

# Stream: Exploring Layer-Fused Mapping of DNNs on Heterogeneous Dataflow Accelerators

Arne Symons and Marian Verhelst

MICAS, Department of Electrical Engineering (ESAT), KU Leuven, Belgium

Paper

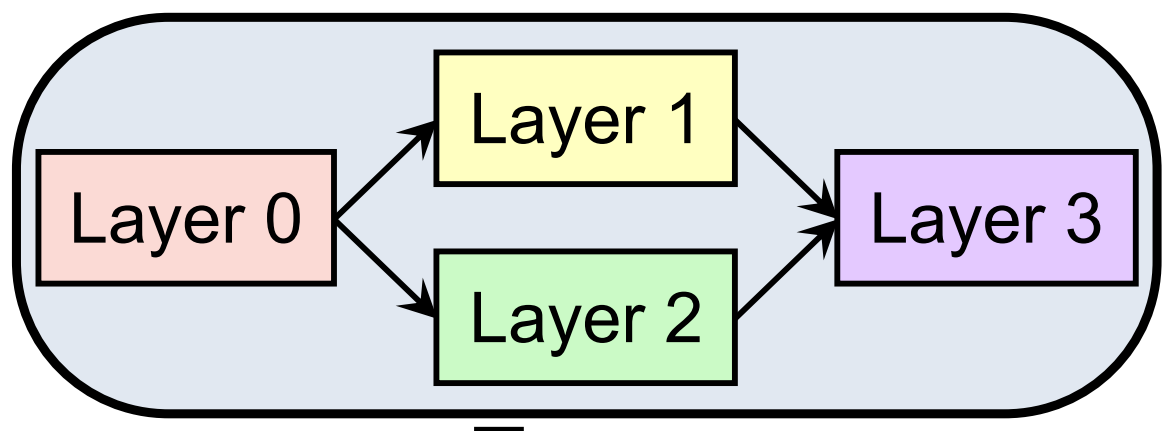


Code



## Layer Fusion

Fuse execution of many DNN layers



Fuse

**Opportunities:**

1. Reduced off-chip accesses

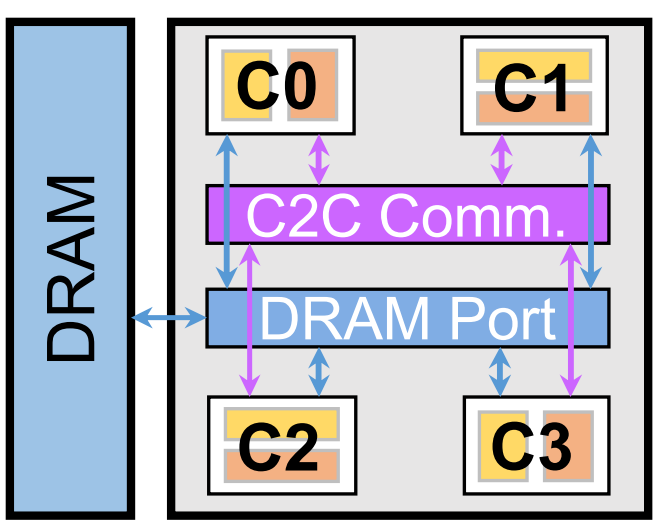
