Network Working Group
Request for Comments: 2098
Category: Informational

Y. Katsube
K. Nagami
H. Esaki
Toshiba R&D Center

February 1997

Toshiba's Router Architecture Extensions for ATM : Overview

Status of this Memo

Abstract

This memo describes a new internetworking architecture which makes better use of the property of ATM. IP datagrams are transferred along hop-by-hop path via routers, but datagram assembly/disassembly and IP header processing are not necessarily carried out at individual routers in the proposed architecture. A concept of "Cell Switch Router (CSR)" is introduced as a new internetworking equipment, which has ATM cell switching capabilities in addition to conventional IP datagram forwarding. Proposed architecture can provide applications with high-throughput and low-latency ATM pipes while retaining current router-based internetworking concept. It also provides applications with specific QoS/bandwidth by cooperating with internetworking level resource reservation protocols such as RSVP.

1. Introduction

The Internet is growing both in its size and its traffic volume. In addition, recent applications often require guaranteed bandwidth and QoS rather than best effort. Such changes make the current hop-by-hop datagram forwarding paradigm inadequate, then accelerate investigations on new internetworking architectures.

Roughly two distinct approaches can be seen as possible solutions; the use of ATM to convey IP datagrams, and the revision of IP to support flow concept and resource reservation. Integration or interworking of these approaches will be necessary to provide end hosts with high throughput and QoS guaranteed internetworking services over any datalink platforms as well as ATM.

New internetworking architecture proposed in this draft is based on "Cell Switch Router (CSR)" which has the following properties.

- It makes the best use of ATM's property while retaining current router-based internetworking and routing architecture.

- It takes into account interoperability with future IP that supports flow concept and resource reservations.

Section 2 of this draft explains background and motivations of our proposal. Section 3 describes an overview of the proposed internetworking architecture and its several remarkable features. Section 4 discusses control architectures for CSR, which will need to be further investigated.

2. Background and Motivation

It is considered that the current hop-by-hop best effort datagram forwarding paradigm will not be adequate to support future large scale Internet which accommodates huge amount of traffic with certain QoS requirements. Two major schools of investigations can be seen in IETF whose main purpose is to improve ability of the Internet with regard to its throughput and QoS. One is to utilize ATM technology as much as possible, and the other is to introduce the concept of resource reservation and flow into IP.

1) Utilization of ATM

Although basic properties of ATM; necessity of connection setup, necessity of traffic contract, etc.; is not necessarily suited to conventional IP datagram transmission, its excellent throughput and delay characteristics let us to investigate the realization of IP datagram transmission over ATM.

A typical internetworking architecture is the "Classical IP Model" [RFC1577]. This model allows direct ATM connectivities only between nodes that share the same IP address prefix. IP datagrams should traverse routers whenever they go beyond IP subnet boundaries even though their source and destination are accommodated in the same ATM cloud. Although an ATMARP is introduced which is not based on legacy datalink broadcast but on centralized ATMARP servers, this model does not require drastic changes to the legacy internetworking architectures with regard to the IP datagram forwarding process. This model still has problems of limited throughput and large latency, compared with the ability of ATM, due to IP header processing at every router. It will become more critical when multimedia applications that require much larger bandwidth and lower latency become dominant in the near future.

Another internetworking architecture is "NHRP (Next Hop Resolution Protocol) Model" [NHRP09]. This model aims at resolving throughput and latency problems in the Classical IP Model and making the best use of ATM. ATM connections can be directly established from an ingress point to an egress point of an ATM cloud even when they do not share the same IP address prefix. In order to enable it, the Next Hop Server [KAT95] is introduced which can find an egress point of the ATM cloud nearest to the given destination and resolves its ATM address. A sort of query/response protocols between the server(s) and clients and possibly server and server are specified. After the ATM address of a desired egress point is resolved, the client establishes a direct ATM connection to that point through ATM signaling procedures [ATM3.1]. Once a direct ATM connection has been set up through this procedure, IP datagrams do not have to experience hop-by-hop IP processing but can be transmitted over the direct ATM connection. Therefore, high throughput and low latency communications become possible even if they go beyond IP subnet boundaries. It should be noted that the provision of such direct ATM connections does not mean disappearance of legacy routers which interconnect distinct ATM-based IP subnets. For example, hop-by-hop IP datagram forwarding function would still be required in the following cases:

- When you want to transmit IP datagrams before direct ATM connection from an ingress point to an egress point of the ATM cloud is established

- When you neither require a certain QoS nor transmit large amount of IP datagrams for some communication

- When the direct ATM connection is not allowed by security or policy reasons

2) IP level resource reservation and flow support

Apart from investigation on specific datalink technology such as ATM, resource reservation technologies for desired IP level flows have been studied and are still under discussion. Their typical examples are RSVP [RSVP13] and STII [RFC1819].

