A High Speed IP Packet Forwarding Architecture over Internet using ATM Technology

May 1997

Hiroshi ESAKI
R&D Center, TOSHIBA Corporation

Abstract

ATM (Asynchronous Transfer Mode) is realized as one of platforms to provide high speed datalink layer service with certain QOS (Quality of Service). The QOS datalink layer service provided by ATM would be realized as an end-to-end, rather than a link-by-link service. ATM network provides both a resource reservation oriented service and a non resource reservation oriented service over the connection oriented VCCs, which could have a certain QOS. Sometimes, it is said that ATM network will provide a router-less large cloud datalink platform, which need not have any routers within the large cloud network. However, we will need routers, even when large ATM cloud datalink platform will be provided. Therefore, physical or logical datalink network segments (i.e., IP subnets) will be interconnected through routers (i.e., through the network layer entities), even when the ATM becomes a major datalink platform.

This paper proposes an architecture to solve two major issues, that the Internet and intranet using ATM technology has to provide. One is a high throughput packet forwarding through the routers over the ATM platform, and the other is a large sale error-free multicast services over the ATM platform.

A framework of IP packet delivery architecture with high throughput and small latency using ATM technology in large scaled internet is proposed, while keeping the current subnet model. The services provided by this new architecture includes a reliable multicast service, which is a new service for the current internet. The router proposed in this paper (i.e., CSR, Cell Switch Router) has the mapping functionality between flow-identifier (e.g., in IPv6 header) and VPI/VCI value to forward IP packets cell-by-cell, rather than the conventional packet-by-packet forwarding. By this cut-thru IP packet forwarding, both resource reservation oriented IP packet flows (e.g., IP packet flow provided by RSVP) and non resource reservation oriented IP packet flows (i.e., best effort service) experience less packet delivery latency and obtain higher aggregated throughput, compared to the conventional hop-by-hop packet forwarding does.

In order to perform the cut-thru IP packet forwarding using cell relaying capability in the router, routers exchange the information how the IP packet flows are aggregated into ATM-VCC. This information exchanging is hop-by-hop base, and the cut-thru decision is a matter of every router's local decision. When all routers along the path, that an IP packet flow takes, perform cell-relaying cut-thru, the soft-state and seamless cell-relaying channel is established to get a high throughput IP packet delivery. With keeping the current subnet model, even in the ATM networks, we can obtain soft-state oriented and scaleable QOS-ed high speed communication platform.

An error-free multicast service will be needed as well as unicast M-Bone type best effort multicast services are. In order to provide a large scale error-free multicast service, a soft-state management policy is applied to and a cell-level forward error correction (FEC) is proposed. Routers maintain down-stream nodes membership and protocol state with a soft-state policy. Also, in order to avoid the implosion of control messages from leave nodes to root node, routers merges control messages transferred from down-stream nodes. We need a cell-level FEC, when we provide large scale error-free multicast service over the ATM platform. In the ATM network, the packet error probability increases due to cell assembling

and due to switch-based packet forwarding. Cell-level FEC can provide sufficiently small packet error probability required for a large scale error-free multicast service over the ATM platform.

Contents

List of Figures

List of Tables

1 Introduction

ATM (Asynchronous Transfer Mode) is realized as one of platforms to provide high speed datalink layer service with certain QOS (Quality of Service), with a unique user network interface [I.150][AFUNI]. Since the user information is handled by the small and fixed sized packet (i.e., called as cell) in the ATM networks, it is said that ATM networks will provide a highly flexible and cost-effective network for handling a wide variety of communications and the applications will be obtain a flexible and cost-effective communication pipe. The cells are transferred through the connection oriented channel, which is called a virtual channel connection (VCC), in order to reduce the protocol overhead. As a result, it is said that an ATM network could provide a high speed data communication channel for end applications.

BISDN (Broadband Integrated Service Digital Network) is the high speed communication platform for public network and for CPN (Customer Premises Network). The standardization and the development of the BISDN has been progressed by ITU-T (International Telecommunication Union Telecommunication Standardization Section) and by the various organizations (e.g., ATM Forum).

As well as for the BISDN, ATM can be the good candidate for a high speed LAN (Local Area Network). With LAN, communication medium is becoming faster and faster; e.g., FDDI or Fiber Channel can provide more than 100Mbps communication channel to the users as the front end communication medium. Computers are getting much faster and powerful data processing capability. And, they are going to process their data through distributed way - distributed data processing (e.g., client-server processing). The distributed data processing requires a lot of data communications among the computers with a small latency. As a result, the distributed computing system requires high speed data communication platform. Also, since the amount of data exchanged among the computers increases constantly as well as the required interface speed, the scaleable communication platform is required. ATM is assumed as the scaleable communication platform to be able to provide such a high speed communication channel, which is said more than hundreds of Mbps to the applications. Also, it is assumed that ATM can deliver data to the destination point (e.g., I/O port of host) with small latency.

ATM can provide variety of service quality (i.e., QOS). Sometimes. the QOS datalink layer service provided by ATM would be realized as an end-to-end, rather than a link-by-link service. However, from the view point of heterogeneous Internet environment, ATM should be realized as a datalink layer service, rather than a network layer service. ATM network provides both a resource reservation oriented service (e.g., RSVP's resource reserved path [RSVP]) and a non resource reservation oriented service (e.g., current IP packet forwarding using connectionless service functions [I.364]) over the connection oriented VCCs, which could have a certain QOS.

Sometimes, it is said that ATM network will provide a router-less large cloud datalink platform, which need not have any routers within the large cloud network. However, we will need routers, even large ATM cloud datalink platform will be provided, e.g., [RFC1937]. For example, as discussed in section 3.1, the IP subnet must not be so large size so as to provide a large scale soft-state multicast service [RSVP]. Therefore, physical or logical datalink network segments (i.e., IP subnets) will be interconnected through routers (i.e., through the network layer entities), even when the ATM becomes a major datalink platform.

Since the recent applications often require reserved bandwidth and QOS rather than the conventional best effort service, the resource reservation oriented network/transport layer protocol, e.g., ST-II and RSVP has been developed in IETF [ST-II][RSVP]. Both ST-II and RSVP establish the flow state in the router(s) along the path where the IP packet flow passes through. ST-II establishes a static (or could say hard-state) state, i.e., the state in the router is established at a session establishment phase and is static during a session. On the contrary, RSVP establishes a dynamic (or could say soft-state) state, i.e., the state in the router is dynamically refreshed during a session. Since the ATM is just one of datalink technologies, we need the network/transport layer level resource reservation protocol to provide a global end-to-end QOS-ed IP packet delivery service. Beyond the QOS-ed IP packet delivery service, the error-free multicast packet delivery service will become the important service provided by the Internet and by intranet, as the next generation service.

This paper proposes an architecture to solve the two major issues that the Internet/intranet using ATM technology has to provide. One is a high throughput router architecture for the ATM platform while providing QOS-ed IP packet delivery services, and the other is a architecture to provide a large scale error-free multicast service over the ATM platform. In other words, this paper focuses on how to transport large amount of resource reservation oriented and non resource reservation oriented IP packet flows over the ATM networks, including the large scale error-free multicast packets delivery.

The hop-by-hop packet based routing at the router by the software processing using some packet scheduling policy with RSVP or ST-II would be one solution to provide QOS-ed IP packet delivery. However, there would be the limitation of processing capability of processors in the routers for the future high speed and large amount of packet flows. In order to scale associated with the required bandwidth by the applications and with the total amount of packet flows handled by the routers, some efficient IP packet forwarding technique (e.g., cut-thru routing) is necessary. The architecture of the router proposed in this paper has a cut-thru IP packet forwarding capability by cell-relaying function, and therefore, it is scaleable for the required bandwidth and for the amount of packet flows.

In order to provide a large scale error-free multicast service, a soft-state management policy is applied to and a cell-level forward error correction (FEC) is proposed. Routers maintain down-stream nodes membership and protocol state with a soft-state policy. Also, in order to avoid the implosion of control messages from leave nodes to root node, routers merges control messages transferred from down-stream nodes. We need a cell-level FEC, when we provide large scale error-free multicast service over the ATM platform. In the ATM network, the packet error probability increases due to cell assembling and due to switch-based packet forwarding. Cell-level FEC can provide sufficiently small packet error probability required for a large scale error-free multicast service over the ATM platform.

The following is the structure of the rest of this paper. Section 2 introduces the overview of ATM technology and internet protocol suite and clarify the technological difference between them. Section 3 discusses issues of the conventional IP over ATM architecture models. Sections 4 and 5 propose and evaluate the appropriate architecture for IP over ATM using CSR (Cell Switch Router). Sections 6 and 7 propose and evaluate an error-free multicast service architecture over large scale ATM networks. Section 8 gives a brief conclusion.

2 ATM Technology and Internet Protocol Suite

2.1 ATM Technology

2.1.1 ATM : Asynchronous Transfer Mode [I.150]

ATM network transfers the user information through a virtual channel connection (VCC). Virtual channel connection is similar to the circuit connection that is used in telephone networks. The network maintains the state of connection, while the VCC is in use. The node (e.g., end-host) requests a VCC establishment to the VCC control and management entity, i.e., signaling. When the VCC establishment request is admitted, the network establishes the states in the network to transfer the user information. The user information is transferred by the fixed size packet (celled cell). Since the network establishes the state in the network, the required address information in the cell header (i.e., VPI/VCI) can be compressed. The cells are routed in the ATM network based on the VPI/VCI.

In the rest of this subsection, four data transfer models are briefly presented.

[1] Shared Media Datalink

Shared media based platform, e.g., Ethernet, is widely used datalink technology in computer networks. In a shared media platform, the broadcast based physical media (e.g., bus, cable or fiber) is used, and the bandwidth is shared among all attached nodes. Some platforms can allocate the network resource (e.g., bandwidth or time slot) exclusively to the certain data flow, however, the widely accepted platforms (e.g., Ethernet or FDDI) never allocate the network resource exclusively to the certain data flow. In the shared media based platform, the datagram exclusively can occupy the physical media shared among all attached nodes in order to transfer the datagram to the other node attached to the same physical media. Whether the datagram can use the physical media is determined by the MAC (Media Access Control) protocol.

The shared network resource is used on-demand, whenever the datagram to be transferred to the other node is generated. This means that the shared media platform does not maintain a status for every data flow (i.e., connection), and that the shared media platform is stateless platform.

The followings are the advantages and disadvantages of shared media platform.

- Advantage
  - High speed datagram transmission interface
    The clock speed to be used for datagram transmission is determined by the operating clock speed of platform (e.g., 100Mbps in FDDI). If the network resource is available, the end-node can send a datagram with the clock speed of platform. As a result, when the offered load to the shared media platform is not high, the available interface speed for the end-node is quite large.
  - Network resource sharing
    The network resource (e.g., bandwidth or buffer space) is occupied, whenever

the packet arrives at the network node (i.e., end-host and router). By the network resource sharing among the connections, the utilization efficiency of network resource will be better than in an STM network presented below.

- Disadvantage
  - Un-specified QOS
    The service quality of datagram delivery provided to the upper layer entity (e.g., IP) can not be specified, in general. Some datagram may experience substantial buffering delay, due to the HOL (Head of Line) blocking. The datagram transmission never start, unless the network resource (e.g., access token in FDDI) is obtained using MAC protocol.
  - Scaling
    The total throughput of shared media platform is determined by the operational clock speed of media (e.g., 100Mbps in FDDI). This is because a *single* network resource (i.e., bandwidth) is completely shared among all attached nodes. In order to get a larger total throughput (i.e., scaling up), either the clock speed is increased or the datalink platform is segmented into multiple platforms. Segmented platforms will be interconnected by bridge or router.

[2] STM (Synchronous Transfer Mode)

STM is equivalent to TDM, Time Division Multiplexing, and can be said STM has come from the conventional cross-bar switching that has been used for telephone networks. The transmission link has a "frame", that has certain number of "time slots". Usually, the length of frame is 125 μsec for all clock speed. Therefore, higher clocked system has larger number of time slots in a frame.

Frame was identified by the frame header field. Switching nodes in the network establish frame synchronization, as well as clock synchronization, to identify the position of time slot in the frame. During a connection establishment procedure, one (or more than two) time slot(s) is(are) exclusively allocated to the connection that is going to be established. Since the allocation of time slot(s) is exclusive, the other connection can not use the time slot(s). This means that the allocated time slot(s) is(are) always reserved for the corresponding connection. When we think that time slot is the network resource, we can say that the network resource is exclusively allocated to each connection. And, since the time slot is always reserved for the corresponding connection, every time slot need not have any explicit information to identify the connection, i.e., which time slot that the connection is allocated is *implicitly* identified by the position of the time slot in the frame. Every STM switching node maintains the mapping information between the connection and the position of corresponding time slot(s).

The switch nodes manage and co-ordinate time slot allocation so that the user signal is relayed end-to-end (i.e., from source end-node to destination end-node). This time slot state information associated with all connections is updated when a connection is newly established or torn down. Except due to some failure status, the time slot state is never changed/re-arranged during a connection. Therefore, the route that the user information signal is transferred is always the same during a connection. Here, the route for the *connection A* that establish at the different time frame may have the different route from the router that the previous connection (*connection B*) has. But, the route of the *connection B* is not changed during the life of *connection B*.

The followings are the advantages and disadvantages of STM.

- Advantage
  - Deterministic Service Quality
    Since the time slot(s) is(are) exclusively allocated to each connection, the service quality (QOS) during a connection are basically associated with the bit error ratio (BER) and with the signal delivery delay. The BER quality and signal delivery delay may be depend on the selected route for the connection. It could say that the BER during a connection for all connection are the same in STM system and that the BER quality is not changed during a connection. Also, it can say that the signal delivery delay is always constant during a connection.
  - Secured Operation
    Since the time slot allocation is exclusive, the user signal flows do not interact to each other. The available network resource for each connection is always constant and is never shared with any other user's signal flow. This means that the user can never use the network resource more than the allocated resource, and some user signal flow (e.g., mis-behaved user) never degrade the other connection's service quality.

    Here, the above discussion is associated with the service quality during a connection. The connection blocking quality will be degraded due to mis-behaved end-node(s).
- Disadvantage
  - Low Resource Utilization
    The utilization efficiency of network resource will be lower than that of ATM and packet network, due to the following two reasons.

    One is due to the existence of idle time slots, that are not allocate to the user signal transmission. The clock speed used in the ATM networks has a hierarchy and granularity, which will be based on SDH. In SDH system, it is 64Kbps (B, basic-rate), 1.5Mbps (T1, primary-rate), 6.0Mbps (super digital), 45Mbps (T3), 155Mbps (OC-3), 622Mbps (OC-12), and 2.4Gbps (OC-48). Due to the clock granularity in the STM system, there are idle slots that are not allocated to the user signal transmission.

    The other is due to the exclusive resource allocation. In general, the application in the end-host does not always transmit the user information that is *actually* used by the peer application. However, in STM system, the network resource (i.e., time slot) is exclusively reserved for each connection, even when the end-host does not transmit the actual user information.

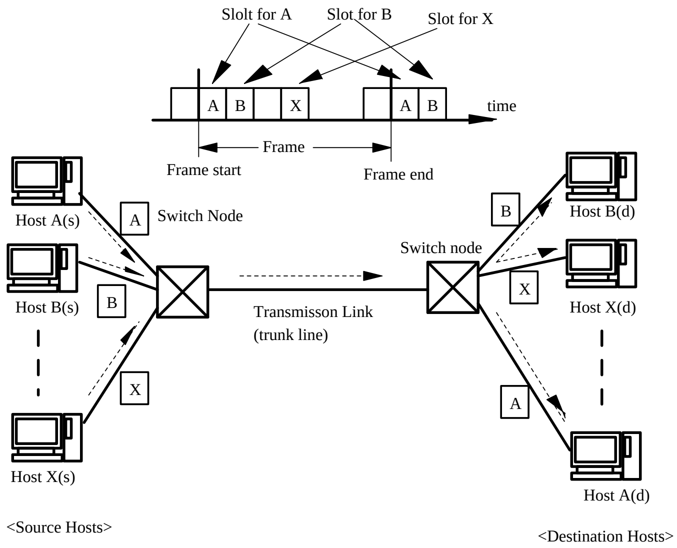


Figure 2-1. STM Architecture Abstraction

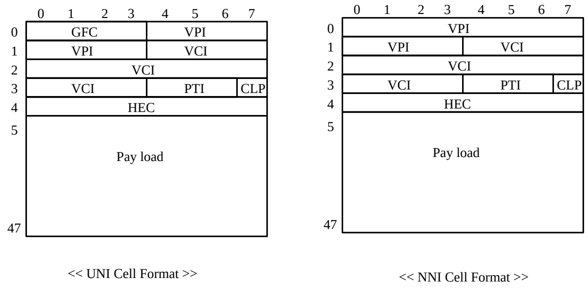
– Granularity of Service Speed
  The clock speed that is available for user network interface (UNI) also has a hierarchy and granularity, as well as the clock speed used among the switch nodes (NNI). For example, when the user application needs 10Mbps interface speed, the end-host has to use T3 (45Mbps) interface. As a result, the network resource corresponding to 35Mbps bandwidth will not be used but reserved for the user application that uses only 10Mbps bandwidth in 45Mbps bandwidth.

[3] ATM (Asynchronous Transfer Mode)

In ATM networks, the user information is fragmented into 53 byte cells, and the destination end-host reassembles the received cells (i.e., fragmented user information) to obtain a user information shipped by the source end-host. The cells are transferred through a virtual channel connection (VCC). 53 Byte cell has a 5 byte header field including VPI/VCI field. Figure 2-2 shows the cell format of UNI and NNI, and figure 2-3 shows an abstraction of cell multiplexing at ATM switching node. The VPI/VCI field is changed at every ATM switching node, and the cell is forwarded based on VPI/VCI information.

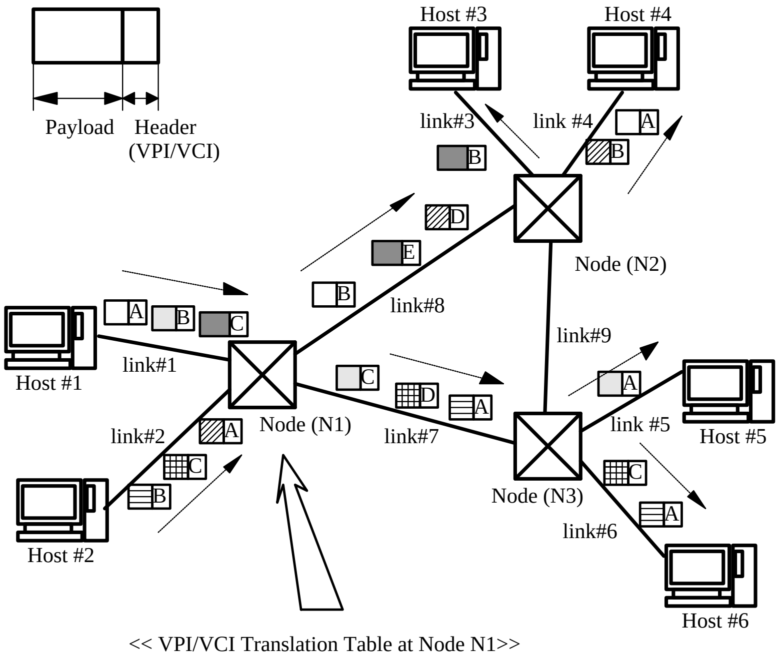
The network maintains the state of VCC during a connection, as well as an STM network does. According to the VCC establishment request, the network establishes the states in the network to transfer the user information. Actually, the ATM switch nodes establish and maintain the VPI/VCI mapping table to relay the cells from the source end-host to the destination end-host. The VPI/VCI mapping table is the VPI/VCI value of egress cell corresponding to the VPI/VCI value of ingress cell.

Since the network establishes the state according to every VCC, the VPI/VCI value corresponding to the certain connection is only unique in the physical interface (ingress and



GFC (4 bits) : General Flow Contol
VPI (8/12 bits) : Virtual Path Identifier
VCI (16 bits) : Virtual Connection Inentidier
PTI (3 bits) : Pay Load-type Identifier
CLP (1 bit) : Cell Loss Priority
HEC (8 bits) : Header Error Control

Figure 2-2. UNI and NNI Cell Format



<< VPI/VCI Translation Table at Node N1>>

Input		Output		Source Host	Destination Host
Link	VPI/VCI	Link	VPI/VCI		
#1	A	#8	B	Host #1	Host #4
	B	#7	C	Host #1	Host #5
	C	#8	E	Host #1	Host #3
#2	A	#7	D	Host #2	Host #4
	B	#8	A	Host #2	Host #6
	C	#8	D	Host #2	Host #6

Figure 2-3. ATM Cell Multiplexing and Relaying

egress link), and the VPI/VCI value can be re-used in the different interface links. This means that VPI/VCI value is not global unique value, and that the required address information length in the cell header (i.e., VPI/VCI) can be less than the packet network requires to the address information in packet header. The cell relaying using the link local unique VPI/VCI value in ATM network could be equivalent to the TCP/IP header compression technique applied in PPP (Point to Point Protocol) [RFC1144][RFC1331].

The ATM network maintains the state information connection-by-connection. This state information is not changed during a connection, without an explicit state renegotiation by the signaling capability (Q.2931). Therefore, in general, the route that the cells travel is always the same during a connection.

The followings are the advantages and disadvantages of ATM networks.

- Advantage
  - Allocated bandwidth hierarchy free
    The clock speed delivered by the network node has a digital hierarchy and granularity, as well as STM network. For example, the referenced clock delivered by the network to the end-host will be OC3 (155Mbps). However, the actually available bandwidth for each connection (or cell flow) can be flexible. The actually available bandwidth for each connection has some granularity, however it is much fine compared to STM network has. For example, the ATM network can provide 10Mbps bandwidth, that can not be provided by STM networks.

    By this finer granularity of available bandwidth, the utilization efficiency of network resource (i.e., bandwidth resource) will be better than in an STM network.
  - Network resource sharing
    ATM network is basically packet oriented platform. The network resource (e.g., bandwidth or buffer space) is used, whenever the cell arrives at the ATM switch node. There will be large diversity associated with resource management (e.g., buffer and bandwidth management) policy. The resource management policy may depend on switch vendor, on network provider, or on service class. For example, CBR (Constant Bit Rate) connection will be exclusively allocated dedicated bandwidth and buffer space. However, some connections, e.g., ABR (Available Bit Rate) connections, will share the common network resource, taking into account of statistical multiplexing effect. By the network resource sharing among the connections, the utilization efficiency of network resource will be better than in an STM network.
  - Date delivery with small latency
    End-to-end data delivery latency is larger than in STM networks, due to the cell buffering at ATM switch nodes. The delay due to the cell buffering at ATM switch node will be depend on the service class, and shall be few μsec (i.e., few cells buffering delay) to few $msec$ (i.e., few thousands cells buffering delay). However, when we compare to the end-to-end data delivery latency in the packet network, it will be much smaller. One is by means of a pipe line data transmission from

the source end-host, and the other is by means of a pipe line data transmission at the ATM switching nodes.

In general, a packet size is larger than cell size and it sometimes corresponds to hundreds or thousands of cells. The source end-host submits a packets after an entire packet is buffered at the interface module, and the packet transmission never start before an entire packet is buffered at the interface module. On the other hand, in ATM networks, the cell transmission can start before an entire packet is buffered at the interface module, i.e., pipe line data transmission.

At the routers in packet networks, an entire packet is buffered whenever it goes through the routers. Therefore, there is a delay due to packet buffering at router. Also, there is a delay due to so called HOL (Head Of Line) blocking. By the HOL blocking, the packet transmission must be halted, until the transmission of previous packet is completed. On the other hand, in ATM networks, the multiple packet transmissions can performed simultaneously, i.e., cells that belong to the different connections can be interleaved on a single transmission link.

- Feasibility by hardware implementation
  ATM is designed so that the hardware implementation should be easier than the packet networks. Since the packet (i.e., cell in ATM networks) size is unique, it will be easier to implement by hardware logic, compared to in packet networks. By means of hardware oriented protocol implementation, we can obtain larger throughput using the same VLSI process rule.
- Scaling
  The total system throughout can be scaled up easily. The followings are the methods to scale the total system throughput.
  * Increase the number of switch interface ports
    Since the ATM is switched oriented network, the total system throughput increases by the increase of switch interface ports. In other words, the total system throughput can be increased even though the system clock speed is not increased.
  * Interconnect ATM switches
    ATM switch nodes can be interconnected without any internetworking entity (e.g., bridge and router), using NNI protocol.

  Therefore, it is said that ATM is scaleable architecture.

  Here, as well as the scaling up of shared media platforms, the ATM network segments that can include multiple ATM switch nodes can be interconnected through routers for scaling up. Which strategy should be taken for system scaling up depends on the requirements and policies of the networks.

• Disadvantage

- Complex traffic control
  As discussed in the following subsection, the traffic control framework in ATM

networks is more complex than the other networks (i.e., STM and packet networks). Packet network basically provides just a best effort service, and STM network provides the same QOS for all connections. However, the ATM network provides various level of QOS classes simultaneously, while performing the network resource sharing.

- Large transmission overhead
  The transmission overhead due to the header/trailer field added to the user information is much larger, compared to the other platforms. In STM networks, the transmission overhead due to the header/trailer field is nothing. In packet networks, the transmission overhead due to the header/trailer field can decrease by the use of larger sized packet. However, in ATM networks, the transmission overhead due to cell header is about 10% ($\simeq 9.4\%$).

The detailed ATM network architecture is discussed in the following subsections in this section.

[4] Packet Transfer (IP Datagram Transmission)

Packet transmission using Internet Protocol is based on connectionless. The data transmission is performed packet-by-packet, not by connection-by-connection. Therefore, in packet networks, the network does not maintain/establish state information associated with every connection, i.e., stateless network. Usually, the packet network maintains the packet transmission route according to the destination point, but this state is not associated with end-to-end packet flow (i.e., connection). Also, the maintained information does not depend on the source address, i.e., the information only depends on the destination address. Every end-host can send a packet whenever it wants without any connection establishment procedure. Since the network does not maintain the state information associated with each connection, each user packet must have a global destination address of the packet. In IPv4, it is 32 bits length address.

The service quality to be provided by the packet networks is usually the best effort service, except for some resource reservation oriented protocols (e.g., ST-II [ST-II] and RSVP [RSVP]). The packet may have some latency due to buffering at some network entity such as router and may be discarded due to buffer overflow at some network entity. However, since the network resource (e.g., bandwidth and buffer space) is shared among all packet flows, the utilization efficiency of the network resource is large, compared to the other platforms (STM and ATM networks).

The followings are the advantages and disadvantages of packet networks.

- Advantage

  - Loose clock synchronization
    Since the packet transmission is based on an asynchronous way using packet buffering, the clock synchronization can be local and the global clock synchronization is not required. For example, at the router, the input interface module, the output interface module, and the packet switching module can operate by the different reference clocks.

- Clock hierarchy free
  The clock speed delivered by the network node has a digital hierarchy and granularity, as well as STM network. However, there is no restriction on the actually available bandwidth for each end-host. If the network resource is available, the end-host can send packet at the interface speed, on the fly packet submission. As a result, the utilization efficiency of network resource (i.e., bandwidth resource) will be better than in an STM network.
- Network resource sharing
  The network resource (e.g., bandwidth or buffer space) is used, whenever the packet arrives at the network node (e.g., router). By the network resource sharing among the connections, the utilization efficiency of network resource will be better than in an STM network.
- Stateless network
  The network need not maintain the state information for every connection, but just maintains the information how the received packet should be forwarded. Also, since each packet has a TTL (Time To Live) field and packet forwarding decision is based on hop-by-hop base, the maintained packet forwarding information could have some transitional in-consistency.
- Datalink independent
  Packet network only specifies network layer protocol (e.g., IP). Therefore, wide variety of datalink technologies can be applied to. ATM technology is just one of these datalink technologies, but ATM will be widely used as a common and major datalink platform.

• Disadvantage

- Un-specified QOS
  The service quality provided to the user applications can not be specified, in general. Some packets may experience substantial buffering delay, and some packets may be discarded due to buffer overflow. However, as discussed in the following subsection, the resource reservation oriented protocols, [RSVP][ST-II], have been under development in IETF (Internet Engineering Task Force).
- HOL (Head Of Line) Blocking
  The network resource (e.g., buffer space and bandwidth) is shared among many packet flows. Some packets are large size and the other packets are small size. In general, we can say that the application requiring small latency for data delivery uses small packets, since both the latency due to packetization (i.e., data packing at source end-host) and the latency due to packet buffering at the intermediate node are smaller than with large packet. On the contrary, the application requiring efficient and high throughput data delivery uses large packets, since the total overhead of packet header decreases and the frequency of process interruption (leading to the degradation of data processing performance of computer system) due to the reception of packet is reduced.

When the packet service policy is a simple FIFO (First In First Out) that is widely adopted in the current packet switch nodes (i.e., routers), the transmission of packet can not start until the transmission of packets queuing before complete, i.e., HOL blocking. The jitter of queuing latency at the queuing point increases, when large packets and small packets share the common queue. This means that the queuing latency at the queuing point will be sometimes large. This is serious for the application using small sized packet in order to get a data transmission with small latency.

2.1.2 ATM Service Capability

ATM platform provides both connection oriented service and connectionless service to the user. Both connection oriented and connectionless service are provided using connection oriented virtual channel connections (i.e., ATM-VCCs). Connectionless service does not require a connection setup procedure, in order to transfer the connectionless datagram (e.g., CLNP packet). In connectionless service, ATM-VCC(s) to provide connectionless service is(are) established in advance between the end-node and CLSs (i.e., ConnectionLess Servers) [I.364]. The ATM-VCC to be used for connectionless could be either PVC (Permanent VCC) or Semi-PVC.

ATM networks provide four of service categories [AF-TM], and provide wide variety of QOS classes. The QOS parameter itself has been still under discussion. The following parameter will be the likely QOS parameters.

- Connection blocking probability
  Connection setup request will be blocked (i.e., rejected), when the available network resource for the connection is insufficient to provide a required QOS. The network provide a sufficient network resource to achieve the performance objective associated with the connection blocking probability.

- Cell Transfer Delay (CTD)
  CTD is an end-to-end cell transfer delay from the source end-node to the destination end-node, and is not a deterministic value but is a statistical value.

- Cell Delay Variation (CDV)
  CDV is associated with a jitter of CTD. CDV quality gives a strong impact for the required buffer space at the destination end-node.

- Cell Loss Ratio (CLR)
  CLR is the probability that the cell is discarded due to the buffer over flow or due to the bit errors in cell header field.

- Cell Mis-delivery Ratio
  Cell mis-delivery ratio is the probability that the cell is not delivered to the correct destination due to some bit error in cell header field.

- Bit Error Ratio (BER)
  BER is the probability that a bit in a cell is errored. BER does not generally depend on service category nor on QOS class.

Table 2-1 shows the service categories defined in ATM Forum [AF-TM], with provided QOS abstraction for each service category.

Table 2-1. Service categories in ATM networks

Attribute	ATM Layer Service Categories				
	CBR	VBR (RT)	VBR (NRT)	ABR	UBR
CLR	specified[1]			specified[2]	unspecified
CTD and CDV	specified		specified[8]	unspecified[6]	unspecified
PCR and CDVT[5]	specified			specified[4]	specified[3]
SCR and BT	n/a	specified		n/a	
MCR	n/a			specified	n/a
Control Information	no			yes	no

[Notes]
[1] For CBR and VBR the Cell Loss Ratio may be unspecified for CLP=1.
[2] Minimized for sources that adjust cell flow in response to control information.
[3] Not subject to CAC, and UPC procedures may use a different values from the other service categories.
[4] Represents the maximum rate at which the source can send as controlled by the control information.
[5] CDVT (CDV Tolerance) is either explicitly or implicitly specified for PVSs or SVCs.
[6] Objective of ABR service is that the network does not excessively delay the admitted cells.
Requirement for explicit specification of the CTD and CDV is for further study.
[7] For non-realtime VBR, the CTD is specified and the CDV is unspecified.

2.1.3 Protocol Structure

Figure 2-4 shows the protocol structure of ATM. An ATM has three planes; user plane, control plane and management plane. User plane is responsible for the transmission of cells containing user information. Control plane is responsible for the control of VCC and VPC (i.e., establish and tearing down of connection) using the signaling protocol (e.g., Q.2931). Management plane is responsible for the management of each connection as well as of the network.