**Challenges:**

1. Which layers to fuse

2. Smaller data reuse in tiles

## Heterogeneous Dataflow Accelerators

Multiple cores with distinct dataflows



**Opportunities:**

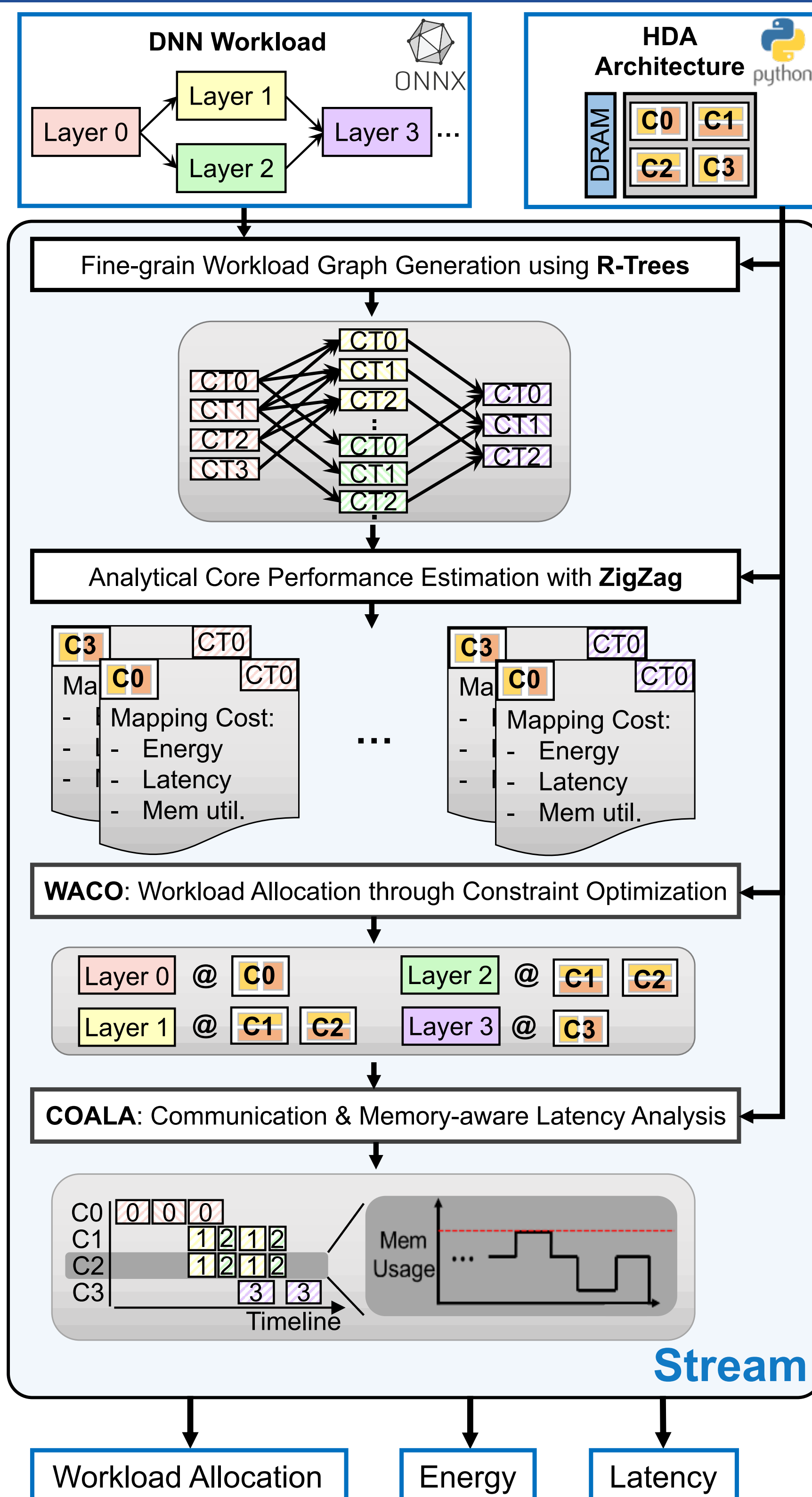
1. Specialized core types

2. Higher utilization for small ops

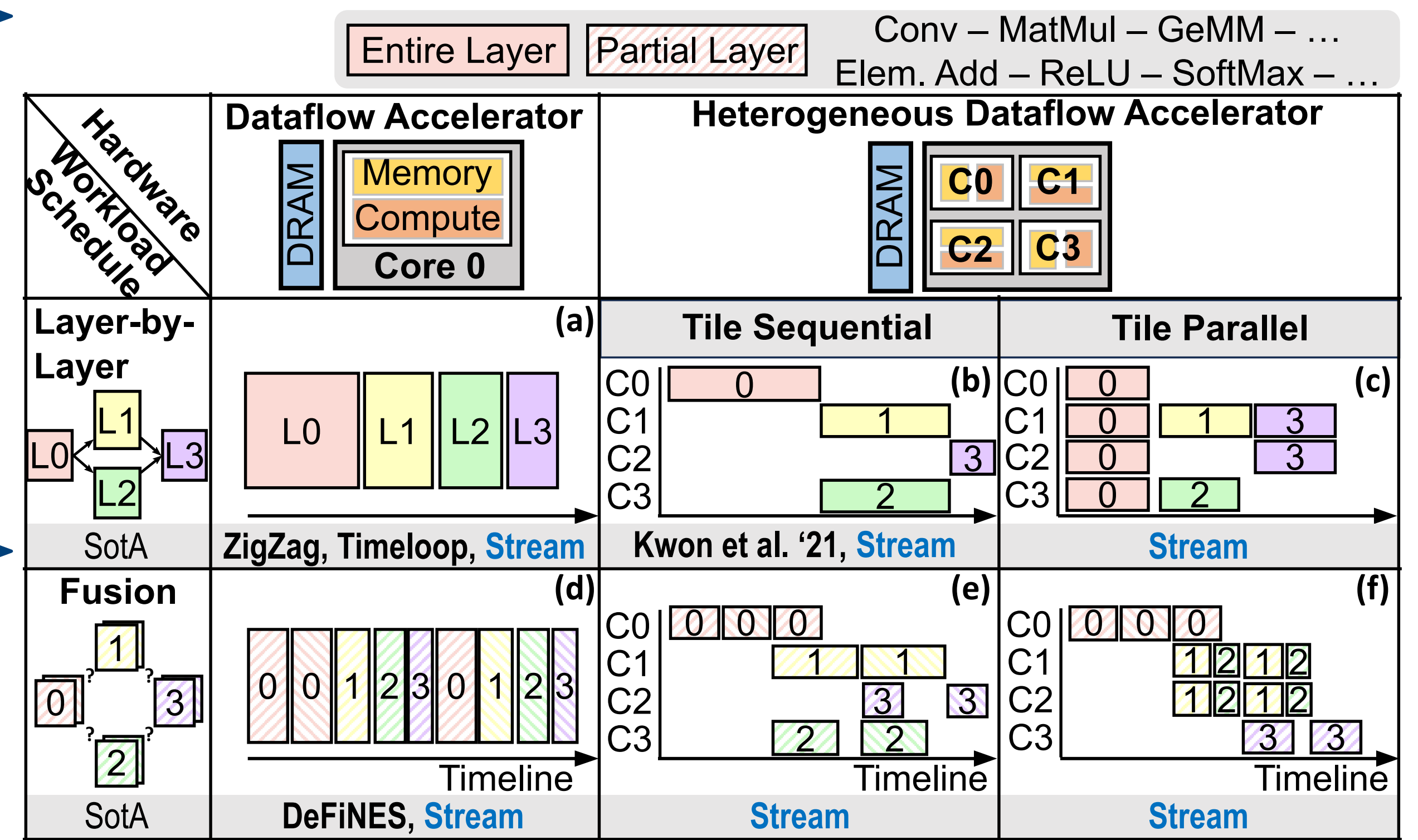
**Challenges:**

1. Larger allocation space

