PAPER

Architecture of Cell Switch Router and Prototype System Implementation

Shigeo Matsuzawa† Ken'ichi Nagami † Akiyoshi Mogi † Tatsuya Jinmei †
Hiroshi Esaki † *and* Yasuhiro Katsube † *Members*

SUMMARY Overview of Cell Switch Router (CSR) and the CSR prototype system are described. CSR can simultaneously support both connection oriented IP flows and connectionless IP flows. CSR contains cell switch fabric and IP packet switch fabric to achieve high throughput IP forwarding. IP packets are forwarded either through a cut-thru packet transmission, in which packet are forwarded without reassembling IP packet nor IP header processing, or through a conventional hop-by-hop IP packet forwarding. This paper describes and proposes the mechanism to forward the connectionless IP packet flows at the CSR.

A CSR prototype system has been developed. The CSR prototype system uses PVC connections to transfer the IP packets. With the CSR prototype system, we can make sure that CSR system can achieve a high throughput, i.e., 2.4 Gbps aggregated throughput. For end-to-end TCP/IP packet transmission, more than 90Mbps can be achieved and realtime video transmission (30Mbps video) can be achieved.
key words: ATM, CSR, cut-thru, hop-by-hop, IP, router, prototype

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 [1][6]. 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.

Sometimes, it is said that ATM network will provide a router-less large cloud datalink platform, which does not have any routers within the large cloud network. However, we will need routers, even large ATM [13] cloud datalink platform will be provided [2]. As discussed in [2], the IP subnet must not be so large size so as to provide a large scale soft-state multicast service [14]. 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 focuses on how to transport conventional connectionless IP packets [11] over the ATM networks. The hop-by-hop packet forwarding at the router by the *software processing* would be one solution. 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. [2], [7] and [4] propose a router architecture containing both conventional IP processing module and cell-switching module. The architecture of the router proposed in this paper is further detailed architecture based on the architecture proposed in [2][7]. [2][7] discuss the architecture to forward both connection oriented and connectionless IP packet flows. On the contrary, the architecture discussed in this paper focuses on the forwarding of conventional connectionless IP packets.

The following is the structure of the rest of this paper. Section 2 briefly discusses issues of the conventional IP over ATM architecture models. Sections 3 and 4 propose and evaluate the IP over ATM architecture using CSR (Cell Switch Router). In section 5, the overview and specification of CSR prototype system using PVC connections. Section 6 gives a brief conclusion.

2. Related Works of IP over ATM Architecture

In this section, the issues of IP service architectures over ATM proposed in IETF and ATM Forum are discussed.

Manuscript received October 1, 1996.
Manuscript revised December 17, 1996.
†The authors are with R&D Center, TOSHIBA CORPORATION, Kawasaki-shi, 210 Japan.

2.1 Classical IP Model [12], LAN Emulation Model [8]

In general, logical IP subnets (LISes) will overlay on a single NBMA ATM network segment. Figure 1 shows the case where packets are transferred from Host A.1 belonging to subnet A to Host B.1 belonging to subnet B. Logical IP Subnet A and subnet B exist on the NBMA ATM network.

LISes are interconnected by the conventional routers. For the intra-IP subnet communication, a direct ATM-VCC between the nodes can be provide. 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. In the case of figure 1, the packet from Host A.1 to Host B.1 is transferred to router A.0 using the direct ATM-VCC. Router A.0 forwards the received packet to the router B.0, which is the next hop router indicated by routing protocol. There may be the direct ATM-VCC between the router A.0 and router B.0, as shown in the figure 1. However, in general, there would be intermediate router(s) between router A.0 and router B.0. The packet received packet by the router B.0 is forwarded to the Host B.0 through the direct ATM-VCC.

Therefore, we can not achieve high throughput IP packet forwarding for inter-subnet communication. This is because, even for the packet transmission within the same NBMA ATM network, the IP packet for inter-subnet communication has to experience more than one IP processing at the router(s).

