Papers on the Next Step of UNIVERGE : Toward the Future

High Reliable Ethernet Technologies: Global Open Ethernet

By Masaki UMAYABASHI,* Nobuyuki ENOMOTO,* Youichi HIDAKA,* Daisaku OGASAHARA,* Kazuo TAKAGI* and Atsushi IWATA*

ABSTRACT The authors propose a high reliable Ethernet technology: GOE (Global Open Ethernet) technology. The GOE technology is a cost-effective and scalable solution for next generation Ethernet VPNs (Virtual Private Networks) that can satisfy simultaneously not only the requirement of high reliability, but also those of flexible network topology, low equipment cost, and low operational cost. In this paper, we focus on the requirement of high reliability for Ethernets. We describe the problems affecting the reliability of current Ethernet technologies, and propose Per Destination - Multiple Rapid Spanning Tree Protocol and In Service Reconfiguration technology to solve these problems. This GOE technology can provide highly reliable and stable networks.

KEYWORDS **Ethernet, VPN (Virtual Private Network), RSTP (Rapid Spanning Tree Protocol), VLAN tag, Failure recovery**

1. INTRODUCTION

Ethernet technology is widely spread in LANs due to its cost-effectiveness and plug-and-play capability. Recently, many carriers and service providers are eager to provide Ethernet-VPN services for MANs (Metro Area Networks). Solving scalability, reliability, controllability, and manageability problems is necessary to satisfy customer demands.

Ethernet over MPLS (EoMPLS[1]), Resilient Packet Ring (RPR[2]), and Extended VLAN (Q-in-Q[3]) are the three main approaches for building a cost-effective Ethernet-based VPN solutions in MANs and solving these problems. Advantages and drawbacks of each approach are as follows.

1) The EoMPLS approach has advantages over all functions that MPLS provides, such as fast failure recovery (almost equivalent to that of SONET/SDH), MPLS-based VPN management, and traffic engineering. However, because LSPs (Label Switch Paths) must be established between any two EoMPLS bridges in a full mesh network, this solution is complicated and not scalable. In addition, an expensive router platform is needed to provide Ethernet VPNs.

2) The RPR approach provides fast protection in a ring topology network, but brings topology constraints that make the network design inflexible.

3) The Q-in-Q approach is a legacy-Ethernet-friendly technology. Q-in-Q improves VLAN ID scalability by adding a VLAN-ID space, but does not solve the reliability and manageability problems.

Thus, although current Ethernet VPN approaches solve some or parts of the problems with the legacy Ethernet, none solves all of the problems.

We have previously proposed the GOE (Global Open Ethernet) architecture as an affordable VPN solution that addresses all these problems[4,5]. The GOE is based on Ethernet technology, but it provides similar functionalities to those of MPLS. In this paper, we focus on improving Ethernet reliability. We propose the PD-MRSTP (Per Destination - Multiple Rapid Spanning Tree Protocol) and ISR (In-Service Reconfiguration) technologies, which are parts of GOE technology, as reliable Ethernet technology. As GOE can provide fast protection almost equivalent to that of SONET/SDH, GOE-based VPN is an alternative to the STM-based leased line services. Therefore, GOE technology would be a driving force for evolving Ethernets into a much wider

*System Platforms Research Laboratories

spreading of Ethernet-based VPNs.

This paper is organized as follows. Section 2 describes reliability problems of current Ethernet technologies. Section 3 addresses these problems by introducing GOE architecture and its components, PD-MRSTP and ISR. Section 4 summarizes the current study.

2. PROBLEMS OF ETHERNET RELIABILITY

Many Ethernet technologies aimed to improving the Ethernet reliability have been proposed and are now in use.

In the standardized technologies specified by the IEEE 802 committee, Link aggregation (LAG)[6] is specified in IEEE 802.3ad as a link redundancy technology, and Spanning Tree Protocol (STP)[7], Rapid STP (RSTP)[8], and Multiple STP (MSTP)[9] are specified in IEEE 802.1D, 1w, and 1s respectively, as link or node failure recovery technologies.

