Minimizing Worst-Case Data Transmission Cycles in Wavelength-Routed Optical NoC through Bandwidth Allocation

Liaoyuan Cheng, Mengchu Li, Tsun-Ming Tseng, Ulf Schlichtmann
{liaoyuan.cheng,mengchu.li,tsun-ming.tseng,ulf.schlichtmann}@tum.de
Technical University of Munich, Munich, Germany

ABSTRACT

With the rapid development of integrated photonic technology, wavelength-routed optical networks-on-chip (WRONoC) is emerging as a high-potential computing architecture due to its low power consumption, high bandwidth, and conflict-free communication advantages. Previous works utilize the multi-resonance properties of the microring resonator (MRR), the key component in WRONoC, to transmit multiple signals on different wavelengths in one transmission path, thereby achieving parallel communication. However, they do not consider the demands of the actual application communication bandwidth. If communications with high bandwidth demands are not allocated with highly parallel transmission paths, they may become the bottleneck of the network, resulting in increased transmission cycles and overall data transmission time. In this work, we propose an optimization strategy that allocates signal wavelengths for each communication based on its actual bandwidth demand to reduce communication time. Specifically, based on the actual bandwidth demands and topology structure, we first map the communication nodes in the target application to the ports of a WRONoC topology. Next, we optimize the radii of MRRs in the topology and allocate the signal wavelengths to each transmission path to minimize the worst-case data transmission cycle. Experimental results show that, compared to methods that only consider communication parallelism, our strategy can reduce the worst-case of data transmission cycles by over five times, thereby significantly decreasing the time required for data transmission.

KEYWORDS

Wavelength-routed optical networks-on-chip, design optimization, bandwidth allocation, microring resonators

ACM Reference Format:
Liaoyuan Cheng, Mengchu Li, Tsun-Ming Tseng, Ulf Schlichtmann. 2024. Minimizing Worst-Case Data Transmission Cycles in Wavelength-Routed Optical NoC through Bandwidth Allocation. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD '24), October 27–31, 2024, New York, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3676536.3676802

ICCAD '24, October 27–31, 2024, New York, NY, USA

ACM ISBN 979-8-4007-1077-3/24/10…$15.00
https://doi.org/10.1145/3676536.3676802

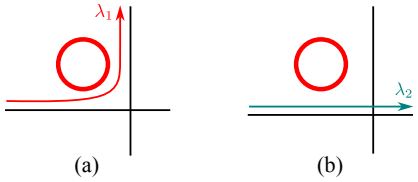

Figure 1: MRR coupling mechanism. (a) The signal on resonant wavelength λ_1 can be coupled to MRR and be turned. (b) The signal on non-resonant wavelength λ_2 cannot be coupled to the MRR.

1 INTRODUCTION

The rise of data-intensive applications requires using many processing cores to perform multiple operations simultaneously for rapid analysis [1, 2]. This need has driven the development of multiprocessor systems-on-chip (MPSoC) to integrate hundreds of processor cores on a single chip [3]. However, traditional electrical interconnects face challenges such as high signal noise and propagation delays[3]. To overcome these limitations, optical network-on-chips (ONoCs) enabled by the rapid advances in integrated photonics have become an attractive solution. Using wavelength-division-multiplexing (WDM) to transmit multiple wavelengths simultaneously in a single waveguide, ONoC provides high-bandwidth, low-latency communications and is considered a key technology for the next-generation MPSoC [4].

As a primary type of ONoC, wavelength-routed ONoC (WRONoC) defines all signal transmission paths between ports during the design phase. In this setup, when electrical signals are converted to optical signals and enter the WRONoC ports, they can reach their communication destinations in parallel and without conflict along pre-designed paths, which significantly reduces the transmission delays and power consumption associated with dynamically configuring paths [5]. In WRONoC, optical waveguides serve as the fundamental signal transmission medium, and microring resonators (MRRs), which are ring-shaped waveguides, are key components for signal routing. Specifically, when a signal traveling through a waveguide encounters an MRR, coupling occurs if the MRR's optical path length, determined by its radius and effective index, matches an integer multiple of the signal's wavelength [6]. As shown in Figure 1(a), the signal on λ_1 is "on-resonance" with the red MRR and thus will be turned by it. This condition is known as "on-resonance", and the corresponding wavelengths are termed "resonant wavelengths". Conversely, if this condition is not met, the signal will not be coupled to, also called "off-resonance" with, the MRR. As shown in Figure 1(b), the signal on λ_2 is "off-resonance" with the red MRR and will not be turned.

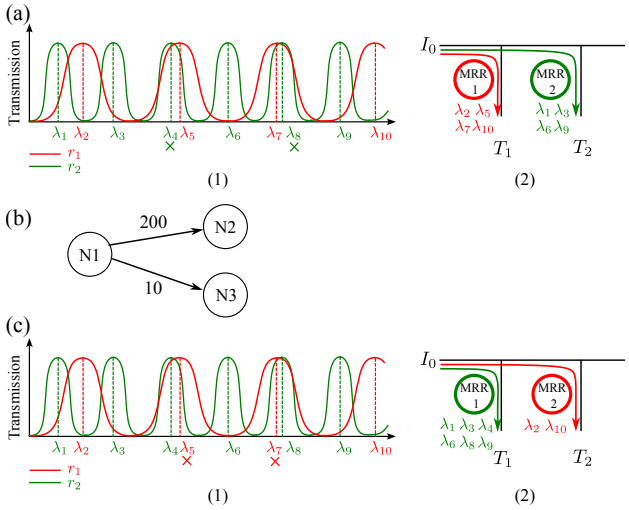

Figure 2: (a) (1) The transmission spectra of MRRs with radius r_1 and r_2, denoted by the red and green line, respectively. The wavelengths corresponding to the peaks represent the resonant wavelengths of the same colored lines. Wavelengths marked with a cross cannot be selected in the parallelism maximization method. (2) The wavelength usage of the parallelism maximization method. (b) Application communication graph with bandwidth demands. (c) (1) The same transmission spectra of MRRs as (a)(1). Wavelengths marked with a cross cannot be selected in the bandwidth allocation method. (2) The wavelength usage of the bandwidth allocation method.