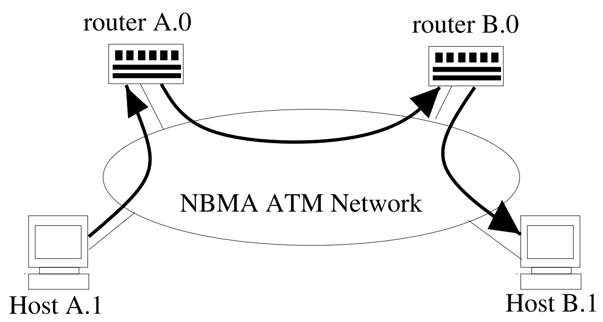

Figure 1. Classical IP / LAN Emulation Model

2.2 NHRP with NBMA Model [9][10]

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. 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 we can not obtain high performed IP packet forwarding beyond the NBMA network.

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 [10], 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. The routers attached to the NBMA domain exchange routing information (e.g., topological information and link attributes) among them. When a router receives an IP packet, the appropriate *next* hop node is resolved by the routing table calculated by routing protocol. ARP will be used, if ARP cache for MAC address of the *next* hop router is flushed. As discussed above, in the usual router's procedure, there is no chance to try to resolve the final destination router to reach the destination host. In order to apply NHRP, the IP forwarding procedure, that is usually implemented in kernel space, has to be substantially modified.

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 [14]. 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.

Figure 2 shows the case where the packet is transferred from Host A.y.1 to Host B.z.1. The two NBMA ATM networks are internetworked through a router. Each NBMA ATM network, subnet A and subnet B, would have multiple LISes. When Host A.x.1 sends a packet to Host B.z.1, Host A.x.1 resolves the ATM address of the egress router leading to Host B.z.1, using the NHRP Server. NHRP Server gives the ATM address of the egress router to forward the packet to Host

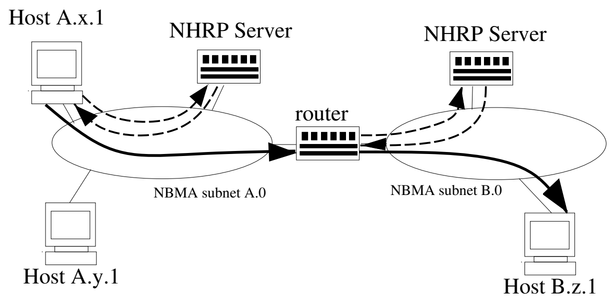

Figure 2. NBMA with NHRP Model

B.z.1, and Host A.x.1 establishes an ATM-VCC to the egress router to send packets to Host B.z.1. When the egress router receives the packet from Host A.x.1, the egress router analyses the destination IP address of the received packet. The egress router analyses the ATM address of Host B.z.1 using the NHRP Server in subnet B NBMA ATM network. The NHRP Server gives the ATM address of the Host B.z.1 to forward the packet to Host B.z.1, and the egress router establishes an ATM-VCC to Host B.z.1 to send packet to Host B.z.1. As shown in the figure, the packet has to experience the IP packet processing, when the packet is transferred to the host located at the different NBMA ATM network.

2.3 IP Switch Model

In the IP switch model [4], routers containing ATM switch fabric in it are interconnected by transparent links.

The architecture model of IP switch is shown in figure 3. Data-link network segment is identical to a transparent link between IP switches. Therefore, IP packet beyond IP subnet must always pass router (i.e. IP switch). In order to transfer the IP packet (both connection oriented IP packet and connectionless IP packet) from host A to host B, host A transfer IP packet to the router1 (i.e. IP switch). Routers (1, 2 and 3) relay the received IP packet toward host B. When the IP packet flow from host A to host B is either active flow or connection oriented IP transport flow, the IP packet will be forwarded without IP forwarding process in the router.

This model can provide high throughput and small latency packet forwarding even for inter-subnet packet exchanging. However, since the link among the IP switches must be transparent links (e.g., physical link or SONET link), we can not use ATM switches in the network, as shown in figure 4.

