PAPER IEICE/IEEE Joint Special Issue on Recent Progress in ATM Technologies

A Distributed Traffic Control Scheme for Large-Scale Multi-Stage ATM Switching Systems

Kohei NAKAI[†], **Eiji OKI**[†], *and* **Naoaki YAMANAKA**[†], *Members*

SUMMARY This paper describes a distributed traffic control scheme for large multi-stage ATM switching systems. When a new virtual circuit is to be added from some source line-interface unit (LU) to a destination LU, the system must find an optimal path through the system to accommodate the new circuit. Conventional systems have a central control processor and control lines to manage the bandwidth of all the links in the systems. The central control processor handles all the virtual circuits, but have trouble doing this when the switching system becomes large because of the limited ability of the central processor to handle the number of virtual circuits. A large switching system with Tbit/s-class throughput requires a distributed traffic control scheme. In our proposed switching system, each port of the basic switches has its own traffic monitor. Operation, administration, and maintenance (OAM) cells that are defined inside the system carry the path-congestion information to the LUs, enabling each LU to route new virtual circuits independently. A central control processor and control lines are not required. The performance of the proposed system depends on the interval between OAM cells. This paper shows how an optimal interval can be determined in order to maximize the bandwidth for user cells. This traffic control scheme will suit future Tbit/s ATM switching systems.

key words: *ATM, switch, multi-stage, large-scale, distributed control*

1. Introduction

Asynchronous transfer mode (ATM) is expected to yield the best high-speed multimedia infrastructure [1]. Recently, data traffic in the public network has been expanding explosively, thus necessitating an expansion of the switching system. In this situation, ATM networks will require switching systems that can offer Tbit/s throughput in a cost-effective manner [2]–[4].

Most ATM switches today use several single-stage switching techniques. Single-stage switches are relatively simple, but are limited in the number of ports and total throughput that they can support effectively. For large systems, multi-stage switching systems are needed. Three-stage switch architectures using unique basic switch elements are attractive. This is because they can be expanded easily using the same functional blocks [5], [6].

We have previously proposed a scalable 3-stage ATM switching architecture that uses optical WDM (wavelength division multiplexing) grouped links and dynamic bandwidth sharing [7], [8]. The former reduces the number of cables necessary. The latter prevents the statistical multiplexing gain from falling as the size of the switching system increases.

Figure 1 shows the scalable 3-stage ATM switching architecture. Each basic switch has N input ports and N output ports. Each port multiplexes M links corresponding to the M wavelengths that are multiplexed into one optical fiber. The total throughput of this system is MN times that of the basic switch. For example, a system capacity of 5.2 Tbit/s can be achieved using 8×8 80-Gbit/s basic switches and 8 wavelengths ($N = 8, M = 8$).

However, multi-stage switching systems need a scheme for controlling traffic. When a new virtual circuit is to be added from some source line-interface unit (LU) to a destination LU, the system must find an optimal path through the system to accommodate the new circuit. Conventional systems have a central control processor and control lines to manage the bandwidth of all the links in the system. The central control processor handles all the virtual circuits, but it has trouble doing this when the switching system becomes large. A large switching system with Tbit/s-class throughput requires a distributed traffic control scheme.

This paper proposes a new distributed traffic control scheme for large multi-stage switching systems in which each port of the basic switches has its own traffic monitor. Operation, administration, and maintenance (OAM) cells that are defined inside the system carry the path-congestion information to the LUs, enabling each LU to route new virtual circuits independently. The performance of the proposed system depends on the interval between OAM cells. This paper shows how an optimal interval can be determined.

The remainder of this paper is organized as follows. Section 2 briefly reviews conventional traffic control schemes for multi-stage switching systems. Section 3 presents a distributed traffic control scheme. Section 4 describes the optimal design, and Sect. 5 concludes the paper.

2. Conventional Traffic Control Schemes for Multi-Stage Switching Systems

Multi-stage switching systems need a scheme for controlling traffic. A 3-stage switching system needs to bal-

Manuscript received June 2, 1999.
Manuscript revised August 26, 1999.

