RONet: Scaling GPU System with Silicon Photonic Chiplet

Chengeng Li[1,†], Fan Jiang[1,†], Shixi Chen[1], Xianbin Li[1], Yinyi Liu[1], Lin Chen[1], Xiao Li[1] and Jiang Xu[1,2,*]
[1]*Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology*
[2] *Microelectronics Thrust, The Hong Kong University of Science and Technology (GZ)*

2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD) | 979-8-3503-2225-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCAD57390.2023.10323762

***Abstract*—Modern GPU systems integrate hundreds of SMs on a single die, and future scaling envisions even more SMs being incorporated. However, the limited number of transistors per die constrains this growth. While current chiplet technology shows promise, its performance is limited by the bandwidth and energy efficiency of existing chiplet interconnect technologies. In contrast, optical interconnects offer ultra-high bandwidth and energy efficiency, making them ideal for high-performance chiplet-based GPUs. This work proposes a novel region-based optical network, called RONet, that divides a chiplet-based GPU system with a 2D Mesh layout into multiple row and column regions, where each region is connected by a separate optical link. Additionally, RONet employs a tuning-free transmission mechanism to further enhance inter-chiplet bandwidth. Experimental results show that RONet achieves 43% improvement on performance and 25.4% reduction on system energy consumption over the baseline.**

***Index Terms*—silicon photonic chiplet, large-scale GPU**

I. Introduction

In the era of big data, machine learning, and other data-intensive applications, parallel workloads demonstrate significant parallelism and have an increasing demand for computational resources [2]. GPU-based computing acceleration is the main alternative driving the performance of high-performance computing (HPC) systems. Commercial GPUs have rapidly evolved over the past decade, with the number of SMs increasing from 16 to 108 [3] and the potential for hundreds or even thousands of SMs in the future. However, integrating additional SMs into a monolithic silicon die presents significant challenges such as integration density, cost, and yield [4].

Several chiplet packaging and interconnect technologies have been proposed, including MCMs, 2.5D integration, and silicon bridges [5]–[9]. Unfortunately, these methods are not optimal alternatives for chiplet-based GPUs for one or more of the following reasons. First, a large number of parallel workloads are memory-intensive, requiring a large inter-chiplet communication bandwidth. However, the inter-chiplet bandwidth cannot meet the GPU requirement due to limited pin density and data rate. Second, the dependency of energy consumption on interconnect length, coupled with centimeter-scale chiplet sizes, only allows interconnecting adjacent chiplets. This leads to high-diameter topologies with high average hop counts where each inter-chiplet hop imposes tens of nanoseconds latency.

Integrated optical interconnects offer properties that can be exploited to overcome the aforementioned challenges of electrical inter-chiplet interconnects [10], [11]. Optical interconnects offer much higher bandwidth density than electrical interconnects especially when wavelength-division multiplexing (WDM) is applied, enabling the transmission of multiple wavelengths in parallel in the same optical link, which promises the bandwidth requirement for GPUs. Optical interconnects support long-distance point-to-point transmission even between two not adjacent chiplet without the need for repeaters or multi-hop transmission, which enables to connect many GPU chiplet together to build a physically-large but logically dense GPU. In addition, the energy consumption of optical interconnects is largely independent of the distance. Thus, optical interconnects are more energy-efficient than electrical interconnects in terms of inter-die communication. All the benefits brought by optical interconnects provide a possibility for large-scale chiplet-based GPU.

Many optical networks [11]–[13] have been proposed in the past, but they cannot be applied to chiplet-based GPU directly due to the following reasons. Firstly, it is crucial to investigate the communication characteristics specific to chiplet-based GPU systems. Optical networks should be customized to these requirements, rather than merely substituting a few electrical links with high-bandwidth optical ones. Additionally, these proposals prioritize the optimization of energy consumption caused by micro resonator (MR) loss and disregard the energy consumption resulting from coupling loss, which is dominant in large-scale chiplet-based GPUs. If not adequately designed, the coupling loss in large-scale chiplet-based GPUs can reach tens or even hundreds of dB, resulting in unacceptably high laser power consumption. Furthermore, previous optical networks utilize a pre-tuning mechanism to reduce the number of receivers and further lower the power consumption of optical networks [14]. Nevertheless, this mechanism results in additional tuning latency, and packets need to be transmitted one by one, decreasing the optical channel utilization and constraining the maximum bandwidth of optical channels.

Our work aims to design a high-performance large-scale GPU through silicon photonic chiplet. We propose a **r**egion-based **o**ptical **net**work, called **RONet**, which effectively enhances inter-chiplet bandwidth and improve system performance. Experimental results show that RONet achieves 43%

† Chengeng Li and Fan Jiang are co-first authors.
* Corresponding author: jiang.xu@ust.hk.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

improvement on performance and 25.4% reduction on system energy consumption over the baseline. Specificlly, our contribution are:

- We systematically analyze traffic distribution and inter-chiplet bandwdith requirement for chiplet-based GPU.
- We propose a novel region-based optical network with tuning-free mechanism for chiplet-based GPU, which effectively improves the system scalability and optical link bandwidth.
- We design an optical channel allocation policy and optical channel mapping scheme for bandwidth balance.
- We quantitatively compare the performance, memory access latency, energy consumption, and scalability of RONet with two representative inter-chiplet networks.

The rest of the paper is organized as follows: Sec. II introduces optical interconnect background. Sec. III analyzes the challenges of designing chiplet-based GPU. Sec. IV details our RONet architecture. Sec. V quantitatively evaluates RONet. Sec. VI discusses related works. Sec. VII makes the conclusion.

