Lane alignment characteristics in massively parallel Ethernet-based transmission systems

Kyosuke Sugiura\textsuperscript{1a)}, Masaki Murakami\textsuperscript{1}, Yoshihiko Uematsu\textsuperscript{1}, Satoru Okamoto\textsuperscript{1}, and Naoaki Yamanaka\textsuperscript{1}

\textsuperscript{1}Graduate School of Science and Technology, Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan
\textsuperscript{a)} kyosuke.sugiura@yamanaka.ics.keio.ac.jp

Abstract: We have proposed the 10 Tbps class dynamic bandwidth flexible Ethernet-based massively parallel media access control (MAC) system; “Dynamic MAC”. The dynamic MAC requires up to 400 lanes parallel transmission and up to 10 ms skew canceling. Therefore, lane alignment among 400 lanes should be established. In this paper, numerical analysis results of lane alignment characteristics; alignment time and link-up time of the dynamic MAC system, are presented.

Keywords: Ethernet, massively parallel, lane alignment, physical layer

Classification: Network system

References

[1] S. Okamoto, K. Sugiura, M. Murakami, E. Oki, and N. Yamanaka, “Research on Dynamic MAC Technologies for Future Massively Parallel and Sliced Optical Networks,” IEICE Technical Report, PN2019-25, pp. 7-14, Nov. 2019 (in Japanese).

[2] K. Sugiura, M. Murakami, Y. Uematsu, S. Okamoto, and N. Yamanaka, “Hierarchical round-robin mapper emulator for evaluating massively parallel Ethernet physical layer developing,” IEICE Communications Express, Vol. 10, No. 5, pp. 248-253, May 2021. DOI: 10.1587/comex.2021XBL0003

[3] H. Zhang, Y. Hou, Y. Chen, and S. Li, “Analysis of simplified frame synchronization scheme for burst-mode multi-carrier system,” IEEE Communications Letters, vol. 23, no. 6, pp. 1054-1056, Apr. 2019. DOI: 10.1109/LCOMM.2019.2913423

[4] D. Choi, “Frame alignment in a digital carrier system - a tutorial,” IEEE Communications Magazine, vol. 28, no. 2, pp. 47-54, Feb. 1990. DOI: 10.1109/35.46685

[5] A. Meurer, C. P. Smith, M. Paprocki, O. Certik, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, V. Roucka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz, “Sympy: symbolic computing in python,” PeerJ Computer Science, vol. 3, p. e103, Jan. 2017.
1 Introduction

We have proposed a dynamic media access control (MAC) concept [1, 2]. The dynamic MAC will provide 10 Tbps class network interface with spatial division multiplexing (SDM) technologies. We have determined 25 Gbps granularity in the dynamic MAC which covers 25 Gbps in 1 lane to 10 Tbps in 400 lanes. 400 lanes will be transported over single core, multiple cores, multiple fibers, and multiple route transmission systems. In the multiple routes case, large transmission delay variations among lanes (i.e. skew) should be cancelled. Current standard Ethernet supports in principle maximum 108 µs skew cancelling (de-skew) with single core transmission. We have determined that the dynamic MAC should be challenged to realize maximum 10 ms de-skew in the multiple route transmission case. In [2], we have introduced a hierarchical round-robin mapping scheme between a MAC layer and a physical coding sublayer (PCS). Therefore, alignment among 400 PCS lanes with maximum 10 ms de-skew should be designed. Current standard Ethernet uses periodical alignment marker (AM) insertion method for aligning lanes and canceling skew. The dynamic MAC also uses the AM insertion method. In this paper, numerical analysis results of AM-based lane alignment characteristics; alignment time and link-up time of the dynamic MAC system, are presented.

