not forward NHRP Resolution Requests to routers in the Exclude List. Instead, the MSS NHS will behave like the last NHS server in the routed path and respond with a NHRP Resolution Reply giving the ATM address of the next hop. Thus, the NHC can establish a *partial short–cut* across ATM i.e. from the NHC to the first router without NHS.

- Security

  Security is one of the important functions provided by routers. In order to preserve router–based security capabilities in NHRP environments, protocol–layer access controls (filters) are applied when generating, forwarding, or responding to NHRP Resolution Requests. This allows shortcuts to be denied in situations where routed path communications is not allowed. Filtering is described in section 1. 2. 2. 5. on page 18.

1. 3. 2. 4. Virtual ATM interfaces

Virtual ATM Interface allows multiple ATM interfaces to be emulated on a single real ATM interface. Thus native ATM protocols like IP and IPX are no longer limited to the number of protocol interfaces supported on a single ATM interface. IP and IPX are supported on virtual ATM interfaces.

Besides additional protocol interfaces, virtual ATM interface also allows IP multicast to work on ATM by putting IP subnets on separate virtual ATM interfaces. Otherwise IP multicast traffic would not be forwarded between subnets.

1. 3. 2. 5. LAN Emulation enhancements

Several enhancements have been made to LAN Emulation that include SuperELAN, support for pre–standard IBM LAN Emulation protocol, a new BUS mode, configurable QOS, and default IP gateway redundancy. These enhancements are described below.

- SuperELAN

  SuperELAN is described in 1. 3. 2. 2. on page 22.

- IBM LAN Emulation

  Prior to ATM Forum standard LAN Emulation, IBM had developed LAN Emulation that was being used in bridges like the IBM 8281. By adding a client for IBM LAN Emulation, the MSS Server can provide routing and bridging services for existing networks that are implemented with IBM LAN Emulation Architecture, while also offering a migration path to ATM Forum LAN Emulation.

- New BUS modes with 100+ Kpps forwarding rates.

  A new BUS mode called *VCC Splicing Mode* enables frames received on one VCC to be immediately forwarded on another VCC without CPU intervention. Another BUS mode called *Adapter Mode* also offers improved BUS performance by keeping packets in ATM adapter memory while allowing the CPU to do some intelligent BUS functions.

- Configurable QoS for LAN Emulation.

  The Configurable QoS feature allows LAN Emulation networks to take advantage of ATM's Quality of Service capabilities. Both ELAN–wide and LEC–based QOS control is provided, along with a negotiation algorithm that picks the parameters *best–suited* for the pair of communicating LECs. ELAN–wide QoS is configured using the LECS which provides the QoS parameters to all LECs in the ELAN that support QoS. LEC–based QoS is configured at the LEC. The QoS parameters that are supported are Maximum Reserved Bandwidth, Traffic Type, Peak Cell Rate, Sustained Cell Rate, and QoS Class

- Default IP Gateway Redundancy for ELANs enhancements

  The Default IP Gateway Redundancy mechanism is no longer coupled with LAN Emulation Redundancy (as described on page 13). In addition to the shared default gateway IP and MAC address, the primary and backup can also be configured with their own unique IP addresses on the same LEC. Thus other routers can discover and use the backup while the primary is still active.

The LECs in the primary and the backup gateway register their unique MAC address as well as the shared MAC address associated with the default IP gateway. The LES allows only one of the LECs to register the shared MAC address. The rejected LEC retries the registration process periodically (the primary retries every 5 seconds and the backup retries every 30 seconds). If the primary gateway fails, the secondary can successfully register and assume the role of the default gateway. When the primary comes back, its attempt to register the shared MAC address fails so it sends a message to the backup. The secondary responds by de–registering the shared MAC address and releasing its VCCs. This allows the primary to register the shared MAC address with the LES and resume the default gateway function.

- LES/BUS/LECS capacity enhancements

  LES/BUS/LECS capacity enhancements are described in 1. 3. 2. 10. on page 31.

1. 3. 2. 6. Classical IP enhancements

The ATM ARP Server redundancy (described on page 16) and default IP Gateway redundancy for LISs (described on page 16) have been improved. The improvements are described below.

- ATM ARP Server Redundancy improvements

  The user can now configure which MSS Server is the primary ARP Server and which is the backup. The MSS Server configured as a backup ARP Server can also simultaneously provide IP routing functions.

  An example of ARP Server redundancy is shown in Figure 12.. The primary and the backup ARP Server are configured with different IP and ATM addresses. The backup is also configured with the ATM address of the primary. The primary ARP Server establishes a *Redundancy VCC* (Red Channel) to the backup. The presence of this VCC indicates that the primary is serving the LIS. When the redundancy VCC is not present, the backup registers the ATM address of the primary ARP Server and starts servicing the LIS. When the primary ARP Server establishes the redundancy VCC, the backup de–registers the primary ARP Server's ATM address, thus allowing the primary to start serving the LIS. While the primary is active, the backup can still perform IP routing because it also has its own unique IP and ATM address.

- Default Gateway Redundancy for LISs.

  The Default Gateway redundancy support for LISs is provided as an optional extension of ARP Server Redundancy. The primary and backup ARP server are configured with an additional IP address, the IP address of the default gateway. While the primary ARP Server is active, it will respond to ARPs for the IP address of the default gateway with its own ATM address. If the primary fails, the backup will take over the primary's ATM address and assume responsibility for the primary's ARP Server as well as default gateway function at that ATM address.

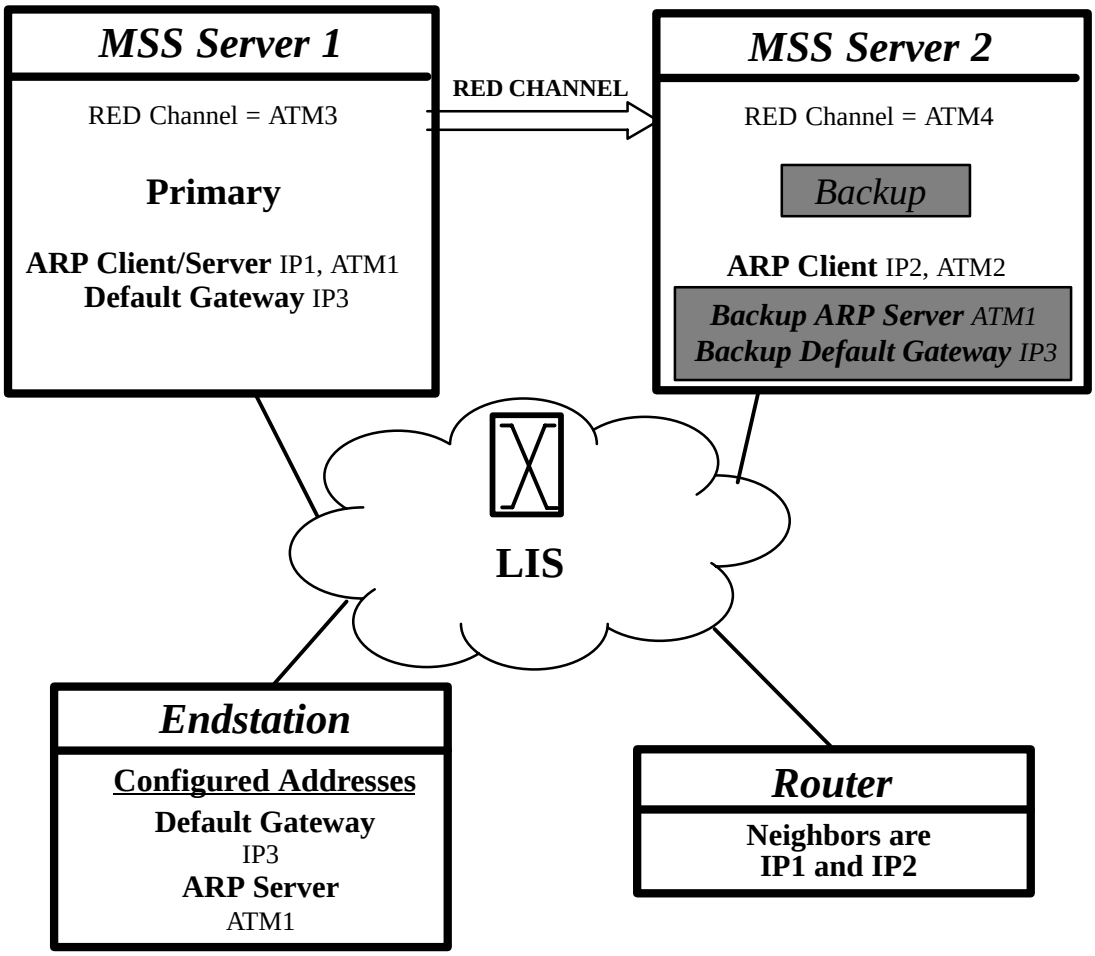

Figure 12. CIP ARP Server and Default Gateway Redundancy

In Figure 12., MSS Server 1 is the primary ARP Server and Default Gateway. MSS Server 2 is configured as a backup ARP Server and backup Default Gateway for MSS Server 1. MSS Server 2 also has its own IP and ATM address (IP2 and ATM2), thus it is visible to the external router and can be used for IP routing even though it is in backup mode.

1. 3. 2. 7. Routing enhancements

A new routing protocol, AppleTalk is supported now. In addition, new function has been added to to support one hop and zero hop IP routing.

- AppleTalk Routing

  AppleTalk routing is supported over FDDI and Emulated LAN interfaces.

- One–hop and zero–hop IP routing

  One–hop and zero–hop IP routing with NHRP are described in 1. 3. 2. 3. on page 24. Zero–hop IP routing with SuperELAN IP cut–through is described in 1. 3. 2. 2. on page 22.

1. 3. 2. 8. Bridging enhancements

Several enhancements to the bridging function have been made in this release.

- SuperELAN

  SuperELAN includes Short–cut Bridging (SCB), Bridge Broadcast Manager (BBCM) and Dynamic Protocol Filtering VLANs (DPF VLANs). SuperELAN functions are described in 1. 3. 2. 2. on page 22.

- SR–TB Bridging for IP

  SR–TB bridging (described in 1. 2. 2. 3. on page 16) now supports translational bridging for IP in addition to LLC based layer–2 protocols like SNA and NetBIOS.

- RFC 1483 Bridging over PVCs

  Bridging support has been added for RFC 1483 PVCs. By encapsulating MAC frames over PVCs, RFC 1483 bridging support enhances the internetworking capabilities of the MSS Server, and thereby provides opportunities for the MSS Server to play a role in networking environments that include Frame Relay/ATM interworking switches.

1. 3. 2. 9. Filtering enhancements

New bridging filters have been added.

- Bridge filters

  Bridging Broadcast Manager (BBCM) and Dynamic Protocol Filtering VLANs (DPF VLANs) are two new types of bridge filters. They are described in 1. 3. 2. 2. on page 22.

1. 3. 2. 10. Capacity improvements

Many capacity related enhancements have been made in this release.

- LES/BUS/LECS capacity enhancements

  Maximum number of LES/BUS instances was increased from 300 to 600. Maximum number of CIP interfaces was increased from 64 to 1000. Maximum number of served CIP clients was increased from 3000 to 10,000. Maximum number of LECS policy values was increased from 1500 to 6000.

- ATM PVC Multiplexing

  ATM PVC Multiplexing allows protocols like IP, IPX and Bridging that can run natively on ATM using RFC 1483 encapsulation to share the same PVC. Multiplexing protocols on a PVC reduces the number of VCC requirements in the MSS Server and the ATM Switch.

- Virtual ATM Interfaces

  Using Virtual ATM Interfaces, the number of native IP and IPX interfaces has been significantly increased. Virtual ATM Interfaces are described in 1. 3. 2. 4. on page 27.

1. 3. 2. 11. Performance improvements.

Better mapping of MSS Server virtual memory increased performance across the board. Dynamic allocation of LE_ARP cache entries enables larger LE_ARP caches and thus improved performance.

1. 4. Overview of MSS 2.0/2.0.1

MSS 2.0/2.0.1 includes:

- a new one–wide MSS Server Module for the 8260/8265 hubs
- a new release of the MSS Server software
- software drivers called *MSS Route Switching Clients* for LAN attached end–stations

These features are described below.

1. 4. 1. Overview of MSS Release 2.0/2.0.1 Hardware

A new MSS Server Module for the 8260/8265 hubs, memory upgrade option for existing MSS Servers, and 8210s with additional memory are being introduced in this release.

1. 4. 1. 1. One–wide A–MSS Server Module

A new MSS Server Module for the 8260/8265 hubs provides several enhancements over the previous MSS Server Module (described in section 1. 2. 1. on page 7):

- Smaller footprint – it occupies a single 8260/8265 slot (vs. 2 slots previously)
- Faster processor – 166 MHz PowerPC 603ev (vs. 100 MHz previously)

  A 66% faster processor improves overall performance especially CPU intensive activities like routing and broadcast processing.

- More memory – 64 MB of processor EDO DRAM standard (vs. 32 MB DRAM previously)

  EDO memory is faster and improves overall performance

- Embedded Ethernet Controller

  An embedded 10 Mbps RJ45 ethernet connection can be used for out–of–band management of the MSS Server. This allows for fast upgrades of the operational code and configurations when compared to the serial port.

- Internal IDE hard disk

  An internal 1.6 GB IDE hard disk is standard and can be used to store multiple operational code images and configurations as well as error logs and dumps. The PCMCIA disk used previously can be plugged into the PCMCIA slots of the new MSS Server Module to transfer its contents to the internal disk.

- PCMCIA flash

  An optional 20 MB plug–in PCMCIA flash memory can be used to store the operational code and configuration. A built–in 1 MB flash memory is used to store the firmware. Previously, built–in 12 MB of flash memory was provided to store firmware, operational code and configuration.

As before, the *MSS Server Module* includes 512 KBytes of L2 cache for the processor, 8 MBytes of buffer memory for ATM, a single 155M ATM interface via the 8260/8265 backplane, a standard serial service port, 2 PCMCIA slots (which can be used for a hard disk, a Voice/Data/Fax modem or flash memory). The serial port, modem and ethernet ports provide access for out of band management of the MSS Server.

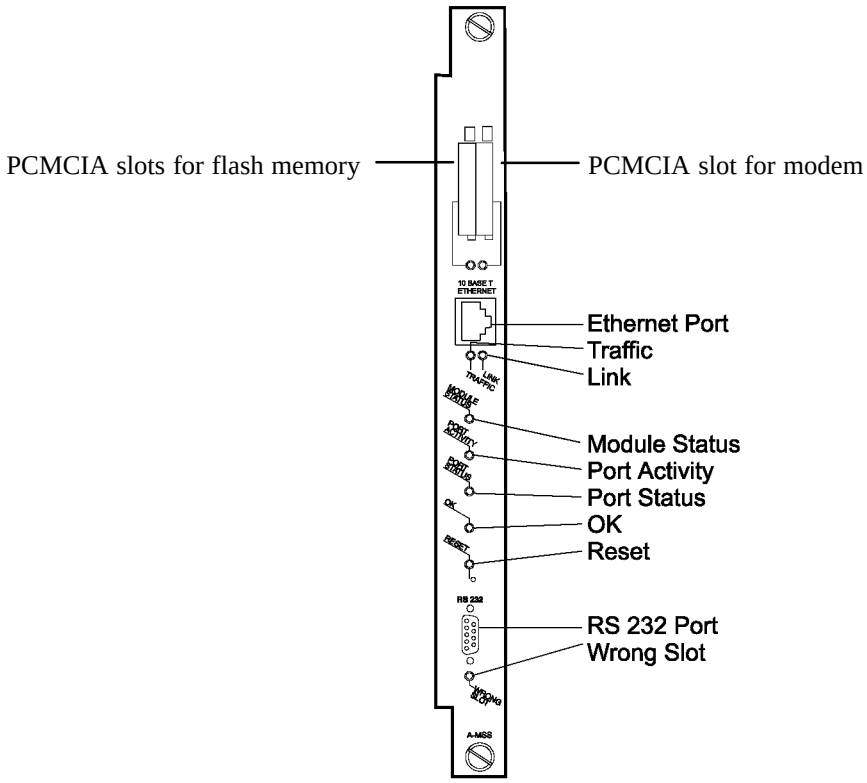

Figure 13. One–wide A–MSS Server Module

Figure 13. shows the face plate of the new one–wide A–MSS Server Module. Up to 7 one–wide A–MSS Server Modules can be installed in a 8260 hub and up to 4 one–wide A–MSS Server Modules can be installed in the 8265 hub in the 4 compatibility slots that support 8260 style modules.

1. 4. 1. 2. MSS Server Memory Upgrade option

A memory upgrade option for the previous 2–slot MSS Server Module and the 8210 Model 001 MSS Server is available to increase processor memory from 32 MB to 64 MB. The 64MB EDO memory

upgrade replaces the previous 32 MB memory. 64 MB of processor memory is required for MSS Server Release 2.0/2.0.1. Updated firmware (v3.0) is also required to support MSS Server 2.0/2.0.1 software.

1. 4. 1. 3. 8210–001 and MSS Server Module upgraded to 64 MB memory

An Engineering Change (EC) has been made to the 8210–001 MSS Server and the 2–wide MSS Server module to include 64MB of EDO memory as standard. 64 MB of processor memory is required for MSS Server 2.0/2.0.1 software.

1. 4. 2. Overview of MSS Release 2.0/2.0.1 Functions

With MSS 2.0/2.0.1, new functions are being provided in

- MSS Server 2.0/2.0.1
- MSS Route Switching Clients

1. 4. 2. 1. Overview of MSS Server R2.0/2.0.1 functions

The MSS Server functions in Release 2.0/2.0.1 include

- Route Switching Server for IP
- SuperELAN II
- Classical IP enhancements
- IP Multicast over ATM
- Routing enhancements
- Bridging enhancements
- LAN Emulation enhancements
- ATM Interface enhancements
- Trouble–shooting enhancements
- Configuration, Monitoring and Management enhancements.

These functions are described below.

1. 4. 2. 1. 1. Route Switching Server for IP.

Route Switching is an IBM extension of NHRP [12] to do *zero–hop routing* between legacy LAN attached end–stations on different IP subnets. To perform zero–hop IP routing, an end–station runs the *MSS RouteSwiching Client* (RSC) software (described on page 45) and communicates with the NHRP Server in the MSS Server using enhanced NHRP. An NHRP Server that has been enabled to do Route Switching is called a Route Switching Server (RSS). NHRP and NHRP Server are described in 1. 3. 2. 3. on page 24. The RSC sends a NHRP Resolution Request with IBM extensions to the RSS. The RSS uses NHRP to find a short–cut to the destination IP address. If a short–cut is available, the RSS sends a NHRP Resolution Reply with IBM extensions to the RSC and includes the destination's MAC address. The RSC uses this information to send packets directly to the destination MAC and bypasses intermediate IP routers (zero–hop routing). If a short–cut is

not possible (due to mis–match in LAN type, MTU size, bridging type etc.), the RSS sends a NHRP Resolution Reply rejecting the request. The different IP subnets may be inter–connected with routers or SuperELAN. The Route Switching Server details are described in section 2. on page 46.

1. 4. 2. 1. 2. SuperELAN II

SuperELAN II is a major enhancement of the SuperELAN function introduced in MSS Server R1.1 (described in section 1. 3. 2. 2. on page 22).

- SuperELAN Bridging (SEB)

  The Short–Cut Bridging (SCB) concept introduced in R1.1 (see 1. 3. 2. 2. on page 22) has been extended to token–rings and is called *SuperElan Bridging (SEB)*. SEBs can bridge between token–ring as well as ethernet ELANs and run independently of the base bridge in the MSS Server. Furthermore, multiple instances of SEBs are supported with each instance running its own spanning tree. The SEB runs a different spanning tree which and does not interact with normal bridge spanning tree BPDUs. Each SEB instance also has its own Bridging Broadcast Manager (BBCM) as well as its own Dynamic Protocol Filter (DPF) VLANs. Thus, service providers that provide ATM LAN Emulation services can use one MSS to support multiple customers and keep the customer networks fire–walled from each other. SEB details are described in section 3. on page 53.

- Dynamic Protocol Filtering VLAN enhancements

  Two additional types of DPF VLANs are now supported. *Sliding Window VLANs* allow the most general VLAN definitions by allowing a user to specify an offset, mask and arbitrary data in the packet as the criteria for VLAN membership. *MAC Address VLANs* allow VLAN membership based on the source MAC address in the packet.

  Also, all DPF VLANs are now enhanced to show active membership of VLANs by end–user MAC addresses. DPF VLAN details are described in section 3. on page 53.

1. 4. 2. 1. 3. Classical IP enhancements

Significant enhancements to Classical IP support include the Distributed ATM ARP Server and the 1577+ Client.

- Distributed ATM ARP Server

  In previous releases, ARP services for a CIP LIS could only be provided by a single ARP Server (see 1. 2. 2. 2. on page 14). Now, multiple ARP Servers can cooperate to share this load on a LIS. CIP clients (1577 and 1577+) can be configured to use different ARP Servers for load distribution. Distributed ATM ARP Servers use the *Server Cache Synchronization Protocol (SCSP)* to synchronize their ARP databases. Thus, a client using the services of one ARP Server can ARP for clients that are being served by other ARP Servers. Distributed ATM ARP Servers provide redundancy for 1577+ clients and load balancing for 1577+ and 1577 clients. For 1577

clients, ARP Server redundancy support was provided in R1.1 and can be combined with the Distributed ATM ARP Server. Similarly, Default IP Gateway Redundancy can also be combined with the Distributed ATM ARP Server. Load balancing and redundancy allow users to build larger CIP LISs. Distributed ATM ARP Server details are described in section 4. on page 62.

