PSION: Combining Logical Topology and Physical Layout Optimization for Wavelength-Routed ONoCs

Alexandre Truppel
Faculdade de Engenharia,
Universidade do Porto
Porto, Portugal
alex.truppel@fe.up.pt

Tsun-Ming Tseng
Chair of Electronic Design
Automation, TUM
München, Germany
tsun-ming.tseng@tum.de

Davide Bertozzi
University of Ferrara
Ferrara, Italy
davide.bertozzi@unife.it

José Carlos Alves
Faculdade de Engenharia,
Universidade do Porto
Porto, Portugal
jca@fe.up.pt

Ulf Schlichtmann
Chair of Electronic Design
Automation, TUM
München, Germany
ulf.schlichtmann@tum.de

ABSTRACT

Optical Networks-on-Chip (ONoCs) are a promising solution for high-performance multi-core integration with better latency and bandwidth than traditional Electrical NoCs. Wavelength-routed ONoCs (WRONoCs) offer yet additional performance guarantees. However, WRONoC design presents new EDA challenges which have not yet been fully addressed. So far, most topology analysis is abstract, i.e., overlooks layout concerns, while for layout the tools available perform Place & Route (P&R) but no topology optimization. Thus, a need arises for a novel optimization method combining both aspects of WRONoC design. In this paper such a method, PSION, is laid out. When compared to the state-of-the-art design procedure, results show a 1.8× reduction in maximum optical insertion loss.

CCS CONCEPTS

• **Theory of computation → Integer programming**; • **Hardware → Emerging optical and photonic technologies**.

KEYWORDS

optical networks-on-chip; silicon photonics; physical layout; design optimization; placement & routing

ACM Reference Format:
Alexandre Truppel, Tsun-Ming Tseng, Davide Bertozzi, José Carlos Alves, and Ulf Schlichtmann. 2019. PSION: Combining Logical Topology and Physical Layout Optimization for Wavelength-Routed ONoCs. In *2019 International Symposium on Physical Design (ISPD '19), April 14–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3299902.3309747

ISPD 2019, DOI: 10.1145/3299902.3309747
http://dl.acm.org/citation.cfm?id=3309747

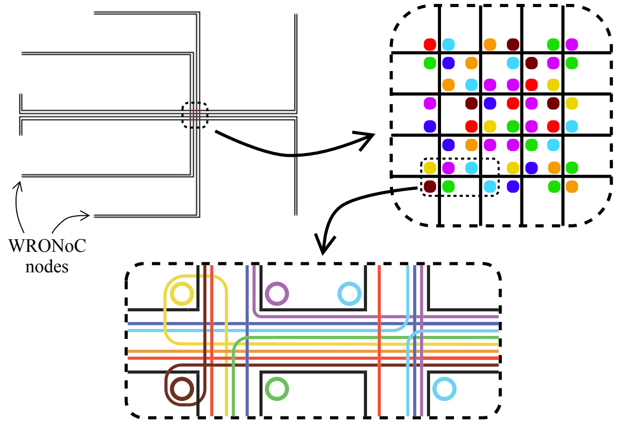

Figure 1: Final design of a WRONoC router for 8 nodes given by PSION. A portion of some message paths is shown (color indicates wavelength).

1 INTRODUCTION

Optical Networks-on-Chip (ONoCs) have been proposed as a solution for the ever-increasing integration requirements of large System-on-Chip designs. Compared to traditional Electrical Networks-on-Chip, ONoCs present not only lower dynamic power consumption but also extremely low signal delay and higher bandwidth [9].

The use of light as opposed to electrical signals to send information between network nodes requires the following four main components on the optical routing plane: 1) *modulators* to convert electrical signals into optical signals at every node (electrical-optical interface) of the optical network, 2) *demodulators* to do the opposite, 3) *waveguides* acting as optical wires and 4) *optical routing elements* to transfer optical signals between waveguides [7].

ONoCs can be organized into two main categories: **1)** *active networks* [3, 12, 17] and **2)** *passive networks.* Active networks require a control layer for routing. Passive networks use routing elements which resonate with different frequencies such that a message is passively routed according to the wavelength of the carrier light. Hence, a message's path is completely defined, at design time, by its origin

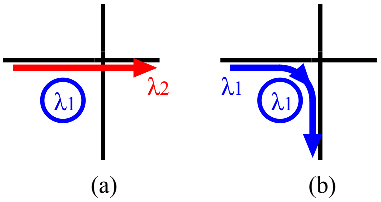

Figure 2: Wavelength routing using an MRR. (a) The light signal is not routed because it has a different wavelength than the MRR. (b) The light signal is routed through the MRR to another waveguide.

and wavelength alone (Figure 1 shows an example of wavelength routing). Thus, passive ONoCs are also termed Wavelength-Routed ONoCs (WRONoCs) [8]. This eliminates network delay resulting from path setup and dynamic power consumption required for the extra control layer.

Multiple light sources of different wavelengths can be used to transmit separate information streams on the same waveguide without interference (wavelength-division multiplexing). This enables conflict-free communications with increased bandwidth. The only requirement is to make sure at design time that no two messages with the same wavelength are allowed to share the same waveguides.

The optical switching element in ONoCs is the Micro-Ring Resonator (MRR). It has a circular silicon structure whose radius defines the resonance frequency. A light signal with a certain wavelength propagating on a waveguide close to an MRR with a matching resonance frequency will be coupled to the MRR and moved onto another waveguide also close to that MRR [10]. Figure 2 shows an example of this behaviour.

