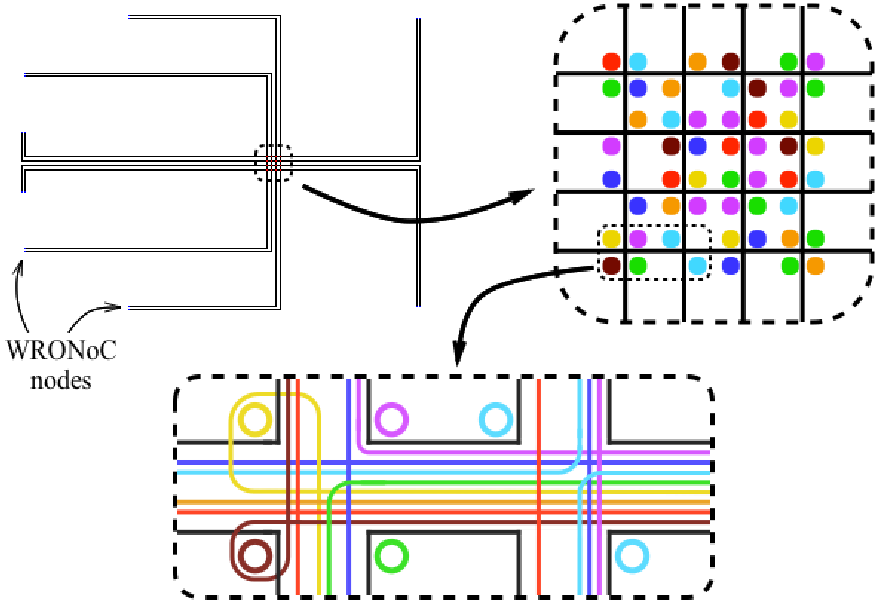
PSION: Combining Logical Topology and Physical Layout Optimization for Wavelength-Routed ONoCs



Alexandre Truppel[+], Tsun-Ming Tseng[+], Davide Bertozzi[*], José Carlos Alves[#], Ulf Schlichtmann[+]

+ Technical University of Munich, Germany

* University of Ferrara, Italy

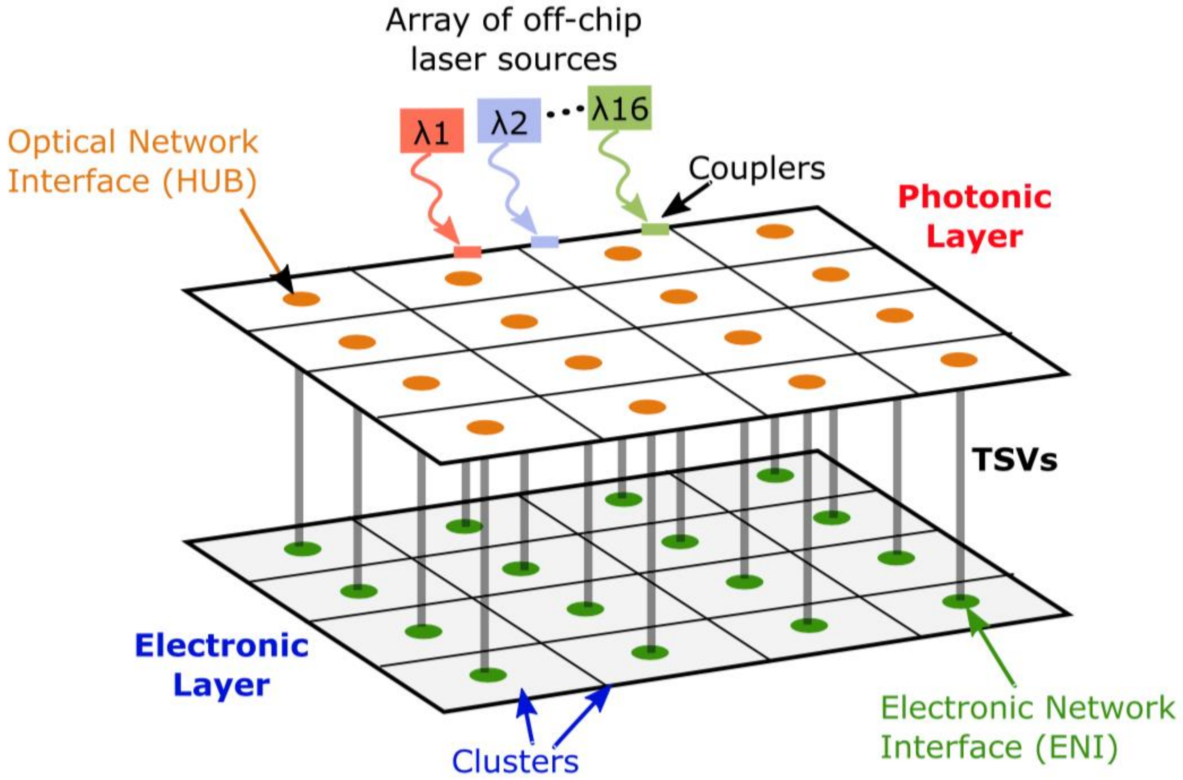
University of Porto, Portugal

Summary

- Brief introduction to ONoCs
- WRONoC design problem & state of the art
- New methodology in PSION
- Optimization algorithm
- Results
- Conclusion

