UDC 621.391.6:627.397.3

IP Router for Next-Generation Network

●Masatoshi Kumagai ●Satoshi Nojima ●Hiroshi Tomonaga

(Manuscript received February 8, 2001)

This paper describes some key technologies for the IP routers of the Next-Generation Network (NGN). A new architecture for packet processing engines is proposed, and its experimental results are reported. A new switch architecture with input queuing scalable up to terabit class capacity is also proposed, and a new scheduling algorithm which enables efficient distributed pipeline processing and fair bandwidth allocation is discussed. Finally, Fujitsu's IP router product is introduced.

1. Introduction

The rapidly expanding Internet is now a common communication platform for social and business communities, and it is being used to provide many services such as banking, education, and even administrative services. In response to this situation, high-speed access lines such as xDSL are quickly spreading and discussions are underway to establish various laws to accelerate the Internet's growth.

In this paper, we discuss the IP-based Next-Generation Network (NGN) and identify the requirements for its IP routers, which are important NGN elements. Then, two key technologies for NGN IP routers are discussed: the packet processing engine and a scalable switch architecture. Finally, we describe one of Fujitsu's IP router products and how it can be expanded to fit into the NGN.

1.1 Next-Generation Network

For decades, the public communication infrastructure consisted of telephone and data networks. However, carriers are currently migrating to a consolidated single network based on IP technology called the Next-Generation Network (NGN) or Broadband Internet infrastructure. The new network will form a high-reliability, high-quality Internet backbone for providing IT services. As shown in **Figure 1**, the NGN consists of the following three layers:

- An optical core network providing high-speed transport of huge volumes of traffic
- A network access layer which forms the entrance to the carrier IP network from subscribers, legacy telephone networks, and the IMT-2000 mobile network
- A network service layer providing call control and network services.

The basic Internet technology is a connectionless scheme that uses the store-and-forward method. However, it is not possible to construct a huge-capacity network using only this scheme and method. For example, there is no signaling for Quality of Service (QoS) negotiation, which is indispensable for supporting QoS services. Also, optical network technologies are needed to improve the network's capacity so that it meets today's demands, but a store-and-forward method cannot be realized with optical network technologies. As a result, as shown in Figure 1, future IP networks are expected to be construct-