The design of a WRONoC router is an optimization process with *two aspects* to consider: the logical topology and the physical layout of the router. The former assigns a wavelength to each message and each MRR and also connects the nodes through waveguides and MRRs such that the communication matrix, which specifies the communication requirements between nodes, is fulfilled. The latter optimally places and routes those elements on the optical plane while considering the physical positions of the nodes and constraints related to the physical placement of the waveguides.

So far both aspects have only been considered separately or with restrictions. Various works have presented specific topologies with few concerns about their layout [7, 13, 14]. Ramini et al. [11] present a topology designed in tandem with placement constraints, yet it results from a manual optimization effort for one set of node positions. Ortín-Obón et al. [9] take into consideration physical constraints, but analyze only the ring topology. Few attempt to optimize for non-complete communication matrices [1, 5]. P&R tools to optimize the second aspect have been developed [2, 15, 16], but all take a topology as input, forcing the designer to choose the topology beforehand.

However, neither aspect can be considered in isolation, as each influences the other [11, 13, 15]. During generation of the logical topology we are unable to accurately predict important physical characteristics, e.g. the number of waveguide crossings, of the final design after P&R. Furthermore, during P&R, if the logical topology has already been chosen and fixed, any subsequent optimization is being done only around a local minimum of the solution space.

Ideally, a design tool would take as inputs the communication matrix and the physical positions of the nodes and, by working on both aspects simultaneously, produce a fully-optimized fully-custom logical topology and matching physical layout [13]. In reality, the problem space of such an optimization is discouragingly vast for any but the simplest cases. Thus, in this paper we propose and solve a constrained version of the complete problem. In this version – PSION – a physical layout template is also given as an input to the optimization. The template mainly consists of MRR placeholders and waveguides already placed and routed on the optical plane, and connects all nodes.

We define the optimization problem in Section 2. Physical layout templates are described in Section 3 and the Mixed Integer Programming (MIP) model used to optimize them is presented in Section 4. Section 5 explains a fast technique to verify the model's feasibility and Section 6 then proposes a 3-step algorithm to efficiently solve it. Finally, Section 7 reveals three layout templates and tests them against the state-of-the-art P&R PROTON+ [15] and PlanarONoC [2] tools.

2 WRONOC DESIGN PROBLEM

We formally define the optimization problem for the design of WRONoC routers as follows:

Input data:

- Communication matrix: a square binary matrix $CM_{i,j} \in \mathbb{R}^{N \times N}$ with N equal to the number of nodes and where $CM_{i,j} = 1$ if node i sends a message to node j.
- Physical positions of the modulators and demodulators of each node on the optical plane.
- Technology parameters: power loss values.

Output data:

- Wavelength of each message and MRR.
- Placement of each MRR.
- Routing of each waveguide.

Minimization objectives. Their choice depends on the technology and the needs of the design. We consider **1)** number of wavelengths, **2)** message insertion loss and **3)** number of MRRs, as in previous publications [9–11, 13–15]. With PSION, the weighting coefficient for each objective can be freely adjusted to meet different designer demands.

Message insertion loss is the sum of seven types of losses: **1)** crossing loss, **2)** drop loss, **3)** through loss, **4)** bending loss, **5)** propagation loss, **6)** modulator loss and **7)** demodulator loss [6, 15]. We consider all except the last two, which are constant and equal for all messages and thus can be ignored from an optimization perspective.

3 PHYSICAL LAYOUT TEMPLATE

We consider a constrained version of the complete problem, where an extra input is required. This input, called a **physical layout *template***, consists of a collection of WRONoC router elements

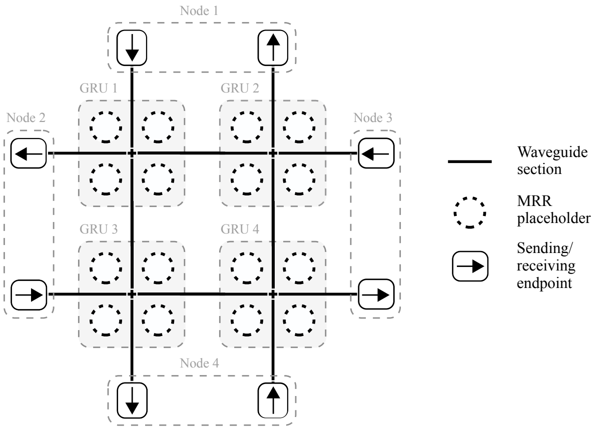

Figure 3: Generalizing the 4x4 GWOR topology [14] using endpoints, GRUs and waveguide sections.

(modulators, demodulators, waveguides and MRR placeholders) already placed and routed on the optical plane.

The role of the solver with this new input is to optimally route the messages defined in the communication matrix through the template and to activate the necessary routing features for the chosen paths.

This way we significantly reduce the complexity of the complete problem while still improving upon the state-of-the-art solutions. Nevertheless, this template does not need to be intricate or sophisticated. In fact, the intuitive knowledge of the designer about the structure of the router to be created is more than enough to provide a good template.

3.1 Template elements

We model layout templates with three layout elements. Together they allow for the design of any WRONoC topology (an example is shown in Figure 3).

Endpoints represent modulators and demodulators. They are placed wherever the (de)modulators for each node are and connect to one waveguide section.