2 Design of the alignment marker for the dynamic MAC

The standard Ethernet inserts one 64-bit AM block in every 16,384 64-bit blocks. The AM contains a 24-bit lane indicator (ID), 8-bit bit interleaved parity (BIP-8), and 32-bit inverted first 32-bit pattern for keeping the direct-current balance. 16,384 blocks are corresponding to about 216 µs in the 100GE (20 PCS lanes and 25 Gbps×4 transmission lanes). Therefore, maximum 108 µs skew can be canceled, and 20 PCS lanes can be aligned with in double AM interval time even in the not good transmission quality (e.g. bit error rate (BER) = 10^-5 without forward error correction) condition. We should extend up to 400 PCS lanes and 10 ms de-skew. Extend the number of lanes will cause extension of the alignment time. To realize the 10 ms de-skew, two possible AM-based lane alignment procedure candidates are evaluated. One is extent the AM insertion interval time from 216 µs to 216×256=55,296 µs (method #1). This leads 27 ms de-skew capability. Second is modify the AM lane ID from 24-bit ID to 16-bit ID with 8-bit counter value (method #2). The AM insertion interval is 216 µs, but it has the effect of equivalently multiplying the AM insertion period by 256. In this paper, we do not discuss with the hardware amount and complexity of the AM insertion method,
only lane alignment characteristics of two methods will be discussed.

3 Numerical evaluation of two alignment methods

Ethernet uses 64B66B coding for physical transmission. Therefore, it is necessary to search for a specific 64-bit AM pattern in the received 66-bit block sequence to all lanes with bit error circumstances. A cross-correlation detector is commonly used to find specific patterns [3]. The frame alignment timing parameters of this AM search process and furthermore alignment procedure are estimated as a probability generating function (PGF) [4]. Let \( l, n, c, k, \) and \( BER \) denote the length of AM, the interval of AM, the number of forward protection steps, the number of lanes, and bit error ratio, respectively. The AM search process compares blocks up to \( n \) times and ends when the AM is found. The signal transfer function \( \tau(z) \) of this process can be obtained by Eq. (1) [4].

\[
\tau(z) = \frac{(1-P_s)^{n-1} z}{(1-P_d)^{n-1}}
\]  

Where, \( P_s \) is a probability of marking non-AM blocks as AM when the bit sequence of regular blocks happens to be the same as AM (=1/2^l), and \( z \) is a dummy variable. The PGF of the alignment process \( P_{RF}(z) \) that combines both search and confirmation processes is expressed by Eq. (2) [4].

\[
P_{RF}(z) = \frac{P_d \tau(z) z^{c-1}}{1-(1-P_d)\tau(z)z^{c-1}z^{-1}}
\]

Where, \( P_d \) is a probability of correctly detecting AM without any bit error \((=1−BER)^l\). From Eq. (2), a probability \( p_{RF}(t) \) that the alignment completes at time \( t \) is calculated as Eq. (3). The unit of the time \( t \) is the AM interval.

\[
p_{RF}(t) = \frac{p_{RF}^{(0)}(z)}{t!}
\]

Where \( (t) \) denotes \( t \)-th differentiating. Eq. (3) can provide an alignment time distribution in a single lane \((k=1)\) case. A probability \( q_{RF}(t) \) that the alignment is completed by time \( t \) in a single lane is provided by Eq. (4).

\[
q_{RF}(t) = p_{RF}(0) + p_{RF}(1) + \cdots + p_{RF}(t) = \sum_{u=0}^{t} p_{RF}(u)
\]

Eq. (4) should be expanded to the \( k \)-lane case. Let \( p_{RF}(i,t) \) denotes as a probability that \( i \)-th lane alignment is completed at time \( t \). A probability \( p_{RF}'(t) \) that alignment of all \( k \) lanes is completed at time \( t=0 \) is expressed by Eq. (5).

\[
p_{RF}'(0) = p_{RF}(1,0) \times p_{RF}(2,0) \times \cdots \times p_{RF}(k,0)
\]

We can assume that all lane alignments are independent events. Therefor Eq. (5) can be rewritten as following:

\[
p_{RF}'(0) = p_{RF}(0)^k
\]

In the same way, we can get a probability \( q_{RF}'(t) \) that all \( k \) lanes are aligned by time \( t \) is expressed by Eq. (7).

\[
q_{RF}'(t) = q_{RF}(1,t) \times q_{RF}(2,t) \times \cdots \times q_{RF}(k,t) = q_{RF}(t)^k
\]

Finally, a probability \( p_{RF}'(t) \) can be got as following:

\[
p_{RF}'(t) = q_{RF}'(t) - q_{RF}'(t-1)
\]
\[
= \left( \sum_{u=0}^{t} p_{RF}(u) \right)^k - \left( \sum_{u=0}^{t-1} p_{RF}(u) \right)^k
\]  
(8)

Eq. (8) gives the alignment time distribution when the number of lanes is massive.