Although an MRR supports multiple resonant wavelengths, most studies of WRONoC designs assume that each communication between ports uses only one signal wavelength, neglecting the potential for parallel communication [7–9]. Recent works [10] and [5] proposed using the multi-resonance properties of MRRs to enhance communication parallelism by optimizing the selection of MRR radii, referred to as the "parallelism maximization method". For example, as shown in Figure 2(a), the red and green lines represent the transmission spectra of MRRs with radii r_1 and r_2, respectively. The wavelengths corresponding to the spectral peaks represent their respective resonant wavelengths. As shown in Figure 2(a)(2), to achieve conflict-free signal transmission from port I_0 to ports T_1 and T_2, the following wavelength assignment conditions must be met: (1) The signal wavelengths for path (I_0, T_1) must resonate with MRR 1. (2) The signal wavelengths for path (I_0, T_2) must resonate with MRR 2 and be off-resonance with MRR 1. To maximize the worst-case communication parallelism of these two paths, MRR 1 can be chosen with radius r_1 so that its resonant wavelengths λ_2, λ_5, λ_7, and λ_{10} can be used for the path (I_0, T_1). Meanwhile, MRR 2 is selected with radius r_2, so that path (I_0, T_2) can use its resonant wavelengths λ_1, λ_3, λ_6, and λ_9. Although λ_4 and λ_8 are also resonant wavelengths of MRR 2, they cannot be used for the path (I_0, T_2) because they partially resonate with MRR 1. In this method, the bit-level communication parallelism for the paths (I_0, T_1) and (I_0, T_2) is uniformly increased to 4.

However, previous works did not consider the actual application communication bandwidth demands between different nodes, which can vary significantly. If the communication parallelism of each transmission path is uniformly distributed, high-demand communications will require more transmission cycles to complete because they are not assigned higher parallelism. In contrast, low-demand communications can be completed much earlier within the same number of cycles, wasting wavelength resources. For example, Figure 2(b) shows the communication graph of an application where vertices N1, N2, and N3 represent the communication nodes within the application, and the edges represent their communication demands. The communication demand between (N1, N2) is much larger than the communication demand between (N1, N3). If N1 is mapped to port I_0, N2 is mapped to port T_1, and N3 is mapped to port T_2 in the topology in Figure 2(a)(2), based on the parallelism maximization method of previous works, the transmission between (N1, N2) will require 200/4 = 50 cycles to complete, while (N1, N3) will require only 10/4 = 2.5 cycles. The total transmission time is determined by the worst-case, 50 cycles. However, if we swap the radii of MRR 1 and MRR 2 so that MRR 1 uses radius r_2 and MRR 2 uses radius r_1 as shown in 2(b)(2), the path for high-demand communication can be allocated more wavelengths. Specifically, to met the wavelength usage conditions, λ_1, λ_3, λ_4, λ_6, λ_8, and λ_9 can be used in path (I_0, T_1), while λ_2 and λ_{10} can be used for (I_0, T_2). Then the path (I_0, T_1) will have a parallelism of 6, and (I_0, T_2) will have a parallelism of 2. With unchanged topology ports for the communication nodes, (N1, N2) will need $200/6 \approx 33.33$ cycles to complete the transmission, and (N1, N3) will need 10/2 = 5 cycles. Thus, the worst-case transmission cycles can be reduced from 50 to 33.33, which significantly shortens the communication time of the application. In other words, by allocating communication bandwidths according to the actual communication demands, network communications can be carried out more efficiently.

In this work, we address the issue of bandwidth allocation in WRONoC for the first time and propose an optimization strategy to determine the signal transmission paths and MRR radii in a given topology based on bandwidth demands to reduce application communication time. Our approach consists of two steps: First, starting from a fully connected WRONoC topology and based on the actual application communication graph with bandwidth demands, we propose an integer linear programming (ILP) model to map the communication nodes to the ports within the given topology. This model prioritizes mapping high-demand communications to transmission paths in a given topology containing fewer MRRs and consuming less power. After specifying the transmission paths to be used in the given topology, based on the bandwidth demands of these paths and the types of MRRs contained, we employ the multi-resonance properties of MRR and construct a progressive ILP model to optimize the MRR radii and allocate wavelengths to each path to minimize the worst-case data transmission cycles. Experimental results show that, compared to existing works that only consider communication parallelism, our strategy can reduce the number of transmission cycles by up to four times, significantly shortening the data transmission time of the application.

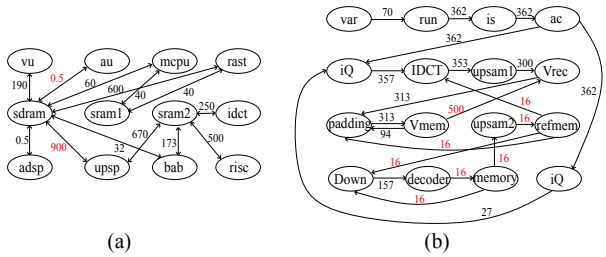

Figure 3: Two actual application communication graphs. The vertices represent the communication nodes. The directed edges represent the communications, and the weight next to each edge is the communication bandwidth demand. (a) An MPEG4 decoder application. (b) A video object plane decoder (VOP) application.