2.1.4 Signaling and Connection Management

A switched ATM connection (SVC ; Switched VC) is established through a signaling procedure (e.g., Q.2931). During a signaling, the end-node indicates the required service category and QOS class with its traffic descriptor. The traffic descriptor represents how the cells are

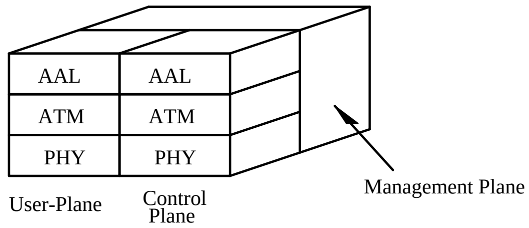


Figure 2-4. Protocol Structure in ATM Networks.

submitted from the end-node to the network. At least, the traffic descriptor includes the PCR (Peak Cell Rate), that is the maximum number of transmitted cells during one second.

Based on the signaling information (i.e., required service category with QOS class and traffic descriptor) from the end-node, the network decides whether the network can accept the ATM-VCC establishment request, taking into account the network status. Once the network decides the ATM-VCC establishment should be accepted, the network establishes the state (e.g., the VPI/VCI mapping table in the ATM switch nodes) associated with the ATM-VCC. The ATM-VCC is torn down due to the connection release message of signaling message. By the ATM-VCC tearing down procedure, the established state associated with the ATM-VCC is eliminated.

As mentioned above, the state is not modified during a life-time of ATM-VCC, but the state is static during a life-time of ATM-VCC. In other words, the management for the connection is based on connection-by-connection. Therefore, we can say that the connection management in the ATM network is a hard-state, instead of a soft-state like in RSVP [RSVP].

2.1.5 Error and Traffic Control Framework

[1] Error Detection and Correction Capability

The purpose of error control in data networks is to provide an error free data-unit delivery between a pair of SAPs. Error detection and error correction are the key functionalities of error control. ATM network fundamentally has only an error detection capability.

Table 2-2 shows the error detection and the error correction capability below CPCS (Common Part Convergence Sub-layer) in ATM networks. As shown, the error(s) both for header field and for payload are detected [I.150][I.363]. On the other hand, the error for payload (CPCS-SDU) is not corrected, excluding the option of AAL Type 1. SSCS of AAL could have an error correction capability for the assured mode operation [I.363]. However, at this time, neither ITU-T nor ATM Forum has been discussed the detailed protocol specification of error correction functionality in SSCS.

The error control is necessary for data delivery. On the contrary, for voice or video communication, the error control may not be necessary. Regarding data delivery service, AAL Type 5 will be used for many cases. Since AAL Type 5 does not have an error correction capability, the dropped or errored data-unit (CPCS-SDU) will be retransmitted by the request of the upper layer protocol (e.g., TCP).

Table 2-2. Error Detection and Correction Capabilities in ATM Networks

		Error Detection		Error Correction	
		Payload(SDU)	Header	Payload (SDU)	Header
PHY		No	Yes(by HEC)	No	1 bit(ATM header)
ATM		No	Yes(by HEC)	No	1 bit
SAR	Type 1	No	Yes(by SNP)	No	1 bit (SN)
SAR	Type 3/4	Yes(CRC-10)	Yes(CRC-10)	No	No
SAR	Type 5	No	No	No	No
CPCS	Type 1	No(optional)	No(optional)	No(optional)	No(optional)
CPCS	Type 3/4	Yes(LI for lost)	No	No	No
CPCS	Type 5	Yes(CRC-32/LI)	Yes(CRC-32)	No	No

[2] Traffic Control Framework

ATM networks basically have two levels of traffic control. One is connection level, and the other is cell level.

Connection level traffic control is a connection admission control (CAC), that includes routing. Routing protocol running among ATM switch nodes decides which route the ATM-VCC's cell flow should take. And, the connection admission control process in each ATM switch node decides whether the ATM-VCC can be accommodated or not. The algorithm of admission control is not subject for standardization. Routing decision and connection admission decision is performed only during an ATM-VCC establish procedure, that is invoked by the signaling (Q.2931).

Cell level traffic control is shaping, policing and reactive flow control. Shaping (i.e., cell transmission scheduling) will be performed at certain point in the ATM network, as well as at the end-node before actually the cells are submitted to the network. The purpose of shaping within the network is to achieve better network resource utilization, while providing desired QOSes. On the contrary, the purpose of shaping at the end-node is to avoid cell discarding due to the violated cell submission pattern against the negotiated traffic descriptor through a signaling. This is because the cell that violates the negotiated traffic descriptor will be discarded or marked (i.e., CLP tagging) by the policing function. Policing control is referred to an UPC (Usage Parameter Control). The CLP (Cell Loss Priority) tagging is a network optional. With the CLP tagging option, the cell with CLP=0 will be modified with CLP=1, if the cell flow violates the negotiated traffic descriptor.

Shaping and policing will be adopted to all service categories. However, the reactive flow control is applied only to the ABR service. In the ABR service, the cell submission rate is regulated according to the control information issued by the ATM switch nodes (and by the destination end-host). The control information sent back to the source end-host is either the congestion indication or the explicit available cell submission rate. Though the reactive flow control *in ATM layer level* applied only to the ABR service, ATM network provides explicit congestion information for all service categories. The upper layer entity can use this congestion information so as to regulate the data submission rate to the network, if necessary.

Another flow control framework, so called frame based intelligent congestion control, will be applied to the ATM network. One is PPD (Partially Packet Discarding), and the other is ERD (Early Random packet Discarding). Both control schemes recognize the frame (e.g., IP packet) in the ATM networks. In congestion status, the cells are discarded based on frame. In a PPD, the whole of succeeding cells after the discarded cell in the frame will be discarded. In ERD, the whole of cells in the frame, which will be randomly picked up among the cell flows that pass through the ERD control point, are discarded, when the ERD control point determines the congestion will be going to occur. Both PPD and ERD are useful for data communication that requires an error-free data transmission between the source and destination end-host. Therefore, whether to apply the PPD/ERD is signaled during a ATM-VCC establishment procedure.

2.1.6 Addressing Structure in ATM Networks

ATM networks use either E.164 address or NSAP address. ITU-T adopts an E.164 address, which will be used in public networks. ATM Forum adopts an NSAP address, which will be at least used in private networks (and may be also used in public networks). Both E.164 and NSAP address will be co-ordinated hierarchically. E.164 has the decimal 15 digits address space, and NSAP has the binary 20 bytes address space. NSAP address can encapsulate E.164 address in it.

2.2 Internet Protocol Suite

2.2.1 IP Service Capability

IP network transfers the user information through a connectionless packet relaying. The user information is transferred by the variable size packet. Since the network does not establish the state for every connection in the network, the required address information in the packet header (i.e., IP address) is a globally unique value. The packets are forwarded based on a global unique IP address.

IP network maintains the state of route how the received IP packet is forwarded to the appropriate next node (router or destination host). And, the maintained information does not depend on the source address, i.e., the information only depends on the destination address. Every end-host can send a packet whenever it wants without any connection establishment procedure.

The service quality to be provided by the packet networks is usually the best effort service, except for some resource reservation oriented protocols (e.g., ST-II [ST-II] and RSVP [RSVP]). The packet may have some latency due to buffering at some network entity such as router and may be discarded due to buffer overflow at some network entity. However, since the network resource (e.g., bandwidth and buffer space) is shared among all packet flows, the utilization efficiency of the network resource is large, compared to the other platforms (STM and ATM networks).

IP network provides basically a best effort service. However, by the use of some resource reservation oriented protocol (e.g., ST-II and RSVP), certain end-to-end QOS quality will be provided to the application. ST-II is the hard-state oriented resource reservation style, as similar to the resource reservation style in ATM networks. RSVP is the soft-state oriented

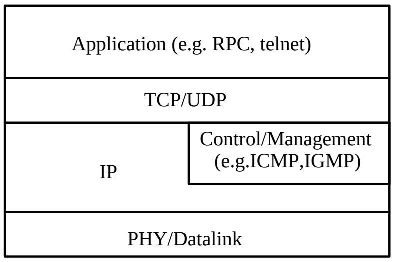


Figure 2-5. Protocol Structure in IP Networks.

resource reservation style, which is different from the resource reservation style in ATM networks rather is similar to the route state maintenance in IP routing protocol. RSVP is soft-state and receiver initiated protocol. On the other hand, ST-II and ATM can be said as the hard-state and sender initiated protocol. In the soft-state, the maintained state information is always refreshed and the state is modified whenever the state is changed. When the refreshment procedure does not work due to some reason (e.g., failure of network entity), the maintained soft-state is eliminated. This procedure is fundamentally the same as the state management in the routing protocol. In the receiver initiated protocol, the (soft-) state is established only when the receiver gives a trigger. And, the state maintenance can be distributed, i.e., the source entity does not have to all state information of receivers. Due to the property of the receiver initiated protocol, the receiver initiated protocol with soft-state policy can accommodate large number of receivers. As a result, it can be said that ST-II is the protocol providing a relatively small scale of QOS-ed multicast service, and that RSVP is the protocol providing any scale of QOS-ed multicast service. Here, the multicast communication service includes the point-to-point communication service.

IPv4 networks provides three communication types, and IPv6 networks provides basically two communication types. IPv4 provides unicast (point-to-point), multicast (multipoint-to-multipoint) and broadcast services. IPv6 provides unicast and multicast services. In IPv6, the broadcast service is realized as the one of multicast service.

2.2.2 Protocol Structure

Internet Protocol suite corresponds to the network layer and transport layer functionalities, when we compared to OSI's protocol reference model. We can see the IP protocol as the protocol simplifying the OSI's protocol reference model.

Figure 2-5 shows the protocol structure of IP suite. There are only four layers in the model, i.e., Physical layer (or datalink layer), IP layer, TCP/UDP layer and application layer.

The function of IP is forwarding the IP packet issued by the source host toward the destination host, based on the routing information that is generally established by some routing protocol. Beside an IP, some control and management protocols (e.g., ICMP) are defined. The followings are the some examples of the control protocols.

- ICMP (Internet Control and Management Protocol)

The task of ICMP is to realize the status of IP forwarding in the network. For example, the error condition (e.g., destination is not reachable) is indicated to the sender node through the ICMP.

- IGMP (Internet Group Management Protocol)
  The task of IGMP is the management of multicast communication group. In order to join(quit) to(from) the multicast service, the receiver end-host has to indicate it through the IGMP.

- RSVP (resource ReSerVation Protocol)
  The task of RSVP is to provide the soft-state resource reserved IP packet forwarding route over the IP networks. The appropriate next hop node is determined based on the routing information established by routing protocol

- ST-II (STream 2 protocol)
  The task of ST-II is to provide the hard-state resource reserved IP packet forwarding route over the IP networks.

- Routing protocol
  The task of routing protocol (e.g., OSPF or BGP) is to decide to which next hop node the received IP packet should be forwarded. The routing information established by routing protocol is generally dynamic and soft-state.

The function of transport protocol layer (e.g., TCP or UDP) is to provide the data stream to the upper layer. The key functionalities of transport protocol are the establishment of session, error control and flow control.

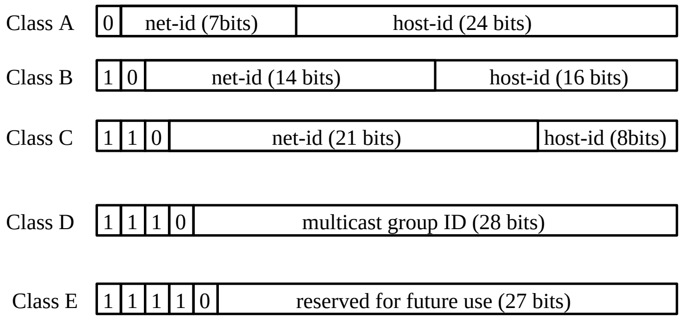
1. Session establishment
   TCP and UDP assign a port-id (16 bits) for each data flow from the application. Each data flows are identified by the combination of source IP address source port-id, destination IP address and of destination port-id.

2. Error control
   TCP provides an error-free data delivery to the application. UDP does not provide an error-free data delivery to the application.

3. Flow control
   TCP provides a dynamic window control to regulate the IP packet submission rate, according to the network status. On the other hand, UDP does not have any flow control function.

2.2.3 Addressing Structure

The current IP is version four (IPv4), and the next generation IP is version six (IPv6). In this subsection, the addressing structures of IPv4 and IPv6 are briefly presented.



Class	Range
A	0.0.0.0 to 127.255.255.255
B	128.0.0.0 to 191.255.255.255
C	192.0.0.0 to 223.255.255.255
D	224.0.0.0 to 239.255.255.255
E	240.0.0.0 to 247.255.255.255

Figure 2-6. IPv4 Address Structure

[1] IPv4 addressing structure

Address field length of IPv4 is 32 bits. IPv4 has basically three address types, i.e., unicast, multicast and broadcast.

The unicast address type has three classes; class A, B and C. Unicast address is constructed by net-id and host-id. Class A address has 8 bits net-id and 24 bits host-id, class B address has 16 bits net-id and 16 bits host-id, and class C has 16 bits net-id and 8 bits host-id. Therefore, the class A network using class A address format can accommodate upto $(2^{24} - 1)$ hosts, the class B network can accommodate upto $(2^{16} - 1)$ hosts, and the class C network can accommodate upto $(2^8 - 1)$ hosts. As described above, IPv4 address had a 8 bits boundary before the CIDR was adopted. CIDR is classless addressing structure. Therefore, both net-id field and host-id field in IPv4 address does not have any octet nor nibble boundary. In other words, any length of net-id less than 24 bits can be allowed. The reason why we must use the CIDR addressing structure is due to the lack of IPv4 address resource to be newly allocated to. By the use of CIDR, the address utilization efficiency can be dramatically improved. The problem due to the use of CIDR is the increase of routing table looking up complexity at the routers. This is because any size of net-id length must be considered as a eligible net-id, though it was only three net-id length types previously.

The multicast address is sometimes called as a class D address. No address structure for multicast address is defined in IPv4. The available multicast address space is 28 bits.

The broadcast address is represented by all "1" bit pattern or all "0" bit pattern. Broadcast address is generally valid only within an IP subnet. Here, IP subnet is defined as the domain which includes the hosts having the same net-id.

[2] IPv6 addressing structure

IPv6 has 128 bits address field, and has unicast address and broadcast address. The features of IPv6 are following. IPv6 address has net-id and host-id fields, as well as IPv4 has.

- Classless address type
  The boundary of net-id field is based on a bit. Therefore, there is no octet nor nibble boundary.
- Private address
  Private IP address, that is valid only within the given network domain (e.g., within a corporate network), can be explicitly indicated by the common address prefix field.
- Scope of multicast address
  Scope where the multicast address is valid can be explicitly specified. Scope field has 4 bits, and can represent from a local to a global scope.
- No broadcast address
  Broadcast is recognized as one of multicast address code point.
- Flow-id
  28 bits of flow-id is defined. In the IP packet processing at the router, the router can treat differently the IP packet flows based on the flow-id, even when the source and destination IP addresses of packets are the same. In IPv4, an IP packet is identified only by the source and destination IP addresses. By the definition of flow-id, source destination end-host pair can define multiple IP packet flows that may be treated as different IP packet flows at the routers.

2.2.4 Autonomous Routing

The management of routing information is distributed and autonomous. In order to establish the routing table, some routing protocol (e.g., OSPF) is applied to the given autonomous system. Within the autonomous system, any routing protocol/policy can be applied. Also, since the autonomous system and routing protocol is recursive, the autonomous system can easily scale up. Associated with the routing protocol, an autonomous domain that may include many nodes within there can be abstracted as one node for the higher level routing hierarchy. And, any management policy (e.g., routing policy) can be applied to every autonomous system.

2.2.5 Flow Management

The conventional IP network does not have any state associated with the end-to-end packet flow (i.e., connection). IP packet forwarding is completely stateless. However, as discussed above, RSVP and ST-II introduce the packet flow management policy into the IP network. ST-II provides a hard-state route and RSVP provides a soft-state route over the IP network.

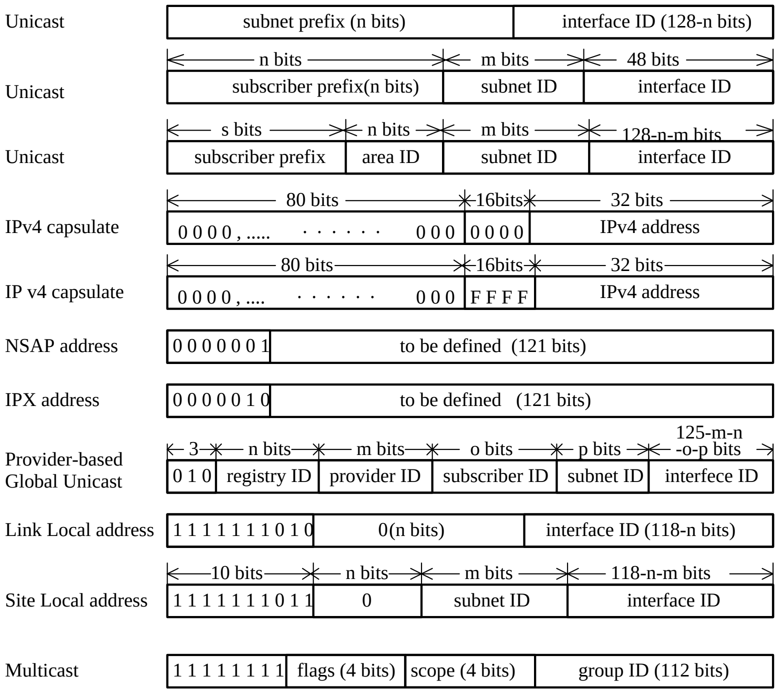


Figure 2-7. IPv6 Address Structure

In hard-state route, the state of route is established before the transmission of IP packets starts. And, the state of route does not changed during a life-time of session. Therefore, the route provided by a hard-state protocol is the static route.

In soft-state route, the state of route is established before the transmission of IP packets starts, as well as a hard-state protocol does. However, the state of route is always refreshed during a session. As a result, when the status of the network is changed, the soft-state route may be changed to the different route. Therefore, the route provided by a soft-state protocol is the dynamic route.

Both ST-II and RSVP reserve some network resource (e.g., bandwidth) for certain service (i.e., unicast or multicast service). This procedure is a state establishment. The established state is associated with a route where the IP packets are forwarded.

2.2.6 Error and Flow Control

The conventional IP network does not provide error control nor flow control. It is a best effort service platform. However, with the resource reservation oriented protocol (e.g., RSVP), IP network will have some flow controls, that will be shaping and policing. The purposes of shaping and policing is similar to the shaping and policing introduced in ATM networks. The purpose of shaping in RSVP is scheduling the IP packet transmission so that the IP packet transmission pattern is compliment with the T-spec that specifies the IP packet transmission pattern from the upstream node(s). The purpose of policing is to avoid the degradation of provided QOS to every service, due to the mis-behaved IP packet transmission from upstream node(s).

The transport protocol operates end-to-end base, not hop-hop base. This means that the transport protocol does not performed at routers. Transport protocol provides an appropriate data stream to the upper layer processes. The tasks of transport protocol are the error control and flow control. TCP and UDP are widely accepted transport protocol in IP network.

TCP provides an error-free data stream to the upper layer process, however UDP does not provide error-free data stream to the upper layer process. In TCP, errored or missed IP packet is re-transmitted from the source host. The policy of IP packet re-transmission is a go-back-n. The IP packet transmission is resumed from the errored or missing IP packet, i.e., going back to the transmission point where the error is observed. Therefore, in TCP, the IP packets that are correctly transferred to the destination host will be transmitted again, when the previously transmitted IP packet from source host is errored or missed.

TCP has an end-to-end closed loop flow control function, however, again, UDP does not have any flow control function. The flow control policy adopted in TCP is a dynamic window control. The transmission window is defined. The IP packets in the transmission window can be transmitted from the source host, without any IP delivery acknowledgment indicated by the destination host. The larger window achieves larger throughput. When IP packet is errored or missed, the transmission window is shrunken to reduce the total amount IP packet transmitted from the source host in a certain period. When there is no further IP packet error or missing, the transmission window is gradually increased.

2.3 Role of ATM in Global Internet

ATM is becoming widely accepted as a technology that can provide a number of architectural benefits including, scaleability and providing a wide variety of QOS classes.

Although IP can provide an end-to-end QOS, as well as a conventional best effort service. In order to provide end-to-end QOS at the network/transport layer level it is preferable that the datalink provide the required QOS for the given packet flow.

Figure 2-8 shows the protocol structure using ATM-VCCs with TCP/IP. The ATM-VCC is terminated at the router which performs the functions required by the network protocol (i.e.IP).

Figure 2-9 shows how the IP packet is fragmented in ATM networks. Since an IP packet is generally larger than an ATM cell, the IP packet is fragmented into multiple cells. In this figure, the fragmentation is performed at the SAR (Segmentation and Reassembly) sublayer and in the CPCS (Common Part of Convergence Sublayer), i.e., streaming mode operation [I.363]. In the message mode operation, the CPCS will not fragment CPCS-SDU (CPCS - Service Data Unit) into multiple CPCS-PDUs (CPCS - Protocol Data Unit).

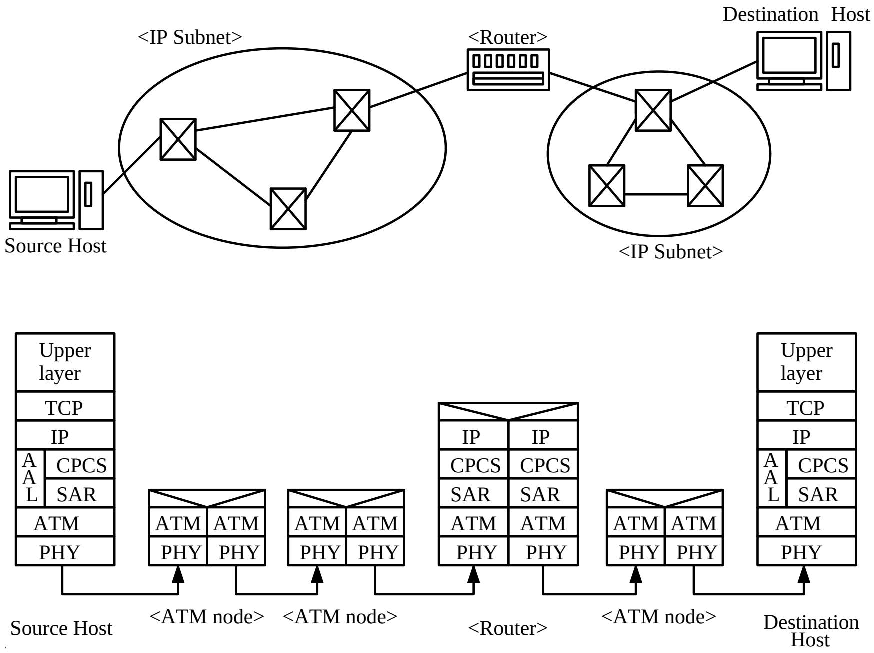


Figure 2-8. TCP/IP Communication over ATM Networks.

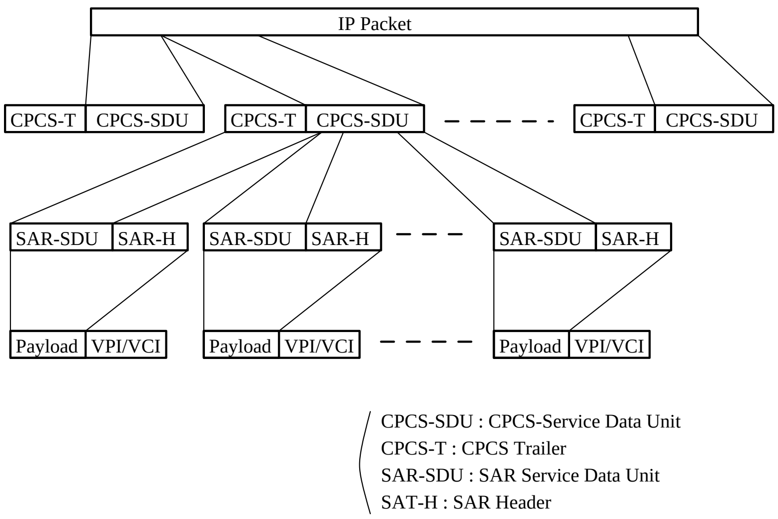


Figure 2-9. IP Packet Fragmentation in ATM Networks.

3 Related Works of IP over ATM Architecture

In this section, the IP over ATM architectures proposed in the research and standardization community is categorized and the issues of IP over ATM architecture model is discussed. In the research and standardization community, such as IETF and ATM Forum, many architectures to provide IP service over ATM has been proposed. However, these architectures still have some issues to solve and can not provide a large scale error-free multicast service. This section clarifies the issues of IP over ATM architecture models proposed by other researchers and clarifies the issues of a large scale error-free multicast over ATM.

3.1 IP Over ATM Architectures

The provision of IP communication service over the ATM platform is one of the keys whether the ATM networks obtain a position as the widely accepted datalink platform for the Internet platform. The extraction of properties of ATM technologies for IP communication service platform is the purpose of the architecture proposed in this paper.

The properties and the benefits using ATM platform will be (1) connection oriented platform providing simultaneous multiple flows between end-nodes pair, (2) QOS for each VCC can be negotiable, and (3) explicit flow identifier is assigned to every cell flow. The provision of multiple cell flows and transmission of IP packets through cells could reduces the delay jitter experienced by the IP packets.

Many IP over ATM architectures are proposed in research and standardization community, such as IETF and ATM Forum. However, every proposed architecture model has certain issue to be solved. This section discusses and clarifies what is the issue for the IP over ATM architecture model. The previously proposed IP over ATM architecture is categorized into the following three architecture models.

- CLPF (CLassical Packet Forwarding) Model
- SCPF (Short-Cut Path Forwarding) Model
- LSTL (Label Switching with Transparent Links) Model

3.1.1 Logical IP Subnet (LIS)

LIS is a set of nodes (hosts and routers) that have the same address prefix associated with IP address. Though the nodes belong to the same LIS have the same IP address prefix, these nodes need not be geographically contiguous. An ATM network segment can contain multiple LISes, which are not contiguously distributed.

The LIS has at least the following functionalities.

- Direct ATM-VCC must be established between the end-nodes in the same LIS.
- The IP packet toward the end-node, that does not belong to the same LIS as the source end-node belongs to, can be forwarded through the router belonging to the LIS.

- Routers belonging to LISes participate in the routing protocols, to decide how the IP packet across the LISes should be forwarded.

- ARP (Address Resolution Protocol) will work to resolve the appropriate data-link layer address from the destination IP address in order to forward the IP packet.

Theoretically, both the numerical and geographical size of LIS could be very large, and the number of LISes in a ATM network segment could be also very large. However, from the view point of practical implementation of LISes over the ATM networks, the number of LISes in the ATM network segment, the number of nodes in the LISes and the geographical scale of LIS must be reasonably small. Especially, the product between the number of LISes in the ATM ATM network segment and the number of nodes in LIS should not be large. The followings are the reasons why the number of nodes in LIS and the number of LISes in the ATM network segment must not be large.

- The amount of information associated with each nodes will increase, according to the increase of product the number of LISes and the number of nodes in LIS. One example information that must be maintained in the database will be the address mapping information between the IP address and the corresponding ATM address, for the provision of ARP functionality. Since there is no geographical correspondence to IP address prefix of LIS, the database will be hard to be geographically distributed. The database could be distributed. However, the corresponding database server for a certain end-node could be very far from it, even if there is some database server that serves to the other LIS. If the database server is not distributed, the amount of maintained information increases linearly according to the product between the number of LISes and the number of nodes in LIS.

- There are the applications that depends on the capability of broadcast service in the IP level (i.e., IP broadcast). In IP broadcast, the broadcast packet must be delivered to all nodes belonging to the IP subnet. When the geographical scale of LIS is large, the LIS would include WAN segment. When the LIS includes WAN segment, the IP broadcast message must always travel WAN segment that is subject for charging.

- In RSVP, the downstream nodes must always send back the $RESV$ message, whenever the $PATH$ message is received. When the number of downstream nodes in LIS is large, the number of $RESV$ messages that the upstream node will receive is also large. The large number of $RESV$ message reception causes large number of process interruption for the upstream node. As well-known, the process interruption leads to a context switching that is costed for the computer system and degrades the processing performance. As a result, in order to avoid the degradation of processing performance at the nodes in the soft-state multicast tree, the number of nodes in LIS should not be large.

3.1.2 CLPF (CLassical Packet Forwarding) Model

CLPF, CLassical Packet Forwarding, model is basically the architecture discussed at IETF IPoverATM Working Group. In general, logical IP subnets (LISes) will overlay on a single

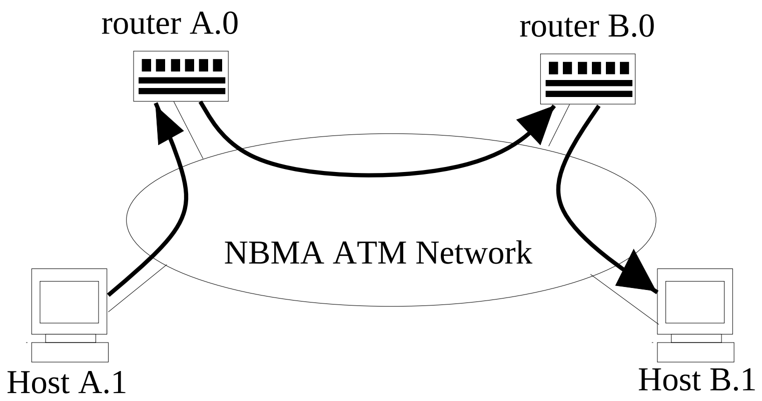


Figure 3-1. CLPF (CLassical Packet Forwarding) Model

ATM network segment, which may be a large scale network. This means that several nodes (i.e., hosts and routers) which generally belong to different IP subnets will exist on a single (physical) datalink segment.

IP subnets are interconnected by the conventional routers. The router in Classical IP model has only IP forwarding capability, but it does not have a cell relaying capability. For the intra-IP subnet communication, the Classical IP model allows a direct ATM connectivity between the nodes (i.e., host and routers). On the other hand, for the inter-IP subnet communication, the IP packet must pass through the router(s). This means that ATM-VCC shall be always terminated at the router. Therefore, it is impossible to provide end-to-end QOS-ed virtual connection using Classical IP model, without some protocol, e.g., RSVP. Here, even when some resource reservation protocol is applied in CLPF model, IP packet forwarding is still performed with an IP packet level (i.e., examination of IP packet header for IP packet routing) and it causes both delay and HOL (Head Of Line) blocking at router. Also, since the IP packet must be always reassembled at the router (i.e., packet-by-packet forwarding), every IP packet experiences IP packet reassembling delay.

However, since the Classical IP model does not violate a subnet model, this model can obtain the benefits (e.g., providing scaleable soft-state multicast service) discussed in the section 4.1.

3.1.3 SCPF (Short-Cut Path Forwarding) Model

SCPF (Short-Cut Path Forwarding) model is basically the architecture discussed at IETF ION (IP over NBMA) Working Group and at ATM Forum MPOA (Multi Protocol Over ATM) Working Group. This architecture mode fundamentally use the ATM address resolution mechanism, that is Next Hop Resolution Protocol (NHRP) using the target IP address of the received packet [NBMA][NHRP].

LIS will overlay on a single ATM datalink segment. NBMA (Non-Broadcast Multiple Access) network is defined as the network domain where end-to-end ATM-VCC can be established. Within the NBMA network, each host could establish a direct ATM-VCC toward any host that is located both at other LIS and at the same LIS without passing through router(s), i.e., short-cut routing. However, when the IP packet should be transferred to outside an NBMA network, IP packet will be transferred to the border router between two NBMA networks. This means that, even if the neighbor NBMA network is also ATM network, it is impossible to establish an end-to-end ATM-VCC using IP address. Then, we can not obtain high performed IP packet forwarding beyond the NBMA network. Here, the large scale NBMA network may be segmented into multiple (and sometimes hierarchically structured) NBMA subnets, due to security concerning or due to obtain a network scaleability.

With NHRP, host or router will resolve the ATM address for a given destination IP address. If the given destination IP address is not directly reachable, the most appropriate exit point's (i.e., router's) ATM address will be resolved. Here, even if we use NHRP, a packet could experience hop(s) of IP processing (passing through router(s)) within a ATM cloud (i.e., NBMA domain). As well known [NHRP], since a direct ATM-VCC can be established regardless of routing topology/hierarchy defined in the routing protocols, there is a possibility to be created a long period routing loop.

