

PACE: A Scalable and Energy Efficient CGRA in a RISC-V SoC for Edge Computing Applications

Vishnu P. Nambiar¹, Yi Sheng Chong¹, Thilini Kaushalya Bandara², Dhananjaya Wijerathne², Zhaoying Li², Rohan Juneja², Li-Shiuan Peh², Tulika Mitra², Anh Tuan Do¹

¹Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore

²School of Computing, National University of Singapore (NUS), Singapore

Introduction

- Coarse-grained reconfigurable arrays (CGRAs) deliver high energy efficiency while maintaining the programmability advantages.
- CGRA is the ideal candidate for efficiently handling loop kernels, which allows it to offload repetitive looping functions such as vector multiplication or hashing algorithms from CPUs.
- It relies on a compiler to convert a given workload into a data flow graph (DFG) which is then mapped onto the hardware in a manner that achieves the highest possible energy efficiency.

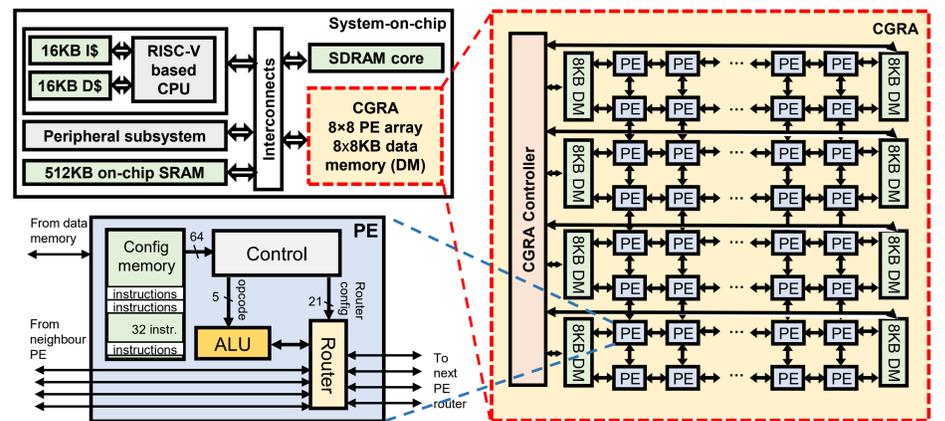


Fig. 1 PACE CGRA integrated in a RISC-V system-on-chip (SoC)

Bypass enabled router

The proposed CGRA enables bypass path in the PE router to enable single-cycle multi-hop capability.

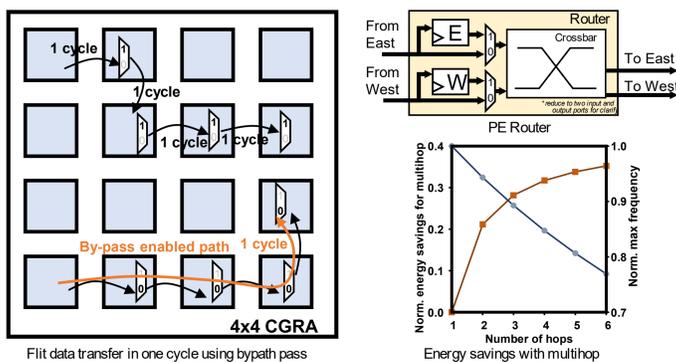


Fig. 2 PE router with bypass pass for single-cycle multi-hop data

Dynamic clock gating in PE

The proposed CGRA has dynamic and static clock gating to suspend idle PEs for reducing power consumption.

