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

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

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

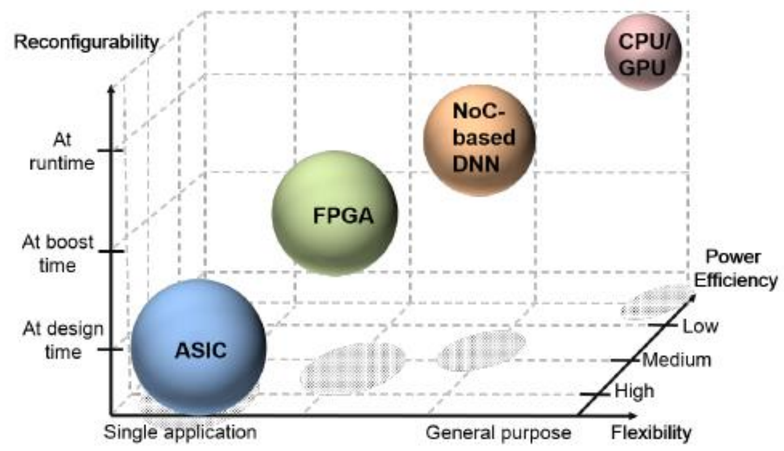
Mohamed Amine Zhiri[1,3], Hana Krichene[1], Chiara Sandionigi[2], Sébastien Pillement[3]

1 Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France
2 Université Grenoble Alpes, CEA, LIST, F-3800, Grenoble, France
3 Nantes Université, CNRS, IETR, UMR 6164, F-44000, Nantes, France
Main contact : mohamed-amine.zhiri@cea.fr

Motivation

NoC-based AI accelerators

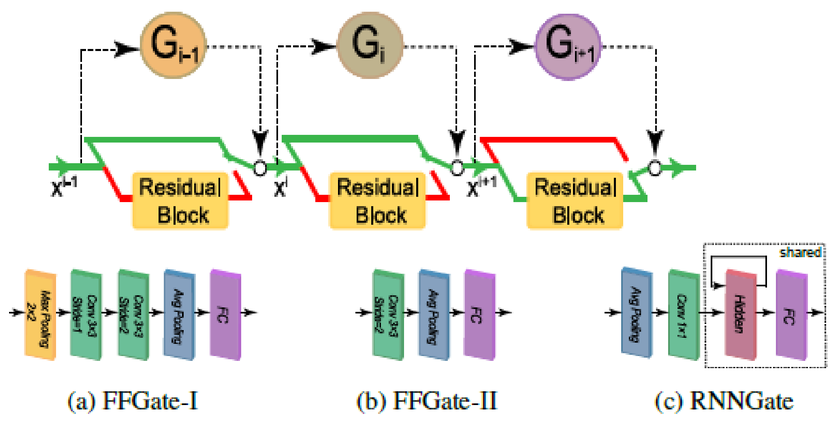
- Rise of AI and dedicated hardware architectures
- NoC-based AI accelerators decouple communication from computation for more flexibility and energy efficicency [1]



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

Dynamic Neural Network

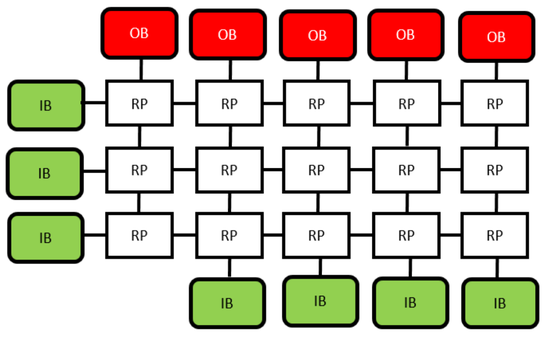
- Dynamic Neural Network : Computational graph or parameters depend on the input [2]
- SkipNet [3] : Layer skipping mechanism



Gating mechanism and Gates design in SkipNet [3]