Since the conventional routing protocol for connectionless IP packet routing is still running among the routers, NHRP may require some major modifications of IP packet forwarding for the routers and hosts. Here, the reason why we need the conventional hop-by-hop IP packet forwarding, even when NHRP is available, is that there are many short-lived communication (e.g., DNS query) for which communications with ATM-VCC establishment procedure is not appropriate and is expensive. The routers attached to the NBMA domain exchange routing information (e.g., topological information and link attributes) among them. Here, in general, the topological information shared and exchanged among the routers will not be full-mesh. When a router receives an IP packet toward the certain host whose location does not belong to one-hop distance from the router (i.e., requiring multiple hops), the appropriate next hop router is resolved. This next hop router's information is obtained by the routing table calculated by routing information advertised among the routers. Usually, the router would pick up the next hop router to forward the received IP packet toward the destination point. ARP will be used, if ARP cache for MAC address of the next hop router is flushed. Here, in the usual router's procedure, there is no chance to try to resolve the final destination router to reach the destination host. It may be said, before looking up routing table (let as table #2) generated by routing protocol, the router attached to ATM cloud will try to look up the other routing table (let as table #1), whose topological information is indicating routers are interconnected in full-mesh. If the ARP cache for the resolved destination router is missed, the router will issue NHRP query to resolve ATM address of the destination (exit) router. In this procedure, the conventional routing protocol would become meaningless, or the routing table #1 and routing table #2 must be identical (i.e., routers in the NBMA domain are assumed to be interconnected with one hop). Another solution is that router decides which table (table #1 or table #2) is referenced for every packet arrival, i.e., some packets will be transferred through a conventional hop-by-hop path and the other packets will be transferred through a short-cut route. Here, though the short-cut route has no hop to reach either at the exit/appropriate router or destination host, the routing loop may be created.

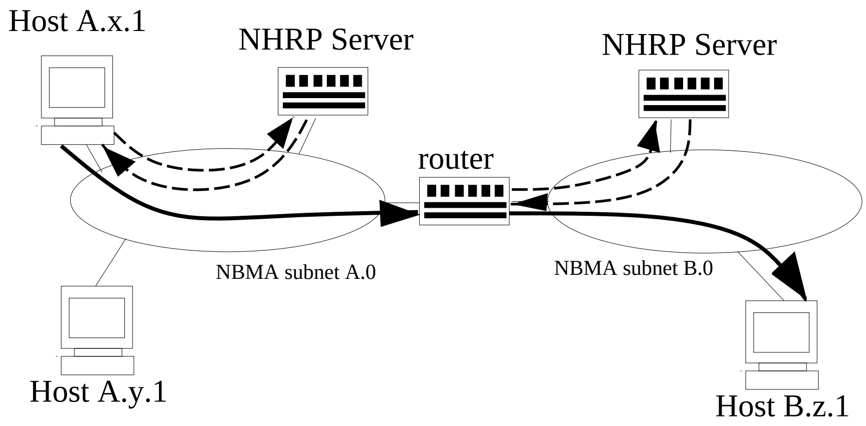


Figure 3-2. SCPF (Short-Cut Path Forwarding) Model

Above issues has come out due to the violation of a subnet model. As proposed in [RFC1937], when we only allow to establish a direct ATM-VCC across the LISs over the ATM cloud for the communication between the source and destination hosts that belong to the same NBMA network, the above issues will disappear. However, since the architecture model proposed in [RFC1937] associated with the IP packet forwarding model at the routers is essentially the same as in Classical IP model, the issue (i.e., the performance of IP forwarding at the routers) to be improved is the same as in CLPF model.

The other issue for NHRP model is the scalability for soft-state multicast service using RSVP. Since the transit routers will be physically bypassed using short-cut routes, the number of receivers in the ATM level multicast connection can be large. In RSVP, each receivers (end-hosts or routers) send RESV message to the up-stream node, which corresponds to the root node of ATM multicast connection, whenever they receive PATH message [RSVP]. Therefore, when the number of receivers in the ATM multicast connection is large, the number of RESV messages that the up-stream node must handle becomes large. Large number of RESV message reception may cause a performance degradation of the up-stream node, due to so many process interruptions.

3.1.4 LSTL (Label Switch with Transparent Links) Model

[1] Label Switching Technology

In a label switching architecture, the layer 3 packets are forwarded either using the datalink label (e.g., VPI/VCI) or using the tag label between the datalink header and the packet header (e.g., tag for Ethernet frame in tag switching architecture [RFC2105]).

The label is used as the information to forward the packets, without analyzing the internetwork layer address (e.g., IP address). This means that the label represents the destination address of t he internetwork layer packets. By using the label instead of the internetwork layer address for packet forwarding, performance is improved as the router does not need to look up in the best-match policy based routing table.

When the internetwork layer packet arrives at the node (i.e., intermediate router or host) located at the egress point of the cut-thru path represented by the label, the conventional packet processing (i.e., analysis of internetwork layer header) is performed.

In a label switching architecture, it is not generally assumed all nodes are interconnected within the datalink giving full-mesh connectivity. Packets are forwarded along the given topology, even for cut-thru packet forwarding. This is a different architectural paradigm from NHRP/MPOA. In NHRP/MPOA, the short-cut virtual connection (i.e., ATM-VCC for ATM) is established irrelevant to the nodes' topology. When a short-cut VC is established to some node, packets to the node do not travel through the intermediate routers. On the other hand, in the label switching architecture, packets are forwarded along exactly the same route, that is given by the routing protocol information. Therefore in label/Tag switching just the conventional packet forwarding process is bypassed, while the conventional route calculation and the packet's forwarding route remain the same.

[2] Label Switch with Transparent Links Architecture

There are three architectures are proposed [Ipsilon][RFC2105][ARIS]. All three architectures have the same basic underlying assumption that the label switching routers are interconnected with transparent links, such as a SONET link.

In [Ipsilon] architecture, the LSTL router establishes the cut-thru path using IFMP (Ipsilon's Flow Management Protocol) [RFC1953]. The cut-thru path is invoked by the down-stream node to the up-stream node on the appearance of the trigger IP packet.

Tag switching proposed in [RFC2105] establishes the cut-thru path using TDP (Tag Distribution Protocol) [TDP]. The trigger to establish the cut-thru path is invoked both by the down-stream node and by the up-stream node, according to the routing entry in the routing table given by the routing protocol. The cut-thru path is established using the topology information obtained by the routing protocol (e.g., BGP). The establishment of the cut-thru path is performed on an end-to-end basis, i.e., end-to-end establishment. For the scaling purpose, any granularity of flow aggregation can be defined. Large (software) processing capability is required for the egress and ingress routers of the given routing domain, when VC-merging function is not supported in the TSR (Tag Switch Router). When the VC-merging function is implemented in the ATM switch module in the TSR, large (software) processing capability may not be required for TSR.

In the ARIS proposed in [ARIS], the cut-thru path (N of multi-point-to-point connections) is established using the topology information obtained by the routing protocol (e.g., BGP). In order to provide multi-point-to-point connection, the VC merging function has to be provided in each ATM switch. To allow scaling any granularity of flow aggregation can be defined.

Since LSTL architecture does not support the standard ATM interface (i.e., ATM Forum UNI 3.0/3.1), the routers and networks applying the LSTL model do not have any interoperability with standard ATM switch and can not use standard ATM links (i.e., ATM-VC) as the link to interconnect LSTL routers.

The LSTL architecture model is shown in figure 3-3. This is a kind of router interconnection using point-to-point links. In this model, since the link among the IP switches must

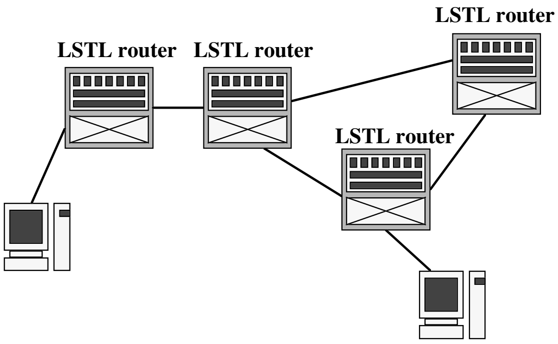


Figure 3-3. LSTL(Label Switch with Transparent Links) Model

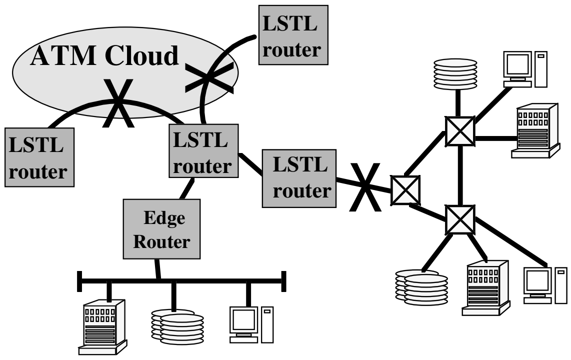


Figure 3-4. Issue of LSTL Model

be transparent links (e.g., physical link or SONET link), we can not use standard ATM switches in the network, as shown in figure 3-4.

LSTL routers can not be interconnected through an ATM public network. LSTL routers can interconnect through an ATM network only with a specific configuration, this is where the ingress VPI/VCI value and egress VPI/VCI value of the ATM link (VP) are the same. This particular configuration is not a likely configuration within a ATM network. Further, LSTL routers can not interconnect with standard ATM switches, since LSTL routers do not support ATM UNI. Since LSTL routers do not have any inter-connectivity with ATM switches, they can not provide a subnet when the network consists of ATM switches

3.2 Large Scale Error-free Multicast Service Architecture

Conventional communications in telephone networks and computer networks is largely dominated by point-to-point communication. However, recently, the requirement of multicast communication, point-to-multipoint (pt-mpt) and multipoint-to-multipoint (mpt-mpt) has increased [Deering]. For example, in the Internet, the M-Bone (Multicast Backbone) experiment has been progressed, providing real-time multicast services for video and voice.

Currently, most large scale multicast communication over the Internet is usually allowing some error or loss of transmitted IP packets, and error-free IP packet delivery to every receivers is not required (i.e., UDP over IP for IP multicast service). This is because the current applications using a large scale IP multicast basically distribute voice and video information, using a loss-tolerant transport layer with the transport protocol RTP on top of UDP [RTP].

However, in the future, other applications (e.g., multicast service of real digital information such as a large scale interactive transaction game over the Internet) will require a *"large scale"* error-free multicast service. For many distributed applications involving multiple transmitters in a group, it is sufficient if the transport system provides reliable pt-mpt communication from every transmitter to all receivers in the group [RFC1301]. Other distributed applications (e.g., distributed database system) may require an ordered and atomic multicast (mpt-mpt) communication service [Birman] [Chang]. However, such a communication service has very poor scaling properties, and is therefore only suitable for small groups.

When a large scale reliable multicast service guaranteeing the Quality of Service (QOS) is required, an IP multicast channel is appropriate that may be operated with reserved resources provided by a reservation protocol (e.g., RSVP). It may be possible to provide such a service through the individual pt-pt communications. However, using the individual pt-pt communication channels causes an extremely large latency of data delivery and requires a huge computational capability for the source entity, when the number of receivers is large.

ATM networks will provide a multicast channel (pt-mpt and mpt-mpt), as well as a unicast channel (pt-pt). Actually, the signaling protocol defined in the ATM Forum UNI specification version 4.0 includes the pt-mpt communication channel establishment capability.

Conventionally, the reliable multicast service was required by distributed computation environments (e.g., distributed database system). The number of receivers for these applications would not be large (several decades of receivers).

The challenges regarding a reliable multicast in the computer networks would be the following items.

1. How to provide an ordered and an atomic multicast
   Ordering of message delivery from multiple senders (i.e., single source ordering, multiple source ordering and multiple group ordering) would be sometimes required for multipoint-to-multipoint communication. There are many approaches to provide the ordered and atomic multicasts : e.g., (a) time-stamping [Schndr][Lamport], (b) centralized token [Chang][RFC1301], (c) distributed two-phase commitment [Birman], (d)

propagation graph algorithm [Garcia]. These approaches will be applicable to the architecture discussed in this paper.

2. How to provide a reliable (i.e., error-free) multicast
   In order to provide an error-free data delivery, some error control mechanisms (i.e. error detection and error recovery) must be applied.

3. How to provide a large scale multicast
   The error control mechanisms must be scaleable regarding both the large number of receivers and the physically widespread senders and receivers.

The framework proposed in this paper does not assume a new routing protocol (e.g., PIM, CBT) nor new transport protocol. This means that the proposed error-free multicast architecture assumes existing routing protocols (e.g., DVMRP) and existing transport protocol (e.g., MTP). For an example, a novel error-free multicast architecture, SRM (Scaleable Reliable Multicast) [SRM], assumes a new transport protocol and has to assume a new multicast routing protocol (e.g., CBT) to scale up.

3.2.1 Error Recovery Mechanisms

For some applications, the message issued by the source process must be delivered to all destination processes without any error. Error detection and error correction are the key functionalities to provide an error-free data delivery. Each layer includes some level of error control functionalities. For example, TCP does have both error detection (by checksum) and error correction (by go-back-N retransmission), but IP does only have error-detection. There are many error-detection and error correction schemes. Below, the existing error correction techniques are described briefly.

- Go-Back-N [Comer]
  IP packets in it's transmission window are consecutively transmitted from sender process without any acknowledgment (positive or negative). When some packet is errored or lost, all the IP packets after the very errored or lost IP packet in the transmission window are subject to retransmission.

- Selective Repeat
  IP packets, that are errored or lost, are selectively retransmitted. Since the order of transmitted IP packets from sender process will not be maintained at the destination process, the destination process must re-order the received IP packets to provide the ordered data delivery to the upper layer application interface. With the MTP [RFC1301], the selective retransmission of erroneous IP packet is performed by the sender process.

  In the SRM [SRM], the selective retransmission is performed by any end-node (i.e., router or host) who can retransmit the erroneous IP packet. In this architecture, the sender does not always have to preserve the IP packets to be transmitted until all receivers receive IP packet correctly.

- (FEC) Forward Error Correction [McAuley][Biers]
  Additional information (or packets) are attached to the transmitted information to recover the erroneous or lost character(s) without data retransmission. Error recovery could be both bit(s) correction and block correction. When the errored data is beyond the correction capability of the applied FEC algorithm, the data (e.g., IP packet) must be retransmitted either by go-back-n or selective repeat policy. Though FEC always requires some overhead to transmit the data, the data error probability observed at destination process can be reduced. The architecture with FEC mechanism proposed in [McAuley] and in [Biers] does not consider for the multicast service. Also, the FEC mechanism proposed does have some issues (e.g., designed for AAL3/4) that are solved in the FEC mechanism proposed in this paper.

In order to perform IP packet retransmission at the sender process, some feed-back signal (acknowledgment) from the destination process is required. There are basically two approaches. One is the positive ACK, and the other is the negative ACK (NACK). In the positive ACK approach, the destination process always sends back the ACK signal to the sender process while it receives IP packet correctly. The errored or lost IP packets are recognized by the sender process both(either) by the lack of the ACK signal from the destination process and(or) by the explicit control signal issued by the destination process. With TCP and major data transmission protocols, the positive ACK approach is applied, because of easy implementation. In the negative ACK approach (e.g., in MTP and SRM), the destination process sends back the NACK signal to the seder process, only when the destination process detects the error or loss of transmitted IP packet(s) from the sender process. Though the control signal between the sender and the destination process is reduced, the method to guarantee the error-free data delivery must combine ACK and NACK, leading to a complex protocol.

3.2.2 Issues of Large Scale Multicast Service Architecture

For multicast service, the scaling is one of the important aspects to solve. There are two aspects for the scaling. One is for the large number of receivers (and senders), and the other is for the geographically widespread senders and receivers. The later one would be regarding the timer design for the feedback loop to recognize error (this is not touched in this paper). And, the former one would be regarding both the implementation complexity at the sender's control process and the actual data transmission performance. This paper mainly focuses on the former issue, rather than on the latter issue.

When the number of destination processes becomes large, the following three issues occur.

1. Increase of Protocol State Information
   When we apply the positive ACK approach, the sender process must maintain all protocol state information for each destination processes. That means that the maintained protocol state information at the sender process will be *order of* N. Here, N is the number of receivers.

   One feasible approach to reduce the maintained protocol state information at the sender process would be using the intermediate layer 4 (e.g., modified TCP for multicast) entities. Between the sender and the receivers, there are the intermediate routers that act as the branching point of the multicast tree. In order to reduce the protocol status information maintained at the sender, the intermediate routers could terminate the layer 4 protocol. Multiple acknowledge packets from the destination process to the sender process could be merged into a single acknowledgment packet at the intermediate router. By the above approach, the protocol state information that should be maintained at the sender can be distributed among the intermediate routers, and it will be the *order of* $log_m N$. Here, m is the number of multicast leafs from the intermediate routers (or the sender process). However, every intermediate router must perform layer 4 protocols for every multicast service, and it must keep many IP multicast packets until the IP multicast packets are correctly delivered to the leaf entities on the down-steam toward the destination processes.

Another approach is proposed in RFC1301 (i.e., MTP : Multicast Transport Protocol). MTP uses the NACK approach with some smart technique (i.e., sending a referenced protocol state information from a special process) [RFC1301][RFC1458]. Using this technique, the maintained protocol state information in the sender process could be *order of constant* $0(c)$. However, this approach could not be applied to the case where number of receivers is large, and the receivers are geographically widespread, because too many control packets (NACK packets) will be transferred to the sender from the receivers as briefly indicated in the following item.

Protocol state management proposed in SRM [SRM] is different NACK approach. Protocol state is maintained in some end-nodes (can include sender node) in order to perform IP packet retransmission according to the reception of NACK control message. In the SRM, the sender node does not always have to maintain the protocol state, since some receiver nodes maintain the protocol state and does IP packet retransmission.

2. Increase of Control Packets from Receivers to Sender
Control packets (ACK/NACK) will be transferred from the receivers to the sender, as well as from the sender to the receivers. The control packets from the sender to the receivers could be done by a multicast channel. On the other hand, the control packets from the receivers to the sender will be issued by each receiver toward the sender. This means that sender would receive many control packets (ACK/NACK) from the receivers. The required bandwidth for the control packets from the receivers and the required processing power for the control packet will be large in a large scale multicast.

   Additionally, when the number of received control packet is large, the sender process will experience frequent interruptions to deal with the control packets. The interruption due to the reception of the control packet invokes a context switching, which is a costly operation in the computer system. The reception of many control packets at the sender process will result in the reduction of processing performance.

   The NACK approach, that is adopted in MTP of RFC1301, can reduce the average amount of control packets. However, in the worst case where all receivers could not correctly receive IP packets, all receivers will transfer the control packets to the sender.

   In SRM approach, some node issues a control packet (NACK message) to the multicast channel to let other nodes know that the control packet (NACK message) is already issued by some node to introduce IP packet retransmission. When an end-node hears a control packet issued by the other node, the end-node missing to receive the IP packet does not issue the control packet to the multicast channel. By means of this operation, the implosion of control packets issued by the end-nodes missing to receive IP packet can avoid. This mechanism basically has to assume having a new multicast routing protocol (e.g., CBT) and has to assume having a new transport protocol.

3. Increase of Observed Packet Error/Loss Probability
   An IP packet issued by the sender process is multicast to the large number of receivers in a large scale multicast. Therefore, the observed (accumulated) IP packet error or loss probability at the sender process increases approximately linearly, according to the number of destination processes [Bhag] [95-0150]. The observed IP packet error or loss probability at the sender process would be of the *order of* $N \times d$. Here, d is the diameter of data-link segments in the multicast tree. When we apply the approach using the layer 4 entities in intermediate routers, we could reduce it to *order of* $m \times d$. However, this approach has the disadvantage of requiring the modification of routers.

   In the ATM networks, the packet error/loss quality in the ATM networks degrades rapidly with increasing the following three factors.

   - Cell loss ratio (or the bit error rate of the medium)
     The packet error/loss quality will degrade by cell loss and the BER (bit error ratio) in ATM networks. The effective cell loss ratio and BER will depend on the following:
     - Physical layer type
       Some media may have a large BER, or BER characteristic of a highly bursty medium than the physical layers currently specified. For example, in wireless LANs, tolerance of BER is higher than that for physical layers of wired media.
     - ATM bearer service type
       In ATM networks, QOS parameters associated with the cell loss ratio may be negotiable, and will also depend on the service type.
     - Congestion status of the network
       During the occurrence of congestion due to congestion in ATM networks, buffer overflows will result in cell losses. It is desirable to avoid the degradation of service quality even in congestion status.

     Reducing the effective BER and cell loss ratio for the upper layer process (e.g., IP) will significantly improve the overall service quality. A number of previous researches have shown that the use of FEC may improve end-to-end ATM performance in terms of effective throughput and latency [Ohta][Ayanog][Carle][Esaki1].
   - Packet (or frame) size
     In general, the transport layer (or the network layer, e.g., IP) does not have a cell-based error correction capability. Therefor, the complete packet must be retransmitted even if the received packet has only one bit in error.

As shown in [Roman] and [94-0914], the loss rate of higher layer packets (e.g., TCP packets) grows linearly with the number of cells composing a packet. Additionally, [95-0151] has shown that the average response time of IPX and TCP protocol stack also degrades rapidly with increasing cell loss. For example, a data-unit of 64KBytes (i.e., the maximum data-unit size of AAL5), the probability that the received data-unit has any bit error is approximately 5×10^{-4}, for uncorrelated errors with a bit error ratio (BER) of 10^{-9}. For packets of 9,180 Byte (i.e., the default MTU size defined in [RFC1626]), the resulting packet error rate is approximately 1×10^{-4}.

The transmission service data-unit (e.g., IP packet) will be segmented into multiple cells. Therefore, a complete packet is assumed to be in error even when only one cell within the received packet is missing. For example, the default MTU size defined in [RFC1626] is 9,180 Byte, corresponding to about 160 cells. The maximum size of an AAL5 CPCS-PDU corresponds to about 1,330 cells (=64KByte). The approximate packet error probability due to the cell loss was provided in [95-0150]. For example, for a 64 KByte data-unit size (i.e., the maximum data-unit size of AAL 5), the probability that the received data-unit is erroneous is about 1.3×10^{-3}, when cell loss ratio (CLR) is 10^{-6}. For a 9,180 Byte (i.e., the default MTU size defined in [RFC1626]), the error probability is about 2×10^{-4}.

- Number of receivers
  Since the ATM networks are switched oriented networks, the actual packet error/loss ratio for the sender process will approximately increase linearly according to the increase of the number of receivers. This is because, even when only one link (or node) within the whole of multicast tree generate packet error due to cell loss or BER, it must be treated for the sender process that the transmitted packet is errored somewhere in the multicast tree. For example, for 10^4 receivers, the actual packet error quality for the sender process will be about 10^{-3}, even when the packet error quality in every link (or node) is 10^{-7} (i.e., cell loss/error ratio is 10^{-9} with 100 cell size of packet).

This issue, i.e., increase of the observed packet loss/error probability, is a serious problem in the ATM network.

4. Management of multicast members
   When we have a large number of receivers associated with a multicast service, it will be hard to manage all receivers at the sender. Therefore, it is advantageous for a large scale multicast to apply a receiver oriented control for joining/quitting a multicast service, as well as to apply the distributed receiver management policy (e.g., in [RSVP]). The receiver will send a request message to the local multicast service management entity (may be located at the branching point in the multicast tree), in order to join to (or quit from) the multicast service.

5. Receiver's unreliable behavior
   Another issue for a large scale multicast service is the unreliable behavior of multicast

receivers. This is very serious for a reliable multicast service. When the number of receivers in the given multicast connection is large, some receivers will suddenly quit from a multicast service without any indication to the multicast service management entity. By this unreliable behavior by the receiver, the sender (or the intermediate management entities) may interpret the receiver did not correctly receive the packet(s) with positive ACK policy, even though it quitted from the multicast service.

6. Service quality assurance
   For a geographically large scale reliable multicast service, it will be hard to provide the sufficient data-link/network service quality along the multicast tree, when we use the conventional best effort service. This is because, for the large scale multicast service, the multicast tree has the large number of branching points and links. Even when only one of entity within a large multicast tree offers a poor service quality, the service quality (e.g., throughput) is degraded due to the entity offering a poor service quality.

   Since the conventional researches were not on large scale "error-free" multicast service, this issue has not considered.

In Sections 6 and 7, the architecture to provide a large scale error-free multicast service is proposed and evaluated.

4 ATM Network Architecture using Cell Switch Router

In this section, the proposed ATM network architecture using CSR (Cell Switch Router) is proposed. The issues, that CSR architecture will solve, are providing high throughput routers to improve packet forwarding capability in the routers, while providing QOS-ed IP packet delivery service.

To operate a network with a lot of hosts and some redundant paths, the network should be divided into several data-link network segments, which are interconnected by routers. It would be recommended that each data-link network segment should accommodate only one IP subnet (a network layer segment). T his is the current network architecture in the Internet. And, both from the distributed application's point of view and from the network management's point of view, the data-link network segment (including ATM-based subnet) should have a broadcast capability. For example, many existing routing protocols require the broadcast (or multicast) channel to all nodes in the link (i.e., in the IP subnet).

Figure 4-1 shows one example of internet model. ATM-based subnet is composed of switch nodes, and the interface with the adjacent network segment is UNI (User Network Interface), even when the adjacent network is an ATM network. Data-link network segments (i.e., IP subnets) are interconnected by routers, and each data-link network segment corresponds to an IP subnet. Regarding ATM networks, since the interconnecting interface is UNI, each ATM-based subnet can operate independently.

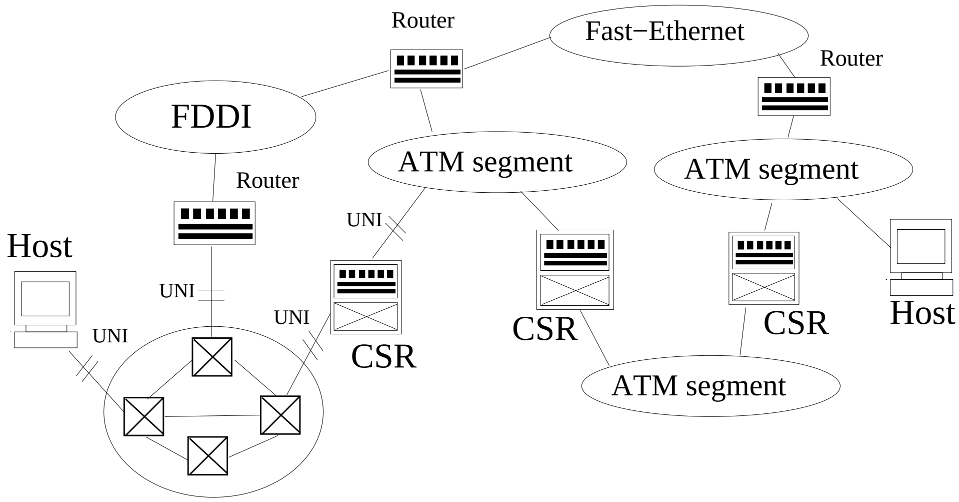


Figure 4-1. Large Scale Internet with Cell Switch Router

4.1 Benefits of Subnet Model

The subnet model should be kept, even when we use the LIS model. The physical or logical IP subnets are interconnected through routers.

The following are the benefits through keeping a subnet model, compared to the large cloud datalink without router or to the NHRP model. Of course, the large cloud datalink networks without router can be interconnected by the routers.

1. Static and Dynamic Route Provision
   P-NNI protocol [PNNI] will provide a capability to establish an end-to-end direct ATM-VCC over the large ATM cloud, even when source host and destination host belong to the different LIS. The established end-to-end ATM-VCC will have a certain QOS, but it is static (i.e., hard-state) route that is not dynamically changed during a session. In other words, with P-NNI, the route arrangement will be performed session-by-session, rather than packet-by-packet or burst-by-burst during a session according to the network status.

   On the contrary, when we use the subnet model (i.e., interconnecting (logical/physical) IP subnets by routers), the route can be dynamically changeable during a session, as well as session-by-session basis. Router can locally pick up the alternate route packet-by-packet, according to the network status. Of course, it is easy to provide static (hard-state) route that is not changed during a session, as well as a dynamic (soft-state) route, by means of the hard-state maintenance by the intermediate routers. As a result, the subnet model can easily provide both static (hard-state) route and dynamic (soft-state) route. Moreover, basically, which mode, static or dynamic route, is applied to each IP flow depends on each router's local decision. Here, when the use of static route is explicitly indicated by some protocol (e.g., by the protocol discussed below), the routers should provide the static route.

2. Change of QOS/flow-spec/service-class during a session
   Integrated Service Model [Shenk1][Shenk2][Shenk3] requires the capability of dynamic modification associated with service class, QOS (i.e., R-spec) and flow-spec (i.e., T-spec) during a session. However, in general, current connection oriented datalink platforms (including ATM) do not have a QOS re-negotiation capability or a traffic descriptor re-negotiation capability. Here, the future ATM signaling (i.e., Q.2931) may have this capability, but it may not include the capability to re-negotiate the service class (e.g., changing from ABR to VBR service class).

   When we use a subnet model, how to map/aggregate the IP packet flows to/into the datalink pipes (e.g., ATM-VCCs) is basically a matter of each router's local decision. Therefore, it would be easy to change R-spec, T-spec and service class during a session through the changing of mapping/aggregation between IP flows and ATM-VCCs.

3. IP packet flow mapping and aggregation
   How to map and aggregate the IP packet flows with/into the ATM-VCCs is basically a matter of router's local decision. Some routers will use a dedicated datalink pipe (i.e., ATM-VCC) for each IP packet flow. But, some other routers will use a dedicated datalink pipe for the IP flows whose requiring QOS class are the same, i.e., IP packet flow aggregation.

   There would be many criteria to decide how the IP packet flows are mapped/aggregated with/into ATM-VCCs. For example, in the LAN environment, one to one mapping between IP packet flow and ATM-VCC would be appropriate, but, in the WAN environment, it would be better that some IP packet flows are aggregated into a single ATM-VCC because ATM-VCC establishment cost may be expensive. Since ATM-VCC is always terminated at the router for a subnet model, even though IP packet flow is actually forwarded by cell-relaying cut-thru as discussed in the next subsection, the flexible use of ATM-VCC will be provided.

4. Scaleable soft-state multicast [RSVP]
RSVP is soft-state protocol, that is, *PATH* and *RESV* messages must be periodically exchanged among the hosts and routers. Up-stream routers and source host(s) periodically advertise *PATH* messages toward down-stream routers or destination hosts. According to the reception of *PATH* message, routers or destination hosts send back *RESV* message to the up-stream routers or source host(s). Therefore, the number of *RESV* messages that branch point node (i.e., intermediate router) or source host receives corresponds to the number of leaves within the corresponding multicast sub-tree. Obviously, when the number of leaves within a multicast sub-tree is large, it would not be possible to deal with many *RESV* messages at the up-stream intermediate router or at the source host.

   ATM will be able to provide large scale multicast datalink service over the large ATM cloud without routers [GJA]. Since the ATM (datalink) cloud will be very large scale, large (or sometimes huge) number of down-stream nodes (routers/hosts) will be directly connected to an up-stream node (router/host). In other words, ATM cloud can provide a multicast-tree that has a large fan-out as a datalink connection. In this case, all down-stream nodes periodically send a *RESV* message to the up-steam node, according to the reception of *PATH* message. Then, a large number (or huge) number of *RESV* messages will be sent back to the up-stream node, and the up-stream node may not be able to handle all the *RESV* messages. As a result, in order to provide a scaleable soft-state multicast service, we should not use large scale datalink level multicast services on a large ATM cloud.

   In order to provide s scaleable soft-state multicast service over the large ATM cloud, we should keep a subnet model. When we keep a subnet model, the multicast sub-trees will be only associated with a single LIS. And, when the scale of the LIS is reasonably small, we can provide a scaleable soft-state multicast service even over the large ATM cloud.

5. Firewall function (e.g., packet filterring)
When we need a packet-by-packet security checking for the IP packet transmission beyond the IP subnet, we must keep a subnet model. However, some domain having a multiple IP subnets may not need packet-by-packet security check for the communication within the domain (i.e., NBMA subnet domain). In this case, we need not keep a subnet model. However, when some IP subnet change a policy to require packet-by-packet security checking, it would be complicated for NHRP model to apply these requirement. As a result, keeping a subnet model seems to be simple and flexible for many types of security policies, that basically depends on the IP subnet.

When the physical IP subnets (i.e., IP subnet and datalink segment are geographically identical) are interconnected by routers, it corresponds to the CATENET model [RFC791]. With the CATENET model, the migration of the datalink platform will be easy. Since the physical datalink segment is physically partitioned by the router, we can easily replace the datalink to the other one that may use the different technology.

As a summary, we propose that we should keep a subnet model, even for the large ATM cloud platform. And, also, it is recommended to keep a CATENET model. However, when we keep a subnet model or NBMA subnet domain (with NHRP) is reasonably small, we can obtain the benefits discussed above.

4.2 Architecture of Cell Switch Router (CSR)

In this subsection, the Cell Switch Router (CSR) is introduced. Though all the routers in the ATM networks need not be CSR, CSR will provide scaleable and high speed IP forwarding capability.

