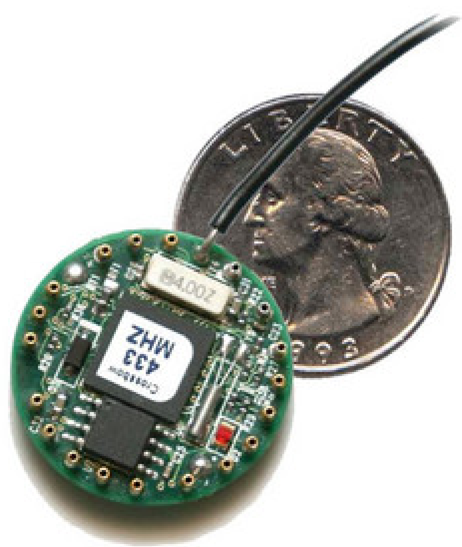
Blocks: Redesigning Coarse Grained Reconfigurable Architectures for Energy Efficiency

M. Wijtvliet
J.A. Huisken
L.J.W. Waeijen
H. Corporaal

Mobile devices



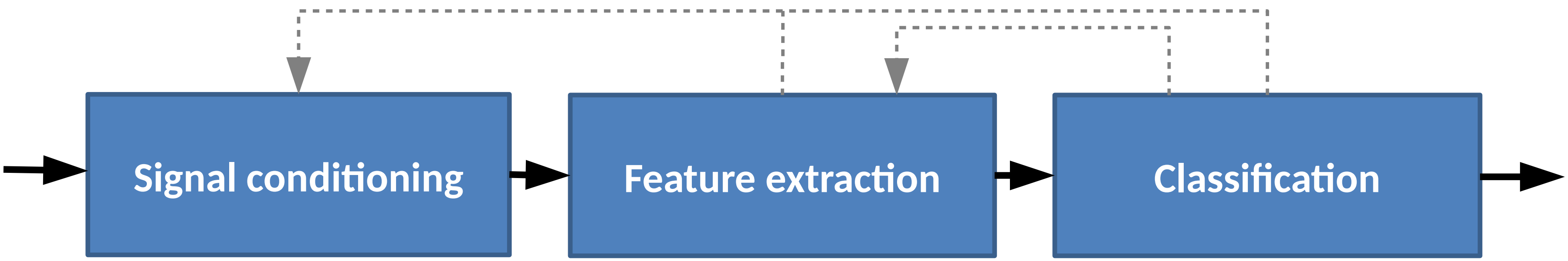
Mobile devices

- Limited size
- Wireless communication
- On-board processing

Algorithms

- Wearable medical devices as an example.
  - Low latency can be important.
  - Often significant compute effort.
  - Updates on a regular basis.
    - This rules out an dedicated ASIC.

Algorithms



Signal conditioning

- FIR
- IIR
- Convolution (Windowing, 2D stencils)
- Peak-suppression (Clipping)
- Amplitude squaring
- Sensor fusion, motion artifact reduction.
- …

Feature extraction

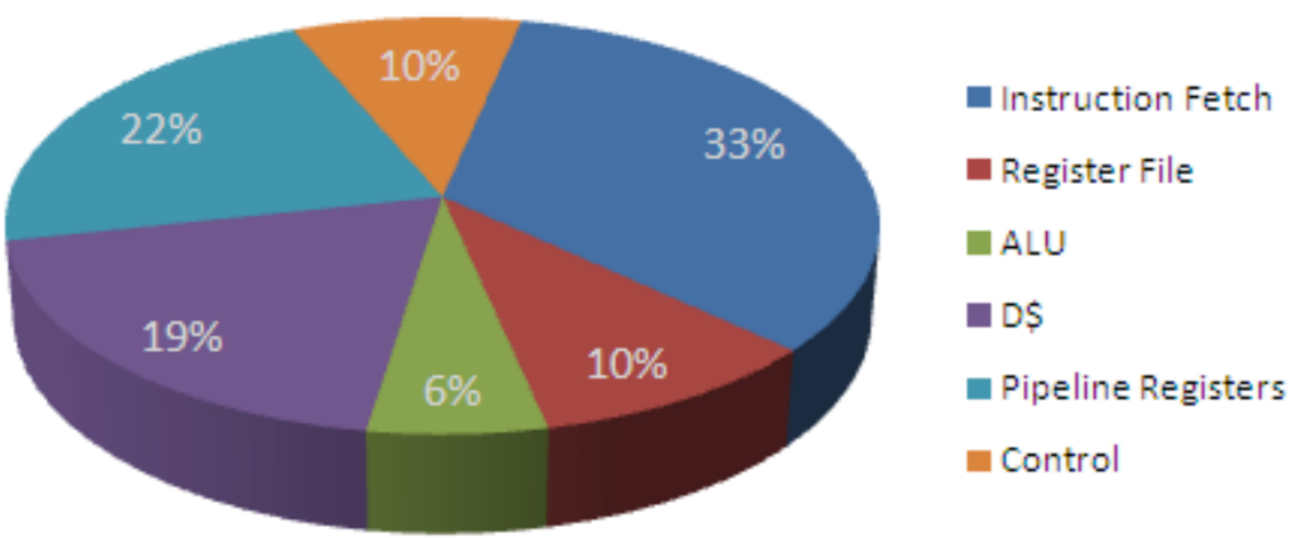
- Mean and Variance
- FFT / DFT
- Continuous Wavelet Transform
- Discrete Wavelet Transform
- Line length
- Root Mean Square
- Number of Maxima and Minima
- Power Spectrum Density
- Band Power
- Derivative
- Edge detection
- …

Classification

- Support Vector Machine
- (Convolutional) Neural Network
- Linear Discriminant Analysis
- Bayesian classifiers (Markov)
- Nearest Neighbors
- Combined (Boosting, Voting, Stacking)
- Independent Component Analysis
- …

Popular architectures

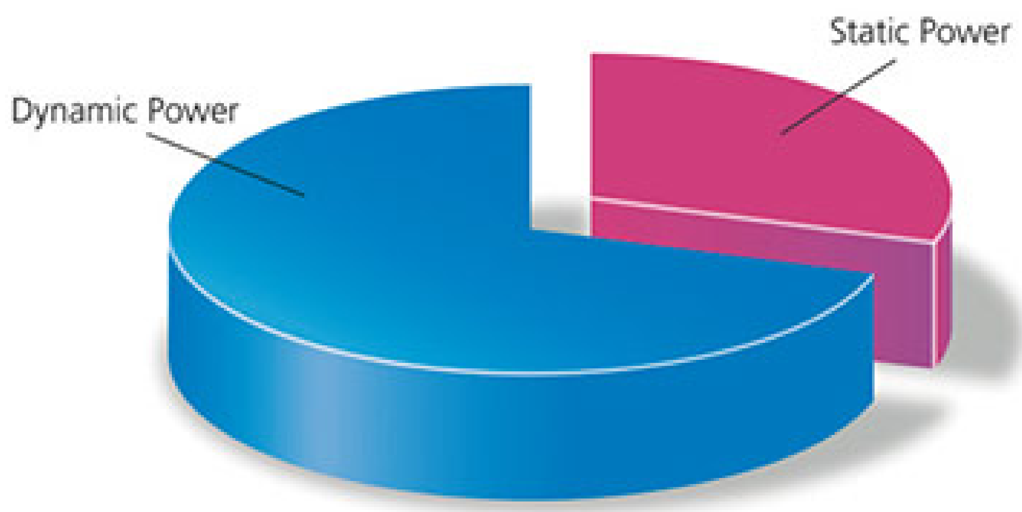
- Microprocessor
  - Quite flexible
  - Reasonable performance
  - Low energy efficiency



[figure: Understanding sources of inefficiency in General-Purpose Chips, Hameed et al.]

