

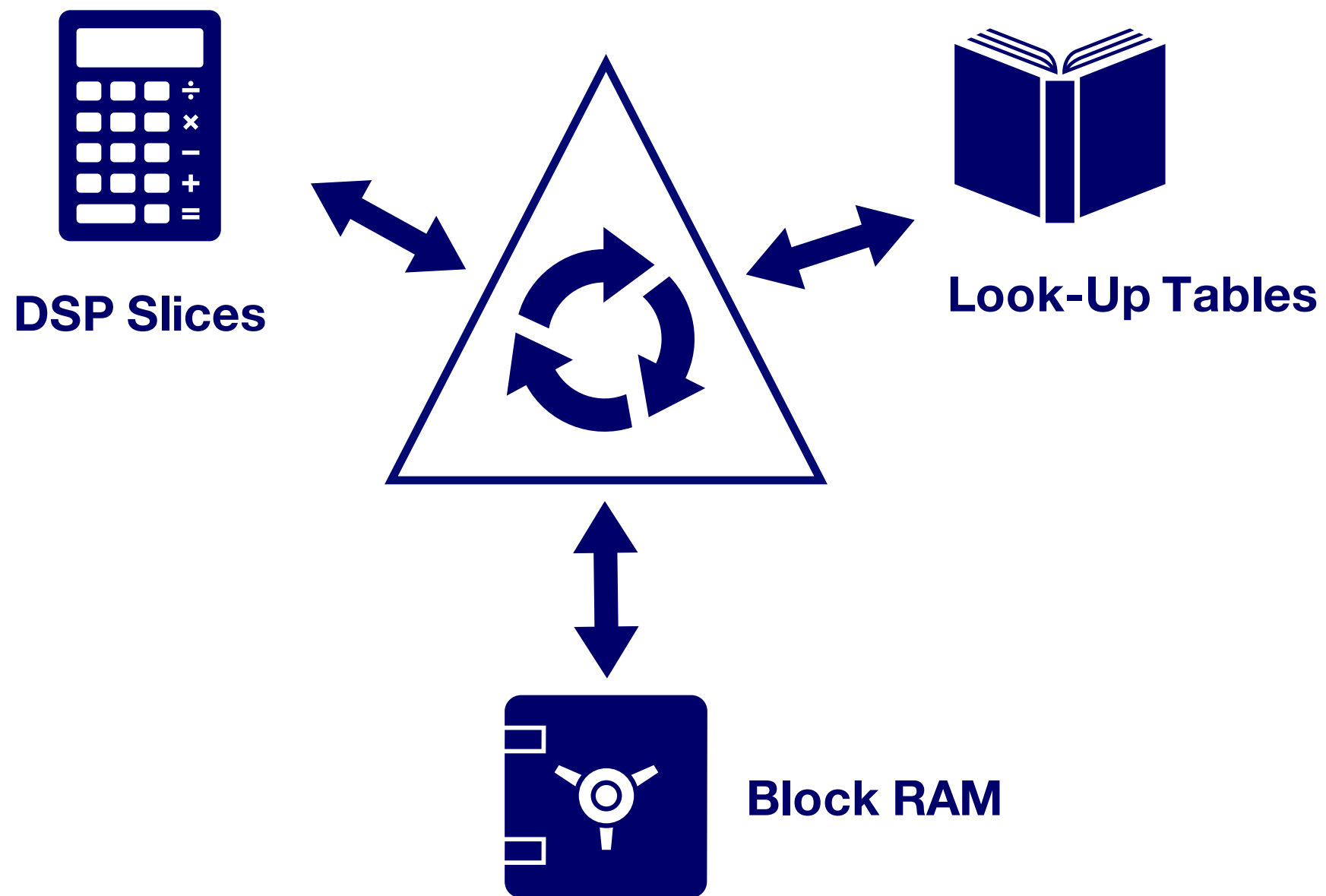
Architectures & Synthesis for Streaming Accelerators on FPGA

John McAllister

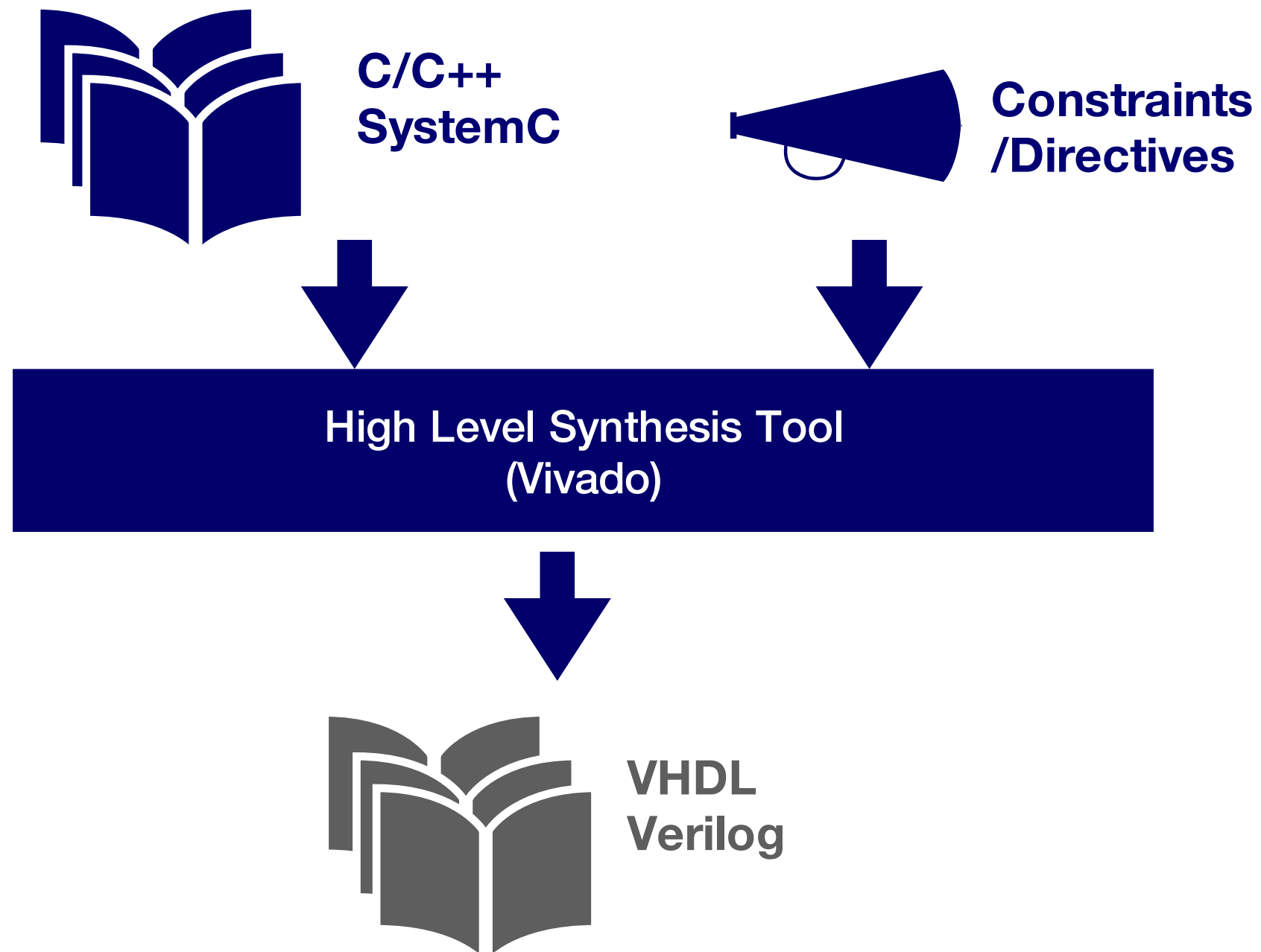
Institute of Electronics, Communications and Information
Technology (ECIT),
Queen's University Belfast

jp.mcallister@qub.ac.uk

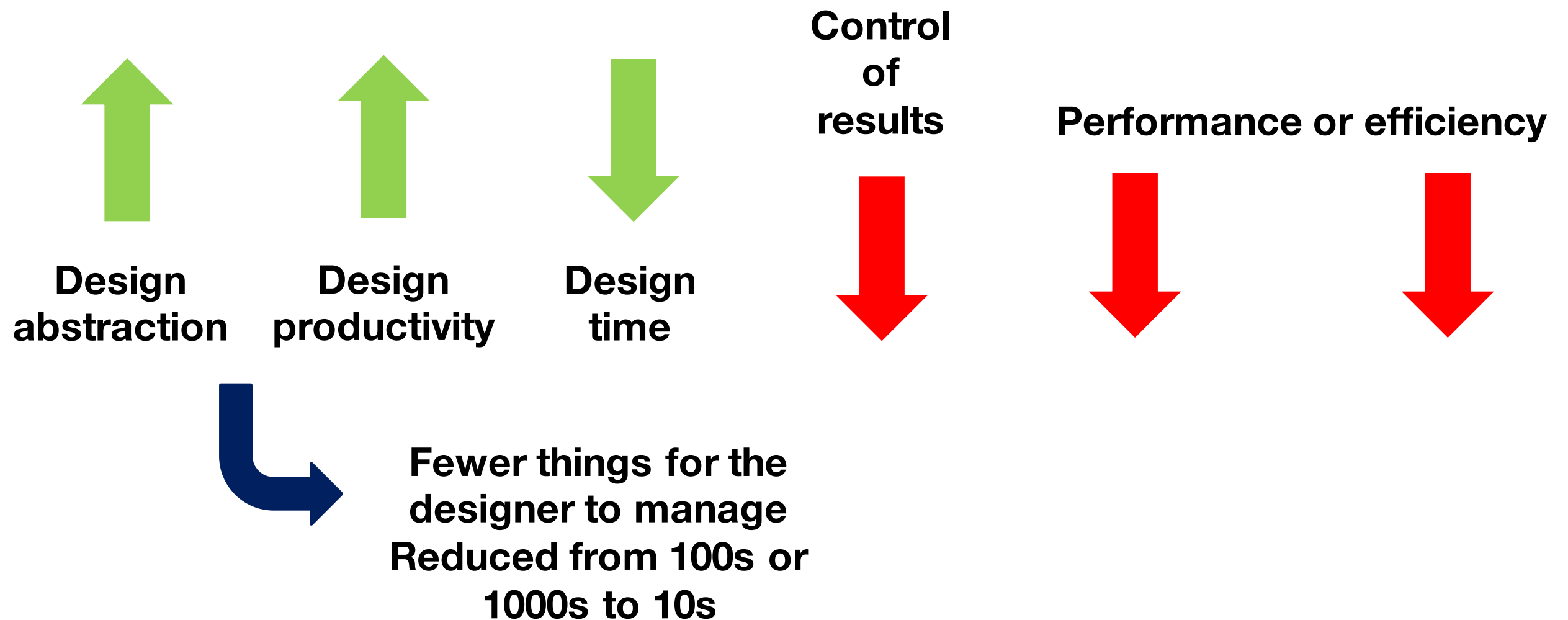
Virtex-Ultrasc



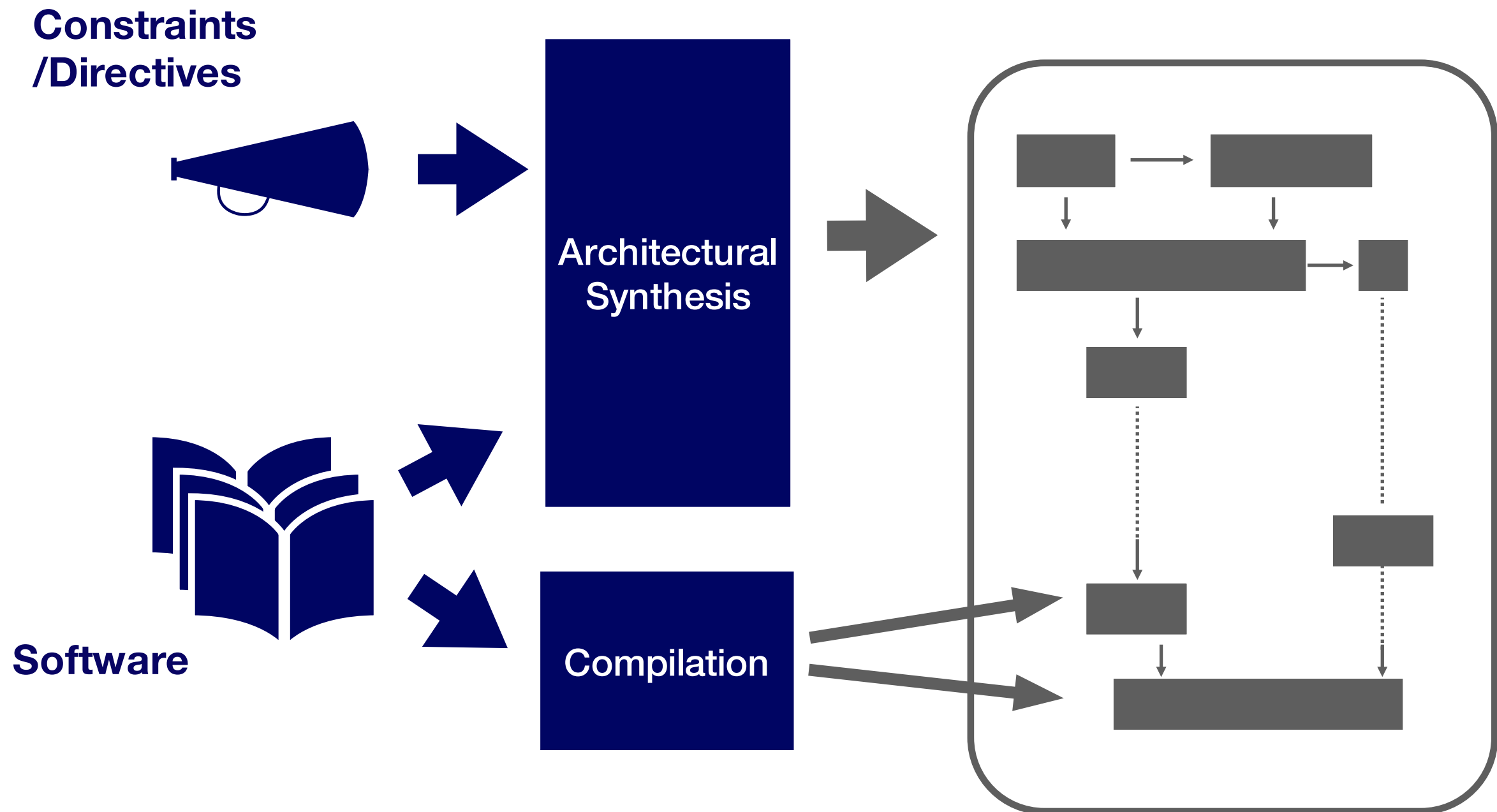
Synthesis



Why HLS?