CSR provides a cell switching capability in addition to the conventional IP packet switching. By the use of cell switching capability, high throughput IP packet transmission can be provided, even when the packet communication goes beyond IP subnet passing through router(s). CSR can forward some IP packet flows while bypassing the packet assembly/reassembly and IP header processing, i.e., the cut-thru packet forwarding. In the cut-thru packet forwarding, IP packet is forwarded based on VPI/VCI and IP header is not examined at the CSR. On the contrary, the conventional hop-by-hop IP packet forwarding based on IP header can be also performed for the other IP packet flows. Therefore, we can see the cell switching capability as a part of packet scheduler or classifier defined in the integrated service model [RFC1633].

The key functionalities of a router interconnecting ATM networks are as follows.

- Hop-by-hop packet forwarding
  The packets, that should be forwarded with usual IP packet processing, are forwarded with hop-by-hop packet forwarding. The packets forwarded with hop-by-hop packet forwarding are transferred through the default-VC, that is established when CSR turns on. In the hop-by-hop packet forwarding, the IP packet is reassembled into the IP packet to analyze the IP header and TCP/UDP header. The CSR decides the next hop node (i.e., router or host) based on destination IP address in the received IP packet, according to the routing information indicated by routing protocol (e.g., OSPF).

- Cut-thru packet forwarding path establishment and cell relaying
  Cells belonging to the packets, that should and could be forwarded without usual IP processing at the CSR, are relayed as the ATM layer's function. The packet forwarding by cell -relaying without IP processing is called as cut-thru packet forwarding, since the IP processing is cut-thru at the CSR. When the CSR receives IP packet, that should and could be forwarded by cut-thru mode, the CSR tries to establish a dedicated-VC between the neighbor nodes (e.g., CSR). When the ingress dedicated-VC and the egress dedicated-VC, regarding a IP packet flow to be cut-thru, are established, these two dedicated-VCs are coupled at the CSR. After the coupling of these two dedicated-VCs,

the cells belonging to this IP packet flow are forwarded in cut-thru mode, i.e., IP packets are forwarded without IP processing at the CSR. Since CSR independently establishes egress and ingress ATM-VCCs, the assignment of ingress and egress VPI/VCI value is independent for each dedicated-VC. Therefore, the CSR has the VPI/VCI translation functionality.

Since IP processing can be bypassed by ATM layer cell relaying at the CSR, the transmission latency to pass through the CSR can be sufficiently small, which is same as an ATM switch node.

- Establishment of dedicated-VC
  A dedicated-VC used for cut-thru packet forwarding is established using a network layer address (i.e., IP address) rather than using an ATM address. An ATM address is used only for the establishment of dedicated-VC within a single data-link segment (i.e., ATM-based subnet) through Q.2931 protocol. When a CSR receives the trigger packet (e.g., RESV packet for RSVP) to establish a dedicated-VC, it selects the appropriate next hop CSR (or the destination end-host) and requests the establishment of (dedicated) VC toward the next hop CSR (or toward the destination end-host). Here, in order to resolve the ATM address of the next hop CSR (and of the destination end-host), the CSR maintains two kinds of mapping information ; (1) between the next hop CSR identifier and the IP address of the next hop CSR, and (2) between the IP address of the next hop CSR (and of the destination end-host) and the ATM address of the next hop router (and of the destination end-host). The former information is obtained by the routing protocol, and the later information is usually maintained at the CSR as a routing cache. When the later information regarding the target entity (the next hop CSR or the destination end-host) is missed by the routing cache, an ARP (Address Resolution Protocol) procedure [RFC1577] will be performed to resolve the ATM address of target entity. The next hop CSR performs the same procedure.

  During a (dedicated) VC establishment procedure, a VPI/VCI value for the established (dedicated) VC is indicated by the network. Since the (dedicated) VC establishment requests are issued to every ATM-based subnet along the selected packet forwarding path, each (dedicated) VC assigns the different VPI/VCI code point for each ATM-VCC. Finally, a CSR will establish a VPI/VCI mapping table using the VPI/VCI values assigned to these ATM-VCCs, in order to concatenate ATM-VCCs at the CSR.

- Security Functionality (if necessary)
  Since every cell (or could be say packet) passes through the CSR, the segmentation of subnet can be fulfilled as well as the current Internet architecture. For instances, the proposed router can equip security functionalities, e.g., packet filtering in fire-wall functions. The proposed router (i.e., CSR) can also equip other security functionalities, that is, unauthorized packets are filtered just as in the case of usual firewalls and the unauthorized reservation request packets for cell relaying are rejected.

As discussed above, the proposed router, CSR, is different from an ATM switch, because an ATM switch does not have IP forwarding capability. And, also, an ATM switch establishes an ATM-VCC based on ATM addresses with Q.2931 protocol, while a router handles IP packet flows based on IP address.

Routers are, in general, network entities which inter-connect several data-link segments. Routers exchange routing information and maintain routing information tables to forward received packets. Logically, each packet is relayed to the optimal interface by looking up the routing tables. But, actually, routers don't have to consult with the routing tables everytime they receive a packet. As an implementation, routers can have an internal cache or a bypass to minimize the packet routing delay of complex table looking-up. Likewise, routers don't have to consult with the routing table everytime they receive a cell. In an implementations, it is possible to have an internal bypass to remove the packet reconstruction delay and packet routing delay of complex table looking-up.

Through bypassing the conventional IP forwarding process using cell-relaying, we could dramatically reduce both the IP packet processing delay and the queuing delay at the router. The IP packet processing delay is not reduced for a non-cut-thru IP packet flow, but is reduced for a cut-thru IP packet flow because the complex table looking-up at the router can be omitted. The queuing delay at the router for a non-bypassed IP packet flow is also reduced, since the total number of IP packets dealt with by each router (especially for core intermediate routers) will be reduced. In the current Internet, the observed IP packet delay at the routers was about from few *msec* to few *sec*, when we fulfill a traceroute. On the contrary, the expected cell delay at ATM cell forwarding module could be from few μsec to *msec*.

Whether an IP packet flow is handled by the cut-thru packet forwarding or is handled by the conventional hop-by-hop packet forwarding will be basically determined by every router's local decision. However, the capability, that the node (router or host) can explicitly request a cell-relaying at all the routers along the routing path of the packet flow, should be provided (e.g., in order to provide a IP packet transmission with IP header encryption discussed below). The IP packet flow that can be forwarded using the cut-thru packet forwarding will be the flow whose required bandwidth/QOS and traffic pattern are known in advance. For example, potentially, the IP packet flow belonging to guaranteed service and predictive service (and maybe controlled delay service) would be able to be forwarded by the cut-thru packet forwarding at CSR, since R-spec and T-spec of the packet flow are specified [Shenk1] [Shenk2][Shenk3].

By using the cut-thru packet forwarding capabilities, the CSR can concatenate two dedicated-VCs (cell flows). If all the CSRs along the route between the source and destination nodes forward an IP packet with the cut-thru packet forwarding, a seamless cell-relaying pipe will be provided. This concatenated (dedicated) VCs is referred to the ATM bypass-pipe or to cut-thru pipe, in this paper. The ATM cut-thru pipe does not have any conventional IP header processing point along it's path.

When we can provide the cut-thru packet forwarding by the CSR, we can obtain the following three additional benefits to the benefits by means of keeping a subnet model.

- High Throughput and Small Latent Packet Forwarding
  By bypassing the IP packet header processing to forward the IP packets, we could obtain the high performance IP packet forwarding. Logically, each IP packet is for-

warded to the optimal interface by looking up the complex IP routing tables. However, actually, CSR does not have to consult with the IP routing table whenever the received IP packet is forwarded using the cut-thru packet forwarding capability. CSR forwards the IP packets by the cell-relaying module whose cell-relaying table is reflected on the IP routing table generated by IP (or some network layer) routing protocol. Since an ATM technology has been developed by the hardware processing, rather than by the software processing, the ATM switch can operate at larger interface speed (e.g., at few Gbps even now).

- High Performance Firewall Function
  When we introduce a resource reservation oriented flow (e.g., ST-II flow), some flows will require security check only when the packet flow path is established (i.e., not packet-by-packet security check). In this case, the cut-thru packet forwarding provides a high performance IP packet transmission.

- IP Forwarding with IP Address Encryption
  For the IP packet flow that explicitly request all the routers along the routing path to carry out cut-thru packet forwarding, the IP addresses in the packet header is basically meaningless, from the view point of IP packet forwarding. When IP packets are directly forwarded by the ATM cell-relaying, there is no need to examine every IP packet header at the CSR. Therefore, we could encrypt the IP header field, when we can use the end-to-end ATM bypass pipe.

  One example is the packet transmission between the certain sites, i.e., the end-to-end ATM bypass pipe is provided between the border routers of each sites. In some cases, we may not want to allow the intermediate routers to be able to recognize detailed information on each IP packet's header (e.g., source and destination IP addresses). By using the scheme mentioned above, the intermediate routers can forward IP packets (cells) via VPI/VCI information without an IP header examination. Therefore, the IP packet header information including source and destination IP addresses can be encrypted.

In order to provide the cut-thru packet forwarding by the CSR, the additional protocol (i.e., FANP) messages discussed below must be defined. Neighbor routers must exchange the information how the IP packet flows are mapped/aggregated with/into datalink pipe (i.e., dedicated-VC).

CSRs will be able to use the equipments that directly interconnects multiple ATM network segments. In this configuration, ATM-based IP subnets are internetworked by the CSRs. The other configuration will be that ATM (or non-ATM) network segments are internetworked by CSRs through the WAN datalink pipes. The WAN datalink pipes can be either the conventional digital links (e.g., OC-3 or DS3 SONET pipes) or the connection oriented datalink pipes (e.g., ATM connections or Frame Relay connections). These configuration will be appropriate for the early stage while an ATM technology starts to be adopted in the LAN (e.g., as a campus backbone network technology). Moreover, these configuration will be adopted, even if the ATM technology is widely applied in the WAN (i.e., public networks). In all configurations, CSR can provide high performance IP packet forwarding capability, even though the IP packet is theoretically forwarded based on the network layer protocol (e.g., IP routing).

4.3 Scalable Implementation of CSR

In this subsection, scaleable CSR implementation examples and the CSR networking to scale packet processing capacity are discussed.

Figure 4-2 shows an example of implementation image of CSR. CSR has both cell switching fabric and packet switching fabric. Incoming cells are switched by the cell switching fabric based on the VPI/VCI value. The cell flow forwarded by the cut-thru packet forwarding is directly switched to the appropriate output port after the translation of VPI/VCI value. On the contrary, the cells belonging to the packet, that should be forwarded by the conventional hop-by-hop packet forwarding, are transferred to the IP switching fabric to be examined IP packet header. After the examination of IP header, the cells are transferred to the appropriate output port with the corresponding appropriate VPI/VCI value. Here, the implementation of packet switching fabric and of cell switching fabric will be implementor's decision. Both cell and packet switching fabric should be scaleable architecture (i.e., number of interface port should be scaleable). In order to provide multiple QOS by the cell switching fabric, cell switching fabric would have some cell scheduling functions (e.g., shaping or priority control based on VPI/VCI).

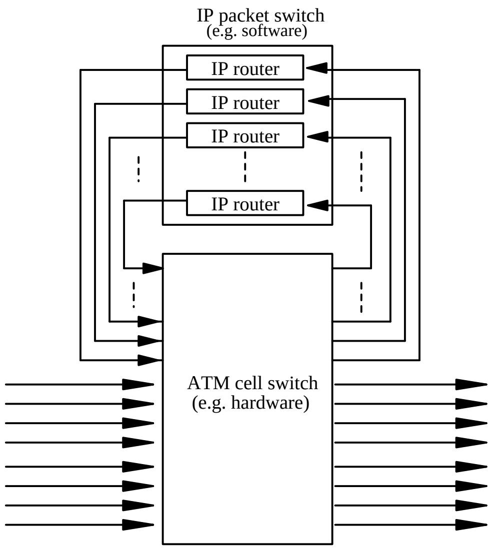


Figure 4-2. Scaleable CSR Implementation Example

In order to scale the IP packet processing capability, multiple IP processing units are implemented in the IP packet switch fabric. The IP packet switch fabric will be able to implement using the multi-processor technlogy applied to the high-end work-stations. However, since IP packet switch fabric has multiple ingress ATM interfaces, some mechanism (e.g., new routing mechanism within the cell switching fabric) has to be developed to distribute the IP packets to each interface.

An alternative scaleable router architecture, shown in figure 4-3, is that each interface port has the IP router module, and these IP router modules are interconnected by cell switch. In this case, the cell switch is used for the packet transmission module among the IP router modules. This architecture, called as RICS (router interconnection by cell switch) in this paper, is also scaleable associated with the number of interface ports (i.e., the total amount of IP packets handled by the router system). Though the RICS model has a scalability associated with the number of interface ports, it may not have the scalability associated with the interface speed of each interface port. The IP packet processing speed at IP router modules at the interface port may become insufficient, when IP packets are handled by software processing. This means that the processing speed of IP router module at each interface port may become a system bottle-neck. In order to solve this concerning, each IP router module at interface port should have both a cut-thru (i.e., cell-relaying with hardware processing) and a hop-by-hop (i.e., packet-relaying with software processing) packet forwarding functionalities. In this case, we can see the IP router module at the interface port as the CSR having two interface ports.

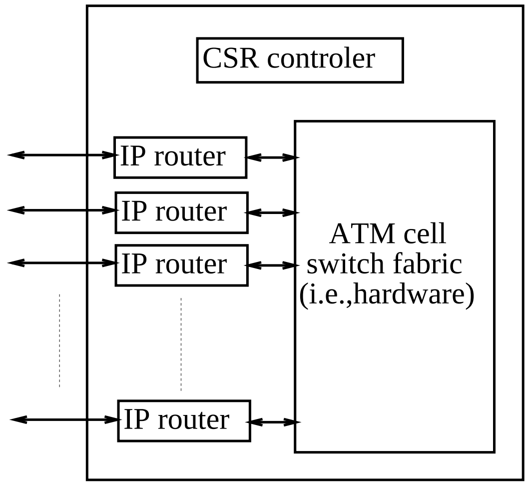


Figure 4-3. RISC architecture model

Yet, another approach to scale the total IP packet processing capability of the CSR system is to interconnecting multiple CSRs, say as the aggregated CSR. By configuring the aggregated CSRs, we can easily achieve larger IP processing capability. Minor problem of this approach would be the physical space that the CSR system requires. The routing among the CSRs could be performed some IP routing protocol (e.g., OSPF), which is effective in the aggregated CSR. There would be a concerning to waste the IP address due to having multiple nodes (i.e., CSRs) in the aggregated CSR. However, we need not waste the IP address space by means of unnumbering technique commonly used for point-to-point links.

4.4 ATM-VCC Management Architecture

In this subsection, the ATM-VCC (i.e., default-VC and dedicated-VC) configuration is discussed.

Figure 4-4 shows an example of the configuration of ATM-VCCs. Between nodes, two types of ATM-VCCs are defined. One is a default-VC and the other is a dedicated-VC. The default-VC transfers both control packets (e.g., the control packets of FANP described in the following subsection) and hop-by-hop based user packets. The IP packets transferred through the default-VC always experience usual IP processing. On the contrary, the dedicated-VC transfers only the user packets that are forwarded by the cut-thru packet forwarding mode at the CSR. Here, the cell flow arrived at a CSR through a dedicated-VC may be transferred by the default-VC to forward the IP packet to the adjacent CSR, and vice versa.

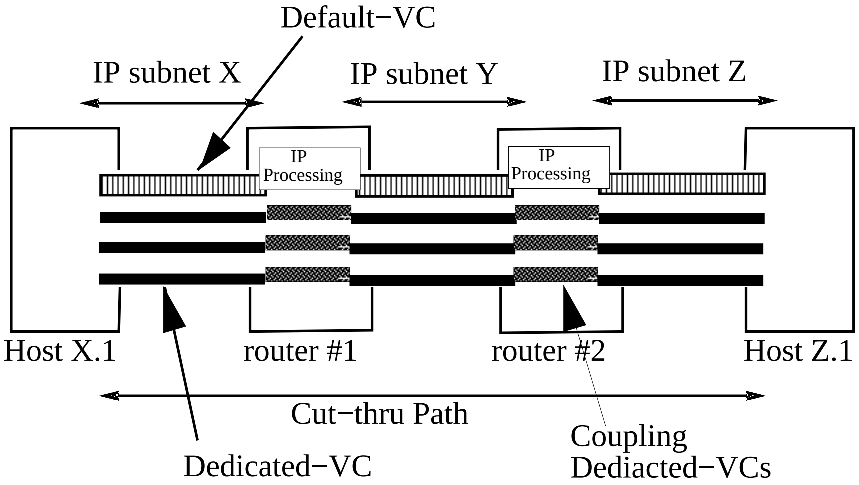


Figure 4-4. ATM-VCC Configuration Example for CSR

The control messages to manage the user packet flows using the dedicated-VCs are transferred through the hop-by-hop based default-VC. Therefore, the management of ATM cut-thru path can be hop-by-hop, and the IP header in the concatenated dedicated-VCs at the CSR need not be examined. By the control packets exchanged among the CSRs, hard-state, soft-state and stateless IP packet forwarding routes are provided. The detail of the protocols and messages for control and management of the user packet flows are not discussed in this paper.

- Hard-state route :
  In hard-state route, the route of IP packet flow is not changed during a session. Therefore, the control packet may be exchanged at route setup phase and route termination phase, and any control packet may not be exchanged for the purpose of state refreshment. The cut-thru path will be established during a route setup phase, and it will not be changed during a session.

- Soft-state route :
  In soft-state route, the route of IP packet flow can be dynamically changed during a session. Therefore, the control packet are continuously exchanged among the CSRs for the purpose of state refreshment, using the default-VC. The cut-thru path will be established during a route setup phase for the connection oriented IP packet flow, and it will be dynamically changed during a session according to the status of network. When the appropriate next hop router indicated by the routing protocol is changed, CSR modifies the cell switch parameter to modify the (soft-state) IP packet flow to the appropriate dedicated-VC leading to the updated neighbor CSR. FANP, discussed below, indicates to the neighbor CSR what type of IP packet flows are conveyed in the dedicated-VC, after the route is actually modified. Therefore, the cut-thru IP packet forwarding will be maintained, even when the route of IP packet flow is dynamically changed.

- Stateless route :
  In stateless route, the route of IP packet can be dynamically changed packet-by-packet. Usually, the cut-thru path will not be established for the stateless route.

Since the management of ATM cut-thru pipes is hop-by-hop, dynamic route changing (i.e., soft-state path) can be provided for the ATM cut-thru pipes, while achieving a high performance IP packet forwarding.

In the ATM cut-thru pipe, TTL (Time To Live) in the packet header is not decremented at the CSR. TTL is the important field to protect routing loop and to define the service scope (e.g., multicast scope). Fortunately, the IP header of control packets, which are transferred through the default-VC, are examined at every CSR. Therefore, the functionalities provided by TTL field can be performed by the control packets, even though the TTL of user packets transferred through the dedicated-VC is not decremented at CSR.

Management of the dedicated-VCs, i.e., establishment, tearing down of the dedicated-VCs, and management of the mapping between IP packet flow and ATM-VCC are matters of every CSR's local decision. Q.2931 will be used to establish or tear-down the dedicated-VCs among the CSRs. The default-VC could be either SVC (Switched VC) or PVC (Permanent VC). Transmission links between the CSRs may be high speed digital links (e.g., OC-3 or DS3 SONET pipe), instead of ATM. This case would be possible when the public network can not provide ATM link between the sites across the public network. In this case, the CSRs interconnected by the high speed digital link may not have to use Q.2931 to establish/tear-down the dedicated-VCs between the CSRs.

4.5 Flow Attribute Notification Protocol : FANP

The purpose of FANP [RFC2129] is establishment and maintenance (i.e., refreshment and flush the state) of dedicated-VCs. Using the FANP, the up-stream node indicates to the neighbor down-stream node how the IP packet flows are aggregated into the datalink pipe (i.e., ATM-VCC for ATM networks). By means of this procedure, the corresponding dedicated-VC will be established. Here, the information exchanged between the router by the FAN is not only useful for CSR, but also is generally useful for the routers attached to the connection oriented datalink platform (e.g., Frame Relay and TDM circuit network). FANP has fundamentally two functionalities ; one is binding the IP flow and dedicated-VC, and the other is negotiating the VCID (Virtual Connection Identifire) negotiation and datalink-level label (e.g., VPI/VCI).

4.5.1 4.5.1 Binding IP Flow and Dedicated-VC

By the flow aggregating information exchanged by the FANP, the router could optimize the packet handling. For the CSR, when the flow aggregation is performed based on the destination IP address, the packet forwarding will be able to be optimized using a cut-thru packet forwarding capability. In general, the FANP indicates the following information associated with how the IP packet flows are aggregated to the down-stream neighbor router.

- Destination Address
  The IP packet flows that have a common destination address (i.e., individual destination address or destination address prefix) could be aggregated into a single datalink pipe (e.g., ATM-VCC), even through the source IP addresses of the IP packet flows are not identical.

- QOS Class
  The IP packet flows that have a common QOS class could be aggregated into a single datalink pipe (e.g., ATM-VCC). Scheduling of IP packet processing at the router may be optimized based on QOS class.

- Flow Identifier
  The IP packet flow that have a common flow identifier (e.g., IPv6 flow-id) could be aggregated into a single datalink pipe (e.g., ATM-VCC). And, the IP packet flow that is identified by the flow identifier (e.g., RSVP's packet flow) may ought to mapped to a single ATM-VCC.

Above three criteria are the orthogonal parameters. Therefore, in the actual case, the above parameters will be combined to aggregate the IP packet flows. The packet flow aggregation policy based on QOS criteria will not appropriate for CSR. Both the packet flow aggregation based on destination address and flow identifier will be appropriate for CSR architecture.

4.5.2 VCID Negotiation Procedure

As discussed in section 3.1.4, LSTL architecture models can not generally work over the ATM platform. Since the LSTL architecture models directly indicate the label (e.g., VPI/VCI value for cell-relaying) information to the adjacent LSTL nodes, the LSTL architecture model can not work over the data-link platform(s), called as label swapping datalink (e.g., ATM platform). Here, in the label swapping datalink platform, the data-link label (e.g., VPI/VCI in the ATM platform) is re-written at every switching node, such as ATM switch. Since the data-link label is re-written at every nodes in the label swapping platform, the

LSTL architecture model can not work over the label swapping data-link platform, such as ATM platform.

In order to let work the label switching paradigm over the label swapping data-link platform(s), the FANP has the VCID (Virtual Connection IDentifier) negotiation procedure [RFC2129]. In the VCID negotiation procedure, adjacent nodes share the VCID information, that could be of local significance between the adjacent nodes, corresponding to the Dedicated-VC. Basically, there are two ways to perform the VCID negotiation procedure. One is through the in-band signaling, and the other is through out-band signaling. The mechanism defined in [RFC2129] uses the in-band signaling for VCID negotiation procedure.

- In-band VCID negotiation procedure

  VCID negotiation procedure is performed using the Dedicated-VC, where the cut-thru packet actually use. In some way, e.g., ARP message format in [RFC2129] mechanism, the CSR has to be able to recognize the packets to perform the VCID negotiation procedure. The CSR node (CSR1) indicates an VCID for the Dedicated-VC through the corresponding Dedicated-VC to the adjacent CSR (CSR2). When CSR2 receives the VCID negotiation packet, CSR2 sends the acknowledgment packet including the VCID information to the CSR1. Through this procedure, the adjacent CSRs (i.e., CSR1 and CSR2) can map the VCID value, that is unique value between the adjacent nodes, and the local VPI/VCI value of the corresponding Dedicated-VC.

- Out-band VCID negotiation procedure

  VCID negotiation procedure is performed using the Default-VC, where the hop-by-hop forwarding packets use. In some way, e.g., end-to-end signaling message in Q.2931/UNI3.1, the CSR has to be able to recognize the VCID for the corresponding Dedicated-VC, that would locally unique between the adjacent CSRs. The CSR node (CSR1) indicates an VCID for the Dedicated-VC through the Default-VC to the adjacent CSR (CSR2). When CSR2 receives the VCID negotiation packet, CSR2 sends the acknowledgment packet including the VCID information to the CSR1, through the Default-VC. Through this procedure, the adjacent CSRs (i.e., CSR1 and CSR2) can map the VCID value, that is unique value between the adjacent nodes, and the local VPI/VCI value of the corresponding Dedicated-VC.

Which mechanism should be applied to depends on the protocol specification of data-link technology. For the ATM platform, that this paper is mainly discusses, both mechanisms can be applied to.

4.6 Packet Forwarding for Connection Oriented IP Flow

Currently, the connection oriented IP transport flow (i.e., a network layer connection oriented service) has been discussed in IETF, e.g., [RSVP][IPv6]. The goal of connection oriented communications is to provide an end-to-end IP transport. Such a transport flow can have a certain QOS. In such IP transport flow, there is an admission policy at the network layer level, that is a similar concept to CAC (Connection Admission Control) in ATM networks. When the IP transport flow set up request is admitted, the resource for the transport flow is reserved at corresponding routers. Here, in order to find out the next network (or can be said the router) that the IP transport flow should be routed to, some kind of routing protocol (e.g., similar to OSPF) is executed among the routers.

Unlike other models of IP over ATM, the architecture proposed in this section assumes QOS-ed communications over not only ATM but also the other types of platforms, using the concept of connection oriented IP transport. The connection oriented IP transport is mapped into data-link layer connection (ATM-VCC) within the ATM networks.

Connection oriented data-link networks (e.g., ATM) can provide a QOS-ed connection between all the data-link SAPs (Service Access Points) within the same data-link network segment through a connection set up procedure . A communication beyond data-link network segment is performed by a QOS-ed IP transport flow to the router. When the data-link network segment is connection oriented, a router or a source end-host establishes a QOS-ed data-link connection (e.g., ATM-VCC) toward the next router. The next router performs the same procedure, until the IP transport flow setting up message reaches at the destination end-host.

After the IP transport flow is established, IP packets are forwarded through the QOS-ed pipe. A unique identifier, called flow-ID (flow identifier), could be used along the IP transport flow [IPv6][RSVP]. Within the ATM network, VPI/VCI can be used instead of flow-ID defined in the IP layer.

The IP packet forwarding procedure is as follows.

1. A packet arrives at router.
2. The flow-ID and the IP address (source/destination) are checked and the next hop router is determined.
3. TTL (Time To Live) is decremented.
4. Unless this router is the destination end-host, the packet is forwarded.

This procedure is not ATM-specific and it is applicable to all other platforms. Three cases are discussed below.

- ATM-segment → ATM-segment
  The router has a mapping table between the ingress VPI/CPI and the egress VPI/VCI for each cell flow according to the appropriate IP transport flow. Then, it is unnecessary to examine the flow-ID and IP address in step (2) : "Re-direction" of the ingress and the egress data-link connection. Here, the mapping table is established during the connection set up procedure.

  1. A cell arrives at a router

2. VPI/VCI of the ingress cell is examined and the next hop VPI/VCI of the egress cell is determined (maybe by hardware)
3. TTL in the IP header may not be decremented
4. Unless this router is the destination end-host, the cell is forwarded.

- ATM-segment → Other-LAN
  A router will just reassemble a IP packet to forward it to the appropriate router or end-host.

  1. A cell arrives at a router
  2. An IP packet is reassembled (e.g., AAL5)
  3. The flow-ID and the IP address (source/destination) is examined and the n ext hop router is determined.
  4. TTL in the IP header is decremented
  5. Unless this router is the destination end-host, the packet is forwarded.

- Other-LAN → ATM-segment
  A router has a mapping table between the ingress flow-ID/IP-address and the egress VPI/VCI for each IP transport flow. Here, again, the mapping table is established during the connection set up procedure. Here, the TTL decrement value in step 3 may not be one. This is because, regarding the IP packets belonging to the cell-relaying cut-thru pipes, the TTL in the IP packet header is not dealt with at the intermediate router(s).

  1. A packet arrives at a router
  2. The flow-ID and the IP address (source/destination) is examined and the n ext hop VPI/VCI of the egress cells is determined.
  3. TTL in the IP header is decremented (may not be one)

When ATM end-hosts communicate purely over ATM segments, a seamless end-to-end cell-relaying virtual channel can be established between them. It should be noted that, even though many ATM-VCCs (data-link connections; dedicated-VCs) are used at the data-link layer, the seamless connection over ATM segments is at a network layer connection (not a data-link layer connection).

4.7 Packet Forwarding of Connectionless IP Flow

There is no difference between the connectionless IP packet transmission in the proposed architecture and that of existing IP forwarding. Within the data-link network segment, end-hosts and routers exchange IP packets using data-link connections (e.g., ATM-VCCs). IP packets beyond the data-link segment are sent to an appropriate router from the source end-host. Routers (i.e., CSRs) exchange routing information and forward IP packets to the appropriate router. The following discussion is regarding the network that has only ATM segments. Regarding toward the other types of platform from an ATM network, ATM-VCC shall be always terminated at the router to execute the conventional IP forwarding process. On the contrary, regarding toward ATM network from the other types of platform, you can see the router as the end-host in the following discussion.

4.7.1 Hop-by-hop ATM-VCC Cacheing for Active IP Flow

The operation of hop-by-hop ATM-VCC cacheing for active IP flow is shown in figure 4-5. When the first IP packet (of the session or after the idle period) arrives at a CSR (attaching to ATM), the next hop router is decided using the IP header, i.e., the routing decision will be done based on the IP address (and the flow-ID in IPv6). After the next router is decided, the received IP packet is forwarded through the default-VC toward the next hop router. CSR analyzes TCP/UDP header, as well as IP header. When the received IP packet belongs to the IP flow, that is likely to be long life session, a dedicated-VC is picked up or is established for the IP packet flow and CSR establishes the mapping table between the incoming VPI/VCI (or some flow-ID in the data-link layer)and the outgoing VPI/VCI. When the packet comes from the conventional platform (e.g., Ethernet/FDDI), the router always examines the IP header in the case of IPv4 (in IPv6, the flow-ID could be mapped to the outgoing VPI/VCI). When the IP packet comes from an ATM network, the successive IP packet is forwarded based on the incoming VPI/VCI value, without any usual IP processing.

If the activity of the IP packet flow becomes low, the mapping information (i.e., the cached information) is pushed out. This means that the pushed out IP packet flow from the cache must be examined by the usual IP processing (looking at the IP header) at the CSR, i.e., transferred through the default-VC. As a result, IP packet flow with a high activity can bypass a usual IP processing at the CSR.

The above procedure is performed at every CSR (e.g., every intermediate router that interconnects ATM networks). Then, an intermediate router can deal with a high activity IP packet flow with a very small latency. This is because IP packet forwarding is done by cell relaying, which will be done by a hardware. The job of CSR is maintaining the VPI/VCI mapping table in the cell switch, or is maintaining the mapping table between the flow-ID and the VPI/VCI, similar to cache (cut-thru) processing for the usual IP forwarding process.

When every data-link segment is an ATM network, the IP packet forwarding with small latency will be provided, even for a connectionless IP packet. This is because an IP packet with a high activity is relayed on the cell-by-cell bases at every CSR.

4.7.2 Bypassed ATM VCC for Active Transaction

When a router knows that it will have or is having a lot of communication with a certain router (or a subnet), it may create a cell-relaying cut-thru path which seamlessly connects the two routers at cell level. This cell-relaying cut-thru path bypasses the IP processing at the intermediate routers along the packet transmission path between the two routers. On the other hand, when a router does not use the cell-relaying cut-thru pipe for IP packet forwarding, the IP packet will be forwarded hop-by-hop, that is completely the same approach as the conventional IP forwarding.

