Spatial Accelerator Architecture for Low Power Embedded Applications

COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

Thomas J. Repetti[1], João P. Cerqueira[2], Martha A. Kim[1], Mingoo Seok[2]
Department of Computer Science[1]
Department of Electrical Engineering[2]

Flexible Application Acceleration in Embedded Systems

Embedded systems typically rely on specialized ASIC accelerator hardware to meet demanding performance and power targets. While optimal for their designed purpose, the inherent lack of flexibility and engineering cost overheads motivate field programmable solutions, the most widely deployed of which area FPGAs. FPGAs suffer from their own drawbacks, however, in terms of performance and efficiency degradation as well as being an unintuitive target for software developers and still requiring considerable hardware expertise to use. We present a prototype for a compiler-targetable low power spatial accelerator that approaches ASIC performance on a variety of workloads while maintaining a familiar programming interface.

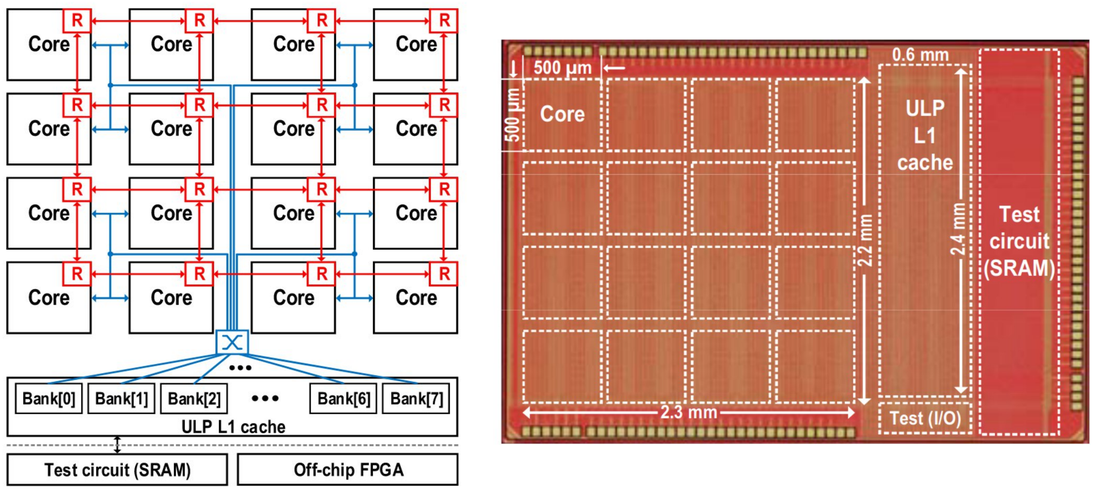


Figure 1. High-level architecture showing spatial (red) and memory subsystem (blue) networks-on-chip and actual chip micrograph.

