Adaptive Data Transmission Control for Multilane-Based Ethernet

Kyeong-Eun Han, Kwangjoon Kim, SunMe Kim, and Jonghyun Lee

We propose a reconciliation sublayer (RS)-based lane and traffic control protocol for energy-efficient 40-G/100-G Ethernet. The RS performs active/inactive lane control and data rate adaptation depending on active lane information received from the upper layer. This protocol does not result in a processing delay in the media access control layer, nor is an additional buffer required at the physical layer for dynamic lane control. It ensures minimal delay and no overhead for the exchange of control frames and provides a simple adaptive data rate.

Keywords: Adaptive traffic control, physical layer protocol, energy-efficient Ethernet.

I. Introduction

Energy conservation has become an important issue in next-generation networks. It is necessary to reduce the energy consumption while providing adequate performance and a high capacity. Networks typically consume a constant amount of energy that does not vary with the network activity. On average, the use of most edge links and the backbone is below 5% and 30%, respectively [1]-[6]. The IEEE 802.3az task force standardized energy-efficient Ethernet (EEE) [7], which is defined as an approximately 100-Mbps to 10-Gbps Ethernet link for energy conservation during an idle status. Meanwhile, the IEEE 802.3ba task force standardized multiple-lane-based 40-Gbps and 100-Gbps Ethernet for local server applications and Internet backbone, respectively [8]. However, the issue of energy efficiency was not considered.

Manuscript received Apr. 23, 2012; revised Aug. 21, 2012, accepted Sept. 6, 2012.

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 (phone: +82 42 860 5758, kehan@etri.re.kr), Kwangjoon Kim (kjk@etri.re.kr), SunMe Kim (kimsunme@etri.re.kr), and Jonghyun Lee (jlee@etri.re.kr) are with the Communications Internet Research Laboratory, ETRI, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.13.0212.0169

The issue of energy efficiency is important for 40-G/100-G Ethernet because when the network speed is higher, more power is consumed. The EEE approach defined by 802.3az is not appropriate for 40-G/100-G Ethernet because of the long turn on/off time of optical devices and because there are multiple lanes per link. This results in a high packet loss and queueing delay during the turn on time of the link. Therefore, a new approach, such as dynamic lane control, is required for efficient energy consumption in a high-capacity link consisting of multiple low-capacity lanes. Dynamic lane control dynamically activates the lanes depending on the load. To reallocate the lanes between two nodes and to transmit the data at a new rate, it is necessary to adaptively determine the appropriate number of transmitting lanes given the state of the network. Additionally, some functions of the physical layer (PHY), including the transceiver, must change to provide dynamic lane operation. While some algorithms have been proposed for dynamic lane control [1]-[4], the issue of lane control and an adaptive data rate for 40-G/100-G Ethernet have not yet been discussed.

We propose a dynamic lane control protocol for energy conservation in 40-G/100-G Ethernet. The reconciliation sublayer (RS) adjusts the transmitting lanes using information about the active lanes from the upper layer. For the control frame, we use a 66-bit sequence-ordered set frame, generated at the RS and transmitted during the interframe time. The RS also performs the data rate adaptation by inserting idle frame blocks into inactive lanes. This RS-based lane control mechanism provides a simple data rate adaptation and a low transmission delay. It imposes no processing delay in the media access control (MAC) layer and no overhead for the exchange of control frames.

II. Multilane-Based 40-G/100-G Ethernet

Figure 1 shows the data transmission in multilane-based

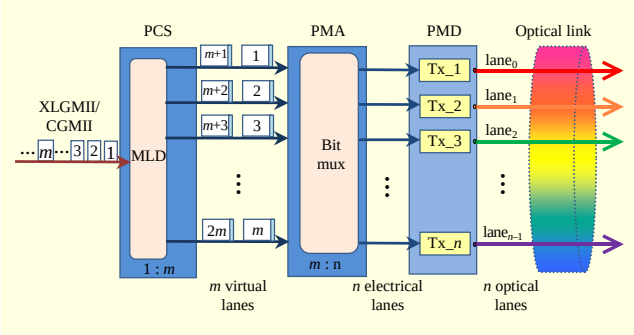

Fig. 1. Multilane-based 40-G/100-G Ethernet.

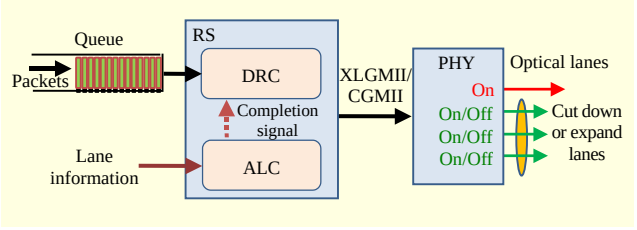

Fig. 2. Functional block diagram for RS.

