100 Gigabit Ethernet

We illustrate the technical overview of 100 Gigabit Ethernet, which will become necessary in the near future with continuous growth in traffic. We also share the knowledge gained through our joint interoperability test.

3.1 Introduction

In this report we first illustrate the technical overview of 100 Gigabit Ethernet (henceforth "GbE"), then report on the 100GbE IX (ISP Internet exchange point) joint interoperability test conducted with Internet Multifeed Co. and NTT Communications Corporation, and finally comment on 100GbE optical transceiver standards and future trends.

3.2 100GbE

100GbE was set out in IEEE 802.3ba*1, and ratified as the standard in June 2010. Here we discuss important points for understanding 100GbE. Because much of the technology established in 10GbE is reused, we assume that the reader has detailed knowledge of 10GbE*2.

■ Basic Principles of 100GbE

The technical hurdles involved in the serial transmission of a signal 10 times that of 10GbE are high, and implementation-related costs are also an issue. For this reason 100GbE is implemented using parallel low speed data transmission such as 10Gbps or 25Gbps (Figure 1). The technology that makes this parallel data transmission possible is one of the key features of 100GbE.

■ Comparison of 100GbE and Link Aggregation

This parallel data transmission technology is called MLD (Multi Lane Distribution), and it is implemented in the physical layer of the OSI reference model (Figure 2). Link aggregation is a similar technology for implementing parallel data transmission that is set out in IEEE 802.3ad. It involves bundling multiple interfaces and presenting them as a single virtual interface, but it differs from 100GbE in the following ways.

1. Because 100GbE conducts parallel data transmission over the physical layer, users do not need to care about parallel data transmission settings or behavior, but for 802.3ad close attention must be paid to link aggregation settings and behavior.
2. Because 100GbE data transmission breaks Ethernet frames down into fixed length to transmit data down parallel channels evenly, there is no risk of a specific channel carrying a disproportionate amount of data. In contrast, 802.3ad, has the problem that the frame distribution method may not disperse frames into channels evenly*3.

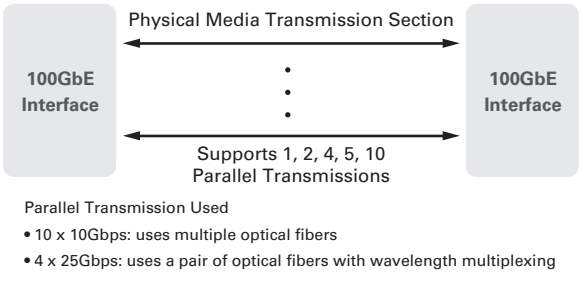

Figure 1: Basic Principles of 100GbE

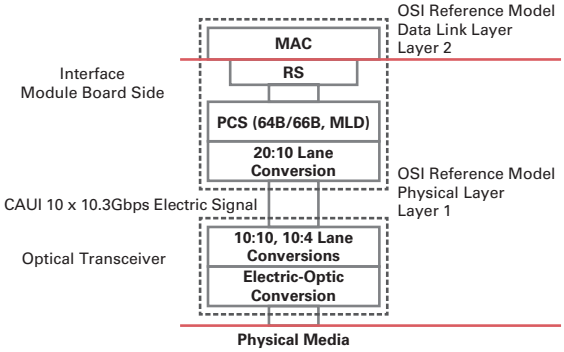

Figure 2: 100GBASE-SR10/LR4/ER4

*1 IEEE802.3ba document (http://standards.ieee.org/about/get/802/802.3.html)

*2 We recommend that readers look over the 2002 first edition of the "10 Gigabit Ethernet Textbook," which was written by Osamu Ishida and edited by Koichiro Seto, and published by IDG Japan.

*3 The distribution method of 802.3ad is frame-based and does not take frame length into consideration. This leads to the uneven utilization of links. 802.3ad also does not specify any detailed algorithm to implement frame-based distribution, so distribution among links may be uneven depending on the equipment.

