HNPU-V2: A 46.6 FPS DNN Training Processor for Real-World Environmental Adaptation based Robust Object Detection on Mobile Devices

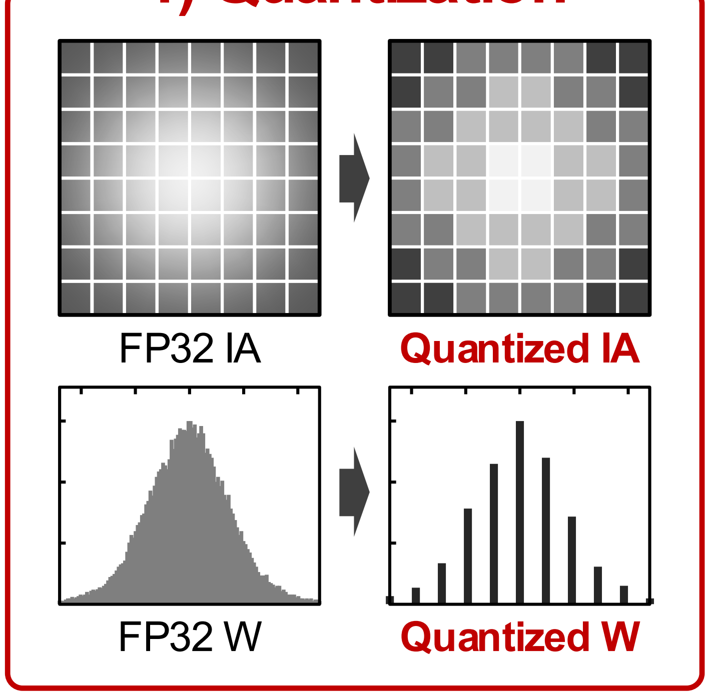
Donghyeon Han, Dongseok Im, Gwangtae Park, Youngwoo Kim, Seokchan Song, Juhyoung Lee, and Hoi-Jun Yoo

Semiconductor System Lab.
School of EE, KAIST

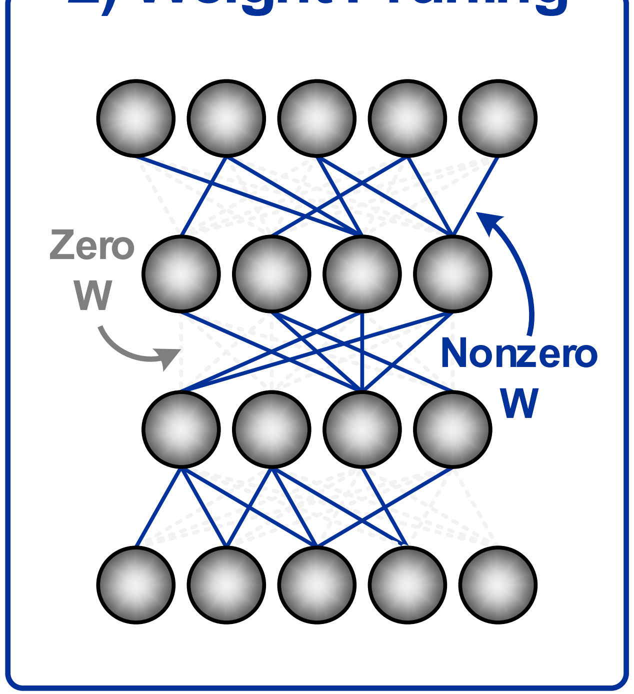
Development of DNN for Mobile Platforms

- **Smarter DNNs: # of Parameter ▲**
- **Lightweight DNNs for Mobile Devices**
  - Quantization, weight pruning, pointwise or depthwise CONV …

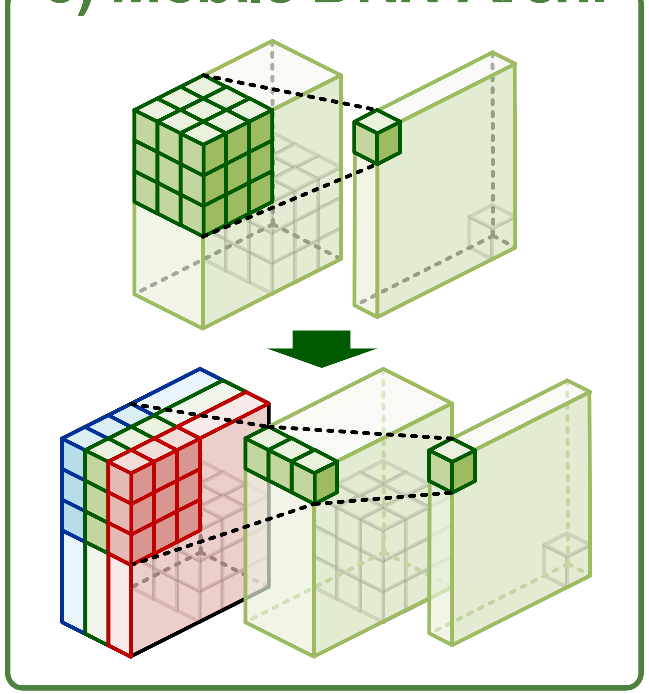
1) Quantization



2) Weight Pruning

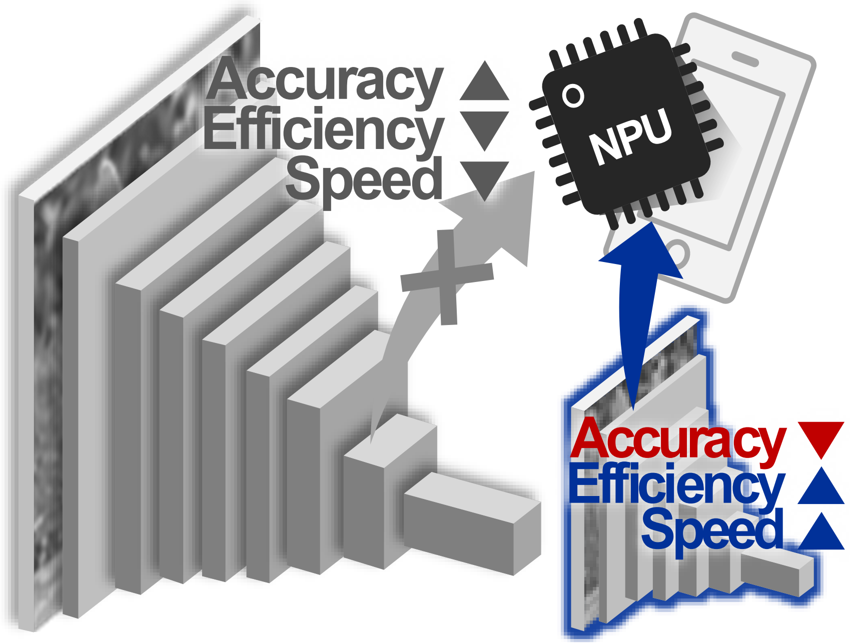


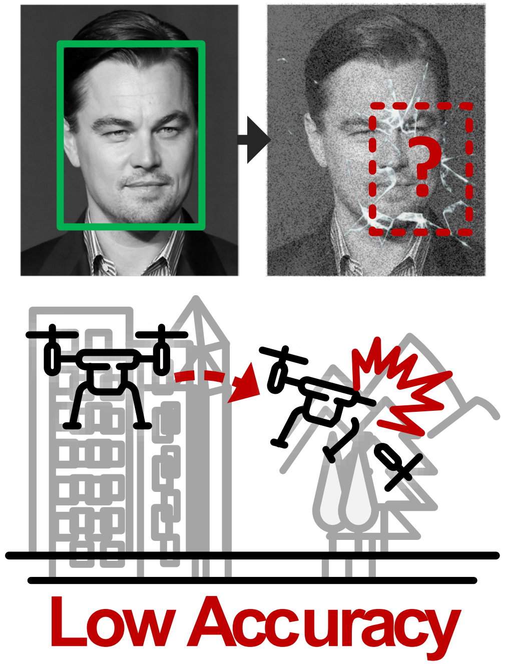
3) Mobile DNN Arch.

