

Carrier-Grade Ethernet Switch for Reliable Wide-Area Ethernet Service

● Tomohiro Ishihara ● Kazuto Nishimura ● Jun Tanaka
● Kazuyuki Miura

(Manuscript received August 5, 2003)

The Wide-Area Ethernet Service is one of the fastest growing services for enterprise users. It connects remote business sites using Ethernet interfaces and provides broader bandwidth at a lower cost than conventional leased-line services. The Wide-Area Ethernet is composed of conventional Ethernet switches that were originally designed for LAN applications. Therefore, its network reliability and QoS (Quality of Service) are insufficient for business applications. In this paper, we show the technical requirements for an Ethernet switch that can be used to build a reliable Ethernet. Especially, we focus on using redundancy to enhance reliability and QoS capabilities. We also describe the basic features of the Wide-Area Ethernet and two new features we are proposing: the Virtual Port and the Directing VWAN (Virtual Wide Area Network). Lastly, we introduce a new line of Fujitsu Ethernet switches that meet these requirements to achieve carrier-grade quality.

1. Introduction

Most enterprises are now seriously interested in using IT (Information Technology) systems to improve their productivity and create new business. IT systems are composed of not only computers but also networks. Wide Area Networks (WANs) that connect headquarters, branch offices, and remote sites are key elements for building regional, nationwide, and worldwide IT systems for enterprises. Enterprises have been using leased-line services provided by telecom operators to connect their remote business sites. The most popular leased-line services are based on technologies such as TDM (Time Division Multiplexing), for example, J1 in Japan, T1 in the US, and E1 in Europe, and ATM (Asynchronous Transfer Mode). Now, a new service in which users connect to remote sites using network services via Ethernet interfaces is attracting attention. This service looks like an extension of an Ethernet LAN (Local Area Network). In Japan, this service is called the Wide-Area Ethernet Service; in the US,

it is called TLS (Transparent LAN Service) or the Metro Ethernet Service. Especially in Japan, the Wide-Area Ethernet Service is one of the fastest growing telecom services. The service is particularly attractive to enterprise IT managers, because it provides broader bandwidth at lower cost compared with legacy leased-line services.

Wide-Area Ethernet Service in Japan is mostly provided over Ethernet-based networks that are constructed using Ethernet switches. Conventional Ethernet switches, which were originally designed for LANs, have poor reliability and QoS (Quality of Service) capabilities. In order to achieve a reliable and feature-rich Wide-Area Ethernet Service, we consider that Ethernet switches must meet new requirements.

In this paper, we discuss the current architecture and issues of the Wide-Area Ethernet. Then, we describe some new requirements for Ethernet switches and propose an architecture and implementation methods for meeting those requirements. We also describe the basic features

of the Wide-Area Ethernet and two new features we are proposing: the Virtual Port and the Directing VWAN (Virtual Wide Area Network). Lastly, we describe the unique Ethernet switches we have developed that meet these new requirements to achieve carrier-grade quality.

2. Wide-Area Ethernet Service

The Wide-Area Ethernet Service provides LAN interconnections that transparently transfer users' Ethernet frames across a WAN. Multiple remote sites are connected via an Ethernet UNI (User Network Interface). **Figure 1** shows a typical network architecture and service. The features of the Wide-Area Ethernet Service are as follows:

- 1) Transparent for any layer-3 protocol

The Wide-Area Ethernet Service is a layer-2 service that enables users to use any protocol of layer 3 and above. This feature is very advantageous for enterprise users because it makes it unnecessary to modify existing network protocols such as IP (Internet Protocol), Apple Talk, and IPX (Internetwork Packet eXchange).

- 2) Not only point-to-point but multipoint

Because Ethernet is a connectionless technology, the Wide-Area Ethernet Service can provide a multipoint-to-multipoint service without special settings. Also, it can provide a broadcast service that can be used for intranet audio/video broadcasting.