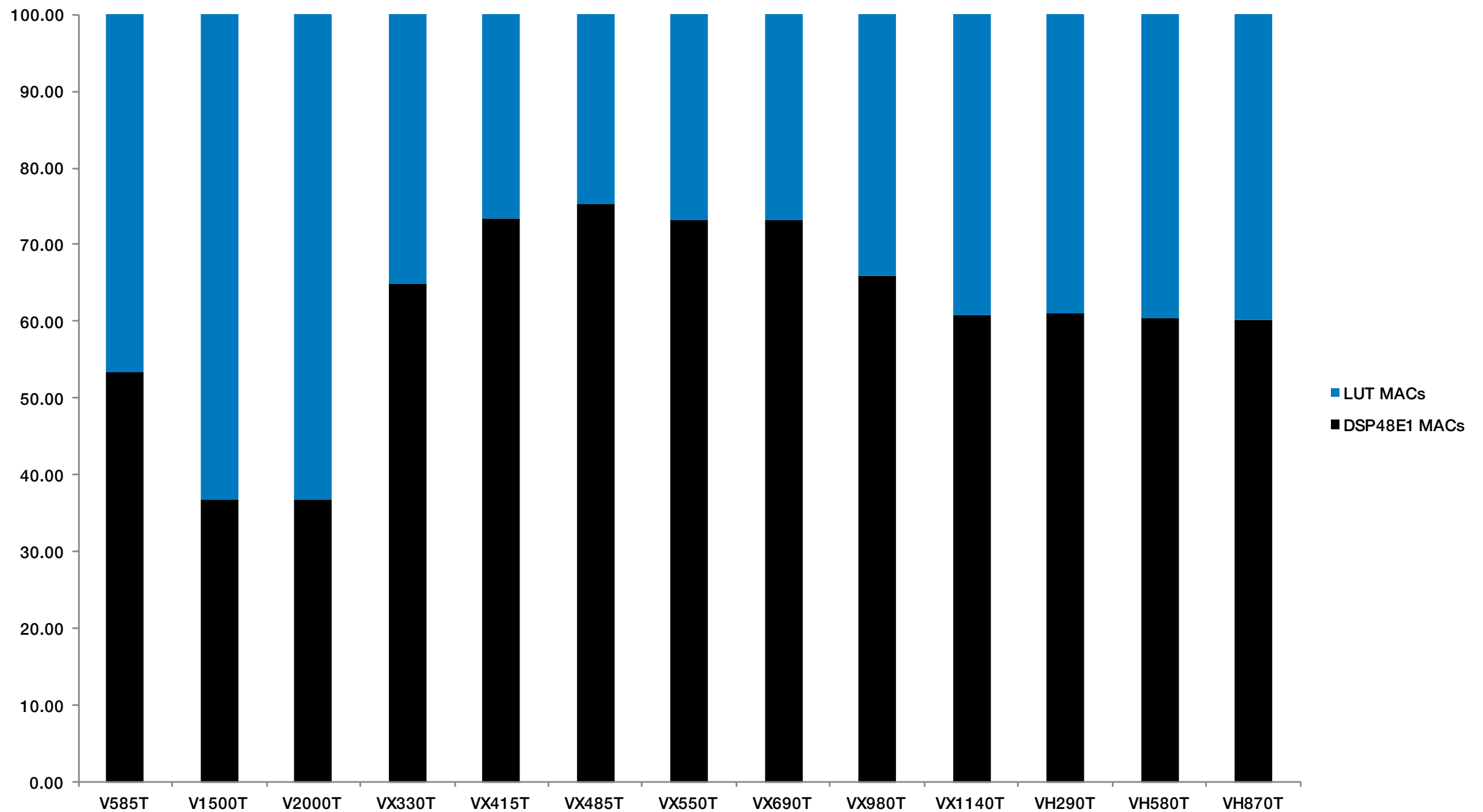


An Alternative

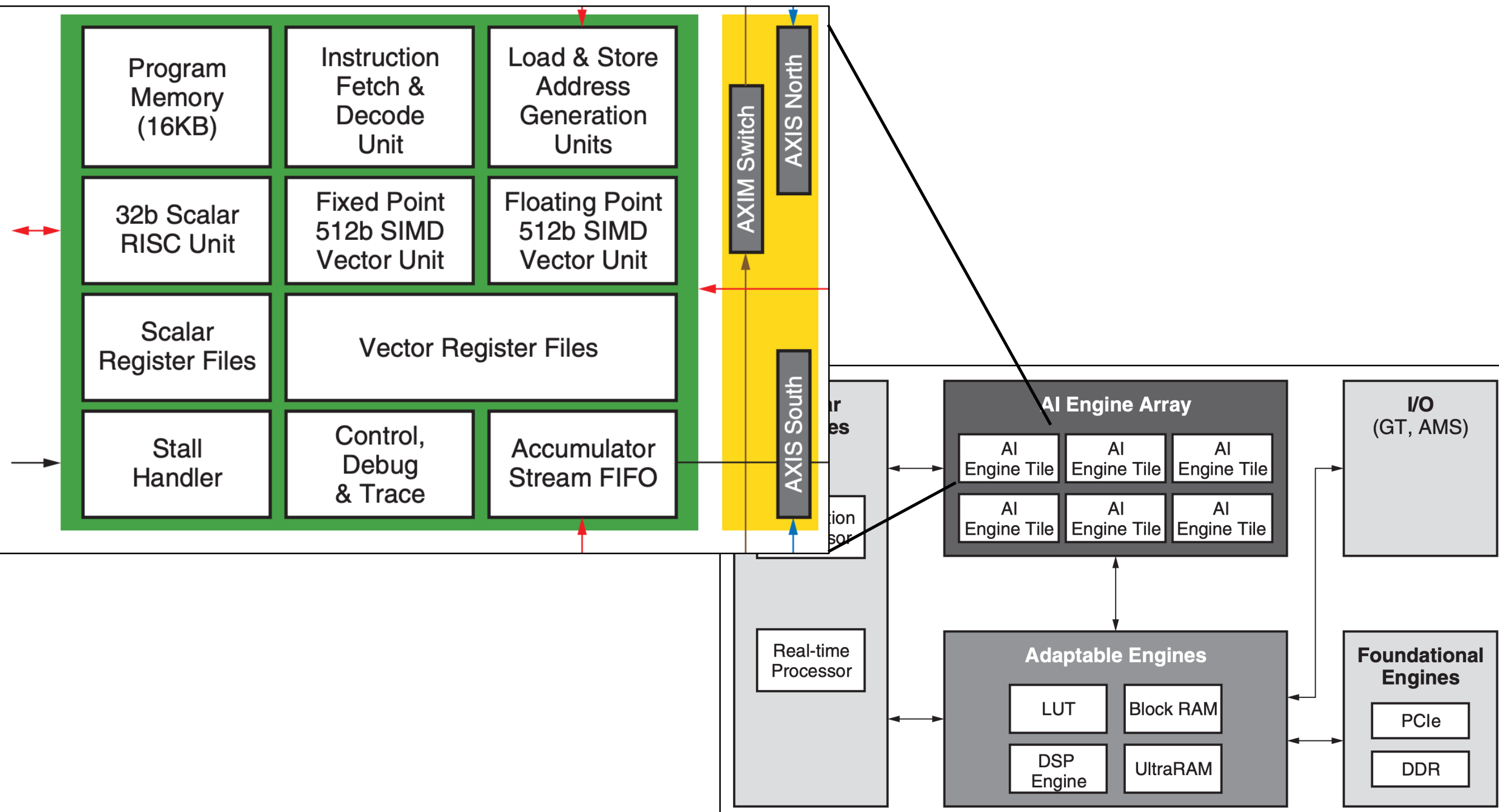


Processing Elements

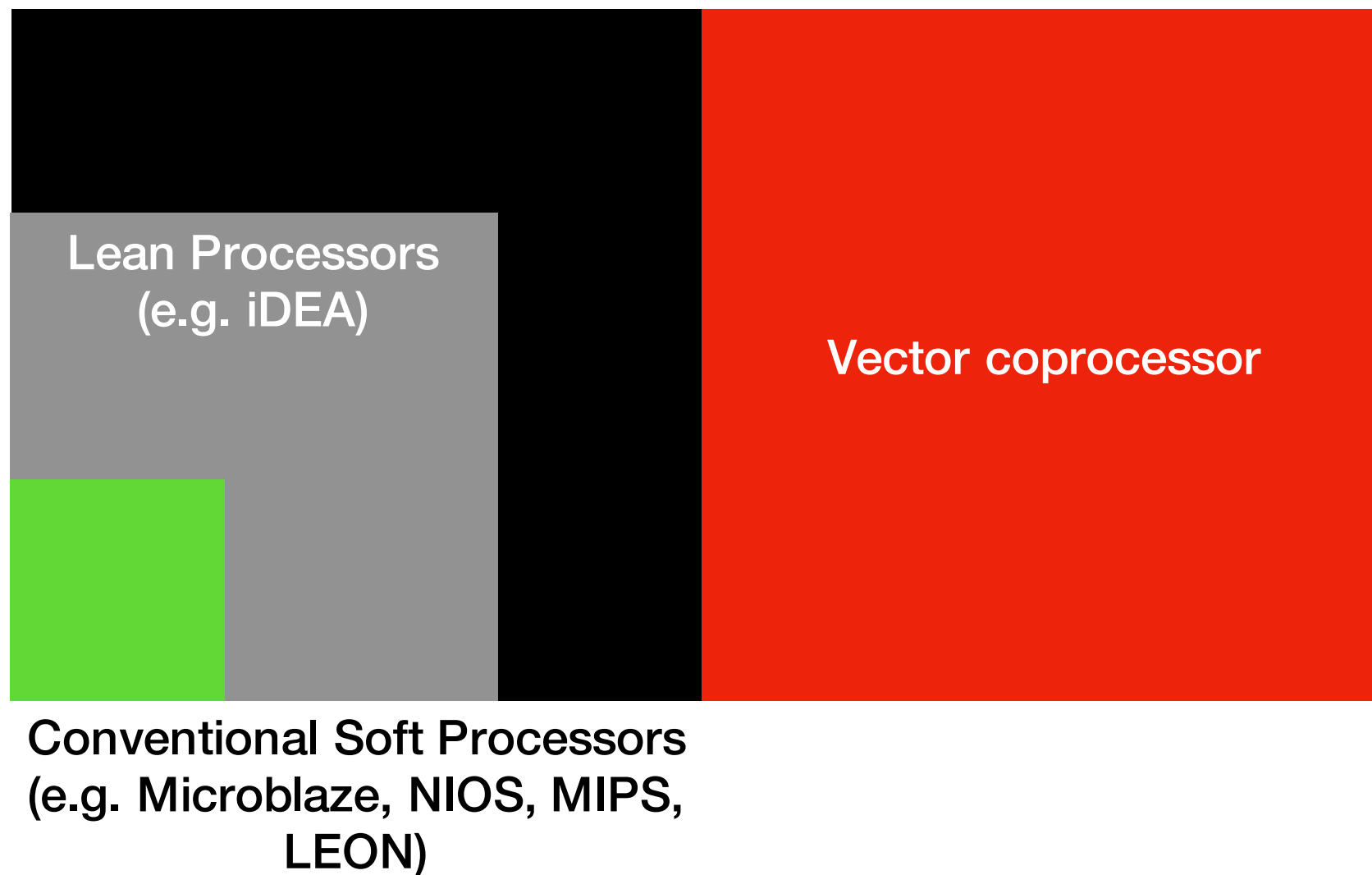
FPGA Compute



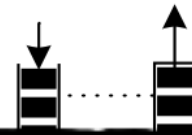
ACAPs



