INVITED PAPER *Special Section on the Next Generation Ethernet Technologies*

Carrier-Grade Ethernet Technologies for Next Generation Wide Area Ethernet

Atsushi IWATA[†a)], ***Member***

SUMMARY This paper describes an overview of overall carrier-grade Ethernet technologies for next generation wide area Ethernet. In recent years, from access network to metro and core network, we can find many areas where communication services are provided by Ethernet technologies. This comes from the fact that operational efficiency and economical efficiency of Ethernet are far better than that of conventional wide area communication technologies such as SONET and ATM. On the other hand, carrier-grade reliability, operations-administration-maintenance (OAM) and quality of service (QoS) are inferior to SONET and ATM. Various standard schemes in IEEE 802 and ITU-T and vendors' proprietary schemes can leave various approaches to solve these problems. In this paper, the author explains a basic architecture of wide area Ethernet service (Q-in-Q tagging for metro network and Mac-in-Mac encapsulation for core network) at first. Various switch control technologies are then discussed which are deployed or are under evaluation in order to improve (i) reliability (i.e., resiliency) to protect subscribers against network failures, (ii) OAM for providers to perform fault and performance management, and (iii) QoS to guarantee subscriber's service level agreement between a carrier and a subscriber. Finally, a new switching architecture, Global Open Ethernet (GOE), is also introduced as one of promising approaches to realize a next generation carrier-grade Ethernet.

key words: *MAN, Ethernet, reliability, QoS, OAM*

1. Introduction

In recent years, Ethernet has begun to be deployed successfully in Wide Area Network (WAN) as well as Local Area Network (LAN), where Ethernet is a dominant and de-facto-standard technology of computer communications. From access network in FTTH to metro and core network, the number of communication service provided by Ethernet technologies are gradually increasing. This comes from the fact that operational efficiency and economical efficiency of Ethernet are far better than that of conventional wide area communication technologies such as SONET and ATM. In addition, by high speed and high capacity service of the maximum link speed 10 Gigabit Ethernet, the spread of wide area Ethernet technologies will be accelerated. On the other hand, carrier-grade reliability, OAM and QoS are inferior to other communication services such as SONET and ATM because Ethernet was started as LAN technology.

Section 2 describes an overview of basic architecture of wide area Ethernet service (Q-in-Q tagging for metro network and Mac-in-Mac encapsulation for core network). In Sects. 3, 4 and 5 various switch control technologies are then discussed which are deployed or are under evaluation in order to improve (i) reliability (i.e., resiliency) to protect subscribers against network failures, (ii) OAM for providers to perform fault and performance management, and (iii) QoS to guarantee subscriber's service level agreement between a carrier and a subscriber. In terms of (i), Sect. 3 explains various reliability control technologies that mostly come from vendor proprietary extensions of Ethernet switch control, some of which have been standardized. In terms of (ii), Sect. 4 then introduces Ethernet OAM control technologies under discussion and raises remaining issues to solve. Since the original Ethernet standard itself does not specify OAM functionalities for WAN, IEEE and ITU-T are now standardizing Ethernet OAM specification in a close relationship. In terms of (iii), in Sect. 5, priority control, bandwidth control, and fairness control are introduced as currently used technologies, and remaining issues are also discussed. Finally, a new switching architecture, Global Open Ethernet (GOE), is also introduced as one of promising approaches to realize a next generation carrier-grade Ethernet.

2. Basic Architecture of Wide Area Ethernet Service

Wide area Ethernet is a technology of "Forwarding Ethernet frame as is over a WAN environment" in communication among remote sites geographically [1]. Now most of user LANs in the enterprise are built by Ethernet, and those LANs can be inter-connected through provider's wide area Ethernet service, where the interconnected Ethernet segments can become a large flat network virtually on layer 2 level. This section describes how such a communication service can be provided.

2.1 Wide Area Ethernet Technology Overview

Wide area Ethernet technologies are categorized into those applied to access network, metro network, and core network, as depicted in Fig. 1. Access network is a domain from a subscriber to a provider's access switch, and metro network is a domain from access switch to a switch connected to a core network, and core network is a domain from a backbone switch to Point of Interest (POI) of an Internet backbone. There have been various technologies proposed for an access network. One of standardized technologies for it is IEEE 802.3ah, or 'Ethernet in the First Mile (EFM),' which utilizes Ethernet with additional link-level OAM functions. Such access network technologies will be

Manuscript received July 17, 2005.
Manuscript revised September 29, 2005.

[†]The author is with System Platforms Research Laboratories, NEC Corporation, Kawasaki-shi, 211-8666 Japan.
a) E-mail: a-iwata@ah.jp.nec.com

DOI: 10.1093/ietcom/e89–b.3.651

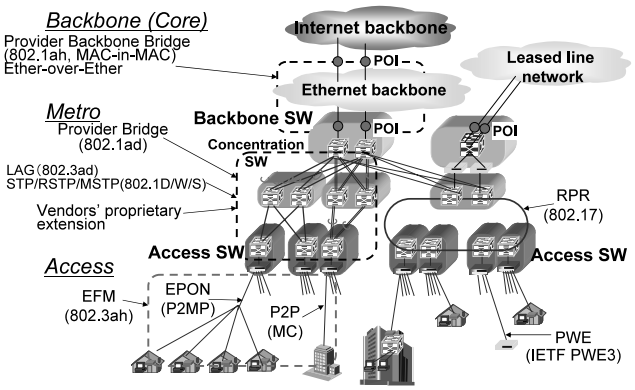


Fig. 1 Wide area Ethernet technologies.

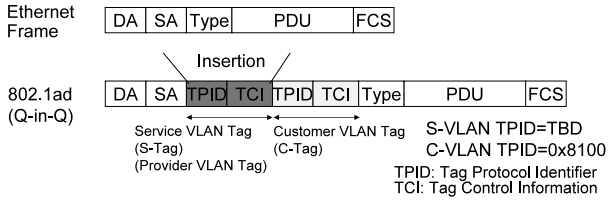


Fig. 2 Frame format of IEEE802.1ad.

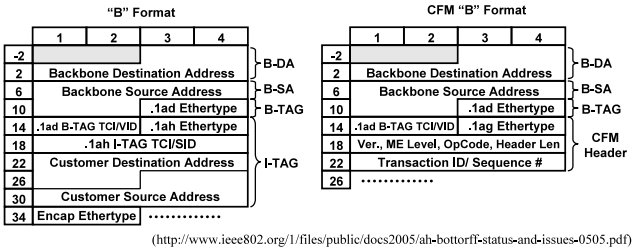
discussed in a separate paper, and metro and core network technologies are highlighted in Sects. 2.2 and 2.3, respectively. Note that technologies themselves do not clearly specify limitations of their applicability to specific target networks, thus either technologies discussed below may be used in either network.