- 3) Layer-2 VPN service

The Wide-Area Ethernet Service is a type of VPN (Virtual Private Network) service that enables each user to construct a virtual dedicated network to interconnect their business sites. The VLAN (Virtual LAN) technology¹⁾ is a key for providing this feature. It securely multiplexes the users' traffic onto a physical medium (e.g., optical fiber). A service provider assigns a VLAN ID (Identifier) for each user and then adds a VLAN tag to the users' Ethernet frames as shown in Figure 1. VLAN-tagged Ethernet frames can only traverse via the designated physical port, so us-

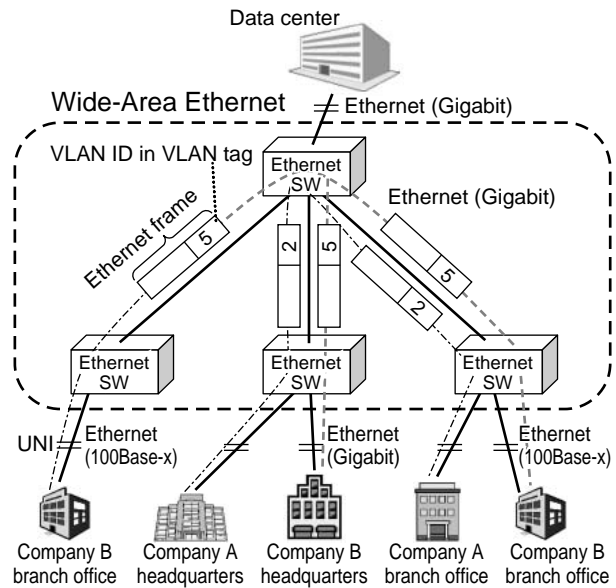


Figure 1 Wide-Area Ethernet and its application.

ers cannot receive the Ethernet frames of another user.

3. Enhancing Wide-Area Ethernet

The Wide-Area Ethernet Service has started to use conventional Ethernet switches originally designed for LAN applications. Therefore, some of the service's features are poor and its overall performance is not as good as that of legacy leased-line services. Especially, the functions related to redundancy and QoS of this service are very limited. We consider that the importance of these functions is increasing because enterprises now require highly reliable IT systems. In this section, we describe the issues related to redundancy and QoS of conventional LAN technology from the viewpoint of the Wide-Area Ethernet Service. Then, we describe the requirements for the new Ethernet switches that will be needed to build reliable networks.

3.1 Redundancy architecture for establishing a highly reliable network

For a system to provide carrier-grade services that meet the requirements of enterprise

users, it must be highly reliable. One practical way to achieve high-reliability in a cost-effective way is to double up the elements; that is, to use a redundancy architecture. Redundancy in networks can be implemented on several levels or layers. Here, we discuss link and node level redundancies because they are key functions in the Wide-Area Ethernet. The following are conventional LAN technologies standardized by IEEE 802 that can be used to implement redundancy:

- 1) Link aggregation.²⁾
- 2) STP (Spanning Tree Protocol) families, including STP,^{3,4)} RSTP (Rapid STP),⁵⁾ and MSTP (Multiple STP).⁶⁾

However, because these functions were not originally developed for redundancy, they do not match the requirements for redundancy in a WAN. The reasons for this are as follows.

Presently, many Ethernet switches perform link aggregation, which establishes a virtual link by bundling physical links (e.g., Fast Ethernet and Gigabit Ethernet) to increase the link capacity. This packet aggregation function can be used for redundancy because, even if one of the bundled physical links fails, the virtual link can continue operating using the other physical links. However, this aggregation function is implemented mostly inside the interface cards, so it cannot be used for interface card redundancy.

The STP and its derivative protocols (RSTP and MSTP) were originally designed to build spanning trees on Ethernet-based networks in order to avoid loops. In other words, the function of the STP families is to build layer-2 logical topologies on physical networks. We can therefore use the STP families to realize redundancy, because they can rebuild a new logical network topology while detouring around faulty links and nodes. However, it takes a relatively long time (seconds or several 10s of seconds) to rebuild a logical topology. This fault-recovery time is much longer than that of traditional leased-line services built on the SONET (Synchronous Optical Network) system, which achieves a 50 ms fault-recovery switching

time on 1+1 APS (Automatic Protection Switching), UPSR (Unidirectional Path Switched Ring), and BLSR (Bi-directional Line Switched Ring).