Popular architectures

- FPGA
  - Very flexible
  - Good performance
  - Medium energy efficiency

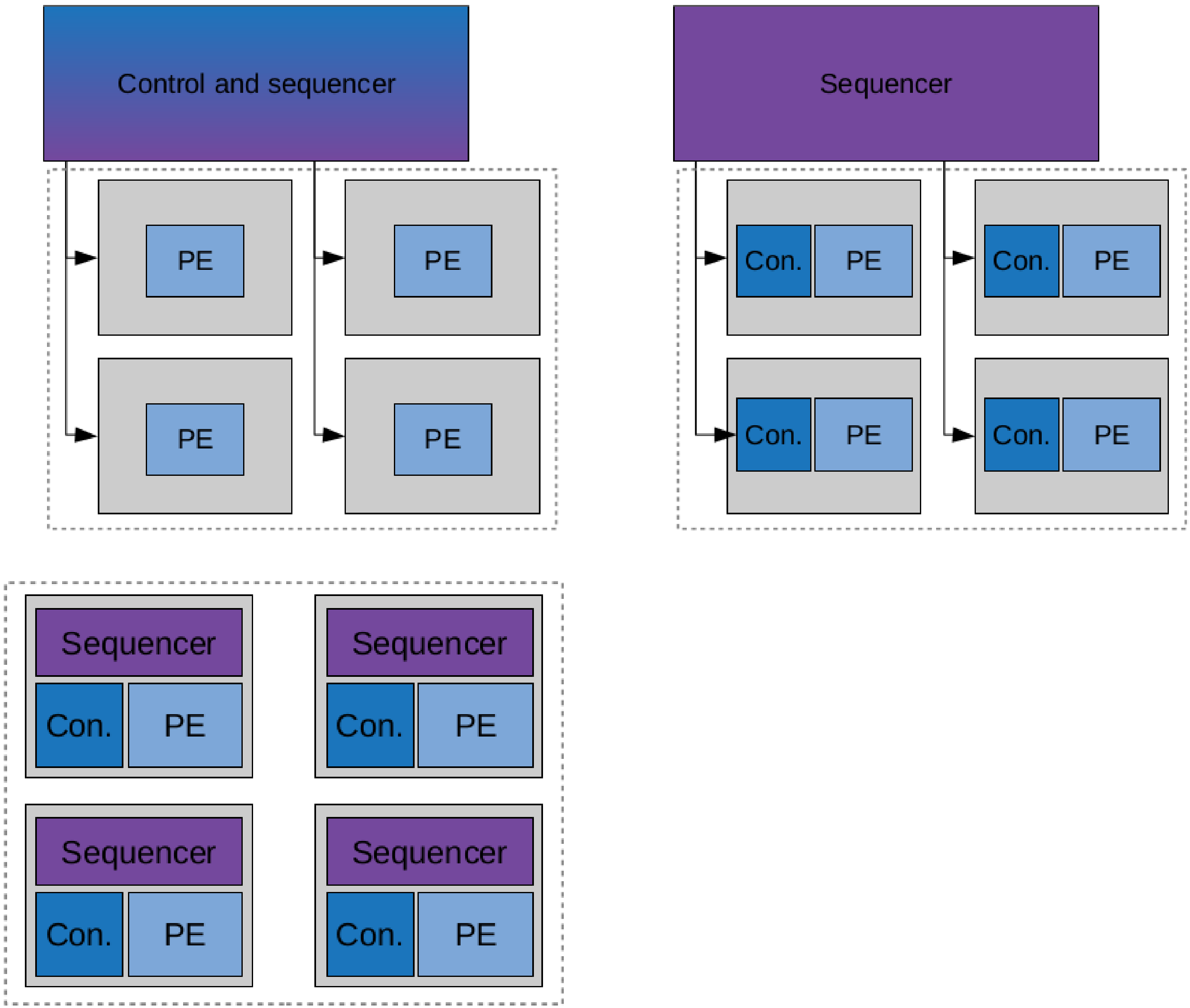


[Figure: ACTEL]

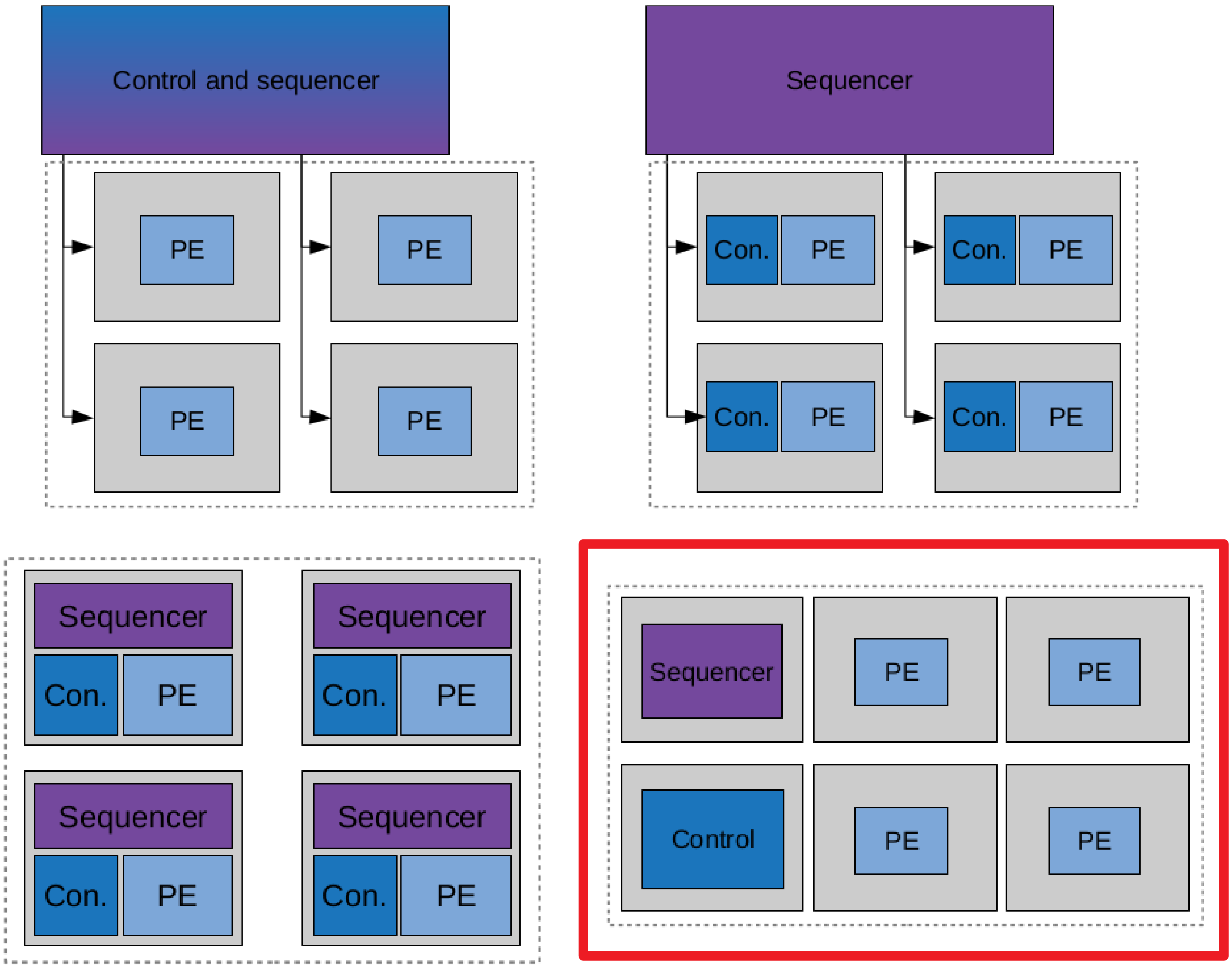
Coarse Grained Reconfigurable Architectures

- Many DSP algorithms do not require bit-level operations.
- CGRAs operate on wider data elements than FPGAs.
- Both spatial and temporal reconfiguration.
- Better energy efficiency possible.