RSVP itself is not a connection oriented technology since datagrams can be transmitted regardless of the result of the resource reservation process. After a resource reservation process from a receiver (or receivers) to a sender (or senders) is successfully completed, RSVP-capable routers along the path of the flow reserve their resources for datagram forwarding according to the requested flow spec.

STII is regarded as a connection oriented IP which requires connection setup process from a sender to a receiver (or receivers) before transmitting datagrams. STII-capable routers along the path of the requested connection reserve their resources for datagram forwarding according to the flow spec.

Neither RSVP nor STII restrict underlying datalink networks since their primary purpose is to let routers provide each IP flow with desired forwarding quality (by controlling their datagram scheduling rules). Since various datalink networks will coexist as well as ATM in the future, these IP level resource reservation technologies would be necessary in order to provide end-to-end IP flow with desired bandwidth and QoS.

aking this background into consideration, we should be aware of several issues which motivate our proposal.

- As of the time of writing, the ATM specific internetworking architecture proposed does not take into account interoperability with IP level resource reservation or connection setup protocols. In particular, operating RSVP in the NHRP-based ATM cloud seems to require much effort since RSVP is a soft-state receiver-oriented protocol with multicast capability as a default, while ATM with NHRP is a hard-state sender-oriented protocol which does not support multicast yet.

- Although RSVP or STII-based routers will provide each IP flow with a desired bandwidth and QoS, they have some native throughput limitations due to the processor-based IP forwarding mechanism compared with the hardware switching mechanism of ATM.

The main objective of our proposal is to resolve the above issues.

The proposed internetworking architecture makes the best use of the property of ATM by extending legacy routers to handle future IP features such as flow support and resource reservation with the help of ATM's cell switching capabilities.

3. Internetworking Architecture Based On the Cell Switch Router (CSR)

3.1 Overview

The Cell Switch Router (CSR) is a key network element of the proposed internetworking architecture. The CSR provides cell switching functionality in addition to conventional IP datagram forwarding. Communications with high throughput and low latency, that are native properties of ATM, become possible by using this cell switching functionality even when the communications pass through IP subnetwork

boundaries. In an ATM internet composed of CSRs, VPI/VCI-based cell switching which bypasses datagram assembly/disassembly and IP header processing is possible at every CSR for communications which lend themselves to such (e.g., communications which require certain amount of bandwidth and QoS), while conventional hop-by-hop datagram forwarding based on the IP header is also possible at every CSR for other conventional communications.

By using such cell-level switching capabilities, the CSR is able to concatenate incoming and outgoing ATM VCs, although the concatenation in this case is controlled outside the ATM cloud (ATM's control/ management-plane) unlike conventional ATM switch nodes. That is, the CSR is attached to ATM networks via an ATM-UNI instead of NNI. By carrying out such VPI/VCI concatenations at multiple CSRs consecutively, ATM level connectivity composed of multiple ATM VCs, each of which connects adjacent CSRs (or CSR and hosts/routers), can be provided. We call such an ATM pipe "ATM Bypass-pipe" to differentiate it from "ATM VCC (VC connection)" provided by a single ATM datalink cloud through ATM signaling.

Example network configurations based on CSRs are shown in figure 1. An ATM datalink network may be a large cloud which accommodates multiple IP subnets X, Y and Z. Or several distinct ATM datalinks may accommodate single IP subnet X, Y and Z respectively. The latter configuration would be straightforward in discussing the CSR, but the CSR is also applicable to the former configuration as well. In addition, the CSR would be applicable as a router which interconnects multiple NHRP-based ATM clouds.

Two different kinds of ATM VCs are defined between adjacent CSRs or between CSR and ATM-attached hosts/routers.

1) Default-VC

It is a general purpose VC used by any communications which select conventional hop-by-hop IP routed paths. All incoming cells received from this VC are assembled to IP datagrams and handled based on their IP headers. VCs set up in the Classical IP Model are classified into this category.

