Technologies for High-speed and Power-efficient Routers and Switches

Masaki Yamada, Hidehiro Toyoda, Takeki Yazaki, and Shinji Nishimura
Central Research Laboratory, Hitachi, LTD.
1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo, Japan

Abstract

We worked on the physical layer model for implementing the optical DQPSK transmission function. In order to enable a high speed DQPSK transmission, a signal multiplexing and de-multiplexing module was needed. The multiplexing and de-multiplexing module exchanges a high frequency 2bit width data signal and a low frequency wide width data signal into each other. For creating such modules, the channel lengths of each low frequency signal must be kept even. We also worked on creating an evaluation board which consists a 64B/66B encoder and decoder, a gearbox, forward-error-correction (FEC) units, and the deskewing units by using two FPGA's. By connecting four 10-Gbps Multiplexer/ Demultiplexer LSIs to this board, a transmission rate of 40 Gbps was achieved.

As a different approach, we placed Dynamic Performance Control that cuts down the power consumption according to the received traffic, as required technologies for future routers and switches. Combining the two technologies together, a high-speed and power efficient routers and switches are achievable.

1. Introduction

As a result of the ubiquitous use of the Internet and the growing data size, traffic over the Internet is growing throughout the world. The average Internet traffic observed in Japan in year 2006 was 637 Gbps. According to the statistics, the Internet traffic in Japan is growing approximately 30% year. Presuming that the traffic keeps on growing at this rate, the average Internet traffic in Japan is estimated to reach 121 Tbps in year 2025[1].

To catch up with the growing traffic, the throughputs of carrier-class routers and switches, too, are increasing rapidly. The increasing rate is approximately 60 percent per year, as shown in Figure 1. Since the number of ports of a router or switch is limited by spatial constraints, the data rate of Ethernet is growing. Accordingly, Ethernet standards of 40 and 100 Gbps are being discussed in the IEEE 802.3 Working Group.

As shown in Figure 2, the power consumption of a single router or switch is increasing at a ratio of 30% per year. This is because of market demands for cutting down the running costs. And since environmental issues and rising OPEX of network devices have become real and substantive problems, market demands for cutting down power consumption are now even stronger.

Figure 3 shows the change of power consumption against throughput for routers and switches on a yearly time scale. The power efficiency of routers and switches has been improving at a rate of 30% per year. We assume that the concentration of semiconductors have been the main factor in the improvement of power efficiency of routers and switches.

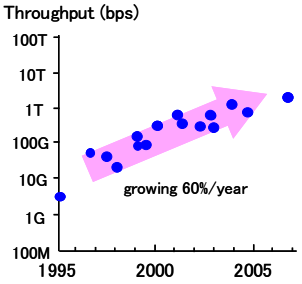

Figure 1. Throughput of routers

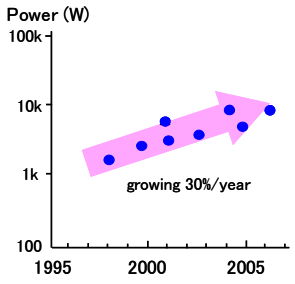

Figure 2. Power consumption of single router

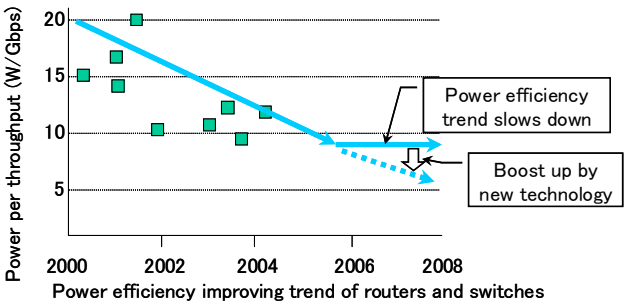

Figure 3. Power consumption per throughput

However, using a semiconductor with a shorter gate length is assumed to be less effective for improving power efficiency, today. This is mainly because of the increase of leakage current inside semiconductors. For such reasons the improving trend is assumed to slow down. To keep up with the traffic increase, we need to work on new power saving technologies such as 100-Gbps single-wavelength serial-transmission formula for 100-Gbps Ethernet and Dynamic Performance Control.

2. Current technology

In this section, we introduce high-speed Ethernet, semiconductor technology, and standby-power-cut-down technology as current technologies.

2.1 High speed Ethernet

High speed optical transmission is one answer for achieving a large throughput and power efficiency over networks. We assume that 100-Gbps-transmission-rate networks are the mainstream for next-generation optical transmission [2-5]. In the standardization of high-speed Ethernet being discussed as IEEE P802.3ba Task Force, multi-lane solutions have been suggested; in particular, either combining ten 10-Gbps lines or combining four 25-Gbps lines is said to be dominant [6, 7].