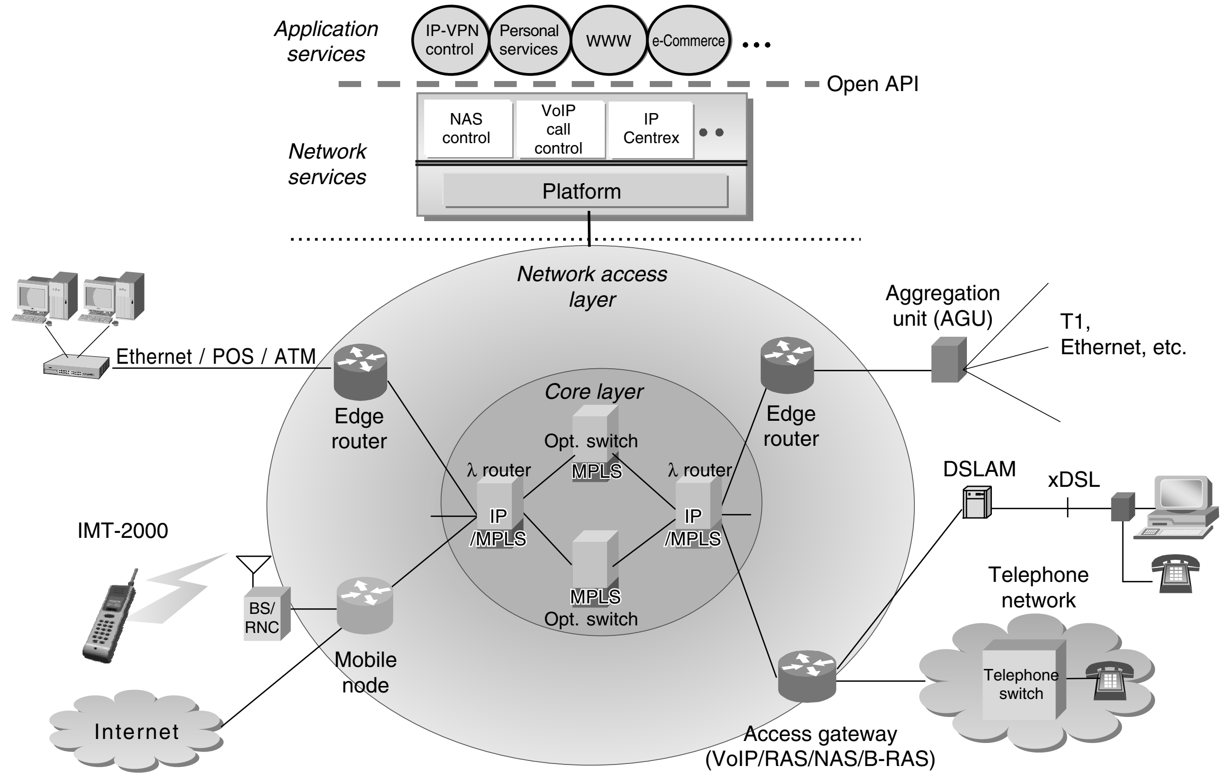


Figure 1
Next-Generation Network (NGN).

ed with a two-tier architecture which consists of a core layer based on optical transmission technologies and a network access layer for realizing IP functions for connections to outside the network. The core network acts as a transmission network between the edges. The edge routers provide various kinds of services, for example, QoS, at the edges of the network.

Since incoming traffic to the core layer has already been multiplexed into a high-speed interface in the network access layer, further multiplexing and processing in the core layer is unnecessary. Hence, it is assumed that, for economic reasons, the internal transmission in the core network will be realized by a simplified protocol and processing such as Multi-Protocol Label Switching (MPLS).[1)]

On the other hand, the edge routers aggregate traffic from various types of access networks, for example, Frame Relay, xDSL, CATV, and leased lines, and forward them to the core network through high-speed links having bandwidths ranging from 2.4 to 10 Gb/s. Moreover, when considering QoS realization, providing transmission paths with QoS service grades is essential in the core network. According to the core network QoS grades, classification of packet and QoS grade decision functions is required at the edge nodes.

1.2 Architecture of IP routers

Compared with core routers, although the edge routers may require less switching capacity, they are required to interface with a high-speed link to the core network. A leading photonic technology for achiving a high-speed link in edge routers is wavelength division multiplexing (WDM), which increases transmission capacity by using multiple wavelengths. A typical WDM bandwidth is 10 Gb/s. Also, a 10 Gb/s high-speed Ethernet has recently been under discussion.

The basic function of an IP packet switch such as an edge router is to reference a routing table using the destination IP address in the packet header. Then, the switch retrieves the next-hop information (i.e., the next-hop router and the outgoing interface) and forwards the packet to the next hop. Note that many advanced features/services are achieved by mapping each packet to a corresponding process using information in the packet header (IP header, TCP/UDP headers) or the upper layer's information. **Figure 2** shows the basic architecture of an edge router.

The edge router not only enables the service features described above but also aggregates various kinds of access lines. It is therefore appropriate to construct an interface unit to maintain several kinds of interfaces. Although the functions of this interface unit will differ depending on the type of line interfaces, after the interface termination, the processing will be almost identical. Therefore, it is necessary to construct two parts: the interface unit and the packet processing unit. The packet processing unit consists of a packet processing engine.

2. Proposed key technologies for IP routers

In this chapter, we discuss the following two key points regarding the future IP router to be used in the network access layer of the NGN:

- The architecture of the packet processing engine
- A switch architecture with scalability up to a multi-terabit capacity.

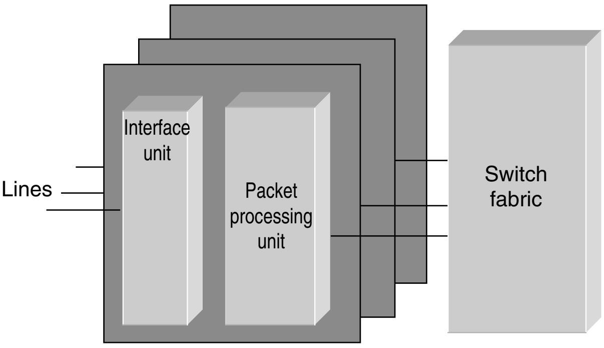


Figure 2
Basic architecture of edge router.

2.1 Packet processing engine

2.1.1 Requirements

The following functions are required for the packet processing engine:

- **QoS function**

As shown in **Figure 3**, this function performs 1) flow detection, which maps each packet into an associated QoS/CoS class (CoS: class of service) by classifying them according to the header information, and 2) QoS management, which measures and polices the flow quality (bandwidth, delay, jitter, etc.) and manages the packet sending order by scheduling.

