An Inter/Intra-Chip Optical Network for Manycore Processors

Xiaowen Wu, *Student Member, IEEE*, Jiang Xu, *Member, IEEE*, Yaoyao Ye, *Student Member, IEEE*, Xuan Wang, *Student Member, IEEE*, Mahdi Nikdast, *Student Member, IEEE*, Zhehui Wang, *Student Member, IEEE*, and Zhe Wang, *Student Member, IEEE*

***Abstract*— Manycore processor system is becoming an attractive platform for applications seeking both high performance and high energy efficiency. However, huge communication demands among cores, large power density, and low process yield will be three significant limitations for the scalability of future manycore processors. Breaking a large chip into multiple smaller ones can alleviate the problems of power density and yield, but would worsen the problem of communication efficiency due to the limited off-chip bandwidth. In response, we propose an inter/intra-chip optical network, which will not only fulfill the intra-chip communication requirements but also address the inter-chip communication, by exploiting the advantages of optical links with high bandwidth and energy efficiency. The network is composed of an inter-chip subnetwork and multiple intra-chip subnetworks, and the subnetworks closely coordinate with each other to balance the traffic. The proposed network effectively explores the distinctive properties of optical signals and photonic devices, and dynamically partitions each data channel into multiple sections. Each section can be utilized independently to boost performance as well as reduce energy consumption. Simulation results show that our network can achieve higher throughput with lower power consumption than alternative designs under most of synthetic traffics and real applications.**

***Index Terms*— inter-chip optical network, manycore processor, optical network-on-chip (ONoC).**

I. Introduction

With CMOS technology scaling down, manycore processor is becoming an attractive platform delivering high performance with limited power budget. It is projected that hundreds or even thousands of cores will be integrated on the chip. In a manycore processor system with so many cores, the communication demand will be so large that conventional electrical interconnects may not be able to fulfill it due to the bandwidth density and energy consumption constraints. The limitation of the communication subsystem will confine the manycore processor performance severely. Another limitation to the future manycore processor is the power density. It is estimated that more than 50% cores on the chip at 8 nm will not be utilized due to the power constraint [1]. The process yield will also confine the chip area and hence the scalability of future manycore processor. Breaking a large manycore processor into many smaller processors may decrease the power density as well as increase the yield. However, it requires enormous inter-chip bandwidth, laying the burdens on the off-chip interconnects, which is already the bottleneck of the system performance. The 3-D technology can be used to stack the chips and support low-latency inter-chip communication. However, the power density becomes even higher.

With the recent progress in silicon photonics, optical interconnects may be adopted to address these issues effectively. Optical interconnects promise ultrahigh bandwidth, low latency, and low energy consumption. They can address both the intra-chip and inter-chip communication requirements with limited power budget. For example, a silicon waveguide on the chip can support a data rate of 10 Gb/s for each light wavelength, and multiple wavelengths can be multiplexed into the single waveguide to achieve extremely high bandwidth. The waveguide can also be connected with off-chip waveguide passively to support ultrahigh off-chip bandwidth.

Optical network-on-chip (ONoC) using optical interconnects has been put forward to replace electronic NoC by many studies [2]–[7]. These works mainly focus on the intra-chip communication, while the work in [8] only deals with the inter-chip network. In this paper, we propose a new inter/intra-chip optical network (I^2CON), which supports both intra-chip and inter-chip communication. It includes multiple intra-chip subnetworks and an off-chip one, where the intra-chip and inter-chip subnetworks are codesigned to balance the bandwidths and the resources as well.

Both intra-chip and inter-chip interconnects are based on silicon photonic devices including modulator [9], photodetector [10], and waveguide [11]. The VCSELs [12] can be used as laser sources. To transmit optical signal from one chip to another in inter-chip interconnects, we use polymer waveguides on board as the transmission medium. The inter-chip interconnects in I^2CON are composed of multiple optical closed loops, which thread the chips together. To save limited network resources, we improve the link sharing by dividing one waveguide into multiple unoverlapped sections such that each section can be independently utilized. There can be multiple concurrent transactions with same wavelengths on a single waveguide without interference. Bidirectional transmission is also supported to further improve the link utilization.

Manuscript received August 8, 2013; revised January 27, 2014; accepted March 31, 2014. This work was supported by the research under Grant GRF620911, Grant GRF620512, and Grant DAG11EG05S.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science of Technology, Hong Kong (e-mail: wxxaf@ust.hk; jiang.xu@ust.hk; yeyaoyao@ust.hk; eexwang@ust.hk; mnikdast@ust.hk; zhehui@ust.hk; zwangag@ust.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2014.2319089

With such segmentation, higher throughput is achieved and the optical path for each transaction is shortened, reducing the power loss on the path as well. For intra-chip interconnects, we propose an optical network which is also composed of multiple optical closed loops similar to the inter-chip interconnects. In the channels, segmentation and bidirectional are also used to enable high utilization of the network resources. The inter-chip and intra-chip subnetworks are interconnected at the sites where the processing cores are located. They cooperate with each other to serve the data transfer from one core on a chip to another core on another chip.

We compared our network with two alternative designs including point-to-point network and limited point-to-point network. The simulation results show that our network can achieve higher throughput with lower power consumption under most of synthetic traffics and real applications. For instance, under all transpose traffics, the throughput of I^2CON is more than six times higher than point-to-point network, and more than 70% higher than limited point-to-point network. The energy comparison shows that under uniform traffic, I^2CON saves 52% and 58% of energy comparing with point-to-point and limited point-to-point networks, respectively. The rest of this paper is organized as follows. We first reviewed on-chip and off-chip optical networks and discussed the differences between our work and related work. Section III shows the overview of the whole architecture. The off-chip network will be detailed in Section IV and the design of on-chip network follows in Section **??**. After that, the simulation and evaluation are conducted in Section VI. Finally, Section VII concludes this paper.

II. Related Work

Based on the silicon photonic technologies, different on-chip network architectures have been proposed. Kirman *et al.* [13] presented an opto-electrical hierarchical bus for future manycore processors with cache-coherence supported. Xu *et al.* [14] proposed a hierarchical optical network and a composite cache coherence protocol, trying to acquire both advantages in snoopy and directory-based protocols. Pasricha and Dutt [15] proposed an optical ring waveguide to replace global pipelined electrical interconnects while preserving the interface with bus protocol standards. O'Connor [3] presented a full connected ONoC based on the special λ-router with WDM technology. Shacham *et al.* [5] proposed a hybrid ONoC combining an optical circuit-switched network with an electrical packet-switched network. Joshi *et al.* [16] presented a photonic clos network in which long electrical links between routers are replaced by optical ones. The proposed network provides more uniform latency and higher throughput compared with mesh network. Cianchetti *et al.* [4] proposed a packet-switched optical network. The packet may pass through multiple routers without being buffered as long as no collision happens. Li *et al.* [17] proposed a hybrid network in which optical network is used to broadcast latency-critical messages and electrical network is used to transfer high bandwidth traffic. Ouyang *et al.* [18] proposed an ONoC based on free-space optical interconnects to reduce power consumption. Psota *et al.* [19] used WDM technology to build contention-free network, which facilitated new programming model. Koohi *et al.* [20] proposed hierarchical optical rings, where local rings are used for intranode communication and global rings are to connect the nodes. In all these designs, only one chip is considered, and the networks are proposed to address the intra-chip communication requirements. In I^2CON, we also use an optical intra-chip subnetwork to support the on-chip communication; but more importantly, we use an inter-chip network, which is highly correlated with on-chip network to address the communication among chips.

The on-chip subnetwork in I^2CON is an optical crossbar network with ring topology. Similar topologies have been proposed in [2], [6], and [21]–[23]. In crossbar design with large network resources, link sharing is important to reduce the resource requirements. For example, Vantrease *et al.* [2] proposed a crossbar, in which a waveguide for data transfer is shared by multiple writers and a single reader. On the other hand, Pan *et al.* [21] proposed a design that a waveguide is shared by single writer and multiple readers. In [6], a waveguide can be further shared by multiple writers and multiple readers (MWMRs). Xu *et al.* [23] proposed a channel borrowing technology to improve the channel utilization and also reduce the power consumption. In all these designs, a waveguide is unidirectional, and at any time, there can be no more than one transaction with same wavelengths in a single waveguide. In I^2CON, bidirectional transmission is supported, and we further improve the resource sharing by allowing concurrent transmissions with same wavelengths on a single waveguide.

Le Beux *et al.* [22] presented an optical ring NoC for both 2-D and 3-D architectures. In their design, a wavelength can be reused in a waveguide such that it can also support multiple transactions to improve the performance as our I^2CON. The wavelength is statically assigned based on the connectivity requirements. In the I^2CON, a single waveguide supports multiple transactions dynamically based on the arbitration. Also, bidirectional transmission is supported for the same link. Morris *et al.* [24] proposed an optical network with 3-D stacking technology. A large crossbar is decomposed into multiple small crossbars on different layers to reduce the power. The idea of decomposing a long link into some shorter links is also adopted in our I^2CON, but we need not physically break the channel and only one optical layer is required. Datta *et al.* [25] proposed segmented optical bus. Buses are segmented to reduce power consumption and they are interconnected by electrical routers. I^2CON segments the bus in more depth and the throughput is higher with efficient arbitration and more independent segments. No electrical switching is required, which consumes large power and area. The optical power is also lower due to the light will only pass the active parts of the link.

Optical interconnects for chip level communication have been proposed for more than a decade. Polymer waveguides on board [26], fiber [27], and free space [28] have been proposed as mediums for light transmission. Among these techniques, the polymer waveguide fabricated on PCB is especially favored for its compatibility with PCB design process.

