Towards Scalable GPU System with Silicon Photonic Chiplet

Chengeng Li[1,†], Fan Jiang[1,†], Shixi Chen[1], Xianbin Li[1], Jiaqi Liu[1], Wei Zhang[1], 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)*

***Abstract*—GPU-based computing has emerged as a predominant solution for high-performance computing and machine learning applications. The continuously escalating computing demand foresees a requirement for larger-scale GPU systems in the future. However, this expansion is constrained by the finite number of transistors per die. Although chiplet technology shows potential for building large-scale systems, current chiplet interconnection technologies suffer from limitations in both bandwidth and energy efficiency. In contrast, optical interconnect has ultra-high bandwidth and energy efficiency, and thereby is promising for constructing chiplet-based GPU systems. Yet, previously proposed optical networks lack scalability and cannot be directly applied to existing chiplet-based GPU systems. In this work, we address the challenges of designing large-scale GPU systems with silicon photonic chiplets. We propose GROOT, a group-based optical network that divides the entire system into groups and facilitates resource sharing among the chiplets within each group. Additionally, we design dedicated channel mapping and allocation policies tailored for the request network and the reply network, respectively. Experimental results show that GROOT achieves 48% improvement on performance and 24.5% reduction on system energy consumption over the baseline.**

I. Introduction

GPU-based computing has emerged as a predominant solution for high-performance computing and machine learning applications [1]. In the last decade, commercial GPUs have experienced significant advancements, witnessing a surge in the number of Streaming Multiprocessors (SMs) from 16 to 108 [2]. However, the computational capacity of modern GPUs still falls short of meeting the demands posed by today's burgeoning applications. To support the ever-increasing computing demand, it is critical to build a large-scale GPU system. Nevertheless, integrating a substantial number of SMs into a monolithic silicon die involves formidable challenges such as integration density, cost, and yield [3].

Chiplet technology is a promising solution for constructing large-scale systems, and numerous chiplet packaging and interconnection technologies have been proposed [4]–[8], including MCMs, 2.5D integration, and silicon bridges. However, applying this kind of electrical-based chiplet technologies to build large-scale GPU systems faces several challenges. First, GPU applications are memory-intensive, which demand a large amount of inter-chiplet bandwidth, which cannot be satisfied by electrical links. Second, the electrical link, whose energy consumption is dependent on interconnect length, only allows interconnecting adjacent chiplets. This constraint implies that packets may undergo multiple hops to reach the destinations, leading to increased latency and energy consumption while increasing the system scale.

Conversely, optical interconnects possess properties that can be harnessed to surmount the aforementioned challenges associated with electrical inter-chiplet interconnects [9], [10]. First, optical interconnect can provide ultra-high bandwidth especially when wavelength-division multiplexing (WDM) is applied. Second, optical interconnect can support direct long-distance communication with relatively constant energy consumption.

Many optical networks [10]–[15] have been proposed in the past, but they cannot be directly applied to chiplet-based GPU. As we will show in the paper, designing a high-bandwidth chiplet-based GPU system via optical interconnect is non-trivial. Conventional approach [12], [13], which connects the L2 caches and SM clusters through an optical link, is not applicable to large-scale chiplet-based GPU system. First, although it can provide high bandwidth, it is at the price of high optical resources. Second, interconnecting a large number of nodes using this approach incurs substantial energy consumption.

To mitigate these issues, we propose GROOT, a group-based optical inter-chiplet network, achieving both high bandwidth and low energy consumption. Experimental results show that GROOT achieves 48% improvement on performance and 24.5% reduction on system energy consumption over the baseline. Specifically, our contributions are:

- We illustrate the inefficiency of electrical chiplet-based GPU architecture in terms of bandwidth, energy consumption, and non-uniformity.
- We establish a power model and an optical device cost model for inter-chiplet optical network.
- We propose a group-based optical network for chiplet-based GPU system, with dedicated optical channel mapping and allocation policies.
- We co-design the intra-chiplet network and propose a crossbar partitioning mechanism to minimize the intra-chiplet network cost.
- We quantitatively compare the performance, memory access latency, and energy consumption of GROOT with two representative inter-chiplet networks.

The rest of the paper is organized as follows: Sec. II describe our motivation. Sec. III introduces the background of optical interconnect. Sec. IV discusses the insight of GROOT.

