An Open-Source DNN Acceleration Fabric with Flexible Inter-Layer Pipelining Support

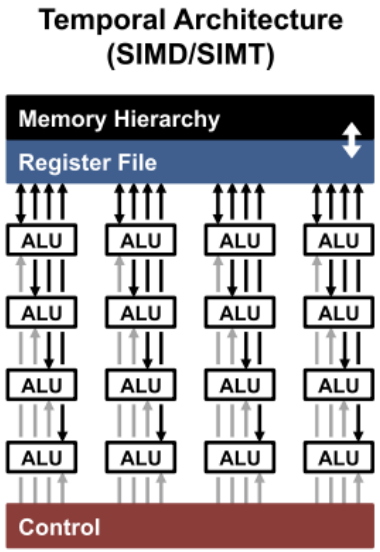
Gabriele Tombesi, Joseph Zuckerman, Je Yang, William Baisi, Kevin Lee, Davide Giri, and Luca P. Carloni

OSCAR 2025

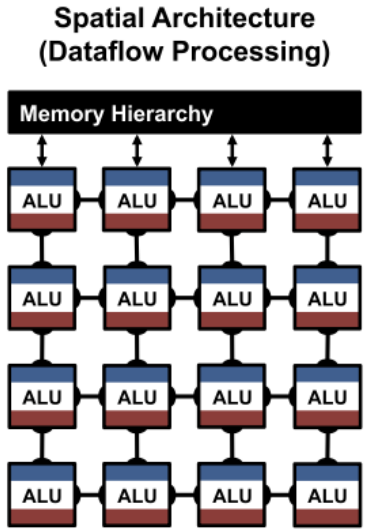
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CS@CU COMPUTER SCIENCE

State of the Art Hardware Accelerators for DL



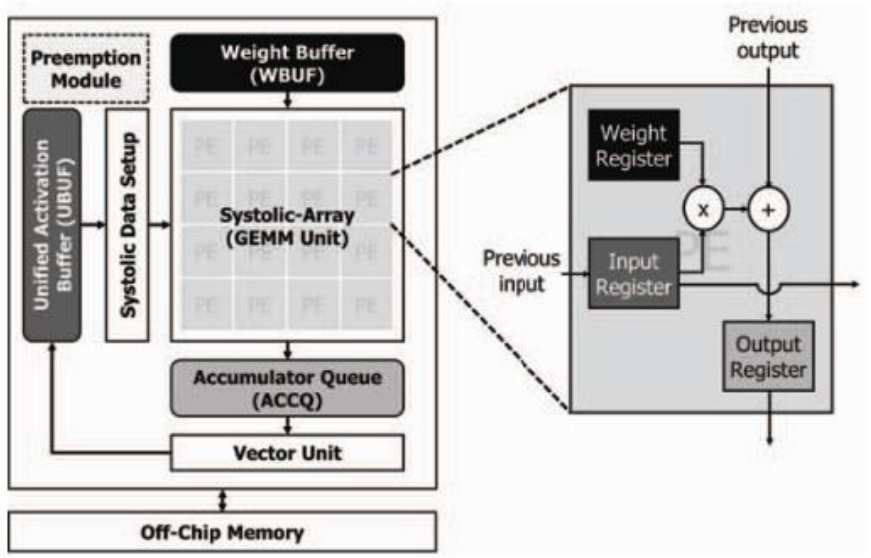
- Mostly in **CPUs/GPUs**
- **Centralized Control** logic
- No Communication across PEs
- Software Libraries (OpenBLAS/cuDNN) and Computational Transformations (FFT/Winograd/Strassen) to reduce computational complexity



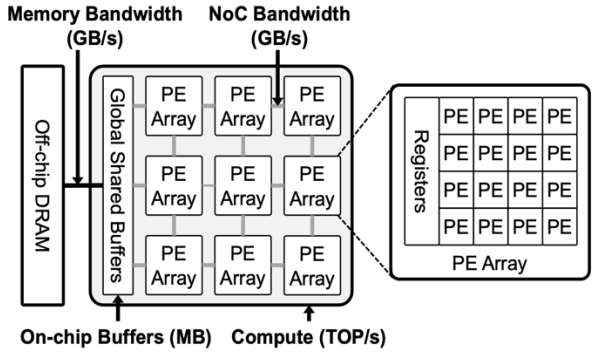
- Accelerators in **ASIC/FPGA-based** designs
- **Dataflow processing + Global Buffer**
- Distributed Control Logic and register-file/scratchpad

[Source: V. Sze et al. , Proc. IEEE ’17]