Comparing with the fiber, waveguide can have smaller pitch width and thus higher bandwidth density. Another feature of waveguide is the possibility to integrate splitters and combiners, which are useful for bus-like structures [29]. In I^{2}CON, polymer waveguides are used for inter-chip communication. Koka *et al.* [8], [30] proposed a new approach to interconnect the chips together. The processor dies are placed on a large SoI substrate on which silicon waveguides are routed. The SoI substrate is with a width of more than 10 cm, which is much larger than a conventional chip and hence reducing the thermal density effectively. Two networks, namely point-to-point network and limited point-to-point networks, had been identified as the two most promising designs in terms of performance and power. In the point-to-point network, each die communicates with all other dies with dedicated channels. There is no routing stage or arbitration required for each channel but at the cost of flexibility. In the limited point-to-point network, electrical routers are employed to break most links into two stages to increase the flexibility. In I^{2}CON, link sharing is explored to support the network flexibility but not introducing large arbitration overhead. The detailed comparison between I^{2}CON and these two networks will be given in the evaluation section.

III. Architecture Overview

I^{2}CON targets an optically connected manycore processor system with multiple chips. On each chip, there are two layers: 1) optical layer and 2) electrical layer. They are stacked together with 3-D stacking technology. There are processor cores in the electrical layer. In optical layer, silicon-photonic devices, including waveguides, optical switches, and photodetectors, will be fabricated to support optical signal transmission. On-chip lasers, VCSELs, are bonded on the chip as light sources. The cores on the electrical layer can access these optical components with through-silicon-vias. The chips are bonded on the board and connected with board-level optical interconnects. The on-chip optical fabrics work together with the on-board waveguides to facilitate not only the intra-chip communication among the cores on the same chip, but also the inter-chip communication among the cores on different chips.

The logical view of I^{2}CON is shown in Fig. 1(a). It is composed of an inter-chip network and multiple intra-chip networks. Each intra-chip network is to interconnect all cores on the same chip plane. And the inter-chip network is to thread the cores in a third dimension, which is perpendicular to the chip plane. In this way, the chips are virtually stacked like a 3-D chip. Optical signals can tolerate much longer distance than electrical signals, given that a longer distance will not introduce much power, throughput, and latency overheads. Therefore, physically, the chips are placed far away from each other as shown in Fig. 1(b). The large distance between chips can essentially reduce the power density. These features can help build a logical dense but physical-large system.

Each chip in the system is a manycore processor with multiple homogeneous cores. Each core is with private L1 data and instruction cache, and every four cores are clustered

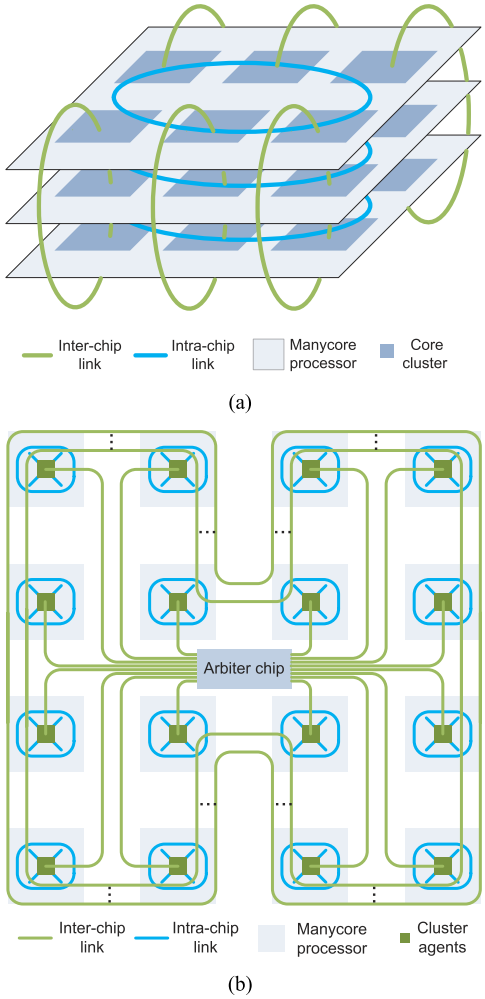

Fig. 1. Conceptual topology and physical floorplan overviews of I^{2}CON. (a) Logical view of I^{2}CON. (b) Physical floorplan of I^{2}CON.

together as a core-cluster sharing an L2 cache. For clarity, we denote the jth core-cluster in ith chip by $CC(i, j)$, and assume there are M chips in our system and each chip is with N core-clusters. All the clusters on a chip, $CC(i, j)$ $\forall j \in [0, N-1]$, are interconnected by an intra-chip network. The chips are interconnected by parallel circular inter-chip optical links, which are controlled by arbiter chip, as shown in Fig. 1(b). Logically, these links are perpendicular to the chip plane, and thus they address the vertical communication in the system. That is to say, each inter-chip link is to connect the core-clusters with the same position on the chips. For example, the kth data channel will interconnect $CC(i, k)$ $\forall i \in [0, M-1]$. Under this arrangement, each channel will be accessed by only one cluster in each chip, and there are totally N ($N =$ number of clusters on a chip) inter-chip data channels. These channels are homogeneous and parallel to each other. With this design, the optical IOs are physically distributed across the whole chip evenly, avoiding a centralized node gathering all IOs at a small place.

The communication within the cores on the same chip can be addressed by intra-chip network alone. But the communication from cores on different chips requires the cooperation between intra-chip and inter-chip networks. For example,

for a transaction $CC(i,u) \rightarrow CC(j,v)$, where $i \neq j, u \neq v$, the packets may take the paths $CC(i,u) \rightarrow CC(i,v)$ and $CC(i,v) \rightarrow CC(j,v)$, in which the former path is on intra-chip network and the latter is on inter-chip network. In the following sections, we will discuss the inter-chip network first and then the intra-chip one.

IV. Inter-Chip Network

Inter-chip network addresses the communication requirements among the chips. The requirements, such as large bandwidth density, low latency, and low power consumption, are difficult to be fulfilled by the conventional electrical wires. By exploiting the inherent properties of optical links, we place the chips far from each other to reduce the power density but still provide high performance and high energy efficiency communication structure.

The inter-chip network is composed of data channels and the accompanying control fabrics. There are N data channels, which are parallel to each other with the same design. They connect the core-clusters in different chips. Payload data are transmitted between clusters on different chips. The control fabric is composed of the control channels and an arbiter chip, as shown in Fig. 1(b). Before accessing the data channel, the clusters are required to send requests to the arbiter chip through the control channels. The arbiter chip will make the arbitration and also configure the data channels by sending out control information to the data channels. The detailed design of data channels will be discussed first, followed by the discussion of the control structure.

A. Inter-Chip Data Channel

The inter-chip data channels are homogeneous and parallel to each other without waveguide crossings. The design of a channel is shown in Fig. 2. The zeroth data channel is used to connect all core-clusters $CC(i,0)$ $\forall i \in [0, M-1]$. The channel is composed of closed-loop waveguides (only one waveguide is shown in the figure) with optical transceivers attached to them. The on-chip optical transceivers interact with the silicon waveguides to get the light out of the data channel or inject the light into the channel. Each closed-loop waveguide is built by bridging silicon waveguides on chip and the polymer waveguides on board. Previous works [31], [32] show that the coupler between silicon and polymer waveguides can be made with very small loss. The coupling is achieved by adiabatic mode transformation. In [32], a coupler with loss around 0.8 dB has been fabricated. Besides the couplers, there are no OE/EO conversions at the chip IO, saving significant power consumption.

1) Optical Transceiver: The optical transceiver is composed of VCSELs, waveguides, photodetectors, and microresonators (MRs). The VCSELs serve as the on-chip laser sources, and arrays of VCSELs can be bonded on top of the chip [12]. Compared with off-chip laser source, the on-chip laser source owns the potential of substantially reducing the static power. The on-chip laser can be powered OFF when there is no data transfer. This will significantly reduce the power consumption if the application load is not high. Another advantage of on-chip laser is that we can dynamically control the output power based on the path loss. The disadvantage of on-chip laser is that it will be thermally affected by the chip. The power efficiency of the laser will drop with high temperature. However, this overhead will be well compensated by saved power, which is verified in our evaluation. When VCSELs are bonded on the chip, the output lights are vertically with respect to the chip. To couple the vertical light to in-plane silicon waveguides, grating technology can be used [33].

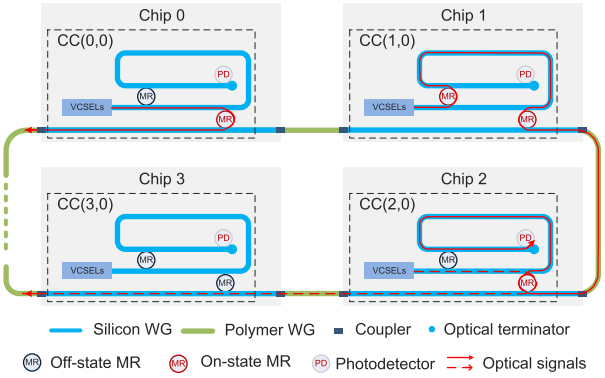

Fig. 2. Illustration of inter-chip data channel 0. One wavelength is illustrated. Both solid and dashed lines are used to distinguish the different transactions at the same wavelength.

The MR is a switching element. It can divert the light with resonance wavelength from one waveguide to the opposing one. The resonance wavelength of the MR can be controlled by electrical voltage. The functionality of the MR in the data channel can be shown in Fig. 2. In the transceiver of cluster $CC(0,\ 0)$, the upper left MR is turned OFF, and the light passes by this MR safely. On the other hand, in the transceiver of cluster $CC(1,\ 0)$, the upper left MR is turned ON and it diverts the light from one waveguide to another parallel one. Therefore, by turning ON/OFF the MR, the light will take different paths.

The MR is wavelength selective, and multiple MRs are used for multiple wavelengths. In I^2CON, we pack W wavelengths into the waveguide for each transaction. Therefore, W MRs are required to multiplex all wavelengths at the source, and there are W MRs at each switching stage and W photodetectors for each receiver. For simplicity, only one MR is shown in Fig. 2, and the MRs used for multiplexing at the source are not shown. The MR with germanium doped is used as a wavelength-selective photodetector. This design of the detector will reduce the capacitance and remove the transimpedance amplifiers [2]. Again, only one photodetector is shown in the figure although W of them are actually deployed.