2.2 Semiconductor technology

The semiconductor process is the dominant factor of the operating frequency of the semiconductor. The improvement of semiconductor technology therefore has held some weight in regards to handling the growth of data throughput.

At the same time, semiconductor technology has contributed to the improvement of power efficiency. The power-consumption factor of a router can be separated as shown in Figure 4. LSIs such as ASIC and FPGA account for 40 to 50%, and memories such as SRAM, DRAM, and CAM account for 25 to 30% of the total power consumption. This fact shows that reducing the power consumption of the semiconductors that compose LSIs and memories has a direct effect on power saving for routers and switches.

Here we show the evolving model of LSI engines used in routers and switches as a example of semiconductor technology. To gain a higher throughput, LSI engines used inside routers and switches are multiplexed, which means the energy consumption will multiply, too. In the second step, by updating the semiconductor process, multiple engines are aggregated into a single chip. The resulting aggregated engine improves both performance and power efficiency.

For such reasons, the power efficiency increasing trend will slow down and other technologies for improving power efficiency are needed.

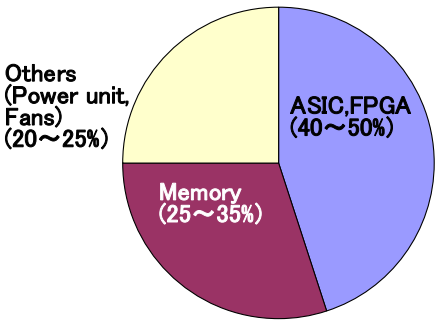

Figure 4. Power consumption rate of router

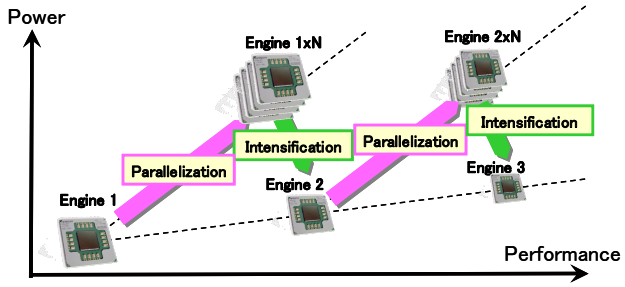

Figure 5. The evolving model of LSI engines

2.3 Standby power cut down

Standby-power cut down is a technology for reducing standby power of routers and switches by turning off the power of unused ports and line interface cards. By using this technology, the power consumption of a router or switch can be saved when traffic is small.

3 Future Requirements

3.1 Physical Upper Layer for DQPSK transmission

To gain a large throughput and to save power 100-Gbps Ethernet is now under standardization. However, the 100-Gbps standard being discussed at IEEE P802.3ba Task Force is a multi-lane solution. Multi-lane solutions need ether multiple wavelength channels or multiple fiber lines to achieve the targeted throughput. This means that the throughput of a network using multi-lane solutions will be smaller than the throughput of a network using a single-wavelength serial transmission formula, as long as the number of fiber lines constructing the network is the same. For such reasons, we believe that a demand for a 100-Gbps single-wavelength serial-transmission formula will rise in the near future, as a solution to achieve a network with larger throughput.

The existing issue for a 100-Gbps single-wavelength serial-transmission formula is the difficulty of developing an optical 100-Gbps OOK ("on-off keying") modulator, which is almost impossible by current technology. This is because multiple issues exist to realize such a modular. The laser emission will be too difficult, the power consumption will be too large, and the manufacturing costs will be too high.

One plan for realization is to use optical multi-level signaling. By increasing the information bits per symbol from one to two or more, optical components with lower operating frequency and power consumption can be used. As for recent studies, a 100-Gbps transmission using an optical Quadrature Phase-Shift Keying (QPSK) has been reported [8, 9].

In our group, we have been working on the physical layer for next-generation optical multi-level transmission. Figure 6 is a model of the physical layer of a 100-Gbps DQPSK (differential QPSK) optical transmission. For DQPSK optical transmission, an interface with 2-bit-width data per symbol at 50 GHz is needed. At the same time, a very fast differential-encoding circuit is needed. Accordingly, in regards to our proposed physical-layer model, the differential encoding inside the forwarding unit is handled by a low-frequency logical circuit, called a "differential precoder". Moreover, by multiplexing the output data of the differential precoder as, for example, a 512-bit-width data at 200 MHz, a 2-bit-width data per symbol at 50 GHz can be achieved.