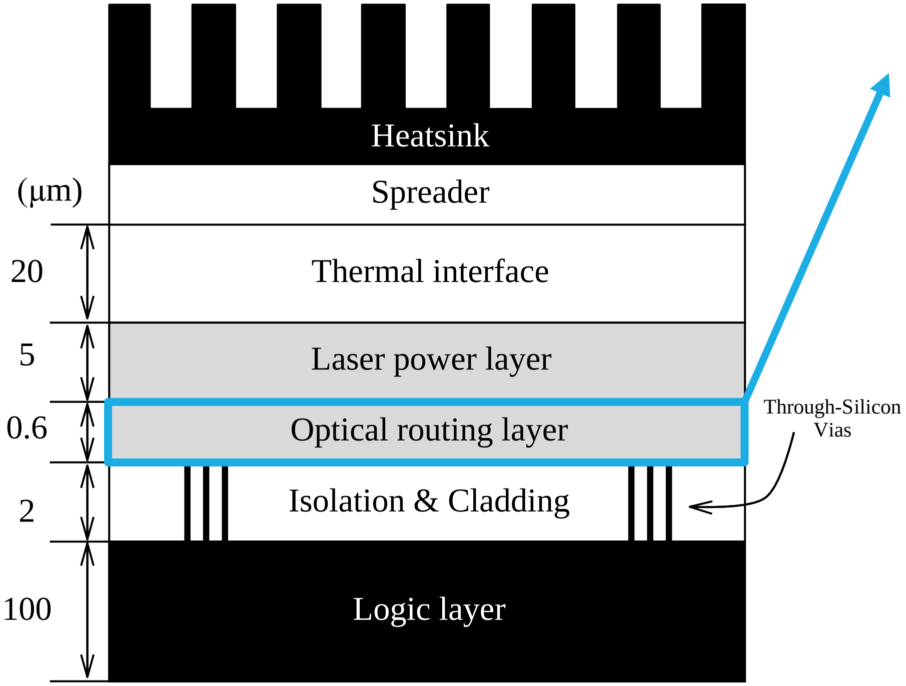
Introduction to ONoCs

- ONoCs – Optical Networks-on-Chip
- Compared to Electrical NoCs, potential for:
  - Lower latency
  - Lower dynamic power consumption
  - Greater bandwidth
- Passive ONoCs use light **w**avelength for **r**outing – **WR**ONoCs

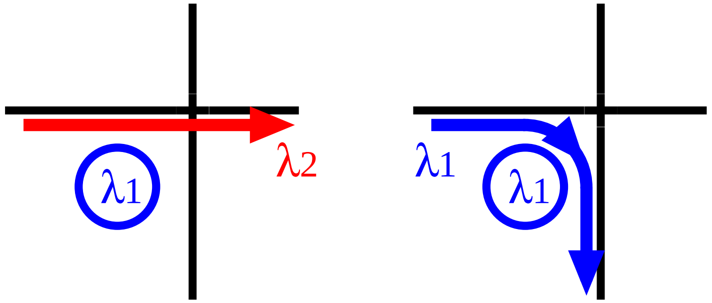


Sources:
(1) Contrasting Laser Power Requirements of Wavelength-Routed Optical NoC Topologies Subject to the Floorplanning, Placement, and Routing Constraints of a 3-D-Stacked System, Marta Ortín-Obón et al. In IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 25, No. 7, July 2017

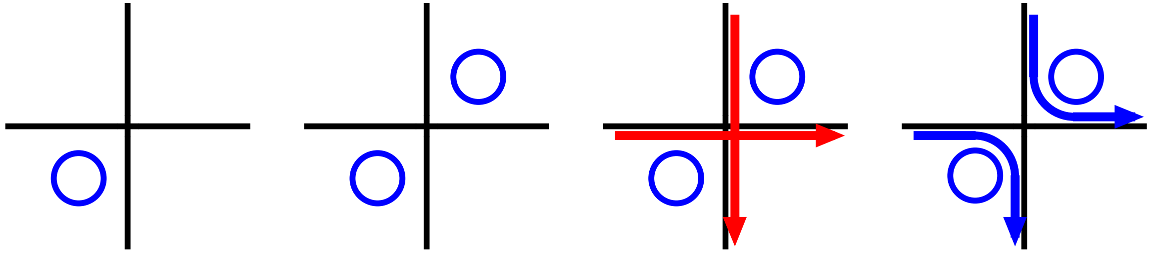
Elements of WRONoCs



- Modulators & demodulators (E⇔O interfaces)
- Waveguides
- Micro-Ring Resonators (MRRs):



- Photonic Switching Elements (PSEs):



1x2 PSE   2x2 PSE   PSE routing

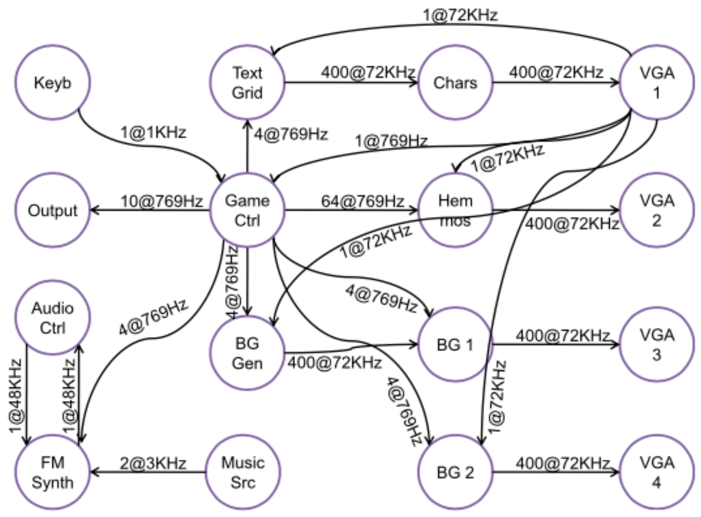
Sources:
(1) Sharing and placement of on-chip laser sources in silicon-photonic NoCs, C. Chen et al. In 2014 Eighth IEEE/ACM International Symposium on Networks-on-Chip (NoCS)

The design & layout optimization problem of Wavelength-Routed ONoCs

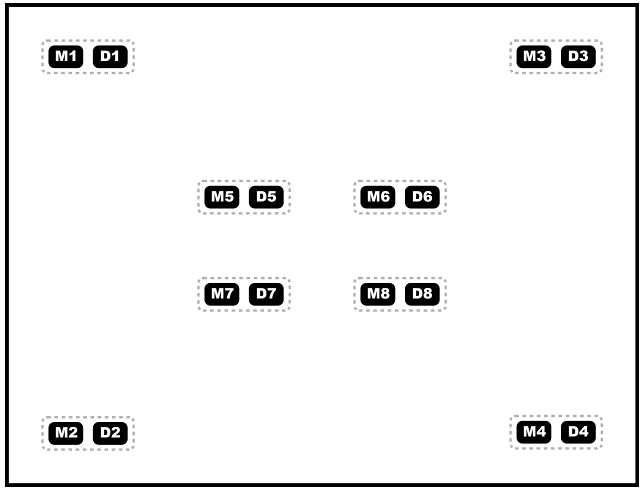
Inputs, outputs & optimization objectives