## Framework Overview



## Goal of Stream: Jointly explore fusion and accelerator types



**Key components of Stream:**

1. Uniform fine-grained layer fusion representation
2. High-level representation of compute cores and interconnects
3. Constraint optimization-based workload allocation
4. Communication & Memory-aware analytical cost model

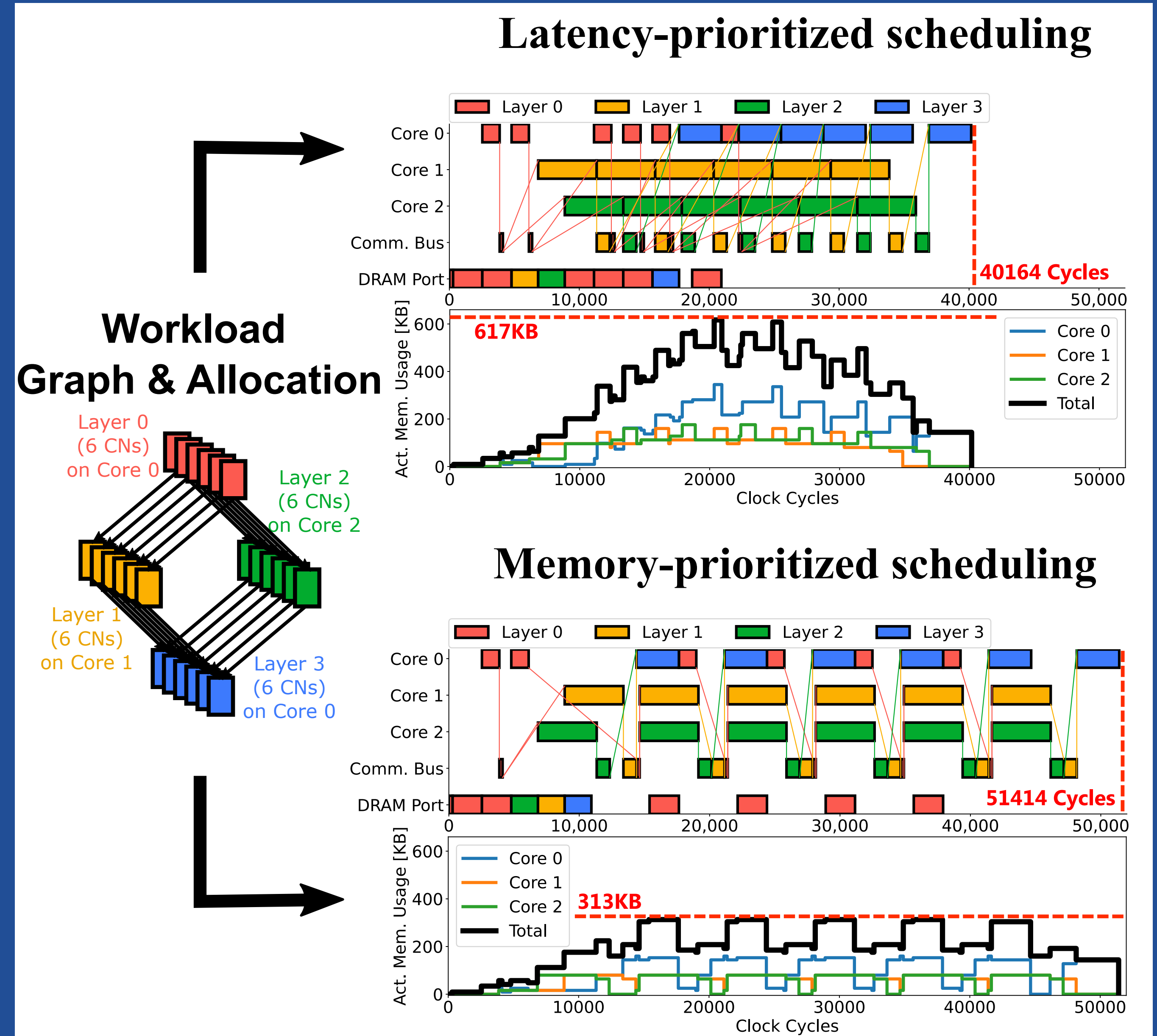
## Validation against 3 SotA chip measurements