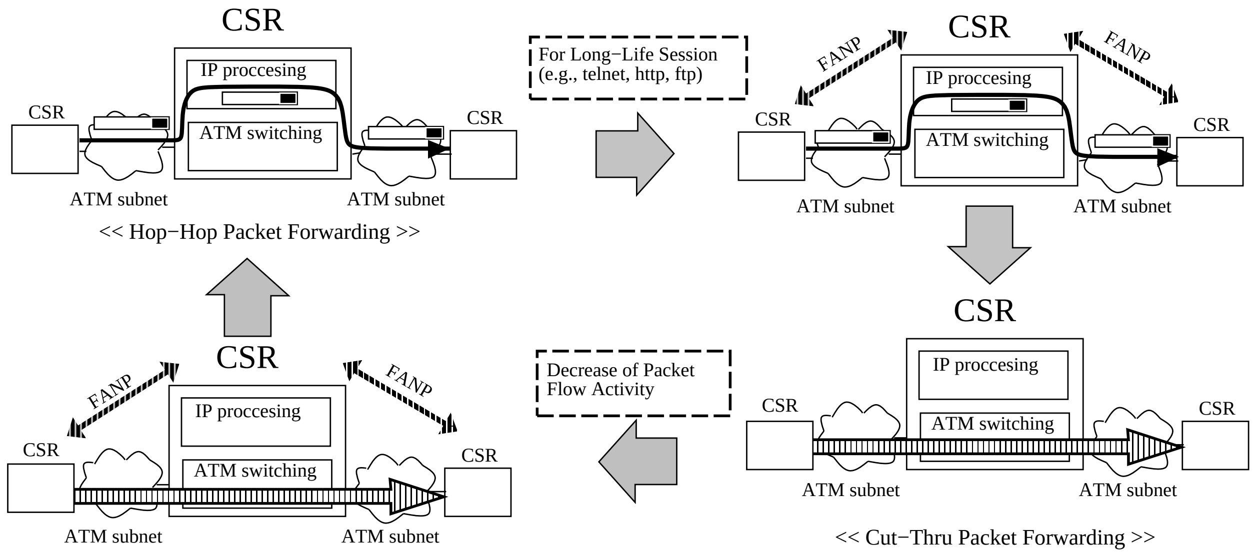


Figure 4-5. Operation of hop-by-hop ATM-VCC cacheing for active IP flow

When a router can establish a seamless cell-relaying route bypassing the intermediate router's network layer protocol processing along the routing path to the destination end-host, we can reduce the end-to-end IP packet delivery latency due to reducing the number of network layer protocol processing points at the intermediate routers. Here, it is unnecessary to propagate the information of such cell-relaying cut-thru pipe for every router. The information of cell-relaying cut-thru pipes shall be maintained by the source and destination routers associated with the cell-relaying cut-thru pipes. This means that the routing protocol and the routing information exchanged by routers is nothing different from the existing routing protocol.

The information and the path to bypass the intermediate router(s) could be seen as a cache image associated with the information and path for hop-by-hop routing (i.e., conventional IP forwarding routing). When the cache is missed (i.e., when there is no cut-thru pipe), the hop-by-hop information is referenced and the IP packet is forwarded to the adjacent router. On the contrary, when the cache is hit (i.e., when a cut-thru pipe exists associated with the destination subnet), the cache information is referenced and the IP packet is forwarded using the cut-thru pipe.

The establishment of the cell-relaying cut-thru pipe will be issued by the following two cases. Here, the establishment procedure of the cell-relaying cut-thru pipe is the same as the establishment procedure of the end-to-end connection oriented data transfer channel set-up discussed in section 4.5. And, the source router and the destination router discussed in this subsection can be an intermediate router, from the view point of end-hosts. This means that the source (or the destination) router may not directly connect to the source (or destination) end-host through a single cell-relaying cut-thru pipe. In other words, in order to reach the source router from the source end-host (or to reach the destination end-host from the destination router), intermediate router(s) may be required.

1. Source router
   The source router has the activity information regarding the destination subnet's. When the IP forwarding activity toward the certain destination subnet is high, the source router establishes the cell-relaying cut-thru pipe route toward the destination subnet's router. When the activity becomes low, the established cell-relaying cut-thru pipe will be torn down (i.e., the cache information is pushed out).

   Every router would establish the cell-relaying cut-thru pipe based on the IP flow's activity information. When there is no merged IP packet flow along the packet transmission path associated with the high activity IP packet flow and all of the IP packets belonging to the high activity IP packet flow take the same path from the source router to the destination router, every router along the packet transmission path will try to establish the cell-relaying cut-thru pipe to the same destination router. However, in general, we could say that there are merged IP packet flows, that have the same destination router, along the high activity IP packet flow. And, we could say that, in hop-by-hop IP forwarding, the IP packet forwarding path would not be the same for all the IP packets. Also, even when some routers (source router and intermediate routers) establish cell-relaying cut-thru pipes routes to the same destination router simultaneously, only one cut-thru pipe will finally survive. The survived cut-thru pipe route will be the route that is established by the farthest router from the destination router. The cut-thru pipes from the other routers will be automatically torn down (i.e., cache information will be pushed out), since the activities of the IP packet flow toward the destination router will be reduced.

   A cell-relaying cut-thru pipe from the source router toward the destination router can be established without the activity information. This means that cut-thru pipe can be established both dynamically and statistically based on the network's topology information. And, of course, whether to establish the cut-thru pipe based on the activity information depends on each router's policy.

2. Intermediate router
   When the intermediate router becomes high loaded, the intermediate router can create cut-thru pipes that are bypassing itself. The intermediate router establishes the cell-relaying cut-thru pipe for the active source and destination router pair. After the establishment of two cell-relaying routes, that are (1) from the source router to the intermediate router and (2) from the intermediate router to the destination router, these two cell-relaying routes are coupled at the intermediate router. This procedure could be seen as a re-direction of cell-relaying : i.e., the cell-relaying coming from the source router to the IP forwarding entity in the intermediate router is "re-directed" to the cell-relaying entity toward the destination router.

   The cell-relaying cut-thru pipes will be torn down by the intermediate router, when the processing load at the intermediate router becomes low. Generally, the threshold value of the processing load to tear down the cell-relaying cut-thru pipe(s) should be smaller than the threshold value to create the cell-relaying cut-thru pipe(s), in order to avoid the oscillation tendency. As a result, the cell-relaying cut-thru pipe(s) established by the high loaded intermediate router would be always between a high active source and destination router pair, when the intermediate router becomes high loaded.

The ideal and optimal case is where full meshed cell-relaying cut-thru pipes are established among all the routers that exist in the routing entry of the routing information.

4.8 IP Header Compression using VPI/VCI

Since ATM networks can define a virtual connection (ATM-VCC) having an unique flow identifier (i.e., VPI/VCI), we can optimize the packet transmission by means of (network / transport layer) header compression using the VPI/VCI. In the proposed architecture, the header compression technique will be applicable for (1) the ATM-VCC between adjacent routers and (2) the ATM cut-thru pipe that explicitly requests all the routers along the routing path to carry out a cut-thru packet forwarding. The benefits of the header compression are as followed.

1. Improvement of Data Transmission Efficiency
   For small size packet transmission, the transmission overhead by the network/transport layer header field is significant and degrades the actual user-data transmission efficiency. By the compression of the network/transport layer header field, the actual user-data transmission efficiency can be improved. The improvement of data transmission efficiency is sometimes an important factor for the data transmission across the WAN.

2. High Throughput Data Transmission
   When the part of or whole of the network/transport layer header is redundant, the redundant header field need not be transmitted to the termination point of the associated datalink pipe. When the whole of the network/transport layer header field can be compressed, it can be equivalent to bypassing the network / transport layer processing in the user-data transmission phase, i.e., it is equivalent to using a native datalink interface from the application.

3. Improvement of Data Transmission Latency
   Especially with small bandwidth communication pipe, the transmission of network/transport layer header field degrades the latency characteristics for real-time (or interactive) applications. By the compression of the network/transport layer header field, the latency characteristics can be improved. For the ATM networks, it would be said that the available bandwidth is not such a small amount, but is large enough. However, the available bandwidth may be small (e.g., less than 64 Kbps) in the following cases : (a) available bandwidth provided by ABR service is severely reduced by the extreme congestion, or (b) allocated bandwidth for long-distance connection will be small because the long-distance connection might be expensive even in ATM.

Since FANP, proposed in the previous subsection, indicates how the IP packet flows are aggregated into the ATM-VCC to the neighbor router, it is easy to define a further message exchanging protocol to perform header compression between the routers. The other method is to apply the similar technique as the PPP, e.g., [RFC1144][RFC1331]. There are following four compression cases.

- Compression #1
  An ATM-VCC conveys only one application flow, that is identified by source/destination IP addresses and by source/destination port-id's. Source/destination IP addresses and source/destination port-id's can be compressed into the corresponding VPI/VCI value.

- Compression #2
  An ATM-VCC conveys the IP packet flow whose source/destination IP addresses are always the same, i.e., only port-id's would be different. Source and destination IP addresses can be compressed into the corresponding VPI/VCI value.

- Compression #3
  An ATM-VCC conveys the IP packet flow whose source (or destination) IP addresses and source (or destination) port-id's are always the same ; IP packet flows aggregation into a single ATM-VCC. Source (or destination) IP address and source (or destination) port-id can be compressed into the corresponding VPI/VCI value.

- Compression #4
  An ATM-VCC conveys the IP packet flow whose source (or destination) IP addresses are always the same : IP packet flows aggregation into a single ATM-VCC. Source (or destination) IP address can be compressed into the corresponding VPI/VCI value.

In all cases, it is assumed that the following header fields can be additionally compressed ; (1) LLC/SNAP field [RFC1483], (2) Version/IHI/Type-Of-Service field in IPv4, (3) Version/Flow-label in IPv6. Here, in some cases, further additional header field will be able to compressed [RFC1144].

Tables 4-1 and 4-2 show the original header field (i.e., TCP/IP with LLC/SNAP encapsulation) length and the compressed header field length.

Table 4-1. Header Compression for TCP/IPv4 with LLC/SNAP Encapsulation

	LLC/SNAP	IPv4	TCP	Total
Original	8 Byte	20 Byte	20 Byte	48 Byte
Compression #1	0 Byte	10 Byte	16 Byte	26 Byte
Compression #2	0 Byte	10 Byte	20 Byte	30 Byte
Compression #3	0 Byte	14 Byte	18 Byte	32 Byte
Compression #3	0 Byte	14 Byte	20 Byte	34 Byte

Table 4-2. Header Compression for TCP/IPv6 with LLC/SNAP Encapsulation

	LLC/SNAP	IPv6	TCP	Total
Original	8 Byte	40 Byte	20 Byte	68 Byte
Compression #1	0 Byte	4 Byte	16 Byte	20 Byte
Compression #2	0 Byte	4 Byte	20 Byte	24 Byte
Compression #3	0 Byte	24 Byte	18 Byte	42 Byte
Compression #4	0 Byte	24 Byte	20 Byte	44 Byte

The Compression #1 for TCP/IPv6 is the most efficient header compression example, i.e., 68 bytes of header field can be compressed into 20 bytes. The Compression #4 for TCP/IPv4 is the least efficient header compression example, i.e., 48 bytes of header field can be compressed into 34 bytes.

4.9 Migration Scenario to the CSR-based Network

4.9.1 Introduce of CSR to Campus and Corporate Networks

Figures 4-6 - 4-9 show the migration scenario to the CSR-based network from a typical current network configuration. Figure 4-6 shows a typical network configuration of current campus network. Many Ethernets are internetworked through an FDDI backbone. Figure 4-7 shows the introduction stage of CSR from the current campus network. When we introduce some servers that require the high throughput among them and among the servers and clients, we need high speed channels among the servers and among the servers and clients. In this stage, CSR would be used as an alternative high speed route to the FDDI route.

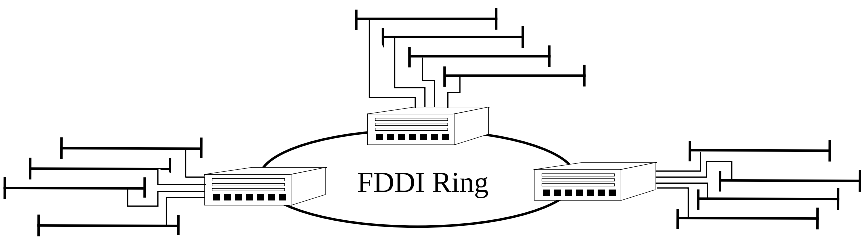


Figure 4-6. Migration scenario to the CSR-based network (current stage)

ATM technology will be a common platform for the subnetworking and then, campus networks will have some ATM-based subnetworks, as shown in figure 4-8. In this stage, we will obtain high throughput networks using the feature of CSR's cut-thru packet forwarding, while we have still legacy networks (e.g., FDDI).

As ATM technology becomes the common platform for sub-networking and the, campus networks will have ATM-based sub-networks, as shown in figure 4-8. In this stage, we will obtain high throughput networks using the CSR's cut-thru packet forwarding, while still retaining the legacy network (e.g., FDDI).

4.9.2 Introduce of CSR to ISP Networks

Figure 4-9 shows a typical network configuration for an enterprise network which is using IP service provided by an ISP, in this example CSR technology is common both in the campus networks and in the public network.

In this example a VPN (Virtual Private Network) has been provided by an ISP (Internet Service Provider). Since we do not have a security concern with a VPN, we may not need to have a firewall router at the border between intranet (i.e., campus network) and Internet (i.e., ISP network). In this particular case, we can directly internetwork CSRs both in the

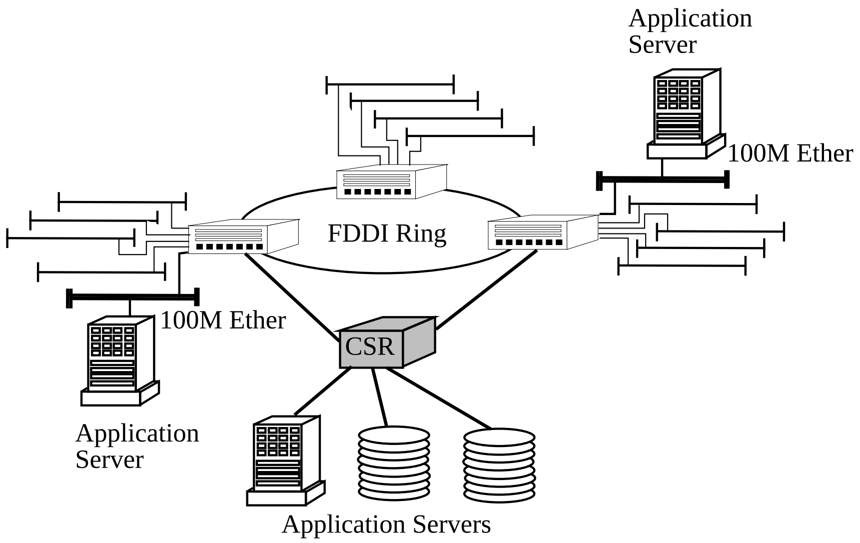


Figure 4-7. Migration scenario to the CSR-based network (stage 1)

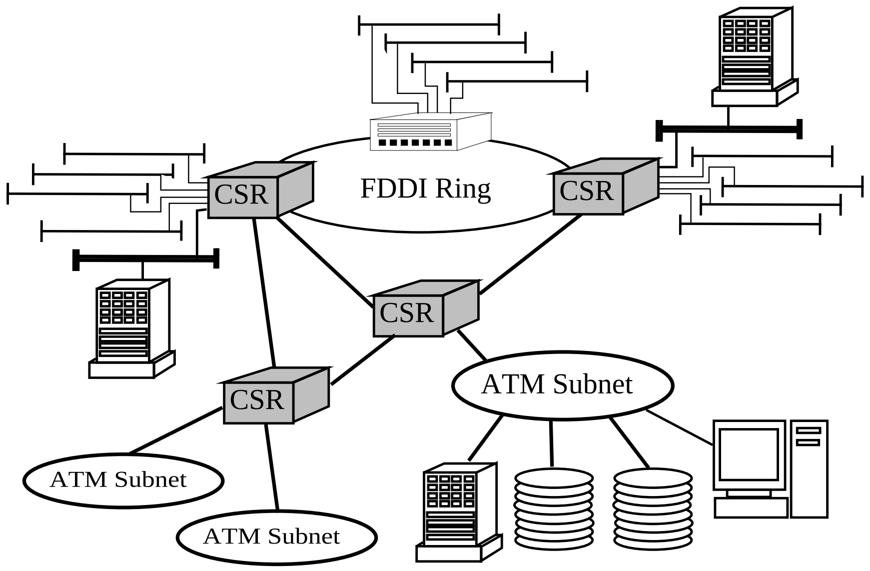


Figure 4-8. Migration scenario to the CSR-based network (stage 2)

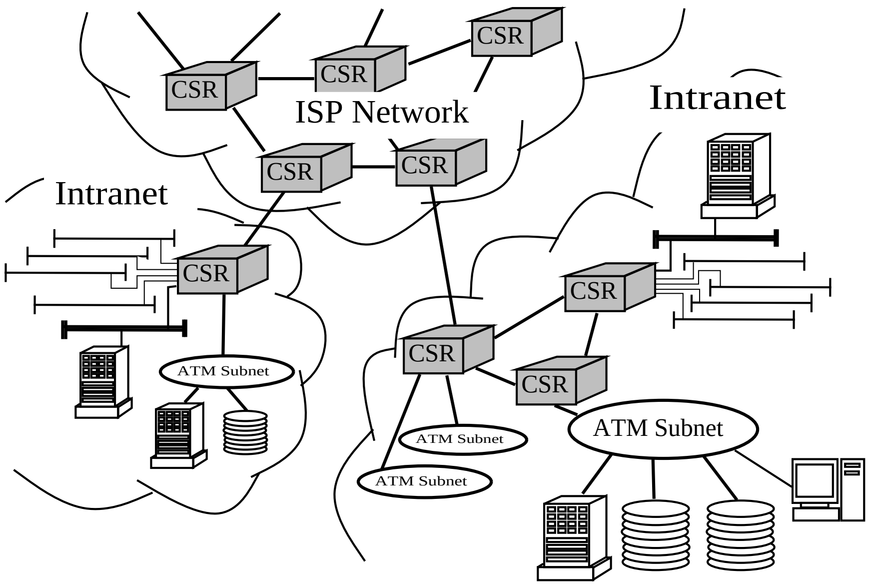


Figure 4-9. Migration scenario to the CSR-based network including ISP

campus network and in the ISP network. Even without VPN's, as discussed in 4.2, CSR could be used as a high throughput firewall router. This is because CSR can provides three levels of security functions, which are no packet filtering, packet filtering per session, and per packet filtering. With the per packet filtering, i.e., every packet is filtered at the router, performance will obviously suffer. However, with per session filtering, i.e., only the first packet in the session is filtered, CSR can achieve a high throughput. This is because we can apply the cut-thru packet forwarding mechanism to these type of sessions.

Figures 4-10 and 4-11 shows an example how CSR can be introduced to the existing ISP networks. Figure 4-10 gives an typical network configuration of ISP. Dial-up links are being used for access are and terminated by an access router. The Access router handles the packets through dial-up links, and multiplexes the packet into Frame Relay link, that leads to high speed routers. High speed routers are interconnected through both Frame Relay links and SONET links. For enterprise networks, the Frame Relay links are directly connected to the high speed routers in the ISP and are used as the access link to the ISP.

Figure 4-11 shows an example how the CSR can be introduced to the existing ISP network shown in figure 4-10. CSR can be introduced as a backbone router. Packet flows from the existing high speed routers can be aggregated by the backbone CSR routers which are connected by high speed SONET/ATM links. As shown, CSR's can be introduced into the existing ISP network, without any drastic configuration change or any equipment replacement.

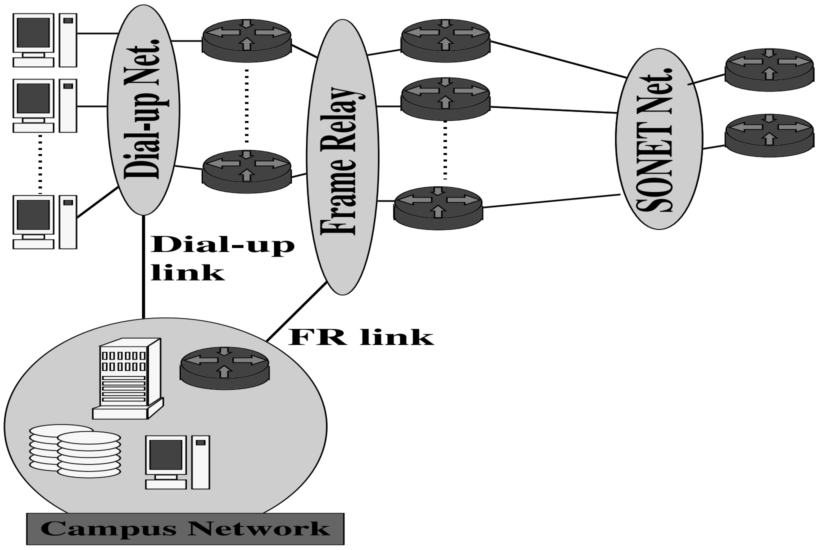


Figure 4-10. Example of Existing ISP Networks

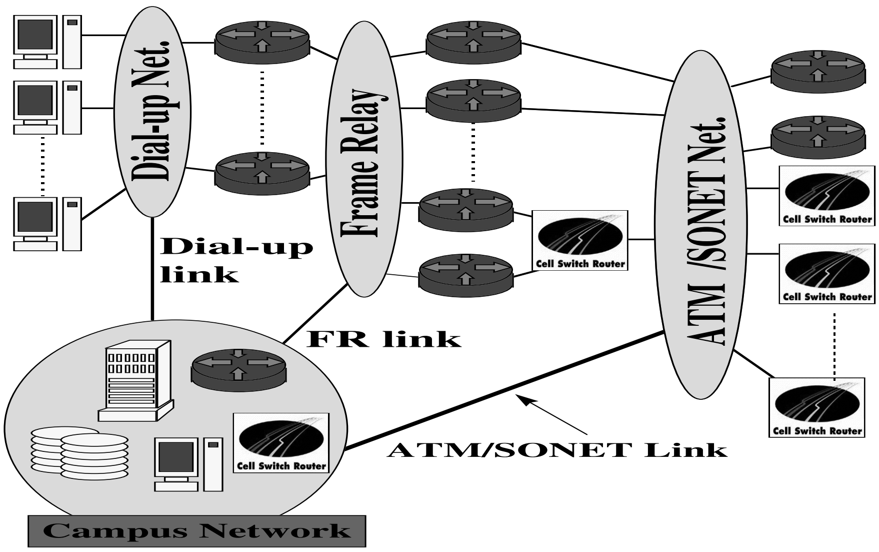


Figure 4-11. Introduce of CSR to Existing ISP Networks

5 Evaluation and Discussion of Cell Switch Router

In this section, the quantitative and qualitative evaluation of cell switch router and the discussions on cell switch router is performed. Through the evaluations, the effectiveness of cell switch router and the implementation feasibility is shown.

Table 5-1 shows the comparison of IP over ATM architecture models discussed in section 3, regarding some major points. As shown, the functionalities and features achieved by the cell switch router is better than the other architecture models.

Table 5-1. Comparison of IP over ATM Architecture Models

Model	CLPF	SCPF	LSTL	CSR
Throughput	poor	fair	Good	Good
Routing loop	No	Yes	No	No
ATM cluster	Yes	Yes	No	Yes
ATM-SW connectivity	Yes	Yes	No	Yes
System Complexity	Simple	Complex	Simple	Simple
System Cost	High	High	Low	Low

5.1 Aggregated System Throughput

The aggregated system throughput of cell switch router depends basically on the product of number and operation clock speed of interface ports. For the larger throughput of cell switching fabric, the required packet processing power of packet processing fabric is larger.

Table 5-2 shows the maximum aggregated system throughput of cell switch router with some configurations.

Table 5-2. Aggregated system throughput of Cell Switch Router

	CSR System Throughput		
	Bit Level	Pakcet Level	
		64B packts	100B packets
OC3−I/F 8 ports	1.2 Gbps	2.4 MPPS	1.5 MPPS
OC3−I/F 16 ports	2.4 Gbps	4.8 MPPS	3.1 MPPS
OC12−I/F 8 ports	4.8 Gbps	9.6 MPPS	6.1 MPPS

(*) MPPS; Million Packets Per Second

For example, with eight ports of OC3 interface, the maximum aggregated system throughput of CSR is about 2.4 million packets per second for 64 bytes average packet length. The throughput of 2.4 million packets per second approximately corresponds to the ten times larger system throughput of current high-end routers. With eight ports of OC12 interface, the maximum aggregated system throughput of CSR is about 9.6 million packets per second for 64 bytes average packet length. The throughput of 9.6 million packets per second approximately corresponds to the 40 times larger system throughput of current high-end routers.

As shown, the cell switch router can easily achieve larger aggregated system throughput than the current high-end routers can. The expected maximum system throughput of CSR can be easily decades times larger than the maximum system throughput of the current high-end routers.

5.2 Delay Variance with CSR

In CSR, IP packets using different dedicated-VCs are transferred in parallel, since IP packets are transferred with a cell-interleaving. On the contrary, in the conventional router, IP packets are transferred packet-by-packet with an FIFO rule. This means that the IP packet has to be in the queue until the all IP packets before it are transferred, i.e., HOL (Head Of Line) blocking. When we compare the packet delay variation (PDV), the PDV with CSR will be smaller than the PDV with the conventional router.

In the following evaluations, it is assumed that the switching speed for ingress IP packets is sufficiently fast than the egress link speed both for CSR and for conventional router. That means IP packets (i.e., cells with CSR) are queued at the egress link to the neighbor node.

The minimum packet transmission delay at the conventional router is smaller than that of CSR. No packet is queued before the IP packet experiencing the minimum packet transmission delay in the conventional router. The packet experiencing the maximum packet transmission delay waits in the queue, until all IP packets before the IP packet in the queue are transferred. With CSR, the PDV will be very smaller than that of the conventional router, though the minimum packet transmission delay of CSR would be larger than that of the conventional router and it will be almost the same as the maximum packet transmission delay of conventional router. This is because CSR does not transfer only one IP packet, but the multiple IP packets are transferred simultaneously with cell-interleaving.

Figure 5-1 shows the maximum and minimum packet transmission delay versus the IP packet size, when 10 packets are simultaneously transferred to the queue at the egress link. Here, the data transmission speed of egress link is 155Mbps. As shown, the PDV of CSR is quite small, however, the PDV of conventional router is large. With CSR, the maximum delay jitter is always 27 μ sec. However, the maximum delay jitter with conventional router increases according to the increase of IP packet size; i.e., the maximum delay jitter with 1 KB packet is 464 μ sec.

As well-known, the conventional router can not achieve sufficient IP packet processing power for large amount of incoming IP packets. In general, the aggregated system throughput of current high-end router would be about 0.2 - 0.3 million packets per second, that corresponds to about 120 - 200 Mbps. However, if the conventional router could have sufficient IP packet processing capability, the average packet transmission delay of conventional router is smaller than that of CSR. Smaller average packet transmission delay is generally good characteristics for many applications. However, some applications (e.g., continuous bit-stream transmission) are sensitive to the PDV rather than to average packet transmission delay. Typical applications, that are sensitive to the PDV, are voice and video signal transmission. With voice and video applications, the required receiving buffer space goes

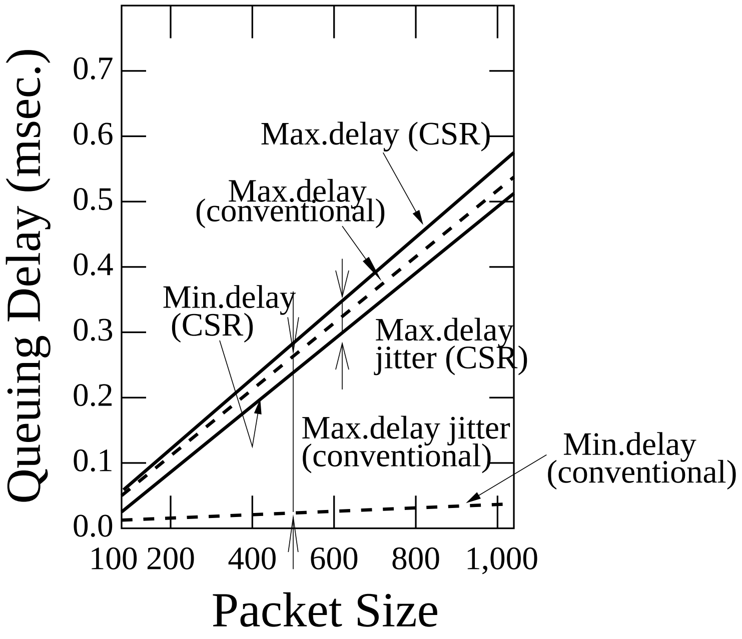


Figure 5-1. Delay Variation of Packet Delivery

to larger for the larger PDV in order to maintain the playback point. This means that the required receiving buffer space for voice and video signal transmission using CSR will be much smaller than that using conventional routers.

5.3 Effectiveness of Cut-thru Packet Forwarding

The effectiveness of cut-thru packet forwarding in the CSR mainly depends on the characteristics of IP packet flows. When we have many long-life sessions generating many IP packets, the cut-thru IP packet forwarding is meaningful and the cost to establish cut-thru paths (i.e., dedicated-VCs) is acceptable.

Tables 5-3 and 5-4 show (a) number of cut-thru sessions that are simultaneously established at the CSR, (b) total amount of IP packets, belonging to the typical applications that would likely be long life sessions, that will be forwarded with cut-thru mode, and (c) the average cut-thru session life-time for typical applications that would likely be long life sessions. CSR decides that four applications, i.e., web, ftp, telnet and nntp, are likely long life session to be forwarded with cut-thru mode. When CSR receives the first IP packet belonging to one of these four application, the dedicated-VC is established to forward the rest of succeeding IP packets with cut-thru mode. The cut-thru mode is terminated to go to the hop-by-hop mode, when the application does not generate any IP packet for two minutes.

As shown, the total amount of IP packets forwarded by cut-thru mode will be more than 86% in the DEC's backbone and will be more than 70% in the Toshiba R&D center backbone. Also, the average life time of cut-thru sessions is more than eight minutes both for DEC's backbone and for Toshiba R&D center backbone. These evaluation results indicate that,

Table 5-3. Cut-thru Packet Forwarding for Current Packet Flows (DEC backbone)

Session	# of Sessions		Traffic Volume	Average Life−Time
	Ave.	Max.		
web	130	161	13.41%	324.7 sec.
telnet	16	22	0.43%	1131.9 sec.
ftp	100	144	32.91%	407.3 sec.
nntp	130	139	39.94%	1058.3 sec.
Total	276	430	86.69%	481.1 sec.

Table 5-4. Cut-thru IP Packet Forwarding for Current Packet Flows
(Toshiba R&D backbone)

Session	# of Sessions		Traffic Volume	Average Life−Time
	Ave.	Max.		
web	63	98	8.88%	504.2 sec.
telnet	15	27	12.50%	897.0 sec.
ftp	4	19	14.15%	238.5 sec.
nntp	4	12	34.81%	260.5 sec.
Total	86	115	70.34%	515.2 sec.

even with the current IP flows, a large amount of IP packets will be able to be forwarded through cut-thru mode and the life time of cut-thru sessions will be long enough to establish dedicated-VC for them.

In the future, long life sessions will increase in the Internet and in the intranets. Voice and video signal transmission generates large amount of IP packets continuously for a long time and the traffic volume of these applications will be large and will increase. By means of using proxy servers, web application will generate larger amount of traffic and will have longer session life time. Using proxy servers, the client host always exchanges the IP packets with its proxy server, even when the client host access the different web site in the Internet. The above tendency will further improve the effectiveness of cut-thru IP packet forwarding provided by CSR.

5.4 Required Number of Dedicated VCs for CSR

The required number of dedicated-VCs provided by the CSR depends on the number of cut-thru sessions established simultaneously in the CSR. When the number of cut-thru sessions established simultaneously is too large, CSR could not provide dedicated-VCs for all IP flows that potentially require cut-thru packet forwarding.

Tables 5-3 and 5-4 shown in the previous subsection indicate the number of dedicated-VCs that CSR has to be able to provide. As for DEC's backbone router, CSR has to be able to provide more than 500 dedicated-VCs. As for Toshiba R&D center backbone router, CSR has to be able to provide more than 120 dedicated-VCs. In the future, the number of dedicated-VCs, that can be provided by the CSR, will increase. The maximum number of dedicated-VCs, that can be provided by the CSR, seems to be few thousand, and it seems to be possible to provide even by the current cell switching fabric.

5.5 Performance Evaluation of Header Compression with VPI/VCI

The performance of header compression using VPI/VCI proposed in section 4.7 is evaluated. The performance evaluation are for the best case (TCP/IPv6 with Compression#1) and the poorest case (TCP/IPv4 with Compression#4). The improvement of transmission efficiency (I) is shown by $(N_{nc} - N_c)/N_{nc}$. Here, N_c is the number of transmitted cells with the header compression, and N_{nc} is that without the header compression.

Tables 5-5 and 5-6 show the improvement of transmission efficiency according to the user data size (i.e., payload of TCP/IP packet). As shown by the evaluation examples, whether we can obtain an sufficient improvement of transmission efficiency by the header compression depends on (1) packet size, and (2) how the IP packet flows are aggregated in the ATM-VCC (i.e., compression type). When the packet size is small (e.g., 64 byte) and one IP packet flow is mapped to one ATM-VCC, we can expect a large gain associated with the data transmission efficiency.