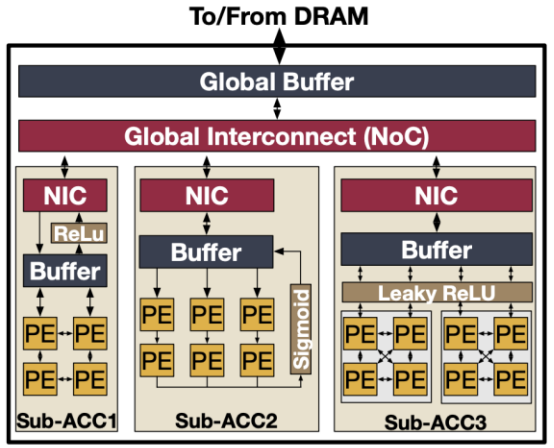
From Monolithic Accelerators to Tiled Architectures



Single-core architecture [1]



Homogeneous multi-core architecture [2]

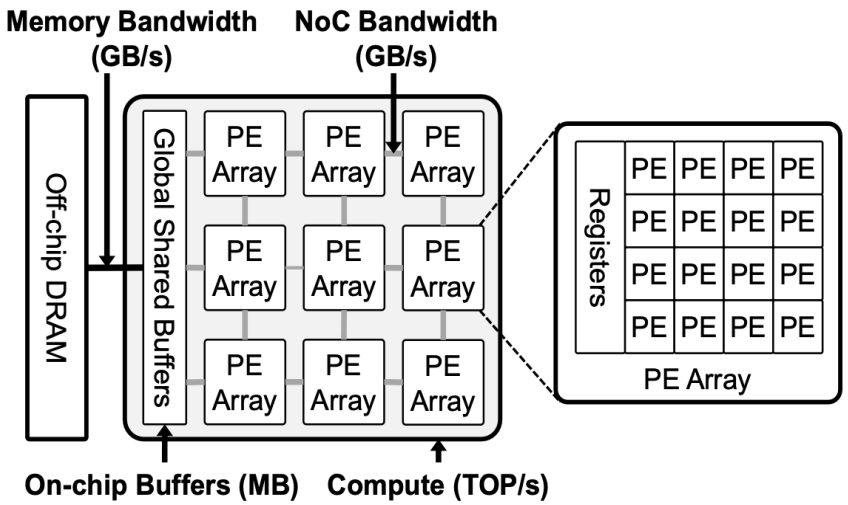


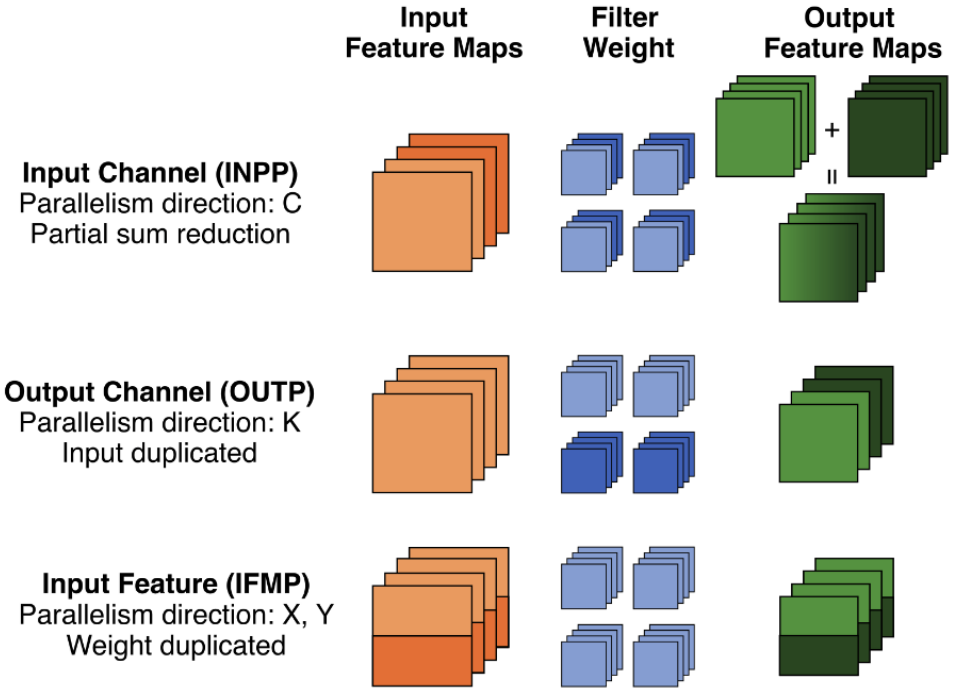
Heterogeneous dataflow accelerators (HDAs) [3]