[†]The authors are with the NTT Network Service Systems Laboratories, Musashino-shi, 180-8585 Japan.

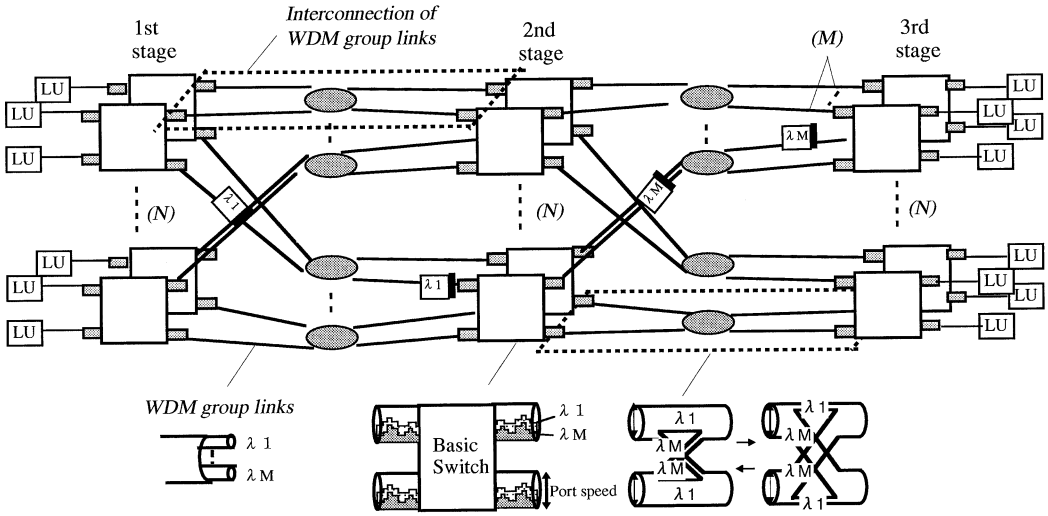


Fig. 1 Scalable 3-stage ATM switch architecture using WDM grouped links.

ance the 2nd-stage switching load. There are two types of traffic control scheme for multi-stage switching systems: those that use cell-based routing and those that use connection-based routing. Here, we assume that the internal link speed between basic switches is not increased. This is because all the same basic switches may be used in the system assuming the modularity of the basic switches.

If the system uses cell-based routing, each cell is routed independently in each input LU. Thus traffic is spread as evenly as possible among all the different paths [3], [9]. Some cells may arrive out of sequence in the wrong order at an output LU because cells belonging to the same virtual circuit may take different paths through the multi-stage switching system. Therefore, cell-resequencing is needed at the output LU. In a conventional cell-resequencing function, cells are time-stamped at the input LUs and the function uses the time-stamp information to reorder the cells arriving at an LU into the correct sequence.

However, the higher the speed of the output link, the harder this cell-resequencing function is to implement. This is because the cell time is shorter. For example, if the link speed of a switching system is 10 Gbit/s, the cell time is only 42.4 ns. This cell-resequencing function will be problematic since it will raise the cost of implementation.

If the system uses connection-based routing, on the other hand, all cells in a given virtual circuit follow the same path through the multi-stage switching system. The system maintains cell ordering directly, but requires explicit path selection. When a new virtual circuit is to be added from some input LU to some output LU, the system must find an optimal path through the system to accommodate the new circuit [10], [11]. Conventional systems have a central control processor and control lines to manage the bandwidth of all the links in the system.