2) Dedicated-VC

It is used by specific communications (IP flows) which are specified by, for example, any combination of the destination IP address/port, the source IP address/port or IPv6 flow label. It can be concatenated with other Dedicated-VCs which accommodate the same IP flow as it, and can constitute an ATM Bypass-pipe for those IP flows.

Ingress/egress nodes of the Bypass-pipe can be either CSRs or ATM-attached routers/hosts both of which speak a Bypass-pipe control protocol. (we call that "Bypass-capable nodes") On the other hand, intermediate nodes of the Bypass-pipe should be CSRs since they need to have cell switching capabilities as well as to speak the Bypass-pipe control protocol.

The route for a Bypass-pipe follows IP routing information in each CSR. In figure 1, IP datagrams from a source host or router X.1 to a destination host or router Z.1 are transferred over the route X.1 -> CSR1 -> CSR2 -> Z.1 regardless of whether the communication is on a hop-by-hop basis or Bypass-pipe basis. Routes for individual Dedicated-VCs which constitutes the Bypass-pipe X.1 --> Z.1 (X.1 -> CSR1, CSR1 -> CSR2, CSR2 -> Z.1) would be determined based on ATM routing protocols such as PNNI [PNNI1.0], and would be independent of IP level routing.

An example of IP datagram transmission mechanism is as follows.

- The host/router X.1 checks an identifier of each IP datagram, which may be the "destination IP address (prefix)", "source/destination IP address (prefix) pair", "destination IP address and port", "source IP address and Flow label (in IPv6)", and so on. Based on either of those identifiers, it determines over which VC the datagram should be transmitted.

- The CSR1/2 checks the VPI/VCI value of each incoming cell. When the mapping from the incoming interface/VPI/VCI to outgoing interface/VPI/VCI is found in an ATM routing table, it is directly forwarded to the specified interface through an ATM switch module. When the mapping in not found in the ATM routing table (or the table shows an IP module as an output interface), the cell is assembled to an IP datagram and then forwarded to an appropriate outgoing interface/VPI/VCI based on an identifier of the datagram.

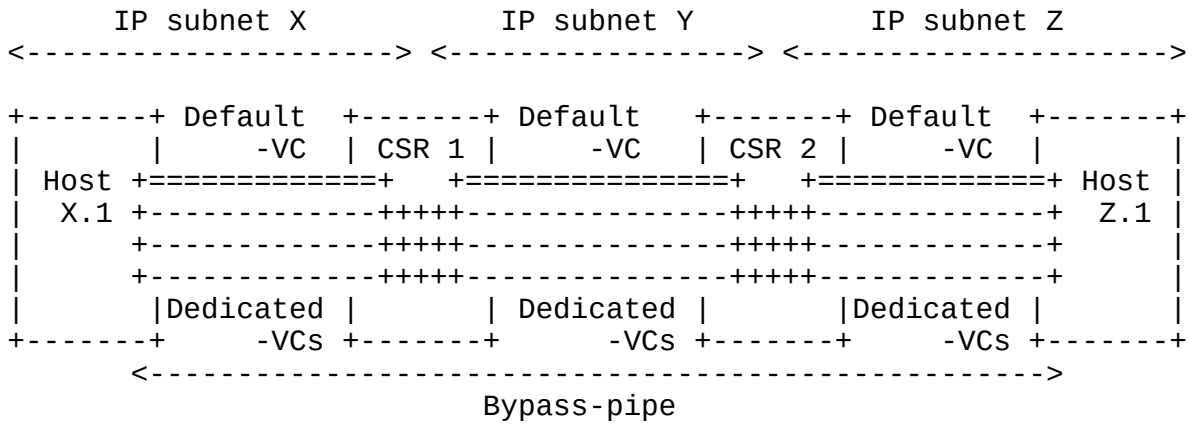

Figure 1 Internetworking Architecture based on CSR

3.2 Features

The main feature of the CSR-based internetworking architecture is the same as that of the NHRP-based architecture in the sense that they both provide direct ATM level connectivity beyond IP subnet boundaries. There are, however, several notable differences in the CSR-based architecture compared with the NHRP-based one as follows.