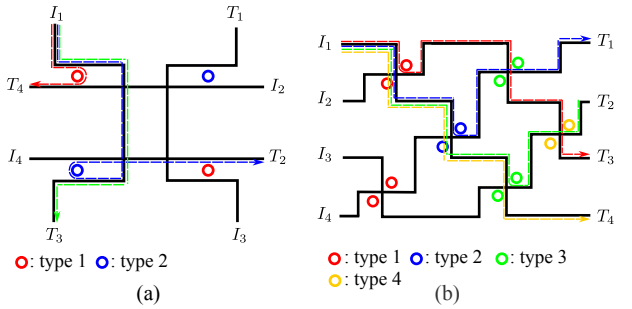

Figure 4: WRONoC topologies. The MRRs with different types are denoted with different colors. The dashed lines represent the signal paths between ports. (a) 4×3 Light. (b) 4×4 λ-Router.

2 BACKGROUND

2.1 Communication Graph

The communication between nodes within the application can be defined by the *communication graph*:

Definition 1. *The communication graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ is a weighted directed graph with each vertex $n_i \in \mathcal{N}$ representing a communication node and the weighted directed edge (n_i, n_j), denoted as $e_{(n_i,n_j)} \in \mathcal{E}$, indicating the communication from node n_i to node n_j, where n_i is referred as **"sender-node"** of n_j and n_j is referred as **"receiver-node"** of n_i. The weight of edge $e_{(n_i,n_j)}$ represents the bandwidth demand of the communication from n_i to n_j, denoted by $\delta_{(n_i,n_j)}$.*

Figure 3 shows the communication graphs of two actual applications. In the MPEG4 application [11], the highest communication demand reaches 910 bandwidth units, while the lowest is only 0.5 units. Similarly, in the VOP application [12], the highest demand can go up to 500 units, while the lowest is 16 units. The data reveals a significant imbalance in bandwidth demand among communications, with the highest demand communication paths often becoming bottlenecks that limit the total communication time of the applications, while other signal paths of lower demand communications remain idle most of the time. The current WRONoC design does not optimize the allocation of wavelength resources according to actual bandwidth demand, resulting in high-demand communications not obtaining sufficient wavelength usage to alleviate their bottleneck effect. To solve this problem, we propose to allocate wavelength usage according to actual communication needs, especially allocating more wavelengths to those communication paths with high bandwidth demands.

2.2 WRONoC Topology

A WRONoC topology consists of a set $\mathcal{S}$ of ports, a set $\mathcal{P}$ of signal paths, and a set $\mathcal{M}$ of MRR types. In WRONoC, the port initiating the signal is referred to as the *"initiator"*, while the signal destination port is referred to as *"target"*. To ensure that each communication node can be mapped to at least one port within the given topology, the size of set $\mathcal{S}$ must be greater than or equal to the size of the node set $\mathcal{N}$. Additionally, since the given communication graph does not necessarily have communications between every pair of nodes, we need to define a subset $\mathcal{PE} \subseteq \mathcal{P}$ to collect the mapped signal paths.

If a signal is designed to be on-resonance with the MRR in its path p, the MRR is referred to as the *"resonance MRR"* for that signal. The set $\mathcal{RM}_p$ denotes the resonance MRR types in the signal path p. Conversely, if a signal is designed to be off-resonance with the MRR in its path, the MRR is referred to as the *"off-resonance MRR"* for that signal. The set $\mathcal{OM}_p$ represents the off-resonance MRR types in the signal path p. The MRRs of the same type are set to have the same radius. As shown in Figure 4, two representative WRONoC topologies, Light [9] and $\lambda-$ router [7], with four ports, are introduced. We use $i \in \mathcal{S}$ to denote the port with $i \in \{1, 2, 3, 4\}$ and I_i represent the initiators, while T_i denotes the targets. The colored dashed lines represent the signal paths between ports, and the signals are designed to be on-resonance with the same-colored MRRs and off-resonance with MRRs of different colors. For instance, Figure 4 (a) shows a Light topology [9] with 4 ports. The set of signal paths of this topology is given as $\mathcal{P} = \{(I_1, T_2), (I_1, T_3), (I_1, T_4), \cdots, (I_4, T_2), (I_4, T_3)\}$. The set of resonance MRR types in path (I_1, T_2) can be denoted as $\mathcal{RM}_{(I_1,T_2)} = \{2\}$, while the set of off-resonance MRR types can be denoted as $\mathcal{OM}_{(I_1,T_2)} = \{1\}$. In contrast to Light, the λ-router in Figure 4 (b) requires more types of MRRs, which are relatively balanced in each signal path.

