Lane Adaptive Recovery for Multiple Lane Faults in Optical Ethernet Link

Kyeong-Eun Han, Sun-Me Kim, and Jonghyun Lee

We propose a lane adaptive recovery scheme for multiple lane faults in a multi-lane-based Ethernet link. In our scheme, when lane faults occur in a link, they are processed not as full link faults but as partial link faults. Our scheme provides a higher link utilization and lower packet loss rate by reusing the available lanes of the link and providing a low recovery time of under a microsecond.

Keywords: Multiple lane faults, multi-lane-based Ethernet link, lane recovery.

I. Introduction

The IEEE 802.3ba task force recently standardized both multi-lane-based 40 G/100 G Ethernet to provide much higher bandwidth while reducing costs [1]–[2]. Here, a link consists of four or ten lanes for 40 G/100 G Ethernet, and data are simultaneously transmitted over all lanes at a rate of 10 Gbps or 25 Gbps per lane. In terms of link failure, 802.3ba only refers to link fault signaling between the local reconciliation sublayer (RS) and the remote RS regardless of the partial lane fault in the link. However, when some of the lanes fail, a partial lane recovery may be more powerful in terms of reducing the packet loss than a link failure recovery in a multi-lane-based transmission architecture [3]–[5]. Although it provides a lower data rate, the use of remaining non-fault lanes may lead to a data transmission during the backup switching time, which may alleviate the performance degradation of the network by providing a low packet loss and a reuse of resources. Currently, there has been little research on this topic, which is the initial stage of this study. In addition, many new functions and methods should be considered for a lane fault recovery.

Many previous studies have been conducted regarding fault management not for multi-lanes but for a link in an optical network [6]–[9]. A few lane fault protection methods for the detection and recovery at the physical medium dependent (PMD) sublayer were proposed in [3]–[5], but their discussion only focused on a single lane fault. However, when we consider the probability of lane faults increasing as the number of composed lanes increase in a multi-lane-based link, a novel lane fault recovery is required to drive multiple lane faults.

In this paper, we propose a lane adaptive recovery scheme that covers not only a single lane fault but also multiple lane faults. For this, the reconciliation sublayer (RS) exchanges the information of faulty lanes with lane fault messages, and replaces component lanes of a link with available lanes. After recovery, the RS retransmits data through employing recovered lanes during backup switching time.

II. Overview

Link fault signaling operates between the remote RS and local RS in 40 G/100 G Ethernet [2]. When a fault occurs between a local RS and remote RS, only an RS originates remote fault signals. While most fault detection is performed on the receive data path of the physical layer (PHY), a fault can be detected on the transmit side of the PHY, which is also indicated by the PHY with a local fault status. The detected fault is delivered to an RS as a local fault. When the RS receives the local fault signal, the RS stops sending media access control (MAC) data and continuously generates a remote fault status on the transmit data path. Upon receiving a remote fault status, the remote RS stops sending MAC data

Manuscript received Nov. 15, 2013; revised Oct. 10, 2014; accepted Oct. 20, 2014.

This work was supported by the IT R&D Program of MKE/KEIT, [10041414, Terabit Optical-Circuit-Packet Converged Switching System Technology Development for Next-Generation Optical Transport Network].

Kyeong-Eun Han (corresponding author, kehan@etri.re.kr), Sun-Me Kim (kimsunme@etri.re.kr), and Jonghyun Lee (jlee@etri.re.kr) are with the Communications & Internet Research Laboratory, ETRI, Daejeon, Rep. of Korea.