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

979-8-3503-4859-0/DATE24/© 2024 EDAA

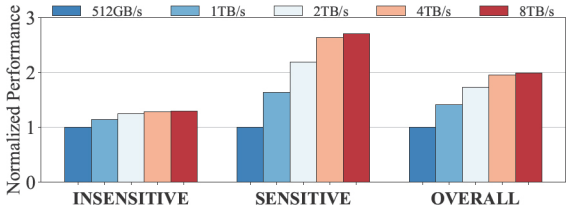


Fig. 1: Normalized performance with respect to inter-chiplet link bandwidth ranging from 512GB/s to 8TB/s for a 16-chiplet EMCM system. We calculate the average performance for bandwidth-insensitive, bandwidth-sensitive, and all applications, respectively.

Sec. V describes the details of GROOT. Sec. VI quantitatively evaluates GROOT. Sec. VII makes the conclusion.

II. Motivation

To facilitate a large-scale system design, chiplet technology has been applied in GPU system [16], denoted as EMCM. Note the original paper only assumes a 4-chiplet system and uses a ring topology. Considering the poor scalability of ring, we extend it using mesh topology for large-scale systems consisting of many chiplets, as shown in Fig. 3(a). In EMCM, each chiplet contains SMs and L2 slices, which are connected using a crossbar. When an SM needs to communicate with a remote L2 cache in the other chiplets, it utilizes the inter-chiplet mesh network. However, this architecture has the following significant drawbacks.

First, due to the limited bandwidth provided by electrical links, the overall system performance is constrained by the inter-chiplet bandwidth. Although electrical links can provide bandwidth of up to 768 GB/s or even 1.5 TB/s [16], these bandwidth levels are inadequate for GPU-based computing, given the memory-intensive nature of GPU applications. As shown in Fig. 1, a substantial performance enhancement is observed as the inter-chiplet bandwidth increases from 512 GB/s to 4 TB/s, eventually plateauing at 8 TB/s.

Second, the energy consumption of electrical link is dependent on interconnect length, which implies only adjacent chiplets can be directly connected. Consequently, electrical-based networks often use large-diameter topologies such as a mesh in large-scale systems, resulting in relatively poor scalability. Packets must traverse multiple hops before reaching their destination, leading to performance degradation and increased energy consumption.

Third, this architecture exhibits non-uniformity in terms of both latency and bandwidth because accessing remote L2 is accomplished via the inter-chiplet network, which shares different bandwidth and latency characteristics with intra-chiplet

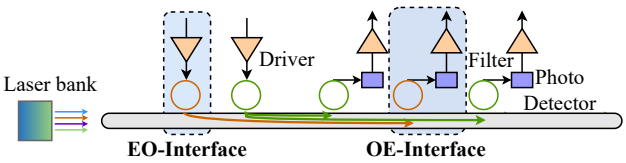


Fig. 2: The structure of a typical optical interconnect. The devices in the same color share the same optical channel. For example, the green sender can send messages to the other two green receivers.

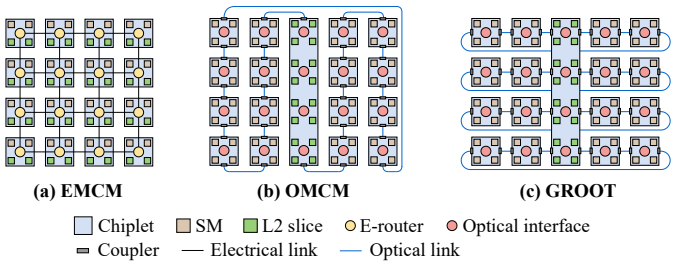


Fig. 3: The architecture overview of three chiplet-based GPU systems.

links. This non-uniformity introduces extra design complexity for both hardware and software stacks [17].

Optical interconnect has high bandwidth and low latency, and thus provides opportunities for solving the aforementioned issues. Nevertheless, the design of a scalable chiplet-based GPU system via optical interconnects demands meticulous consideration. In Sec. IV, we examine the inefficiencies of conventional designs and highlight the key idea of our work.

