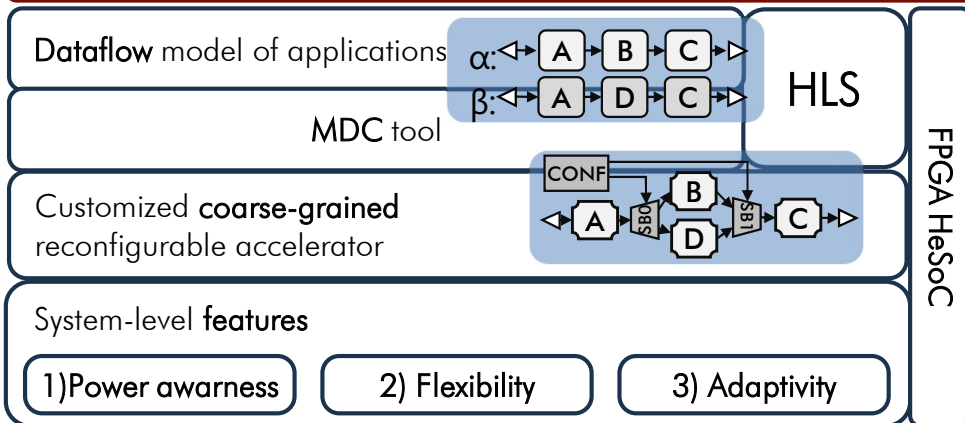


Design methodologies and architectures for application-specific coarse-grain reconfigurable accelerators



Ph.D.: Francesco Ratto (francesco.ratto@unica.it) – Università degli Studi di Cagliari (IT)
Supervisors: Prof. Luigi Raffo and Prof. Francesca Palumbo

Motivation and Goals

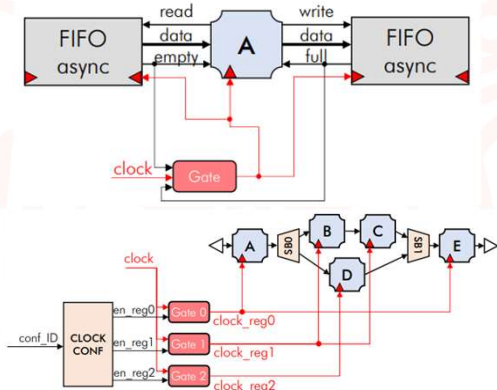


To take full advantage of the efficiency and performance of FPGA HeSoCs, effective programming models and tools are needed. This means going beyond the limits of the available HLS tools. These tools are effective in compiling a single functionality, but they still struggle to manage system-level design features. The research described focuses on: — power awareness, intended as the ability to have a clear view of the impact of design choices on the final power consumption; — flexibility, intended as the capability of the system to execute various/different workloads; — adaptivity, the ability to change the execution mode to adapt to internal or external changes.

Contributions

1) Mutual impact between clock gating (CG) and HLS in coarse-grained reconfigurable systems

Explored the combination of CG granularity and adopted HLS tool.

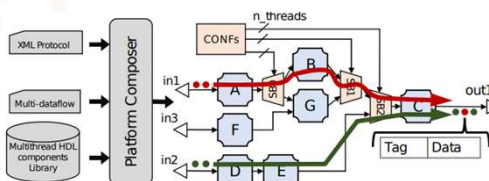


Evaluation on HEVC filter

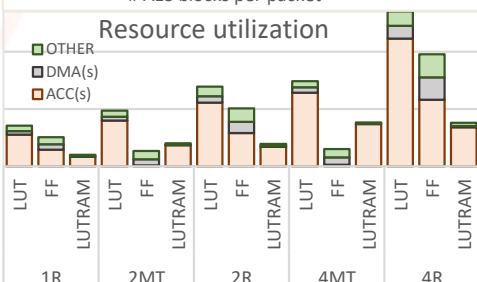
HLS	design	Dynamic power [mW]		
		8 taps	5 taps	3 taps
CAPH	base	132	104	75
	Actor	0%	-5%	-8%
	Region	.8%	-3%	-8%
	Multi	.8%	-4%	-8%
STRATUS	base	107	80	60
	Actor	0%	-6%	-15%
	Region	2%	-5%	-16%
	Multi	0%	-6%	-18%
VIVADO	base	128	99	70
	Actor	-6%	-8%	-11%
	Region	-2%	-6%	-11%
	Multi	-5%	-6%	-10%

2) Multi-threading and multi-tasking combination through tagged-dataflow modelling

Trade-off between a single coprocessor and a set of parallel replicas. Design automation through MDC tool and complete system integration.

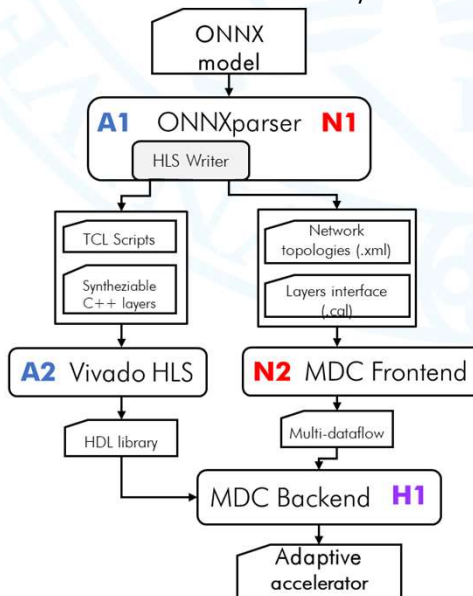


Evaluation on AES encryption



3) Runtime adaptive acceleration of CNNs through non-function reconfigurability

ONNX-to-HW toolchain that exploits an HLS template architecture for CNN layers.



The goal is to achieve adaptivity through accuracy vs energy consumption trade. Integration with QONNX is ongoing.

- Ratto, F., Fanni, T., Raffo, L., & Sau, C. (2021). Mutual impact between clock gating and high level synthesis in reconfigurable hardware accelerators. *Electronics*, 10(1), 73.

- Ratto, F., Esposito, S., Sau, C., Raffo, L., & Palumbo, F. (2022). Multithread Accelerators on FPGAs: A Dataflow-Based Approach. In *Proceedings of PARMA-DITAM Workshop 2022*, co-located with HPEAC, Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

- Ratto, F., Máinez, Á. P., Sau, C., Meloni, P., & Palumbo, F. (2023). An Automated Design Flow for Adaptive Neural Network Hardware Accelerators. *Journal of Signal Processing Systems*.

- Manca, F., & Ratto, F. (2023). ONNX-to-Hardware Design Flow for the Generation of Adaptive Neural-Network Accelerators on FPGAs. *arXiv preprint arXiv:2309.13321*. In *Proceedings of the Cyber-Physical Systems Workshop 2023 (CPS'23)*.