The functionality of the optical transceiver is based on the coordination of the lasers, MRs, and photodetectors. The laser injects modulated light into the waveguide; the MRs switch the light from one waveguide to another; finally, the photodetector receives the light from the waveguide. For example, in the transceiver of $CC(0,\ 0)$, the upper left MR is turned OFF and the lower right MR is turned ON. The light emitted from the VCSEL is thus injected into the data channel counterclockwise. For the cluster $CC(1,\ 0)$, both MRs are turned ON, and thus the light is injected into the data channel clockwise.

For the cluster $CC(2, 0)$, the transceiver turns on the lower right MR to couple the light from data channel [sent by $CC(1, 0)$] into the transceiver and later receives it with photodetectors. For cluster $CC(3, 0)$, all MRs are turned OFF and thus the light passes through this cluster safely.

2) Bidirectional Transmission: In conventional design, each link is single directional and two links are required to support the communication between the two communicating nodes. Due to the imbalance property of the real traffics, it is often the case that one unidirectional link is busy with heavy traffic burden while the opposite link is idle with no data transmission, wasting the network resources.

Motivated by this observation, we design the channel such that each link supports bidirectional transmission as shown in Fig. 2. For example, cluster $CC(1, 0)$ can send data to $CC(2, 0)$ as illustrated by powering ON the upper left MR in $CC(1, 0)$ and powering OFF the upper left MR in $CC(2, 0)$. On the other hand, by powering ON the upper left MR in $CC(2, 0)$ and powering OFF the upper left MR in $CC(1, 0)$, $CC(2, 0)$ will be able to send data to $CC(1, 0)$ with the same optical link between them. This flexibility in direction can well handle the heterogeneous real traffics.

3) Channel Segmentation: Another feature of the channel is that, the channel is virtually segmented into multiple sections, and these sections can work independently and concurrently. This can effectively improve the link utilization. For simplicity, we use $S_0[i, j]$ to denote the channel section from cluster CC(i, 0) to cluster CC(j, 0), in clockwise direction. As shown in Fig. 2, when $CC(0, 0)$ sends data out to the CC(M$-$1, 0) (not shown in the figure), the cluster $CC(1, 0)$ can send data to $CC(2, 0)$ simultaneously. That is, $S_0[M-1, 0]$ and $S_0[1, 2]$ can work independently although they are in the same channel. In conventional data channel design [2], [6], a data channel can support only one transaction at a time, while our channel supports M (M is the number of chips) concurrent transactions in the best case, improving the throughputs by M times. Please be noted that, multiple consecutive sections can also work together to form a large section. For example, by turning OFF the MRs in $CC(3, 0)$, $S_0[2, 3]$ and $S_0[3, 4]$ are connected as $S_0[2, 4]$, preserving the flexibility of the channel.

The segmentation feature can be further facilitated by the bidirectional feature in improving the resource utilization. Since the channel is a circle, a long link used by a transaction can be replaced by a short one in opposite direction such that the unused long link can be utilized by other transactions. For example, in Fig. 2, $CC(0, 0)$ sends data to cluster $CC(M-1, 0)$ (not shown in the figure) in counterclockwise direction occupying the $S_0[M-1, 0]$, leaving most of sections free to work. If only single direction (e.g., clockwise) is allowed, $S_0[0, M-1]$ would be occupied, leaving no sections for other clusters. It is also possible that we can choose an opposite direction for a transaction to prevent collisions. For example, if $CC(0, 0)$ wants to send data to $CC(3, 0)$ given that $CC(1, 0)$ is sending data to $CC(2, 0)$, we can let $CC(0, 0)$ choose the counterclockwise direction such that no collision happens. This cannot be achieved if bidirectional transmission is not supported.

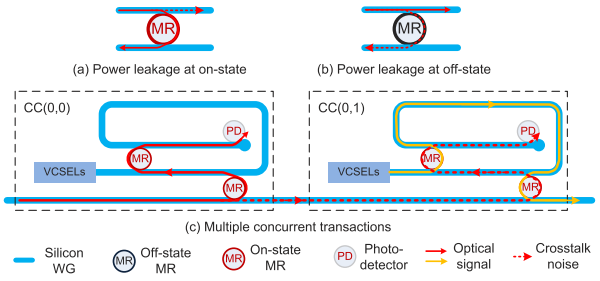

Fig. 3. Crosstalk noise illustration.

4) Crosstalk Noise Analysis: Crosstalk can be a threat to the network performance and scalability as our previous work showed in [34]. In this section, we will show our network is immune to such threat. As shown in Fig. 3(a), when the optical signal is switched by MR from one waveguide to another waveguide, some residual power will be left in the original waveguide with an extinction ratio K_{ON}. On the other hand, when the optical signal bypasses an OFF-state MR, some fraction power will be switched to another waveguide, and the extinction ratio is K_{OFF}. Here, we assume K_{ON} is -25 dB and K_{OFF} is -20 dB [35]. The unintended leakage power can be the noise to other transactions. However, the leakage power is so small that it would not hurt the signal until the noise is accumulated. Fortunately, when we have multiple concurrent transactions on the same channel, the noise will not accumulate because a transaction who makes noise will absorb the noise produced by others. As shown in Fig. 3(c), although $CC(0, 1)$ is sending data and thus creating noise to other downstream clusters on the channel, it is removing noise from upstream transactions. In the worst case that there are N concurrent transactions on the channel, the maximum accumulated noise on the channel is

$$P_{\text{noise}} * \left(1 + K_{\text{ON}} + K_{\text{ON}}^2 + \cdots + K_{\text{ON}}^N\right) < P_{\text{noise}}/(1 - K_{\text{ON}}) \approx P_{\text{noise}} \quad (1)$$

where P_{noise} is the leakage power from ON-state MR, and K_{ON} in percentage is 0.3%.

B. Control Subsystem

The data channel requires a conflict resolution scheme to prevent two transactions overlapping at the same channel section. Also, the path is required to be set up before the transfer of payload data. Since each channel is independent with the others, each channel is controlled by a separate control unit. This will help decompose the complexity of the arbiter. We put all the control units in a special chip called arbiter chip, as shown in Fig. 1. All the manycore processor chips are optically connected to this arbiter chip with on-board waveguides.

1) Control Protocol: Before accessing the data waveguide, the cluster will send a request to its control unit with the information, including destination ID, request ID, and packet size. Destination ID is used to identify the receiver cluster. The request ID is attached for each request so that the cluster can send multiple requests out before receiving grant information.

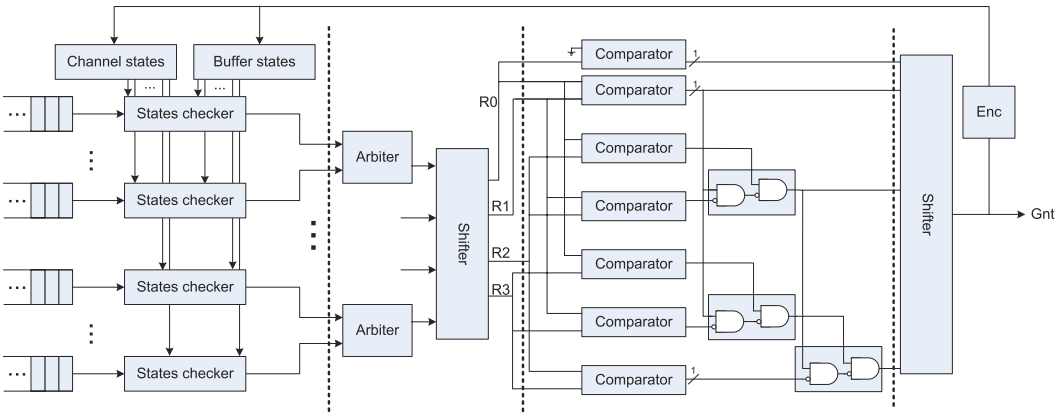

Fig. 4. Control unit of the inter-chip network. The units are homogeneous and independent with each other, and thus only one unit is shown. The comparator in the figure will output true if there is no collision between two input requests.

This will help to boost the throughput of the control system through pipelining, especially when the round trip delay is large. Variable packet size for each transaction is supported and thus the size information is required.

After receiving the request, the control unit will check the channel states, try to reserve a channel section for this request and finally send the grant packet back to the cluster. At the same time, the control unit will also send the grant information to the destination cluster, telling it to configure the receiver to detect the incoming data. After receiving the grants, the source cluster will send data out on the assigned channel while the destination cluster will configure the MRs to detect the coming signals.

Credit-based flow control is used in I^2CON, which is facilitated by the control units. Each control unit has the initial number of tokens corresponding to the number of buffer slots of each receiver. It counts down the tokens each time a packet is sent. On the other hand, the receiver cluster will send the new tokens back to the control unit via the optical link if the buffer slots are emptied. If no token is left, the requests will not be processed by the control unit.

2) Control Unit: Each control unit is to process the arbitration for all the requests on the same data channel. It is optically connected to the clusters via on-board waveguides. Since the cluster will send both requests and buffer tokens to the agent, two types of packets are required. One bit is enough to identify the difference. From the agent to the cluster, there are also two types of information: 1) the grants answering the request and 2) the grants informing the receiver that new packets are coming.

After receiving the requests, the arbitration process is divided into four stages, and the corresponding hardware design is shown in Fig. 4. The first stage is to select the requests from the request pool. The selected requests should satisfy two conditions: the intended link should be idle and the destination's buffer is not full. To check these two conditions, channel states and buffer states are kept in the registers. The channel states are in the granularity of section. For example, if $M = 16$, the states of zeroth data channel is composed of all sections from $S_0(0, 1)$ to $S_0(15, 0)$. If a request represents a transaction from cluster 4 to cluster 8, all the sections $S_0(i, i + 1)$ $i \in [4, 7]$ should be idle. Each section can be tagged by a valid bit, and checking multiple sections can be done by single OR operation.

