IEEE Computer Society Annual Symposium on VLSI
ISVLSI 2026

O-POPE: HIGH-FREQUENCY PIPELINED OUTER PRODUCT BASED GEMM ACCELERATION WITH MINIMAL BUFFERING OVERHEAD

Danilo Cammarata
Paper 169
Dr. Angelo Garofalo, Prof. Dr. Luca Benini
ETH Zurich, Università di Bologna

7th - 10th July, 2026
ITC Sonar, Kolkata, India

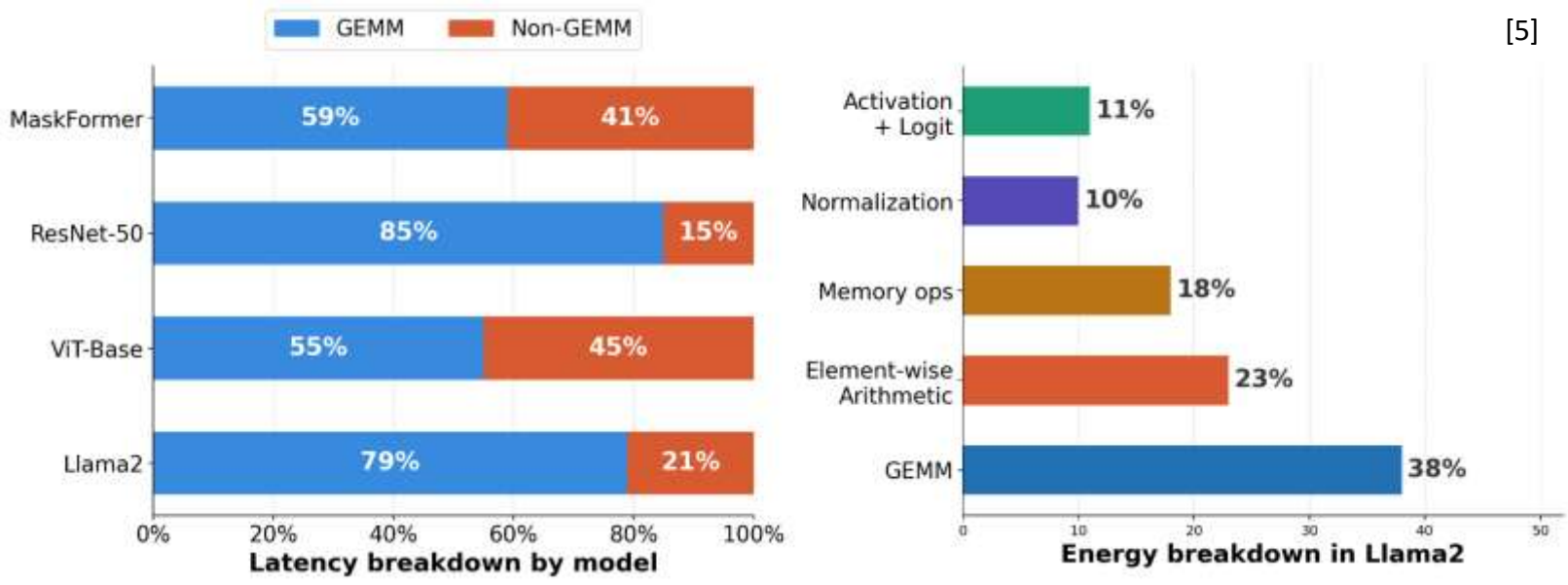
Efficiency in FP GEMM is essential

A Study of BFLOAT16 for Deep Learning Training [1]

FP8 Quantization:
The Power of the Exponent [2]

Mixed Precision Training [4]

FP8 Formats for Deep Learning [3]

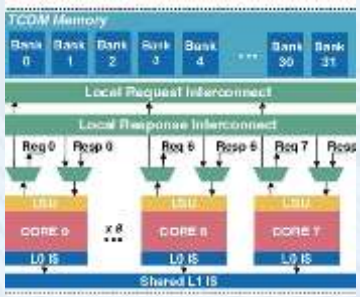


[5]

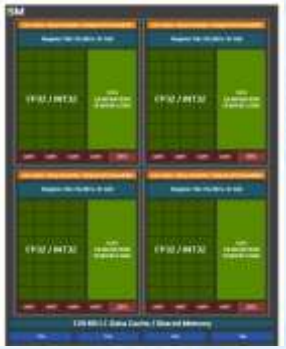
Danilo Cammarata
dcammarata@iis.ee.ethz.ch

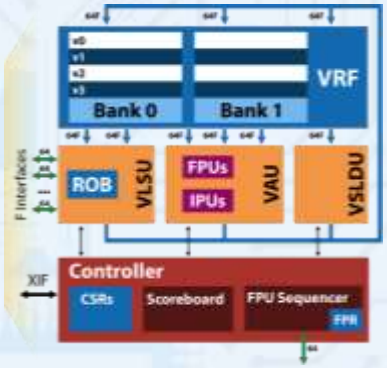
8 Jul 2026

DSAs minimize PPA

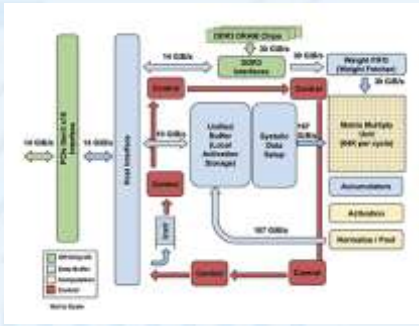


[6]


