



**HAL**  
open science

## Data traffic management in a reconfigurable Network-on-Chip for Dynamic Neural Networks

Mohamed Amine Zhiri, Hana Krichene, Chiara Sandionigi, Sébastien  
Pillement

► **To cite this version:**

Mohamed Amine Zhiri, Hana Krichene, Chiara Sandionigi, Sébastien Pillement. Data traffic management in a reconfigurable Network-on-Chip for Dynamic Neural Networks. Journée Nationale GDR SoC2 - 17ème Colloque du GDR SoC2, Jun 2023, Lyon, France. , 2023. hal-04119393v2

**HAL Id: hal-04119393**

**<https://hal.science/hal-04119393v2>**

Submitted on 22 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mohamed Amine Zhiri<sup>1,3</sup>, Hana Krichene<sup>1</sup>, Chiara Sandionigi<sup>2</sup>, Sébastien Pillement<sup>3</sup>

<sup>1</sup> Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France

<sup>2</sup> Université Grenoble Alpes, CEA, LIST, F-3800, Grenoble, France

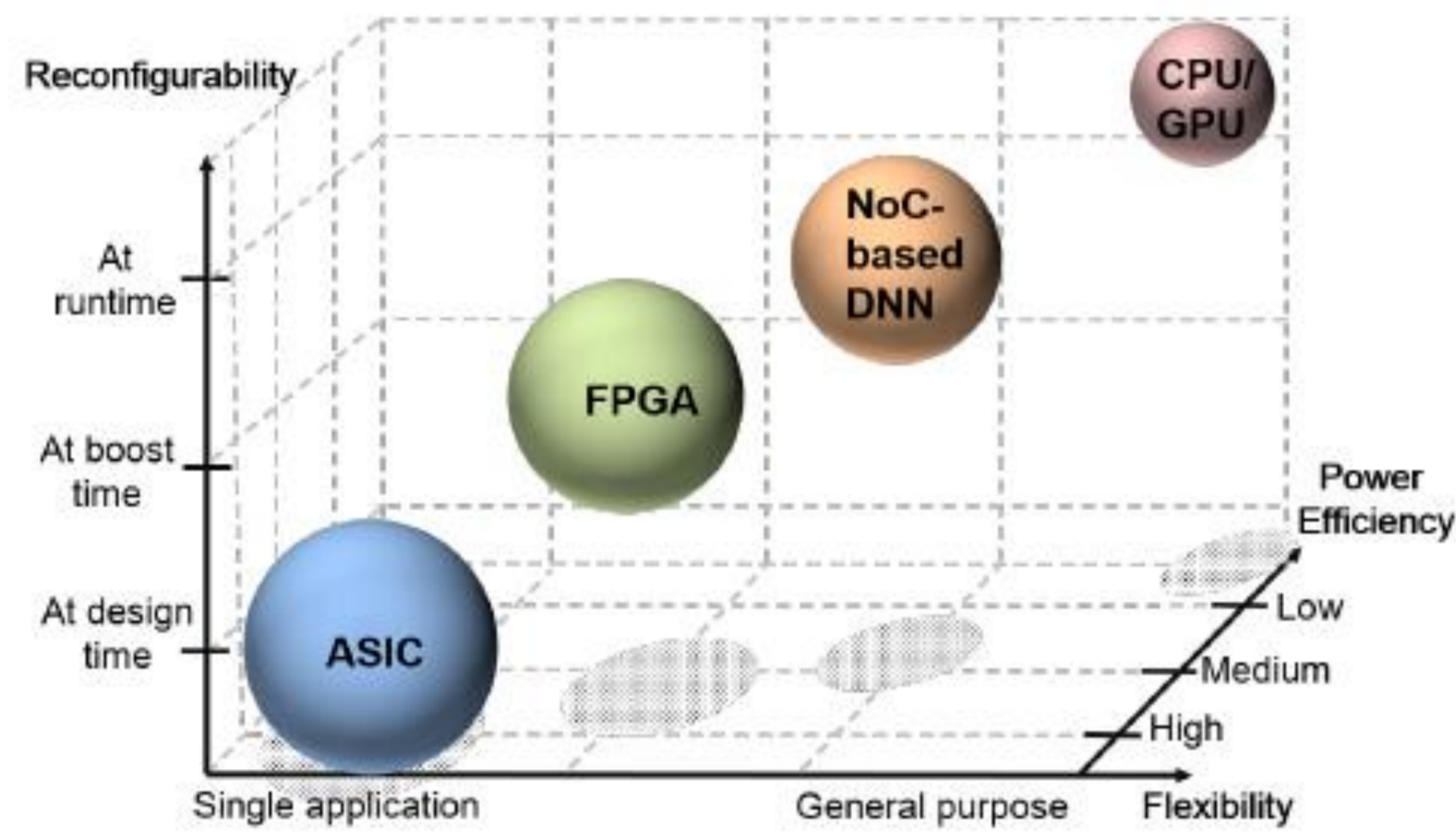
<sup>3</sup> Nantes Université, CNRS, IETR, UMR 6164, F-44000, Nantes, France