The second stage is to reduce the collision possibility among the clusters. Each channel is logically divided into R ($R = 4$ in Fig. 4) regions, and only one request is selected from each region. The selection can be based on round-robin or other schemes preserving fairness. We do not select more than one requests due to the fact that the requests in the same region are quite likely to collide with each other. The selected four requests are then permutated by the shifter as shown in the figure. The outputs of the second stage are four permutated requests, which will be sent into the third stage that is with fix priority. The fairness among the regions can be preserved by changing the permutation scheme in this stage.

The third stage is to check the collisions among the selected requests. The two requests, with intended sections $S_i(a, b)$ and $S_i(c, d)$, respectively, collide if and only if $d \leqq a \leqq c \parallel d \leqq b \leqq c$. The selected set should include maximum requests in which there is no collision. Checking them serially may save resources but takes long time, which is linearly to R (the number of regions). To save check time, the system is designed in a parallel way. Every requests are compared in pair to check the collision, and each comparison is made in the comparator units as shown in the figure. The output of this stage is the valid set of requests. The invalid request is marked as 0 and 1 otherwise.

The last stage is first to permutate the requests back to the original order as in input of the second stage. The shifter here should implement a permutation scheme, which is reverse to that in the second stage. The grant information for the selected requests will be sent out to the clusters. And the update information will also be sent back to update the buffer states and channel states.

V. Intra-Chip Network

In I^2CON, multiple homogenous chips are used. The architecture of the chip with intra-chip network is shown in Fig. 5. All the core-clusters are interconnected with each other by

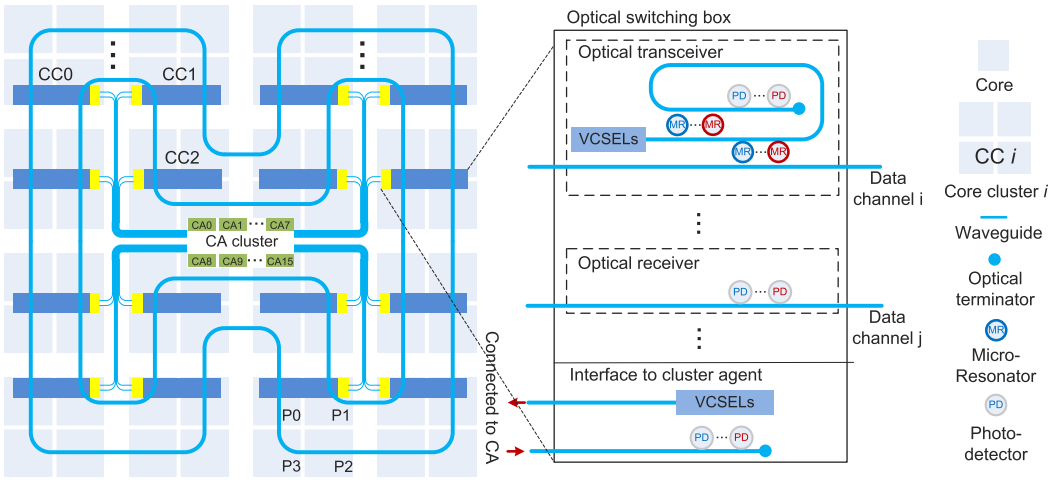

Fig. 5. Overview of intra-chip architecture and its floorplan. There are multiplexers following the VCSELs, which are not shown in the figure.

the parallel closed-loop channels. The cluster accesses the channels with optical switching box shown in the right side of Fig. 5. The communication among the cores are coordinated by the control subsystem. In the control subsystem, each cluster is assigned with a cluster agent (CA). All agents are located at the chip center, forming cluster agent cluster as shown in Fig. 5. They are interconnected with short local electrical wires, while each cluster agent is connected to the corresponding cluster with dedicated optical waveguides. This approach utilizes the advantages of optical interconnects in long distance communication and electrical interconnects in short distance.

A. Data Channel Design

On-chip data channels are composed of multiple parallel waveguides, which are aligned as closed loops and pass through all clusters on the chip. Each cluster accesses all channels with optical switching box, as shown in the right side of Fig. 5. The switching box includes many optical transceivers and they are designed in the same way as the inter-chip ones. For on-chip network, we also pack W wavelengths into the waveguide for each transaction. At the source, W MRs are used to multiplex all wavelengths into single waveguide, which is not shown in the figure for simplicity. W MRs are also used at each switching stage and there are W photodetectors for each receiver as shown in the right side of figure.

A complete data channel is a waveguide with all transceivers attached to it, which is similar to inter-chip channel design, as shown in Fig. 2. The only difference is that, the inter-chip channel is constructed by connecting silicon waveguides and on-board polymer waveguides, while the on-chip channel is a single closed-loop silicon waveguide.

The similarity between on-chip and inter-chip channel design implies that on-chip channel also owns two important properties: 1) bidirectional transmission and 2) channel segmentation. To illustrate the benefits of the properties, we compare our design with an alternative on-chip channel design called MWMR channel [6], as shown in Fig. 6. For simplicity, we use CC_i to denote the ith cluster on the chip, and use $S[i, j]$ to denote the section from CC_i to CC_j in clockwise

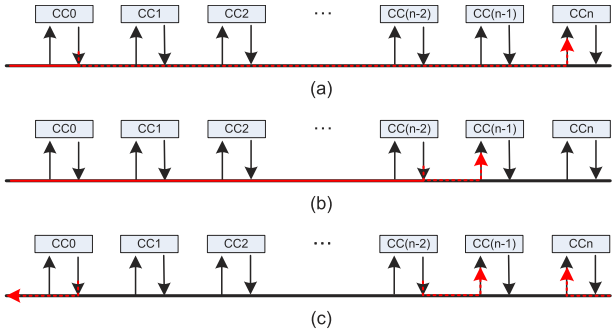

Fig. 6. Illustration of different optical channel designs. Red solid lines denote the continuous waves and dotted lines denote the modulated light signals. (a) Transaction 1 in MWMR. (b) Transaction 2 in MWMR. (c) Two concurrent transactions in I^{2}CON.

direction. In MWMR channel design, each channel is accessible for all writers and readers as in I^{2}CON. In Fig. 6(a) and (b), there are two different transactions, but only one transaction can be supported at a time in MWMR channel. Therefore, two transaction periods are required to complete them. In contrast, the I^{2}CON channel can support multiple concurrent transactions and bidirectional design as shown in Fig. 6(c). With bidirectional transmission, the original long path $S[0, n]$ of transaction 1 is replaced by a shorter path $S[n, 0]$. With segmentation, the left links $S[0, n]$ can be utilized to support other transactions, such as transaction 2. As a result, two transactions can be finished in a single transaction period. In best case, I^{2}CON can support N (the number of clusters on a chip) concurrent transactions, improving the throughput by N times.

Besides throughput improvement, power reduction can also be achieved by bidirectional transmission and channel segmentation. On the data channel, multiple senders and receivers are attached to the waveguide, introducing power loss for the light passing through them. The waveguide itself will also introduce some loss. In MWMR channel design, the light has to go through the whole link, encountering large power loss. In contrast, the light in I^{2}CON only passes through necessary link sections, avoiding unnecessary power loss. Further, bidirectional transmission can help to find a potential shorter path with smaller loss in opposite direction. The power saving

is more significant when the network size grows. As shown in Fig. 6, with bidirectional transmission property, an original transaction 1 with path $S[0, n]$ in MWMR channel is replaced by much shorter path $S[n, 0]$ in I^{2}CON. The power loss is reduced from n link sections to one section, and thus the power saving is significant when n is large. The effective distance of transaction 2 in MWMR channel is $S[n-2, n-1]$, but the light has to go through the path $S[0, n-1]$. In contrast, the light will only passes $S[n-2, n-1]$ in I^{2}CON with segmentation property, saving large power. Also, the power consumption of transaction 2 in MWMR is proportional to the network size n, while it is constant in I^{2}CON. With these properties, the network power in I^{2}CON well scales with network size. The evaluation results, based on the parameters in Table II, show that compared with 16-cluster intra-chip network, the 32-cluster network will only consume 12% more power for each bit transmission in average. In contrast, in MWMR design, 110% more power is required.

B. Control Subsystem

Each transaction requires path setup before payload data transmission. Each cluster is assigned with a cluster agent that is responsible for processing the requests from this cluster. A cluster agent needs negotiate with other agents to make sure the channel is idle and destination buffer is not full. Therefore, we put all agents close with each other in the chip center. They communicate using short local electrical wires. The relatively large distance between the agent and cluster is offset by the dedicated optical links, which provides low communication delay. The connection between each agent and the corresponding cluster is composed of two unidirectional waveguides: one for transmitting requests from the cluster to the agent, and the other one for grant information from the agent to the cluster. The latency between cluster and agent is within one clock cycle.

Before accessing the data channel, the cluster will send a request to its agent (called source agent) with the information, including destination ID, request ID, and packet size. After receiving the request, the source agent will check with the other agents, try to reserve a channel section for this request and finally send the grant packet containing the channel ID back to the cluster. At the same time, the destination cluster's agent (called destination agent) will also send the grant information to the destination cluster. After receiving the grants, the source cluster will send data out on the assigned channel (identified by the channel ID) while the destination cluster will open the receivers to detect the data. Similar credit-based flow control is adopted to prevent buffer overflowing.

For on-chip network, it is possible to use the same arbitration scheme as the inter-chip network. However, for intra-chip network, the number of channels is larger, and there are more clusters ($N \gg M$) attached to each channel. Therefor, the on-chip network resources are more abundant but the arbitration overhead is also much larger than inter-chip networks. On the other hand, the area budget for on-chip network controller is much less compared with a separated control chip used for inter-chip network. Based on these observation, we proposed channel grouping to effectively reduce the arbitration overhead.