[7]



[8]


[9]

MIMD **SIMT** **SIMD** **DSA**

Programmability

Efficiency

- **Minimal area and energy consumption**
- **High arithmetic utilization**
- **High operating frequency and scalability**

8 Jul 2026
dcammarata@iis.ee.ethz.ch
Danilo Cammarata

State-of-the-art

Accelerator	Dataflow	Max Frequency (node)	Buffer size
RedMulE [10]	Inner Product	***613 MHz* (22 nm)**	9088 B
Sauria [11]	Outer Product	*555 MHz* (22 nm)	65536 B
Gemmini [12]	Outer Product	*550 MHz* (12 nm)	**5120 B**
Axon [13]	Outer Product	*550 MHz* (7 nm)	NR‡

Existing GeMM accelerators cannot achieve at the same time high frequency and minimal area/energy overhead

‡ Not Reported

8 Jul 2026
dcammarata@iis.ee.ethz.ch
Danilo Cammarata

Challenges

- **Minimal area and energy consumption**

  ➡Output Stationary Outer Product Dataflow

  (minimizes memory access and input buffers size)

- **High Utilization**

  ➡Double-buffer the output buffer

- **High Frequency**

  ➡Make use of pipeline registers in MAC unit

Danilo Cammarata dcammarata@iis.ee.ethz.ch

The key idea

How to design an efficient FP GEMM accelerator given these constraints ?

Reuse MAC unit pipeline registers as double-buffers

Danilo Cammarata dcammarata@iis.ee.ethz.ch 8 Jul 2026

Contributions

Danilo Cammarata dcammarata@iis.ee.ethz.ch 8 Jul 2026

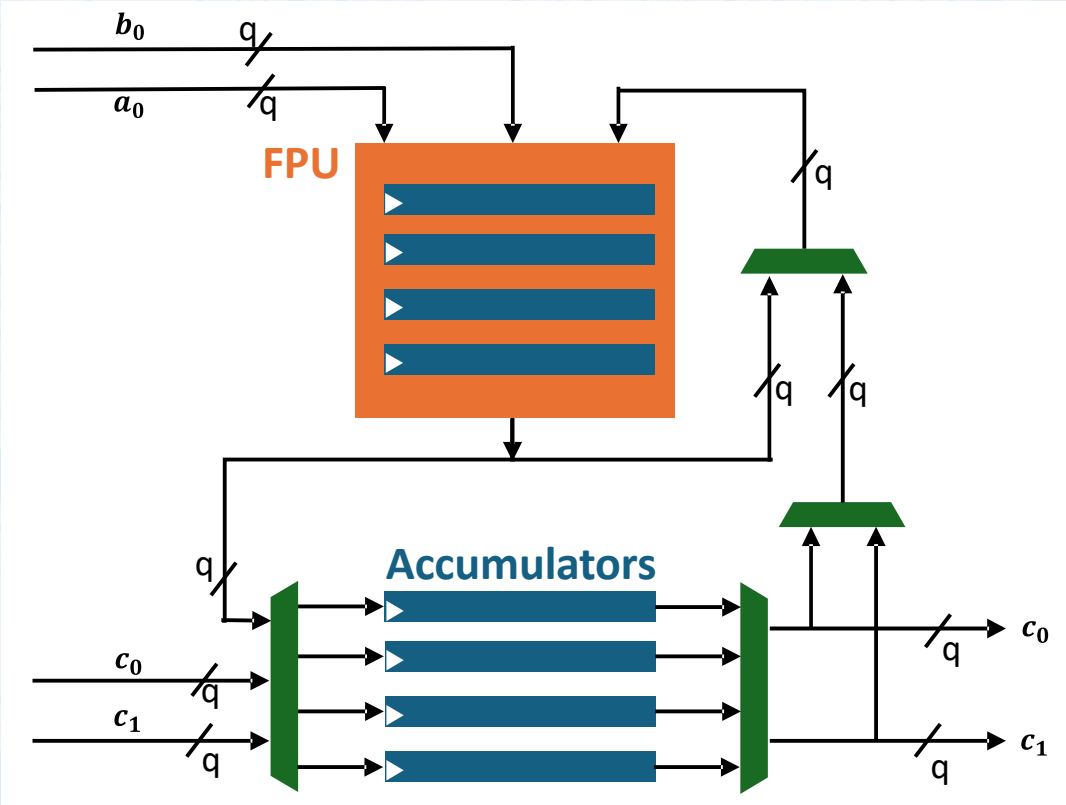
O-POPE:
High-Frequency Pipelined Outer Product-based GEMM acceleration with minimal buffering overhead

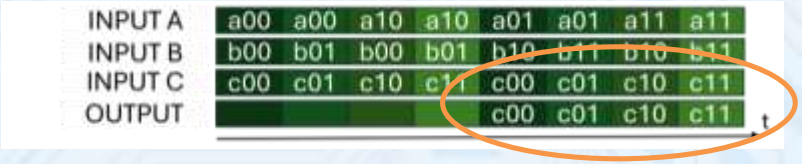
- A compact, efficient PE leveraging the FPUs' pipeline registers for buffering
- A scalable semi-systolic‡ architecture with minimal buffering overhead, achieving 99.97% utilization
- A comprehensive PPA evaluation