■ MLD

The RS (Reconciliation Sub-layer) receives Ethernet frames from the data link layer (MAC), and passes them on to lower layers after breaking them down into 64-bit units. In the PCS (Physical Coding Sub-layer), 64B/66B blocks used in physical media transmission are made at first by adding a 2-bit header to the 64-bit data. The MLD function then sends these blocks in rotation down parallel data transmission channels called virtual lanes*4 (Figure 3).

■ Multiplexing and Demultiplexing of Virtual Lanes

In the course of transmission, virtual lanes can be aggregated in stages, and in this case data between lanes is multiplexed at bit level (Figure 4). If it is possible to use multiplexing for lane aggregation, the number of lanes can be changed as necessary, meaning transmission adapted to a variety of physical media can be supported*5. However, when multiplexing is used at bit level partway down the transmission channel, the lanes no longer align between the sender and recipient (see lane 5 in Figure 4). For this reason an alignment marker system is used to inform the recipient of the alignment of lanes between the sender and recipient.

■ Alignment Markers

For every 16383 64B/66B blocks of data sent, data block transmission is temporarily suspended, and each alignment marker is sent down all lanes at the same time (Figure 5). For example, the alignment marker for physical lane 5 on the sender side contains information identifying it as lane 5. If this is received over physical lane 1 on the recipient side, it is interpreted as data corresponding to virtual (sender side physical) lane 5 by checking the ID.

■ Bit Error Monitoring for Virtual Lanes

When conducting parallel transmission using multiple lanes, data may be transmitted over different physical media for each lane. For this reason, the BIP (Bit Interleaved Parity) function is implemented to monitor the line quality of each lane (Figure 6). The BIP function calculates the bit parity of the 16384 64B/66B blocks including the previous alignment marker for

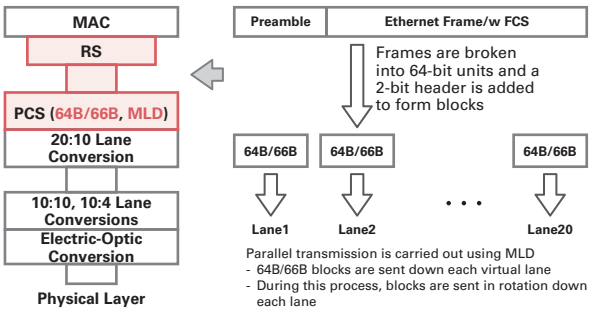

Figure 3 RS+PCS (64B/66B, MLD)

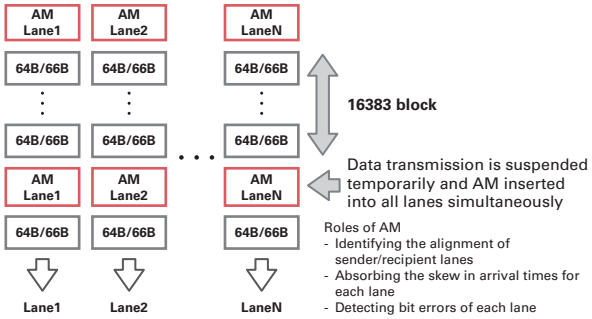

Figure 5: Alignment Markers (AM)

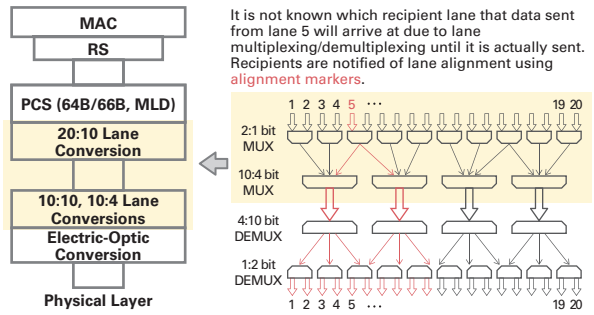

Figure 4: Relationship Between Sender/Recipient Lanes During Lane Multiplexing/Demultiplexing

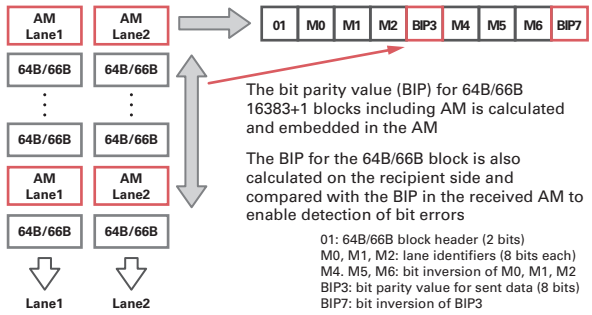

Figure 6: Relationship Between Alignment Markers and BIP

*4 100GbE uses 20 lanes.