From the viewpoint of achieving redundancy, not only the switching time but also the bandwidth handling are critical issues in the STP families. It is very difficult to preserve the service's bandwidth using the STP families, because neither the service nor the bandwidth is within the scope of the STP specifications. An example is shown in **Figure 2**. Here, the capacity of each link is 100 Mb/s and the bold lines indicate an active spanning tree built by the STP running on the network. Also, the service guarantees its bandwidth on the network. The bandwidths of services are 70 Mb/s for user 1 and 60 Mb/s for user 2. In the normal state, the required bandwidth for each service is reserved on the network by routing each service on different physical links as shown in Figure 2 (a). If a link failure occurs between node B and D, the STP builds a new tree as shown in Figure 2 (b). On the new tree, the service for user 1 is now routed along node A, C, and D. At this time, the 70 Mb/s and 60 Mb/s services both run on the same 100 Mb/s link, which

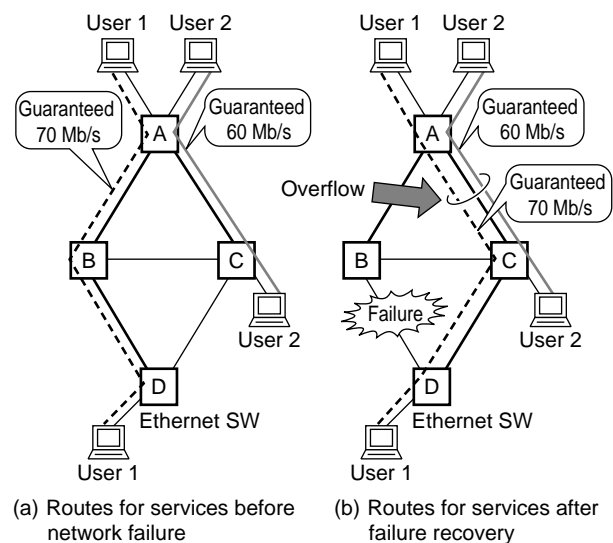


Figure 2 Inconsistency between STP and guaranteed bandwidth service.

is too small to handle the combined 130 Mb/s bandwidth. This means that the system cannot provide a guaranteed bandwidth service. In order to avoid this kind of overbooking, the network operator must route services and assign bandwidth in a way that considers all possible spanning tree topologies that can arise due to link failures. However, this is not a practical approach, so STP-implemented redundancy is not a practical way of providing a guaranteed bandwidth service.

Taking account of the above issues, we propose the following redundancy architecture (**Figure 3**):

- 1) Full-redundancy for the most critical elements, for example, the control processors, fans, and power units.
- 2) No interruption of Ethernet frame switching and forwarding during suspensions of the control CPU.
- 3) 1+1 Ethernet APS to achieve fast-recovery and preserve the service bandwidth.

Here “1+1 Ethernet APS” is our newly developed technology, which achieves rapid (less than 50 ms) fault-recovery switching between working and protection (backup) Ethernet ports. It is a link-layer unidirectional switching technology in which two sender ports (working and protection) send the same packet stream and one of the ports is chosen at the receiver side.

Using the above architecture, an Ethernet-based, high-reliability network can be achieved.

3.2 QoS capabilities

The QoS capability of Ethernet is relatively poor. Although it is sufficient for LAN applications, the IEEE 802.1p only defines priority-control. On the other hand, carrier-provided services like Wide-Area Ethernet must have richer QoS capabilities to carry data, voice, video, and other kinds of traffic. In this section, we discuss the QoS capabilities required for carrier-grade services.

There are two types of QoS: priority-based QoS and bandwidth-based QoS. Priority-based QoS is achieved by identifying the packet forwarding priority that is set based on a certain policy. Bandwidth-based QoS is achieved by reserving buffer-memory capacity and reserving memory read-out cycles to provide guaranteed bandwidth.

These two types of QoS are used not exclusively but simultaneously. For instance, a 10 Mb/s service connects two business sites, and a 1 Mb/s band within the service bandwidth is given a higher priority than the other bands so it can be used for low-latency forwarding of VoIP (Voice over IP) traffic. Therefore, the Ethernet switch for carrier-grade services must provide these two types of QoS simultaneously and independently.