1) Relationship between IP routing and ATM routing

In the NHRP model, an egress point of the ATM network is first determined in the next hop resolution phase based on IP level routing information. Then the actual route for an ATM-VC to the obtained egress point is determined in the ATM connection setup phase based on ATM level routing information. Both kinds of routing information would be calculated according to factors such as network topology and available bandwidth for the large ATM cloud. The ATM routing will be based on PNNI phase1 [PNNI1.0] while the IP routing will be based on OSPF, BGP, IS-IS, etc. We need to manage two different routing protocols over the large ATM cloud until Integtrated-PNNI [IPNNI96] which takes both ATM level metric and IP level metric into account will be phased in in the future.

In the CSR model, IP level routing determines an egress point of the ATM cloud as well as determines inter-subnet level path to the point that shows which CSRs it should pass through. ATM level routing determines an intra-subnet level path for ATM-VCs (both Dedicated-VC and Default-VC) only between adjacent nodes (CSRs or ATM-attached hosts/routers). Since the roles of routing are hierarchically subdivided into inter-subnet level (router level) and intra-subnet level (ATM SW level), ATM routing does not have to operate all over

the ATM cloud but only in individual IP subnets independent from each other. This will decrease the amount of information for ATM routing protocol handling. But an end-to-end ATM path may not be optimal compared with the NHRP model since the path should go through routers at subnet boundaries in the CSR model.

2) Dynamic routing and redundancy support

A CSR-based network can dynamically change routes for Bypass-pipes when related IP level routing information changes. Bypass-pipes related to the routing changes do not have to be torn down nor established from scratch since intermediate CSRs related to IP routing changes can follow them and change routes for related Bypass-pipes by themselves.

The same things apply when some error or outage happens in any ATM nodes/links/routers on the route of a Bypass-pipe. CSRs that have noticed such errors or outages would change routes for related Bypass-pipes by themselves.

3) Interoperability with IP level resource reservation protocols in multicast environments

As current NHRP specification assumes application of NHRP to unicast environments only, multicast IP flows should still be carried based on a hop-by-hop manner with multicast routers. In addition, realization of IP level resource reservation protocols such as RSVP over NHRP environments requires further investigation.

The CSR-based internetworking architecture which keeps subnet-by-subnet internetworking with regard to any control protocol sequence can provide multicast Bypass-pipes without requiring any modifications in IP multicast over ATM [IPMC96] or multicast routing techniques. In addition, since the CSR can handle RSVP messages which are transmitted in a hop-by-hop manner, it can provide Bypass-pipes which satisfy QoS requirements by the cooperation of the RSVP and the Bypass-pipe control protocol.

4. Control Architecture for CSR

Several issues with regard to a control architecture for the CSR are discussed in this section.

4.1 Network Reference Model

In order to help understanding discussions in this section, the following network reference model is assumed. Source hosts S1, S2, and destination hosts D1, D2 are attached to Ethernets, while S3 and

D3 are attached to the ATM. Routers R1 and R5 are attached to Ethernets only, while R2, R3 and R4 are attached to the ATM. The ATM datalink for subnet #3 and subnet #4 can either be physically separated datalinks or be the same datalink. In other words, R3 can be either one-port or multi-port router.

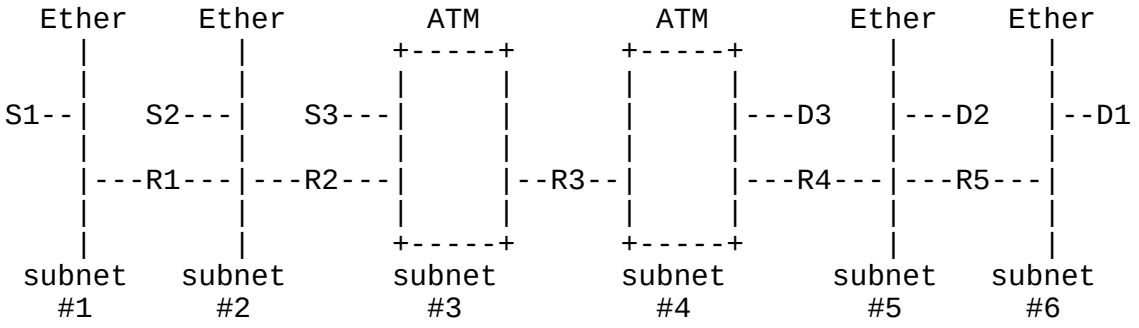

Figure 2 Network Reference Model

Bypass-pipes can be configured [S3 or R2]-->R3-->[D3 or R4]. That means that S3, D3, R2, R3 and R4 need to speak Bypass-pipe control protocol, and means that R3 needs to be the CSR. We use term "Bypass-capable nodes" for hosts/routers which can speak Bypass-pipe control protocol but are not necessarily CSRs.