*5 The current IEEE standard stipulates that data transmission between the interface board and the optical transceiver module uses 10 lanes, so 1, 2, 4, 5, or 10 lanes are available for physical media. In actual practice, the 4-lane 100GBASE-LR4/ER4 and the 10-lane 100GBASE-SR10, 10x10 MSA are used for physical media in 100GbE.

each lane, and sends this value. The same method is used on the recipient side to calculate the bit parity of the data received, and this is compared with the BIP value in the alignment marker to confirm whether any bit errors occurred in the 16384 blocks. The BIP function makes it possible to determine whether bit errors occurring in the transmission channel affect all parallel transmission channels, or only transmission channels carrying some of the lanes.

■ Summary

Here we have explained the features important for achieving parallel data transmission using 100GbE. We recommend that those of you who would like to learn more look over the 802.3ba document.

3.3 100GbE IX Joint Interoperability Test

On June 1 we issued a press release regarding the 100GbE IX (ISP Internet exchange point) joint interoperability test conducted by three companies*6. We also presented some of the details of this joint interoperability test at the JANOG28 Meeting on July 15*7. In this report we provide a brief summary of the information that has been released regarding the details of this joint interoperability test.

The purpose of the test was to confirm stability and verify operational issues in preparation for implementing 100GbE. We also examined interconnection issues for IX connections in the multi-vendor equipment environment. We gained the following information from this test.

1. There were no serious interconnection issues in the multi-vendor equipment environment
2. Care must be taken regarding some operational differences to conventional 10GbE
   - Be aware that the optical power for sending and receiving with 100GBASE-LR4 is strong due to the wavelength multiplexing
   - When you want to measure the optical level of each 100GBASE-LR4 wavelength, you must use a dedicated power meter or a command line interface if it is supported.
   - For many vendors there is no function for checking the BIP counter to monitor the quality of each lane
   - 100GBASE-LR4 is highly susceptible to errors from even minor contamination, so it was necessary to use a cleaning product to clean both the CFP optics and fiber ends

3.4 100GbE Optical Transceivers

■ CFP MSA and 10x10 MSA Standards

Currently, the two types of optical transceivers can be used with 100GbE. One is the CFP modules prescribed in the CFP MSA*8 in compliance with IEEE standards, and the other is the transceivers built to 10x10 MSA*9 independent standards that are not in compliance with IEEE standards. Figure 7 compares the brief architecture of these optical transceivers.

There is an obstacle to 100GBASE-LR4/ER4 CFP optical transceivers being popularized, because they include two expensive components: 4:10 lane conversion GearBox chips and 4 x 25.8Gbps high speed laser diodes. In light of this,

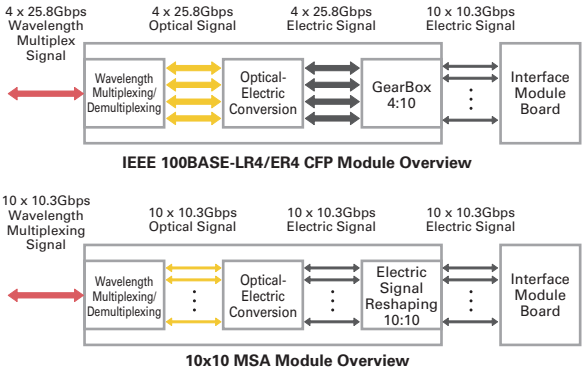

Figure 7: Comparison of 100GBASE-LR4/ER4 CFP and 10x10 MSA

*6 IIJ Press Release "Success of Industry's First High Speed 100 Gigabit Ethernet Joint Interoperability Test at IX (Internet Exchange Point)" (http://www.iij.ad.jp/en/news/pressrelease/2011/0601-02.html)