2.2 Provider Bridge (Q-in-Q) in Metro Network

Provider bridge technology under standardization (IEEE 802.1ad [2]: Q-in-Q) is most commonly used by current wide area Ethernet providers as a basic bridging architecture. Simple reliability, OAM, and QoS control are also utilized as additional functionalities.

Q-in-Q extends user VLAN technology in IEEE 802.1Q [3], and doubles VLAN tag field (i.e., VLAN tag stacking) to increase the number of subscribers or customers in a provider network. The double VLAN tagging consists of two kinds of tags, Service-VLAN tag (S-Tag) for a provider and Customer-VLAN tag (C-Tag) for a customer. S-Tag is assigned by the providers to identify the customers' ID and to partition customer's network resources from other customers. C-Tag is assigned by customers' themselves to identify VLAN IDs allocated to different organizations within a customer. S-Tag is inserted before C-Tag of a user-frame at an ingress edge switch of a provider as shown in Fig. 2. Both S-Tag's VLAN ID (or customer ID) and customer destination MAC address (DA) are only referred to decide a route within a provider's network. S-Tag is then stripped off at an egress switch, and the user frame arrives at a destination VLAN.

Using this method, every switch in the provider network must keep a huge size of MAC address learning table, which stores all the number of active MAC addresses in whole subscribers, thus it leads to network scalability problems. In addition, a customer ID space in S-Tag is limited to 4094, and needs to be extended for accommodating a large number of customers. There is also another limitation of data transfer. If a customer wants to transmit transparently bridged PDU (BPDU) of spanning tree protocol (STP), for example, over a Q-in-Q network, specific MAC DA used in BPDU cannot be transmitted through the network. This is due to the fact that the same MAC DA may be used to control "provider" switches and that frames having such MAC DA from customers have to be discarded at the provider's edge switch. Note that although this can be solved by a special technique using MAC address translation functions at the provider edge, MAC dependent data transfer is a nuisance problem.



(http://www.ieee802.org/1/files/public/docs2005/ah-bottorff-status-and-issues-0505.pdf)

Fig. 3 Frame format of IEEE802.1ah.

2.3 Provider Backbone Bridge (MAC-in-MAC) in Core Network (Backbone)

Since Q-in-Q has a scalability problem of number of MAC entries used by customers, provider backbone bridge (IEEE802.1ah [4]: Mac-in-Mac), which is now being standardized in IEEE, has been proposed to solve this problem.

MAC-in-MAC basically uses a provider-MAC based tunneling scheme where an ingress provider-edge switch encapsulates a customer's frame by a provider MAC header, whose SA is an ingress provider-switch MAC address and whose DA is an egress provider-switch MAC address. The frame format, Backbone Format ("B" Format) is depicted in Fig. 3. In a core network, a core switch does not refer to a customer's MAC address, but refer only to the provider MAC address. Thus, the number of MAC entries that provider switches have to support can be dramatically reduced to the magnitude of the number of provider edge switches. In addition, in terms of supporting the number of customer ID space, MAC-in-MAC can also increase a customer ID space from 4 k to 16 M by using two VLAN-tags (.1ad B-Tag and .1ah I-Tag) for identifying customers.

In conjunction with IEEE802.1ah, OAM and Connectivity Failures Management (CFM) for a network-level fault management are also being discussed in 802.1ag [5]. In addition, PoweredCom has proposed Ethernet over Ethernet (EoE) technology [1], [6] whose basic concept is the same as that of 802.1ah, and is currently operating this. EoE sup-

ports unique features, Time to Live (TTL) in MAC level to prevent a network meltdown due to loops of bridges and various OAM (i.e., EoE ping and traceroute), although it uses a proprietary different frame format from 802.1ah.

3. Reliability Control Technologies

To improve service availability for wide area Ethernet, various switch control technologies have been proposed; (1) link failure detection control, (2) link redundancy control, (3) loop prevention control, (4) optimized path control, (5) node redundancy control, (6) ring network redundancy control, (7) non-stop network operation control.

3.1 Link Failure Detection Control

When a uni-directional link failure occurs, it may cause a loop of bridges in a network (i.e., causing network meltdown), unless a status of that interface is immediately changed to the link-down in both directions. To prevent such a loop, it is important to reduce a detection time of a uni-directional link failure and to use strong fault monitoring functions that can be used in various network designs.

A fault monitoring function on a physical layer, there are Auto-negotiation/Remote Fault (RF) function [32] in Gigabit Ethernet (GbE) and the Link Fault Signaling (LFS) function [33] in 10 GbE. On the other hand, along the link segmented by a media converter or a transmission system, failure information may not be notified beyond segments. A fault monitoring function depending only on the physical layer is not sufficient. Therefore, the same function implemented in a MAC layer has been proposed as a vendor proprietary function, or Uni-Directional Link Detection (UDLD) method [7]. This method sends a "keep-alive" signal by using MAC frame between both ends of the link each other. When "keep-alive" signal does not arrive during a specific timeout, the switch let the interface to link-down immediately in both directions.

3.2 Link Redundancy Control

Link Aggregation (LAG) (IEEE802.3ad) is a well-known function that uses multiple links (2-to-8) as a virtualized single fat link for increasing a capacity of a link between adjacent switches. Since this function can also be used for load distribution along the multiple links, it can contribute to recover a link failure in an automatic manner, which can be regarded as a link redundancy control.

3.3 Loop Prevention Control

As already mentioned, when broadcast and unknown frames get loop in Ethernet bridges, it leads to a serious network failure. To solve this problem, several loop prevention functions, STP (802.1D) [8], RSTP (802.1w) [9], MSTP (802.1s) [10], have been standardized in IEEE and are being used to generate a loop-free tree topology. However, since STP and RSTP have several corner cases where they are not sufficiently stable, and they are also not originally designed for a large network, several vendor proprietary functions have been proposed to make them more reliable, stable, and scalable to solve these problems. Here, let me assume that the network operators use STP/RSTP for creating a loop-free topology in the provider network.

3.3.1 Root Guard

The standard STP/RSTP does not allow network operators to configure a tree topology in a static manner. If an internal provider bridge is mis-configured to let it become a root bridge of a spanning tree, a new root bridge election is started to change the tree topology, which may cause to drop the customer's traffic until the election is ended. Root guard is a function of allowing a network operator to manually specify a root bridge of the core network to stop automatic election of root node so that mis-configuration of provider bridges can be solved [11].

