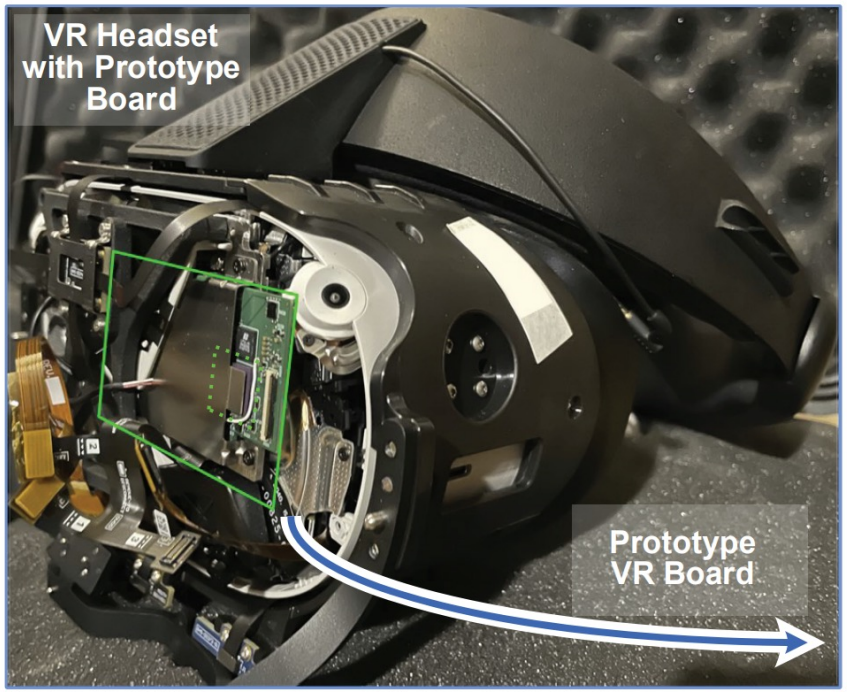
Heterogeneous Dataflow Accelerators for Multi-DNN Workloads

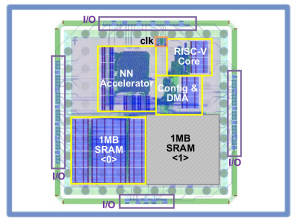
HPCA 2021 Paper

DL Compiler Study
Constant Park (박상수)
2022-08-08

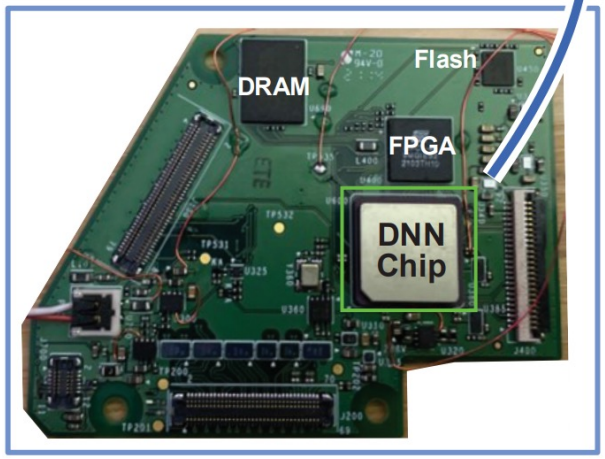
Multi-DNN Workloads

- Multiple DNN for various tasks to collaboratively
  - Sub-tasks (VR) such as object-detection, hand/eye tracking[1], etc.,





Custom built low-power DNN Accelerator Chip in 7nm

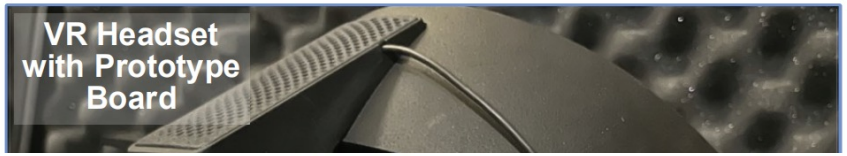


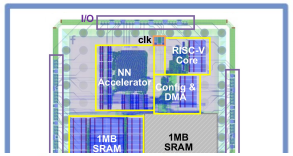
Prototype Printed Circuit Board Featuring DNN Accelerator Chip

[1] System-Level Design and Integration of a Prototype AR/VR Hardware Featuring a Custom Low-Power DNN Accelerator Chip in 7nm Technology for Codec Avatars, *IEEE CICC*, 2022

Multi-DNN Workloads

- Multiple DNN for various tasks to collaboratively
  - Sub-tasks (VR) such as object-detection, hand/eye tracking[1], etc.,