When differential encoding is done, each symbol of the signal handed over, is encoded continuously against time. The time continuity of the encoded signals must be kept until it is handed over to the DQPSK transmitter. This is because each symbol of the signals that are demodulated at the DQPSK receiver also needs to be continuous against time. For such reasons, the continuity of the signals at the N:2 multiplexer must be kept, when the incoming signal of the differential precoder is multiplexed. The N:2 multiplexer multiplexes a large-width data, for example, a data with a width of 512 bits (N=512), into a 2-bit-width data. The continuity of the signals is kept by adjusting the channel length of all the channels between the differential precoder and multiplexer at a same length. Since there is a technical assignment to realize this large-width and high speed (up to 50Gbps) Multiplexer by a single LSI, it is natural to implement by using multistage Multiplexers. However, for connecting multistage Multiplexers, there is an issue that it is nearly impossible to step down the skews

between channels, because 10Gbps class signals need to be handled between Multiplexers and the latitude of the board trace is heavily restricted. To solve such problems, we placed a pre-deskew circuit which relaxes the limit of the board trace caused by the skew and cancels the skew ahead.

On the other hand, a skew depending on the channel length may occur when the 2-bit-width data per symbol (which is handed over from the DQPSK receiver) at 50-GHz frequency is demultiplexed. However, as for low-frequency data, for example, a 512-bit-width data at 200 MHz, a complex deskewing is possible. In our proposed model, we suppose the deskewing method shown in reference [2] to be used.

We worked on an implementation test of our physical-layer model at a transmission rate of 40 Gbps. Using two FPGAs, one for the transmitter and the other for the receiver, we built the PMA sublayer (physical medium attachment), consisting of a 64B/66B encoder and decoder, a gearbox, forward-error-correction (FEC) units, and the deskewing units. The transmitter FPGA hands over parallel data to the four 10-Gbps Multiplexer/ Demultiplexer LSIs, and four 10-Gbps signals are multiplexed by the 40-Gbps multiplexer. The multiplexed data (at 40 Gbps) is handed over to the PMD sublayer (physical medium dependent). We made an evaluation board by arranging the FPGAs and Multiplexer/ Demultiplexer LSIs as shown in Figure 7.

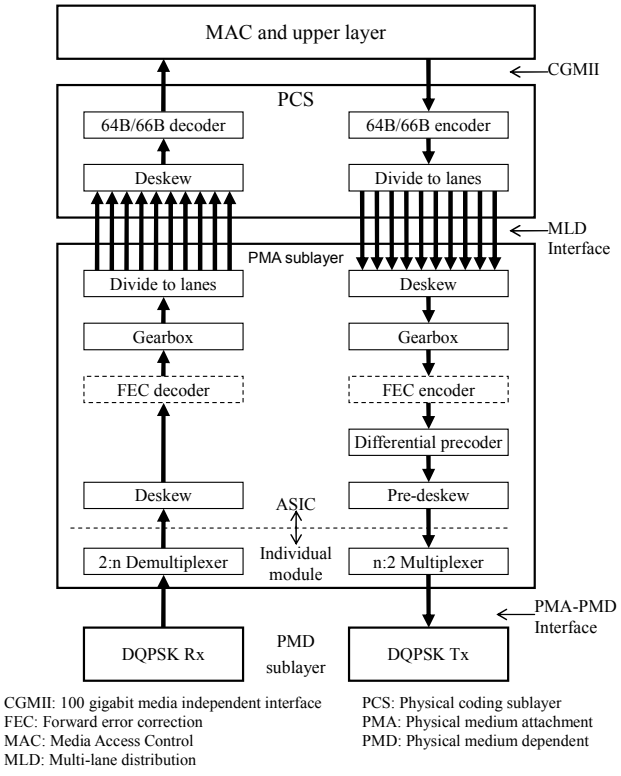

Figure 6. 100 GB/s Ethernet physical layer on DQPSK transmission

Figure 8 is a block diagram of the evaluation board. We have omitted the MLD interface in order to allocate the PCS and the PMA sublayer on a single FPGA. Additionally we have omitted the deferential precoder, too, because the DQPSK PMD is not connectable for the mean while.

Using this evaluation board, we performed a test by inputting Ethernet frames from the upper layer and looping back the signals at the PMA-PMD interface. The test showed that the time continuity of the signals was kept by the pre-deskew unit, and the Ethernet frames were successfully received at the upper layer interface.