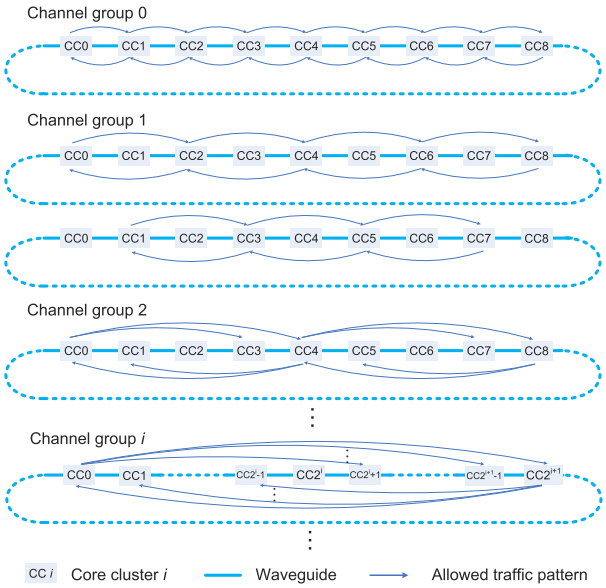

Fig. 7. Channel grouping. Channels are classified into multiple groups based on the allowed transmission patterns.

1) Channel Grouping: To cope with the arbitration complexity in cluster agent, we set some access rules in data channel at first place. This will eliminate some traffic patterns on a specific channel such that these patterns do not need to be considered for that channel at all. Although imposing access rules may sacrifice some flexibility of the network, it effectively reduces the arbitration overhead and thus the processing delay, potentially achieving even higher performance in reality.

To impose the access rule, we classify the data channels into groups according to the allowed patterns on the channel. In each group, only special traffic patterns are allowed. Here, the pattern is referred to the transaction distance and the distance in turn is measured as the minimum number of hops between two clusters. For example, the distance between cluster CC_1 and CC_0 in Fig. 5 is one hop instead of 15 hops since both clockwise and counterclockwise directions are supported.

As shown in Fig. 7, in the ith group, only allowed transactions are within distance $\in (2^{i-1}, 2^i]$ hops. With this classification, intergroup collisions are skipped in arbitration. For example, the traffic with three hops will be assigned to the group 2, and will not be interfered by transactions with distance larger than four hops or smaller than three hops. Within a group, we further classify each data channel based on the allowed intervals on the channel. For example, as shown in the figure, on the first channel in group 1, the permitted senders are the clusters with even labels. And on the second channel, only odd labeled clusters are allowed. Formally, on the jth channel in group i, the allowed senders are the clusters with labels $(j + k \times 2^i)\%N$, where k here is any nonnegative integer and N is the total number of clusters. And there are 2^i waveguides in group i. In this way, each channel in group i is divided into $N/2^i$ sections with length 2^i hops. All transactions are confined within the sections such that no cross section collisions exist. The arbitration left is relatively simple since only two neighboring senders may conflict with

each other. It is also necessary to mention that, due to the accessing rules, many senders/receivers can be omitted on the data channel. In Fig. 7, take the first waveguide in group 2, for example, there is no sender or receiver attached to the waveguide at cluster 2. The senders in cluster 1 and 3 are also omitted in this waveguide.

2) Cluster Agent Design: With accessing rules in channel grouping, a cluster agent will only need to negotiate with two neighboring agents for each transaction, the complexity of the arbitration algorithm is reduced to $O(1)$. The agent receives the packet from cluster and then decodes it. If it is a packet containing requests, the requests will be stored in the request pool. If it is a packet with buffer tokens, the tokens will be sent to the related sender agents. For each request in the request pool, it has to undergo three steps before being granted. The first step is to check whether the destination buffer is full, which is achieved in flow controller unit. It is followed by the channel collision solver, which checks whether the channel segment is available. Finally, the state of the destination agent is checked in the destination checker. It is to make sure that the destination agent is able to inform the destination cluster to open the detector on time. If the three steps are passed successfully, the request is granted and the grant information is sent back to the cluster. Otherwise, the request will be stored back into the request pool. Multiple requests are processed in parallel and the process of each request is well pipelined to improve the throughput of the controller.

The cluster agents share some information but the negotiation is limited such that delay is well confined. First of all, buffer information is shared among the agents. Given the credit-based flow control, one single wire between two agents is enough for buffer information transfer: 1 indicates one new buffer slot is available and 0 means null. Buffer information transfer is unidirectional and there is no arbitration or broadcasting required. Furthermore, variable packet size is fully supported by this design. The second shared information is channel states. The link sharing is confined between two neighboring senders, given that a section can only be utilized by two senders at the two endpoints. The arbitration can be round-robin between these two senders. The last shared information is the state of the cluster agent. In our protocol, the destination agent is required to inform the receiver cluster to listen on a specific data channel at the right time. It is possible that the destination agent is busy and cannot process the request. The checking protocol is NAK-based; a negative signal is sent back to source agent from the destination agent, otherwise no signal is transmitted back. This will save power since the case of busy is not common due to the sufficient link bandwidth allocated.

C. Connecting Intra/inter-chip Networks

The intra-chip and inter-chip networks intersect at the core-clusters. Each core-cluster is a buffering point for inter-chip communications. We denote the uth cluster on chip i by $CC(i,u)$, and $CC(i,u) \rightarrow CC(j,v)$ represents the traffic from $CC(i,u)$ to $CC(j,v)$. If $i = j \| u = v$, $CC(i,u) \rightarrow CC(j,v)$ is a pure inter-chip or intra-chip transaction and will be served by inter-chip or intra-chip network, respectively. If $i \neq j \&\& u \neq v$, then the transaction has to be split into two transactions: $CC(i,u) \rightarrow CC(i,v) \cup CC(i,v) \rightarrow CC(j,v)$ or $CC(i,u) \rightarrow CC(j,u) \cup CC(j,u) \rightarrow CC(j,v)$. The packet can take intra-chip path or inter-chip path first. But to avoid deadlock, we make the packet always take intra-chip path first. The whole process is as follows. $CC(i,u)$ first sends a intra-chip request to its agent and then sends data to $CC(i,v)$ after receiving the grant information. $CC(i,v)$ receives the packet, finds out it is an inter-chip packet, and thus issues a request to its control unit on arbiter chip. After receiving the grant from the control unit, $CC(i,v)$ sends out the packet to $CC(j,v)$, completing the whole transaction.

VI. Quantitative Analysis and Simulation Results

In this section, we evaluate the performance and power efficiency of I^{2}CON and compare it with the related works. Although there are lots of works studying optical on-chip network, few of them have been focused on off-chip network. The work in [8] discussed the multichip systems, and proposed some nanophotonic networks for the system. Two networks, namely point-to-point network and limited point-to-point networks, had been identified as the two most promising designs in terms of performance and power. Therefore, we compare our work with these two designs in this section.

In both point-to-point and limited point-to-point networks, the processor dies are placed as a 2-D array on a large SoI substrate. Each processor die is a cluster with four processing cores [30]. In the point-to-point network [8], each die communicates with all other dies with dedicated channels. There is no routing stage or arbitration required for each channel, but it is at the costs of flexibility and scalability. The waveguides are aligned horizontally or vertically, and each point-to-point channel uses two wavelengths for data transmission. In the limited point-to-point network [8], the dies in the same row/column are still connected in a point-to-point fashion, but each die is associated with an electrical router such that a die can communicate with a die not in the same row/column. With an electrical router on the path, the flexibility is increased but it is at the cost of extra OE/EO conversion power and also the electrical switching power.

To compare I^{2}CON with the point-to-point network and the limited point-to-point network, we consider a system with 64 clusters and each cluster is with four processing cores, which is also assumed in [30]. In I^{2}CON, there are four chips with each chip 16 clusters. For each chip, we use two sets of channel groups, and each set includes four groups from group 0 to group 3. For inter-chip interconnects, each channel connecting all four chips are set to have six waveguides. Be noted that the number of group sets and the number of waveguides for each inter-chip channel can be varied, but we intentionally set a matched network such that we can compare our network with the other two designs. The resource comparison summary is given in Table I.

In all three designs, the clock frequency is assumed as 5 GHz and the data rate of each wavelength in all three designs is assumed to be 10 Gb/s. We also assume that eight wavelengths are multiplexed into a single waveguide for each

TABLE I
NETWORK RESOURCES

	I^{2}CON	point-to-point	limited point-to-pint
Core	256	256	256
Cluster	64	64	64
Tx/cluster	120	126	112
Rx/cluster	264	126	112

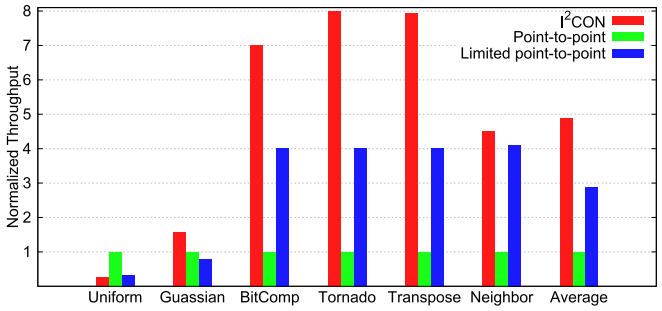

Fig. 8. Maximum throughputs of three designs.

transaction. To estimate the propagation delay of the optical signal, the cluster size is assumed to be 8 mm^2 [30], and the group refractive index of the silicon waveguide is 4.2. In point-to-point and limited point-to-point networks, the clusters are placed in 2-D arrays, the distance between two neighboring cluster is around 1.3 cm [8]. In I^{2}CON, the chip distance is as far as 5 cm, reducing the thermal density effectively.

I^{2}CON can scale to higher WDM channel count by integrating more lasers, microresonators, and photodetectors for each channel, given these devices are wavelength selective. Given the relative abundant area budget on optical layer, more optical devices can be integrated to support higher WDM channel count. Higher bit rate can be achieved by improving the modulation speed at the source and detecting frequency at the receiver with the cost of higher power consumption. We will explore the impact of higher WDM channel count and higher bit rate in our future work.