- Systolic Array of PEs to exploit different types of *data reuse* (**WS** - Google TPU style/**OS** - ShiDianNao style)
- **Challenges** from fast evolving DL applications:
  - Bigger layer sizes -> scalability issues
  - Intra-model Heterogeneity -> resources underutilization

[Source: [1] Y. Choi , HPCA '20 [2] Y. Oh, HPCA '21 [3] H. Kwon, HPCA '21]

Tiled Architectures – Coarse-Grained Intra-Layer Parallelism



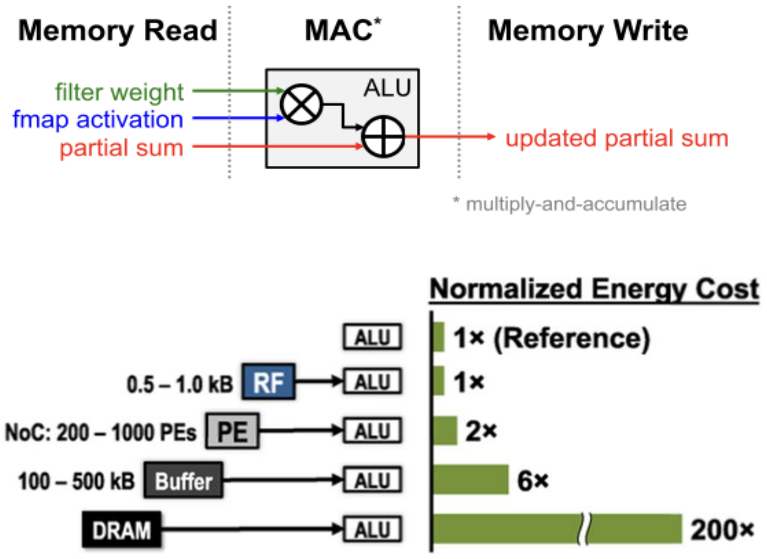


- Each tile assigned to a distinct spatial or channel partition of the target DNN layer
- Different **parallelism schemes** depending on the architecture and layer parameters

	Filter Partitioning	Input Channel Partitioning	Input Fmap Partitioning	Partial Output Sums
INPP	✓	✓		✓
OUTP	✓			
IFMP			✓	

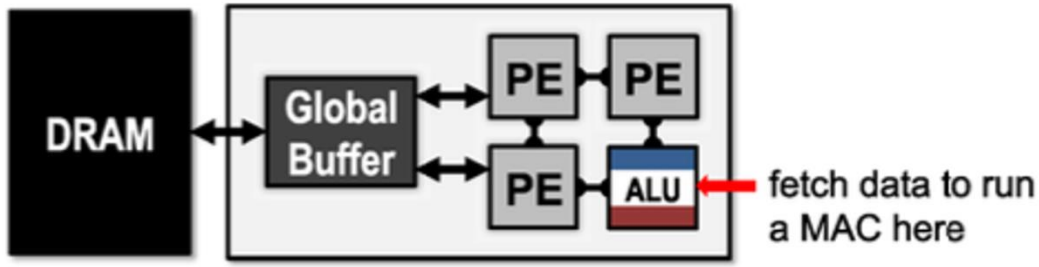
Intra-Layer Parallelism: The Memory Wall Problem

Fundamental Component of FC/CONV layers: MAC (multiply and accumulate)



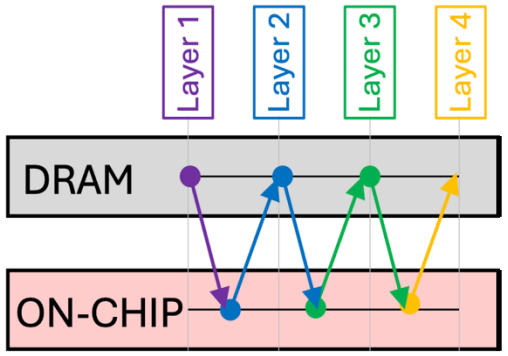
Tiled architectures address this problem by **maximizing the reuse of data** in low-level memories: **low-cost** but **limited-capacity**

Spatial Architectures

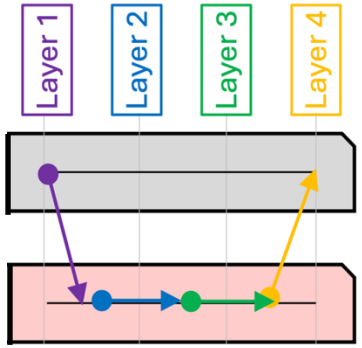


[Source: V. Sze et al. , Proc. IEEE ’17]