Main contact : mohamed-amine.zhiri@cea.fr

## Motivation

### NoC-based AI accelerators

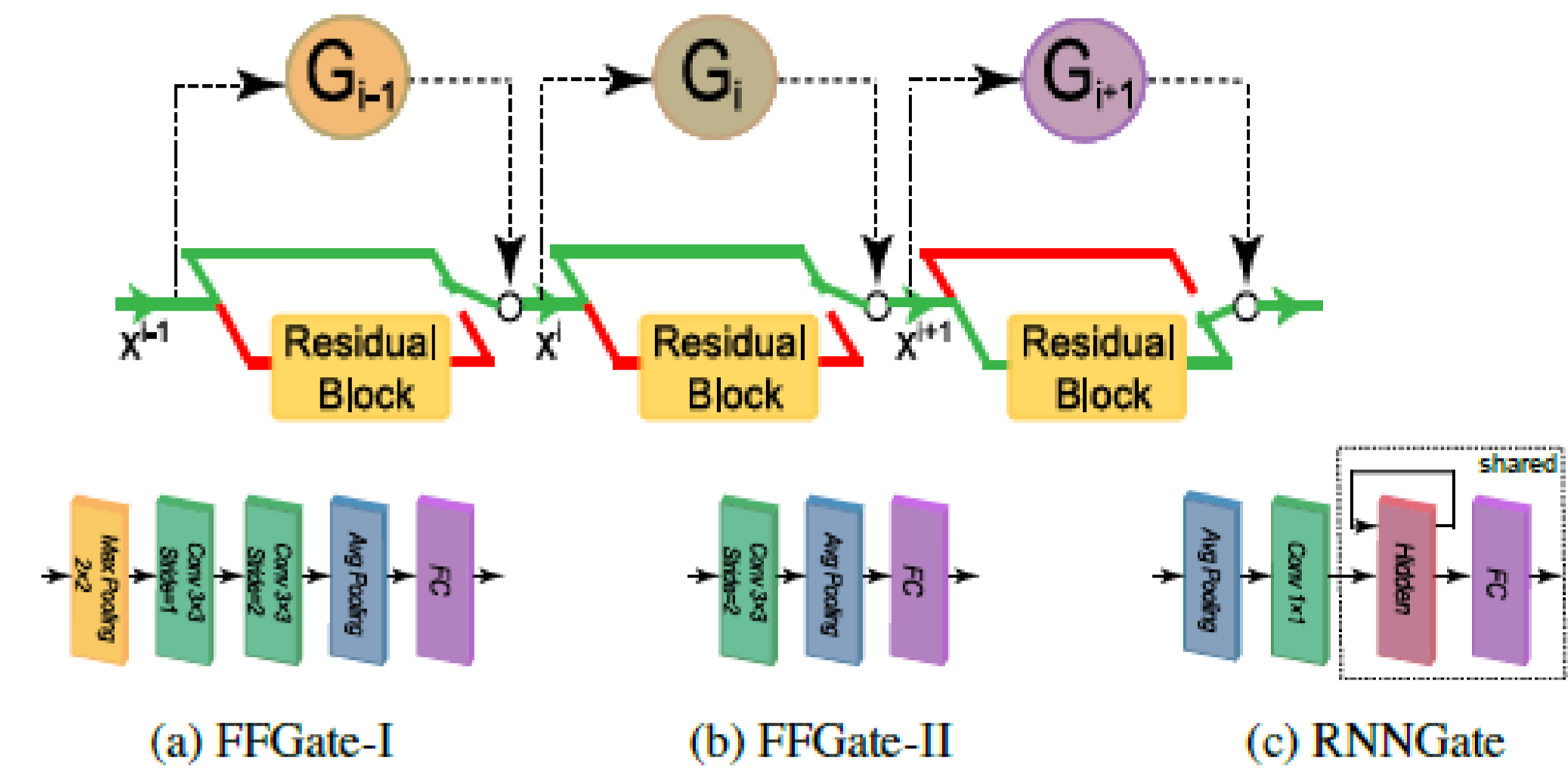
- Rise of AI and dedicated hardware architectures
- NoC-based AI accelerators decouple communication from computation for more flexibility and energy efficiency [1]