- 1577+ Client

  The 1577+ client, also know as the *Classic2 client*, is an enhancement of the 1577 Classical IP client (described on page 15). The 1577+ client eliminates the single point of failure associated with the 1577 client's ARP services when such services are provided by Distributed ATM ARP Servers. 1577+ clients have the ability to switch to backup ARP Servers in the same Logical IP Subnet (LIS) if their current ARP Server fails. When the primary ARP Server becomes active again, the 1577+ Client will switch back to the primary. Thus, 1577+ clients have continuous connectivity across the Logical IP Subnets (LIS) in cases of ARP Server failure. Note, the MSS Server already provides redundancy for 1577 Clients (as described on page 16), but it requires that the primary and backup ARP server be connected to the same ATM switch. For 1577+ clients, the primary and backup ARP servers can be connected to different switches. The 1577+ Client is described in section 5. on page 66.

1. 4. 2. 1. 4. IP Multicast over ATM

Multicasting is the simultaneous transmission of a single data element to multiple destinations. IP multicasting is a layer 3 protocol based on RFC 1112 [20] and is used for transmitting IP datagrams from one IP source to many IP destinations in a local or wide–area network. It is the Internet abstraction of hardware multicasting. IP multicasting and broadcasting has a number of uses, but the most well known use is for distributing audio and video. IP multicasting and broadcasting are also used by major routing protocols (RIP, OSPF, BGP etc.) and networking services such as BOOTP and DLS. In IP multicasting, the source of the traffic knows nothing about the receiving destinations. The source simply sends IP datagrams to a particular IP multicast group address. In IP version 4, addresses in the range 224.0.0.0 and 239.255.255.255 are termed *Class D* or *Multicast Group Addresses*. The set of destinations receiving the multicasted IP datagrams is known as a host group or multicast group. Membership into the multicast group is dynamic and initiated by the multicast group members. Hosts can join and leave at any time. There are no restrictions on the location or number of members in a multicast group. The IP Multicast solution assumes that the link layer provides a connectionless transport service and some form of broadcast or multicast addressing.

ATM networks using UNI 3.0 and UNI 3.1 signalling provide neither the connectionless transport nor the multicast addressing capability needed for IP multicasting. The point–multipoint VC capability of ATM can be used for IP multicasting, but the key limitation is that the sender must have prior knowledge of the ATM address of each destination. In order for ATM attached devices to use the layer 3 IP multicast services, a mapping layer or emulation service needs to be provided by the ATM network. The IETF IP over ATM Work Group has proposed RFC 2022 as a standard for providing IP multicast services natively over an ATM network [21]. The MSS implementation of IP Multicast over ATM is based on services outlined in [21]. The MSS implementation of IP Multicast over ATM includes the Multicast Address Resolution Server (MARS), MARS Client and MultiCast Server (MCS). IP Multicast over ATM is described in section 6. on page 68.

- MARS Server

  The Multicast Address Resolution Server (MARS) is an extended analog of the ATM ARP Server (introduced in 1. 2. 2. 2. on page 14). It provides the necessary connection and addressing services required by IP Multicast and Broadcast services over ATM networks. This service is similar to the ATM ARP Server, but while the ATM ARP Server provides a one–to–one mapping between an IP unicast address and an ATM address, the MARS server provides a one–to–many mapping between an IP multicast/broadcast address and one or more ATM addresses. Each MARS Server can serve a cluster of ATM endpoints (MARS Clients) within a single Logical IP Subnet (LIS). Each LIS can be served by a single MARS Server. Up to 32 MARS Servers can be defined in a single MSS Server. MARS Server details are described in section 6. on page 68.

- MARS Client

  The MARS client is an extension of the ATM ARP Client (introduced in 1. 2. 2. 2. on page 14). A MARS client is an ATM end–point that can support IP multicasting natively over ATM. Each MARS client is configured with the ATM address of one primary MARS server and zero or more backup MARS servers. The MARS client sends control messages to the MARS server when joining or leaving a particular multicast group and when sending traffic to a particular multicast group. However, the MARS server does not get involved with the multicast data path, only the creation of it. A point-to-point VC (P2P VC) is established between the MARS server and each MARS client for private control messages. A point-to-multipoint (P2MP) Cluster Control VC (CCVC) is used by the MARS server to asynchronously update all MARS clients of changes in the multicast group. Each MARS client is served by a single MARS Server. One MARS Client can be defined per ATM interface (real or virtual). MARS Client details are described in section 6. on page 68.

- MultiCast Server (MCS)

  The MultiCast Server (MCS) is like a proxy that can forward IP multicast traffic on behalf of other MARS clients. The source of the multicast traffic establishes a P2P VC to the MCS which is then responsible for establishing a P2MP VC to send traffic to all members of the multicast group. The MCS can be beneficial when multiple sources exist in a multicast group. Without a MCS, each source must establish a P2MP VC to each destination. With a MCS, each source only needs to establish one data P2P VC to the MCS, thus eliminating the need for meshes of VCs. Another advantage of MCS is that very little signalling is required when group membership changes, as compared to the mesh approach. A drawback of the MCS is that it can be a bottleneck. The MARS architecture allows for the multicast data path to be created from meshes of P2MP VCs or the use of MultiCast Servers (MCS) on a per multicast group basis. The choice of which service to use is configurable and based on application and network requirements. The use of both meshes of VCs and MCSs is supported by the same MARS Server. The only requirement is that a particular multicast group be of one type, not both. A single MCS can serve multiple group addresses, but a group address can only be served by one MCS. MCS details are described in section 6. on page 68.

1. 4. 2. 1. 5. Routing enhancements

Routing support for two new protocols, APPN and Banyan VINES has been added. Previously routing support was limited to IP, IPX and AppleTalk. IP routing function has also been been enhanced.

- APPN Routing

  APPN Routing support is being provided for ethernet and token–ring LECs, as well as natively on ATM interface (real or virtual) using RFC 1483 encapsulation [8]. The MSS Server can be configured as an APPN Network Node (NN) that can provide Directory services, Routing services and Management services to APPN end–nodes and APPN Low–entry networking (LEN) end–nodes. The NN can provide Dependent LU Requestor (DLUR) services to legacy PU 2.0 nodes containing dependent LUs. The NN also exchanges network topology information with other NNs.

  The NN can act as an intermediate node for session data from adjacent nodes. This is called Intermediate Session Routing (ISR) and is supported on ethernet and token–ring LECs. The NN also supports High Performance Routing (HPR) by using Automatic Network Routing (ANR) instead of using ISR. HPR reduces processing overhead in intermediate routing nodes and moves error recovery and flow control to the end–points of a HPR connection. As an end–point of a HPR connection, the MSS uses the Rapid Transport Protocol (RTP) to take advantage of HPR. HPR is supported on ethernet and token–ring LECs as well as the native ATM interface. APPN also allows direct communication between nodes, thus eliminating the routing by intermediate NNs, when the nodes are connected to the same shared transport facility (eg. LAN). This type of network is called an APPN Connection Network and is supported on ethernet and token–ring LECs as well as the native ATM interface.

  APPN is loaded into MSS memory only if APPN is configured (see Dynamic linking and loading on page 44). APPN is described in section 7. on page 75.

- Banyan VINES Routing

  Banyan VINES Routing support is being provided for ethernet and token–ring LECs. VINES Internet Protocol (VINES IP) is used to route packets, VINES RouTing Update Protocol (VINES RTP) is used to exchange routing information with other VINES routers, VINES Address Resolution Protocol (VINES ARP) is used to assign internet addresses to clients, VINES Internet Control Protocol (VINES ICP) provides diagnostics and support functions. Banyan VINES routing is described in section 8. on page 76.

- IP enhancements

  RIP v2 support has been added and is compliant with IETF RFC 1723 [19]. RIP v2 is an extension of the RIP protocol that allows routers to share important additional information such as subnet–mask and authentication key. Thus, RIP v2 can reliably learn subnets and is inherently more secure. RIP v2 also reduces interruptions to other stations on the network by advertising

routes to a well known IP multicast address instead of broadcasting them. RIP v2 is backward compatible with the existing RIP implementations. RIP v2 is described in 9. 6. on page 80.

The number of IP interfaces supported on a single network interface is unlimited now. Previously, only 32 IP interfaces could be configured on a single network interface.

1. 4. 2. 1. 6. Bridging enhancements

Base bridging function has been enhanced and a new type of bridge is being introduced.

- SuperELAN Bridge

  A new type of bridge called the SuperELAN Bridge (SEB) is being introduced. SEB (described on page 37) runs independently of the base bridge (described on page 16).

- Duplicate MAC address support in SR–TB environments

  The SR–TB Translational bridge is being enhanced to support duplicate MAC addresses on the SRB network. SR–TB Translational bridges are typically used to allow communication between ethernet and token–ring stations when routing between the two is not possible or desirable. Previously, the SR–TB bridge did not work in networks with duplicate MAC addresses on the SRB network. Duplicate MAC addresses on SRB networks are extensively used in SNA networks to provide redundancy and load–balancing across multiple IBM 3745 communication controllers. The SR–TB has been enhanced to support 2 instances of up to 7 MAC Addresses on the SRB network. Additionally, the enhanced SR–TB can load balance traffic from ethernet stations destined to a MAC address that is duplicated on the SRB network. Thus, ethernet attached stations can now get redundancy and load–balancing when accessing mainframes connected to a SRB network. Duplicate MAC address support for SR–TB environments is described on page 81.

- 1483 SVC Bridging Support

  Previously, the MSS Server provided the capability to transmit bridged traffic natively over ATM using LLC/SNAP encapsulation as specified in RFC 1483. The support was limited to transmitting and receiving bridged frames over PVCs (see 1. 3. 2. 8. on page 31). 1483 Bridging support is being extended to include SVCs. With SVC support, the MSS Server uses signalling to find the destination ATM address. Now, the user can simply specify the destination ATM address instead of specifying a VPI/VCI and configuring the PVC in the intermediate ATM switches. 1483 SVC support for bridging is described in section 9. 2. on page 77.

- Bridging Broadcast Manager (BBCM) support for NetBIOS NameSharing

  The OS/2 LAN Server has a feature called *NameSharing*, which allows the same NetBIOS name (file server name) to be used on multiple LAN interfaces of the server. NameSharing is used to overcome the NetBIOS limitation of 254 sessions per LAN interface. Without this technique, a file server could not be accessed by more than 254 clients at one time. Another benefit of

NameSharing is that it allows a LAN Server's clients to be distributed across multiple LAN interfaces of the server, thus balancing network traffic across the server's interfaces. In previous releases of the MSS Server, if Broadcast Manager (BCM) or Bridging Broadcast Manager (BBCM) was enabled for NetBIOS, it associated each learned NetBIOS name with a single unicast MAC address. Subsequently, if BCM/BBCM transformed a NetBIOS broadcast, it always directed the packet to the associated unicast MAC address. However, this defeated the purpose of NameSharing. With R2.0/2.0.1 of the MSS Server, BCM/BBCM can support networks with NameSharing servers. BBCM support for NameSharing servers is available with the normal bridge (described on page 16) as well as the SuperELAN bridge (described on page 37). BCM is described on page 11 and BBCM is described on page 23. Details of NameSharing support are described on page 79.

1. 4. 2. 1. 7. LAN Emulation enhancements

Enhancements have been made to LAN Emulation to improve performance and network reliability. BCM/BBCM has been enhanced to support duplicate NetBIOS names.

- LE–ARP Cache enhancements

  The LEC maintains databases that map MAC addresses to ATM addresses, and ATM addresses to VCCs. These databases are now stored in binary trees for fast searches instead of doing linear searches in arrays, as was done previously. An option is also being provided to allow users to configure permanent entries (i.e. they won't age out) in the MAC to ATM address database. This can be used to configure addresses of heavily used devices and avoid the LE ARP process. The maximum LE–ARP cache size has also been significantly increased. LE–ARP cache enhancements are described in section 9. 3. on page 77.

- LES initiated pacing during congestion

  In large networks, ATM switches may not have the processing power or memory to handle the addition of large numbers of LECs when they join an ELAN simultaneously. This is sometimes the case if the LANE Service or switched network recovers from a fault or is restarted. If this congestion occurs in the network, the Add Party messages from the LES or BUS to the ATM switch to add LECs may be explicitly rejected or dropped, which generally forces the LECs to attempt to join the ELAN again. If the LECs do not randomly delay their attempt to rejoin (and many do not), additional network congestion occurs which can prevent, or at least greatly delay, stabilization of the network.

  In order to reduce the rate of signalling messages in the ATM network, the MSS LES and BUS will randomly delay the addition of joining LECs to the ELAN when signalling congestion has been detected in the network. This algorithm will spread out the signalling traffic associated with stations joining the ELAN, which will allow the ATM switches to handle a much larger network of ATM attached devices. LES initiated pacing is described in section 9. 4. on page 78.

- BCM and BBCM support for NetBIOS NameSharing

  BBCM support for NetBIOS NameSharing is described in 1. 4. 2. 1. 6. on page 41.

1. 4. 2. 1. 8. Performance enhancements

Performance has been improved with faster hardware and tuning of the LAN Emulation software.

- Faster hardware

  The new single–wide A–MSS Server Module (described in 1. 4. 1. 1. on page 33) has a 66% faster processor that improves overall performance of all MSS functions.

- LE_ARP Cache enhancements

  The LE_ARP cache databases are now stored in binary trees for fast searches instead of doing linear searches in arrays, as was done previously. This improves the performance of all protocols that use LAN Emulation. LE_ARP cache enhancements are described on page 42.

1. 4. 2. 1. 9. ATM Interface enhancements

The ATM interface has been enhanced for more efficient use by the protocols.

- ATM LLC Multiplexing

  The native ATM interface has been enhanced to support ATM LLC Multiplexing which multiplexes connections (SVCs and PVCs) and ATM addresses for the new native ATM protocols. Previously, only PVC multiplexing was supported (described on page 31). Fewer ATM addresses and VCCs means reduced signalling load on the network. The operation of ATM LLC multiplexing is transparent to users and is automatically used when these protocols are configured. ATM LLC Multiplexing is supported for SCSP and APPN. ATM LLC multiplexing is described in section 9. 1. on page 77.

1. 4. 2. 1. 10. Trouble–shooting enhancements

Several functions have been added to the MSS Server to enhance its trouble–shooting capabilities. These are described in 9. 8. on page 82.

- IP TraceRoute

  The IP TraceRoute function has been enhanced to allow the user to specify the source IP address, data size, number of probes to send, time between probes and the maximum time to life (TTL).

- IPX Ping

  The *IPX Ping* function has been enhanced to allow the user to specify the source network, source node, data size and the time between pings.

- IPX TraceRoute

  IPX TraceRoute is a new function that is similar to IP TraceRoute that allows a user to find different routed paths in the IPX network.

- IPX RecordRoute

  IPX RecordRoute is a new function that allows the user to trace a specific routed path taken by an IPX ping packet to the destination. It also traces the path taken by the response from the destination.

1. 4. 2. 1. 11. Configuration, Monitoring and Management enhancements

Several functions have been added to the MSS Server to enhance its configuration, monitoring and management capabilities.

- Dynamic Reconfiguration (DR)

  Dynamic Reconfiguration allows users to change the configuration of the MSS Server without requiring a reboot to activate the changes. DR is critical for providing high availability services. DR is described in section 9. 9. on page 83.

- Dynamic Linking and Loading (DLL)

  Dynamic Linking and Loading allows selective loading of functions from the operational code image into processor memory. If a function is not configured, it is not loaded into memory, thus making more memory available to the configured functions. In this release, the APPN function is available as a DLL module. DLL is described on page 83.

- Time activated re–boot

  An option is provided to automatically re–boot the MSS server on a user specified date and time. The user can also select the operational code and configuration to load from non–volatile storage when the box is automatically re–booted. Time activated re–boot is described on page 83.

1. 4. 2. 2. Overview of MSS Route Switching Clients

To use Route Switching (described on page 36), a LAN attached end–station runs the MSS Route Switching Client (RSC) software and uses the Route Switching Server (RSS), which is in the MSS Server, as its default gateway. The RSC communicates with the RSS to learn layer–2 short–cuts for layer–3 destinations.

The RSC software is specific to the layer–3 protocol, LAN type, end–station operating system and network interface characteristics. In this release, RSC support is being provided for IP on ethernet and token–ring attached PCs. For ethernet attached PCs, the RSC is implemented as a *Protocol Driver* between the station's IP Protocol stack and the Network Device Driver. The RSC is being made available as software that would work with ethernet cards and IP stacks from different vendors. For token–ring attached PCs, the RSC is embedded in the token–ring device driver and is being made available for specific token–ring adapters. The RSC software can be downloaded from the World Wide Web. RSC details are described in section 2. on page 46.

2. Route Switching

Route Switching is an *IP switching* technique used by the MSS Server to allow *zero–hop routing* between legacy LAN attached end–stations on different IP subnets. The terms Route Switching, IP Switching and Zero–Hop Routing are used inter–changeably in this document. To switch IP, an end–station runs the *MSS Route Switching Client (RSC)* software which communicates with the *MSS Route Switching Server (RSS)* using the NHRP protocol [12] with IBM extensions. The RSS resides in the MSS Server with the IP router. The RSC in the end–station requests a shortcut path for switching IP traffic directly to the destination. If the RSS responds with a shortcut path, the RSC sends packets directly to the destination and the RSS/Router is no longer in the forwarding path. Thus, traffic starts flowing at switching speed instead of being delayed by routers and the need for expensive routers is significantly reduced. Furthermore, the destination is not required to participate in Route Switching. In fact, the destination is completely unaware that packets from another subnet are being switched to it. This makes Route Switching ideal for high volume server to client traffic. RSC software can be deployed on relatively few servers that transmit a large volume of data while clients continue to use routers and require no changes. Another benefit of Route Switching is that it does not impose any special requirements on intermediate network devices like LAN switches. Route Switching can work with any LAN switch or bridge. Furthermore, there is no user configuration required to run the RSC in the end station. Once installed, all discovery and switching is automatic.

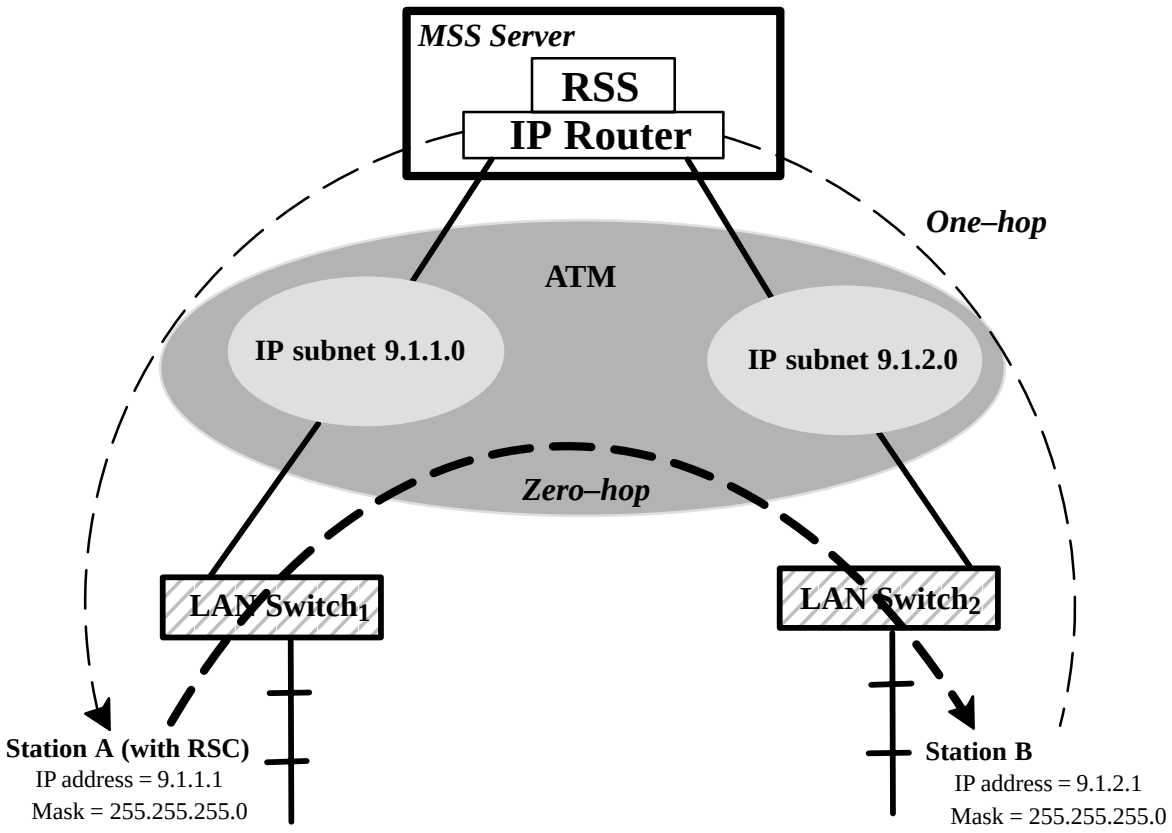

Figure 14. Zero–hop routing with Route Switching

Route Switching is a technique whereby a token–ring or ethernet attached end–station can resolve a layer–3 network address on a subnet other than its own to a layer–2 MAC address. Route Switching can effectively create a switched connection for traffic, thus enabling an end station to bypass one or more routers when sending data. By moving the routing function to the edge of the network, the price/performance of switching is maximized and latency is minimized. With the routing function distributed in end–stations, the need for traditional routing is reduced and one does not have to over–provision routing capacity in anticipation of demand, as is the case with stand–alone routers. Routing capacity is automatically available where it is needed.