II. Optical Device Background

Fig. 1 illustrates the structure of a typical optical interconnect, which includes four main parts: light resource, E-O interface, optical link ,and O-E interface. The light resource is generated by the off-chip laser bank, which reduces the on-chip power and area burden compared to an on-chip laser. At the sender end, electrical signals are imprinted into the laser lights through an E-O interface that absorbs and passes the light for signal '0' and '1' respectively. We adopt optical fiber as the optical link for the optical signal, which ensures tremendously low optical propagation loss over the long-distance transmission (0.2 dB/Km). In addition, optical fiber supports Wavelength Division Multiplexing (WDM) technology, which allows laser lights of multiple wavelengths to be transmitted in parallel inside the same fiber, greatly improving network bandwidth. Fibers are coupled with the chiplets through couplers, each of which leads to about 1.5 dB coupling loss. Thus, we need to minimize the number of chiplets coupled with an optical link. We group several laser lights as an optical channel that can transmit the whole packet in one cycle without the need for packet segmentation. At the receiver end, the laser light with a specific wavelength is

TABLE I: Optical device parameters

Parameter	Value
MR passing loss	0.01 dB
MR dropping loss	1 dB
MR heat tuning power	0.65 mW
Waveguide propagation loss	0.5 dB/cm
Optical pin coupling loss	1 dB
Receiver sensitivity	-20 dBm
Laser power conversion efficiency	25%
Data rata per wavelength	32 $Gbps$

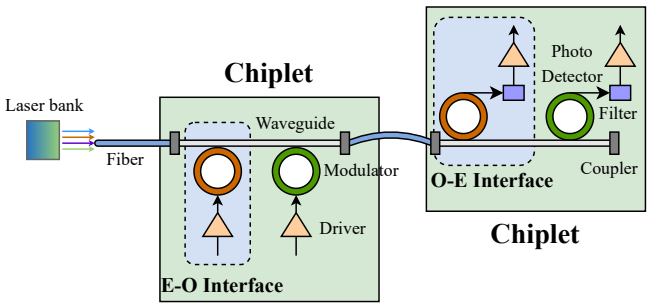


Fig. 1: The structure of optical interconnect.

extracted by the optical filter and then transfer to a photo-detector. The filter is implemented by micro resonator (MR) that can switch laser light with a specific wavelength. Finally, the photo-detector in the O-E interface converts optical signals into electrical ones which are then passed to amplifiers. To lower the E-O/O-E transmission latency, we adopt the optical weaving E-O/O-E interface [15]. Tab. I shows the important parameters of optical devices used in RONet. Based on the optical device parameter and some previous work [16]–[18], [36], we estimate the latency of an optical transmission in electrical clock domain: 3 cycles for E-O conversion, 2 cycles for optical signal propagation and 2 cycles for O-E conversion.

III. Motivation

To facilitate a large-scale system design, chiplet technology is now applied in GPU system [19]. The GPU system is divided into several modules and each module is fabricated into a separate chiplet. Each chiplet contains serveral SMs with their L1s and memory partitions consisting of L2s and memory controllers (MC), which are connected using an intra-chiplet crossbar. All the chiplets are connected through an inter-chiplet electrical-based network. However, designing a high-performance and low-power large-scale chiplet-based GPU faces many challenges.

A. Bandwidth requirement

In chiplet-based GPU systems, all memory partitions among the chiplets typically provide a globally shared memory address space. Addresses are finely interleaved across the physical memory partitions, allowing the operating system and programmers to be isolated from the fact that a single logical GPU may consist of multiple GPU chiplets working together. However, this design also leads to a significant amount of inter-chiplet memory traffic as chiplets need to fetch data from memory partitions located on other chiplets. In a 16-chiplet GPU system, assuming an equal probability of accessing each memory partition, we know that approximately 15/16 of the packets are inter-chiplet packets for a specific chiplet. Our experiments confirm this observation, with over 90% of the traffic being inter-chiplet traffic, albeit with slight variations across different applications. The predominant inter-chiplet packets exert considerable pressure on the inter-chiplet network which could become a bottleneck if without careful

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Summary of state-of-the-art chiplet interconnect techniques

Inteconnect Technology	Multi-chip Module [19]	2.5D Interposer [33]	Silicon Bridge [31]	Silicon Photonic [32]
Pin Pitch (um)	6	2	2	5
Pin Bandwidth ($Gbps$)	20	28	28	hundreds even thousands
Energy efficiency ($pJ/bit/Gbps$)	0.027 (4.5 mm)	0.0114 (3.5 mm)	0.035 (1 mm)	0.017(several cm)

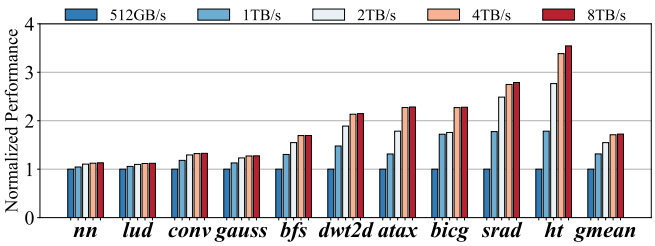


Fig. 2: Relative performance sensitivity to inter-chiplet link bandwidth for a 16-chiplet, 512-SM GPU system.

design. Moreover, the memory-intensive nature of GPU applications exacerbates this issue. To gain a deeper understanding of the impact of the inter-chiplet network on system performance, we analyze performance in relation to the inter-chiplet bandwidth. We increase the inter-chiplet link bandwidth of the baseline electrical chiplet-based GPU system from 512 GB/s to 8 TB/s and plot the performance improvement in Fig. 2. The detailed configuration of the baseline system is provided in Sec. V-A. The results indicate that system performance steadily increases with the increment in inter-chiplet link bandwidth. Notably, a dramatic improvement in performance is observed when the bandwidth increases from 512 GB/s to 4 TB/s. However, as the bandwidth reaches 8 TB/s, most applications' performance improvement plateaus, suggesting that the bandwidth becomes saturated, and further increments no longer yield benefits. We can see that different applications have different sensitivities to bandwidth increment. Although bandwidth-insensitive applications such as nn and lud show relatively modest performance improvement, they still benefit from bandwidth increases. Therefore, it is preferable to design the bandwidth to exceed 4 TB/s. However, for electrical inter-chiplet interconnects, the bandwidth is constrained by pin density, area, and data rate limitations. Achieving an inter-chiplet link bandwidth of 1.5 TB/s for a single electrical chiplet is challenging [19]. As a result, system performance may be limited by the constrained bandwidth. High-bandwidth optical interconnects can overcome these limitations and provide sufficient bandwidth. In this study, we investigate the integration of optical links to design an optical network for chiplet-based GPUs.