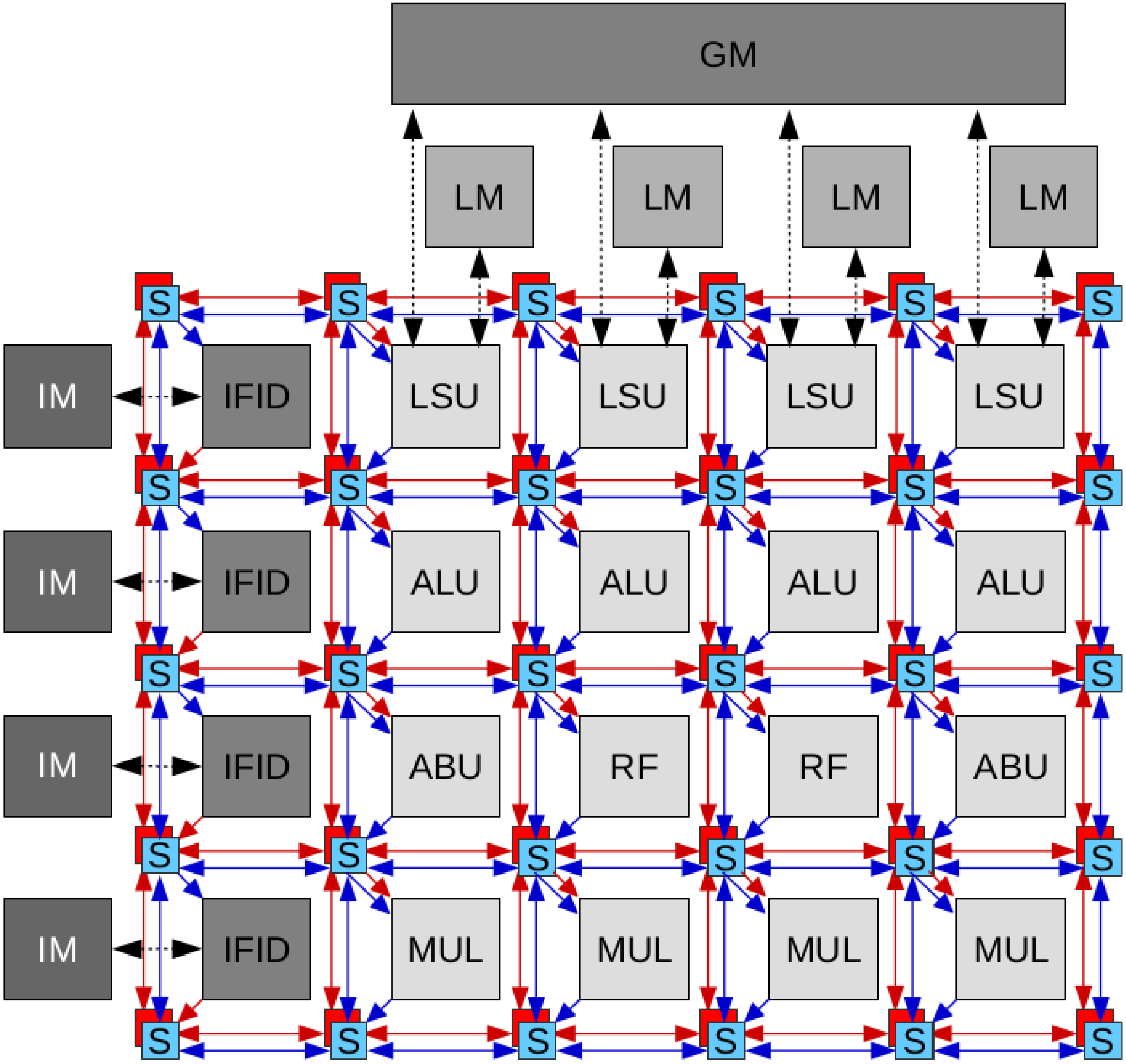
Temporal reconfiguration



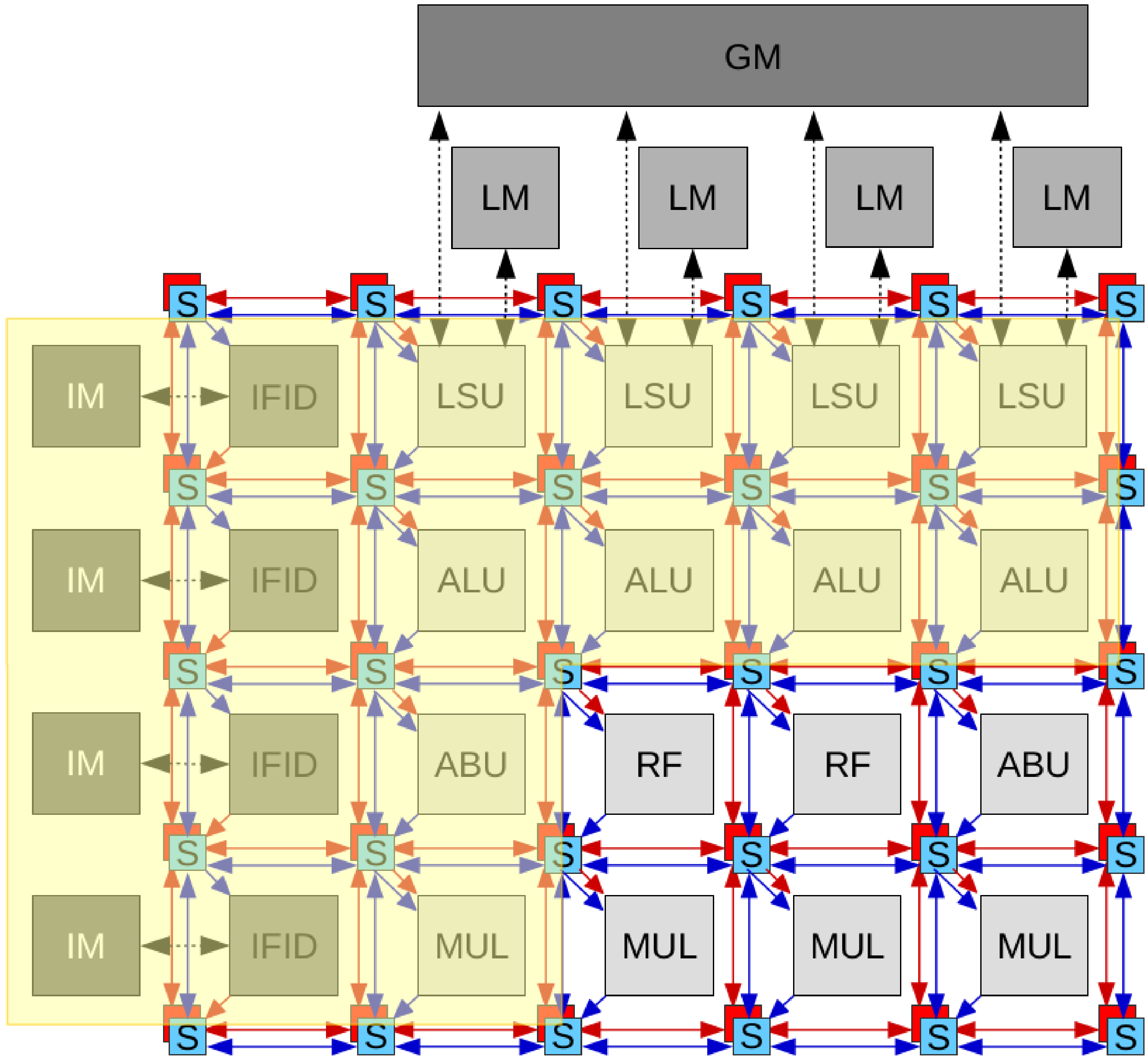
Temporal reconfiguration



Spatial reconfiguration



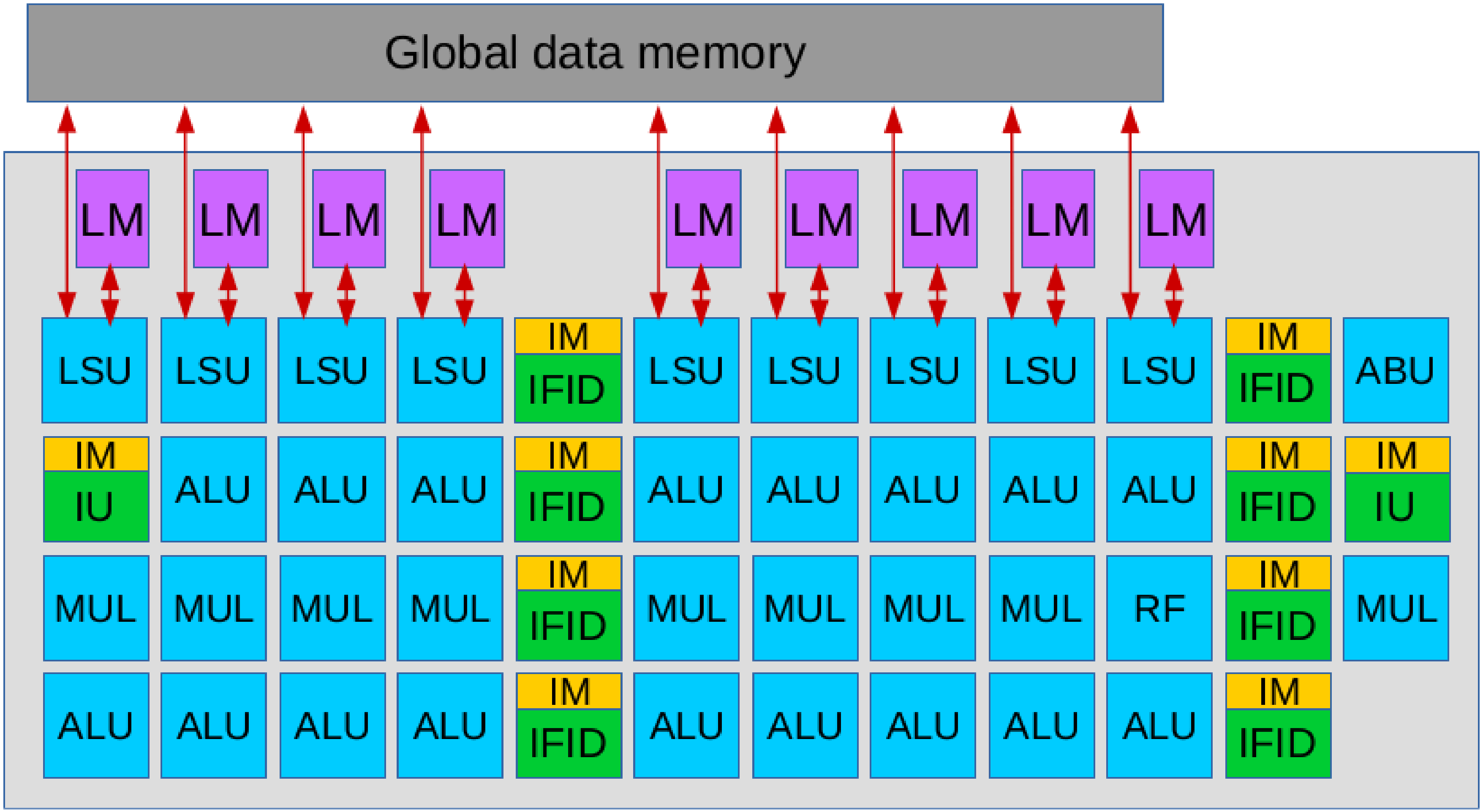
Spatial reconfiguration



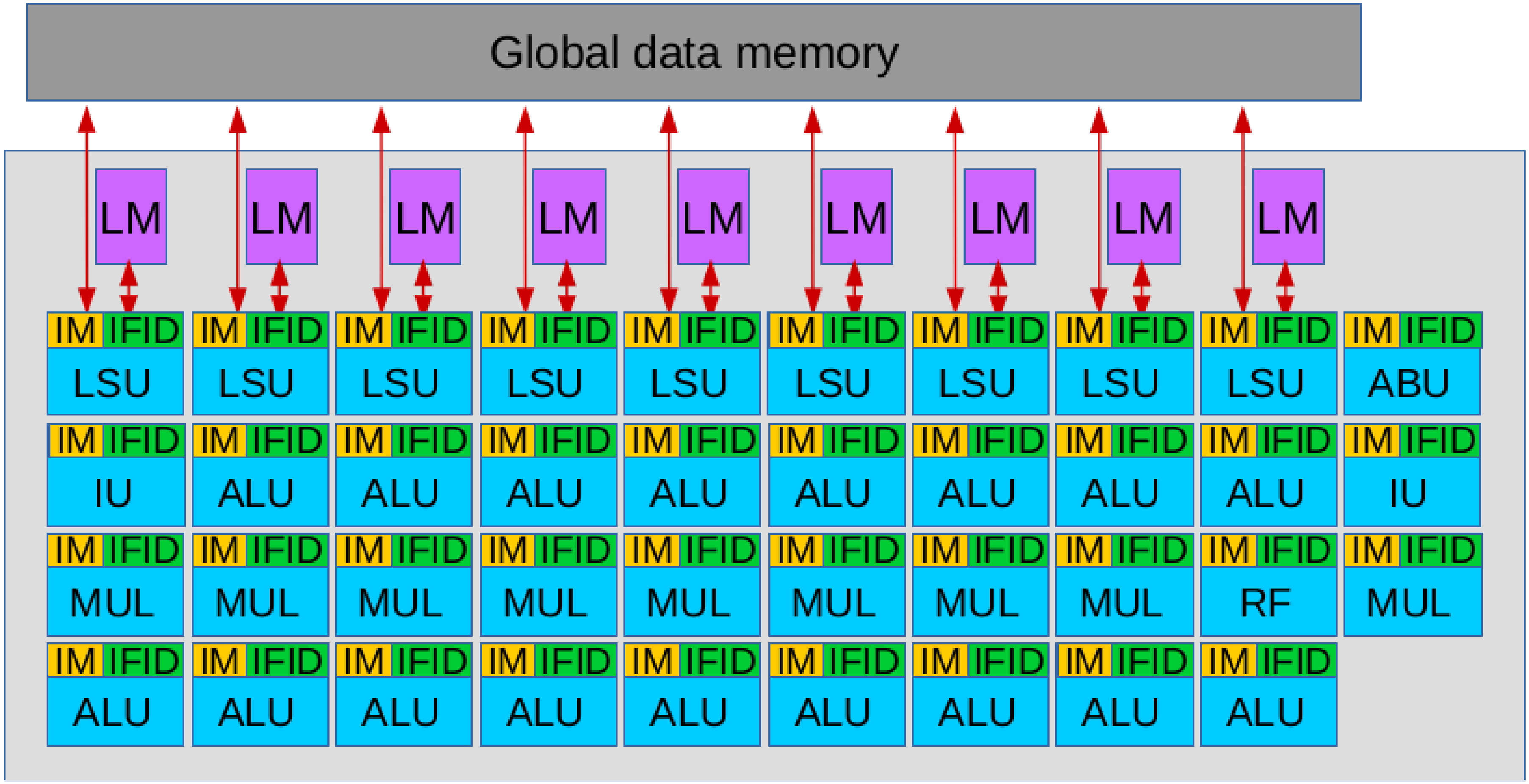
Evaluation

- Numbers obtained with post place-and-route simulation and power estimation on commercial 40nm, low-power.
- Memories included (TSMC)
- For three architectures:
  - Blocks
  - Traditional CGRA
  - Application specific processor
- Eight benchmark kernels