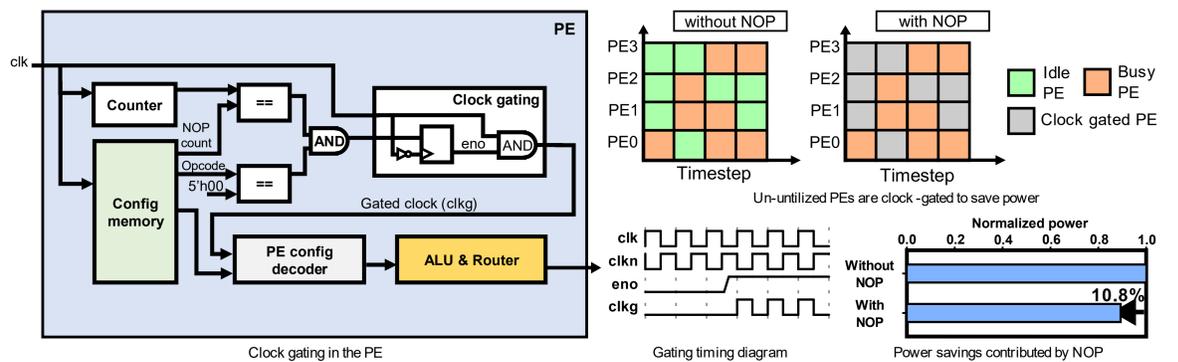


Fig. 3 Dynamic clock gating for power savings

Software toolchain

An end-to-end toolchain (written in Python and C++) is developed to map various applications onto our CGRA.

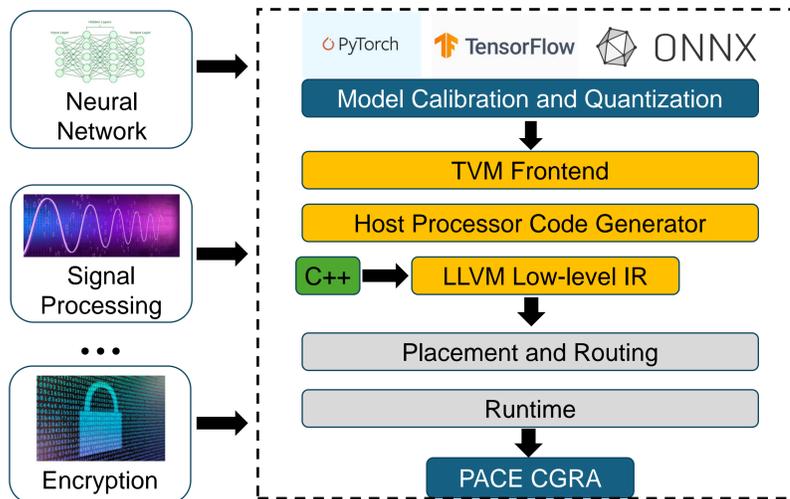


Fig. 5 End-to-end compiler toolchain

Demonstration: The microspeech application

The proposed CGRA executes the convolution and fully-connected layers for wake word detection.

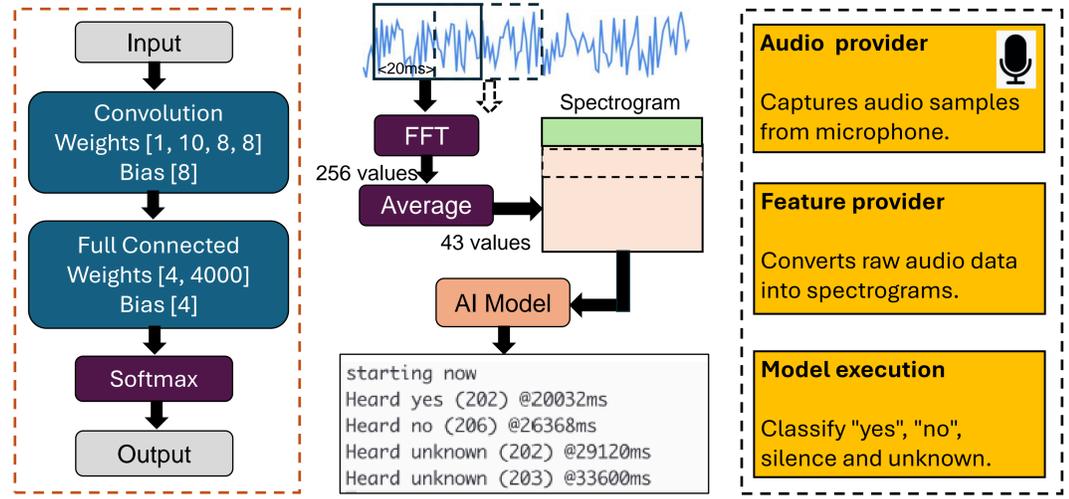


Fig. 6 Microspeech neural network model and execution flow

Performance evaluation

The proposed CGRA delivers a peak efficiency of 360 GOPS/W, which is 1.2 to 4.6 times higher than the state-of-the-art.