Custom built low-power DNN Accelerator Chip in 7nm

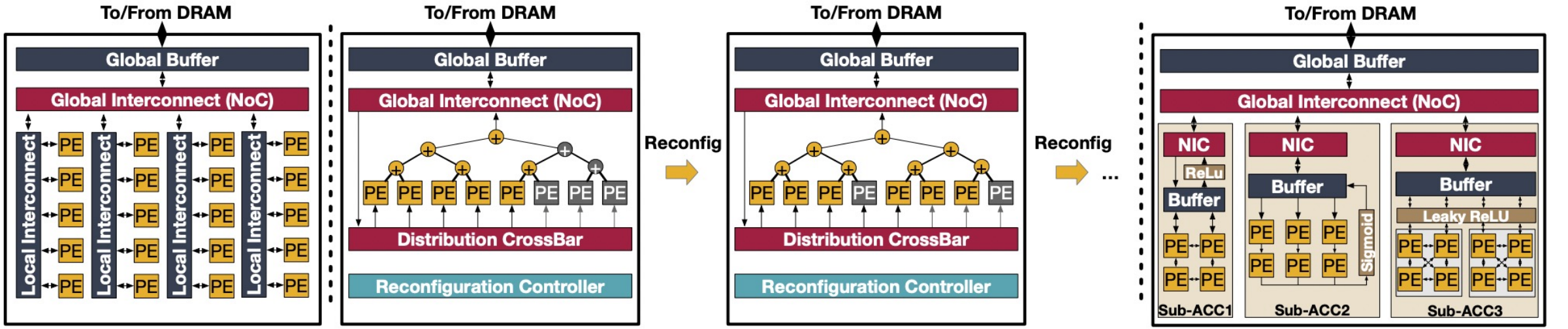
Task	Model	Channel-Activation Size Ratio	Layer Operations
Object Detection	MobileNetV2 [10]	Min: 0,013, Median: 13.714, Max: 1280	CONV2D, PWCONV, DWCONV, Skip-Con.
Object Classification	Resnet50 [9]	Min: 0.013, Median: 18.286, Max: 292.571	CONV2D, FC, Skip-Con.
Hand Tracking	UNet [11]	Min: 0.002, Median: 1.855, Max: 34.133	CONV2D, FC, UPCONV, Concat.
Hand Pose Estimation	Br-Q HandposeNet [16]	Min: 0.016, Median: 1024, Max: 1024	CONV2D, FC
Depth Estimation	Focal Length DepthNet [17]	Min: 0.013, Median: 4.571, Max: 4096	CONV2D, FC, UPCONV

Accelerator Chip

[1] System-Level Design and Integration of a Prototype AR/VR Hardware Featuring a Custom Low-Power DNN Accelerator Chip in 7nm Technology for Codec Avatars, *IEEE CICC*, 2022

Various Dataflows inside NPU

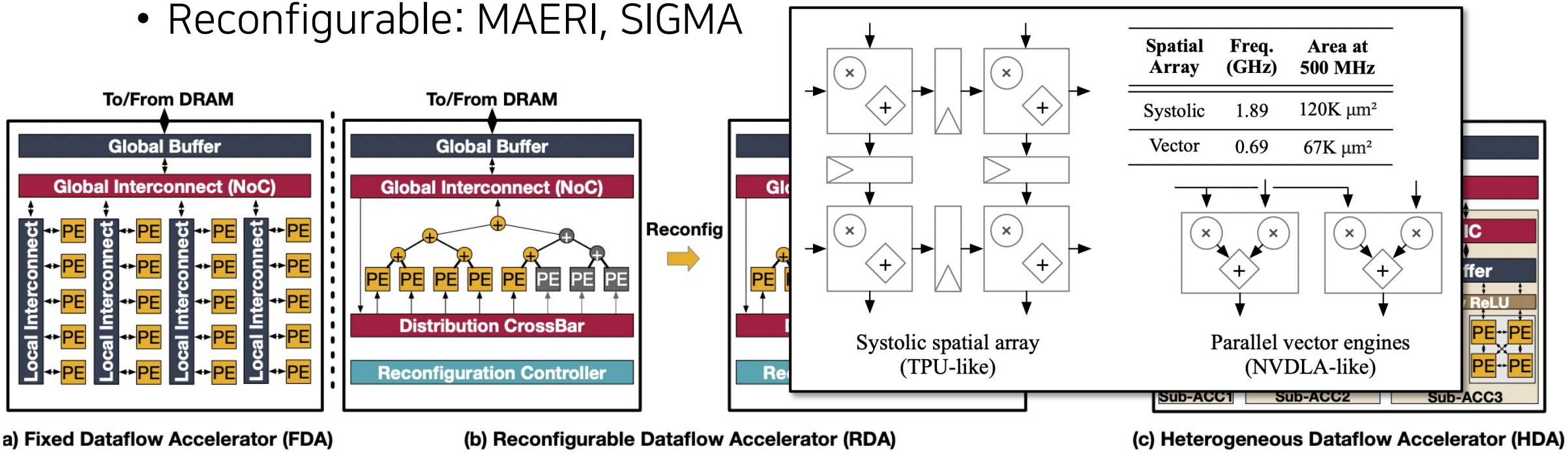
- Fixed, reconfigurable, and heterogeneous dataflows
  - Fixed dataflow: Systolic array (Eyeriss, TPU)
  - Reconfigurable: MAERI, SIGMA



[2] Gemmini: An Agile Systolic Array Generator Enabling Systematic Evaluations of Deep-Learning Architectures, *DAC,* 2021