In general, IP switches can not be interconnected through the ATM least lines. IP switched can interconnected through the ATM least line only with the specific configuration, where ingress VPI/VCI value and egress

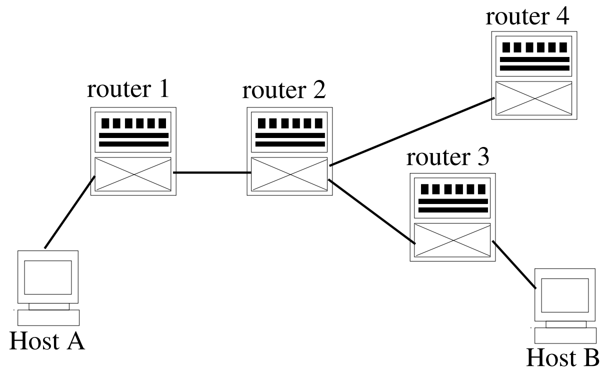

Figure 3. IP Switch Model

VPI/VCI value of the ATM link (VP) are the same. This particular configuration is not a general configuration in the ATM network. Further, IP switches can not interconnect with standard ATM switches, since IP switch does not support ATM UNI. Since IP switch does not have an interconnectivity with ATM switches, the network with IP switch can not provide a subnet constructed by ATM switch(es).

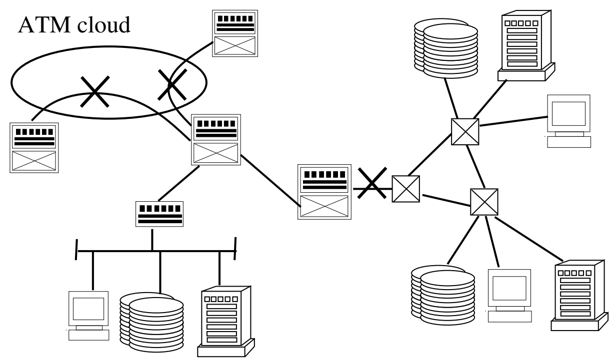

Figure 4. Issue of IP Switch Model

3. ATM Network Architecture using Cell Switch Router

3.1 Architecture of Cell Switch Router (CSR)

In this section, the proposed ATM network architecture using CSR is discussed. As described in [2][7], CSR supports both connection oriented and connectionless IP flows. This paper focuses on further discussion on how CSR supports connectionless IP packet flows and gives some evaluations on the proposed architecture.

Figure 5 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.

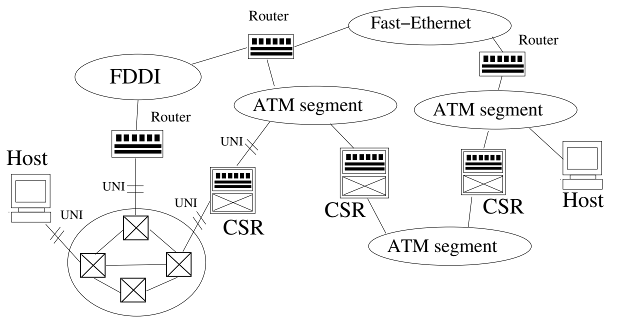

Figure 5. Large Scale Internet with Cell Switch Router

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 forwarding can be achieved. 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 applied to the other IP packet flows.

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

- 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. 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, i.e., cut-thru packet forwarding. 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. Since egress and ingress VCCs are independently established, CSR has to have the VPI/VCI translation functionality.

  Here, 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.

Through bypassing the conventional IP forwarding process using cell-relaying, we could dramatically increase the IP processing capacity of the router.

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

3.2 ATM-VCC Management Architecture

Figure 6 shows an example of the configuration of ATM-VCCs and IP packet flows with the cut-thru packet forwarding at CSRs. All ATM-VCCs, i.e., VCCs between CSRs and between CSR and host, are PVC (Permanent Virtual Connection). During the boot procedure of each node (CSR and host), the PVCs are established to the adjacent nodes.

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. On the contrary, the dedicated-VC transfers only the user packets that are forwarded through the cut-thru packet forwarding mode. 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 default-VC. The IP packets transferred through the default-VC are always performed the usual IP processing. Therefore, the management of cut-thru path can be hop-by-hop, and the IP header in the cut-thru path does not have to be examined. By the control packets (e.g., ST2, RSVP messages) 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.