‡ Semi-systolic: a systolic-like architecture with some non-local connections among processing elements [14]

A compact and efficient PE

C += A * B

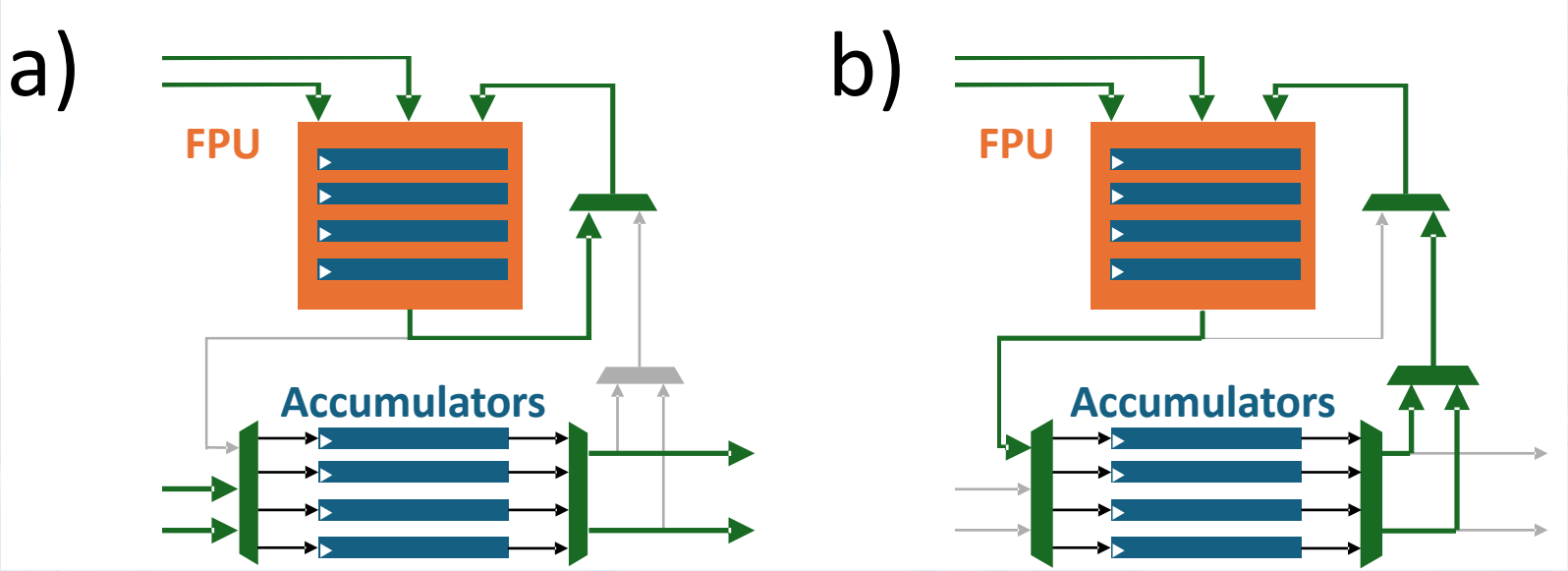




High-frequency + minimal output buffer overhead

Danilo Cammarata
dcammarata@iis.ee.ethz.ch

PE coupling-decoupling



DECOUPLING PHASE

- Accumulator free to be read/written externally
- FPU locked in accumulation

COUPLING PHASE

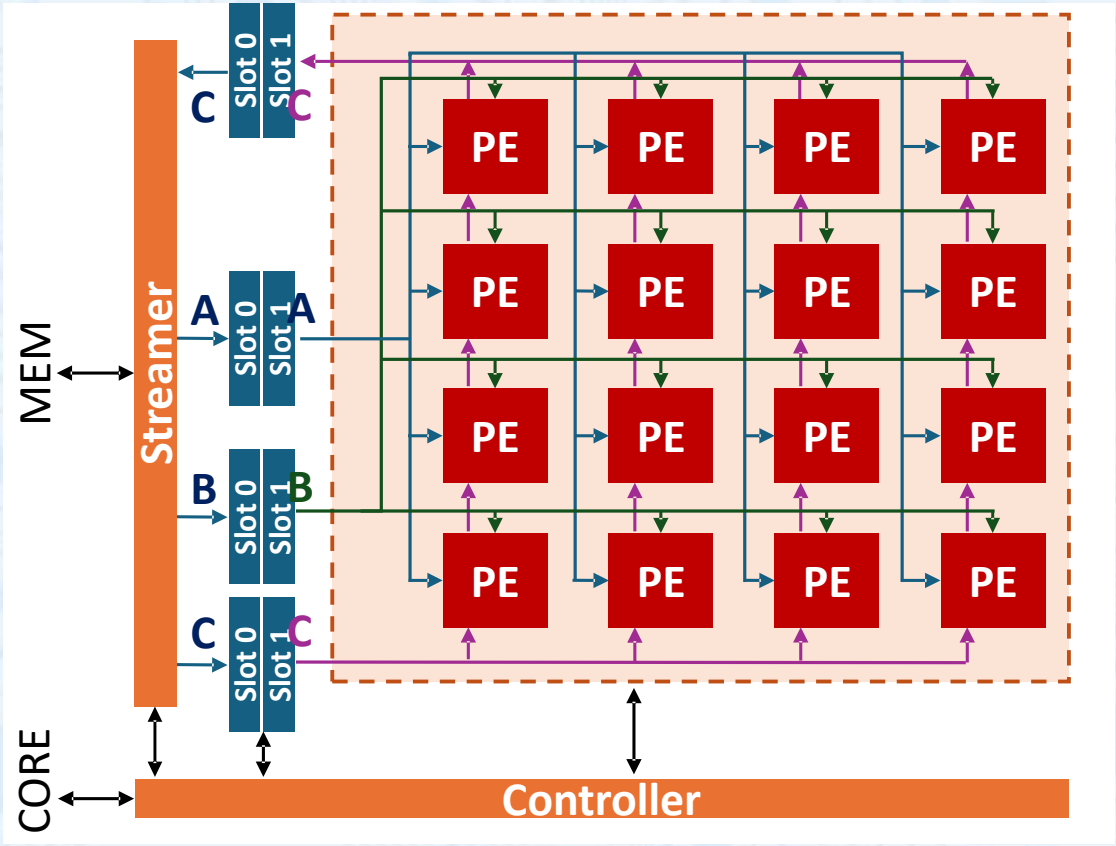
- Accumulator provides the new initial value
- FPU writes the computed value