Various Dataflows inside NPU

- Fixed, reconfigurable, and heterogeneous dataflows
  - Fixed dataflow: Systolic array (Eyeriss, TPU)
  - Reconfigurable: MAERI, SIGMA



[2] Gemmini: An Agile Systolic Array Generator Enabling Systematic Evaluations of Deep-Learning Architectures, *DAC,* 2021

Various Dataflows inside NPU

- Fixed, reconfigurable, and heterogeneous dataflows
  - Various configurations between MAC array and Memory

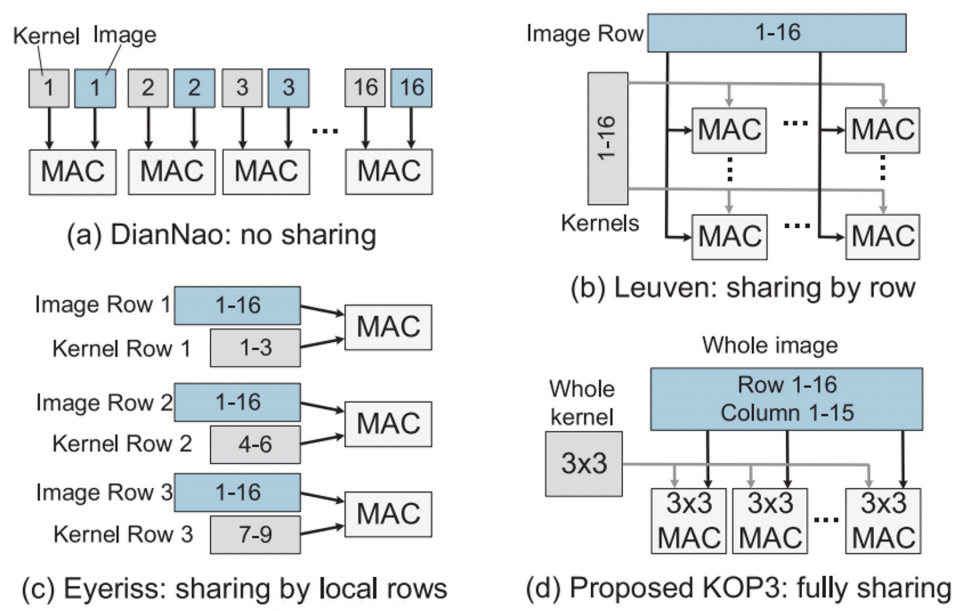


Fig. 3. Conv-Array structures of (a) DianNao [1] (b) Leuven [6] (c) Eyeriss [2] (d) Proposed KOP3.

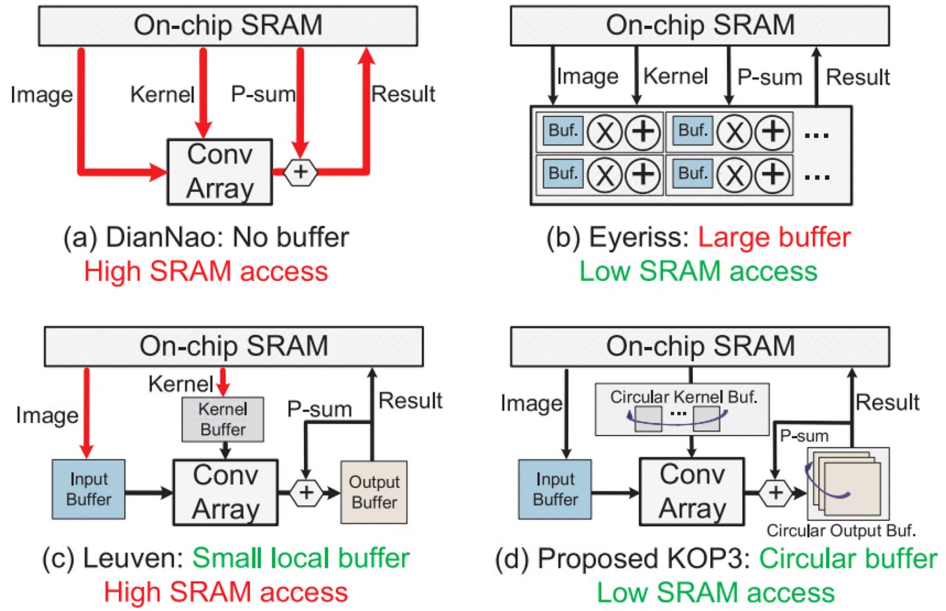
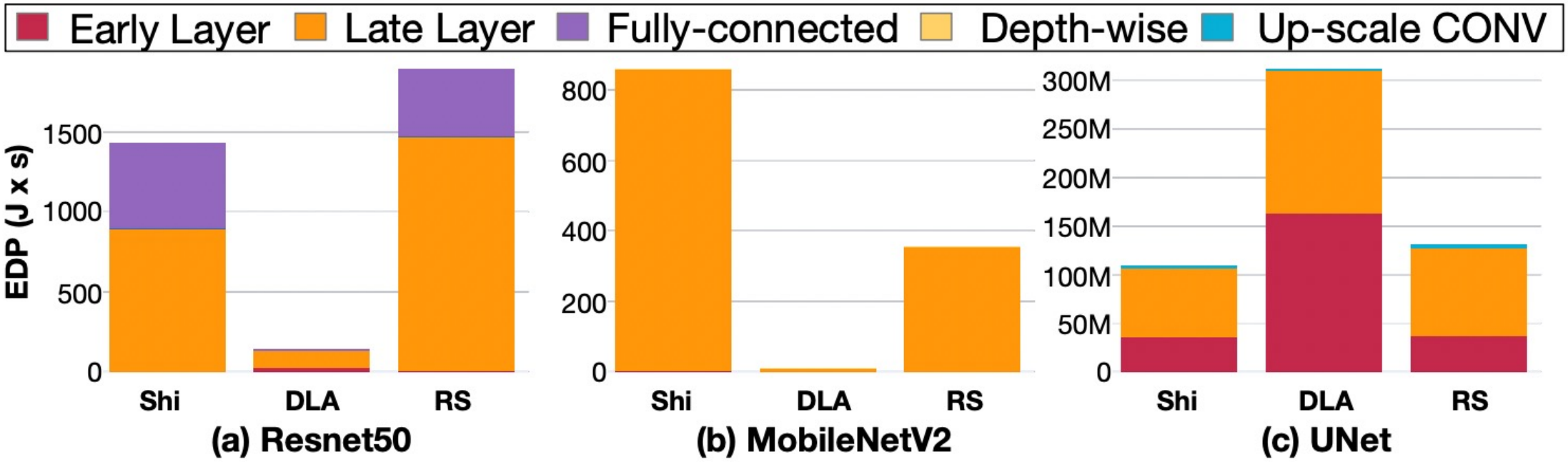