General Routing Units (GRUs) are elements that connect to multiple waveguide sections (the *edges* of the GRU) and contain MRR placeholders, to be populated by the solver as needed. They are the routing building blocks of the template and are described further in the next section.

Waveguide sections connect two GRUs or a GRU and an endpoint. Each section has two associated parameters: *length* and *extraloss*. The latter is used to describe sections with other constant sources of insertion loss besides length, such as sections with 90° bends.

Our method can solve for any template, i.e., any arrangement of endpoints, GRUs and waveguide sections.

3.2 General Routing Unit

Photonic Switching Elements (PSEs) are commonly applied in WRONoC routers [7, 11, 13, 14]. For PSEs, MRR locations and wavelengths are explicitly specified and the waveguide structure is fixed.

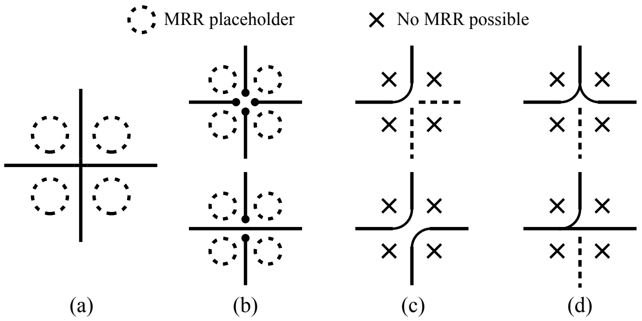

Figure 4: Internal structure of a GRU. (a) 4 MRR placeholders and a crossing. (b) Avoiding the crossing, when possible (c) Valid corner bending states. (d) Invalid corner bending states.

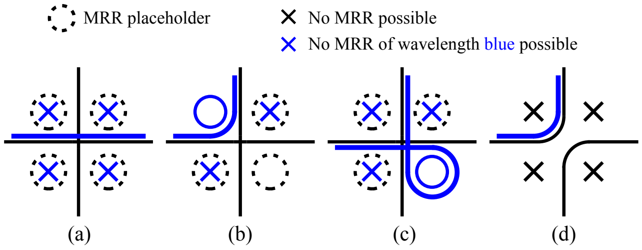

Figure 5: Routing possibilities through a GRU. (a) Direct path. (b)(c) Routing through an MRR. (d) Routing through a bend.

GRUs are the routing building blocks for the proposed layout template and, in contrast to PSEs, GRUs are not inherently constrained to a specific internal structure. Instead, only MRR placeholders are predefined in a GRU. Thus, different MRR placement and wavelength configurations can happen for each GRU, as well as different edge connection arrangements. This provides more flexibility in the resulting WRONoC design.

3.2.1 *Structure.*

Figure 4(a) shows the structure of a GRU: the four waveguide sections form a crossing where any of the four corners on that crossing can have an MRR. Sometimes the crossing can be avoided, leading to the variations in Figure 4(b).

We also consider an additional structure variation called *corner bending*. When active, the GRU contains no MRRs and some corners may be replaced by a bend between the two edges in that corner, as in Figure 4(c).

Note that two corners connected to the same edge of a GRU *cannot* be both bent. Therefore, if two edges are connected through a corner bend, the other two edges must be bent through the opposite corner if they have messages going through. Figure 4(d) shows two invalid configurations.

This extra variation proves useful for sparser templates (low ratio of the number of messages to the number of MRR positions), or in cases where multiple messages must be routed through the same corner.

3.2.2 Routing. Figure 5 shows the routing possibilities through a GRU. If no MRRs of the same wavelength as the message are present and corner bending is not activated, the message will have no direction change, as shown in Figure 5(a).

For wavelength routing, the message can be routed through an MRR with the same wavelength in the closest corner, as shown in Figure 5(b), or in the opposite corner, as shown in Figure 5(c).

With corner bending, since the two waveguides become connected, all messages in any of the two waveguides are routed through that corner, regardless of wavelength, as shown in Figure 5(d).

A message's path through a GRU is always independent of its direction, i.e., all routing features are bidirectional. Also, the four MRRs on a GRU can have different wavelengths (examples are shown in Figure 1). This allows for intricate multi-message routing capabilities per waveguide crossing which have not yet been optimized to their full potential.

3.3 Communication Matrix

Given a layout template, the communication matrix can be translated into a set of messages (one for each nonzero entry), where each message is associated with two endpoints on that template: the sender and the receiver.

4 MATHEMATICAL MODEL

We solve the constrained problem using a Mixed Integer Programming model. Advantages of MIP models include:

(1) A MIP model can give optimal solutions, or at least an upper/lower bound to the optimal value of the optimization function.
(2) The same MIP can be used to optimize different objectives, therefore giving the designer more flexibility.
(3) MIP models are flexible, so new GRU designs, routing features or other modifications can easily be added.

The model constants and indices are outlined in Table 1. Constants L_{wg}, L^E_{wg} and indices W^*_i collectively describe the physical layout template and indices E^*_m define the communication matrix. Table 2 lists all model variables.

We now specify the constraints and the optimization function (note that similar constraints for multiple directions or corners are omitted). Finally, we present some model reduction techniques.