FPGA Processors

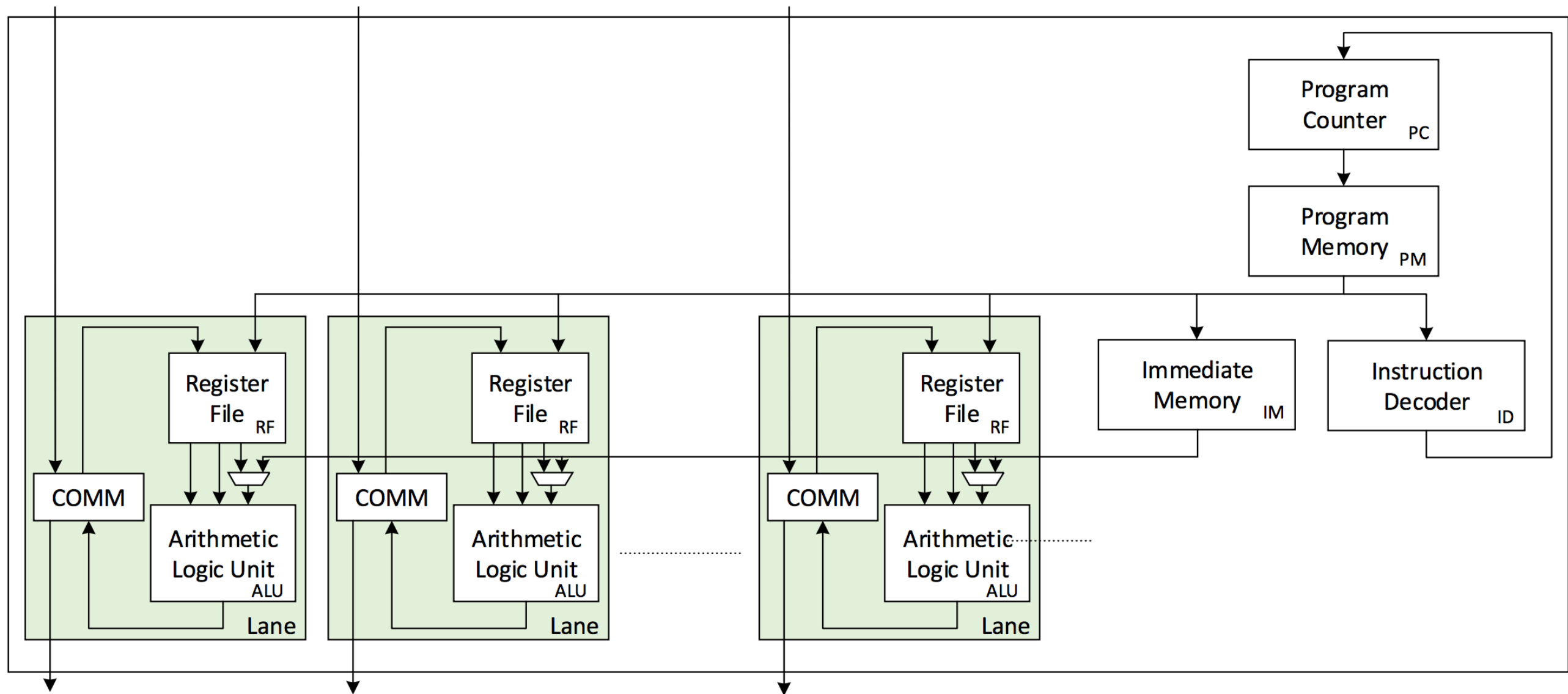


'Smart Dust' Processing Elements

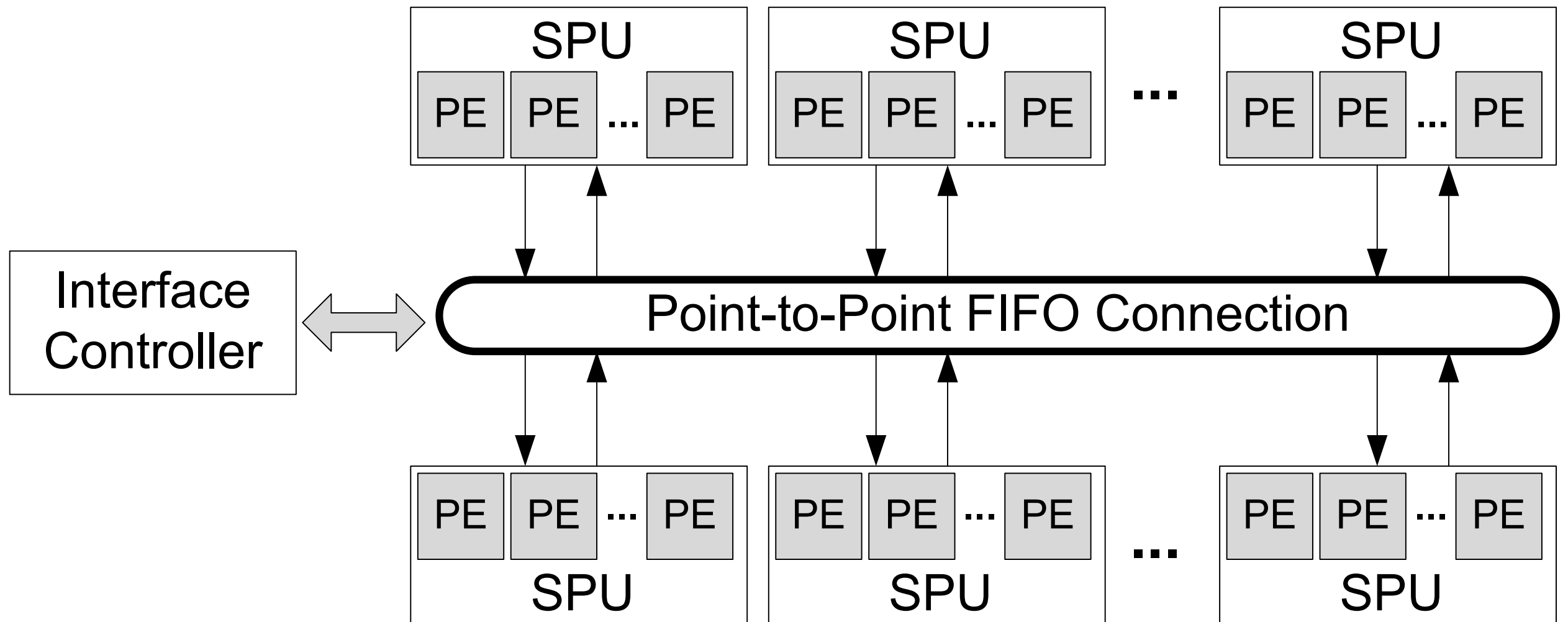


90 LUTs

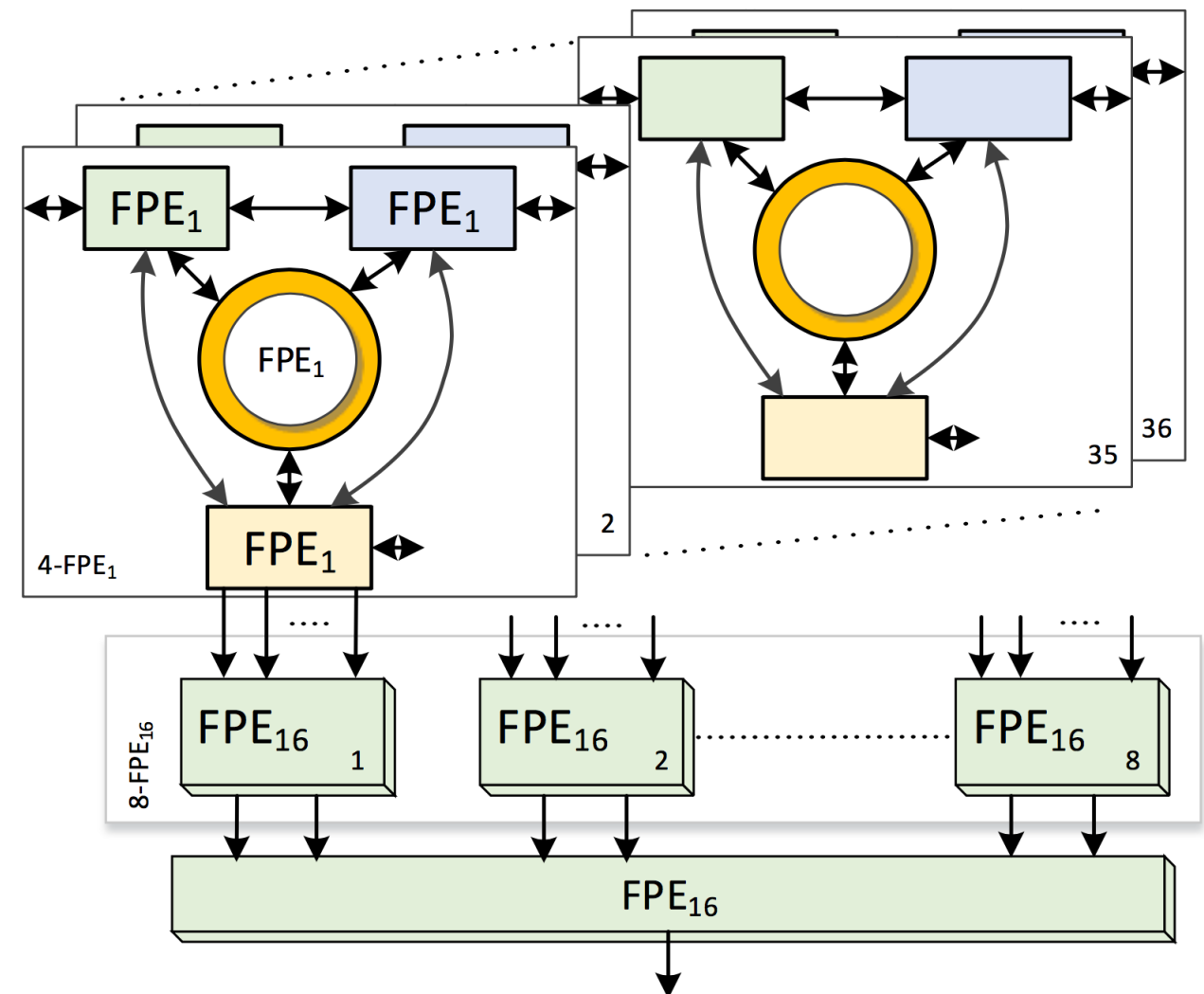
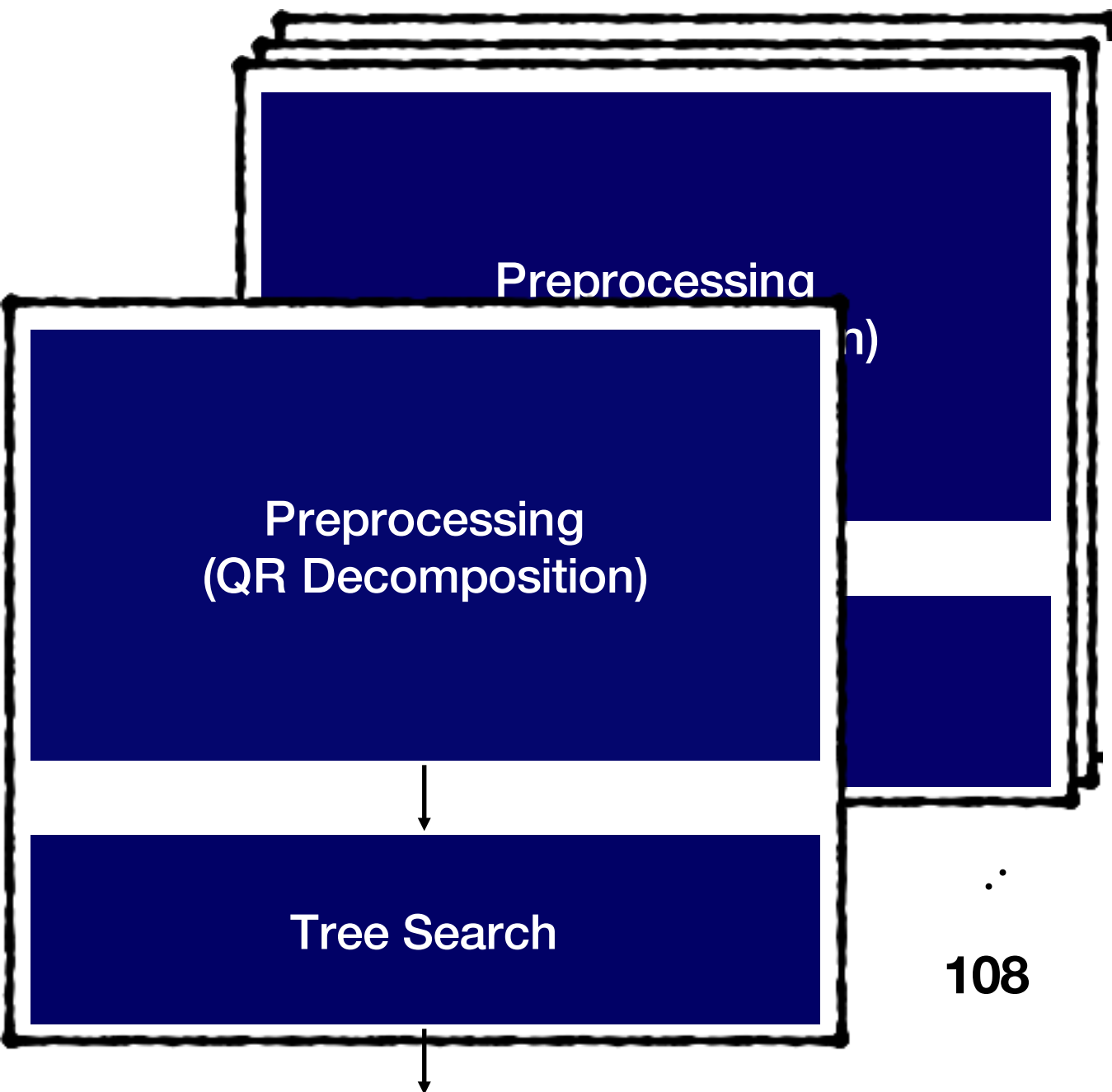
Scaling Up