Architecture	Latency Validation		
	Measured (cc)	Stream (cc)	Accuracy (%)
DepFiN (Digital Core)	$6.18 \times 10^6$	$5.65 \times 10^6$	91
Jia et al. (4x4 AiMC Cores)	$3.66 \times 10^5$	$3.68 \times 10^5$	99
DIANA (Digital Core + AiMC Core)	$8.12 \times 10^5$	$7.83 \times 10^5$	96

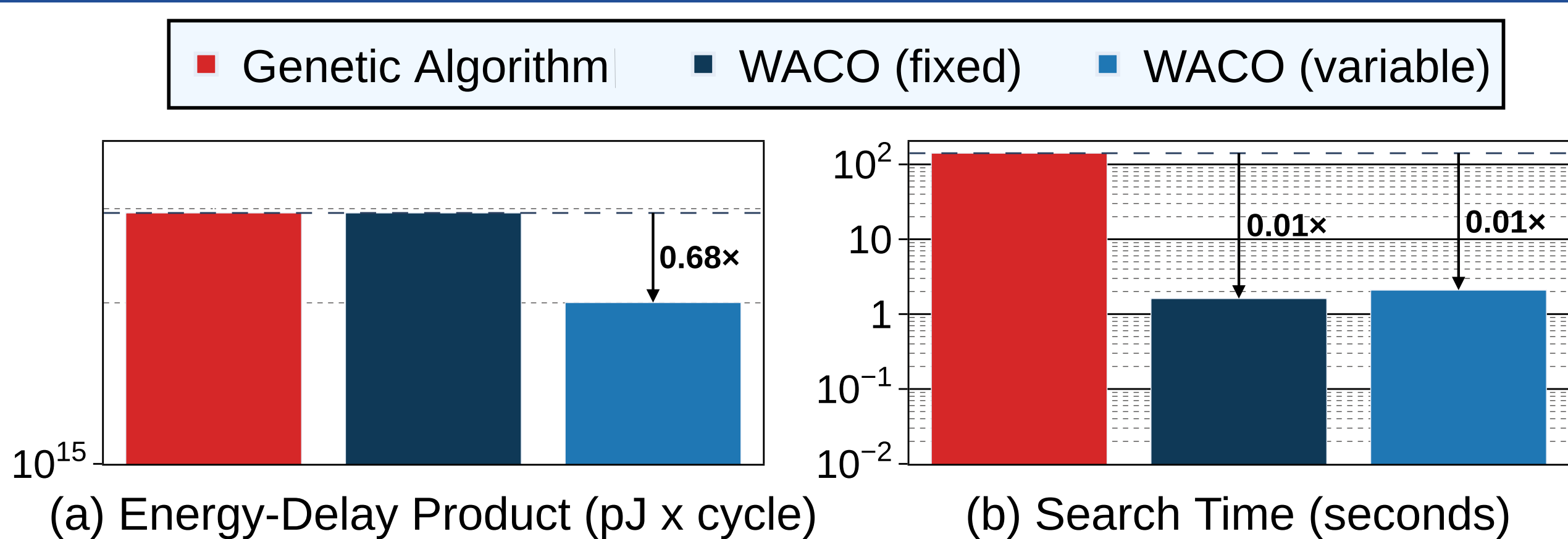
  

Architecture	Memory Usage Validation		
	Measured (KB)	Stream (KB)	Accuracy (%)
DepFiN (Digital Core)	238	244	97
Jia et al. (4x4 AiMC Cores)	N/A	16.5	N/A
DIANA (Digital Core + AiMC Core)	134	137	98

## Highlight: Hardware-aware scheduler



## Highlight: Constraint optimization allocation



## Design space exploration using Stream across multiple architectures and workloads