Inter-Layer Pipelining

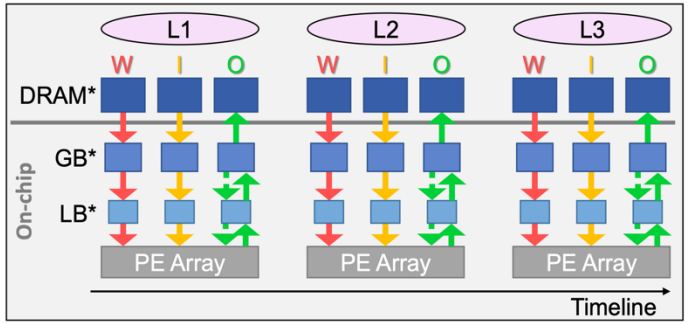


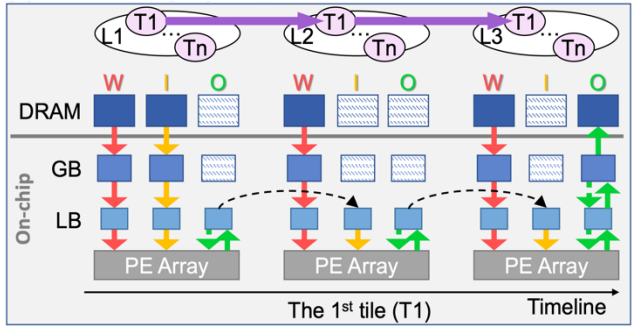
Layer-Wise Execution



Segment Execution [1]

Monolithic Systolic Array





[2]

[Source: [1] T. Glint. , DAC'24 [2] L. Mei, HPCA'23]

This Work

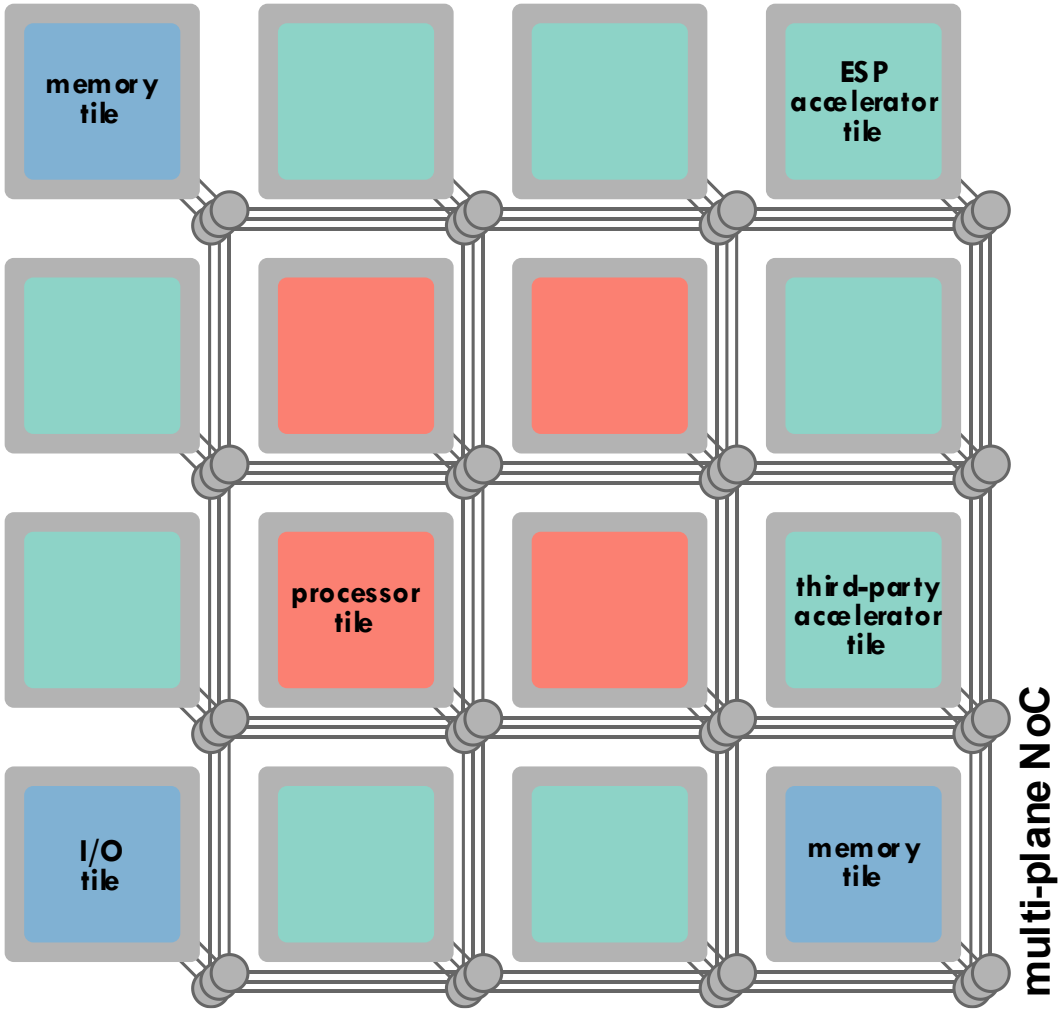
Problem: Existing work focuses on specialized tiled architectures without inter-layer pipelining support *or* highly optimized inter-layer pipelining solutions for monolithic accelerators

