Multi-mode Network-on-Chip for AI dataflow accelerators

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

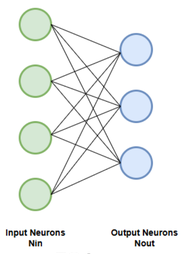
Multi-mode Network-on-Chip for AI dataflow accelerators

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 École doctorale : MaSTIC N°641 Démarrage : Octobre 2022

Motivation



FC layer

$$Y^{Nout} = W^{Nout \times Nin} \times X^{Nin}$$

X^{Nin} : Input neurons vector

$W^{Nout \times Nin}$: Weight matrix

Y^{Nout}: Output neurons vector

- Found in many AI models
  - FFN blocks of transformers
  - Classification layers of CNNs
- **Bandwidth bound**
- Propose High Throughput Network-on-Chip (HT-NoC) : a reconfigurable NoC to accelerate FC layers execution.
- HT-NoC shows promising results for some CONV layers also.

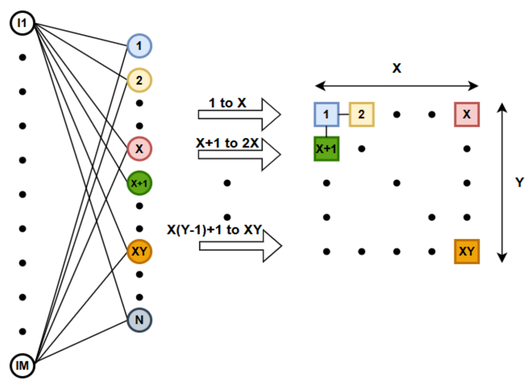
Related works

Family	Method	Shortcomings
Resource duplication	Wider links[1]	Large area overhead
	Muliple planes[2]	
	Multiple routers[3]	
Resource reuse	Channel reuse [4]	Limited impact on FC layer traffic pattern
	Buffer sharing [5]	

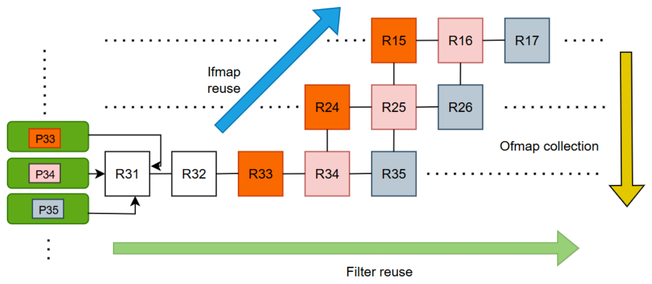
Techniques for Bandwidth Optimization and Latency Reduction in NoCs

HT-NoC Strategy → **Add limited amount of logic to fully reuse all available router resources**