An example of zero–hop routing using Route Switching is shown in Figure 14.. Station A on subnet 9.1.1.0 is sending data to station B on subnet 9.1.2.0. The initial flow from station A to station B (not shown) goes through the IP router in the MSS Server. The RSC software in station A detects that inter–subnet data is being sent to station B via the default gateway router in the MSS Server. The RSC sends a NHRP Resolution Request to the RSS in the MSS Server requesting a short–cut to the IP address of the destination (B). The RSS determines the MAC address of the destination and sends it to station A in a NHRP Resolution Reply. For subsequent packets addressed to B's IP address, the RSC in station A switches the destination MAC address from the IP router to the MAC address of B. The packets are now directly switched from LAN $Switch_{1\ to}$ LAN $Switch_2$ without passing through the IP router.

2. 1. Route Switching Server and Clients

The RSS function in the MSS Server is implemented as an extension of the NHRP Server (described on page 25). To activate the RSS, the **Route–Switching** option under NHRP configuration must be enabled. Multiple RSSs (one per MSS Server) can be active in the network.

The RSC is implemented as end–station software available free of charge on the World Wide Web at url **http://www.networking.ibm.com/nes/nesswitc.htm** for many different platforms. For ethernet attached PCs, the RSC is implemented as a *Protocol Driver* between the station's IP Protocol stack and the Network Device Driver. The following operating systems and network interfaces are supported for ethernet:

- Windows NT 4.0 or higher (NDIS 4.0)
- Windows 95 (32 bit ODI)
- DOS, Windows, Windows for Workgroups (16 and 32 bit ODI)
- Novell IntraNetware Server R.3.12 and later (32 bit server ODI)

For token–ring attached PCs, the RSC is embedded in the token–ring device driver and is being made available for specific token–ring adapters. The following operating systems and network interfaces are supported for token–ring:

- Windows NT 4.0 or higher

- Windows 95
- Novell IntraNetware Server
- OS/2 2.11 or higher
- DOS, Windows, Windows for Workgroups (16 bit ODI)

2. 2. Route Switching Process

This section examines the Route Switching process in more detail. During initialization, the RSC discovers the MAC address of the RSS which is also its default gateway. The RSC broadcasts a NHRP Resolution Request packet with the destination IP address set to loopback (127.0.0.1) and the source IP set to 0. The RSS responds with a NHRP Resolution Reply and the RSC saves the source MAC address of the RSS from the reply. If multiple replies are received, the RSC saves multiple MAC addresses (between 8 and 16 depending on the platform). When examining an outgoing IP packet, the RSC compares the destination MAC address to the RSS MAC address (or list of MAC addresses). If a match is found, the RSC selects that MAC address as the address of the default gateway and discards the rest of the gateway MAC addresses.

When inter–subnet traffic is being forwarded to the default gateway router's MAC address, the RSC issues an NHRP Resolution Request to determine the layer–2 information associated with the destination IP address. The layer–2 information consists of the destination MAC address, and for token–ring networks, the MAC address and Routing Information Field (RIF). The RSS on the source subnet (also called the ingress RSS) determines if the destination subnet is directly attached. If the destination can only be reached via other routers/RSSs, then the RSS sends a NHRP Resolution Request to the next RSS until the RSS on the destination subnet (also called the egress RSS) is reached. The egress RSS finds out the layer–2 information associated with the destination using ARP and returns it to the ingress RSS in a NHRP Resolution Reply via the routed path. The ingress RSS builds a NHRP Resolution Reply that includes the destination layer–2 information and forwards it to the RSC. The RSS also specifies a holding time for which the layer–2 information is valid.

The RSC caches the supplied layer–2 information to build the data link header for frames transmitted to the associated destination IP address (instead of the usual procedure of sending the frames using the router's layer–2 information in the data link header). Frames to the destination are then delivered via the normal layer–2 procedures which result in a shortcut VCC between the ingress and egress LAN switch. The RSC refreshes active cache entries by sending out NHRP Resolution Requests before the holding time provided by the RSS expires. If the layer–2 information associated with the destination changes, the egress RSS sends a NHRP Purge to the ingress RSS, which in turn sends a NHRP Purge to the RSC to invalidate its cached layer–2 information for the target IP host. If the network topology changes, the ingress RSS initiates the NHRP Purge.

If the RSC does not receive a NHRP Resolution Reply from the RSS, then the RSC will not re–attempt to find a short–cut to this destination for 16 seconds. If the reply is received but it indicates that a short–cut is not available to the destination, then the RSC will not re–attempt to find

a short–cut to this destination for 15 minutes. This ensures that an excessive number of short–cut requests are not sent to the RSS.

Although, the example cited in Figure 14. only shows LANs (and ELANs), the network topology between the ingress and egress RSS is a routed topology which can be arbitrarily complex. For zero hop routing, the only requirement is that intermediate routers must also be RSSs or MSS based NHRP Servers. If an intermediate router is not a RSS or MSS based NHRP Server, then a partial short–cut is used up to this intermediate router. For end–stations, the requirements are that the source and destination be of the same LAN type (ethernet or token–ring), support the same type of bridging (transparent or source–route bridging), and MTU size of the source should not be greater than the MTU size of the destination. If the end–station conditions are not met, the RSS will deny a short–cut to the RSC and the RSC will continue to use the normal routed path. In either case client, packets continue to be forwarded without loss. It is important to note that the RSS's decision to grant or deny a short–cut is based on a case by case basis. Figure 15. shows Route Switching in a network with multiple RSSs with a Classical IP LIS between the ingress and egress RSS. Since the source and destination MTU sizes are different, only RSC A can establish a short–cut to destination B. RSC B has a smaller MTU so it is denied a short–cut by RSS–2 and uses its default gateway (IP Router–2) for forwarding. However, since RSS–2 is NHRP enabled, it can bypass RSS–1 and short–cut directly to the destination (RSC A), resulting in a one–hop path.

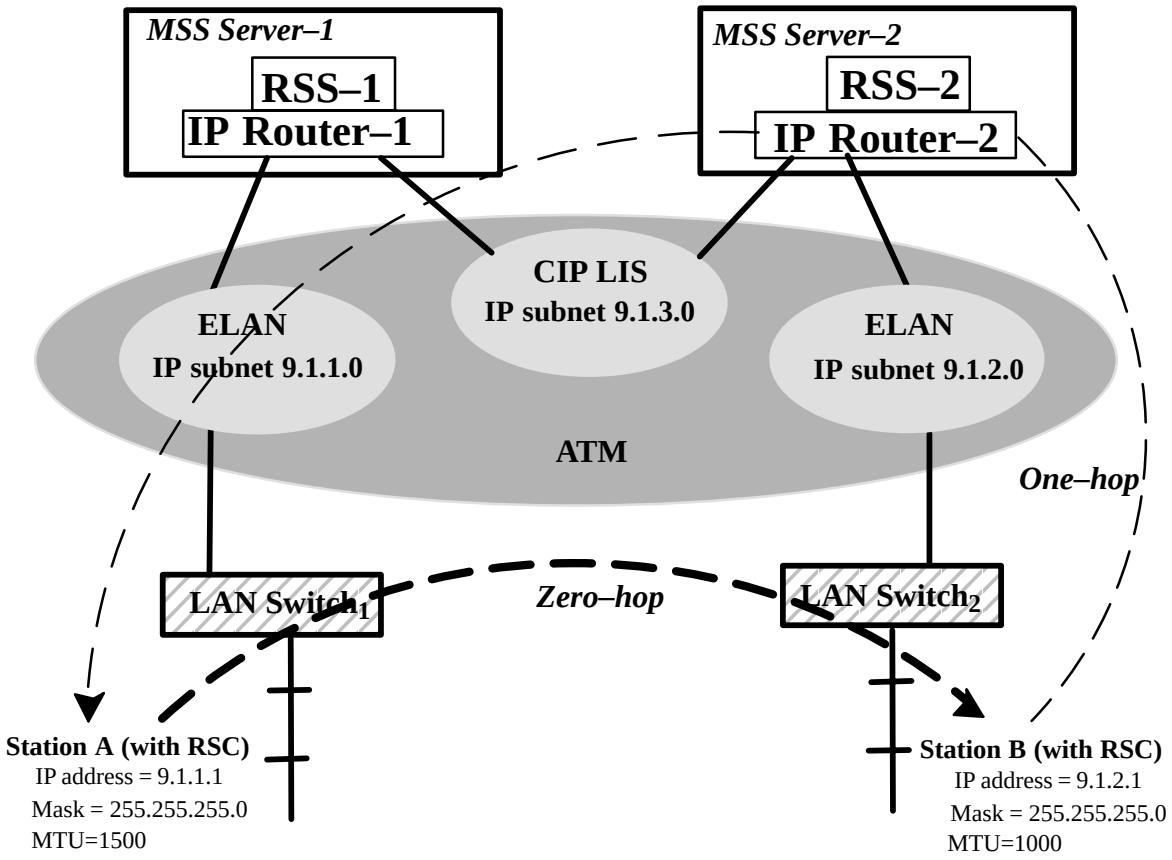

Figure 15. Route Switching in a heterogeneous network

2. 3. Route Switching Considerations

Having discussed the basic Route Switching mechanisms, lets focus on some important details. One significant network design issue is to decide if an independent SuperELAN bridge path (described on page 37) should exist between the different subnets in addition to the IP router path. While a router only path is supported by Route Switching, it has some limitations. Foremost is the fact that in SuperELAN enabled networks, the ingress LAN switch can set up a short–cut VCC to the egress LAN switch on a different ELAN, thus allowing Route Switching to work automatically. Without SuperELAN, this type of short–cut VCC is not possible. To make Route Switching work in non–SuperELAN networks, the ingress RSS registers the destination's MAC[2] and ATM address with the ingress LES/BUS. This allows an ingress LAN switch to find the destination using normal LAN Emulation address resolution (described on page 10). The LES/BUS in the MSS Server has also been modified to correctly handle LAN emulation flush and unknown frames in this environment. Thus, in non–SuperELAN Route Switching environments, the ingress LES/BUS must be in a MSS Server R2.0 or higher. Furthermore, a maximum of 50 such registrations are allowed per ELAN, which limits the number of active short–cuts to 50 per ELAN in this environment[3].

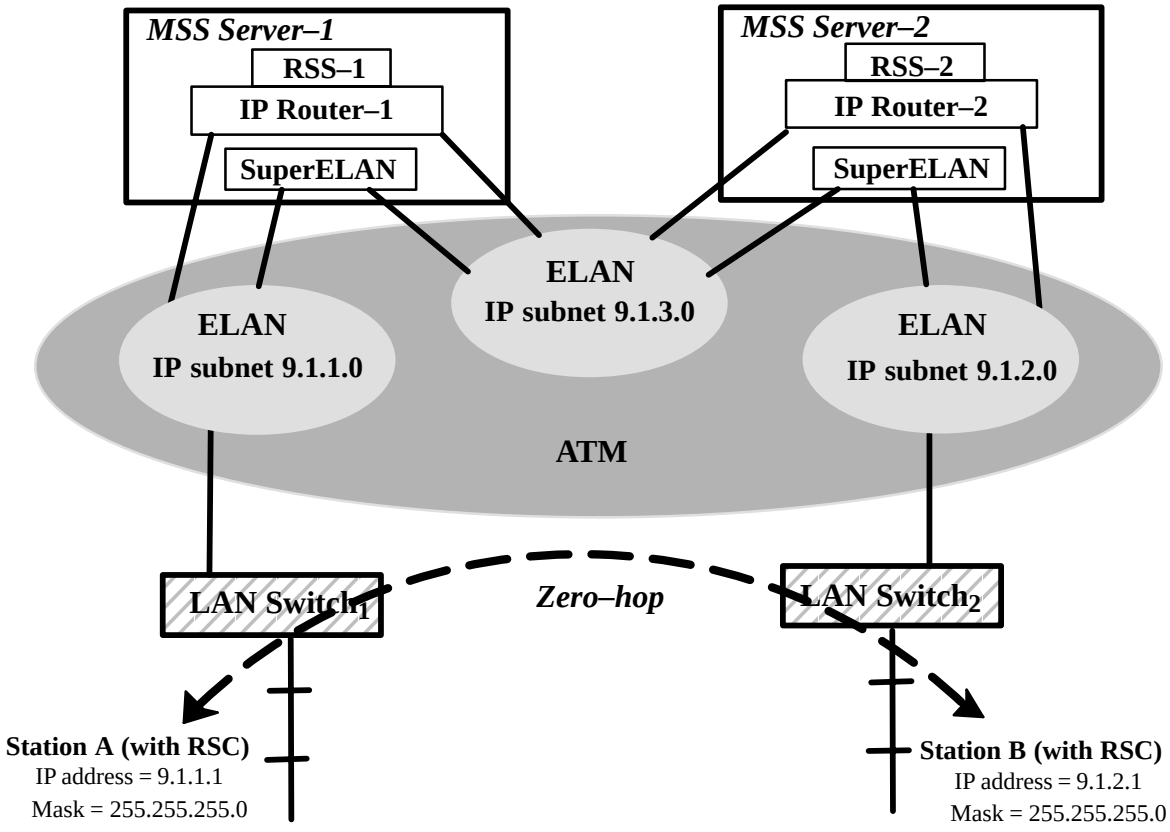

Figure 16. Route Switching in a heterogeneous network

2. For non–SuperELAN token–ring networks, the ingress RSS will register a virtual route descriptor with the ingress LES/BUS instead of the destination's MAC address.

3. For non–SuperELAN token–ring networks, a maximum of 50 virtual route descriptors can be registered but the number of short–cuts permitted is not limited to 50. This is because many end–stations can be reached using a single route descriptor.

Figure 16. shows a network that is routed as well as SuperELAN bridged. Route Switching works automatically and no registration with the LES/BUS is required.

Another issue to consider is the kind of LAN switches being used in the network. Some LAN switches can snoop on IP traffic and are capable of blocking inter–subnet traffic. Obviously, these types of LAN switches would prevent an inter–subnet short–cut from being established.

2. 4. Comparison between MSS Route Switching and 3Com's Fast IP

Fast IP is a software based IP switching technique available from 3Com corporation. Fast IP is similar to MSS Route Switching in that it is used to switch IP traffic (zero–hop routing) between end–stations on different subnets. Fast IP, like Route Switching eliminates the router bottleneck by performing the routing function in the end–station. Both techniques use the routed path until a switched path is discovered. They will continue to use the routed path if a switched path cannot be established. Both techniques use NHRP to discover the switched path. Both techniques require software in the end–station which is a simple layer running as an extension of the LAN adapter device driver and is available at no charge. However, there are some important differences between the two techniques.

The most important difference is that to switch IP traffic, Fast IP requires the Fast IP client software to be running on the source as well as the destination. Route Switching on the other hand allows a source end–station to establish a switched connection to the destination even if the destination end–station is not running the Route Switching software. The destination in this case will send traffic back via the normal routed path (asymmetric paths). Requiring Fast IP software at both ends eliminates the need for a server like the RSS, but it also precludes asymmetric switching. Asymmetric switching feature is attractive because in large networks it might not be feasible or desirable to upgrade all end–stations. Furthermore, it might not even be possible to install this software on some end–systems because appropriate drivers might not exist for the end–system platform or operating system. With Route Switching, the software can be installed on servers that generate the bulk of traffic, thus allowing a large volume of network traffic to be switched with no changes to end–stations.

Another significant difference between Route Switching and Fast IP is that with Fast IP the destination replies directly to the source's MAC address. This means that the destination must be able to directly reach the source using a layer–2 address which implies that both source and destination are in the same layer–2 broadcast domain. However, most routed networks are segmented to prevent layer–2 broadcasts from flooding indiscriminately. In segmented networks, Fast IP cannot establish inter–subnet switched connections. One way to get around this problem is to use *VLAN Tag Switching*, but that requires intermediate switches to support VLAN tagging, which most switches available today do not. MSS Route Switching solves this problem in two ways. First, with SuperELANs (described on page 53) and protocol VLANs (described on page 59), the broadcast domain is segmented into subnets to match the IP subnet infrastructure. This preserves the broadcast domains, yet allows inter–subnet and inter–ELAN switched connections to be established. The second way Route Switching addresses this problem is by allowing a *router mode* in the Route Switching Server (RSS). In routed mode, the RSS allows inter–subnet and inter–ELAN connections to be established by registering the destination with the ingress LES.

Other differences between the two techniques are that Route Switching supports ethernet as well as token–ring end stations, while Fast IP only supports ethernet end stations today. Route Switching client software is available for Windows 95, Windows NT, Windows 3.x, DOS, Novell Netware and OS/2. Fast IP client software is only available for Windows 95 and Windows NT today. Route Switching is supported on IBM as well as non–IBM adapters. Fast IP is only supported on 3Com adapters today.

3. SuperELAN II

The SuperELAN concept was first introduced in MSS Server 1.1 but its short–cut bridging was limited to a transparent bridge environment (see 1. 3. 2. 2. on page 22). Additionally, the SuperELAN was limited to a single spanning tree. These restrictions existed because short–cut bridging (SCB) was coupled with the ASRT transparent bridging function of the MSS Server. In MSS Server 2.0/2.0.1, the SuperELAN bridge runs independently of the ASRT transparent bridge function and also supports source–route token–ring networks. The new bridge is referred to as a SuperELAN Bridge (SEB). Multiple SEBs can exist on a single MSS Server, each running an independent spanning tree.

The BBCM and DPF features of the previous release have been extended to the new SuperELAN bridge and multiple instances of BBCM and DPF VLANs can now be configured (one per SuperELAN). New DPF VLAN functions have also been added. *Sliding window VLANs* can be used to create VLANs based on user–defined policies and *MAC address VLANs* can be used to restrict VLAN membership by MAC addresses.

SuperELAN Bridging, Bridging BroadCast Manager, and Dynamic Protocol Filtering VLANs combine to provide a reliable distributed ELAN with strong controls on the propagation of broadcast frames.

3. 1. SuperELAN Bridging (SEB)

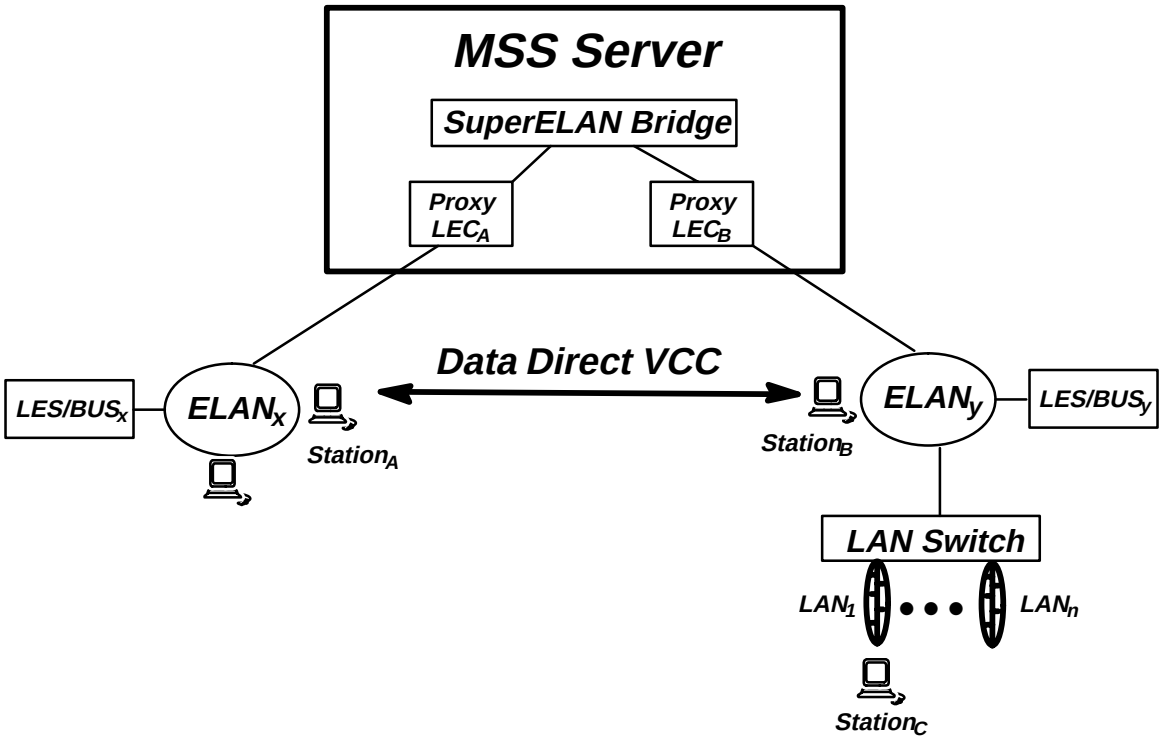

Figure 17. Simple Short–Cut Bridge Example

SuperELAN bridging extends the forwarding mechanisms of a bridge so that certain LE control frames are forwarded along with data frames. The information in the control frames allows clients to

establish inter–ELAN Data Direct VCCs. The SuperELAN bridge supports MAC address filters in addition to BBCM and DPF VLANs.