4.1 Constraints

Message routing. A path with the correct beginning and end must be guaranteed for each message. For that we apply the following three sets of constraints:

(1) A message must be on the waveguide of the endpoints it is sent from and received by.

$$mwg_{m,W^E_{E^S_m}} = 1 \quad mwg_{m,W^E_{E^R_m}} = 1 \qquad \forall m = 1...N_m$$

(2) If an endpoint does *not* send or receive a given message, that message *cannot* be present on its waveguide section.

$$mwg_{m,W^E_{ep}} = 0 \quad \forall ep = 1...N_{ep} \setminus \{E^S_m, E^R_m\}$$
$$\forall m = 1...N_m$$

Table 1: Model constants & indices

Constants	
N_{gru}, N_{wg}, N_m, N_{ep}, N_λ	Total number of GRUs, waveguide sections, messages, endpoints and wavelengths
L^P, L^C, L^B, L^D, L^T	Values for propagation, crossing, bending, drop and through loss
L_{wg}, L^E_{wg}	Length and extra loss of waveguide section *wg*
Indices	
W^T_g, W^B_g, W^L_g, W^R_g	Waveguide section connected to GRU *g* to the top, bottom, left and right
W^E_{ep}	Waveguide section connected to endpoint *ep*
E^S_m, E^R_m	Sending and receiving endpoints for message *m*

(3) A message is exactly on 0 or 2 edges of a GRU.

$$mwg_{m,W^T_g} + mwg_{m,W^R_g} + mwg_{m,W^B_g} + mwg_{m,W^L_g} \in \{0,2\}$$
$$\forall m = 1...N_m, g = 1...N_{gru}$$

It is possible for a message to be on all four edges of a GRU, but this was neglected because it appearing on an optimized solution is highly unlikely, and not including it simplifies the model and the problem space. The reason is that a message routing through all 4 edges (enter through edge 1, leave through 2, enter through 3, leave through 4) can also route through 2 edges (enter through 1, leave through 4) with half the loss on that GRU and a shorter path.

Wavelength exclusion. Each waveguide section has at most one message going through it for each wavelength. First, each message must use exactly one wavelength:

$$\sum_{\lambda=1}^{N_\lambda} mwl_{m,\lambda} = 1 \qquad \forall m = 1...N_m$$

Then the value of mwe_{m_1,m_2} is set accordingly:

$$mwl_{m_1,\lambda} \wedge mwl_{m_2,\lambda} \Rightarrow mwe_{m_1,m_2}$$
$$\forall \lambda = 1...N_\lambda$$
$$\forall m_1, m_2 = 1...N_m : m_2 \neq m_1$$

Now enforce exclusivity of wavelengths on all waveguides:

$$mwe_{m_1,m_2} \Rightarrow (mwg_{m_1,wg} + mwg_{m_2,wg} \leqslant 1)$$
$$\forall m_1, m_2 = 1...N_m : m_1 \neq m_2$$
$$\forall wg = 1...N_{wg}$$

Activation of routing features. A path is chosen for each message but, to make that path take effect, constraints are needed to enforce the activation of the routing features responsible for it.

If a message takes the direct path through a GRU, no features need to be turned on. However, if a message is present on adjacent edges of a GRU, then one of the three options from Figure 5(b-d)

Table 2: Model variables

Binary	
$cb_{g,p}$	Corner p on GRU g is bent
wlu_{λ}	At least one message uses wavelength λ
$mwl_{m,\lambda}$	Message m uses wavelength λ
mwe_{m_1,m_2}	Messages m_1 and m_2 use the same wavelength
$mwg_{m,wg}$	Message m goes through waveguide section wg
$cl_{g,m}$, $bl_{g,m}$	Message m has crossing/bending loss on GRU g
$tl_{g,p,m}$	Message m has through loss due to MRR p in GRU g
$rum_{g,p,m}$	MRR on GRU g, corner p, used by message m
$ru_{g,p}$	MRR on GRU g, corner p, used by a message
mch_g, mcv_g	GRU g has at least one message going through the center crossing horizontally/vertically
Integer	
nwl	Number of used wavelengths
Continuous	
mil_m	Insertion loss for message m
$maxil$	Maximum insertion loss over all messages

Index $p \in \mathbb{P}$, $\mathbb{P} = \{TL$: Top-Left, TR : Top-Right, BL : Bottom-Left, BR : Bottom-Right$\}$.

must be active:

$$mwg_{m,W_g^T} \wedge mwg_{m,W_g^L} \Rightarrow rum_{g,TL,m} \vee rum_{g,BR,m} \vee cb_{g,TL}$$
$$\forall \text{ 4 corners}, m = 1...N_m, g = 1...N_{gru}$$

Each MRR can only be used for one message. The following constraints both set the value of $ru_{g,p}$ and enforce that restriction:

$$ru_{g,p} = \sum_{m=1}^{N_m} rum_{g,p,m} \qquad \forall g = 1...N_{gru}, p \in \mathbb{P}$$

Corner bending. The following three sets of constraints are required[1]:

(1) A GRU cannot have corners bent and MRRs active.

$$cb_{g,p_1} + ru_{g,p_2} \leqslant 1 \qquad \forall p_1, p_2 \in \mathbb{P}, g = 1...N_{gru}$$

(2) Corners for the same edge cannot be bent at the same time for the same GRU.

$$cb_{g,TL} + cb_{g,TR} \leqslant 1 \qquad cb_{g,TR} + cb_{g,BR} \leqslant 1$$
$$cb_{g,TL} + cb_{g,BL} \leqslant 1 \qquad cb_{g,BL} + cb_{g,BR} \leqslant 1$$
$$\forall g = 1...N_{gru}$$