III. Optical Device Background

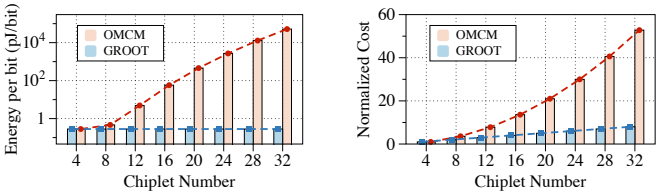
In this section, we give a quick introduction of the background of optical link. Fig. 2 illustrates a typical optical link including four main parts: off-chip laser resource, E-O interface (sender), optical link, and O-E interface (receiver). At the sender end, electrical signals are modulated into laser lights of specific wavelengths by micro resonators (MRs). Optical fibers serve as the optical link for its low optical loss. Laser lights of multiple wavelengths can be transmitted in parallel inside fiber to provide ultra-high bandwidth (WDM). At the receiver end, the laser light with a specific wavelength is extracted by the optical filter. The filters are also implemented by MRs that can switch laser light with a specific wavelength. Finally, the optical signals are converted to electrical signals by photo-detectors.

IV. Insight of Group-based Optical Network

A. *Inefficiency of conventional design*

Instead of placing SMs and L2s within a chiplet, we can separate the L2 caches and SMs, forming L2 chiplets and SM chiplets accordingly. We can then connect these chiplets using an optical link, where each chiplet is assigned several optical channels, and these optical channels are shared by the SMs or L2s within a chiplet via an intra-chiplet crossbar. We denote this architecture as OMCM, shown in Fig. 3(b). This is the extension of the representative optical network for GPU [12] to chiplet scenario if we view the SMs or L2 caches within a chiplet as a cluster. In such a system, with optical interconnect, the communication latency and bandwidth between any SM and L2 pair is constant. Hence, there is no non-uniformity issue. Note that this kind of system is not achievable in electrical-based systems, since electrical links cannot support such high bandwidth demanded by L2 chiplet and long-distance direct connections.

However, this kind of design is not applicable to large-scale chiplet-based GPU systems due to the excessive power consumption of laser and high hardware cost. The energy consumption of optical networks mainly comes from the laser,



(a) (b)

Fig. 4: (a) Energy per bit of an optical transmission in OMCM and GROOT. (b) Optical device cost of OMCM and GROOT, which are normalized to the cost of 4-chiplet OMCM.

which is given by Equ. 1. The laser power (P_l) is influenced by detector sensitivity (P_s), power conversion efficiency (E_l), fiber optical loss ($loss_f$), MR dropping loss ($loss_d$), wavelength number per optical channel (W), the number of chiplet connected to the link (N), optical channel number (M), MR passing loss ($loss_p$), and coupling loss ($loss_c$). According to the equation, we can find the laser power increases exponentially with the number of chiplet.

$$P_l = P_s/E_l \times 10^{(loss_f + 2\times loss_d + (W\times N\times M-2)\times loss_p + 2\times N\times loss_c)} \quad (1)$$

Equ. 2 provides the cost of optical devices, primarily determined by the quantity of MRs.

$$Cost = M \times W \times (N+1) \quad (2)$$

To keep the allocated bandwidth per chiplet at a constant level while scaling up the number of chiplets (N), the optical channel number (M) must also increase in tandem with N. Consequently, the cost experiences a quadratic escalation with the expansion of the system scale.

B. Key idea of group-based network

Considering the limitations of energy consumption and hardware cost of OMCM, we propose a group-based optical network named GROOT. The whole system is divided into several groups, and the chiplets within the same group are connected through a separated optical link (see Fig. 3(c)). In this way, the optical energy consumption and hardware cost are related to the number of chiplets within a group (group size) instead of the total number of chiplets. A smaller group size could lead to lower energy consumption and hardware cost, but the group size cannot be infinitely small because a smaller group could also reduce the degree of bandwidth sharing. Considering the trade-off between bandwidth and cost, We set the group size to 4 in this work. From Fig. 4, we can see GROOT exhibits significantly lower energy consumption and optical hardware cost compared to OMCM, especially when the chiplet number is large.