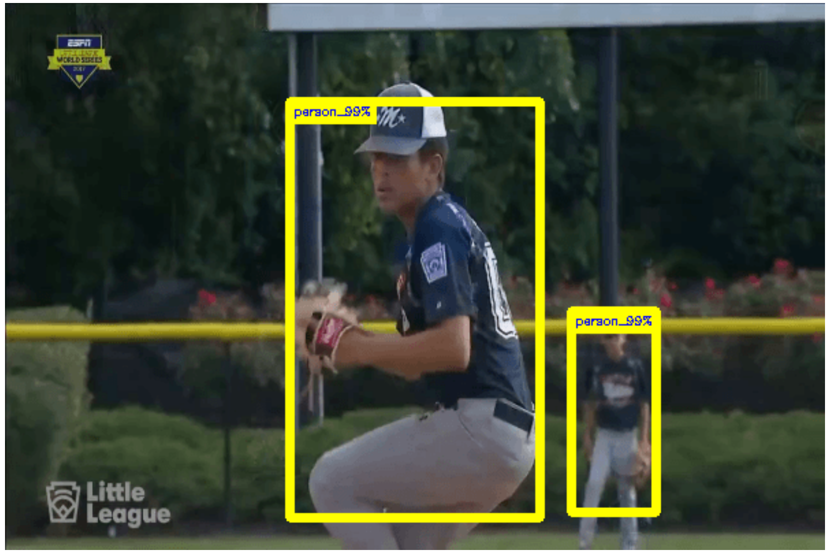


Disadvantages of Mobile-oriented DNNs

- **Low Detection Accuracy in Practice**
- **Performance Degradation After Unexpected Situations**
  - – Low network capacity → Loosing generality → Sensitive to accident

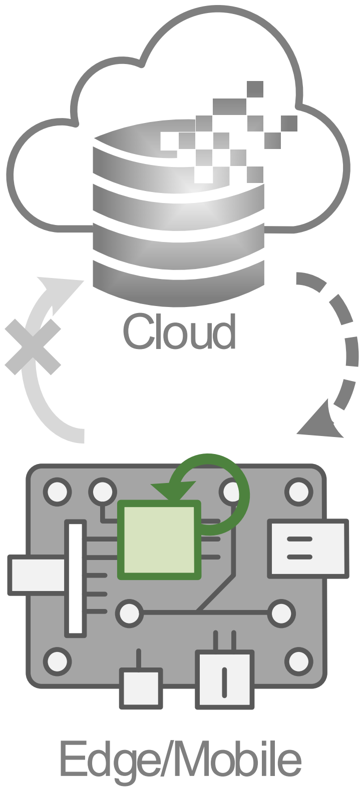


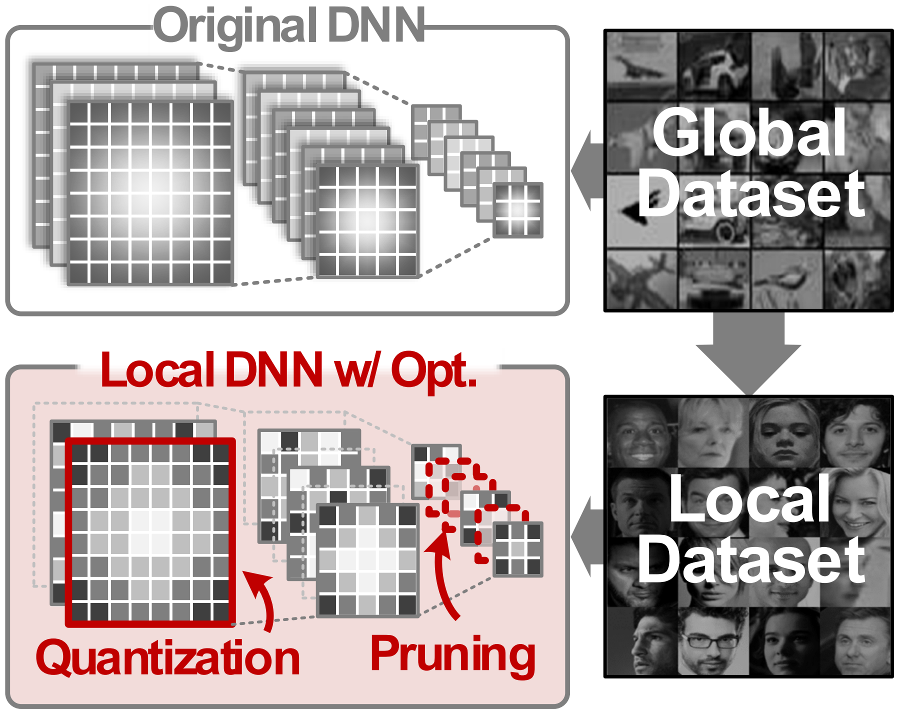




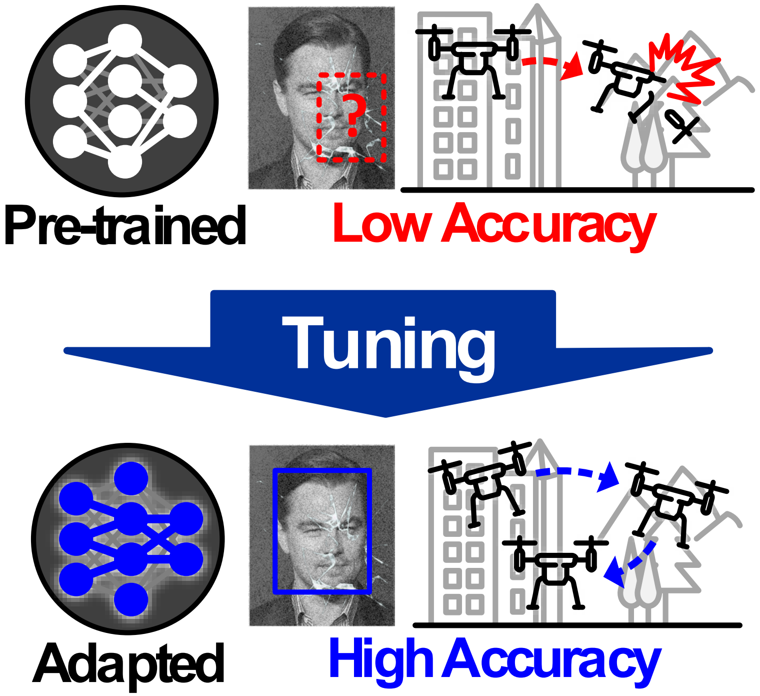
Promising Solution: On-device DNN Training

- **Personalization:** High Accuracy only for User-specific Task
- **Adaptation:** Performance Recovery using Online Tuning





1) Optimization w/ Pruning/Quant.