We use the transmission intensity formula from [6] and [13] to calculate the resonant wavelength of each MRR with a given radius. For a given MRR radius, denoted as r, and a given signal wavelength, denoted as λ, the transmission intensity of the signal at a resonance MRR with radius r, denoted as $T_d(r, \lambda) \in [0, 1]$ can be formulated as follows:

$$T_d(r, \lambda) = \frac{(1 - t_1^2)(1 - t_2^2)a}{1 - 2t_1 t_2 cos(\phi(r, \lambda))a + (t_1 t_2 a)^2} \tag{1}$$

where $t_1, t_2 \in [0, 1]$ denote the self-coupling coefficients, $a \in [0, 1]$ denote the single-pass amplitude transmission, and $\phi(r, \lambda)$ is the phase shift, which can be calculated as:

$$\phi(r, \lambda) = \beta(\lambda) 2\pi r. \tag{2}$$

$\beta(\lambda)$ is defined as the waveguide propagation constant function for a given λ and can be calculated as:

$$\beta(\lambda) = \frac{(2.57 - 0.85(\lambda \cdot 10^6 - 1.55))2\pi}{\lambda}. \tag{3}$$

We assume symmetric coupling in the transmission, i.e., $t_1 = t_2$, and ignore the propagation and bending losses occurring in the MRR, setting $a = 1$. Thus, Equation (1) can be written as:

$$\begin{aligned} T_d(r,\lambda) &= \frac{(1-t_1^2)^2}{1-2t_1^2 cos(\phi(r,\lambda)) + t_1^4} \\ &= \frac{1-2t_1^2+t_1^4}{1-2t_1^2 cos(\phi(r,\lambda)) + t_1^4} \end{aligned} \tag{4}$$

The cosine function, $\cos(\phi(r,\lambda))$, ranges from -1 to 1. From Equation (4), it can be observed that if $cos(\phi(r,\lambda)) = 1$, $T_d(r,\lambda)$ will reach its maximum value of 1. The wavelength corresponding to this maximum value is the resonant wavelength, which will be derived in Section 3.2.

As stated in [5], if the distance between the resonant wavelengths of the resonance MRR type and the off-resonance MRR type is less than the safety channel spacing, these wavelengths are considered overlapping and cannot be used for communication. In most WRONoC topologies, each signal path contains at most one resonance MRR type, and the upper limit of the communication parallelism for that path is determined by the number of resonant wavelengths of that resonance MRR type within the given wavelength range. Paths containing only off-resonance MRR types can use all wavelengths that do not conflict with other off-resonance MRR types, so their theoretically achievable communication parallelisms are much higher than paths restricted by the number of resonant wavelengths of resonance MRRs.

2.3 Transmission Cost

To achieve high parallelism in signal paths with high bandwidth demand, we should prioritize mapping high-demand communications to signal paths that only include off-resonance MRRs. If such ideal paths are unavailable, we should consider paths that include resonance MRRs but have a minimal number of off-resonance MRRs. This strategy helps reduce the overlapping of resonant wavelengths between MRRs, thereby enhancing communication parallelism.

The cost of communication is influenced by several factors, including the number of MRRs on the signal path and the insertion loss of the signal path. The insertion loss is 0.5 dB at a resonance MRR, 0.005 dB at an off-resonance, and 0.04 dB at a waveguide crossing [9]. We will also use these loss coefficients for this work. Based on these considerations, we should devise an optimization strategy that selects signal paths with fewer MRRs and lower MRR insertion losses for high-demand communications. This reduces communication costs and lays a foundation for minimizing subsequent transmission cycles.

2.4 Transmission Cycle

In [5], an ILP model has been proposed to achieve maximal communication parallelism of signal paths in WRONoC by optimizing MRR radii and wavelength usages. In this work, we adopt the constraints from this model to calculate the communication parallelism of each signal path p, denoted by $v_{para,p}$.

Since the communications will be mapped to signal paths in the topology, their bandwidth demands are also mapped to these paths. We denote the bandwidth demand of a signal path p as δ_p. Together with the communication parallelism, we can then model the transmission cycle of signal path p, denoted by $v_{c,p}$, as:

$$v_{c,p} = \frac{\delta_p}{v_{para,p}} \tag{5}$$

3 METHODOLOGY

In this section, we first develop an ILP model to map communication nodes to the ports within a given topology. This model aims to map communications with higher demands to signal paths in the topology that contain fewer MRRs and lower MRR insertion loss. Subsequently, we derive formulas for calculating resonant wavelengths to prepare sets of resonant wavelengths associated with the radius options. We then use the ILP model from [5] to calculate the communication parallelism of each signal path. Building on this, we propose new constraints based on bandwidth demands to calculate the transmission cycles for each signal path and minimize the cycle times in the worst-case scenarios. To accelerate the optimization process, we introduce an iterative algorithm that progressively reduces the problem space of the ILP model. This algorithm starts with the highest-demand signal path and gradually reduces the radius options for each MRR, thereby enabling faster searching of optimal MRR radii.

3.1 Communication Nodes Mapping

3.1.1 Input Preparation. For a given application graph, we define a set $\mathcal{N}$ to represent the nodes and a set $\mathcal{E}$ to represent the communications. The communication from node $n_s \in \mathcal{N}$ to $n_t \in \mathcal{N}$ is denoted by $(n_s, n_t) \in \mathcal{E}$. The set of receiver-nodes for node $n_s \in \mathcal{N}$ is denoted as $\mathcal{N}_{n_s}$. The bandwidth demand of (n_s, n_t) is denoted by $\delta_{(n_s,n_t)}$. For a given WRONoC topology, we introduce a set $\mathcal{S}$ to denote the ports, with each pair (s_i, s_j) representing the signal path from port $s_i \in \mathcal{S}$ to port $s_j \in \mathcal{S}$. The parameter $\epsilon_{(s_i,s_j)}$ represents the signal power loss in dB, while $\eta_{(s_i,s_j)}$ denotes the number of MRRs in the path (s_i, s_j).