For example, in RSVP,[2] we need the mapping from the flow identifier (e.g., source/destination IP addresses, protocol types, source/destination port numbers) to the associated QoS parameters. Also, packets should be managed by an appropriate scheduling algorithm, for example, weighted fair queuing (WFQ).

- **Packet routing functions**

After the above QoS has been decided, the packet has to be transmitted within the network according to a routing scheme, which searches a table to find a routing. Then, the packet headers must be modified in accordance with the routing.

From the viewpoint of performance, the following requirements should be considered:

- **High-speed processing**

As mentioned before, it is essential that edges can provide a high-speed link to the core network. A wire-speed processing of 10 Gb/s is a

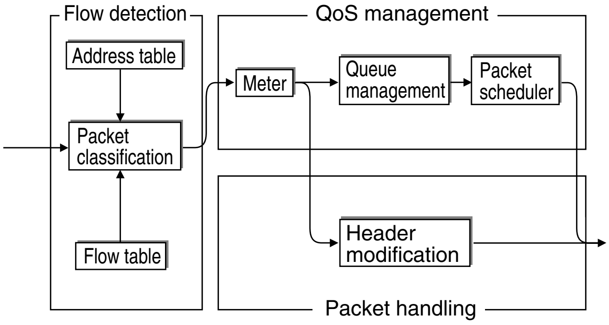


Figure 3
QoS processing flow.

current milestone.

- **Flexible programmability of processing**

The most basic function of packet processing is the classification function, which performs packet differentiation and interpretation based on the packets' header information. However, the decision and searching processes require a variety of parameters that depend on several protocol and service features that networks provide.[3),4)] Therefore, to adapt to the evolution of services/protocols and network individuality, it must be possible to flexibly define and modify the classification functions.

2.1.2 Proposed architecture

Realizing a flexible processing definition by software is desirable, but this cannot satisfy the 10 Gb/s performance requirement. Hence, we propose the following configuration:

1) Basic structure

To achieve both high-speed packet processing and complex processing with flexibility, the packet engine is constructed from multiple programmable processing units (PUs) arranged in a pipeline-manner as shown in **Figure 4**. Each processing unit is specialized for table lookup operation and packet header processing. These processing units are controlled by micro-code, which can be programmed to meet the customers' requirements. Implementing multiple processing units may increase the circuit-size; but given the recent advances in LSI technology, this will not be a significant issue.

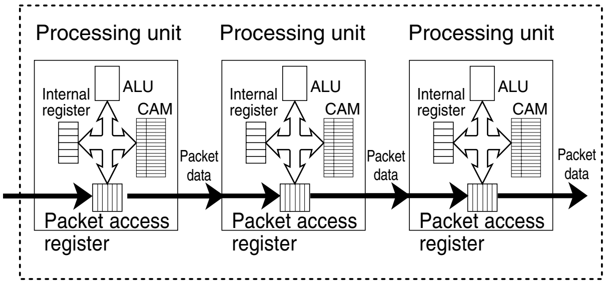


Figure 4
Packet processing engine.

2) Processing architecture

The proposed packet processing unit architecture is specialized to improve the processing speed and to enable pipelined processing. A major difference between this architecture and a general-purpose processor architecture is the packet access registers in the processing units that the packets go through. These packet access registers form a cascaded packet transmission route. The incoming packets are transmitted out in synchronization with clocks with no additional buffering delay time. The processing units can perform operations on the packets only when they are passing through their packet access registers. Therefore, the amount of processing that can be done is limited by the transmission time. To overcome this limitation, the processing is distributed among multiple processing units.

We propose the packet classifier architecture shown in Figure 4. It provides 1) high-speed processing, 2) the packet classification function essential for QoS processing, and 3) the ability to flexibly define these functions.

2.1.3 Experimental results

We implemented a prototype of the packet processing system described above using field programmable gate arrays (FPGAs). We also evaluated some sample coding. Based on our prototype, we estimate that IP routing requires about 20 instruction steps. Because the processing latency of a 10 Gb/s Ethernet is about 70 ns with a 150 MHz clock, each PU has 10 clock cycles to perform packet processing. If IP routing requires 20 cycles, less than 10% of the total steps available in a 24-PU environment is needed. We estimate that each PU can be constructed using about 100 k gates. We expect that 24 PUs and an 8 k × 128-bit content addressable memory (CAM) for the lookup table can be integrated into a single application-specific integrated circuit (ASIC) using a 0.18 micron process.