However, conventional systems have trouble doing this when the switching system becomes large. The central processor can only handle a limited number of connections. The centralized control scheme has to find a new path in a very short time, and the system need to be able to handle a lot of information, such as the traffic characteristics of accommodated connections and their routes. This information is needed to manage the bandwidth of all the internal links. To find a new path, the cell loss ratio must be estimated after a new connection is accommodated. For example, assuming that each LU get 500 requests a second, the average period to find a new path is only 4 μs in a system with 512 LUs. Within only 4 μs, therefore, the system has to estimate the cell loss ratio and search for appropriate available routes. To do these heavy tasks, a very high-speed processor would be needed. We think that this centralized control approach would be too expensive. Thus, a centralized control scheme would not allow the switching system to be expanded in a cost-effective manner.

Therefore, we need a new traffic control scheme that does not require the cell-resequencing function used in the cell-based routing and that does not need a central control processor to manage all the links in connection-based routing. A traffic control scheme that works in a distributed manner is required.

3. Proposed Traffic Control Scheme

Our distributed traffic control scheme uses connection-based routing. We think that this scheme can handle CBR (constant bit rate) and VBR (variable bit rate) traffic. Our scheme does not require a central control processor or control lines. Instead, each port of the basic switches has its own traffic monitor. OAM cells carry the path-congestion information to the LUs, en-

abling each LU to route new virtual circuits independently.

One way to estimate the residual bandwidth is to count the cells arriving at a port. We can use an easy-to-implement simple traffic monitor described in [12]. If the traffic descriptors of connections such as average bit rate and peak bit rate are given, we can also employ a bandwidth management scheme [15] instead of the measurement-based estimation scheme. This scheme estimates the residual bandwidth that satisfies the specified cell-loss probability by generating a series of virtual requests for connection. Both of these estimation schemes support VBR traffic take into account statistical multiplexing effects as well as CBR traffic.

The threshold bandwidth, which is defined as B_r, is set for the residual bandwidth to judge whether the port is congested, as shown in Fig. 2. B_r is designed by taking into account the traffic characteristics of connections to satisfy QoS of connections and overbooking probability. The overbooking probability will be described in the next section.

Note that we use the term "port" to refer to the port of each basic switch. B_r must be set to an optimal bandwidth. If it is set too wide, the bandwidth assigned by LUs decreases; if it is set too narrow, many cells are lost in the port. Section 4 describes this in more detail.

If the residual bandwidth is lower than the threshold, the congestion-information (CI) bit of arriving cells is set to 1. LUs route virtual circuits using this information. The congestion probability at port i, which is defined as $P_c(i)$ is given by the traffic condition and switching system structure.

OAM cells are used in the system to check route performance. An LU has an input module and an output module. A virtual circuit is routed from the input module of a source LU to the output module of a destination LU. There are many routes from the source LU to the destination LU, so forward OAM cells are sent from the source LU along each route in turn. The destination LU returns the path-congestion information using backward OAM cell to the source LU through the reverse route, as shown in Fig. 3. The system does not require control lines because it uses OAM cells.

The CI bit of the OAM cell sent from a source LU is set to 0. If forward OAM cells arrive at a congested port, the CI bit is set to 1. When a load is low, we can assume that each $P_c(i)$ is independent. If a load increases to some degree, we can also assume that it is independent when there are many alternative routes in the system. However, when a load becomes very heavy, the traffic does not follow the Poisson distribution and $P_c(i)$ is not independent. As a first step in this study, we assume that the traffic follows the Poisson distribution and $P_c(i)$ is independent. In the next study, we should take into account the dependence of $P_c(i)$ on the heavy traffic condition. If there are P ports on a route, the worst congestion probability of the route P_r is given by Eq. (1). In Fig. 3, there are 4 ports on a route that may be congested.

$$P_r = 1 - \prod_{i}^{P}(1 - P_c(i)). \tag{1}$$

Each LU sends forward OAM cells along each route every T_{oam} seconds, as shown in Fig. 4. If a route is available, the LU rewrites the routing-lookup table which indicates currently available routes. If a route is congested, it immediately sends an OAM cell along another route. The probability that a route is congested is less than P_r. If an LU can not get the backward OAM

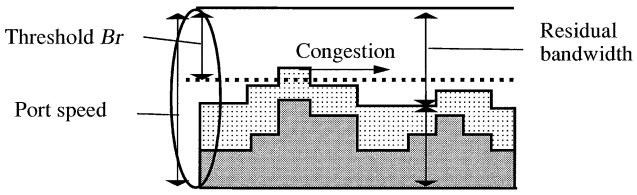


Fig. 2 Traffic monitor in port of basic switches.