Fig. 5 show an overview of GROOT-based 16-SM-chiplet GPU system. 512 SMs are distributed across 16 SM chiplets and we put 128 L2 cache slices in a separated L2 chiplet. The SM chiplets are divided into 4 groups and each group contains 4 chiplets. In consistency with other works, we use two sub-networks: one for request messages (SMs to L2 caches), and the other for reply messages (L2 caches to SMs). Within each

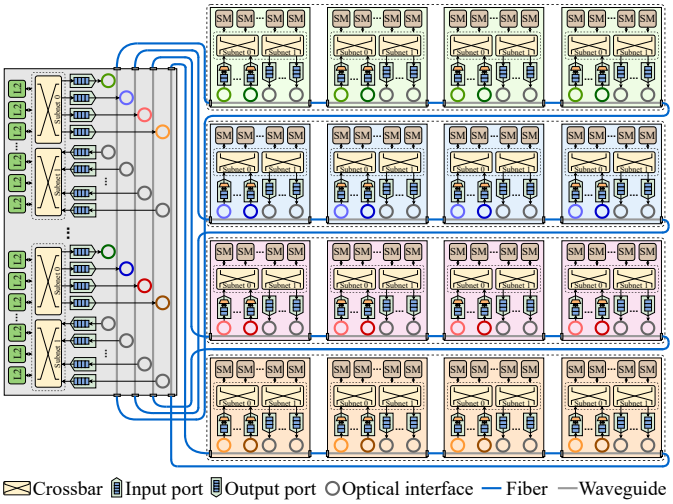


Fig. 5: The architecture of a GROOT-based 16-SM-chiplet GPU system. The chiplets in the same group are labeled in the same color. Two sub-networks are adopted: Subnet 0 for reply packets and Subnet 1 for request packets.

chiplet, there are a set of input ports and output ports equipped with O/E or E/O interface. One end of these ports connects to the intra-chiplet crossbar, while the other end connects to the inter-chiplet optical channels.

After the whole system is divided into several groups, it is still crucial to thoroughly address the design specifics of the optical network such as port assignment and channel mapping under bandwidth requirements while minimizing the cost. In next section, we detail our architecture and describe the co-design method of our inter-chiplet and intra-chiplet network. Furthermore, taking into account the distinctive characteristics of reply and request traffic, we introduce dedicated designs tailored to the reply and request networks, respectively.

V. GROOT Architecture Details

A. Port assignment and channel mapping

In pursuit of load balancing and cost-effectiveness, we propose a pre-allocated channel mapping scheme between the ports in SM chiplets and L2 chiplet. This mapping ensures the existence of a dedicated path for each SM and L2 pair. From now on, let us assume there are L L2 cache slices in the L2 chiplet, S SM chiplets and they are divided in to G groups, and the group size is K (as mentioned in Sec. IV-B, K is 4).

Reply network. For the reply network, within a chiplet group, the L2 chiplet serves as the sole source node, while all the SM chiplets act as destination nodes. Hence, to avoid bandwidth waste and enable optical resource sharing, we adopt single-write-multiple-read (SWMR) optical channels [14] in the reply network to connect the L2 chiplet and SM chiplets within a group. Specifically, within a group, for each channel, each chiplet has an attached port to send/receive packets, namely output port in L2 chiplet (denoted as L2-OP) and input port in SM chiplet (denoted as SM-IP), respectively. For each channel, the L2 chiplet can transmit reply messages through its corresponding L2-OP, while the K SM chiplets can receive these messages via their respective SM-IP.

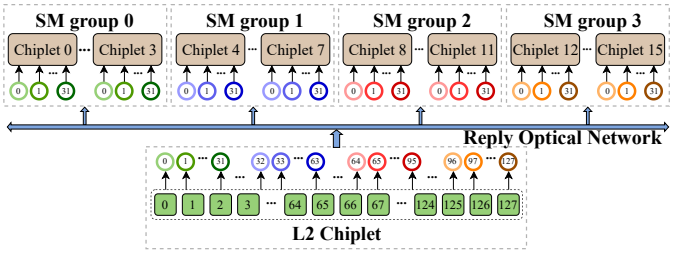


Fig. 6: Illustration of optical channel mapping for the reply network of a 16-chiplet system. The output-input pair in the same channel is labeled in the same color.

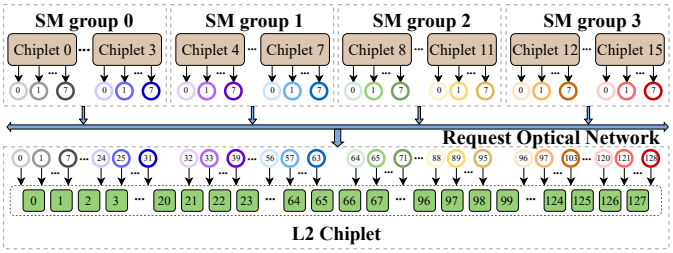


Fig. 7: Illustration of optical channel mapping for the request network of a 16-chiplet system. The output-input pair in the same channel is labeled in the same color.