Some vendors propose their proprietary technologies. Extreme Network proposes ESRP (Extreme Standby Routing Protocol)[10] and Foundry Network proposes VSRP (Virtual Switch Redundancy Protocol)[11] as their node redundancy technologies for the tree topology. For the ring topology, EAPS (Extreme Automatic Protection Switching)[12] of Extreme and MRP (Metro Ring Protocol) [13] of Foundry are in use.

These current technologies can actually provide protection in some cases in several tens of milliseconds, but in general, they have the following problems.

(1) Increased Protection Time Required to Flush the Information of a Filtering Database (FDB)

When a failure occurs, the network is reconfigured to recover based on the current Ethernet reliable technologies. In an Ethernet including such technologies, each bridge transmits frames to a learned port recorded on the FDB. If a bridge transmits frames based on the FDB created before a failure has occurred, the frames cannot reach the destination bridge, because the FDB is not updated after the failure. Therefore, in all current technologies, when a failure occurs, the information of the FDB should be flushed. In order to flush the FDB, the MAC addresses must be deleted from FDB entries, and then re-learned and restored as FDB entries. In WANs, such as carrier networks and large enterprise networks, where the bridges have many stored MAC address entries, an FDB flush takes a long time so that protection also takes a long time. According to our performance evaluation, the FDB flush requires much more time compared with that spent for reconfiguration. To minimize the time spent for the FDB flush requires reducing the number of MAC entries or to use another technology that eliminates the FDB flush.

(2) Increased Recovery Time for Root Bridge Failure

The RSTP/MSTP can recover failures, except for root bridge failure, within a second in any topology. This is because of the rapid state transition achieved by using handshake procedure and a pre-calculated alternate port as a secondary root port. When a root bridge goes down, in RSTP/MSTP, a new root bridge has to be elected from among all of the bridges and the spanning tree must be stabilized under the new root bridge. This process takes several seconds in the best case. Though RSTP/MSTP does provide fast protection in cases of link/node failures, it does not provide very fast recovery from root bridge failure.

(3) Broadcast Storm in Loop Condition

The STPs and node redundancy technologies basically create loop-free logical networks from physical networks that include loops. However, loops may occur even when using such technologies if the control frames are discarded in the following cases. If the CPU processing load increases too much, the bridge cannot send control frames in adequate time. Then, the opposite bridge cannot receive the control frames, and that means discarding the control frames. In other cases, the degradation of fiber or optical modules causes loss of the control frames. If the control frames are lost for a specified continuous time, the receiver side bridge may open a port that should remain closed to prevent loops. As a result, a loop is created. Once the loop is created, a broadcast storm occurs in which each switch in the loop continues to forward broadcast traffic repeatedly, and the networks go down.

Some vendors propose their proprietary technology to solve loop problems. For example, EoE (Ethernet over Ethernet) technology includes TTL (Time to Live) in frames. The TTL prevents the looped traffic from being forwarded permanently. Cisco proposes their Loop Guard, a function to avoid loops. However, a configuration is very complicated to use the Loop Guard and the other functions simultaneously. Thus, these technologies are not sufficient enough to prevent broadcast storms caused by loops.

(4) Packet Loss When Network Configuration Changes

In the RSTP/MSTP, once a new bridge is added or an existing bridge is removed, the spanning tree algorithm must work again. Until the new spanning tree is created and becomes stable, existing packets may be discarded in locations where the spanning tree direction has to be changed. If there are N locations with a change in tree direction, packets are discarded N times in an interval of several seconds (This interval could be several tens of seconds in total.) To minimize the affect of packet losses, network operators upgrade network configurations at midnight or on weekends when the traffic volume is small. This increases network operation costs.

