High speed IP packet forwarding over internet using ATM technology

Hiroshi Esaki
(hiroshi@ctr.columbia.edu)

R&D Center, Toshiba Corporation
c/o CTR, Columbia University, New York, New York, 10027

ABSTRACT

Framework of IP packet delivery with high throughput and small latency using ATM technology in large scaled internets is proposed, while keeping the current subnet model.

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 using 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 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. With the hop-by-hop cut-thru IP packet forwarding, soft-state oriented and scalable QoS-ed high speed communication platform can be provided.

Keywords: ATM, IP, cut-thru, cell-relaying, scalable, soft-state, internet

1 INTRODUCTION

ATM is realized as one of platforms to provide high speed datalink layer service with certain QoS (Quality of Service)[1,2]. Sometimes, it is said that ATM network will provide a router-less large cloud datalink platform, which need not have any routers within the cloud. However, we will need routers, even when large ATM cloud datalink platform will be provided [3]. And, the IP subnet must not be so large size so as to provide a large scale soft-state multicast service [4]. 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 [4,5]. Both ST-II and RSVP establish the flow state in the router(s) along the path where the IP packet flow passes through. 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.

This paper focuses on how to transport both resource reservation oriented and non resource reservation oriented IP packet flows over the ATM networks, effectively. 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 scalable for the required bandwidth and for the amount of packet flows.

Section 2 discusses issues of the IP over ATM architecture models proposed so far. Section 3 proposes the appropriate architecture for IP over ATM using CSR (Cell Switch Router). Section 4 gives a brief summary.

2 IP OVER ATM MODELS IN IETF AND ATM FORUM

Models developed at ATM Forum (LAN Emulation) and at IETF (Classical IP model[6] and NHRP model[7]) are discussed in this subsection. Here, LAN Emulation model is not explicitly discussed, since the LAN Emulation model and Classical IP model have the same issue associated with inter-IP subnet communication.

2.1 CATENET model[8]

The network should physically be divided into several datalink network segments, which are interconnected by routers. It would be recommended that each datalink segment should accommodate only one IP subnet (i.e. network layer segment). This is the current network architecture in the Internet.

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.

2.2 Classical IP model[6]

Classical IP model has been developed by IETF IPoverATM Working Group. In general, logical IP subnets (LISs) will overlay on a single 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 does not have a cell relaying capability, but it has only an IP forwarding 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 Classical IP 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.

2.3 NHRP with NBMA model

NHRP (Next Hop Resolution Protocol) with NBMA (Non-Broadcast Multiple Access) model has been developed by IETF ROLC (Routing Over Large Cloud) Working Group. LIS will overlay on a single ATM datalink segment. NBMA network is defined as the network domain where end-to-end ATM-VCC can be established [9]. Within the NBMA network, each host can 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 boader 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 performance 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 to obtain a network scalability.

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). In 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 [7].

3 PROPOSED IP OVER ATM ARCHITECTURE

3.1 Benefits of subnet model

The subnet model should be kept, even when we use LIS model. The physical or logical IP subnets are interconnected by 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 [10] 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.

   On the contrary, when we use the subnet model (i.e. interconnecting IP subnets by routers), the route is dynamically changeable during a session. 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 dynamic (soft-state) route. 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 a static route.

2. Change of QoS/flow-spec/service-class during a session
   Integrated service model[11,12,13], 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. Scalable soft-state multicast[4]

   RSVP is soft-state protocol, that is, *path* and *reserve* 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 *reserve* message to the up-stream routers or source host(s). Therefore, the number of *reserve* 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 *reserve* 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 [14]. 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 *reserve* message to the up-steam node, according to the reception of *path* message. Then, a large number (or huge) number of *reserve* messages will be sent back to the up-stream node, and the up-stream node may not be able to handle all the *reserve* messages.

   In order to provide a scalable soft-state multicast service over the large ATM cloud, we should keep a subnet model. When the scale of the LIS is reasonably small, we can provide a scalable 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.