3.1.2 ILP Model. For each communication node $n \in \mathcal{N}$ and each port option $s \in \mathcal{S}$, we define a binary variable b_s^n to indicate whether node n is mapped to port s. To ensure that each communication node is uniquely assigned to one WRONoC port, we implement the following constraint:

$$\forall n \in \mathcal{N} : \sum_{s \in \mathcal{S}} b_s^n = 1. \tag{6}$$

To ensure that each port is assigned to at most one communication node, we introduce the following constraint:

$$\forall s \in \mathcal{S} : \sum_{n \in \mathcal{N}} b_s^n \leq 1. \tag{7}$$

For each node $n \in \mathcal{N}$ and $u \in \mathcal{N}_n$, and for every two different ports $s, t \in \mathcal{S}$, we introduce a binary variable $b_{s,t}^{n,u}$. Communication between nodes n and u is considered to be transmitted through the signal path (s, t) if node n is mapped to port s and node u is

mapped to port t. This condition is represented by setting $b_{s,t}^{n,u} = 1$. Conversely, if either node n is not mapped to port s or node u is not mapped to port t, then $b_{s,t}^{n,u} = 0$, indicating that the communication does not occur on the path (s,t). This can be modeled with the following constraints:

$$\forall n \in \mathcal{N}\ \forall u \in \mathcal{N}_n\ \forall s \in \mathcal{S}\ \forall t \in \mathcal{S}:$$
$$b_s^n + b_t^u \leq 1 + b_{s,t}^{n,u} \cdot M \tag{8}$$
$$b_s^n + b_t^u \geq 2 + (b_{s,t}^{n,u} - 1) \cdot M \tag{9}$$

where M is an extremely large auxiliary constant. Thus, if $b_s^n + b_t^u = 2$, $b_{s,t}^{n,u}$ will be forced to take 1. Otherwise, $b_{s,t}^{n,u}$ will be set to 0.

Besides, we need to ensure that for each communication, there is exactly one signal path is assigned to it, which can be formulated as follows:

$$\forall n \in \mathcal{N}\ \forall u \in \mathcal{N}_n:\ \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}} b_{s,t}^{n,u} = 1. \tag{10}$$

Next, we use the continuous variable v_u^n to indicate the summation of the transmission cost and the bandwidth demand of each communication from node n to node u, which can be modeled as:

$$\forall n \in \mathcal{N}\ \forall u \in \mathcal{N}_n:$$
$$v_u^n = \alpha \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}} b_{s,t}^{n,u} \cdot \epsilon_{(s,t)} + \beta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}} b_{s,t}^{n,u} \cdot \eta_{(s,t)} + \delta_{(n,u)} \tag{11}$$

where α and β are coefficients that represent the weights for the signal power loss and the number of MRRs in the signal path (s,t), respectively. The term $\delta_{(n,u)}$ denotes the bandwidth demand of the communication between nodes n and u. To minimize v_u^n, communication with higher bandwidth demand will be forced to choose signal path with lower transmission cost.

Then, a new continuous variable v_w is introduced to calculate the maximal summation of the transmission cost and the bandwidth demand of all communications, which can be formulated as:

$$\forall n \in \mathcal{N}\ \forall u \in \mathcal{N}_n:\ v_w \geq v_u^n \tag{12}$$

Our goal is to minimize the maximal summation of the transmission cost and the bandwidth demand of all communications to enable high-demand communications to select paths with lower power loss and fewer MRRs. Thus, the optimization model can be formulated as follows:

$$\text{Minimize:} \quad v_w\,,$$
$$\text{Subject to:} \quad (6) - (12).$$

After the optimization, the case of $b_{s,t}^{n,u} = 1$ implies that node n is mapped to port s and node u is mapped to port t. Thus, the signal paths for minimizing transmission cycles are determined, which can be denoted by the set $\mathcal{PE}$. This also means that the bandwidth demands of the communication between n and u are mapped to each corresponding signal path of port s and t. We denote the bandwidth demand of signal path $(s,t) \in \mathcal{PE}$ as $\delta_{(s,t)}$ and use this parameter in the next sections for calculations of the transmission cycles.

3.2 Derivation of Resonant Wavelengths

As the input preparation for the transmission cycle minimization, we will first derive the resonant wavelengths for each radius option using Equations (2)-(4). As stated before, the resonant wavelength corresponds to the condition where the cosine of the phase,$\cos(\phi(r,\lambda))$, equals 1.

Given MRR radius r, the resonant wavelength λ_{res} can be derived as follows:

$$\begin{aligned}
&\cos(\phi(r,\lambda_{\text{res}})) = 1\\
&\Rightarrow \cos(\beta(\lambda_{\text{res}})2\pi r) = 1\\
&\Rightarrow \cos((\frac{(2.57 - 0.85(\lambda_{\text{res}} \cdot 10^6 - 1.55))2\pi}{\lambda_{\text{res}}})2\pi r) = 1\\
&\Rightarrow \frac{(2.57 - 0.85(\lambda_{\text{res}} \cdot 10^6 - 1.55)) \cdot 2\pi \cdot 2\pi r}{\lambda_{\text{res}}} = 2l\pi, l \in \mathbb{Z}\\
&\Rightarrow \lambda_{\text{res}} \cdot 2l\pi = (2.57 - 0.85(\lambda_{\text{res}} \cdot 10^6 - 1.55)) \cdot 2\pi \cdot 2\pi r\\
&\Rightarrow \lambda_{\text{res}}(4\pi^2 r \cdot 0.85 \cdot 10^6 + 2l\pi) = 4\pi^2 r \cdot (2.57 + 0.85 \cdot 1.55)\\
&\Rightarrow \lambda_{\text{res}} = \frac{4\pi^2 r(2.57 + 0.85 \cdot 1.55)}{4\pi^2 r \cdot 0.85 \cdot 10^6 + 2l\pi}\\
&\Rightarrow \lambda_{\text{res}} = \frac{7.775\pi r}{l + 1.7 \cdot 10^6 \pi r}
\end{aligned} \tag{13}$$

Thus, given $l \in \mathbb{Z}$, wavelength lower bound λ_{low} and upper bound λ_{upper}, the resonant wavelengths within the given wavelength range $[\lambda_{\text{low}}, \lambda_{\text{upper}}]$ can be calculated as the input options for the transmission cycle optimization.