We propose GOE technologies as the next generation Ethernet VPN technology. The PD-MRSTP and ISR technologies that are parts of GOE technology can solve the problems described here, and provide reliable Ethernet networks. The details of each of these technologies are explained in the next section.

3. GOE SOLUTION

The authors propose PD-MRSTP and ISR technologies that can solve the problems of legacy Ethernets. In this section, we discuss the PD-MRSTP, which can solve problems described in (1), (2) and (3) in Section 2, and the ISR which can solve those described in (4).

3.1 PD-MRSTP Technology

A GOE network consists of GOE edge and core bridges. Each bridge uses the MSTP which can create a spanning tree of RSTP for VLAN independently. Each edge bridge creates a spanning tree, with itself as its root, by using MSTP to establish forwarding routes to it from any other GOE edge bridges to itself.

Figure 1 shows an example of a spanning tree root bridge of which is edge A. In the GOE scheme, when the edge bridges send Ethernet frames destined to edge bridge A, they transport the frames along this spanning tree. To transport the frames along the spanning tree, the ingress bridge and the core bridges send the frames destined for edge bridge A to the root port of the spanning tree.

Since the spanning tree always provides the most cost-effective (the shortest) route between the root bridge and any other bridges, the frames are transported along a cost-effective route using this scheme. Since a spanning tree appears to be created for each destination bridge, we call this spanning tree configuration technology Per Destination - Multiple Rapid Spanning Tree Protocol (PD-MRSTP), which is completely compatible with the legacy Ethernet standards.

Figure 1 also shows the frame forwarding mechanism when using PD-MRSTP. The ingress bridge D receives customer Ethernet frames with or without an IEEE 802.1Q VLAN tag and inserts a GOE header after an Ethernet source MAC address field. The GOE header has a flexible and extensible length structure based on conventional VLAN tag stacking technology. The ingress edge D resolves the identification of the destination edge bridge to which the received Ethernet frame is transported by referring to a destination MAC address and receiving port, and then stores the destination edge bridge ID in the forwarding tag field in the GOE header. After that, the edge bridge D sends the frame to the root port of the spanning tree with the destination edge A as the root bridge. The core bridges forward it along the spanning tree after referring to the forwarding tag field only. The egress edge bridge A removes the GOE header, and then sends the Ethernet frame to the destination terminal.

When a failure occurs in the network, the network is reconfigured based on the RSTP procedure. After creating the new spanning tree, the root port of the new spanning tree is set as the new output port, and then the frames are transported again along a new route.

As described above, in PD-MRSTP, each core bridge refers only to the bridge ID of the egress bridge stored in the forwarding tag, and forwards the frames to the root port of the spanning tree. Therefore, in PD-MRSTP, since each bridge has no need to refer to the

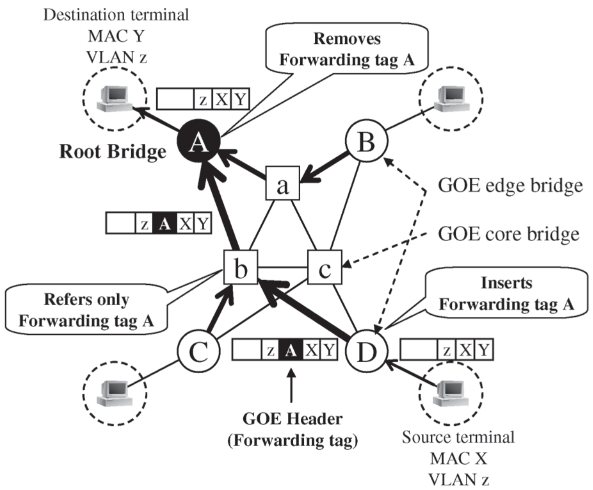

Fig. 1 Frame forwarding mechanism in PD-MRSTP.