As shown in this reference model, Bypass-pipe can be configured from host to host (S3-->R3-->D3), router to host (R2-->R3-->D3), host to router (S3-->R3-->R4), and router to router (R2-->R3-->R4).

4.2 Possible Use of Bypass-pipe

Possible use (or purposes) of Bypass-pipe provided by CSRs, in other words, possible triggers that initiate Bypass-pipe setup procedure, is discussed in this subsection.

Following two purposes for Bypass-pipe setup are assumed at present;

a) Provision of low latency path

This indicates cases in which end hosts or routers initiate a Bypass-pipe setup procedure when they will transmit large amount of datagrams toward a specific destination. For instance,

- End hosts or routers initiate Bypass-pipe setup procedures based on the measurement of IP datagrams transmitted toward a certain destination.

- End hosts or routers initiate Bypass-pipe setup procedures when it detects datagrams with certain higher layer protocols such as ftp, nntp, http, etc.

Other triggers may be possible depending on the policy in each network. In any case, the purpose of Bypass-pipe setup in each of these cases is to reduce IP processing burden at intermediate routers as well as to provide a communication path with low latency for burst data transfer, rather than to provide end host applications with specific bandwidth/QoS.

There would be no rule for determining bandwidth for such kinds of Bypass-pipes since no explicit information about bandwidth/QoS requirement by end hosts is available without IP-level resource reservation protocols such as RSVP. Using UBR VCs as components of the Bypass-pipe would be the easiest choice although there is no guarantees for cell loss quality, while using other services such as CBR/VBR/ABR with an adequate parameter tuning would be possible.

b) Provision of specific bandwidth/QoS requested by hosts

This indicates cases in which routers or end hosts initiate a Bypass-pipe setup procedure by triggers related to IP-level bandwidth/QoS request from end hosts. The "resource management entity" in the host or router, which has received bandwidth/QoS requests from applications or adjacent nodes may choose to accommodate the requested IP flow to an existing VC or choose to allocate a new Dedicated-VC for the requested IP flow. Selecting the latter choice at each router can correspond to the trigger for constituting a Bypass-pipe. When both an incoming VC and an outgoing VC (or VCs) are dedicated to the same IP flow(s), those VCs can be concatenated at the CSR (ATM cut-through) to constitute a Bypass-pipe. Bandwidth for the Bypass-pipe (namely, individual VCs constituting the Bypass-pipe) in this case would be determined based on the bandwidth/QoS requirements by the end host which is conveyed by, e.g., RSVP messages. The ATM service classes; e.g., CBR/VBR/ABR; that would be selected depends on the IP-level service classes requested by the end hosts.

Bypass-pipe provision for the purpose of b) will surely be beneficial in the near future when related IP-level resource reservation protocol will become available as well as when definitions of individual service classes and flow specs offered to applications become clear. On the other hand, Bypass-pipe setup for the purpose of a) may be beneficial right now since it does not require availability of IP-level resource reservation protocols. In that sense, a) can be regarded as a kind of short-term use while b) is a long-term use.

4.3 Variations of Bypass-pipe Control Architecture

A number of variations regarding Bypass-pipe control architecture are introduced. Items which are related to architectural variations are;

- o Ways of providing Dedicated-VCs
- o Channels for Bypass-pipe control message transfer
- o Bypass-pipe control procedures

Each of these items are discussed below.

4.3.1 Ways of Providing Dedicated-VCs

There are roughly three alternatives regarding the way of providing Dedicated-VCs in individual IP subnets as components of a Bypass-pipe.

a) On-demand SVC setup

Dedicated-VCs are set up in individual IP subnets each time you want to set up a Bypass-pipe through the ATM signaling procedure.

b) Picking up one from a bunch of (semi-)PVCs

Several VCs are set up beforehand between CSR and CSR, or CSR and other ATM-attached nodes (hosts/router) in each IP subnet. Unused VC is picked up as a Dedicated-VC from these PVCs in each IP subnet when a Bypass-pipe is set up.

c) Picking up one VCI in PVP/SVP

PVPs or SVPs are set up between CSR and CSR, or CSR and other ATM-attached nodes (hosts/routers) in each IP subnet. PVPs would be set up as a router/host initialization procedure, while SVPs, on the other hand, would be set up through ATM signaling when the first VC (either Default- or Dedicated-) setup request is initiated by either of some peer nodes. Then, Unused VCI value is picked up as a Dedicated-VC in the PVP/SVP in each IP subnet when a Bypass-pipe is set up. The SVP can be released through ATM signaling when no VCI value is in active state.