Minimizes FPU utilization loss

Danilo Cammarata
dcammarata@iis.ee.ethz.ch
8 Jul 2026

O-POPE architecture

$$C \mathrel{+}= A * B$$



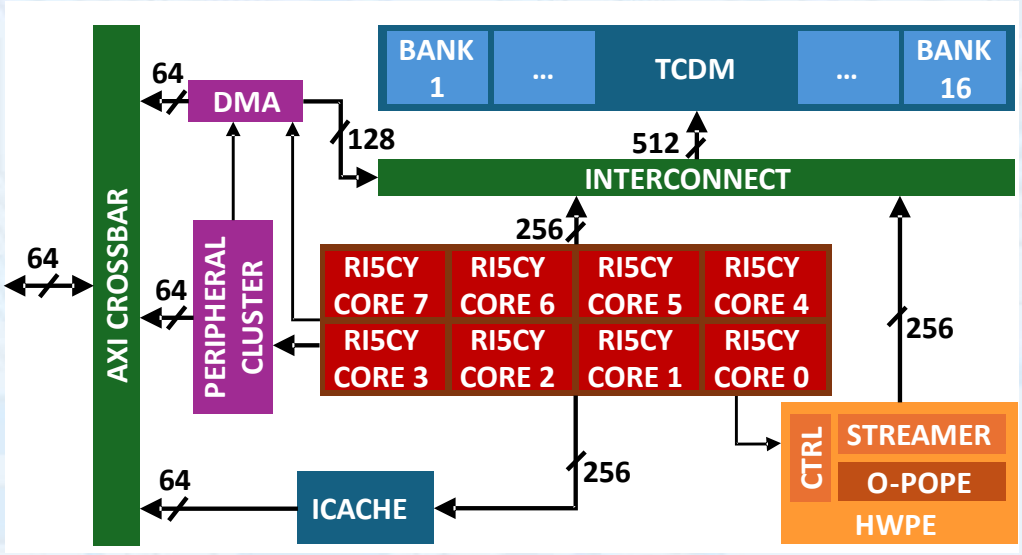
Minimize MEM bandwidth requirement

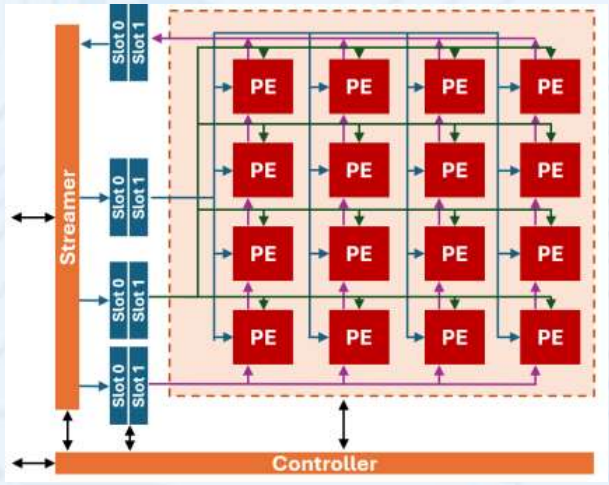
Scalability through systolic write-in/read-out of C