In source routing environments, the SEB does not serve as a true source route bridge, but as a "source route aware" bridge. In effect, the SEB learns the locations of route descriptors just as it learns the locations of MAC addresses. All rings directly attached to the SEB have the same ring number.

Figure 17. illustrates a simple SuperELAN bridging configuration. To communicate with $Station_B$, $Station_A$ issues an LE_ARP request for the MAC address of $Station_B$. $Station_A$ may also send unicast data frames destined for $Station_B$ to the BUS for $ELAN_X$. *MSS* LEC_A receives the LE_ARP request, and either answers the request (if it can) or forwards the request to the LES for $ELAN_Y$ through *MSS* LEC_B. *MSS* LEC_A also receives any data frames forwarded by the $ELAN_X$ *BUS*, and the SuperELAN Bridge forwards those frames to the BUS for $ELAN_Y$ through *MSS* LEC_B. Assuming LES_Y answers the LE_ARP request, the LE_ARP response is received by *MSS* LEC_B, and forwarded to LES_X by *MSS* LEC_A. LES_X forwards the the LE_ARP response to $Station_A$, and $Station_A$ then sets up a Data Direct VCC to $Station_B$.

Like a transparent bridge, the SEB learns which LAN destination are on which ports. Known LAN destinations need only be forwarded out a single port, while unknown LAN destinations must be transmitted on all ports. Figure 18. contains a slightly more complex example of SuperELAN bridging. To prevent a loop, the spanning tree protocol deactivates one of the ports on one of the MSS Servers. Assume that *MSS Server 1* has been elected root of the spanning tree, that *SCB–*LEC_f on *MSS Server 3* has been deactivated by the spanning tree protocol, and that LEC_B wishes to communicate with LEC_C.

As in the first example, LEC_B sends an LE_ARP for LEC_C, and may send some initial unicast data frames destined for LEC_C to the BUS for $ELAN_Y$. If this were a transparently bridged network, all data frames from LEC_B to LEC_C would pass through *MSS Server 1* and *MSS Server 2*. For a SuperELAN bridged network, only a few initial unicast data frames traverse this path. After receiving the LE_ARP response, LEC_B can setup a Data Direct VCC to LEC_C, and future communications bypass the SEBs.

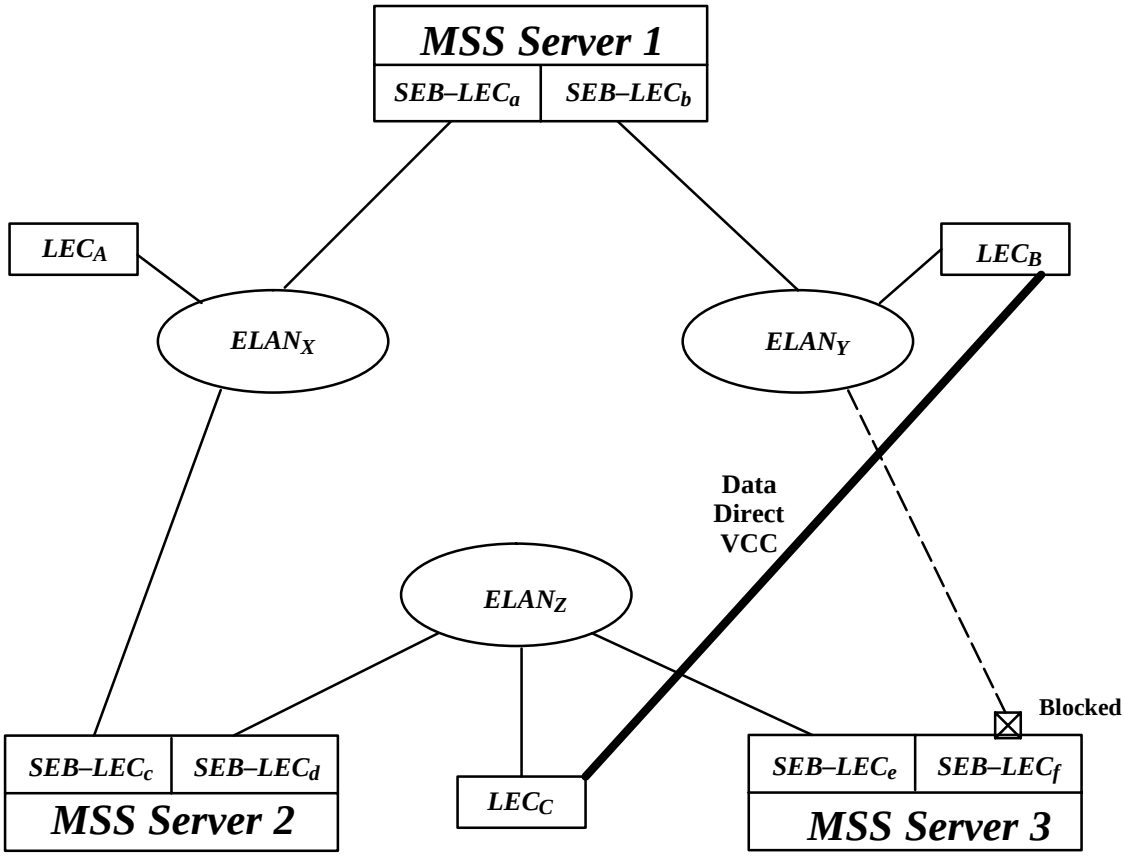

Figure 18. More Complex Short–Cut Bridge Example

SuperELANs can be connected using a traditional bridge or router. Figure 19. depicts a simple configuration with two SuperELANs connected via an MSS Server Bridge. This might be necessary when connecting a Token–Ring SuperELAN with an Ethernet SuperELAN. In this example, data direct VCCs can be setup between ELANs A and B and between ELANs C and D. Data traffic between either ELANs A or B to ELANs C or D will flow through the MSS Server bridge.

SEB LECs may be configured to route MSS Server supported protocols. SEB LECs route a protocol only if the LEC is configured with an address for that protocol. This allows protocol subnetting within a SuperELAN. This is useful for subnetting one protocol, for example IP, while allowing other protocols, like SNA, the performance advantage of having a large flat LAN infrastructure.

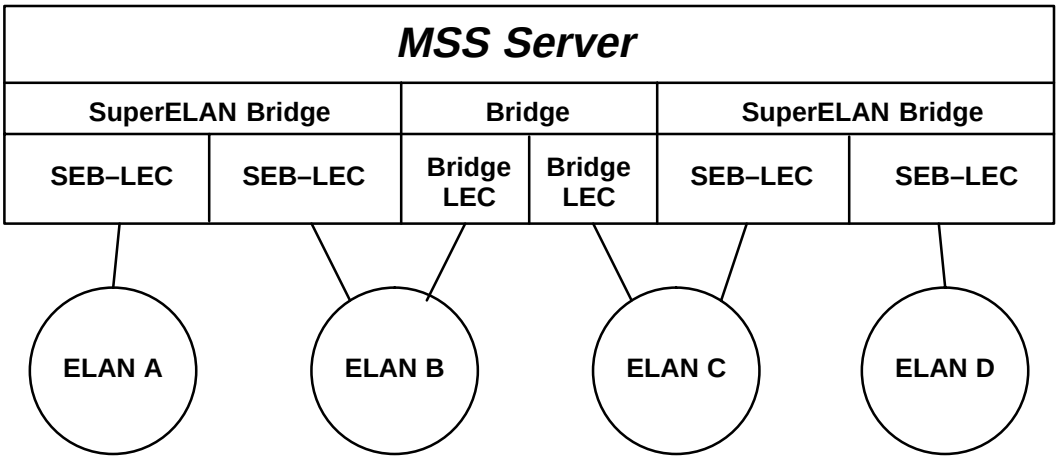

Figure 19. Routing with Subnets Partitioned Across ELANs

3. 1. 1. SuperELAN Data Frame Forwarding Rules

Forwarding of data frames on SEB ports is now done based on the destination MAC address or the Routing Information Field (RIF). If a data frame without a RIF is received and it has a broadcast or multicast destination, the frame is forwarded over all bridge ports in forwarding state, except the source port. If the frame has a unicast destination and the destination is unknown, then the frame is forwarded over all bridge ports in forwarding state. If the destination is known to be on the source port, then the frame is discarded. Otherwise, the frame is forwarded over the port associated with the destination MAC address.

3. 1. 2. SuperELAN Control Frame Forwarding Rules

Inter–ELAN DDVCCs are established as a result of forwarding certain control frames between ELANs. Control frame forwarding affects the following frame types: LE ARP requests and responses, LE NARP requests, and LE FLUSH requests and responses. Other control frame types are *not* forwarded by any LECs. The control frame forwarding rules from MSS Server Release 1.1 have been enhanced to support forwarding based on route descriptors.

3. 1. 3. SuperELAN Learning

If a data frame with a RIF is received and it has a broadcast or multicast destination, then the frame is forwarded over all bridge ports in forwarding state. If the frame has a unicast destination and the local ring number is the last ring number in the RIF, the frame is forwarded based on the destination MAC address as described above. If the frame has a unicast destination and the local ring number is not the last ring number in the RIF, the frame is forwarded based on the next Route Descriptor (RD) in the RIF. If the next route descriptor is unknown, then the frame is forwarded over all bridge ports in forwarding state. If the next route descriptor is known to be on the source port, then the frame is discarded. Otherwise, the frame is forwarded over the port associated with the next route descriptor.

The SEB does not modify the RIF in any way. RIF and MAC learning occurs on every incoming SEB frame. If the frame does not contain a RIF, the source MAC address is learned. If the frame contains a RIF, the SEB learns the source MAC address only if the frame originated from a locally attached ELAN. If the frame contains a RIF and the frame did not originate from a locally attached

ELAN, the SEB learns the source RD associated with the RIF. The SEB also learns its local ring number from incoming frames with a RIF.

3. 1. 4. SuperELAN Spanning Tree

SuperELAN bridges use a modified IEEE 802.1d spanning tree algorithm to maintain the SuperELAN topology. Only SuperELAN bridges participate in the SuperELAN spanning tree protocol. The SuperELAN spanning–tree protocol is based on the IEEE 802.1d Spanning tree with the following exceptions:

- The SuperELAN Bridge Protocol Data Unit (BPDU) and Topology Change Noticication (TCN) frame does not use SAP 0x42. Instead, the frame is SNAP encapsulated. The SNAP encapsulation uses an IBM OUI of 0x10–00–5A and an ethertype of 0x80–D7.

- A two–byte SuperELAN ID field and a 20–byte LES address field are appended to the end of BPDU and TCN frames. These fields are used by the SuperELAN bridge to detect loops caused by external non–SuperELAN bridges. If such a loop exists, the SuperELAN BPDUs and TCNs will be forwarded by the external bridge and the SuperELAN bridge will detect the loop. In this case, the SuperELAN bridge will continue forwarding and force the non–SuperELAN bridge to block and prevent the loop. If a redunant SuperELAN bridge is configured, then it will participate in the SuperELAN spanning tree and will block its ports to prevent a loop.

A MSS Server may contain multiple logical SuperELAN bridges, each with its own instance of the SuperELAN spanning tree. Encapsulation and STP frame formats are the same for both Ethernet and Token–Ring SuperELANs.

3. 1. 5. SuperELAN Migration

R1.1 style SuperELANs using SCB are still supported and migration to R2.0 style SuperELAN using SEB is optional. However, both types of SuperELAN personalities cannot be active in the same MSS Server. A utility has been provided to ease the migration from R1.1 SuperELAN configuration to R2.0 SuperELAN.

3. 2. Bridging BroadCast Manager (BBCM)

Broadcast frames are generally sent by a source station to find the MAC address of the destination station. Examples of such frames are IP ARP requests and NetBIOS NAME_QUERY commands. In these examples, the source station is trying to find a specific higher level address or name that maps to some unknown MAC address. Since the destination MAC address is unknown, all stations are interrupted and queried for the higher layer address or name. Most stations discard the request because it does not pertain to them, and only one station replies to the query.

These broadcast frames are harmful in bridged environments because every broadcast frame is transmitted on every LAN segment. This causes an interruption at each station on the bridged LAN, consumes precious network bandwidth, and uses processing cycles. BroadCast Manager (BCM, described on page 11) was designed to limit the effects of such broadcast frames. For some of the most common types of these broadcasts, BCM attempts to turn the broadcast frame into a unicast frame. BCM was implemented as an optional extension for the BUS in R1.0 of the MSS Server. In R1.1, Bridging BroadCast Manger (BBCM, described on page 23) extended the BCM concept to a bridged environment, providing broadcast management for IP and NetBIOS within the bridge of the MSS Server. These broadcast frames, which are normally transmitted on every active bridge port, are transformed into unicast frames and handed to the bridge for forwarding. If the unicast destination exists in the bridge database, then the frame only goes out one port. If the unicast destination does not exist in the bridge database, then the frame must be transmitted out every active bridge port. However, even in this case, the unicast frame does not interrupt every station on all segments. Bridging BroadCast Manger (BBCM) was introduced in MSS Server Release 1.1 as a method to limit the effects of broadcast frames in a bridged or SuperELAN environment. In MSS Server 2.0/2.0.1, BBCM is extended to apply to SuperELAN bridges as well as the ability to have multiple logical instances of BBCM within a single MSS Server.

3. 3. Dynamic Protocol Filtering Virtual LANs

Protocol VLANs (PVLANs, described on page 23) were introduced in Release 1.1 of the MSS Server to dynamically learn and control protocol and subnet specific broadcasts in a SuperELAN environment. Protocols supported were IP, IPX and NetBIOS. In MSS Server Release 2.0/2.0.1, DPF VLANs have been extended to the new SuperELAN II environment. Multiple instances of DPF VLANs can now be configured (one per SuperELAN). New DPF VLAN functions have also been added that are not protocol specific. *Sliding window VLANs* can be used to create VLANs based on user–defined policies and *MAC address VLANs* can be used to restrict VLAN membership by MAC addresses. An enhancement has also been made for all VLANs to view current VLAN membership by end–station MAC addresses.

DPF PVLANs are based on protocol and subnet, monitoring traffic over each bridge port and learning the protocols and subnets being used on that port. For each configured subnet, the subset of bridge ports on which traffic for that subnet is being received is referred to as the *forwarding domain* of that subnet. DPF VLANs manage the forwarding domains for each subnet. Broadcast/multicast frames for a particular subnet are not forwarded on bridge ports that are not in the forwarding domain of that subnet.

In MSS Server 2.0/2.0.1, DPF VLANs are extended to include sliding window VLANs and MAC address VLANs. These VLANs are not protocol–specific. DPF again monitors the forwarding domain of each VLAN and limits broadcast/multicast frames to the forwarding domain of the VLAN.

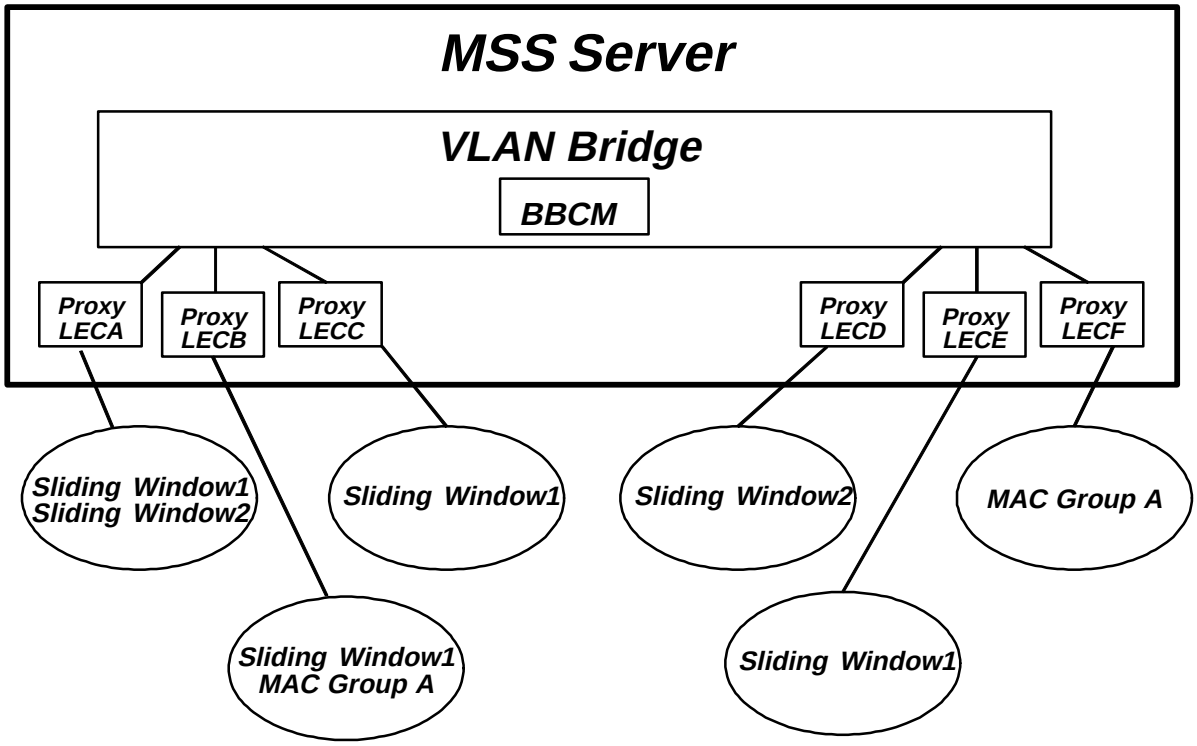

Figure 20. Dynamic Protocol Filtering Example

Figure 20. shows an example DPF VLAN configuration. In the example environment, VLANs are intermixed across a bridged environment. Assume that the sliding window VLANs *Sliding Window1* and *Sliding Window2* are configured for DPF, as is the MAC address VLAN *MAC Group A*. *MSS* LEC_A and *MSS* LEC_C are identified as members of *Sliding Window1's* forwarding domain based on traffic seen on those ports. Multicast traffic matching *Sliding Window1* is then restricted to those two ports. Similarly, the *MAC Group A* VLAN is learned to be on *MSS* LEC_B and *MSS* LEC_F, and multicast frames sourced from a MAC address in *MAC Group A* is limited to these two ports.

DPF VLANs, including sliding window VLANs and MAC address VLANs, process only multicast and broadcast frames. If a frame matches multiple VLANs, then it is forwarded over those ports in the forwarding domain of *all* matched VLAN(s). DPF VLAN filtering occurs before any manually configured outbound filters.

3. 3. 1. Sliding Window VLANs

Sliding window VLANs are the most general PVLAN definition. The user specifies an offset into the frame, a mask, and value. Any frames matching the specified mask and value at the given offset are included in the sliding window VLAN, and forwarded only on those ports in the forwarding domain of the sliding window VLAN.

Each frame passed to the DPF VLAN processing logic is compared against all defined sliding window filters. Each matching sliding window filter results in the set of ports in the forwarding domain of the sliding window filter. The output of the sliding window DPF VLAN processing routine is the logical OR of the forwarding domains of all matched sliding window filters.

3. 3. 2. MAC Address VLANs

MAC address VLANs allow the user to define a collection of MAC addresses which form a VLAN. Ports receiving frames with a source MAC address in the VLAN definition are included in the forwarding domain of the VLAN. All multicast frames sent by such sources are limited to the MAC Address VLAN's forwarding domain.

Each frame passed to the DPF VLAN processing logic is compared against all defined MAC address groups. Each matching MAC address group results in the set of ports in the forwarding domain of the MAC address group. The output of the DPF VLAN processing routine for MAC address groups is the logical OR of the forwarding domains of all matched MAC address group filters. Note that a single MAC address may appear in multiple MAC address groups. The number of configured MAC addresses in a MAC address VLAN is limited only by available memory.

3. 3. 3. VLAN Membership

All DPF VLANs have been enhanced with the ability to display the VLAN membership by end–station MAC addresses. VLAN membership can be accessed via the console or SNMP.

4. Distributed ATM ARP Server

An ATM ARP Server is used by IP clients in an ATM environment to resolve IP addresses into ATM addresses. Native transmission of IP over ATM is known as Classical IP (CIP) which is standardized by the IETF in RFC 1577 [9]. Classical IP support was introduced in MSS Server 1.0 (described in 1. 2. 2. 2. on page 14). The ATM ARP Server as proposed by the IETF in RFC 1577 is a single point of failure in CIP networks. However, the MSS Server implementation of the ARP Servers includes an option to configure redundant ARP Servers. In MSS Server 2.0/2.0.1, the ARP Server has been further enhanced to provide distributed ATM ARP Services in CIP environments. In the distributed ARP Server model, multiple ARP Servers can be actively serving clients in a single Logical IP Subnet (LIS). When used in conjunction with 1577+ clients (described on page 66), distributed ARP Servers provide load balancing as well as redundancy. In a 1577+ environment, distributed ARP Servers can be located anywhere in the network and are not restricted to the same ATM switch, as is the case when redundancy is provided using MSS value add services (described on page 16). For 1577 clients, distributed ARP Servers provide load balancing while redundancy can be provided using MSS value add services. The MSS implementation of the distributed ARP Server adheres to the *Server Cache Synchronization Protocol" (SCSP)* IETF Internet Draft [11]. SCSP is a general purpose protocol for distributing server databases over ATM.

