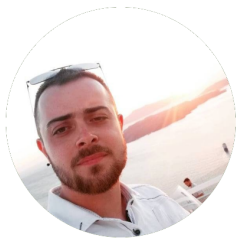
Velos - Virtualizing CGRAs

National Technical University of Athens, CSLab

WRC 2026 – Krakow, Poland







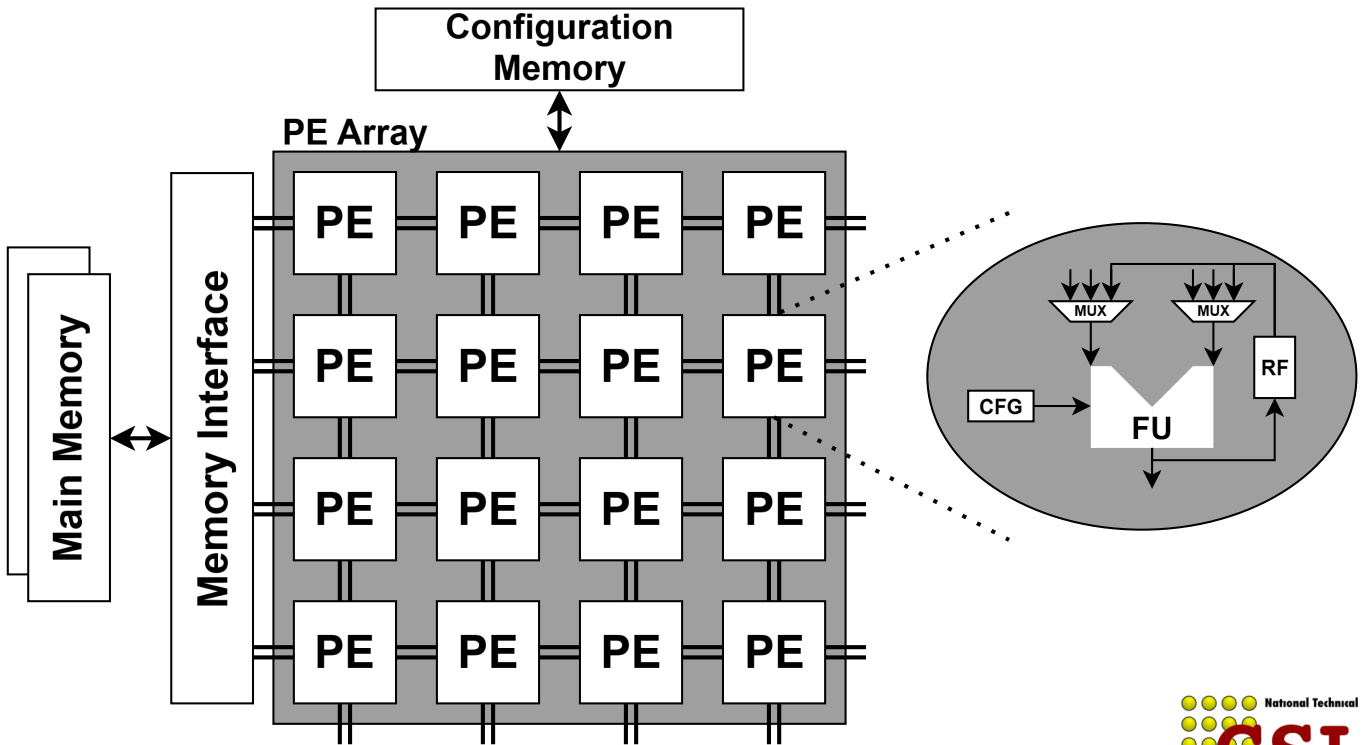
A. Kyriazis P. Miliadis D. Pnevmatikatos

What is Velos?

Velos is an end-to-end attempt to integrate CGRAs into modern compute systems.

Why CGRAs?