B. Energy and scalability

The energy consumption of electrical interconnects is highly dependent on their length, making long-distance (several cm) electrical interconnects unfeasible due to their high energy consumption. Chiplets, which are typically large (approximately 1 cm^2) and laid out on a two-dimensional planar floorplan, cannot be directly connected with a single electrical link if they are not adjacent, due to the long distance. To address this issue, most proposed chiplet-based system designs use a high-diameter network, such as Ring and Mesh, which leads to multi-hop transmission and results in inter-chiplet communication energy consumption being proportional to the transmission hops, making it not scalable in a large-scale chiplet-based system. However, unlike electrical links, optical links have ultra-high bandwidth density, and the energy consumption of optical interconnects is relatively independent of distance, enabling a chiplet to connect to many other chiplets without inter-chiplet bandwidth limitations. Additionally, a chiplet can connect to a non-adjacent chiplet with low energy consumption, making the chiplets logically close. Tab. II provides a summary and comparison of the optical inter-chiplet interconnect and some typical state-of-art electrical inter-chiplet interconnects. From Tab. II, we can observe that the electrical inter-chiplet interconnect is limited to millimeter-level while optical interconnect can support centimeter-level communication with high energy efficiency (0.017 $pJ/bit/Gbps$). In addition, optical interconnect achieves hundreds of GB/s per pin, which is one order of magnitude higher bandwidth density compared to electrical interconnect and can largely alleviate the bandwidth bottle of chiplet-based GPU.

IV. RONet Architecture

In this section, we introduce the overview of the RONet architecture, and then detail three key design points: region-based optical network, optical channel allocation and mapping, and tuning-free mechanism.

A. Architecture overview

Fig. 3 provides an overview of RONet-based 16-chiplet GPU system. Each chiplet comprises 32 SMs and 8 memory partitions. Throughout the rest of the paper, we will use the terms "memory partition" and "L2" interchangeably. Within each chiplet, there is a 32×18 intra-chiplet crossbar that connects the SMs to a set of 18 ports (explained in detail in Section IV-C). Among these ports, 2 are intra-ports responsible for communication with local L2 caches within the same chiplet, while the remaining 16 are inter-ports for communication with remote L2 caches in other chiplets. Each chiplet integrates E-O/O-E optical interfaces, and all GPU chiplets are interconnected through our proposed region-based optical network (detailed in Sec. IV-B) to support efficient inter-chiplet packet transmission.

B. Region-based optical network

A typical topology together with its channel allocation used in previous optical networks [14], [20] is shown in Fig. 4a, where all the chiplets are connected by a single

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

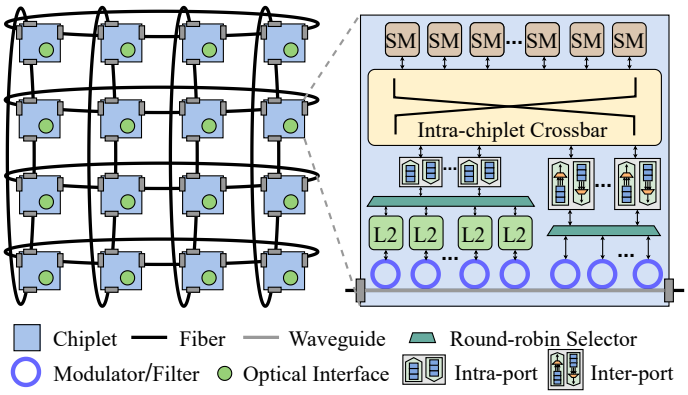


Fig. 3: Overview of a RONet-based 16-chiplet GPU system.

optical link. Each chiplet is assigned several optical channels with a certain number of optical lights, and a packet sent by this chiplet via these optical channels can be received by the other chiplets simultaneously. However, this network is not scalable for large-scale chiplet-based GPUs due to excessive power consumption from the laser. The power consumption of laser is influenced by optical loss, which can be calculated by Euq. 1, where P_l denotes the laser power, P_s denotes detector sensitivity,$loss_f$ denotes the fiber optical loss, $loss_{MR}$ denotes MR loss, $loss_{coup}$ denotes coupling loss.

$$P_l = P_s/E_l * 10^{(loss_{MR}+loss_f+loss_{coup})} \quad (1)$$

As shown in the equation, the lase power increases exponentially with optical loss which includes fiber optical loss, MR loss, and coupling loss. The fiber optical loss is less than 0.1 dB, which is negligible. MR loss and coupling loss are given by Equ. 2 and Equ. 3, respectively, where $loss_d$ denotes MR dropping loss, W denotes the number of wavelength per optical channel, N denotes the number of chiplet, $loss_p$ denotes MR passing loss, $loss_c$ denotes optical pin coupling loss.

$$loss_{MR} = 2 * loss_d + (W * N^2 - 2) * loss_p \quad (2)$$

$$loss_{coup} = 2 * N * loss_c \quad (3)$$

If we use a single optical link to connect all of the chiplets, the loss increases linearly and the laser power rises exponentially with the system scale, even if channel sharing [32] is employed, as shown in Fig. 4c and Fig. 4d. Notably, the energy consumption per bit becomes unacceptable when the number of chiplets exceeds 16. We can observe when the chiplet number reaches 16, the laser energy consumption even approaches 100 pJ/bit. Thus, an optical network should be carefully designed to reduce the number of chiplets connected into a single optical link. According to the above power model and analysis, we propose a scalable region-based optical network for chiplet-based GPUs, as demonstrated in Fig. 4b with a 16-chiplet example.