Table 5-5. Improvement of Transmission Efficiency for TCP/IPv6 with Compression#1

Payload (Byte)	N_{nc}	N_c	Improvement (I)
0 - 20	2	1	50.0%
21 - 68	3	2	33.3%
69 - 116	4	3	25.0%
117 - 164	5	4	20.0%
165 - 212	6	5	16.7%
213 - 260	7	6	14.3%
261 - 308	8	7	12.5%
309 - 356	9	8	11.1%
357 - 404	10	9	10.0%
405 - 452	11	10	9.1%
453 - 500	12	11	8.3%
501 - 548	13	12	7.7%
1,024	23	22	4.4%
8,192	172	171	0.6%
65,536	1367	1366	0.1%

Table 5-6. Improvement of Transmission Efficiency for TCP/IPv4 with Compression#4

Payload (Byte)	N_{nc}	N_c	Improvement (I)
0 - 6	2	1	50.0%
7 - 40	2	2	0.0%
41 - 54	3	2	33.3%
55 - 88	3	3	0.0%
89 - 102	4	3	25.0%
103 - 136	4	4	0.0%
137 - 150	5	4	20.0%
151 - 184	5	5	0.0%
185 - 198	6	5	16.7%
199 - 232	6	6	0.0%
233 - 246	7	6	14.3%
247 - 280	7	7	0.0%
473 - 484	12	11	8.3%
485 - 520	12	12	0.0%

5.6 Internetworking Capability

IP over ATM architectures, e.g., CLPF, SCPF, LSTL, can be internetworked through the conventional router. However, when we internetwork different IP over ATM segments (e.g., SCPF network and LSTL network) through the conventional router, the IP packet forwarding throughput beyond the IP over ATM segment will not be sufficiently large.

All different IP over ATM model networks can be internetworked through the conventional router, however CLPF segment, LSTL segment and SCPF segment will not be able

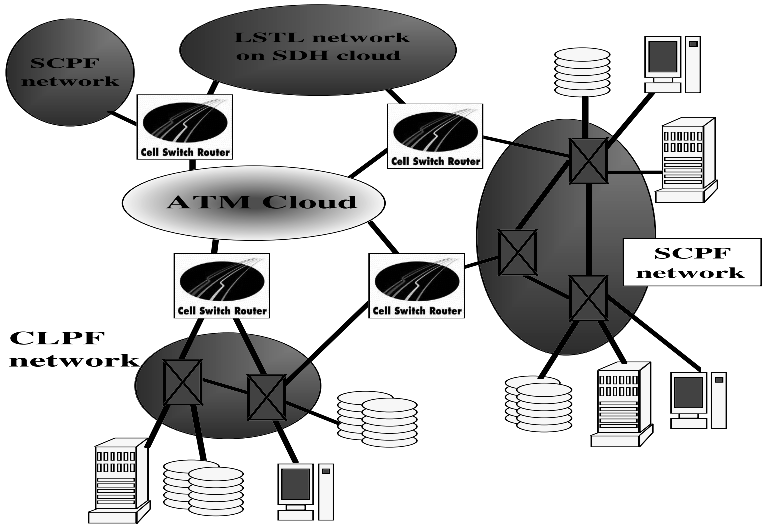


Figure 5-2. Internetworking of IP over ATM segments through CSR

to internetworked with high throughput. On the contrary, CSR will be able to achieve high throughput internetworking with all other IP over ATM segments. By means of introduction of CSR, IP packets will be able to be forwarded through cell-by-cell relaying (i.e., cut-thru IP packet forwarding), even when the IP packets are forwarded across the different IP over ATM segments. Figure 5-2 shows the case where different IP over ATM segments are internetworked through CSRs. Through the internetworking with CSRs, we will be able to achieve sufficient IP packet forwarding capability among the different IP over ATM segments.

5.7 Discussion

In the following subsections, the issues, that require further consideration and evaluation, are discussed.

5.7.1 Cut-thru Parameters

Cut-thru IP packet forwarding is initialized by TCP/UDP information in the received IP packet and it is released when no packet is generated during a give period (T_{term}). The performance and effectiveness of cut-thru packet forwarding depends on these parameters, i.e., (1) which TCP/UDP ports are the trigger to initiate the cut-thru forwarding and (2) how long T_{term} should be.

When we pick up many applications (i.e., many TCP/UDP ports) to be as cut-thru trigger, the total amount of cut-thru packets will be going to be large. However, the dedicated-VC establishment overhead may be too large, since many short-life sessions may picked up as a cut-thru trigger. Therefore, we have to choose the appropriate applications to achieve an effective cut-thru packet forwarding. Though we may have some default rule which applications should be cut-thru trigger, each CSR can independently determine which applications are cut-thru trigger. When we introduce a new application, each CSR decides whether it should be treated as a cut-thru trigger or not. Regarding the current applications over the Internet and intranet, ftp, telnet, web, nntp, voice and video seem to be appropriate applications to be cut-thru trigger.

Parameter T_{term} is also important parameter to optimize the effectiveness of cut-thru packet forwarding. When T_{term} is so long, the utilization efficiency of cut-thru pipe (i.e., dedicated-VC) will be poor, while the percentage of packets forwarded with cut-thru mode is large enough. On the contrary, when T_{term} is so short, the utilization efficiency of cut-thru pipe may be good. However, the overhead of dedicated-VC establishment may be too large. And, the percentage of packets forwarded with cut-thru mode may be small, since the first packet after the termination of dedicated-VC is always transferred through the hop-by-hop mode. The optimal value of T_{term} will be unique for each applications and unique for each user. However, we can not define T_{term} for each applications and for each user in the actual operation, i.e., we have to pick up one T_{term} value for each CSR. This value T_{term} should be determined either by the default value provided by the vendor or by the actual operation (i.e., try and error).

The detailed evaluation on these parameters should be further study item.

5.7.2 Cut-thru IP Packet Flow Aggregation

When the CSRs are interconnected through an ATM link across the wide area networks (WAN), it may be better to multiplex multiple IP flows using the dedicated-VC into a single ATM-VCC, in order to transfer IP packets to the CSR across the WAN. This would happen when the cost of an additional ATM-VCC establishment across the WAN is more expensive, than using an ATM-VCC which has a sufficiently large bandwidth. In this case, multiple cell flows using different dedicated-VCs must be merged into a single dedicated-VC. In general, we must re-construct IP packets in the multiple dedicated-VCs in order to merge them into a single dedicated-VC. However, if we have a cell switching fabric that can schedule a cell transmission from the output port so as to avoid a cell interleaving associated with the packets that belong to the different ingress ATM-VCC but using the same egress ATM-VCC (i.e., using the same VPI/VCI), we do not have to relay on the packet switching fabric with the reconstruction of IP packets. This will be possible when the internal data-unit transferred in the cell switching fabric has the ingress VPI/VCI value and the input port identifier. If we have such a cell switching fabric, the packet switching fabric can transfer the IP packet in pipe-line (i.e., IP packet transmission will be able to start before re-assembly of IP packet is completed) and could aggregate the multiple cut-thru IP packet flows into a single cut-thru path.

5.7.3 Route change

The IP packet routing path is changed according to the change of network status. Both hop-by-hop path with default-VC and cut-thru path with dedicated-VC will dynamically reflect to the packet forwarding path indicated by the routing protocol (e.g., OSPF). When the routing path is changed in the CSR system, one or two packets would be dropped at the edge node of cut-thru path. This is because all cells in the packet will not take the same path to be reordered. The one or two packet drops causes the degradation of end-to-end throughput for the old version of TCP due to a TCP windows shrink. However, since the current version of TCP generally implements a fast retransmission mechanism, one or two packet drops will not cause the throughput degradation.

When the routing path is changed, the packet forwarding goes to a hop-by-hop mode and each CSR tries to establish a cut-thru path based on the TCP/UDP port information of the first packet to be transferred to a new path. When the routing path frequently change, we can not achieve the effective operation of cut-thru packet forwarding. However, actually, since many routing protocol (e.g., OSPF, BGP) has a pinning option to avoid frequent path flapping, we can expect that we will be able to achieve effective use of cut-thru forwarding path.

5.7.4 Flow Mapping Guideline

CSR has to map a flow-spec (R-spec and T-spec) of IP packet flow and an ATM traffic parameters for each cut-thru path. Though the mapping rule between flow-spec and ATM traffic parameters is each CSR's local decision, it would be better to have some mapping guideline. The following would be a possible guideline.

- Conventional best effort service; ABR or UBR
- Control Load (CL) service, e.g., M-bone; CBR or VBR
- Guaranteed Delay service; CBR

Regarding RSVP, the heterogeneous resource reservation could be allowed to each down-stream node. How to map the ATM traffic parameter to the down-stream nodes requesting different resource reservation should be for further study item.

5.7.5 TTL Issue

At the CSR, TTL does not decrease with cut-thru packet forwarding mode. And, the TTL is decreased by one at the egress node of cut-thru path. Therefore, the decreased TTL experienced cut-thru path is smaller than the decreased TTL with conventional hop-by-hop packet forwarding. The actual purposes of TTL would be (1) avoiding permanently existing packet on the routing loop, and (2) define the scope where the packet can travel.

The former purpose would be satisfied with CSR, since TTL is decreased at the egress point of cut-thru path anyway. Only problem associated with item (1) is that the period before discarding the packet on the routing loop will be large due to slow decrease of TTL.

The later purpose will be satisfied with CSR, since routing protocol is performed with hop-by-hop path (default-VC) and the first packet in the session is never forwarded with cut-thru mode. Either by the routing protocol or by the first packet's hop-by-hop packet forwarding, the purpose of TTL field to define IP packet's traveling scope will be satisfied.

5.8 Summary and Conclusion

As shown above, the CSR (Cell Switch Router) will be able to achieve high throughput and small latency packet delivery with small cost a nd will be able to achieve effective utilization of cut-thru paths. Also, the internetworking capability of CSR is sufficient to internetwork any type of IP over ATM segment (e.g., SCPF network). Some points need further evaluation or operational tuning, which could be solved without major difficulty.

As a result, the CSR can be a scaleable and high throughput platform for the future realtime Internet and intranet.

6 Large Scale Error-free Multicast Service Architecture over ATM Networks

In this section, a large scale error-free multicast architecture over ATM networks is proposed. The proposed architecture does not require a new transport protocol and a new multicast routing protocol. T he appropriate transport protocol is MTP (Multicast Transport Protocol) defined in [RFC1301], and DVMRP (Distance Vector Multicast Routing Protocol) would be applied as the multicast routing protocol. In the following subsections, the proposed architecture to solve each issues discussed in section 3.2.

1. Control Packet and Protocol State Management
2. Resource and Error Management
   - Resource Management for Multicast Connection
   - Cell-Level FEC Control Policy
   - Retransmission Policy
3. Receiver (Membership) Management
4. Further Scaling Up Strategy

6.1 Control Packet and Protocol State Management

Protocol state information should be maintained at each branch point entities (maybe branch point routers) within the multicast tree. The branching point entities need not buffer the IP packets, but just maintain the protocol state and merge the control packets. The control packets from the down-stream toward the sender process should be merged into one control packet, that is transferred toward the sender process (i.e., toward up-stream). The method of the protocol state maintenance at the branch point entities is fulfilled based on a soft-state state maintenance technique with NACK (negative ACK) control message, which is defined in RFC1301. Here, the soft-state state maintenance means that the protocol state is checked and refreshed only when a control message is received.

The source process or some dedicated processes in the multicast tree advertises the referenced protocol state information (e.g., the information sent from the source process to the receivers). When the receivers identify the miss of some packets either according to the referenced protocol state information or due to sequence number inconsistency associated with the received packets, a control packet to be sent toward the source process is generated by the receiver process. The control packet to be sent toward the sender process is also generated, when the receiver process gets an errored packet. As a result, the generation of the control packet to be sent toward the sender process is an on-demand and is not generated in the normal status (i.e., in no packet loss and error). When no control message is received by the sender process within the given period, no-action is fulfilled and it is assumed that the receiving entity, which does not send back the control message (i.e. NACK message), correctly receives the packets or quit from the multicast service. When the fan-out (number of leaves that branching point has) at the branching point entities is m, the protocol state information maintained in the branching point entity will be *order of m*.

In this approach, a timer, that determines whether no control packet is issued by the receiver for error-free packet delivery, must be defined. This timer must be larger than the maximum round-trip delay between the source and destination receivers. Therefore, the maximum expected round-trip delay within the multicast tree must be specified and the multicast connection must provide this delay quality.

In order to work well even if receivers quit from the multicast service without any quitting indication to the multicast service management entity, a negative ACK policy (i.e. NACK policy) is applied to so as to get a packet reception status of the receivers. In a NACK policy, when the NACK packet is not received, it is assumed that the end-host receives the packet correctly, i.e., soft-state management policy. As a result, the end-host without sending a quit indication can be assumed that it always receives packets correctly.

Finally, the branching point entities merge the control packets. The control packets from the receiver processes toward the sender should be merged into one control packet at the branching entities. By this approach, we can avoid the increase of returning control packet from the receivers toward the sender process, even for large number of receiver processes. Once again, in this approach, the timer that determines whether no control packet is issued by the associated down-stream entities (i.e., receiver processes or branching entities) must be defined. Therefore, the maximum expected round-trip delay between the branching entity (or sender process) and the associated leaves of branching point entities (and/or receiver processes) must be specified and the multicast tree must provide this delay quality.

6.2 Resource and Error Management

6.2.1 Resource Management for Multicast Connection

In order to provide a sufficient service quality for the multicast service and to operate soft-state NACK based protocol state management, some resource reservation oriented protocol (e.g., RSVP) should be applied to the reliable multicast service. With the use of a resource reservation oriented protocol, we can expect some level of service quality (i.e., packet error/loss ratio and delay quality) over the multicast tree. Packet error/loss ratio quality is mainly associated with the cell loss ratio quality that is provided by the ATM networks. Packet delay quality is mainly associated with the service class that is provided by the routers [Shnk1][Shnk2][Shnk3], and with the delay by the buffering in the intermediate ATM switches.

As discussed in the previous subsection, we need specified delay quality over the multicast tree. The delay quality provision is required for both between the source process and the receiver processes and between the branch point entity (or source process) and the associated branch entities (or receiver processes). The service category to be used for the reliable multicast service by the routers will be guaranteed, predictive or controlled delay service class [Shnk1][Shnk2][Shnk3]. The service category to be used for the reliable multicast service by the ATM platform should be generally either CBR or VBR. UBR and ABR services could not be used because neither UBR nor ABR has any delay quality objective.

6.2.2 Cell-level FEC Control Policy

Cell-level FEC control policy can be applied to obtain a sufficient packet error/loss quality for a large scale multicast service [Carle][95-0325][95-0326]. Basically, the processing of cell-level FEC control scheme is performed only at end-nodes (i.e., source node and destination node). However, in some cases (e.g., a reliable multicast over the large could ATM network), the some intermediate routers will perform the cell-level FEC control scheme, wherever the IP packet is re-constructed from the multiple cells. Here, the intermediate router within the ATM cloud need not always terminate ATM connection to re-construct IP packet. For example, for a cut-thru router, some IP packet flow (e.g., resource reserved ST-II [ST-II] flows) can be forwarded cell-by-cell without IP packet re-construction [Esaki2][RFC2098], even when the IP packet passes through the router. Without cell-level FEC control, the packet error/loss quality in the ATM networks degrades rapidly with increasing the following three factors as discussed in section 3.2.

- Cell loss ratio (or the bit error rate of the medium)
  The packet error/loss quality will degrade by cell loss and the BER (bit error ratio) in ATM networks. Reducing the effective BER and cell loss ratio for the upper layer process (e.g., IP) will significantly improve the overall service quality.

- Packet (or frame) size
  The loss rate of higher layer packets (e.g., TCP packets) grows linearly with the number of cells composing a packet.

- Number of receivers
  Since the ATM networks are switched oriented networks, the actual packet error/loss ratio for the sender process will approximately increase linearly according to the increase of the number of receivers.

Using cell-level FEC control for the multicast service over the ATM networks, the above three issues will be able to be solved. By applying a cell-level FEC control, the actual packet error ratio will be significantly reduced. For example, for 10^{-9} cell error/loss ratio with 100 cell size packet, when one cell loss (or error) in 10 user cells is corrected by the FEC control, the actual packet loss/error ratio will be 10^{-12} even for 10^4 receivers. In this case, as mentioned above, the packet error ratio for the sender process without the cell-level FEC control will be about 10^{-3}.

Figure 6-1 shows the proposed protocol structure. AAL type is AAL Type 5. Since the proposal is about SSCS of AAL Type 5, only the processing sequence of SSCS is briefly described below.

SSCS-SDU (SSCS-Service Data-Unit), which corresponds to AAL-IDU (AAL-Interface Data-Unit) may be equivalent to an IP packet, whose default MTU size would be 9,180 bytes in ATM networks [RFC1626]. SSCS-SDU will be divided into multiple FEC-SDUs. Each FEC-SDU is stored in the FEC Frame, and the FEC code is calculated to be attached, e.g., [McAuley][I.363]. The writing order and the reading order for the FEC frame are the same order, in order to avoid an unnecessary user data interleaving. When the data (SSCS-S-SDU) is read out from the FEC Frame, the FEC frame header field, containing the

SN (sequence number) field, is inserted in order to identify which SSCS-S-SDU(s) is(are) dropped at the SSCS entity at the destination end-host. Also, the FEC frame header field contains the CRC-10 error check sequence, in order to detect the bit(s) error in the SSCS-S-SDU. The FEC-SSCS described in appendix A can transmit the FEC-SDU in pipe line, since the writing order and the reading order for the FEC frame are the same order. When the actual SSCS-SDU (i.e., AAL-IDU) or FEC-SDU is smaller than the size of data part in the FEC frame that is defined in the FEC calculation algorithm, the remaining field of FEC frame's data part is not have to be sent.

The FEC-SSCS uses the Reed-Solomon error recovery algorithm of the block erasure mode [I.363]. The FEC frame has a user-data part and an FEC code part. The variable length of FEC frame (both user-data part and FEC code part) can be allowed for the efficient transmission of small sized SSCS-SDU. A user-data part containing the user information will be more than one cell, and an FEC code part will be one or more than one cells. The symbols in the FEC code part are generated by the generator polynomial using the user-data symbols. The error detection for the FEC frame is performed by the AAL5's CRC-32 error check sequence, and the error detection for each SSCS-S-SDU is performed by the CRC-10 error check sequence in the FEC frame header field. The CRC-10 error check sequence is adopted, in order to improve the error correction capability of the Reed-Solomon algorithm.

When the number of errored or dropped SSCS-S-PDUs is larger than the correction capability of the applied FEC, certain retransmission policy of SSCS-S-PDUs is fulfilled. For example, whole of FEC Frame will be retransmitted, or selected SSCS-S-PDUs will be retransmitted.

The size of FEC frame for the FEC calculation algorithm should be determined aligning to the IP packet length and should be determined based on the requirement of the end-to-end IP packet delivering latency. The larger FEC frame causes larger latency for FEC code calculation, when there is bit(s) error or cell(s) discarding in the received FEC frame. When there is no bit error nor cell loss in the received FEC frame, the FEC calculation at the receiver is unnecessary.

The detailed FEC-SSCS specification is described in the appendix A and in [95-326]. The FEC mechanism applied to in this paper is the Reed-Solomon with Block-Erasure mode discussed in [MaAuley]. The other FEC mechanism could be applied to the proposed architecture in this paper, however the Reed-Solomon with Block-Erasure mode would suitable to the ATM platform environment. Since the information is dropped in burst (i.e., by 48 bytes unit) in the ATM platform rather than is errored randomly bit-by-bit like in the legacy data-links (such as SONET links), the FEC algorithm has to deal the burst error. From this point of view, the Reed-Solomon algorithm would be one of good mechanism to applied to. Also, since the bit error ratio will be far smaller than the cell loss ration in the ATM platform, applying the block-erasure mode defined in [McAuley] would be appropriate. By the applying the block-erasure mode for Reed-Solomon algorithm, we can reduce the amount of calculation to recover the erroneous information at the receiver node.

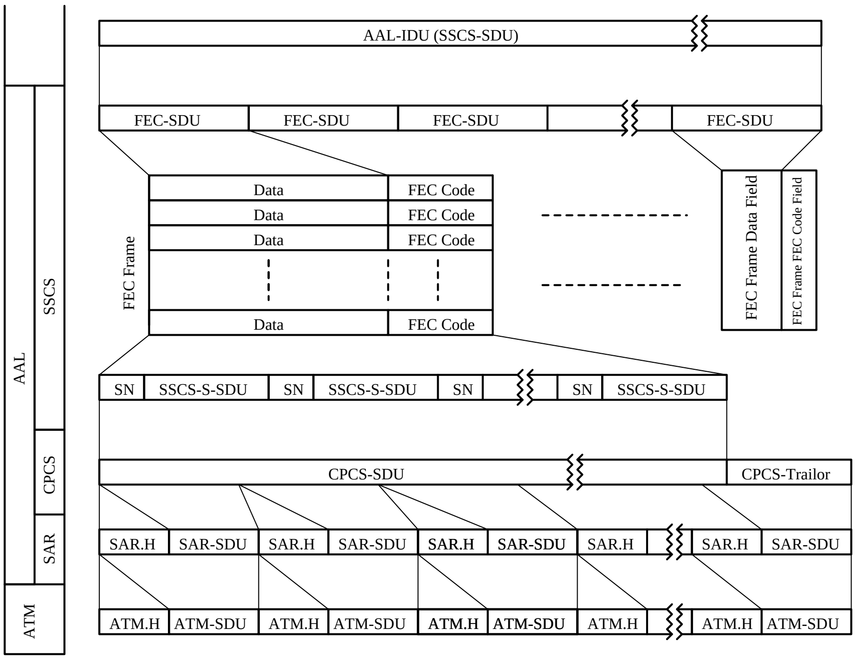


Figure 6-1. Protocol Structure Applying FEC-SSCS

6.2.3 Retransmission Policy

Basically, the branching point entities, that maintain protocol state information, do not perform any user packet retransmission to recover the error or missing packets. In other words, the packet retransmission to recover the error or missing packets is basically performed by the sender process. Therefore, we need not to preserve any user packet within the networks, for the purpose of error/missing packet recovery. There are many proposal suggesting to perform a local retransmission for the errored or missing packets, instead of the sender process. This approach requires the preservation of the user packets at the branching point entities [Sanjoy]. This approach should be applied only for the following cases.

1. The expected packet error or missing probability in the given multicast sub-tree region is not small enough.

2. The packet retransmission by the sender process will be expensive, e.g., packet transmission across the WAN.

3. The expected round trip delay of the control packets is unacceptably large.

4. The expected packet error probability for the sender process is unacceptably large, e.g., 10^{-3}, due to the extremely large number of receivers.

Cases 1 and 2 are completely beside on the local decision, and these two operations does not give any bad interaction to the proposed packet retransmission architecture. On the contrary, the cases 3 and 4 are rather the essential issue and are the strategy for the further scaling up of the multicast service.

Regarding a packet (or cell) retransmission policy for the packet that can not be recovered by cell-level FEC control scheme, the following items should be considered.

- Cell-based retransmission and Frame-based retransmission
  When the receiver can not recover the correct packet because cell losses and bit errors is beyond the FEC control capability, the receiver requests the data (frame or cells) retransmission in order to get a correct packet. The performance of cell-based retransmission scheme and frame-based retransmission scheme has been evaluated in [Carle]. Total number of retransmitted cells for the cell-based retransmission scheme will be always smaller than that for the frame-based retransmission scheme will be. However, the implementation of cell-based retransmission scheme is much complex than that of frame-based retransmission scheme is. The difference of total number of retransmitted cells is larger at the larger cell loss ratio case, and is smaller at the smaller cell loss ratio case. Since the probability that the source node must perform a frame retransmission becomes sufficiently small by the cell-level FEC control, it would not be necessary to apply the cell-level retransmission scheme for the un-recovered frame by the cell-level FEC scheme.

  Therefore, in the performance evaluation at the next section, it is assumed that the frame-based retransmission scheme is applied.

- FEC capability for retransmitting data (packet or cells)
  Loss of retransmitted frame will significantly degrade QOS associated with packet delivery latency. In order to obtain a sufficient quality for retransmitted frame, more redundant cells that is larger than the usual frame transmission includes could be attached [McAuley]. This scheme requires the further implementation complexity. The effectiveness of this scheme should be for further study item.

  Here, this approach may lead to reduce the totally transmitted cells from the sender process, when the probability that the frame will experience retransmission is sufficiently small. Therefore, this approach may be applied, when the transmission bandwidth between source process and receiver processes are precious.

  In the performance evaluation at the next section, it is assumed that the FEC capability (i.e., the number of redundant cells to be appended) is not changed for the transmission of retransmitting packet.

6.3 Receiver (Membership) Management

In order to join to (or quit from) the multicast service, each receiver will send a request message (i.e., IGMP) to the local multicast service management entity that may be located at the branching point of the multicast tree.

Membership management is locally performed by the local multicast membership management entities (LMME), i.e., distributed membership management. In other words, each LMME only manages the leaf entities (i.e., branching point entities or receiver processes) associated with its multicast sub-tree. When the number of receiver is not large, membership management could be performed by the sender process (or single management entity) associated with all receivers in the multicast tree, i.e., centralized membership management. However, for the large scale multicast, the centralized membership management has significant shortcomings.

In the distributed membership management, there are two membership management strategies. One is a soft-sate management, and the other is a hard-state management. In the soft-state membership management, the LMME periodically refreshes the membership by the membership control message, that are similar to a hello protocol. The receivers processes (or branching point entities) that reply to the membership control message can keep the membership. In this strategy, we need not assume a reliable operation of each leaf entities. In the hard-state membership management, the LMME does not refresh the membership. Therefore, the leaf entities does not receive the membership control message, whenever they keep the membership. In order to terminate the membership, the leaf entity must send an explicit control message indicating the quitting from the multicast membership. Both membership management schemes should be supported. Soft-state membership management is preferable for relatively unreliable system. On the contrary, the hard-state membership management could be applied to the reliable system.

6.4 Further Scaling Up Strategy

The proposed multicast architecture described above will be able to support large scale multicast service, e.g., 10^6 receiver processes. However, for the further large scale multicast, the proposed architecture will not be able to provide a sufficient service quality. One would be due to the increase of round-trip delay between the sender process and the receiver process. The other would be due to the huge number of receiver processes (e.g., 10^{10}).

In order to support such a large scale multicast, the localized user packet preserving and packet retransmission for the errored/missing packets must be performed [Carle][Sanjoy]. Some dedicated branching point entities within the large multicast tress have the responsibility to retransmit the errored/missing packets.

7 Performance Evaluation of Error-Free Multicast Architecture over ATM Networks

In this section, the performance of the proposed error-free multicast service architecture is evaluated. The transmission quality of user packets (i.e., excluding control packets) associated with observed loss/error probability, the number of control packets sent from the receiver processes to the sender process and the transmission overhead (i.e., additional information and retransmitted information) is also evaluated in this section.

7.1 Evaluation Model

Figures 7-1 and 7-2 show the evaluated models. (a) in figure 7-1 is the multicast service without cell-level FEC policy and (b) in figure 7-2 is the proposed approach applying cell-level FEC policy in AAL-SSCS. IP multicast packets are transferred from the sender to N receivers. An IP packet is segmented into M ATM cells. The evaluation is by the pt-to-mpt, rather than by the mpt-to-mpt. This is because we could evaluate the performance of the mpt-to-mpt case through the performance of the pt-to-mpt case.

With cell-level FEC control, the IP packet will be segmented into FEC-SDU(s), and FEC Frame (FEC-SDU plus FEC Code) is segmented into f ATM cells. An ATM cell will be errored or lost at each data-link segment with a probability β. Without cell-level FEC control, an IP packet is errored when any cell belonging to the IP packet is errored or lost. FEC frame is errored, when the number of errored or lost cell are beyond FEC's correction capability. In the evaluation, it is assumed that, when more than one cell in FEC frame are errored or lost, the correct FEC-SDU can not be delivered. Then, the IP packet is errored. The detailed calculation methods are described in Appendices. Here, it is assumed that cells are randomly dropped or errored in the evaluation, even though cell will not dropped or errored at random in the actual system.

7.2 IP Packet Error or Loss Probability

Figures 7-3, 7-4 and 7-5 show the IP packet error or loss probability observed at the sender versus the number of receivers N. Approximately, $M = 1300$cells corresponds to 64 Kbyte IP packet, $M = 160$cells corresponds to 8 Kbyte IP packet and $M = 10$cells corresponds to 500 byte IP packet. The diameter of data-link segment d, which is the number of data-link segments between the sender and the receivers, is 5. FEC frame size f is 10 cells, i.e., FEC-SDU is 9 cells. Figure 7-3 is the case where the cell error or loss probability (β) in data-link segment is 10^{-3}, figure 7-4 is the case where β is 10^{-6} and figure 7-5 is the case where β is 10^{-9}.

For example, when the cell error or loss probability (β) in data-link segment is large (i.e., 10^{-3}), the IP packet error or loss probability without cell-level FEC policy is larger than 0.5 even for 10 receivers. On the contrary, with cell-level FEC policy, the IP packet error or loss probability is less than 0.05 even for 100 receivers. And, when the cell error or loss probability (β) in data-link segment is 10^{-6}, the IP packet error or loss probability without FEC policy is larger than 0.5 for 10^3 receivers with 160 cells IP packet size. With cell-level

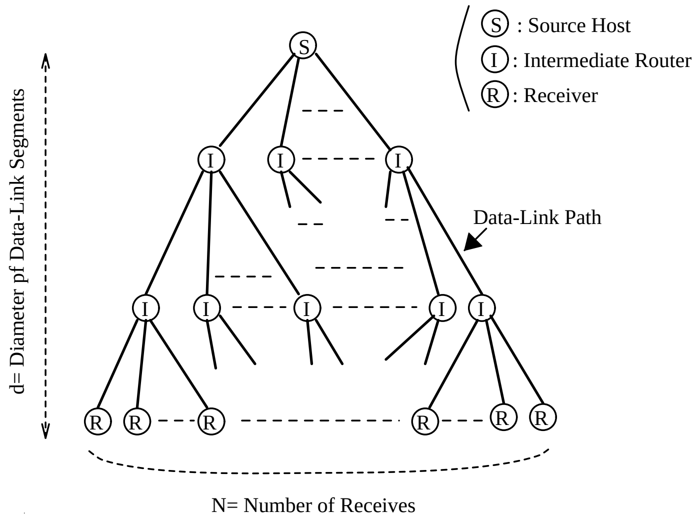


Figure 7-1. Evaluation Model (IP Multicast-Tree)

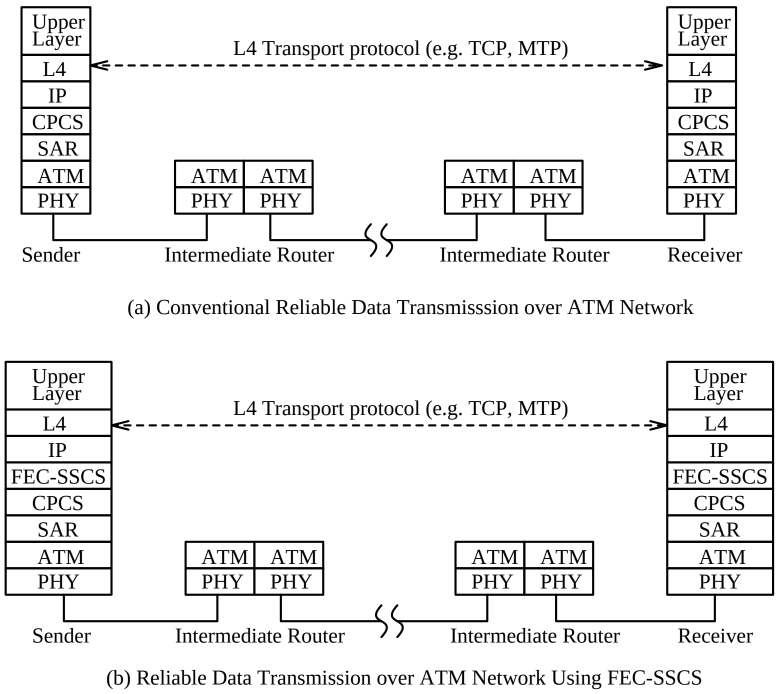


Figure 7-2. Evaluation Model (IP Multicast Protocol Structure)

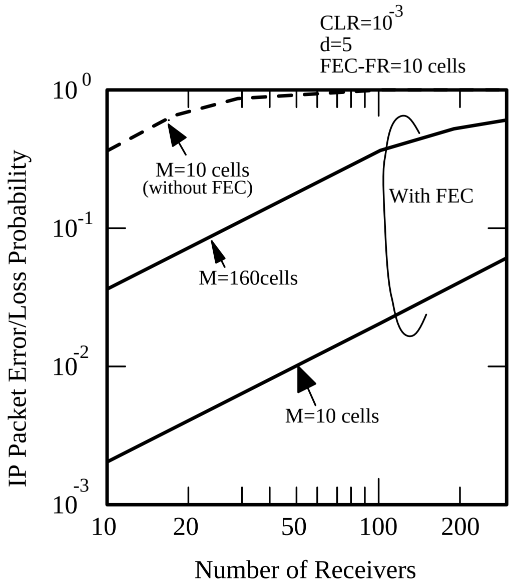


Figure 7-3. IP Packet Error/Loss Prob. vs. Number of Receivers ($CLR = 10^{-3}$)

FEC policy, the IP packet error or loss probability is less than 10^{-5} for 10^3 receivers, and it is less than 10^{-2} even for 10^6 receivers. These results indicate that the sender can take care of IP packet retransmission procedure without the intermediate TCP entities between the sender and the receivers, when we apply the cell-level FEC policy. Without the cell-level FEC policy, the intermediate TCP entities between the sender and the receivers seems to be required.