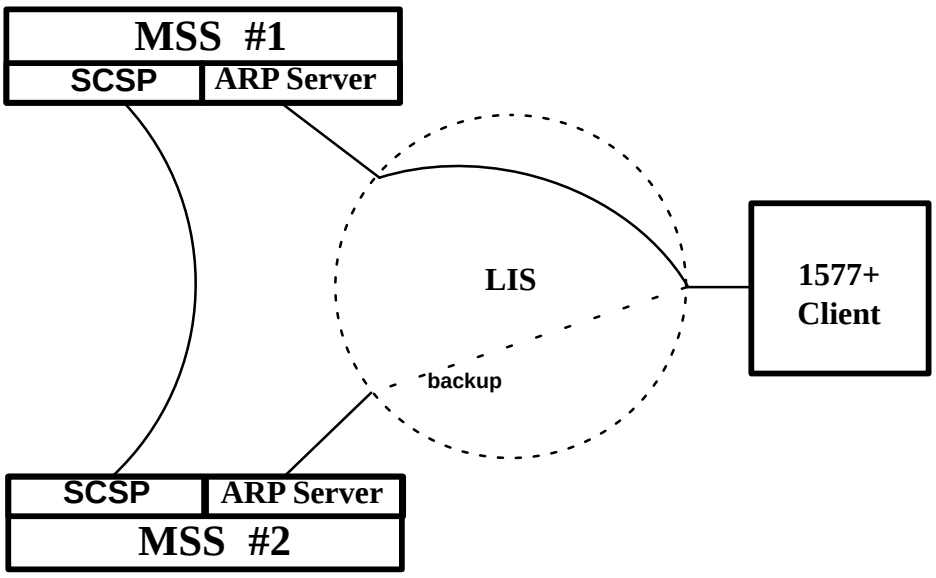

Figure 21. Simple Distributed ARP Server Configuration

In Figure 21., two distributed ARP servers are defined on a Logical IP Subnet (LIS), one in each of the two MSSs. The SCSP entities in each MSS are configured with the ATM address of the other ARP server. A private VCC is established between the two MSSs to exchange ARP database information using the SCSP protocol. The SCSP entity in each MSS interacts with the local ARP Servers to get and report cache changes. The client supports 1577+ and is configured with the ATM

address of both ARP servers, MSS #1 as primary and MSS #2 as backup. If the client loses contact with it's primary, it will register with MSS #2. At this point, MSS #2 will have the full ARP resolution database (as did MSS #1) and will provide ARP resolution service to the client.

Although, any number of distributed ARP Servers can serve a LIS, in practice two or three are sufficient. Servers must be configured with the ATM addresses of their Directly Connected Servers (DCSs). All servers in the Server Group (SG) need not be directly connected (i.e. meshed). There are no restrictions on the topology of these servers as long as there is some graph which connects all servers in the group. A fully meshed topology would provide the best redundancy at the expense of more channels and increased traffic. A ring topology provides redundancy for a single server failure and limits the amount of propagated traffic.

Figure 22. shows three ARP servers configured in a ring (as well as mesh) topology on one LIS. MSS #1 is configured with two DCSs, MSS #2 and MSS #3. MSS #2 is configured with two DCSs, MSS #1 and MSS #3. MSS #3 is configured with two DCSs, MSS #1 and MSS #2. With this configuration, the ARP database is duplicated among all three servers. Client #1 is configured with MSS #1 as it's ARP Server. Client #2 is configured with MSS #3 as it's primary server. Both clients are configured with MSS #2 as their backup server. Client #1 can ARP MSS #1 for the address of Client #2 and get resolution even though Client #2 is registered with MSS #3. Similarly, Client #2 can ARP MSS #3 for the address of Client #1.

If MSS #1 fails, client #2 can switch to MSS #2 as the server with no loss of connectivity. Similarly if MSS #3 goes down, Client #2 can switch to MSS #2 without loss of connectivity.

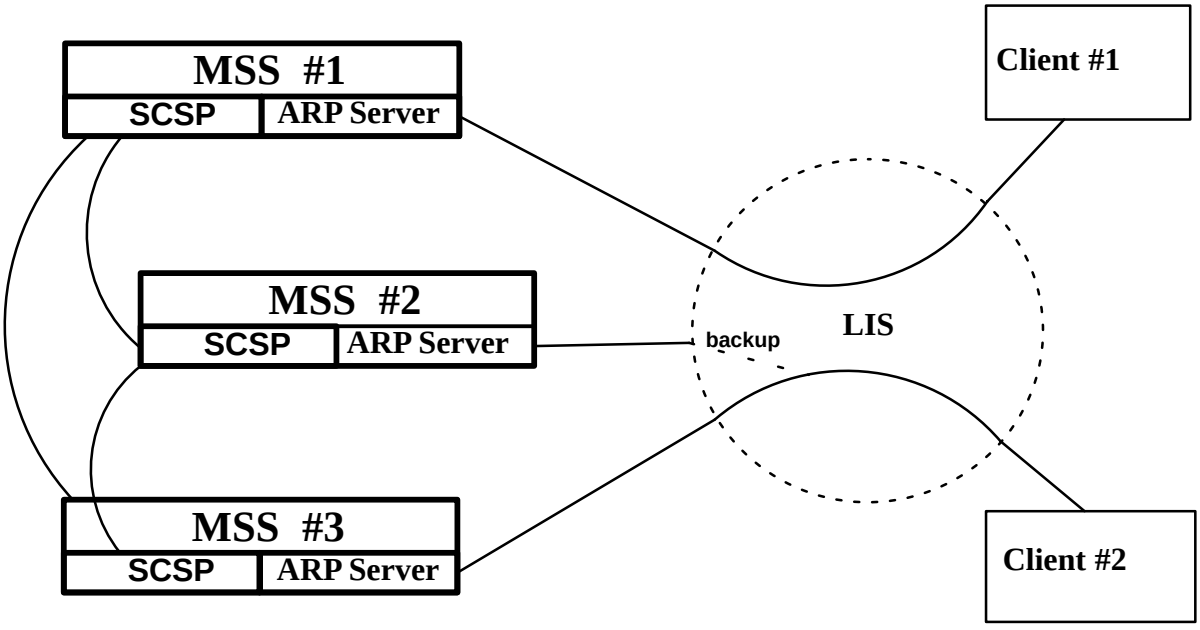

Figure 22. Three ARP Servers

4. 1. Combining Distributed ARP Server and Redundancy

To provide load balancing and redundancy support for 1577+ as well as 1577 clients, the distributed ATM ARP Server can be combined with the redundancy services of MSS R1.1. An ARP server can be configured as a backup redundant server as well as a distributed server. Figure 23. shows an example of combining distributed ARP Servers and R1.1 style redundancy. Two distributed ARP Servers are actively providing ARP service to clients on a LIS. The RFC 1577 compliant client uses *ARP Server$_x$* as its ARP server, while the RFC 1577+ compliant client is configured with A*RP Server$_y$* as the primary and *ARP Server$_x$* as the backup ARP server. The databases of the two ARP Servers are synchronized via the SCSP protocol. If *ARP Server$_y$* were to fail, the 1577+ client would switch to its backup server, *ARP Server$_x$*.

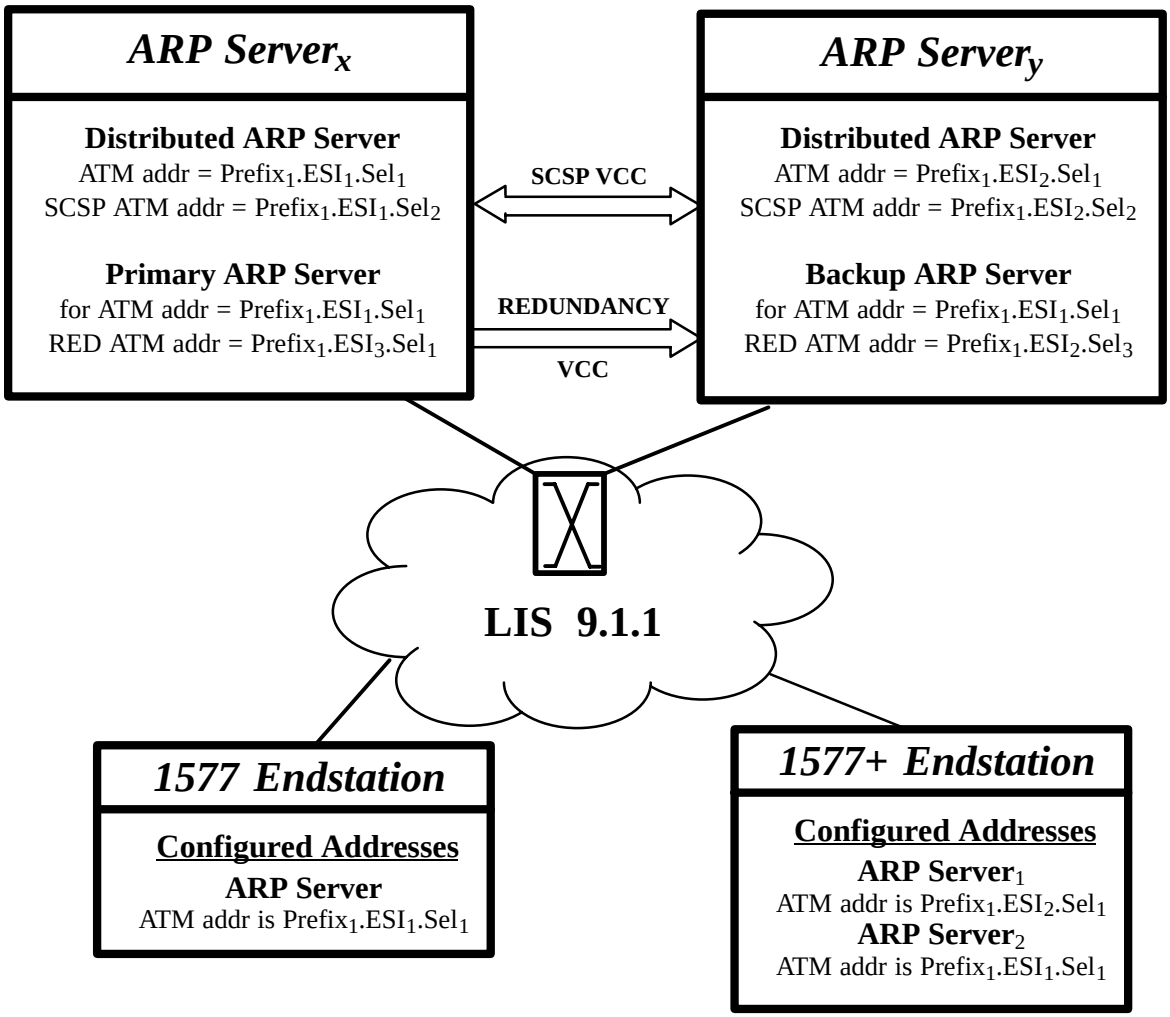

Figure 23. Distributed ARP services for LIS with 1577 and 1577+ Clients

In order to provide ARP Server redundancy for the RFC 1577 compliant client, *ARP Server$_x$* is designated as the primary ARP Server and *ARP Server$_y$* is designated as the backup ARP Server. The primary and backup ARP Servers must be connected to the same ATM switch. The primary and

backup ARP Servers provide redundancy support by employing the same basic mechanisms as in Release 1.1 i.e., establishment of Redundancy VCC etc. (described on page 16). However, there is one significant difference. If *ARP Server*$_x$ fails and *ARP Server*$_y$ takes over, *ARP Server*$_y$ will register the ATM Address of *ARP Server*$_x$ in addition to its other ATM addresses; thus, *ARP Server*$_y$ will simultaneously represent both ARP Server ATM Addresses. As in the previous release, if *ARP Server*$_x$ subsequently recovers and re–establishes the Redundancy VCC to *ARP Server*$_y$, *ARP Server*$_y$ will de–register *ARP Server*$_x$'s ATM Address so that *ARP Server*$_x$ can resume its role as one of the active ARP Servers for the LIS.

4. 2. SCSP Protocol

The Distributed ARP Server implementation depends upon the *Server Cache Synchronization Protocol (SCSP)* to synchronize the existing ARP database with other ARP servers in the Server Group (SG). A SG is thus the set of ARP Servers serving one Logical IP Subnet (LIS).

ARP Server distribution is defined in each MSS participating in the Server Group. Each ARP Server on the LIS is given the same Server Group ID (SGID). Each server is also given a unique Local Station ID (LSID). The value for the LSID is the IP address of the corresponding ARP Server. Each participating ARP Server also contains a list of Directly Connected Servers (DCS) with which this server attempts to synchronize.

Once an ARP Server in an MSS is initialized, the SCSP will attempt to bring up a session with all of the configured DCSs. SCSP has a two phase initialization with each DCS:

1) Establish a communication path from the Local Server (LS) to the DCS. The SCSP ”hello” protocol is used to establish and maintain this connection.

2) Perform initial ”cache alignment” between the LS and DCS. This itself is accomplished with a three step process in each LS and DCS:

   i) Negotiate a Master/Slave relationship.

   ii) Send a cache summarization to the partner. Filter the received cache summarization for ”newer” entries.

   iii) Request and supply full cache updates to partner. Process received cache updates.

SCSP synchronizes the ARP cache as changes occur. The ARP server notifies the local SCSP when changes occur and SCSP propagates these changes to the DCSs. When SCSP receives a cache update from a DCS, it notifies the LS so that it can update its cache. It also propagates the update to other DCSs.

An ARP Server leaves a Server Group by simply stopping communication with its DCSs. The hello protocol will discover that the Local Server is no longer exchanging SCSP packets and declare the LS down. The DCSs will periodically check if the LS is back by sending Hello messages.

Details of the SCSP protocol are described in [11].

5. 1577+ Client

The 1577+ client, also know as the *enhanced 1577 client* or the *Classic2 client*, is an enhancement of the RFC 1577 [9] compliant Classical IP client. The 1577+ client eliminates the single point of failure associated with the 1577 client's ARP services when such services are provided by Distributed ATM ARP Servers (described in section 4. on page 62). 1577+ clients have the ability to switch to backup ARP Servers in the same Logical IP Subnet (LIS) if their primary ARP Server fails. When the primary ARP Server becomes active again, the 1577+ Client switches back to the primary. Thus, it is possible for 1577+ clients to have continuous connectivity across the Logical IP Subnets (LIS) in cases of ARP Server failure. It is also possible to distribute the load on the ARP Servers in a network with distributed ARP Servers if groups of 1577+ clients in the same LIS have connectivity to different primary ARP Servers for their ATM addresss resolution services.

The 1577+ client allows the user to configure a primary ARP Server and several backup ARP Servers. The 1577+ client always tries to connect to its primary ARP Server, but if it cannot, it will connect to a backup ARP Server. The 1577+ client is responsible for initiating the registration process with the ARP Server. Previously, the ARP Server initiated the registration process via an Inverse ARP Requests. With 1577+, the client registers via an ARP Request with the source and target protocol addresses set to the client's IP address. The 1577+ client re–registers with the ARP Server periodically (the re–registration period is a configurable value with a default of 15 minutes). The client is also responsible for refreshing entries in its ARP cache by issuing ARP requests to the ARP Server. The MSS Server implementation of the 1577+ client is based on IETF Draft [10].

In Figure 24. , two ARP servers are defined on a Logical IP Subnet (LIS), one in each of the two MSSs. The ARP servers are configured such that they will duplicate each other's ARP database (see Distributed ARP Server on page 62 for details). The 1577+ Client is configured to have two ARP servers, MSS #1 as the primary and MSS #2 as the backup. If the client loses contact with it's primary, it will register with MSS #2. At this point, MSS #2 will have the full ARP resolution database (as did MSS #1) and will provide ARP resolution service to the 1577+ Client. The 1577+ Client will switch to the primary ARP Server (MSS# 1) as soon as the primary becomes active again.

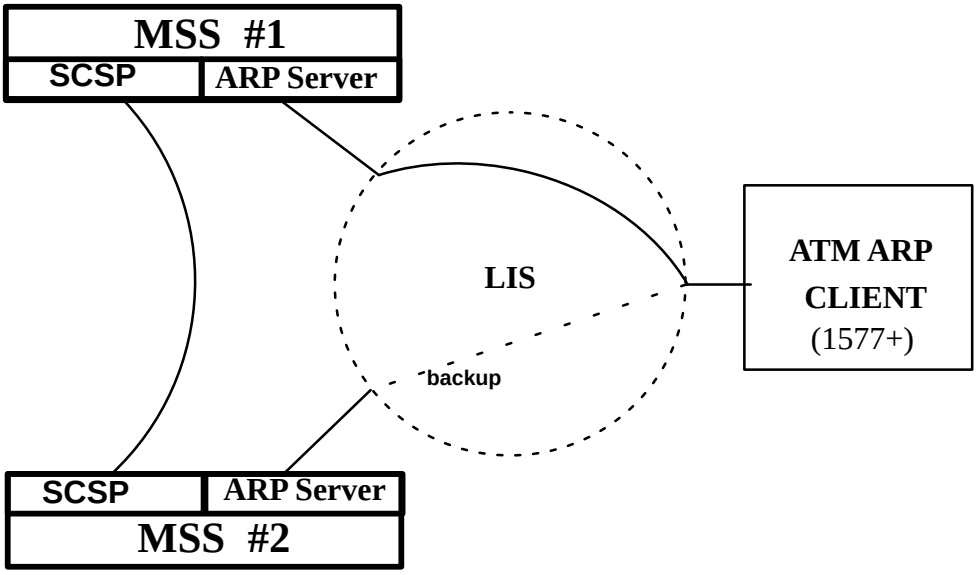

Figure 24. Simple 1577+ Client Configuration

5. 1. Switching to a backup ARP Server

When an ARP Server fails, the client may detect the failure through the switch via a 'Disconnect Call' to bring down the VCC. In this case, the next ARP Server in the client's ARP Server list (relative to the failed one) will be tried.. However, it is not necessarily true that an ARP Server failure would translate into a 'Disconnect Call'. The ARP Server could be in a crippled state but the VCC to the client could remain active. In this case, the client detects that the ARP Server has failed when its re–registration attempt with the ARP Server fails. If the re– registration fails twice, the client assumes that the ARP Server has failed and it proceeds to switch to a backup ARP Server.

5. 2. SDU/MTU negotiation

The MSS 1577+ Client can participate in limited Service Data Unit (SDU) size negotiation with other 1577+ Clients. The MSS 1577+ Client cannot adjust it's own SDU size but it allows the remote client to adjust it's SDU size. The MSS 1577+ Client performs SDU negotiation if the remote client's forward SDU size is greater than MSS Server's backward SDU size or the remote client's backward SDU size is less than the MSS Server's forward SDU size. The MSS 1577+ Client has the same SDU size for both forward and backward directions. The SDU size can be configured on a per client basis and the default is 9188 bytes. Different clients on an MSS Server's ATM interface can have different SDU sizes, but the IP MTU is defined on a per ATM interface. Therefore, the IP MTU is automatically set to the minimum of all the clients SDU sizes on an ATM interface less 8 bytes for the LLC SNAP header.

6. IP Multicast over ATM

Multicasting is the simultaneous transmission of a single data element to multiple destinations. IP multicasting is a layer 3 protocol used for transmitting IP datagrams from one IP source to many IP destinations in a local or wide–area network. It is the Internet abstraction of hardware multicasting. IP multicasting and broadcasting has a number of uses, but the most well known use is for distributing audio and video. IP multicasting and broadcasting are also used by major routing protocols like RIP, OSPF, and BGP as well as networking services such as BOOTP and DLS. In IP multicasting, the source of the traffic knows nothing about the receiving destinations. The source simply sends IP datagrams to a particular IP multicast group address. In IP version 4, addresses in the range 224.0.0.0 and 239.255.255.255 are termed *Class D* or *Multicast Group Addresses.* The set of destinations receiving the multicasted IP datagrams is known as a host group or multicast group. Membership into the multicast group is dynamic and initiated by the multicast group members. Hosts can join and leave at any time. There are no restrictions on the location or number of members in a multicast group. The IP Multicast solution assumes that the link layer provides a connectionless transport service and some form of broadcast or multicast addressing. The IETF has standardized IP Multicasting in RFC 1112 [20].

ATM networks today using UNI 3.0 and UNI 3.1 signalling provide neither the connectionless transport nor the multicast addressing capability needed for IP multicasting. The point–multipoint VC capability of ATM can be used for IP multicasting, but the key limitation is that the sender must have prior knowledge of the ATM address of each destination. In order for ATM attached devices to use the layer 3 IP multicast services, a mapping layer or emulation service needs to be provided by the ATM network. The IETF IP over ATM Work Group has proposed RFC 2022 [21] as a standard for providing IP multicast services natively over an ATM network. Many vendor proprietary implementations of IP multicast over ATM are available today but the MSS implementation of IP Multicast over ATM is based on services outlined in IETF RFC 2022. The following services are provided by the MSS Server.