For future IP networks, network devices with both high performance and versatile functionality are required. Our proposed Flexible Network

Node, which has programmable packet engines, can achieve various requirements in future IP networks. Packet engines are constructed by pipelining multiple PUs and packet access methods and can perform packet processing fast enough for throughputs of up to about 10 Gb/s. Furthermore, the open software architecture makes it easy to implement customer-specific features such as packet classification and other packet-by-packet processing.

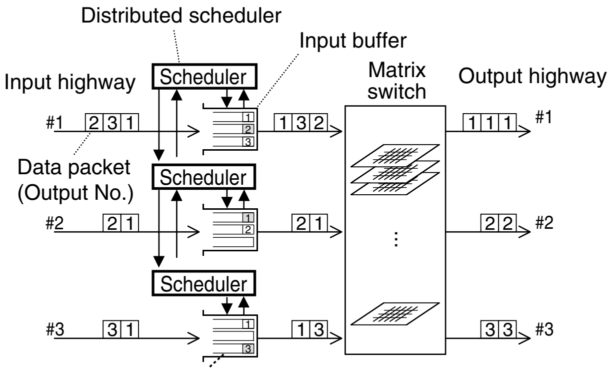


Figure 5
Proposed switch configuration (3 × 3).

2.2 Scalable switch architecture

2.2.1 Overview of switch configuration

To produce a switch with a terabit-class performance, we considered using input queuing, which is a promising technique that leads to fewer hardware bottlenecks than other switching techniques. To employ output queuing on the other hand, it is necessary to multiplex packets from multiple inputs at a higher bit rate. Eventually, limits on the bus clock rate and buffer memory access rate may cause a bottleneck in the system. Conventional first in/first out (FIFO) input queuing has a throughput limit of up to 58.6% due to head-of-line (HOL) blocking. HOL blocking, however, can be eliminated by separating the input queues for virtual output queues (VOQs) and reading the queued packets while avoiding contention by employing a suitable scheduling algorithm. Several scheduling algorithms have been proposed,[5)-7)] but a simpler and higher-performance scheduling algorithm is required to achieve a terabit class switch.

Figure 5 shows the configuration of our proposed switch.[8)] The input buffer contains a VOQ for each output port. The VOQs are controlled by distributed schedulers employed for one or more input lines. Scheduled packets can be read from the input buffer without contention on the bufferless matrix switch. Because the schedulers are not implemented centrally and can be added according to the number of lines being used, they do not cause bottlenecks in the switch. The crossbar matrix can be implemented with minimal hardware using bit-sliced or packet-sliced switch chips. The initial cost of an application with several line ports can also be minimized. As a result, this architecture enables a scalable switch to be built at low cost by adding necessary line cards.

2.2.2 Scheduling algorithm

A scheduler is an obstacle to achieving a terabit class switch. For example, in a 1.2 Tb/s switch composed of 128 input lines each having a 10 Gb/s bandwidth, 16 384 VOQs must be scheduled fairly and efficiently without contention among the input-output ports every single time unit. To fulfill this demand, we propose a novel scheduling algorithm called Sequential Round Robin (SRR). SRR has the following three key features:

1) Simple control, because round robin is employed for the output port and there is a consistent ordered rotation for the input port
2) High performance and scalability through the use of a distributed pipeline
3) Fair bandwidth allocation by rearranging the round robin according to load observations.

We describe our algorithm in more detail below.

2.2.3 Basic operation of SRR algorithm

The SRR algorithm searches the VOQ in a round robin way to determine a suitable output port for the input port with the highest priority and selects a queue to send the packet through. This procedure is repeated over all input queues to ensure that the same output port queue is not

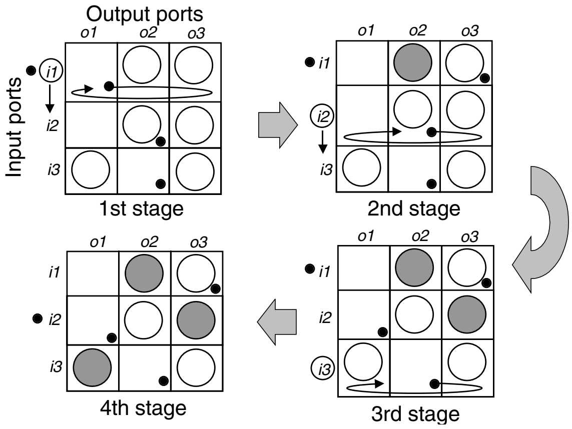


Figure 6
Example of basic SRR operation (3 × 3).

selected twice. Thus, packets from different input ports are sent to different output ports at each time unit.