(3) If a corner is bent then messages present on one of the edges of that corner must be present on the other.

$$cb_{g,TL} \Rightarrow mwg_{m,W_g^T} = mwg_{m,W_g^L}$$
$$\forall \text{ 4 corners}, m = 1...N_m, g = 1...N_{gru}$$

Crossing loss. A message suffers crossing loss when going through a crossing with a perpendicular waveguide. Two things must happen for a message to have crossing loss on a GRU: **1)** the message must take a direct path through the GRU and **2a)** the perpendicular direct path must be taken by at least one other message *or* **2b)** there must be at least one message taking the path on Figure 5(c). For any other case the crossing on the GRU can be avoided, as exemplified in Figure 4(b), and no crossing loss exists.

First set the values of the variables mch_g and mcv_g:

$$mwg_{m,W_g^L} \wedge mwg_{m,W_g^R} \Rightarrow mch_g$$
$$\forall \text{ 2 directions}, m = 1...N_m, g = 1...N_{gru}$$
$$mwg_{m,W_g^T} \wedge mwg_{m,W_g^L} \wedge rum_{g,BR,m} \Rightarrow mch_g \wedge mcv_g$$
$$\forall \text{ 4 corners}, m = 1...N_m, g = 1...N_{gru}$$

The value of $cl_{g,m}$ follows:

$$mwg_{m,W_g^T} \wedge mwg_{m,W_g^B} \wedge mch_g \Rightarrow cl_{g,m}$$
$$\forall \text{ 2 directions}, m = 1...N_m, g = 1...N_{gru}$$

Through loss. If a message is going through the direct path on a GRU, then it has through loss for each MRR present on that GRU.

$$mwg_{m,W_g^L} \wedge mwg_{m,W_g^R} \wedge ru_{g,p} \Rightarrow tl_{g,p,m}$$
$$\forall \text{ 2 directions}, m = 1...N_m, p \in \mathbb{P}, g = 1...N_{gru}$$

Bending loss. A message has bending loss on a GRU if it routes through a corner that is bent.

$$mwg_{m,W_g^T} \wedge mwg_{m,W_g^L} \wedge cb_{g,TL} \Rightarrow bl_{g,m}$$
$$\forall \text{ 4 corners}, m = 1...N_m, g = 1...N_{gru}$$

Drop loss. Proportional to the number of MRRs used by each message.

Propagation loss. Proportional to the length of the waveguides the message goes through.

Message insertion loss. The total insertion loss of a message over all waveguides and GRUs is a weighted sum.

$$mil_m = \sum_{i=1}^{N_{wg}} (L^P * L_i + L_i^E) * mwg_{m,i} + L^T * \sum_{g=1}^{N_{gru}} \sum_{p \in \mathbb{P}} tl_{g,p,m}$$
$$+ \sum_{g=1}^{N_{gru}} (L^C * cl_{g,m} + L^B * bl_{g,m} + L^D * \sum_{p \in \mathbb{P}} rum_{g,p,m})$$
$$\forall m = 1...N_m$$

[1] This feature can be turned off, if needed, by adding constraints to set all $cb_{g,p}$ variables to zero.

4.2 Objective function

Calculating the number of wavelengths is done with the following constraints:

$$wlu_\lambda \geqslant mwl_{m,\lambda} \qquad \forall m = 1...N_m, \lambda = 1...N_\lambda$$

$$nwl = \sum_{\lambda=1}^{N_\lambda} wlu_\lambda$$

Determining the maximum insertion loss over all messages is done with the following constraints:

$$maxil \geqslant mil_m \qquad \forall m = 1...N_m$$

Finally, the following objective function is minimized:

$$\alpha_1 * nwl + \alpha_2 * maxil + \alpha_3 * \sum_{m=1}^{N_m} mil_m + \alpha_4 * \sum_{g=1}^{N_{gru}} \sum_{p \in \mathbb{P}} ru_{g,p}$$

where α_i are optimization weights chosen by the designer.

Since the value for the insertion loss of each message is available through the mil_m variables, functions other than the maximum or the sum of the insertion loss can also be added to the model and used for optimization.

4.3 Model reduction techniques

4.3.1 Restrictions on usage of wavelengths. The following constraints can be added:

$$mwl_{m,\lambda} = 0 \qquad \forall \lambda = (m+1)...N_\lambda \qquad \forall m = 1...N_m$$

They restrict the possible wavelengths for each message: message 1 uses wavelength 1, message 2 uses wavelengths 1 or 2, etc. This way, some meaningless variations around the same effective solution are removed. The optimal solution, however, is not removed from the solution space.

4.3.2 Restrictions on usage of MRRs. Empirically we find that minimizing the insertion loss favors optimal solutions where each message uses a low total number of MRRs. Following this reasoning, constraints can be added to the model that force a maximum number of MRRs per message (R^{max}):

$$\sum_{g=1}^{N_{gru}} \sum_{p \in \mathbb{P}} rum_{g,p} \leqslant R^{max} \qquad \forall m = 1...N_m$$

This reduces the set of paths considered by the solver by removing poor, convoluted paths while keeping the more direct paths between endpoints.