Pile-em Up



When It Worked



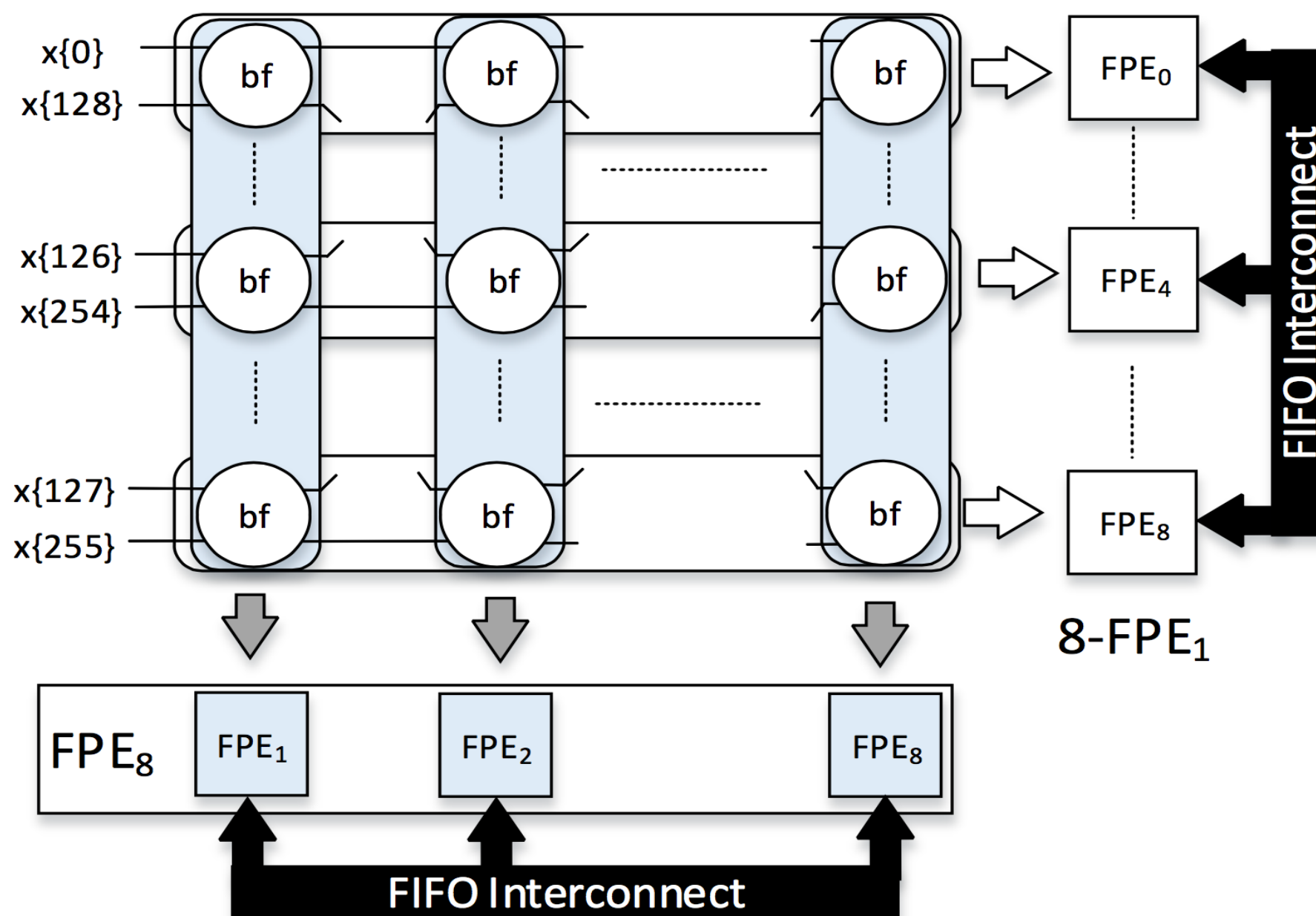
108

As Good As Custom RTL Circuits

Realisation	Throughput (Mbps)	DSP48e	LUTs	BRAM
FPE	502.5	144	16,601	0
Barbero & Thompson, ICC '08	600	160	13,197	49
Qi & Chakrabarti, SiPS '10	200	64	18,893	12
Wu & Masera, Euromicro DSD '10	27.7	0	6,587	0

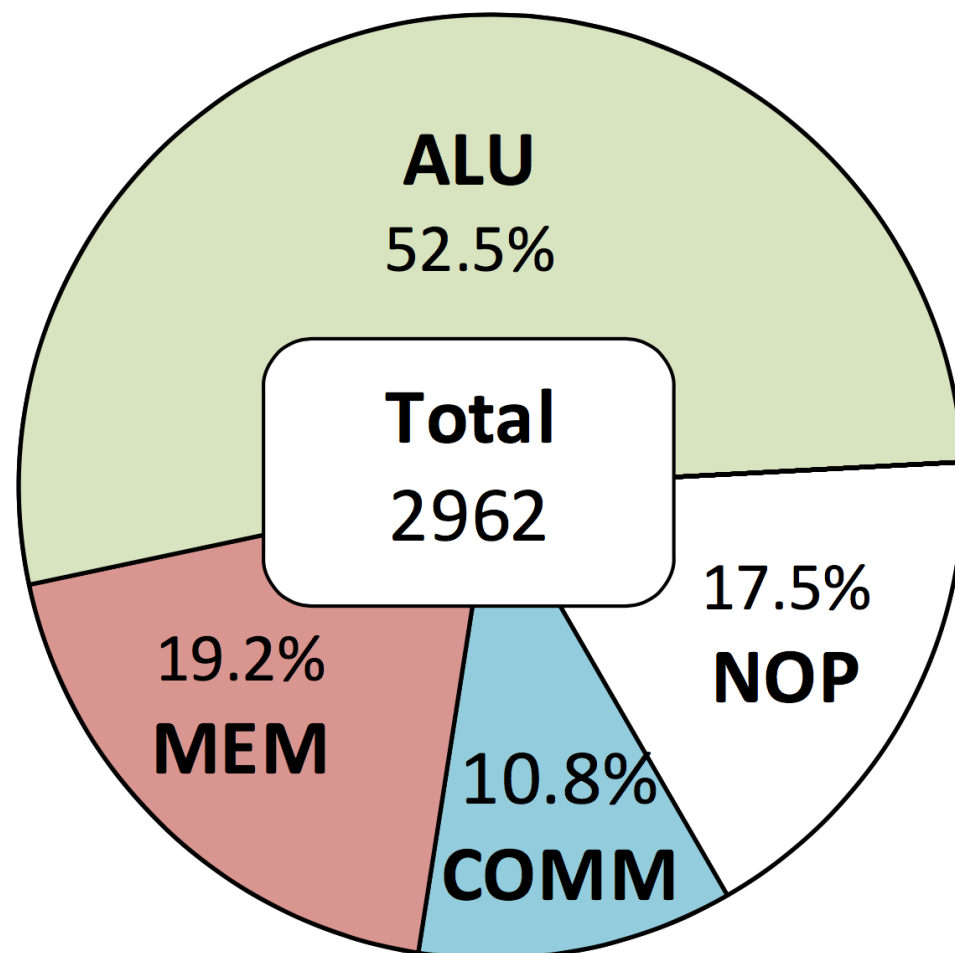
When It Didn't:

Low Compute/Communication Ratio

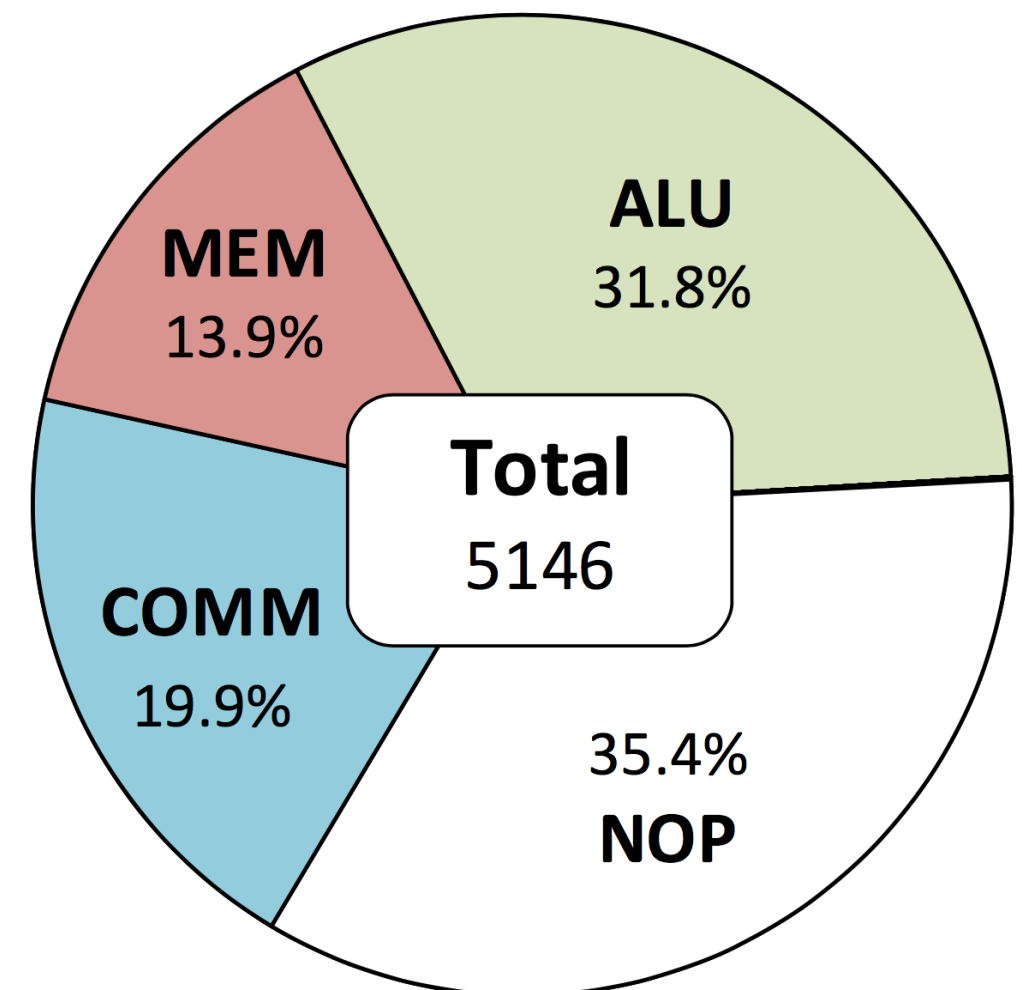


Low Compute/Data Access Ratio Is A Problem

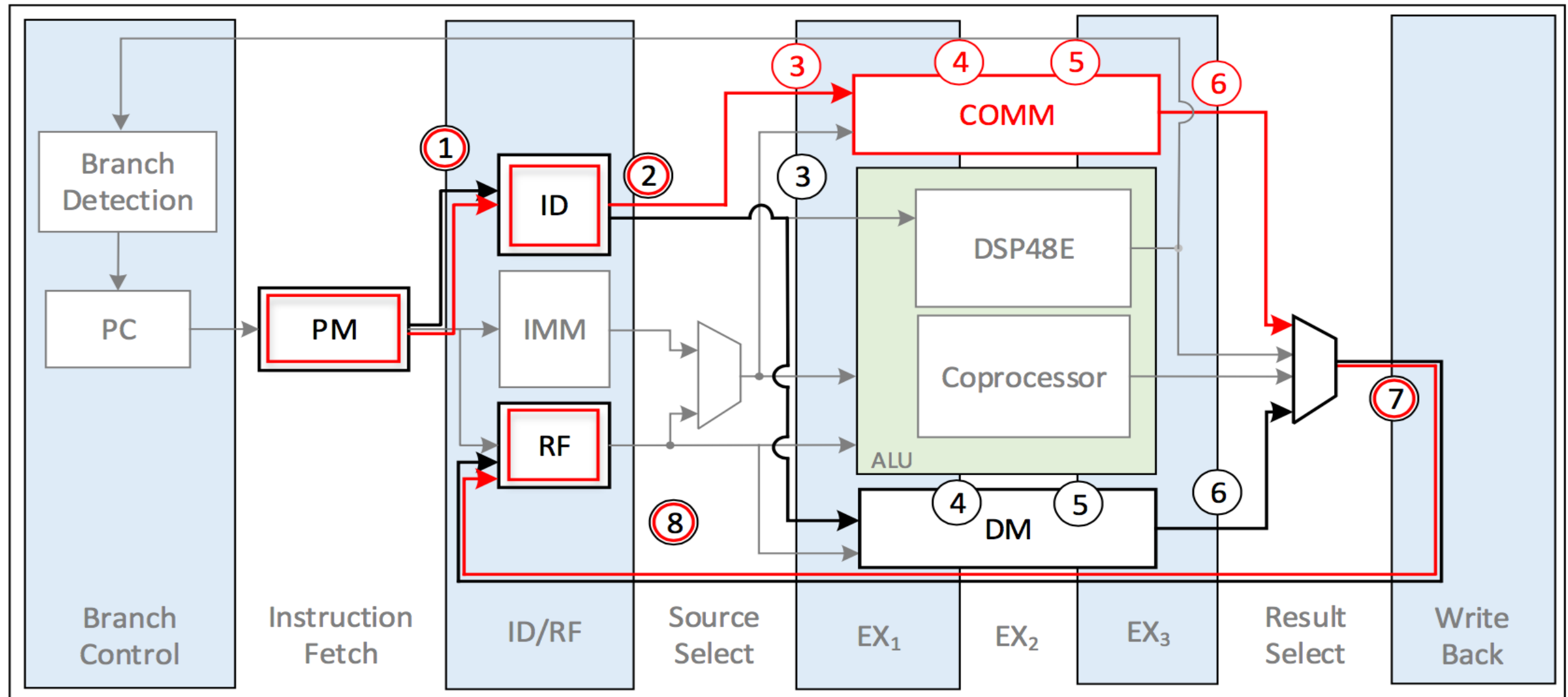
SIMD FFT



MIMD FFT

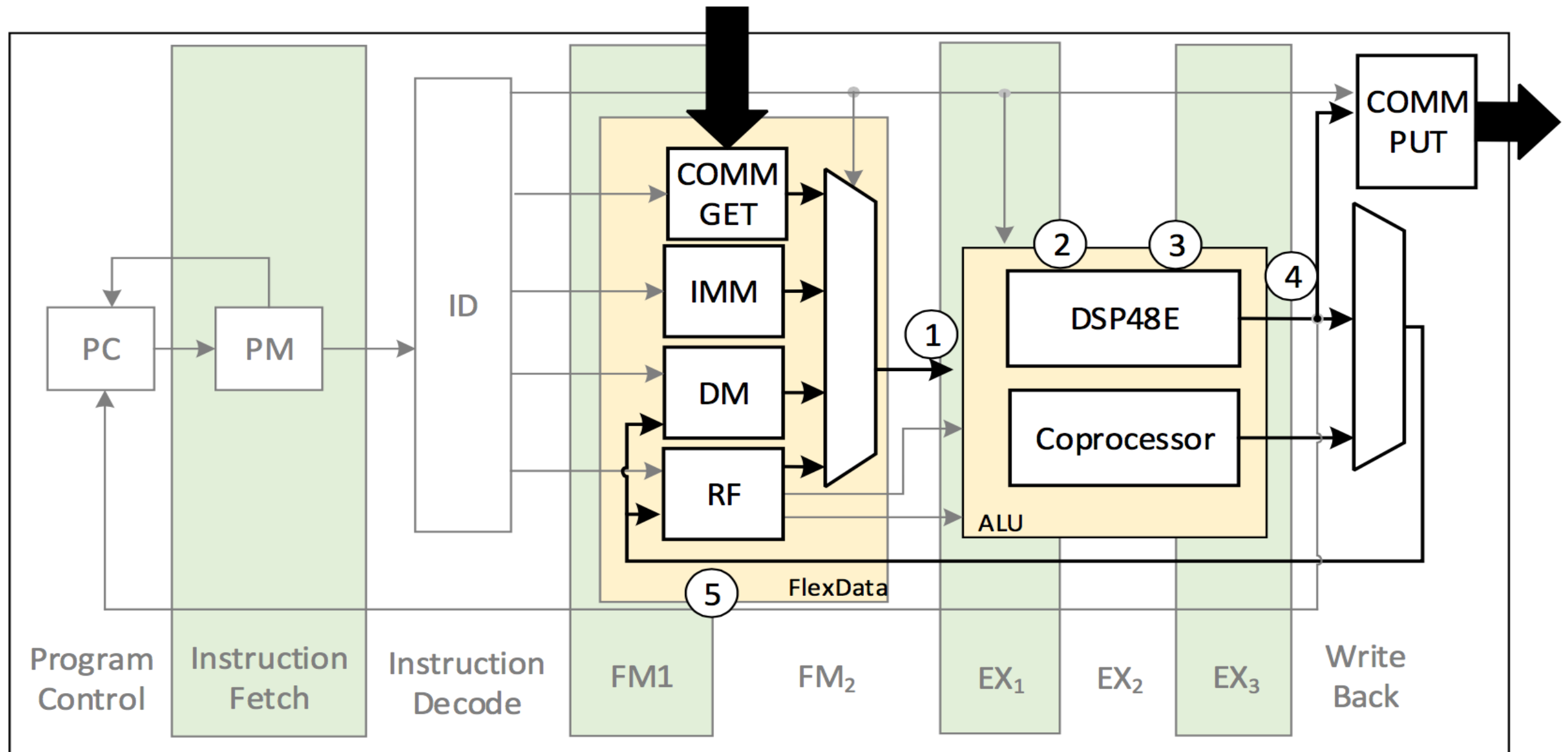


The Issue



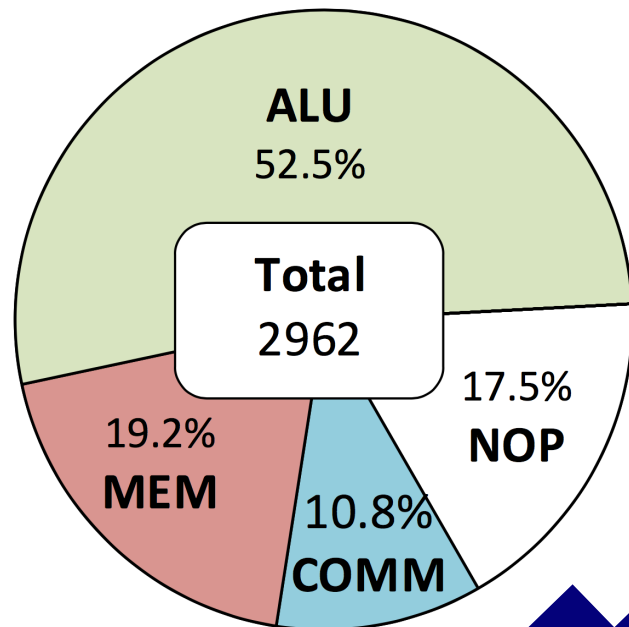
Streaming Processing Elements

Stream Processing

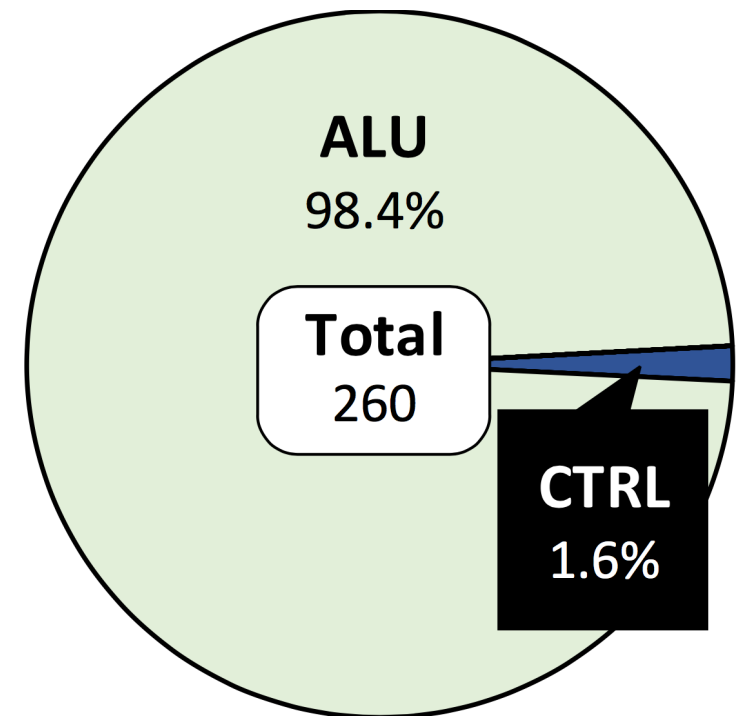
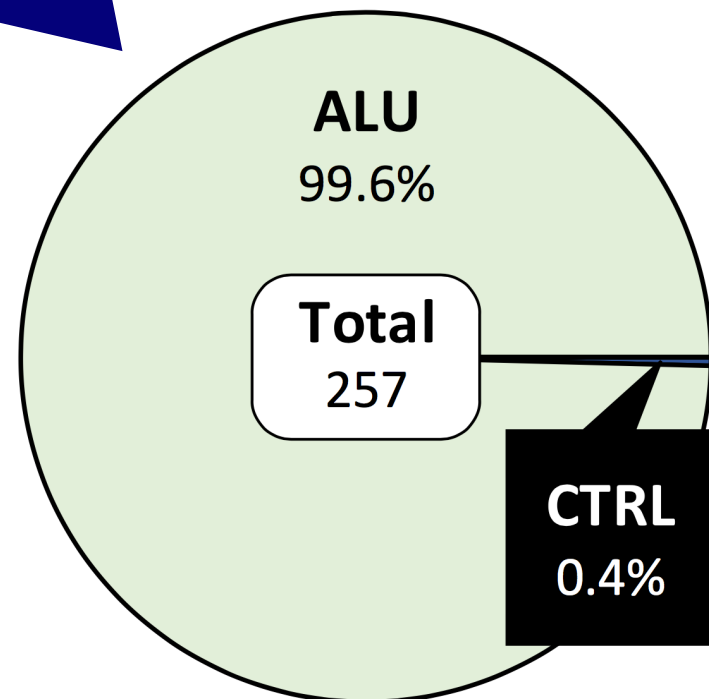
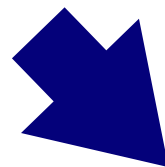
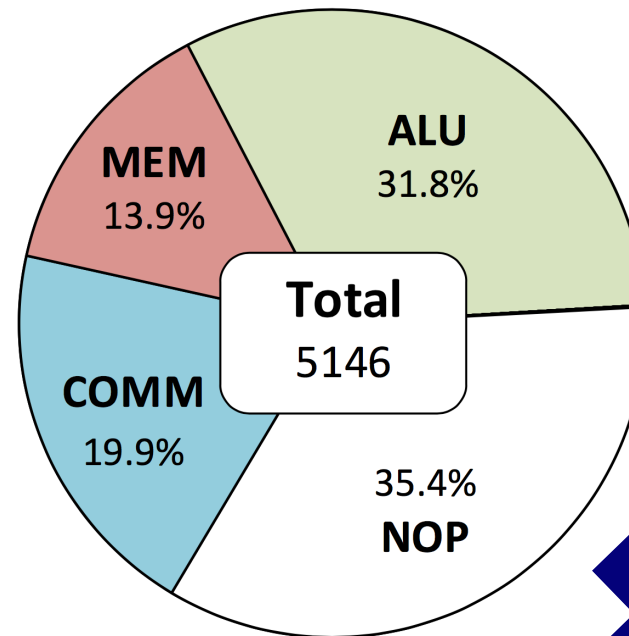