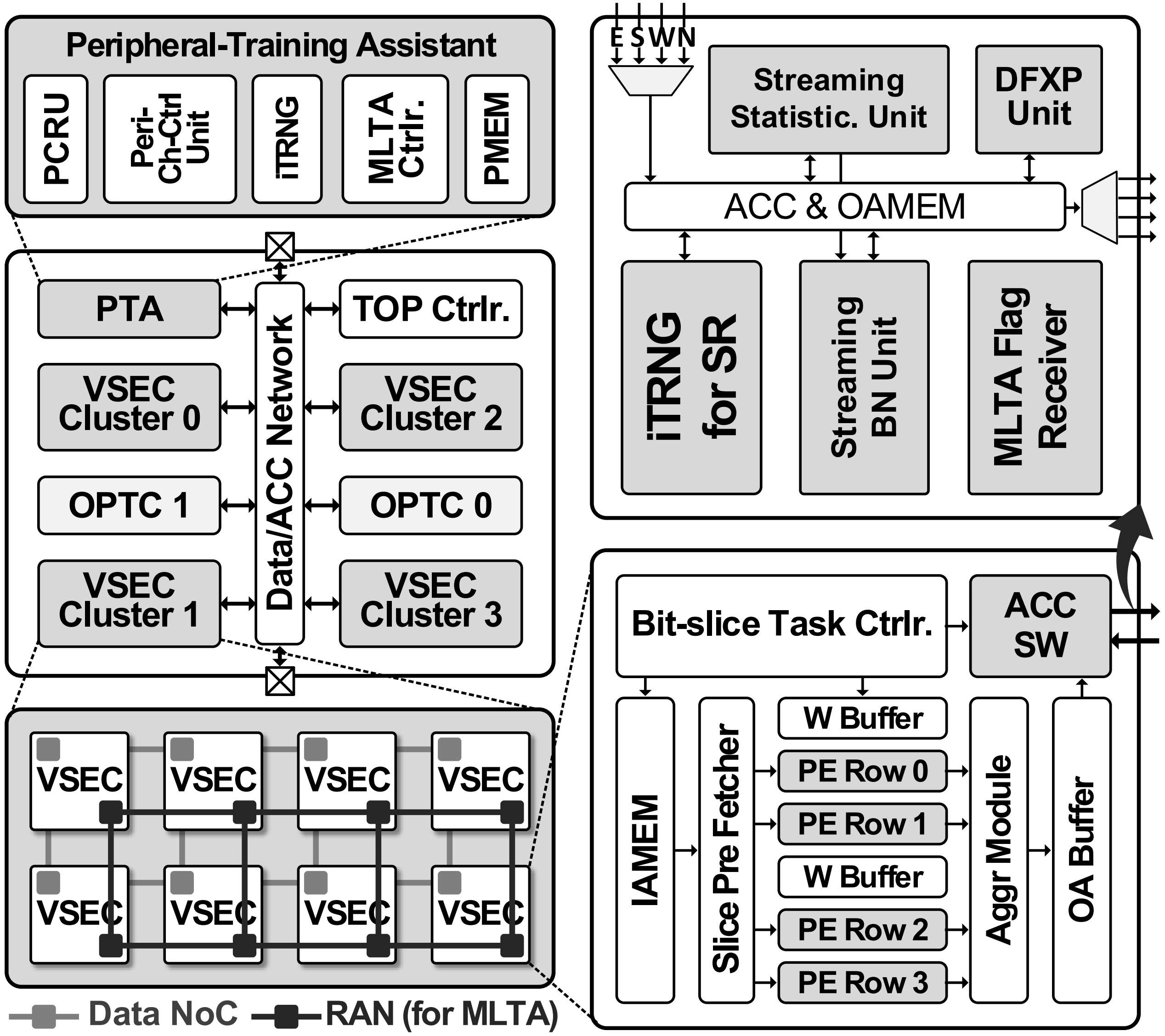


2) Low-latency DNN Tuning

Overall Architecture of HNPU-V2

**Dynamic Fixed-point*
*** Optimizer Core*
**** Peripheral Training Assistant*

- **32 Versatile Sparsity Exploitation Cores (VSEC)**
  - Bit-slice = 4b
  - Computing unit: 4b×4b
  - Support (4,8,12,16)-bit
  - DFXP* + Stochastic rounding
  - Input slice skip
  - Weight skip
- **2 OPTC**s & 1 PTA*****
- **2-D Mesh NoC**



Overall Architecture of HNPU-V2

*True-Random Number Generator
** Pruning-aware Channel Reordering Unit
***Reconfigurable Accumulation Network

1. intrinsic-TRNG*

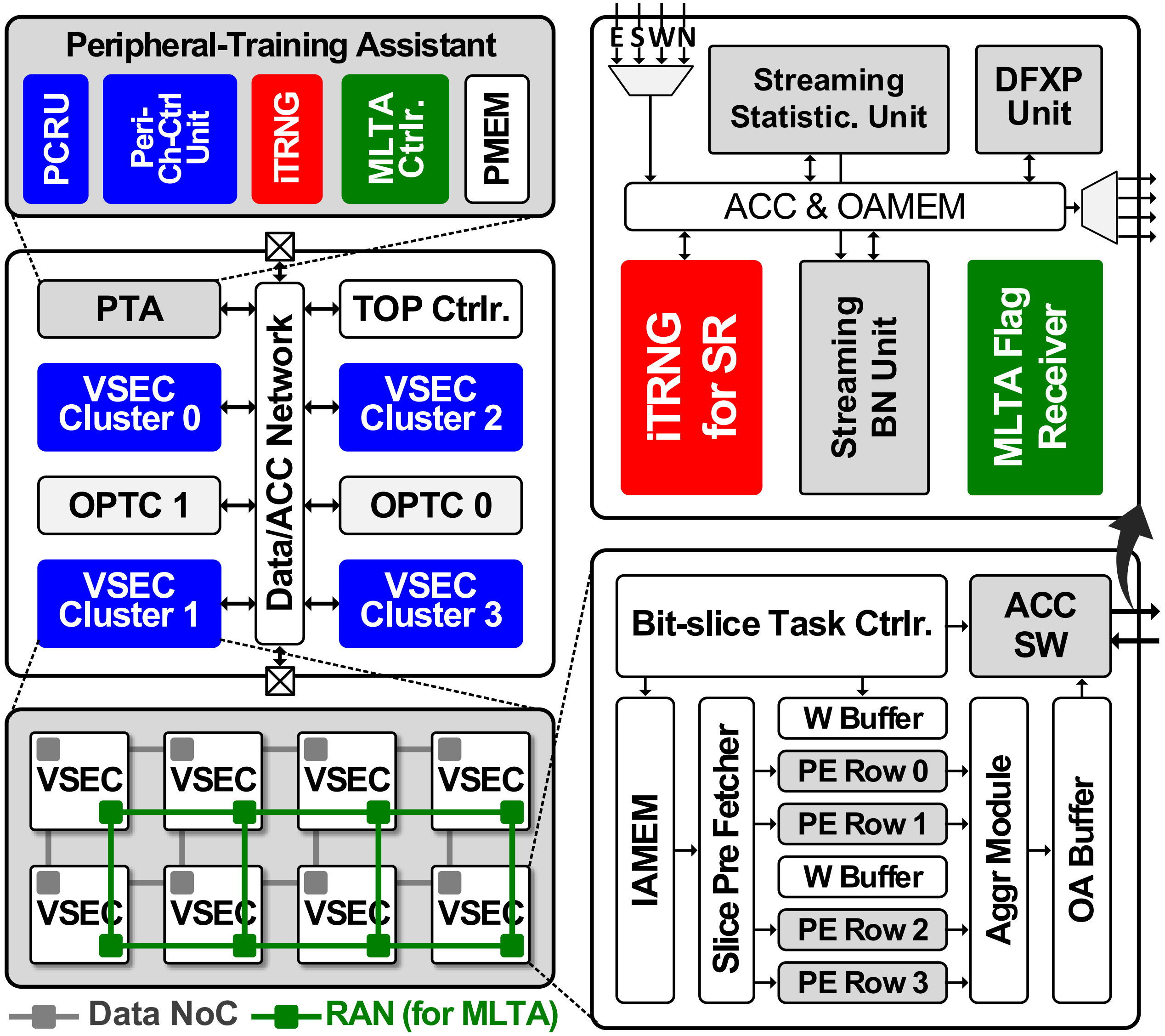
- Truly random bit-streams
- Stochastic rounding for low-precision training