3.2 Internetworking by CSR

In this subsection, the Cell Switch Router is introduced. Though all the routers in the ATM networks need not be CSR, CSR will provide scalable 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, a high throughput IP packet transmission with small latency can be provided, even when the packet communication goes beyond IP subnet passing 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 [15].

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). Potentially, the IP packet flow belonging to guaranteed service and predictive service (and maybe controlled delay service) would be able to be forwarded by cut-thru packet forwarding at CSR, since R-spec and T-spec of the packet flow are specified [11,12,13].

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

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

- High throughput packet forwarding with sall latency
  Logically, each IP packet is forwarded 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 cut-thru packet forwarding capability. CSR forwards the IP packets by 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.

- High performance firewall function
  When we introduce a resource reservation oriented flow (e.g. by ST-II), 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 ATM cell-relaying, there is no need to examine every IP packet header at CSR. Then, 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 boader 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 briefly below, must be defined. Neighbor routers must exchange the information how the IP packet flows are mapped (aggregated) with (into) datalink pipe (i.e. ATM-VCC).

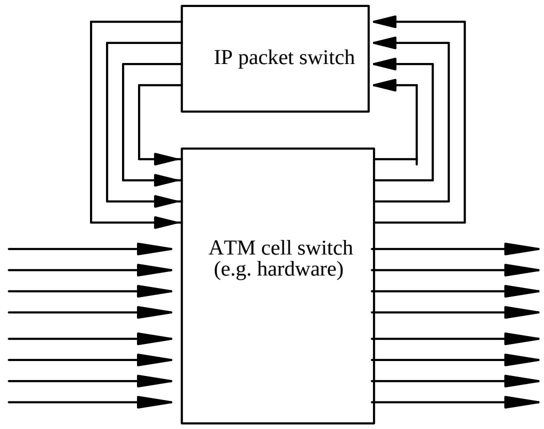

Figure 1: Example of CSR's implementation image

CSRs will be able to use the equipments that directly interconnects multiple ATM network segments. 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).

3.3 Implementation example and control signals

In this subsection, an example of CSR implementation image and control signals is discussed.

Figure 1 shows an example of implementation image of CSR, based on a CATENET model. CSR has both cell switching fabric and packet switching fabric. Incoming cells are switched by cell switching fabric based on the VPI/VCI value. The cell flow forwarded by 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 matters of implementor's decision. However, both cell and packet switching fabric should be scalable architecture (i.e. number of interface port should be scalable). 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 2 shows an example of the configuration of ATM-VCCs and IP packet flows with cut-thru packet forwarding at CSRs. Between nodes, two types of ATM-VCCs are defined. One is a default-VC and the other is dedicated-VC. The default-VC transfers both control packets (e.g. the control packets of FANP described below) and hop-by-hop based user packets. On the contrary, the dedicated-VC transfers only user packets that are forwarded by cut-thru packet forwarding 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.

The control messages to manage the user packet flows using the dedicated-VCs are transferred through the hop-by-hop based default-VC, which is different pipe from the dedicated-VCs. Therefore, the management of ATM bypass-pipes can be hop-by-hop, and the IP header in the concatenated dedicated-VCs at the CSR need