5 PROOF OF FEASIBILITY

It is possible that the chosen layout template cannot satisfy the entire communication matrix (for example, if the template is too small). For those cases, the model above will be unfeasible. Verifying the existence of a solution can be done much faster using a simplified version of the model. For that we consider $N_\lambda = N_m$ and uniquely assign a wavelength to each message by adding these constraints:

$$mwl_{m,\lambda} = 1 \qquad \forall m = 1...N_m, \lambda = m$$
$$mwl_{m,\lambda} = 0 \qquad \forall m = 1...N_m, \lambda \neq m$$

The resulting model can be solved much faster but, if the solver is unable to find a feasible solution for this simplified model, the complete model is also unfeasible.

Proof. Assume a feasible solution exists. It will have $nwl \leqslant N_m$. From that solution build another where each message uses its own wavelength (thus either maintaining or increasing nwl). Any message that changes its wavelength must also change the wavelength of the MRRs it uses. This is always possible because each MRR routes only one message. Furthermore, the wavelength exclusion rule is always satisfied. Hence, the feasibility of the complete model implies the existence of a solution for the simplified version. □

6 3-STEP OPTIMIZATION

Section 4 introduced a MIP model that is capable of solving the constrained problem for *any* layout template. Therefore, programming the model as presented on any MIP solver and solving it directly for the chosen minimization objective is enough to obtain the optimal solution. However, due to the nature of the problem, it is possible to slightly alter the optimization process yielding more control and faster results. This leads to the proposed 3-step optimization process used in PSION, where each step optimizes a slightly different version of the model and produces a solution used at the start of the next step.

In the **first step** we consider $N_\lambda = N_m$ and apply the feasibility proof from Section 5. In this way we can generate the first feasible solution much faster if one exists. It can then be used as a warm start, which decreases optimization times substantially. This has the added bonus of stopping the process as quickly as possible if unfeasible.

In the **second step** we only minimize the number of wavelengths, for two reasons. Firstly, the designer will most likely want to use fewer wavelengths than the number of messages, thus making this optimization problem hierarchical, i.e., minimizing wavelengths has a higher priority than minimizing insertion loss or #MRRs. Secondly, because, after completing this step, a feasible solution for a smaller number of wavelengths is then available, so the model can again be simplified by eliminating from it the $N_m - nwl$ unused wavelengths. To make this simplification, the following constraints are added:

$$mwl_{m,\lambda} = 0 \qquad \forall m = 1...N_m, \text{ unused wavelengths } \lambda$$

The designer might be willing to use more wavelengths than the minimum needed. In that case it is up to the designer to know the maximum acceptable number of wavelengths. The second step can be stopped earlier once a solution is found within that acceptable range.

In the **third step** we consider the model with the needed amount of wavelengths only and further optimize the last solution using the chosen function (*maxil*, for example). We have now reached the final solution.

Using this process we can notably simplify the problem space during the optimization. However, because the model reductions are always done according to the hierarchical characteristics of the optimization goals, the optimal solution is never missed.

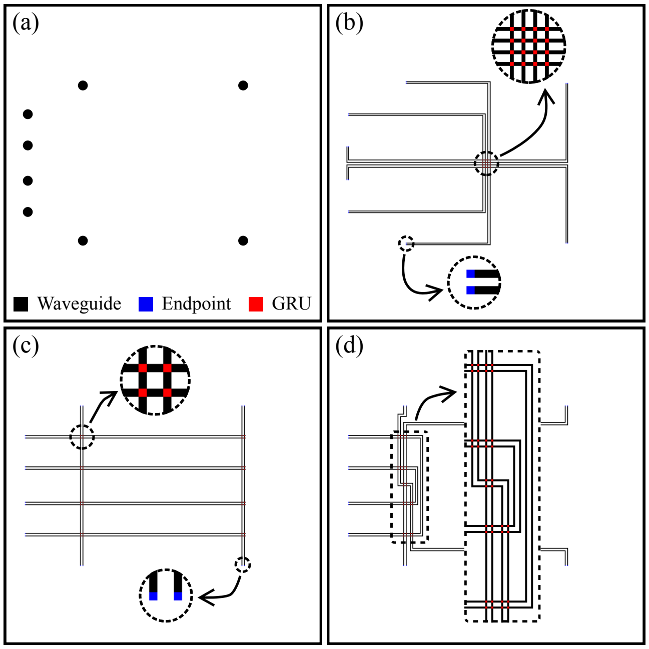

Figure 6: (a) Location of the eight nodes that produces the best result in PROTON+. (b) A centralized grid template connecting those nodes. (c) A distributed grid template. (d) A custom template.

7 RESULTS

The MIP model and 3-step optimization algorithm are programmed in C++ and make use of Gurobi [4], a MIP solver, on a 2.6 GHz CPU.

We tested our model and optimization procedure against the state-of-the-art PROTON+ and PlanarONoC P&R tools. Most of their result analysis is dedicated to an 8 node test case with 44 messages. We solved the same test case considering the same communication matrix, node placement, die size, crossing size and loss parameters.

PROTON+ and PlanarONoC compare results originating from P&R of three logical topologies (8x8 λ-Router, 8x8 GWOR and 8x8 Standard Crossbar). PROTON+ also considers five different sets of node positions and various permutations of solver parameters, which results in a range of values for the results. We used the node positions that produced the best result over all presented in PROTON+, shown in Figure 6(a). We manually designed three simple layout templates, presented in Figure 6(b-d), that connect to these node positions. The last step of the optimization was set to minimize the max. insertion loss (*maxil*), just like PROTON+ and PlanarONoC.