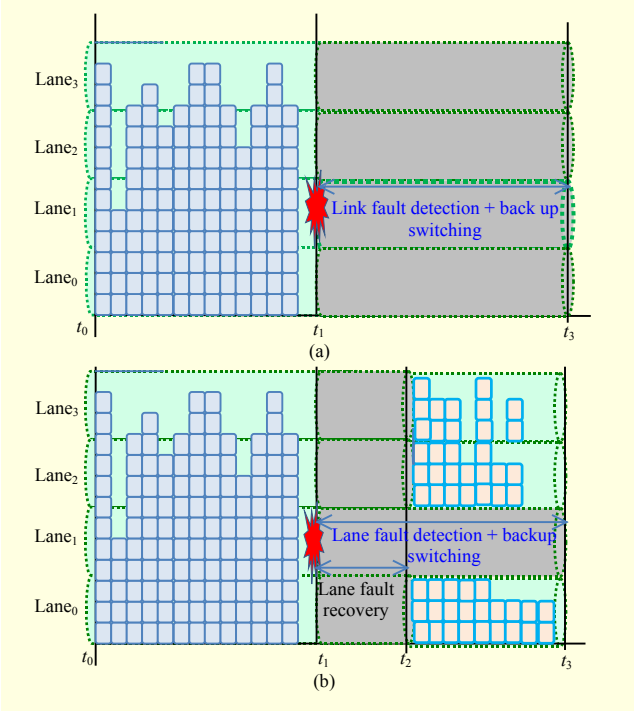

Fig. 1. Application of (a) link and (b) lane fault mechanism in multi-lane-based Ethernet link.

and continuously generates idle control characters. The RS returns to a normal status and transmits data when it no longer receives fault status messages. The fault status is signaled with a sequence ordered_set control character, which is used for transmitting the control information and link status.

Figure 1 shows the application of a link and lane fault mechanism in a multi-lane-based optical Ethernet link. A single data stream is simultaneously distributed and transmitted through all lanes of a link during a period of [t_0 - t_1]. Figures 1(a) and 1(b) show the recovery time and link reuse of the link and fault mechanism, respectively, when a lane failure occurs in lane$_1$ and the RS detects the fault at t_1. In Fig. 1(a), the link fault is detected at t_1 and the data transmission is stopped until the backup switching is completed at the upper layer at t_3. Therefore, the failed link is not used during the period t_1 - t_3, and the input traffic is queued during this period. This causes more packet loss according to a successful backup switching time, link capacity, and buffer size. In Fig. 1(b), the lane fault detection time and backup switching time are the same as in Fig. 1(a), a lane fault recovery is performed during the period t_1 - t_2, which takes less time than the backup switching time. After finishing the recovery of a faulty lane, the data are transmitted over the available lanes during the period t_2 - t_3. The packet loss can be greatly alleviated and the reuse of the link can be maximized, while at the same time, the lane fault recovery time is shorter.

Unlike a link fault, the lane fault recovery requires information on the lane location to correctly transmit normal data while removing a faulty lane. Therefore, the RS should support a new function and protocol for these requirements, as well as recovering both multiple and single lane faults to maximize the utilization of limited resources, and minimize the packet loss.

III. Lane Adaptive Recovery Scheme

Figure 2 shows the flexible lane adaptive recovery when multiple faults occur in a link. For the given n lanes of a link, if k lanes have failed ($1 \le k \le n-1$), then the data are transmitted over new transmission lanes that are recomposed with the ($n-k$) available lanes between the local and remote nodes. The available lanes can be controlled between one and $n-1$ as a partial lane fault according to the number of faulty lanes.

To support this scheme, the RS should provide link/lane fault management and data rate control as a new function for fault management. The former classifies the link fault and lane fault, and exchanges the control frames for informing the remote lane faults, recomposing the transmission link, as well as and confirming the start of transmission through a newly recovered link. The latter only generates and inserts idle characters on removed faulty lanes, and sends the data on available lanes. Upon receiving a local lane fault signal, the local RS stops sending MAC data and sends the remote lane fault status using a sequence ordered_set frame to the remote RS. The remote RS stops sending MAC data and transmits idle control characters instead. It also reports the lane fault to a higher layer and adjusts the transmitting lane by removing the failed lane according to the received lane fault information. After adjusting the lane, the remote RS sends the ACK to the local RS and retransmits data over the available lanes when receiving the confirmed frame as a response of the ACK. The data rate control, referred to in [1], simply controls the transmission rate by inserting idle characters on the corresponding failed lanes,

