A Chip-Level Optical Interconnect for CPU

Qinfen Hao, Mengyuan Qin, Nan Qi, *Member, IEEE*, Haiyun Xue, Meng Han, Xiaolin Li, Kai Hao, Xingmao Niu, Limin Xiao, Dongrui Fan, and Kazuhiko Kurata, *Member, IEEE*

***Abstract*— With the rapid growth of electronic chip's performance, electric signal limits chip I/O in power consumption, reachability, and signal quality. In this letter, we propose a new chip-let architecture for chip-level optical interconnect. An optic I/O chip-let including an ultra-small optic transceiver and electronic components is implemented. Our analysis shows that the optical interconnect based on this architecture can achieve 1/3 power consumption and 1/2 area compared with traditional board-level optical interconnect in Ethernet NIC application. By adopting the architecture we proposed, the optic I/O chip-let can support any payload IC such as CPU, GPU, switch to have optic I/O.**

***Index Terms*— Optical interconnections, digital integrated circuits, very high speed integrated circuits, chip scale packaging, system integration.**

I. Introduction

IMPLEMENTING an optical interconnect (OI) on a chip level is better than on a board level or a chassis level. The trace on the print circuit board (PCB) will be shorter, contributing to lower system power consumption, module size, and manufacturing cost [2].

Many letters discussed chip-level OI [3], [4], [6], [8], but few of them take a system point of view, such as integrating an optical transceiver with electronic chips to build a complete solution.

Recently, several complete chip-level OI works have been published. In [5] and its later work, a chip-level OI design for FPGA chips is implemented. Intel [11] developed an optical transceiver chip to be integrated with its switch chip. These works are different in most aspects, so it is challenging to share technology among different applications, which hinders the development of optical interconnects in the long run.

Manuscript received January 14, 2021; revised May 15, 2021; accepted May 20, 2021. Date of publication May 31, 2021; date of current version July 29, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB2203004 and in part by the Beijing Science and Technology Program under Grant Z191100004819006. *(Corresponding author: Qinfen Hao.)*

Qinfen Hao, Mengyuan Qin, Xiaolin Li, and Dongrui Fan are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: haoqinfen@ict.ac.cn).

Nan Qi and Kai Hao are with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China (e-mail: qinan@semi.ac.cn).

Haiyun Xue and Xingmao Niu are with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China (e-mail: xuehaiyun@ime.ac.cn).

Meng Han and Limin Xiao are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China.

Kazuhiko Kurata is with AIO Core Company Ltd., Tokyo 112-0014, Japan (e-mail: k-kurata@aiocore.com).

Digital Object Identifier 10.1109/LPT.2021.3084945

In this letter, we proposed a chip-let architecture for chip-level optical interconnect and developed an optic I/O chip-let including 100Gbps physical layer circuit and an ultra-small optical transceiver. The optic I/O chip-let can be used to integrate with any payload IC such as CPU, GPU, and switches.

II. The Architecture

Nowadays, ISO/IEC OSI 7-layer communication protocol model is widely used by the industry. There is a physical (PHY) layer which is divided into three sublayers: a physical coding sublayer (PCS), a physical medium attachment sublayer (PMA), and a physical medium dependent sublayer (PMD) [1]. When physical media is optic, the PMD will be implemented as an optical module. In a board-level OI application, PMD is separated from PMA and PCS. However, in chip-level OI application, all the components are very close to each other, so they are integrated very tightly. In [5] and [11], the PMA is integrated with TIA/Driver. In [12], the CPU design integrates PMA and TIA/Driver circuit. Although these designs improve the integration density, they are too specific to provide a general architecture for the future development of a chip-level OI.

A. A Chip-Let Architecture for Chip-Level OI

We proposed a general chip-let architecture for chip-level OI application, composed of a payload chip-let and an optic I/O chip-let, as shown in Figure 1. Considering that the payload IC (Integrated Circuit) is sensitive to the foundry process while the PMA and the PMD are not, PMA and PMD are integrated as an optic I/O chip-let, the connection between payload IC and optic I/O chip-let are standardized to achieve decoupled development. Therefore, payload IC can benefit more from the advanced CMOS process, while PMA and PMD can stay on an old process to keep low cost. For example, CPU can be taped out in 28nm CMOS while PMA can be taped out in the 40nm CMOS process. PMD may use process older or non-CMOS process.

B. Design of a PCS Sublayer and Internal Interface

To make two chip-lets develop independently, we need to define an interface between them. We choose the PCS sublayer to achieve this target because there is no exact definition between PCS and PMA in standards such as Ethernet so that the change will have no impact on the existing application.