6. 1. MARS Server

The Multicast Address Resolution Server (MARS) is an extended analog of the ATM ARP Server (introduced on page 15). MARS provides the necessary connection and addressing services required by the IP Multicast and Broadcast services over ATM networks. This service is similar to an ATM ARP Server, but while the ATM ARP Server provides a one–to–one mapping between an IP unicast address and an ATM address, the MARS server provides a one–to–many mapping between an IP multicast/broadcast address and one or more ATM addresses. Each MARS Server can serve a cluster of ATM endpoints (MARS Clients) within a single Logical IP Subnet (LIS). Each LIS can be served by a single MARS Server.

6. 2. MARS Client

The MARS client is an extension of the ATM ARP Client (introduced on page 15). A MARS client is an ATM end–point that can support IP multicasting natively over ATM. Each MARS client is configured with the ATM address of one primary MARS server and zero or more backup MARS

servers. The MARS client sends control messages to the MARS server when joining or leaving a particular multicast group and when sending traffic to a particular multicast group. However, the MARS server does not get involved with the multicast data path, only the creation of it. A system view of the network and the VCs used for control information can be seen in Figure 25.. A point-to-point VC (P2P VC) is established between the MARS server and each MARS client for private control messages. A point-to-multipoint (P2MP) Cluster Control VC (CCVC) is used by the MARS server to asynchronously update all MARS clients of changes in the multicast group. Each MARS Client can be served by a single MARS Server. One MARS Client can be defined per ATM interface (real or virtual).

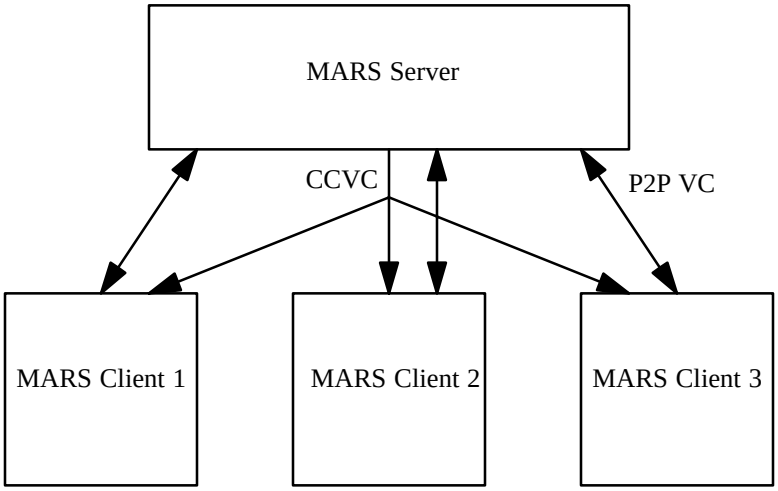

Figure 25. System view of ATM with MARS support.

A simple multicast group with one source and three destinations is shown in Figure 26.. If other endpoints also need to transmit to the multicast group, they establish a P2MP VC to each member of the multicast group, thus creating meshes of P2MP VCs (as shown in Figure 27.).

Note: A source need not be a member of the multicast group to send data to the multicast group.

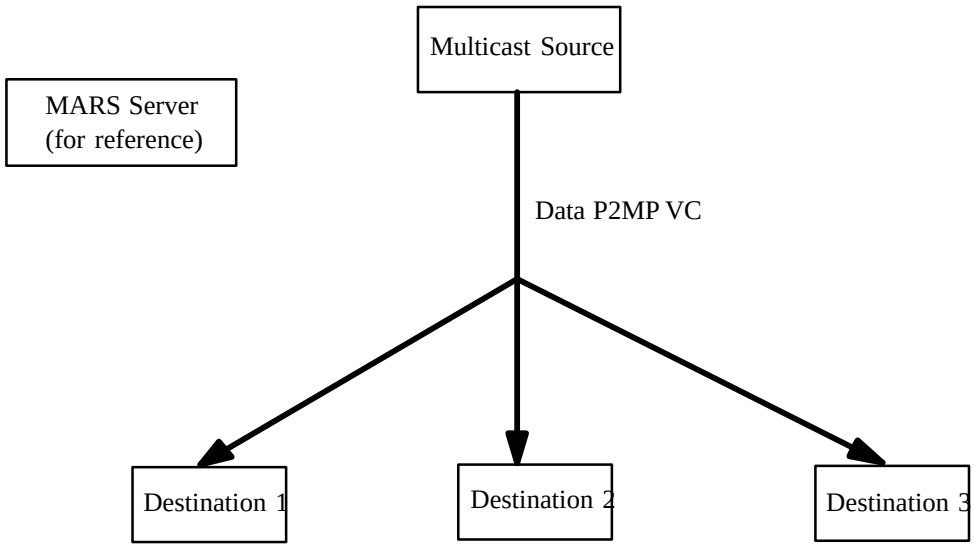

Figure 26. Multicast Data Path with meshes of VCs.

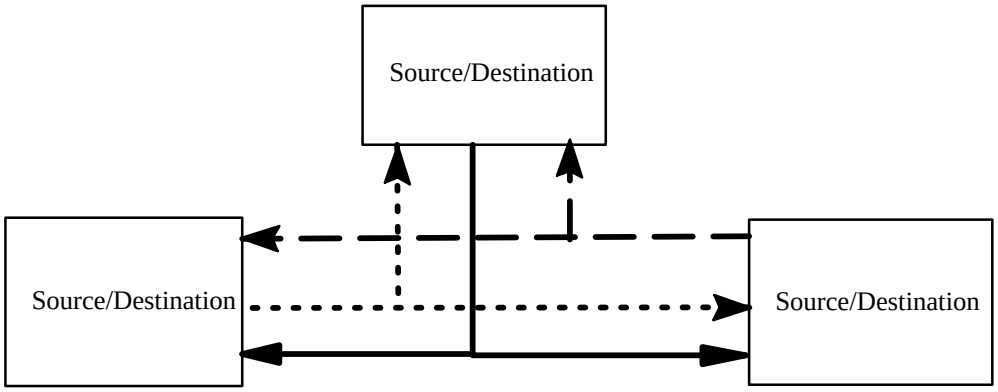

Figure 27. Meshes of P2MP VCs

6. 3. MultiCast Server (MCS)

The MultiCast Server (MCS) is like a proxy that can forward IP multicast traffic on behalf of other MARS clients. The source of the multicast traffic establishes a P2P VC to the MCS which is then

responsible for establishing a P2MP VC to send traffic to all members of the multicast group. The MCS can be beneficial when multiple sources exist in a multicast group. Without a MCS, each source must establish a P2MP VC to each destination, as shown in Figure 27.. With a MCS, each source only needs to establish one data P2P VC to the MCS, thus eliminating the need for meshes of VCs. Another advantage of MCS is that very little signalling is required when group membership changes, as compared to the mesh approach. A drawback of the MCS is that it can be a bottleneck. The MARS architecture allows for the multicast data path to be created from meshes of P2MP VCs or the use of MultiCast Servers (MCS) on a per multicast group basis. The choice of which service to use is configurable and based on application and network requirements. The use of both meshes of VCs and MCSs is supported by the same MARS Server. The only requirement is that a particular multicast group be of one type, not both. A single MCS can serve multiple group addresses, but a group address can only be served by one MCS.

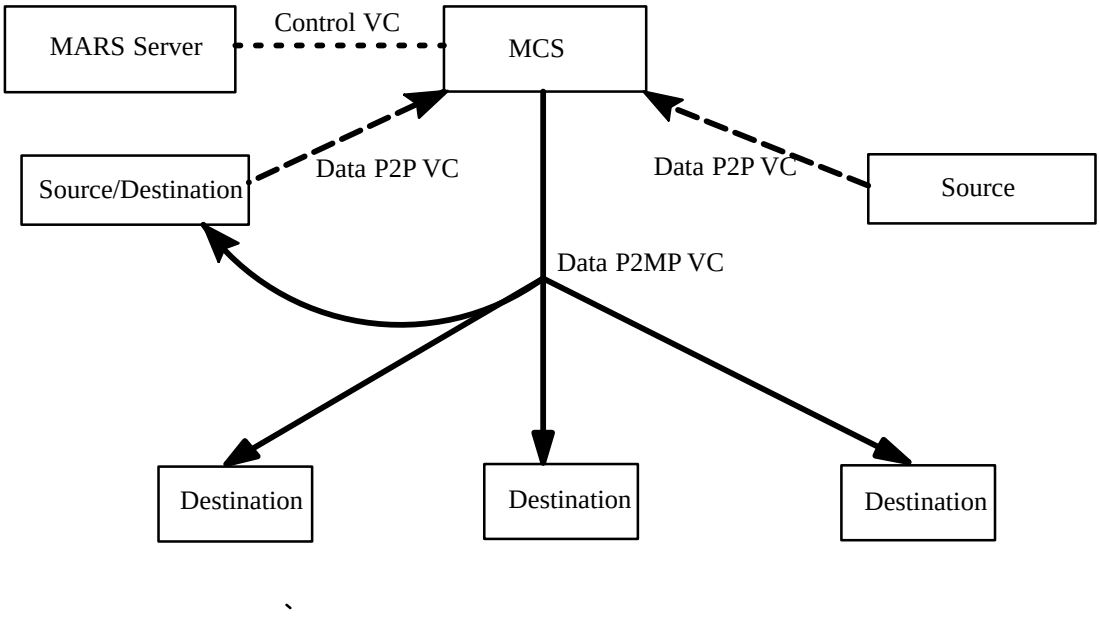

Figure 28. Multicast data path with MCS.

A system view of the data path used for a MCS can be seen in Figure 28.. The two sources of the multicast traffic establish a P2P VC to the MCS. The MCS is then responsible for establishing the P2MP VC used to send traffic to all members of the multicast group. Neither the traffic source or destination endpoints know anything about the MCS. When the traffic source establishes a connection to the MCS, the source thinks that the MCS is the only member of the multicast group. The MCS determines which destination endpoints are members of the multicast group via a private control VC between the MCS and the MARS server. The MCS sets up a data P2MP VC to each destination.

6. 4. MARS Architecture

The MARS architecture is based on each MARS Server managing a *cluster* of MARS Clients. A cluster is a group of ATM endpoints within a Logical IP Subnet (LIS) that communicate

multicast group membership information with the same MARS Server. A cluster is further restricted to be a proper subset of a LIS. Each cluster/LIS must be serviced by a separate MARS Server.

There can be multiple MARS Servers in one physical ATM network. A single MSS Server can contain one or more MARS Server as long as each MARS Server has a unique ATM address assigned to it. This can be accomplished by having multiple ATM interfaces or a different SEL field in the NSAP address of a single interface. Each MARS Server serves a separate cluster and multicast routers forward traffic between clusters. Multicast routers can use the IGMP protocol [20] to learn active multicast groups in a LIS. If the multicast router supports MARS, it gets group membership information directly from the MARS server instead of using IGMP. Multicast routers use protocols like DVMRP and MOSPF to propagate multicast topology information to other multicast routers.

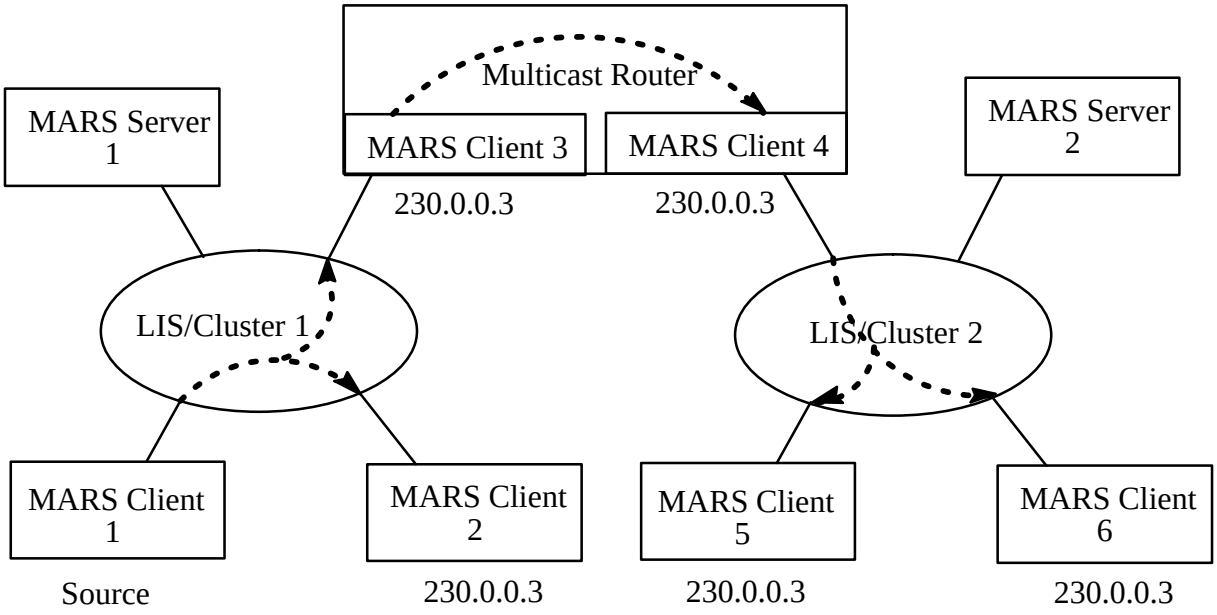

Figure 29. Inter–cluster communication using a Multicast Router

An example of inter–cluster communication using a multicast router is shown in Figure 29.. MARS clients 1,2 and 3 are in cluster 1 and served by MARS Server 1. Similarly, MARS clients 4, 5 and 6 are in cluster 2 and served by MARS Server 2. Cluster 1 and 2 are inter–connected by a multicast router which has interfaces MARS client 3 in cluster 1 and MARS client 4 in cluster 2. MARS Client 1 is the source of the multicast traffic and the destination is the IP multicast address 230.0.0.3. The source sends the multicast traffic to all destinations in cluster 1 and the multicast router forwards it to all destinations in cluster 2.

The MARS Server distributes group membership update information to cluster members over the ClusterControlVC (CCVC) shown in Figure 30.. A specific MARS Server utilizing MCSs

establishes a P2MP VC to each MCS for distributing server group membership updates. This second P2MP VC is known as the ServerControlVC (SCVC). Both VCs can be seen in Figure 30..

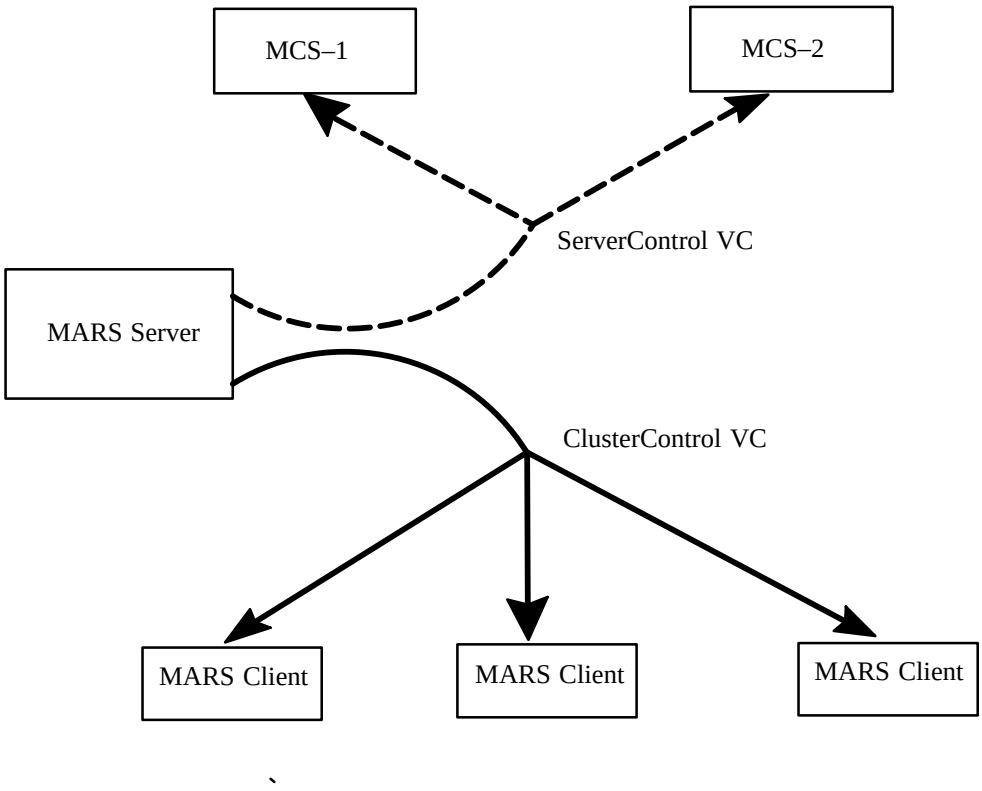

Figure 30. ClusterControlVC and ServerControlVC

An ATM endpoint which is a cluster member of a particular MARS Server will always be a leaf member of that MARS Server's ClusterControlVC. An endpoint (eg. router or multihomed host) can also be a member of multiple MARS Servers. An endpoint that is a member of multiple clusters will be added as a leaf node to the ClusterControlVC of each MARS Server. Each cluster member is assigned a unique Cluster Member ID (CMI) by the MARS Server when the cluster member registers with the MARS Server. In addition, all MCSs registered to a particular MARS Server are also leaf nodes to the MARS Server's SCVC.

6. 5. MARS Redundancy

The MARS architecture has backup and redundant features built into its messaging format and state machine variables. A cluster member can be configured with a list of backup MARS Servers or learn them dynamically via MARS_REDIRECT_MAP messages. The MARS_REDIRECT_MAP message is sent from the MARS Server over the ClusterControlVC

and ServerControlVC periodically. The message contains the default MARS Server for the cluster to use and the list of backup MARS Servers. A cluster member that determines that a problem exists with the default MARS Server will attempt to register with a backup MARS Server. The MARS_REDIRECT_MAP message can also be used to force cluster members from one MARS Server to another by changing the value of the default MARS in the message.

6. 6. Interaction between MARS Server and MultiCast Server (MCS)

The support of MCS for a particular multicast group address is initiated when the MCS registers with the MARS Server. The MCS issues a MARS_MSERVE for the particular group address it wishes to serve. The MCS will issue a MARS_UNSERV if it no longer wishes to support the group.

6. 7. MARS Configuration

Presently, MARS can only be configured using the console or web interface. MARS configuration using the GUI Configuration Program will be provided at a later date.

7. APPN Routing

Advanced Peer–to–Peer Networking (APPN) is an extension of IBM's SNA architecture which allows SNA PU Type 2.1 nodes to communicate directly without requiring the services of a SNA host computer. APPN Routing support is being provided for ethernet and token–ring LECs, as well as the native ATM interface (real or virtual) using RFC 1483 encapsulation. The MSS Server can be configured as an APPN Network Node (NN) that can provide Directory services, Routing services and Management services to APPN end–nodes and APPN Low–entry networking (LEN) end–nodes.

The NN can also provide Dependent LU Requestor (DLUR) services to legacy PU 2.0 nodes containing dependent LUs. The NN also exchanges network topology information with other NNs and can act as an intermediate node for session data from adjacent nodes. This is called Intermediate Session Routing (ISR) and is supported on ethernet and token–ring LECs. The NN also supports High Performance Routing (HPR) by using Automatic Network Routing (ANR) instead of using ISR. HPR reduces processing overhead in intermediate routing nodes and moves error recovery and flow control to the end–points of a HPR connection. As an end–point of a HPR connection, the MSS uses the Rapid Transport Protocol (RTP) to take advantage of HPR. HPR is supported on ethernet and token–ring LECs as well as the native ATM interface. APPN also allows direct communication between nodes, thus eliminating the routing by intermediate NNs, when the nodes are connected to the same shared transport facility (eg. LAN). This type of network is called an APPN Connection Network and is supported on ethernet and token–ring LECs as well as the native ATM interface.

APPN is loaded into MSS memory only if APPN is configured (see Dynamic linking and loading on page 83).

8. Banyan VINES Routing

Banyan VINES is a service–based network operating system which uses the VINES networking protocols. Banyan VINES Routing support is being provided for ethernet and token–ring LECs. The following VINES protocols are supported:

- VINES Internet Protocol (VINES IP)

  VINES IP is used to route packets through the VINES internetwork. It provides a connectionless delivery service that delivers packets without errors.

- VINES RouTing Update Protocol (VINES RTP)

  VINES RTP is used to exchange routing information with other VINES routers.

- VINES Address Resolution Protocol (VINES ARP)

  VINES ARP is used by VINES clients to dynamically get a unique VINES internet addresses from a VINES server or router.

- VINES Internet Control Protocol (VINES ICP)

  VINES ICP provides diagnostics and support functions.

9. Miscellaneous MSS Server Enhancements

This section contains several additional enhancements included in MSS Server R2.0/2.0.1.