To fully utilize L2 bandwidth, we assign the number of L2-OPs in a 1:1 ratio to L2 slices. A larger number of L2-OPs would remain underutilized, while a smaller number of L2-OPs could lead to congestion in the L2 queue. Consequently, a 1:1 ratio guarantees that reply packets from L2 caches can be promptly dequeued, minimizing queuing delay. The L2-OPs and the L2 slices can be connected using $L \times L$ intra-chiplet crossbar. However, a crossbar of this scale is prohibitively expensive. In Section V-C, we will present a crossbar partitioning mechanism designed to mitigate this cost. For the sake of our present discussion, we assume the utilization of a $L \times L$ crossbar.

For the purpose of load balance, we divide the L2-OPs evenly among the groups, i.e., each group is allocated L/G L2-OPs. Accordingly, each SM chiplet is assigned L/G SM-IPs to receive the reply packets. Under this design, if a L2 slice with $L2_{ID}$ wants to send a packet to an SM chiplet with SC_{ID}, the corresponding $L2OP_{ID}$ and $SMIP_{ID}$ can be calculated based on the following equation, respectively. A 16-chiplet example of the channel mapping for the reply network is shown in Fig. 6.

$$L2OP_{ID} = SC_{ID}||K \times (L/G) + L2_{ID}\%(L/G) \quad (3)$$

$$SMIP_{ID} = L2_{ID}\%(L/G) \quad (4)$$

Request network. For the request network, within a group, the L2 chiplet serves as the only destination, while all the SM chiplets act as the source nodes. Although we can use a multiple-write-single-read (MWSR) optical channel [13] to support transmission from SM chiplets within a group to L2 chiplet, which enables bandwidth sharing of the MWSR optical channel among SM chiplets, the channel contention can happen. To address the issue, an arbitration system (for example, a token-based approach [13]) is needed to serialize the packets and guarantee only one packet is sent to the optical channel. This kind of arbitration system poses extra cost and design complexity. Considering the request packet is short and not bandwidth-demanding, thus each SM chiplet is connected to the L2 chiplet through several separate point-to-point optical channels in our design. In other words, each channel is exclusively utilized by a specific input port of an SM chiplet (SM-OP) and an output port of the L2 chiplet (L2-IP).

We assign M L2-IPs in the L2 chiplet to receive request packets from all the SM chiplet, and also divide the L2-IPs evenly among each chiplet, so every L/S L2-IP is responsible for receiving request packets from an SM chiplet. Correspondingly, each chiplet has L/S SM-OPs. If an SM chiplet with SC_{ID} wants to send a request packet to an L2 slice with $L2_{ID}$, the ID of the L2-IP can be calculated based on the following equation. A 16-chiplet example of the channel mapping for the request network is shown in Fig. 6.

$$SMOP_{ID} = L2_{ID}\%(L/S) \quad (5)$$

$$L2IP_{ID} = SC_{ID} \times (L/S) + L2_{ID}\%(L/S) \quad (6)$$

B. Optical channel allocation

In this section, we introduce the channel allocation scheme for our inter-chiplet optical link. The aim of channel allocation is to satisfy the inter-chiplet bandwidth requirements while maintaining the balance between intra-chiplet and inter-chiplet bandwidth.

Given that the reply messages are long as they carry the data, congestion frequently occurs within the reply network, leading to performance bottleneck, namely reply bottleneck [18]. To mitigate this, we should assign a larger bandwidth for the reply network. Hence, here we set both the L2 intra-chiplet crossbar link width and inter-chiplet optical channel width to the reply packet size (144B in our setting), which guarantees that a reply packet can be popped out from L2 slices promptly. Remember that each channel is shared by four SM-IPs, which means each SM-IP is allocated 1/4 equivalent bandwidth if we assume load balance. Hence, we set the SM intra-chiplet crossbar to 36B.

Different from reply packets, request packets are short, so the request network is not bandwidth-starvation. Hence, we set both the inter-chiplet optical channel width and intra-chiplet crossbar link width to 32B.