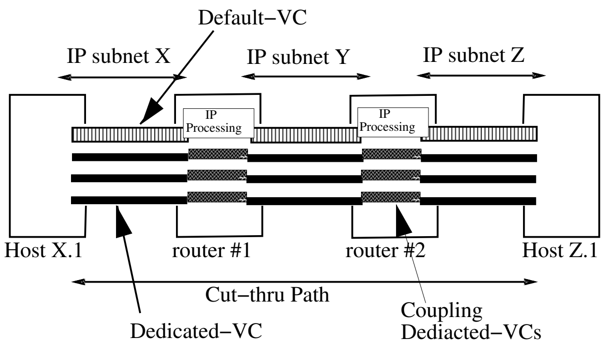

Figure 6. ATM-VCC Configuration Example for CSR

- Hard-state route (e.g., ST2 flow):
  The route of cut-thru path is not changed during a session. Therefore, the control packet may be exchanged only at route setup phase and at route termination phase. Any control packet may not be exchanged for the purpose of state refreshment.

- Soft-state route (e.g., RSVP flow, cut-thru flow):
  The control packet are continuously exchanged among the CSRs for the purpose of state refreshment using the hop-by-hop default-VC. The dedicated VCs are managed hop-by-hop, and the cut-thru path can dynamically reflect to the routing path indicated by the routing protocol.

- 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.

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. When we use digital links (e.g., OC-3 or DS3 SONET pipe) to interconnect CSRs, Q.2931 will not be used to establish/tear-down the dedicated-VCs.

3.3 Flow Attribute Notification Protocol : FANP [3]

The purpose of FANP 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 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 down-stream node.

- 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). The source IP addresses of the IP packet flows may not be 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 [5]) 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.

Here, in the prototype CSR system described in below, the IP packet flow is identified by the pair of source and destination IP addresses.

3.4 Packet Forwarding of Connectionless IP Flow

IP packets beyond the data-link segment are sent to an appropriate router from the source end-host. Routers exchange routing information and forward IP packets to the appropriate router.

The operation of connectionless IP packet forwarding is shown in figure 7. When the first IP packet (of the session or after the idle period) arrives at a CSR through the dedicated VC, 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 cut-thru a usual IP processing at the CSR.

The above procedure is performed independently at every CSR. 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.

4. 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 1 shows the comparison of IP over ATM architecture models, regarding some major points. As shown, the functionalities and features achieved by the cell switch router is better than the other architecture models. The meaning of "ATM link (VC/VP) availability" is whether the routers can be interconnected through the *standard* ATM links (VCC/VPC) provided by the ATM network. The meaning of "ATM switch connectivity" is whether the router can be directly connected with the *standard* ATM switch.

4.1 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 2 shows the maximum aggregated system throughput of cell switch router with some configurations.

Table 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.

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.

