This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

Routing and Wavelength Assignment for Multiple Multicasts in Optical Network-on-Chip (ONoC)

Wen Yang, Yawen Chen, Zhiyi Huang, Haibo Zhang, and Huaxi Gu

***Abstract*—Optical Network-on-Chip (ONoC) is an emerging chip-scale optical interconnection technology to realize high-performance and power-efficient inter-core communication for many-core processors. Multicast communication is popularly used in parallel applications on chip. However, existing researches for multicast in ONoC mainly focus on the optimization of one multicast. This limits the practical applications of the research outcomes because we often face the dynamic formation of multiple multicast groups in real network systems. In this paper, we define the problem of routing and wavelength assignment for multiple multicasts in ONoC with the objective of minimizing the number of wavelengths required. To solve the problem, we first formulate it as an integer programming model for general topologies. Then we design routing policies for special instances that optimally use only one wavelength on mesh topology. For general instances, we design a Group-Partitioning Routing algorithm for Multiple Multicasts (GPRMM). GPRMM decouples a group of multicasts into a number of sub-groups, each of which matching one of the special instances. Theoretical results show that the number of wavelengths required by GPRMM is no more than the Destination Density σ_d, i.e., the maximum number of multicasts with destinations in the same row or column. Moreover, we find the upper bound and the lower bound on the number of wavelengths required for GPRMM. The wavelength requirement is also upper bounded by the network size n for an $n \times n$ mesh network. Simulation results show that GPRMM can reduce the number of wavelengths by 26.7% compared with previous methods. GPRMM has the advantages of low routing complexity, low wavelength requirement, low power consumption, and good scalability.**

***Index Terms*—Optical Network-on-Chip, Multiple Multicasts, Routing and Wavelength Assignment.**

I. Introduction

With the development of manufacturing technologies in integrated circuits industry, many-core Chip Multi-Processors (CMPs) are becoming the mainstream computational platform for cloud computing, data center, and supercomputing applications [1][2]. Electrical Network-on-Chip (ENoC) has been proposed to handle the interconnect parallelization provided by many-core CMPs. As thousands of cores will fit on one chip [3], the inherent problems of ENoC, such as wire delay, bandwidth, power dissipation and signal interference, will deteriorate the performance of CMPs. In order to overcome the drawbacks of ENoC, Optical Network-on-Chip (ONoC), a chip-scale inter-core optical network, has been proposed [4]. Compared with ENoC, ONoC shows obvious advantages (e.g., low end-to-end communication latency, distance-independent power consumption and high bandwidth by Wavelength-Division-Multiplexing (WDM) [5]), as it inserts silicon nanophotonics into on-chip interconnection networks.

W. Yang is with Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, China, 443002; College of Computer and Information Technology, China Three Gorges University, Yichang, China, 443002. E-mail: yangwen0720@163.com.

Y. Chen, Z. Huang and H. Zhang are with the Department of Computer Science, University of Otago, Dunedin, New Zealand, 9016.

H. Gu is with State Key Laboratory of ISN, Xidian University, Xi'an, China.

Manuscript received ####, 2022; revised ####, 2022.

As the number of cores integrated into a chip increases, inter-core communication in ONoC is progressively becoming a significant and challenging problem in the development of many-core processors [6]. Multicast communication, in which packets from one source need to be transmitted to multiple destinations simultaneously, widely exists in many applications of CMPs, such as barrier synchronization [7], clock synchronization [8], replication [9] and multi-reading programs in distributed shared memories. Previous researches have shown that multicast communication contributes to a large proportion of total traffic in various cache coherence protocols such as token coherence and directory-based coherence [10]. Fig.1 shows the percentage of different communication patterns (unicast, multicast, many-to-1 communication) for PARSEC benchmark applications in a 64-core system [11]. It can be seen that multicast traffic takes about 15% and 45% respectively in average for AMD HyperTransport and Token Coherence connections.

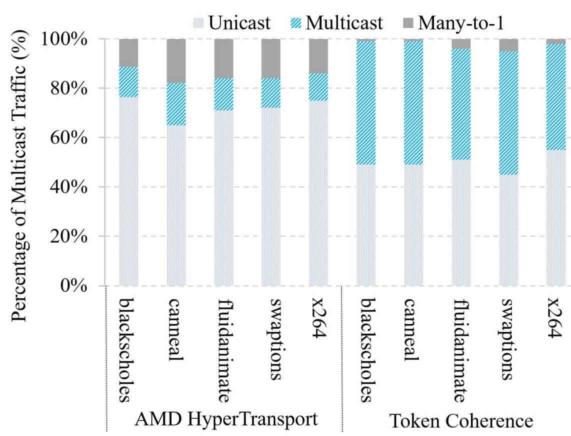


Fig. 1. Traffic percentage for a set of standard PARSEC benchmark applications for Token Coherence and HyperTransport in a 64-core system [11].

For multicast communication in ONoC, there are three challenges to be addressed. One is how to transmit packets to each destination effectively considering the particular characteristics of ONoC. As ONoC's physical properties are different from the traditional ENoC (e.g., no optical buffer, limited number of wavelengths), existing multicast communication schemes for electrical interconnections cannot be used directly to ONoC. The second challenge is how to utilize the limited network resources (e.g., the number

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

of wavelengths) efficiently. WDM enables multiple optical signals to be transmitted in a single waveguide simultaneously by using different carrier wavelengths, offering ONoC with ultra-high throughput and low transmission latency. However, the maximum number of wavelengths that each waveguide can support is limited in realistic scenarios (e.g., at most 62 wavelengths under 10 Gbps data rate [12]), since the maximum optical power that can be injected into the optical interconnect without non-linear effects is limited. With the increasing number of cores in one chip, the supported number of wavelengths is not enough for large sized network (e.g., 32×32 mesh ONoC) with hundreds of multicasts. In addition, laser source and MR tuning power is proportional to the total number of wavelengths used [13]. Hence, using more wavelengths will incur more power consumption, which in turn causes high heat dissipation and affects the reliability of ONoC. Moreover, reducing the number of wavelengths can reduce the hardware cost of ONoC. In ONoC, the optical routing paths are dynamically established by configuring the optical routers with wavelength-specific MRs. Thus, the fewer wavelengths are used, the fewer wavelength-specific MRs are required. Although reducing the number of wavelengths may increase the calibration power from the device-level aspect, several current techniques have shown significant savings in overall calibration power by utilizing the enhanced connectivity to maintain sufficient throughput with fewer rings, or disabling the tuning circuitry for resonators not on the active communication path. Therefore, reducing the number of wavelengths has important impact on network performance, energy consumption, hardware complexity and reliability.

The third challenge is how to implement inter-core communication when multiple multicasts occur during the same time period. If there are multiple applications running on the ONoC simultaneously, each application can generate a multicast request during execution, and thus multiple applications can generate multiple multicasts simultaneously. For some parallel applications such as Deep Neural Networks (DNNs), the data exchange between workers could be implemented by intensive multiple multicast communications [14]. For such applications, efficient methods to support multiple multicasts can largely improve the execution performance. Unlike single multicast solutions, the problem of multiple multicasts should not only optimize an individual multicast but also consider the whole set of multicasts as a combined optimization problem. Existing researches on multicast in ONoC have rarely taken into account all the above challenges. Our work will bridge these gaps. In this paper, we propose an efficient routing and wavelength assignment approach for multiple multicasts to reduce the number of wavelengths used in ONoC.

The proposed methods include two routing and wavelength assignment procedures supporting multiple multicasts in ONoC. One is an optimal routing and wavelength assignment scheme for special instances based on special distributions of source and destination nodes on a mesh ONoC. The other is a group-partition based heuristic routing and wavelength assignment scheme for general instances without contraints on the distribution of multicast nodes. The novelty in the multiple multicasts routing procedure is a distribution-based routing algorithm that adopts a resource combination strategy to make non-overlapping routing paths share the same wavelength. Our main contributions can be summarized as follows.

- We propose an optimal routing and wavelength assignment algorithm for special distributions of multicast nodes on a mesh ONoC, where only one wavelength is required.
- Based on the proposed algorithm for special distributions, we propose GPRMM - a Group-Partitioning Routing algorithm for Multiple Multicasts (GPRMM) for random distributions of multicast nodes. This algorithm is proposed to find non-overlapping multicast groups that can share the same wavelength.
- We derive the upper bound and lower bound on the number of required wavelengths.
- We carry out extensive simulations to evaluate GPRMM, using both real and synthetic traffic traces. The simulation results demonstrate that GPRMM can achieve significant reduction on the number of wavelengths required and has good scalability.

The rest of the paper is organized as follows. Section II introduces related work and motivation. Section III presents the problem definition. Section IV gives the routing algorithm for particular instances. Section V presents a heuristic routing method for multiple multicasts with general instances. Section VI presents the lower bound on the number of wavelengths. Section VII illustrates the system implementation for the given method. Section VIII evaluates the performance of GPRMM through simulations. Finally conclusions are given in Section IX.

II. Related work and Motivation

A. Related Work

Multicast communication widely exists in many applications of CMPs. Unicast-based multicast [15] is a scheme for multicast communication without hardware support, in which a multicast packet is replicated multiple times and transmitted to each of the destinations separately. However, this scheme will increase the network congestion and serialization delay, since transmitting multiple copies of the same packet into the network not only causes a significant amount of traffic, but also introduces a large latency (every copy of the message suffers from startup latency at the source node). Apart from this, redundant packets transmitted in the network will consume more power which is an important consideration in the design of CMPs. Current research about the multicast communication in an ONoC can be classified into two classes: hardware-based design and the design based on routing and wavelength assignment.

The hardware-based design studies the multicast support architectures, such as topologies and on-chip devices. In [16], a hybrid hierarchical architecture, called Firefly, is designed. In Firefly, electrical signal is used for intra-cluster communication, and multiple optical crossbars are used for inter-cluster communication. In order to avoid the global switch arbitration, the crossbar in Firefly is partitioned into multiple

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

smaller crossbars, so the arbitration is localized. The authors in [17] proposed ZMesh, which is a scalable and energy efficient NoC topology. Two mapping techniques are designed to map applications onto nodes effectively. It showed that multicast can reduce energy consumption using the proposed multicast routing algorithm. In [18], the authors derived a novel reconfigurable silicon-photonic NoC architecture, SwiftNoC, which can achieve higher performance and energy efficiency by efficiently utilizing MWMR waveguides with its improved multicast-enabled channel sharing, bandwidth transfer mechanism, and cluster priority adaption. A new on-chip network architecture, called VRNOC, was proposed in [19]. Based on VRNOC, an adaptive routing method for multicast was designed, which can find alternative routing paths without increasing the path length. In [20], the authors proposed a new 7×7 non-blocking optical router based on the Dimension Order Routing (DOR) algorithm, which can decrease the crosstalk and insertion loss of the network.