To reduce the optical energy consumption and loss, we divide the chiplets into several row and column regions. From Fig. 4b, we can see there are eight regions and each region has four chiplets, which is connected by a separate optical

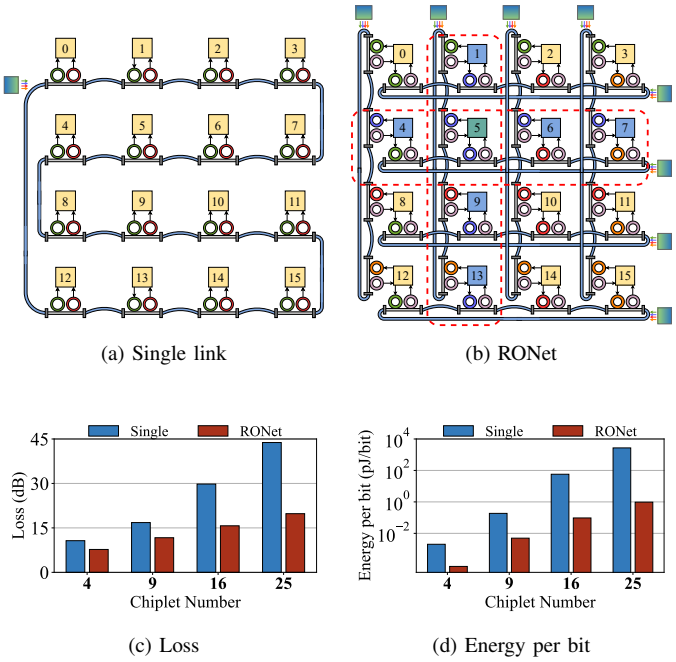


Fig. 4: (a), (b) illustrate topology and channel allocation policy of conventional optical network and RONet, respectively. (c), (d) plot the optical loss and laser energy consumption per bit of conventional optical network and RONet, respectively.

link. Each chiplet is affiliated with two regions, a row region and a column region. We denote the chiplet inside the same region as logical neighbors. Thus, a chiplet has six neighbors. As illustrated in Fig. 4b, take chiplet 5 as an example, the neighbors of it are chiplet 1, 4, 6, 7, 9, and 13. A chiplet can communicate with the neighbors directly using the regional optical network. On the other hand, in those cases that a chiplet wants to communicate with those chiplets belonging to different different regions, the packet is first forwarded to an intermediate node and then sent to the destination.

The routing algorithm of RONet is illustrated in Fig. 5, which aims to minimize the link traversal and always choose the links that keep the energy and latency as low as possible. We classify the routing into three cases based on the source and destination.

- case 1: The source and destination are located in the same chiplet. The packet are sent through the intra-chiplet crossbar network directly.
- case 2: The source and destination are located in the same row or column region, as shown in Fig. 5a. In the case, the packet is first sent to the E-O interface via inter-port, and then sent to the O-E interface at the destination chiplet via the row or column optical link. Finally, the packet is sent to the destination through intra-chiplet crossbar.
- case 3: The source and destination are neither in the same row region nor in the same column region, as shown in Fig. 5b. In this case, after the source E-O interface receives the packet, the packet is sent to an intermediate chiplet residing in the same row region through the row optical link. Then, the packet is forwarded to the O-E interface

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

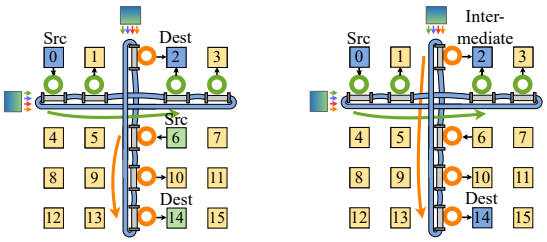


Fig. 5: Two routing cases in RONet: (a) Source and destination are located at the same region. (b) Source and destination are located at different regions.

in the destination chiplet through the column optical link.

In our proposed routing algorithm, each packet undergoes at most two optical transmissions. It is true that some packets may require one additional hop compared to the single optical link proposal. However, considering the low optical transmission latency (less than 10 cycles) and the bandwidth-demanding nature of GPU applications, the latency caused by the extra hop is negligible compared to other delays, such as network queuing delay caused by limited bandwidth.

Our design has ultra-low power consumption compared with conventional optical network for large-scale system. Since we divide the whole system into a couple of regions, the number of chiplet connected to an optical link is reduced. Originally, one optical link is used to connect all of the 16 chiplets. Now we have 8 optical link, where each link is used to connect 4 chiplet. In this way, the loss in an optical transimission is reduced from 29.8 dB to 15.7 dB, and thus the laser power decreases by orders of magnitude.

C. Optical channel allocation and mapping

Channel allocation. To provide sufficient bandwidth for L2 caches in each chiplet, and to reduce network queuing delay, we propose an optical channel allocation scheme for RONet. For each L2, we assign 2 optical channels for inter-chiplet data transmission, one for the row region and the other for the column region. Each optical channel contains 36 laser lights which enables a data packet to be popped out in one cycle. Thus, each chiplet is assigned 16 optical channels in total, which guarantees that there is no contention among L2 caches for optical transmission.

Channel mapping. In a conventional single-die GPU, a crossbar is used to connect SMs and L2 caches. However, in our chiplet-based system, L2 caches are distributed across each chiplet. We refer to the L2 caches located within a specific chiplet as local L2 caches and those in other chiplets as remote L2 caches. Consequently, the crossbar is responsible for communication with both local and remote L2 caches in chiplet-based system. To facilitate this communication, we have designed 18 output ports: 2 for local communication (called intra-port) and 16 for remote communication (called inter-port). The 2 intra-ports are shared by the 8 local L2 caches, while the 16 inter-ports are linked to optical channels.

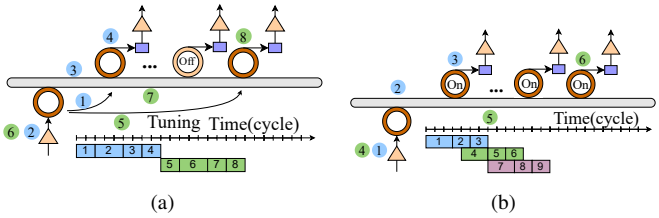


Fig. 6: The transmission process and timing diagram of (a) Conventional optical link. (b) Our tuning-free optical link.