Modeling

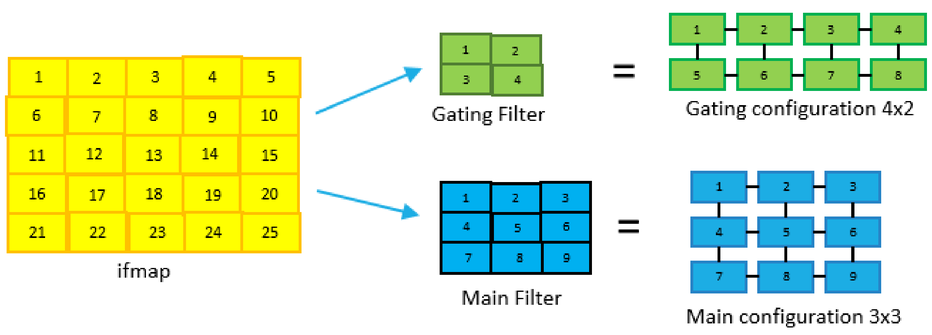
AINoC based architecture



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

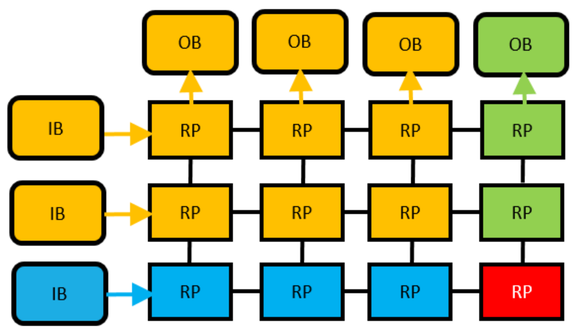
SkipNet Mapping on the architecture



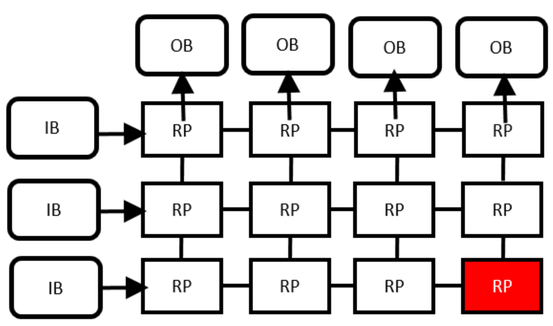
Mapping example of SkipNet over the architecture

Comparative study

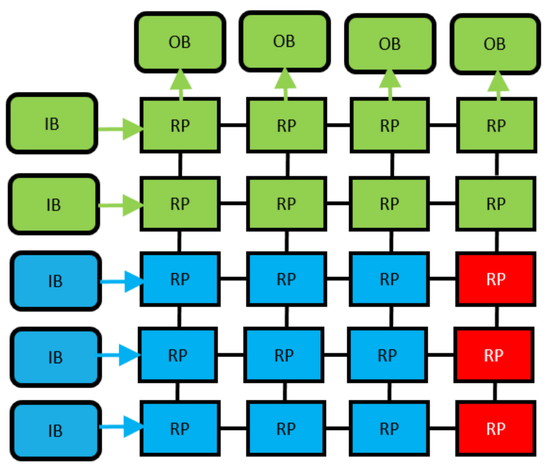
Mapping techniques



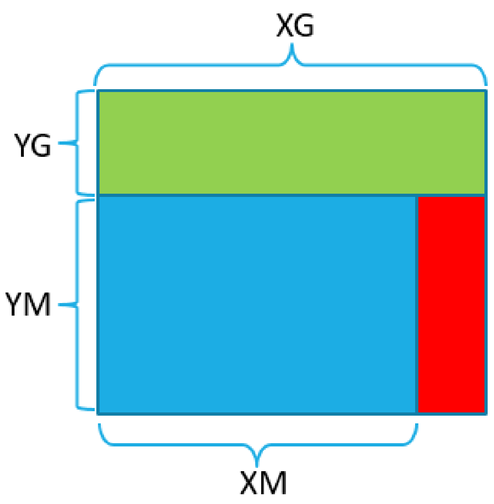
Shared resource + parallel execution



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 YulinWang. 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, September8-14, 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.