Ultra-high-speed Interconnect Technology for Processor Communication

● Yoshiyasu Doi ● Samir Parikh ● Yuki Ogata ● Yoichi Koyanagi

In order to improve the performance of storage systems and servers that make up the cloud, it is essential to have high-bandwidth interconnects that connect systems. Fujitsu has already marketed a CMOS high-speed interconnect product that works between CPUs for the UNIX server (SPARC M10). Its data transfer rate per signal line is 14.5 Gb/s. Further, Fujitsu is currently engaged in studies of a CMOS interconnect that can operate at over 32 Gb/s per signal line to achieve a higher bandwidth interconnect. To realize higher speed, we developed a high-speed interconnect with a loss-compensation performance higher than 30 dB at 32 Gb/s by integrating a series of features: a communication method requiring no high-frequency operation in a transmitter circuit, a wideband loss-compensation equalizer circuit in the receiver circuit, and a data-receiving method that does not require generation of a highly accurate sample clock. When these technologies are implemented in a CPU chip, it is possible to improve the entire performance of server systems to a level more than twice to a level more than twice as high as the current one. This paper introduces the features of these ultra-high-speed interconnect technologies for 32-Gb/s serial interconnects implemented by using 28-nm CMOS technology.

1. Introduction

In expectation of an improvement of data centers that support cloud computing, further enhancement of the data processing performance of servers is needed. Against this background, the amount of data that can be processed with a single chip has been drastically increased by minimizing the semiconductor process on a yearly basis and adopting a multi-core CPU structure. The data being processed in a CPU are transmitted to and received from external units via an I/O circuit inside the chip. However, because the number of signal lines in a chip is limited, it is difficult to improve the data transfer rate by using a bus. Therefore, a high-speed interconnect with an extremely high communication capacity per signal line (higher than 10 Gb/s) is needed [**Figure 1 (a)**]. The high-speed interconnect is used not only to establish communication among CPU chips but also to establish connection among CPUs, memories and storages virtualized by resource pool architectures [**Figure 1 (b)**]. Its application range is expanding. While it is said that the required data rate for data communication will be doubled every three years, the performance improvement in CMOS technology remains as slow as approx. 5% annually, indicating that the discrepancy between required performance and current performance is growing every year. Fujitsu has already developed and commercialized a CMOS high-speed interconnect that enables data transmission at 14.5 Gb/s per signal line in communication between CPUs in the UNIX Server "SPARC M10".[1)] However, further R&D is needed, and the bandwidth required for next-generation servers is expected to be more than twice.

2. Overview of and challenges in developing high-speed interconnects

A high-speed interconnect connecting the systems is packaged together with CPU chips in order to carry out data communication among the CPU chips. For instance, with regard to the SPARC64 X, multiple CPU cores are arranged on a chip, and a high-speed interconnect macro is arranged on each side (right and

left) of the chip to ensure data input to and output from the CPUs [Figure 1 (a)]. The high-speed interconnect is comprised of a transmitter circuit and a receiver circuit. The transmitter circuit transfers the internal data generated by the CPU, etc. to a data transmission channel after conducting a parallel-serial conversion of the input data. The receiver circuit transfers 1/0 signals to the CPU again as internal data after converting the input data through serial-parallel converter [**Figure 1 (c)**]. The transmitter circuit and receiver circuit are connected via transmission channels comprised of packages, print circuit boards and cables, etc. Because each transmission channel has a specific loss depending on its material and geometry, degradation of the signal quality is more serious when the frequency component of the transmitted signal is higher. For this reason, the waveform of 1/0 signals output from a transmitter circuit will degrade until the signals reach the receiver circuit, which makes it difficult to determine the 1/0 of the data concerned. This phenomenon is called inter symbol interference (ISI). Further, because the transmitter circuit and receiver circuit are operated synchronously to a local clock source respectively, each circuit has its timing error depending on the frequency difference and fluctuation specific to each clock source. Therefore, the receiver circuit has a loss-compensation feature that compensates the signal degradation, and a feature to establish a synchronous operation between the receiver circuit and transmitter circuit (CDR: Clock and Data Recovery). It does this by detecting the timing information of the transmitter circuit from the receiver circuit input data and adjusting the receiver's timing.