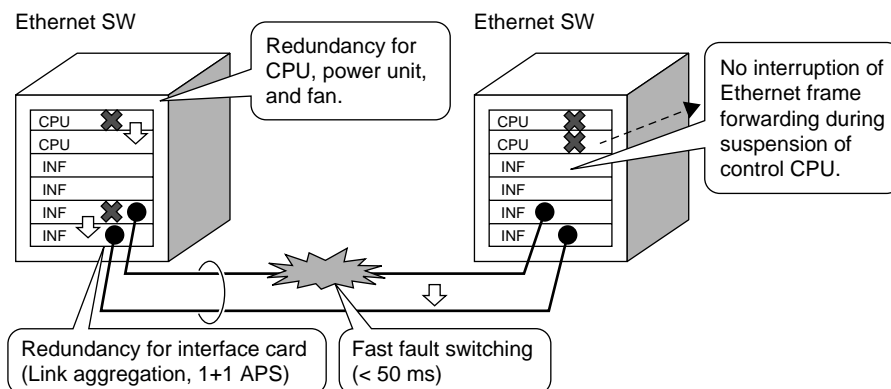


Figure 3
Ethernet switch architecture for achieving high reliability.

3.2.1 Priority-based QoS: requirements and implementation methods

There are a number of priority controls based on prioritized parameters. The two most popular ones are latency priority control and discard priority control.

To achieve latency priority control, queues are established for each priority class and the read-out cycle for each queue is uniquely differentiated. There are two ways to differentiate between read-out cycles: 1) absolute preference and 2) the weighted read-out cycle (e.g., WFQ [Weighted Fair Queuing]). Taking into account the fact that VoIP (Voice over IP) traffic should be carried with the lowest possible delay, absolute preference is better than WFQ. This is why WFQ makes the highest-priority packets wait until the lower-packets have left the queue, adding a delay to the highest priority traffic. **Figure 4** shows a three-queue example in which a read-out scheduler is used for latency priority control.

Discard priority control differentiates between discarded lower-priority and higher-priority packets when network congestion reaches a certain level. For example, in Figure 4, two threshold levels are set up to perform discard priority control as follows:

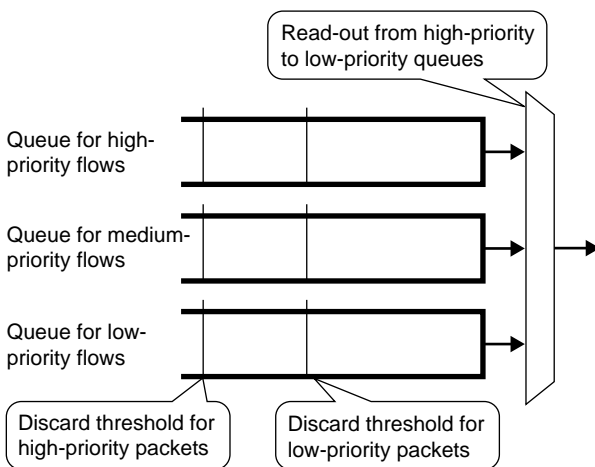


Figure 4 Queuing architectures for priority-based QoS.

- 1) When packet accumulation reaches a low-level, only the lower-priority packets are discarded.
- 2) When packet accumulation reaches a high-level, the higher-priority packets are discarded as well as the lower-priority ones.

Discard priority control when combined with, for example, a policer at the ingress, provides a committed-rate (or minimum-rate guarantee) service. The policer measures the incoming traffic of each service and marks a packet as green, yellow, or red according to its rate.⁷⁾ Green packets are under committed, yellow packets are over committed but under the peak rate, and red packets are over the peak rate. Red packets are discarded immediately at the ingress, while green and yellow packets enter the network. At the packet queue in the network, only yellow packets are discarded when packet accumulation reaches the low threshold. In other words, green packets can go through the queues even at this congestion level, thus ensuring a committed-rate service.

3.2.2 Bandwidth-based QoS: requirements and implementation methods

The service bit-rates of the initial stage of the Wide-Area Ethernet Service are the same as the interfaces' physical bit-rates (e.g., 100 Mb/s or 1 Gb/s). These services provide too much bandwidth and are costly for medium and small business sites. Therefore, smaller-bandwidth services, for example, from several to several 10s of Mb/s, should be used for those applications. To provide these services, the physical port bandwidth must be divided into smaller bands. To achieve this bandwidth control, packet-flow policing and shaping are fundamental functions. Such functions have been implemented on existing Ethernet switches, but there are only a few Ethernet switches that fulfill the requirements for carrier-grade services. Their granularity, policing accuracy, and shaping accuracy are not fine enough. In this section, we discuss the capabilities that Ethernet switches must have to provide

carrier-grade services.