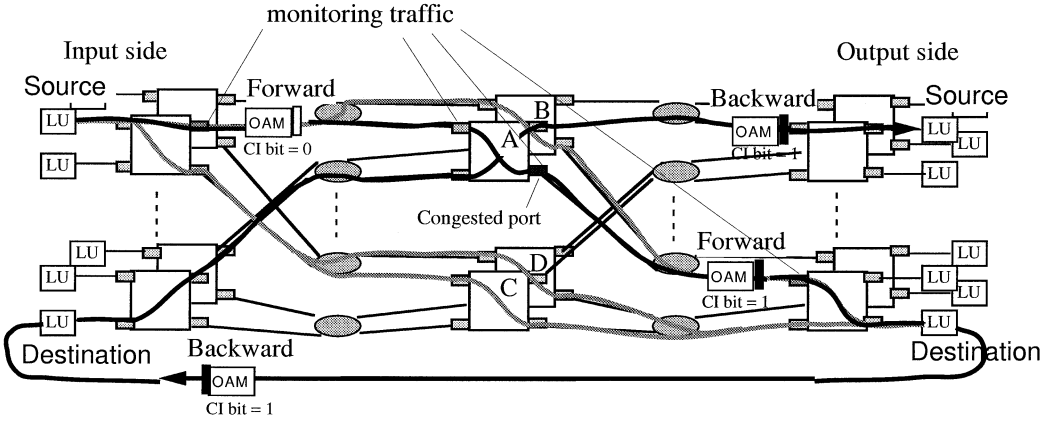


Fig. 3 Route-performance check using OAM cells.

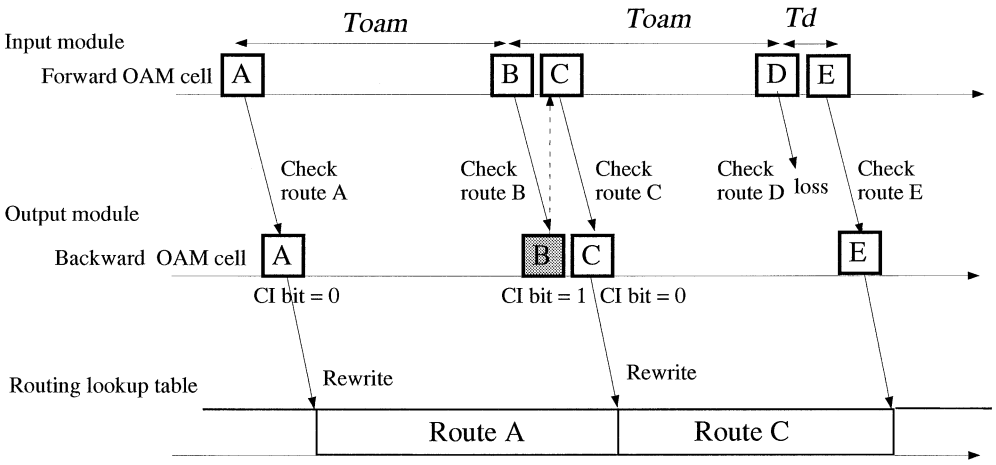


Fig. 4 Path selection by an LU.

cell within a maximum delay time (T_d), the OAM cell is considered to have been dropped or to still be waiting in a long queue. This situation means that the route is not available. In this case, the LU sends an OAM cell along another route.

Each LU is allowed to search available routes R times. If R routes are congested, the LU refuses a new requested virtual circuit. We call this a internal blocking. The internal blocking ratio in the system is less than P_r^R. The maximum time when an LU sends new forward OAM cells until it change the routing-lookup table is $R\ T_d$.