=> In this work, we propose an acceleration fabric that combines coarse-grain intra-layer parallelism from tiled architectures with inter-layer pipelining, using the open-source ESP platform

ESP Architecture

- RISC-V Processors
- Many-Accelerator
- Distributed Memory
- Multi-Plane NoC

The ESP architecture implements a **distributed** system, which is **scalable**, **modular** and **heterogeneous**, giving processors and accelerators similar weight in the SoC



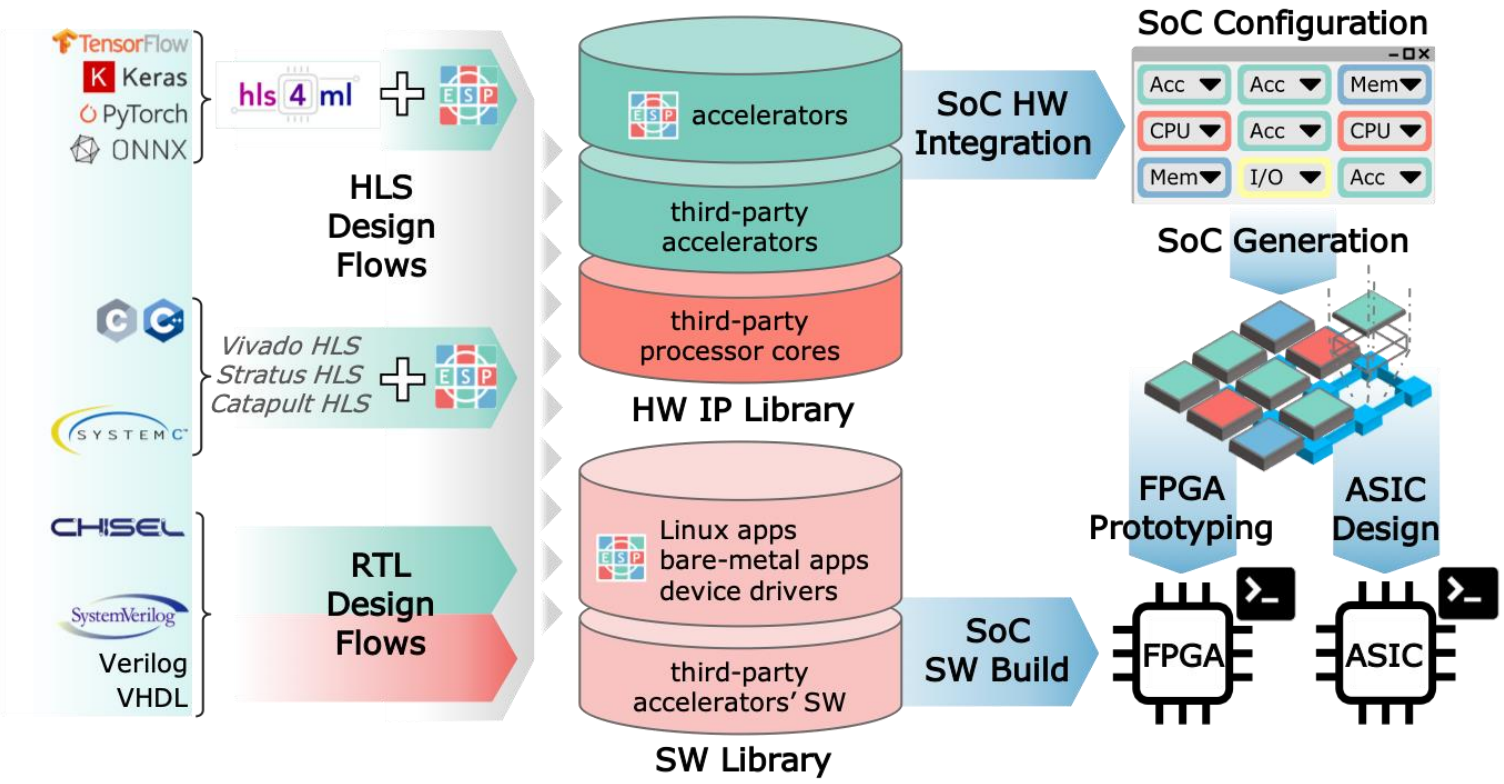
ESP Methodology

Accelerator Flow

- Simplified design
- Automated integration

SoC Flow

- Mix&match floorplanning GUI
- Rapid FPGA prototyping



[Source: https://esp.cs.columbia.edu/]