The best choice will be a) with regard to efficient network resource usage. However, you may go through three steps, ATMARP (for unicast [RFC1577] or multicast [IPMC96] in each IP subnet), SVC setup (in each IP subnet) and exchange of Bypass-pipe control message in this case. Whether a) is practical choice or not will depend on whether

you can allow larger Bypass-pipe setup time due to three-step procedure mentioned above, or whether you can send datagrams over Default-VCs in a hop-by-hop manner while waiting for the Bypass-pipe set up.

In the case of b) or c), the issue of Bypass-pipe setup time will be improved since SVC setup step can be skipped. In b), each node (CSR or ATM-attached host/router) should specify some traffic descriptors even for unused VCs, and the ATM datalink should reserve its desired resource (such as VCI value and bandwidth) for them. In addition, the ATM datalink may have to carry out UPC functions for those unused VCs. Such burden would be reduced when you use UBR-PVCs and set peak cell rate for each of them equal to link rate, but bandwidth/QoS for the Bypass-pipe is not provided in this case. In c), on the other hand, traffic descriptors which should be specified by each node for the ATM datalink is not each VC's but VP's only. Resource reservations for individual VCs will be carried out not as a functionality of the ATM datalink but of each CSR or ATM-attached host/router if necessary. A functionality which need to be provided by the ATM datalink is control of VPs' bandwidth only such as UPC and dynamic bandwidth negotiation if it would be widely available.

4.3.2 Channels for Bypass-pipe Control Message Transfer

There are several alternatives regarding the channels for managing (setting up, releasing, and possibly changing the route of) a Bypass-pipe. This subsection explains these alternatives and discusses their properties.

Three alternatives are discussed, Inband control message, Outband control message, and use of ATM signaling.

i) Inband Control Message

When setting up a Bypass-pipe, control messages are transmitted over a Dedicated-VC which will eventually be used as a component of the Bypass-pipe. These messages are handled at each CSR, and similar messages are transmitted to the next-hop node over a Dedicated-VC along the selected route (based on IP routing table). Unlike outband message protocol described in ii), each message does not have to indicate a Dedicated-VC which will be used since the message itself is carried over "that" VC.

The inband control message can be either "datagram dedicated for Bypass-pipe control" or "actual IP datagram" sent by user application. Actual IP datagrams can be transmitted over Bypass-pipe after it has been set up in the former case. In the latter case, on the other hand, the first (or several) IP datagram(s) received from

an unused Dedicated-VC are analyzed at IP level and transmitted toward adequate next hop over an unused Dedicated-VC. Then incoming Dedicated-VC and outgoing Dedicated-VC are concatenated to construct a Bypass-pipe.

In inband control, Bypass-pipe control messages transmitted after a Bypass-pipe has been set up cannot be identified at intermediate CSRs since those messages are forwarded at cell level there. As a possible solution for this issue, intermediate CSRs can identify Bypass-pipe control messages by marking cell headers, e.g., PTI bit which indicates F5 OAM cell. With regard to Bypass-pipe release, explicit release message may not be necessary if individual CSRs administer the amount of traffic over each Dedicated-VC and deletes concatenation information for an inactive Bypass-pipe with their own decision.

ii) Outband Control Message

When a Bypass-pipe is set up or released, control messages are transmitted over VCs which are different from Dedicated-VCs used as components of the Bypass-pipe. Unlike inband message protocol described in i), each message has to indicate which Dedicated-VCs the message would like to control. Therefore, an identifier that uniquely discriminates a VC, which is not a VPI/VCI that is not identical at both endpoints of the VC, need to be defined and be given at VC initiation phase. However, an issue of control message transmission after a Bypass-pipe has been set up in inband case does not exist.

Four alternatives are possible regarding how to convey Bypass-pipe control messages hop-by-hop over ATM datalink networks.

1) Defines VC for Bypass-pipe control messages only.