We explain how connection admission control (CAC) is executed using the distributed traffic control scheme described above. When a new connection is requested to be connected, the switching system has to check whether cell-loss probabilities at both the external link and the internal links of the switching system after the new connection is accepted are smaller than each specified probability. If both estimated cell-loss probabilities are satisfied with the conditions, the connection is accepted. The check for the cell loss probability at the external link is executed using well-known cell-loss estimation schemes, for example, those presented in [16]–[18], and [19]. For the internal links, on the other hand, each LU manages its own routing-lookup table which expresses whether available routes that satisfied the specified cell-loss probability exist between the source LU and the destination LU. We note that the cell loss probability is taken into account by using the residual bandwidth estimation scheme presented in [12] and [15].

4. Optimal Design Scheme for the Proposed System

Switching systems need to be designed in a cost-effective manner [13]. We should design bandwidth for user cells in a port (B_{usr} bit/s) to maximize the system

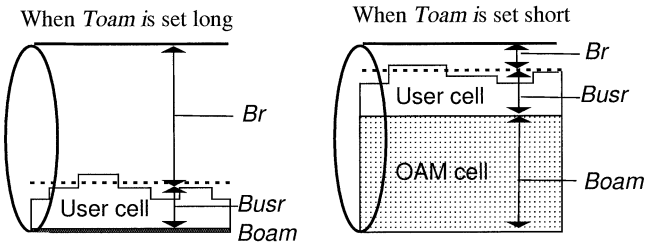


Fig. 5 Bandwidth for user cells in a port.

efficiency. B_{usr} is given by

$$B_{usr} = C - B_r - B_{oam}. \tag{2}$$

B_{usr} depends on the interval T_{oam} between OAM cells. Figure 5 shows the bandwidth for user cells in a port. If we set it too long, we must set B_r to a wide bandwidth. But if we set it too short, the bandwidth for OAM cells in a port (B_{oam} bit/s) increases while B_r is reduced. We show how an optimal interval between OAM cells can be determined.

If we set T_{oam} too long, LUs assign many virtual circuits into a currently "available" route that was given when LUs received the last OAM cell. However, before the next available route is given, the currently "available" route may be congested, and it will thus not be an available route. Therefore, the total bandwidth allocated in a port exceeds the speed of the switch port. We say that the port is overbooked [14]. Figure 6 shows the relationship between the residual bandwidth and the overbooking probability. Here, we assume that T_d is much smaller than T_{oam} in order to simplify the discussion on the relationship between T_{oam} and the overbooking probability. We must set the threshold B_r to an optimal bandwidth to ensure the specified overbooking probability. This is because if a port is overbooked, many cells are lost in the port. Figure 7 shows B_r for assuring that the overbooking probability is less than 10^{-9}. As T_{oam} becomes longer, the optimal B_r be-

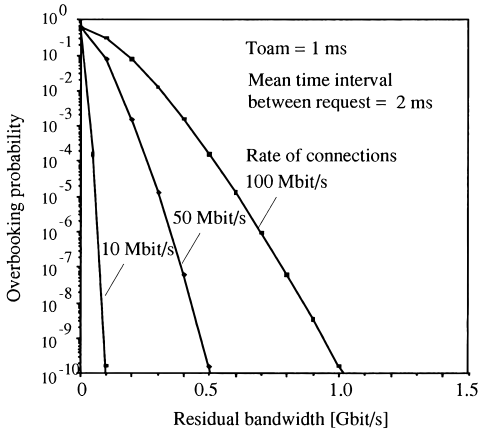


Fig. 6 The overbooking probability of CBR traffic.