9. 1. ATM LLC Multiplexing

ATM LLC multiplexing allows sharing of ATM addresses and channels (both SVCs and PVCs) between protocols. SCSP (described on page 62) and APPN over ATM (described on page 75) can take advantage of ATM LLC multiplexing. Other protocols (such as CIP, IPX, 1483 bridging, ARP and NHRP), which can run natively on ATM do not share ATM addresses or SVCs, however they may share PVCs.

ATM LLC multiplexing is transparent to users and requires no explicit configuration. However, it does have its own ELS monitoring system (described on page 20). ATM LLC Multiplexing gets implicitly configured when the protocols which use it are configured. The main benefits of using ATM LLC multiplexing is reduced resource consumption due to the sharing of ATM addresses and VCCs and potential performance gain because it implies a lower amount of signalling load due to fewer VCCs.

9. 2. 1483 SVC support for bridging

Previously, the MSS Server provided the capability to transmit bridged traffic natively over ATM using LLC/SNAP encapsulation as specified in RFC 1483 [8]. The support was limited to transmitting and receiving bridged frames over PVCs. In R2.0/2.0.1, 1483 Bridging support is being extended to include SVCs. With SVC support, the MSS Server uses signalling to find the destination ATM address. Now, the user can simply specify the destination ATM address associated with a 1483 bridge port instead of specifying a VPI/VCI and configuring the PVC in the intermediate ATM switches.

A 1483 bridge port remains disabled while the SVC associated with it is not active. When the call is placed successfully, the port progresses eventually to the blocked or the forwarding state (due to the spanning tree). If the SVC is disconnected for any reason (switch reset or disconnect from the remote end), the port moves to the POSTCONFIGURING state. Since all configured SVCs are retried at fixed intervals if they are not active, the port comes back up when the call is placed successfully. The SVC is configured at both ends, thus two VCCs get created for each bridge port using SVC support. The state of a port depends solely on the state of the locally placed SVC. Each port transmits frames on the locally placed VCC and receives frames on the other.

9. 3. LAN Emulation ARP Cache enhancements

The following LAN Emulation ARP cache enhancements have been made:

- Improved ARP cache performance

The LEC maintains databases that map MAC addresses to ATM addresses, and ATM addresses to VCCs. These databases are now stored in binary trees for fast searches instead of doing linear searches in arrays, as was done previously.

- Allow for the configuration of ARP cache entries

  An option is being provided to allow users to configure permanent entries (i.e. they won't age out) in the MAC to ATM address database. This can be used to configure addresses of heavily used devices and avoid the LE ARP process.

- Age out ARP cache entries per the LANE specification.

  LE ARP cache entries are now aged out if a frame has not been sent to the destination (MAC or RD) for the period specified in the aging timer (C17). The aging timer is configurable.

- Increase max ARP Cache size.

  The maximum ARP cache size for a LEC has been increased from 2048 to 65,535. The default ARP cache size for a LEC has been increased from 1024 to 5000.

9. 4. LES initiated pacing during congestion

The LAN Emulation (LANE) architecture (described on page 9) emulates traditional LAN technologies over a switched ATM network. Two components of the LANE architecture, the LAN Emulation Server (LES) and the Broadcast and Unknown Server (BUS), control the addition of LAN Emulation Clients (LECs) to the emulated LAN (ELAN). LECs are added to point–to–multipoint connections that carry control and data frames to the LECs on the ELAN. The process of adding LECs to these multipoint connections requires ATM switches in the network to process special control (Add Party) messages.

In large networks, switches may not have the processing power or memory to handle the addition of large numbers of LECs when they join an ELAN simultaneously, as is sometimes the case if the LANE Service or switched network recovers from a fault or is restarted. If this congestion occurs in the network, the Add Party messages will either be explicitly rejected or dropped, which generally forces the LECs to attempt to join the ELAN again. If the LECs do not randomly delay their attempt to rejoin (and many do not), additional network congestion occurs which can prevent, or at least greatly delay, stabilization of the network.

In order to reduce the rate of signalling messages in the ATM network, the MSS LES and BUS can randomly delay the joining of LECs to the ELAN when signalling congestion has been detected in the network. The network is deemed congested if the Add Party message transmitted by the LES or BUS to add a LEC to the Control Distribute or Multicast Forward connection is either not responded to (times out) or is rejected for one of the following reasons: User Busy, No User Responding, Temporary Failure, Too Many Pending Add Party Requests.

In congested state, the LES and BUS will randomly delay the initial and subsequent Add Party messages until congestion has not been detected for a period of time (e.g. 30 seconds). Add Party attempts that fail in the manner described above will be retried until the Control Time–out timer expires (in the case of the LES) or the maximum number of retries is exceeded (in the case of the BUS). This technique spreads out the signalling traffic when many stations attempt to join the ELAN at the same time, which allows the ATM switches to handle a much larger network of ATM attached devices.

9. 5. BCM and BBCM Support for NetBIOS NameSharing

The OS/2 LAN Server has a feature called *NameSharing*, which allows the same NetBIOS name (file server name) to be used on multiple LAN interfaces of the server. NameSharing is used to overcome the NetBIOS limitation of 254 sessions per LAN interface. Without this technique, a file server could not be accessed by more than 254 clients at one time. Another benefit of NameSharing is that it allows a LAN Server's clients to be distributed across multiple LAN interfaces of the server, thus balancing network traffic across the server's interfaces.

In previous releases of the MSS Server, if BroadCast Manager (BCM, described on page 11) or Bridging BroadCast Manager (BBCM, described on page 23) was enabled for NetBIOS, it associated each learned NetBIOS name with a single unicast MAC address. Subsequently, if BCM/BBCM transformed a NetBIOS broadcast, it always directed the packet to the associated unicast MAC address. However, this defeated the purpose of NameSharing. With MSS Server 2.0/2.0.1, BCM/BBCM can support networks with NameSharing servers. BCM/BBCM timestamps a NetBIOS name entry in it's cache when a NetBIOS Name.Query for that name is processed. If a Name.Recognized response to the Name.Query is not seen within a specific time period, the NetBIOS name is aged out of the BCM/BBCM cache. A subsequent Name.Query to the server (seen by BCM/BBCM after the timeout period), will not be transformed by BCM/BBCM and will thus reach all interfaces of the server. The server will respond to the Name.Query on another interface (if that interface has not exceeded it's session limit). Thus, clients get distributed across multiple interfaces of the server and traffic gets balanced.

For BCM, the ageout to support NetBIOS NameSharing is 1 second and cannot be changed. For BBCM, the *Duplicate Frame Filter Timeout* is used for this purpose. Duplicate Frame Filter Timeout is configurable and it defaults to 1.5 seconds. BBCM support for NameSharing servers is available with the ASRT bridge (described on page 16) as well as the SuperELAN bridge (described in section on page 37).

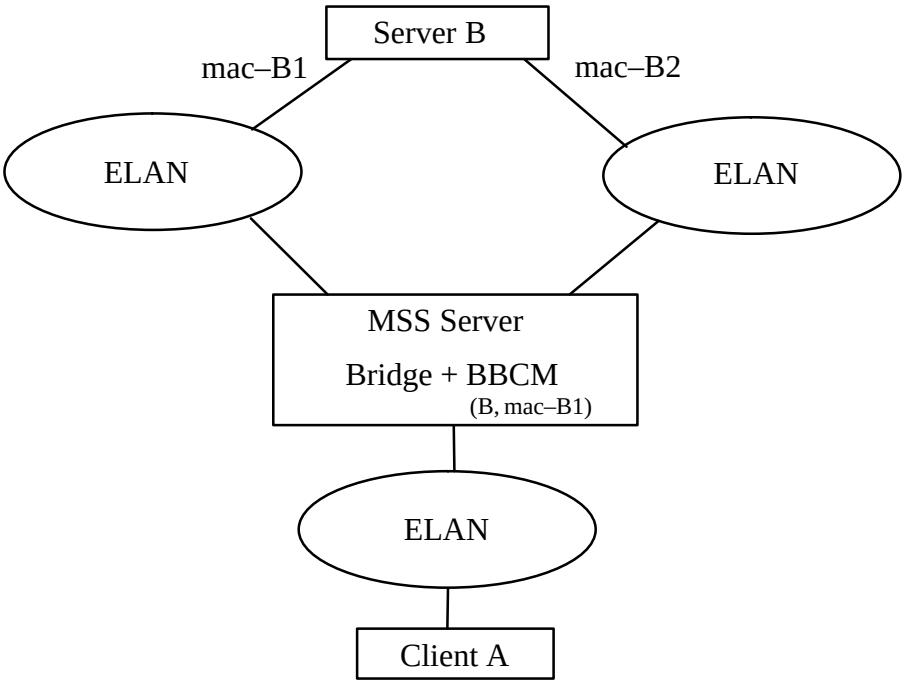

Figure 31. NetBIOS NameSharing

Figure 31. shows an example of NetBIOS NameSharing support. Assume BBCM in the MSS Server has already learned server B's NetBIOS name and associated it with one of its MAC addresses (mac–B1). When client A sends a Name.Query to server B for a call setup, BBCM transforms the Name.Query into a unicast frame addressed to mac–B1. BBCM also timestamps the entry associated with server B's NetBIOS name. If the number of NetBIOS sessions on interface mac–B1 has reached the maximum, server B won't respond with a Name.Recognized over that interface. Client A will resend the Name.Query to server B but this time BBCM detects that a previous Name.Query to B went unanswered, so BBCM will remove B from its cache. The second Name.Query to B will not be transformed by BCM/BBCM and will reach all interfaces of the server. Server B will respond to the Name.Query on interface mac–B2 (if mac–B2 interface has not exceeded it's session limit).

9. 6. IP enhancements

• RIP v2

RIP v2 support has been added and is compliant with IETF RFC 1723 . RIP v2 is an extension of the RIP protocol that allows routers to share important additional information such as subnet mask and authentication key. Thus, RIP v2 can reliably learn subnets and is inherently more secure. RIP v2 also reduces interruptions to other stations on the network by advertising routes to a well known IP multicast address instead of broadcasting them. RIP v2 is backward compatible with the existing RIP implementations.

• IP Interfaces

The number of IP interfaces supported on a single network interface is unlimited now. Previously, only 32 IP interfaces could be configured on a single network interface.

9. 7. Duplicate MAC Address Support for SR–TB Bridging

The SR–TB Translational bridge (described on page 17) is being enhanced to support duplicate MAC addresses on the SRB network. SR–TB Translational bridges are typically used to allow communication between ethernet and token–ring stations when routing between the two is not possible or desirable. Previously, the SR–TB bridge did not work in networks with duplicate MAC addresses on the SRB network. Duplicate MAC addresses on SRB networks are extensively used in SNA networks to provide redundancy and load–balancing across multiple IBM 3745 communication controllers. The SR–TB has been enhanced to support 2 instances of up to 7 MAC addresses on the SRB network. Additionally, the SR–TB can load balance traffic from ethernet stations destined to a MAC address that is duplicated on the SRB network. Thus, ethernet attached stations can now get redundancy and load–balancing when accessing mainframes connected to a SRB network.

The SR–TB bridge learns MAC addresses on the SRB network and can associate multiple Routing Information Fields (RIFs) with a MAC address. When sending data from an ethernet source to a MAC address on token–ring, the SR–TB selects a RIF using the source stations MAC address as a key. Thus ethernet stations get distributed across multiple instances of a MAC address, thereby achieving load balancing.

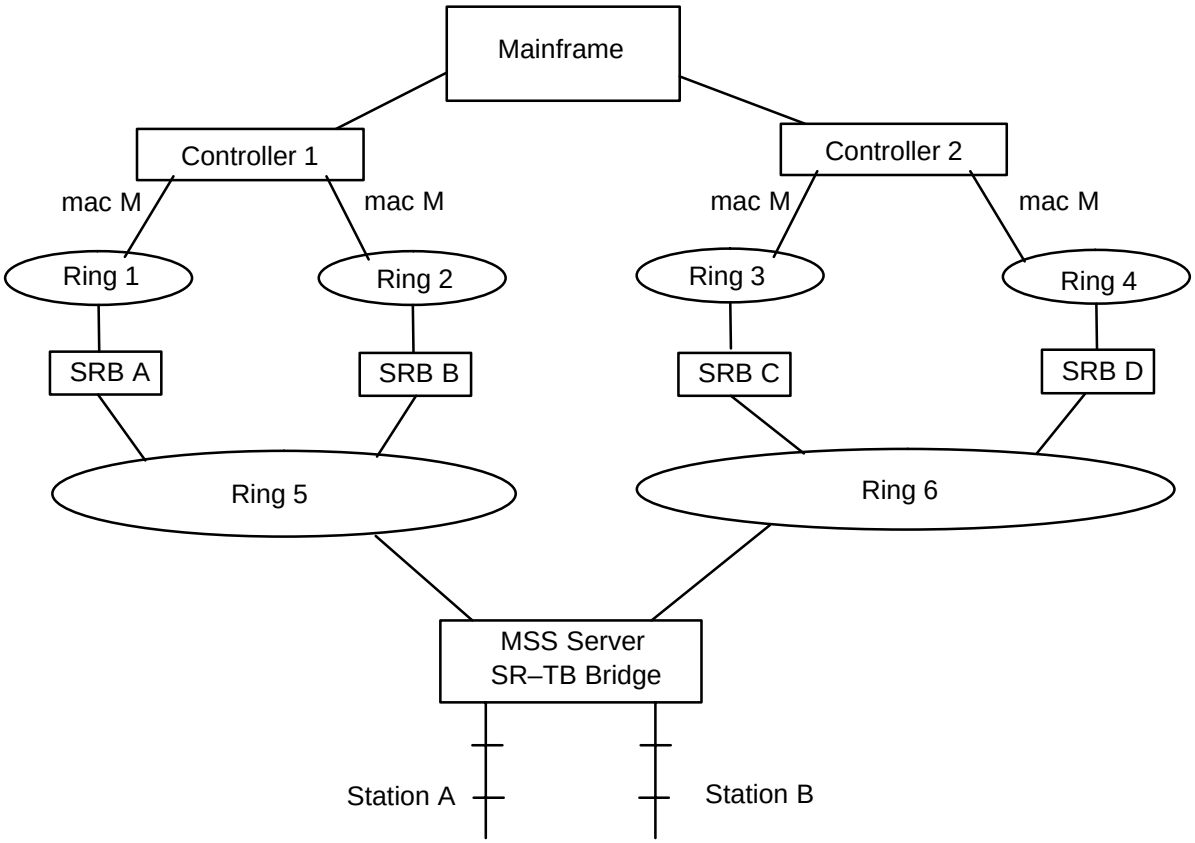

Figure 32. SR–TB bridged network with duplicate MAC addresses

Figure 32. shows an example of duplicate MAC address support in the MSS Server's SR–TB bridge. The mainframe is attached to a token–ring SRB network via 4 MAC interfaces in controllers 1 and 2. Each of the mainframe's 4 MAC interfaces is configured with the same MAC address (mac M). Stations A and B are attached to ethernet networks. The SR–TB bridge in the MSS Server bridges between the ethernet networks and the token–ring network. Since there are 4 paths between clients on ethernet and the mainframe, the SR–TB bridge will distribute the ethernet stations across the 4 paths.

9. 8. Trouble–shooting enhancements

Several functions have been added to the MSS Server to enhance its trouble–shooting capabilities.

- IP TraceRoute

  The IP TraceRoute function has been enhanced to allow the user to specify the source IP address, data size, number of probes to send, time between probes and the maximum time to life (TTL).

- IPX Ping

The *IPX Ping* function has been enhanced to allow the user to specify the source network, source node, data size and the time between pings.

- IPX TraceRoute

  IPX TraceRoute is a new function that is similar to IP TraceRoute that allows a user to find different routed paths in the IPX network. It allows the user to specify the source address, data size, number of probes to send, time between probes and the maximum number of router hops.

- IPX RecordRoute

  IPX RecordRoute is a new function that allows the user to trace a specific routed path taken by an IPX ping packet to the destination. It also traces the path taken by the response from the destination. It allows the user to specify the source address, time between record route requests and the number of requests to send.

9. 9. Dynamic Reconfiguration (DR)

Dynamic Reconfiguration allows users to change the configuration of the MSS Server without requiring a reboot to activate the changes. DR is critical for providing high availability services. In previous releases, some MSS configuration parameters could be changed dynamically, but most critical parameters could not be changed without rebooting. For example, it was not possible to dynamically add new bridging or routing interfaces. Now, the user can create spare interfaces for which protocols can be configured and activated. The spare interfaces can be activated without rebooting. Note, not all functions of the MSS Server can be dynamically reconfigured and DR still has some restrictions. DR capability will be further enhanced in subsequent releases.

9. 10. Dynamic Linking and Loading

Dynamic Linking and Loading (DLL) allows selective loading of functions from the operational code image into processor memory. If a function is not configured, it is not loaded into memory, thus making more memory available to the configured functions. In this release, only the APPN function load module can be dynamically loaded.

9. 11. Time activated re–boot

An option is provided to automatically re–boot the MSS server on a user specified date and time. The user can also select the operational code and configuration to load from non–volatile storage when the box is automatically re–booted.

10. Capacity Characteristics

The capacity characteristics described in this section are based on MSS Server 1.1 software running on a MSS Server with 32 MB of memory. MSS Server 2.0/2.0.1 capacity characteristics are currently being measured on the new MSS Server module with 64 MB of memory. This document will be updated once these measurements are completed.

The capacity of an MSS Server is dependent upon several factors including design constraints, VCC resources, memory availability, and performance considerations. In the *Quick Guide to MSS Server Release 1.0 [1]*, capacity limitations due to design and resource constraints were identified independent of performance considerations. The same approach is employed in this report. Table I below summarizes the design, VCC, and memory constraints for four different hardware configurations. Table entries in **bold** print indicate that the upper bound has increased since Release 1.0. Performance measurements are presented separately in section 11. on page 86.

Table I identifies upper bounds for the number of LECS instances, the number of LECS policy values, the number of LES/BUS instances, the number of LECs that can be served by a single LES/BUS instance, the number of internal LECs or ILECs that may be executing for bridging/routing/management purposes, the number of bridge ports, the number of internal Classical IP Clients or Client/Server pairs, and the number of external CIP Clients that can be served. All of the upper bounds assume a single MSS Server, but it is important to recognize that each capacity characteristic was analyzed independently, and all the maximums cannot be supported simultaneously. In particular, the analysis of the maximum number of clients that can be served (either LECs or CIP Clients) assumes that the majority of the VCC resources available to the MSS Server are devoted to this function.

The four hardware configurations for which the analysis was performed are:

1) MSS Server Module for IBM 8260 ATM Switch, or 8210 with single ATM adapter connected to a 8260 ATM port,

2) 8210 with two ATM adapters connected to different ATM port modules resident in a single 8260 hub,

3) 8210 with two ATM adapters connected to different 8260 hubs, and

4) 8210 connected to ATM switch network with no VCC resource constraints.

Note that some of the table entries include an alphabetic suffix, which indicates that the upper bound(s) are qualified. These qualifiers are described in the table legend.

Table I
Capacity Bounds due to Design, VCC, and Memory Constraints

Capacity Characteristics	Hardware Configuration			
	(1)	**(2)**	**(3)**	**(4)**
LECS instances	1	1	1	1
LECS Policy Values	**3000** / **6000** *(a)*	**3000** / **6000** *(a)*	**3000** / **6000** *(a)*	**3000** / **6000** *(a)*
LES/BUS instances	31–63 *(b)*	31–63 *(b)*	62–126 *(b)*	**600** *(c)*
Served LECs	450 / 1500 *(d)*	900 / 1500 *(d)*	900 / 3000 *(d)*	5000
Served LECs on one ELAN	450 / 1500 *(d)*	450 / 1500 *(d)*	450 / 1500 *(d)*	5000
Internal LECs or ILECs	63	63	126	252
Bridge Ports	254	254	254	254
CIP Clients or Client/Server pairs	**450** (*e*) / **1000** *(d)*	**900** (*e*) / **1000** *(d)*	**900** (*e*) / **1000** *(d)*	**1000**
Served CIP Clients	900 / **4000** *(d)*	1800 / **6000** *(d)*	1800 / **8000** *(d)*	**10,000**

Legend for Table I

Qualifiers on Upper Bounds

a) the first number is for the *ATM Address* and *ELAN Name* assignment policies, which require twice as much configuration record storage as the other assignment policies, and the second number is for any combination of the other assignment policies including the new *ESI–Selector* policy

b) dependent upon configuration parameters (i.e., number of PtMP VCCs used by LES and BUS instances)

c) depending upon the configuration parameters, memory constraints can reduce this upper bound

d) the first number is for hardware configurations with 8260 modules that support 992 connections and the second number is for hardware configurations with 8260 modules that support 4K connections, the second number also assumes the A–CPSW Module is equipped with 16 MBytes of memory and Version 2.1 (or higher) of the Control Point Software

e) assumes two VCCs per CIP Client or Client/Server pair

11. Performance Characteristics

The performance characteristics described in this section are based on the MSS Server 1.1 software running on a MSS Server with a 100 MHz PowerPC 603 processor and 32 MB of memory. MSS Server 2.0/2.0.1 capacity characteristics are currently being measured on the new MSS Server Module with a 166 MHz PowerPC 603 processor and 64 MB of memory. This document will be updated once these measurements are completed.