Layer Mapping



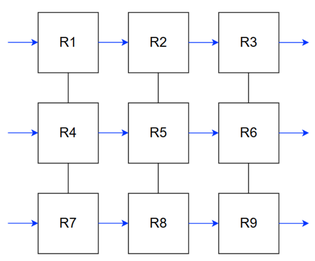
FC layer mapping in 2D mesh



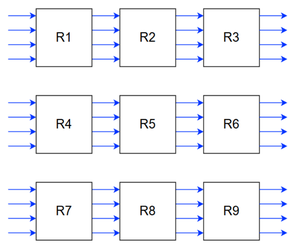
CONV layer mapping in 2D mesh

HT-NoC

❖ Operating modes

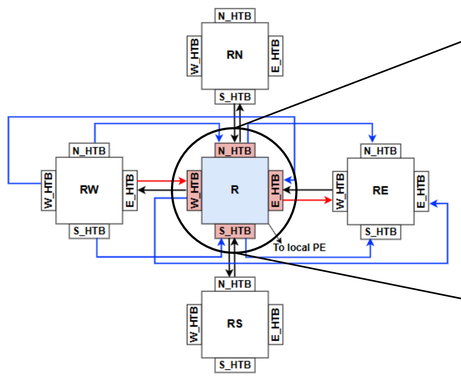


Normal mode

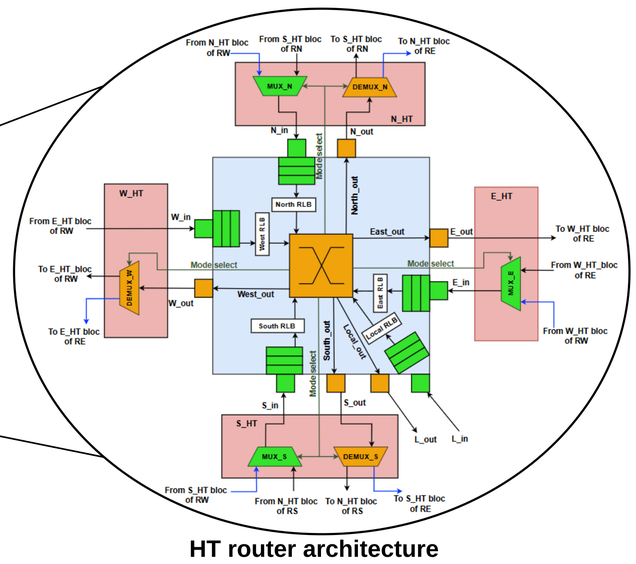


HT mode

❖ HT router



HT router connections



HT router architecture

Higher bandwidth utilization

Higher resources utilization

- Normal mode : bandwidth evenly distributed across all output ports
- HT mode : all available bandwidth is allocated for West-to-East communication
- HT blocks switch from one operating mode to another
- HT mode : router ports and internal resources uitlization are maximized

❖ Communication phases

Phase	FC layers		CONV layers	
	Comm. Type	Mode	Comm. type	Mode
Input neurons propagation	Broadcast	Normal	Diagonal multicast	HT
Weight propagation	Unicast	HT	Horizontal multicast	Normal
Computation	No communication			
Output neurons collection	Unicast	Normal	Unicast	Normal

Execution phases of FC and CONV layers

Evaluation results

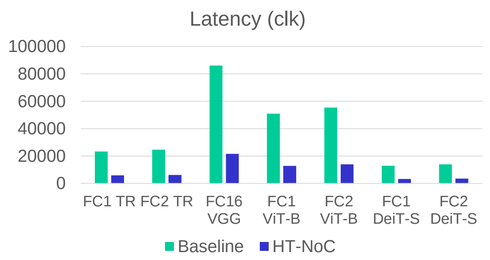
❖ Synthesis

	Baseline router	HT router	Baseline NoC (12x12)	HT-NoC (12x12)
CLB LUTs	2064	2137	290810	339549
Flip Flops	1142	1146	158620	164996

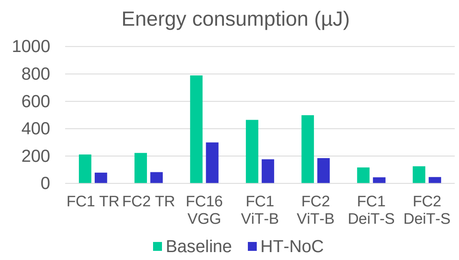
Synthesis results on Versal XCVC1902 FPGA at 100MHz

- HT-NoC : +17% LUT usage compared to mesh NoC (12x12 configuration)

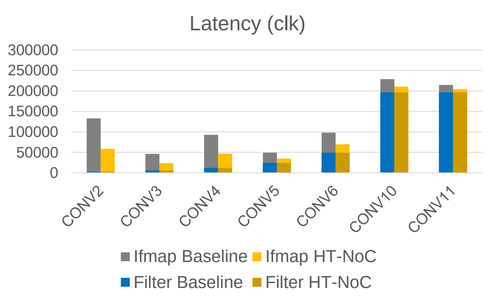
❖ Data propagation results



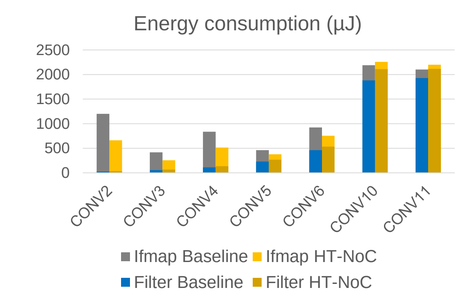
FC latency in clock cycles



FC dynamic energy in (µJ)



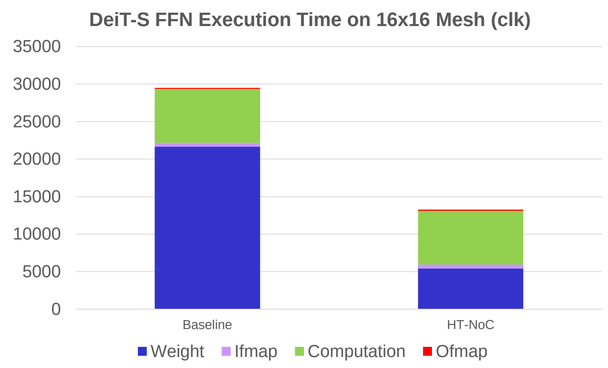
CONV latency in clock cycles

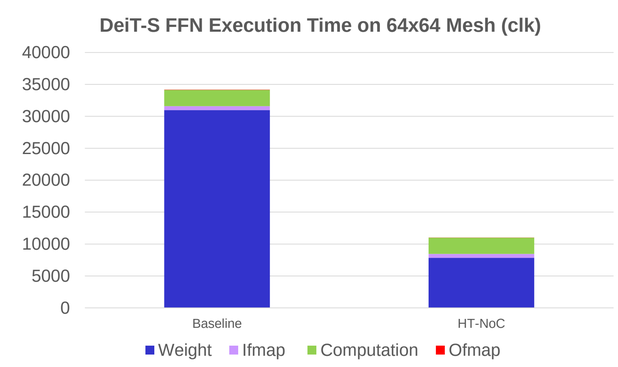


CONV dynamic energy in (µJ)