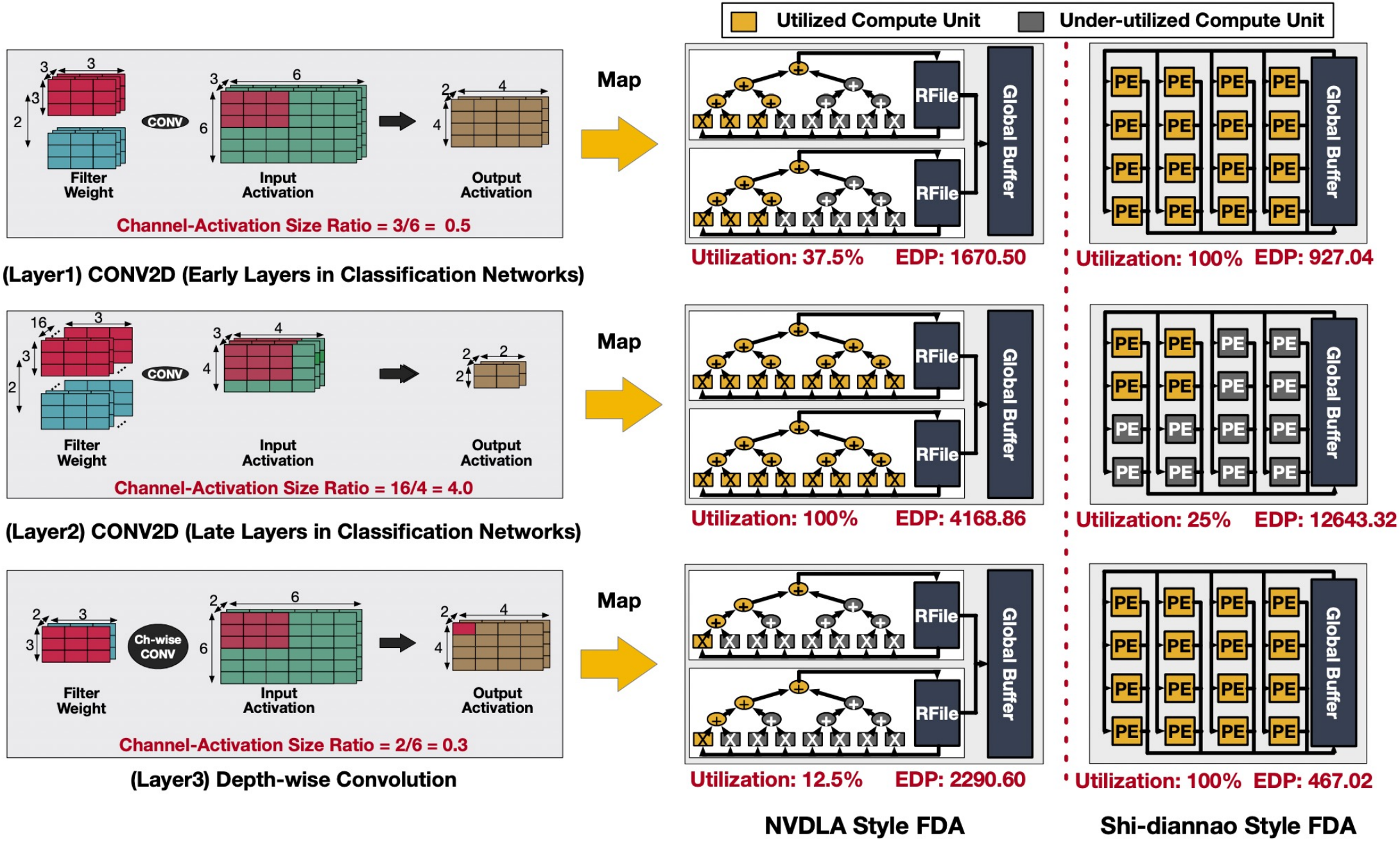
Fig. 5. Scheduling strategies of (a) DianNao [1] (b) Eyeriss [2] (c) Leuven [6] (d) Proposed KOP3.