3.3 Iterative Transmission Cycle Minimization

Section 2.4 introduces the calculation of transmission cycles for each signal path, which is determined by dividing the bandwidth demand by the communication parallelism. As stated in Section 2.2, paths containing only off-resonance MRR types have theoretically higher communication parallelism and thus have negligible impact on the worst-case transmission cycle. Thus, we do not calculate the transmission cycles for such paths in this work and remove these paths from $\mathcal{PE}$ resulting from Section 3.1. Additionally, any MRR type not used as a resonance MRR in any path will be removed from the topology. We adopt the method from [5] to establish the communication parallelism for each signal path, represented by the variable $v_{para,p}$ for $p \in \mathcal{PE}$.

However, in ILP models, the variable $v_{para,p}$ cannot appear in the denominator because this will make the equations non-linear. Thus, we transform the original equation into its reciprocal form to address this issue. By maximizing the reciprocal of the transmission cycle (i.e., the ratio of communication parallelism to bandwidth demand), we achieve the equivalent of minimizing the transmission cycles themselves. Thus, we use variable $v_{rc,p}$ to denote the reciprocal of the transmission cycle for signal path p, which can be calculated as:

$$\forall p \in \mathcal{PE}: v_{rc,p} = \frac{v_{para,p}}{\delta_p} \tag{14}$$

where δ_p is the bandwidth demand of signal path p calculated from Section 3.1.

Since the worst-case transmission cycle of all signal paths is the bottleneck of communication, we need to find the one with the most transmission cycles, equivalent to the path with the minimal reciprocal transmission cycles. We denote the minimal reciprocal transmission cycle by a continuous variable v_{wrc} and model it with

the following constraint:

$$\forall p \in \mathcal{PE} : \ v_{wrc} \leq v_{rc,p} \tag{15}$$

Thus, our target is to maximize the worst-case reciprocal transmission cycle v_{wrc} to minimize the worst-case transmission cycle of all signal paths. The objective can be formulated as follows:

$$\begin{aligned} &\text{Maximize:} \quad v_{wrc} \ , \\ &\text{Subject to:} \quad (14) - (15). \end{aligned}$$

Now, we have established the ILP model for minimizing the transmission cycle, which we will refer to as the "ILP_Model_Cycle". In [5] has shown that although the ILP model for maximizing the communication parallelism has proven optimality, the solution speed significantly decreases as the topology scales up and the number of input signal paths increases. To accelerate the optimization process for transmission cycles, we run the modeling and solving process using several iterations, each of which uses the result of the last one as input. In particular, this iterative algorithm sorts at the beginning of the signal paths by their demands. It conducts a series of primary iterations corresponding to the number of signal paths. In the i-th primary iteration, the algorithm optimizes all the MRR radii within the signal paths ranked among the top i in demands. Each iteration uses the optimized MRR radii set from the previous iteration as input options. Additional secondary iterations are conducted within each primary iteration to optimize and collect the MRR radii solutions for each path involved. In each iteration, we attempt to maximize the reciprocal of the transmission cycle and repeat this process, progressively storing all MRR solutions generated under the current number of paths until all signal paths have been iteratively solved. The pseudo-code for iterative minimizing the transmission cycles using the ILP model is shown in Algorithm 1. The loop in lines 6-30 creates the ILP model, and the number of given input signal paths is the count of secondary iterations. Inside each of these iterations, the input options for MRR radii gradually decrease. These iterative improvement steps are used to overcome the limitations of the ILP model.

Algorithm 1 Progressively minimize transmission cycles using an ILP model with iterations

1: **Input:**
$\mathcal{PE}$: The set of signal paths with bandwidth demands and MRR types;
$\mathcal{R}$: The radius options for MRRs;
Ω_r: The resonant wavelength sets of MRR with radius r;
$\mathcal{M}$: The types of MRRs;
$\mathcal{M}_p$: The types of MRRs being used in path p;
2: **Output:**
The radius of each MRR type;
The wavelength usage for each signal path;
The transmission cycle for each signal path;
3: **Variables:**
$\mathcal{PE}_i$: The set of signal paths with the top i ranking in demand;
$\mathcal{M}_{\mathcal{PE}_i}$: The set of MRR types in $\mathcal{PE}_i$;
$\mathcal{RM}_{\mathcal{PE}_i}$: The set of resonance MRR types in $\mathcal{PE}_i$;
optimized_mrr[m]: The set of optimized radius solutions for MRR type m of all iterations;
optimized_mrr_iteration[i][m]: The set of optimized radius solutions for MRR type m in iteration i;
4: Rank all signal path $p \in \mathcal{PE}$ by bandwidth demands in descending order;
5: Initialize *optimized_mrr[m]* as empty radius set of type m;
6: **for** $i = 1$ to $|\mathcal{PE}|$ **do**
7: Insert the signal paths ranked among the top i in bandwidth demands in $\mathcal{PE}_i$;
8: Insert the MRR types occurs in $\mathcal{PE}_i$ in $\mathcal{M}_{\mathcal{PE}_i}$;
9: Insert the resonance MRR in $\mathcal{RM}_{\mathcal{PE}_i}$;
10: Initialize *optimized_mrr_iteration[i][m]* as empty radius set of type m;
11: **for** $j = 1$ to $|\mathcal{PE}| - i + 1$ **do**
12: **for** $m \in \mathcal{M}_{\mathcal{PE}_i}$ **do**
13: **if** $m \in \mathcal{M}_{\mathcal{PE}_{i-1}}$ **then**
14: **if** $m \in \mathcal{RM}_{\mathcal{PE}_i}$ **then**
15: Use *optimized_mrr[m]* \ *optimized_mrr_iteration[i][m]* as the input radius options;
16: **else**
17: Use *optimized_mrr[m]* as the input radius options;
18: **end if**
19: **else**
20: **if** $m \in \mathcal{RM}_{\mathcal{PE}_i}$ **then**
21: Use $\mathcal{R}$ \ *optimized_mrr_iteration[i][m]* as the input radius options;
22: **else**
23: Use $\mathcal{R}$ as the input radius options;
24: **end if**
25: **end if**
26: **end for**
27: Solve ILP_Model_Cycle;
28: Insert solutions for MRR type m in *optimized_mrr[m]* and *optimized_mrr_iteration[i][m]*;
29: **end for**
30: **end for**
31: **return** Output;