For the design based on routing and wavelength assignment, tree-based and path-based routing methods are two major approaches used in ONoC. In the tree-based routing [21], a multicast packet is delivered along a spanning tree from root (the source node) to individual leaves (all destination nodes). In [22], two tree-based routing methods for multicast, Optimize Tree (OPT) and Left-xy-Right-Optimized Tree (LXYROPT), were proposed. OPT is an optimized tree-based routing using the west-first turn model to achieve the deadlock free. To reduce the network latency caused by OPT, LXYROPT was designed. The authors in [23] proposed a Recursive Partitioning Multicast routing method (RPM) to implement multicast communication in NOC, which can intelligently select appropriate replication points for multicast packets according to the global distribution of destination nodes. In [24], a routing method based on the Minimum Directed Spanning Tree was proposed, with the objective of reducing power consumption. Overall, the tree-based routing method can obtain low network latency since it constructs the tree by shortest paths. However, the packet is replicated at branching nodes, which may result in the blockage of packets.

In the path-based routing, a packet is transmitted along a Hamiltonian path [25] without being replicated, so it can reduce the packet congestion. In [26][27], two adaptive routing algorithms were proposed respectively, called HAMUM and HOE. In this method, several rules about permitting and prohibiting turns were derived to achieve deadlock-free routing. Choosing proper routing paths according to the congestion situation of the network can improve routing flexibility, thus achieving higher adaptiveness. The authors in [28] proposed a hybrid deadlock-free multicast routing scheme by combining the path-based and tree-based method, which can hold large packets without additional virtual channels or large buffers. The authors in [29] proposed an efficient WDM mesh-based ONoC mapping approach based on a particle swarm optimization algorithm. It investigated the tradeoff between the number of wavelengths and the network size, and it is found that increasing the network size can reduce the number of wavelengths needed. In [30], the authors addressed the parallel implementation of a bitonic sorting on ONoC with bus topology, where a wavelength-saving strategy was proposed.

Although these multicast-based routing methods can improve network performance (e.g., reduce latency, avoid deadlock), they still have several limitations: (1) they were initially designed for ENoC, which cannot be migrated to ONoC directly because of the different physical properties; (2) they only focused on improving the conventional criteria of routing design, such as transmission delay or shortest path, which are not suitable to ONoC. Conversely, those important routing criteria for ONoC, such as the network resource consumptions (e.g., physical links and wavelengths), are not considered; (3) only one multicast was considered. For example, DWRMR [31] is one of the state-of-the-art multicast routing schemes for ONoC, which can reduce the packet latency and the number of wavelengths significantly by reusing the multicast rings. However, it only considers the optimization of one multicast, without considering the optimization of multiple multicasts. When there are multiple simultaneous multicasts in ONoC, methods like this are very likely to cause high contentions without considering other multicasts, resulting in waste of wavelengths.

Therefore, we need to optimize the wavelength allocation for multiple multicasts to efficiently use the limited wavelengths in ONoC. In this paper, we target optimizing the utilization of wavelengths for routing multiple simultaneous multicasts requested from applications.

B. Motivation Example

The following example compares the number of wavelengths used by different routing schemes in order to accommodate all given multicasts. It illustrates that the number of wavelengths for multiple multicasts by the existing methods can be further reduced. Fig. 2 presents a motivation example by showing two multicasts with different routing schemes in a 4×4 ONoC. In Fig. 2 (a), (b) and (c), the multicast routing schemes that were designed for single multicast are used. In Fig. 2 (a), the unicast-based routing method is used, in which source nodes 10 and 7 produce 4 copies of packets respectively. Each copy is transmitted to the destination by 3 wavelengths. Fig. 2 (b) shows the tree-based routing method, where packets are transmitted from the source nodes to the destination nodes along two spanning trees with 2 wavelengths. In Fig. 2 (c), the path-based routing method is used that needs 2 wavelengths. These existing routing schemes used in Fig. 2 (a), (b) and (c) only consider single multicast, without considering the optimization of the multiple multicasts problem in terms of utilization of network resources, such as wavelengths, links, and etc. It may be feasible if the network resources are sufficient with a small number of multicasts. However, when the network resources become insufficient and the number of multicasts increases, these methods are not efficient enough with the possibility of using more wavelengths and consuming more energy. In order to solve this problem, we should consider the multiple multicasts as a whole and design a routing scheme from the global perspective. This problem has not been well studied as far as we know. Therefore, in this paper, the main focus is to design an effective routing and wavelength assignment scheme to accommodate multiple

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

multicasts using minimum number of wavelengths. In Fig. 2 (d), only one wavelength is needed in this example by using our proposed routing algorithm.

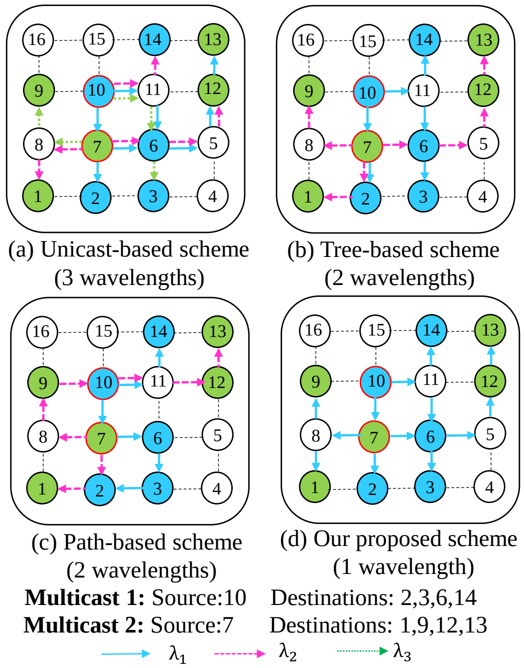


Fig. 2. Routing examples for multiple multicasts in a 4×4 ONoC.

An illustration of our proposed routing and wavelength assignment algorithm for multiple multicasts is shown in Fig. 3. The main idea is to partition all multicast nodes into a minimum number of groups, with each group only holding non-overlapping routing paths. Thus, all routing paths in one group can share the same wavelength. For the given three multicasts in Fig. 3, two groups are derived (Fig. 3 (b), (c)), each of which has distinct routing method (e.g., XY routing for group 1; YX routing for group 2) derived on the basis of an optimal routing and wavelengths assignment algorithm for special distribution of multicast nodes (detailed in Section IV). Two wavelengths are needed in this example by assigning only one distinct wavelength to each group.

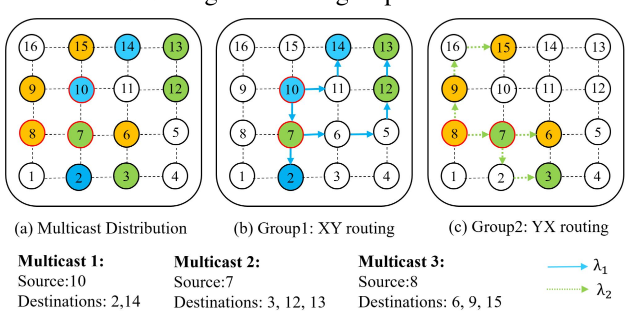


Fig. 3. Illustration of the proposed multiple multicasts routing scheme.

III. Problem Description for Multiple Multicasts in ONoC

For multiple multicasts in ONoC, the objective is not only to deal with each individual multicast but also to consider the whole set of multicasts as a combined problem. The problem of **Routing and Wavelength Assignment for Multiple Multicasts in ONoC** (RWA-MM-ONoC) is defined as: *given an ONoC and a set of multicasts, the RWA-MM-ONoC problem is to find the best routes and assign proper wavelength(s) for each multicast so that the total number of wavelengths is minimized.* In this section, we formulate it as a Non-linear Integer Programming (NIP) model.

Some general terms used in this paper are introduced firstly. A **Path** in ONoC is a set of links, which is established for delivering a packet from a source node to a destination node. A **Multicast Path Set** is the set of paths for one multicast. **Routing** is a process of selecting paths for packets transmission, which determines the directions of the transmitted packets. **Wavelength assignment** is to assign a proper wavelength to every established path. Two paths can share the same link, provided that two different wavelengths are used.

We model the topology of an ONoC as a directed graph G=(V, E), where vertices in $V = \{v_1, v_2, ..., v_N\}$ represent the nodes and edges in E denote the links. N is the total number of nodes in G. Each link e_{ij} in E is the unidirectional optical interconnect from node v_i to an adjacent node v_j. We denote the set of multicasts as $M = \{m_i | 1 \leq i \leq C\}$, where C is the total number of multicasts and m_i is the ith multicast. The nodes involved in multicast m_i include its source s_i and its destination set $D_i = \{d_{i,j} | 1 \leq j \leq |D_i|\}$, where $d_{i,j}$ is the jth destination and $|D_i|$ is the total number of destinations for m_i. So, the total number of destinations to be reached for M is $K = \sum_{i=1}^{C} |D_i|$. We denote the path set as $P = \{p_k | 1 \leq k \leq K\}$, where p_k is the kth path that reaches the kth destination. The set of wavelengths is denoted by Λ=$\{\lambda_1, \lambda_2, \ldots, \lambda_W\}$ and W is the total number of available wavelengths.

We use a 0-1 integer programming formalism to formulate RWA-MM-ONoC. We first introduce the variables.

Link usage variable: Let $x_{i,j,k}$ denote the link usage by a routing path from a source to a destination, defined as

$$x_{i,j,k} = \begin{cases} 1, & \text{if path } p_k \text{ passes through link } e_{ij}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $i \in [1, N]$, $j \in [1, N]$, and $k \in [1, K]$.

Wavelength assignment variable: Let $y_{i,j,k,w}$ denote the wavelength assignment for multicast paths, defined as

$$y_{i,j,k,w} = \begin{cases} 1, & \text{if path } p_k \text{ is assigned by } \lambda_w \text{ on link } e_{ij}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $i \in [1, N]$, $j \in [1, N]$, $k \in [1, K]$ and $w \in [1, W]$.

Multicast distinction variable: Let $z_{k,h}$ denote which multicast each path belongs to, defined as

$$z_{k,h} = \begin{cases} 1, & \text{if path } p_k \text{ belongs to } m_h, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $k \in [1, K]$, $h \in [1, C]$.