2. VSEC w/ PCRU**

- Input zero-slice skipping
- Pruned Ch skip

3. Multi-Learning Task Aalloc.

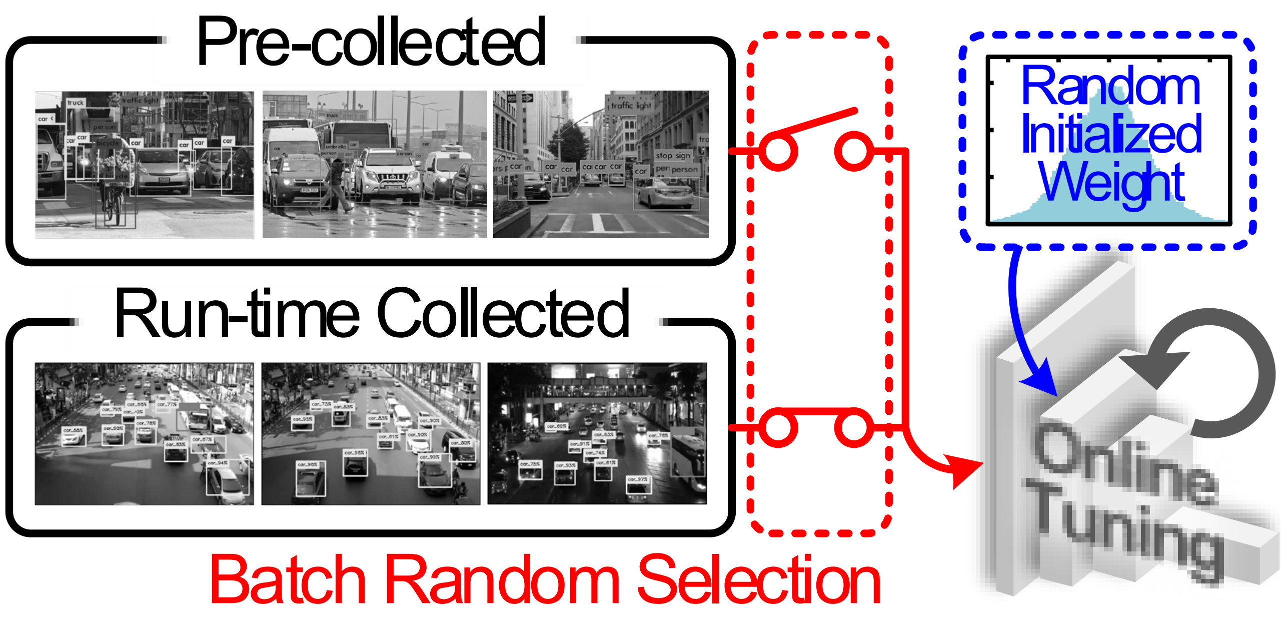
- Flag based RAN*** control
- To support backward unlocking

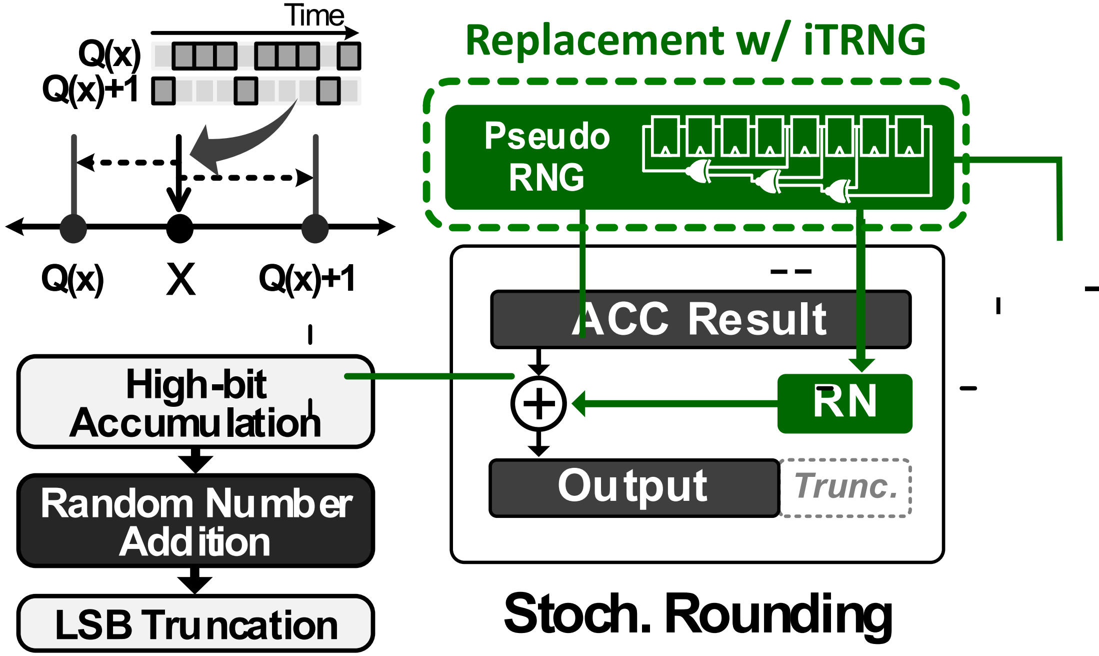


Various Usages of RNG

*D. Han, JSSC'21

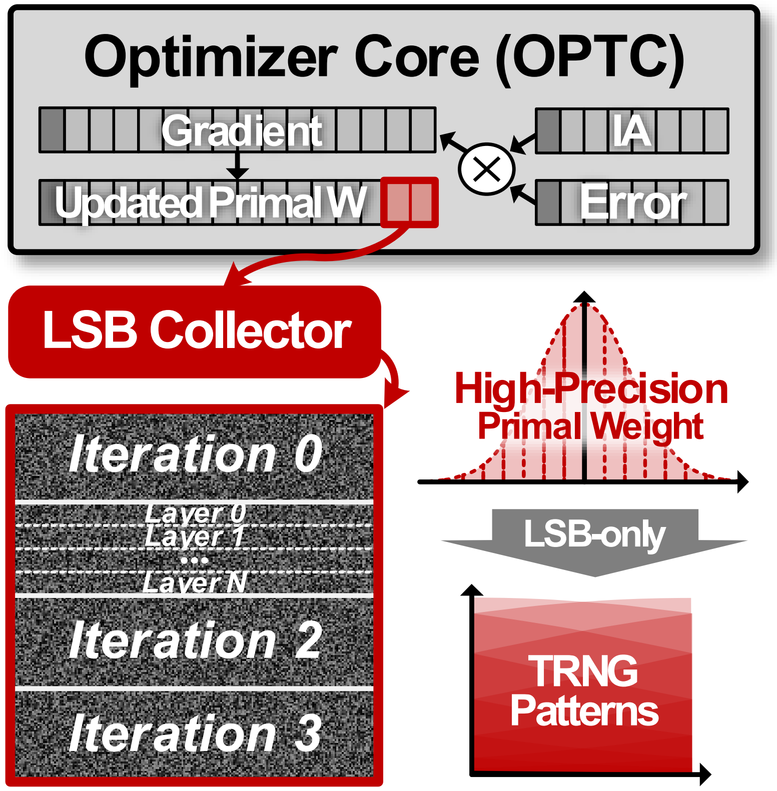
- **Example 1: Basic Training Functionality**
  - – Ex) Weight initialization, batch selection,
- **Example 2: Stochastic Rounding***
  - – For Low-precision computing during the FXP based DNN training