3.3.2 Loop Guard

When a BPDU timeout of STP/RSTP occurs by CPU processing overload or a uni-directional link failure of the switch, a loop may occur. Loop guard function solves this problem. Loop guard is configured at ports (usually blocked-link status) toward a root bridge not to change the port status (rather notifying an event of loop mismatching), even if the switch does not receive BPDU from adjacent switches [12].

3.3.3 SuperSPAN

SuperSPAN is a function to partition a large STP/RSTP domain into several sub-domains (core and edge domains) to minimize an impact of a network failure by localizing it in each domain. STP/RSTP runs in each domain independently, and in case of communication between edge domains, STP/RSTP of edge domains can be connected via tunneling (Q-in-Q tagging and BPDU tunneling) in a core domain, and STP/RSTP function works only in Edge-to-Edge as shown in Fig. 4 [13].

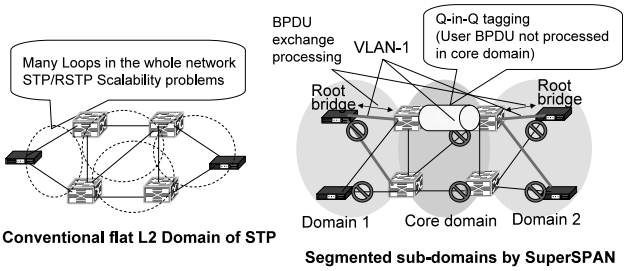


Fig. 4 Conventional STP single domain and sub-domains by SuperSPAN.

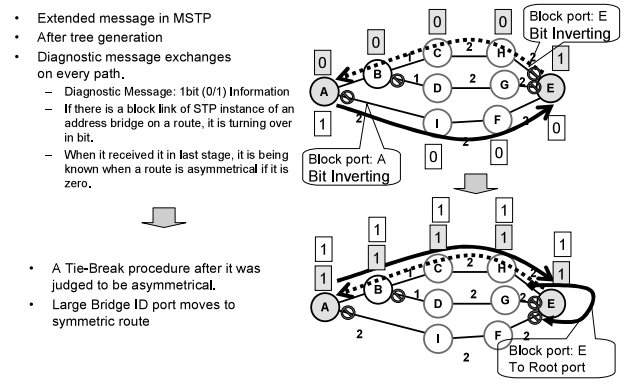


Fig. 5 Solution of asymmetrical link in MSTP Reflection Vector.

3.4 Optimized Path Control

STP, RSTP and MSTP can find the optimized-path from a root bridge to other leaf bridges, based on a link cost using an inverse of physical link bandwidth, in default, along a path. This path computation is a bandwidth-intensive one and is regarded as a static traffic engineering based on a physical link bandwidth. This is not a dynamic traffic engineering (such as a residual-link-bandwidth based path computation) that MPLS can provide. Although the path from a root bridge to each leaf bridge in a network is always the shortest path or optimal path, some of paths between leaf bridges may not be so.

In order to establish the optimal paths among all bridges, venders' proprietary approaches, per-destination multiple rapid STP (PD-MRSTP) [14] in Global Open Ethernet (GOE), as described in Sect. 6, and MSTP Reflection Vector (as shown in Fig. 5) [15] have been proposed. For finding the optimal paths, both approaches create multiple spanning trees having different VLAN instance, where every bridge becomes a root for its own VLAN instance. In the Ethernet bridge processing, a bi-directional communication must be done along a symmetrical path, MSTP Reflection Vector extends MSTP to include additional procedure, by which an asymmetrical path can be shifted to a symmetrical path, as shown in Fig. 5.

3.5 Node Redundancy Control

Loop-free network-topology-configuration method such as STP and RSTP can also provide a link-failure recovery mechanism in a network. However, the failure recovery time for STP may need 30 seconds and that of RSTP may need a few seconds in the worst case (i.e., root bridge failure). Thus, vendors' proprietary node redundancy controls, active/active and active/standby redundancy controls, have been proposed to accelerate the failure recovery performance.

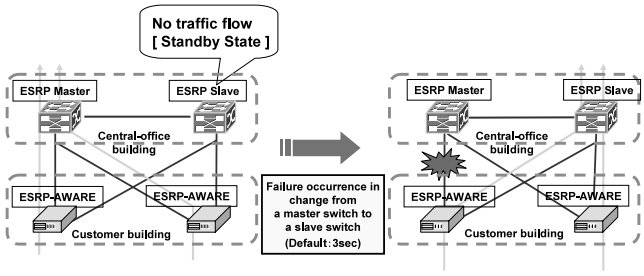


Fig. 6 Node redundancy configuration in active/standby.

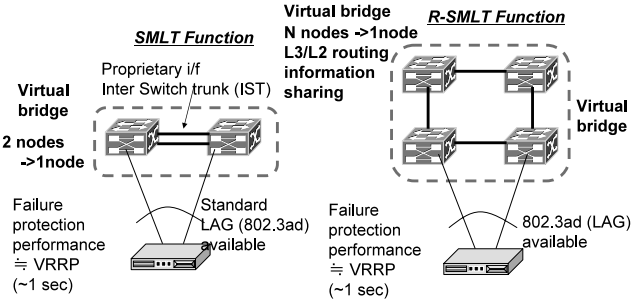


Fig. 7 Node redundancy configuration in active/active.

3.5.1 Active/Standby Redundancy Control

ESRP (Extreme Standby Routing Protocol) [16] is a well-known active/standby redundancy control, as shown in Fig. 6, which works in Layer 2 level and behaves like Virtual Router Redundancy Protocol (VRRP) of IP routers. A standby (or slave) bridge can take over the master function of an active (or master) bridge, once detecting a failure of the active bridge by exchanges of periodical keep-alive messages between both bridges. This mechanism allows us to localize the influence of any network failure to a limited segment (ESRP segment). However, since this protocol has a corner case where both bridges become master (dual-master's problem) and a network gets loop, it further requires a loop-prevention function (i.e., ELRP) to solve this problem.