Because 6.25Gbps is close to the limitation of single-ended signal, also power-efficient and mature for differential signal, we choose to implement a 16 × 6.25Gbps date rate

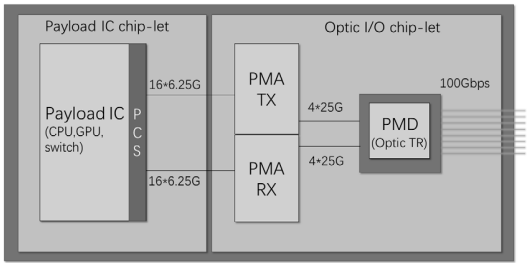


Fig. 1. A chip-let architecture for chip-level OI application.

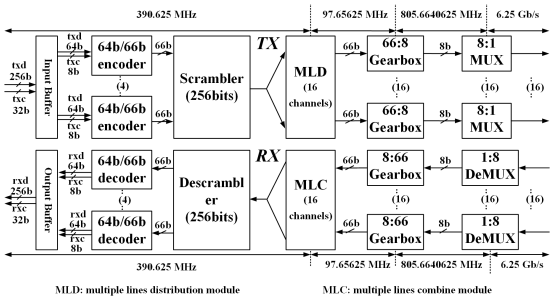


Fig. 2. Block diagram of PCS sublayer.

between PCS and PMA. It will act as an interface between the payload IC such as CPU, and the optic I/O chip-let. If an application focuses on low latency, it can choose to use the parallel single-ended signal to implement 6.25Gbps I/O between PCS and PMA. If not, SerDes is a good candidate for such speed I/O.

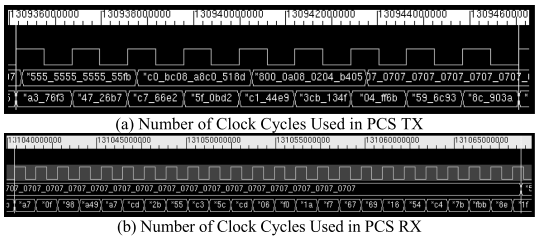
We designed a PCS sublayer and connected it to the MAC layer of the payload IC. The internal structure of the PCS sublayer is shown in Figure 2.

To improve PCS's performance, we use several design approaches. By implementing parallel detection, PCS RX can distinguish all possible positions of the 2-bit synchronize header faster. A synchronous buffer is added in each lane to remove the time skew. An I/O buffer has been implemented to dynamically adjust idle characters to keep the data rate after the alignment marker is inserted at the interval of 16383 blocks. A verification module has been implemented in PCS RX to detect errors earlier so RX can reduce time spent on error handling.

Figure 3 is the simulation results by using VCS software from Synopsis corporation, which show the TX and RX can handle data at 100Gbps in 9 clock cycles (23.04 ns) and 23 clock cycles (58.88 ns), respectively. The power consumption of PCS is 42.5207mW, which is shown in Figure 12a.

To test PCS we developed, we taped it with a CPU design in 28nm CMOS process. We added a pin to the chip to measure its bandwidth. The pin is designed to receive a pulse signal after a MAC frame is being processed. When the chip starts to work, we measure the pin with an oscilloscope and get four pulses during every 500ns interval. Therefore the bandwidth calculation is as following:

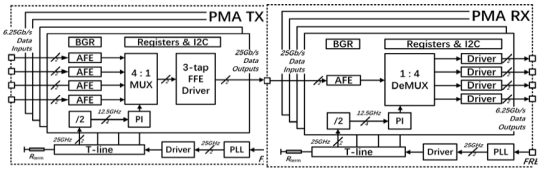
$$Bandwidth = \frac{4 \times (1512 + 20) \times 8 \times \frac{66}{64}}{500} \quad (bits/ns)$$



(a) Number of Clock Cycles Used in PCS TX

(b) Number of Clock Cycles Used in PCS RX

Fig. 3. Simulation result of PCS circuit.



(a) PMA Transmitter (TX) (b) PMA Receiver (RX)

Fig. 4. Block diagrams of the PMA schematic.

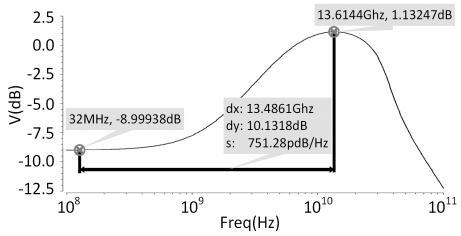


Fig. 5. Simulation result of RX AFE.

where 1512 is the length of the frame we send in bytes, and 20 is the header size of the frame in bytes. 8 means there are 8 bits in a byte, and 66/64 means we add a 2-bit synchronize header for every 64 bit, which is the requirement of Ethernet protocol standard. The calculation result of this formula shows that the bandwidth of this design can be up to 101.112Gbps.