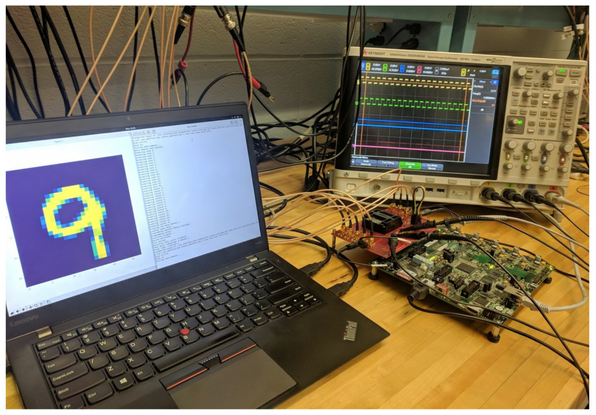
		Kilocore [1]	3D-MAPS [14]	Centip3De [15]	Catena (this work)		
HW. Spec.	Process tech.	32-nm PD-SOI	130-nm CMOS	130-nm CMOS	65-nm LP CMOS		
	Architecture	Spatial array architecture	3D multiprocessor system	3D multiprocessor system	Spatial array architecture		
	Number of cores	1000	64	64	16		
	Area	59.97 mm²	2 x 25 mm² (2 layers)	2 x 63.3 mm² (2 layers)	6.50 mm²		
	Norm. area (65-nm)	247.43 mm²	2 x 6.25 mm² (2 layers)	2 x 15.825 mm² (2 layers)	6.50 mm²		
	Supply voltage	0.76 V – 1.10 V	0.90 V – 1.90 V	0.65 V – 1.50 V	V_{DDC} = 0.37 V – 1.15 V, V_{DDL} = 0.21 V – 0.65 V, V_{DDH} = 0.50 V – 1.20 V		
	Energy/cycle (LP)	15.15 nJ *, V_{DD} = 0.76 V	2.6 nJ, V_{DD} = 0.90 V	5.07 nJ, V_{DD} = 0.65 V	227 pJ, V_{DDC}=0.54 V, V_{DDL}=0.21 V, V_{DDH}=0.77 V		
Workload		FFT (simulation) †	Matrix multiplication	Instruction mix	FFT	FIR	MNIST
	Algorithm	Fixed-point, complex, 4096-pt (12 transforms)	Cannon's algorithm for distributed matrix mult.	Not reported	Fixed-point, complex, 256-pt	Fixed-point, 11-tap, low pass	784x120x 120x120x10
	Bitwidth	16-bit complex	32-bit inputs	32-bit inputs	16-bit complex	16-bit input/coeff	8-bit weight
	Energy consump.	11.8 nJ/sample	13.68 nJ /cycle	5.07 nJ/cycle	9.02 nJ/sample	2.10 nJ/sample	6.95 µJ/pred
	Norm. energy ‡	6.15 µJ/FFT	6.84 nJ/cycle	2.53 nJ/cycle	2.31 µJ/FFT	2.10 nJ/sample	6.95 µJ/pred
	IPC/core	0.23	0.32	Not reported	0.49	0.88	0.61
	Number of cycles	8.85k	Not reported	Not reported	7.70k	4.39k	14.4k
	Throughput	824 Msample/s	Not reported	Not reported	21.7 Ksample/s	133.2 Ksample/s	42 pred/s
	Supply voltage	1.10 V	1.50 V	V_{DDCore} = 0.65 V, $V_{DDCache}$ = 0.80 V	V_{DDC} = 0.54 V, V_{DDL} = 0.21 V, V_{DDH} = 0.77 V		
	Clock frequency	Not reported	277 MHz	10 MHz	1.0 MHz		

* Estimated based on the plots shown in the paper [1].
† The results for FFT and other applications presented in [1] are estimated using simulations while the results shown for Catena are actual measurements.
‡ Norm. energy per FFT = (256 / FFT size) * (Energy per FFT) / {(Technology / 65 nm) * [(2 / 3) * (Wordlength / 16) + (1 / 3) * (Wordlength / 16)²]}.
For the remaining workloads, the energy is normalized to the technology node.

Figure 2. Comparison to state of the art spatial accelerator work.

Diverse Set of Workloads

After designing and fabricating the prototype chip, we tested a wide variety of workloads on our prototype chip: AES-128 CTR (cryptography), FFT and FIR (DSP) and MNIST-trained MLP DNN (Machine Learning). Runtime power consumption and performance measurements for an ultra-low power operating point demonstrated efficiency approaching ASIC performance across several of these workloads and superior performance compared to other state-of-the-art spatial architectures (Figure 2).



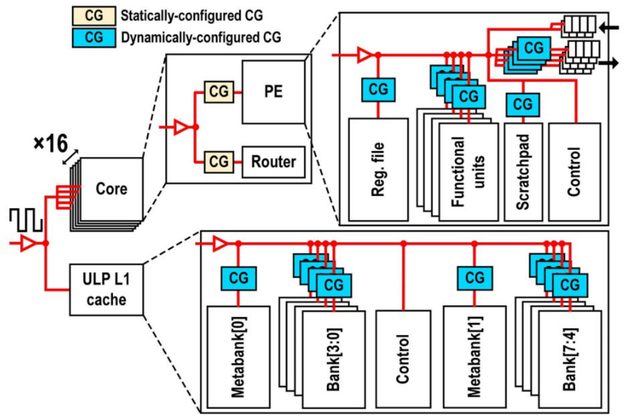


Figure 3. Live evaluation of MNIST workload and the fine grain clock-gating architecture that dramatically reduced power consumption.

Ongoing Work

Current research related to this class of architecture aims to improve programmability and support the OpenCL programming standard for heterogeneous programmable systems.

Acknowledgements

We gratefully acknowledge the support of C-FAR (a program center sponsored by DARPA and the Semiconductor Research Corporation), and a 2017 Qualcomm Innovation Fellowship.

References

1. Catena: A Near-Threshold, Sub-0.4-mW, 16-Core Programmable Spatial Array Accelerator for the Ultralow-Power Mobile and Embedded Internet of Things, IEEE Journal of Solid State Circuits, Aug., 2020.
2. Catena: A 0.5-V Sub-0.4-mW 16-Core Spatial Array Accelerator for Mobile and Embedded Computing, Symposium on VLSI Circuits, 2019.
3. Pipelining a Triggered Processing Element, IEEE/ACM International Symposium on Microarchitecture, 2017.