Flow-conservation constraint: According to the flow-conservation constraint [35], the total number of paths entering a vertex must be equal to that leaving that vertex, except that each source s_i has $|D_i|$ outgoing paths and each destination $d_{i,j}$ has 1 incoming path. The constraint can be formulated as

$$\sum_{t=1}^{N}\sum_{k=1}^{K} x_{t,q,k} - \sum_{g=1}^{N}\sum_{k=1}^{K} x_{q,g,k} = \begin{cases} -|D_i|, \text{if } (v_q{=}s_i), \\ 1, \text{if } (v_q{=}d_{i,j}), \\ 0, \text{otherwise,} \end{cases}$$
$$\forall m_i \in M, \forall v_q \in V, j \in [1, |D_i|]. \tag{4}$$

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

Wavelength continuity constraint: All links along a path from the source to the destination should be assigned the same wavelength, so this constraint can be satisfied by

$$y_{i,j,k,w} = y_{o,s,k,w},\ \forall e_{ij} \in p_k,\ \forall e_{os} \in p_k,\ w \in [1, W]. \quad (5)$$

Distinct wavelength constraint: If multiple paths in the same multicast have shared links, these segments can be combined into one path since a light splitter can be used to split an input signal into multiple outputs. As illustrated in Fig. 4, the shared segments from a to b in the three paths from the same multicast are combined into one path, and

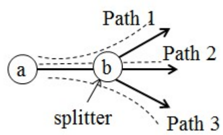


Fig. 4. Illustration of the light splitter in the network.

then be split into multiple outputs through the splitter b. Let $\bar{P} = \{p_k | 1 \le k \le K'\}$ be the set of paths after combination. For any two paths in $\bar{P}$, they must belong to different multicasts if they have shared links. This constraint enforces that, if multiple paths belonging to different multicasts pass through the same link simultaneously, different wavelengths must be assigned to the paths to avoid conflicts, as formulated by

$$\sum_{h=1}^{C} \sum_{k=1}^{K'} x_{i,j,k} z_{k,h} y_{i,j,k,w} \le 1. \quad (6)$$

Splitter constraint: This constraint can be added if light splitters are available. It enforces that the signal after split should be transmitted with the same wavelength used by the original signal before split. This constraint can be represented as

$$\begin{aligned} & x_{i,j,k} z_{k,h} y_{i,j,k,w} = x_{i,j,r} z_{r,h} y_{i,j,r,w}, \\ & \quad \forall p_k, p_r \in P, p_k \ne p_r, \forall w \in [1, W]. \end{aligned} \quad (7)$$

The objective function of RWA-MM-ONoC is to achieve the minimum number of wavelengths, which can be formulated as:

$$\text{Minimize} \ \sum_{w=1}^{W} \left(\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} y_{i,j,k,w}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} y_{i,j,k,w} + \alpha} \right) \quad (8)$$

such that (1)-(7) are satisfied.

In the objective function, if the wavelength λ_w is used, $\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} y_{i,j,k,w}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} y_{i,j,k,w} + \alpha}$ is 1, otherwise[1] it is 0. Since the problem is formulated as a 0-1 integer NIP programming model that is NP-hard in general, it is not possible to solve all instances of RWA-MM-ONoC in polynomial time. In this paper, we first identify special instances on mesh ONoC that can achieve optimal results with only one wavelength for a set of multicasts. Then, we extend the results obtained from these special instances to solve the general instances.

As far as we know, there is no formal formulation presented in the previous researches for Routing and Wavelength Assignment for Multiple Multicasts in ONoC (RWA-MM-ONoC). Moreover, such a formulation will enable us to obtain the optimal solution for small problem sizes that can be used as benchmarks to compare our heuristic solution. It is also worth noting that this is a very general formulation without any constraint on network topology. Even though we use mesh topology in this paper because of its good scalability, the formulation can be used to investigate the same RWA-MM-ONoC problem on other typologies such as torus.

[1] α is a minimal number to avoid the divide-by-zero error.

IV. Optimal RWA of Multiple Multicasts for Special Instances

In this section, we identify special instances for RWA-MM-ONoC on mesh ONoC, and propose a routing algorithm which can achieve the optimal solutions for those special instances.

A. *Multiple Multicasts Distribution for Special Instances*

Since routing selections depend on the distributions of source and destination nodes on a mesh network, we need to investigate the relative distributions among the source nodes and destination nodes of the multicasts. We identify some special node distributions as follows:

- Same-row: all nodes have the same Y-axis coordinate;
- Same-column: all nodes have the same X-axis coordinate;
- Different-rows: all nodes have different Y-axis coordinates;
- Different-columns: all nodes have different X-axis coordinates.

For a set of multiple multicasts, there are two groups of nodes: source node group and destination node group. Since each group of source and destination nodes may satisfy one of those 4 special distributions, there are 4×4=16 special distributions for a set of multiple multicasts, as shown in Table I. For example, distribution ① in Table I indicates that all sources in a set of multiple multicasts are in the same row and all destinations are also sharing one row. **Different rows (columns) for the sources** in Table I means that the source nodes are distributed on different rows (columns), while **Different rows (columns) for the destinations** represents that each row (column) can only be occupied by destinations belonging to the same multicast, i.e., the destinations of any two multicasts do not share any rows (columns).

TABLE I
16 Special Distributions of Nodes for Multiple Multicasts

Sources / Destinations	Same row	Same column	Different rows	Different columns
Same row	①	⑤	⑨	⑬
Same column	②	⑥	⑩	⑭
Different rows	③	⑦	⑪	⑮
Different columns	④	⑧	⑫	⑯

Among the 16 special distributions, some of them have the same features that can be merged. For example, distribution ① is the special case of distribution ⑯, because nodes must be in different columns if those nodes are in the same row. So, ①, ④, ⑬, ⑯ can be merged to one distribution, called **Instance 1**. Similarly, ⑤, ⑧, ⑨, ⑫ can be merged to one distribution, as **Instance 2**. ②, ③, ⑭, ⑮ can be merged to one distribution as **Instance 3**, and ⑥, ⑦, ⑩, ⑪ as **Instance 4**. Hence, we reduce the 16 distributions to 4 special instances and propose the routing schemes for each instance as follows.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

B. Routing Schemes for Special Instances

According to the distribution of source and destination nodes of multiple multicasts, we propose a routing algorithm which can select a proper routing scheme from the basic routing schemes (i.e., XY, YX, XYX, YXY) [36] for each special instance using only one wavelength. We derive the following theorems for these special instances.

Theorem 1. *(for Instance 1) Given a set of multicasts $M = \{m_1, ..., m_i, ..., m_C\}$, if the source and destinations of m_i do not share any* **columns** *with the source and destinations of m_j for any i, j $(1 \le i \le C, 1 \le j \le C, i \ne j)$, only* **one wavelength** *is required by using* **YXY routing** *to achieve M simultaneously.*

Proof. Since any multicast in the set does not share any columns with any other multicasts, each column hosts at most one multicast and the links of the column can be used by the multicast exclusively. It can also be deduced that there are at most n such multicasts for an $n \times n$ ONoC. Therefore, we can assign one row to each multicast as its dedicated row to avoid the overlaps among multicasts. For any multicast in the set, the YXY routing can be used via its dedicated row to find its multicast path that is non-overlapping with other multicasts, as follows. First, route from the source along the Y-axis to find the dedicated row, then route along the X-axis of the dedicated row, and find the column of each destination of the multicast, and finally route in the Y-axis to reach the destinations. Since the columns and the rows host the multicast exclusively, the resulting multicast path will not overlap with the paths of any other multicasts in the set. Therefore, only one wavelength is needed for routing a set of multicasts that satisfies the condition of this theorem. □

Theorem 2. *(for Instance 2) Given a set of multicasts $M = \{m_1, ..., m_i, ..., m_C\}$, if the source of m_i does not share any* **rows** *with the source of m_j for any i, j $(1 \le i \le C, 1 \le j \le C, i \ne j)$ and the destinations of m_i do not share any* **columns** *with the destinations of m_j for any i, j, only* **one wavelength** *is required by using* **XY routing** *to achieve M simultaneously.*

Theorem 3. *(for Instance 3) Given a set of multicasts $M = \{m_1, ..., m_i, ..., m_C\}$, if the source of m_i does not share any* **columns** *with the source of m_j for any i, j $(1 \le i \le C, 1 \le j \le C, i \ne j)$ and the destinations of m_i do not share any* **rows** *with the destinations of m_j for any i, j, only* **one wavelength** *is required by using* **YX routing** *to achieve M simultaneously.*

Theorem 4. *(for Instance 4) Given a set of multicasts $M = \{m_1, ..., m_i, ..., m_C\}$, if the source and destinations of m_i do not share any* **rows** *with the source and the destinations of m_j for any i, j $(1 \le i \le C, 1 \le j \le C, i \ne j)$, only* **one wavelength** *is required by using* **XYX routing** *to achieve M simultaneously.*

The proof of Theorem 2 is similar to the proof of Theorem 1, so we omit it here. The proof for Theorem 3 and 4 are also similar by rotating the instance with 90 degree on the network.

The above theorems provide routing solutions for special instances of RWA-MM-ONoC problem, which have the following advantages: **(i) Only one wavelength is needed.** If the distribution of multicast nodes satisfies one of the 4 special instances, optimal results can be achieved by using the corresponding routing theorems to transmit packets with only one wavelength. **(ii) At most two turn-around counts is needed.** Low number of turn-around counts can not only reduce the power consumption but also avoid the high micro-ring resonator insertion loss in ONoC. XY and YX routing have only one turn-around count, while XYX and YXY routing have two. **(iii) Routing complexity is low.** XY, YX, XYX, and YXY are minimal-path routing algorithms which are easy to implement without using any routing tables.

By taking advantages of the results obtained from the special instances, we further extend the design to general instances, where the distribution of multicasts is random in the following section.

V. Heuristic RWA of Multiple Multicasts for General Instances

In this section, we propose a heuristic routing algorithm to solve general instances of RWA-MM-ONoC. The main idea of this algorithm is to partition all multicast nodes into a number of sub-groups, with each sub-group satisfying one of the special instances using the corresponding routing theorems. We call this algorithm Group-Partitioning Routing algorithm for Multiple Multicasts (GPRMM).

In GPRMM, we select proper sources and destinations to form a group, with two selection policies: source selection policy and destination selection policy. Source (or destination) selection policy is to scan and select the sources (or destinations) row by row (row-based) or column by column (column-based) according to some selection factors, which are defined as follows:

Definition 1. Multicast Density (MD) is the maximum number of multicasts in a row (column), denoted as σ^r (σ^c).

Definition 2. Source Density (SD) is the maximum number of multicasts whose sources are in a row (column), denoted as σ_s^r (σ_s^c).