C. Crossbar partitioning

As mentioned previously, the crossbar in the L2 chiplet is $L \times L$ (128×128 in the 16-chiplet example) with 144B width. The power consumption is prohibitively high for such a crossbar. Based on our estimation, it consumes more than $100W$ for 128×128 $144B$ crossbar. To mitigate this issue, we propose a crossbar partitioning scheme to reduce hardware cost, as explained below.

For the reply network, our channel mapping scheme reveals that each port is exclusively designated for communication with chiplets within a specific group. Therefore, based on the destination group ID, the L2-OPs are categorized into four types. If we select one port from each type (resulting in a total of four ports) and package these four into a group, then this group of L2-OPs can facilitate access to any SM chiplet from any group. Correspondingly, we evenly divide the L2 slices into

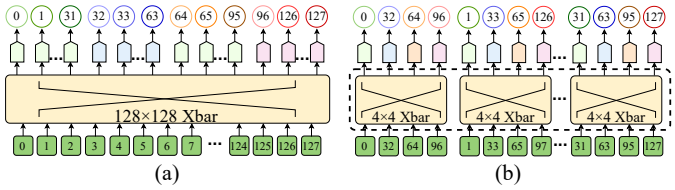


Fig. 8: Illustration of crossbar partitioning mechanism within the L2 chiplet of the reply network.

groups of size four. Each four L2 slices establishes a crossbar connection with the four L2-OPs. This results in a total of 32 smaller crossbars (4×4), as shown in Fig. 8. Such a design still ensures that each L2 can be accessed by any SM chiplet, but the cost is much lowered. Similarly, the L2-IPs in the request network can be categorized into 16 types, each matching an SM chiplet, and thus we can partition the 128×128 crossbar into eight 16×16 crossbars, correspondingly.

VI. Evaluation

In this section, we evaluate our GROOT architecture and compare it with the two architectures EMCM [16] and OMCM [12] mentioned in the previous sections. We quantitatively investigate the performance, memory access time, and energy consumption of the three systems.

A. Experimental setup

We use Accel-Sim [19] to conduct the simulation experiment. In this evaluation, we mainly focus on the 16-chiplet scale under 23 nm technology node. It is important to note that the design methodologies outlined in Sec. V are general and can be applied to other system scales, including 8-chiplet, 32-chiplet, etc. For EMCM, we set the electrical inter-chiplet link the same as the original paper, i.e., 32 cycles per hop and 768 GB/s bi-directional bandwidth. GROOT maintains configurations identical to the architectural details discussed in previous sections. As mentioned earlier, OMCM necessitates significantly larger optical resources to achieve bandwidth parity with GROOT. For a fair comparison, we allocate approximately double the optical

TABLE I: GPU configuration

SM	16 chiplets, 32 SMs per chiplet @2GHz Greedy-then-oldest (GTO) scheduler
CTA Scheduling	Two-level round-robin
L1 caches / SM	32KB 4-way Data, 2KB 4-way Instruction 12KB 2-way Constant, 24KB 24-way Texture
L2 Cache	128 L2 cache slices, 256 KB per slice 16-way, 128B cacheline

TABLE II: Optical device parameters

Parameter	Value	Parameter	Value
MR passing loss	0.01 dB	MR dropping loss	1 dB
MR tuning power	0.65 mW	Waveguide loss	0.5 dB/cm
Coupling loss	1 dB	Receiver sensitivity	-20 dBm
Laser power efficiency	25%	Data rate	64 $Gbps$

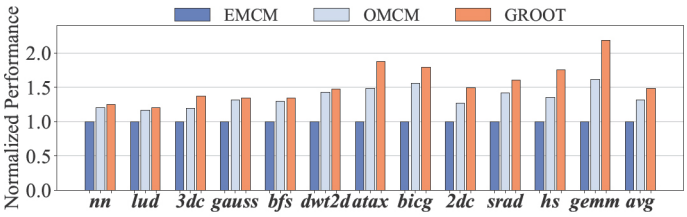


Fig. 9: Performance of 16-chiplet EMCM, OMCM, and GROOT.