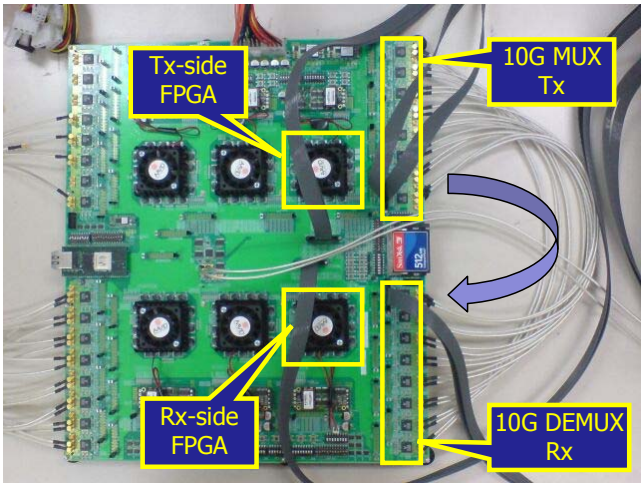

Figure 7. Loop back test performed on evaluation board

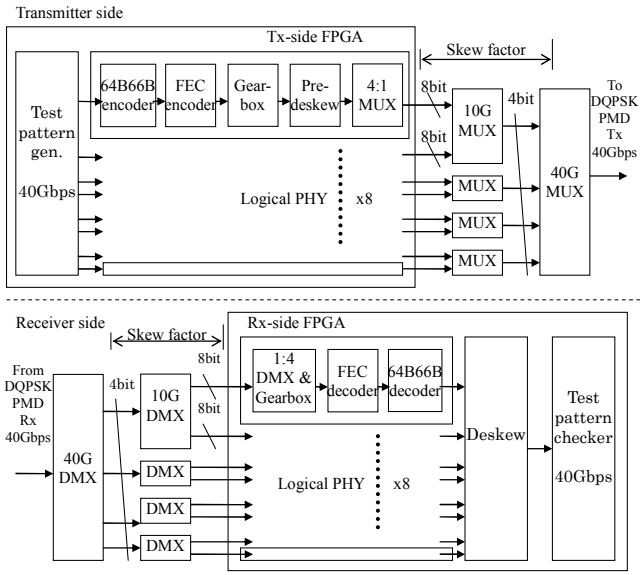

Figure 8. Logical block diagram for evaluation

3.2 Dynamic Performance Control

Dynamic Performance Control is a technology to save power by changing the performance of routers and switches in proportion to the amount of traffic. By using this technology, the power that had been cast away when traffic was small, will be saved.

The left figure in figure 9 shows a model of a fluctuating traffic for a network in a day, and the performance of a router handling such traffic. The router will change its performance according to the traffic, but as the change will take some time after the actual increase or decrease of traffic, the performance is set to be able to handle a slightly larger traffic.

The figure on the right shows the comparison of the power consumption between current and suggesting routers. As shown, the power consumption of an existing router or switch stays constant with no relation with the

traffic volume. By using Dynamic Performance Control, the power consumption will be decreased to a amount in relation with the performance.

The technical issue which is being handled now is to choose an appropriate performance to avoid or to minimize quality of service degrades. To overcome this issue, an accurate traffic forecasting and a fast and precise performance control are needed.

Each management technology handles different targets. The “Energy Efficient Ethernet” (EEE) is a new standard under discussion in IEEE 802.3 Ethernet Working Group [10]. By changing the PHY speed according to the traffic, from 10Base-T to 10GBase-T, this technology will save power at the Ethernet interface.

As for an additional usage of the Dynamic Performance Control, Link Aggregation which dynamically controls the number of links under operation will be capable. Figure 10 shows the model of the Link Aggregation. By reducing the number of active links belonging to the Link Aggregation group when traffic is small, a power efficient management is enabled. [11].