The policer must operate on each packet flow independently, and the flow must be identified not only by user but also by service. The Wide-Area Ethernet Service uses VLAN ID or its extension (stacked VLAN tag) in Ethernet packets for user identification. At the same time, multiple service classes are provided on a user's packet flow. Therefore, policing must identify VLAN IDs and services. The policer must also measure the flow as accurately as possible, because the measurement results are the most critical factors that define QoS. The reference model for measuring accuracy is the token-bucket based on a one-byte token. Granularity is another critical parameter. Taking account of the application of the Wide-Area Ethernet Service, the required granularities are sub-Mb/s for narrowband services to small sites and 1 Mb/s for broadband services over 1 Mb/s.

The shaper must operate on the shaping queues that correspond to the output ports and are independent of the priority queues. There are two methods of associating an individual shaping queue with a flow. The first method is to associate the queue with an individual user flow indicated by a VLAN ID. This method is used when a strictly-guaranteed bandwidth service can be provided for each user. The second method is to associate the queue with a user flow group or VLAN-group. The second method is used when bandwidth sharing between users is provided for a committed and best-effort service. The accuracy and granularity of the shaper's buffer read-out control must be the same as those of the policer.

3.2.3 Our proposed implementation method

Taking account of the above discussion, we propose the QoS implementation method shown in **Figure 5**. This method features the following:

- 1) Policing of both users' traffic flows and service flows.
- 2) Policing at the peak and committed rate (judgment of green, yellow, or red).
- 3) Accurate policing and shaping using one-byte

tokens.

- 4) Granularity of sub Mb/s domain for policing and shaping.
- 5) Shaping on shaping queue corresponding to output port.

The incoming packet flow first reaches a policer that has features 1) to 3). Then, each packet is allowed to go forward (green or yellow) or is discarded (red). The output packet flow goes through a shaper that has features 4) and 5) so that latency and discard priority control are achieved. Using this architecture, we can provide carrier-grade QoS services.

4. Advance features

In Section 3, we described the fundamental requirements and implementation methods for achieving a carrier-grade Ethernet service. In addition to those features, we have developed two unique features for improving Ethernet services: the Virtual Port and the Directing VWAN.

4.1 Virtual Port

A practical solution for providing a cost-effective service to MTUs (Multiple Tenant Units) is the "remote concentration" configuration. In

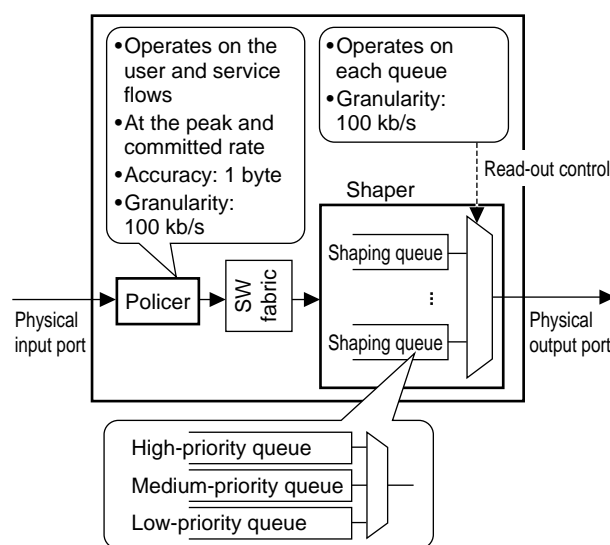


Figure 5 Queuing architecture for bandwidth-based QoS.

this configuration, the tenants share, with other subscribers, a subscriber line (optical fiber) that connects their building to the service provider's office. **Figure 6** shows an example of this configuration realized using conventional port VLAN technology. Ethernet switch A is located in the user's building to multiplex all of the tenants' traffic onto a subscriber line. The physical port of switch A is used to identify the traffic source (e.g., physical port 1 is for office A of tenant 1). Then, switch A adds VLAN tags to identify each Ethernet frame (e.g., the VLAN tag of tenant 1 is 59) and multiplexes all incoming frames onto a fiber.