Minimize input buffers overhead

Danilo Cammarata dcammarata@iis.ee.ethz.ch

Experiment setup



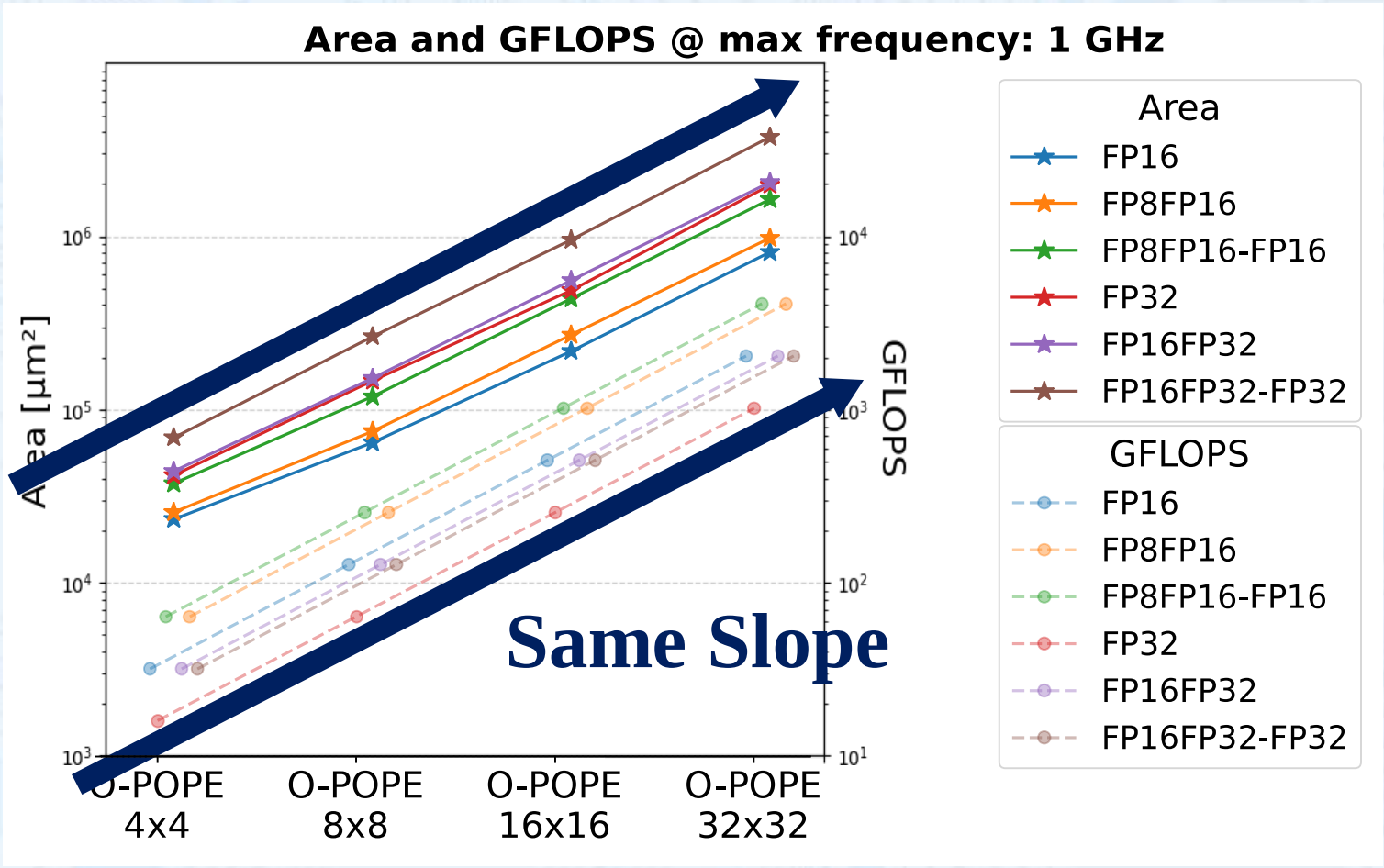


- **Evaluation across multiple FPUs and O-POPE mesh sizes**

- Cycle-accurate RTL simulation with *QuestaSim 2022.3*
- Synthesis with *Synopsys Design Compiler 2021.06*
- Physical Design with *Cadence Innovus 21.17*
- Power evaluation with *Synopsys PrimeTime 2022.03*