4 OPTIMIZATION RESULTS

4.1 Input Settings

We implement our approach, referred to as *"bandwidth allocation method"* in C++ and use an optimization solver named Gurobi [14] to solve the integer linear programming (ILP) models. In Algorithm 1, we set the solution pool size in Gurobi to 10. Our work addresses the communication graphs of five applications: Filter (5 nodes) [15], DSP (5 nodes) [16], MPEG4 decoder (12 nodes) [11], VOPD (16 nodes) [12], MWD (12 nodes) [17]. For transmission cost, we set α = 100 and β = 100 in Constraint (11). We apply two full-connectivity WRONoC topologies: Light [9] and λ-router [7] with 8 and 16 ports as our input topologies. The radius options for MRRs are the same as in [5], ranging from [5 μm, 30 μm] with increments of 0.25μm, in total 101 radius options. The wavelength range is given as [1500 nm, 1600 nm]. The safety channel spacing for two resonant wavelengths is set as 0.8 nm [5]. This spacing means that if the distance between two resonant wavelengths is less than 0.8 nm, they are considered to be in an overlapped state. Additionally, we implement the method from [5], referred to as the *"parallelism maximization method"*, with the same input options as a baseline for comparison and set the Gurobi search time for this method to 20 hours. We run both approaches on an Apple M1 Pro 10-core CPU. The results are shown in Table 1 and Figure 5.

Table 1: Experiment results

App.	$\|\mathcal{N}\|$	$\|\mathcal{S}\|$	$\|\mathcal{M}\|$	$\|\mathcal{PE}\|$	hd.	ld.	T	PM.	BA.	R.	time
Filter	5	8	8	7	200	100	Light	14.29	6.67	2.144	105s
							λ-R.	14.29	6.45	2.21	49s
DSP	5	8	8	7	600	200	Light	40	19.35	2.07	33s
							λ-R.	42.86	20	2.14	151s
MWD	12	16	16	13	128	64	Light	18.29	32	0.57	3753s
							λ-R.	25.6	32	0.8	42298s
VOPD	16	16	16	21	500	16	Light	125	40.22	3.11	7936s
MPEG4	12	16	16	26	910	0.5	Light	227.5	43.33	5.25	42354s

App.: the application communication graph;
$|\mathcal{N}|$: the number of the communication nodes;
$|\mathcal{S}|$: the number of the topology ports;
$|\mathcal{M}|$: the number of MRR types;
$|\mathcal{PE}|$: the number of communications;
hd.: the highest bandwidth demand of all communications;
ld.: the lowest bandwidth demand of all communications;
T: the topology type;
λ-R.: the $\lambda-$ router;
PM.: the worst-case transmission cycle results from the parallelism maximization method;
BA.: the worst-case transmission cycle results from our bandwidth allocation method;
R.: the ratio of PM results to BA results;
time: the optimization time.

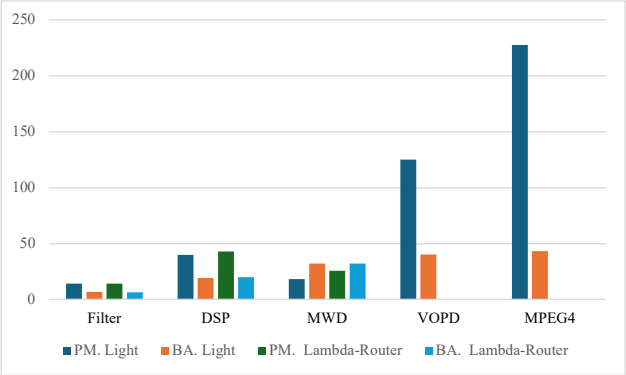

Figure 5: The comparison results of the worst-case transmission cycle between parallelism maximization method and our approach in Light and λ-Router (Lambda-Router).