State of the art procedure: description & issues

1) A communication graph/matrix

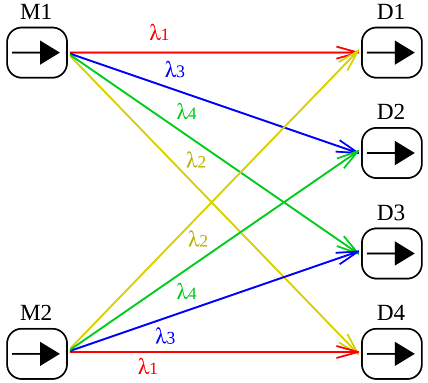


2) The physical location of the nodes (modulators & demodulators) on the optical plane

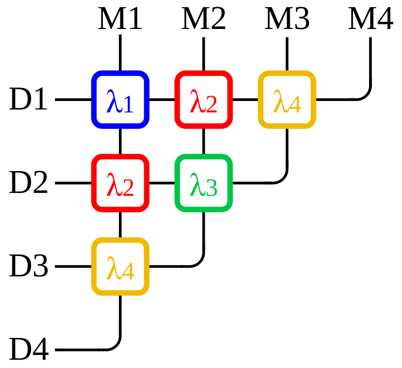


1) The logical topology of the router

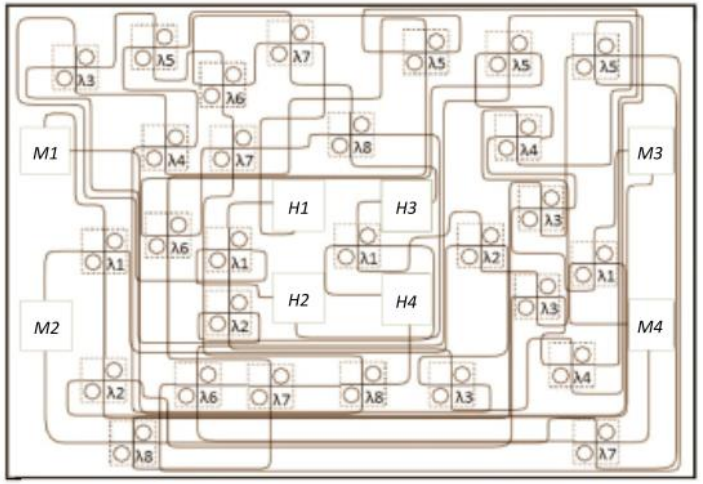
1.1) An assignment of a wavelength to each message



1.2) The logical connections between PSEs and nodes



2) The physical layout of the router



Sources:

(1) A scalable, non-interfering, synthesizable network-on-chip monitor – extended version, Alhonen et al. In Microprocessors and Microsystems, 2013.

(2) Proton+: A placement and routing tool for 3d optical networks-on-chip with a single optical layer, Beuningen et al. In Emerg. Technol. Comput. Syst., December 2015.

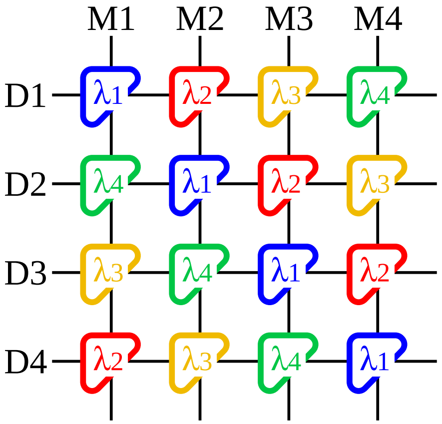
Minimization goals

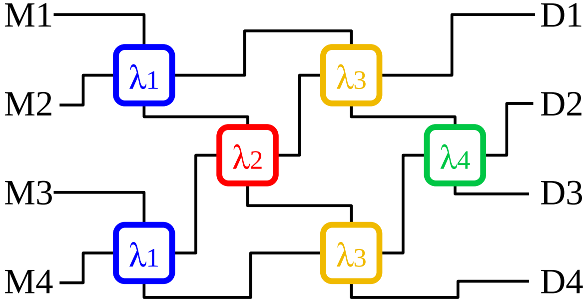
- Message insertion loss → directly impacts power usage of the laser sources
- Number of wavelengths used
- Number of unique/total MRRs used

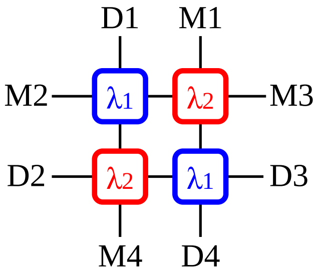
- Why?
  - Power usage
  - Performance (throughput & bandwidth)
  - Required physical resources

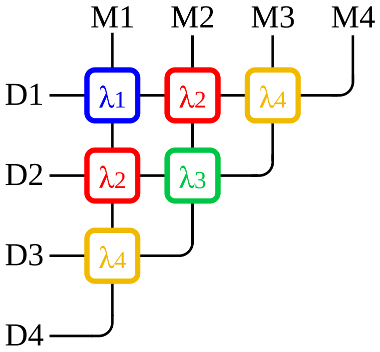
State of the art: 2-step procedure

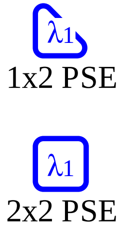
1. **Choose a logical topology** (for example Lambda, GWOR, standard crossbar…)



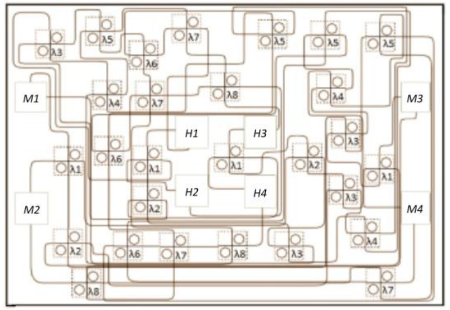








2. **Place and route (P&R)** all waveguides and PSEs/MRRs of that topology
   - **Proton+** and **PlanarONoC** are the state of the art tools for this step



State of the art: pitfalls

- **During topology choice/synthesis**: unable to predict P&R results.
- **During P&R**: topology is already fixed and only local optimum can be obtained.

- An example of results that may be caused by the asynchronization of the 2 steps:
  - After choosing topology: **28** total crossings (for 8x8 Lambda-router)
  - After physical design: **90** total crossings (after Proton+)
  - The main motivation for this work!