7.3 Control Packets

By the merging of the control packets (i.e., NACK packets) at the branching point entities, the maximum number control packets sent back to the sender process from the down-stream entities (i.e., receiver processes) can be reduced dramatically. The maximum number of returning control packet from the receiver processes without the NACK packet merging is N, that is the number of receiver processes. However, the maximum number of returning control packets from the branching point entities (or the receivers processes) is only m, that is the number of the corresponding leaf entities.

The required bandwidth for the returning NACK packets from the receiver processes for the proposed architecture would be about m/N compared to the required bandwidth without the merging of the NACK packets at the branching point entities. m would be decades and N would be a large number such as 10^4. Then, the required bandwidth for returning NACK packets for the proposed architecture is about 1/100 of the required bandwidth without control message merging at the branching entities.

For the reference, lets have a rough evaluation, when the positive acknowledgment policy is adopted. The control packet will be small and would be in one or two cells. On the contrary, the user packet would be in few hundreds cells. Usually, the control packet from

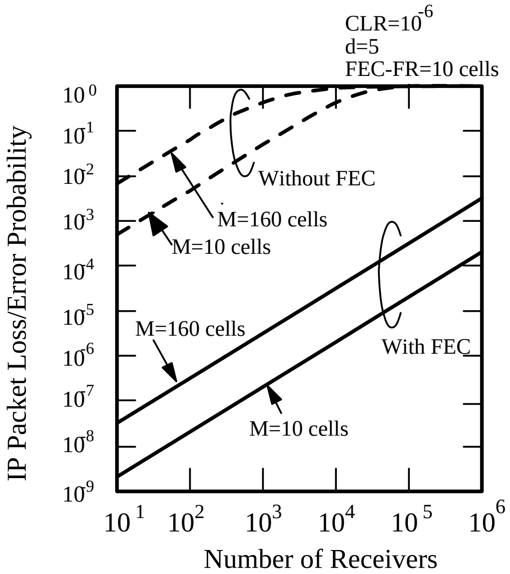


Figure 7-4. IP Packet Error/Loss Prob. vs. Number of Receivers ($CLR = 10^{-6}$)

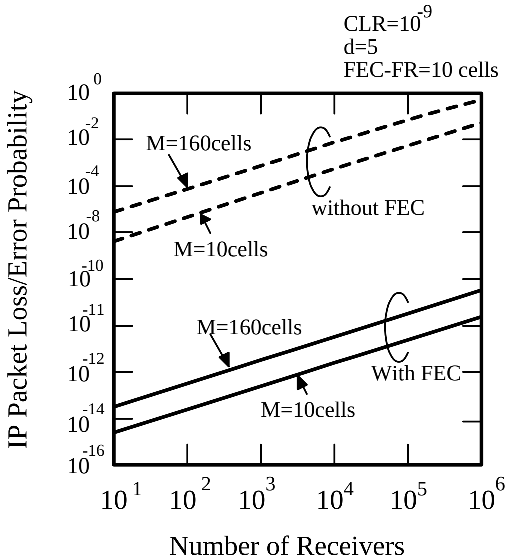


Figure 7-5. IP Packet Error/Loss Prob. vs. Number of Receivers ($CLR = 10^{-9}$)

the receiver process is not sent for every received packet, but would be sent, say, one control packet for ten received packets. In this case, the required bandwidth for the control packets for each receiver process is about 1/1000 of the required bandwidth for the user packets. This means that, for 10^4 receivers, the required bandwidth for the control packets will be ten time larger than the required bandwidth for the user packets. On the contrary, the required bandwidth for the control packets in the proposed architecture will be about 1/10 compared to the required bandwidth for the user packets, even when the number of leaf entities within the multicast sub-tree is 100.

7.4 Retransmission Overhead and FEC Overhead

In this subsection, the transmission overhead in order to provide an error-free packet delivery is evaluated. Transmission overhead is the following two factors.

- FEC code
  In order to perform the cell-level FEC policy, the FEC code is additionally transferred from the sender. Transmission overhead due to the cell-level FEC policy is given by $1/f$. For example, the transmission overhead is 10% for 10 cell size of FEC frame.

- Retransmission of errored or lost IP packet
  The errored or lost IP packet is subject to retransmission. Since TCP applies the go-back-N policy, the IP packets, which are correctly delivered to the receivers, would be subject to retransmission. However, in the evaluation, the retransmission of the IP packets, which are correctly delivered but are retransmitted due to the go-back-N policy, has not been taken into account. This means that the evaluated retransmission overhead corresponds to the case where the selective retransmission policy for errored or lost IP packet(s) is applied (i.e., IP packet retransmission policy of MTP in RFC1301).

Figures 7-6 and 7-7 show the transmission overhead in order to provide an error-free packet delivery versus the number of receivers N. The diameter of data-link segments d is 5 and the FEC frame size f is 10 cells. Figure 7-6 is the case where the cell error or loss probability at data-link segment β is 10^{-3}, and figure 7-7 is the case where β is 10^{-6}.

When the cell error or loss probability in data-link segment β is large (i.e., 10^{-3}), the transmission overhead without the cell-level FEC policy is about 100% even for *only 10 receivers* with 10 cells sized IP packet. On the contrary, the transmission overhead with the cell-level FEC policy can be always about 10%, which is almost same as the overhead due to the FEC code, *even for 100 receivers* with 10 cells sized IP packet. That means that the transmission overhead due to the IP packet retransmission can be sufficiently small. When the cell error or loss probability in data-link segment β is 10^{-6}, the retransmission overhead without the FEC policy is about 100% for 10^4 receivers with 10 cells sized IP packet. This means that throughput of this case will be less than 50%. On the contrary, the transmission overhead with the cell-level FEC policy can be again always about 10%. This means that the transmission overhead due to the IP packet retransmission can be sufficiently small, even for 10^6 receivers.

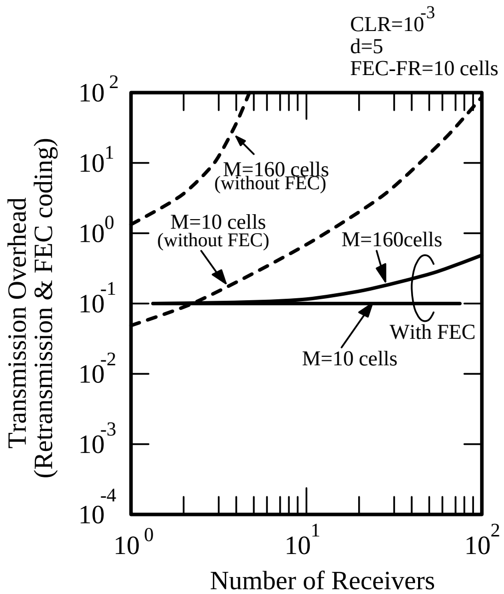


Figure 7-6. IP Packet Error/Loss Prob. vs. Transmission Overhead ($CLR = 10^{-3}$)

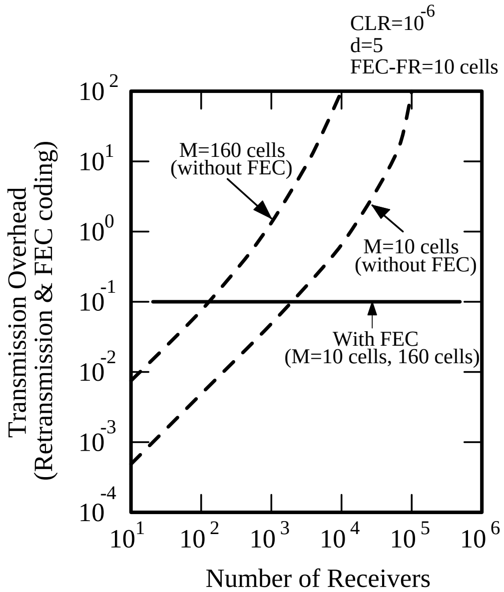


Figure 7-7. IP Packet Error/Loss Prob. vs. Transmission Overhead ($CLR = 10^{-6}$)

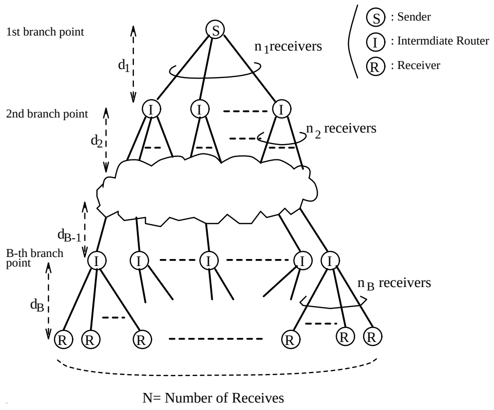


Figure 7-8. Evaluation Model of Link Sharing Effect

Since the probability of IP packet retransmission is sufficiently small with the cell-level FEC policy, we can provide a high throughput IP multicast packet delivery service with small latency by applying the cell-level FEC policy.

7.5 Discussion

7.5.1 Impact of Data-Link Sharing at Intermediate Links

The analytical evaluation above corresponds to the case where every IP packet delivery from the sender to receivers uses the separate paths. But, in the actual network, some links are shared, i.e., one down-stream IP packet flow will duplicated at the branching points (i.e., intermediate router). Also, some destinations (i.e., intermediate routers or end-receivers) could share the same physical transmission channel (e.g., pt-mpt connection without any server, e.g., Ethernet). Here, the number of destinations that shares a physical multicast channel could be usually less than 100.

Regarding the former factor, when $d = 1$, the expected packet error or loss probability could be the smallest. Let analyze the following case, shown in figure 7-8. Here, the detailed analysis is described in appendix B.4. The i-th ($1 \leq i \leq B$) intermediate router multicasts the received IP packet to n_i of destinations (i.e., receivers or intermediate routers). And the diameter of data-links between i-th branching point and $(i+1)$-th branching point is d_i. Here, the first branching point corresponds to the root of multicast tree.

Then, the packet error or loss probability (P_{share}) for sender is given by the following equations. Here, Q is the packet error or loss probability for d_i diameter's data-links, and r is the packet error or loss probability for a single data-link.

- Without FEC policy

$$P_{wofec} \simeq d \times N \times M \times \beta \tag{1}$$

$$P_{share} \simeq \sum_{k=1}^{k=B} [d_k \prod_{i=1}^{i=k} n_i] \times M\beta \tag{2}$$

$$< \quad d_B \times N \times M \times \beta \tag{3}$$

- With FEC policy

$$P_{wfec} \simeq d \times N \times m' \times (f-1)^2\beta^2 \tag{4}$$

$$P_{share} \quad \simeq \quad \sum_{k=1}^{k=B}[d_k \prod_{i=1}^{i=k} n_i] \times m' \times (f-1)^2\beta^2 \tag{5}$$

$$< \quad d_B \times N \times m' \times (f-1)^2\beta^2 \tag{6}$$

The difference is "$(\sum_{k=1}^{k=B}[d_k \prod_{i=1}^{i=k} n_i] < N \times d_B)$" of P_{share} and "$(N \times d)$" of P_{wodec} and P_{wfec}.
Here, $\sum_{k=1}^{k=B} d_k = d$ and $\prod_{i=1}^{k=B} n_i = N$.

As shown in the previous subsection, the contribution of d to the IP packet error or loss probability is small. Therefore, the contribution of data-link sharing in multicast tree could be also small as the evaluation results in the previous subsection.

Regarding the later factor, you can modify the number of receivers in the evaluation results based on the number of receivers sharing the physical multicast channel, which could be from 10 to 100. But, regarding ATM networks, many ATM networks may provide multicast service using a multicast server. In this case, the multicast service is provided by a multiple point-to-point connections between the server and every client. Therefore, when a multicast service is provided by a multicast server in the (ATM-based) data-link segment, the later factor is not effective, i.e., the evaluation in 7.2 could be true for the actual network.

7.5.2 Impact of FEC Policy for Point-to-Point Communication

This paper focuses on a multicast communication (pt-mpt) to realize that the cell level FEC policy can provide an error-free IP multicast service with a simple architecture. The cell level FEC policy could also achieve the benefits for a point-to-point (pt-pt) communication, as well as a multicast communication.

Either(both) when cell loss/error probability is not sufficiently small or(and) when IP packet size is large, the FEC policy will yield quite a higher throughput compared to the case without the FEC policy.

The IP packet loss/error probabilities are given by the following equations.

- Without FEC policy

$$P_{wofec} = 1 - (1-q)^d \tag{7}$$

where,

$$q = 1 - (1-\beta)^M \tag{8}$$

When β is sufficiently small (i.e., $\beta \ll 1.0$), P_{wofec} is almost equivalent to the following equation.

$$P_{wofec} \simeq d \times M \times \beta \tag{9}$$

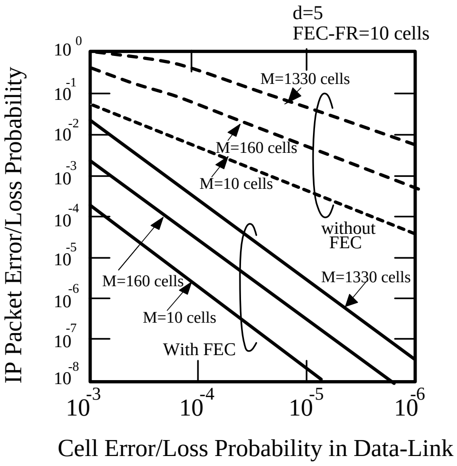


Figure 7-9. IP Packet Loss/Error Ratio for Point-to-Point Communication

- With FEC policy

$$P_{wfec} = 1 - (1-q)^d \quad (10)$$

where,

$$\begin{aligned} q &= 1-(1-s)^{m'} && (11)\\ s &= 1-(1-\beta)^{(f-1)} \times [1+(f-1)\beta] && (12)\\ m' &= \lceil M/f \rceil && (13) \end{aligned}$$

When β is sufficiently small (i.e., $\beta \ll 1.0$), P_{wfec} is almost equivalent to the following equation.

$$P_{wfec} \simeq d \times m' \times (f-1)^2 \beta^2 \quad (14)$$

Figure 7-9 shows the IP packet loss/error probability versus the cell error/loss probability of the each data-link segment for a point-to-point communication. Here, the FEC Frame size (f) is 10 cells, and the diameter of data-link segment (d) is 5.

When the IP packet size is large (e.g., 1330cells for 64Kbyte), the IP packet loss/error probability without the FEC policy will be large, that is larger than 0.6% at $\beta = 10^{-6}$. Since the throughput of TCP will be suddenly degraded around $10^{-2} - 10^{-3}$, it would be expected that we could not obtain a sufficient throughput at TCP level without the FEC policy, when IP packet size is large. On the contrary, the IP packet error/loss probability is 2.99×10^{-8} at $\beta = 10^{-6}$ with the FEC policy. Therefore, it is expected that we can obtain a sufficient throughput with the FEC policy, even for the large sized IP packet.

When β, cell loss/error probability in data-link segment, is not small (e.g., 10^{-3}), we could not obtain a sufficient throughput without the FEC policy. In general, for the large error/loss probability in data-link, IP packet size should be small. However, even when the IP packet size is 10 cells (for 500byte packet), the IP packet loss/error probability will be 4.8%.

With 4.8% of the IP packet loss/error probability, it is expected that the throughput at TCP level will be poor. On the contrary, with the FEC policy, the IP packet error/loss probability will be 2.24×10^{-4}. As a result, it is expected that we could obtain a sufficient throughput with the FEC policy, even when the cell loss/error probability in data-link segment is not small.

7.5.3 Performance of Cell-level FEC with Correlated Cell Loss

The analytical evaluation performed above is assuming random cell loss rather than correlated cell losses. There is a concerning that, with a correlated cell losses, the packet error probability will increase, compared to with a random cell losses. However, as shown in [96-0173], when we use an appropriate size of FEC frame and an appropriate number of redundant cells to the user cells, the expected packet error probability will be sufficiently small. [96-0173] shows the effectiveness of cell-level FEC for unicast service by a analytical evaluation and by a computer simulation.

The appropriate size of FEC frame and the number of redundant cells for each FEC frame depends on network characteristics (i.e., cell loss characteristics) and on the traffic pattern generated by each application. Therefore, these parameters (i.e., FEC frame size and the number of redundant cells) will be determined through the actual operation. Fortunately, the proposed FEC mechanism in this paper can easily modify these parameters session by session. As a result, the proposed cell-level FEC will achieve sufficiently small packet error probability even with correlated cell losses.

7.5.4 Implementation Complexity and Throughput

The error recovery algorithm of the proposed FEC method in this paper is based on the Reed-Solomon erasure code (RSE), which was originally proposed in [McAuley]. The RSE is based on the Reed-Solomon burst error correcting code (RSC) and uses the same polynomial division circuit employed by the RSC. The difference between these two code architecture is the decoding procedure. The RSC can correct both errors and erasures, but the RSE is designed to correct only erasures and to detect errors. Due to dropping error correction in the RSE, the required computation at the decoder can be much simpler and easier compared to the RSC. In the ATM environment, it could be said that the RSE has the same error characteristics as the RSC has. Regarding error correction for symbol erasures, the error correction capability of RSE is the same as RSC. Regarding error correction for symbol errors, the RSE can correct a half of erroneous symbols RSC can correct. The detailed discussion and evaluation can be found out in [95-1162].

As shown in [Tong], we can achieve 320 Mbps throughput encoding and decoding with RSC even with 1990 process. Since RSE requires less amount of gates and memory than RSE requires, we could implement the proposed FEC algorithm achieving sufficient throughput (i.e., few hundred Mbps).

Implementation through a software solution would also be possible. The throughput with software implementation may be less than that with hardware solution. With software processing solution, we will be able to achieve more than few Mbps. Actually, for example, [Bleszy] shows that ATM protocol (PHY, ATM and AAL)can be processed at 150 Mbps with a conventional software implementation. Even if the throughput with software processing is not high (e.g., few hundreds Kbps), the relatively slow error-free multicast channel is still very useful for many applications. This is because there will be many applications, that does not need such a high throughput error-free multicast service.

As a result, both cell-level FEC implementations with software and with hardware are useful for the applications that use the error-free multicast service. Also, it will be possible to achieve sufficient high throughput with a hardware implementation.

8 Conclusion

A framework of IP packet delivery architecture with high throughput and small latency using ATM technology in large scaled internets is proposed, while keeping the current subnet model. The proposed architecture is solving the two major issues that the Internet/intranet using ATM technology has to provide. One is a high throughput router architecture for the ATM platform while providing QOS-ed IP packet delivery services, and the other i s a architecture to provide a large scale error-free multicast service over the ATM platform.

CSR (Cell Switch Router), that has both cell switching fabric and packet switching fabric, architecture is proposed to achieve high throughput packet forwarding over the ATM platform. The CSR can forward some IP packets cell-by-cell based on VPI/VCI without examining IP header, rather than the conventional packet-by-packet forwarding. By this cut-thru IP packet forwarding, both resource reservation oriented IP packet flows (e.g., IP packet flow provided by RSVP) and non resource reservation oriented IP packet flows (i.e., best effort service) experience less packet delivery latency and obtain higher throughput, compared to the conventional hop-by-hop packet forwarding does. In order to perform the cut-thru IP packet forwarding using the cell relaying capability in the router, routers exchange the information how the IP packet flows are aggregated into ATM-VCC. This information exchanging is hop-by-hop base, and the cut-thru decision is a matter of every router's local decision. When all routers along the path, that an IP packet flow takes, perform the cell-relaying cut-thru, the soft-state and seamless cell-relaying channel is established to get a high throughput IP packet delivery with small latency. With keeping the current subnet model, even in the ATM networks, we can obtain soft-state oriented and scaleable QOS-ed high speed communication platform.

Also, a system architecture, that can provide a scaleable error-free multicast service over ATM networks, is proposed and evaluated. The proposed architecture is based on soft-state management in order to apply to the large scale multicast services. With the architecture framework discussed and proposed in this paper, the scaleable and high performed ATM platform for IP service (both unicast and multicast services) can be provided.

Acknowledgment

The author thanks to all persons who give valuable suggestions associated with this work. Especially, the author appreciates the suggestions by and discussions with Prof. Tadao Saito (University of Tokyo) , Prof.Hitoshi Aida (University of Tokyo), Prof.Shoichiro Asano (University of Tokyo), Prof. Mitsuru Ishizuka (University of Tokyo), Prof. Masao Sakauchi (University of Tokyo), Prof. Shoogo Ueno (University of Tokyo), Dr.Masataka Ohta (Tokyo Institute of Technology), Prof.Daniel Duchamp (Columbia University), Dr.David Clark (Laboratory of Computer Science, MIT), Dr.Georg Carle (University of Karlsruhe), Dr.Aloke Guha (Network Systems Corporation), Dr.Tim Dwight (MCI Telecommunication Corporation), Dr.Toshikazu Kodama (R&D Center, Toshiba Corporation), Dr.Takashi Kamitake (R&D Center, Toshiba Corporation), Mr.Yasuhiro Katsube (R&D Center, Toshiba Corporation), Dr.Yasuro Shobatake (R&D Center, Toshiba Corporation), Mr.Takeshi Saito (R&D Center, Toshiba Corporation), Mr.Ken-ichi Nagami (R&D Center, Toshiba Corporation) and Keiji Tsunoda (R&D Center, Toshiba Corporation). Also, the discussion on the electrical mailing lists (i.e., mailing lists at IETF and JAIN) has helped the exploration of this work.

References

[AF-TM] ATM Forum : "Traffic Management Specification version 4.0", August 1995

[AFUNI] ATM Forum : "ATM User Network Interface Specification version 3.1", September 1994

[ARIS] R.Woundy, A.Viswanathan, N.Feldman, R.Boivie : "ARIS: Aggregete Route-Based IP Switching", IETF Internet-Draft, draft-woundy-aris-ipswitching-00.txt, November, 1996.

[Ayanog] E.Ayanoglu, R.D.Gitlin, and N.C.Oguz : "Performance Improvement in Broadband Networks Using Forward Error Correction," Journal of High Speed Networks, pp.287-304, vol. 2, 1993

[Bhag] P.Bhagwat, P.Mishra, S.Tripathi : "Effect of Topology on Performance of Reliable Multicast Communication", Infocom94, 5b.1.1, June 1994.

[Biers] E.Biersack : "Performance Evaluation of Forward Error Correction in an ATM Environment", IEEE Journal on Selected Areas in Communication, Volume 11, Number 4, pp. 631-640, May 1993

[Birman] K.P.Beirman, T.A.Joseph : "Reliable communication in the presence of failures", ACM Transactions on Computer Systems, vol.5, No.1, February 1987

[Bleszy] R.Bleszynski : "Handling ATM Protocols with an Embedded CPU", Computer Design, January 1990.

[Carle] G.Carle : "Error Control for Reliable Multicast Communications in ATM Networks", ICCCN'94, September 1994

[Chang] J.Chang, N.F.Maxemchuk : "Reliable broadcast protocols", ACM Transactions on Computer Systems, vol.2, No.1, February 1987

[Comer] D.E.Comer : "Internetworking with TCP/IP", Prentice Hall, 1991

[Deering] S.Deering : "Host Extensions for IP Multicasting", IETF RFC1112, August 1989

[Esaki1] H.Esaki, Y.Tsuda, K.Kanai : "Evaluation of High Speed Multimedia Communication Architecture in ATM Networks", IEICE Transactions on Communications, Special Issue on Distributed Architecture for Next Generation Communication Networks, Vol.E77-B, No.11, November 1994.

[Esaki2] H.Esaki, K.Nagami, M.Ohta : "High Speed Datagram Delivery over Internet using ATM Technology", IEICE Transactions on Communications, Vol.E78-B No.8, August 1995

[Esaki3] H.Esaki, Y.Tsuda, T.Saito, S.Natsubori : "Class D Service Architecture in ATM-Internet", ICC'94, May 1994.

[Esaki4] H.Esaki, Y.Tsuda, T.Saito, S.Natsubori : "Datagram Delivery in an ATM-Internet", IEICE Transactions on Communications, Special Issues on Future Private Networks, Vol.E.77-B, No.3, March 1994.

[GJA] G.Armitage : "IP Multicast over UNI 3.0 based ATM Networks", (Internet-Draft) draft-armitage-ipatm-ipmc-03.txt, January, 1995.

[IPng] F.Kastenholz, C.Partridge : "Technical Criteria for Choosing IP:The Next Generation (IPng)", IETF Internet-Draft, draft-kastenholz-ipng-criteria-02.txt, May 1994.

[Ipsilon] P.Newman, T.Lyan, G.Minshall : "Flow labelled: Connectionless ATM Under IP", Engineer Conference, Networld+Interop'96 Las Vegas, April, 1996.

[IPv6] S.Deering, R.Hinden : "Internet Protocol, Version 6 (IPv6), Specification", IETF Internet-Draft, draft-ietf-ipngwg-ipv6-spec-01.txt, March 1995

[I.150] ITU-T Recommendation I.150 : "B-ISDN Asynchronous Transfer Mode", 1990

[I.363] ITU-T Recommendation I.363 : "B-ISDN ATM Adaptation Layer (AAL) Specification", 1993

[I.364] ITU-T Recommendation I.364 : "Support of Broadband Connectionless Data Service on BISDN", 1992

[McAuley] A.J.McAuley : "Reliable Broadband Communication Using a Burst Erasure Correcting Code", ACM SIGCOMM90, September 1990

[LANE] ATM Forum : "LAN Emulation Over ATM Specification version 1.0", April 1995 Sept., 1990.

[NBMA] J.Heinanen : "NBMA Address Resolution Protocol", IETF RFC1735, December 1994

[NHRP] D.Katz and D.Piscitello : "NBMA Next Hop Resolution Protocol (NHRP)", IETF Internet-Draft, draft-ietf-rolc-nhrp-03.txt, November 1994

[Ohta] H. Ohta and T.Kitami : "A Cell Loss Recovery Method Using FEC in ATM Networks," IEEE JSAC, vol. 9, pp. 1471-1483, December 1991

[PNNI] ATM Forum P-NNI SWG : "P-NNI Draft Specification", ATM Forum Technical Contribution 94-0471R8, June 1995

[RFC791] ISI-USC : "Internet Protocol : DARPA Internet Program Protocol Specification", IETF RFC791, September 1981

[RFC1144] V.Jacobson : "Compressing TCP/IP Headers", IETF RFC1144, January 1990

[RFC1301] S.Armstrong, A.Freier, K.Maezullo : "Multicast Transport Protocol", IETF RFC 1301, February 1992

[RFC1331] W.Simpson : "The Point-to-Point Protocol", IETF RFC1331, May 1992

[RFC1458] R.Brudes, S.Zabele : "Requirements for Multicast Protocols", IETF RFC1458, May 1993

[RFC1483] J.Heinanen : "Multiprotocol Encapsulation over ATM Adaptation Layer 5", IETF RFC1483, July 1993

[RFC1577] M.Laubach : "Classical IP and ARP over ATM", IETF RFC1577, October 1993

[RFC1626] R.Atkinson : "Default IP MTU for use over ATM AAL5", IETF RFC1626, May 1994

[RFC1633] R.Braden, D.Clark, S.Shenker : "Integrated Services in the Internet Architecture ; an Overview", IETF RFC1633, June 1994

[RFC1937] Y.Rekhter and D.Kandlur : "Local/Remote Forwarding Decision in Switched Datalink Subnetworks", IETF RFC1937, May, 1996.

[RFC2098] Y.Katusbe, K.Nagami, H.Esaki : " Toshiba's Router Architecture Extensions for ATM : Overview", IETF RFC2098, February, 1997.

[RFC2105] Y.Rekhter, B.Davie, D.Katz, E.Rosen, G.Swallow : "Cisco System's Tag Switching Architecture Overview", IETF RFC2105, February, 1997.

[RFC2129] K.Nagami, Y. Katsube, Y. Shobatake, A. Mogi, S. Matsuzawa, T. Jinmei, H. Esaki : "Toshiba's Flow Attribute Notification Protocol (FANP) Specification", IETF RFC2129, April, 1997.

[Roman] Romanov, A. : "Some Results on the Performance of TCP over ATM", Second IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems HPCS'93, Williamsburg, Virginia, U.S.A., September 1993

[RSVP] L.Zhang, R.Braden :"Resource ReSerVation Protocol (RSVP)", IETF Internet-Draft, draft-ietf-rsvp-spec-05.txt, March 1995

[RTP] H.Schulzrinne, S.Casner, R.Frederick, V.Jacobson : "RTP : A Transport Protocol for Real-Time Application", IETF Internet-Draft, draft-ietf-avt-rtp-07.txt, March 1995

[Sanjoy] Sanjoy Paul, K.Sabnani, D.Kristol : "Multicast Transport Protocols for High Speed Networks", International Conference on Network Protocols, October 1994

[Schmit] A.Schmidt, R.Campbell : "Internet Protocol Traffic Analysis with Applications for ATM Switch Design", Computer Communications Review, vol.23, No.2, pp.39-52, April 1993

[Shenk1] S.Shenker, C.Partridge : "Specification of Controlled Delay Quality of Service", IETF Internet-Draft, draft-ietf-intserv-controlled-delay-svc-01.txt, July 1995

[Shenk2] S.Shenker, C.Partridge : "Specification of Predictive Quality of Service", IETF Internet-Draft, draft-ietf-intserv-predictive-svc-00.txt, March 1995

[Shenk3] S.Shenker, C.Partridge : "Specification of Guaranteed Quality of Service", IETF Internet-Draft, draft-ietf-intserv-guaranteed-svc-01.txt, July 1995

[SRM] S.Floyd, V.Jacobson, S.McCane : "A Reliable Multicast Framework for Light-weight Session and Application Level Framing", ACM SIGCOMM95, pp.342-356, September, 1995.

[Tong] P.Tong : "A 40-MHz Encoder-Decoder Chip Generated by a Reed-Solomon Code Compiler", IEEE Custom Integrated Circuits Conference, May, 1990.

[ST-II] L.Delgrossi and L.Berger : "Internet STream Protocol Version 2 (ST-II)", IETF RFC1819, August 1995

[TDP] P.Doolan, B.Davie, D.Katz, Y.Rekhter, E.Rosen : "Tag Distribution Protocol", IETF Internet-Draft, draft-doolan-tdp-spec-00.txt, September, 1996.

[94-0914] T.Chen, L.Jones, S.Liu, and V. K.Samalam : "Effect of ATM Cell Loss on TCP Packet Loss", ATM Forum technical contribution 94-0914, September 1994

[95-0150] H.Esaki, K.Kanai, Y.Tsuda : "Necessity of Cell Level FEC Scheme for Data Transmission Service", ATM Forum technical contribution 95-0150, February 1995

[95-0151] H.Li, K.-Y.Siu, H.-Y.Tzeng : "IPX and TCP Performance over ATM Networks with Cell Loss", ATM Forum technical contribution 95-0151, February 1995

[95-0325] G.Carle, A.Guha, H.Esaki, K.Tsunoda, K.Kanai : "Necessity of an FEC Scheme for ATM Networks", ATM Forum Technical Contribution, 95-0325, April 1995

[95-0326] G.Carle, A.Guha, H.Esaki, K.Tsunoda, K.Kanai : "Draft Proposal for Baseline-text of FEC-SSCS for AAL5", ATM Forum Technical Contribution, 95-0326, April 1995

[95-1162] K.Tsunoda, K,Kanai, H.Esaki : "A Reed-Solomon Erasure Code and Its Application to AAL", ATM Forum Technical contribution, 95-1162, October, 1995

[96-0173] A.Guha, T.S.Chang : "Network Conditions Under Which FEC is Effective", ATM Forum Technical Contribution, 96-0173, February, 1996

A FEC-SSCS Specification

A.1 FEC Frame Format

The format of FEC frame, that is actually transmitted from the source entity to the destination entity is described in Figure A-1.