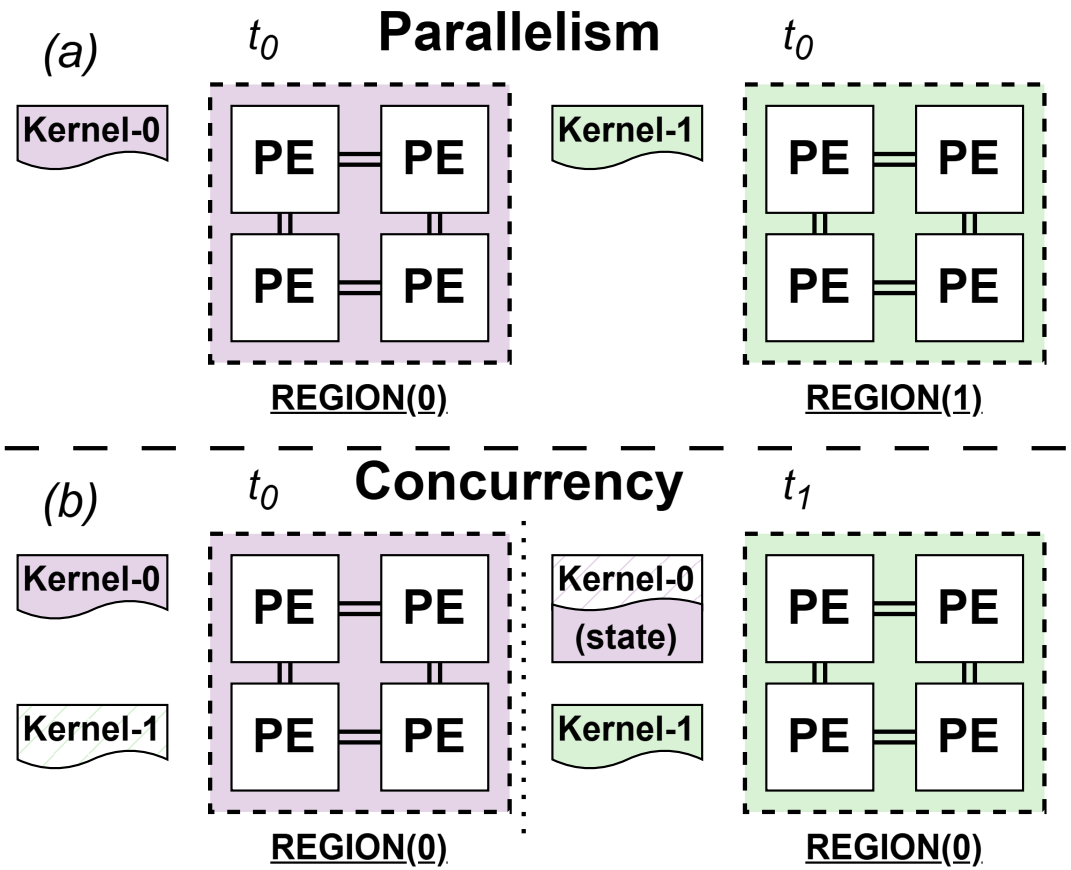
1. CGRAs offer close to ASIC energy efficiency and FPGA-like degree of post-fabrication flexibility.
2. As CGRA scale increases, compilers fall behind [1].
3. CGRA virtualization is quite similar to FPGA virtualization but is much less studied [2].



Desired Execution Model

A CGRA to enable both **parallel** and **concurrent** execution

a. Parallel execution - CGRA as shared platform among kernels
b. Concurrent execution & Context-Switching - traceable execution progress.



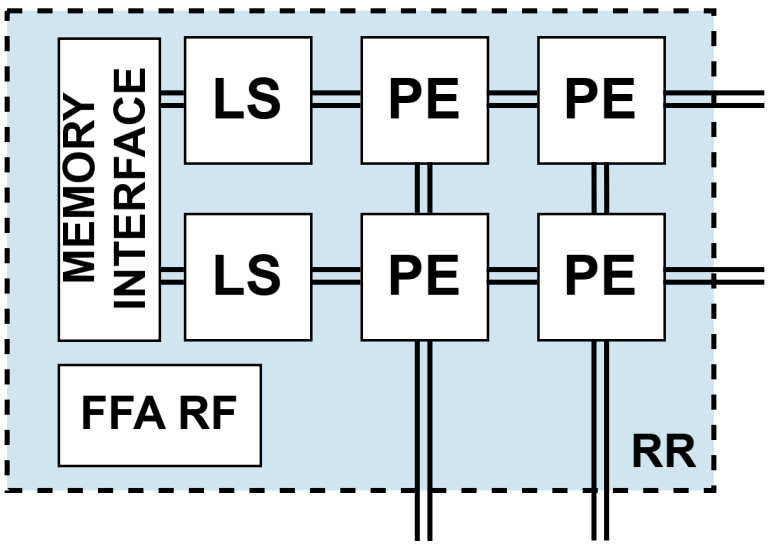
Our Contribution[1]

1. A **virtualization-ready** CGRA architecture

a. Load-Store and Compute PEs arranged on a grid and connected using a 2D elasticized network.

b. Neighboring PEs are grouped into flexible regions that can be dynamically merged with other neighboring.

c. Each regions is equipped with two Load/Store and four Compute PEs.