MAC addresses, the PD-MRSTP can recover from failure without the FDB flush and MAC address learning, whereas the current Ethernet technologies need these. Since the time spent on FDB flush increases in proportion to the number of MAC address entries, the failure recovery time increases severely in large scale networks, accommodating a large number of MAC address entries, such as carrier network or large enterprise network. However, PD-MRSTP can provide rapid failure recovery in such large scale networks.

PD-MRSTP can also solve the problem of root bridge failures. In PD-MRSTP, root bridge failure is a destination bridge failure, and there is no available access point, in single-homing access. Thus, selecting a new root bridge and reconfiguring this spanning tree is unnecessary. On the other hand, in dual-homing access, the frames destined to a failed destination bridge should be recovered by another destination bridge. Moving from an old destination bridge to a new dual-homed destination bridge takes less than a second, which is a significantly smaller than that of RSTP. Therefore, PD-MRSTP provides rapid failure recovery in the case of a root bridge failure.

In PD-MRSTP, each bridge forwards the frames to their root port. This is unidirectional transport toward the root port direction. As described in Section 2, when the BPDUs (Bridge Protocol Data Units), which are control frames of the STPs, are lost continuously, the bridge opens a port which should to be kept closed to eliminate loops. As a result, a loop is created. In such a loop condition, the opened port is generally a designated port, but not a root port. Thus, if such a logical loop topology is created, the PD-MRSTP can avoid transporting Ethernet frames into the loop due to its unidirectional transport characteristics. In such a case, the root port direction of each bridge may form a loop. In that case, since a GOE header includes TTL, the TTL mechanism is running to eliminate the loop traffic. Therefore, PD-MRSTP can provide a mechanism that avoids broadcast storm due to loops, and provides stable networks.

As explained above, PD-MRSTP can drastically decrease the failure recovery time by eliminating the FDB flush and topology reconfiguration time under root bridge failure. In addition, PD-MRSTP can prevent broadcast storms even in a logical loop topology. Therefore, PD-MRSTP can provide highly reliable and stable networks.

3.2 ISR Technology

GOE can provide network reconfiguration with zero-packet discarding with ISR technology. **Figure 2** shows an example of ISR applied to a network model. In ISR, two bridge IDs (default ID and alternate ID) are assigned to each edge bridge. The default ID is used by spanning trees to transmit data traffic before a new bridge is added, and the other is used by spanning trees for reconfiguring the topology.

Once a new core bridge is added, a new spanning tree based on the alternate ID is reconfigured. The existing spanning tree based on the default ID continues to work while the new spanning tree is being

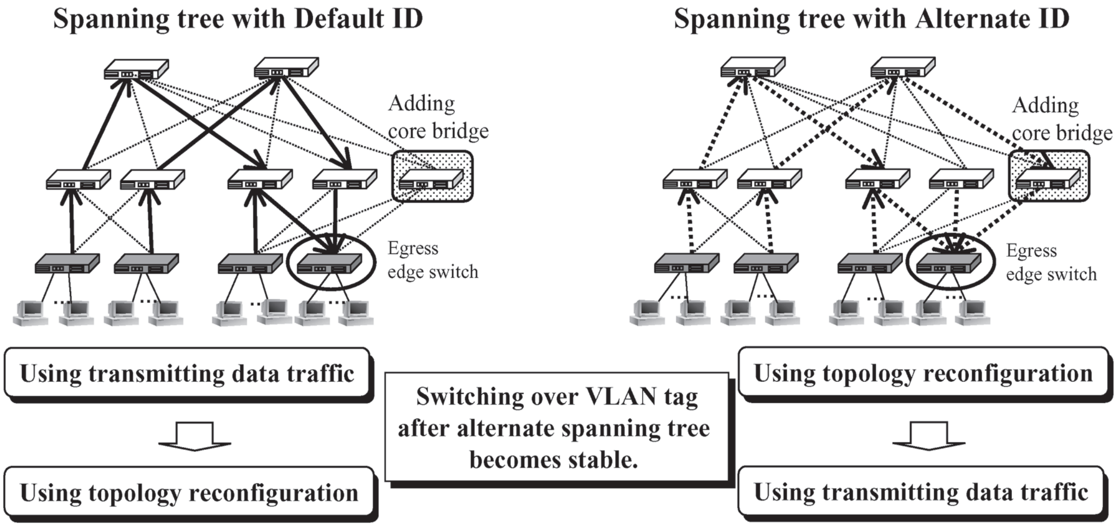
Spanning tree with Default ID
Adding core bridge
Egress edge switch
Using transmitting data traffic
Using topology reconfiguration
Switching over VLAN tag after alternate spanning tree becomes stable.
Spanning tree with Alternate ID
Adding core bridge
Egress edge switch
Using topology reconfiguration
Using transmitting data traffic