[3] A 3.77TOPS/W Convolutional Neural Network Processor With Priority-Driven Kernel Optimization, *IEEE TCAD*, 2019

Various Dataflows inside NPU

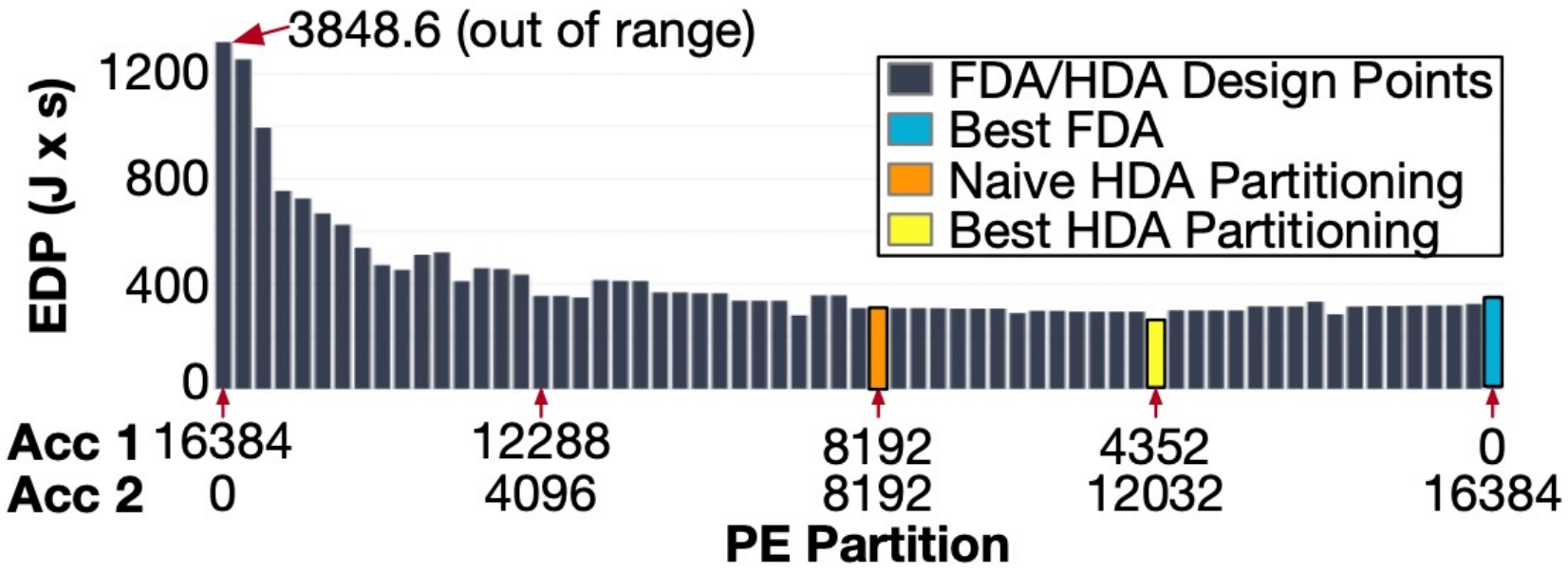
- Dataflow and Efficiency
  - Uniform dataflow leads to low throughput and energy efficiency
  - Dataflow: ShiDianNao, NVDLA, Eyeriss (Row-Stationary)




Utilized Compute Unit
Under-utilized Compute Unit
3
3
3
2
CONV
6
3
6
2
4
4
Filter Weight
Input Activation
Output Activation
Channel-Activation Size Ratio = 3/6 = 0.5
Map
RFile
RFile
Global Buffer
PE
Utilization: 37.5%
EDP: 1670.50
Utilization: 100%
EDP: 927.04
(Layer1) CONV2D (Early Layers in Classification Networks)
16
3
3
2
CONV
3
4
4
2
2
2
Filter Weight
Input Activation
Output Activation
Channel-Activation Size Ratio = 16/4 = 4.0
Map
Utilization: 100%
EDP: 4168.86
Utilization: 25%
EDP: 12643.32
(Layer2) CONV2D (Late Layers in Classification Networks)
2
3
3
Ch-wise CONV
2
6
6
2
4
4
Filter Weight
Input Activation
Output Activation
Channel-Activation Size Ratio = 2/6 = 0.3
Map
Utilization: 12.5%
EDP: 2290.60
Utilization: 100%
EDP: 467.02
(Layer3) Depth-wise Convolution
NVDLA Style FDA
Shi-diannao Style FDA