Flexibility and reconfigurability of current DNN accelerator design paradigms [1]

### Dynamic Neural Network

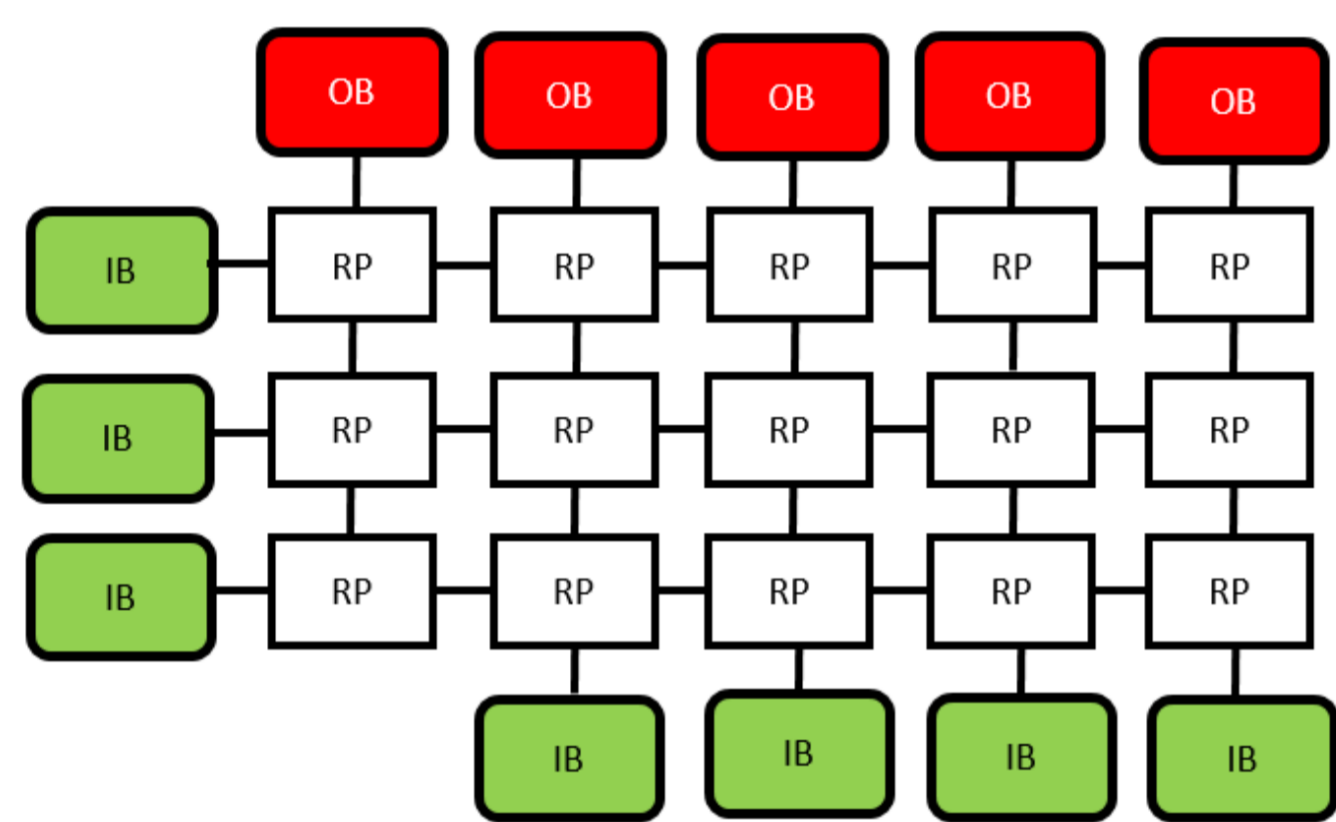
- Dynamic Neural Network : Computational graph or parameters depend on the input [2]
- SkipNet [3] : Layer skipping mechanism