3.5.2 Active/Active Redundancy Control

As methods of active/active redundancy bridge controls, there are Split Multi-Link Trunking (SMLT) (as shown in Fig. 7) [17], Expandable Resilient Networking (XRN) [18] and StackWise [19]. Each method provides a virtual bridge function where multiple physical bridges can be interconnected or be stacked via proprietary inter-switch links to emulate a single virtual bridge in total. When one of physical bridges has a failure, a proprietary control protocol fixes the interconnection to continue to work. Since LAG (802.3ad) can be executed on any ports of different physical bridges inside of the virtual bridge, it is easy to provide a link/node redundancy function. In addition to layer 2 bridging information, layer 3 routing information is also distributed among

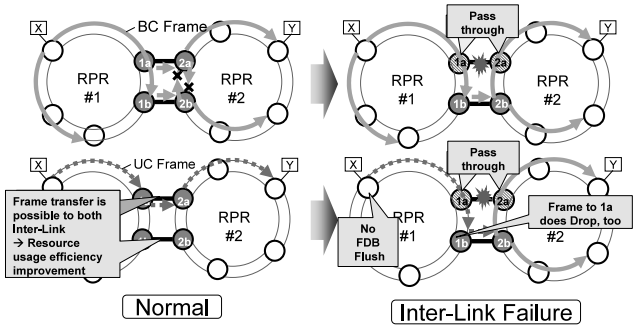


Fig. 8 Inter-Link redundancy in multiple-ring RPR.

those bridges to behave a single virtual router control.

3.6 Ring Network Redundancy Control

A ring network is often used in an environment where a failure protection delay is critical and sufficiently small, since an Ethernet-based mesh or tree networks cannot always provide such a short recovery time. The IEEE 802.17 Resilient Packet Ring (RPR) [20], which uses a ring-intensive different MAC layer from 802.3 Ethernet, is a superior method as a packet ring control for this failure protection performance. Although RPR has been deployed in provider networks for current services, it still needs to solve the new requirements from customers: (i) a large-scale ring network via a multiple-ring topology, (ii) localization of link/node failure, (iii) high-speed failure recovery in multiple-ring environment, and (iv) load balancing functions. We propose new scalable multiple-ring architecture [21] for RPR, where a 50 msec protection can be simply realized in any failures even in inter-link bridges as shown in Fig. 8.

On the other hand, EAPS (Extreme Automatic Protection Switching) [22] and MRP (Metro Ring Protocol) [23] have been proposed as a 802.3 MAC based ring network with additional control procedures which provides a modest failure recovery (around 1–2 seconds), and also prevents a network loop.

3.7 Non-stop Network Operation Control

We have proposed an In-Service Reconfiguration (ISR) function [1], [24] in GOE system, where a network operation can be continued without any interruptions (i.e., zero-loss) even during a topology change via adding or removing a bridge in the network. ISR maintains dual spanning trees, active and standby trees. The active tree is continued to use for a current operation. The standby tree, on the other hand, can be re-configured for a new topology without any impacts to the active tree. Once the standby spanning tree becomes stable, a network administrator can switch the active tree to the standby one without any packet loss.

3.8 Applicability Statement of Reliability Controls Combinations

Link failure detection in Sect. 3.1, link redundancy control in Sect. 3.2, and ring network redundancy control in Sect. 3.6 can be used independently with other reliability control schemes. Loop prevention control in Sect. 3.3, optimized path control in Sect. 3.4 and non-stop network operations in Sect. 3.7 can coexist with each other and work on a basis of STP/RSTP/MSTP network. Node redundancy control in Sect. 3.5, on the other hand, can work only with proprietary redundant protocols and cannot coexist with reliability control schemes using STP/RSTP/MSTP.

4. OAM Control Technologies

The current Ethernet OAM functionality is not sufficient to support carrier-grade network operations in terms of (i) failure detection and notification in multipoint-to-multipoint (MP-to-MP) network, (ii) recognization of accurate failure location, (iii) acceleration of failure recovery. The ITU-T and IEEE are standardizing the Ethernet-based OAM technologies in a close relationship. The SG13 and SG15 of ITU-T studies Ethernet OAM (Y.17ethoam) for (i) and (ii), and Ethernet protection (G.ethps) for (iii), respectively. IEEE 802.1ag WG studies Connectivity Fault Management (CFM), or a network level OAM for (i) and (ii), while IEEE802.3ah (EFM) has already standardized a link-level OAM. The network level OAM is supposed to utilize the standard link-level OAM in some cases. Here we discuss the network level OAM of ITU-T and IEEE, especially for (i) and (ii).

4.1 OAM Mechanisms

Ethernet OAM (Y.17ethoam) [25] is categorized into 8 functions: Ethernet Continuity Check (ETH-CC), Ethernet Alarm Indication Signal (ETH-AIS), Ethernet Loopback (ETH-LB), Ethernet Link Trace (ETH-LT), Ethernet Remote Defect Indication (ETH-RDI), Ethernet Frame Loss Data Collection (ETH-LM), Ethernet Loopback State Request (ETH-LS) and ETH-TEST. Ethernet OAM frame format is shown in Fig. 9. ETH-CC/AIS/RDI/LM aim at failure detection and notification functions in MP-to-MP network, and ETH-LB/LT/LS/TEST aim to be used to recognize accurate failure location. This section describes ETH-CC and ETH-LT as a basic operation of Ethernet-based OAM.

4.1.1 Ethernet Continuity Check (ETH-CC)

This function is a receiver-driven failure detection method. "Maintenance Entity" Group End Point (MEP) [25] on a provider bridge periodically sends a CC frame (as a Unidirectional Heartbeat), which has a multicast DA, within a bridge segment. If a receiver side of MEP cannot receive the CC frame within a period of 3.5 times of the Heartbeat

- **OAM Ethernet Type**:
  - OAM frame Identifier
- **Version**:
  - OAM protocol version ID. Current: 0x00
- **ME Level**:
  - OAM Maintenance Domain ID.
  - 0x00-0x02: User domain
  - 0x03-0x04: Provider domain
  - 0x05-0x07: Operator domain
- **OpCode**:
  - Command ID. Unknown op-codes MUST be discarded. Vendor specific op-code is provided
- Transmission/Sequence ID:
  - Request/reply relation ID
- Transmission Timestamp:
  - Available by OpCode type
- Service ID
  - Upper service Identifier

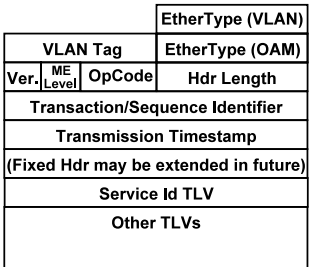


Fig. 9 Ethernet OAM frame format of ITU-T SG13 Q.5/13.

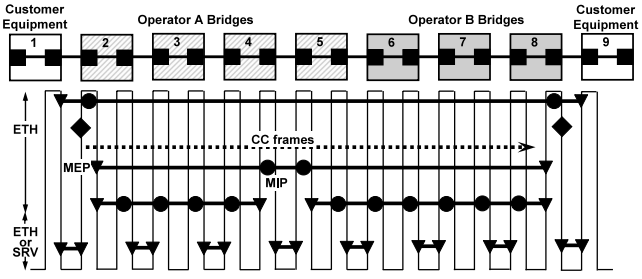


Fig. 10 Example of Ethernet Continuity Check operation.

cycle, it triggers an alarm to notify a failure, as shown in Fig. 10.