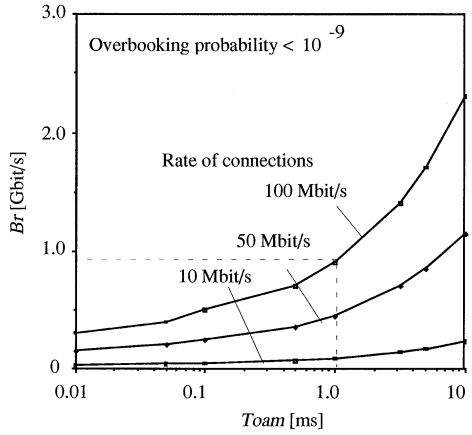


Fig. 7 B_r to ensure overbooking probability $< 10^{-9}$.

comes wider. If we set T_{oam} to 1 ms, then B_r must be set to 0.9 Gbit/s, assuming that the rate of connections is 100 Mbit/s. The overbooking probability is estimated by using maximum peak bit rate in terms of a conservative design. However, if we can predict the traffic characteristic of new connections that will requested to be connected, we can design the B_r more effectively due to multiplexing effect of VBR connections. The B_r for VBR traffic is estimated less than that of CBR traffic, as shown in Fig. 8, assuming that a certain cell loss ratio is allowed. If VBR connections and CBR connections coexist, B_r can be designed between the B_r for CBR and the B_r for VBR.

On the other hand, if we set T_{oam} too short, the number of OAM cells in the system increases. B_{oam} is inversely proportional to T_{oam}, as shown in Fig. 9. B_{oam} is given by

$$B_{oam} = \frac{2FL\sum_{i=0}^{R} Pr^i}{T_{oam}} \tag{3}$$

where F is the cell length in bits and L is the number of output LUs. On the right side of Eq. (3), the coefficient of 2 means both forward and backward OAM cells, and

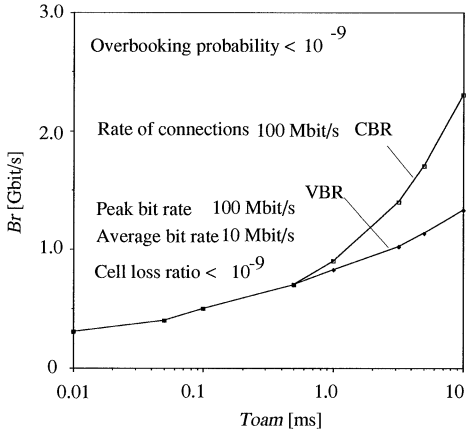


Fig. 8 B_r for CBR and VBR traffic.

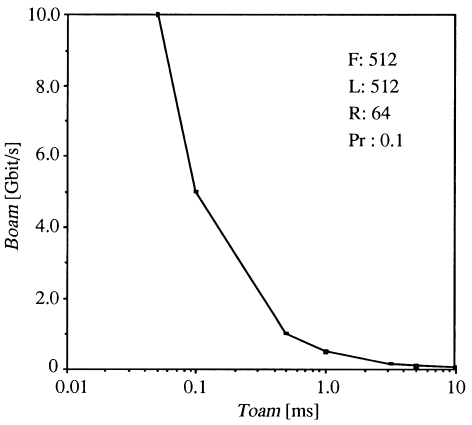


Fig. 9 Relationship between T_{oam} and B_{oam}.

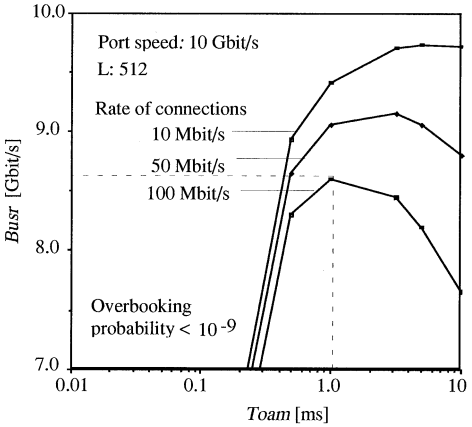


Fig. 10 Relationship between T_{oam} and B_{usr}.

$\sum_{i=0}^{R} Pr^i$ is the average number of transmissions until an available route is found up to a maximum of R times.