2) Uses Default-VC and discriminates Bypass-pipe control messages from user datagrams by an LLC/SANP value in RFC1483 encapsulation.

3) Uses Default-VC and discriminates Bypass-pipe control messages from user datagrams by a protocol field value in IP header.

4) Uses Default-VC and discriminates Bypass-pipe control messages from user datagrams by a port ID in the UDP frame.

When we take into account interoperability with Bypass-incapable routers, 1) will not be a good choice. Whether we select 2) or 3) 4) depends on whether we should consider multiprotocol rather than IP only.

In the case of IP multicast, point-to-multipoint VCs in individual subnets are concatenated at CSRs consecutively in order to constitute end-to-end multicast tree. Above four alternatives may require the same number of point-to-multipoint Defalut-VCs as the number of requested point-to-multipoint Dedicated-VCs in multicast case. The fifth alternative which can reduce the necessary number of VCs to convey control messages in a multicast environment is;

5) Defines point-to-multipoint VC whose leaves are members of multicast group 224.0.0.1. All nodes which are members of at least one of active multicast group would become leaves of this point-to-multipoint VC.

Each upstream node may become a root of the point-to-multipoint VC, or a sort of multicast server to which each upstream node transmits cells over a point-to-point VC may become a root of that. In any case, Bypass-pipe control messages for every multicast group are transmitted to all nodes which are members of either of the group. When a downstream node has received control messages which are not related to a multicast group it belongs, it should discard them by referring to a destination group address on their IP header. Donwstream node would still need to use point-to-point VC to send control messages toward upstream.

iii) Use of ATM Signaling Message

Supposing that ATM signaling messages can convey IP addresses (and possibly port IDs) of source and destination, it may be possible that ATM signaling messages be used as Bypass-pipe control messages also. In that case, an ATM connection setup message indicates a setup of a Dedicated-VC to an ATM address of a desirable next-hop IP node, and also indicates a setup of a Bypass-pipe to an IP address (and possibly port ID) of a target destination node. Information elements for the Dedicated-VC setup (ATM address of a next-hop node, bandwidth, QoS, etc.) are handled at ATM nodes, while information elements for the Bypass-pipe setup (source and destination IP addresses, possibly their port IDs, or flow label for IPv6, etc.) are transparently transferred to the next-hop IP node. The next-hop IP node accepts Dedicated-VC setup and handles such IP level information elements.

ATM signaling messages can be transferred from receiver to sender as well as sender to receiver when you set zero Forward Cell Rate and non-zero Backward Cell Rate as an ATM traffic descriptor information element in unicast case, or when Leaf Initiated Join capabilities will become available in multicast case.

Issues in this method are,

- Information elements which specify IP level (and port level) information need to be defined, e.g., B-HLI or B-UUI, as an ATM signaling specification.

- It would be difficult to support soft-state Bypass-pipe control which transmits control messages periodically since ATM signaling is a hard-state protocol.

4.3.3 Bypass-pipe Control Procedures

This subsection discusses several items with regard to actual procedures for Bypass-pipe control.

a) Distributed trigger vs. Centralized (restricted) trigger

The first item to be discussed is whether the functionality of detecting a trigger of Dedicated-VC/Bypass-pipe control is distributed to all the nodes (including CSRs and hosts/edge devices) or restricted to specific nodes.

In the case of the distributed trigger, every node is regarded as having a capability of detecting a trigger of Bypass-pipe setup or termination. For example, every node detects datagrams for ftp, and sets up (or fetches) a Dedicated-VC individually to construct a Bypass-pipe. After setting up or fetching the Dedicated-VCs, messages which informs (or requests) the transmission of the IP flow over the Dedicated-VC are exchanged between adjacent nodes. That enables peer nodes to share the same knowledge about the mapping relationship between the IP flow and the Dedicated-VC. There is no end-to-end message transmission in the Bypass-pipe control procedure itself, but transmission between adjacent nodes only.