To further increase the data transfer rate in future (e.g., up to 32 Gb/s), changing to high-speed transmitter and receiver circuits is not the only challenge to

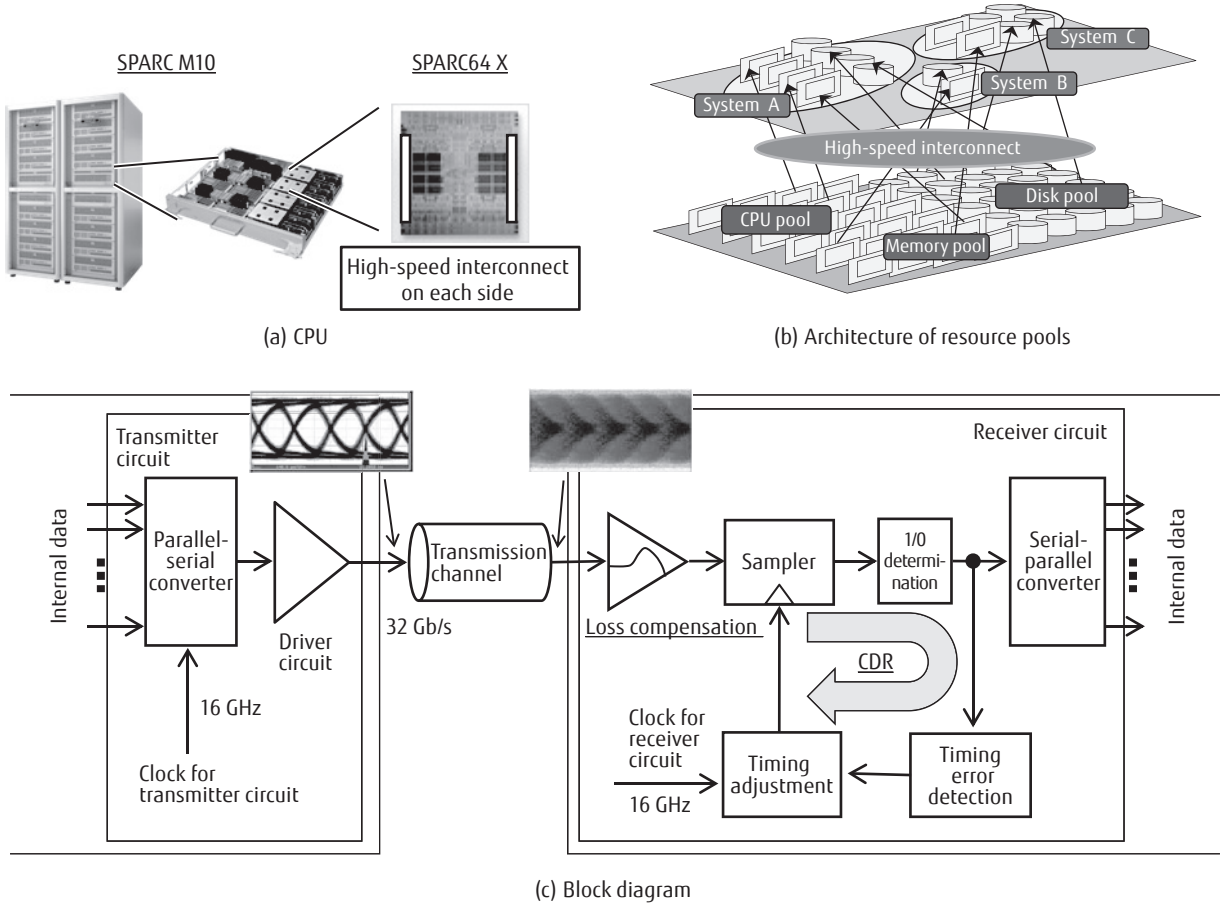


Figure 1
Overview of high-speed interconnect.

overcome. Because severe degradation of the signal quality associated with transmission channels, such as electric wiring on a print circuit board, poses a serious issue, it is imperative to improve the performance of a loss-compensation circuit so as to compensate for the degradation. Further, in proportion to the data transmission speed, the accuracy of synchronization timing for the receiver circuit should be also enhanced. For instance, to realize a 32-Gb/s interconnect, an adjustment accuracy of less than 1 ps (10^{-12} sec) is needed.

In this report, three types of CMOS high-speed interconnect technologies are introduced that realize 32 Gb/s in next-generation communication.