- Thus, the **optimal solution** can only be reached when:

a. **Both optimization steps are taken together**, not one by one

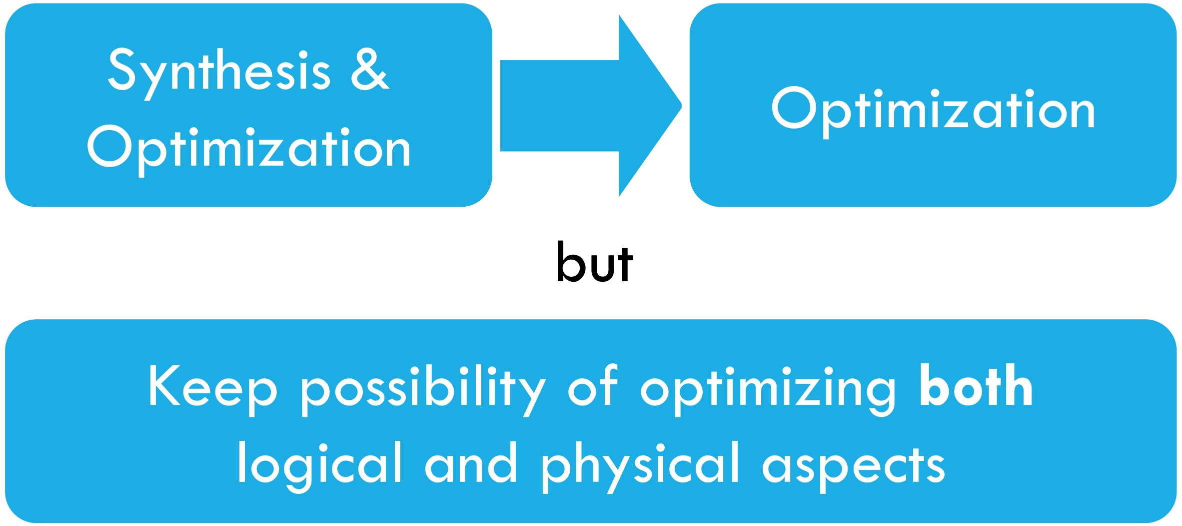
b. Corollary: Must consider the two inputs to the problem (CM & node positions)

Proposed methodology

Theoretical approach
Optimization algorithm

Constrain the problem

- Choosing logical topology first constrains the problem
- Basis of our approach is to also constrain the problem, but do it **better**:



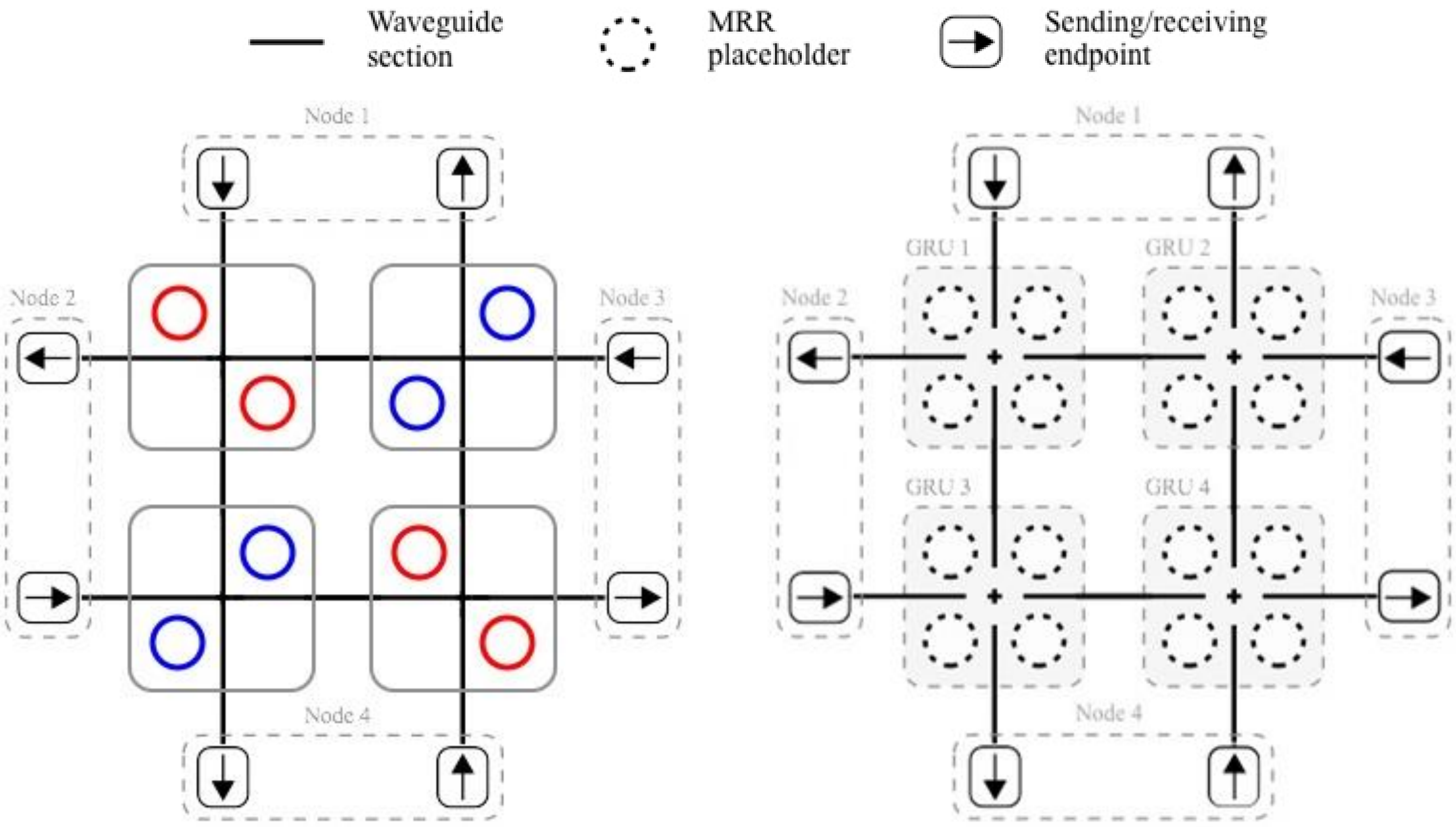
- Use a **physical layout template**.

Physical layout template

- A collection of WRONoC router elements already placed and routed on the optical plane.
  - Positions of nodes are automatically considered in the template.
- The template can be **created manually**.
- The template is an **input to the optimization procedure**.
- The solution must conform to this template.
  - An optimization algorithm will never be asked to place any new elements in new locations.