Definition 3. Destination Density (DD) is the maximum number of multicasts whose destinations are in a row (column), denoted as σ_d^r (σ_d^c).

In other words, MD is the number of multicasts in the most dense row or column, since this row or column accommodates the largest number of multicasts compared with other rows or columns. Similarly, SD is the number of multicasts whose sources are in the most dense row or column, and DD is the number of multicasts whose destinations are in the most dense row or column. SD and DD are used as the criteria to choose the proper selection policy, as showed in Table II. For example, when $\sigma_s^r = \sigma_s^c$ and $\sigma_d^r \ge \sigma_d^c$, we use row-based and column-based policies to select sources and destinations respectively to form a group.

The design of the selection policy is explained as follows. The column-based selection policy is to scan and select the

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

TABLE II
SOURCES AND DESTINATIONS SELECTION POLICY BASED ON SD AND DD

	$\sigma_d^r \geq \sigma_d^c$	$\sigma_d^r < \sigma_d^c$
$\sigma_s^r = \sigma_s^c$	Source row-based Destination column-based	Source column-based Destination row-based
$\sigma_s^r > \sigma_s^c$	Source column-base/ Destination row-based	
$\sigma_s^r < \sigma_s^c$	Source row-based/ Destination column-based	

nodes column by column to form a group, with at least σ_d^r destinations and σ_s^r sources selected for each group. This is because at least one source can be selected from each of the columns whose X-axis coordinate is the same with the source calculated in σ_s^r, and at least one destination can be selected from each of the corresponding columns calculated in σ_d^r. Similarly, the row-based selection policy is to scan and select the nodes row by row to form a group, with at least σ_d^r destinations and σ_s^r sources selected for each group. When $\sigma_d^r > \sigma_d^c$, column-based selection can select more destinations than row-based selection. Likewise, when $\sigma_d^r < \sigma_d^c$, row-based selection can select more destinations than column-based selection. Since the objective of GPRMM is to obtain minimum number of groups, the proper selection policy can make the first selected group to contain as many multicast nodes as possible. It is worth noting that we only need to use this selection policy when $\sigma^r > 1$ and $\sigma^c > 1$, because the distributions of multicast nodes satisfy the conditions of Theorem 2 and Theorem 1 respectively when $\sigma^r = 1$ and $\sigma^c = 1$ using only one wavelength by the corresponding routing theorems.

The steps of GPRMM to form the groups are:

Step 1 **Sort all multicasts**. Sort the multiple multicasts in an ascending order according to their number of multicast nodes.

Step 2 **Assign priorities**. Assign to the multicasts unique priorities from high to low according to the above sorted order. The nodes in a multicast inherit the priority of the multicast.

Step 3 **Select sources for group** i. Select sources with the highest priority in each row (or column if the source selection policy is column-based according to the above policy) and mark them as selected. Mark the sources that have not been chosen and the corresponding destinations as unselected.

Step 4 **Select destinations for group** i. Select destinations in each row (or column if the destination selection policy is column-based as the above policy) and mark them as selected.

Step 5 **Check sources**. For any selected source, if it still has some destinations unselected, mark the source as unselected.

Step 6 **Repeat**. Repeat steps 3 to 5 until all multicast nodes are selected.

The pseudocode for GPRMM is given in Algorithm 1 which has the computation complexity of $O(MN^2)$ on a mesh-based ONoC with N nodes and M multicasts. It can be seen that GPRMM is a polynomial time algorithm with low time complexity, which is related to the network size and the number of multicasts. The following results are derived for the wavelength requirement of GPRMM.

Lemma 1. *The number of wavelengths by GPRMM is no more than the number of groups.*

Proof. Since each group formed by GPRMM satisfies one of the distributions for special instances, we can use the corresponding routing policies in Theorem 1-4 by only one wavelength for each group. If the routing paths of each group have conflicts with any other groups, each group is assigned one distinct wavelength. Hence, the number of wavelengths is equal to the number of groups. Otherwise, some groups can share the same wavelength. So the number of wavelengths is no more than the number of groups. □

Lemma 2. *In an $n \times n$ mesh ONoC, the number of groups by GPRMM is at most $max\{\sigma_d^r, \sigma_d^c\}$.*

Proof. As the number of destinations is more than the number of sources for a group of multicasts, the key issue of partitioning all multicast nodes to different groups is how to partition all destinations into different groups. If we always use column-based selection policy to select destinations, at most σ_d^c groups are derived. Because only destinations belonging to one multicast can be selected for each group considering the column with σ_d^c multicasts, σ_d^c multicasts should be partitioned into at most σ_d^c groups using column-based selection policy. Similarly, at most σ_d^r groups are derived if we always use row-based selection policy to select destinations. If we use column-based and row-based selection policies alternately, multiple multicasts in the row (column) with σ_d^r (σ_d^c) may be selected simultaneously. This makes σ_d^r (σ_d^c) multicasts in a row or column be partitioned into Q ($Q < \sigma_d^r$ or $Q < \sigma_d^c$) groups, which is less than $max\{\sigma_d^r, \sigma_d^c\}$. Therefore, the number of groups derived by GPRMM is no more than $max\{\sigma_d^r, \sigma_d^c\}$. □

By Lemma 2, we can see that the wavelength requirement of GPRMM is dependent on DD, i.e., the number of multicasts whose destinations are in the most dense row or column.

Theorem 5. *In an $n \times n$ mesh ONoC, the upper bound on the number of wavelengths is $max\{\sigma_d^r, \sigma_d^c\}$ for any multiple multicasts by GPRMM. The number of wavelengths is also upper bounded by the network size n for an $n \times n$ mesh network considering no nodes overlapping.*

Proof. Let λ' be the number of wavelengths derived by GPRMM and T' be the number of groups. By Lemma 1 and Lemma 2, we have $\lambda' \leq T'$ and $T' \leq max\{\sigma_d^r, \sigma_d^c\}$. So $\lambda' \leq max\{\sigma_d^r, \sigma_d^c\}$. If there is no multicast nodes overlapping, i.e., any node involved in a multicast is either a source or a destination of the multicast, we have $max\{\sigma_d^r, \sigma_d^c\} \leq n$ in an $n \times n$ mesh. Hence, $\lambda' \leq n$. Therefore, the maximum number of wavelengths derived by GPRMM for an $n \times n$ mesh network is n in this situation. □

For tree-based and path-based routing schemes, we can always find some instances that need more than n wavelengths when the number of multicasts increases in an $n \times n$ mesh network.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

Algorithm 1: Group-Partitioning Routing Algorithm for Multiple Multicasts (GPRMM)

```
Input  : Source Set S, Destination Set D
Output: Group
C ← ∅; F ← ∅; Group = 0; D′ = D;
Sort the multiple multicasts in an ascending order
  according to their number of nodes;
Assign to the multicasts unique priorities from high to
  low according to the above sorted order;
do
    if (σ_s^r = σ_s^c and σ_d^r = σ_d^c) or (σ_s^r = σ_s^c and
     σ_d^r > σ_d^c) or (σ_s^r < σ_s^c) then
        Source_row_selection();
        Destination_column_selection();
    end
    else if (σ_s^r = σ_s^c and σ_d^r < σ_d^c) or (σ_s^r > σ_s^c) then
        Source_column_selection();
        Destination_row_selection();
    end
    Check_destinations();
while C ≠ D;
Function Destination_row_selection()
for each destination in D′ in every row do
    if destinations have the highest priority then
        Put them to F;
        C ← C ∪ F; D′ = D′ − F;  F ← ∅;
          Group++;
    end
    return D′;
end
Function Destination_column_selection()
for each destination in D′ in every column do
    if destinations has the highest priority then
        Put them to F;
        C ← C ∪ F; D′ = D′ − F;  F ← ∅;
          Group++;
    end
    return D′;
end
Function Source_row_selection()
for each source in every row do
    if a source has the highest priority then
        Put it to S′;
        Mark the sources that are not in S′ and the
          corresponding destinations as unselected;
    end
end
Function Source_column_selection()
for each source in every column do
    if a node has the highest priority then
        Put it to S′;
        Mark the sources that are not in S′ and the
          corresponding destinations as unselected;
    end
end
Function Check_destinations()
while the source in the group still has some
  destinations unselected do
    Keep the source in the mesh network;
end
```

Theorem 6. *GPRMM is deadlock free.*

Proof. Deadlock can occur if packets are allowed to hold some resources (e.g., buffer, channel) while requesting others. In an all-optical ONoC, there is no optical buffer because of the intrinsic properties of optical device, so we only need to consider whether channel usage will cause deadlock. XY, YX, XYX, YXY routings are dimension-ordering routings used in our proposed scheme, which are deadlock free [37][38]. Within each group derived by GPRMM, only one of these four routing schemes is used, so the routing for each group is deadlock free. Among multiple groups, different groups use different wavelengths. There is no situation that one group needs to wait for a channel that is being held by others, because multiple groups can use the same link simultaneously by different wavelengths. Therefore, the proposed routing algorithm is deadlock free. □

VI. Lower bound on the number of wavelengths

In this section, we derive the lower bound on the number of wavelengths required for a set of multicasts based on the concept of network cut [39]. In graph theory, a *cut* is a partition of the vertices of a graph into two disjoint subsets (V_1, V_2). Any *cut* determines a *cut-set*, the set of edges that have one endpoint in each subset of the partition. In ONoC, packets are transmitted from the source to the destination through one lightpath. If the source and the destination are in different subsets, a lightpath must pass through the cut-set edges. For a multicast with d destinations, one subset must contain the source node and i destination nodes $(1 \leq i \leq d)$ while the other subset contains $d-i$ destination nodes. As long as there is a destination in the different subset with the source, at least a lightpath is passing through cut-set edges. In order to obtain the lower bound of wavelengths for multiple multicasts, we need to consider the total number of transmissions passing through the cut-set edges and the number of cut-set edges. Assume the sum of lightpaths that sources and destinations communicate through cut-set is *H*. Since communications between V_1 and V_2 must pass through the edges in cut-set *K*, the number of distinct wavelengths required is at least

$$W(K,G) = \lceil \frac{|H|}{|K|} \rceil, \tag{9}$$

where $|H|$ is the number of lightpaths in the multiple multicasts passing through K and $|K|$ is the number of edges in cut-set.