The FEC frame has user-data-part and FEC-code-part. Each part has the FEC-frame-header field. The user-data-part contains the part (or whole) of FEC-SSCS-SDU with the attached FEC-frame header field. The FEC-code-part contains the FEC redundant information field with the attached header field.

- user-data field
  The user-data field is used to carry the part (or whole) of FEC-SSCS-SDU. The "n", the length in vertical direction, is determined so as that the length of the vertical line in the FEC frame, {(t,1),(t,2),(t,3),...(t,n)}, is 46/a octet. Here, "a" is a integer more than zero and it is usually one. In other words, the length of the vertical line in FEC frame is usually 48 octet, that is the payload length of AAL5 SAR-PDU. The maximum length of horizontal line in the user-data must be less than $[2^p - q]$ symbols. Here, p is the symbol size in bit, and q is the horizontal length of FEC-code-part (s).

- FEC redundant data field
  The FEC redundant data field is used to carry the appended data to protect the user-data field information against bit error and cell loss. The data in FEC redundant data field is calculated based on the defined FEC algorithm with the negotiated parameter (i.e., p and q). The calculation algorithm of the FEC redundant data vector, {(M+1,t),(M+2,t),...(M+S,t)}, is specified based on the Reed-Solomon error recovery algorithm.

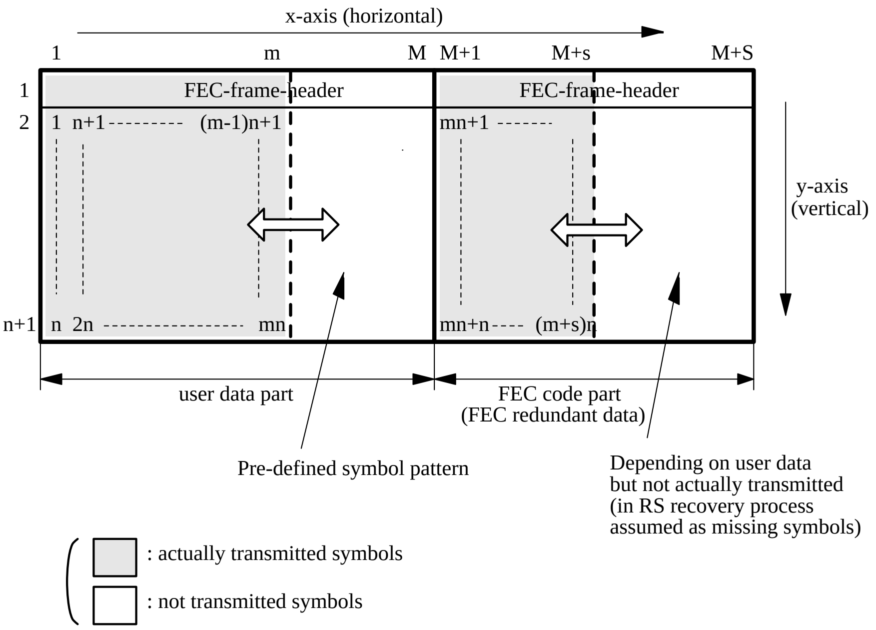
- FEC-frame-header field
  The FEC-frame-header field is used to identify the position of missing or errored vertical line associated x-axis and to detect the bit error in each vertical line. The length of the FEC-frame-header is 2 octets in each vertical line. The detailed format and coding rule of the FEC-frame-header field is specified below.

The writing order and the reading order have the same order, to avoid the re-ordering procedure at the destination FEC-SSCS entity that causes some processing latency.

Each horizontal line (i.e., the horizontal vector) in the user-data-part corresponds the user data protected by the FEC redundant information, that located at the same vertical position in the FEC-code-part. "m" of symbols {(k,1),(k,2),....(k,m)} ($m \leq M$) in data-part are protected by the "s" of symbols {(k,M+1),(k,M+2),...(k,M+s)} ($s \leq S$) in FEC-code-part.

A.2 Format and coding rule of FEC-frame-header field

The function of the FEC-frame-header field is as followed.



[Writing Order]

(1,2), (1,3), (1.4), ... (1,n+1),
(2,2), (2,3), (2,4), ... (2,n+1),
(3,2), (3,3), (3,4), ... (3,n+1),
: : :
(m,2), (m,3), (m,4), ... (m,n+1),
(M+1,2), (M+1,3), ... (M+1,n+1),
(M+2,2), (M+2,3), ... (M+2,n+1),
: : :
(M+s,2), (M+s,3), ... (M+s,n+1).

[Reading Order]

(1,1), (1,2), (1,3), (1.4), ... (1,n+1),
(2,1), (2,2), (2,3), (2,4), ... (2,n+1),
(3,1), (3,2), (3,3), (3,4), ... (3,n+1),
: : :
(m,1), (m,2), (m,3), (m,4), ... (m,n+1),
(M+1,1), (M+1,2), (M+1,3), ... (M+1,n+1),
(M+2,1), (M+2,2), (M+2,3), ... (M+2,n+1),
: : :
(M+s,1), (M+s,2), (M+s,3), ... (M+s,n+1).

Figure A-1. FEC Frame Format

1. Identification of the position of the errored or missing vertical lines, associated with horizontal direction.
   In order to recover the errored or missing symbols in the user-data-part, the FEC algorithm specified in this paper requires the exact position of the vertical line with bit error or symbol missing.

2. Error detection in the vertical line.
   In order to improve the error correction capability by the FEC algorithm, the FEC algorithm specified in this draft requires the detection of bit error in the vertical line, as well as requires the exact position of the vertical line with bit error. In SLC mode operation, this function can be skipped.

In order to provide the above function, the following field and coding rule is provided. All fields exist in each vertical line in the FEC frame. The length of FEC-frame-header field of each vertical line is 2 octets.

1. Sequential Number (SN) field
   The SN field is used to identify the position of the vertical line associated with horizontal direction. When the length of SN field is "q", 2^q must be larger than FCPL, the number of vertical lines in FEC-code-part. The SN of the first vertical line both in user-data-part and in FEC-code-part always starts from zero (0).

2. User/FEC (U/F) field
   The U/F field is used to identify which part, i.e., user-data-part or FEC-code-part, the received vertical line belongs to. The U/F field has one bit and takes two values. "U" represents user-data-part, and "F" represents FEC-code-part. The U/F field in user-data-part is always "U", and the U/F field in FEC-code-part is always "F".

3. Parity (P) field
   The P field is used to perform the parity check for the FEC-frame-header field excepting the CRC field. The purpose of the P bit is to detect the bit error in the FEC-frame-header field. The one bit error in the FEC-frame-header field excepting CRC field can be detected.

4. CRC field
   The CRC field is used to detect the bit error in the vertical line.

In figure A-2 , the format of the FEC-frame-header field and the example of field pattern is described.

A.3 FEC Frame Mapping to CPCS-PDU

In order to avoid the large padding field in the CPCS-PDU for the transmission of FEC-code field, the following FEC frame transmission method is applied. If the FEC frame (will have 48 octets alignment) is transmitted as a single CPCS-SDU, each CPCS-SDU have 40 octets padding field.

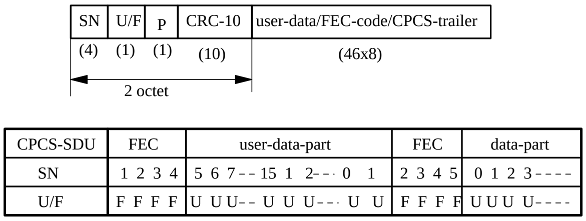


Figure A-2. FEC-frame-header Field

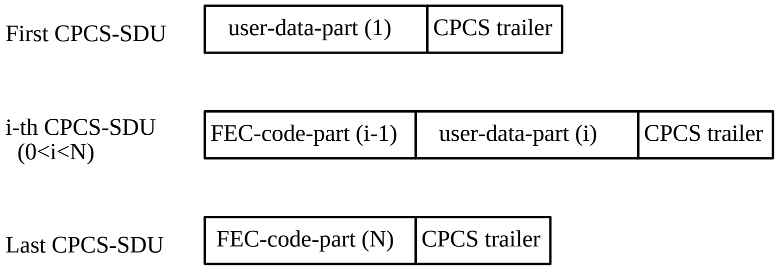


Figure A-3. FEC Frame and CPCS-SDU

By the method shown in figure A-3, only the last CPCS-SDU will have a large padding field. The padding field length of other CPCS-SDUs depends on the length of user-data-part, that is generally variable length. The CPCS-SDU, excepting the first and the last CPCS-SDU in the burst, includes the i-th user-data-part of the (i-1)-th FEC-code-part.

A.4 FEC Algorithm

A.4.1 Symbol Length, FEC-frame Size, and Correction Capability

In the FEC frame specified above, one Read-Solomon block (RS-block) corresponds to one horizontal line. Therefore, the FEC frame has n of RS-blocks.

The FEC uses a Reed-Solomon (M, M+S) code which is able to correct up to [S/2] errored symbols or s erased (missing) symbols. Here, m is equal to or smaller than M, and s is equal to or smaller than S. m and s is the number of vertical lines in FEC frame. m is the number of vertical lines for data-part, and s is the number of vertical lines for FEC-code-part. m+s vertical lines are actually transmitted from the source entity to the destination entity.

The RS(M, M+S) code can correctly recover the received data when the following equation is satisfied.

$$a + 2b \leq S \tag{15}$$

Here, "a" is the number of erased (missing) symbols. And, "b" is the number of errored symbols, whose position in the FEC frame is not identified. Since the specified FEC frame has the CRC's bit error detection capability in each vertical line, the actual error correction capability of the FEC algorithm based on RS(M,M+S) is as followed.

$$a + b \leq S \tag{16}$$

The symbols in a vertical line are assumed to be erased when bit errors occurred in this cell. This means that the erasure mode is used for the correction of dummy symbols corresponding to cell loss locations.

Here, if the total number of vertical lines with bit errors is larger than "s", then the RS-Code may be used for bit error detection in every row of symbols. In this case, the FEC frame can be recovered as long as not more than s/2 symbols per row are errored. Whether to implement the RS-Code's bit error detection is not mandatory and depends on vendor's decision. Also, the receiving entity of FEC-SSCS entity may use the error correction capability of CRC-10 to correct a single bit error per cell.

Reed-Solomon codes to be used are built over Galois Field (2^N), and the generator polynomial $G(X)$ is given by :

$$G(X) = \prod_{i=0}^{S-1} (X - \alpha^{i+k}) \tag{17}$$

where α is a root of the primitive polynomial, N is the size of symbol in bits, and k is the base exponent of the generator polynomial.

Since the symbol length is determined by (i) feasibility of implementation, and (ii) the allowable maximum number of (m+s). From the view point of implementation feasibility, the symbol length should be 4 bits, 8 bits, 16 bits, 32 bits and 64 bits. The allowable maximum number of symbols in RS-block is given by $2^N - 1$. Here, N is the symbol length in bit. For the current IP (IPv4), 4 bits symbol or 8 bit symbol would be preferable. For an native ATM environment, 8 bit symbol or 16 bit symbol would be preferable. 16 bits symbol can cover the maximum SDU size of AAL type 5 by the single FEC frame. And, 8 bits symbol can cover the default MTU for ATM by the single FEC frame.

A.4.2 Calculation of FEC Redundant Symbols

The FEC redundant symbols are calculated by the specified generator polynomial and user-data symbols.

- Generator Polynomial : $G(X)$
- RS-block : $\{C[[M+S-1], C[M+S-2], ...C[S], C[S-1], ...C[1], C[0]\}$
- User-data : $\{U[m], U[m-1], ...U[1]\}$
- FEC-data : $\{F[s-1], F[s-2], ...F[0]\}$

The transmitted data is User-data $\{U[m], U[m-1], ...U[1]\}$ and FEC-data $\{F[s-1], F[s-2], ...F[0]\}$. m is smaller than or equal to M, and s is smaller than or equal to S.

$C[M+s-i] (1 \leq i \leq M)$ is give by the following equation.

$$C[M+s-i] = \begin{cases} U[m-i+1] & (1 \leq i \leq m) \\ P & (m < i \leq M) \end{cases} \tag{18}$$

Here, symbol P is pre-determined symbol, that is shared between the source and destination FEC-SSCS entities. The symbols $C[i](m < i \leq M)$ are only used for the calculation of FEC-data, and they are not transmitted to the destination FEC-SSCS entity.

$\{C[S-1], C[S-2], ...C[0]\}$ is given by the following equation.

$$M(X) \times X^S = Q(X) \times G(X) + C[S-1] \times X^{S-1} + C[S-2] \times X^{S-2} + ...C[1] \times X + C[0] \tag{19}$$

Here,

$$M(X) = C[M+S-1] \times X^{M-1} + C[M+S-2] \times X^{M-2} + ...C[S+1] \times X + C[S] \tag{20}$$

$F[i]$ $(0 \leq i \leq s-1)$ is give by the following equation.

$$F[s-i] = C[S-i](1 \leq i \leq s) \tag{21}$$

A.4.3 FEC Error Recovery Algorithm

The correct symbols are obtained by the following equation, when the received symbol vector is $\{R[m+s-1], R[m+s-2], ...R[s], R[s-1], ...R[1], R[0]\}$. The received symbol vector have missing symbols and symbols with bit errors. The positions of missing symbols and symbols with bit errors can be identified by the sequence number (SN). Here, the errored/missing symbols in user-data part is represented by $\{EU[1], EU[2], ...EU[q]\}$ $(1 \leq q \leq m)$, and the errored/missing/un-transmitted symbols in FEC-code part is represented by $\{EF[1], EF[2], ...EF[r]\}$ $(S-s \leq r \leq S)$. $EF[i]$ $(1 \leq i \leq S-s)$ is the symbol in FEC-code part, that is not actually transmitted from the source entity, and $EF[i]$ $(S-s \leq i \leq r)$ is the symbols in FEC-code part, that is errored/missed during cell transmission.

The symbols in FEC-code part, that is not actually transmitted from the source entity, $EF[i]$ $(1 \leq i \leq S-s)$ are assumed as the missing/errored symbols in FEC error recovery algorithm. FEC error recovery algorithm is not triggered when there is no errored/missing symbols in user-data part, even when there is errored/missing/un-transmitted symbol in FEC-code part.

When $q+r$ is equal to or smaller than S, the errored user date $\{EU[1], EU[2], ...EU[q]\}$ can be correctly recovered. The errored/missing symbols, $R[x_i]$ $(1 \leq i \leq q+r)$, are the un-known variables, and they are obtained by the following equation.

$$R(X) = Q(X) \times G(X) \tag{22}$$

Here,

$$R(X) = \sum_{i=1}^{m} R[m+s-i] \times X^{M+S-i} + \sum_{i=m+1}^{M} P \times X^{M+S-i} - \sum_{i=1}^{s} R[s-i] \times X^{S-i} - \sum_{i=s+1}^{S} EF[i-s] \times X^{S-i} \tag{23}$$

The generator polynomial ($G(X)$) and the pre-determined symbol (P) are shared between source and destination FEC-SSCS entities. Since the $G(X)$ has "S" independent solutions, $\{a[1], a[2], ...a[S]\}$, the following "S" equations are obtained.

$$R(a[i]) = 0 \tag{24}$$

where, $G(a[i]) = 0$ $(1 \leq i \leq s)$.

Since the length of user is indicated by LI (Length Indicator) in the CPCS-trailer, the destination FEC-SSCS entity can identify the originally transmitted user-data $\{U[m], U[m-1], ...U[1]\}$ $(m \leq M)$.

B IP Packet Error or Loss Probability

B.1 IP Packet Error or Loss Probability without FEC

The following is the parameters to evaluate the IP packet error or loss probability without the FEC policy.

- L2 data-unit error or loss probability in data-link segment : β
  (corresponding to Cell Loss Ratio in ATM network)
- Diameter of data-link segments : d
- Number of L2 data-unit's in one IP packet : M units
  (in ATM networks, it is M cells)
- Number of receivers : N
- Packet error or loss probability in data-link : q
- Packet error or loss probability for receiver : p
- Packet error or loss probability for sender : P_{wofec}

The packet error or loss probability in data-link (q) is given by the following equation, when the cell error or loss is occurred in random with the probability β.

$$q = 1 - (1 - \beta)^M \tag{25}$$

Then, the packet error or loss probability for sender (P_{wofec}) is given by the following equation.

$$P_{wofec} = 1 - (1 - p)^N \tag{26}$$

When β is sufficiently small (i.e., $\beta \ll 1.0$), the equations are almost equivalent to the followings.

$$\begin{aligned} q &\simeq M \times \beta && (27) \\ p &\simeq d \times M \times \beta && (28) \\ P_{wofec} &\simeq N \times d \times M \times \beta && (29) \end{aligned}$$

B.2 IP Packet Error or Loss Probability with FEC

The FEC capability is indicated by "f". This means that when the error or loss L2-data-units within f L2-data-units (called as FEC-Frame) is less than two, the IP packet is correctly transferred.

Therefore, the error or loss probability of FEC-Frame, "s", is given by the following equation.

$$\begin{aligned} s &= 1 - [(1-\beta)^f + f \times \beta(1-\beta)^{(f-1)}] && (30)\\ &= 1 - (1-\beta)^{(f-1)} \times (1 - \beta + f\beta) && (31)\\ &= 1 - (1-\beta)^{(f-1)} \times [1 + (f-1)\beta] && (32) \end{aligned}$$

Let m' is $\lceil M/f \rceil$, which corresponds to the number of FEC-Frames in one IP packet. Then, q, p and P_{wfec} are given by the following equations. Here, P_{wfec} is the packet error or loss probability for sender, when FEC policy is applied to.

$$\begin{aligned} q &= 1 - (1-s)^{m'} && (33)\\ p &= 1 - (1-q)^d && (34)\\ P_{wfec} &= 1 - (1-p)^N && (35) \end{aligned}$$

When β is sufficiently small (i.e., $\beta \ll 1.0$), the equations are almost equivalent to the followings.

$$\begin{aligned} s &\simeq (f-1)^2 \times \beta^2 && (36)\\ q &\simeq m' \times (f-1)^2 \times \beta^2 && (37)\\ p &\simeq d \times m' \times (f-1)^2\beta^2 && (38)\\ P_{wfec} &\simeq N \times d \times m' \times (f-1)^2\beta^2 && (39) \end{aligned}$$

Therefore, the packet error or loss probability (P_{wfec}) will be about $(f-1) \times \beta$ times compared to P_{wofec}, that is for without FEC policy.

$$P_{wfec} \simeq (f-1) \times \beta \times P_{wofec} \qquad (40)$$

B.3 Re-transmission Overhead to Provide Error-Free Delivery

When the packet error or loss probability at sender is P, the probability (α_k), that IP packet is successfully transferred by the k-th packet transmission (excluding first IP packet transmission), is given by the following equation.

$$\begin{aligned} \alpha_k &= P \times P^{(k-1)}(1-P) && (41)\\ &= (1-P)P^k && (42) \end{aligned}$$

Then, the expected re-transmission number (R) is given by the following equation.

$$R = \sum_{k=1}^{\infty} k \times \alpha_k \qquad (43)$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} k \times (1-P)P^k && (44)\\
&= (1-P)\sum_{k=1}^{\infty} kP^k && (45)\\
&= (1-P) \times P/(1-P)^2 && (46)\\
&= P/(1-P) && (47)
\end{aligned}
$$

B.4 Impact of Data-Link Sharing at Intermediate Links

The evaluation model is shown in figure 16. The packet error or loss probability (P_{share}) for the sender is given by the following equation. Here, Q is the packet error or loss probability for the d_i diameter's data-links, and r is the packet error or loss probability for a single data-link segment.

$$P_{share} = \prod_{k=1}^{k=B} \{1-(1-Q_k)^{\prod_{i=1}^{i=k} n_i}\} \quad (48)$$

$$Q_k = 1-(1-r)^{d_k} \quad (49)$$

Here,

$$
\begin{aligned}
r &= 1-(1-\beta)^M \qquad \text{(for without FEC)} && (50)\\
r &= 1-(1-s)^{m'} \qquad \text{(for with FEC)} && (51)\\
s &= 1-(1-\beta)^{(f-1)} \times [1+(f-1)\beta] && (52)
\end{aligned}
$$

When, β is sufficiently small ($\beta \ll 1.0$), the above equations can be almost equivalent to the followings.

$$
\begin{aligned}
P_{share} &\simeq \sum_{k=1}^{k=B} \{Q_k \prod_{i=1}^{i=k} n_i\} && (53)\\
&\simeq \sum_{k=1}^{k=B} \{r \times d_k \times \prod_{i=1}^{i=k} n_i\} && (54)
\end{aligned}
$$

Here,

$$
\begin{aligned}
Q_k &\simeq d_k \times r && (55)\\
r &\simeq M\beta \qquad \text{(for without FEC)} && (56)\\
r &\simeq m' \times s \qquad \text{(for with FEC)} && (57)\\
s &\simeq (f-1)^2 \times \beta^2 && (58)
\end{aligned}
$$

Then,

$$P_{share} \simeq \sum_{k=1}^{k=B}\{d_k \prod_{i=1}^{i=k} n_i\}M\beta \qquad \text{(for without FEC)} \tag{59}$$

$$P_{share} \simeq \sum_{k=1}^{k=B}\{d_k \prod_{i=1}^{i=k} n_i\}m'(f-1)^2\beta^2 \qquad \text{(for with FEC)} \tag{60}$$

Now, let compare with the evaluation in B.1 and B.2.

- Without FEC policy

$$P_{wofec} \simeq d \times N \times M \times \beta \tag{61}$$

$$\begin{aligned} P_{share} &\simeq \sum_{k=1}^{k=B}\{d_k \prod_{i=1}^{i=k} n_i\} \times M\beta & (62) \\ &< d_B \times N \times M \times \beta & (63) \end{aligned}$$

- With FEC policy

$$P_{wfec} \simeq d \times N \times m' \times (f-1)^2\beta^2 \tag{64}$$

$$\begin{aligned} P_{share} &\simeq \sum_{k=1}^{k=B}[d_k \prod_{i=1}^{i=k} n_i] \times m' \times (f-1)^2\beta^2 & (65) \\ &< d_B \times N \times m' \times (f-1)^2\beta^2 & (66) \end{aligned}$$

List of Acronyms

- AAL : ATM Adaptation Layer
- ABR : Available Bit Rate
- ACK : Acknowledgment
- ARP : Address Resolution Protocol
- ATM : Asynchronous Transfer Mode
- BER : Bit Error Ratio
- BGP : Border Gateway Protocol
- BISDN : Broadband Integrated Digital Service Network
- BT : Burst Tolerance
- CAC : Connection Admission Control
- CBR : Constant Bit Rate
- CBT : Core Base Tree
- CDV : Cell Delay Variation
- CDVT : CDV Tolerance
- CIDR : Classless Inter-Domain Routing
- CLNP : ConnectionLess Network Protocol
- CLP : Cell Loss Priority
- CLPF : CLassical Packet Forwarding
- CLR : Cell Loss Ratio
- CLS : ConnectionLess Server
- CPN : Customer Premises Network
- CPCS : Common Part Convergence Sub-layer
- CSR : Cell Switch Router
- CTD : Cell Transfer Delay
- DNS : Domain Name Server
- ERD : Early Random Discarding
- FANP : Flow Attribute Notification Protocol
- FEC : Forward Error Correction
- FDDI : Fiber Distributed Data Interface
- FFOL : FDDI Flow On LAN
- FIFO : First In First Out
- HOL : Head of Line
- ICMP : Internet Control and Management Protocol
- IETF : Internet Engineering Task Force
- IGMP : Internet Group Management Protocol
- IGP : Interior Gateway Protocol
- IHI : Internet Header Length
- ION : IP Over NBMA
- IP : Internet Protocol
- IPv4 : Internet Protocol version 4
- IPv6 : Internet Protocol version 6
- IPX : Internet Packet eXchange

- ISP : Internet Service Provider
- ITU-T : International Telecommunication Union - Telecommunication Standardization Sector
- LAN : Local Area Network
- LIS : Logical IP Subnet
- LLC : Logical Link Control
- LSTL : Label Switch with Transparent Links
- LMME : Local multicast membership management entity
- MAC : Media Access Control
- M-Bone : Multicast Back-Bone
- MCR : Minimum Cell Rate
- MTP : Multicast Transport Protocol
- MTU : Maximum Transfer Unit
- NACK : Negative Acknowledgment
- NBMA : Non-Broadcast Multiple Access
- NHRP : Next Hop Resolution Protocol
- NSAP : Network Service Access Point
- OSI : Open System Interconnection
- OSPF : Open Shortest Path First
- PCR : Peak Cell Rate
- PDU : Protocol Data Unit
- PDV : Packet Delay Variance
- PIM : Protovol Independent Multicast
- P-NNI : Private Network to Network Interface
- PPD : Partial Packet Discarding
- PPP : Point to Point Protocol
- PVC : Permanent VC
- QOS : Quality of Service
- RICS : Router Interconnection by Cell Switch
- RFC : Request for Comments
- ROLC : Routing Over Large Cloud
- R-spec. : Reservation specification
- RSVP : resource ReSerVation Protocol
- RTP : Realtime Transmission Protocol
- SAP : Service Access Point
- SCPF : Short-Cut Path Forwarding
- SAR : Segmentation and Reassembly
- SCR : Sustainable Cell Rate
- SDU : Service Data Unit
- SNAP: SubNetwork Attachment Point
- SONET : Synchronous Optical NETwork
- SSCS : Service Specific Convergence Sub-layer
- SRM : Scaleable Reliable Multicast
- STM : Synchronous Transfer Mode

- ST-II : STream Protocol version 2
- SVC : Switched VC
- TCP : Transmission Control Protocol
- T-spec. : Traffic specification
- TTL : Time To Live
- UBR : Un-specified Bit Rate
- UDP : User Datagram Protocol
- UNI : User Network Interface
- UPC : Usage Parameter Control
- VBR : Variable Bit Rate
- VCC : Virtual Channel Connection
- VCI : Virtual Connection Identifier
- VPI : Virtual Path Identifier
- VPN : Virtual Private Network
- WAN : Wide Area Network

Selected Published Papers

Transactions Full Papers

1. **Hiroshi Esaki**, Yoshiyuki Tsuda, Takeshi Saito, Shigeyasu Natsubori: "Datagram Delivery in an ATM-Internet", IEICE Transactions on Communications, Special Issues on Future Private Networks, Vol.E77-B, No.3, March, 1994.

2. **Hiroshi Esaki**, Masataka Ohta, Ken-ichi Nagami: "High Speed Datagram Delivery over Internet using ATM Technology", IEICE Transactions on Communications, Vol.E78-B, No.8, August, 1995.

3. **Hiroshi Esaki**, Masataka Ohta, Ken-ichi Nagami: "High Speed Datagram Delivery over Internet using ATM Technology", Toshiba's Selected Papers on Science and Technology, Vol.8, No.1, January, 1996.

4. **Hiroshi Esaki**, Yoshiyuki Tsuda, Kumiko Kanai: "Evaluation of High Speed Multimedia Communication Architecture in ATM Networks", IEICE Transactions on Communications, Special Issue on Distributed Architecture for Next Generation Communication Networks, Vol.E77-B, No.11, November 1994.

5. **Hiroshi Esaki**, Yoshiyuki Tsuda, Takeshi Saito, Shigeyasu Natsubori: "Datagram Delivery in an ATM-Internet", Toshiba's Selected Papers on Science and Technology, Vol.7, No.1, January, 1995.

6. **Hiroshi Esaki**, Takeo Fukuda : "Reliable IP Multicast Communication Over ATM Networks Using Forward Error Correction Policy", IEICE Transactions on Communications, Vol.E78-B, No.11, November, 1995.

7. **Hiroshi Esaki**, Kazuaki Iwamura, Toshikazu Kodama, Takeo Fukuda : "Connection Admission Control in ATM Networks", IEICE Transactions on Communications, Vol.E77-B, No.1, January, 1994

8. Katsumi Yamato, **Hiroshi Esaki** : "Congestion Control for ABR Service Based on Dynamic UPC/NPC", IEICE Transactions on Communications, Vol.E79-B, No.3, March, 1996.

9. Shoogo Ueno, **Hiroshi Esaki**, Koosuke Harada : "Combustion Process under Strong DC Magnetic Field", IEEE Transactions on Magnetics, Vol.MAG-21, No.5, September, 1985.

IETF RFC (Request For Comments

1. K.Nagami, Y. Katsube, Y. Shobatake, A. Mogi, S. Matsuzawa, T. Jinmei, **H. Esaki** : "Toshiba's Flow Attribute Notification Protocol (FANP) Specification", IETF RFC2129, April, 1997.

2. Y.Katusbe, K.Nagami, **H.Esaki** : "Toshiba's Router Architecture Extensions for ATM : Overview", IETF RFC2098, February, 1997.

International Conference Papers

1. **Hiroshi Esaki**, Kazuaki Iwamura, Toshikazu Kodama : "A Simple and Effective Admission Control Method for an ATM Network", IEEE Global Telecommunications Conference 1990 (GLOBECOM'90), 300.5, pp.28-33, December, 1990.

2. **Hiroshi Esaki** : "Call Admission Control Method in ATM Networks", IEEE International Conference on Communications 1992 (ICC'92), 354.4, pp.1628-1633, June, 1992

3. **Hiroshi Esaki**, Yoshiyuki Tsuda, Takeshi Saito, Shigeyasu Natsubori: "Class D Service Architecture in an ATM-Internet", IEEE International Conference on Communications 1994 (ICC'94), pp.1312-1318, May, 1994.

4. **Hiroshi Esaki**, Masataka Ohta, Ken-ichi Nagami: "High Speed Datagram Delivery over Internet using ATM Technology", Networld+Interop'95 Engineer Conference, EN12.1, March, 1995.

5. **Hiroshi Esaki**: "High Speed IP Packet Forwarding over Internet using ATM Technology", conference on Emerging High-Speed Local-Area Networks and Wide-Area Networks, SPIE Photonics East '95 Symposium, October, 1995.

6. Katsumi Yamato, **Hiroshi Esaki** : "Dynamic UPC/NPC for Provision of Best Effort Service", ITU-T TELECOM95 Technology Summit, October, 1995.

7. Yasuhiro Katsube, Ken-ichi Nagami, **Hiroshi Esaki**: "Cell Switch Router - Basic Concept and Migration Scenario - ", Networld+Interop'96 Engineer Conference, April, 1996.

8. Kumiko Kanai, Reto Grueter, Keiji Tsunoda, Takeshi Saito, **Hiroshi Esaki**: "Forward Error Correction Control on AAL5; FEC-SSCS", *(to be appeared)* IEEE International Conference on Communications 1996 (ICC'96), June, 1996.

IETF Drafts

1. Masataka Ohta, **Hiroshi Esaki**, Ken-ichi Nagami : "Conventional IP over ATM", IETF Internet-Draft, draft-ohta-ip-over-atm-01.txt, July, 1994.

2. **Hiroshi Esaki**, Ken-ichi Nagami, Masataka Ohta : "Connection Oriented and Connectionless IP Forwarding Over ATM Networks", IETF Internet-Draft, draft-esaki-co-cl-ip-forw-atm-00.txt, October, 1994.

3. Yasuhiro Katsube, Ken-ichi Nagami, **Hiroshi Esaki** : "Router Architecture Extensions for ATM : Overview", IETF Internet-Draft, draft-katsube-router-atm-overview-00.txt, March, 1995.

[RESUME]

Name; Hiroshi ESAKI
Date of Birth; January 18, 1963
Bachelor of Electronic Engineering; March 1985 (Kyushu Univ.)
Master of Electronic Engineering; March 1987 (Kyushu Univ.)

January 1963; Born in Fukuoka, Japan

April 1981; Entered to Kyushu University, Fukuoka, Japan

March 1985; Received BE from Kyushu University, Fukuoka, Japan
Reserching on the effect of strong DC magnetic field on combustion processes.

March 1987; Received ME from Kyushu University, Fukuoka, Japan
Reserching on the effect of strong DC magnetic field on combustion processes.

April 1987; Entered to R&D Center, TOSHIBA Corporation.
Researching on traffic control for ATM networks

April 1990 - October 1991; Residential researcher at Bellcore, New Jersey, USA.
Researching on routing control for ATM networks and on internet control technologies.

July 1994 - April 1996; Visiting Scholar at Columbia University, New York, USA.
Researching on ATM internet architecture and on mobile computing architecture.
Contributing to IETF (Internet Engineering Task Force) and to ATM Forum.
- April 1995; Chairing the BoF(Birds of Feather) on CSR (Cell Switch Router) at IETF.
- October 1995; December 1995, February 1996; Chairing BoF on ATM level FEC (Forward Error Correction) at ATM Forum.

Affiliation;
R&D Center, TOSHIBA Corporation,
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, 210, Japan
TEL: +81-44-549-2238
FAX: +81-44-520-1806
E-mail: hiroshi@isl.rdc.toshiba.co.jp