Physical layout template elements

- Endpoints
  - Modulators & demodulators
- Waveguide sections
- General Routing Units (GRUs)
  - Similar to PSEs
  - Contain MRR placeholders

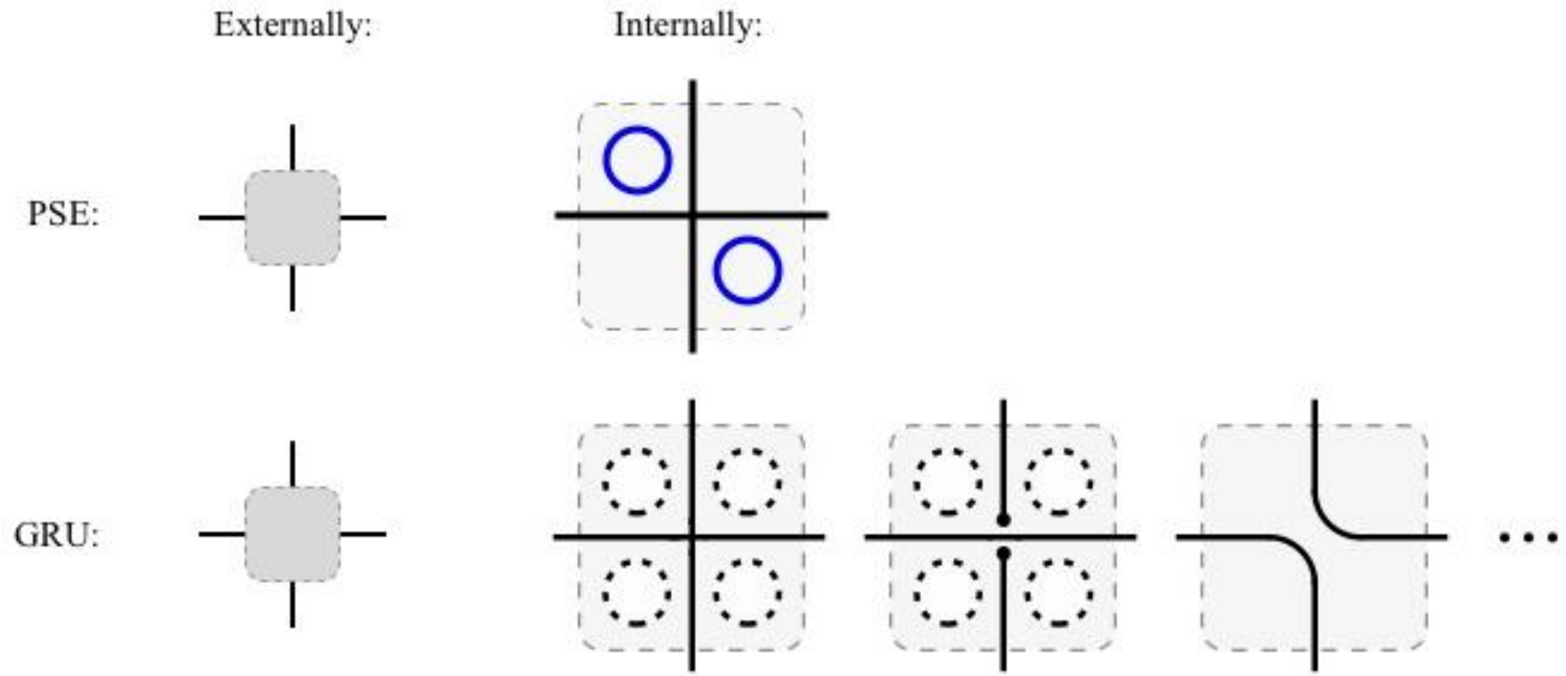


Currently: GWOR router using PSEs

New: Physical Layout Template using GRUs

General Routing Unit (GRU)

- Externally, equal to a PSE.
- Internally, **many different structures possible**.
  - Different wavelengths for MRRs, crossing avoidance, corner bending. More structures possible in the future.



Optimization algorithm

- Must perform the following tasks:
  a. **Assign wavelengths** to messages
  b. **Route messages** through the template
  c. **Activate routing features** of GRUs
- This is a **combinatorial optimization problem** with a **linear optimization function**.
- One important detail: **feasible solutions are hard to find** and **iterating through the solution space is difficult**.

Use Mixed Integer Programming!

- Many advantages…
- Most importantly, **MIP gives optimal solutions**:
  - If fast enough → no other algorithm is needed
  - If too slow → provides a baseline comparison in speed & solution quality for other algorithms
  - Thus: good starting point

MIP speed-up techniques

- Explored some techniques to speed up the MIP solving procedure.
- A model reduction technique (doesn't remove optimal solutions).
- A heuristic (may *possibly* remove optimal solutions):
  - Restrictions on usage of MRRs (4.5x faster)
- Feasibility proof:
  - Very quickly find a feasible solution or prove infeasibility.
- 3-step optimization (2.5x faster):
  1. Use feasibility proof to find the first feasible solution very quickly
  2. Optimize only number of wavelengths → reduce problem space without harming optimality
  3. Perform optimization for the chosen optimization function