III. Design of an Optic I/O Chip-Let

We implemented the optic I/O chip-let which consists of a PMA circuit and an ultra-small optic transceiver, and integrate them through co-package technology.

A. Design of PMA Sublayer

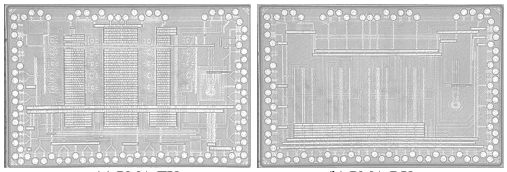
The PMA serializes 16 × 6.25Gbps data into 4 × 25Gbps in TX while deserializes 4 × 25Gbps data back into 16 × 6.25Gbps in RX.

Figure 4a shows all building blocks in the TX channel. Figure 4b depicts the schematic of the PMA RX. Besides the reverse data flow, RX Analog Front-end Equalizer (AFE) circuit implementation is different from TX AFE. In order to compensate for the channel loss and intrinsic bandwidth, RX AFE employs series inductive peaking for the input termination and the continuous-time linear equalizer for tunable equalization.

Figure 5 shows the simulation result of RX AFE by using Virtuoso software from Cadence Corporation, which shows RX AFE can provide up to 10 dB equalization at Nyquist Frequency for 25Gbps full-rate inputs. According to Table I, the power consumption of TX for one lane can be calculated as around 160mW. RX consumes the same power as TX,

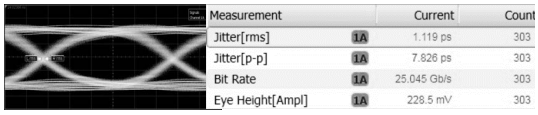
TABLE I
SIMULATION RESULT OF POWER CONSUMPTION

Power supply	Current	Power Consumption
4-Lane AFE	144.6mA	144.6mW
Clock	152.0mA	152.0mW
Driver	35.2mA	56.3mW
CDR+PRBS+MUX	282.0mA	282.0mW
Total	N/A	634.9mW



(a) PMA TX (b) PMA RX

Fig. 6. Microscope picture of PMA chips.



(a)Eye Diagram (b)Values of Some Indicators

Fig. 7. Eye diagram of PMA TX chip.

so power consumption per lane including TX and RX is 160mW + 160mW = 320mW.

To test the PMA design, we taped it out in 40nm CMOS process. Figure 6 is a microscope picture of PMA chips. TX chip is on the left, and RX chip is on the right.

Figure 7 is the eye diagram of one lane on the TX chip. We can see our TX chip reaches a 25.045Gbps bit rate while keeping good signal quality.

B. An Ultra Small Optic Transceiver

Optical I/O core is a silicon photonics chip designed by the Japan AIO core corporation [9]. It is an all-in-one chip that integrates driver circuit, quantum-dots laser, and rest parts such as modulator and photodetector. With all components integrated, its size without a package is only 5mm × 5mm, a minimal size compared to other optical transceivers whose parts are separated. Because it uses quantum-dots laser, it can also tolerate temperatures over 100°C. It also has a socket design so that it can be pluggable even in tight integration. For better adaptation in the data center, it adopts 1310nm wavelength. So it is an ideal candidate to be integrated with electronic chips.

C. Co-Package PMA With Optical Transceiver

The substrate's size is 3cm × 2.5cm in the package with a thickness of 1.49mm. To reduce signal skew, we control the intra-length difference of each pair of differential lines within 100μm. The inter-length difference between the four adjacent sets of differential lines is controlled within 200μm for the same reason (as shown in figure 9a). Two PMA chips and an optic I/O core chip are integrated into one module by co-package. Figure 9b is a picture of packaged module. Figure 9c is the substrate with two PMA chips and an installation slot for an ultra-small optical transceiver.

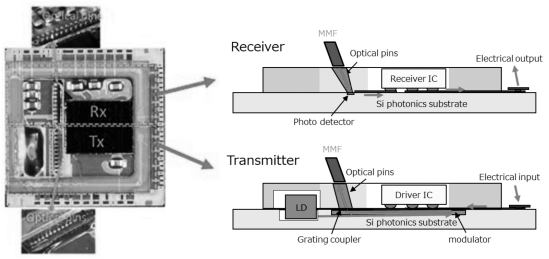
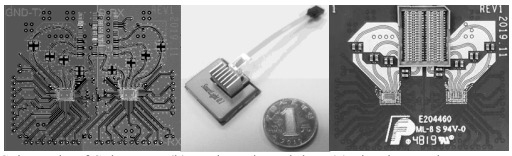
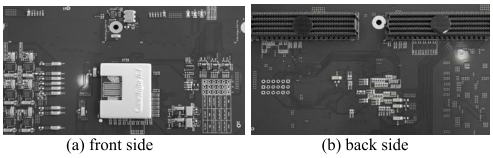


Fig. 8. Structure of optical I/O core.