To calculate Eq. (8) as parameters \(l, c, n, \) and \(k,\) a Python library for symbolic mathematics “SymPy” [5] is applied. Figure 1 shows the alignment time distribution under \(k = 1, 20,\) and \(400,\) BER =10\(^{-3}\), 10\(^{-5}\), and 10\(^{-10}\) conditions. The average alignment time from Fig. 1 is shown in Table 1. Parameters are set to \(l =48,\) \(c =2,\) and \(n =16,384.\) Note that \(l =48\) corresponds to 24-bit lane ID and 24-bit inverted lane ID, and the unit of time axis in Fig. 1 is the AM insertion interval. In case of \(k = 20\) in Fig. 1(a) means that the alignment complete after two AM intervals with a probability of 90.8%. Fig. 1(a) and Fig. 1(b) show that the alignment time distribution shifts to the right in the high-error environment. When BER=10\(^{-3}\), the average alignment time of \(k =400\) was 1.6 times higher than of \(k =20.\) When BER=10\(^{-5}\), this was reduced to 1.2 times. We can confident that 400 lanes can be applicable in the actual field. Under good transmission conditions (i.e. BER=10\(^{-10}\), the average alignment time is almost the same in all \(k\) cases.

![Graphs showing alignment time distribution under different BER conditions](image)

**Fig. 1.** Alignment time distribution under different BER. (a) BER=10\(^{-3}\). (b) BER=10\(^{-5}\). (c) BER=10\(^{-10}\).

**Table 1.** Average alignment time under different BER.

| \(k\) | BER=10\(^{-3}\) | BER=10\(^{-5}\) | BER=10\(^{-10}\) |
|------|-----------------|-----------------|-----------------|
| 1    | 2.150           | 2.001           | 2.000           |
| 20   | 3.667           | 2.029           | 2.000           |
| 400  | 5.765           | 2.494           | 2.000           |

To evaluate the two alignment methods, parameter \(n =16,384 \times 256 =\)
4,194,304 should be applied to the method #1. While SymPy can evaluate a floating-point expression with arbitrary precision, calculating derivatives and factorials is time-consuming, making it challenging to calculate Eq. (8) when \( n \) is large. Therefore, we used the finite difference approximation [6] to derivatives to perform calculations quickly at the cost of certain errors. We evaluated the error of the approximate solution in the case of \( l = 48, c = 2, n = 16,384 \), and \( BER = 10^{-5} \), from \( t = 2 \) to \( 11 \). As a result, maximum 4.2 \% relative error was obtained at \( t = 11 \). Note that \( p_{RR}(11) = 1.3 \times 10^{-16} \). This means that the approximation may not cause a critical calculation error. Figure 2 shows the evaluation of the alignment time distribution for the two methods. The values of both graphs are almost identical within three significant digits. However, the horizontal axis of method #1 is multiplied by 256 to keep the alignment time scale the same for methods #1 and #2.

![Fig. 2. Alignment time distribution in proposed de-skew methods (BER=10^{-5}). (a) Distribution in the method #1 (\( n = 4,194,304 \)). (b) Distribution in the method #2 (\( n = 16,384 \)).](image)

As shown in Fig. 2, the method #2 was able to reduce the alignment time to 1/256 compared to the method #1. This means that the method #2 can reduce the expected link-up time which can be calculated as skew value for waiting time in buffer plus the expected lane alignment time. Here, the link-up is defined as a situation where 400 PCS lanes are synchronized and the upper sublayer can work. Under good transmission conditions, the expected alignment time of each method is 512 (=2×256) AM time (110.592 ms) and 2 AM time (432 \( \mu s \)), respectively. Therefore, the method #2 can reduced by nearly 110 ms compared to the method #1. Under condition of \( BER=10^{-5} \), the difference expanded to 137 ms; [skew + 137.9 ms] vs. [skew + 540 \( \mu s \)].

4 Conclusion

We evaluated the two alignment methods for 400 lane dynamic MAC systems. It was confirmed that the alignment time of all 400 lanes was within 2.5 AM time (540 \( \mu s \)) on average under condition of \( BER=10^{-5} \). The method #2 provides faster link-up time (actually around 540 \( \mu s \) plus skew; max. 10 ms) than the method #1.

Acknowledgments

This work is partly supported by the “Massively Parallel and Sliced Optical Network (MAPLE)” project funded by the National Institute of Information and
Communications Technology (NICT), Japan.