This section contains a small set of performance measurements for Release 1.1 of the MSS Server. Results are presented for routing, LE Service, and ATMARP Server throughput.

11. 1. ROUTING THROUGHPUT

The routing throughput of the MSS Server for IP is illustrated in Figure 33. below. Results are shown for two cases: (1) routing IP traffic from one LIS to another using two Classical IP Clients, and (2) routing IP traffic from one Token–Ring ELAN to another Token–Ring ELAN using two LECs. Throughput when routing IP between two Ethernet ELANs is very similar to that shown for Token–Ring ELANs. With respect to Release 1.0 performance, throughput has increased for small packet sizes; with 64–byte packets, IP routing throughput has improved by 34% for Classical IP and 18% for ELANs.

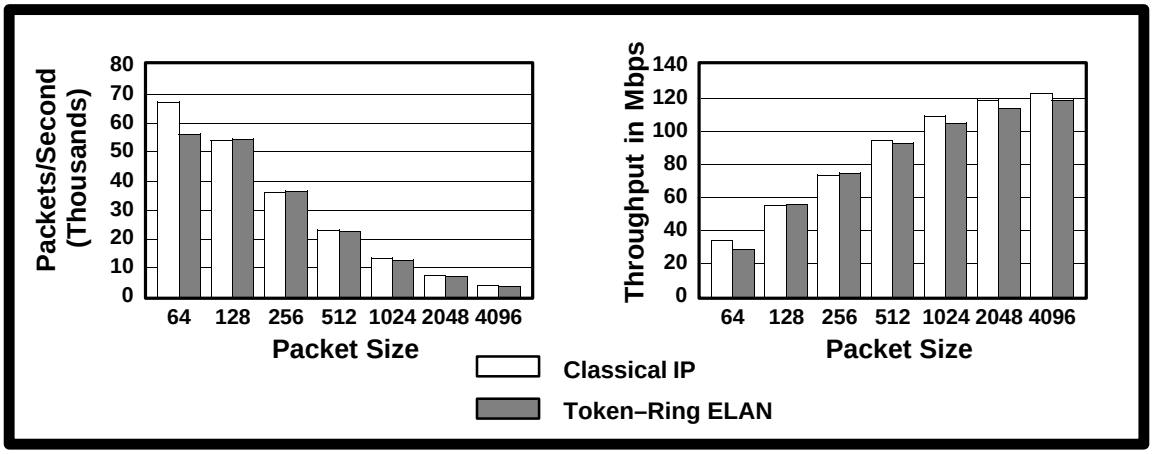

Figure 33. IP Routing Throughput

Routing throughput results for IPX and AppleTalk are shown in Figure 34. and Figure 35., respectively. In both figures, results are provided for two cases: (1) routing from one Ethernet ELAN to another Ethernet ELAN, and (2) routing between two Token–Ring ELANs. As was the case with IP, IPX routing throughput has increased for small packet sizes, especially Token–Ring; with 64–byte packets, throughput over Ethernet ELANs has improved by 7%, and throughput over Token–Ring ELANs is up by 64%.

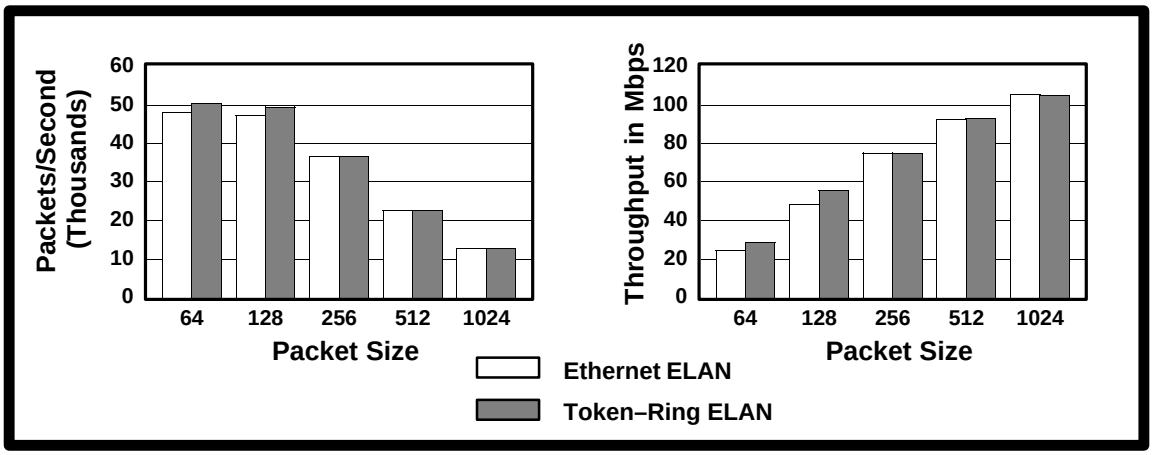

Figure 34. IPX Routing Throughput

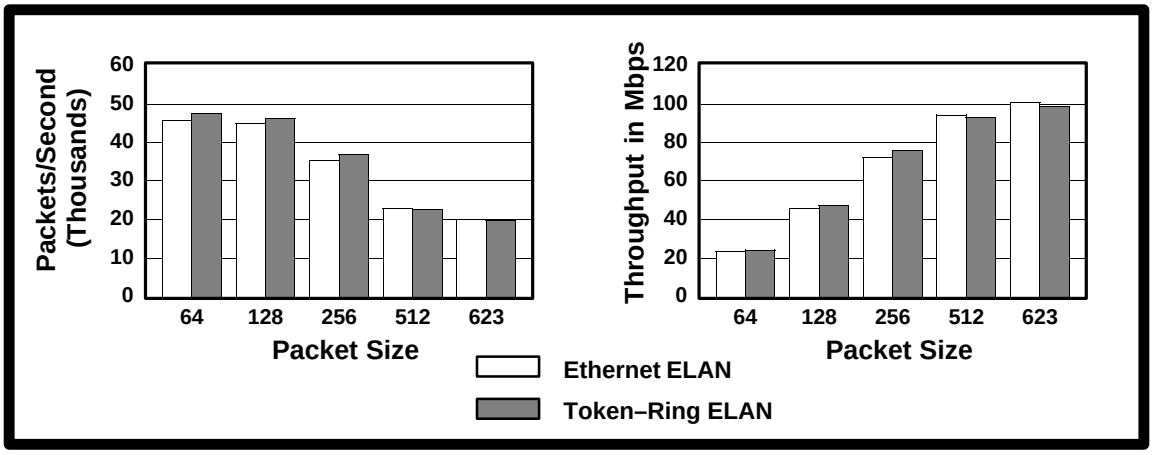

Figure 35. AppleTalk Routing Throughput

MSS Server throughput when routing over an FDDI interface is illustrated in Figure 36.. Results are shown for routing IP traffic between a Token–Ring ELAN and an FDDI LAN. Media speed is achieved on the FDDI LAN when packet sizes are 512 bytes or larger.

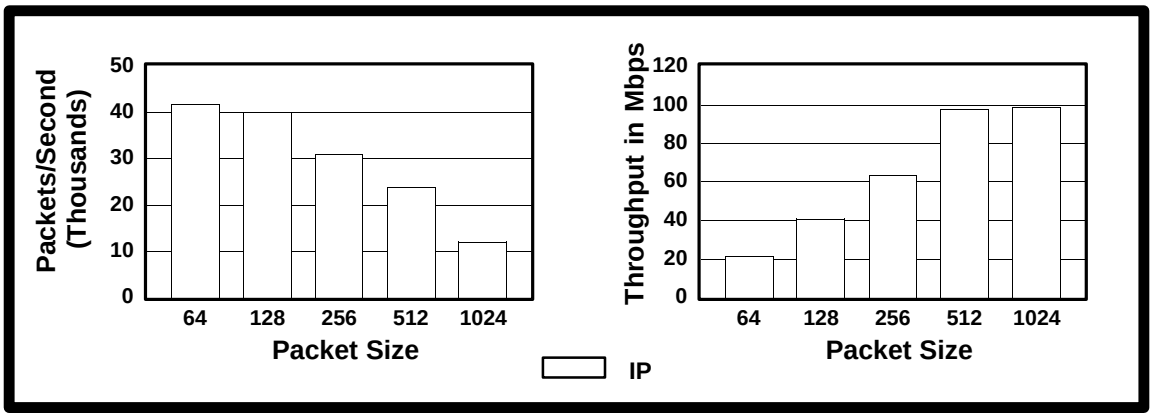

Figure 36. Routing Throughput over FDDI

11. 2. LAN Emulation Service Performance

Performance results for the LECS, LES, and BCM components of the LE Service are presented in the following subsections.

11. 2. 1. BUS Throughput

When frames are transferred into system memory, the BUS can forward 64–byte packets at rates up to ~71,000 packets per second (pps), while forwarding rates in excess of 100,000 pps are achievable in either adapter or VCC splicing mode. Media speed is achieved in VCC splicing mode when packets are 1024 bytes are larger, and in adapter mode when packets are 2048 bytes or larger. Even though, in Release 1.1, comparable forwarding throughput may be achieved with either adapter or VCC splicing mode, VCC splicing mode does offer one advantage in that forwarding does not require any system bus bandwidth or processor cycles. This enables cumulative forwarding power to scale linearly with the addition of a second ATM adapter. Thus, cumulative BUS forwarding rates in excess of 200,000 pps are possible when two ATM adapters are installed in an IBM 8210 MSS Server.

The BUS mode is configurable on a per ELAN basis.

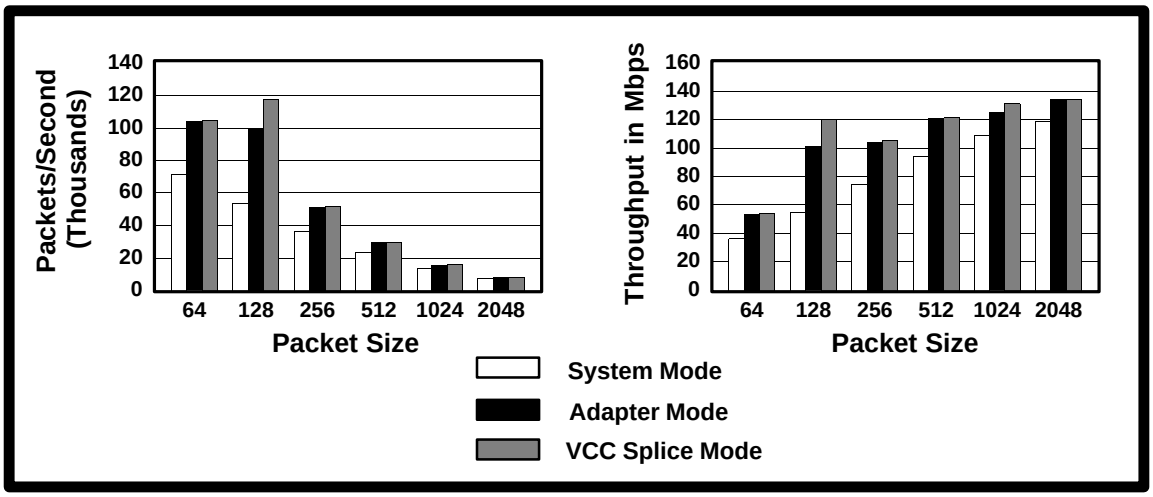

Figure 37. BUS Forwarding Throughput

11. 2. 2. LECS Throughput

Performance of the LECS has more than doubled in Release 1.1. With a simple ELAN Name assignment database, the MSS Server LECS can respond to ~60,000 LE_CONFIGURE requests per second.

11. 2. 3. LES Throughput

In Release 1.1, the MSS Server LES can answer up to 59,000 LE_ARP requests per second, which represents a 5% improvement over Release 1.0.

11. 2. 4. BCM Throughput

Figure 38. shows the throughput when BCM is actively managing every frame, which should be viewed as worst–case performance (i.e., every received frame is a broadcast frame that BCM converts to a unicast). Results are shown for IP ARPs and NetBIOS Datagrams when BCM's database contains 4096 entries. BCM throughput has increased for both protocols in Release 1.1. For the illustrated case, throughput has doubled for IP, and improved by 22% for NetBIOS.

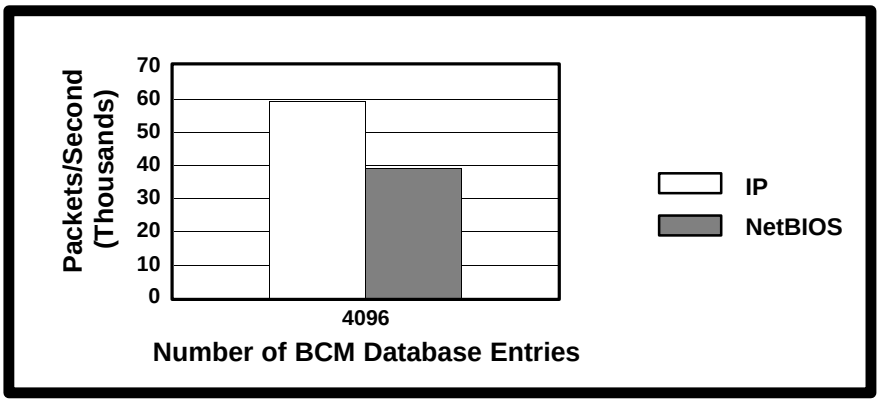

Figure 38. BCM Throughput

11. 3. Classical IP ATM ARP Server Throughput

The Release 1.1 MSS ATMARP Server can answer up to 30,000 ATMARP requests per second, which is a 15% improvement over Release 1.0.

GLOSSARY OF TERMS AND ACRONYMS

AAL ATM Adaptation Layer, the layer that adapts user data to/from the ATM network by adding/removing headers and segmenting/reassembling the data into/from cells

AAL–5 ATM Adaptation Layer 5, one of several standard AALs, AAL–5 was designed for data communications and is used by LAN Emulation and Classical IP

A–CPSW ATM Control Point and SWitch, ATM switch module for IBM 8260 hub

ARP Address Resolution Protocol, IP ARP translates network addresses into hardware addresses, LE ARP translates LAN Destinations into ATM addresses

ATMARP ARP in Classical IP

ATM Asynchronous Transfer Mode, a connection–oriented, high–speed networking technology based on cell switching

BBCM Bridging BroadCast Manager, a bridge enhancement designed to limit the effects of broadcast frames

BCM BroadCast Manager, an IBM extension to LAN Emulation designed to limit the effects of broadcast frames

BUS Broadcast and Unknown Server, a LAN Emulation Service component responsible for the delivery of multicast and unknown unicast frames

CIP Classical IP, an IETF standard for ATM–attached devices to communicate using IP

CIPC Classical IP Client, a Classical IP component that represents users of the Logical IP Subnet (LIS)

DLCI Data Link Connection Identifier, identifies a specific virtual connection on a frame relay link

DPF Dynamic Protocol Filtering, a bridge enhancement supporting PVLANs

ELAN Emulated Local Area Network, a LAN segment implemented with ATM technology

ESI End System Identifier, a 6–byte component of an ATM address

FDDI Fiber Distributed Data Interface, standard for 100 Mbps Token–Ring Local Area Networks

IBUS Intelligent BUS, a LAN Emulation optimization designed to limit the scope of unknown unicast frames sent to the BUS

ICMP Internet Control Message Protocol, a protocol for communicating control information over IP

IEEE Institute of Electrical and Electronic Engineers, an organization involved in establishing Local Area Network standards

IETF Internet Engineering Task Force, an organization that produces Internet specifications

ILMI Interim Local Management Interface, SNMP–based procedures for managing the User–Network Interface (UNI)

IP Internet Protocol, a widely–used network layer protocol specified by the IETF

IPX Internet Packet Exchange, a network layer protocol that is commonly used by personal computer systems

ISO International Standards Organization, an organization that specifies international communication standards

LANE LAN Emulation, an ATM Forum standard supporting communication between legacy LAN applications over ATM networks, the terms LANE and LE are often used interchangeably

LE LAN Emulation, an ATM Forum standard supporting communication between legacy LAN applications over ATM networks, the terms LE and LANE are often used interchangeably

LEC LAN Emulation Client, a LAN Emulation component that represents users of the Emulated LAN

LECS LAN Emulation Configuration Server, a LAN Emulation Service component that centralizes and disseminates configuration data

LES LAN Emulation Server, a LAN Emulation Service component that resolves LAN Destinations to ATM Addresses

LIS Logical IP Subnet, an IP subnet implemented with Classical IP

LLC Logical Link Control, the top sublayer of the Data Link Layer, which is layer 2 of the ISO model

LNNI LANE Network–Network Interface, the interface between distributed LE Service entities

LSI LANE Shortcut Interface, component of the MSS Server's NHRP implementation that establishes shortcut VCCs to LAN Emulation Clients

LUNI LANE User–Network Interface, the interface between LAN Emulation Clients and the LAN Emulation Service

MAC Medium Access Control, the bottom sublayer of the Data Link Layer, which is layer 2 of the ISO model

MIB Management Information Base, a network management database supporting the monitoring and control of network elements

MOSPF Multicast Open Shortest Path First, a multicast routing protocol specified by the IETF

MSS Multiprotocol Switched Services, a component of IBM's Switched Virtual Networking (SVN) framework

MTU Maximum Transmission Unit, the maximum amount of data that can be transmitted as a single unit (frame) on a communications link (does not include layer 2 headers)

NBMA Non–Broadcast Multi–Access, attributes characterizing a set of networking technologies that includes ATM

NHC Next Hop Client, an NHRP entity that initiates establishment of shortcut routes

NHRP Next Hop Resolution Protocol, an IETF protocol for bypassing routers on NBMA networks

NHS Next Hop Server, an NHRP entity that provides the protocol address–to–NBMA address mappings necessary for establishment of shortcut routes

OSPF Open Shortest Path First, a routing protocol specified by the IETF

PCMCIA Personal Computer Memory Card International Association, an organization involved in establishing hardware standards that are often associated with miniaturized peripherals

PCR Peak Cell Rate, maximum transmission rate on an ATM Virtual Connection

PPS Packets Per Second, units commonly used in throughput performance measurements

PtP Point–to–Point, a VCC between two parties

PtMP Point–to–MultiPoint, a unidirectional VCC that allows transmissions from the source party to be received by multiple destination parties

PU Physical Unit, represents a physical device and its associated resources in SNA networks

PVC Permanent Virtual Circuit, a configured VCC

PVLAN Protocol VLAN, a specific type of VLAN where membership is based on use of a common communications protocol or participation in a particular protocol subnetwork

QoS Quality of Service, attribute of networks capable of delivering preferred/guaranteed service to specified traffic flows

RIF Routing Information Field, sequence of ring and bridge numbers specifying the path between two stations in a Source Route Bridge network

RIP Routing Information Protocol, a vector–distance routing protocol, versions of RIP are used with IP and IPX

RSC Route Switching Client, an extension of the NHRP Client (NHC) to allow legacy LAN attached devices to request short–cuts from a Route Switching Server (RSS)

RSS Route Switching Server, an extension of the NHRP Server (NHS) to extend short–cuts to legacy LAN attached devices.

SAP Service Advertising Protocol, an IPX protocol used to advertise the location of available services

SCB Short–Cut Bridge, a bridging technique used by the MSS Server to allow inter–ELAN data–direct VCC to be established between LECs on different transparently bridged ELANs

SEB SuperELAN Bridge, an enhanced implementation of the SCB bridging technique used by the MSS Server that can support transparent (TB) as well as source route (SR) bridged networks.

SCR Sustained Cell Rate, maximum sustained transmission rate on an ATM Virtual Connection

SDU Service Data Unit, data as it appears at the interface between a layer and the layer immediately above

SLIP Serial Line IP, an IETF standard for running IP over serial communication links

SNA Systems Network Architecture, a networking architecture developed by IBM with a large base of installed systems

SNAP SubNetwork Attachment Point, an LLC header extension that identifies the protocol type of a frame

SNMP Simple Network Management Protocol, an IETF standard protocol that uses MIBs to control and monitor network elements

SR Source Routing, a bridging protocol that is primarily used with Token–Ring Local Area Networks

SRB Source Route Bridge, a bridging device that supports the Source Route (SR) protocol

SRT Source–Route Transparent, a bridging protocol for Local Area Networks specified in the IEEE 802.1d standard, SRT bridges support both source–route and transparent bridging on the same port

SR–TB Source Route–Transparent Bridge, a bridge that connects SR and TB ports

SVC Switched Virtual Connection, a VCC that is dynamically established via signalling protocols

SVN Switched Virtual Networking, the name of IBM's framework for building and managing switch–based networks

TB Transparent Bridging, a bridging protocol for Local Area Networks specified in the IEEE 802.1d standard

TLV Type/Length/Value, a generalized information element that may be present in LAN Emulation and NHRP packets

UNI User–Network Interface, the interface between user equipment and an ATM switch network

VCC Virtual Channel Connection, a connection between parties communicating via an ATM network

VCI Virtual Channel Identifier, the VPI/VCI pair uniquely identifies a specific ATM connection on a given link

VINES VIrtual NEtworking System, Banyan's networking operating system and associated protocols

VLAN Virtual Local Area Network, a logical grouping of hosts forming a broadcast domain that is independent of the physical topology

VPI Virtual Path Identifier, the VPI/VCI pair uniquely identifies a specific ATM connection on a given link