(a) Schematic of Substrate (b) Packaged Module (c) circuit on substrate

Fig. 9. Co-package of PMA chips and optical transceiver.



(a) front side (b) back side

Fig. 10. Chip-level OI on a PCB board.

In order to isolate the mutual interference between different power supplies, all the power voltages are supplied independently. Thus, the parasitic inductance between the power plane and the ground plane is also minimized, and the fast switching noise on the power planes is decreased. For optical I/O core chip and PMA chips, the power planes and ground planes are divided into several individual parts, respectively.

IV. THE TEST

To test the optic I/O chip-let we designed, we implemented a MAC sublayer to generate test data.

We put the MAC and PCS code on an FPGA card, then connect it to a PCB board by FMC connector(shown in Figure 10b). There is an optic I/O chip-let we designed located on this PCB (shown in Figure 10a).

Using Vivado—a software developed for programming FPGA board from Xilinx Corporation, we proved the communication between PCS TX and PCS RX is correct, as shown in Figure 11.

V. A COMPARISON

We compared chip-level OI (C-OI) and board-level OI (B-OI) regarding their area occupation and power consumption. The calculation includes both electronic chip and optical transceiver. Chip-level OI comprises a CPU chip-let and an optic I/O chip-let that includes PMA and an ultra-small optical transceiver. In contrast, board-level OI is composed of an electronic ASIC chip and a pluggable optical transceiver. They implemented the same Ethernet NIC function.

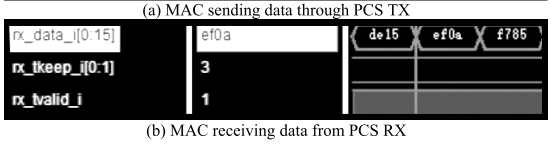
A. Comparation of Area Occupation

In chip-level OI implementation, CPU and optic I/O chip-let are integrated before packaging, so it can be easily put into a

TABLE II
PPA COMPARISON BETWEEN C-OI AND B-OI

Category	Chip-level OI	Board-level OI
Performance	100Gbps	100Gbps
Area Occupation	3.3cm x 3.3cm =10.89 cm^2	3.3cm x 3.3cm + 2cm x 6cm =22.89 cm^2
Power Consumption	1.17W+1.28W+1.5W =3.95W	12W+2.5W=14.5W

(a) MAC sending data through PCS TX



(b) MAC receiving data from PCS RX

Fig. 11. Communication test by FPGA board.

Power Group	Total Power	(%)
io_pad	0.0000	(0.00%)
memory	10.7287	(25.23%)
black_box	0.0000	(0.00%)
clock_network	0.0000	(0.00%)
register	0.0000	(0.00%)
sequential	21.7559	(51.17%)
combinational	10.0361	(23.60%)
Total	42.5207 mW	

(a) Power Consumption of PCS.

Power Group	Total Power	(%)
io_pad	22.5379	(1.92%)
memory	272.4408	(23.24%)
black_box	7.6371	(0.65%)
clock_network	402.1244	(34.30%)
register	415.3081	(35.43%)
sequential	0.5975	(0.05%)
combinational	51.6903	(4.41%)
Total	1.1723e+03 mW	

(b) Power Consumption of CPU

Fig. 12. DC synthesis report.

package with size 3.3cm × 3.3cm, which is a regular size for an electronic chip after package such as Chelsio Corporation's T6 chip [10]. Meanwhile, the area a typical board-level OI occupied is determined by the total area of the electronic ASIC chip and the pluggable optical transceiver. For the electronic ASIC chip, we again use 3.3cm × 3.3cm = 10.89cm^2 as the area it occupied. For optical transceiver, the area a Quad Small Form-factor Pluggable (QSFP) optical transceiver occupied is 2cm × 6cm = 12cm^2. So the area a board-level OI occupied will be 10.89cm^2 + 12cm^2 = 22.89cm^2. We assume a very short trace between the electronic ASIC chip and the pluggable optical transceiver to simplify the discussion, which means a board-level OI implementation may occupy more area than our calculation.