In RONet, each chiplet is assigned 8 optical channels for packet transmission within each region, as mentioned earlier. Therefore, in an $n \times n$ 2D Mesh chiplet-based GPU system, a chiplet can receive packets from other $2 \times (n-1)$ neighbors via $2 \times 8 \times (n-1)$ optical channels. In our channel mapping scheme, considering hardware costs, instead of using a single crossbar network, we map inter-ports to groups of $(n-1)$ optical channels. Each inter-port is separately mapped to a specific group of optical channels. For load balancing purpose, the $(n-1)$optical channels within each group are sourced from $(n-1)$ different chiplets respectively. With the system scale increasing, more optical channels share a single inter-port. Meanwhile, the traffic sent by an optical channel is reduced. Hence, the total traffic load received via a inter-port does not increase with the system scale and there is no scalability issue about the port, as we demonstrate in Sec. V-F.

D. Tuning-free mechanism

In conventional optical networks, a tuning step is performed before packet transmission to turn on the MR filters in the destination node and turn off the MR filters in other nodes, as shown in Fig. 6a. Typically, the tuning step takes 2 cycles. This ensures that only the specified destination node receives signals from the optical channel for the current packet. As a result, no other packets can be sent until the current packet completes transmission because the MR filters in other nodes are turned off and cannot receive signals. Thus, the packets are transmitted one by one, which leads to ultra-low bandwidth utilization. Additionally, a control network is needed to send the tuning message during the tuning step, which brings additional overhead. To address this issue, we propose a tuning-free transmission mechanism (shown in Fig. 6b), where the MR filters are always on for all optical channels. When a packet is sent to an optical channel, all other nodes can receive the packet. However, the packet is required by a specific node only, so we integrate a comparator in the optical interface. The comparator compares its chiplet ID with the destination chiplet ID in the packet header before sending the received packets to the buffer. If they match, the packet is sent to the buffer; otherwise, it is dropped. Although this mechanism increases power consumption due to an increased number of receivers, our region-based design has a small number of receivers in each region (i.e., only several chiplet nodes), making the additional power consumption acceptable. Taking the additional receivers into consideration, our energy

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

consumption is still lower than conventional optical networks, as we demonstrate in Sec. V-D.

V. Evaluation

In this section, we compare our RONet-based GPU system to the two most representative GPU system: MCM [19] and the latest proposed GPUOPT system [20]. MCM employs a Ring-based electrical inter-chiplet network for a 4-chiplet system. However, Ring topology is not scalable for large-scale system. For fair comparison, we modify the Ring topology to Mesh topology, but maintain the same inter-chiplet link bandwidth as the original paper. For GPUOPT, it is originally designed for single-die GPU. Here, we apply its design methodology for chiplet network (i.e., using SWMR to connect the chiplets). In our experiment, these three GPU systems have the same SM number, SM type, cache configuration, and other GPU-related configurations. In addition, the optical channel width of RONet and GPUOPT keeps the same. We quantitatively evaluate the performance, memory access latency, energy consumption, and scalability of these three systems.

A. Experimental setup

We evaluate the performance of three systems by a cycle-accurate simulator Accel-Sim [21]. We modified the Intersim in Accel-sim to implement these three chiplet-based GPU systems. Tab. III summarizes the detailed configurations of the GPU systems in our experiment. For MCM, we set the electrical inter-chiplet link latency and bandwidth to 32 cycles per hop and 1 TB/s, respectively. In GPUOPT, the optical transmission latency is set to 9 cycles (2-cycle tuning, 3-cycle E-O, 2-cycle propagation, and 2-cycle O-E). In RONet, the optical transmission latency is 7 cycles because it is tuning-free. There is a 0.65mW heater placed close to each optical interface to stabilize the temperature and tolerate variations in the optical channel. To estimate power dissipated by SM cores, caches, and memory controllers when running the application workloads, we use AccelWattch [22]. In addition, we use DSENT [23] to estimate the intra-chiplet crossbar in these three systems. The energy consumption of the optical interface are analyzed by OEIL [15] based on the parameters in Tab. I. The power consumption of optical links includes the power of laser sources, E-O/O-E interfaces, heaters, and thermal tuning devices. We calculate the power consumption of laser sources by finding the worst-case power loss of any possible

TABLE III: GPU configuration

Parameter	Value
Number of chiplet	9, 16, 25
Number of SMs	32 per chiplet
GPU frequency	1 GHz
Max number of warps	64 per SM
Warp scheduler	Greedy then Round Robin
L1 data cache	128KB per SM, 128B lines, 4 ways
L2 cache	4MB per chiplet (8 slices, 16-way, 256-set)
Intra-chiplet crossbar	32×18, 16B flit width

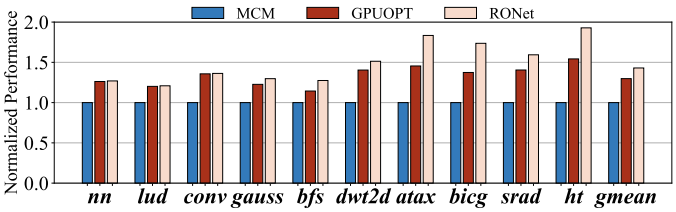


Fig. 7: Performance of 16-chiplet MCM system, GPUOPT system and RONet system

optical channels. To outline the great scalability of RONet, we also evaluate three different system scales: 9-chiplet, 16-chiplet, and 25-chiplet. We select a variety of applications from two benchmark suites: nn, lud, $3DConv$ ($conv$), $guassian$ ($guass$), bfs, $dwt2d$, $srad-v2$ ($srad$) and $hotpsot$ (ht) from rodinia-3.1 [34], and $bicg$ and $atax$ from polybench [35]. For all of the following results, we use the electrical network-based system MCM as a baseline to get the normalized results for all applications.