4.1.2 Ethernet Link Trace (ETH-LT)

This function is used for fault localization and for tracing a data path in Ethernet bridge network. Although it is similar to IP traceroute function, it rather uses a multicast request/response due to Ethernet bridge characteristics, in contrast to a unicast request/response in IP traceroute. A source bridge transmits the Request (RQ) message in multicast, which includes a transaction ID with an effective period of five seconds, toward the target bridge periodically or each time Loss-of-Connectivity (LOC) is detected. The intermediate bridges relay RQ message, in which TTL value is decremented by one, to the target bridge direction. It sends back the RP message to the source bridge after a specific period (0 to 1 sec). Finally, when the target bridge receives the RQ message, it sends back the RP message to the source bridge. Note that this behavior is different from IP traceroute. This operation is shown in Fig. 11.

4.2 Further Issues of Ethernet OAM

As described above, Ethernet-based OAM is a bit different from IP-based OAM, in which both connectivity-check and link trace are done by IP ping and IP traceroute command. IP control is based on IP address and IP routing, and Ethernet control, on the other hand, is based on MAC address and MAC bridging. Due to bridging functionality, there are

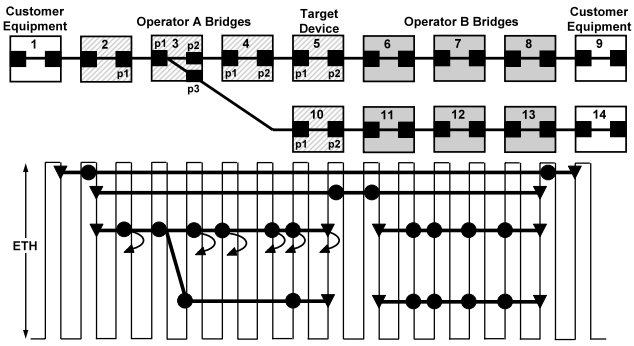


Fig. 11 Example of Ethernet Link Trace operation.

two cases of "before" and "after" MAC address learning at an intermediate bridge. After a MAC address is learned, Ethernet OAM can be the same as IP OAM basically. However, before a MAC address is learned, or just after MAC address learning table is flushed out due to a time-out, the link-trace cannot be performed, even though a network operator wants to analyze it. Thus, Ethernet OAM requires a periodical ETH-CC to force the intermediate switch to keep the MAC address entries, which is a big difference from IP-OAM. ETH-CC, however, leads to a network scalability problem due to heavy periodical traffic, which will be an issue of Ethernet OAM [26].

In terms of applicability statement of Ethernet OAM, any reliability control mechanisms described in Sect. 3 can be used simultaneously with Ethernet OAM, since Ethernet OAM is implemented in MAC frame format and can be processed in any Ethernet network device.

5. QoS Control Technologies

The Service Level Agreement (SLA) that most of Ethernet service providers currently use consists of bandwidth and delay contract. In terms of bandwidth contract, bandwidth menu for customers can be categorized into (i) guaranteed bandwidth in 1 Mbps unit, (ii) shared bandwidth for multiple subscribers, and (iii) ATM/SONET-service granular bandwidth via ATM/SONET-to-Ethernet bridging. In terms of delay contract, a provider defines a SLA, where a monthly average round-trip transfer delay of an IP packet is kept a value less than N milliseconds [27]. If the monitored results go beyond the contract value, the provider will have to pay back a part of service charge [28]. This section gives QoS control technologies to realize these communication services.

5.1 Bandwidth Guarantee Control Technologies

There are two kinds of bandwidth guarantee, (i) peak-rate guarantee (discarding traffic of exceeded bandwidth by a rate limiter) and (ii) committed-rate (minimum-rate) guarantee (drop-precedence based priority control). A typical scheme to support these kinds of bandwidth guarantee is Two Rate Three Color Marker (trTCM), as shown in Fig. 12

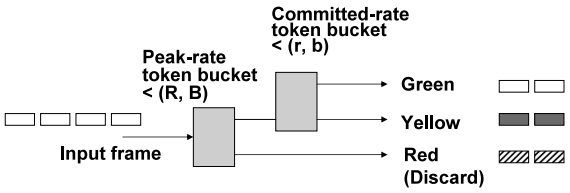


Fig. 12 trTCM mechanism.

[29]. In trTCM, traffic exceeding a peak-rate is discarded, and traffic between a committed-rate and the peak-rate is marked in yellow, and traffic below the committed rate is marked green. Yellow traffic may be discarded in a later processing unless a network capacity is sufficient.

When using a token bucket model for packet accounting, the token bucket parameter (B, b) must be defined based on the burst tolerance for incoming traffic. If a buffer is small, TCP application (such as FTP) that customers use may get a poor performance due to unnecessary packet losses causing rate-reductions. Thus, Ethernet service providers have to take into account of even higher layer traffic (i.e., TCP/UDP/RTP) characteristics to design their bridge network, although an Ethernet service itself means Layer 2 services independent of higher layer applications.

5.2 Layer 2 Packet Marking

Ethernet 802.3 standard frame format has a field to specify traffic classes by IEEE 802.1p [30], but does not have a field of drop precedence (DP), which is used to mark a packet in "yellow" on the same class of "green" traffic. To support variable bit rate (VBR) traffic with the committed-rate guarantee and peak-rate limitation, DP field is necessary in the Ethernet frame. Thus, a layer 2 (L2) packet marking scheme is now being discussed in 802.1ad.

There were initially two schemes proposed for DP in 802.1ad, and finally a hybrid scheme of both ones has been accepted in 802.1ad/Draft 4.0 Amendment4 of Feb. 2005, as described below. One of initial two proposals is using Canonical Format Indicator (CFI) field in an explicit manner. Although the drop precedence control can be simple, there are two problems. It is necessary to change a hardware of an existing bridge since the CFI (1 bit) was originally used to identify a Token Ring traffic and to keep a backward compatibility for that traffic. It cannot also be inter-connected to a MPLS service having only Priority bit field (3 bit), which does not include a DP field. Another one is using Code point (3 bits) field of 802.1p in an implicit manner. Although the existing hardware can be used, the drop precedence allocation and configuration control becomes complex. The number of traffic classes has to be reduced, since some of traffic classes have to be specified as drop precedenced classes.