Switch B in the provider's office demultiplexes the incoming frames using VLAN tags.

This architecture has the limitation that tenant 1's traffic for office A and B cannot be distinguished from each other because it has the same user's VLAN tag. Therefore, the offices cannot be allocated different bandwidths (e.g., 10 Mb/s for office A and 5 Mb/s for office B). To solve this problem, we propose the "Virtual Port" method shown in **Figure 7**. In this method, a partial identifier (Virtual Port [Vport] ID) is assigned for each physical port at switch A, and switch B converts the Vport IDs into VLAN IDs (i.e., the user

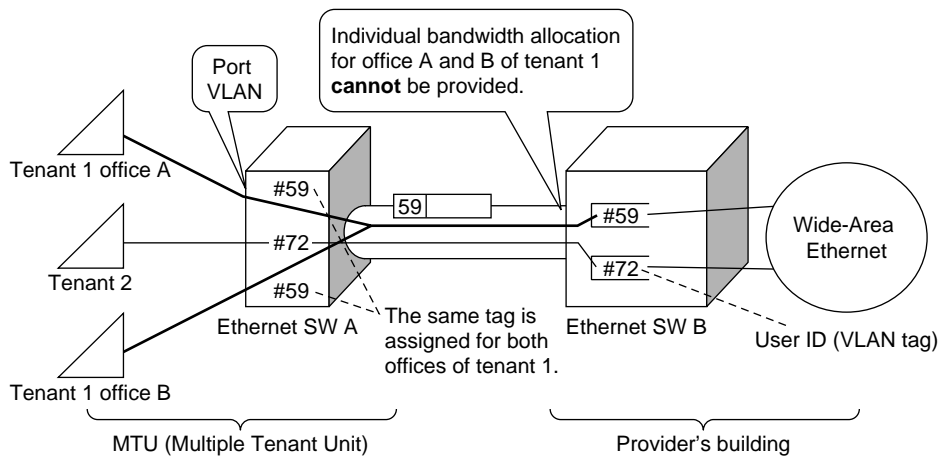


Figure 6 Remote concentration using port & tag VLAN.

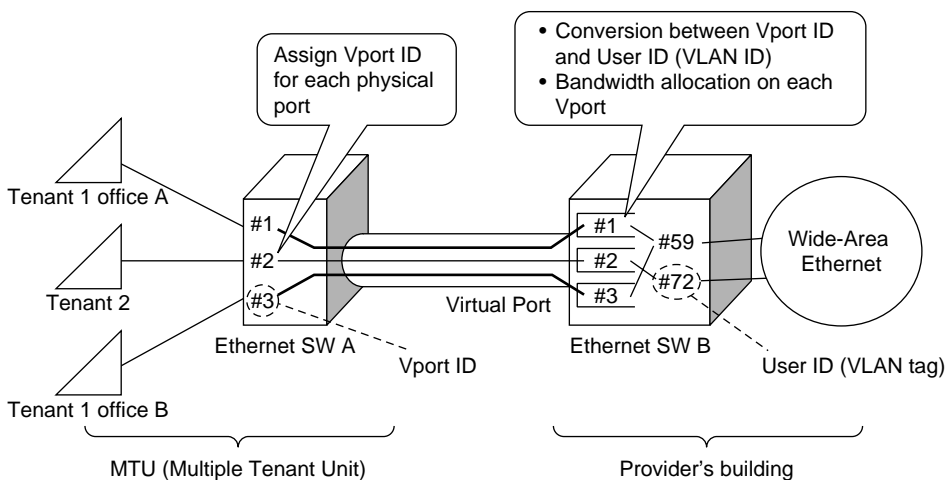


Figure 7 Remote concentration using Virtual Port.

IDs used in the Ethernet service network). The key here is that switch B performs QoS control (e.g., bandwidth assignment) on the Vport IDs and not on the VLAN tags. Therefore, QoS control for individual services can be performed on a remote concentration application.

4.2 Directing VWAN

In a tag VLAN application, VLAN IDs are only used for identifying users and the system cannot distinguish between different services used by the same user. For instance, as shown in **Figure 8**, service A-B (between site A and B) and service A-C (between site A and C) cannot be identified on the transit link because all packets have the same VLAN ID. However, in such an application, they

must be identified because service A-B and service A-C may have different properties (e.g., different bandwidths and priorities). Using conventional Ethernet technologies, the only way to identify services in a transit link is to use separate physical links for each user, which is a very costly solution.