B. Performance

Fig. 7 illustrates the normalized performance of a 16-chiplet GPU with MCM, GPUOPT, and our proposed RONet. Performance is inversely proportional to execution time. We can observe that both RONet system and GPUOPT system outperform the MCM system. The main reason is that there is a large amount inter-chiplet traffic in chiplet-based GPU, and both RONet and GPUOPT utilize optical interconnects for chiplet-level communication, resulting in a substantial enhancement of the inter-chiplet link bandwidth and alleviation of network congestion. Notably, bandwidth-sensitive applications such as $hotspot$ and $srad$ demonstrate greater improvement. These applications frequently access data from memory partitions, resulting in heavier network traffic, more frequent network congestion, and stalls in memory controllers. As a result, these applications benefit more from the ultra-high bandwidth of optical interconnects. On the other hand, bandwidth-insensitive applications such as nn and lud also demonstrate decent speedup because of the low transmission latency provided by the optical network. On average, GPUOPT and RONet exhibit 1.30× and 1.43× speedup, respectively, compared to MCM. Although both GPUOPT and RONet adopt optical interconnects with the same inter-chiplet link bandwidth, RONet outperforms GPUOPT. The primary reason for this difference lies in our tuning-free mechanism that enables pipeline transmission in the optical link, thereby increasing the maximum bandwidth of the optical network. Consequently, RONet shows significant performance improvement for bandwidth-demanding applications, while having similar performance for bandwidth-insensitive applications, compared to GPUOPT.

C. Average memory access time

To further understand the performance differences among the MCM, GPUOPT, and RONet systems, we analyze their average memory access time (AMAT), as shown in Fig. 8.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

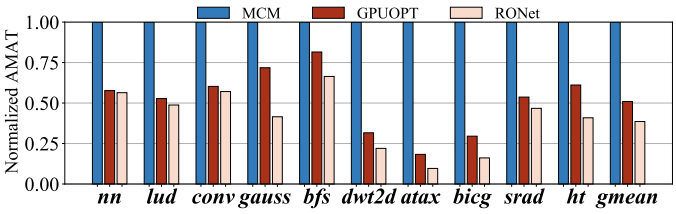


Fig. 8: Average memory access latency of 16-chiplet MCM system, GPUOPT system and RONet system.

Memory access time refers to the time it takes for SMs to fetch data from the memory system, and it directly impacts system performance. On average, we observe that GPUOPT and RONet achieved 49.1% and 61.5% reduction on AMAT, respectively, compared to MCM. Two main reasons account for this observation. First, most of GPU applications are memory-intensive and the abundant traffic causes congestion in the network and memory controllers. To address this issue, both RONet and GPUOPT employ high-bandwidth optical interconnect, which effectively reduces the queuing delay in the network, thereby reducing the average memory access latency. In addition, RONet adopts a tuning free mechanism that further increases network bandwidth. Therefore, RONet achieves a greater reduction in average memory access latency compared to GPUOPT. Secondly, the latency of an inter-chiplet optical transmission is significantly lower than that of inter-chiplet electrical transmission, taking less than 10 cycles, compared to 32 cycles. Moreover, it takes only one optical transmission in GPUOPT and no more than two optical transmissions in RONet to transfer inter-chiplet packet, while it takes much more hops in MCM (2D Mesh) for inter-chiplet communication.

D. Network energy

This subsection presents a comparison of the network energy among MCM, GPUOPT, and RONet. The network energy is divided into two parts: intra-chiplet crossbar and inter-chiplet network. The architecture of the intra-chiplet crossbar is the same for all three systems, while the inter-chiplet networks vary. The energy consumption of optical inter-chiplet network comes from the following parts: laser resources, E-O/E-O interfaces, thermal tuning, heaters, and electrical peripheral circuits, which are all counted in our calculation. The network energy comparison of these three

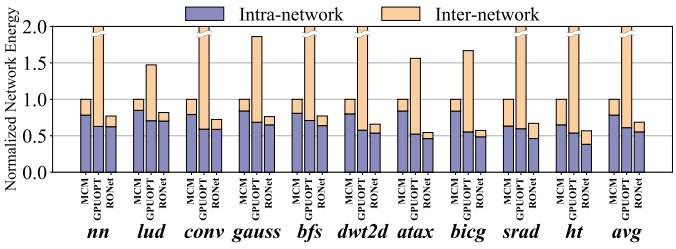


Fig. 9: Network energy breakdown of 16-chiplet MCM system, GPUOPT system and RONet system.

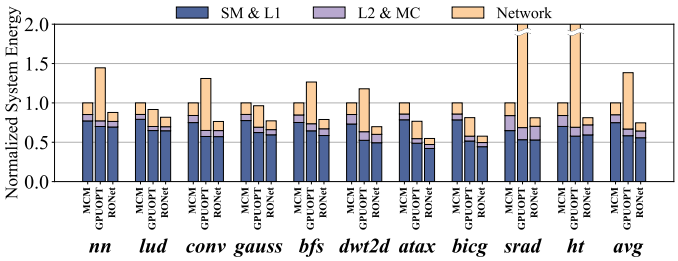


Fig. 10: System energy breakdown of 16-chiplet MCM system, GPUOPT system and RONet system.

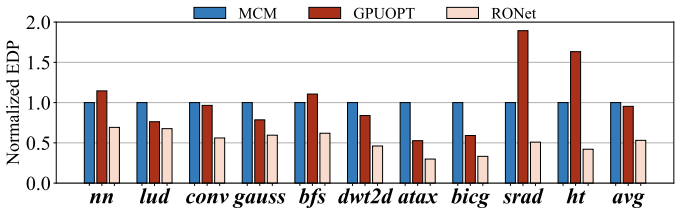


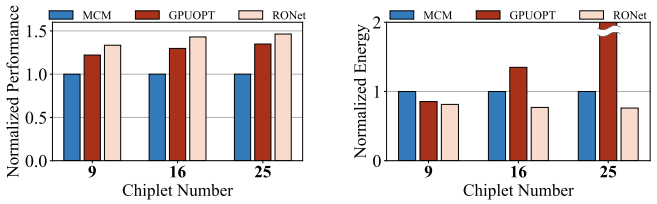
Fig. 11: EDP of 16-chiplet MCM system, GPUOPT system and RONet system.

systems is shown in Fig. 9. Overall, RONet consumes the least energy, whereas GPUOPT consumes the most energy among these three systems. There are two main reasons why RONet exhibits lower energy consumption. Firstly, RONet has lower energy per bit and fewer hops, resulting in reduced inter-chiplet power. Secondly, RONet has a shorter execution time, thereby reducing static energy consumption. On the other hand, GPUOPT employs a single optical link to connect all the chiplets, and its laser energy consumption increases exponentially with the number of chiplets, which is deemed unacceptable. On average, RONet can save 31.4% network energy while GPUOPT consumes 4.60× network energy, compared to MCM.