Gating mechanism and Gates design in SkipNet [3]

## Modeling

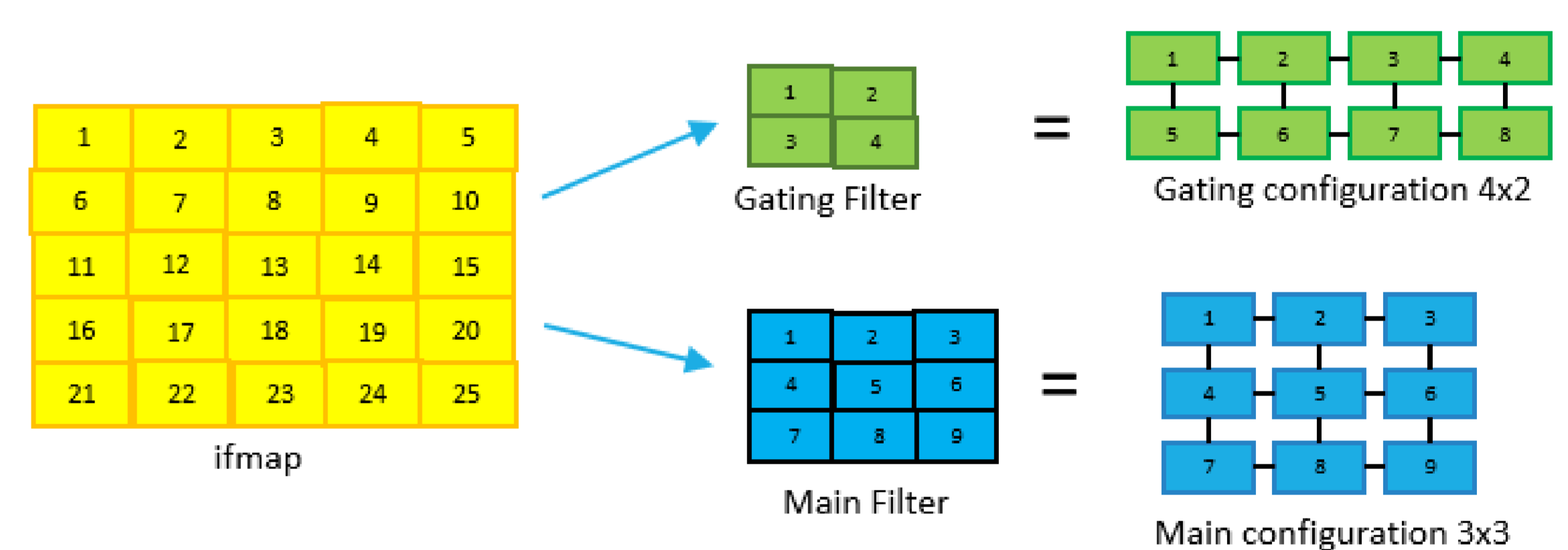
### AINoC based architecture



5x3 configuration of AINoC

- AINoC[4]: reconfigurable NoC
- Row stationary dataflow [5]:
  - Filters: horizontal multicast
  - Input feature map: diagonal multicast
  - Output feature map: vertical accumulation