4.2 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

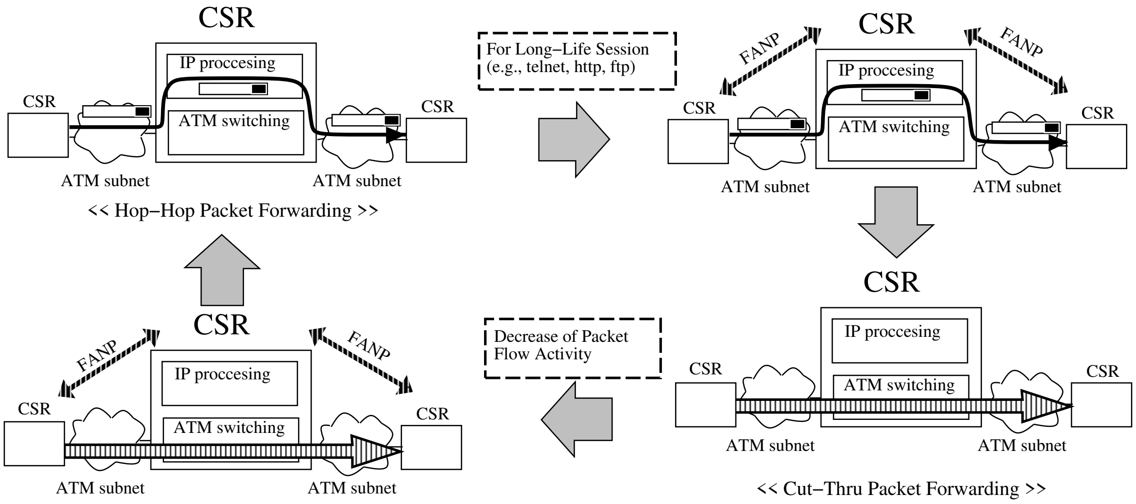

Figure 7. Operation of hop-by-hop ATM-VCC cacheing for active IP flow

Table 1. Comparison of IP over ATM Architecture Models

	Software router	NHRP /MPOA	IP Switch	CSR
Throughput	poor	fair	Good	Good
Routing loop possibility	No	Yes	No	No
ATM link (VP/VC) availability	Yes	Yes	No	Yes
ATM switch connectivity	Yes	Yes	No	Yes
System Complexity	Simple	Complex	Simple	Simple

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 would 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 8 shows one example of 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.

Some applications (e.g., continuous bit-stream transmission) are sensitive to the PDV rather than to average packet transmission delay. Typical applica-

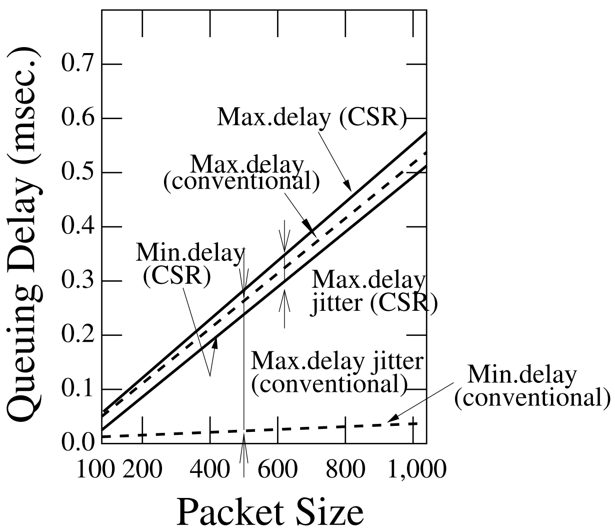

Figure 8. Delay Variation of Packet Delivery

tions, that are sensitive to the PDV, are voice and video signal transmission. With voice and video applications, the required receiving buffer space goes 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 very high speed conventional routers.

Since the CSR provides less chance of conventional packet processing at the router than the other architectures (e.g., NHRP, LANE) do, the PDV characteristics using CSR will be better than that using other architectures. This is because LANE always has to perform the conventional packet processing at the router and NHRP has to perform the convetional packet processing at the egress point of NBMA ATM network.

4.3 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 3 and 4 show (a) number of cut-thru sessions that will be 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. In these evaluations, four applications, i.e., web, ftp, telnet and nntp, are assumed as the sessions to be forwarded with cut-thru mode. When CSR receives the first IP packet belonging to one of these four application, dedicated VC will be established to forward the rest of succeeding IP packets with cut-thru mode. The cut-thru mode will be terminated to go to the hop-by-hop mode, when the application does not generate any IP packet for two minutes. Figure 3 is based on the packet transmission traces, which is available with http://town.hall.org/pub/ITA/html/contrib/DEC-PKT.html, for each one hour's worth of all wide-area packet transmission between Digital Equipment Corporation and the rest of the world. Figure 4 is based on the packet transmission traces for eight hour's worth of all packet transmission between the R&D Center (about thousand peoples) of Toshiba Corporation and the rest of the Toshiba's corporate network.

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

Table 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 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.

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 to the proxy server. This is because 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.

4.4 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 3 and 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.

4.5 High Performance Firewall Function

CSR may provide 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, we could not achieve a high throughput. With per session filtering, since only the first packet in the session has to be filtered, the rest of the packets can be forwarded in cut-thru mode without any IP packet filtering. This is means that we can apply the cut-thru packet forwarding mechanism to this type of sessions. As a result, when the traffic volume of the sessions, that require the per packet filtering, is not large portion in the aggregated traffic handled by CSR, CSR with a firewall function could achieve a sufficient high throughput.

4.6 Discussion

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

4.6.1 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 given 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.

4.6.2 Scaling up of IP Processing Capability

Theoretically, the packet processing capacity of CSR is scaleable according to the scalability of cell switch fabric. However, in the actual CSR system, the processing capability of IP processing module will be the bottleneck of system performance. When the cell switching fabric scales, the processing capability of IP processing module has also to scale.