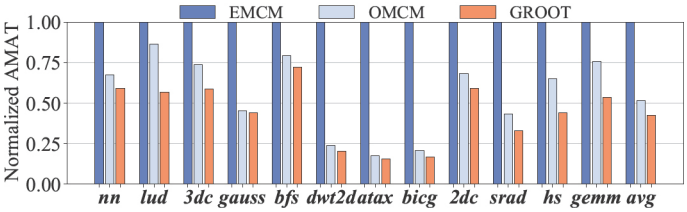


Fig. 10: AMAT of 16-chiplet EMCM, OMCM, and GROOT.

resources to OMCM compared to GROOT. The device parameters [20]–[22] used in optical networks are shown in Tab. II. We use a mix of applications taken from two benchmark suits: nn, lud, $guassian$ ($guass$), bfs, $dwt2d$, $srad-v2$ ($srad$) and $hotpsot$ (hs) from rodinia [23], and $atax$, $bicg$, $2DConv$ ($2dc$), $3DConv$ ($3dc$), and $gemm$ from polybench [24]. All the following results are normalized to the EMCM system.

B. Performance

Fig. 9 shows the performance of the three systems. We can see that both OMCM and GROOT outperform EMCM system. As previously mentioned, EMCM system encounters performance bottlenecks stemming from the constrained inter-chiplet bandwidth provided by electrical links. In contrast, optical links offer a higher inter-chiplet link bandwidth potential. However, the resource efficiency of OMCM system is relatively poor. Even though it uses more optical resources compared with GROOT, the inter-chiplet bandwidth is still lower than GOORT. Our GROOT system has a higher inter-chiplet bandwidth, thereby yielding superior performance compared to the other two systems. For bandwidth-sensitive applications such as $gemm$, our system achieves a remarkable speedup of up to 2.18×. On average, OMCM achieves a speedup of 1.32×, while GROOT attains a higher speedup of 1.48×.

C. Average memory access time

To gain a deeper understanding, in this subsection, we study the average memory access time (AMAT) of the three systems. Memory access time refers to the time it takes for SMs to retrieve data from the memory system, and it plays an important role in impacting the overall system performance. From Fig. 10, we can observe that GROOT has the lowest AMAT, compared with EMCM and OMCM. The reasons are twofold. Firstly, GROOT achieves the highest inter-chiplet bandwidth, which could largely mitigate network congestion, consequently reducing congestion delay. Secondly, the lower optical transmission latency also contributes to the reduced AMAT. On average, we observe that OMCM and GROOT achieve 48.2% and 57.4% reduction on AMAT, respectively, compared to EMCM.

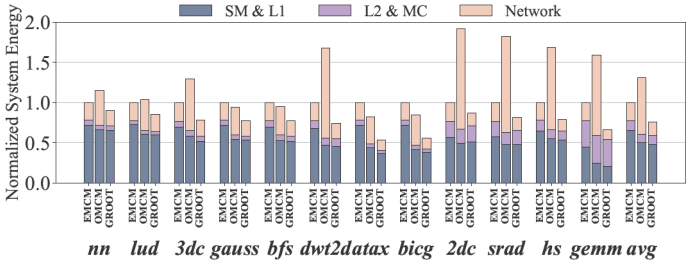


Fig. 11: System energy breakdown of 16-SM-chiplet EMCM, OMCM, and GROOT.

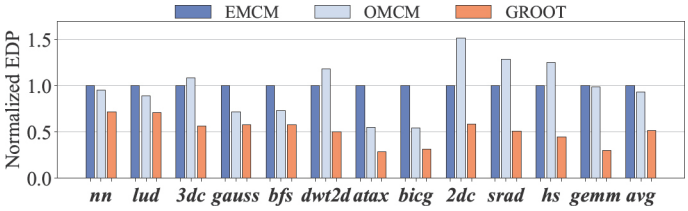


Fig. 12: EDP of 16-chiplet EMCM, OMCM, and GROOT.

D. System energy and EDP

In this subsection, we evaluate the energy consumption of the three systems. We breakdown the energy of the system into three parts: SMs and L1s, L2s and MCs, and network. The network energy consists of both intra-chiplet and inter-chiplet network (including both electrical and optical link). We use AccelWattch [25] to evaluate the energy of SMs, caches, and MCs, and use DSENT [26] to evaluate the network energy. Fig. 11 demonstrates the energy breakdown of the three systems. We can see our GROOT system consumes the least energy, while OMCM consumes the most energy. There are three main reasons. First, GROOT and OMCM can reduce the execution time, leading to lower static energy consumption. Second, GROOT, with its group-based topology, significantly reduces optical energy consumption, while OMCM incurs substantial laser-related energy expenditure. Third, the crossbar partitioning mechanism further contributes to energy savings. Overall, our GROOT system achieves an energy savings of 24.5%, compared with the baseline EMCM. We also plot the energy-delay product (EDP) in Fig. 12. A lower EDP indicates a higher energy efficiency. We see that GROOT has the lowest EDP among the three systems, with an EDP of only 51.4% of the baseline EMCM system.