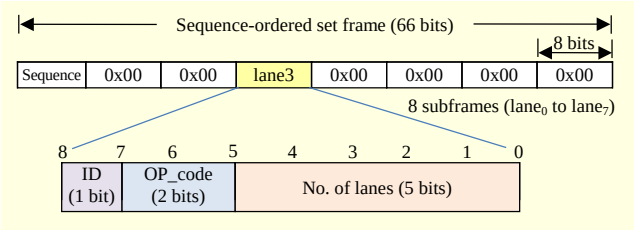

Fig. 3. Format of lane control frames.

40-G/100-G Ethernet [8]. The serial data transmitted through the XLGMII/CGMII interface is encoded and scrambled with 66-bit blocks and distributed to *m* virtual lanes via multilane distribution, which is a round-robin distribution. This allows the physical coding sublayer (PCS) to support multiple physical lanes in the physical medium dependent (PMD) sublayer. The 66-bit blocks transmitted over the *m* PCS virtual lanes are mapped to *n* electrical lanes in the physical medium attachment (PMA) and then transmitted to the receiver over the *n* optical lanes of the optical link. Here, the number of virtual and physical lanes {*m*, *n*} is {4, 4} for 40-G Ethernet, {20, 4} for four-lane 100-G Ethernet, and {20, 10} for ten-lane 100-G Ethernet. In this architecture, all lanes are active regardless of the link utilization.

As mentioned earlier, the 802.3ba standard uses all lanes of a link for transmission, and the PHY statically carries out its function for a fixed number of lanes. In this letter, however, we use partially dynamic lane control [4] as a dynamic lane control method that identifies whether all lanes are dynamically used (which we call "fully dynamic lane control"). In partially dynamic lane control, with a given *n* total number of lanes, some of them (*i* lanes) are used statically and the rest (*n*–*i* lanes) are used dynamically for a data transmission depending on the link utilization. This may alleviate the performance degradation of networks. Note that the network has a better transmission performance when using all lanes than when using only some of the lanes, although the latter situation results in greater energy efficiency. In this letter, we consider one static lane, which is always used for transmitting data, and a control frame, with the remaining lanes used dynamically.

III. RS-Based Dynamic Lane Control Protocol

Figure 2 shows a functional block diagram of the RS for dynamic lane and traffic control. The RS generally splits a packet into a sequence of 64-bit frames and transmits them with 8-bit control signals to the PCS. The RS contains two parts of an active lane controller (ALC) and a data rate controller (DRC), which are newly defined for dynamic lane control and adaptive data rate control. Initially, the DRC carries out a defined number of default lanes. If the RS receives new lane information from a higher layer, the ALC creates 64-bit control frames and transmits them during the interframe time, which is the idle period between the transmission of data frames. Note that the transmission of very short control frames through the interframe period allows for less delay and no overhead for exchanging the control frames. After completing a lane change, the ALC sends a completion signal of the lane change to the DRC. The DRC inserts idle frames to be mapped into inactive electrical lanes. These inserted idle frames are dropped at the turn off transmitters, and the data frames are transmitted over the active lanes. However, in partially dynamic lane control, one lane is used as a static lane and the others are cut down or expanded depending on the link utilization.

The control messages are based on an 8-byte sequence-ordered set frame that is generally used to send both control and status information (for example, the fault status) over the link. Three control frames are defined for the lane control: LANE_CTRL_REQ, LANE_CTRL_ACK, and LANE_CTRL_BEGIN. Figure 3 shows the lane control message format. It consists of eight 8-bit subframes ($lane_0$ to $lane_7$). The first subframe ($lane_0$) is set to a sequence control character representing the sequence-ordered set frame. The fourth subframe ($lane_3$) contains the lane control information, and the others are set to data characters of 0x00. There are three fields that indicate the type of message and number of lanes to use:

- ID (1 bit) is the identification bit; 1 indicates a lane control message.
- OP_code (2 bits) indicates the type of lane control frame.