3. Data interleaved driver circuit

The transmitter circuit converts low-speed parallel data outputs from a CPU into a single high-speed serial signal and transmits it to a receiver using a driver circuit. To generate 32-Gb/s serial data, the 2:1 parallel-serial converter of the final phase should be operated by a high-speed clock of 16 GHz. Namely, the circuit requires the highest speed of operation among the transmitter circuits. Due to the difficulty of implementing a 2:1 parallel-serial converter, this stage makes it hard to achieve high-speed transmission on the entire transmitter circuit. Further, because a circuit consumes more electric power when it operates at higher speed, the power consumption of the 2:1 parallel-serial converter accounts for a high percentage of the total power consumption of the transmitter circuit.

To address this issue, the 2:1 parallel-serial converter has been removed from conventional transmitter circuits and a new data interleave method has been developed. In this method, two output signals of 4:2 parallel-serial converter that are shifted by half a cycle are added to the voltage direction via a driver circuit (**Figure 2**).[2] While the conventional method involves a signal compression in the time direction, the data interleave method involves signal compression in the voltage (vertical) direction. Such a signal compression method is called amplitude modulation (AM) in general.

In the conventional method, the 4:2 parallel-serial converter that generates input signals for the 2:1 parallel-serial converter will output two signals based on synchronous timing. On the other hand, the 4:2 parallel-serial converter output generates two signals that are shifted by half a cycle in this method. A signal created by synthesizing the output from each driver circuit that uses the aforementioned two signals is output to outside of the chip. This output signal is transferred via a transmission channel as a ternary signal and reaches the receiver circuit, because it is created by synthesizing two signals in the voltage direction. The counterpart receiving circuit normally contains a decision feedback equalizer (DFE) as a function to compensate for the signal loss. Because the transmitted ternary signals are those under the influence of mutual interference of adjacent 1/0 signals at an equivalent power rate (1:1), they can be considered equivalent to the ISI caused by signal loss generated in the transmission channel. Because the main function of DFE is to eliminate ISI associated with the loss generated in the transmission channel per se, it is possible to eliminate the superimposed adjacent signal components from the signals by regarding these ternary signals as ISI. Thus, it is also possible to recover the original binary signals.

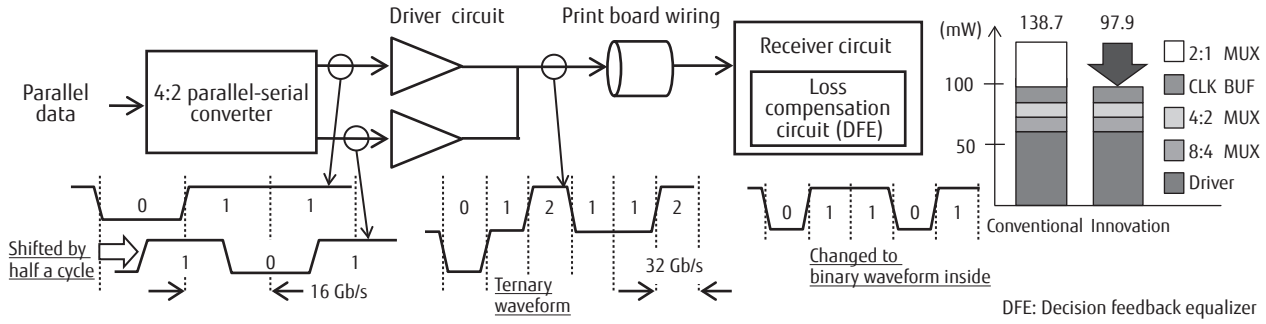


Figure 2
Interleave driver circuit.

By using this technology, an innovative transmitter circuit approach has been realized that does not use any 2:1 parallel-serial converter requiring high-speed operation. As a consequence, high-speed operation at 32 Gb/s and a reduction of power consumption to 70% versus the level in the conventional method could be achieved.

4. Data interpolator circuit

The CDR of the receiver circuit is a feature to correctly determine 1/0 of the input signal by detecting the timing error in the time direction of the input signals and correcting the timing error. Because there are fine frequency offset and frequency fluctuations between the transmitter circuit and receiver circuit, the sampling timing is continuously changed to determine 1/0 signals. Therefore, it is imperative to have a CDR function integrated in the receiver circuit. In the conventional method, the timing errors are detected by using the data which are determined as 1/0, and the timing of the clock used for input signal sampling is adjusted automatically. The timing errors are detected by comparing the data (D[n], D[n+1]) and transition (B[n]) between such data. For instance, if D[n]=0, D[n+1]=1 and B[n]=0, a judgment is made that the sample clock is faster than the input data. Therefore, an adjustment is made to retard the phase of the sample clock. In general, 1/0 determination is made in this way by using the phase-adjusted sample clock and sampling the signal amplitude at the data center. However, a more accurate adjustment circuit is still needed, because the time per bit is shorted to 1/2 in each generation with the growing request for bandwidth. For instance, with regard to 1/0 in 32 Gb/s, the time for 1 bit is 31.25 ps, where a 6-bit resolution (less than 0.5 ps) is requested for a single bit as an adjustment accuracy of the sample clock. Because it is necessary to realize both high-speed operation and high-accuracy timing adjustment in the sample clock adjustment circuit, it makes it hard to increase the speed of the entire receiver circuit.