E. System energy and EDP

In this subsection, we assess the system's energy consumption and energy production delay (EDP) for three systems: MCM, GPUOPT, and RONet. The energy consumption is categorized into three components: SMs and their L1 caches, L2 caches along with memory controllers, and the network. Fig. 10 illustrates that RONet system can achieve 25.4% energy savings, while GPUOPT system consumes 38.4% more energy, compared to MCM. Approximately 80% of the energy is consumed by the SMs, L2 caches, and memory controllers. RONet not only reduces the energy consumed by the network, as demonstrated earlier, but also effectively decreases the energy consumption of SMs compared to MCM. This is due to its high network bandwidth, which reduces execution time and subsequently lowers the static energy consumption of SMs and L2s. Fig. 11 highlights that the RONet system exhibits the lowest EDP among the three systems, with an EDP of only 53.1% of the MCM system. These results demonstrate that RONet significantly reduces energy consumption across the

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.



(a) Performance (b) System energy

Fig. 12: Performance and system energy of MCM system, GPUOPT system and RONet system at three scales: 9-chiplet, 16-chiplet, 25-chiplet.

network and other system components, achieving both high performance and energy efficiency.

F. Scalability

In the previous sections, we have demonstrated the performance and energy characteristics of a 16-chiplet system. In this subsection, we further investigate the scalability of our RONet design and highlight its superiority across different system scales. Our RONet performs consistently well, regardless of the scale. In addtion to the 16-chiplet system, We also evaluate the system with smaller (9-chiplet) and larger (25-chiplet) configurations, and the corresponding average performance and energy results for the 9, 16, and 25-chiplet systems are depicted in Fig. 12. It is important to note that all the results for RONet and GPUOPT are normalized to the MCM system of the same scale. We observe that RONet outperforms both MCM and GPUOPT at all scales. Additionally, as the system scale increases, the performance of RONet continues to improve. The speedup of RONet at 9, 16, and 25-chiplet are 1.33×, 1.43×, 1.46×, respectively. This is attributed to RONet's consistent number of hops. Moreover, the normalized energy consumption of RONet compared to MCM decreases as the system scale increases, owing to two reasons. Firstly, MCM incurs more hops and consequently higher energy consumption. Secondly, the improved performance and reduced execution time of RONet contribute to an overall decrease in energy consumption, albeit a slight increase in energy consumption per bit of optical link. Furthermore, from the figure we can also observe that the energy scalability of GPUOPT is poor, as we mentioned earlier. Although the system energy consumption of GPUOPT at 9-chiplet is lower than MCM, yet as the system scale increases, its energy consumption worsens, reaching an unacceptable level (24.2× at 25-chiplet system), because of the ultra-high laser power consumption.

G. Area

Tab. IV shows the main optical elements in RONet and GPUOPT, including fibers, waveguides, and rings. We can observe RONet needs more fibers and waveguides to connect the chiplets because a chiplet in RONet is connected to two regions optical network, while a chiplet in GPUOPT is only connected to a global optical network. In contrast, RONet saves about half of the number of MRs, because a chiplet only needs to receive packets from other chiplets within the same region in RONet while a chiplet needs more receivers for all the other chiplets in GPUOPT. We use 4 um pitch waveguides and 10 um diameter miro-rings in both RONet and GPUOPT. Thus, the area occupied by optical devices in RONet and GPUOPT are 3.1 mm^2 and 6.1 mm^2, respectively. The area of a 32-SM GPU chiplet is more than 200 mm^2 [14], [19], thereby RONet taking about 0.1% on-chip area in a 16-chiplet GPU system.

TABLE IV: Optical device cost

Network	Fiber	Waveguide	MR
RONet	24	32	38912
GPUOPT	16	16	77824

VI. Related Work

Alleviating GPU bandwidth pressure. A significant obstacle that hampers GPU performance is the limited network bandwidth. Several proposals exist to alleviate bandwidth pressure. One is to use memory coalescing [24], [25], which combines request messages to the same address into one and sends only one reply message back, resulting in a reduced total packet count. Shared L1 cache has been proposed to eliminate replication and reduce the L1 cache miss rate, thereby decreasing L1 and L2 communication [26]. Another approach focuses on utilizing remote core bandwidth to reduce L1 and L2 communication [27], [28]. These methods are complementary to our work and can be integrated to further enhance performance.

Optical interconnect. Optical interconnect has been explored in many previous works. Optical networks-on-chip have been proposed in CPU system with a primary focus on reducing network latency [12], [13]. They replace large-diameter electrical network with small-diameter optical network to eliminate multi-hop transmission. High-bandwidth optical interconnect has been employed to enhance network bandwidth in single-die GPU systems, thereby improving GPU performance [14], [20]. Additionally, optical interconnect has been utilized to facilitate chiplet communication [29], [30], which aligns with our objectives. However, in these works, the scalability of the system is limited by laser power due to the significant coupling loss.

VII. Conclusion

In conclusion, we propose RONet, a region-based optical network with tuning-free mechanism to achieve inter-chiplet bandwidth improvement and optical loss reduction. We also propose an optical channel allocation policy and mapping scheme for bandwidth balance. Compared to previous proposals, RONet shows a 43% improvement on performance, 61.5% reduction on average memory access latency, 25.4% reduction on energy consumption, and better scalability.

VIII. Acknowledgement

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0013).

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Su L, " Delivering the future of high-performance computing" 2019 IEEE Hot Chips 31 Symposium (HCS). IEEE Computer Society, 2019: 1-43.

[3] Choquette J, Gandhi W, "Nvidia a100 gpu: Performance & innovation for gpu computing" 2020 IEEE Hot Chips 32 Symposium (HCS). IEEE Computer Society, 2020: 1-43.

[4] Naffziger S, Beck N, Burd T, et al, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families: Industrial product" 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2021: 57-70.

[5] Patterson D, De Sousa I, Achard L M, "The future of packaging with silicon photonics" Chip Scale Rev, 2017, 21(1): 1-10.