B. Comparation of Power Consumption

We calculate the power of chip-level OI by using DC software from Synopsis corporation. The CPU we used includes 4 CPU cores, cache, bus, MAC, and PCS components. From the DC report (as shown in Figure 12b), its power consumption is 1.17W. In section III.A, we can see the power consumption of PMA is 320mw × 4 = 1.28W. According to AIO's datasheet [9], the power consumption of AIO's optical transceiver is 1.5W. So total power consumption of chip-level OI will be 3.95W. In board-level OI, the power consumption of a typical electronic ASIC chip on 100G Ethernet NIC is 12W [10], and the power consumption of a pluggable optical transceiver is 2.5W [7]. So total power consumption will be 12W + 2.5W = 14.5W. The comparison result is shown in Table II.

It is not easy to acquire the components power of an electronic ASIC chip in a commercial 100G Ethernet NIC, so we cannot compare the power of chip-level OI with board-level OI at a deeper level to see which parts contribute to power saving. Nevertheless, we consider there are two factors that may contribute to less power consumed by our chip-level OI. One is that the PMA chips we designed work on such a short reach so that less power will be consumed on sending and receiving 100Gbps data. The other is that we choose a CPU architecture to implement 100Gbps Ethernet data processing, which significantly reuses its existing circuit elements, makes the electronic chip smaller, and saves power.

VI. CONCLUSION

We proposed a general chip-let architecture for chip-level optical interconnect application, which includes CPU chip-let and optic I/O chip-let. We redesigned the interface to keep two chip-let developing independently. We developed an optic I/O chip-let based on a pair of PMA chips and an ultra-small optical transceiver through co-package technology. We also taped out a CPU chip-let with MAC and PCS in 28nm CMOS process. This architecture can support any payload chip such as CPU, GPU, and switches to have optic I/O.

Compared with board-level OI [7], our work can achieve 1/3 of power consumption and 1/2 area occupation (shown in Table I) in Ethernet NIC application. By using chip-level OI, an Ethernet NIC can be much smaller. Moreover, this work can implement a direct optical connection on server PCB without installing an add-on optical Ethernet NIC.

REFERENCES

[1] *Part 3: CSMA/CD Access Method and Physical Layer Specifications*, Standard 802.3ba-2010, Jun. 2010, pp. 77–86.
[2] D. A. B. Miller, "Optical interconnects to electronic chips," *Appl. Opt.*, vol. 49, no. 25, p. F59, Sep. 2010.
[3] M. B. Riester, R. Houbertz-Krauss, and S. Steenhusen, "Chip-to-board interconnects for high performance computing," *Proc. SPIE* vol. 8630, Feb. 2013, Art. no. 863002.
[4] K. Schmidtke, F. Flens, and D. Mahgarefteh, "Taking optics to the chip: From board-mounted optical assemblies to chip-level optical interconnects," in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2014, pp. 1–3.
[5] C. Sun *et al.*, "A monolithically-integrated chip-to-chip optical link in bulk CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 4, pp. 828–844, Apr. 2015.
[6] K. Chen *et al.*, "Wavelength-multiplexed duplex transceiver based on III-V/Si hybrid integration for off-chip and on-chip optical interconnects," *IEEE Photon. J.*, vol. 8, no. 1, pp. 1–10, Feb. 2016.
[7] (2017). *100G QSFP28 Active Optical Cable Products Specification*. [Online]. Available: https://www.finisar.com/sites/default/files/downloads/finisar_fcbn425qe1cxx_100g_quadwire_qsfp28_active_optical_cable_productspecrevb5.pdf
[8] M. Moralis-Pegios *et al.*, "Chip-to-chip interconnect for 8-socket direct connectivity using 25Gb/s O-band integrated transceiver and routing circuits," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2018, pp. 1–3.
[9] *OPTICAL I/O CORE PAA-XW80001-ESA (Engineering Sample)*. Accessed: Sep. 21, 2020. [Online]. Available: http://aiocore.com/pdf/Optical_IO_Core_Leaflet_draft
[10] *Chesio T6 ASIC*. Accessed: Jan. 4, 2021. [Online]. Available: https://www.chelsio.com/terminator-6-asic/
[11] H. Li *et al.*, "12.1 A 3D-integrated microring-based 112Gb/s PAM-4 silicon-photonic transmitter with integrated nonlinear equalization and thermal control," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 208–210.
[12] B. Wang, W. V. Sorin, P. Rosenberg, L. Kiyama, S. Mathai, and M. R. T. Tan, "4×112 Gbps/fiber CWDM VCSEL arrays for co-packaged interconnects," *J. Lightw. Technol.*, vol. 38, no. 13, pp. 3439–3444, Jul. 1, 2020.