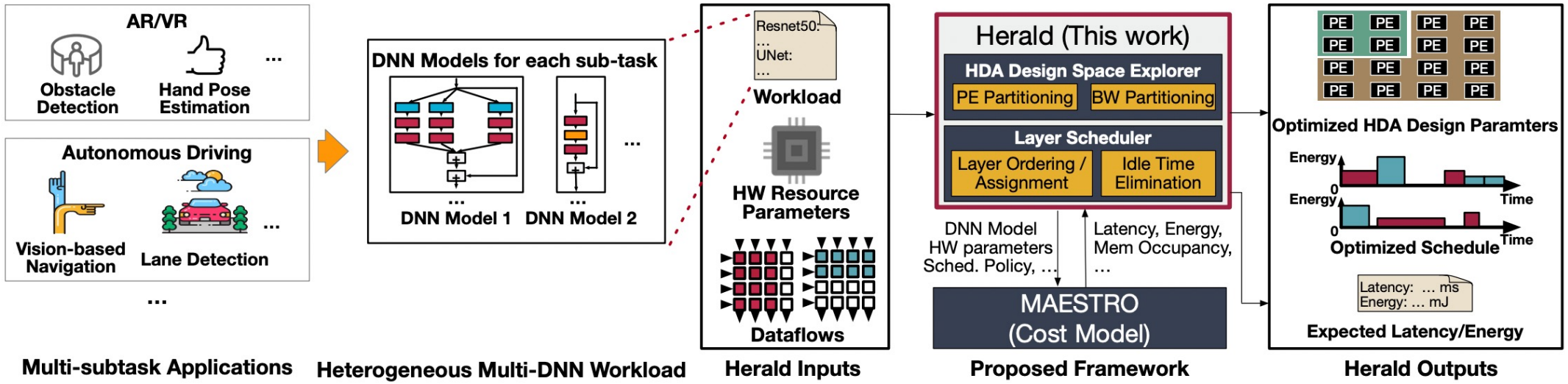
Heterogeneous Dataflows inside NPU

- Considerations of heterogenous dataflow accelerator (HAD)
  - HW resource partitioning: # of PE, NoC Bandwidth, Memory
  - Dataflow selection and Layer scheduling



Heterogeneous Dataflows inside NPU

- Implementation of HDA optimization framework
  - Optimization framework for exploring design space of HDA and schedules
  - Fixed HW configurations (NVDLA, ShiDianNao, Eyeriss)



Heterogeneous Dataflows inside NPU

- Implementation of HDA optimization framework
  - Optimization framework for exploring design space of HAD and schedules
  - Fixed HW configurations (NVDLA, ShiDianNao, Eyeriss)

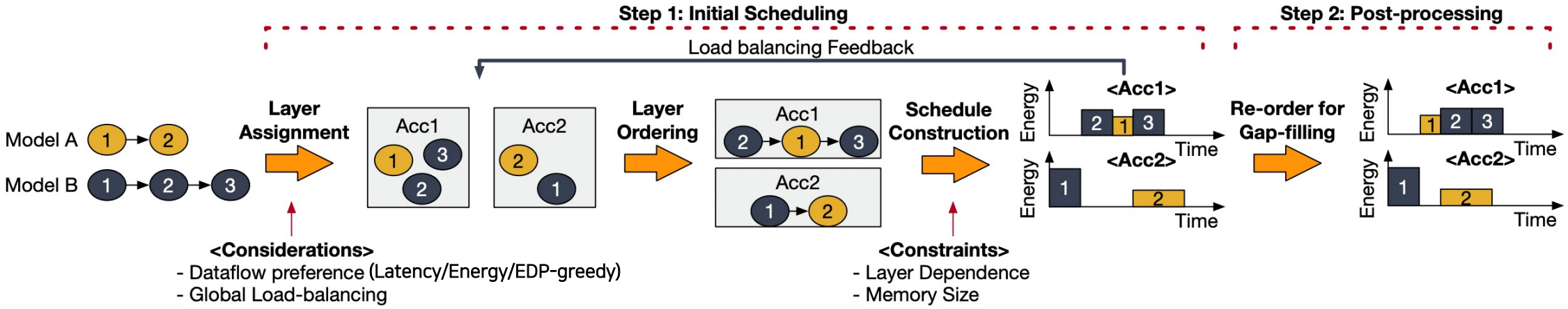
Workload	Model	# of batches
AR/VR-A	Resnet50	2
	Unet	4
	MobileNetV2	4
AR/VR-B	Resnet50	2
	Unet	2
	MobileNetV2	4
	BR-Q Handpose	2
	Focal Length DepthNet	2
MLPerf	Resnet50	1 (and 8 for batch size study)
	MobileNetV1	1 (and 8 for batch size study)
	SSD-Resnet34	1 (and 8 for batch size study)
	SSD-MobileNetV1	1 (and 8 for batch size study)
	GNMT (RNN)	1 (and 8 for batch size study)