Now, the current standardized method (as shown in Fig. 13) is a hybrid method, where both CFI field and priority field of Service VLAN Tag are utilized. The CFI field itself can be used as Drop Eligible Bit (DE), and at the same

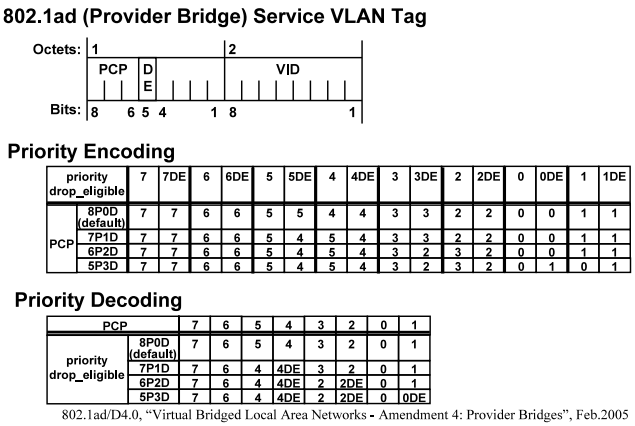


Fig. 13 Layer 2 packet marking encoding and decoding.

time, the DE information is also mapped to a range of Priority 0–7. By using this scheme, the bridge can recognize drop precedence information by looking at either only DE or only priority bit.

5.3 Fairness Control

When a customer uses a committed-rate guarantee service or a best effort service, a traffic exceeding the committed-rate or the best effort traffic are marked "yellow" in DP bit, and may go through a bottleneck link between areas or between providers. In this scenario, it is very effective to achieve fair-share rate control of DP-marked traffic before simply dropping, so that each customer can maximally utilize the bottleneck link in a fair manner. Unless having the fair-share traffic control, a single customer may consume almost all of the bottleneck link bandwidth by sending or receiving a heavy traffic, causing unfair bandwidth utilization among different customers.

5.4 Further Issues of QoS Control

As described above, the QoS control of priority control, bandwidth control, and fairness control has been used practically in current wide area Ethernet services, even though Ethernet cannot support finer quality than ATM.

The current problems of QoS control are that there are no standardized specification (i.e., detailed parameters for QoS control and its conformance for career grade Ethernet QoS). Thus, current implementation of Ethernet-based QoS control is vendors' proprietary implementations, and sometimes has interoperability problems of QoS parameter definitions and its performance between different vendors.

In addition, there are various access environments such as a leased line, an ATM leased line, and FTTH Ethernet access, and they sometimes need to be interconnected or bridged to wide area Ethernet network. Since the bandwidth control and the delay control are different in each access network, the seamless interconnection of QoS control still needs to work in terms of various control parameters.

In terms of applicability statement of QoS control, any

QoS control mechanisms can be used simultaneously with reliability controls described in Sect. 3 and Ethernet OAM described in Sect. 4.

6. Next Generation Wide Area Ethernet Switch Technology: Global Open Ethernet (GOE)

Section 2 describes the currently proposed standard network architecture for wide area Ethernet. Sections 3, 4, and 5 explain the currently available technologies for reliability, OAM and QoS control, which include standard technologies and vendor proprietary ones.

This section introduces next generation wide area Ethernet architecture, Global Open Ethernet (GOE), which we have already proposed in [31]. It also highlights especially GOE reliability control mechanisms that improve current mechanisms described in Sect. 3.

GOE employs a new network architecture compared with the one (i.e., 802.1ad and 802.1ah) described in Sect. 2. It inherently improves high-speed switching performance via GOE tag switching described below and also improves reliability functions in terms of loop prevention control, optimized path control, and node redundancy control in Sect. 3.

GOE has VLAN-tag related unique features of (i) VLAN-based tag switching (i.e., MAC-less switching) which provides a high speed switching as well as a fast protection switching, (ii) IP-router like shortest path routing, more specifically a shortest widest path routing (bandwidth-intensive path routing) via a routing tag, and (iii) hitless or non-stop network reconfiguration operation via active/standby routing tags. The feature (i) (ii) (iii) contribute to improvement of loop prevention control (in Sect. 3.2), optimized path control (in Sect. 3.4), and node redundancy control (in Sect. 3.5), respectively. The detailed points of each improvement are discussed below.

6.1 New Network Architecture via GOE

The proposed GOE provides new network architecture via a novel tag stacking scheme. It extends Q-in-Q tag stacking scheme in Sect. 2.1 (consisting of S-Tag and C-Tag) by adding scalability, reliability, and operationability functions into another tag field, while the S-Tag in Q-in-Q is only used to identify and separate customers. GOE uses two separate provider tags, GOE Forwarding Tag (FW-Tag) and GOE Customer Tag (GC-Tag), whose format are backward compatible with current Q-in-Q tags. FW-Tag and GC-Tag, both of which are called GOE tag in total, are used to control the routing and protection switching within a provider bridge network using GOE. Both tags are inserted at the provider bridge and are processed within the GOE network as shown in Fig. 14.

FW-Tag is an important tag for improvement of overall reliability functions, to achieve three unique features (i), (ii) and (iii) as described above. GOE provides a loop-free routing mechanism (for feature (i)) and a shortest widest path routing mechanism (for feature (ii)) by combination of using FW-Tag and a novel per-destination multiple rapid STP (PD-MRSTP) described in Sect. 6.3. GC-Tag, on the other hand, contributes to increase the number of customer ID space from 12 bits to 20 bits that are sufficiently large and to include TTL to prevent loops of Ethernet bridges (for feature (i)). Note that the role of GC-Tag is similar to S-Tag in Q-in-Q.

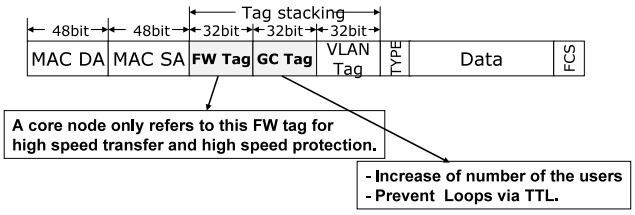


Fig. 14 Ethernet frame format with GOE tag.

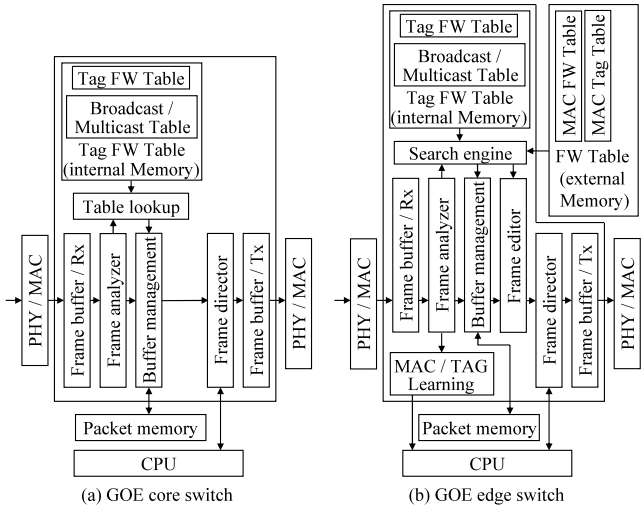


Fig. 15 GOE switch architecture.