7.1 Physical templates

All templates share some common features:

(1) Each node has two endpoints: a modulator and a demodulator.
(2) The power distribution network – not shown in these templates – can always be routed from the outside such that no other crossings in the router exist besides those considered by the template.

The **centralized grid** template is a $w \times h$ grid of GRUs where $w + h$ equals the number of nodes. Each node is connected with waveguides to two ports on the grid (one for sending, the other for receiving), which are next to each other. This router can be thought of as a different generalization of the 4x4 GWOR router in Figure 3. The grid itself was placed on the center of the die, the ports used by each node were chosen as to remove any crossings external to the grid and the waveguides connecting the nodes to the grid were manually routed to minimize bends.

The **distributed grid** template was built by placing horizontal or vertical pairs of waveguides starting at each node, with a GRU on each crossing.

The **custom** template was built specifically for this test case (i.e., these node positions and communication matrix). In particular, no message needs to use more than one MRR. Therefore, R^{max} was set to 1 for this template while the grid templates were solved with $R^{max} = 2$.

Table 3: Results for 8 nodes, 44 messages

	#WLs	**Max IL**	#MRRs	Time	
PROTON+				T_{total}	
λ-Router	**8**	**6.6 - 9.0**	56	134	
GWOR	**7**	**8.1 - 11.3**	48	79	
Std. crossbar	**8**	**10.5 - 13.0**	64	602	
PlanarONoC				T_{total}	
λ-Router	**8**	**5.2**	56	<1	
GWOR	**7**	**6.4**	48	<1	
Std. crossbar	**8**	**7.4**	64	<1	
PSION				T_{opt}	T_{total}
Centralized	**8**	**3.1**	52	178	271
Distributed	**8**	**3.6**	48	37	376
Custom	**7**	**4.1**	40	<1	6

T_{opt} is time to find the optimal solution, T_{total} is total execution time (for PSION: $T_{total} = T_{opt}$ + time to prove optimality; for others: the time that produces the best result). Time in seconds, insertion loss in dB.

7.2 Comparison to the state-of-the-art

Figure 1 shows the result for the centralized grid router and Table 3 presents the various comparisons. Most important are the number of wavelengths and maximum insertion loss, but #MRRs and execution time are also given. Results from PSION are optimal solutions for the given templates.

Number of wavelengths. The communication matrix in these tests requires an absolute minimum of 7 wavelengths when using one modulator per node. The custom template matches this value, but the grid templates require an actual minimum of 8. However, PSION can reduce this number if given a smaller communication matrix, in contrast to the presented logical topologies.

Max. insertion loss. PSION produces results that are **2.7×** better compared to PROTON+ and **1.8×** better compared to PlanarONoC. Some intuitive reasons are available to justify these outcomes:

- We combined logical topology and physical layout optimization.
- We used templates, which automatically removes many suboptimal solutions compared to a conventional P&R solution space.
- We used GRUs, which support up to four MRRs per crossing, whereas PSEs only support two. Thus, fewer GRUs are used in our templates than PSEs are used in logical topologies such as the λ-Router. This increases the density of our designs which decreases the total number of crossings.
- We drastically reduced the number of crossings outside PSEs/GRUs.
- We obtain the optimal solution within the specified template.

MRR usage. This was not an optimization objective in these tests, but the comparison to both PROTON+ and PlanarONoC remains favourable.

Time. Grid templates have a total execution time comparable with PROTON+. PlanarONoC is still two orders of magnitude faster. The custom template is much better, however, mostly because of the technique from Section 4.3.2.

Furthermore, the optimal solution is consistently reached in half or less than the total execution time. Thus, a designer not requiring proof of optimality can end the optimization once a satisfactory solution is found which, based on these results, is likely to appear quickly and be close to optimal.

7.3 Further comments

We also solved the MIP models from these tests by directly minimizing $100 \times nwl + 1 \times maxil$ – which assures the same hierarchical optimization – and got the same final results, but found that using the 3-step procedure is **2.5×** faster on average. Likewise, we ran the same tests without any of the reduction techniques from Section 4.3. The results were the same, but using the techniques was **4.5×** faster on average.

Finally, the grid templates are entirely straightforward and can be used in virtually any WRONoC, which speaks to the potential of PSION even when no effort is spent in designing the template. The custom template, however, was built for this case. The fact that it achieves even better results in some areas also shows the promising possibilities available through careful template synthesis.

8 CONCLUSION

In this work we defined the WRONoC design problem and presented PSION, a novel method for solving it. This method uses a physical layout template to combine logical topology and physical layout optimization. We also presented a new, flexible, routing element, the GRU. We used a MIP model and a 3-step optimization procedure to solve for the optimal solution. These combined efforts produce results superior to the state of the art. In future work the proposed method can be extended to include optimization of the power distribution network and other GRU designs. Also, the runtime characteristics of MIP modelling may yet be improved with further reduction techniques. Finally, template synthesis methods should also be explored.

REFERENCES

[1] Sébastien Le Beux, Ian O'Connor, Gabriela Nicolescu, Guy Bois, and Pierre Paulin. 2013. Reduction methods for adapting optical network on chip topologies to 3D architectures. *Microprocessors and Microsystems* 37, 1 (2013), 87 – 98. https://doi.org/10.1016/j.micpro.2012.11.001

[2] Yu-Kai Chuang, Kuan-Jung Chen, Kun-Lin Lin, Shao-Yun Fang, Bing Li, and Ulf Schlichtmann. 2018. PlanarONoC: Concurrent Placement and Routing Considering Crossing Minimization for Optical Networks-on-chip. In *Proceedings of the 55th Annual Design Automation Conference*. ACM, Article 151, 6 pages.