Figure 9 shows an example of implementation image of CSR. 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 technology 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 10, 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. When IP packets are handled by software

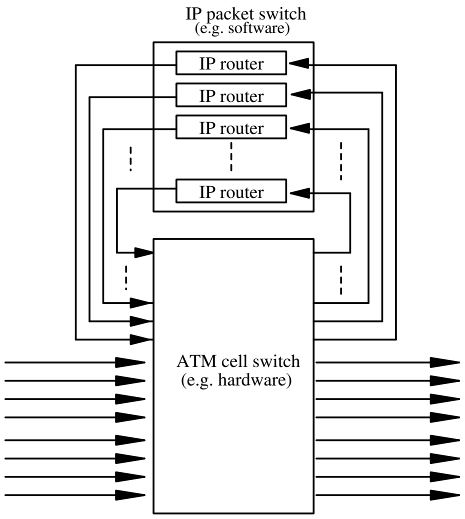

Figure 9. Scaleable CSR Implementation Example (I)

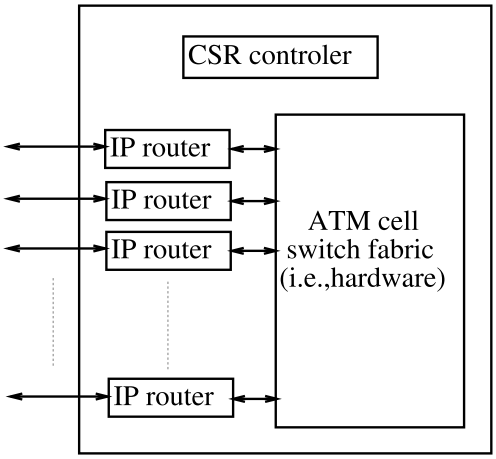

Figure 10. Scaleable CSR Implementation Example (II)

processing, the IP packet processing speed at IP router modules at the interface port may become insufficient. 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 forwarding 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.

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.

4.6.3 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.

4.6.4 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.

4.6.5 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. CSR Prototype System

Figure 11 shows the overview of the CSR prototype system using PVC connections to transfer the IP packets. Four ATM switched are interconnected through a CSR. Each ATM switch has it's own IP subnet. Therefore, the CSR in the figure internetworks four IP subnets constructed by ATM switch. One of four ATM based subnet is internetworked with CSR and with 100M Ethernet through a PC-based router. As a result, the prototype system has four IP subnets constructed by ATM switch and one 100M Ethernet IP subnet.

The following is the specification of developed CSR prototype system shown in figure 11.

- Maximum throughput of CSR; 4.8 MPPS for 64B packets (2.4 Gbps)

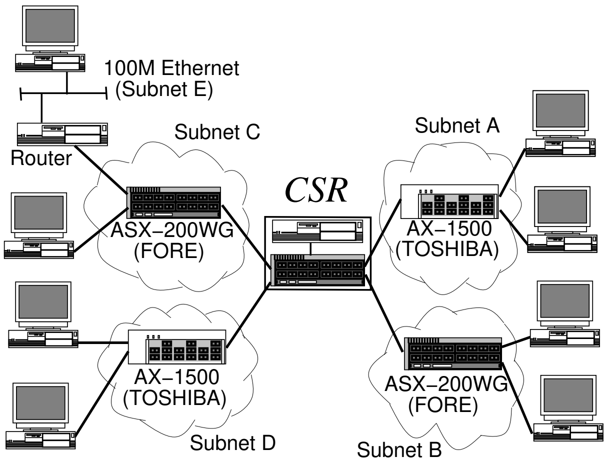

Figure 11. CSR prototype System Overview

- End-to-end TCP throughput;
  - ATM host to ATM host; 90 – 100Mbps
  - 100M Ethernet host to ATM host; 40Mbps
- Realtime video (NTSC signal) transmission; 30Mbps
- Remote CSR monitoring and management with SNMP
- Internetworking with 100M Ethernet
- IP packet forwarding using PVCs
- CSR IP packet switch module;
  - CPU; Pentium 100MHz
  - OS; BSD/OS v2.1
- CSR cell-switch module; FORE ASX-200WG
- Cut-thru Parameters (configurable)
  - Applications to be cut-thru: web, ftp, telnet, nntp
  - Cut-thru time out (T_{term}): two minutes

6. Conclusion

Over view of CSR architecture and CSR prototype system focusing on the connectionless IP packet transmission is discussed. CSR has both cell switching fabric and packet switching fabric. Since 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, CSR can achieve high throughout IP forwarding.