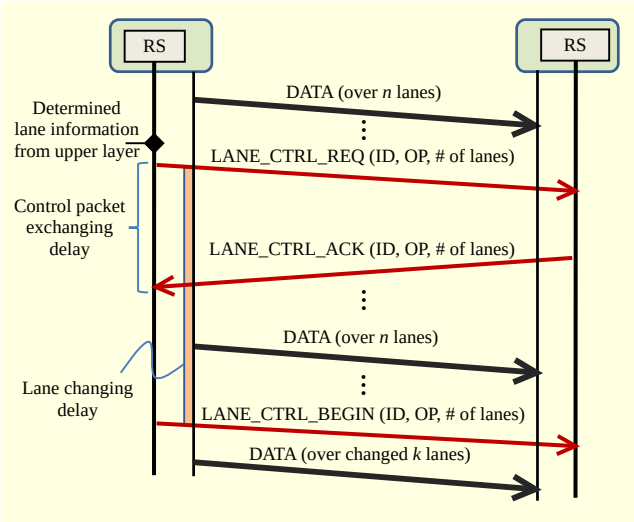

Fig. 4. Procedure for dynamic lane control

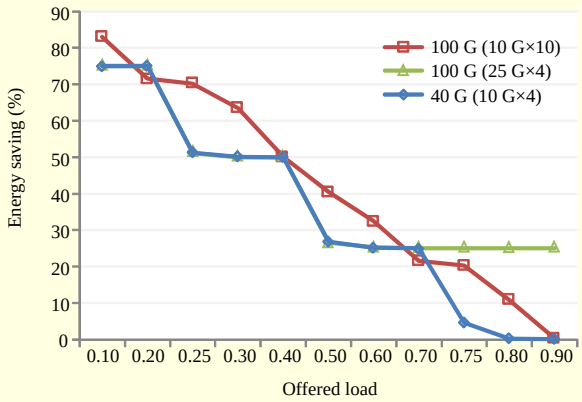

Fig. 5. Energy reduction.

The settings for the LANE_CTRL_REQ, LANE_CTRL_ACK, and LANE_CTRL_BEGIN messages are 10, 01, and 11, respectively.

• The number of lanes (5 bits) is the number of active lanes for a data transmission.

This control message format may expand to indicate another lane status, such as a lane fault, and the physical identification of lanes more easily. However, we do not discuss this in this letter, as it is another issue beyond the scope of our topic.

Figure 4 shows the control message exchange in dynamic lane control. After receiving the lane information, the RS sends the LANE_CTRL_REQ over the current n active lanes to the corresponding RS. The receiving RS sends the LANE_CTRL_ACK as a response and adjusts the set values for synchronization and alignment depending on the number of lanes to be used. After completing this adjustment, the RS transmits the LANE_CTRL_BEGIN to indicate the start of a data transmission over the new k active lanes. The corresponding RS then indicates to the PHY the number of active lanes on the path.

A change of data rate for transmission is required as the number of active lanes is changed. For this, the RS sends the data and idle frames into active and inactive lanes, respectively. For example, if two lanes are active and two lanes are inactive, gradually, the RS sends two data frames and two idle frames, repeatedly. This makes avoiding the packet drop of the PHY caused by traffic transmitted from the higher layer possible. These inserted idle frames are only used for traffic control and dropped at the transmitters.

IV. Performance Evaluation

Using the OPNET simulator, we design a lane decision algorithm, our lane control protocol, and a multilane-based Ethernet link and evaluate the performance in terms of the control packet exchange delay, lane change delay, and energy reduction. We consider a 40-G/100-G optical link consisting of four or ten lanes. We assume that there is one active default lane and one static lane, and the turn on/off times are 100 ms and 100 μs, respectively [9]. The packet length is variable and follows an exponential distribution with a mean of 1,045.94 bytes for traffic generation [10]. The offered load varies from 0.1 to 0.9. We use the algorithm proposed in [1] to determine the number of active lanes depending on the status of the network. It employs the average input traffic load, queue threshold, current queue size, and increase and decrease rates of the queue as decision parameters for applying network status. For the decision algorithm, we assume that the queue size is 30 ms, traffic monitoring time (T) is 500 ms, weight of traffic (α) is 0.6, queue occupation rate (β) is 0.2, and increase and decrease threshold (δ) is 2.0. The simulation time is 100 s.

Figure 5 shows a graph of the energy reduction against the load in a link. The energy efficiency decreases with a lane operation unit such as 10 G or 25 G because the number of lanes required increases to cover the increased traffic amount. On the other hand, the energy efficiency increases considerably when the link has many lanes with a smaller lane operation unit at the same link capacity. As mentioned above, when we assume that the utilization is less than 30%, the maximum energy reduction is more than 50% and 70%, respectively, for 40-G/100-G Ethernet. Energy reduction is a relative value for each link, and the actual power reduction depends on the number of off lanes. For example, assuming that a given optical device consumes 1,000 mw when turning on a lane, saving 75% in energy indicates a reduction of 3,000 mw and 7,000 mw for a 4-lane 100-G link and a 10-lane100-G link, respectively.