Fig. 2 Switching over the spanning tree in ISR.

reconfigured. After the new spanning tree becomes stable, the root bridge sends a trigger message to each bridge to switch over from the existing spanning tree with the default ID to the new spanning tree with the alternate ID. Since the network upgrade (node add/delete) is operated using the new spanning tree with the alternate ID, the existing traffic is always being forwarded on an existing stable spanning tree.

In terms of the ingress bridge side, switching over the spanning tree can be accomplished by changing the VLAN tag ID from the default to the alternate ID, because the spanning tree using frame transmission is identified by a VLAN tag ID inserted at the ingress bridge. On the egress bridge side, the egress bridge can always receive both the frames with the inserted default ID and those with the alternate ID. Therefore, ISR can be used to upgrade network configuration without any packet losses.

Moreover, zero-packet-loss can be achieved with simple procedures without synchronizing the timing of switching over VLAN tag IDs at each ingress bridge because bridges can receive frames with either default or alternate IDs inserted. Although there may be packet reordering when the delay through the existing spanning tree is longer than that through the new spanning tree, packet reordering problems can be solved by enforced buffering at the ingress bridge until the last packet in transit reaches the egress bridge.

We evaluated ISR performance in our prototype system, and proved that the network upgrade (node add/remove) is performed with zero packet loss. This result was achieved under the severe conditions that the transmission rate was 1Gbps full rate and the packet size was 64 Bytes. Thus, ISR technology can provide zero-packet-loss network upgrading under severe traffic conditions, and can solve the problem (4) described in Section 2. Using the ISR technology, network operators can upgrade network configuration or maintain switches at any time, thus network operation costs can be reduced.

4. CONCLUSION

In this paper, the authors have proposed PD-MRSTP and ISR technology, which are components of GOE technology that can solve problems of Ethernet reliability.

The PD-MRSTP method, or extended standardized RSTP/MSTP, can decrease the failure recovery time by eliminating the need to flush the FDB and recover from root bridge failures. PD-MRSTP can also solve broadcast problem due to the network loops. The ISR method provides a network upgrade capability without any packet loss. Therefore, this GOE technology can provide highly reliable and stable networks.

REFERENCES

[1] L. Martini, et al., “Transport of Layer 2 Frames Over MPLS,” IETF Internet Draft, draft-martini-l2circuit-trans-mpls-14.txt, Jun. 2004.

[2] IEEE 802.17 Resilient Packet Ring Working Group, http://www.ieee802.org/17.

[3] IEEE 802.1ad Provider Bridges Working Group, http://www.ieee802.org/1/pages/802.1ad.html.

[4] A. Iwata, et al., “Global Open Ethernet Architecture for a Cost-Effective Scalable VPN Solution,” *IEICE Trans. on Communications*, **E87-B**, 1, pp.142-151, Jan. 2004.

[5] A. Iwata, et al., “PERFORMANCE EVALUATION OF GLOBAL OPEN ETHERNET (GOE) PROTOTYPE SYSTEM,” National Fiber Optics Engineers Conference 2003, pp.997-1006, Dallas, Sept. 2003.

[6] ANSI/IEEE Standard 802.3ad, “Link aggregation protocol,” “Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications.”