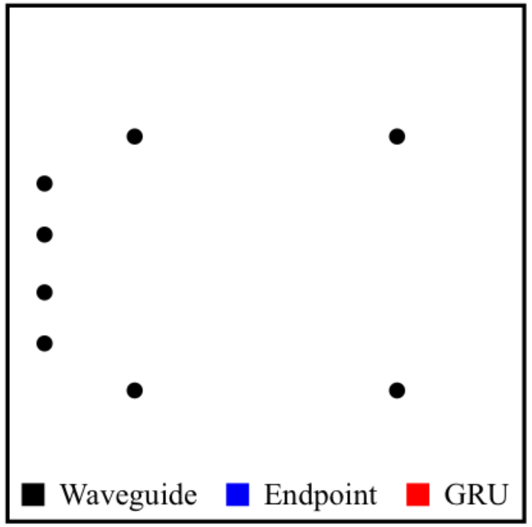
Results

Comparison to state of the art

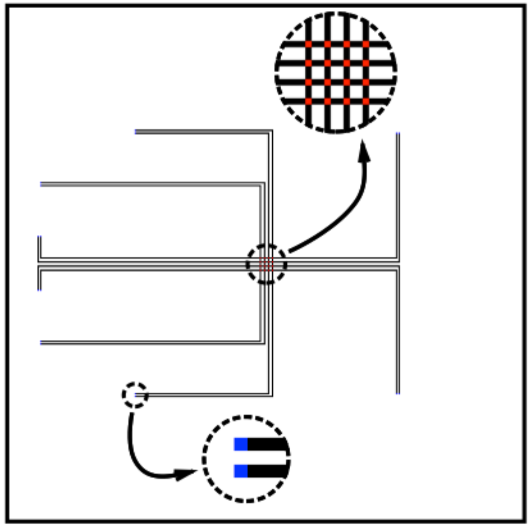
Example results

Comparison to state of the art

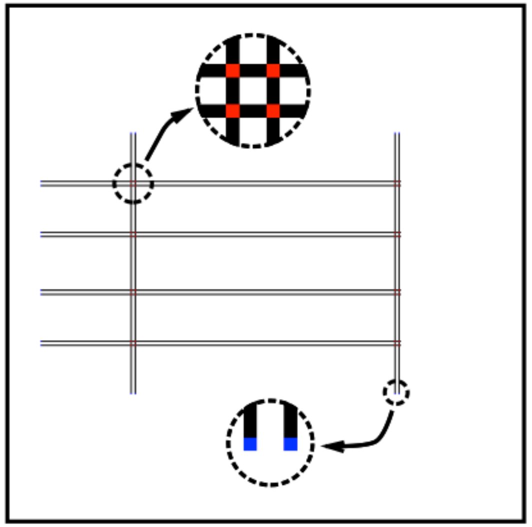
- **Proton+** and **PlanarONoC** are the state of the art tools for P&R of WRONoCs.
- Compared this new method against the best results available with Proton+/PlanarONoC for an **8 node, 44 message** test case (from Proton+).
- Node positions (from Proton+) and layout templates used:



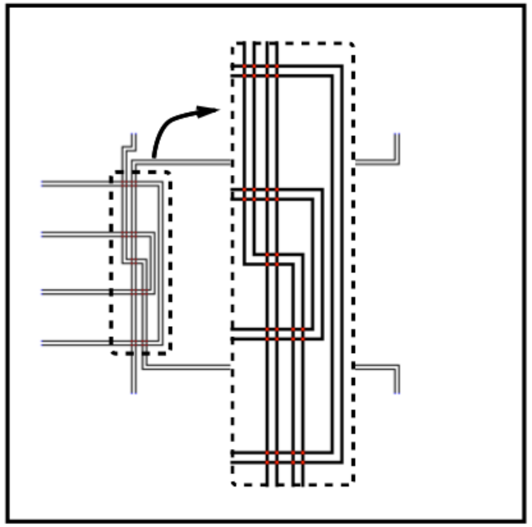
9x9 mm die size



Centralized grid router



Distributed grid router



Custom router

Verdict: major improvements

	#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

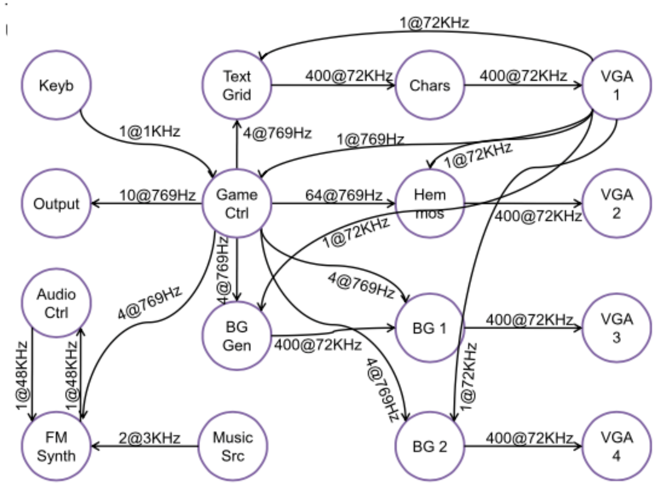
- **1.8x to 2.7x reduction** in maximum insertion loss.
- **Equal or better** number of wavelengths and MRRs.
- **Equivalent** optimization time to Proton+. *Custom* template takes **only 6 seconds** due to judicious use of solver heuristics.
- **Fast solution convergence.** Optimal (not proven) solution available in less than half of the total time.

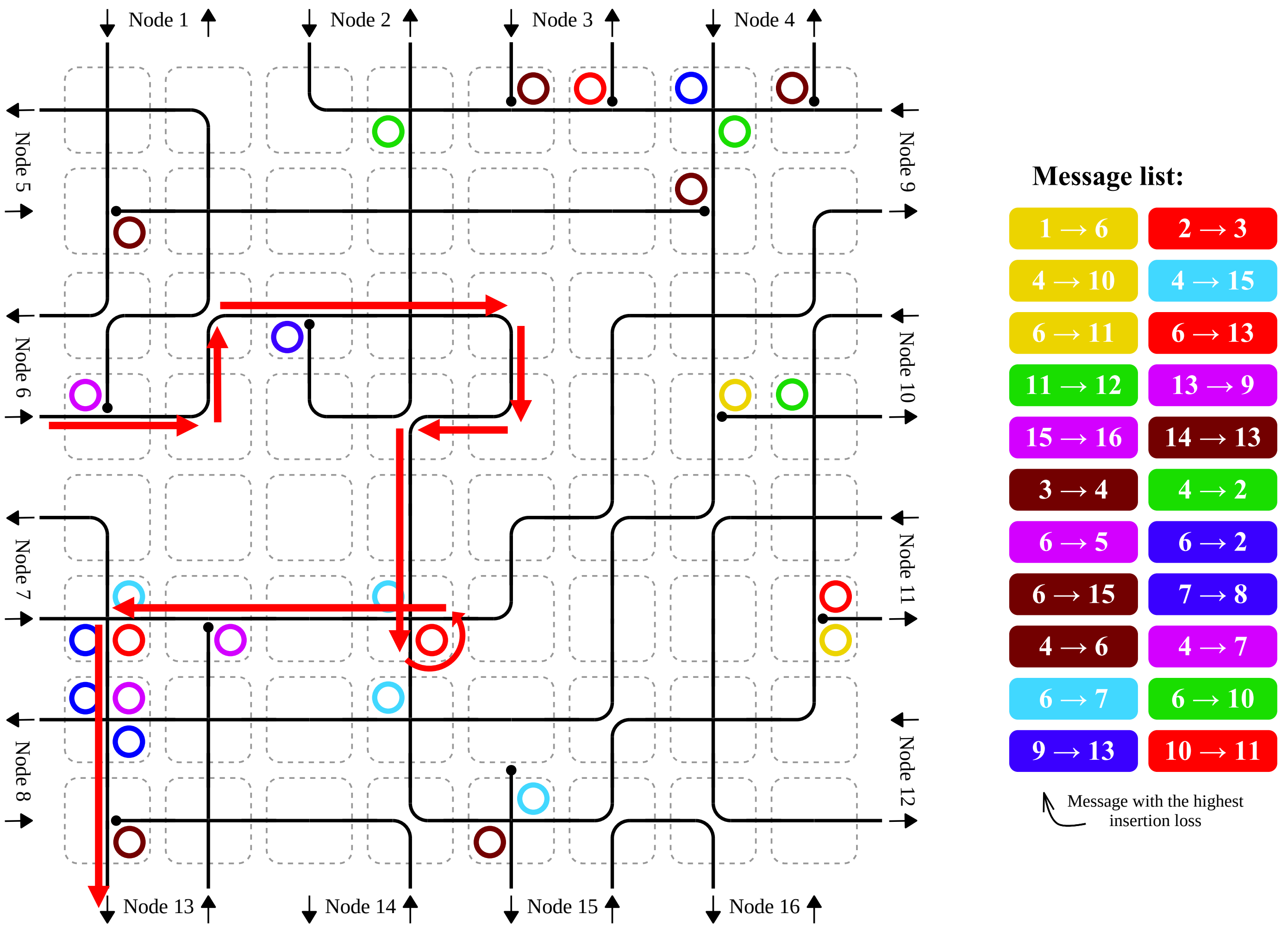
Verdict: major improvements (cont'd)