E. Optical device cost analysis

Here we show the optical device cost of OMCM and GROOT (see Tab. III). OMCM consumes approximately double the resources compared to GROOT, but the bandwidth and performance are still lower than GROOT, as discussed before.

TABLE III: Optical device cost

Network	Fiber	Waveguide	MR
GROOT	20	20	25088
OMCM	17	17	41216

VII. Conclusion

In conclusion, we propose GROOT, a group-based optical network for chiplet-based GPU systems. Compared to previous proposals, GROOT shows a 48% improvement on performance, 57.4% reduction on average memory access latency, 24.5% reduction on energy consumption, and better scalability.

VIII. Acknowledgement

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

References

[1] L. Su, "Delivering the future of high-performance computing," in *HCS*, 2019.
[2] J. Choquette and W. Gandhi, "Nvidia a100 gpu: Performance & innovation for gpu computing," in *HCS*, 2020.
[3] S. Naffziger *et al.*, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families: Industrial product," in *ISCA*, 2021.
[4] D. Patterson *et al.*, "The future of packaging with silicon photonics," *Chip Scale Rev*, 2017.
[5] J. W. Poulton *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.
[6] E. Beyne, "The 3-d interconnect technology landscape," *IEEE Design & Test*, 2016.
[7] S. Ramalingam, "Hbm package integration: Technology trends, challenges and applications," in *HCS*, 2016.
[8] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (emib)–a high density, high bandwidth packaging interconnect," in *ECTC*, 2016, pp. 557–565.
[9] G. Chen *et al.*, "Predictions of cmos compatible on-chip optical interconnect," in *SLIP*, 2005.
[10] Z. Wang *et al.*, "Improve chip pin performance using optical interconnects," *TVLSI*, 2015.
[11] Y. Demir *et al.*, "Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects," in *ICS*, 2014.
[12] J. Bashir and S. R. Sarangi, "Gpuopt: Power-efficient photonic network-on-chip for a scalable gpu," *JETC*, 2020.
[13] A. K. K. Ziabari *et al.*, "Leveraging silicon-photonic noc for designing scalable gpus," in *ICS*, 2015.
[14] C. Li *et al.*, "Accelerating cache coherence in manycore processor through silicon photonic chiplet," in *ICCAD*, 2022.
[15] P. Fotouhi *et al.*, "Enabling scalable chiplet-based uniform memory architectures with silicon photonics," in *MEMSYS*, 2019.
[16] A. Arunkumar *et al.*, "Mcm-gpu: Multi-chip-module gpus for continued performance scalability," in *ISCA*, 2017.
[17] B. Pratheek *et al.*, "Designing virtual memory system of mcm gpus," in *MICRO*, 2022.
[18] K. H. Kim *et al.*, "Packet coalescing exploiting data redundancy in gpgpu architectures," in *ICS*, 2017.
[19] M. Khairy *et al.*, "Accel-sim: An extensible simulation framework for validated gpu modeling," in *ISCA*, 2020.
[20] R. Morris *et al.*, "Dynamic reconfiguration of 3d photonic networks-on-chip for maximizing performance and improving fault tolerance," in *MICRO*, 2012.
[21] Q. Xu *et al.*, "Micrometre-scale silicon electro-optic modulator," *NATURE*, 2005.
[22] C. L. Schow *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.
[23] S. Che *et al.*, "Rodinia: A benchmark suite for heterogeneous computing," in *IISWC*, 2009.
[24] S. Grauer-Gray *et al.*, "Auto-tuning a high-level language targeted to gpu codes," in *InPar*, 2012.
[25] V. Kandiah *et al.*, "Accelwattch: A power modeling framework for modern gpus," in *MICRO*, 2021.
[26] C. Sun *et al.*, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *NOCS*, 2012.