Figure 6 shows an example operation of the basic SRR algorithm for a 3 × 3 switch. The four stages show the HOL states of the VOQs. A large open circle indicates a queued packet, and the small circles indicate the round robin pointer for the input port and output port which decide the priority of readout. In the 1st stage, the queued packet of the top priority input port, *i1*, is processed by searching from the *o1* output port, which corresponds to the pointer value, and output port *o2* is selected (indicated by large gray circle in the 2nd stage in Figure 6). Next, the pointer value for the output ports is updated to *o3* to ensure that the selected port has the lowest priority. In the 2nd stage, the queued packet of the next input port, *i2*, is processed by searching from output port *o2*, which corresponds to the pointer value, for the output ports that were not selected in the first stage. Since output port *o2* was selected earlier for input port 1, output port *o3* is selected. This procedure is repeated for the 3rd stage. In the 4th stage, the pointer value for the input port is shifted from 1 to 2 to prepare for the next procedure.

2.2.4 Distributed pipeline processing in SRR

Pipeline processing can be applied to SRR. Applying pipeline processing enables SRR to schedule all ports at high-speed without imposing speed penalties on the individual scheduling processes.

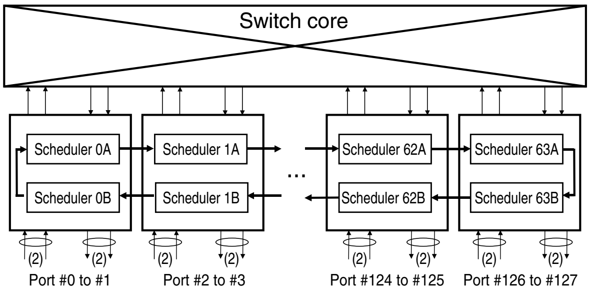


Figure 7
Distributed pipeline processing.

Figure 7 shows an overview of how distributed pipeline processing is applied in SRR when 128 input ports are divided into pairs and each pair is controlled by two schedulers, A and B, in each component. Schedulers A are interconnected over the upper route, and schedulers B are interconnected over the lower route. First, scheduler 0A, which has the highest priority, selects an output port and the information about the selected output port number is transported to the adjacent scheduler, 1A, on the right side. In the same way, the scheduling information is sent to scheduler 63A, which is located at the rightmost position. Next, the scheduling information is returned to the origin through scheduler B. Since each scheduler can perform the next processing step after sending the scheduling information to its neighbor, pipeline processing for all 128 input ports becomes possible. In this example, pipeline processing brings a dramatic 64-fold increase in available scheduling time.

2.2.5 Fair bandwidth allocation

Although pure SRR works well under uniform traffic, it causes a reduction in the throughput of the respective ports when the traffic pattern is asymmetrical. There are two types of asymmetry to consider (**Figure 8**). One is an asymmetry in the number of available paths between input ports and output ports. The other is an asymmetry in the load between input ports and

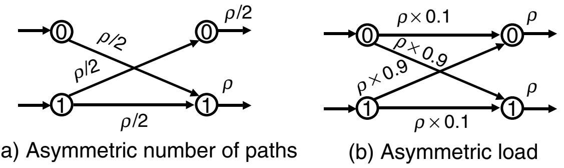


Figure 8
Two types of asymmetric traffic patterns.

output ports.

SRR rotates the input port of the highest priority in constant order, so a simple cyclic sequence causes an unbalanced selection when the number of paths is asymetrical. In rearranging the round robin, SRR alternately repeats a forward sequence and a backward sequence for input port rotation. This equalizes the selection of the paths. Replacement of the sequence can be achieved by exchanging the active ports between scheduler A and B within the same component.

The load observation function is efficient with asymmetric loads, because a simple round robin allocates bandwidth evenly, which causes throughput reductions for unbalanced loads. Load observation is done by counting the number of arriving packets for each virtual output queue at a constant frequency and providing an opportunity to read out according to the counted number.

3. IP switching node

This chapter describes Fujitsu's GeoStream R940 IP switching node, which is the first commercialized component of the GeoStream R900 series routers for large-scale IP backbone networks for carriers, ISPs, and other organizations.

3.1 Background of development

We analyzed the followng requirements of end users and carriers/ISPs and then identified the following requirements for the products:

1) End users' requirements

End users require more bandwidth for rapid and comfortable Web browsing and new services such as moving pictures. End users also require less expensive and higher quality services.