Since different cuts K on the network may result in different values of $W(K,G)$, the greatest $W(K,G)$ can be identified as a lower bound on the number of wavelengths, denoted by $W_{LB}(K,G)$, which can be calculated by:

$$W_{LB}(K,G)) = \max_{\forall K} W(K,G) = \max_{\forall K} \lceil \frac{|H|}{|K|} \rceil. \tag{10}$$

The lower bound obtained in (10) provides a general result which can be applied to any ONoC topologies. For mesh network, a cut can be classified into *row-cut* and *column-cut*. A *row-cut* is a cut that vertices of a network are partitioned

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

by the horizontal cut. Similarly, a *column-cut* is a cut that vertices of a network are partitioned by the vertical cut. A lower bound of required wavelengths for multiple multicasts on mesh ONoC, denoted as $W_{LB}(K,M)$, can be derived by the following Theorem.

Theorem 7. *In an $n\times n$ mesh ONoC, the lower bound on the number of wavelengths for multiple multicasts is:*

$$W_{LB}(K,M) = max\{W^r_{min}, W^c_{min}\}, \tag{11}$$

$$W^r_{min} = \max_{\forall row-cut} \lceil \frac{\sum_{i=1}^{C} \delta_{i,s/d}}{n} \rceil, \tag{12}$$

$$W^c_{min} = \max_{\forall column-cut} \lceil \frac{\sum_{i=1}^{C} \delta_{i,s/d}}{n} \rceil, \tag{13}$$

$$\delta_{i,s/d} = \begin{cases} 1, \text{ if } s_i \in V_1, \exists d_{i,j} \in V_2 \text{ or } s_i \in V_2, \exists d_{i,j} \in V_1, \\ 0, \text{ otherwise.} \end{cases} \tag{14}$$

W^r_{min} is the minimum number of wavelengths based on row-cut and W^c_{min} is the minimum number of wavelengths based on column-cut. $\delta_{i,s/d}$ is an index to represent whether there is a lightpath between two subsets for multicast m_i. For a given multicast, if at least one destination is located in the different subsets with the corresponding source, a lightpath is needed with $\delta_{i,s/d} = 1$. Otherwise, $\delta_{i,s/d} = 0$ if the source and destinations are all in the same subset. The number of cut-set edges for an $n \times n$ mesh network is n.

In order to prove the result, we only need to prove that the minimum number of wavelengths for any cut is:

$$W \geq \lceil \frac{\sum_{i=1}^{C} \delta_{i,s/d}}{n} \rceil. \tag{15}$$

Proof. We use mathematical induction to prove it.

Basis: When C=1, one multicast only needs one wavelength, so the left-hand side (W) is simply equal to 1. In the right-hand side of the inequality, $\lceil \frac{\sum_{i=1}^{C} \delta_{i,s/d}}{n} \rceil = \lceil \frac{\delta_{1,s/d}}{n} \rceil$. There are 2 situations to be considered: (1) In this cut, there is at least one destination locating in the different subsets with the corresponding source. So, $\delta_{1,s/d} = 1$ resulting in the right-hand side equal to 1. (2) In this cut, all destinations are in the same subset with the corresponding source with $\delta_{1,s/d} = 0$, so the right-hand side is equal to 0. As the left-hand side is larger than or equal to the right-hand side in both situations, the statement is true for $C = 1$.

Inductive step: Assume that the statement is true when $C = m$. So, $W_m \geq \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d}}{n} \rceil$. When $C = m+1$, in the left-hand side of the inequality, $W_{m+1} \geq W_m$, which means $W_{m+1} = W_m$ or $W_{m+1} = W_m + 1$. In the right-hand side,

$$\lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil = \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d} + \delta_{m+1,s/d}}{n} \rceil. \tag{16}$$

For the $(m+1)th$ multicast, if all destinations are in the same subset with the corresponding source, $\delta_{m+1,s/d} = 0$. So $\lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil = \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d}}{n} \rceil$. According to the assumption, $W_{m+1} \geq \lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil$, so the statement is true. On the other hand, if at least one destination is in the different subset with the corresponding source, $\delta_{m+1,s/d} = 1$. In this situation, the right-hand side is:

$$\begin{aligned} \lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil &= \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d} + \delta_{m+1,s/d}}{n} \rceil \\ &= \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d} + 1}{n} \rceil \geq \frac{\sum_{i=1}^{m} \delta_{i,s/d}}{n} + \frac{1}{n}. \end{aligned} \tag{17}$$

The left-hand side is:

$$\begin{aligned} W_{m+1} = W_m + 1 &\geq \lceil \frac{\sum_{i=1}^{m} \delta_{i,s/d}}{n} \rceil + 1 \\ &\geq \frac{\sum_{i=1}^{m} \delta_{i,s/d}}{n} + 1. \end{aligned} \tag{18}$$

By (17) and (18),

$$W_{m+1} - \lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil \geq 1 - \frac{1}{n} \geq 0. \tag{19}$$

So, $W_{m+1} \geq \lceil \frac{\sum_{i=1}^{m+1} \delta_{i,s/d}}{n} \rceil$, and the statement is true.

Since both the basis and the inductive step have been performed, by mathematical induction, the statement for any multicast is true. □

This lower bound may not always be achieved, since the routing of paths within the network is not determined. It is a very useful measure to verify the efficiency of the routing and wavelength assignment scheme.

VII. System Implementation

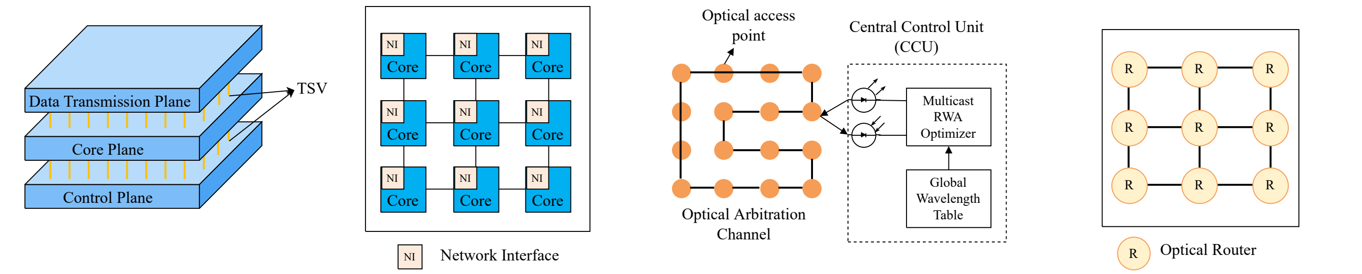
To implement GPRMM in an ONoC system, we use a hierarchical architecture that was designed in [31]. The hierarchical architecture contains three planes: core plane, control plane and data transmission plane (as shown in Fig. 5 (a)). The functions of the corresponding planes work as follows.

(1) Core plane. In the core plane (Fig. 5 (b)), each core connects with the network through a network interface (NI), by which each core also connects the optical control plane with an optical access point and connects the data transmission plane with an optical router. The network interface works as the coordination with the other two planes, such as sending the multicast routing requests to the optical control plane and configuring the connected optical router in the data transmission plane to construct the routing paths.

(2) Control plane. The control plane (Fig. 5 (c)) is a key module in the ONoC architecture, which consists of a Centralized Control Unit (CCU) and a cyclic optical arbitration channel. CCU gathers the multiple multicast requests coming from the core plane, carries out multicast routing and wavelength allocation on a global view of wavelength utilization, and dynamically configures the multicast paths in the data transmission plane. CCU consists of a global wavelength table and a multicast Routing and Wavelength Assignment (RWA) optimizer. The global wavelength table maintains the wavelength utilization of each optical link, while the multicast RWA optimizer optimizes the multicast routing path by optimizing the wavelength usage. When multicast paths belonging to the same group are allocated, CCU updates the global wavelength table and sends out the configuration packets to all the cores located in the multicast path.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

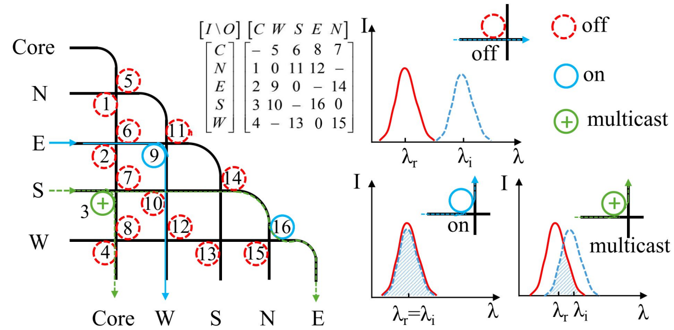


(a) Three-plane Architecture (b) Core Plane (c) Control Plane (d) Data Transmission Plane

Fig. 5. A three-plane ONoC architecture for implementing GPRMM.

The optical arbitration channel only transmits the control packets between the network interfaces and CCU, e.g., multicast routing requests and multicast configuration. Thus, each network interface in the core plane has an optical access point in the optical control channel. Generally, the multicast packet size depends on the application. For a cacheline invalidation message, the multicast packets generally have small size (e.g., 4 bytes). For other applications like real data traces blackscholes run on PARSEC benchmark, the packet size is about 9 bytes. For the multicast construction, the arbitration channel only needs to implement an N-to-1 and an 1-to-N optical buses (i.e., N is the total number of nodes in the network). In the N-to-1 optical bus, all the network interfaces of N cores use different wavelengths to send multicast requests to CCU in parallel, while in the 1-to-N optical bus, CCU uses different wavelengths to simultaneously send configuration packets to the network interfaces of cores located on the allocated multicast path. Therefore, the maximum number of cores that can send their multicast requests depends on the maximum number of wavelengths. If the number of wavelengths transmitting the multicast requests is not enough, we can use multiple time slots to transmit the multicast requests.

(3) Data transmission plane. The data transmission plane (Fig. 5 (d)) is a configurable wavelength-routed optical communication network that transmits multicast packets passively from the source core to the destination cores in accordance with the configured routing paths and the allocated wavelengths. It consists of multicast-enabled optical routers and bidirectional optical links. As shown in Fig. 6 (a), each optical router has five pairs of input/output ports to connect with four neighboring routers in different directions (North, East, South, West) and a local core. The optical router utilizes the active Micro-ring Resonators (MR) to implement light splitting and configurable optical switching for dynamically established routing paths. Each MR in the router has a unique resonant wavelength as a wavelength-selective filter, and it can be tuned by using thermal-optical or electrical-optical effects [24], which works as follows shown in Fig. 6 (b): if the resonant wavelength of MR (λ_i) is tuned to mismatch with the optical input signal (λ_r) at off-state (i.e., $\lambda_i \neq \lambda_r$), the optical signal will transmit along the original waveguide; if the resonant wavelength is tuned to fully match the input signal at on-state, (i.e., $\lambda_i = \lambda_r$), the optical signal will be coupled into the MR and transmit along the other waveguide. Moreover, in order to achieve multicast routing, the resonant wavelength of MR can be tuned to partially match the input signal at multicast-state (i.e., $\lambda_i \cap \lambda_r$), thus only a part of optical signal is coupled to MR and it can be output to both waveguides. In order to provide wavelength-routed multicast communication, the MRs in the optical router can be configured according to the routing matrix as shown in Fig. 6 (a). For example, to implement the light splitting, if the optical signal inputs from the south port and outputs to the connected core and the east port, MR 3 is tuned to multicast state and MR 16 is tuned to on-state. Some communications, e.g., from the north input to the west output, do not need to tune any MR, so they are labelled to 0 in the routing matrix.



