35th IEEE International Conference on Application-specific Systems, Architectures and Processors 2024

Leveraging MLIR for Efficient Irregular-Shaped CGRA Overlay Design

Mohamed Bouaziz, Suhaib A. Fahmy
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Motivation

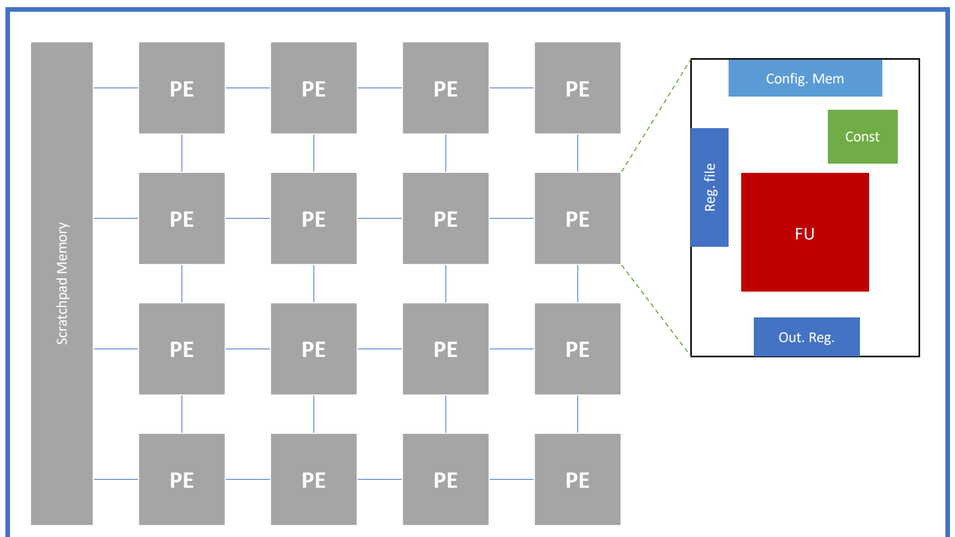

Fig. 1: Generic CGRA architecture

- Efficient as datapath architectures ✓
- Flexible as FPGAs ✓
- Communicates at word level ✓
- Computes using efficient FUs ✓
- Regular-shaped and homogenous ✗
- Doesn't fit HPC data patterns ✗

Example:

- HW underutilization of top right PE & bottom left PE
- Oblivious compiler to data access patterns
- Much complex search space

Generally, it applies to:

- Interconnects
- Memory and registers structure
- Functional units operations

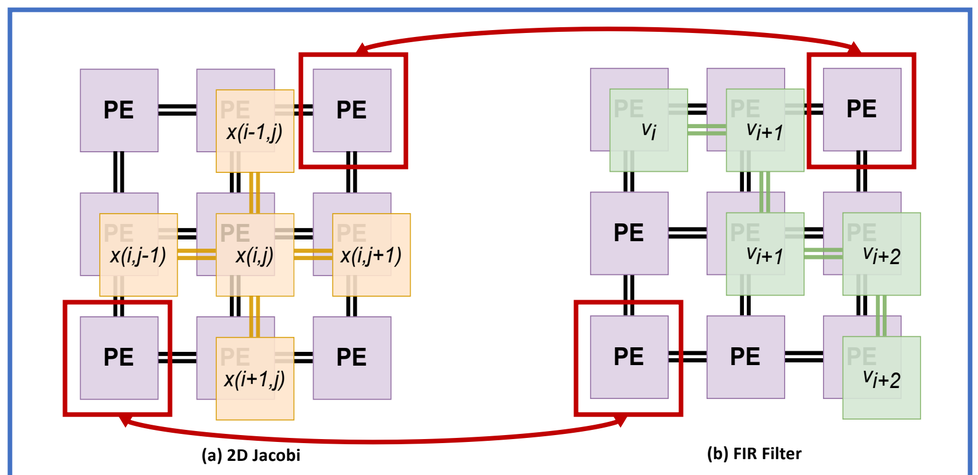

Fig. 2: Mapping Example

Poor hardware reuse & Increased deployment effort

Related Work & Proposed Approach

	HW optimization	*SW optimization*	*Overlay Generation*	*Multi-targeting*
REVAMP, ASPLOS'22	+	-	-	+
FlexC, ArXiv'23	-	+	-	+
OverGen, Micro'22	±	±	+	-
This work	+	+	+	+

Tab. 1: Comparison with Related Work

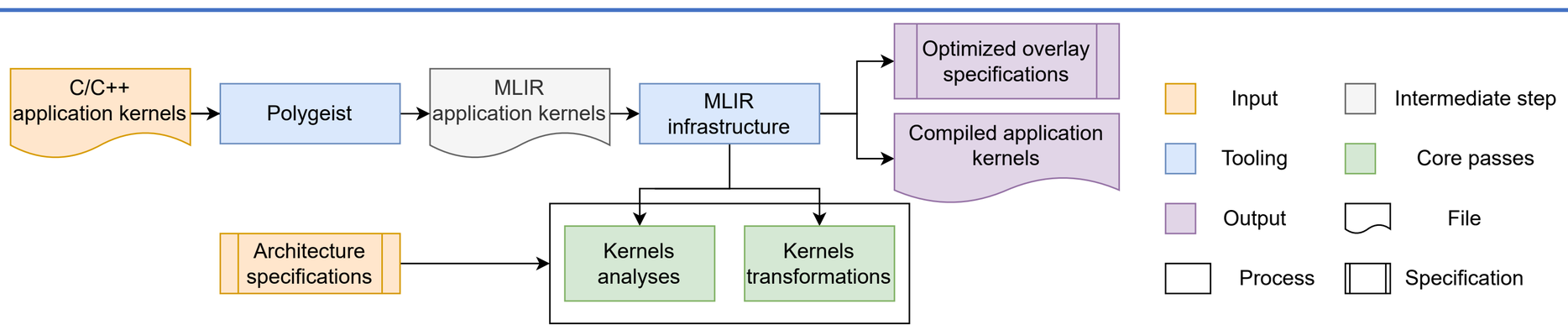

Fig. 3: Optimized Overlay Generation Process

Preliminary Results

MLIR Built-in Passes

```
affine-loop-tile="cache-szie=32"
affine-loop-unroll="unroll-full=1,unroll-full-threshold=3"
affine-super-vectorize="virtual-vector-size=8"
```

MLIR-AIE Pass

```
aie-vectorize
```

	Tile Size	Vector Update Ops.	Multiply Op.	Mac Ops
2D Convolution	16 x 16	6	1	8
2D Jacobi	16 x 16	6	1	4

Tab. 2: Resource Utilization

Same data access pattern ***Run on RISC-V cores, stored in prog. mem.***

Same generated overlay

Kernels analyses | Kernels transformations

Architecture specifications

8-lane SIMD core/PE

32KB SPM/PE

Fig. 4: AMD Versal ACAP VCK5000

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

accelerated Connected Computing Lab