The Effect of Streaming

SIMD FFT

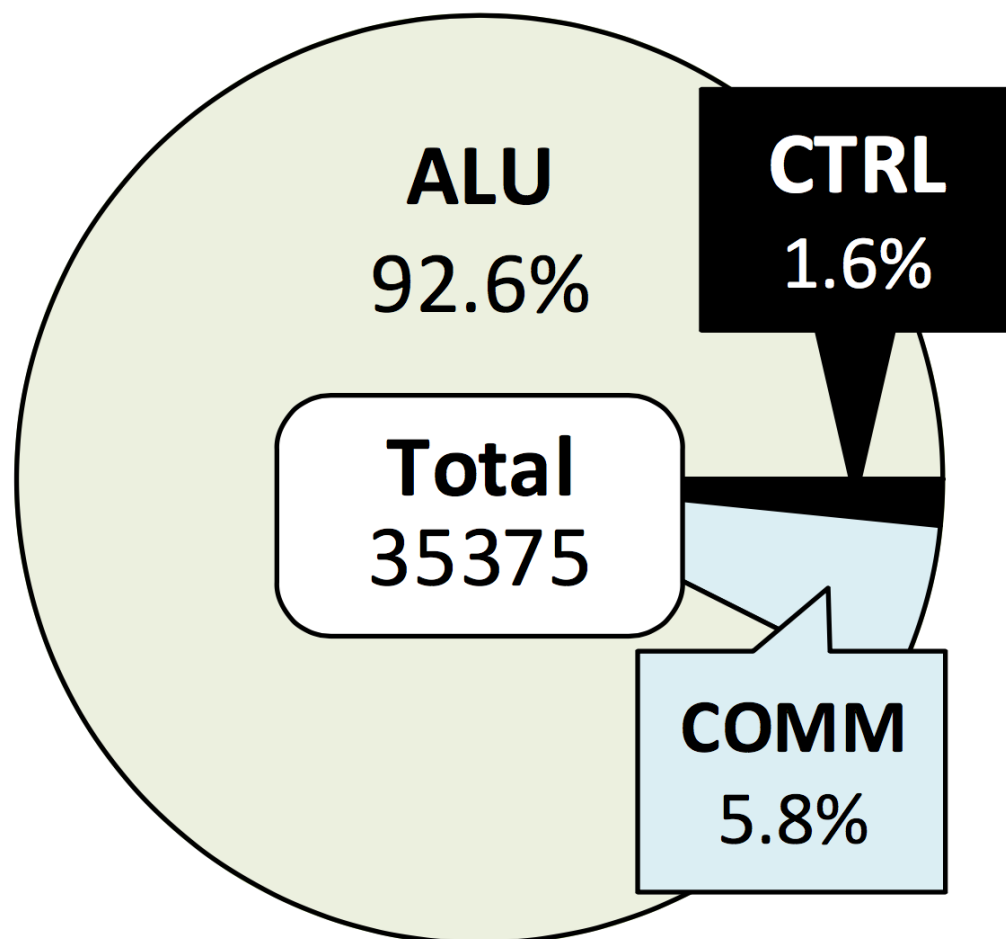


MIMD FFT

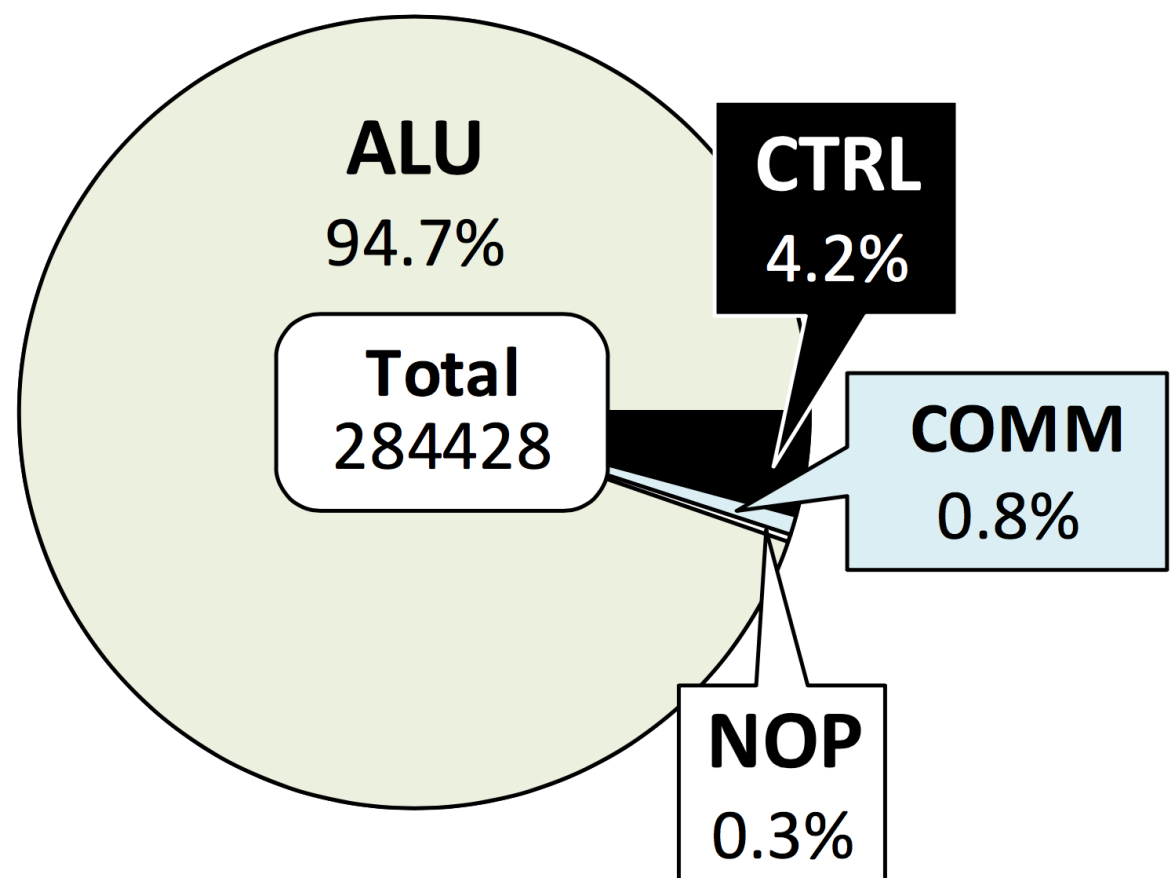


When It Didn't: Large Data Objects

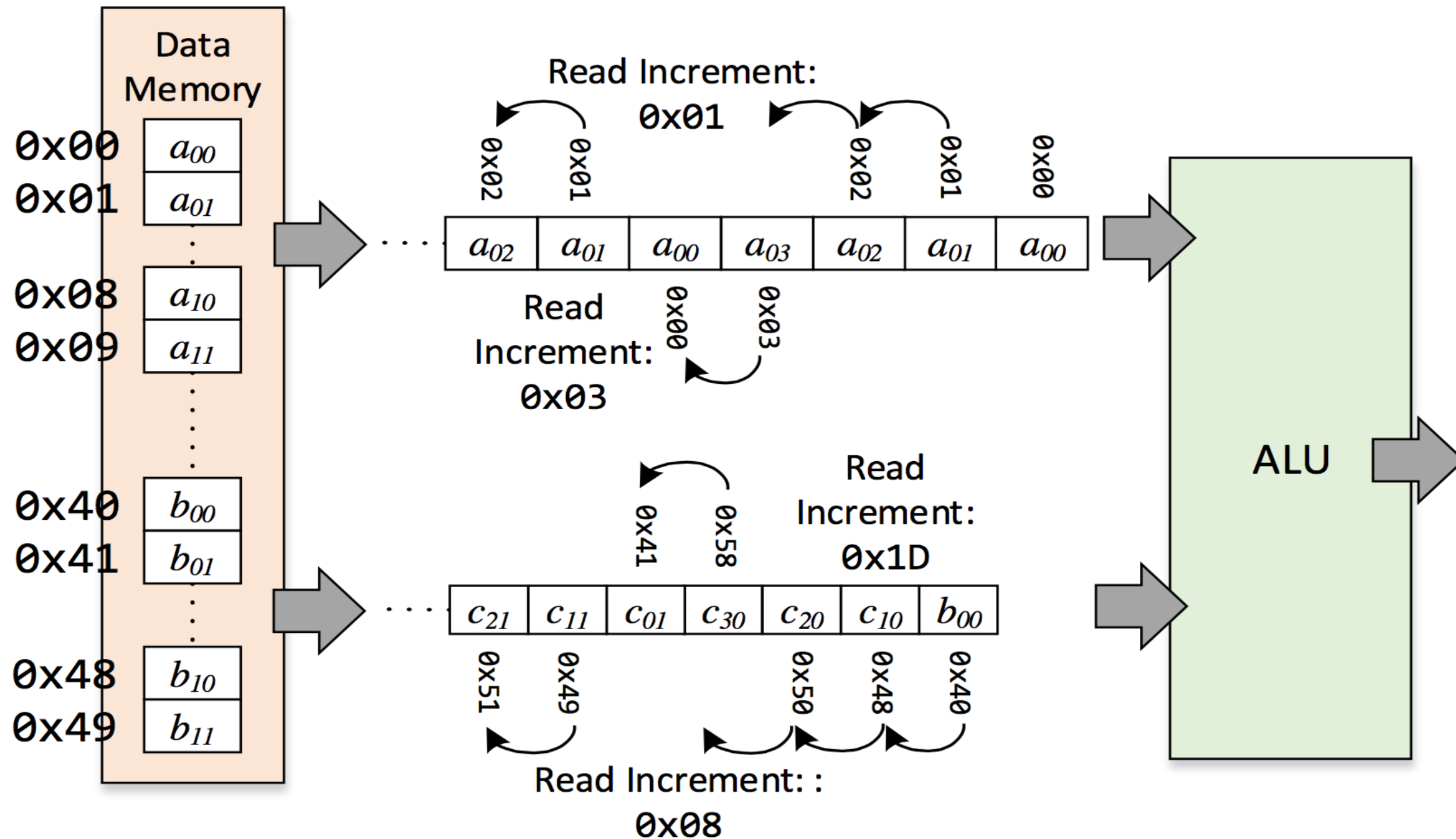
1024² Matrix-Matrix Multiplication



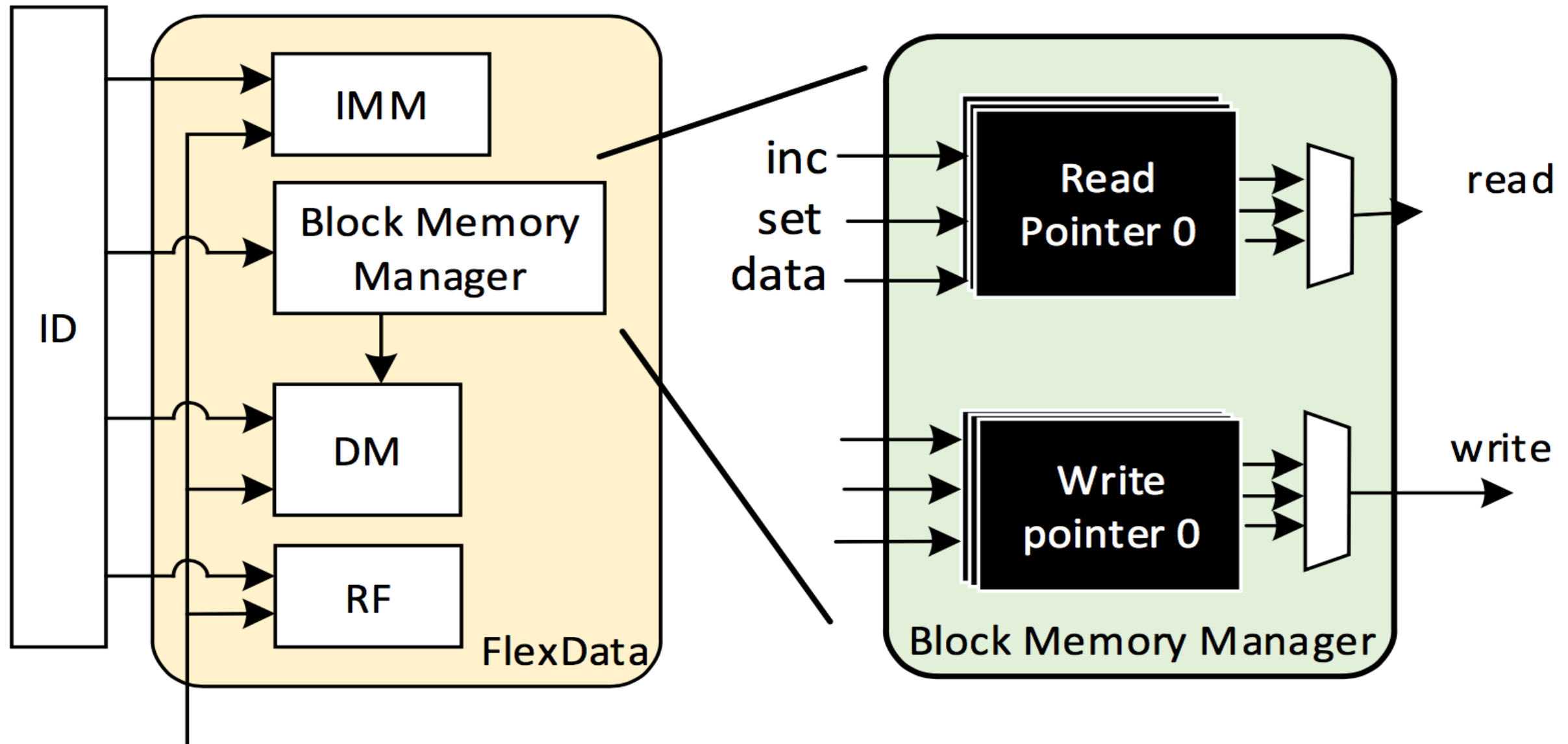
CIF Full Search Motion Estimation



Token Processing



Block Memory Access & Zero-Overhead Repeat



Dramatic Reductions in No. of Instructions

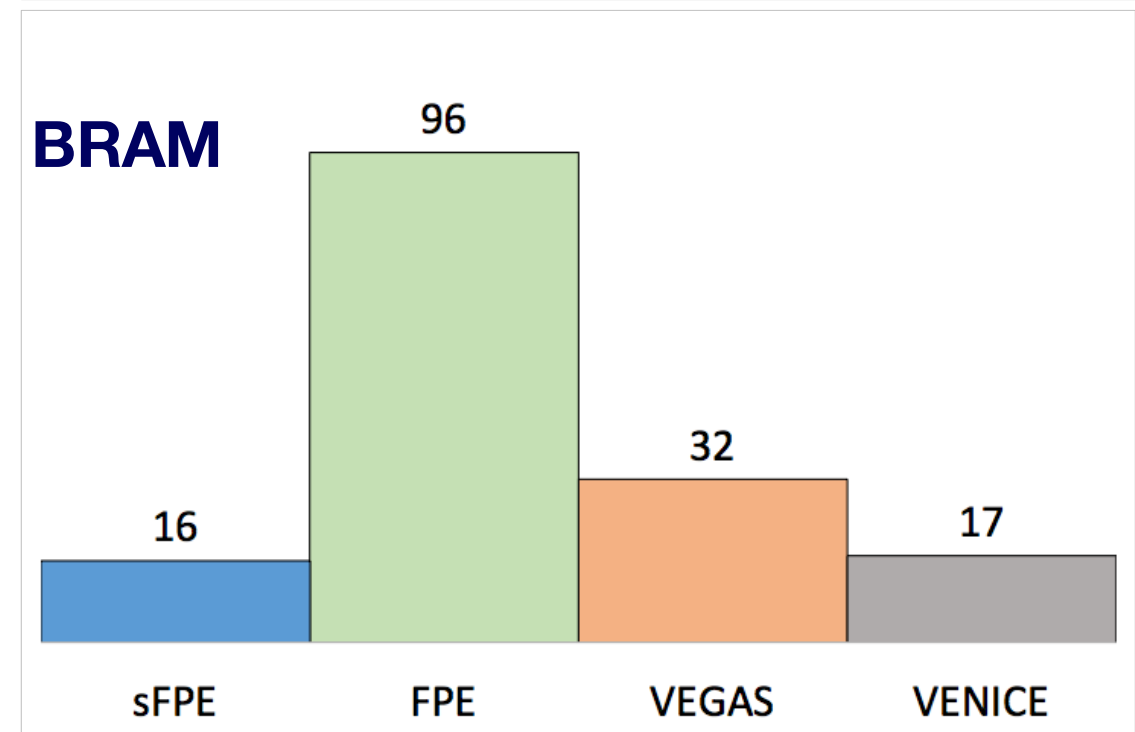