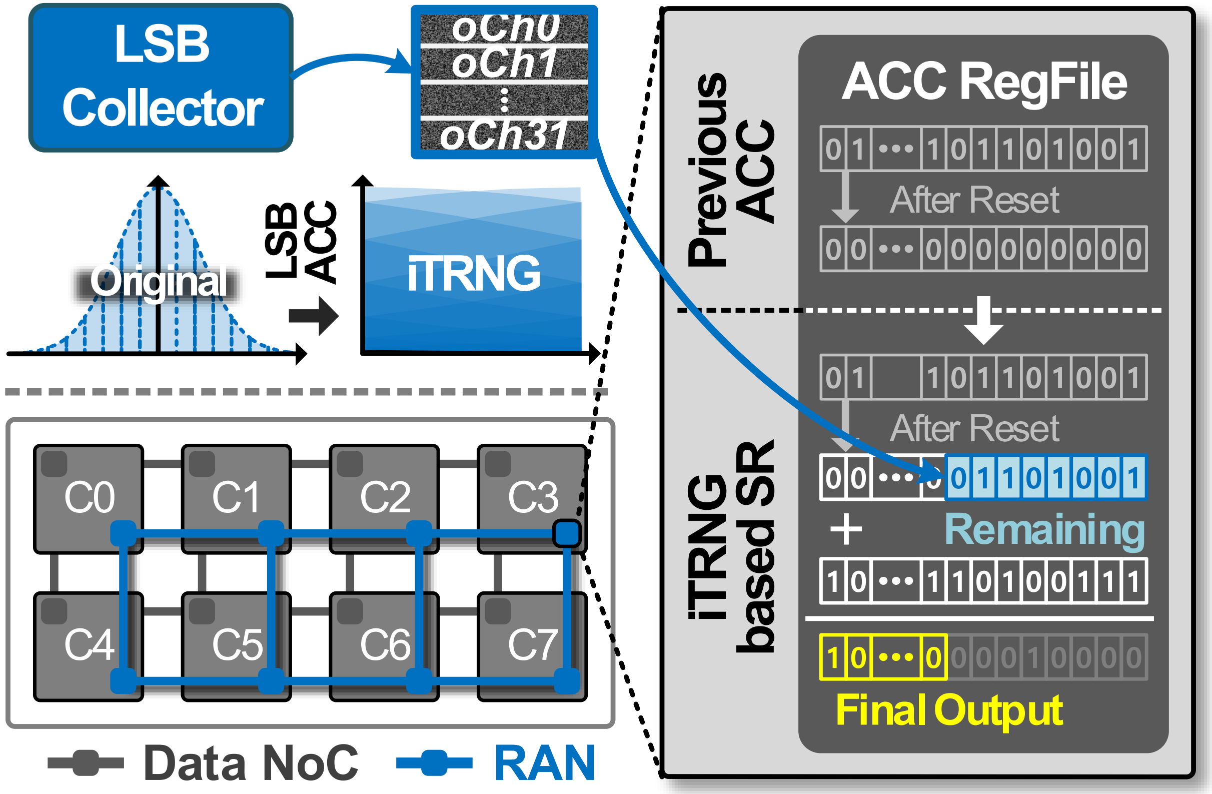


Two Different Types of iTRNGs in HNPU-V2

- **1st iTRNG: Placed in PTA**
  - Extracting & cumulating LSB bit-stream of primal weight



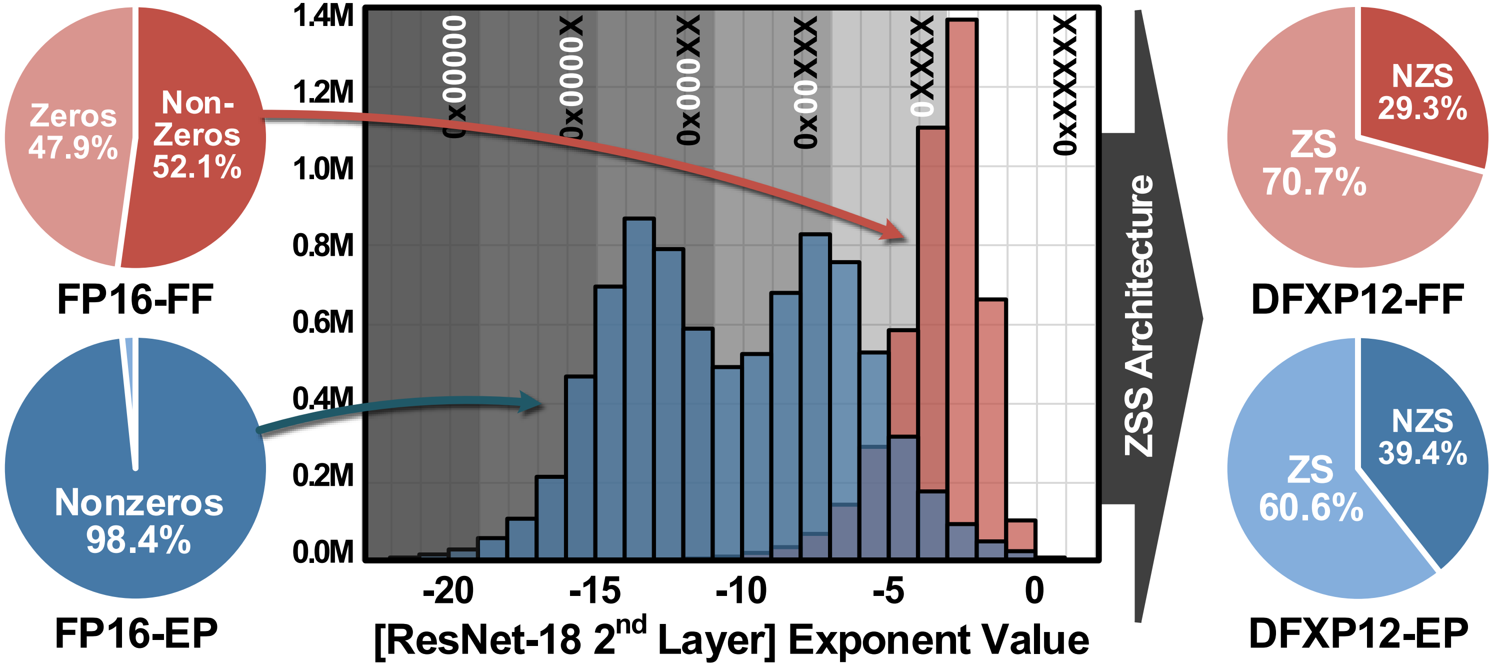
- **2nd iTRNG: Placed in ACC SW**
  - Adding random noise → Remaining LSB bit-stream of accum results