- Post-implementation simulation
- FC layers : 4x acceleration and 2.75x energy reduction
- CONV layers : overall performance depends on the layer. Better results on early layers

❖ Integration into an AI dataflow accelerator

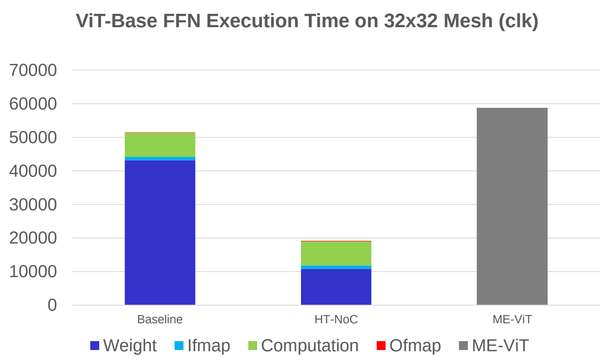




Execution time breakdown of FFN layers of DeiT-S on 16x16 and 64x64 configurations of the accelerator[6]

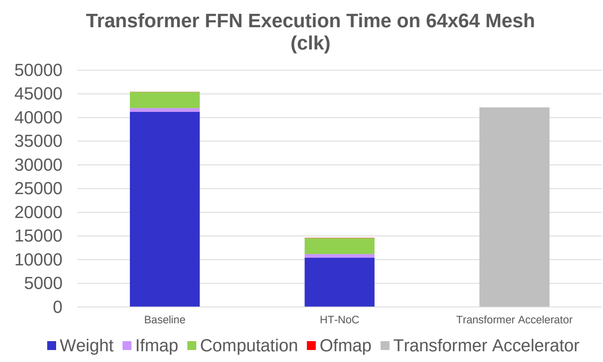
- HT-NoC : acceleration of weight propagation phase => Better overall performance

❖ Comparison with SoA accelerators



Execution time of FFN layers of ViT-Base on baseline NoC, HT-NoC and ME-ViT[7]

2.7x acceleration



Execution time of FFN layers of the transformer model on baseline NoC, HT-NoC and the transformer accelerator[8]

3x acceleration

References

[1] T. Fischer, M. Rogenmoser, M. Cavalcante, F. K. Gürkaynak, and L. Benini: FlooNoC: A Multi-Tb/s Wide NoC for Heterogeneous AXI4 Traffic. IEEE Design & Test, vol. 40, no. 6, pp. 7–17, 2023, doi: 10.1109/MDAT.2023.3306720.
[2] L. Wang, Y. Wang, and X. Wang: An Approximate Multiplane Network-on-Chip. In 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 234–239, 2020, doi: 10.23919/DATE48585.2020.9116377.
[3] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze: Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 2, pp. 292–308, 2019, doi: 10.1109/JETCAS.2019.2910232.
[4] Y.-C. Lan, S.-H. Lo, Y.-C. Lin, Y.-H. Hu, and S.-J. Chen: BiNoC: A bidirectional NoC architecture with dynamic self-reconfigurable channel. In 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip, pp. 266–275, 2009, doi:10.1109/NOCS.2009.5071476.
[5] H. Farrokhbakht, H. Kao, and N. E. Jerger: UBERNoC: Unified Buffer Power-Efficient Router for Network-on-Chip. In Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip, articleno. 1, pp. 1–8, 2019, doi: 10.1145/3313231.3352362.
[6] H. Krichene, R. Prasad, A. Mouhagir (2023). AINoC: New Interconnect for Future Deep Neural Network Accelerators. In: Chavarrías, M., Rodríguez, A. (eds) Design and Architecture for Signal and Image Processing. DASIP 2023. Lecture Notes in Computer Science, vol 13879. Springer, Cham.
[7] K. Marino, P. Zhang, and V. K. Prasanna: ME-ViT: A Single-Load Memory-Efficient FPGA Accelerator for Vision Transformers. In 2023 IEEE 30th International Conference on High Performance Computing, Data, and Analytics (HiPC), 2023, pp. 213-223, doi: 10.1109/HiPC58850.2023.00039.
[8] S. Lu, M. Wang, S. Liang, J. Lin, and Z. Wang: Hardware Accelerator for Multi-Head Attention and Position-Wise Feed-Forward in the Transformer. In 2020 IEEE 33rd International System-on-Chip Conference (SOCC), 2020, pp. 84-89, doi: https://doi.org/10.1109/SOCC49529.2020.9524802.