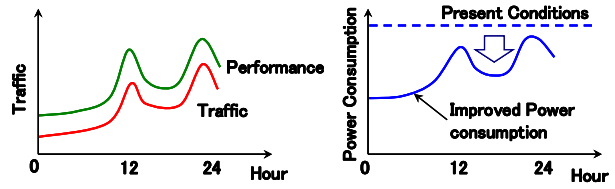

Figure 9. Effect of Dynamic Performance Control

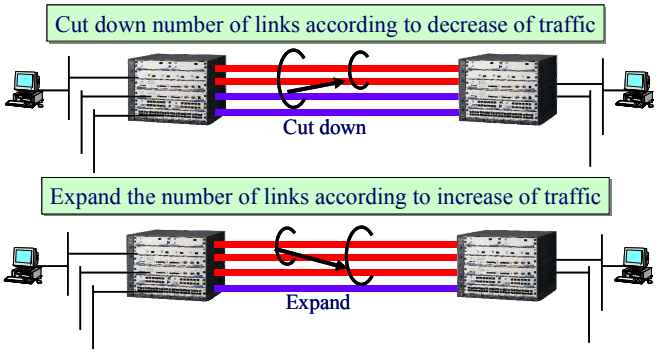

Figure 10. Performance control of Link Aggregation

4 Conclusions

As traffic has increased, improving throughputs and saving power have become essential for routers and switches. For handling the next-generation standards of Ethernet, we targeted single-wavelength serial 100-Gbps optical transmission using DQPSK as a solution. We focused on the physical layers of the I/O interface to enable DQPSK under such a high data rate. By placing a differential precoder and adjusting the channel length between the multiplexer and the encoder, high-speed DQPSK will be possible. Using FPGAs, we then focused on a test implementation of a physical-layer model for DQPSK. Although the test device was a rather simplified model, a transmission speed of 40 Gbps was achieved.

To reduce power consumption, since semiconductor technology will become less effective in saving power in future technological generations, we need new technologies. We placed Dynamic Performance Control, which controls power consumption according to the incoming traffic, as required technologies for the future.

We believe that by combining these technologies, high-speed power-efficient routers and switches will be achieved.

Acknowledgements

Part of this work was supported by "Lambda Access Technologies" project, National Institute of Information and Communications Technology (NICT), Japan.

References

[1] T. Hoshino, Plenary Talk, Green IT Symposium 2007, Japan, August 4 2007.

[2] H. Toyoda, H. Nishi, and S. Nishimura, “Signal Transmission and Coding Architecture for Next Generation Ethernet,” IEICE Transaction on Information and Systems, Special Issue on New Technologies in the Internet and their Applications, E86-D, No. 11, pp. 2317-2324, 2004.

[3] H. Toyoda, S. Nishimura, M. Okuno, R. Yamaoka, and H. Nishi, “A 100-Gb-Ethernet subsystem for next-generation metro-area network,” ICC 2005 in Seoul (IEEE International Conference on Communications), GC10-3, May 2005.

[4] H. Toyoda, S. Nishimura, M. Okuno, K. Fukuda, K Nakahara, and H. Nishi, “100-Gb/s physical-layer architecture for next-generation Ethernet,” IEICE Transactions on Communications, Special Section on the Next Generation Ethernet Technologies, Vol. E89-B No. 3, pp. 696-703, Mar. 2006.

[5] H. Toyoda, S. Nishimura, M. Okuno, and M. Terada, “A 100-Gb/s-physical-layer architecture for higher-speed Ethernet for VSR and backplane applications,” IEICE Trans. Electron., Vol.E90-C, No.10, pp.1957-1963, October 2007.

[6] S. Nishimura, H. Toyoda, and M. Shishikura, “PMD architecture with skew compensation mechanism for parallel link,” IEEE 802.3 Higher Speed Study Group, Nov. 2006. http://www.ieee802.org/3/hssg/public/nov06/nishimura_01_1106.pdf

[7] M. Gustlin, “100GE and 40GE PCS and MLD Proposal,” IEEE, IEEE P802.3ba, January 2008. http://grouper.ieee.org/groups/802/3/ba/public/jan08/gustlin_01_0108.pdf

[8] P. J. Winzer, G. Raybon, and M. Duelk, “107-Gb/s Optical ETDM Transmitter for 100G Ethernet Transport,” ECOC 2005, Th4.1.1, Bell Laboratories, Lucent Technologies, September 2005.

[9] M. Daikoku, I. Morita, H. Taga, H. Tanaka, T. Kawanishi, T. Sakamoto, T. Miyazaki, and T. Fujita “100Gbit/s DQPSK Transmission Experiment without OTDM for 100G Ethernet Transport,” OFC/NFOEC 2006, PDP36, March 2006.

[10] H. Barass, M. Bennet, W. Diab, D. Law, B. Nordman, and G. Zimmerman “Energy Efficient Ethernet: An Overview” IEEE, IEEE P802.3az, 16 July 2007. http://www.ieee802.org/802_tutorials/july07/IEEE-tutorial-energy-efficient-ethernet.pdf

[11] S. Aibara, Y. Fukuda, K. Kawahara, Y. Oie, “Power Saving Architecture with Link Aggregation on Ethernet Switch”, Technical Report of IEICE, IN2006-147, pp.55-60, January 2007.