Figure 6 shows a graph of the average time for the exchange

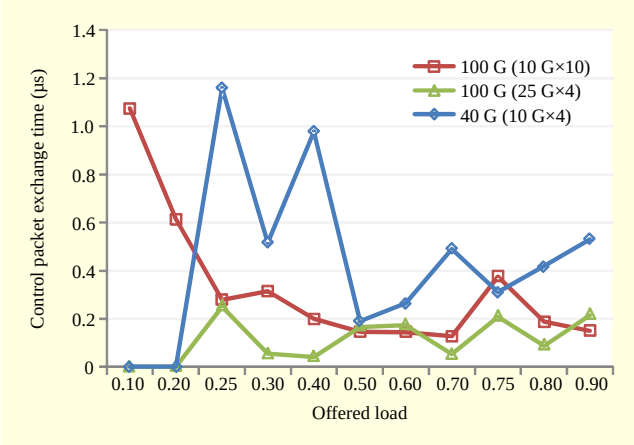

Fig. 6. Time for exchange of control packets.

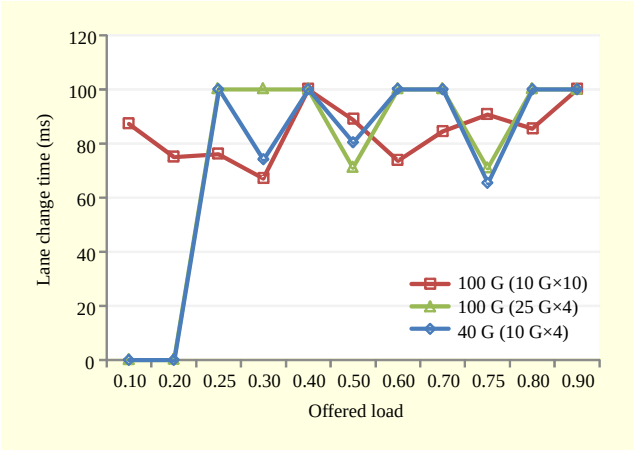

Fig. 7. Time for lane change.

of control packets against the load. Lane changes occur depending on the lane operation unit based on the traffic status. Four-lane 40-G/100-G Ethernet can provide adequate support with one lane when the load is less than 0.25; thus, there is no lane changing. The time for the exchange of control packets for 100-G Ethernet is less than that for 40-G Ethernet because when more active lanes are required, the data transmission rate increases. This is affected not by the network traffic but by the length of the transmitting data packet because the control packets are sent during the interframe time. There is a maximum 1.2-μs delay for the exchange of the control packets.

Figure 7 shows a graph of the time for a lane change against the load. This includes the time between the completion of the lane change and the start of transmission over the active lanes. The turn on/off time is the main component of the delay: the laser on/off time ranges from 65 ms to 100 ms, whereas the exchange of control packets is rapid. However, the rapid delivery of new lane information is important when we consider the reconfiguration time for synchronization and alignment in accordance with the total number of active lanes at the PHY. The limitation of the on/off time may be overcome by the scheduling algorithm and devices themselves.

V. Conclusion

We proposed an RS-based dynamic lane control protocol and partially dynamic lane control for energy reduction in 40-G/100-G Ethernet. The RS controls the active/inactive lanes and adaptively sends data at the changed rate. Our protocol is simple, with little overhead and no processing delay at the MAC layer. The simulation results showed that it provides a reduction in energy consumption with only a small delay in the exchange of control packets. A reduction in the long turn on time of the optical device will further reduce the energy consumption.

References

[1] K.E. Han and K.J. Kim, "Dynamic Lane Decision Method for Energy Efficiency in 40G/100G Ethernet," *Proc. CEIC*, Dec. 2011, pp. 25-28.

[2] H. Imaizumi et al., "Power Saving Technique Based on Simple Moving Average for Multi-channel Ethernet," *Proc. OECC*, FT3, Aug. 2009, pp. 1-2.

[3] P. Reviriego et al., "Improving Energy Efficiency in IEEE 802.3ba High-Rate Ethernet Optical Links," *IEEE J. Sel. Topics Quantum Electron.*, vol. 17, no. 2, 2011, pp. 419-427.

[4] 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.

[5] T. Yang, C. Zhao, and D. Chen, "Feedback Analysis of Transcutaneous Energy Transmission with a Variable Load Parameter," *ETRI J.*, vol. 3, no. 4, Aug. 2010, pp. 548-554.

[6] M.C. Domingo, "Packet Size Optimization for Improving the Energy Efficiency in Body Sensor Networks," *ETRI J.*, vol. 33, no. 3, June 2011, pp. 299-309.

[7] http://www.ieee802.org/3/az/index.html

[8] IEEE Std. 802.3ba-2010, "Part3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, Amendment 4: Media Access Control Parameters for 40Gb/s and 100Gb/s Operation," June 2010.

[9] http://www.cfp-msa.org

[10] S. McCreary and K. Claffy, "Trends in Wide Area IP Traffic Patterns," Cooperative Association for Internet Data Analysis (CAIDA), University of California, San Diego (UCSD), La Jolla, CA, USA. http://www.caida.org