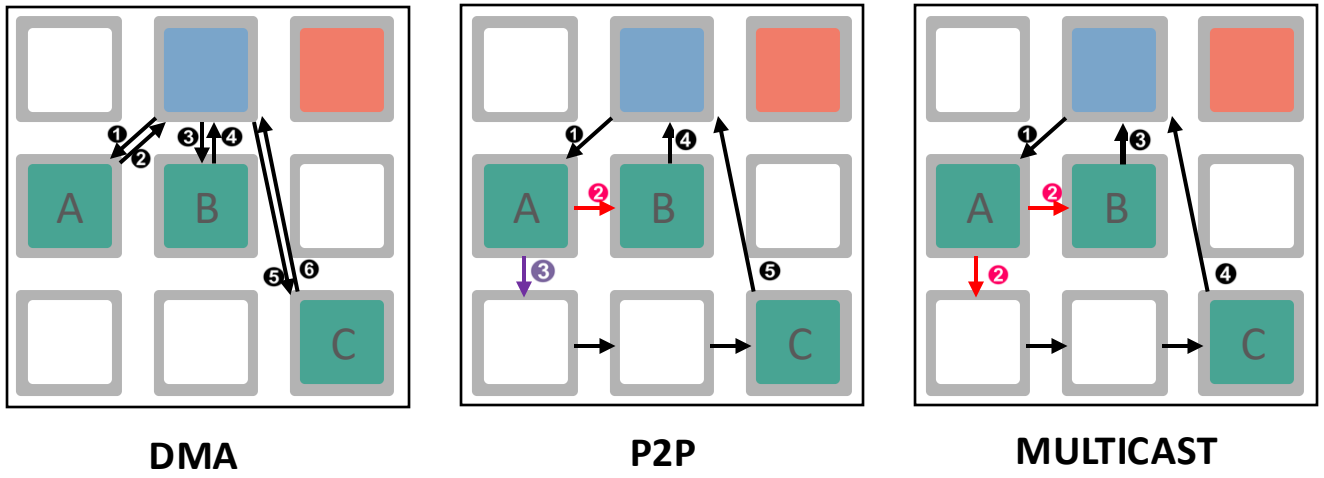
ESP-DNN Acceleration Fabric

Acceleration fabric:

- 2 types of **accelerator tiles**:
  - Compute Tile
  - Reduction Tile

- 3 **Data-transfer primitives**:
  - Direct Memory Access
  - Point-to-Point
  - Multicast

Example Dataflow: **A** → **B**, **C**



ESP-DNN Acceleration Fabric – Segment Mapping

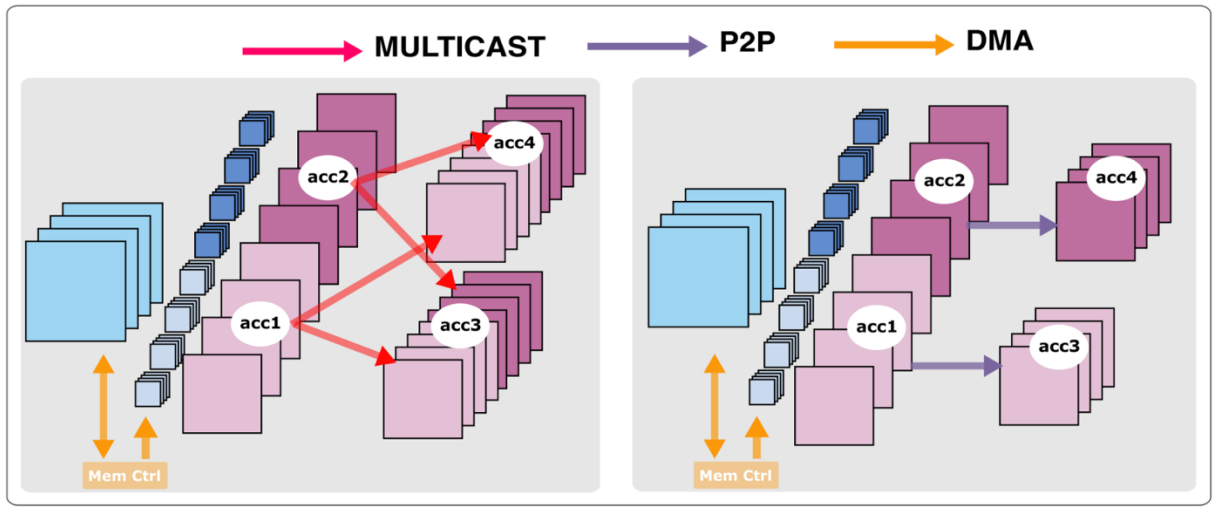
- **Segment** = *Sequence of adjacent layers executed concurrently by distinct groups of accelerator tiles.*