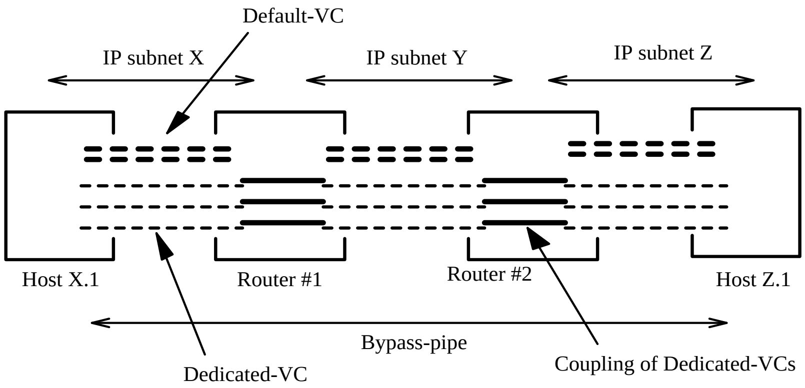

Figure 2: ATM-VCC configuration example for CSR

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. The cut-thru path will be established during a route setup phase, and it will not be changed during the session.
- Soft-state route :
  In soft-state route, the route of IP packet flow can be dynamically changed during a session. The control packet are periodically exchanged among the CSRs for the purpose of state refreshment, using the hop-by-hop default-VC.
- 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 stateless route.

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

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. Here, 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).

3.4 Flow attribute notification protocol : FANP

The purpose of FANP (Flow Attribute Notification Protocol) is indicating to the neighbor router how the IP packet flows are aggregated into the datalink pipe (i.e. ATM-VCC for ATM networks). This information is required to perform cut-thru packet forwarding at the CSR. Here, the information exchanged between the router by the FANP 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).

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 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 [16]) 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.

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 CONCUSION

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.

CSR (Cell Switch Router), which has both cell switching fabric and packet switching fabric is introduced. 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 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 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 small latency. With keeping the current subnet model, even in the ATM networks, we can obtain soft-state oriented and scalable QoS-ed high speed communication platform.

5 REFERENCES

1. ITU-T Recommendation I.150 ; "B-ISDN Asynchronous Transfer Mode", 1990
2. ATM Forum ; "ATM User Network Interface Specification version 3.1", Sept, 1994
3. Y.Rekhter and D.Kandlur : "IP Architecture Extensions over ATM", IETF Internet-Draft, draft-rekhter-ip-atm-architecture.txt, Jan., 1995.
4. L.Zhang ;"Resource ReSerVation Protocol (RSVP)", IETF Internet-Draft, draft-ietf-rsvp-spec-02.ps, May, 1994
5. L.Delgrossi and L.Berger : "Internet STream Protocol Version 2 (ST-II)", IETF Internet-Draft, draft-ietf-st2-spec-02.ps, Feb., 1995.
6. M.Laubach : "Classical IP and ARP over ATM", IETF RFC1577, Oct., 1993.
7. D.Katz and D.Piscitello : "NBMA Next Hop Resolution Protocol (NHRP)", IETF Internet-Draft, draft-ietf-rolc-nhrp-03.txt, Nov., 1994.
8. ISI-USC ; "Internet Protocol : DARPA Internet Program Protocol Specification", IETF RFC791, Sept., 1981.
9. J.Heinanen : "NBMA Address Resolution Protocol", IETF RFC1735, Dec., 1994.
10. ATM Forum P-NNI SWG : "P-NNI Draft Specification", ATM Forum Technical Contribution 94-0471R4, Nov., 1994.
11. S.Shenker, C.Partridge : "Specification of Controlled Delay Quality of Service", IETF Internet-Draft, draft-ietf-intserv-controlled-delay-svc-00.txt, March, 1995.
12. S.Shenker, C.Partridge : "Specification of Predictive Quality of Service", IETF Internet-Draft, draft-ietf-intserv-predictive-svc-00.txt, March, 1995.
13. S.Shenker, C.Partridge : "Specification of Guaranteed Quality of Service", IETF Internet-Draft, draft-ietf-intserv-guaranteed-svc-00.txt, March, 1995.
14. G.Armitage ;"IP Multicast over UNI 3.0 based ATM Networks", IETF Internet-Draft, draft-armitage-ipatm-ipmc-03.txt, Jan., 1995.
15. R.Braden, D.Clark, S.Shenker : "Integrated Services in the Internet Architecture ; an Overview", IETF RFC1633, June, 1994.
16. S.Deering, R.Hinden : "Internet Protocol, Version 6 (IPv6), Specification", IETF Internet-Draft, draft-ietf-ipngwg-ipv6-spec-01.txt, March 1995