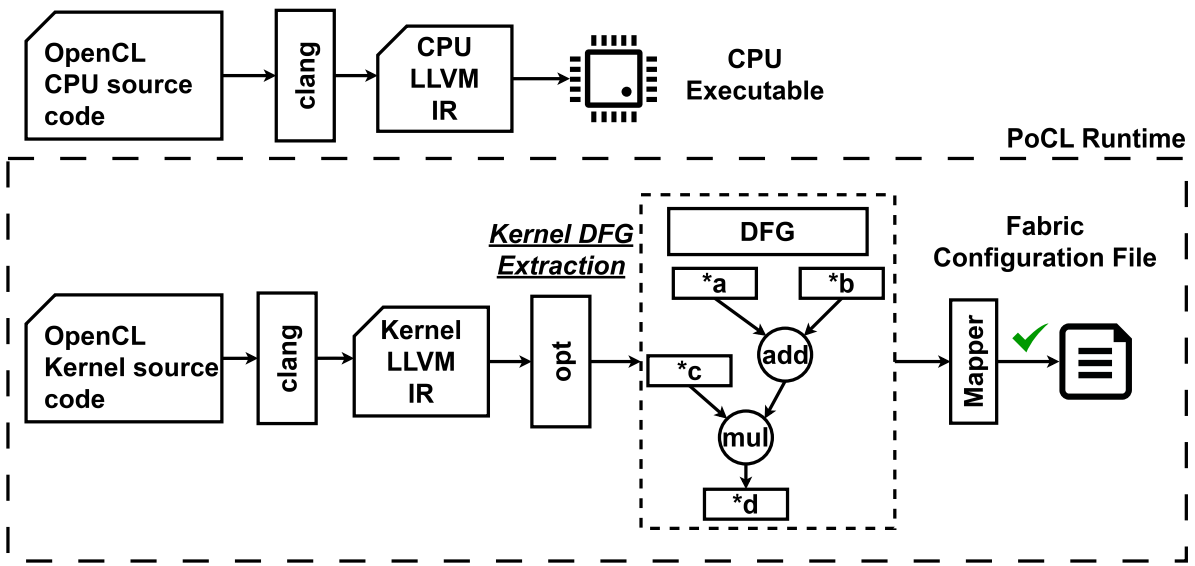


Our Contribution[2]

2. An **abstraction layer**

a. Runtime kernel scheduling via dynamic management of regions.
b. Communication and configuration of the CGRA's regions.
c. Data exchange between device global memory and host.