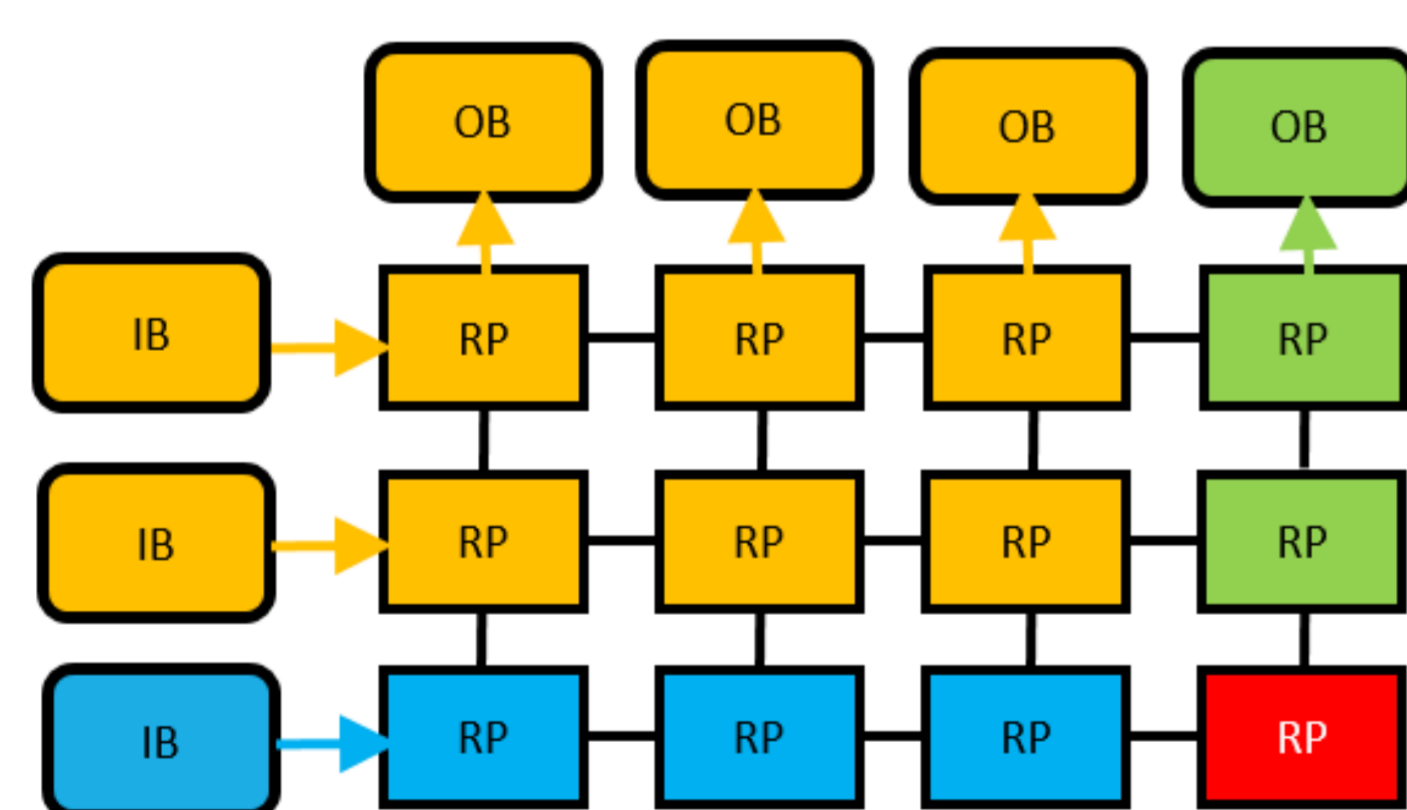
### SkipNet Mapping on the architecture