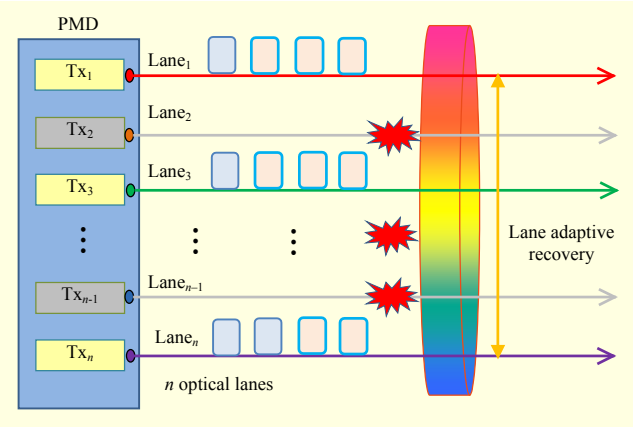

Fig. 2. Lane adaptive recovery for multiple lane faults.

and transmits data taken from the queue through available lanes. For example, if the first and third lanes fail in a four-lane link, then the RS transmits in order of idle, data, idle, and data on the first, second, third, and fourth lanes, respectively. The RS transmits data through a lower data rate employing available lanes until it obtains a backup indication signal from the upper layer. The RS returns to a normal status and transmits data at a full rate when it receives an indication that backup switching has been successfully completed from a higher layer.

Figure 3 shows the control-message exchange for multiple lane fault signaling and adaptive lane recovery. After receiving the local fault signal from the PHY, the local RS sequentially sends a remote lane fault message to the remote RS based on the number of faulty lanes. When all remote lane fault messages are received, the remote RS adjusts the new transmission lanes with the available lanes by excluding faulty lanes. After adjusting the lanes, the remote RS sends the same number of ACK messages as the number of remote fault messages received. The local RS knows that the remote node is ready to receive data over the available lanes in a blocking link, and finally sends a confirmation message to the remote RS. The remote RS again begins transmitting data over the lane-recovered link. The remote lane fault and confirmation messages are transmitted at a full rate because the transmission path is in a normal state in the local node, and all messages are transmitted over all lanes concurrently. However, the ACK message is transmitted at a lower rate one or more times according to the number of failed lanes. This may lead to an additional transmission delay during the message exchange time. The frequency of an ACK message transmission (N_t) can be determined as follows:

$$N_t = \left\lfloor \frac{k}{i} \right\rfloor,$$

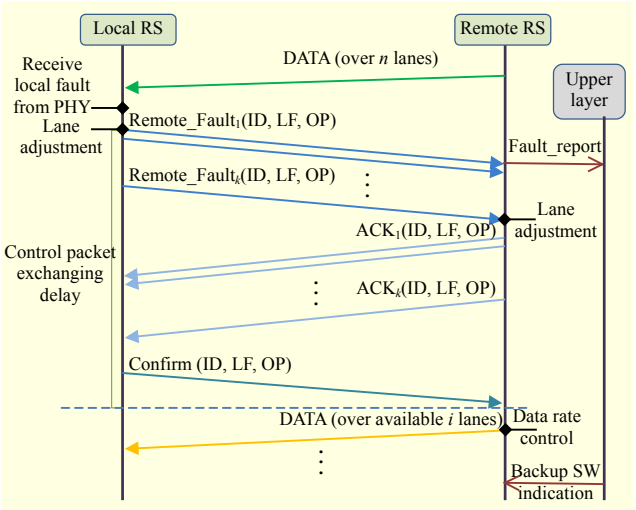

Fig. 3. Message exchange procedure.