A. Synthetic Traffics

In performance evaluation, we use both synthetic and real traffics. For the synthetic traffic, six traffic patterns, including uniform, Gaussian, transpose, tornado, bit complement, and neighbor traffics are considered [36]. In all traffics, the packet size is assumed to be a constant value of 512 bits, mimicking a cache line. Under uniform traffic, each cluster will send packets to all other clusters with the same probability. Bisectional bandwidth of the network is the critical factor under this traffic. Under Gaussian traffic, the probability distribution of the destination follows a Gaussian distribution, simulating the locality feature of real traffic. Neighbor traffic only allows packets with one hop, simulating the well-mapped tasks with communications with high locality. The bit complement, transpose, and tornado traffics are all permutation traffics. These traffics will stress the load balance of the network [36].

The throughput comparison is shown in Fig. 8. I^{2}CON outperforms the other two designs under most of traffic patterns except for uniform traffic. I^{2}CON has lower bisectional bandwidth and thus achieves lower throughput under the uniform traffic. Uniform traffic specially favors point-to-point network because all channels can be fully utilized. Under uniform traffic, each cluster will send data to all other clusters with the same possibility, and the data can be transferred through the dedicated channels connecting to each cluster. However, this dedicated channels will suffer from low utilization if the traffic is not uniform since a channel can only serve for a single source–destination traffic but not be shared by other traffics. As shown in Fig. 8, the throughputs of point-to-point network under permutation traffics, including transpose, permutation, and tornado traffic patterns, are very low compared with the other two designs. Under permutation traffics, a cluster will only send data to a fix destination, which implies that most of channels (62 out of 63) are failed to be utilized.

Limited point-to-point network alleviates the problem by using an electrical switching router at each cluster. The packets destined to the clusters in the same column/row can share a channel starting from the source cluster, and the packets destined to the same cluster will share a channel if they are from the same column/row. This sharing can help improve the performance under ununiform traffic as shown in the figure. It achieve less throughput under uniform traffic due to that around 75% packets will transfer two hops and thus the utilization of the links is around 50%. The head-of-line problem at the electrical router further reduces the throughput.

I^{2}CON achieves very stable throughput under all kinds of traffic, implying that the network can well accommodate different traffic patterns. In particular, the throughputs under permutation traffics stay high compared with the other two designs. Under bit-complement traffic, the throughput of I^{2}CON is six times higher than point-to-point network, and 74% higher than limited point-to-point network. Under both tornado and transpose traffics, the throughput of I^{2}CON is eight times of that of the point-to-point network, and twice of the limited point-to-point network. For Gaussian traffic, we set the standard deviation to be four, implying that around 68% of traffic is destined to the neighboring eight clusters. Under this traffic, both point-to-point network and limited point-to-point network achieves lower throughout than them under uniform traffic. The reason is that each cluster will only send data to some but not all other clusters, leaving many channels unused. In contrast, I^{2}CON achieves even higher throughput under Gaussian traffic than under uniform traffic, showing that I^{2}CON favors locality in the traffic. Similar results under neighboring traffic are shown in the figure.

The performance gain of I^{2}CON is mainly achieved by the link sharing and link segmentation. For each inter-chip link, it is shared by all clusters it interconnects, and more importantly, it is shared by all other clusters which try to send packets to these clusters. We can consider this sharing as time-division multiplexing. Segmentation, on the other hand, help to support the sharing by providing multiple concurrent transactions on each link, which we call as space-division multiplexing. Similar case is for intra-chip links. A link connects all clusters on the chip, and the segmentation applies on

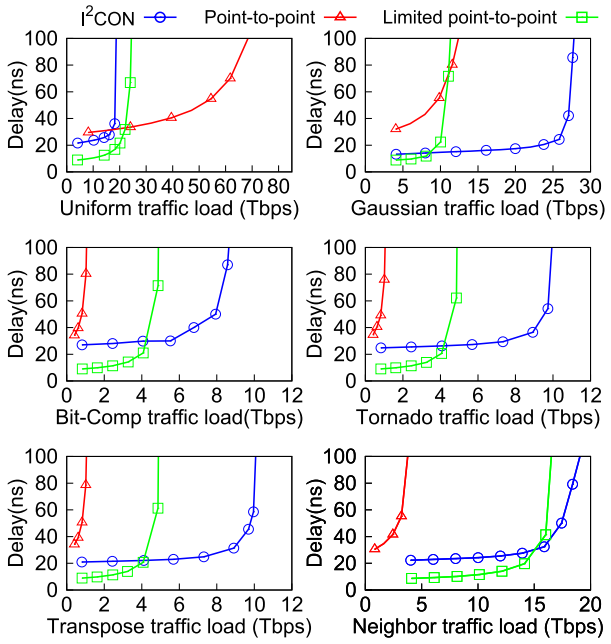

Fig. 9. Average latency versus the offered load for synthetic traffics.

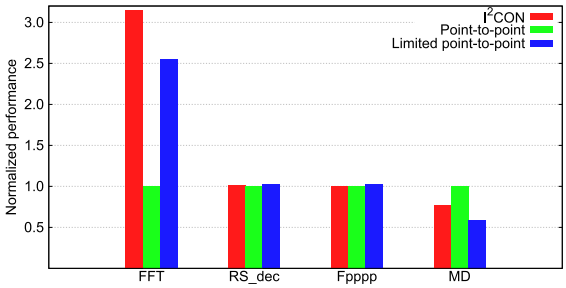

Fig. 10. Performance comparison under real applications.

the link. Under segmentation, spacial locality is favored since a link can support more transactions if the transactions are with short distance. For example, a link connects N clusters can support N concurrent transactions at most. This explains that I^{2}CON achieves higher throughput under Gaussian and neighboring traffics.

The average latency of the packets in all three designs is shown in Fig. 9. The delay climbs dramatically after saturation point for all designs, and the figure also shows that point-to-point network has more than twice zero-load latency compared with I^{2}CON and limited point-to-point network. Larger delay comes from larger serialization latency since that each channel in the point-to-point is with the bandwidth of only one fourth of the other two.

B. Real Applications

Besides synthetic traffics, real applications are also used in the evaluation. We adopt the MCSL NoC benchmark suits [37], and the included applications in the evaluation are fast Fourier transform (FFT), Reed–Solomon code decoder (RS_dec), SPEC95 Fpppp (FPPPP), and molecular-dynamics (MD) simulation.

The performance results are shown in Fig. 10. Under FFT traffic, the performance of I^{2}CON is 214% higher than that of point-to-point network, and 23% higher than that of limited point-to-point network. Although point-to-point network shows very high throughput under uniform traffic, it achieves very low performance under this real application. That is due to that the traffic loads of this application are far from even. With unbalanced traffic in real applications, the resource sharing and flexibility are critical for the performance achievement. Under MD traffic, the performance of I^{2}CON is 23% lower than point-to-point network but 32% higher than limited point-to-point network. The MD traffic pattern is even and thus it favors the point-to-point network. Under the other two traffics, three networks achieve very similar performances. A closer look at the applications reveals that these traffics are with very low injection rates and do not stress any of the networks.

TABLE II
OPTICAL LOSS

Component	Loss
Laser coupling loss	1 dB
MR passing loss	0.001 dB
Filter drop loss	1.5 dB
Silicon waveguide loss(on thin SOI)	1dB/cm
Routing waveguide loss(on thick SOI)	0.1dB/cm
Polymer waveguide loss(on board)	0.07dB/cm
Splitter	0.2 dB
Coupler	0.45 dB
Bending	0.005dB/90^o

C. Power Consumption

To show the power consumption in the architectures, we adopt the nanophotonic power model proposed in [16] for I^{2}CON. There are various losses along the optical path and we list them in Table II. We assume the receiver power is 50 fJ/bit [35], modulation power is also 50 fJ/bit [38], and thus, the EO/OE conversion power is 100 fJ/bit, which is in the estimation range of [16]. The sensitivity of the photodetector is assumed to be 10 μW in [16], and the photodetector with similar sensitivity has been demonstrated in [39]. Therefore, we have assumed the same photodetector sensitivity. The thermal tuning power is assumed to be 20 μW per MR, and the MR switching power is assumed to be 50 μW per MR as in [16]. The power efficiency of off-chip laser is assumed to be 30% as in [16]. The lasers for I^{2}CON are bonded on-chip, and the power efficiency will degrade due to higher temperature. However, many works [40], [41] show that the power efficiency at 80 °C is still higher than half of the efficiency at room temperature. Here, we assume that the power efficiency of on-chip laser is only half of the off-chip laser. The coupler between laser and waveguide had been fabricated with a loss of 1.5 dB in [33], and the simulated result showed 0.3-dB loss could be achieved. We assume the loss is 1 dB for three networks. Besides these on-chip components, the single-mode polymer waveguide and the coupler are required for inter-chip communication. The polymer waveguide loss is 0.07 dB/cm [42]. In [32], a coupler with loss around 0.8 dB has been fabricated, and it is explained

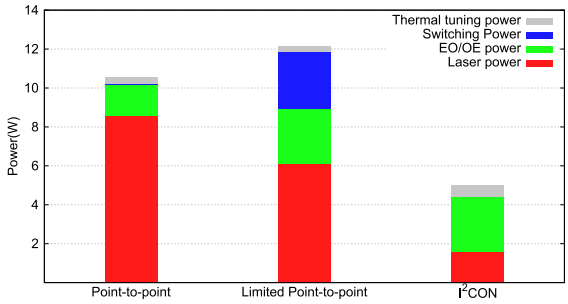

Fig. 11. Power breakdown of three designs under uniform traffic.