The CDR[3] using a data interpolator circuit in the current development eliminates the need for a function to adjust sample clock phase, because it is operated based on a sample clock that is non-synchronous with the sample clock in the transmitter circuit (**Figure 3**). The data interpolator circuit is comprised of a data sampler (gm) and a switched capacitor based on variable capacitance. The data sampler is driven non-synchronously from the transmitter circuit. In this approach, sampled voltage is regenerated through charge-sharing interpolation of the voltage level at the center of the data by using the switched capacitor on which the ratio of the variable capacity is controlled. "Charge-sharing" in this context represents the sharing of electric charge Q accumulated in each of the multiple capacities by the total capacity of these multiple capacities. For instance, if each of two voltages (V_1, V_2) are applied to each of two individual capacities (C_1, C_2), the accumulated electric charges (Q_1, Q_2) are expressed as $Q_1 = C_1V_1$ and $Q_2 = C_2V_2$ respectively. If these two capacities are connected, the entire electric charge Q is expressed as $Q = Q_1 + Q_2$, and the total capacity is expressed as C_1+C_2. Therefore, the voltage V of the capacity newly generated after the connection is expressed by $V= (C_1V_1+C_2V_2) / (C_1+C_2)$. The capacity ratio for the charge-sharing can be also controlled from the information on the data center and the data transition in the same way as with the adjustment of sample clock timing. The resolution necessary for

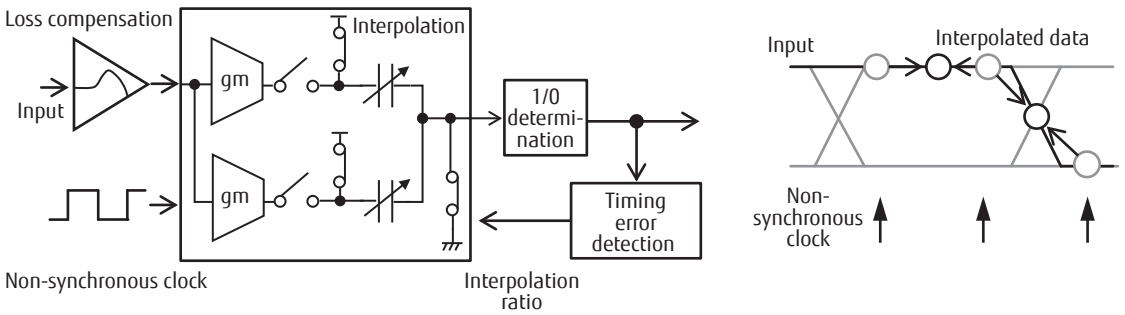


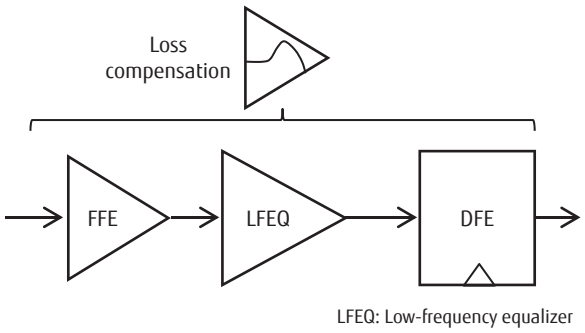
Figure 3
Data interpolator circuit.

voltage interpolation by switched capacity is 6 bits per 1UI. While this accuracy is the same as that requested for the clock timing adjustment, the resolution level is approx. 3 mV after conversion into voltage level. This resolution level is regarded as "not particularly demanding" in the area of the analog–digital converters (ADC) based on similar technology. Therefore, it is easy to realize high accuracy in the voltage direction.