4.2 Results Analysis

- The experimental results clearly show that, except for the case of MWD, the worst-case transmission cycle of our method is lower than that achieved by the parallelism maximization approach. In the case of MWD, the communications are evenly distributed across multiple nodes, and the demand differences are relatively small. This results in the early selected solution being unable to fully support the parallelism of higher-demand but lower-ranked communications, resulting in longer transmission cycles. The results of the other four cases indicate that our bandwidth allocation method can effectively accelerate application communication and significantly alleviate the bottleneck effect in high-demand communications.
- In the larger application cases, VOPD and MPEG4, both methods cannot obtain a feasible solution within 20 hours for $\lambda-$ router. Compared with λ-router, Light topology has a higher efficiency and better bandwidth allocation optimization effect. This is reflected in table columns "R." where Light has a higher ratio. This ratio is calculated by dividing the worst-case transmission cycle of the parallelism maximization method by the worst-case transmission cycle for the bandwidth allocation. This difference is due to the imbalanced distribution of the number of MRR types between signal paths in the Light topology, resulting in unbalanced transmission costs on different signal paths. This means that paths with high bandwidth demands can be prioritized over signal paths with lower transmission costs, allowing for greater parallelism. In contrast, the λ-router has a uniform MRR type distribution over all signal paths, making it difficult to obtain a feasible solution for different demands in large cases, and so does not show as obvious an improvement as Light. Therefore, a topology with an uneven distribution of MRR types, such as Light, will provide faster communication for applications with unbalanced requirements.
- As shown in Figure 5, using the same topology, the greater the difference between the highest and lowest demands in the same application, the more significant the bandwidth allocation method improvement will be. For example, the MPEG4 application's highest communication demand is 910, while the lowest demand is only 0.5. Compared with the parallelism maximization method, our approach can reduce the worst-case transmission cycle by up to more than five times, greatly improving transmission efficiency and proving the practical applicability of our method.
- The optimization time increases as the number of MRR types and signal paths grows. Our work demonstrates a significant improvement over the parallelism maximization method, which mainly addresses the cases where port size is less or equal to 8. Compared to this, we can solve problems using our method involving 16 nodes in shorter times and find feasible, high-quality solutions more efficiently.

5 CONCLUSION

In this paper, we propose an optimization method that allocates wavelength usages based on data communication demands, aiming to minimize the worst-case communication transmission cycles in applications. Specifically, we propose a mapping strategy of nodes within an application to a given topology, establishing signal paths. We then leverage the multi-resonance properties of MRRs to optimize the selection of MRR radii and allocate wavelength usages to signal paths based on their bandwidth demands, thereby reducing transmission cycles. Compared to the parallelism maximization method that does not consider bandwidth demands, our approach can significantly reduce the worst-case data transmission cycle in large-scale topologies, achieving reductions of up to five times or more. In conclusion, our method can effectively allocate the wavelength usages and significantly shorten application data transmission time.

ACKNOWLEDGMENTS

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 496766278.

REFERENCES

[1] Aditya Narayan, Ajay Joshi, and Ayse K. Coskun. Bandwidth allocation in silicon-photonic networks using application instrumentation. In *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–2, 2020.

[2] Lieven Vandersypen and Antoni van Leeuwenhoek. 1.4 quantum computing - the next challenge in circuit and system design. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 24–29, 2017.

[3] J. Luo, A. Elantably, V. D. Pham, C. Killian, D. Chillet, S. Le Beux, O. Sentieys, and I. O'Connor. Performance and energy aware wavelength allocation on ring-based wdm 3d optical noc. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, pages 1372–1377, 2017.

[4] Tsun-Ming Tseng, Alexandre Truppel, Mengchu Li, Mahdi Nikdast, and Ulf Schlichtmann. Wavelength-routed optical nocs: Design and eda — state of the art and future directions: Invited paper. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–6, 2019.

[5] Mengchu Li, Tsun-Ming Tseng, Mahdi Tala, and Ulf Schlichtmann. Maximizing the communication parallelism for wavelength-routed optical networks-on-chips. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 109–114, 2020.

[6] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets. Silicon microring resonators. *Laser & Photonics Reviews*, 6(1):47–73, 2012.

[7] M. Briere, B. Girodias, Y. Bouchebaba, G. Nicolescu, F. Mieyeville, F. Gaffiot, and I. O'Connor. System level assessment of an optical noc in an mpsoc platform. In *2007 Design, Automation Test in Europe Conference Exhibition*, pages 1–6, 2007.

[8] Mengchu Li, Tsun-Ming Tseng, Davide Bertozzi, Mahdi Tala, and Ulf Schlichtmann. Customtopo: A topology generation method for application-specific wavelength-routed optical nocs. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8, 2018.

[9] Zhidan Zheng, Mengchu Li, Tsun-Ming Tseng, and Ulf Schlichtmann. Light: A scalable and efficient wavelength-routed optical networks-on-chip topology. In *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 568–573, 2021.

[10] Andrea Peano, Luca Ramini, Marco Gavanelli, Maddalena Nonato, and Davide Bertozzi. Design technology for fault-free and maximally-parallel wavelength-routed optical networks-on-chip. In *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8, 2016.

[11] L. Benini. Network-on-chip architectures and design methods. *IEE Proceedings - Computers and Digital Techniques*, 152:261–272(11), March 2005.

[12] S. Murali and G. De Micheli. Bandwidth-constrained mapping of cores onto noc architectures. In *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, volume 2, pages 896–901 Vol.2, 2004.

[13] Lukas Chrostowski and Michael Hochberg. *Silicon Photonics Design: From Devices to Systems*. Cambridge University Press, 2015.

[14] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.

[15] Srinivasan Murali, Paolo Meloni, Federico Angiolini, David Atienza, Salvatore Carta, Luca Benini, Giovanni De Micheli, and Luigi Raffo. Designing application-specific networks on chips with floorplan information. In *2006 IEEE/ACM International Conference on Computer Aided Design*, pages 355–362, 2006.

[16] S. Murali and G. De Micheli. Bandwidth-constrained mapping of cores onto noc architectures. In *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, volume 2, pages 896–901 Vol.2, 2004.

[17] A. Jalabert, S. Murali, L. Benini, and G. De Micheli. /spl times/pipescompiler: a tool for instantiating application specific networks on chip. In *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, volume 2, pages 884–889 Vol.2, 2004.