To solve this issue, we propose a new feature for VLAN ID called the Directing VWAN. As shown in **Figure 9**, the Directing VWAN adds an additional VLAN ID2 to identify the directions of the service and VLAN ID1 is used to identify users as in the conventional method.

The Directing VWAN provides not only direction identification but also output shaping for each direction. Stated differently, Directing VWAN assigns a kind of end-to-end path for each

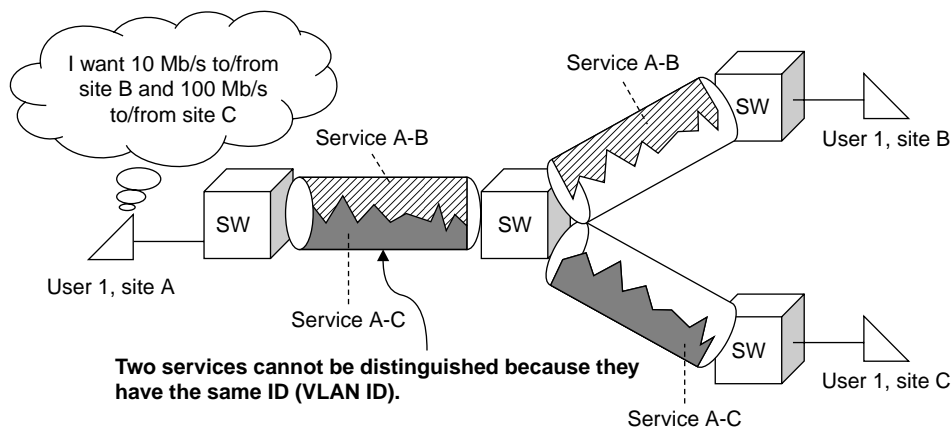


Figure 8
No individual QoS control on service A-B and A-C using conventional VLAN.

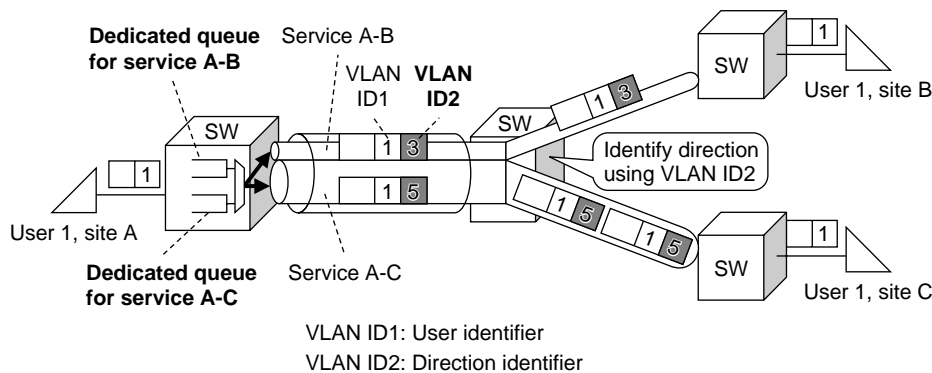


Figure 9
Individual QoS control on service A-B and A-C using Directing VWAN.

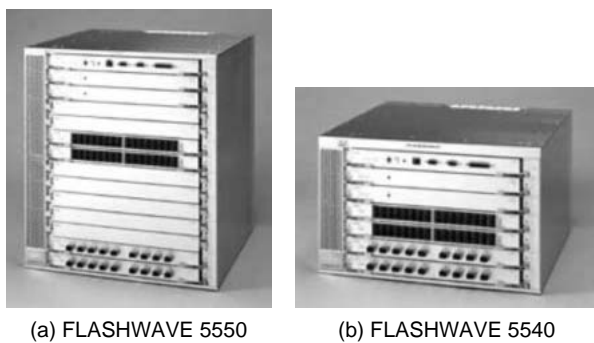


Figure 10
Fujitsu carrier-grade Ethernet switch.

service. Its concept is similar to that of EoMPLS (Ethernet over Multi-Protocol Label Switching), but it has the following advantages over EoMPLS:

- 1) It is much easier to implement than EoMPLS.
- 2) No IP routing protocol is needed. (EoMPLS needs complicated routing protocols.)
- 3) Tandem switches can be constructed using simple Ethernet switches that just perform VLAN tag switching.