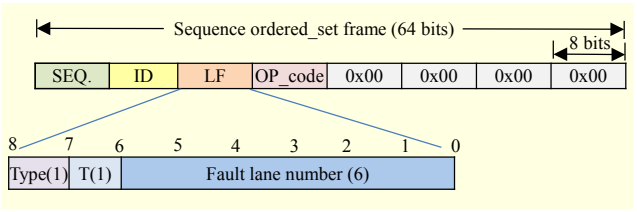

Fig. 4. Frame format of multiple lane fault recovery.

where k is the number of failed lanes and i is the number of available lanes.

Figure 4 shows the lane fault recovery message format. It consists of eight 8-bit fields. The sequence field of the first 8 bits is set to a sequence control character of 0x9c, as mentioned earlier. The ID field indicates the type of fault, and the link fault is set to 0x00, while the lane fault is set to 0x01. The LF field for the lane fault information is separated into the type, T, and fault lane number. It is only used for a lane fault and indicates the type of lane fault and information of a faulty lane as follows:

- The type (one bit) represents whether multiple faults exist: zero and one indicate single and multiple faults, respectively.
- T (one bit) is valid when multiple faults occur. It is used to indicate either the continuity or termination of the control frame: one indicates a termination. The number of faulty lanes can be inferred by this bit.
- The fault lane number (six bits) represents the location of a faulty lane by setting the identification of a failed lane. This lane is removed from the transmission lanes during the lane protection process.

This LF field is only used for a lane fault (that is, ID = 0x01) and is ignored for a link fault (that is, ID = 0x00). The OP_code field represents the type of message for fault signaling and lane control for lane fault protection. The settings for the local fault, remote fault, ACK, and confirmation messages are 0x01, 0x02, 0x03, and 0x04, respectively. The others are set to data characters of 0x00. This frame format is backward-compatible with link fault signaling.

IV. Performance Evaluation

Using an OPNET simulator, we examined the performance of the proposed lane adaptive recovery scheme. For this simulation, we considered a 100-Gb Ethernet optical link with a queue size of 100 Mbytes and a backup switching time of 50 ms. The packets are generated with an exponential distribution, having a mean size of 1045.94 bytes [1].

Figure 5 shows the lane recovery time according to the number of lane faults in both four- and ten-lane Ethernet links. The lane recovery time increases as the number of lane faults increases because each lane fault signaling message is transmitted with the fault information for each lane. In our

scheme, the lane recovery time is very short at under 0.12 μs, and this prompt recovery provides a lower packet loss and higher link reuse.

Figure 6 shows the packet loss rate according to the offered load. It typically increases as the offered load and number of failed lanes increase. Because our scheme can recover until nine out of ten lanes fail with a very short recovery time, the packet loss rate is much lower than the link fault method although the number of faulty lanes is nine. Even when the number of faulty lanes is under three, packet loss hardly occurs.

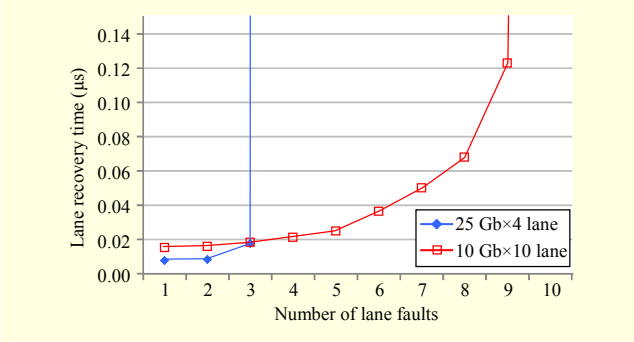

Fig. 5. Lane recovery time.

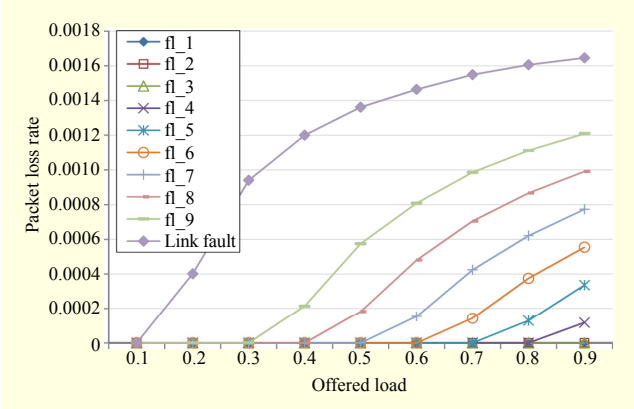

Fig. 6. Packet loss rate.