2) Carrier/ISP requirements

In response to users' requests, carriers and ISPs are looking for high-capacity, high-reliability, and low-cost equipment. Easy maintenance and operation without service interruption are important requirements in such euipment.

3) Development of GeoStream R940

We developed the GeoStream R940 for carriers and ISPs by drawing on our extensive experience in the public telephone, ATM switch, and enterprise IP router/switch fields. It provides carrier class reliability and maintenance features as well as full IP router features. The advantages of GeoStream R940 compared to other systems are described in detail in the next section.

3.2 Advantages

GeoStream R940 has the following advantages:

1) Large capacity, high speed, and scalability

GeoStream R940 has a large port capacity of 40 Gb/s and supports high-speed interfaces of up to 2.4 Gb/s. Its flexible architecture allows its capacity and interface speed to be increased in the future.

Wire speed forwarding is realized with no degradation, even when services such as filtering and VPN are enabled. This is indispensable for carriers to carry out reliable network engineering and provide a service level agreement (SLA) for their customers.

2) High reliability, high availability

The IP network is widely used as the backbone network of enterprises and public organizations, many of which cannot tolerate service interruptions. Carriers are moving from existing telephone and data networks to the consolidated IP network, which requires carrier-class reliability. In response to this severe requirement, customers are looking for reliable and robust router equipment.

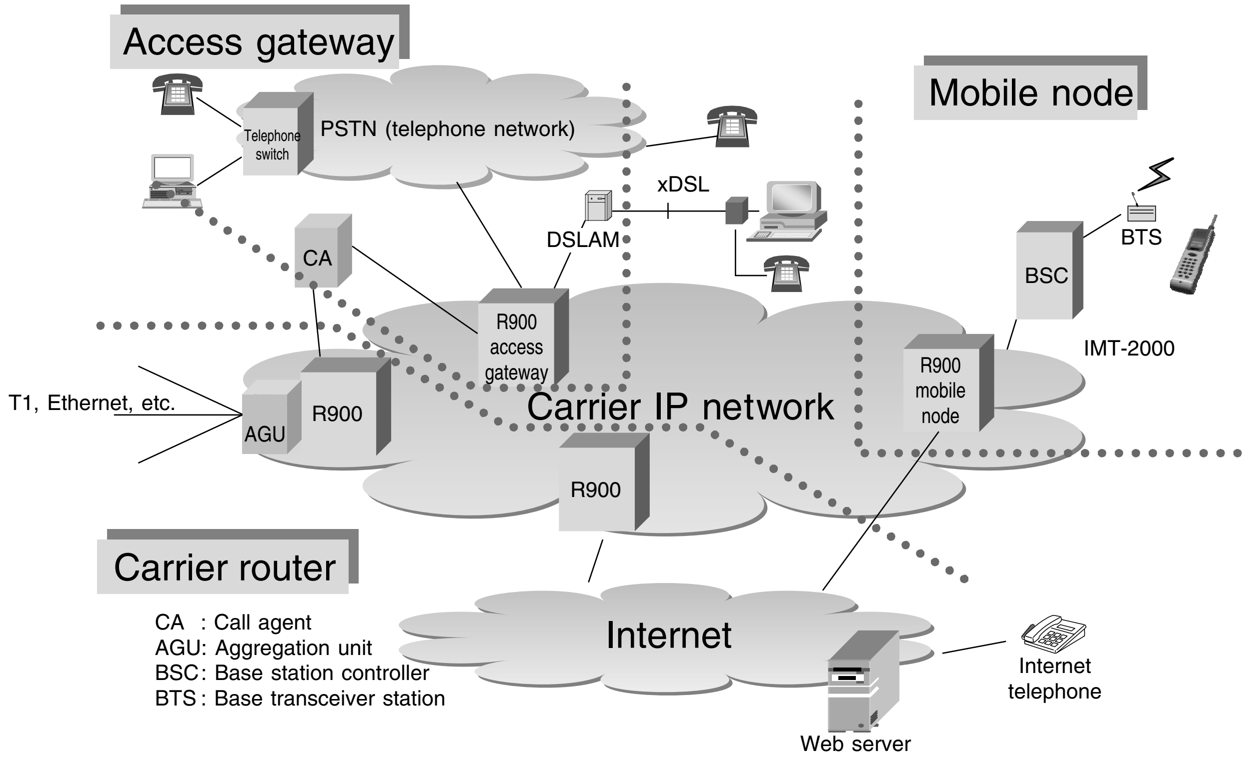
The GeoStream R940 IP switching node answers this requirement. It was developed under the same design criteria as those defined by Telcordia (formally Bellcore) for broadband switching systems for public networks.[9] Based on a reliability analysis calculation using the Markov model,[10] the total hardware unavailability with a full-duplex configuration is less than 0.0001% (more than 99.9999% availability), which conforms to Telcordia's criteria of a total downtime of less than 0.4 minutes/year).[9] Furthermore, well designed maintenance features will remarkably decrease service interruptions due to maintenance. For example, hardware extensions and software upgrades can be done without service interruption. Furthermore, there are two program banks and two configuration data banks to enable immediate recovery from unexpected failures after software upgrades and configuration changes.