[6] Poulton J W, Dally W J, Chen X, et al, "A 0.54 pJ/b 20 Gb/s ground-referenced single-ended short-reach serial link in 28 nm CMOS for advanced packaging applications" IEEE Journal of Solid-State Circuits, 2013, 48(12): 3206-3218.

[7] Beyne E, "The 3-D interconnect technology landscape" IEEE Design & Test, 2016, 33(3): 8-20.

[8] Ramalingam S, "HBM package integration: Technology trends, challenges and applications" 2016 IEEE Hot Chips 28 Symposium (HCS). IEEE, 2016: 1-17.

[9] Mahajan R, Sankman R, Patel N, et al, "Embedded multi-die interconnect bridge (EMIB)–a high density, high bandwidth packaging interconnect" 2016 IEEE 66th Electronic Components and Technology Conference (ECTC). IEEE, 2016: 557-565.

[10] Chen G, Chen H, Haurylau M, et al, "Predictions of CMOS compatible on-chip optical interconnect" Proceedings of the 2005 international workshop on System level interconnect prediction. 2005: 13-20.

[11] Wang Z, Xu J, Yang P, et al, "Improve chip pin performance using optical interconnects" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2015, 24(4): 1574-1587.

[12] Pan Y, Kim J, Memik G, "Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar" HPCA-16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture. IEEE, 2010: 1-12.

[13] Vantrease D, Schreiber R, Monchiero M, et al, "Corona: System implications of emerging nanophotonic technology" ACM SIGARCH Computer Architecture News, 2008, 36(3): 153-164.

[14] Ziabari A K K, Abellán J L, Ubal R, et al, "Leveraging silicon-photonic noc for designing scalable gpus" Proceedings of the 29th ACM on International Conference on Supercomputing. 2015: 273-282.

[15] Wang Z, Xu J, Yang P, et al, "A holistic modeling and analysis of optical–electrical interfaces for inter/intra-chip interconnects" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2016, 24(7): 2462-2474.

[16] Corning. 2014. Corning® Single-Mode Optical Fiber. Technical Publication (2014)

[17] Schow C L, Doany F E, Rylyakov A V, et al, "A 24-channel, 300 Gb/s, 8.2 pJ/bit, full-duplex fiber-coupled optical transceiver module based on a single "Holey" CMOS IC" Journal of Lightwave Technology, 2011, 29(4): 542-553.

[18] Qianfan Xu et al, "Micrometre-scale silicon electro-optic modulator" nature (2005). 435, 7040 (2005), 325–327.

[19] Arunkumar A, Bolotin E, Cho B, et al, "MCM-GPU: Multi-chip-module GPUs for continued performance scalability" ACM SIGARCH Computer Architecture News, 2017, 45(2): 320-332.

[20] Bashir J, Sarangi S R, "GPUOPT: Power-efficient photonic network-on-chip for a scalable GPU" ACM Journal on Emerging Technologies in Computing Systems (JETC), 2020, 17(1): 1-26.

[21] Khairy M, Shen Z, Aamodt T M, et al, "Accel-Sim: An extensible simulation framework for validated GPU modeling" 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2020: 473-486.

[22] Kandiah V, Peverelle S, Khairy M, et al, "AccelWattch: A power modeling framework for modern GPUs" MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture. 2021: 738-753.

[23] Sun C, Chen C H O, Kurian G, et al, "DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling" 2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip. IEEE, 2012: 201-210.

[24] Wang L, Zhao X, Kaeli D, et al, "Intra-cluster coalescing to reduce GPU NoC pressure" 2018 IEEE International parallel and distributed processing symposium (IPDPS). IEEE, 2018: 990-999.

[25] Kim K H, Boyapati R, Huang J, et al, "Packet coalescing exploiting data redundancy in GPGPU architectures" Proceedings of the International Conference on Supercomputing. 2017: 1-10.

[26] Ibrahim M A, Kayiran O, Eckert Y, et al, "Analyzing and leveraging decoupled L1 caches in GPUs" 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2021: 467-478.

[27] Ibrahim M A, Liu H, Kayiran O, et al, "Analyzing and leveraging remote-core bandwidth for enhanced performance in GPUs" 28th International Conference on Parallel Architectures and Compilation Techniques (PACT). IEEE, 2019: 258-271.

[28] Zhao X, Eeckhout L, Jahre M, "Delegated Replies: Alleviating Network Clogging in Heterogeneous Architectures" 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2022: 1014-1028.

[29] Li C, Jiang F, Chen S, et al, "Accelerating Cache Coherence in Manycore Processor through Silicon Photonic Chiplet" Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design. 2022: 1-9.

[30] Demir Y, Pan Y, Song S, et al, "Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects" Proceedings of the 28th ACM international conference on Supercomputing. 2014: 303-312.

[31] Mahajan R, Sankman R, Patel N, et al, "Embedded multi-die interconnect bridge (EMIB)–a high density, high bandwidth packaging interconnect" 2016 IEEE 66th Electronic Components and Technology Conference (ECTC). IEEE, 2016: 557-565.

[32] Wang Z, Wang Z, Xu J, et al, "CAMON: Low-cost silicon photonic chiplet for manycore processors" IEEE transactions on computer-aided design of integrated circuits and systems, 2019, 39(9): 1820-1833.

[33] Beyne E, "The 3-D interconnect technology landscape" IEEE Design & Test, 2016, 33(3): 8-20.

[34] Che S, Boyer M, Meng J, et al, "Rodinia: A benchmark suite for heterogeneous computing" 2009 IEEE international symposium on workload characterization (IISWC). Ieee, 2009: 44-54.

[35] Grauer-Gray S, Xu L, Searles R, et al, "Auto-tuning a high-level language targeted to GPU codes" 2012 innovative parallel computing (InPar). Ieee, 2012: 1-10.

[36] Morris R, Kodi A K, Louri A, "Dynamic reconfiguration of 3d photonic networks-on-chip for maximizing performance and improving fault tolerance" 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE, 2012: 282-293.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on September 24,2024 at 09:01:28 UTC from IEEE Xplore. Restrictions apply.