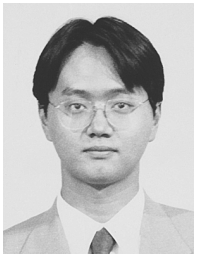
B_{usr} has a maximum at a certain value of T_{oam}. Figure 10 shows the relationship between T_{oam} and B_{usr}. When the rate of connections is 100 Mbit/s, we can maximize B_{usr} (8.6 Gbit/s) if we set T_{oam} to 1 ms.

5. Conclusions

We have proposed a distributed traffic control scheme for large multi-stage ATM switching systems. When a new virtual circuit is to be added from some source LU to a destination LU, the system must find an optimal path through the system to accommodate the new circuit. Conventional systems have a central control processor and control lines to manage the bandwidth of all the links in the system. The central control processor handle all the virtual circuits, but it has trouble doing this when the switching system becomes large. A large switching system with Tbit/s-class throughput requires a distributed traffic control scheme. In our switching system, each port of the basic switches has its own traffic monitor. OAM cells carry the path-congestion information to the LUs, enabling each LU to route new virtual circuits independently. Our system does not require a central control processor and control lines. The performance of the proposed system depends on the interval between OAM cells. We showed how an optimal interval can be determined in order to maximize the bandwidth for user cells. This control scheme will suit future Tbit/s ATM switching systems.

References

[1] H. Ohnishi and K. Miyake, "Issues in ATM network service development, standardization and deployment," IEICE Trans. Commun., vol.E81-B, no.2, pp.152–163, Feb. 1998.
[2] J. Turner and N. Yamanaka, "Architectural choices in large scale ATM switches," IEICE Trans. Commun., vol.E81-B, no.2, pp.120-137, Feb. 1998.
[3] T. Chaney, J.A. Fingerhut, M. Flucke, and J.S. Turner, "Design of a gigabit ATM switch," Proc. IEEE INFOCOM'97, pp.2–11.
[4] K. Genda and N. Yamanaka, "TORUS: Terabit-per-second ATM Switching system architecture based on distributed internal speed-up ATM switch," IEEE J. Sel. Areas Commun., vol.15, no.5, 1997.
[5] Y. Kamigaki, T. Nara, S. Machida, A. Hakata, and K. Yamaguchi, "160 Gbit/s ATM switching system for public network," Proc. IEEE GLOBCOM'96, pp.1380–1387.
[6] N. Yamanaka, S. Yasukawa, E. Oki, and T. Kawamura, "OPTIMA: Tb/s ATM switching system architecture: Based on highly statistical optical WDM interconnection," Proc. IEEE ISS'97, System Architecture, 1997.
[7] N. Yamanaka, "Breakthrough technologies for the high-performance electrical ATM switching system" IEEE J. Lightwave Technol., Dec. 1998.
[8] K. Nakai, E. Oki, and N. Yamanaka, "Scalable 3-stage ATM switch architecture using optical WDM grouped links based on dynamic bandwidth sharing," IEICE Trans. Commun., vol.E82-B, no.2, pp.265–270, Feb. 1999.
[9] M. Prycker and M. Somer, "Performance of a service independent switching network with distributed control," IEEE J. Sel. Areas Commun., vol.5, no.8, 1987.
[10] R. Melen and J. Turner, "Nonblocking multirate distribution networks," IEEE Trans. Commun., vol.41, no.2, Feb. 1993.
[11] S.C. Liew, M. Ng, and W.C. Chan, "Blocking and nonblocking multirate clos switching network," IEEE Trans. Networking, vol.6, no.3, June 1998.
[12] K. Shiomoto and N. Yamanaka, "An admission control scheme based on measurement of instantaneous utilization," IEICE Trans., vol.J80-B-I, no.12, pp.950–960, Dec. 1997.
[13] P. Coppo, M. D'Ambrosio, and R. Melen, "Optimal cost/performance design of ATM switches," IEEE Trans. Commun., vol.1, no.5, Oct. 1993.
[14] N. Yamanaka, K. Shiomoto, and H. Hasegawa, "ALPEN: A simple and flexible ATM network based on multiprotocol emulation at edge nodes," IEICE Trans. Commun., vol.E79-B, no.4, pp.611–615, 1996.
[15] E. Oki and N. Yamanaka, "Performance of high-speed admission control in ATM networks based on virtual request generation," IEICE Trans. Commun., vol.E82-B, no.3, pp.473–480, 1999.
[16] T. Murase, H. Suzuki, S. Sato, and T. Takeuchi, "A call admission control scheme for ATM networks using a simple quality estimate," IEEE J. Sel. Areas Commun., vol.9, no.9, pp.1461–1470, April 1991.
[17] H. Saito, "Call admission control in an ATM network using upper bound of cell loss probability," IEEE Trans. Commun., vol.40, no.9, pp.1512–1521, 1992.
[18] H.G. Perros and K.M. Elsayed, "Call admission control schemes: A review," IEEE Commun. Mag., pp.82–91, Nov. 1996.
[19] T. Lee, K. Lai, and S. Duann, "Design of a real-time call admission controller for ATM networks," IEEE/ACM Trans. Networking, vol.4, no.5, pp.758–765, 1996.