Danilo Cammarata
dcammarata@iis.ee.ethz.ch
8 Jul 2026

Area & Frequency Scalability

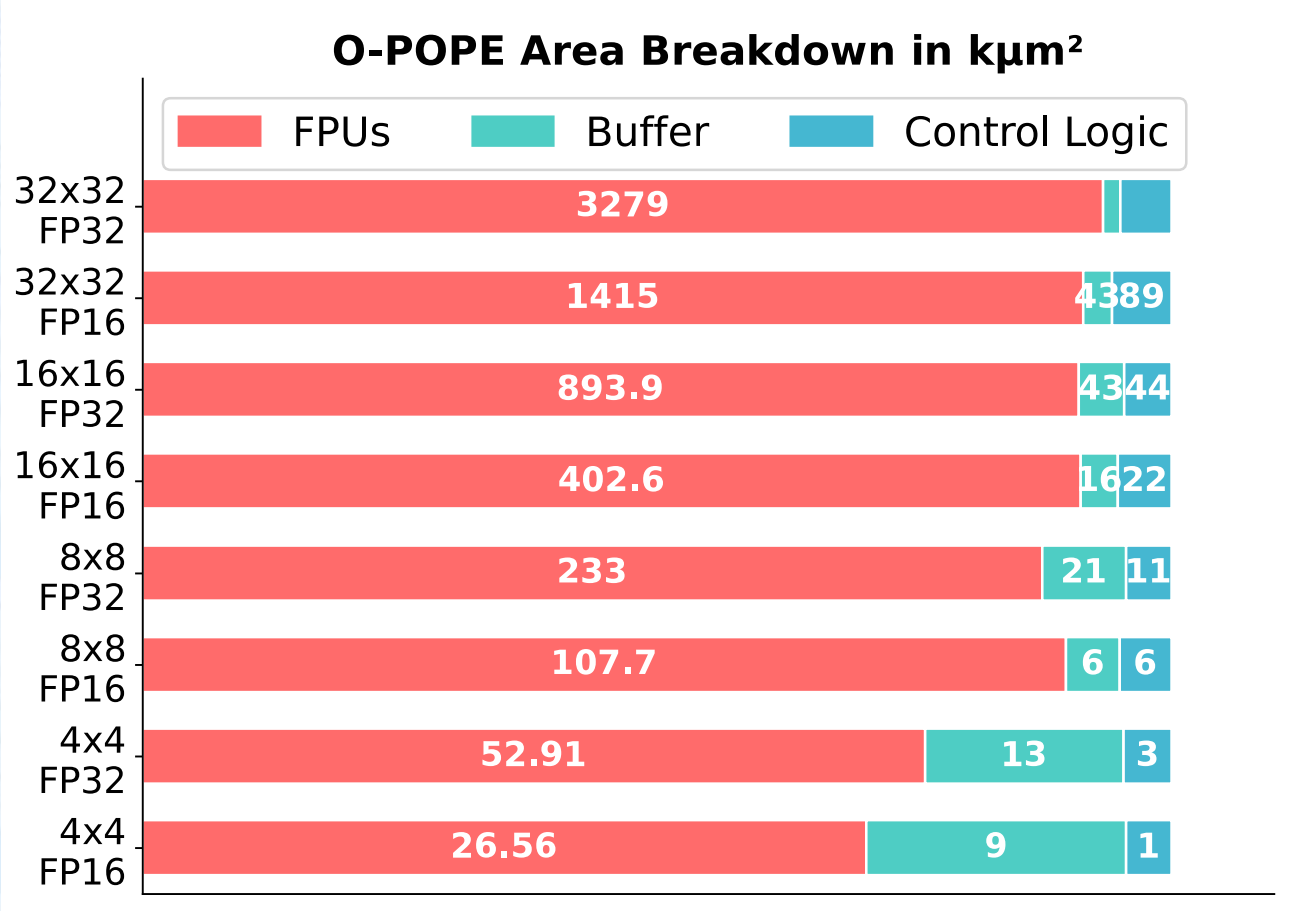


Technology: GlobalFoundries' 12LP+ FinFET

Corner : TT, 0.8 V, 25°C

Danilo Cammarata dcammarata@iis.ee.ethz.ch 8 Jul 2026

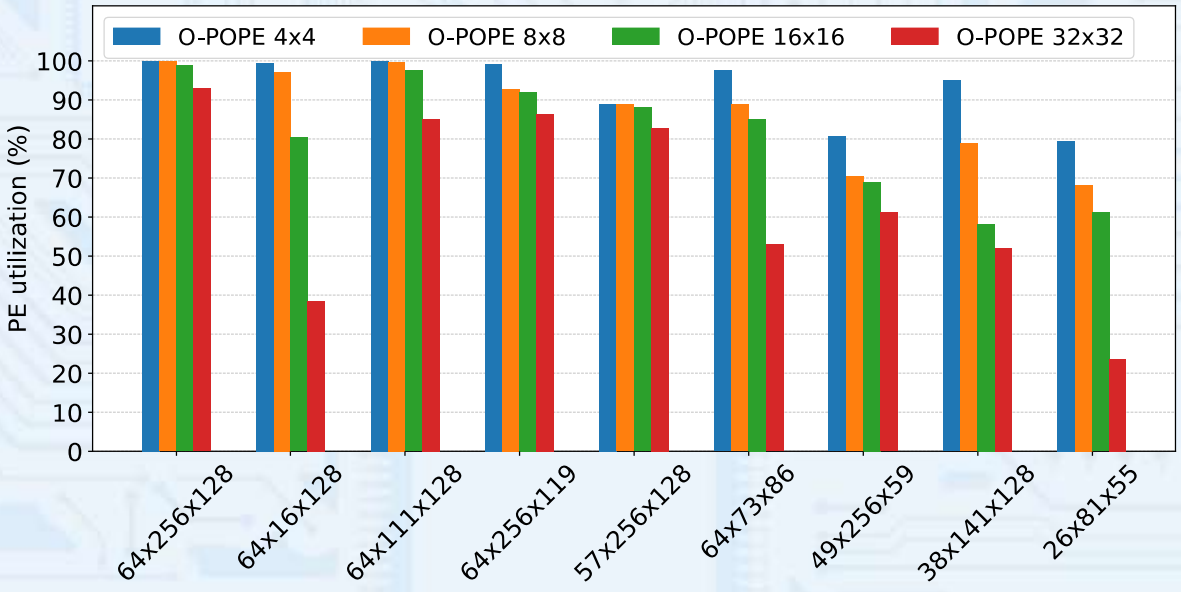
Area Breakdown



In Outer Product engines, input buffers scale linearly while compute scales quadratically