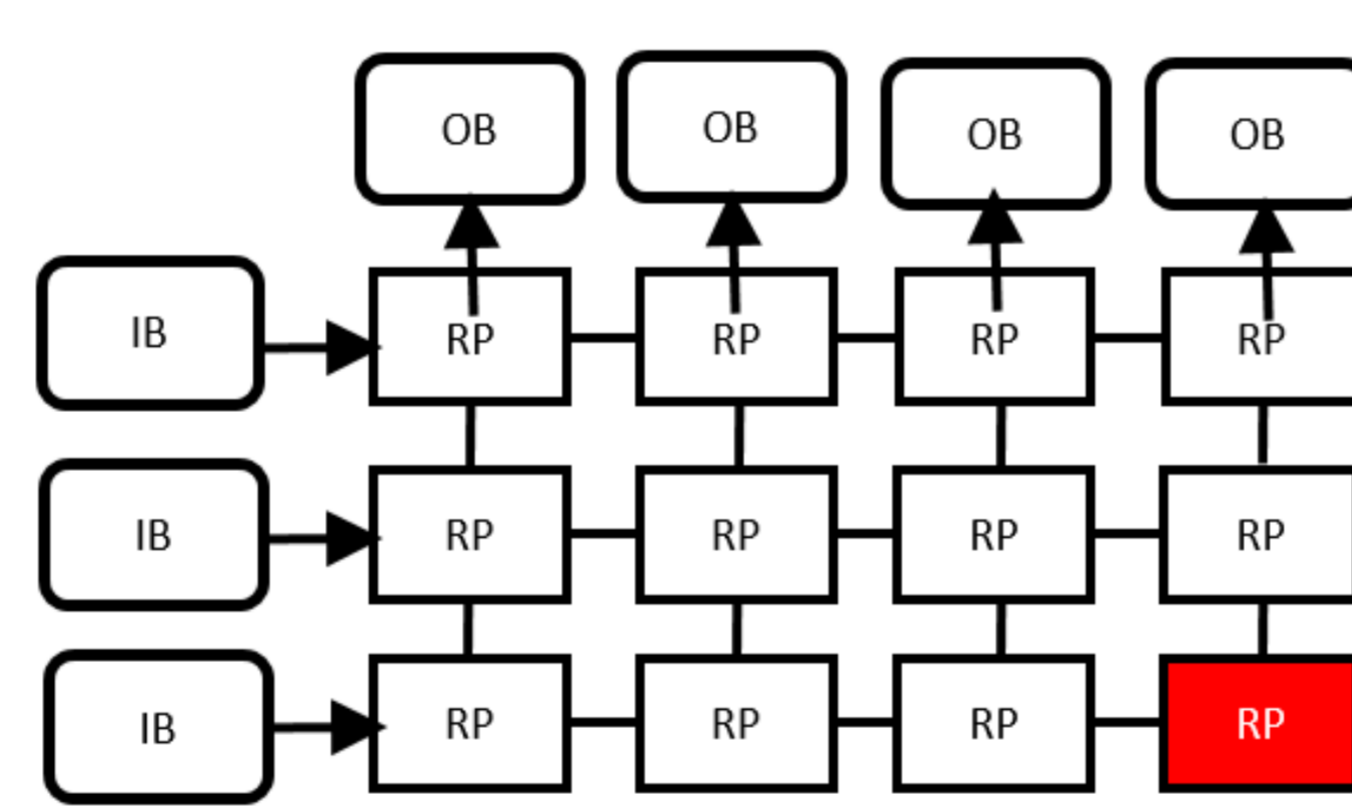
Mapping example of SkipNet over the architecture