3) Multi-service platform

Some important factors for carriers regarding future NGN deployment are the seamless integration of existing telephone and data networks and the provision of services such as voice and an Intelligent Network (IN) in the IP network. The ability to quickly introduce new services and the provisioning of high-quality services will be decisive factors in a network operator's future economic success. An IP network based on the GeoStream R900 series will enable an integration of legacy networks and services for a seamless migration from today's telecommunication network into an NGN. The GeoStream R900 series will also support the IMT-2000 next-generation mobile network.

Figure 9 shows the multi-service feature of the GeoStream R900 series IP switching node. To support a seamless migration of existing telephone networks to NGN and integration with the IMT-2000 mobile network, the following three configurations will be supported:

1) High-reliability conventional router used as a carrier edge router
2) Access gateway for interworking with existing telephone networks for smooth migration to NGN
3) ALL-IP mobile node for the IMT-2000 network.



CA : Call agent
AGU: Aggregation unit
BSC: Base station controller
BTS: Base transceiver station

Figure 9
Multi-service feature of GeoStream R900 series.

3.3 System configuration and specifications

1) System configuration

Figure 10 shows the system configuration of the GeoStream R940.

All circuits can be duplicated to realize high reliability and availability. Maintenance personnel can carry out maintenance work such as hardware replacement and software upgrades thanks to this duplication mechanism.

The physical line termination module and processing module are separated from each other for flexibility and reliability. There are several types of processing modules, for example, an IP packet forwarding module for interfacing to an Ethernet and packet over SONET/SDH (POS) and an ATM cell forwarding module for interfacing to an ATM. Other modules will be provided in the future to support new services. The interfaces that are supported are determined by the combinations of physical line termination modules and processing modules. For example, OC-3c POS is supported by the combination of an OC-3c line interface card and an IP packet forwarding module. Even if the line part and physical line interface are not duplicated, processing modules can be duplicated for high reliability.

The control module controls the entire equipment. It also handles the IP routing protocols, generates routing information, and sends it to the processing modules.

2) System specification

Table 1 shows the main specifications of GeoStream R940.

3.4 Future enhancements

There is a strong demand for larger models to support the rapidly increasing IP traffic. To support higher speed interfaces such as OC-192c/STM-64, we will develop a new packet processing engine based on the study results described in Section 2.1. To support increased capacity, a multiterabit class switch, a new switch architecture, and the scheduling algorithms described in Section 2.2 will be studied.

4. Conclusion

This paper proposed some new architectures for two key components of NGN IP routers: the packet processing engine and the scalable switch. We proposed a new packet processing engine with multiple programmable processing units arranged in a pipeline-manner that can realize high speed and flexible processing. We also proposed a new scalable switch architecture with an SRR scheduling algorithm that enables high-speed pipeline processing and fair bandwidth allocation. Fujitsu