[7] ANSI/IEEE Standard 802.1D, 1998 Edition, “Media access control (MAC) Bridges.”

[8] ANSI/IEEE Draft Standard 802.1w-2001, “Part 3: Media Access Control (MAC) Bridges, Amendment 2- Rapid Reconfiguration,” http://standards.ieee.org/getieee802/802.1.html.

[9] ANSI/IEEE Draft Standard 802.1s/D11.2, “Amendment 3 to 802.1Q Virtual Bridged Local Area Networks: Multiple Spanning Trees,” http://www.ieee802.org/1/pages/802.1s.html.

[10] “ExtremeWare7.2 User Guide,” http://www.extremenetworks.com/services/documentation/.

[11] “Foundry Switch and Router Installation and Basic Configuration Guide,” http://www.foundrynet.com/services/documentation/sribcg/Metro.html.

[12] S. Shah, et al., “Extreme Networks’ Ethernet Automatic Protection Switching (EAPS) Version 1,” http://www.ietf.org/rfc/rfc3619.txt.

Received September 10, 2004

* * * * * * * * * * * * * * *

Masaki UMAYABASHI received his B.E. and M.E. degrees in electrical engineering from Keio University, Japan, in 1995 and 1997, respectively. In 1997, he joined NEC Corporation, Kanagawa, Japan, and is a Research Staff member at System Platforms Research Laboratories. His current research interest are the design and analysis of traffic control and protocols for communication networks.

Mr. Umayabashi is a member of the IEICE of Japan.

Nobuyuki ENOMOTO received his B.E. and M.E. degrees in School of Science and Engineering, Waseda University, Tokyo, Japan, in 1999 and 2001, respectively. He joined NEC Corporation in 2001, and is a Research Staff member at System Platforms Research Laboratories, NEC Corporation, Kanagawa, Japan. His current research interest is the design and analysis of network architectures, routing algorithms and protocols for computer communication networks.

Mr. Enomoto is a member of the IEICE of Japan.

Youichi HIDAKA received his B.E. and M.E. degrees in information systems science from Soka University in 1995 and 1997, respectively. He joined NEC Corporation, as a member of the hardware engineering department and he was engaged in the development and evaluation of LAN network devices and system from 1997 to 2001. More recently, he joined System Platforms Research Laboratories, NEC Corporation, Kanagawa, Japan, in 2002. His current research interest is the design and analysis of network architectures, hardware architectures, protocols for computer communication networks.

Mr. Hidaka is a member of the IEICE of Japan.

Daisaku OGASAHARA received his B.E. and M.E. degrees in electrical engineering from Kyoto University, Japan, in 1998 and 2000, respectively. He is a research staff member at System Platforms Research Laboratories, NEC Corporation, Kanagawa, Japan. His current research interest is the design of next generation network architectures.

Kazuo TAKAGI received his B.E. and M.S. degrees in electrical engineering from Keio University, in 1989 and 1991, respectively. He joined NEC Corporation in 1991, and is a research staff member at System Platforms Research Laboratories, NEC Corporation, Kanagawa, Japan. Since joining NEC, he has researched and developed optical ATM switches, all optical access systems, ATM access systems, and WDM ring systems in their Networking Research Laboratory. He worked in the Network Product Research Department in C&C Research Labs., NEC Laboratories America, from 2002 to 2003. His current interest is the design of next generation Ethernet architectures.

Atsushi IWATA received his B.E., M.E., and Ph.D. degrees in electrical engineering from the University of Tokyo, Japan, in 1988, 1990, and 2001, respectively. In 1990, he joined NEC Corporation, Kanagawa, Japan, as a member of the Research Staff and he engaged in research and development of ATM-LAN systems. From 1997 to 1998, he was a Visiting Researcher at the University of California, Los Angeles. He is currently the Senior Manager of the System Platforms Research Laboratories, focusing on high-speed broadband and computer networking systems.

* * * * * * * * * * * * * * * *