## Comparative study

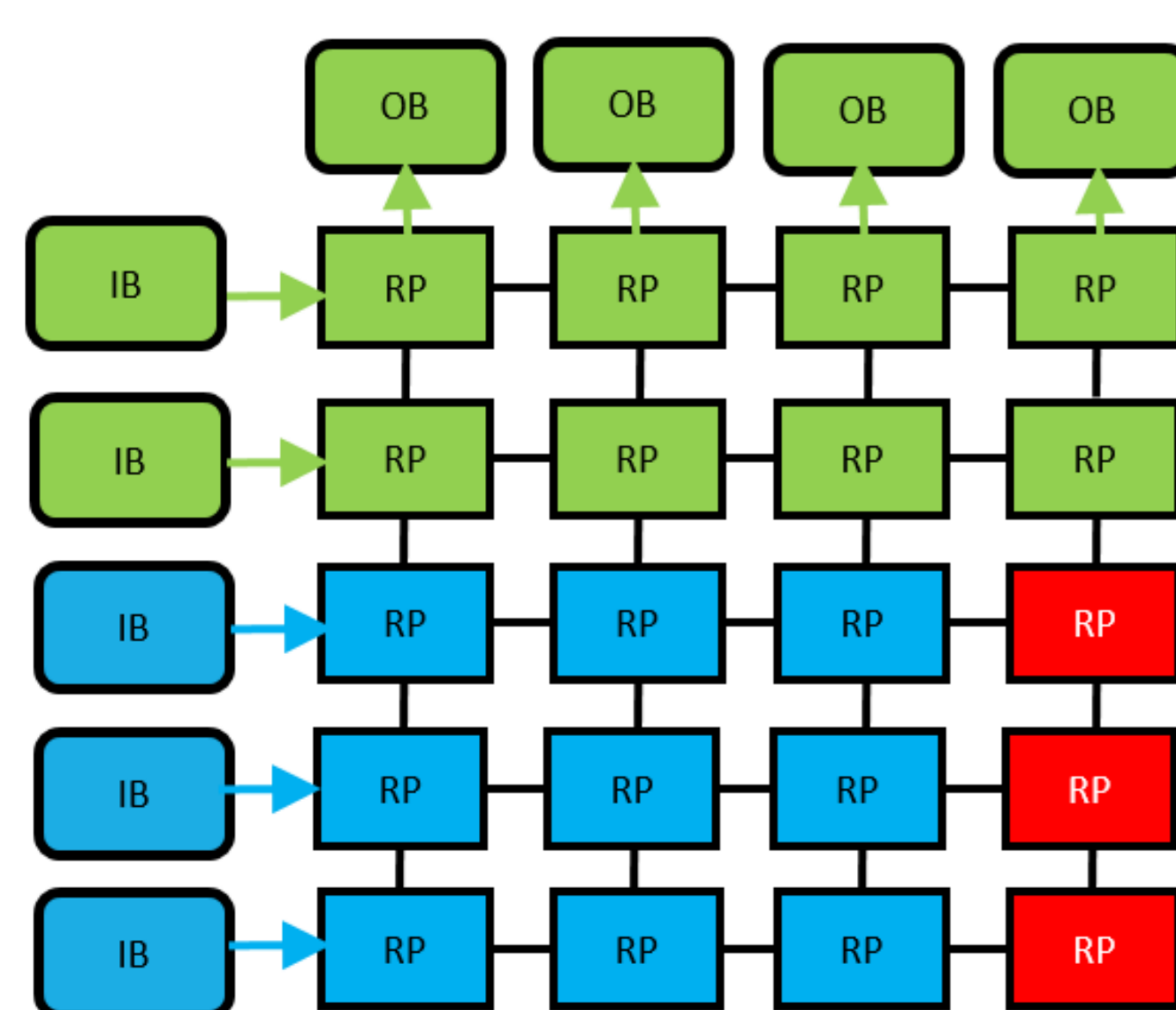
### Mapping techniques



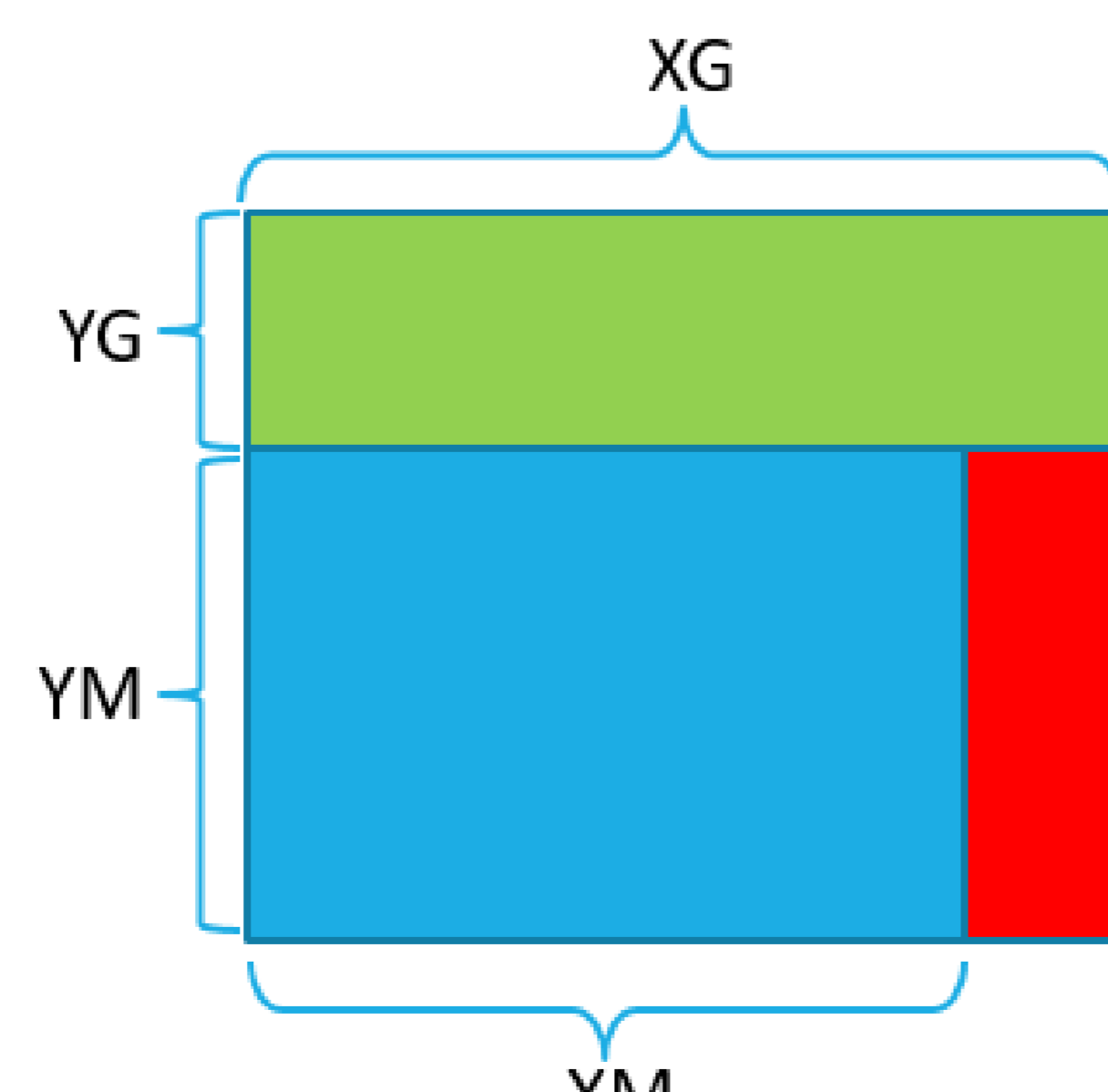
Shared resource + parallel execution



Shared resource + sequential execution



Clustering



Clustering method analysis

■ Gating convolution only ■ Main convolution only ■ Simultaneous gating and main convolutions ■ Non used component

### Latency analysis

	KO	OK
Clustering	XG+YG+XM	YG+max(XM+YM,XG)
Shared parallel	XG+YG+XM	max(XG+YG,XM+YM)
Shared sequential	XG+YG	XG+YG+XM+YM

### Resource analysis

	Non used PE %	Resource size
Clustering	High	Small
Shared parallel	Low	Large
Shared sequential	Low	Small

## References

- [1] Chen, Kun-Chih & Ebrahimi, Masoumeh & Wang, Ting-Yi & Yang, Yuch-Chi. (2019). NoC-based DNN accelerator: a future design paradigm. NOCS '19: Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip.
- [2] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey, 2021
- [3] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. 2018. SkipNet: Learning Dynamic Routing in Convolutional Networks. In Computer Vision –ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Springer-Verlag, Berlin, Heidelberg, 420–436.
- [4] Krichene, H., Prasad, R., Mouhagir, A. (2023). AINoC: New Interconnect for Future DeepNeural Network Accelerators. In: Design and Architecture for Signal and Image Processing, DASIP 2023. Lecture Notes in Computer Science, vol 13879. Springer, Cham.
- [5] Y. -H. Chen, J. Emer and V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, Korea (South), 2016, pp. 367-379, doi: 10.1109/ISCA.2016.40.