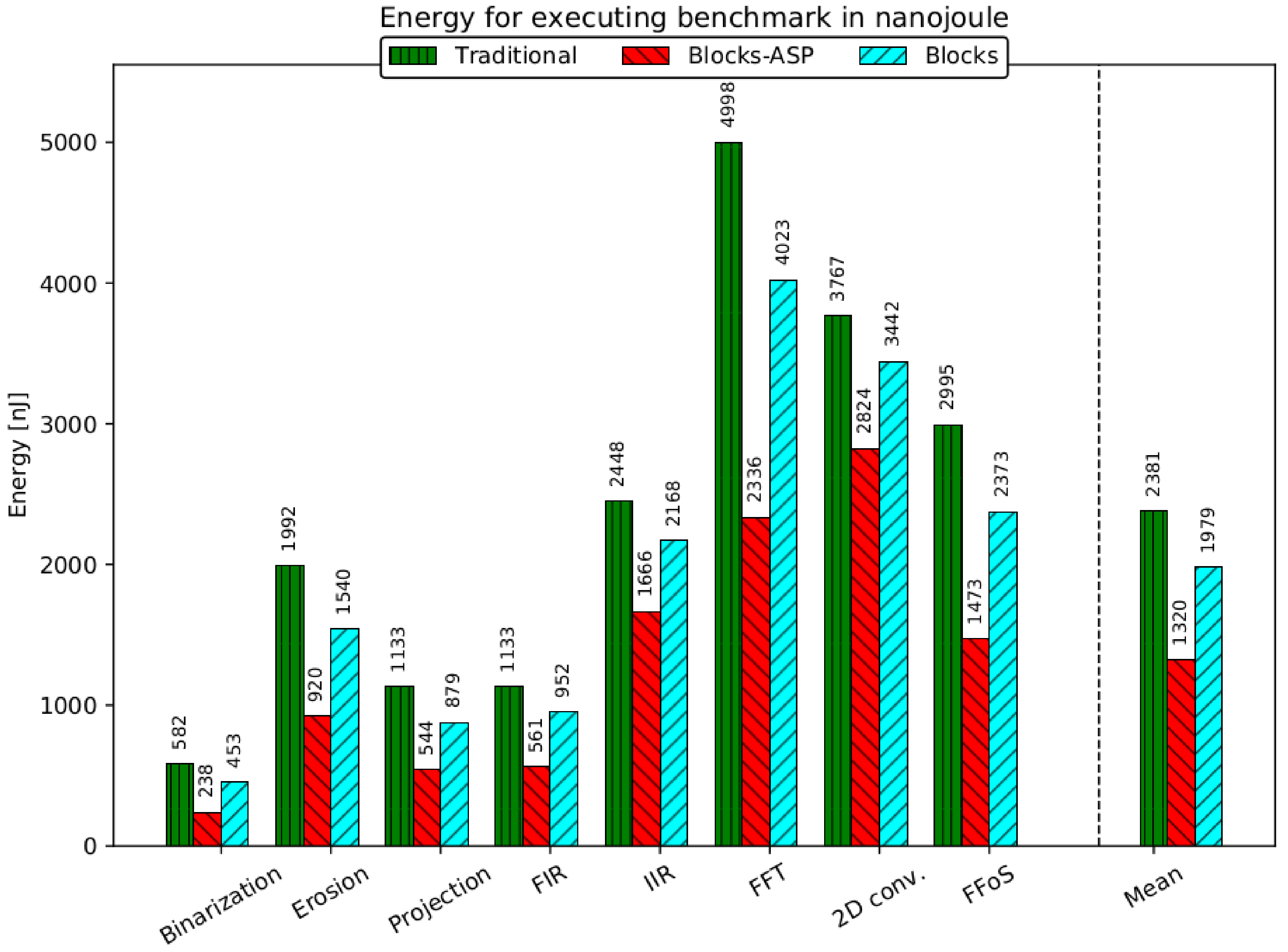
Reconfigurable architectures



Reconfigurable architectures

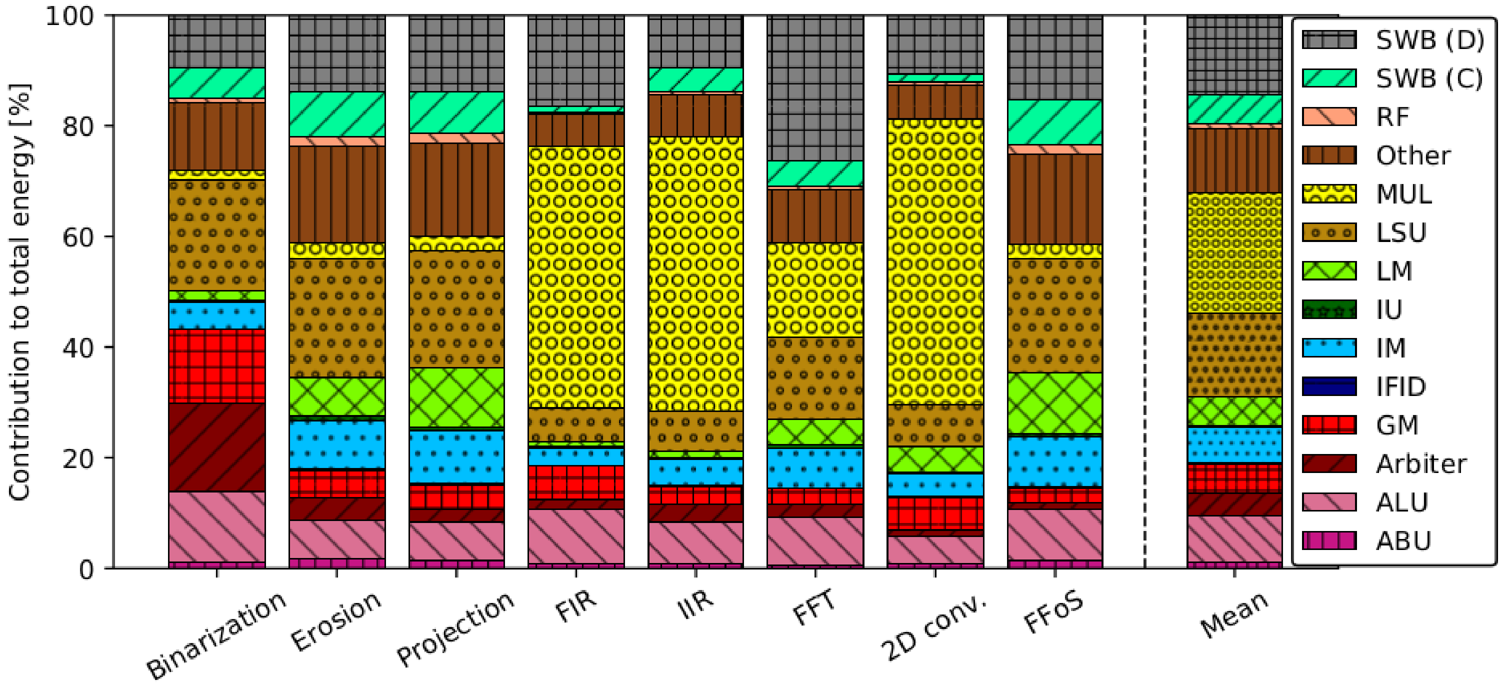


Results

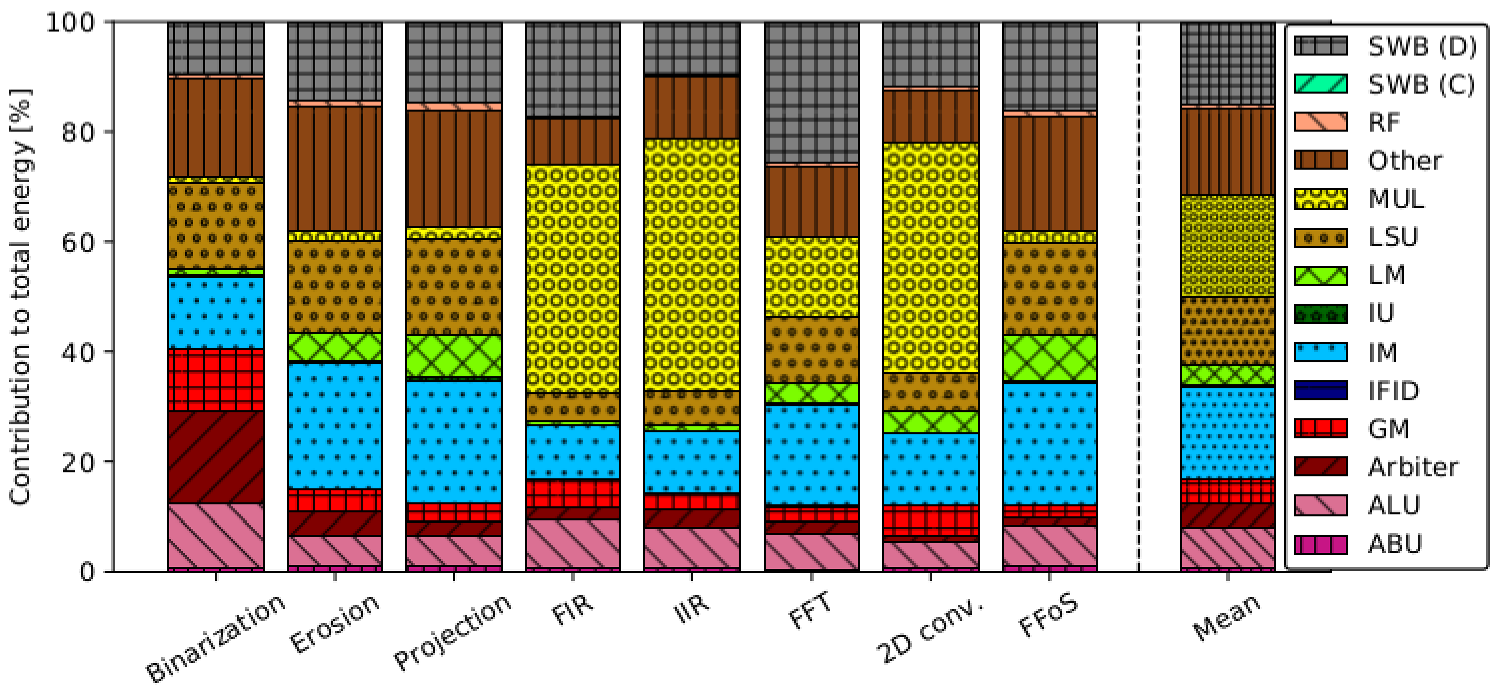


Energy for executing benchmark in nanojoule
Traditional
Blocks-ASP
Blocks
Energy [nJ]
0
1000
2000
3000
4000
5000
582
238
453
1992
920
1540
1133
544
879
1133
561
952
2448
1666
2168
4998
2336
4023
3767
2824
3442
2995
1473
2373
2381
1320
1979
Binarization
Erosion
Projection
FIR
IIR
FFT
2D conv.
FFoS
Mean