- **Segment Mapping** onto the ESP acceleration fabric:
  1. #accs per layer
  2. Intra-layer parallelsim per layer
  3. Inter-layer communication pattern

⇒Example: 2 alternative segment mappings for a **2-layer segment of a CNN**:

1. 2 accs per layer

(a) O-O => MUTLICAST+DMA
(b) O-I => P2P + DMA

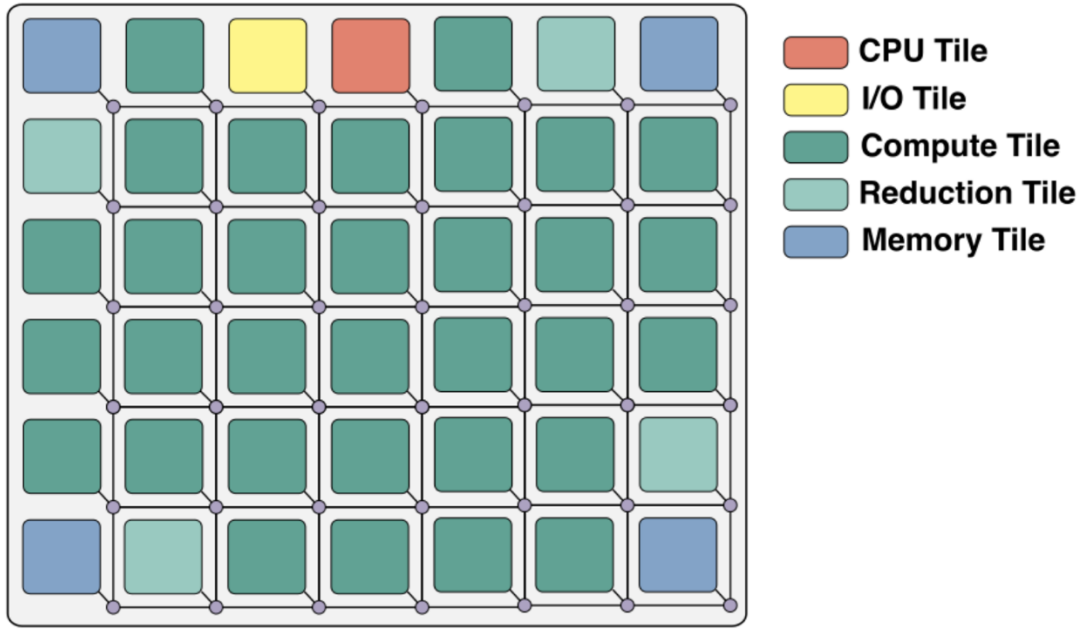


(a) (b)

Experimental Evaluation

FPGA Prototype:

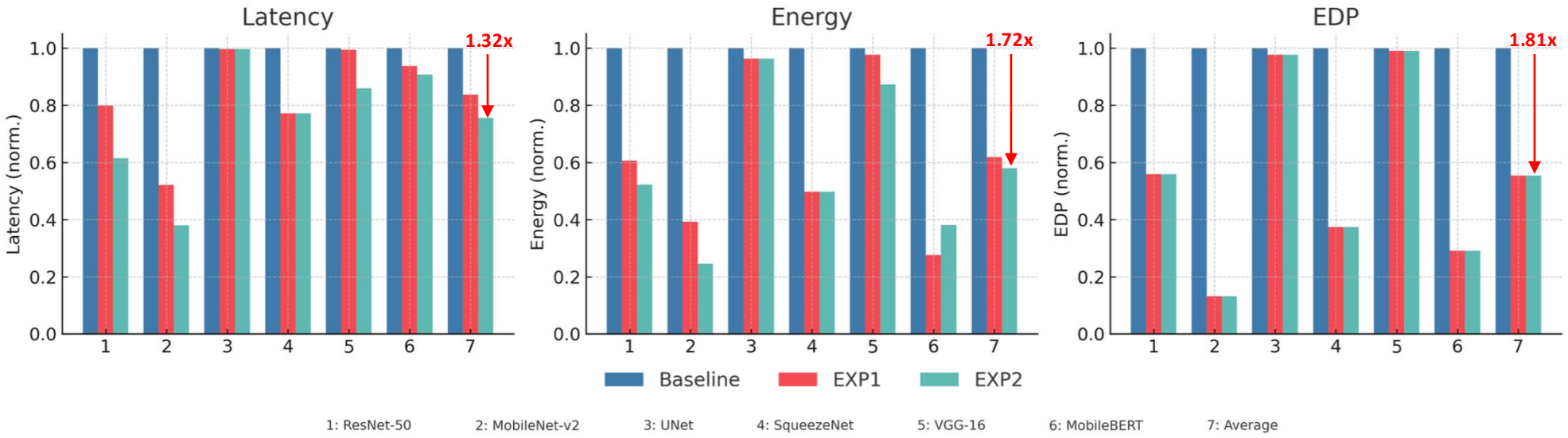
- 6x7 ESP instance
- Xilinx XCVU19P board
- Resources:
  - **1 CPU tile** for workload scheduling
  - **1 IO tile** for peripherals
  - **4 memory tiles**
  - **36 accelerators tiles:**
    - 32 compute tiles
    - 4 reduction tiles



Experimental Evaluation

- **Benchmark:**
  - 6 representative networks for **vision/speech/language tasks** (ResNet-50, MobileNet, U-net, SqueezeNet,VGG16, MobileBERT)
- **3 Deployment modes:**
  - **Baseline:** all 32 compute and 4 memory tiles active + NO inter-layer pipelining
  - **EXP-1:** # compute and memory tiles customized + NO inter-layer pipelining
  - **EXP-2:** EXP-1 + inter-layer pipelining enabled.

Results:



Conclusions

- We implemented an acceleration fabric to flexibly combine **coarse-grain intra-layer parallelism** and **inter-layer pipelining** using the Open Source ESP architecture.

- Results demonstrate consistent **PPA improvement** across the DNN models of our benchmark suite, when multi-layer segments are enabled.

- **Future work**: extend the fabric to support **multi-model execution** with multiple segments sharing the on-chip compute resources and off-chip memory bandwidth:
  - Optimization framework to explore the scheduling/mapping problem.
  - Spatial independence assumption (guaranteed by NoC + tiled architecture)

Tomorrow: Tutorial on Agile Design of Secure and Resilient AI-Centric Systems

- Full Day: 8AM – 3PM
- Location: 121-B1F-113 (Building 121, Floor B1F, Room 113)
- Co-organized by IBM Research and Columbia University

8:00 - 8:30 AM	**Tutorial Introduction** Pradip Bose (IBM Research)
8:30 - 9:15 AM	**ESP Mini-Tutorial** Luca Carloni (Columbia University)
9:15 - 10:00 AM	**Illustrative Use of ESP to Design Efficient CAV SoCs (EPOCHS) and Beyond** Joseph Zuckerman (Columbia University) & Karthik Swaminathan (IBM Research)
10:00 - 10:30 AM	*Coffee Break*
10:30 - 11:30 AM	**Introduction to FHE Algorithms and Architectures** Charanjit Jutla (IBM Research)
11:30 - 12:00 PM	**Fourier Transform Accelerators Using Integrated Photonics for Fully Homomorphic Encryption** Imon Kundu (Optalysys)
12:00 - 1:00 PM	*Lunch Break*
1:00 - 1:30 PM	**Security and Resilience Challenges in AI-Centric Systems** Naorin Hossain, Karthik Swaminathan, Pradip Bose (IBM Research)
1:30 - 2:00 PM	**IBM's SARA SoC/SiP Project: Application-Driven High Level View** Pradip Bose *et al.* (IBM Research)
2:00 - 3:00 PM	**HELayers Driven Software Stack for AI/FHE Appliances** Eyal Kushnir *et al.* (IBM Israel Research Laboratory, Haifa)
3:00 PM	**End of SARA Tutorial** (see you next year!)

Thank you from the ESP team!

 sld.cs.columbia.edu  ColumbiaSld  esp.cs.columbia.edu  sld-columbia/esp  c/ESP-platform

An Open-Source DNN Acceleration Fabric with Flexible Inter-Layer Pipelining Support

Gabriele Tombesi, Joseph Zuckerman, Je Yang, William Baisi, Kevin Lee, Davide Giri, and Luca P. Carloni

OSCAR 2025