Danilo Cammarata
dcammarata@iis.ee.ethz.ch
8 Jul 2026

Cycle Runtime Analysis



Proportionally lower utilization for

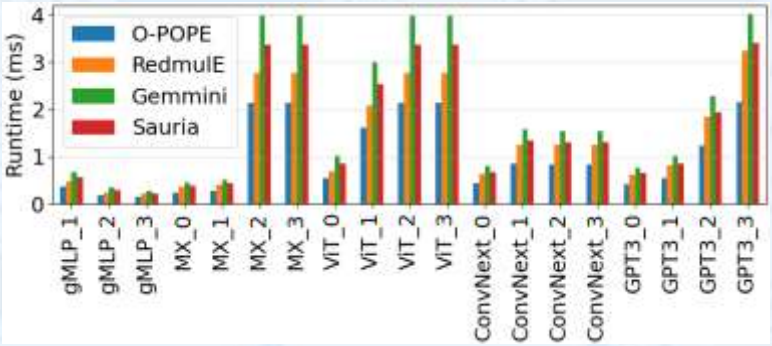
- Larger engines
- Workloads with smaller inner dimensions
- Workloads with output tiles not multiple of O-POPE output buffer

but up to 99.97% utilization for "regular" GEMM workloads

Danilo Cammarata dcammarata@iis.ee.ethz.ch

State-of-the-art comparison

Comparison with SoA **16 × 16 GEMM accelerators** supporting **FP16 → FP16** MAC units in **12 nm**



Accelerator	GFLOPS	GFLOPS/mm²	TFLOPS/W
RedMulE [10]	384	2134	2.74
Sauria [11]	333	1036	2.95
Gemmini [12]	280	749	NM‡
This work	**512**	**2336**	**3.18**

‡ Not Measured

Danilo Cammarata
dcammarata@iis.ee.ethz.ch

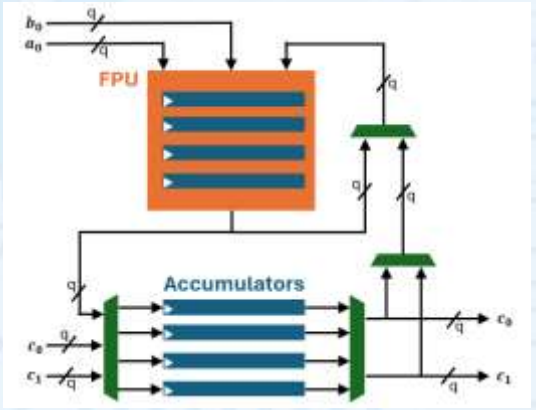
Conclusion

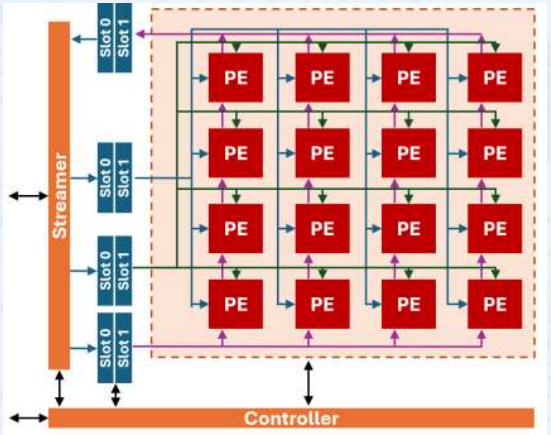
By leveraging

- **FPUs pipeline registers as double-buffers**
- **a scalable semi-systolic architecture**

O-POPE achieves

- **1 GHz operation frequency**
- ✓ **up to 99.97% arithmetic utilization**
- ✓ **minimal buffer area (<4% from 16x16)**
- ✓ **2048 GFLOPS (32x32) and 512 GFLOPS (16x16)**
- ✓ **2336 GFLOPS/mm²**
- ✓ **3.18 TFLOPS/W**
- ✓ **Open Sou**





github.com/pulp-platform/opope

Danilo Cammarata dcammarata@iis.ee.ethz.ch

THANK YOU

References

[1] D. Kalamkar et al.,"A Study of BFLOAT16 for Deep Learning Training"

[2] A. Kuzmin et al.,"FP8 Quantization: The Power of the Exponent"

[3] P. Micikevicius et al.,"FP8 Formats for Deep Learning"

[4] P. Micikevicius et al.,"Mixed Precision Training"

[5] R. Karami et al.,"Understanding the Performance Horizon of the Latest ML Workloads with NonGEMM Workloads"

[6] https://pulp-platform.org/docs/lugano2023/terapool_lugano2023.pdf

[7] https://images.nvidia.com/aem-dam/Solutions/geforce/blackwell/nvidia-rtx-blackwell-gpu-architecture.pdf

[8] https://github.com/pulp-platform/spatz/blob/main/docs/fig/spatz_arch.png

[9] P. J. Norman et al.,"In-Datacenter Performance Analysis of a Tensor Processing Unit "

[10] Y. Tortorella et al.,"RedMule: A Mixed-Precision Matrix-Matrix Operation Engine for Flexible and Energy-Efficient On-Chip Linear Algebra and TinyML Training Acceleration"