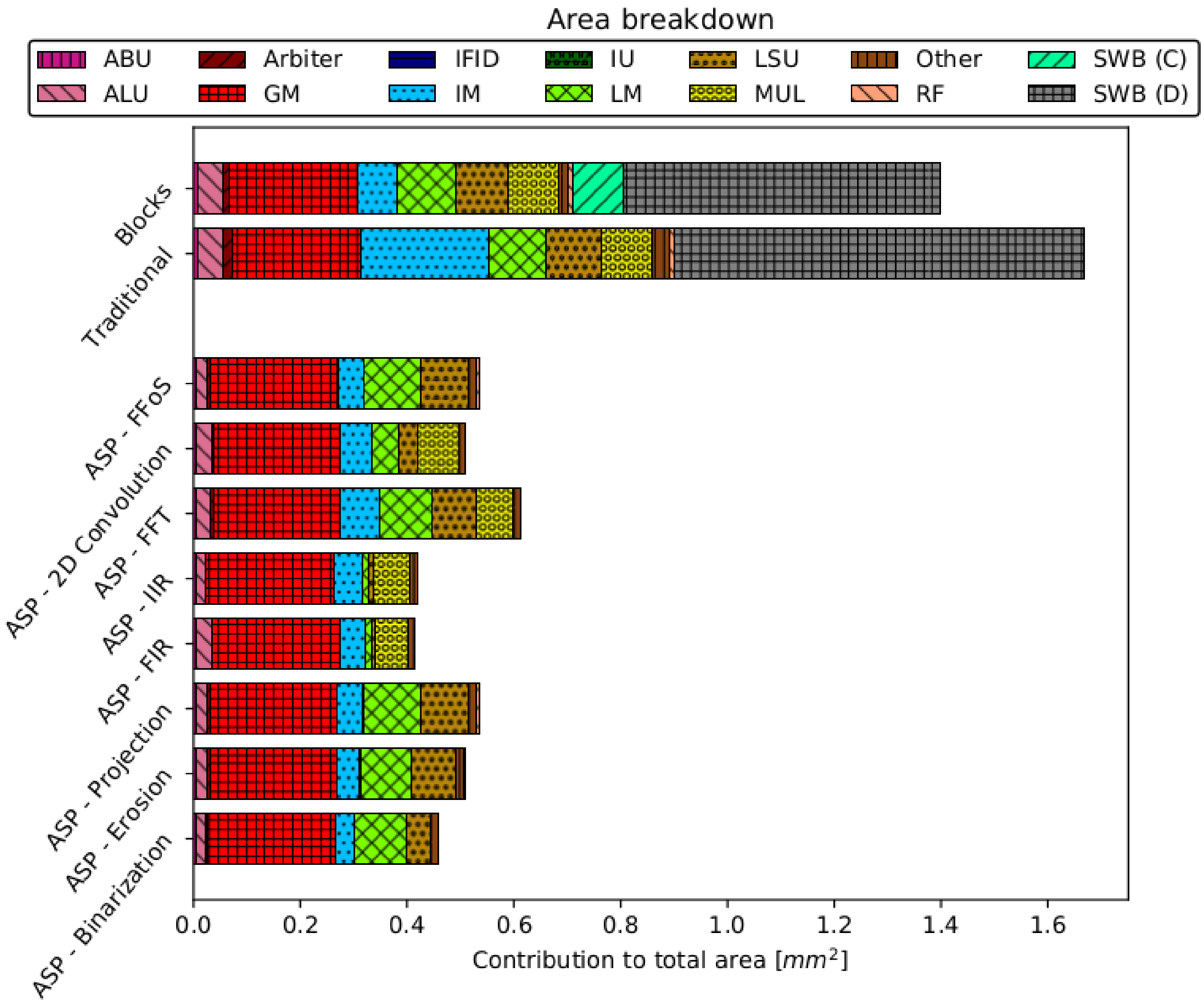
Blocks



Traditional



Results



Area breakdown
ABU
Arbiter
IFID
IU
LSU
Other
SWB (C)
ALU
GM
IM
LM
MUL
RF
SWB (D)
Blocks
Traditional
ASP - FFoS
ASP - 2D Convolution
ASP - FFT
ASP - IIR
ASP - FIR
ASP - Projection
ASP - Erosion
ASP - Binarization
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Contribution to total area [mm^2]

Conclusions

- Blocks allows CGRAs to match required parallelism types closely, leading to:
- Total energy reduction between 9% and 29%.
- Overhead reduction between 46% and 76%.
- A small version of Blocks was taped-out and shown working. Larger version in progress.

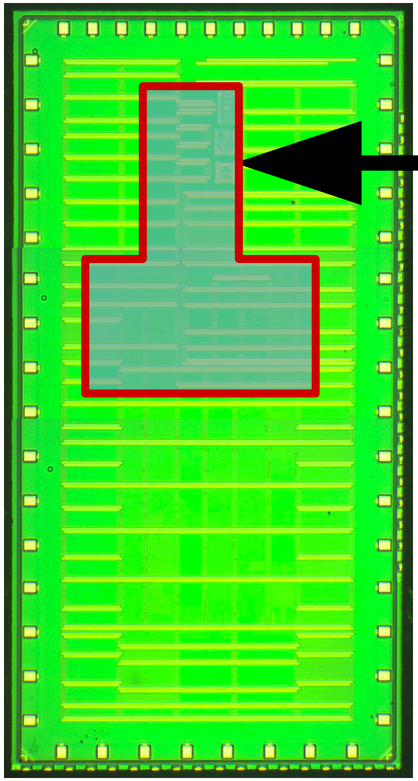
Thank you for your attention

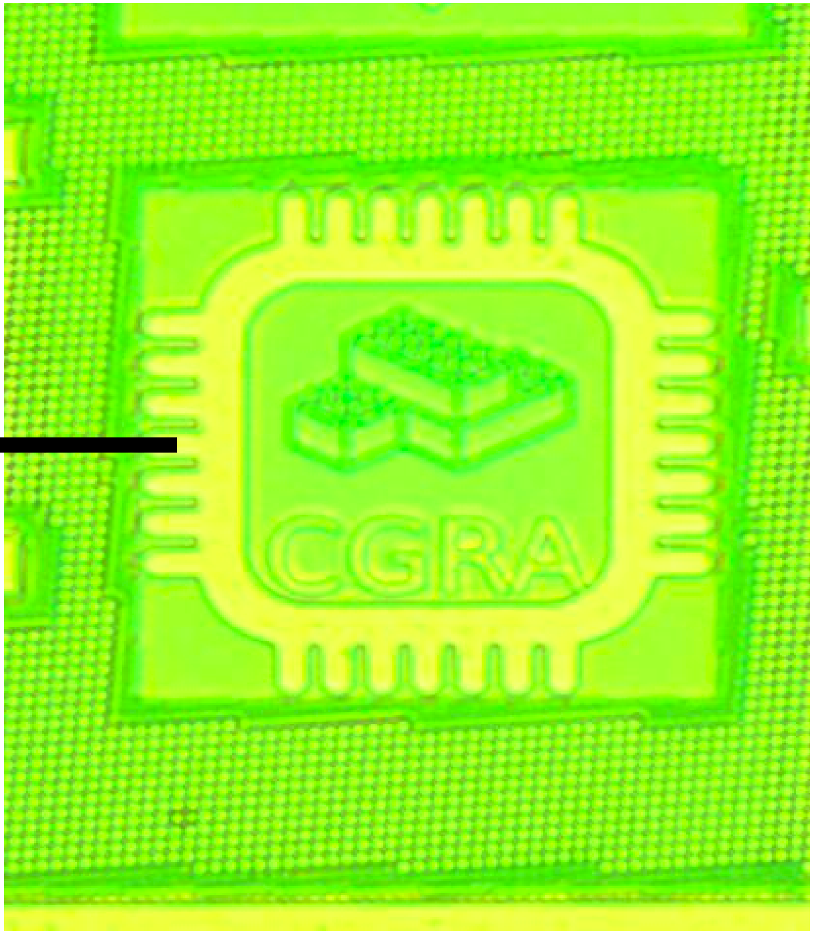
- Any questions?

Chip info:

Area: 1.92 x 3.84mm
Power: 157.2 mW

3*ARM Cortex-M0
1*CGRA





Backup slides

Toolflow

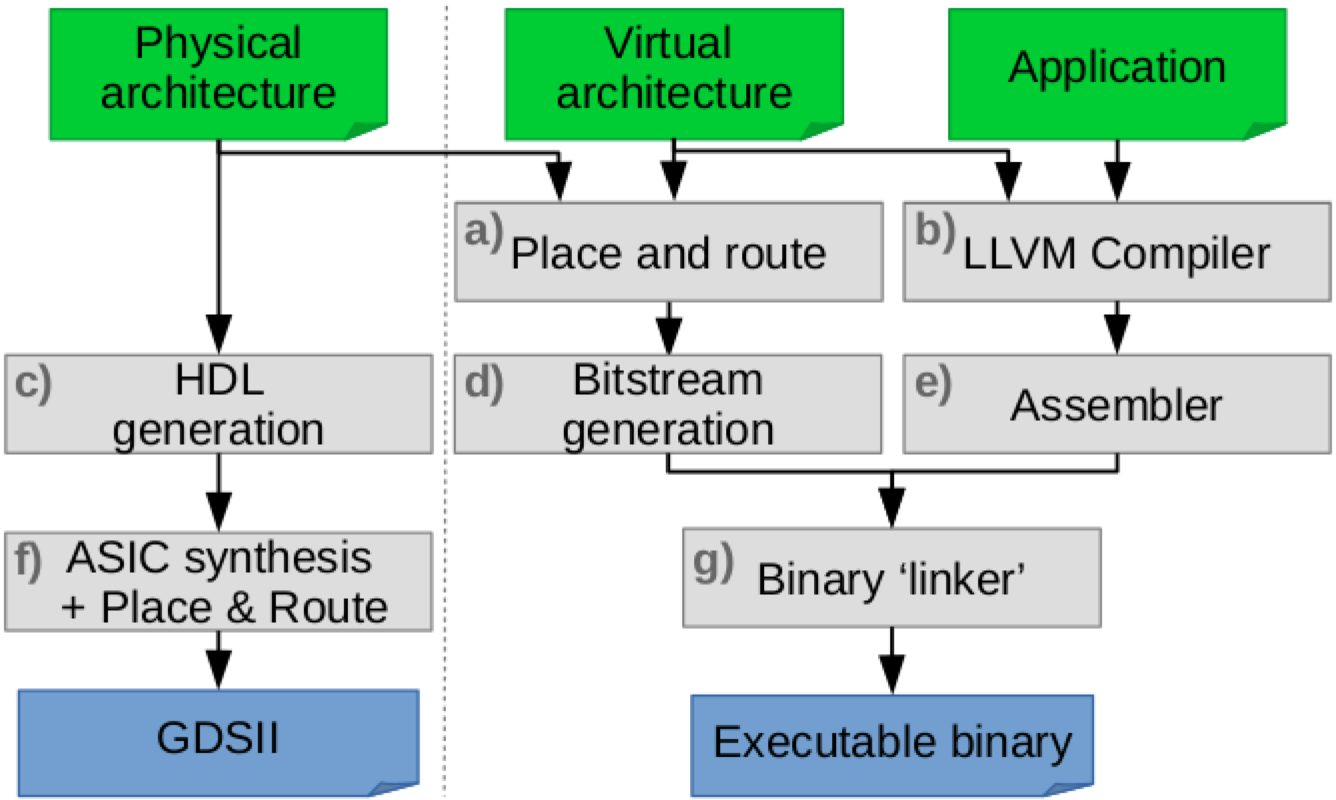


Fig. 3: The Blocks tool-flow

Inside a function unit

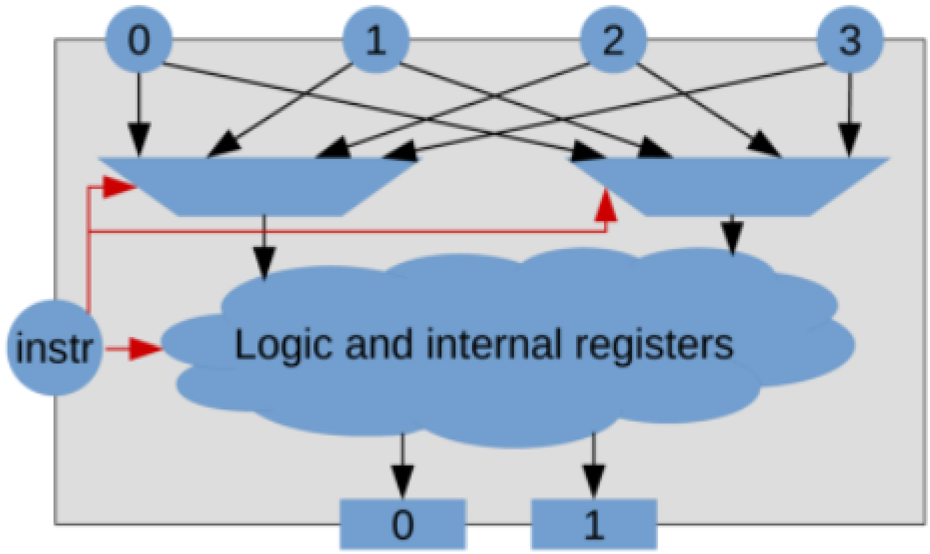


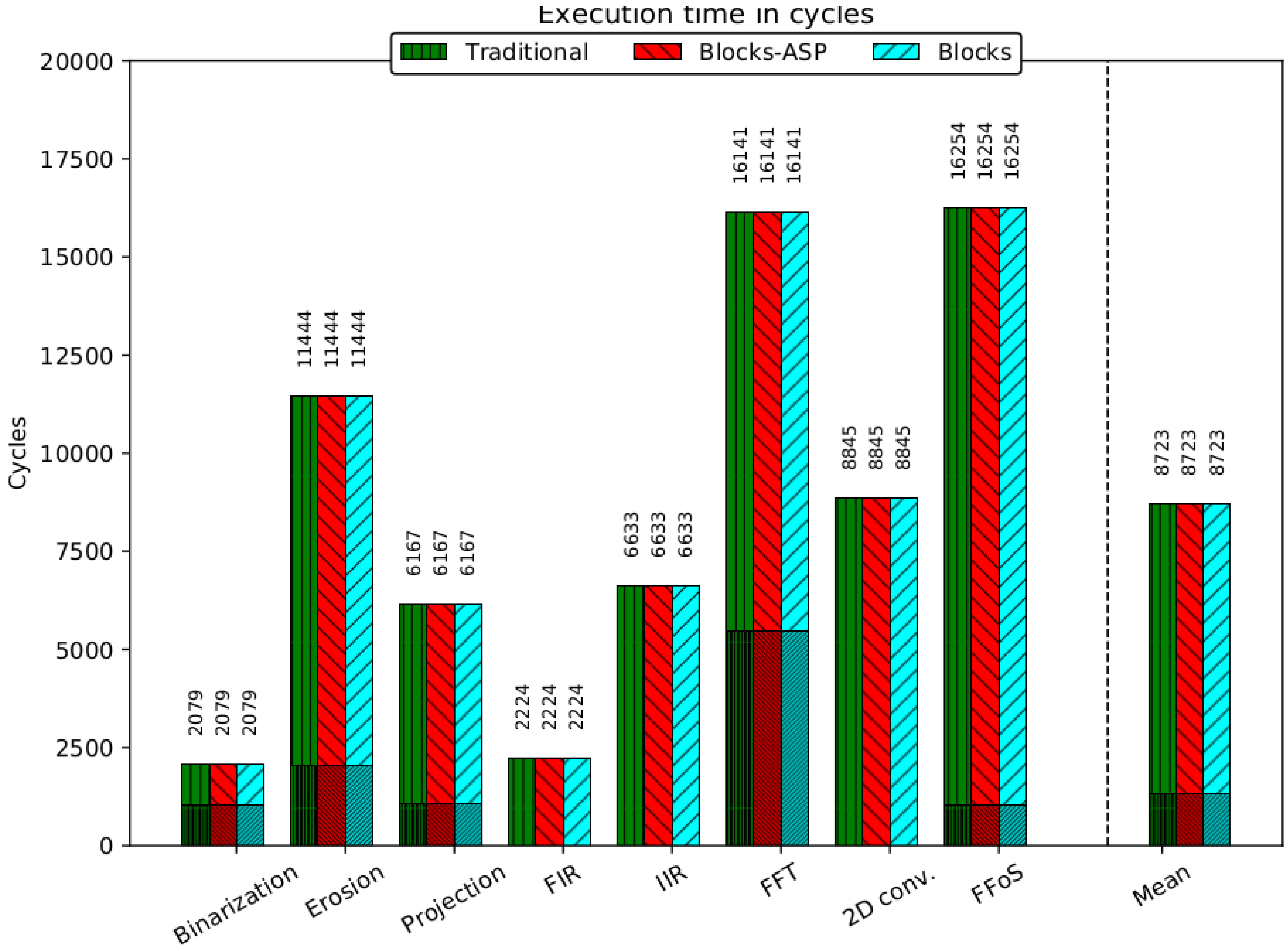
Fig. 2: Functional unit with 4 inputs and 2 outputs.

Benchmark kernels

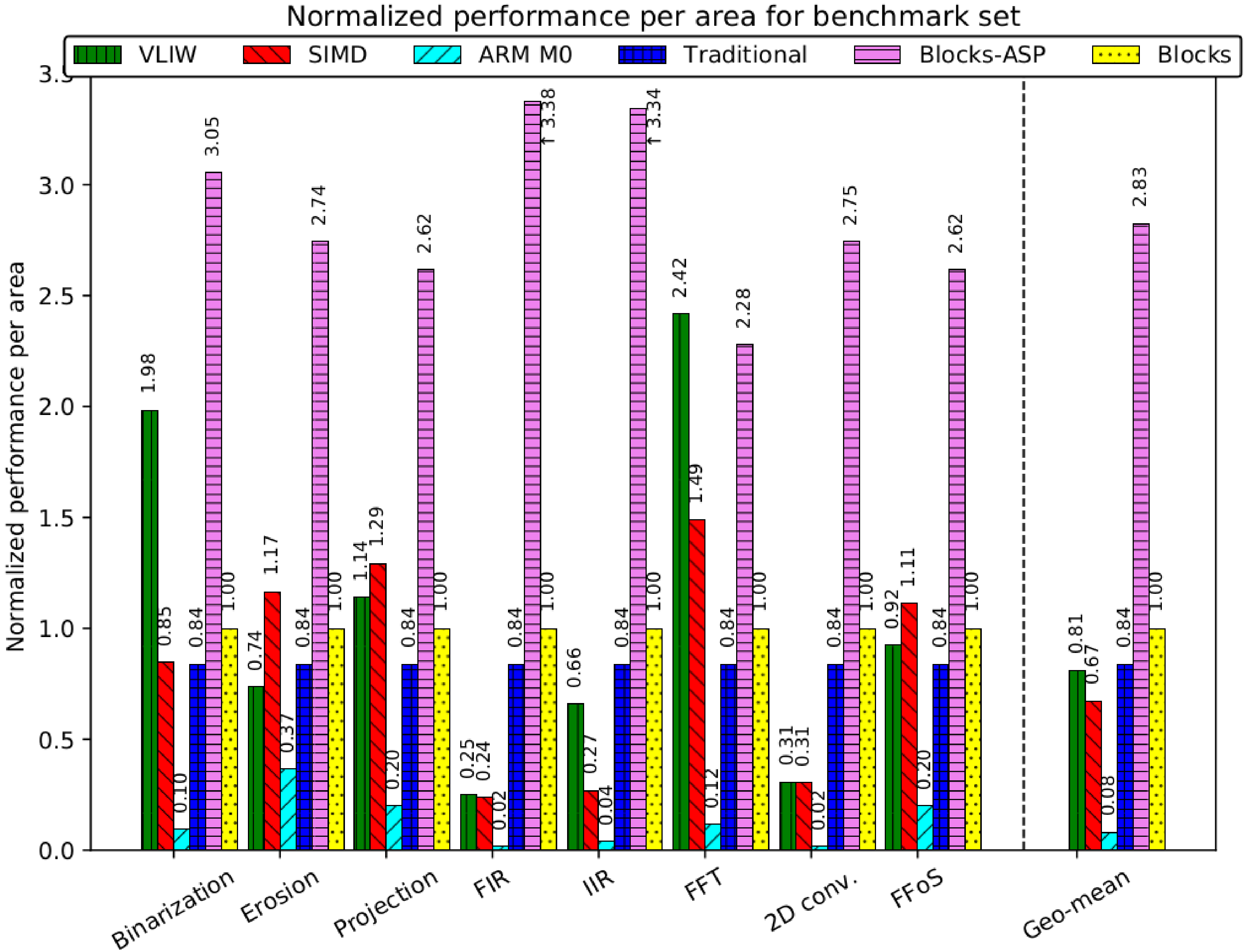
Benchmark	Description	Type	Data size
Binarization	Thresholding on greyscale image	scalar to scalar	128*64 pixels (8 bit)
Erosion	Noise removal by AND-ing neighbouring pixels	3*3 window to scalar	128*64 pixels (8 bit)
Projection	Sums each horizontal and vertical row/column	vector to scalar	128*64 pixels (8 bit)
FIR	8-tap low-pass FIR filter on input signal	8*1 convolution	2200 samples (32-bit)
FIR	3rd order low-pass filter on input signal	2*1 + 2*1 convolution	2200 samples (32-bit)
FFT	8-point complex FFT on audio signal	8*1 to 8*1 vector (complex)	2200 samples (32-bit)
2D convolution	Gaussian blur on image	3*3 window to scalar	128*64 pixels (8 bit)
FFoS	Industrial vision application	image to 2 8*1 vectors	128*64 pixels (8 bit)