*7 JANOG28 "IX and 100 Gigabit Ethernet" (http://www.janog.gr.jp/en/index.php?JANOG28%20Programs#baef9550)

*8 An abbreviation of "C Form-factor Pluggable Multi Source Agreement." Because the IEEE does not provide detailed specifications for the configuration or functions of 100GbE optical transceiver modules, optical module vendors came together to define common specifications and manufacture products based on these. (http://www.cfp-msa.org/)

*9 Ten by Ten Multi Source Agreement. Equipment vendors and their users came together to draw up independent optical transceiver standards that make low cost production possible. (http://www.10x10msa.org/)

10x10 MSA reuses inexpensive 10GbE technology that is already produced in volume to directly convert 10 x 10.3Gbps electric signals into 10 x 10.3Gbps optical signals, creating an independent specification that reduces component costs. However, 10x10 MSA is not in compliance with IEEE standards, so it must be noted that it is only supported by some vendor equipment.

■ Usable Physical Media and Operating Distance

Table 1 shows the interface standards that can be used with 100GbE. The entries in red are optical transceivers that were usable at the time of the joint interoperability test. The three main types of physical media that are usable are copper cable, multi-core multi-mode fiber (MMF), and single mode fiber (SMF). However, for actual connection between core network equipment, the only options are 100GBASE-LR4/ER4 and 10x10 MSA, which can be used with existing single mode fiber as-is. It is also difficult to construct metro area networks using 100GbE and dark fiber because the operating distance of the optical transceivers currently usable cannot be extended very far, and at this point medium and long distance transmissions are problematic without carrier-class transmission equipment. In the future it should become possible to use 100GBASE-ER4 (30km) and 10x10-40km optical modules that enable transmission over medium distances, but considering the current price of CFP it is not clear whether this will be feasible from a cost perspective.

■ Future Transceiver Standards

Current CFP-type optical transceivers are extremely large at 78 x 13.6 x 144 (mm), and it is not possible to increase the density of ports that can be installed in an interface module board. However, in the near future CFP2 and CFP4 modules with significantly reduced size are planned to be released. After CFP2 the speed of electric signals between interface module board and CFP2/4 will be changed to 4 x 25.8Gbps (Figure 8). This is expected to remove the need for GearBox chips in CFP as with 10x10 MSA, and bring prices down below CFP.

3.5 Conclusion

At present there is very little comprehensive information regarding 100GbE in Japanese, so we hope that this report serves to aid the understanding of 100GbE. Currently 100GbE products are extremely expensive and there are few products that support it, so we believe it will still be some time before it spreads to common use. IIJ will actively engage in assessment of incorporating 100GbE into our backbone network in preparation for future increases in traffic.

	IEEE 100G	10x10 MSA 100G
Backplane 1 m	100G BASE-KR4 (Planned)	N/A
Copper Cable 5 m	100G BASE-KR4 (Planned)	N/A
Copper Cable 7 m	100G BASE-CR10	N/A
MMF (OM3) 100 m	100G BASE-SR10 100G BASE-SR4 (Planned)	N/A
SMF 2 km	100G BASE-FR4 (Planned, WDM)	10x10-2 km (WDM)
SMF 10 km	100G BASE-LR4 (WDM)	10x10-10 km (WDM)
SMF 30 km/40 km	100G BASE-ER4 (WDM)	10x10-40 km (WDM)

Table 1: List of Interfaces Defined in IEEE/10x10 MSA

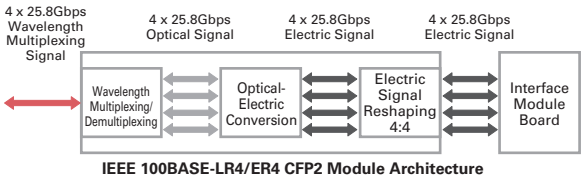

Figure 8 100GBASE-LR4/ER4 CFP2

Author:
Munenori Ohuchi
Ph.D. Ohuchi works in the Network Service Department of the IIJ Service Division. Since joining IIJ, Ph.D. Ohuchi has been engaged in the testing of equipment used in the IIJ backbone network, and the survey, research, and development of new Internet-related technologies.