(a) (b)

Fig. 6. The principle of multicast-enabled optical router, (a) the router architecture for a single wavelength; (b) different switch status by tuning the resonant wavelength of MR.

Based on the ONoC architecture, the communication process is executed by the following 3 steps:

Step 1 (Multicast request delivery): In the core plane, the cores having multicast requests send the communication requests to the control plane via the corresponding optical access point in a fixed time slot (e.g., 1 cycle). The communication requests include the addresses of source cores and destination cores. Then, the corresponding access point of the source core sends the multicast packets to CCU through the optical channel in the control plane.

Step 2 (Routing and wavelength computation): After receiving the multicast requests, CCU periodically calculates the routing paths and allocates optical wavelengths using GPRMM, according to the distribution of destination cores and the global wavelength usage of each link. CCU updates the global wavelength table after the routing paths are allocated, and transmits the configuration packets to all the cores allocated on the multicast paths.

Step 3 (Network configuration): When the cores involved in a multicast routing path receive the configuration packets, the corresponding interface changes the interconnection state

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

of the connected optical router to establish the routing. The MRs in the corresponding routers can be configured with the assigned wavelength and appropriate split ratio based on the number of destinations in the path, where the split ratio per ring should be dynamically configured per multicast request. Hence, each MR has at most one configuration for wavelength and split ratio once a path is set up. After a multicast path is torn down, all splitters along that path should be reset.

VIII. Performance Evaluation

In this section, GPRMM is evaluated through extensive simulations using synthetic multicast traffic and real data traces. As the number of wavelengths is one of the most important factors affecting energy consumption and complexity of the chip, firstly, we focus on the comparison of the wavelength requirement in the simulation results. We compare our results with baseline multicast routing schemes, including tree-based routing (TB) and path-based routing (PB) methods. Then, we evaluate the overheads of GPRMM.

A. Simulation Setup

In order to evaluate the performance of GPRMM, a simulator is developed by using C++, which implements the proposed routing and wavelength assignment scheme. Specifically, the simulator consists of three parts: kernel network architecture, routing and wavelength allocator, and performance analyzer. The kernel network architecture is an $n \times n$ mesh-based ONoC constructed by the components of cores, optical routers and optical links. The routing and wavelength allocator implements the proposed routing and wavelength assignment scheme. The performance analyzer calculates the performance parameters, such as the number of wavelengths, according to the optical routing paths and wavelength assignment from the routing and wavelength allocator. The simulation parameters for ONoC are configured in the proposed simulator.

B. Synthetic-based Simulations

In the synthetic-based simulations, the multicast traffic is subjected to the following settings: (1) Each node produces the multicast packets independently with a data rate of ϕ packets/cycle/node. ϕ follows a Poisson distribution ($\phi \in [0, 1]$); (2) The source node and the destination nodes of each multicast are uniformly distributed. It is worth noting the multicast traffic for each node is generated according to the above distribution in advance and stored in the separate traffic files. The same multicast traffic files are used to evaluate the performance of different multicast routing schemes; (3) GPRMM is compared with other routing methods for different network sizes and different multicast proportions. The multicast proportion is defined as the percentage of multicast nodes participating in multicast communication to all nodes in the network. For example, in a 16×16 ONoC, 76 nodes will participate in the multicast communication when the multicast proportion is 30%. Because each multicast should have at least 3 nodes (i.e., one source and two destinations), there are at most 25 multicasts without nodes overlapping. In this paper, multiple multicasts means there are more than one multicast in the network. Hence, 2 to 25 multicasts (involving the addresses of source and destination nodes) can be selected randomly from the derived multicast traffic files. The size of mesh network is set to 8×8, 16×16 and 32×32. The lower bound (LB) is also presented in this simulation. Fig. 7 presents the average number of wavelengths required for different multicast proportions (30%, 50%, 90%) under different network sizes.

1) Multicast proportion is 30% (Fig. 7 (a)): In this case, there are at most 19, 76 and 307 multicast nodes without overlapping in an 8×8 ONoC (64 nodes), 16×16 ONoC (256 nodes) and 32×32 ONoC (1024 nodes) respectively. Accordingly, there are at most 7, 25 and 102 multicasts respectively. From Fig. 7 (a), we can see that GPRMM uses the least number of wavelengths compared with traditional routing schemes (i.e., TB and PB). The number of required wavelengths can be reduced by 22% and 37.6% in average compared to TB and PB respectively.

2) Multicast proportion is 50% (Fig. 7 (b)): There are at most 32 multicast nodes and at most 10 multicasts without overlapping in an 8×8 ONoC in this case. Similarly, there are at most 42 and 170 multicasts in 16×16 and 32×32 ONoC. Compared with TB and PB, GPRMM still use the least number of wavelengths under different network sizes. It can be seen that the number of required wavelengths is reduced by 17.7% and 26.2% in average compared with TB and PB respectively.

3) Multicast proportion is 90% (Fig. 7 (c)): When the multicast proportion is 90%, there are at most 58 multicast nodes and at most 19 multicasts in an 8×8 ONoC. Likewise, there are at most 230 (921) nodes and 76 (306) multicasts in a 16×16 (32×32) ONoC. GPRMM still requires the least number of wavelengths compared with other routing schemes. The number of required wavelengths is reduced by 9.8% and 17.8% in average compared with other routing schemes respectively.

In general, GPRMM shows obvious advantages over the TB and PB. This is because the TB and PB only consider each multicast individually, with routing paths fixed and unable to avoid link conflicts. While GPRMM considers multiple multicasts as a whole, the routing paths are established based on the distribution of multicast nodes. This can make GPRMM route packets via less congested links and alleviate the link sharing probability. Meanwhile, GPRMM has better scalability than traditional routing schemes (i.e., TB and PB). As can be seen from Fig. 7, GPRMM can always use the least number of wavelengths to transmit packets with the increasing of network size. Moreover, the wavelength used by GPRMM is the closest to the optimal value of wavelengths utilization (OPW), which can be obtained by the NIP model in LINGO solver on only small instance of 8×8 ONoC.

The lower bound on the number of wavelengths is not tight enough in Fig. 7. This is because the lower bound is derived using the network cut theory, on which we assume a lightpath passes through the cut-set as long as one destination is located in the different subsets with the corresponding source. When the number of destinations for each multicast is smaller, this lower bound is closer to the tight bound, while the quality of

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edi
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

the lower bound is reduced if the number of destinations for each multicast is large. Therefore, the lower bound needs to be improved by considering the relationship between the number of cut-set edges and the number of destinations located in the different subsets with the corresponding source.

C. Simulation with real data traces

In the trace-based simulations, the multicast communication is filtered from the inter-core communication of a 64-core system running PARSEC benchmark [40]. If a node transmits multiple packets with the same type to different destinations in the successive clock cycles, then it is considered as a multicast communication. The addresses of the source node and all the destination nodes are recorded in the multicast trace files. Similarly, the same multicast trace files are used in the simulations of different multicast routing schemes. Fig. 8 gives the simulation results with the real data traces.

It can be seen that GPRMM can reduce the number of wavelengths significantly in all the applications. In average, GPRMM can reduce the number of wavelengths by 31.5% compared with TB and PB. This is because TB and PB only consider one multicast. When there are multiple multicasts in ONoC simultaneously, they are very likely to cause high link conflicts, thus the advantage of GPRMM based on multiple multicasts can be well exploited in these applications. Therefore, GPRMM can achieve much better performance than the other multicast schemes.

D. Overhead Analysis

1) Transmission, Configuration, and Processing Latency: Network latency for multicast communication in ONoC is the time interval that a packet is transmitted from the source node until it is received by the destination node, which includes transmission latency, configuration latency and processing latency. In GPRMM, when a specific distribution of multiple multicasts is given in a time period, the average network latency (denoted by D), can be calculated as [41]:

$$D = D_{oe} + N_{hop}/V_{os} + D_{eo} + D_t + D_c, \tag{20}$$

where D_{oe} and D_{eo} are the latency for O-E and E-O conversions, respectively, N_{hop} is the average length of routing paths counted by the hops of routers, and V_{os} is the transmission speed of optical signal in a waveguide in ONoC. D_t is the processing latency for routing and wavelength allocation scheme by GPRMM. D_c is the network configuration latency, which includes the time that the local configuring unit changes the interconnection state of the connected optical router to establish the routing paths and the multicast packets modulated to optical signals and transmitted from the source core to all the destination cores. The parameter settings for all the schemes are summarized in Table III. In this simulation, GPRMM is compared with two basic routing methods (i.e., PB, TB), as well as ENoC with XY routing scheme. We can see that the average latency for different routing schemes are similar from Fig. 9 (a). This is because the network latency is mainly related to the average length of routing paths in ONoC from Equation (20). Due to the high speed of optical signal (e.g., $V_{os} = 8\ hops/cycle$ for 8×8 mesh-based ONoC in a $20\ mm \times 20\ mm$ chip working at the system clock $2GHz$ [13]), the difference of latency among different routing methods is very tiny. It can be also seen that the network latency for ENoC is much higher than GPRMM and other routing methods in ONoC. This is because ENoC should transmit packets hop by hop from the source node to the destination node in each communication process, while ONoC can use the pre-configured non-blocking optical routing paths between nodes.