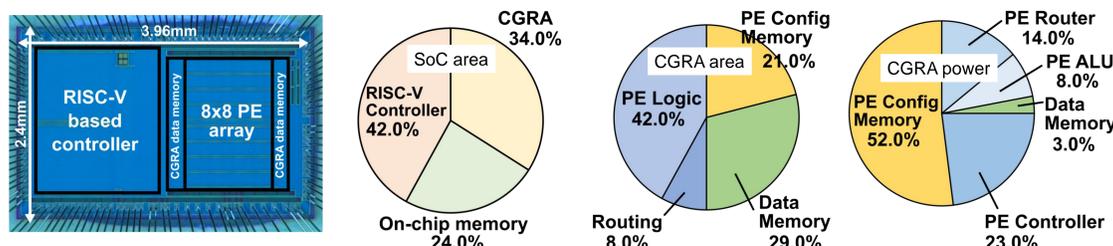
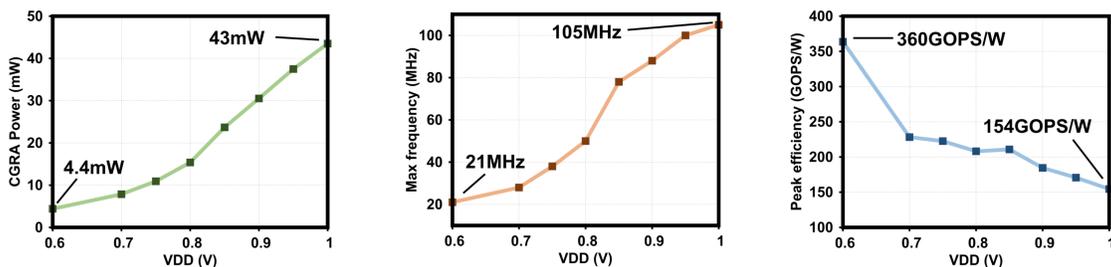


Fig. 7 CGRA efficiency across VDD, chip micrograph, area and power breakdown of the SoC and CGRA

	Amber [2]	SSCL [4]	ISSCC [3]	TVLSI [5]	Hycube [6]	JSSC [7]	This work
Year	2022	2020	2019	2018	2019	2020	2023
Tech (nm)	16	28	22	55	40	28	40
CGRA area (mm ²)	20.1	3.9	4.9	5.19	2.87	4.80	3.02
#PEs	384	120	15	30	16	64	64
Voltage (V)	1.29	0.9	0.8	N/A	1.1	0.9	1.0
Freq (MHz)	955	89	36	450	853	800	100
Throughput (GOPS)	367 (INT16)	14.1	145	77.4	6.48	0.88	64
Power (mW)	N/A	45.9	N/A	1526	72	537	43@1V
Efficiency (GOPS/W)	538@1.29V (INT16)	307@0.9V	978@0.48V	50.8	90	196	154@1V, 360@0.6V
Memory	4500KB	234KB	690KB	54KB	7KB	320KB	80KB
Norm. area (mm ²)	50	5.5	3.2	3.67	2.87	6.86	3.02
Norm. area per PE (mm ²)	0.13	0.05	0.21	0.12	0.18	0.11	0.05
Norm. efficiency (GOPS/W)	86	150	296	96	90	96	360

Norm. efficiency = efficiency × (node_{40nm})²

Norm. area = area × (40nm/node)

Fig. 8 Performance comparison with state-of-the-art

Acknowledgements This research is supported by the National Research Foundation, Singapore (CRP23-2019-0003).