In order to perform the cut-thru IP packet forwarding using the cell relaying capability in the CSR, CSRs exchange the information how the IP packet flows are aggregated into dedicated-VC through FANP. This information exchanging with FANP 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 cut-thru packet forwarding, the soft-state and seamless cell-relaying channel is established to get a high throughput IP packet delivery.

Also, we can make sure that CSR can operate correctly and can achieve high throughput IP forwarding by the prototype system. Even with prototype system, The maximum aggregated throughput of prototype system is 2.4Gbps, which corresponds to 4.8 million packets per second with 64B packets.

As the future works, CSR will use SVC (Switched VC), as well as PVC, and will support RSVP to provide the end-to-end QoS (Quality of Service) over the Internet. Also, CSR will support the multicast service using the hardware-based multicast within the CSR, in order to provide high performance multicast services.

References

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

[2] 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

[3] Y.Katsube, K.Nagami, H.Esaki :"Mapping of IP Flow to Datalink Layer Connection", IETF Internet-Draft, draft-katsube-flow-mapping-dl-conn-00.txt, May, 1995

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

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

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

[7] Y.Katusbe, K.Nagami, H.Esaki : "Router Architecture Extensions for ATM : Overview", IETF Internet-Draft, draft-katsube-router-atm-overview-00.txt, March 1995

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

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

[10] J.V.Luciani, D.Katz, D.Piscitello and B.Cole : "NBMA Next Hop Resolution Protocol (NHRP)", IETF Internet-Draft, draft-ietf-rolc-nhrp-10.txt, October 1996

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

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

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

[14] L.Zhang, R.Braden, S.Berson, S.Herzog, S.Jamin :"Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification", IETF Internet-Draft, draft-ietf-rsvp-spec-13.txt, August, 1996

Shigeo Matsuzawa received the B.E. degree from Waseda University, Japan, in 1990. He joined Research and Develelopment Center, Toshiba Corporation where he has been working on communication system. He is currently interested in internet technology for high speed networks.

Ken-ichi Nagami received the B.E. degree from Chiba University, Japan, in 1990 and M.E. degree from Tokyo Institute of Technology, Japan, in 1992. He joined Research and Development Center, Toshiba Corporation where he has been working on communication system. He is currently interested in internet technology for high speed networks.

Akiyoshi Mogi received the B.E. degree from Waseda University, Japan, in 1989 and M.E.degree from Waseda University, Japan, in 1991. He joined Research and Development Center, Toshiba Corporation in 1991. He has been working on computer system software such as compiler and operating system. He has been working with Cell Switch Router project from 1995.

Tatsuya Jinmei received the B.S. degree from Kyoto University, Japan, in 1993 and M.E. degree from Kyoto University, Japan, in 1995. He joined Research and Development Center, Toshiba Corporation where he has been working on communication system. He is currently interested in internet technology for high speed networks.

Hiroshi Esaki received the B.E. and M.E. degrees from Kyushu University, Fukuoka, Japan, in 1985 and 1987, respectively. In 1987, he joined Research and Development Center, Toshiba Corporeation, where he engaged in the research of ATM systems. He has been at Bellcore in New Jersey (USA) as a residential researcher from 1990 to 1991, and has engaged in the research on high speed

computer communications. From 1994, he has been at CTR (Center for Telecommunications Reserch) of Columbia University in New York (USA) as a visiting scholar. He is currently interested in a high speed internet architeture and a mobile computing.

Yasuhiro Katsube received the B.E. and M.E degrees in electrical engineering from Keio University, Yokohama, Japan, in 1984 and 1986, respectively. In 1986, he joined Toshiba Research and Development Center, Kawasaki, Japan, where he has been engaged in the research and development of ATM network systems. From 1992 to 1994, he was a visiting researcher at Bell Communications Research. He is currently at the Communication and Information Systems Research Labs, Toshiba Research and Development Center.