Kohei Nakai received the B.E. and M.E. degrees in electronic engineering from the University of Tokyo, Japan, in 1995 and 1997, respectively. In 1997, he joined Nippon Telegraph and Telephone Corporation's (NTT's) Network Service Systems Laboratories, Tokyo Japan. He is currently researching high-speed ATM switching systems at NTT Network Service Systems Laboratories. He is a member of IEEE Communication Society.

Eiji Oki received B.E. and M.E. degrees in instrumentation engineering and a Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1991, 1993, and 1999, respectively. In 1993, he joined Nippon Telegraph and Telephone Corporation's (NTT's) Communication Switching Laboratories, Tokyo Japan. He has been researching multimedia-communication network architectures based on ATM techniques and traffic-control methods for ATM networks. He is currently developing high-speed ATM switching systems in NTT Network Service Systems Laboratories as a Research Engineer. Dr. Oki received the Switching System Research Award and the Excellent Paper Award from the IEICE in 1998 and 1999, respectively. He is a member of the IEEE Communication Society.

Naoaki Yamanaka was born in Sendai-city, Miyagi prefecture, Japan, on July 22, 1958. He graduated from Keio University, Japan where he received B.E., M.E. and Ph.D. degrees in engineering in 1981, 1983 and 1991, respectively. In 1983 he joined Nippon Telegraph and Telephone Corporation's (NTT's) Communication Switching Laboratories, Tokyo Japan, where he was engaged in research and development of a high-speed switching system and high-speed switching technologies such as ultra-high-speed switching LSI, packaging techniques and interconnection techniques for Broadband ISDN services. Since 1989, he has been active in the development of Broadband ISDN based on ATM techniques. He is now researching future ATM based broadband ISDN architecture, and traffic management and performance analysis of ATM networks. He is currently a senior research engineer, research group leader in Broadband Network System Laboratory at NTT. Dr. Yamanaka received Best of Conference Awards from the 40th and 44th IEEE Electronic Components and Technology Conference, TELECOM System Technology Prize from the Telecommunications Advancement Foundation, IEEE CPMT Transactions Part B: Best Transactions Paper Award, MES98 Best Conference Paper Award, IEM/IMT Japan Outstanding paper award from IMPAS/IEEE, and Best Transaction paper award from IEICE in 1990, 1994 ,1994, 1998, 1999 and 1999 respectively. Dr. Yamanaka is Broadband Network Area Editor of IEEE Communication Surveys, Editor of IEICE Transaction, IEICE Communication Society International Affairs Director as well as Secretary of Asia Pacific Board at IEEE Communications Society. Dr. Yamanaka is a senior member of IEEE.