Table 6.1: *Overview of benchmark kernels*

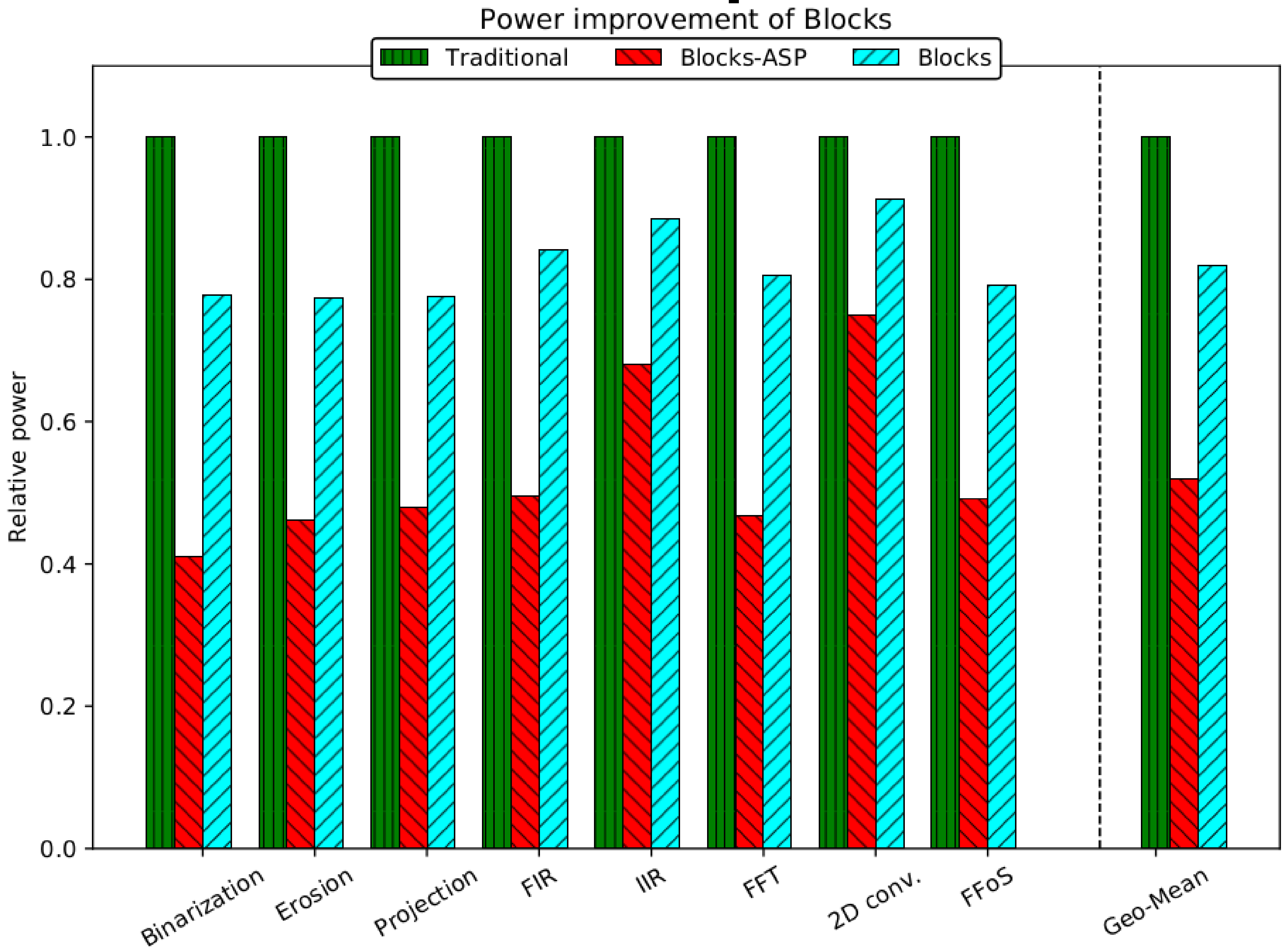
Performance



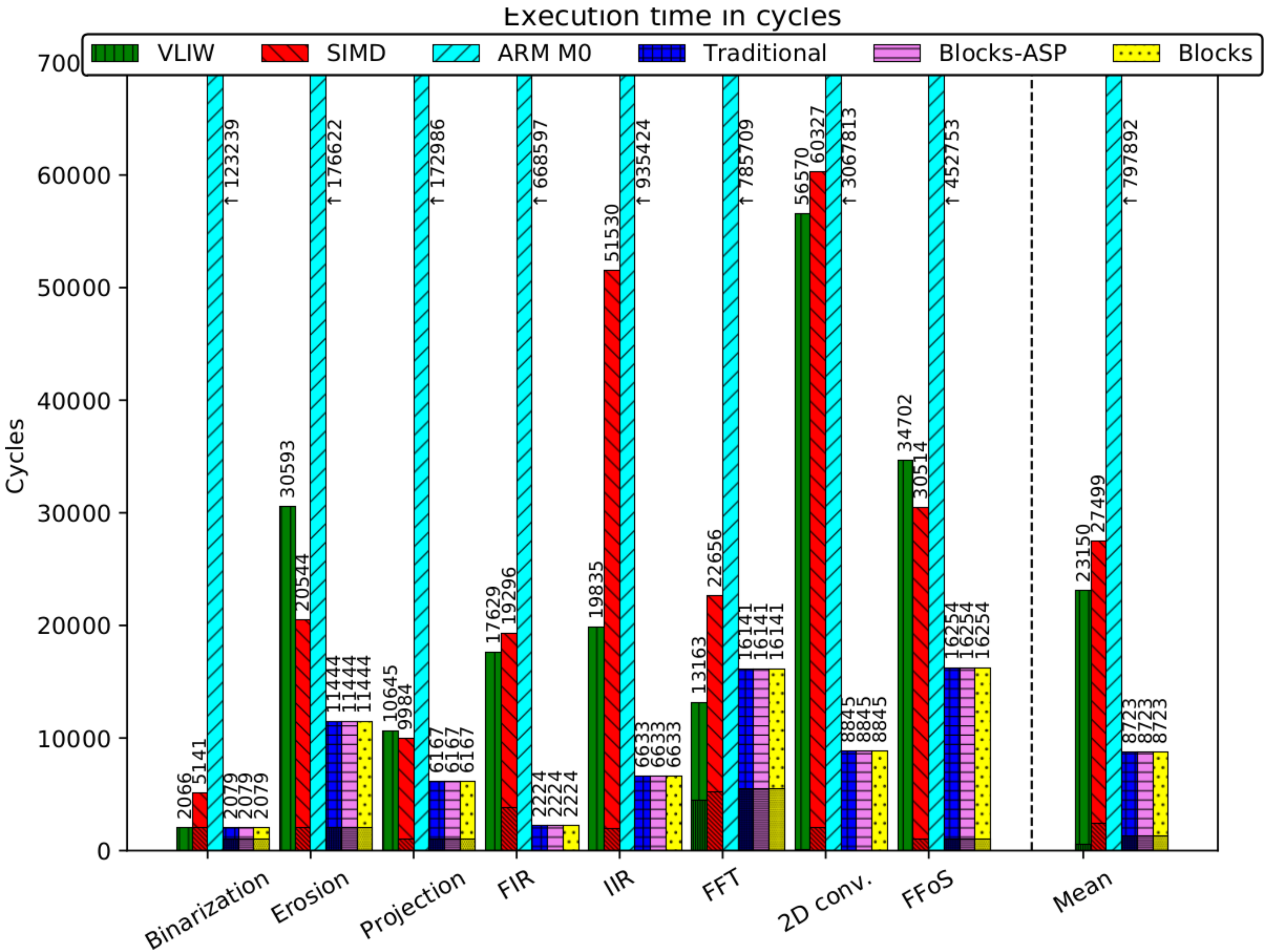
Performance per area



Relative power



Performance (SIMD + VLIW+ ARM M0)



Energy (SIMD + VLIW+ ARM M0)