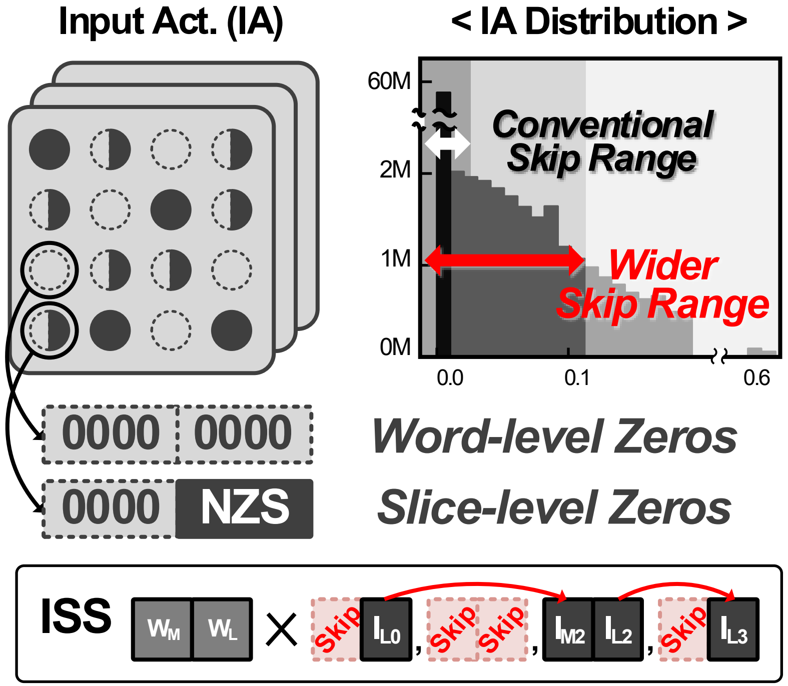


Input-slice Skipping* (ISS)

**D. Han, JSSC'21*

- **Bit-slice-level Sparsity Exploitation**
  - Most of data: placed near zero (Gaussian-like distribution)
  - Giving possibility of skipping MSB zeros even with non-ReLU / EP Stage

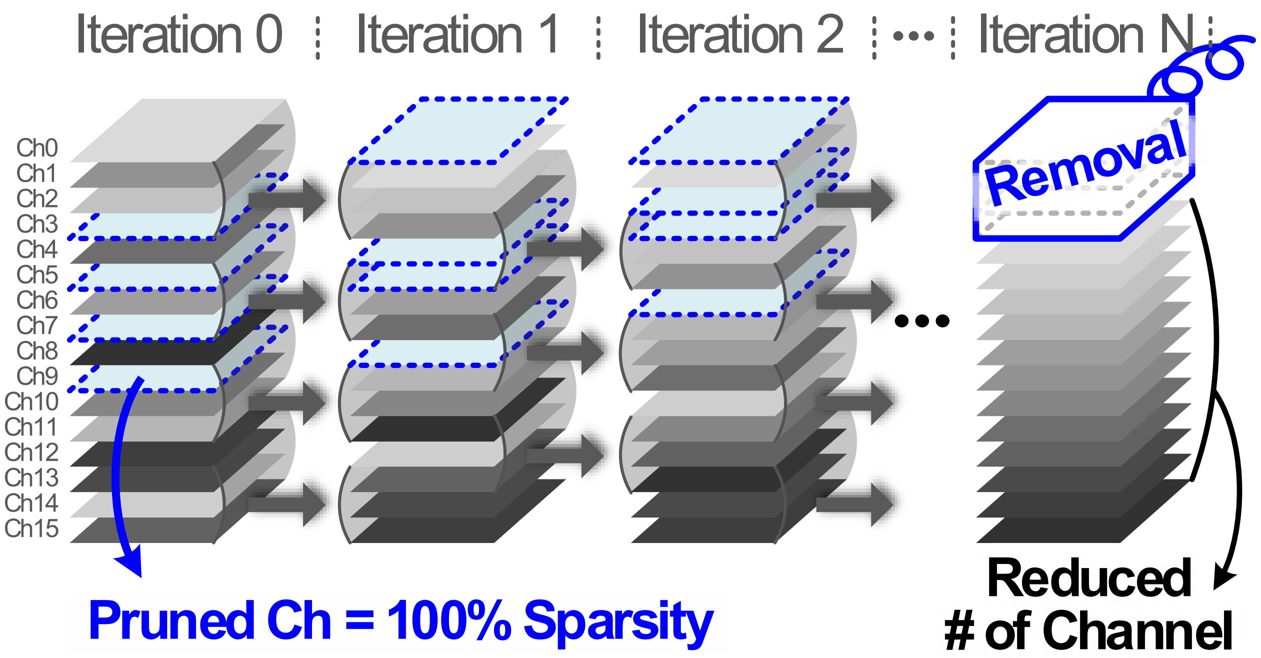


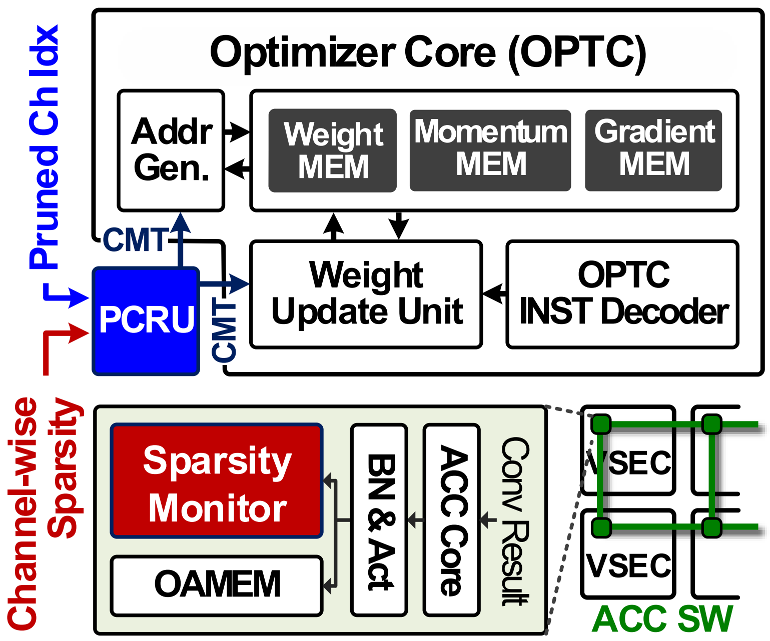


Pruning-aware Ch Reordering Unit (PCRU)

- **Supporting Weight Pruning w/ Channel Removal**
  - Receiving pruned channel index → Considering Ch as 100% sparsity
  - OPTC: updating weights by referring the pruning-aware Ch mapping table