TABLE III
LATENCY PARAMETERS

Parameter	Value	Parameter	Value
Channel bandwidth	$10\ Gbps/\lambda$	Clock frequency	$2\,GHZ$
Transmission speed	$8\,routers/cycle$	O-E/E-O delay	$1\,cycle$

2) Power Consumption: The power consumption for multicast communication in an ONoC includes two parts: optical power consumption P_o and electrical power consumption P_e. *(i)* The optical power consumption P_o incurred by the laser source and power attenuation of the optical devices along the routing paths, which can be calculated by [41]:

$$P_o = \frac{1}{\gamma} \times N_\lambda \times 10^{P_{ds}} \times 10^{IL_{wil}} \times N_{mn}, \tag{21}$$

where γ is the power efficiency of laser source, P_{ds} is the sensitivity of photodetector, N_{mn} is the number of nodes participating in the multicast communication, and IL_{wil} is the worst-case insertion loss of all optical devices along the optical routing paths, such as optical routers, waveguides and etc. According to (21), P_o is decided by the number of wavelengths N_λ and the worst-case insertion loss IL_{wil}, as γ and P_{ds} are constant parameters of optical devices. For a specific optical routing path, the insertion loss IL can be calculated by [41]:

$$IL = IL_l \times (N_{thop} - 1) + IL_r \times N_{thop} + IL_{eo} + IL_{oe}, \tag{22}$$

where N_{thop} is the total number of hops in the routing path; IL_l and IL_r are the insertion losses of one optical link and one optical router, respectively; IL_{eo} and IL_{oe} are the insertion losses of E-O and O-E converters, respectively. Because IL_l, IL_r, IL_{eo}, and IL_{oe} are constant parameters for specific optical devices and router structure, the worst-case insertion loss IL_{wil} is decided by the maximum length of the routing paths between the cores. Hence, the optical power consumption P_o is mainly determined by the number of wavelengths (N_λ) and the maximum length of the optical routing paths ($max(N_{thop})$). *(ii)* The electrical power consumption P_e mainly includes the static power for thermally tuning the MRs, and the dynamic power for modulation and photodetection. P_e can be calculated by [41]:

$$P_e = N_{MR} \times N_\lambda \times P_{MT} + (E_{EO} + E_{OE}) \times B_o \times N_\lambda \times \sum \theta_i, \tag{23}$$

where N_{MR} is the total number of MRs used per wavelength; P_{MT} is the power for tuning one MR; E_{EO} and E_{OE} are the energy costs for modulation (E-O conversion) and photodetection (O-E conversion) respectively; B_o is the bandwidth of optical link per wavelength; θ_i is the actual traffic load of a multicast node, $\theta_i \in [0, 1]$.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

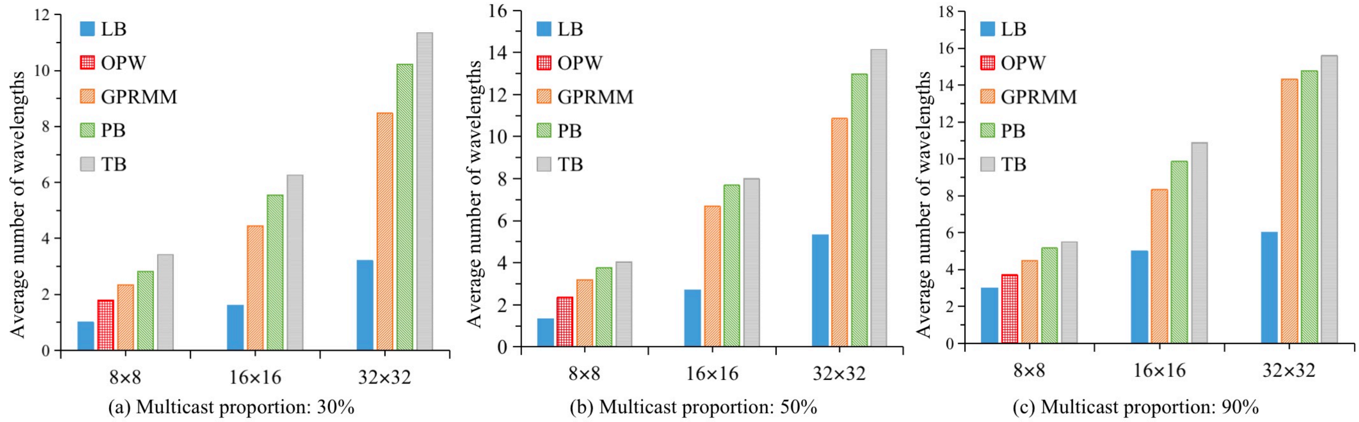


Fig. 7. Average number of wavelengths of GPRMM, TB, PB in 8×8, 16×16, 32×32 mesh ONoC under different multicast proportions.

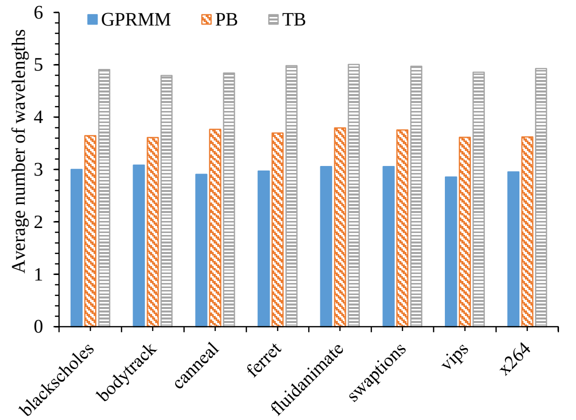


Fig. 8. Average number of wavelengths in different applications of trace-based simulations.

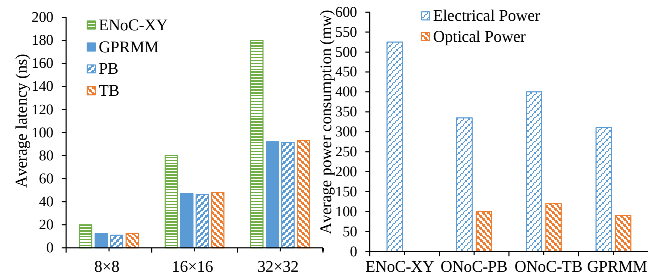


Fig. 9. Performance comparison for different routing schemes, (a) average latency under different network sizes; (b) power consumption under 8×8 mesh network.

To analyze the overall power consumption, we use the typical power parameters for the optical devices as in [42][43], which are listed in Table IV. Fig. 9 (b) shows the average power consumption for different routing schemes (i.e., GPRMM, PB, TB) in an ONoC and XY routing in an ENoC (ENoC-XY) in the 8×8 mesh network. We can see that GPRMM consumes the least overall power compared with other routing schemes. Specifically, ENoC has much higher power consumption since the packets are routed hop by hop and buffered in the electrical routers, especially when the distances between nodes increase. For PB and TB routings in the ONoC, they consume higher electrical power to tune the resonant wavelength of MRs, since they need more wavelengths than GPRMM. As shown in Fig. 9 (b), the total power consumption of GPRMM is about $410mw$, which is 24.7% less than the other ONoC schemes.

TABLE IV
POWER CONSUMPTION PARAMETERS OF OPTICAL DEVICES

Parameter	Value	Parameter	Value
Modulator	$85fJ/bit$	MR drop	$0.5dB/MR$
Photodetector	$50fJ/bit$	MR pass	$0.005dB/MR$
Receiver sensitivity	-26dBm	Thermal tuning	$26\mu W/MR$
Optical bandwidth	$10Gbps/\lambda$	Waveguide passing	$1.5dB/cm$

IX. CONCLUSIONS

In this paper, we first give the problem definition of routing and wavelength assignment for multiple multicasts on ONoC (RWA-MM-ONoC) and formulate it as an integer programming model. As far as we know, this is the first time this problem is formally defined. We investigate this problem from studying optimal routing methods for special instances on mesh ONoC according to the distribution of multicast nodes, which only need one wavelength for a group of multicasts. We can easily route the group of multicasts with just one wavelength as long as the distribution of the source and destination nodes satisfies the identified conditions in Theorem $1, 2, 3$ or 4. To solve the general instances of RWA-MM-ONoC, we propose a Group-Partitioning Routing algorithm for Multiple Multicasts (GPRMM) by partitioning multiple multicasts into a number of groups, where each group satisfies the conditions in Theorem $1, 2, 3$ or 4 using one wavelength. In addition, we provide the upper bound and lower bound on the number of wavelengths of GPRMM. The simulation results show that our GPRMM outperforms other routing schemes in terms of the number of wavelengths and power consumption. Compared with other traditional routing schemes (tree-based and path-based routings), GPRMM shows obvious advantages under different network sizes and different multicast proportions as summarized below: (i) GPRMM has low wavelength requirement. It uses the least number of wavelengths for different multicast proportions and network sizes. The number of wavelengths used by GPRMM is no more than the destination density, which is always smaller than the network size n for an $n \times n$ mesh network considering no multicast nodes overlapping, while traditional routing schemes cannot guarantee this. (ii) GPRMM has low complexity. It not only has a low polynomial time computation complexity, but also adopts simple routing policies (XY/YX/XYX/YXY) with at most 2 turn-around counts and easy implementation on

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edit
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

ONoC. (iii) GPRMM has good scalability. With the increasing of network sizes, the number of wavelengths required is always the least compared with traditional methods. (iv) GPRMM has low power consumption. GPRMM can achieve much higher communication performance and lower power consumption than the corresponding ENoC, and also achieve much lower power consumption than the traditional ONoC schemes by reducing the number of required wavelengths.

In our future work, we will extend the proposed methods to other topologies, such as torus, to achieve higher performance for multiple multicasts in ONoC.

REFERENCES

[1] "192-core CSX700 processor", *http://www.clearspeed.com/products/csx700.php*, 2012.

[2] "TrueNorth chip sign of new possibilities in brain-like computing", *http://techxplore.com/news/2015-08-truenorth-chip-possibilities-brain-like.html*, 2015.

[3] "International Technology Roadmap for Semiconductors (ITRS)", *http://www.itrs2.net/*.

[4] A.Shacham, K.Bergman, and L.Carloni, "Photonic Network on Chip for future generations of chip multiprocessors", in *IEEE Trans. Comput.*, vol. 57, pp. 1246-1260, Sep. 2008.

[5] D.Pile, "Integrated Photonics: Compact Multiplexing", in *Nature Photonics*, vol. 9, no. 2, pp. 78, Feb. 2015.

[6] F.Liu, H.Zhang, Y.Chen, Z.Huang and H.Gu, "Wavelength-Reused Hierarchical Optical Network on Chip Architecture for Manycore Processors", in *IEEE Trans. Sustainable Computing*, vol. 4, no. 2, pp. 231-244, 2019.

[7] H.Xu, P.K.Mckinley and L.M.Ni, "Efficient implementation of barrier synchronization in wormhole-routed hypercube multicomputers", in *J. Parallel and Distributed Computing*, pp. 172-184, 1992.

[8] M.Azevedo, and D.Blough, "Fault-Tolerant Clock Synchronization of Large Multicomputers via Multistep Interactive Convergence", in *IEEE Conf. ICDCS*, pp. 249-257, 1996.

[9] P.K.Mckinley, H.Xu, E.T.Kalns and L.M.Ni, "Compass: Efficient communication services for scalable architectures", in *IEEE Conf. Supercomputing*, pp. 478-487, Nov. 1992.