Accelerator Class	Num. of PEs	NoC BW	Glob. Memory
Edge	1024	16 GB/s	4 MiB
Mobile	4096	64 GB/s	8 MiB
Cloud	16384	256 GB/s	16 MiB

Scenario	BW Partitioning (NVDLA / Shi)	PE Partitioning (NVDLA / Shi)
AR/VR-A, Edge	4 / 12	128 / 896
AR/VR-A, Mobile	40 / 24	1792 / 2304
AR/VR-A, Cloud	224 / 32	9728 / 6656
AR/VR-B, Edge	4 / 12	128 / 896
AR-VR-B, Mobile	48 / 16	1536 / 2560
AR/VR-B, Cloud	128 / 128	12032 / 4352
MLPerf, Edge	4 / 12	64 / 960
MLPerf, Mobile	32 / 32	1280 / 2816
MLPerf, Cloud	160 / 96	8192 / 8192

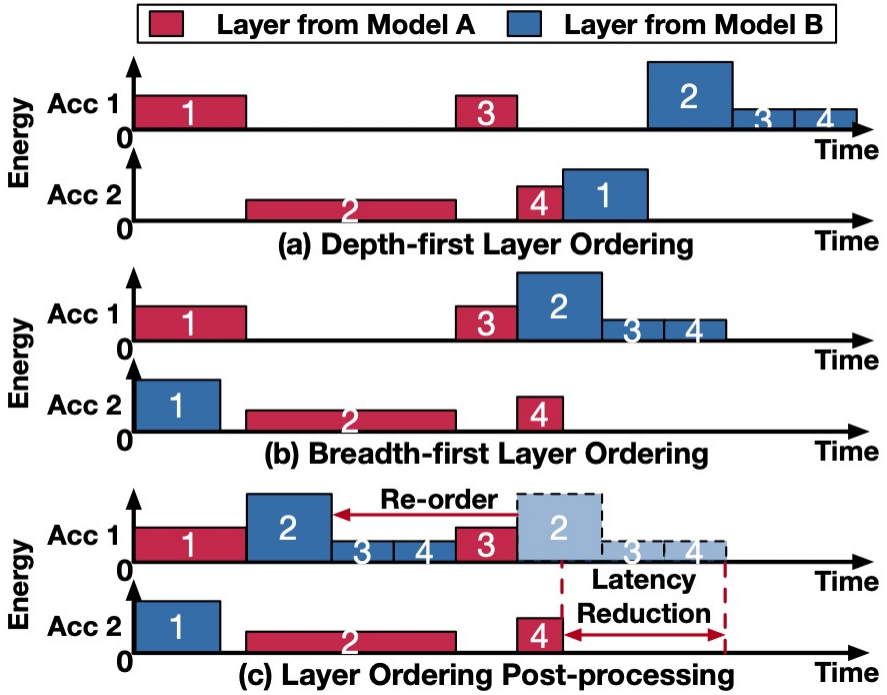
Heterogeneous Dataflows inside NPU

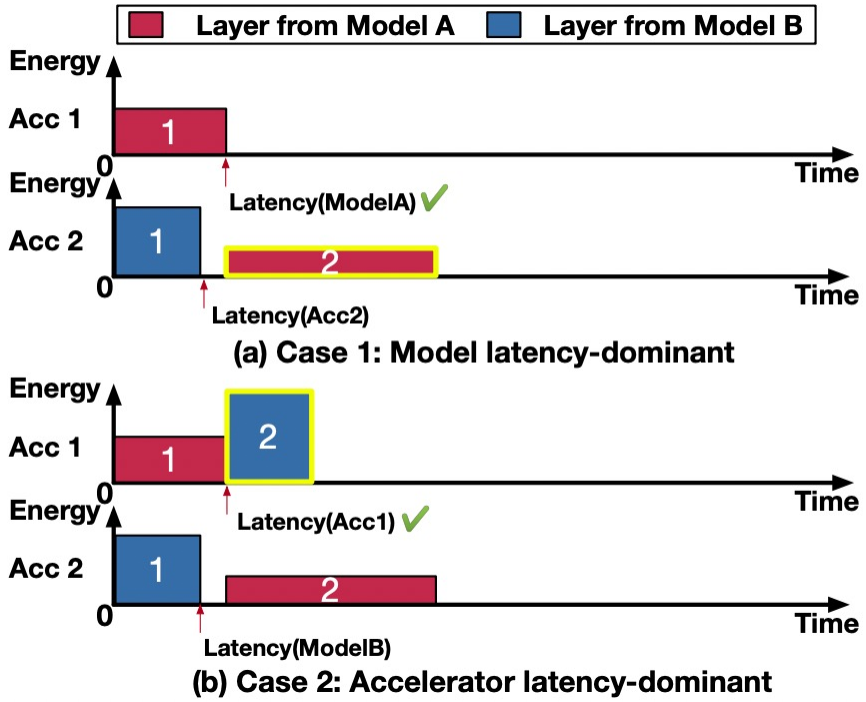
- Implementation of HDA optimization framework
  - Layer scheduling: Layer **assignment**/ordering and timeline construction
    - Latency/Energy/EDP based greedy method