    → New weight: changed order & excluded Ch



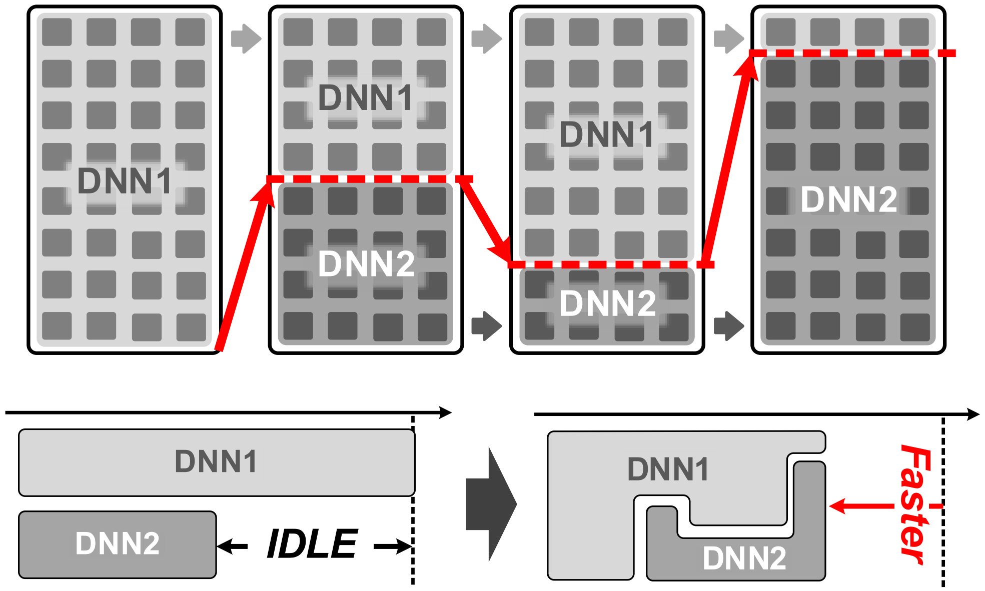


Opposite Properties of Two Processors

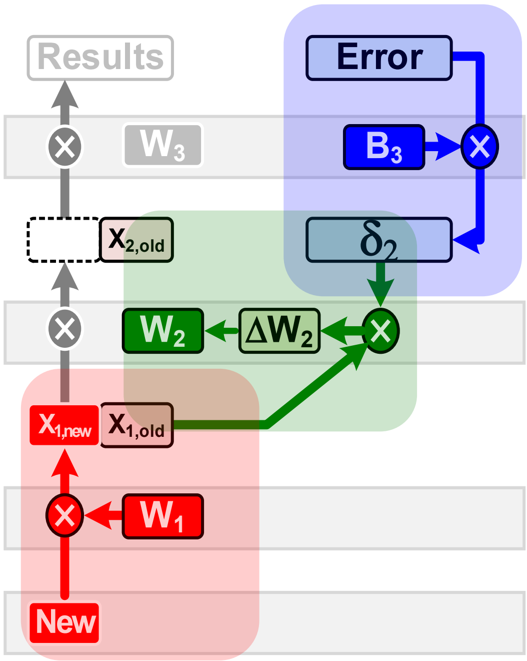
**S. Kang, ISSCC'21*
***D. Han, S.VLSI'19*

- **Problems of GANPU***
  - High Reconfigurability
  - Single-LT Supporting
    - Back-propagation (FF → EP → WG)

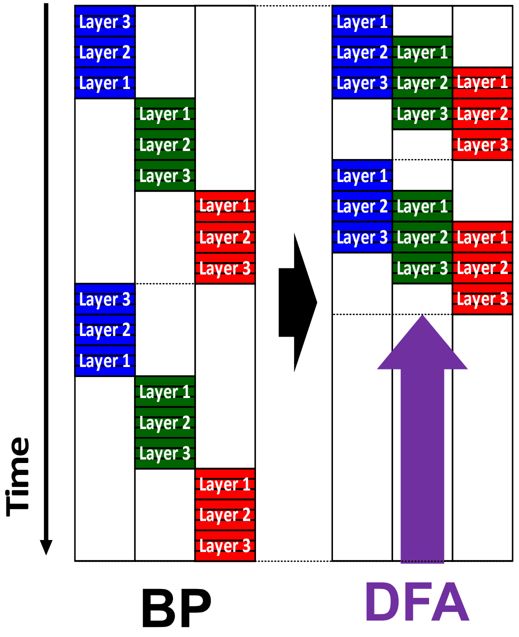
- **Problems of DF-LNPU****
  - Low Reconfigurability
  - Multi-LT Allocation
    - Backward unlocking (e.g. DFA)



Workload Opt. w/ Dynamic Core Alloc.