In the case of the centralized (or restricted) trigger, capability of detecting a trigger of Bypass-pipe setup or termination is restricted to nodes which are located at "the boundary of the CSR-cloud". The boundary of the CSR-cloud signifies, for individual IP flows, the node which is the first-hop or the last-hop CSR-capable node. For example, a node which detects datagrams for ftp can initiate Bypass-pipe setup procedure only when its previous hop is non-ATM or CSR-incapable. In this case, Bypass-pipe control messages are originated at the boundary of the CSR-cloud, and forwarded hop-by-hop toward another side of the boundary, which is similar to ATM signaling messages. The semantics of the messages may be the request of end-to-end Bypass-pipe setup as well as notification or request of mapping relationship between the IP flow and the Dedicated-VC.

b) Upstream-initiated control vs. Downstream-initiated control

The second item to be discussed is whether the setup of a Dedicated-VC and the control procedure for constructing a Bypass-pipe are initiated by upstream side or downstream side.

In the case of the upstream-initiated control, the upstream node takes the initiative when setting up a Dedicated-VC for a specific IP flow and creating the mapping relationship between the IP flow and the Dedicated-VC. For example, a CSR which detects datagrams for ftp sets up (or fetches) a Dedicated-VC toward its downstream neighbor and notifies its downstream neighbor that it will transmit a specific IP flow over the Dedicated-VC. This means that the downstream node is requested to receive datagrams from the Dedicated-VC.

In the case of the downstream-initiated control, the downstream node takes the initiative when setting up a Dedicated-VC for a specific IP flow and creating the mapping relationship between the IP flow and the Dedicated-VC. For example, a CSR which detects datagrams for ftp sets up (or fetches) a Dedicated-VC toward its upstream neighbor and requests its upstream neighbor to transmit a specific IP flow over the Dedicated-VC. This means that the upstream node is requested to transmit the IP flow over the Dedicated-VC.

c) Hard-state management vs. Soft-state management

The third item to be discussed is whether the control (setup, maintain, and release) of the Bypass-pipe is based on hard-state or soft-state.

In hard-state management, individual nodes transmit Bypass-pipe control messages only when they want to notify or request any change in their neighbors' state. They should wait for an acknowledgement of the message before they change their internal state. For example, after setting up a Bypass-pipe, it is maintained until either of a peer nodes transmits a message to release the Bypass-pipe.

In soft-state management, individual nodes periodically transmit Bypass-pipe control messages in order to maintain their neighbors' state. They do not have to wait for an acknowledgement of the message before they changes its internal state. For example, even after setting up a Bypass-pipe, either of a peer nodes is required to periodically transmit refresh messages to its neighbor in order to maintain the Bypass-pipe.

5. Security Considerations

Security issues are not discussed in this memo.

6. Summary

Basic concept of Cell Switch Router (CSR) are clarified and control architecture for CSR is discussed. A number of methods to control Bypass-pipe will be possible each of which has its own advantages and disadvantages. Further investigation and discussion will be necessary to design control protocol which may depend on the requirements by users.

7. References

[IPMC96] Armitage, G., "Support for Multicast over UNI 3.0/3.1 based ATM Networks", RFC 2022, November 1996.

[ATM3.1] The ATM-Forum, "ATM User-Network Interface Specification, v.3.1", Sept. 1994.

[RSVP13] Braden, R., et al., "Resource ReSerVation Protocol (RSVP), Version 1 Functional Specification", Work in Progress.

[IPNNI96] R. Callon, et al., "Issues and Approaches for Integrated PNNI", The ATM Forum Contribution No. 96-0355, April 1996.

[NHRP09] Luciani, J., et al., "NBMA Next Hop Resolution Protocol (NHRP)", Work in Progress.

[PNNI1.0] The ATM-Forum, "P-NNI Specification Version 1.0", March 1996.

[RFC1483] Heinanen, J., "Multiprotocol Encapsulation over ATM Adaptation Layer 5", RFC 1483, July 1993.

[RFC1577] Laubach, M., "Classical IP and ARP over ATM", RFC 1577, October 1993.

[RFC1819] Delgrossi, L, and L. Berger, "Internet STream Protocol Version 2 (STII) Protocol Specification Version ST2+", RFC 1819, August 1995.

8. Authors' Addresses

Yasuhiro Katsube
R&D Center, Toshiba
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210
Japan
Phone : +81-44-549-2238
EMail : katsube@isl.rdc.toshiba.co.jp

Ken-ichi Nagami
R&D Center, Toshiba
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210
Japan
Phone : +81-44-549-2238
EMail : nagami@isl.rdc.toshiba.co.jp

Hiroshi Esaki
R&D Center, Toshiba
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210
Japan
Phone : +81-44-549-2238
EMail : hiroshi@isl.rdc.toshiba.co.jp