6.2 GOE Switch Architecture (Core and Edge)

Each bridge has own bridge ID, whose instance ID (FW-Tag ID) is distributed as a destination bridge address via PD-MRSTP, which creates a shortest widest path route on the provider bridge network. The core bridge refers only to the FW Tag (including destination bridge ID) in the frame and can switch the frame toward the destination edge bridge. The edge bridge, on the other hand, has a function which performs MAC address learning and associates its destination MAC address with a FW-Tag (including a destination bridge ID). The actual implementation is explained below.

For feature (i), in terms of high speed switching and routing, a core bridge as shown in Fig. 15(a) analyzes GOE tag (frame type and tag) via a frame analyzer, and look-up at both a tag forward table and a broadcast/multicast table and decides an output port for the FW-Tag. The edge switch as shown in Fig. 15(b), on the other hand, there are a MAC forward table and a MAC/tag learning in addition to the ta-

ble of GOE core switch. The main difference of behaviors between GOE core switch and edge switch are (a) MAC-tag learning and mapping and (b) GOE-tag pushing and popping on a customer frame.

6.3 Optimized Routing of PD-MRSTP Protocol

For feature (ii), GOE utilizes existing MSTP to create multiple spanning trees for different VLANs, and also simultaneously utilizes existing RSTP for quick failure recovery. We call this per-destination multiple rapid spanning tree protocol (PD-MRSTP), which is completely interoperable with legacy Ethernet bridges.

With PD-MRSTP, a spanning tree is created from each destination bridge, to which a different VLAN-ID (or node ID) is allocated, to other bridges. This tree looks like a reverse spanning tree to each destination, where a root bridge of this tree is always a destination bridge. PD-MRSTP distributes and creates a destination bridge ID (tag ID) based routing table on all of the provider bridges. This routing table is referred during the MAC frame forwarding using FW-Tag ID. The generated spanning tree becomes IP-router like shortest path route, more specifically a shortest widest path route (bandwidth-intensive path route) on the network, as shown in Fig. 16. Thus, PD-MRSTP can generate multiple optimized spanning trees, whose total number is the same as the number of GOE edge bridge, among any pair of bridges, as described in Sect. 3.4.

Since using FW-Tag with PD-MRSTP creates a unidirectional routing path on the spanning tree toward destination bridge ID, it can also provide a loop free routing (for feature (i)) in any network topology (including ring topology) at the same time. Thus, it solves the loop prevention control in Sect. 3.3 more effectively and completely.

It also accelerates rapid spanning tree protocol itself to get a highest failure recovery time (2 msec to 50 msec [31]), by implementing a protection behavior on a hardware. In case of network failure recovery, Q-in-Q and MAC-in-MAC technologies require three procedures to recover network failures: (a) topology change (via RSTP etc.), (b) MAC-address flushing, and (c) MAC-address re-learning. Since a provider bridge network has to maintain (or learn) a huge number of MAC addresses on the bridges, time required for procedures (b) and (c) increases in linear to the number of MAC address entries. Thus, the time for (b) and (c) becomes a rather large value on the bridge network [24]. GOE, on the other hand, does not process (b) and (c) but process only (a), which implies that the protection switching time can always be guaranteed to a very small value (2 msec to 50 msec) in dependent of number of MAC addresses in the network. Thus, GOE can provide very short protection switching (for feature (iii)) as an alternative improvement of node redundancy control in Sect. 3.5, compared with current available Ethernet technologies.

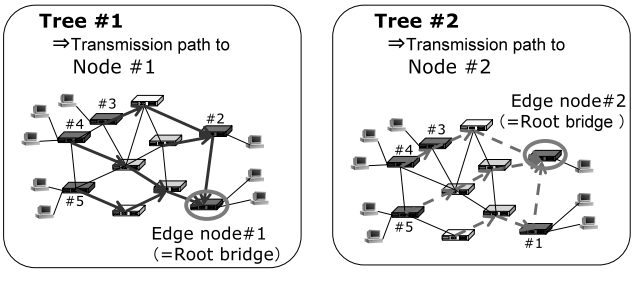


Fig. 16 Optimized multiple routing trees by PD-MRSTP.

6.4 Non-stop Network Operation of ISR Protocol

For feature (iii), as explained in Sect. 3.7, GOE also provides In-Service Reconfiguration (ISR) [1], [24] function where a network operation can be continued without any interruptions (i.e., zero-loss) even during a topology change via adding or removing a bridge in the network. ISR maintains dual spanning trees, active and standby trees, which are assigned different Tag IDs to PD-MRSTP. The active tree is continued to use for a current operation. The standby tree, on the other hand, can be re-configured for a new topology without any impacts to the active tree. Once the standby spanning tree becomes stable state, a network administrator can switch the active tree to the standby one without any packet loss. The ISR function can provide a perfect node redundancy control in Sect. 3.5 without any service interruptions.

7. Conclusion

This paper describes an overview of overall carrier-grade Ethernet technologies, especially from the viewpoint of reliability, OAM, and QoS control, and introduces GOE as one of promising approaches of a next generation carrier-grade Ethernet. Although there are still technical issues to solve to make it carrier-grade, most of Ethernet providers have already provided real services even in the current technologies. It implies that wide area Ethernet technology is not perfect yet, but can be used in a practical way. Since the IEEE and ITU-T are making efforts in a close relationship to strengthen Ethernet technologies to carrier-grade one, the technology will be expected to mature in a near future.

Acknowledgments

The authors would like to thank everyone at System Platform Research Laboratories, NEC Corporation for their invaluable comments, as well as the many others who gave generous advice regarding this manuscript.

References

[1] A. Arutaki, A. Iwata, K. Seto, H. Sono, K. Hyoudou, S. Mori, and M. Ando, Wide Area Ethernet Technology Overview, IEICE book, ISBN: 4-88552-211-0, 2005.
[2] IEEE802.1ad, http://www.ieee802.org/1/files/private/ad-drafts/d5/802-1ad-d5-0.pdf, April 2005.