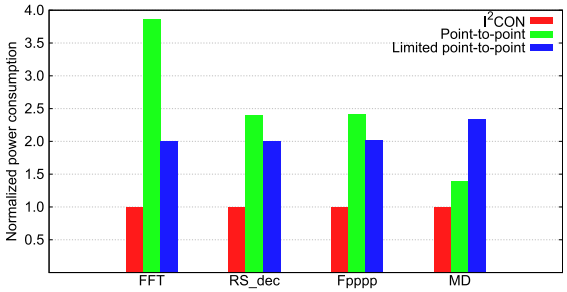

Fig. 12. Power consumptions of three designs under real applications.

that further improvement can be made by optimizing the design parameters. Shu *et al.* [31] shows that the loss can be smaller than 0.4 dB. In our design, we assume the coupler loss is 0.45 dB. For point-to-point network and limited point-to-point networks [8], the special routing waveguides on thick SoI are with smaller loss than general waveguides as shown in the table. The coupling loss between routing layers in these two networks is also assumed as 0.45 dB. We have also synthesized the cluster agent with 45-nm library and scaled it to 17 nm. It runs at 5 GHz, consuming 213 μW with a switching rate of 15%. The area is 3517 μm^2.

We analyze the power consumption of all architectures under uniform traffic with injection rate of 0.1, that is, in average, each cluster injects 51.2 bits into the network. Fig. 11 shows the power breakdown of the three designs. For the total power consumption, I^{2}CON saves 52% and 58% of energy comparing with point-to-point and limited point-to-point networks, respectively. The high energy efficiency of I^{2}CON is mainly contributed by the low power consumption of lasers, even though we have assumed the power efficiency of on-chip laser (I^{2}CON) is only half of the off-chip laser (the other two designs). The low power requirement for the lasers in I^{2}CON is achieved by efficient segmentation and bidirectional transmission technologies, which drastically reduce the power loss on the path. Also, the on-chip laser can be powered OFF when no data are transferred. This cannot be done if off-chip laser is adopted as the case in the other two designs. Another significant portion of the power is consumed in the OE/EO conversion. This implies that OE/EO conversion should be minimized as much as possible. Given that optical signal cannot be buffered or processed easily, OE/EO conversion is generally required on the path in the optical network for buffering, processing, and switching. The conversion helps to set up a network with flexibility. Both I^{2}CON and limited point-to-point networks require two OE/EO conversions for a large portion of the packets and thus, they consume more conversion power than point-to-point network. On the other hand, the flexibility of point-to-point network is limited as shown in the previous performance evaluations. In limited point-to-point network, electrical switching is required and it also consumes significant portion (23%) of the total power. No such kind of switching is required in I^{2}CON and limited point-to-point network.

Fig. 12 shows the power comparison of three networks under real applications. I^{2}CON achieves lowest power consumption for all applications. Under RS_dec and FPPPP traffics, I^{2}CON saves 57% and 51% power compared with limited point-to-point and point-to-point networks. Under MD traffic, I^{2}CON reduces 28% and 57% power compared with the other two networks, respectively. Under FFT applications, I^{2}CON can finish the task much faster and thus can achieve even lower power consumption by reducing the static power.

TABLE III
VALUE CHANGES OF COMPONENTS

Component	Original	Alternative
MR passing loss(MR_P)	0.001 dB	0.1 dB
MR drop loss(MR_D)	1.5 dB	0.5 dB
Inter-layer coupler loss(CP)	0.45 dB	1.4 dB
Silicon waveguide loss(S_WG)	1dB/cm	0.274 dB/cm
Polymer waveguide loss(P_WG)	0.07dB/cm	0.5 dB/cm
Photodetector sensitivity(PD)	-20 dBm	-14.2 dBm
On-chip Laser efficiency(Laser)	15%	10%
OE/EO conversion power(OE/EO)	100 fJ/bit	500fJ/bit
Tuning power(Tuning)	20 μW	100 μW

D. Silicon Photonic Technology Discussion

Silicon photonic technology is still in the early stage that, a comprehensive optical on-chip network has not been fabricated yet, though individual devices has already been demonstrated with potentially very good property. In our network design, we have tried to select the technology-compatible devices such that they can work together properly. There are alternative technologies, and thus in this section, we discuss how the network power is affected by changing the selection of devices. Table III is a summary of all changes, and Fig. 13 shows re-evaluation results corresponding to the changes.

We first consider how the power loss of the optical components can change. MR passing loss generally is very small. However, the loss can be accumulated easily by a large number of MRs on the path in a network. Here, we first increase MR passing loss from 0.001 [6] to 0.1 dB [8]. Evaluation results show that 2.4-W more power is required for I^{2}CON. Similar results are shown for the other two networks, given all networks are with same WDM channel count. Light switching with MR is common in the network, and thus, the MR drop loss can be significant in deciding the network power. We change the MR drop loss from 1.5 to 0.5 dB as demonstrated in [43]. Around 20% of power can be saved in I^{2}CON and limited point-to-point networks. We assumed the coupling loss between silicon and polymer waveguide is 0.45 dB. This coupling loss value was also

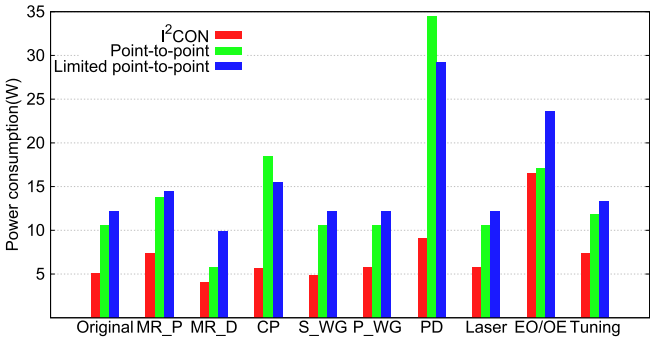

Fig. 13. Network power consumptions with different component value. The labels at x-axis denote the value changes listed in Table III.

assumed for the other two networks for interlayer coupling. Soganci *et al.* [32] demonstrated a coupler with 0.8-dB loss, and the ±2-um lateral misalignment can induce extra 0.6-dB loss. Here, we assume the coupling low is 1.4 dB for all three networks. I^{2}CON consumes 12% more power, while point-to-point network requires 75% more power. The relative low power increment in I^{2}CON is due to that only small number of packets require coupling multiple times. The silicon waveguide propagation loss can be reduced from 1 to 0.274 dB/cm with the cost of larger waveguide width [11]. On the other hand, polymer waveguides are generally fabricated with around 0.5-dB/cm loss [32], [44]. We use these values in new power model. The reduction of silicon waveguide loss shows not much power saving for I^{2}CON due to the short length of on-chip data channel. On the other hand, increasing the polymer waveguide loss shows 13% increment of power consumption.

Although the photodetector with sensitivity of −18.9 dbm has been demonstrated [39], there are many photodetectors with much lower sensitivity. Here, we assume the sensitivity is −14.2 dbm [10]. The large sensitivity difference will cause much higher optical power requirement as shown in the figure. However, I^{2}CON shows much lower increment compared with two other networks, due to that power loss has been sufficiently reduced. Similarly, if we decrease on-chip laser efficiency from 15% to 10%, 15% more power would be consumed by I^{2}CON. For EO/OE conversion power, although 50-fJ/bit modulation power has been demonstrated in [38], the fabricated photodetector is still at around 690 pJ/bit [39]. Here, we assume the power is 500 pJ/bit. The power consumption of all three networks increases significantly. Finally, the MR tuning power is assumed 100 uW/ring [35], instead of 20 uW/ring as assumed before. 47.5% more power for I^{2}CON is required due to the tuning power is a significant portion of total power consumption, as shown in Fig. 11.

The analysis above shows that, the power consumption of I^{2}CON is not very sensitive to the parameter change, and more importantly, I^{2}CON always outperforms the other two designs in terms of power consumption.

VII. Conclusion

The advances in nanophotonics have motivated us to exploit the benefits of optical interconnects for future manycore processor with a large number of cores. In this paper, we propose an inter/intra-chip optical network called I^{2}CON, which supports high-throughput and low-latency communication for the multichip system. The proposed network effectively explores the distinctive properties to boost performance as well as reduce energy consumption. The comparison with the alternatives shows that I^{2}CON achieves promising throughput with good energy efficiency.

References

[1] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, May/Jun. 2012.

[2] D. Vantrease *et al.*, "Corona: System implications of emerging nanophotonic technology," in *Proc. 35th ISCA*, Jun. 2008, pp. 153–164.

[3] I. O'Connor, "Optical solutions for system-level interconnect," in *Proc. 2004 Int. Workshop SLIP*, New York, NY, USA, pp. 79–88.

[4] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi, "Phastlane: A rapid transit optical routing network," in *Proc. 36th Annu. ISCA*, New York, NY, USA, 2009, pp. 441–450.

[5] A. Shacham, K. Bergman, and L. P. Carloni, "Photonic networks-on-chip for future generations of chip multiprocessors," *IEEE Trans. Comput.*, vol. 57, no. 9, pp. 1246–1260, Sep. 2008.

[6] Y. Pan, J. Kim, and G. Memik, "FlexiShare: Channel sharing for an energy-efficient nanophotonic crossbar," in *Proc. IEEE 16th Int. Symp. HPCA*, Jan. 2010, pp. 1–12.

[7] S. Bartolini and P. Grani, "A simple on-chip optical interconnection for improving performance of coherency traffic in CMPs," in *Proc. 2012 15th Euromicro Conf. DSD*, Sep., pp. 312–318.

[8] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy, "Silicon-photonic network architectures for scalable, power-efficient multi-chip systems," in *Proc. 37th Annu. ISCA*, New York, NY, USA, 2010, pp. 117–128.

[9] S. P. Q. Xu, B. Schmidt, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," *Nature*, vol. 435, pp. 325–327, May 2005.

[10] G. Masini, G. Capellini, J. Witzens, and C. Gunn, "A 1550 nm, 10 Gbps monolithic optical receiver in 130 nm CMOS with integrated Ge waveguide photodetector," in *Proc. 4th IEEE Int. Conf. Group IV Photonics*, Sep. 2007, pp. 1–3.

[11] P. Dong *et al.*, "Low loss silicon waveguides for application of optical interconnects," in *Proc. 2010 IEEE Photonics Society Summer Topical Meeting Series*, Jul., pp. 191–192.