[10] S.Ma, N.Jerger, and Z.Wang, "Supporting efficient collective communication in NoCs", in *IEEE Conf. HPCA*, pp. 1-12, Feb. 2012.

[11] A.Karkar, T.Mak, K.F.Tong and A.Yakovlev, "A survey of Emerging Interconnects for On-chip Efficient Multicast and Broadcast in Manycores", in *IEEE Circuits Syst. Mag.*, pp. 58-72, 2016.

[12] K.Preston, N.Sherwood-Droz, J.S.Levy and M.Lipson, "Performance guidelines for WDM interconnects based on silicon microring resonators", in *IEEE Conf. CLEO*, pp. 1-2, May. 2011.

[13] J.Liu, J.Yang, and R.Melhem, "GASOLIN: Global Arbitration for Streams of Data in Optical Links", in *IEEE Int. Parallel and Distributed Processing Symposium (IPDPS)*, pp. 93-102, 2015

[14] K.Chen, M.Ebrahimi, T.Wang, and Y.Yang, "NoC-based DNN accelerator: A future design paradigm", in *Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip*, pp. 1-8. 2019.

[15] T.Krishna, L.S.Peh, B.M.Beckmann and S.K.Reinhardt, "Towards the ideal on-chip fabric for 1-to-many and many-to-1 communication", in *IEEE Conf. MICRO-44*, pp. 71-82, 2011.

[16] Y.Pan, P.Kumar, J.Kim, G.Memik, Y.Zhang, and A.Choudhary, "Firefly: Illuminating Future Network-on-Chip with Nanophotonics", in *ACM SIGARCH Computer Architecture News*, vol. 37, pp. 429-440, 2009.

[17] N.Prasad, P.Mukherjee, S.Chattopadhyay, and I.Chakrabarti, "Design and evaluation of ZMesh topology for on-chip interconnection networks", in *J. Parallel and Distributed Computing*, 113, pp.17-36, 2018.

[18] S.V.R.Chittamuru, S.Desai, and S.Pasricha, "Swiftnoc: a reconfigurable silicon-photonic network with multicast-enabled channel sharing for multicore architectures", in *ACM J. Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 4, p.58, 2017.

[19] M.F.Chatmen, A.Baganne, and R.Tourki, "Low latency multicasting architecture implemented using new network topology", in *Microprocessors and Microsystems*, vol. 68, pp.47-57, 2019.

[20] F.Meaad, H.Gu, and W.Wei, "DORR: a DOR-based non-blocking optical router for 3D photonic network-on-chips", in *IEICE TRANSACTIONS on Information and Systems*, vol. 104, no. 5, pp. 688-696, 2021.

[21] F.Samman, T.Hollstein, and M.Glesner, "Adaptive and Deadlock-free Tree-based Multicast Routing for Networks-on-Chip", in *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, pp. 1067-1080, Jul. 2010.

[22] W.Hu, Z.Lu, A.Jantsch and H.Liu, "Power-efficient tree based multicast support for Networks-on-Chip", in *Proc. the 16th Asia and Pacific Design Automation Conference (ASP-DAC)*, 2011.

[23] L. Wang, Y. Jin, H. J. Kim and E. J. Kim, "Recursive Partitioning Multicast: A Bandwidth-Efficient Routing for Networks-On-Chip," in *Int. Symp. Networks-on-Chip (NOCS)*, pp. 64-73, May 2009.

[24] A.K.Shoukath, P.Samapth, S.Elango, S.P.Philip, "Modified Multicast Routing Algorithm for Network-on-Chip", in *Int. J. Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, pp. 208-210, 2019.

[25] M.Ebrahimi, M.Daneshtalab, P.Liljeberg, J.Plosila, J.Flich and H.Tenhunen , "Path-based partitioning methods for 3D Networks-on-Chip with minimal adaptive routing", in *IEEE Trans. Comput.*, vol. 63, pp. 718-733, Mar. 2014.

[26] M.Ebrahimi, M.Daneshtalab, P.Liljeberg and H.Tenhunen, "HAMUM-A novel routing protocol for unicast and multicast traffic in MPSoCs", in *IEEE Conf. PDP*, pp. 525-532, 2010.

[27] P.Bahrebar, D.Stroobandt, "Improving Hamiltonian-based methods for on-chip networks: a true model approach", in *IEEE Conf. DATE*, pp. 1-4, Mar. 2014.

[28] C.W.Wu, K.J.Lee, and A.P.Su, "A Hybrid Multicast Routing Approach with Enhanced Methods for Mesh-based Networks-on-Chip". in *IEEE Trans. Comput.*, vol. 67, no.9, pp.1231-1245, 2018.

[29] W.Xu, N.Wu, and F.Ge, "Minimizing the number of wavelengths in a cluster WDM mesh-based ONoC through application-specific mapping", in *IEICE Electronics Express*, vol. 17, no. 6, pp. 20200045, 2020.

[30] J.Zhang, W.Zhang, J.Yuan, and H.Wang, "Implementing bitonic sorting on optical network-on-chip with bus topology", in *Photonic Network Communications*, vol. 39, no. 2. pp. 129-134, 2020.

[31] F.Liu, H.Zhang, Y.Chen, Z.Huang and H.Gu,"Dynamic Ring-Based Multicast with Wavelength Reuse for Optical Network on Chips", in *IEEE Conf. MCSoc*, pp. 153-160, 2016.

[32] S.Werner, J.Navaridas, and M.Lujan, "A survey on optical network-on-chip architectures", in *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p.89, 2018.

[33] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, "A 5-GHz Mesh Interconnect for a Teraflops Processor", in *IEEE Micro*, vol. 27, no. 5, pp. 51-61, 2007.

[34] D. Wentzlaff, P. Griffin, H. Hoffmann, L. Bao, B. Edwards, C. Ramey, M. Mattina, C.-C. Miao, J. F. B. III, and A. Agarwal, "On-Chip Interconnection Architecture of the Tile Processor", in *IEEE Micro*, vol. 27, no. 5, pp. 15-31, 2007.

[35] T.H.Cormen, C.E.Leiserson, R.L.Rivest and C.Stein, "Introduction to Algorithms (2nd ed.)", *MIT Press and McGraw-Hill*, p.580, 2001.

[36] W.Yang, Y.Chen, Z.Huang, H.Zhang, "RWADMM: Routing and Wavelength Assignment for Distribution-based Multiple Multicasts in ONoC", in *IEEE Conf. ISPA/IUCC*, pp. 550-557, 2017.

[37] P.N.Chopkar, and M.A.Gaikwad, "Review of XY routing algorithm for 2D torus topology of NoC architecture". in *Int. J. Computer Applications*, 975(8887), pp. 22-26, 2013.

[38] Y.Jin, E.J.Kim, and K.H.Yum, "Design and analysis of on-chip networks for large-scale cache systems", in *IEEE Trans. Comput.*, vol. 59, no. 3, pp. 332-344, 2009.

[39] T.H.Cormen, C.E.Leiserson, R.L.Rivest and C.Stein (2001), "Introduction to Algorithms (2nd ed.)", *MIT Press and McGraw-Hill*, pp. 563,655,1043.

[40] J. Hestness, B. Grot, and S. Keckler, "Netrace: dependency drive trace-based network-on-chip simulation", in *IEEE Proc. NoCArc*, 2010.

[41] F.Liu, "High-Performance and Wavelength-Reused Optical Network on Chip (ONoC) Architectures and Communication Schemes for Manycore Processor (Doctoral dissertation, University of Otago)", 2017.

[42] J.Chan, G.Hendry, A.Biberman, K.Bergman, and L.P.Carloni, "Phoenixsim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks", in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 691-696, March, 2010.

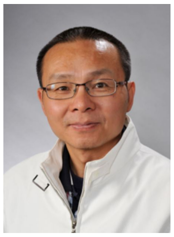
[43] X.Zhang, and A.Louri, "A multilayer nanophotonic interconnection network for on-chip many-core communications", in *Design Automation Conference*, pp. 156-161, June, 2010.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. This is the author's version which has not been fully edi
content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2023.3274951

Wen Yang received her PhD degree in Computer Science from the University of Otago in New Zealand in 2019, and Master degree of Communication and Information System in 2011 and Bachelor degree of Electronic Information Science and Engineering in 2008 from Central China Normal University, China. From 2019 to 2020, she worked as an assistant research fellow in University of Otago, New Zealand. She is currently a Lecturer with the College of Computer and Information Technology, China Three Gorges University, China. Her research interests include Network on Chip, Optical Network on Chip (ONoC), multicast communication, and routing algorithm, etc.

Yawen Chen obtained her PhD degree in Computer Science from The University of Adelaide in Australia in 2008, and Bachelor degree in Computer Science in 2002 and Master degree in 2004 from Shandong Normal (Teachers) University in China. She worked as a researcher in Japan Advanced Institute of Science and Technology (JAIST) in 2005. She is currently a Senior Lecturer with the Department of Computer Science, University of Otago, New Zealand. Her research interests include resource optimization and performance evaluation in computer networking and computer architecture (optical network, interconnection network, green computing).



Zhiyi Huang is a Professor at the Department of Computer Science, University of Otago. He received the BSc degree in 1986 and the PhD degree in 1992 in Computer Science from the National University of Defense Technology (NUDT) in China. He was a visiting professor at EPFL and Tsinghua University in 2005, and a visiting scientist at MIT CSAIL in 2009. His research fields include parallel/distributed computing, multicore architectures, operating systems, signal processing, green computing, cluster/-grid/cloud computing, high-performance computing, bio-engineering, and computer networks. He has more than 100 publications.

Haibo Zhang received the M.Sc. degree in Computer Science from Shandong Normal University, China, in 2005 and the PhD degree in Computer Science from the University of Adelaide, Australia, in 2009. From 2009 to 2010, he was a Postdoctoral Research Associate with the Automatic Control Lab at Royal Institute of Technology in Sweden. He is currently an Associate Professor with the Department of Computer Science, University of Otago, New Zealand. He has served on a number of technical program committee boards and co-chaired several international conferences. His current research interests include real-time industrial wireless communication, optical network-on-chips, wireless body sensor networks, delay-tolerant networks, green computing, distributed algorithms, and protocol design.

Huaxi Gu received his PhD degree in telecommunication and information system from Xidian University in 2005. He is currently a Professor in the state key lab of ISN, Xidian University, Xian, Shanxi, China. His current research interests include interconnection networks, networks-on-chip and optical interconnect, and data center networks. He has more than 100 publications in many international journals and conferences.

Authorized licensed use limited to: University of Ottawa. Downloaded on October 30,2023 at 17:39:37 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.