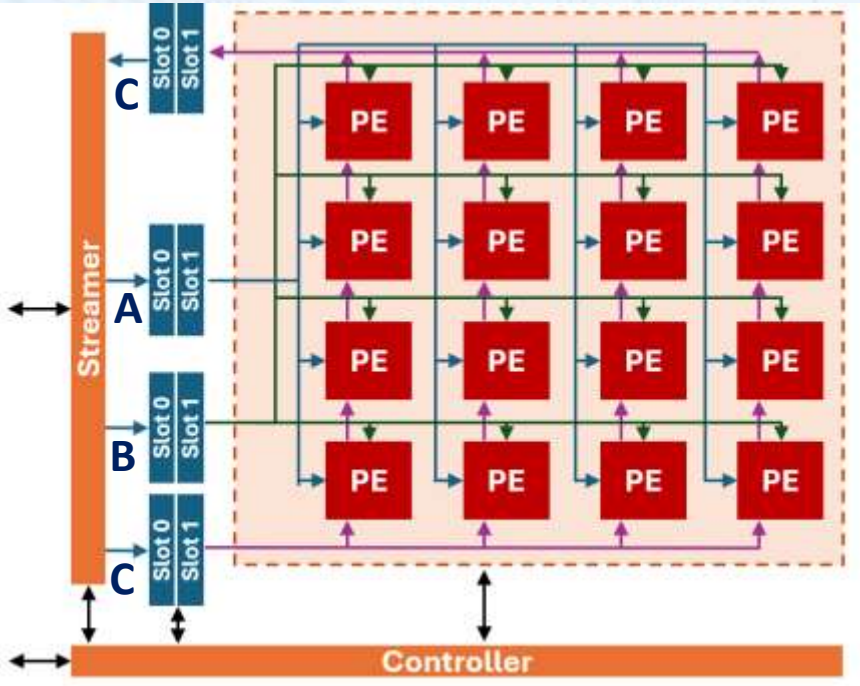
[11] J. Fornt et al.," An Energy-Efficient GeMM-Based Convolution Accelerator With On-the-Fly im2col"

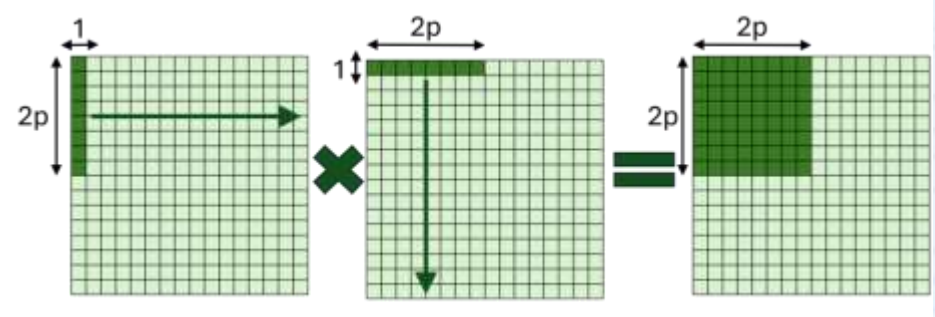
[12] H. Genc et al.," Gemmini: Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration"

[13] M. M. R. Nayan et al.," Axon: A Novel Systolic Array Architecture for Improved Run Time and Energy Efficient GeMM and Conv Operation with On-Chip im2col"

[14] H. T. Kung, "Why systolic architectures?"

O-POPE FSM


Streamer
C
A
B
C
Slot 0
Slot 1
PE
Controller


1
2p
2p
1
2p
2p


STREAMER
LOAD C
LOAD A , B, C
LOAD A , B
LOAD A , B; STORE C
LAST ITERATION ?
LOAD A , B
STORE C
ACCUMULATORS
SAVE C
SAVE next C
IDLE
SAVE FPU RESULTS
IDLE
LAST ITERATION ?
IDLE
SAVE FPU RESULTS
IDLE
FPUs
IDLE
BUSY
LAST ITERATION ?
BUSY
IDLE

The Proof

(a) Output Stationary dataflow

```
1. for i:m:M
2.   for j:n:N
3.     load C_(m×n)×γ
4.     for l:k:K
5.       load A_(m×k)×α
6.       load B_(k×n)×β
7.       C_m×n += A_m×k × B_k×n
8.     end
9.     store C_(m×n)×γ
10.  end
11. end
```

(b) Input Stationary dataflow

```
1. for i:m:M
2.   for l:k:K
3.     load A_(m×k)×α
4.     for j:n:N
5.       load B_(k×n)×β
6.       load C_(m×n)×γ
7.       C_m×n += A_m×k × B_k×n
8.       store C_(m×n)×γ
9.     end
10.  end
11. end
```

(c) Weight Stationary dataflow

```
1. for l:k:K
2.   for j:n:N
3.     load B_(k×n)×β
4.     for i:m:M
5.       load A_(m×k)×α
6.       load C_(m×n)×γ
7.       C_m×n += A_m×k × B_k×n
8.       store C_(m×n)×γ
9.     end
10.  end
11. end
```

$$OS : M \times N \times K \times \left(\frac{2}{K} + \frac{\alpha \times m + \beta \times n}{\gamma \times m \times n} \right)$$

$$IS : M \times N \times K \times \left(\frac{1}{N} + \frac{\beta \times k + 2 \times \gamma \times m}{\alpha \times m \times k} \right)$$

$$WS : M \times N \times K \times \left(\frac{1}{M} + \frac{\alpha \times k + 2 \times \gamma \times n}{\beta \times n \times k} \right)$$

- Matrices are large enough in ML
- Same number of bytes kept stationary
- Input datawidth is smaller than or equal to output datawidth