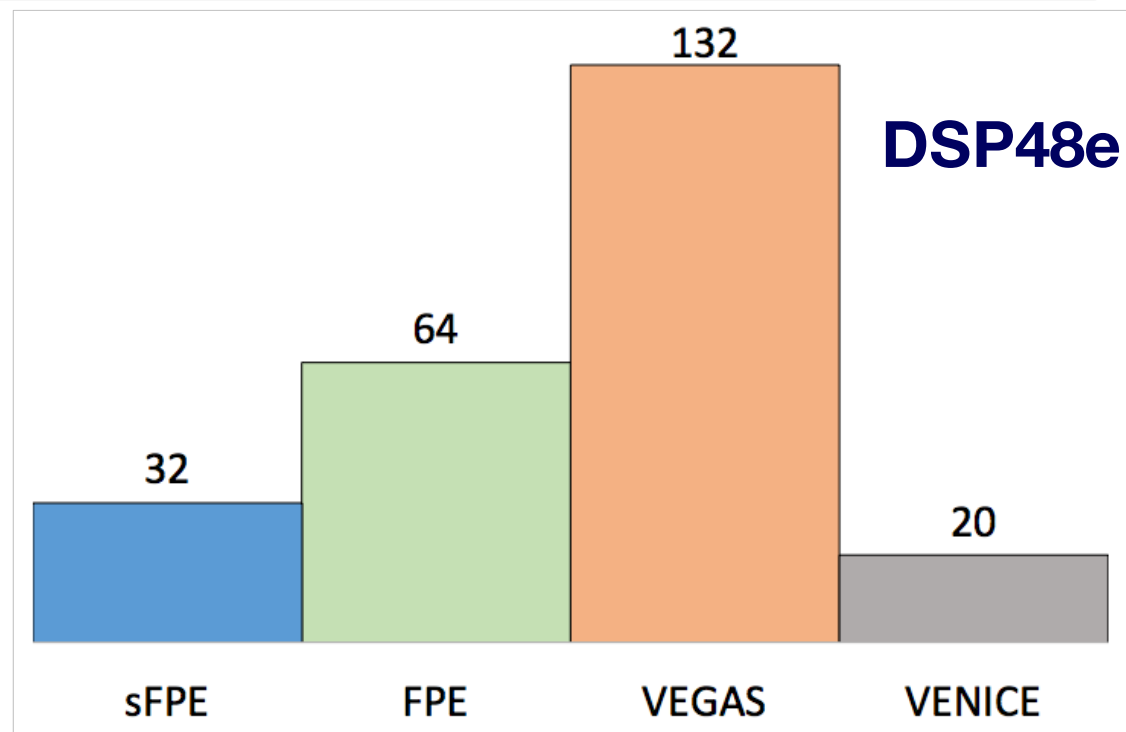
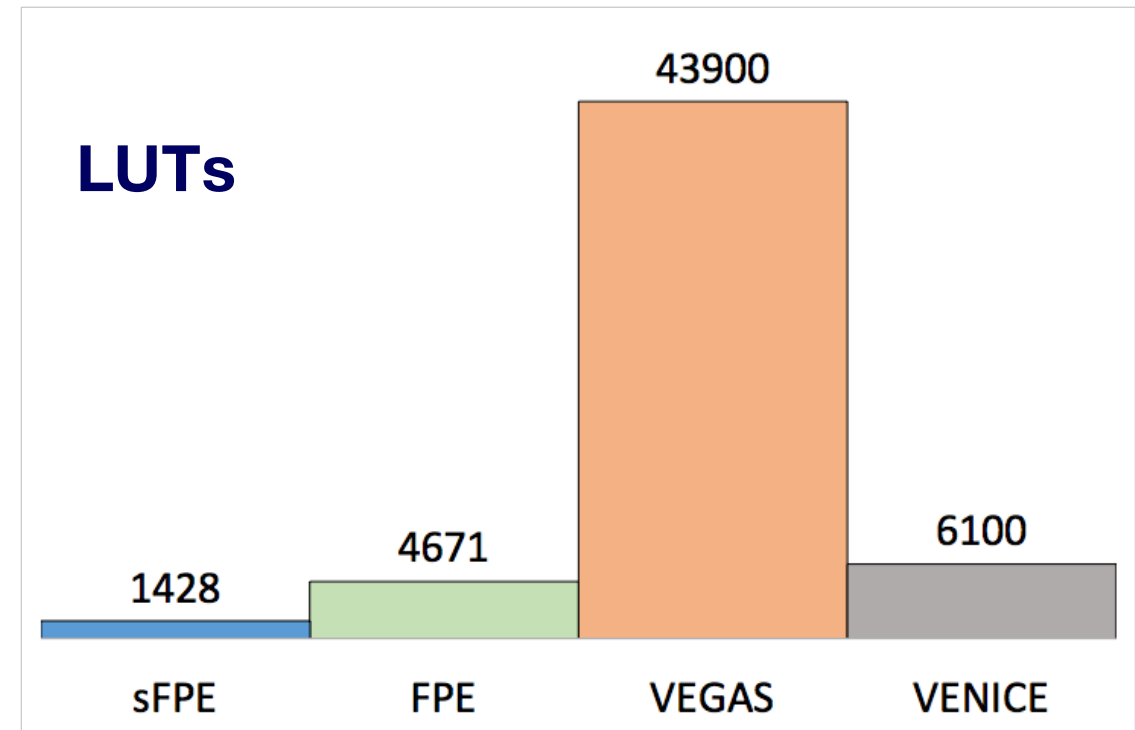
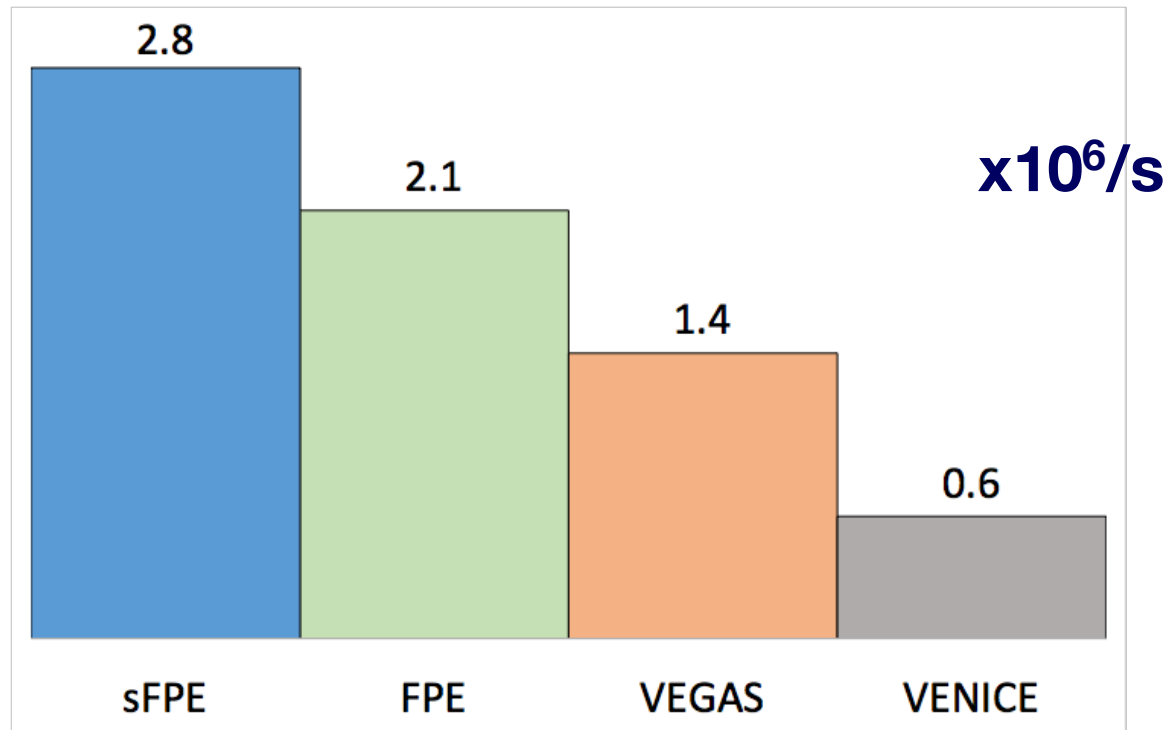
1024² Matrix-Matrix Multiplication

Class	FPE	sFPE	$\delta(\%)$
ALU	32768	32	-99.9
COMM	2048	6	-99.7
CTRL	559	4	-99.7
NOP	0	4	
Total	32375	54	-99.8

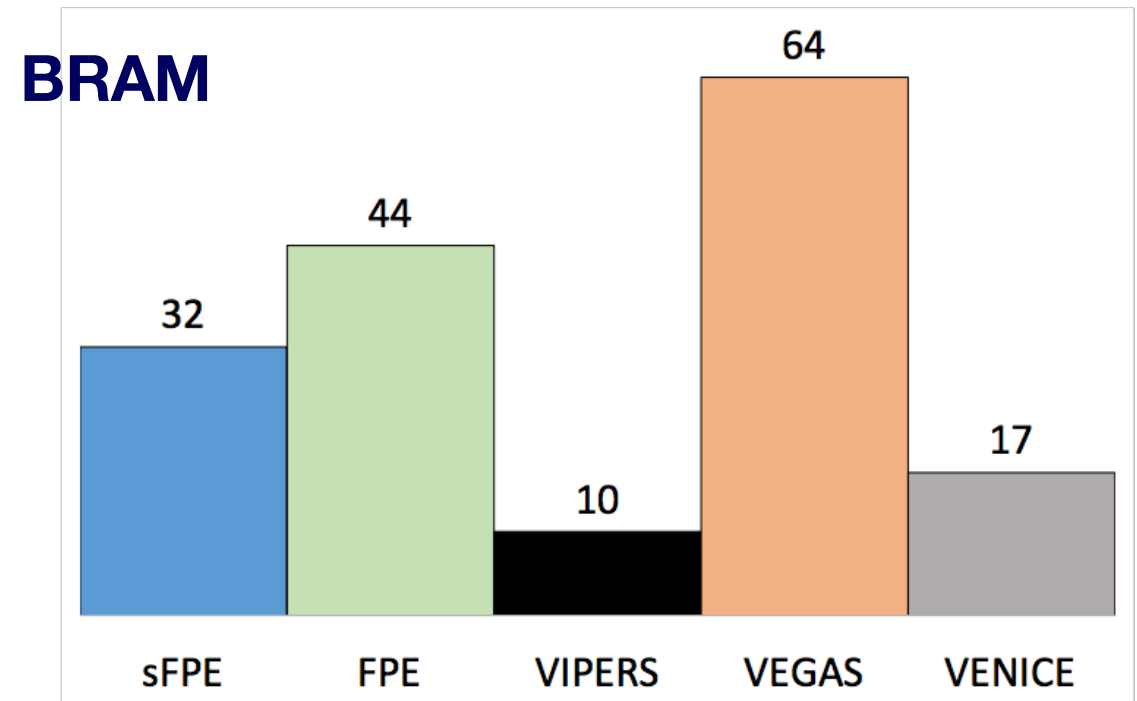
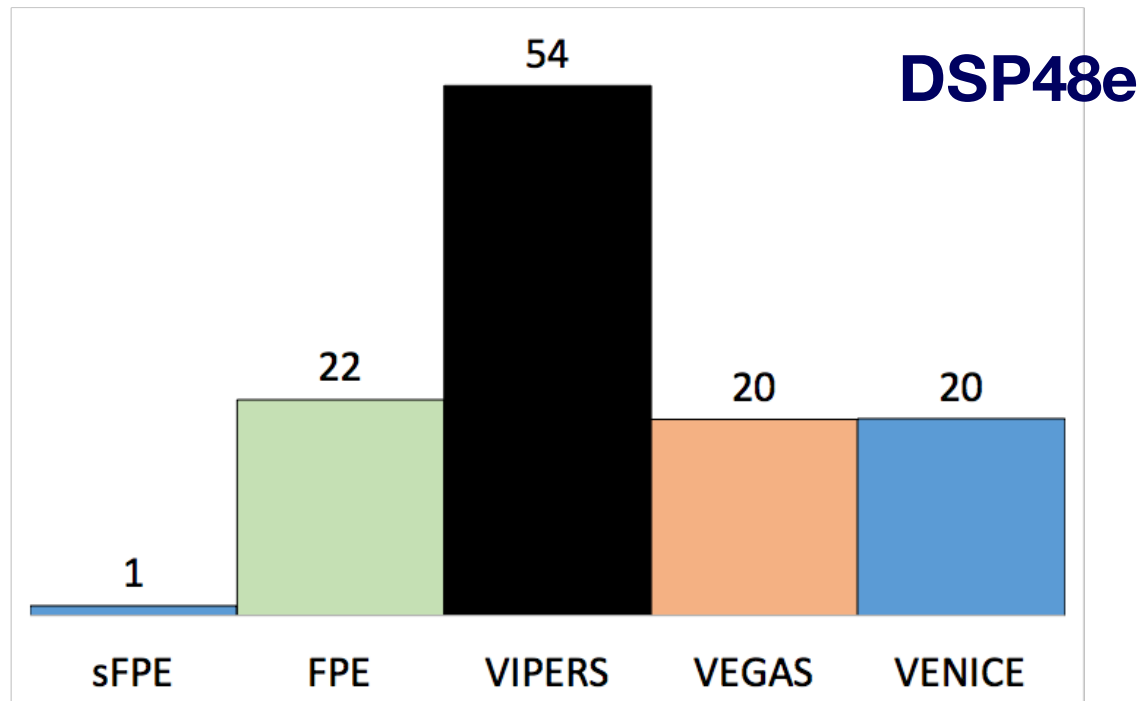
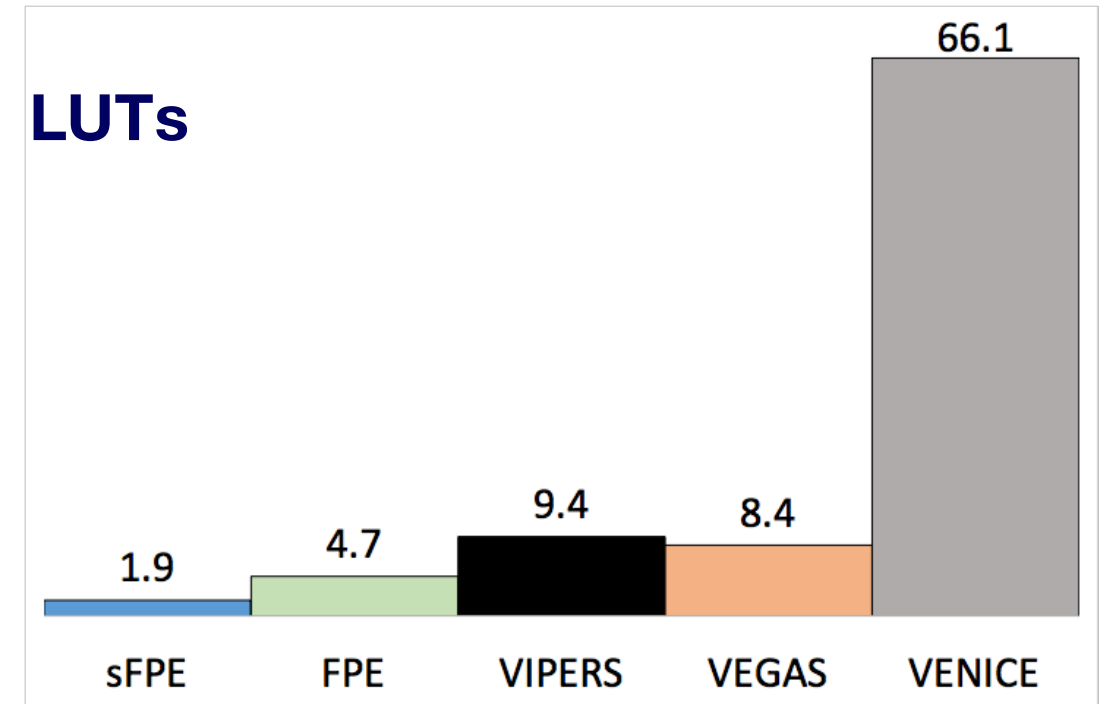
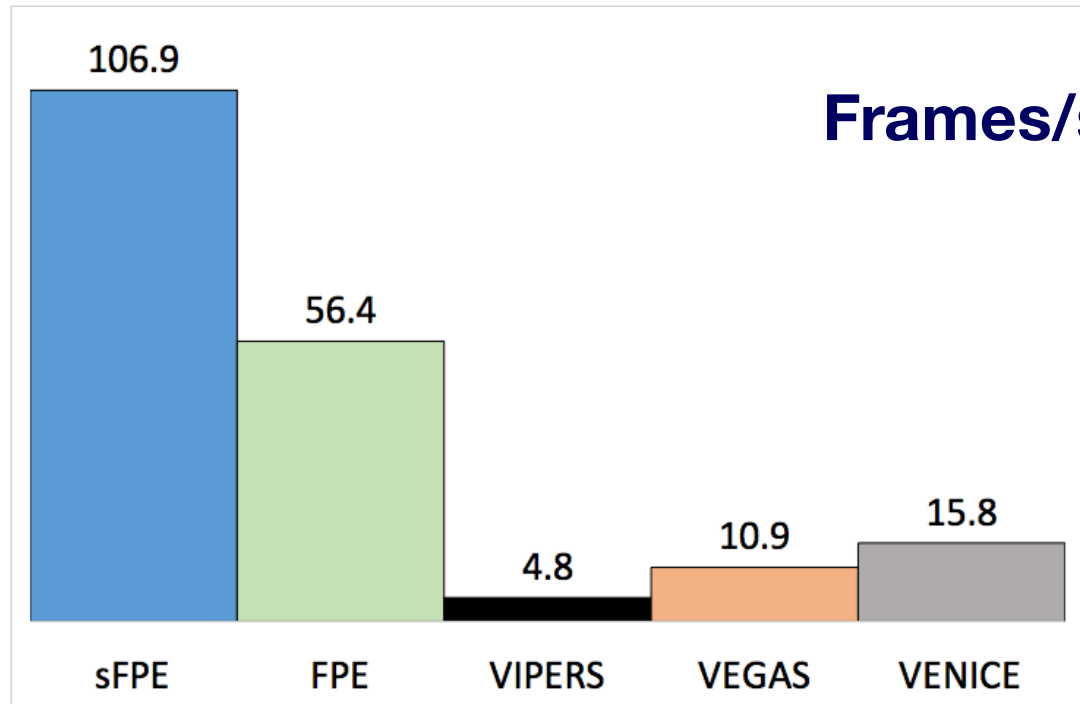
CIF Full Search Motion Estimation