	#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

- **We target application-specific design.** For sparser communication matrices:
  - Insertion loss, #WLs and #MRRs are reduced with our method.
- Proton+/PlanarONoC are physical design tools only:
  - Communication matrix may change but logical topology is unchanged
  - Results are unchanged with sparser CMs.

Final example





- 16 nodes, 22 messages
  - Full CM would have 240 messages
- **240 MRRs** would be required with the Lambda-router
- Here only **27 MRRs** are used

Sources:
(1) A scalable, non-interfering, synthesizable network-on-chip monitor – extended version, Alhonen et al. In Microprocessors and Microsystems, 2013.

Conclusion

Major contributions

Future work

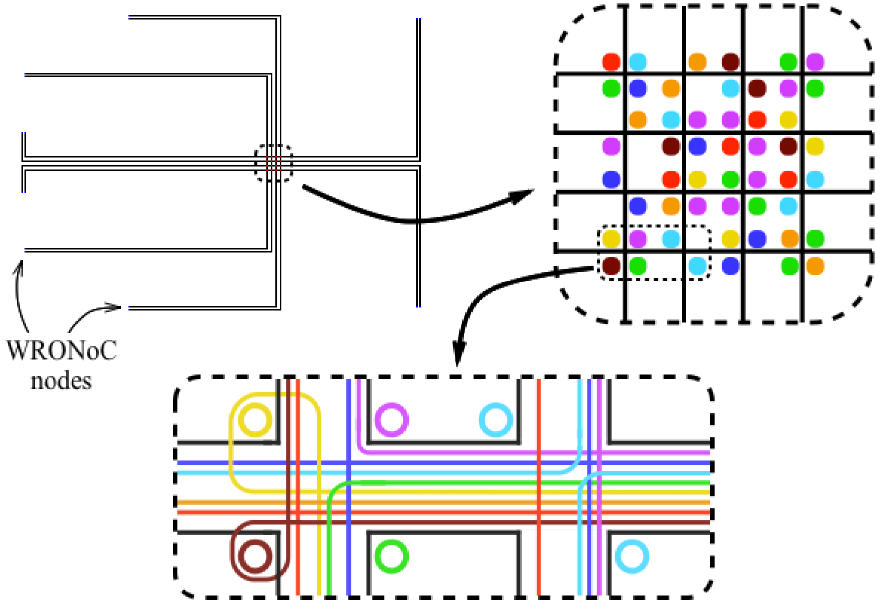
Major contributions

- Solved the WRONoC design problem differently using a **physical layout template**.
- Considered more physical routing possibilities with **Generic Routing Units (GRUs)**.
- Designed a fast algorithm to solve the problem using **MIP** and developed **multiple heuristics and reduction techniques to speed up optimization**.
- Got **results superior to state of the art**.