[12] Y. Rao, "InP-based long wavelength VCSEL using high contrast grating," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Dec. 2012.

[13] N. Kirman *et al.*, "Leveraging optical technology in future bus-based chip multiprocessors," in *Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Washington, DC, USA, 2006, pp. 492–503.

[14] Y. Xu, Y. Du, Y. Zhang, and J. Yang, "A composite and scalable cache coherence protocol for large scale CMPs," in *Proc. ICS*, New York, NY, USA, 2011, pp. 285–294.

[15] S. Pasricha and N. Dutt, "ORB: An on-chip optical ring bus communication architecture for multi-processor systems-on-chip," in *Proc. ASPDAC*, Mar. 2008, pp. 789–794.

[16] A. Joshi *et al.*, "Silicon-photonic clos networks for global on-chip communication," in *Proc. 3rd ACM/IEEE Int. Symp. Networks-on-Chip*, May 2009, pp. 124–133.

[17] Z. Li *et al.*, "Spectrum: A hybrid nanophotonic—Electric on-chip network," in *Proc. 46th ACM/IEEE DAC*, New York, NY, USA, Jul. 2009, pp. 575–580.

[18] J. Ouyang, C. Yang, D. Niu, Y. Xie, and Z. Liu, "F2BFLY: An on-chip free-space optical network with wavelength-switching," in *Proc. ICS*, New York, NY, USA, 2011, pp. 348–358.

[19] J. Psota *et al.*, "ATAC: Improving performance and programmability with on-chip optical networks," in *Proc. 2010 IEEE ISCAS*, May/Jun., pp. 3325–3328.

[20] S. Koohi, M. Abdollahi, and S. Hessabi, "All-optical wavelength-routed NoC based on a novel hierarchical topology," in *Proc. 5th 2011 IEEE/ACM Int. Symp. Networks Chip*, May, pp. 97–104.

[21] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, "Firefly: Illuminating future network-on-chip with nanophotonics," in *Proc. 36th ISCA*, 2009, vol. 37, no. 3, pp. 429–440.

[22] S. L. Beux, J. Trajkovic, I. O'Connor, G. Nicolescu, G. Bois, and P. Paulin, "Optical ring network-on-chip (ORNoC): Architecture and design methodology," in *Proc. DATE Conf. Exhibition*, Mar. 2011, pp. 1–6.

[23] Y. Xu, J. Yang, and R. Melhem, "Channel borrowing: An energy-efficient nanophotonic crossbar architecture with light-weight arbitration," in *Proc. 26th ACM ICS*, New York, NY, USA, 2012, pp. 133–142.
[24] R. W. Morris, A. Kodi, A. Louri, and R. Whaley, "Three-dimensional stacked nanophotonic network-on-chip architecture with minimal reconfiguration," *IEEE Trans. Comput.*, vol. 63, no. 1, pp. 243–255, Jan. 2014.
[25] I. Datta, D. Datta, and P. P. Pande, "BER-based power budget evaluation for optical interconnect topologies in NoCs," in *Proc. 2012 IEEE ISCAS*, May, pp. 2429–2432.
[26] G. Van Steenberge *et al.*, "MT-compatible laser-ablated interconnections for optical printed circuit boards," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2083–2090, Sep. 2004.
[27] S. H. Hwang *et al.*, "Passively assembled optical interconnection system based on an optical printed-circuit board," *IEEE Photon. Technol. Lett.*, vol. 18, no. 5, pp. 652–654, Mar. 1, 2006.
[28] A. Apsel, Z. Fu, and A. G. Andreou, "A 2.5-mW SOS CMOS optical receiver for chip-to-chip interconnect," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2149–2157, Sep. 2004.
[29] C. Berger, M. Kossel, C. Menolfi, T. Morf, T. Toifl, and M. Schmatz, "High-density optical interconnects within large-scale systems," *Proc. SPIE*, vol. 4942, pp. 222–235, Apr. 2003.
[30] A. V. Krishnamoorthy *et al.*, "Computer systems based on silicon photonic interconnects," *Proc. IEEE*, vol. 97, no. 7, pp. 1337–1361, Jul. 2009.
[31] J. Shu, C. Qiu, X. Zhang, and Q. Xu, "Efficient coupler between chip-level and board-level optical waveguides," *Opt. Lett.*, vol. 36, no. 18, pp. 3614–3616, Sep. 2011.
[32] I. M. Soganci, A. L. Porta, and B. J. Offrein, "Flip-chip optical couplers with scalable I/O count for silicon photonics," *Opt. Exp.*, vol. 21, no. 13, pp. 16075–16085, Jul. 2013.
[33] C. Gunn, "CMOS photonics for high-speed interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58–66, Mar./Apr. 2006.
[34] Y. Xie *et al.*, "Formal worst-case analysis of crosstalk noise in mesh-based optical networks-on-chip," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 10, pp. 1823–1836, Oct. 2013.
[35] J. Chan, G. Hendry, K. Bergman, and L. P. Carloni, "Physical-layer modeling and system-level design of chip-scale photonic interconnection networks," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 10, pp. 1507–1520, Oct. 2011.
[36] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Mateo, CA, USA: Morgan Kaufmann, 2003.
[37] W. Liu *et al.*, "A NoC traffic suite based on real applications," in *Proc. IEEE Computer Society Annu. Symp. VLSI*, Jul. 2011, pp. 66–71.
[38] P. Dong *et al.*, "Low V_{pp}, ultralow-energy, compact, high-speed silicon electro-optic modulator," *Opt. Exp.*, vol. 17, no. 25, pp. 22484–22490, Dec. 2009.
[39] X. Zheng *et al.*, "A sub-picojoule-per-bit CMOS photonic receiver for densely integrated systems," *Opt. Exp.*, vol. 18, no. 1, pp. 204–211, 2010.
[40] W. Hofmann *et al.*, "22-Gb/s long wavelength VCSELs," *Opt. Exp.*, vol. 17, no. 20, pp. 17547–17554, Sep. 2009.
[41] A. Syrbu, A. Mereuta, V. Iakovlev, A. Caliman, P. Royo, and E. Kapon, "10 Gbps VCSELs with high single mode output in 1310 nm and 1550 nm wavelength bands," in *Proc. OFC/NFOEC*, Feb. 2008, pp. 1–3.
[42] S. Uhlig and M. Robertsson, "Limitations to and solutions for optical loss in optical backplanes," *J. Lightw. Technol.*, vol. 24, no. 4, pp. 1710–1724, Apr. 2006.
[43] S. Xiao, M. H. Khan, H. Shen, and M. Qi, "Multiple-channel silicon micro-resonator based filters for WDM applications," *Opt. Exp.*, vol. 15, no. 12, pp. 7489–7498, Jul. 2007.
[44] L. Eldada and L. W. Shacklette, "Advances in polymer integrated optic," *IEEE J. Sel. Topics Quantum Electron.*, vol. 6, no. 1, pp. 54–68, Jan./Feb. 2000.

Xiaowen Wu (S'12) received the B.Sc. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008. He is currently pursuing the Ph.D. degree in electronic and computer engineering with the Hong Kong University of Science and Technology, Hong Kong.

His current research interests include embedded systems, multiprocessor systems, and network-on-chip.

Jiang Xu (S'02–M'07) received the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2007.

He was with Bell Labs, NJ, USA, as a Research Associate, from 2001 to 2002. He was a Research Associate with NEC Laboratories America, NJ, USA, from 2003 to 2005. He joined Sandbridge Technologies, Tarrytown, NY, USA, from 2005 to 2007, where he developed and implemented two generations of network-on-chip (NoC)-based ultralow-power multiprocessor systems-on-chip (SoC) for mobile platforms. He is an Associate Professor with the Hong Kong University of Science and Technology, Hong Kong, where he is the Founding Director with the Xilinx-HKUST Joint Laboratory, and establishes Mobile Computing System Laboratory. He has authored and co-authored more than 70 book chapters and papers in peer-reviewed journals and international conferences. His current research interests include NoC, multiprocessor SoC, optical interconnects, embedded system, computer architecture, low-power VLSI systems design, and hardware–software codesign.

Dr. Xu serves as the Area Editor of NoC, SoC, and GPU for *ACM Transactions on Embedded Computing Systems* and an Associate Editor of the IEEE TRANSACTION ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He is an ACM Distinguished Speaker and an IEEE Distinguished Lecturer. He served on the steering committees, organizing committees, and technical program committees of many international conferences.

Yaoyao Ye (S'09) received the B.S. degree in electronic information science and technology from the University of Science and Technology of China, Hefei, China, in 2008. She is currently pursuing the Ph.D. degree in electronic and computer engineering with the Hong Kong University of Science and Technology, Hong Kong.

Her current research interests include network-on-chip, multiprocessor system-on-chip, and embedded system.

Xuan Wang (S'12) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

His current research interests include embedded system, multiprocessor system, network-on-chip, and fault tolerant design and reliability issues in very deep submicrometer technologies.

Mahdi Nikdast (S'10) received the B.Sc. (Hons.) degree in computer engineering from Islamic Azad University, Esfahan, Iran, in 2009. He is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

His current research interests include embedded systems, multiprocessor system-on-chip, network-on-chip, and computer architecture.

Dr. Nikdast was a recipient of the Second Best Project Award from the 6th Annual AMD Technical Forum and Exhibition in 2010.

Zhehui Wang (S'11) received the B.S. degree in electrical engineering from Fudan University, Shanghai, China, in 2010. He is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.

His current research interests include embedded system, multiprocessor systems, network-on-chip, and floorplan design for network-on-chip.

Zhe Wang (S'14) received the B.S. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. He is currently pursuing the Ph.D. degree in electronic and computer engineering with the Hong Kong University of Science and Technology (HKUST), Hong Kong.

He is currently with the Mobile Computing System Laboratory, HKUST. His current research interests include embedded systems, multiprocessor system-on-chip, network-on-chip, hardware/software co-design, and design space exploration techniques.