[3] H. Gu, K. H. Mo, J. Xu, and W. Zhang. 2009. A Low-power Low-cost Optical Router for Optical Networks-on-Chip in Multiprocessor Systems-on-Chip. In *2009 IEEE Computer Society Annual Symposium on VLSI*. 19–24. https://doi.org/10.1109/ISVLSI.2009.19

[4] Gurobi Optimization, Inc. 2018. *Gurobi Optimizer Reference Manual*. http://www.gurobi.com.

[5] Mengchu Li, Tsun-Ming Tseng, Davide Bertozzi, Mahdi Tala, and Ulf Schlichtmann. 2018. CustomTopo: A Topology Generation Method for Application-Specific Wavelength-Routed Optical NoCs. In *Proceedings of the 37th International Conference on Computer-Aided Design*.

[6] M. Nikdast, J. Xu, L. H. K. Duong, X. Wu, X. Wang, Z. Wang, Z. Wang, P. Yang, Y. Ye, and Q. Hao. 2015. Crosstalk Noise in WDM-Based Optical Networks-on-Chip: A Formal Study and Comparison. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 23, 11 (Nov 2015), 2552–2565. https://doi.org/10.1109/TVLSI.2014.2370892

[7] I. O'Connor, M. Brière, E. Drouard, A. Kazmierczak, F. Tissafi-Drissi, D. Navarro, F. Mieyeville, J. Dambre, D. Stroobandt, J.-M. Fedeli, Z. Lisik, and F. Gaffiot. 2005. Towards reconfigurable optical networks on chip. *ReCoSoC'05* (2005), 121–128.

[8] Marta Ortín-Obón, Luca Ramini, Herve Tatenguem Fankem, Víctor Viñals, and Davide Bertozzi. 2014. A Complete Electronic Network Interface Architecture for Global Contention-free Communication over Emerging Optical Networks-on-chip. In *Proceedings of the 24th Edition of the Great Lakes Symposium on VLSI*. ACM, 267–272. https://doi.org/10.1145/2591513.2591536

[9] M. Ortín-Obón, L. Ramini, V. Viñals Yúfera, and D. Bertozzi. 2017. A tool for synthesizing power-efficient and custom-tailored wavelength-routed optical rings. In *Asia and South Pacific Design Automation Conference (ASP-DAC)*. 300–305. https://doi.org/10.1109/ASPDAC.2017.7858339

[10] A. Peano, L. Ramini, M. Gavanelli, M. Nonato, and D. Bertozzi. 2016. 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)*. 1–8. https://doi.org/10.1145/2966986.2967023

[11] L. Ramini, P. Grani, S. Bartolini, and D. Bertozzi. 2013. Contrasting wavelength-routed optical NoC topologies for power-efficient 3d-stacked multicore processors using physical-layer analysis. In *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 1589–1594. https://doi.org/10.7873/DATE.2013.323

[12] M. Ashkan Seyedi, Antoine Descos, Chin-Hui Chen, Marco Fiorentino, David Penkler, François Vincent, Bertrand Szelag, and Raymond G. Beausoleil. 2016. Crosstalk analysis of ring resonator switches for all-optical routing. *Opt. Express* 24, 11 (May 2016), 11668–11676. https://doi.org/10.1364/OE.24.011668

[13] M. Tala, M. Castellari, M. Balboni, and D. Bertozzi. 2016. Populating and exploring the design space of wavelength-routed optical network-on-chip topologies by leveraging the add-drop filtering primitive. In *2016 Tenth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*. 1–8. https://doi.org/10.1109/NOCS.2016.7579331

[14] X. Tan, M. Yang, L. Zhang, Y. Jiang, and J. Yang. 2011. On a Scalable, Non-Blocking Optical Router for Photonic Networks-on-Chip Designs. In *2011 Symposium on Photonics and Optoelectronics (SOPO)*. 1–4. https://doi.org/10.1109/SOPO.2011.5780550

[15] Anja von Beuningen, Luca Ramini, Davide Bertozzi, and Ulf Schlichtmann. 2015. PROTON+: A Placement and Routing Tool for 3D Optical Networks-on-Chip with a Single Optical Layer. *J. Emerg. Technol. Comput. Syst.* 12, 4, Article 44 (Dec. 2015), 28 pages. https://doi.org/10.1145/2830716

[16] Anja von Beuningen and Ulf Schlichtmann. 2016. PLATON: A Force-Directed Placement Algorithm for 3D Optical Networks-on-Chip. In *Proceedings of the 2016 on International Symposium on Physical Design*. ACM, 27–34. https://doi.org/10.1145/2872334.2872356

[17] Yiyuan Xie, Mahdi Nikdast, Jiang Xu, Wei Zhang, Qi Li, Xiaowen Wu, Yaoyao Ye, Xuan Wang, and Weichen Liu. 2010. Crosstalk Noise and Bit Error Rate Analysis for Optical Network-on-chip. In *Proceedings of the 47th Design Automation Conference*. ACM, 657–660. https://doi.org/10.1145/1837274.1837441