Future work

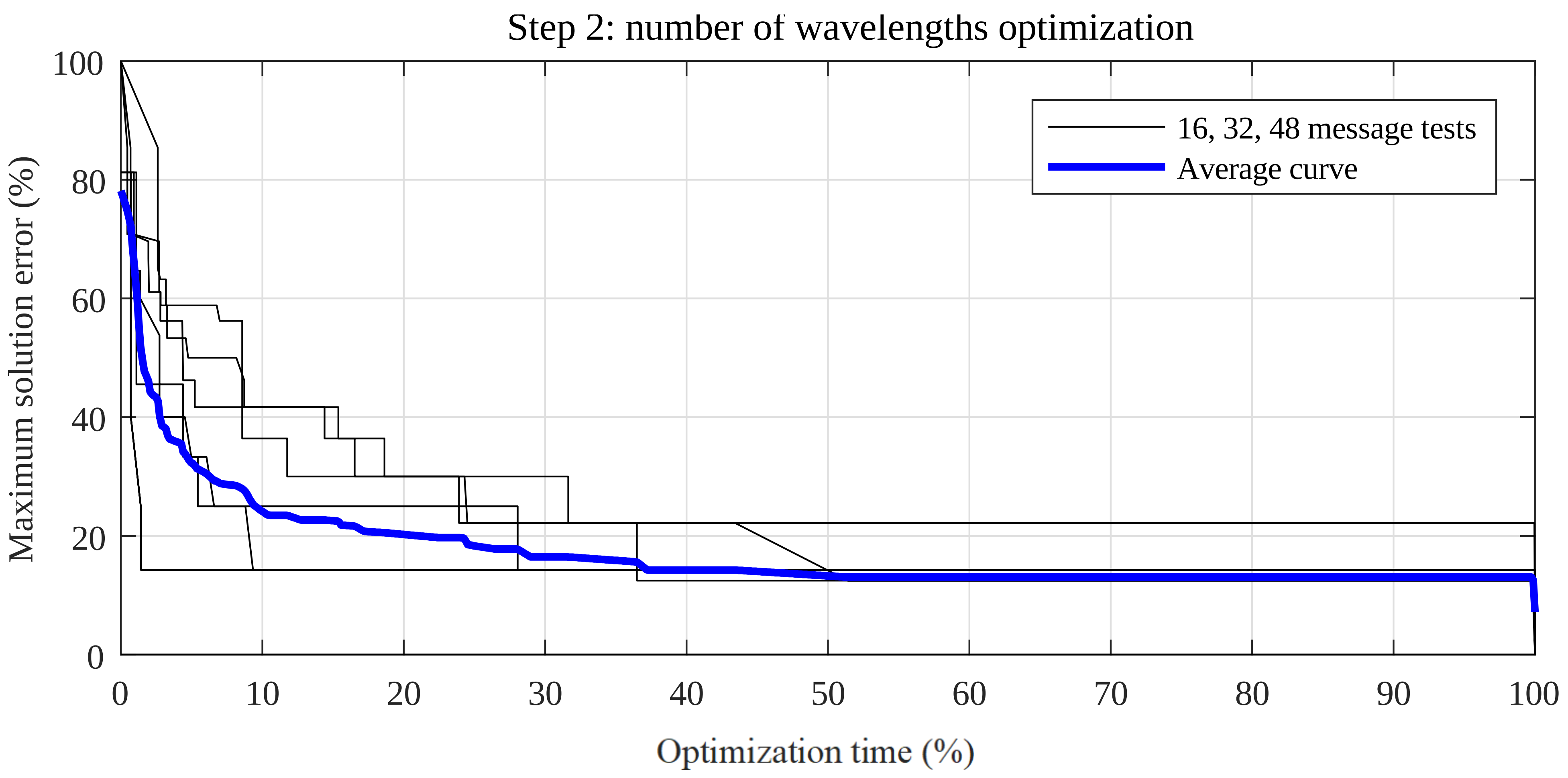
- Improve solver runtime with more heuristics
- Other GRU designs
- Optimize optical Power Distribution Network
- Consider other objectives (eg: crosstalk, thermal awareness...)
- Analyse other layout templates (eg: ring templates...), develop template synthesis tools

Thank you for your attention!

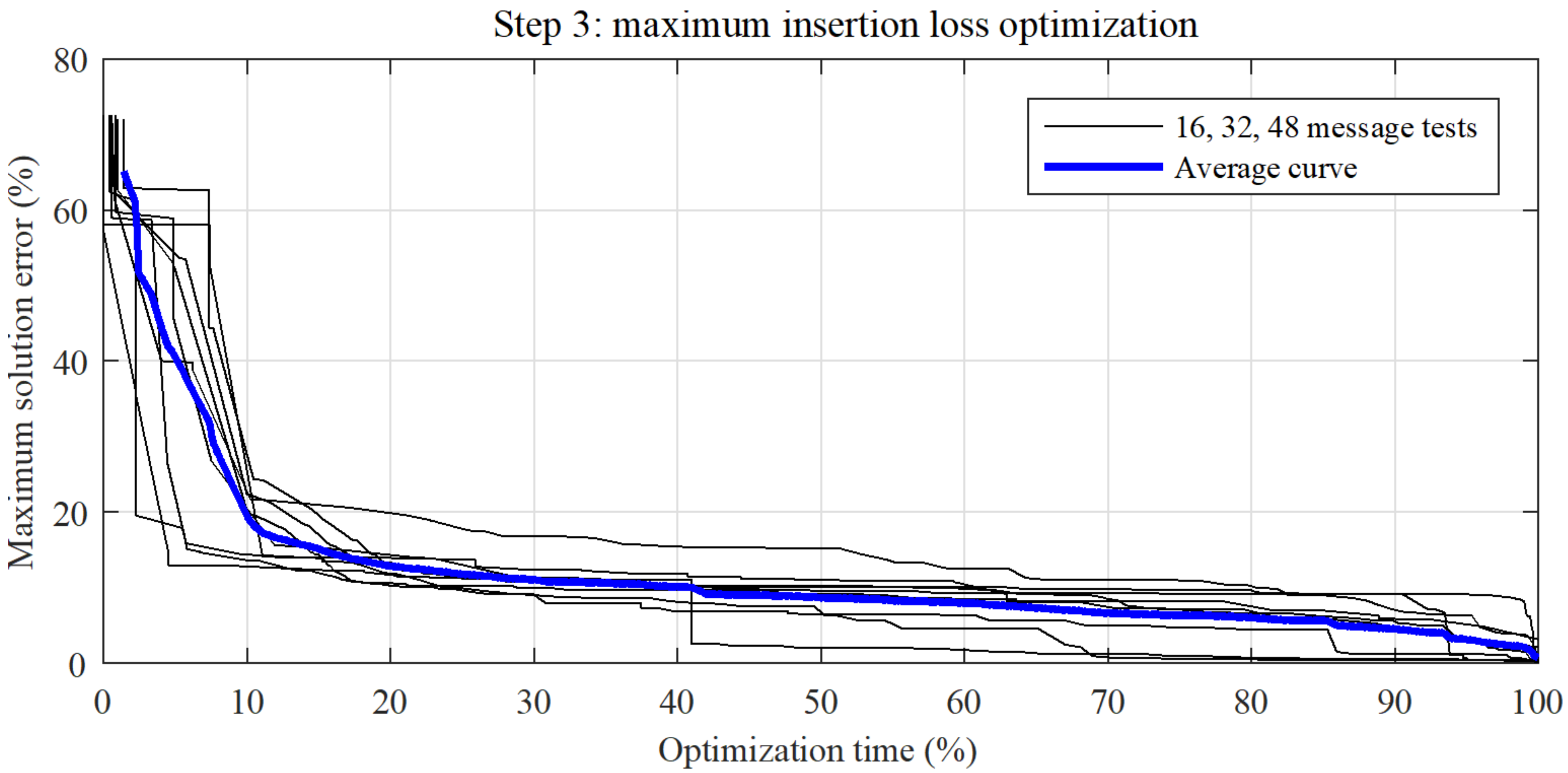


Any questions?

Fast solution convergence



Fast solution convergence



Advantages of application-specific design