Heterogeneous Dataflows inside NPU

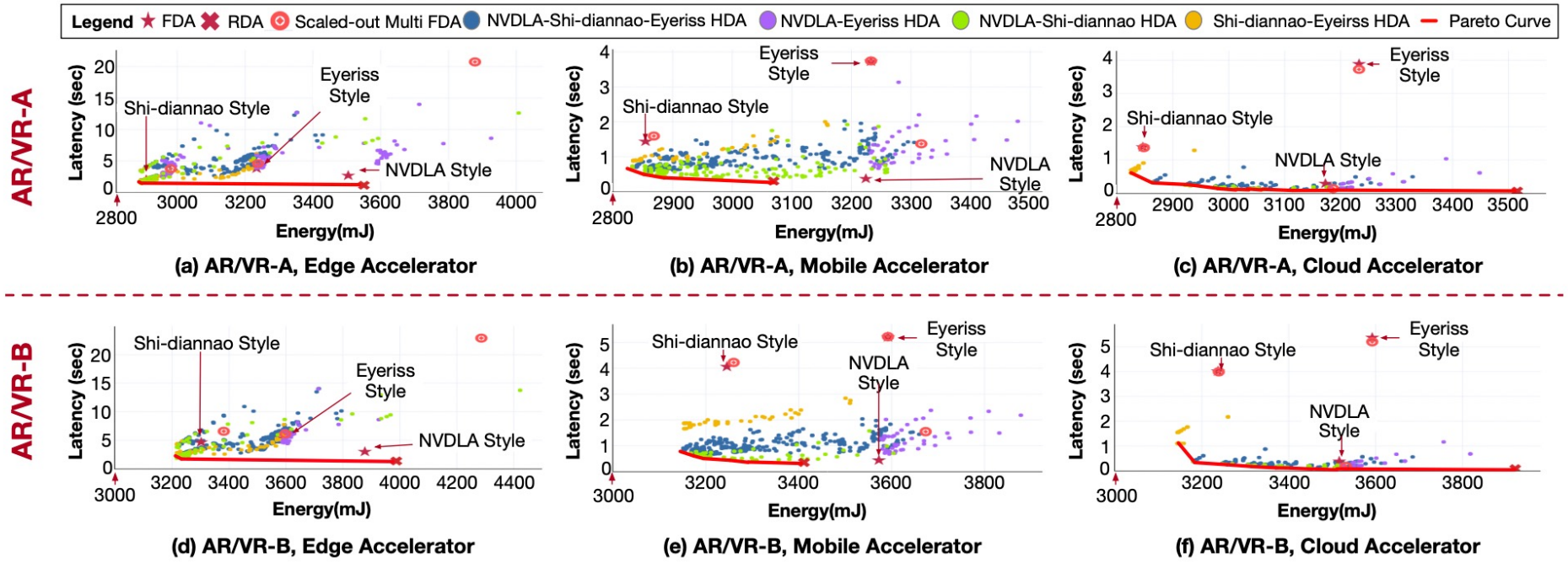
- Implementation of HDA optimization framework
  - Layer scheduling: Layer assignment/**ordering** and timeline construction
    - Depth (First model and Second model), Second (Interleaving method)





Heterogeneous Dataflows inside NPU

- Evaluation
  - FDA/RDA (Fixed/Reconfigurable), Scaled-out (Dual), Pareto (Optimal)
    - HDA shows 65.3% and 55.0% lower latency and energy than that of best FDA



Heterogeneous Dataflows inside NPU

- Evaluation
  - FDA/RDA (Fixed/Reconfigurable), Scaled-out (Dual), Pareto (Optimal)
    - HDA shows 65.3% and 55.0% lower latency and energy than that of best FDA

LATENCY AND ENERGY GAIN AGAINST THE FDA AND RDA WITH THE BEST EDP ON VARIOUS BATCH SIZES ON MLPERF WORKLOAD.

Acc. Class	Batch Size	Latency Gain (vs FDA / vs RDA)	Energy Gain (vs FDA / vs RDA)
Edge	1	12.4% / -8.2%	0.2% / 20.4%
	8	21.28% / 26.7%	10.8% / 22.9%
Mobile	1	12.4% / -8.2%	0.2% / 17.1%
	8	56.0% / 76.1%	1.3% / 43.5%
Cloud	1	20.2% / 25.7%	10.8% / 26.8%
	8	63.9% / 80.4%	1.34% / 41.3%

Heterogeneous Dataflows inside NPU

- Evaluation
  - Scheduling time
    - I9-9880H, 16GB memory environment
    - On average, 11.09ms per layer and per HAD design point

AVERAGE TIME REQUIRED FOR SCHEDULING EACH WORKLOAD ON HDAS.

Workload	# Layers	# sub-accelerators	Scheduling Time (s)
AR/VR-A	448	2	2.89
		3	4.32
AR/VR-B	618	2	3.98
		3	10.74
MLPerf	181	2	1.61
		3	3.22

Thank you