[3] IEEE Computer Society, IEEE Std 802.1Q, 2003 Edition, IEEE Standards for Local and Metropolitan Area Networks Virtual Bridged Local Area Networks, http://standards.ieee.org/getieee802/download/802.1Q-2003.pdf, May 2003.
[4] IEEE802.1ah, http://www.ieee802.org/1/files/private/ah-drafts/d0/802-1ah-d0-2.pdf, March 2005.
[5] IEEE802.1ag, http://www.ieee802.org/1/files/private/ag-drafts/d3/802-1ag-d3-0.pdf, April 2005.
[6] M. Ando, "Nowadays VLAN technology trend," http://www.janog.gr.jp/meeting/janog10/pdf/janog10-l2-ando.pdf, July 2002.
[7] Cisco Systems, Understanding and Configuring the Unidirectional Link Detection Protocol Feature, Document ID: 10591, http://www.cisco.com/warp/public/473/77.pdf, April 2005.
[8] IEEE Computer Society, IEEE Std 802.1D-2004, IEEE Standard for Local and Metropolitan Area Networks Media Access Control (MAC) Bridges, http://standards.ieee.org/getieee802/download/802.1D-2004.pdf
[9] IEEE802.1w (incorporated into IEEE Std 802.1D-2004)
[10] IEEE802.1s (incorporated into IEEE Std 802.1Q, 2003 Edition)
[11] Cisco Systems, Spanning-Tree Protocol Root Guard Enhancement, Document ID: 10588, http://www.cisco.com/warp/public/473/74.pdf
[12] Cisco Systems, Spanning-Tree Protocol Enhancements using Loop Guard and BPDU Skew Detection Features, Document ID: 10596, http://www.cisco.com/warp/public/473/84.pdf, May 2005.
[13] Foundry Networks, Superspan, A Break-Through for Layer 2 Ethernet Networks, http://www.foundrynet.com/solutions/appNotes/PDFs/SuperSpan.PDF, Nov. 2001.
[14] A. Iwata, Y. Hidaka, M. Umayabashi, N. Enomoto, A. Arutaki, K. Takagi, D. Cavandish, and R. Izumailov, "Global open Ethernet architecture for a cost-effective scalable VPN solution," IEICE Trans. Commun., vol.E87-B, no.1, pp.142–151, Jan. 2004.
[15] N. Finn, MSTP Reflection Vector, http://www.ieee802.org/1/files/public/docs2005/new-nfinn-mstp-vector-0305.ppt, March 2005.
[16] Extreme Networks, Extreme Standby Router Protocol and Virtual Routing Redundancy Protocol, http://www.extremenetworks.com/libraries/whitepapers/technology/VRRPvsESRP_WP.pdf, Sept. 2002.
[17] Nortel Networks, What is Split Multi-Link Trunking?, http://www.nortel.com/products/01/passport/8600_rss/collateral/nn108460-060304.pdf, 2004.
[18] 3Com, Introduction to XRN: A New Direction for Enterprise Networking, http://www.3com.com/other/pdfs/legacy/en_US/xrn_intro_whitepaper.pdf, 2002.
[19] Cisco Systems, Cisco StackWise Technology, http://www.cisco.com/warp/public/cc/pd/si/casi/ps5023/prodlit/stkws_wp.pdf, Aug. 2003.
[20] IEEE Computer Society, IEEE Std 802.17—2004, IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements, Part 17: Resilient packet ring (RPR) access method and physical layer specifications, http://standards.ieee.org/reading/ieee/std/lanman/restricted/802.17-2004.pdf, Sept. 2004.
[21] M. Sakauchi, D. Ogasahara, K. Takagi, K. Fukuchi, and A. Iwata, "Interlink redundancy scheme for multiple RPR networks," Proc. IEICE Gen. Conf. '05, B-7-97, p.251, March 2005.
[22] S. Shah and M. Yip, Extreme Networks, Ethernet Automatic Protection Switching (EAPS) Version 1, http://www.ietf.org/rfc/rfc3619.txt, Oct. 2003.
[23] Foundry Networks, Metro Ring Protocol (MRP), http://www.foundrynet.com/services/documentation/sribcg/Metro.html#60958, 2005.
[24] M. Umayabashi, A. Iwata, Y. Hidaka, N. Enomoto, D. Ogasahara, K. Takagi, and A. Arutaki, "Smooth migration toward global open Ethernet networking," National Fiber Optics Engineers Conference 2005, NWG3, Anaheim, CA, March 2005.
[25] ITU-T, Draft Recommendation Y.17ethoam—OAM Functions and Mechanisms for Ethernet Based Networks, http://www.ieee802.org/1//files/public/docs2005/ag-liaison-y-17-ethoam-0105.pdf, Sept. 2004.
[26] A. Iwata, "Switching and traffic management trend for Metro Area Ethernet," Proc. IEICE Gen. Conf. '05, BT-3-3, pp.SS-21–22, March 2005.
[27] PoweredCom, Powered Ethernet: Introduction and Features, http://www.poweredcom.net/service/ethernet/index.html
[28] PoweredCom, Powered Ethernet: Service Level Agreement (SLA), http://www.poweredcom.net/service/ethernet/sla.html
[29] J. Heinanen and R. Guerin, A Two Rate Three Color Marker, http://www.ietf.org/rfc/rfc2698.txt, Sept. 1999.
[30] IEEE802.1p (incorporated into IEEE802.1D).
[31] A. Iwata, Y. Hidaka, M. Umayabashi, N. Enomoto, and A. Arutaki, "Global Open Ethernet (GOE) system and its performance evaluation," IEEE J. Sel. Areas Commun., vol.22, no.8, pp.1432–1442, Oct. 2004.
[32] IEEE 802.3-2002, IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, http://standards.ieee.org/getieee802/download/802.3-2002.pdf, March 2002.
[33] IEEE Computer Society, IEEE 802.3ae-2002, IEEE Standard for Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications-Media Access Control (MAC) Parameters, Physical Layer and Management Parameters for 10 Gb/s Operation, http://standards.ieee.org/getieee802/download/802.3ae-2002.pdf, Aug. 2002.

Atsushi Iwata was born at Fukuoka, Japan, in 1964. He received the B.E. and M.E. degrees in electrical engineering from the University of Tokyo, Japan, in 1988 and 1990, respectively. He joined NEC Corporation in 1990. From 1997 to 1998, he was a Visiting Researcher at the University of California, Los Angeles. He received Ph.D. degree in electrical engineering from the University of Tokyo, Japan, in 2001. He is currently the Senior Manager of System Platforms Research Laboratories, NEC Corporation, and focusing on high-speed broadband and computer networking systems. He received the Best Paper Award from the IEICE Switching Systems Technical Group in 1999, and IEICE Network Systems Technical Group in 2004. He is an author of a book "Wide Area Ethernet Technology Overview" published by IEICE in June 2005.