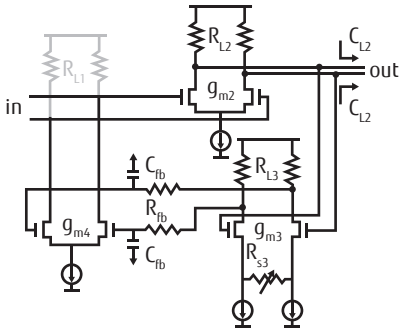
By using this technology, the need to generate a high-speed and high-accuracy clock has been eliminated in the receiver circuit, resulting in further acceleration of the data transfer rate to higher than 32 Gb/s.

5. Wide bandwidth loss compensation

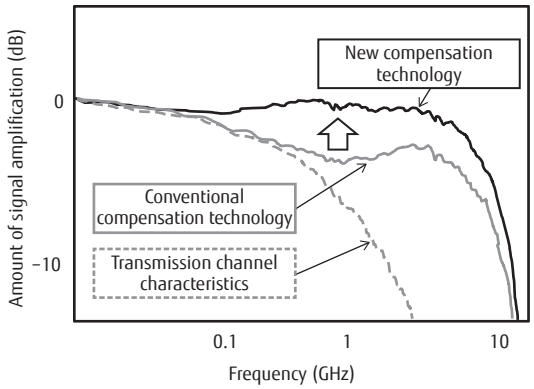
To eliminate ISI generated by the transmission channel and to reproduce the signal quality necessary for data reception, the receiver circuit integrates a series of equalizing circuits such as a feed forward equalizer (FFE) and DFE. However, the loss in the transmission channel increases when the frequency becomes higher, and this poses a serious challenge in realizing a high-speed signal system. Further, due to increase in frequency spectrum associated with high-speed signal transmission and reception, it has become difficult to neglect the loss associated with the signal components that have been considered negligible so far because of their low frequency nature. The gray solid lines in **Figure 4 (c)** represent the amount of signal amplification achievable with the conventional compensation technology, assuming a data transfer rate of 32 Gb/s. There is no loss when the amount of signal amplification is 0, indicating high signal quality. With the increase of value in the negative direction, the loss will grow and the degradation of the signal quality will become worse, which makes it more difficult to determine 1/0 on the data frame. Our analysis revealed that the loss compensation achieved



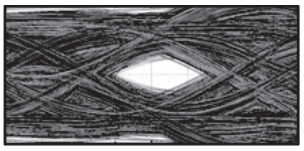
(a) New loss-compensation technology



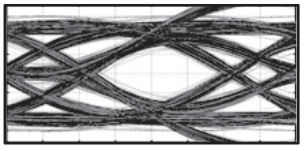
(b) LFEQ circuit diagram



(c) Frequency characteristics



Conventional loss-compensation technology applied



New loss-compensation technology applied

(d) Improvement effect

Figure 4
Wide bandwidth loss-compensation circuit.

by only restoring the desired signal components in high frequency is insufficient, because the distortion caused by roll-off of the amount of signal amplification from the low frequency range to 1 GHz has a significant impact on the degradation of the signal quality in the receiving circuit. In our data transmission and reception approach, the amplitude level in the section immediately before DFE (from the transmitter circuit to FFE and low-frequency equalizer (LFEQ) via the transmission channel) should remain flat to the frequency level equivalent to 1/4 of the data transfer rate. In our innovative compensation technology, a solution has been developed to carry out not only the conventional loss compensation in a high-frequency range but also loss compensation in a low-frequency range by using the LFEQ [Figure 3 (a)]. **Figure 4 (b)** shows the actual scheme of the LFEQ circuit. Based on this technology, loss compensation in the transmission channel across a wide frequency range has become possible, from a low-frequency to a high-frequency range [black solid line in Figure 4 (c)].[4] **Figure 4 (d)** shows a comparison of the signal quality of the conventional technology and new technology based on a simulation of the time range. The improvement in signal quality has been almost doubled compared to the improvement achieved with the conventional approach.

By introducing this technology, high-speed signal transmission has become feasible for a distance of 80 cm on a backplane required for inter-server processer communication while satisfying the error rate requirement (10^{-12}) in the evaluation using an actual 32-Gb/s system. **Figure 5** indicates a transmitter and receiver circuit chip integrating the above-explained three high-speed interconnect technologies implemented by using the 28-nm CMOS standard process.