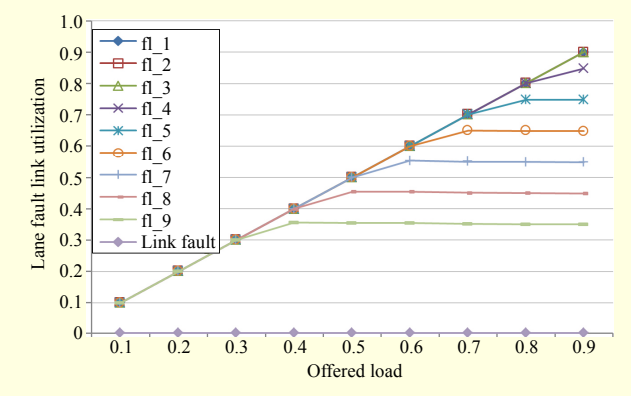

Fig. 7. Utilization of link with faulty lane.

Figure 7 shows the utilization of a link containing a faulty lane according to the offered load. When the lane fault is processed as a link fault, the link utilization is zero regardless of the offered load and number of composing lanes. In lane fault recovery, the link utilization increases as the offered load increases, but decreases as the number of lane faults increase. When the number of lane faults is less than two, the link utilization is very high, and the minimum usage is around 35% when the offered load is more than 0.4.

V. Conclusion

We proposed a lane adaptive recovery scheme for multiple lane faults in a multi-lane-based Ethernet link. For multiple lane fault recovery, lane fault messages with fault type and faulty lane number are exchanged between RSs. Simulation results show that the proposed method provides a link reuse utilization of around 60% and a packet loss of less than 10^{-4}.

References

[1] K.E. Han et al., "An Energy Saving Scheme for Multilane-Based High-Speed Ethernet," *ETRI J.*, vol. 34, no. 6, Dec. 2012, pp. 807–815.

[2] IEEE Std 802.3ba-2010, *Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, *Amendment 4: Media Access Contr. Parameters for 40 Gb/s and 100 Gb/s Operation*, June 2010.

[3] T. Malpass, "Partial Fault Protection," presented at the IEEE 802.3 Higher Speed Study Group, San Francisco, CA, USA, Sept. 10–13, 2007.

[4] W.B. Jiang et al., "Multi-lane PMD Reliability and Partial Fault Protection (PFP)," presented at the IEEE 802.3ba Task Force, Portland, OR, USA, Jan. 23–25, 2008.

[5] L. Zeng et al., "PHY OAM and Lane Fault Monitoring," presented at the IEEE 802.3ba Task Force, Portland, OR, USA, 2008.

[6] S. Ramamurthy, L. Sahasrabuddhe, and B. Mukherjee, "Survivable WDM Mesh Networks," *J. Lightw. Technol.*, vol. 21, no. 4, Apr. 2003, pp. 870–883.

[7] K.-K. Lee, J.D. Ryoo, and S. Min, "An Ethernet Ring Protection Method to Minimize Transient Traffic by Selective FDB Advertisement," *ETRI J.*, vol. 31, no. 5, Oct. 2009, pp. 631–633.

[8] L. Guo et al., "Enhanced Dynamic Segment Protection in WDM Optical Networks under Reliability Constraints," *ETRI J.*, vol. 28, no. 1, Feb. 2006, pp. 99–102.

[9] B. Wu, P.-H. Ho, and K.L. Yeung, "Monitoring Trail: On Fast Link Failure Localization in All-Optical WDM Mesh Networks," *J. Lightw. Technol.*, vol. 27, no. 18, Sept. 2009, pp. 4175–4185.