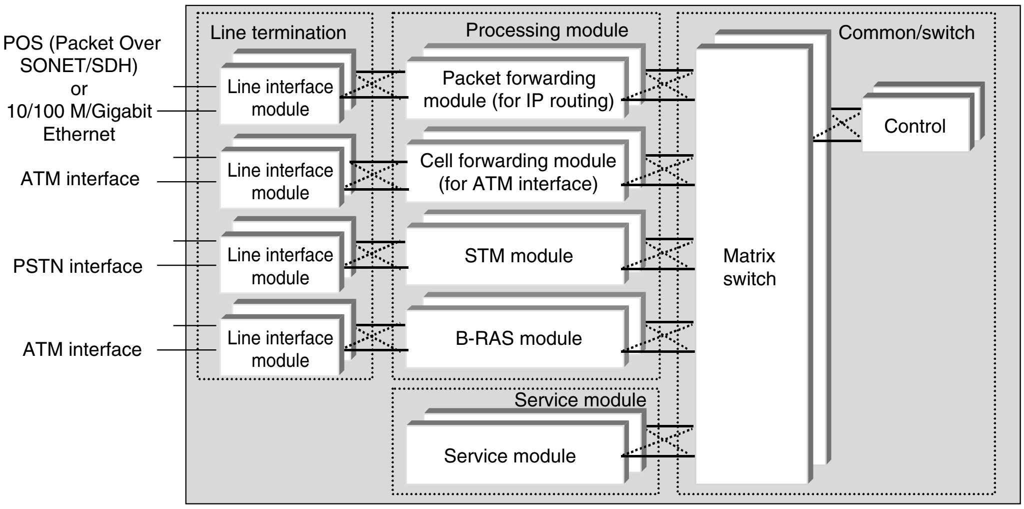


Figure 10
System configuration of GeoStream R940.

Table 1
Main system specifications of GeoStream R940.

Item		Specification
Switch		Matrix switch
Port capacity (half duplex)		40 Gb/s
IP packet handling capacity		48 Mpps[note)]
Number of slots		2.5 G × 8
Redundancy		Single or duplicated (Main control, Switch, Forwarding module), 1 + 1 APS (LTM)
Interface	POS	OC-48c/STM-16, OC-12c/STM-4, OC-3c/STM-1
	ATM	OC-12c/STM-4, OC-3c/STM-1
	Ethernet	1000Base-SX/LX, 10/100Base-T
	STM	Channelized OC-3/STM-1, OC-12/STM-4 (future)
Service	IP-QoS	DiffServ
	L3 switch	Wire speed, filtering (IP address, port number)
	VPN	MPLS, L2TP
	Security	IPSec
Routing protocols		RIP/OSPF/BGP4, IGMP/PIM-SM, IPv6
Network management		SNMP agent (MIB II, RMON), telnet, CLI
Power		−48 VDC
Mounting		19-inch rack mount
Regulation		UL, FDA, FCC, VCCI, NEBS level 3, CE marking

note) For packet over SONET/SDH (POS) interface, 40 bytes/packet

will continue to enhance its GeoStream IP routers by employing these new key technologies to provide further IT solutions for end users.

References

1) E. Rosen, A. Viswanathan, and R. Callon: Multiprotocol Label Switching Architecture. Internet Drafts<draft-ietf-mpls–arch-06.txt>, August 1999.

2) R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, and S. Jamin: Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. RFC2205, September 1997.

3) T. Li and Y. Rekhter: A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE). RFC2430, October 1998.

4) K. Nichols, S. Blake, F. Baker, and D. Black: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. RFC2474, December 1998.

5) N. McKeown, M. Izzard, A. Mekkittikul, B. Ellersick, and M. Horowitz: The Tiny Tera: A Packet Switch Core. Hot Interconnects V, Stanford University, August 1996.

6) R. Schoenen, G. Post, and G. Sander: Prioritized Arbitration for Input-queued Switches with 100% Throughput. Proc. of IEEE ATM Workshop'99, pp.253-258, May 1999.

7) J. Hurt, A. May, Z. Zhu, and B. Lin: Design and Implementation of High-Speed Symmetric Crossbar Schedulers. Proc. of IEEE ICC'99, Vol.3, pp.1478-1483, June 1999.

8) K. Kawarai, H. Tomonaga, N. Matsuoka, T. Kato, and A. Hakata: Terabit Switch Architecture using Input Queuing Technique. Proc. of ICCC'99, T-15-04, September 1999.

9) Broadband Switching System (BSS) Generic Requirements. GR-1110-CORE Issue 1, Rev.5, Bellcore, October 1997.

10) Bellcore IMAP Software (Systems Reliability Analysis Software) MS-DOS Version 2.2 Program Description and User's Guide. SP-NPL-000046 Issue 1, Bellcore, January 1987.

Masatoshi Kumagai received the B.E. degree in Chemical Engineering and the M.E. degree in Materials Chemistry from Tohoku University, Sendai, Japan in 1978 and 1980, respectively. He joined Fujitsu Ltd., Kawasaki, Japan in 1980, where he has been engaged in research and development of telecommunication systems.

Satoshi Nojima received the B.S. and M.S. degrees in Electrical Engineering from Waseda University, Tokyo, Japan in 1976 and 1978, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1978, where he has been engaged in research and development of computer communication network systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Hiroshi Tomonaga received the B.S. degree in Electrical and Electronic Engineering from Tokyo Institute of Technology, Tokyo, Japan in 1990. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1990, where he has been engaged in research and development of ATM switching systems and high-speed packet switching systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.