Class	FPE	sFPE	$\delta(\%)$
ALU	268353	26	-99.9
COMM	2467	14	-99.4
CTRL	12582	12	-99.9
NOP	1026	6	-99.6
Total	284428	58	-99.9

1024² Matrix Multiplication

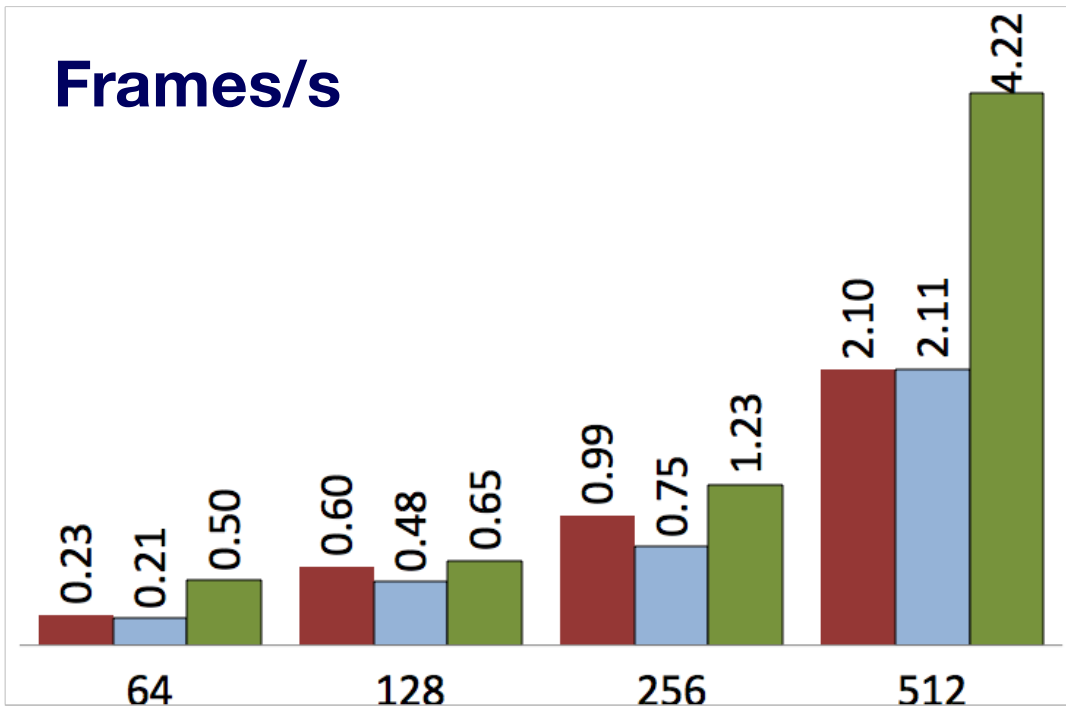


Full Search Motion Estimation

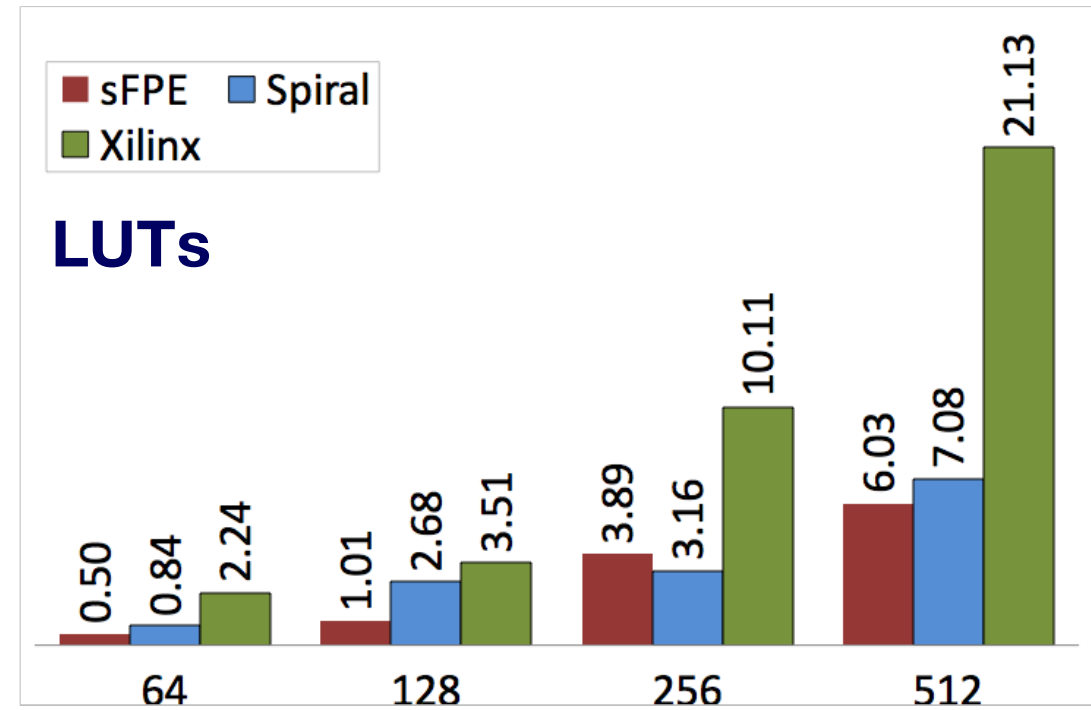


FFTs

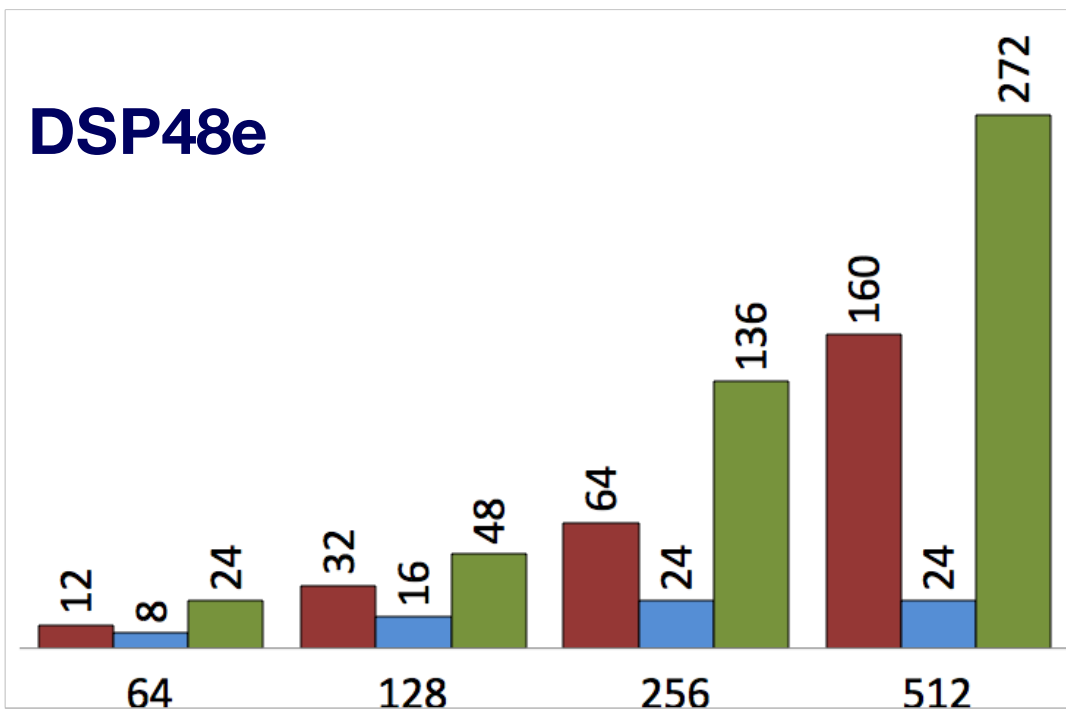
Frames/s



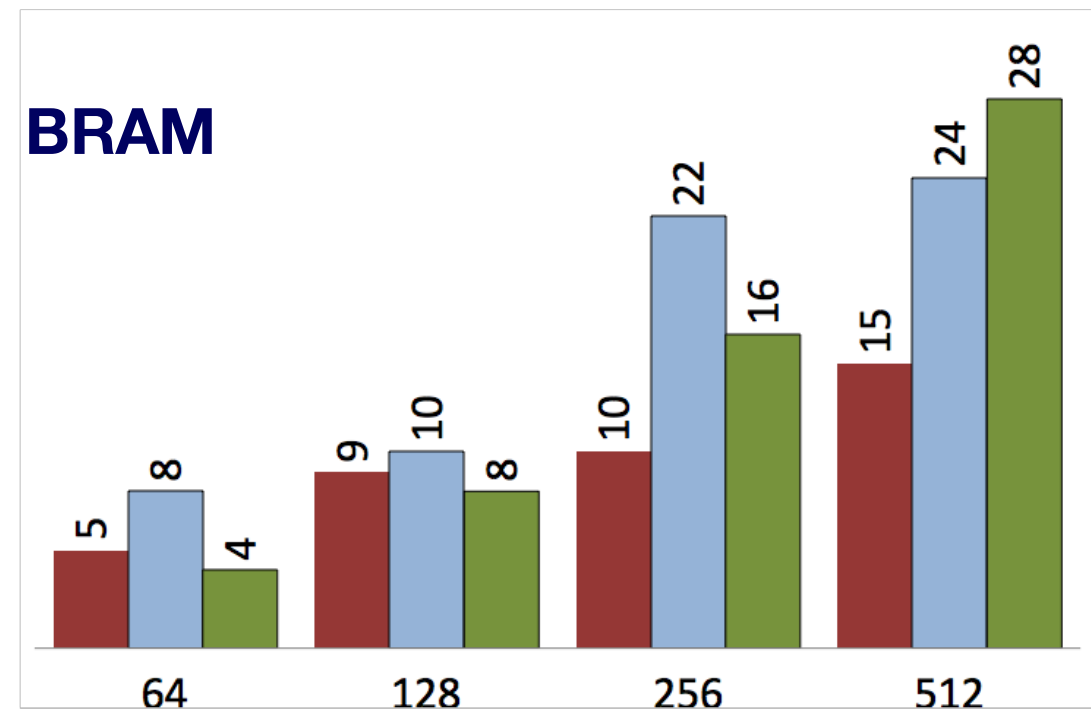
LUTs



DSP48e



BRAM



Summary

HLS undermines
a key reason for
using FPGA.

Goal: productivity gain
with performance/cost
benefit.

Are there others?

Domain-specific, configurable
and programmable RTL
components?

Good performance/cost,
much greater
productivity

One instance: multicores,
requiring a designer to handle
tens of components