3. A **runtime** for Portable Compute Language (PoCL)[3]

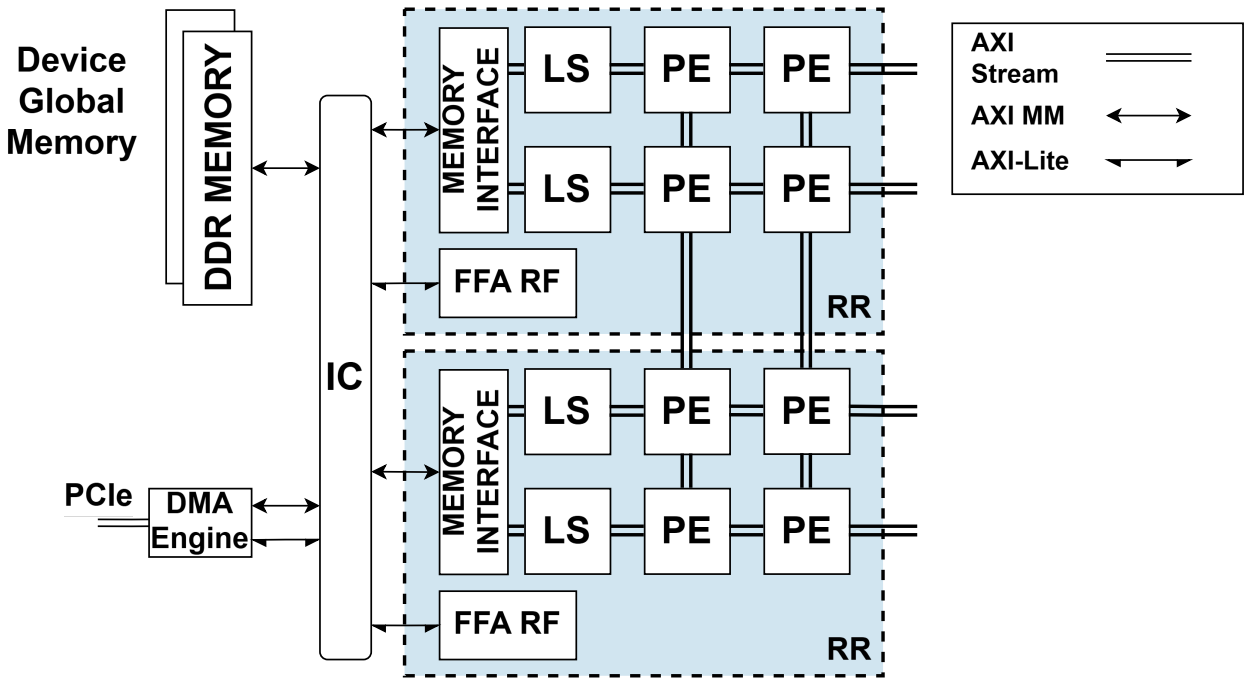


Evaluation

Implemented and verified on an Alveo U280 FPGA

1. The CGRA consists of 4 regions.
2. Host/Device communication is handled by the XDMA core IP.
3. Global Buffer implemented using the on-board DDR4 module.

- Velos costs 7% of the total FPGA's LUTs and operates at 250MHz.
- Total chip power measured at 4.8 W.



Evaluation

Cost of statefull

Snapshot-Capture and Execute-From-Snapshot induces a time overhead of 28% relative to region configuration.

Performance

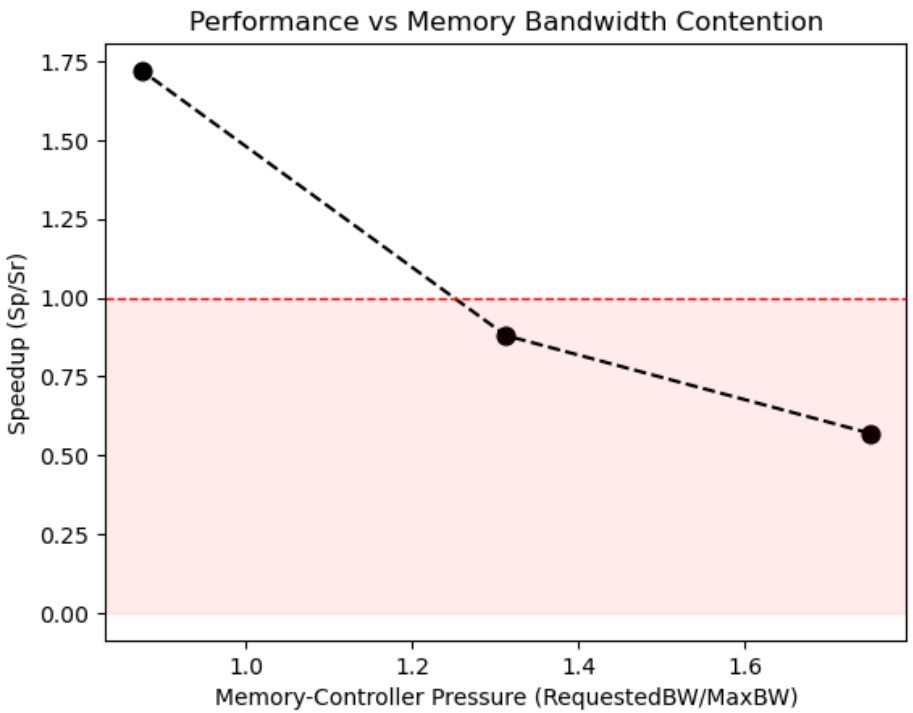
Evaluated using common ML operations (`ReLU`, `Vector/Matrix-Vector/Matrix`)

- 34% speedup under shared memory controller.
- up to 97% with isolated execution.

Evaluation

Virtualizing only compute PEs is not enough!

A central buffer creates choke points that lead to performance degradation.



Thank you for your attention

&

See you at the poster session!

akyriazis@cslab.ntua.ece.gr

CSLab – National Technical University of Athens