6. Conclusion

In this report, we introduced our 32-Gb/s ultra-high-speed CMOS interconnect technologies for realizing high-speed data transmission between servers or CPUs. By designing a loss-compensation circuit to realize loss compensation in both high- and low-frequency ranges, large-capacity data transmission has become feasible even in the transmission

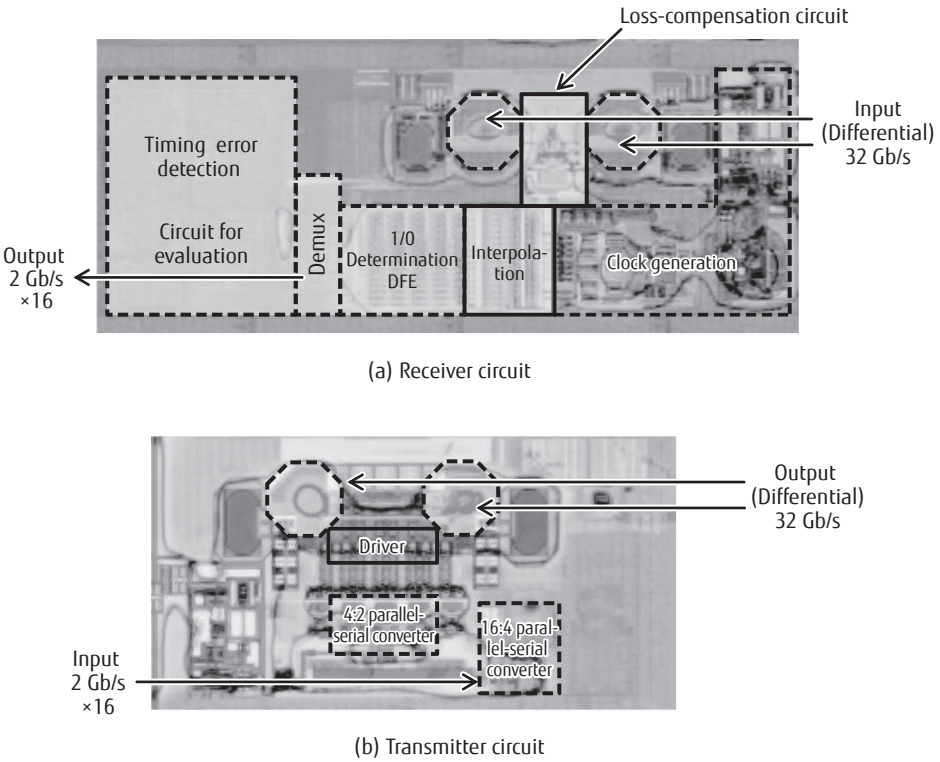


Figure 5
Photos of chips.

channels using a print circuit board which are larger than 80 cm. Further, by integrating a data interleaved driver circuit and data interpolator circuit, the need for a high-speed and high-accuracy element circuit has been eliminated. This makes it possible to reduce power consumption and allows server system designs to be made more freely. By implementing high-speed interconnect technologies in a CPU, the data transfer rate between CPUs can be improved and, as a consequence, the entire performance of a server system can be improved to a level twice as high as that achieved by conventional technologies. We are committed to further improving the performance of next-generation servers and super-computers.

References

1) R. Kan et al.: A 10th Generation 16-Core SPARC64 Processor for Mission-Critical UNIX Server. ISSCC Dig. Tech. Paper, pp. 60–61, Feb. 2013.
2) Y. Ogata et al.: 32Gb/s 28nm CMOS Time-Interleaved Transmitter Compatible with NRZ Receiver with DFE. ISSCC Dig. Tech Paper, pp. 40–41, Feb. 2013.
3) Y. Doi et al.: 32Gb/s Data-Interpolator Receiver with 2-Tap DFE in 28nm CMOS. ISSCC Dig. Tech. Paper, pp. 36–37, Feb. 2013.
4) S. Parikh et al.: A 32Gb/s Wireline Receiver with a Low-Frequency Equalizer, CTLE and 2-Tap DFE in 28nm CMOS. ISSCC Dig. Tech. Paper, pp. 28–29, Feb. 2013.

Yoshiyasu Doi
Fujitsu Laboratories Ltd.
Mr. Doi is currently engaged in studies related to high-speed interconnects for server systems.

Yuki Ogata
Fujitsu Laboratories Ltd.
Mr. Ogata is currently engaged in studies related to high-speed interconnects for server systems.

Samir Parikh
Fujitsu Laboratories of America Inc.
Mr. Parikh is currently engaged in studies related to high-speed interconnects for server systems.

Yoichi Koyanagi
Fujitsu Laboratories Ltd.
Mr. Koyanagi is currently engaged in studies related to high-speed interconnects for server systems.