5. FLASHWAVE 5540/5550

We have developed Ethernet switches that meet the requirements described in the previous sections. FLASHWAVE 5540 and 5550 are feature-rich Ethernet switches for achieving reliable Ethernet services and are very different from conventional Ethernet switches designed for LAN applications. **Figure 10** shows a photograph of our new switches, and their specifications are shown in **Table 1**.

Our switches are now used by telecom operators to provide Wide-Area Ethernet services in Japan.

6. Conclusion

In this paper, we discussed the requirements for achieving reliable Ethernet services, including redundancy and QoS capabilities, and some example implementation methods for meeting these requirements. We also described two new

Table 1
FLASHWAVE 5550/5540 specifications.

Description	Specification	
	FLASHWAVE 5550	FLASHWAVE 5540
Performance	80 Gb/s switch fabric	160 Gb/s switch fabric
Interface	Fast Ethernet: 10/100 Base -TX × 32 ports Gigabit Ethernet: 1000 Base - SX/LX × 8 ports	
Redundancy (modules)	Full redundancy for CPU, power, fan units	
Redundancy (interface)	Link aggregation between interface cards 1+1 APS between interface cards	
Priority control	3-class queues, 2 discard priorities, absolute preference	
Policer	Peak & committed rate policing Accuracy: 1 byte, Granularity: 500 kb/s minimum	
Shaper	Range: 500 kb/s to 1 Gb/s Granularity: 500 kb/s minimum	

features we have developed called the Virtual Port and the Directing VWAN.

Lastly, we introduced a new line of Fujitsu Ethernet switches that meet these requirements. Our new Ethernet switches are being deployed and are providing very reliable Ethernet services.

References

- 1) Virtual Bridged Local Area Networks—Amendment 4: Provider Bridges, IEEE Draft 802.1ad.
- 2) Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications, IEEE std 802.3-2002.
- 3) Media Access Control Bridges, IEEE std 802.1D-1998.
- 4) Media Access Control Bridges: Technical and Editorial Corrections, IEEE std 802.1t-2001.
- 5) Media Access Control Bridges--Amendment 2--Rapid Reconfiguration, IEEE std 802.1w-2001.
- 6) Virtual Bridged Local Area Networks--Amendment 3: Multiple Spanning trees, IEEE std 802.1s-2002.
- 7) J. Heinanen and R. Guerin: A Two Rate Three Color Marker. RFC 2698, 1999.



Tomohiro Ishihara received the B.E. and M.E. degrees in Electronics and Communication Engineering from Waseda University, Tokyo, Japan in 1983 and 1985, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1985, where he has been engaged in research and development of high-speed optical transmission systems, broadband optical access systems, and Ethernet-based transport

systems. He was a visiting researcher at the University of California at Berkeley, USA for a year between 1992 and 1993. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the IEEE.

E-mail: t.ishihara@jp.fujitsu.com



Jun Tanaka received the B.E. and M.E. degrees in Electronics and Communication Engineering from Tohoku University, Sendai, Japan, in 1987 and 1989, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1989, where he has been engaged in research and development of digital video coding systems, ATM-based transmission systems, broadband access systems, and Ethernet-based transport systems.

He was a visiting researcher at the University of California at Berkeley, USA for a year between 1995 and 1996. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the IEEE.

E-mail: tanaka.jun.777@jp.fujitsu.com



Kazuto Nishimura received the B.E. and M.E. degrees in Communication Engineering from Osaka University, Osaka, Japan in 1994 and 1996, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1996, where he has been engaged in research of ATM traffic control. He is currently engaged in research and development of Ethernet-based transport systems.

He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

E-mail: nisimura.kazuto@jp.fujitsu.com



Kazuyuki Miura received the B.E. degree in Electronics and Communication Engineering from Waseda University, Tokyo, Japan in 1982. He joined Fujitsu Ltd., Kawasaki, Japan in 1982, where he has been engaged in development of synchronous/asynchronous multiplexer systems. He is currently developing Ethernet transport systems.

E-mail: miura.kazuyuki@jp.fujitsu.com