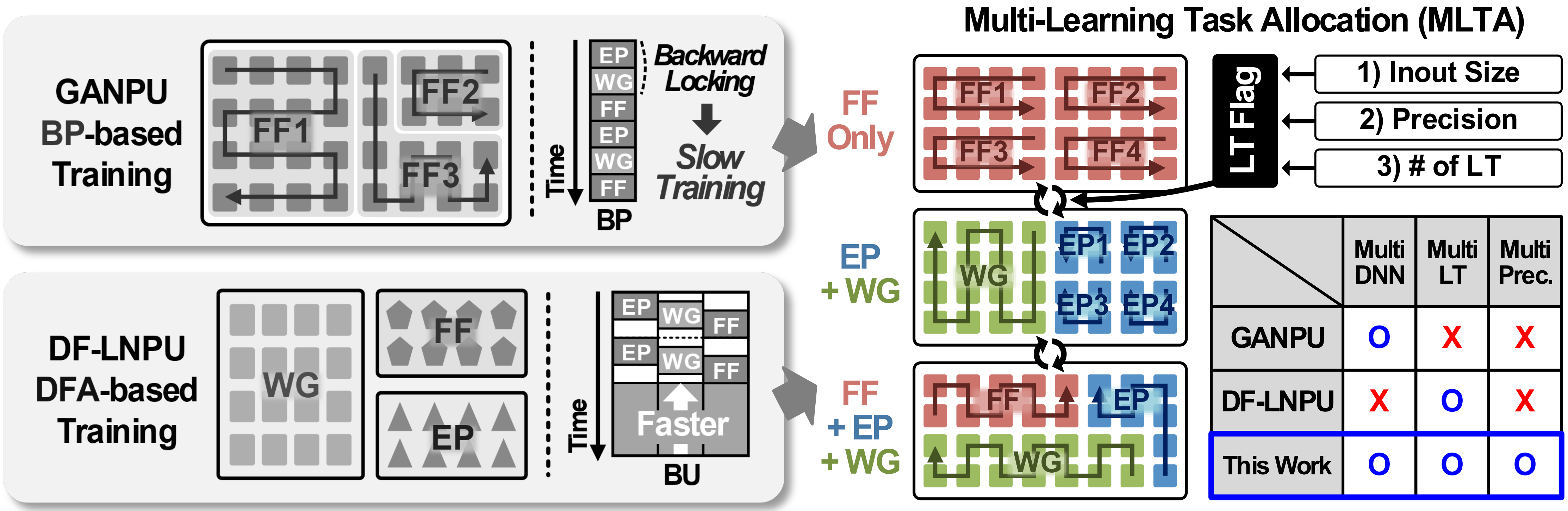


Training Timeline



Multi-Learning Task Allocation in HNPU-V2

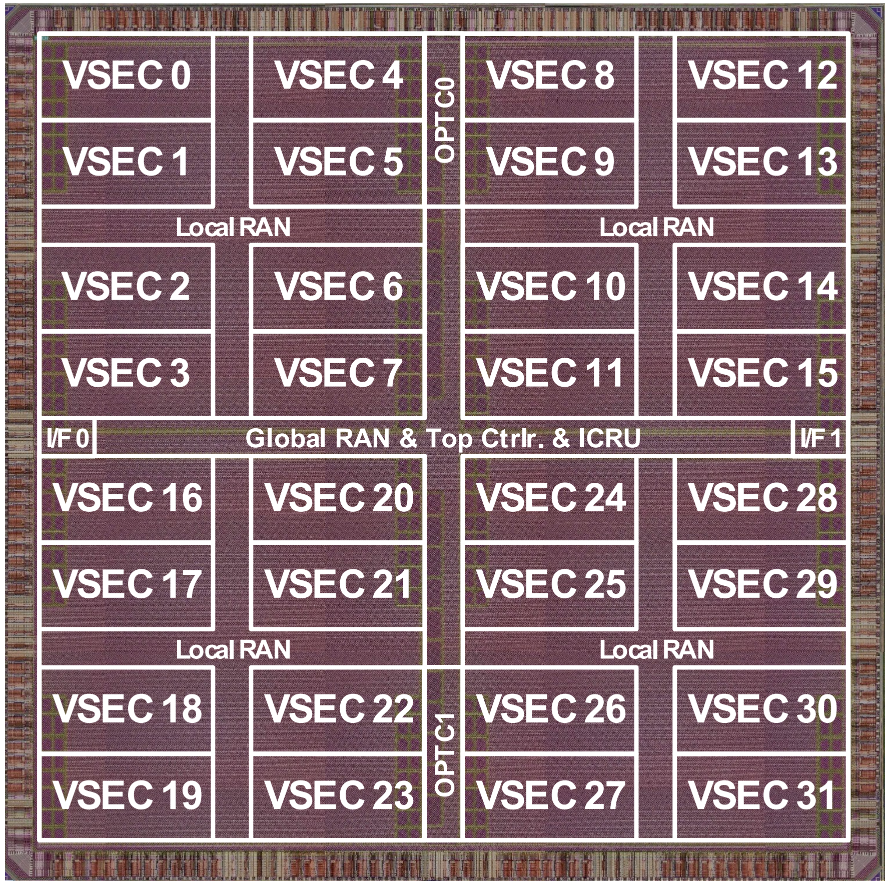
- **RAN w/ Learning-Task-flag → Indicating Training Stages**
- **Dynamic Core Allocation according to Three Parameters**
  - 1) Inout Size, 2) Bit-precision, 3) Learning tasks (Training stages)



	Multi DNN	Multi LT	Multi Prec.
GANPU	O	X	X
DF-LNPU	X	O	X
This Work	O	O	O

Chip Summary

- **Chip Photograph & Performance Summary**



I/O Voltage = 1.8V, 1 MAC = 2 OP

	Specifications			
Technology	Samsung 28nm 1P8M CMOS			
Die Area	3.6mm × 3.6mm (12.96mm^2)			
Supporting BU	**BP, DFA [6], ... (Programmable)**			
Op. Condition	0.63V (@ 10MHz) ~ 1.0V (@ 250MHz)			
Data Type	DFXP + **SR** (4/8/12/16)-bit × (4/8/12/16)-bit			
Area Efficiency	59-to-9334 GOPS/mm^2			
	Sparsity (IS, W) [%]	**(0,0)**	**(50,50)**	**(90,90)**
Power [mW]	250MHz, 1.0V	1032	850	616
	10MHz, 0.63V	24.1	19.8	14.6
Energy Efficiency @ 10MHz [TOPS/W]	16b×16b	2.04	6.01	98.1
	8b×8b	6.81	19.9	220
	4b×4b	20.4	49.8	**332**

Chip Performance Comparison

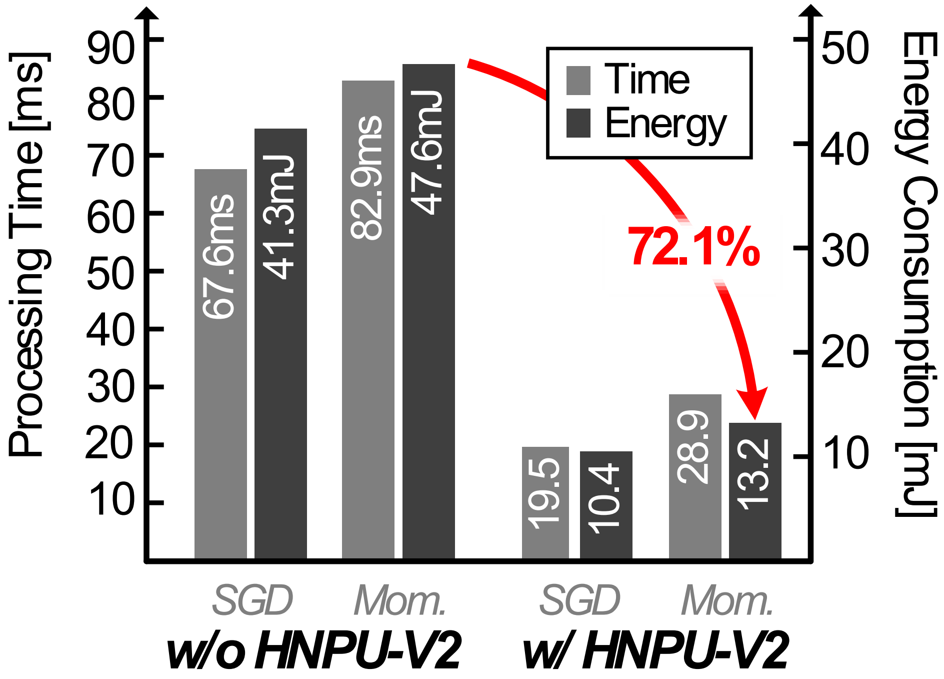
	JSSC'20	ISSCC'20	S.VLSI'20	HNPU-V1	**HNPU-V2**	
Backward Unlocking	O	X	X	X	**O**	
Low-precision Training	DFXP	X	X	SDFXP	**DFXP+SR**	
Robustness for Non-ReLU	X	X	X	Zero-slice Skip	**Zero-slice Skip**	
Technology	65nm	65nm	65nm	28nm	**28nm**	
MAX Core Frequency	200MHz	200MHz	200MHz	250MHz	**250MHz**	
Supporting Precision	FXP 13/16	FP 8/16	FP 8/16	FXP 4/8/12/16	**FXP 4/8/12/16**	
Throughput [GOPS]	155	1080	763	4526	**7072**	
Area Efficiency* [(GOPS or GFLOPS)/mm^2]	26.9	33.3	47.7	349	**545**	**56% ▲**
Energy Efficiency* [(TOPS or TFLOPS)/W]	0.62	1.67	1.79	4.74	**7.89**	**67% ▲**

**Measured @ Object Detection Scenario (Tiny-yolo-v3 w/ 20% Pruning)*

Object Detection Performance Comparison

- **Highest Framerate: 46.6 FPS**
- **Lowest Energy Consumption: 0.95 mJ/frame**

Processing Time & Energy Consumption



< Object Detection Comparison Table >

	TCAS-I '18	JSSC '20	A-SSCC '21	HNPU V1	**HNPU V2**
# of Detection*	1	1	> 1	> 1	**> 1**
Backward Unlocking	X	O	X	X	**O**
Operating Frequency	200 MHz	200 MHz	200 MHz	200 MHz	**100 MHz**
Framerate (FPS)	30.4	34.4	25.2	26.7	**46.6**
Energy per Frame [mJ/frame]	4.14	4.88	2.13	1.68	**0.95**

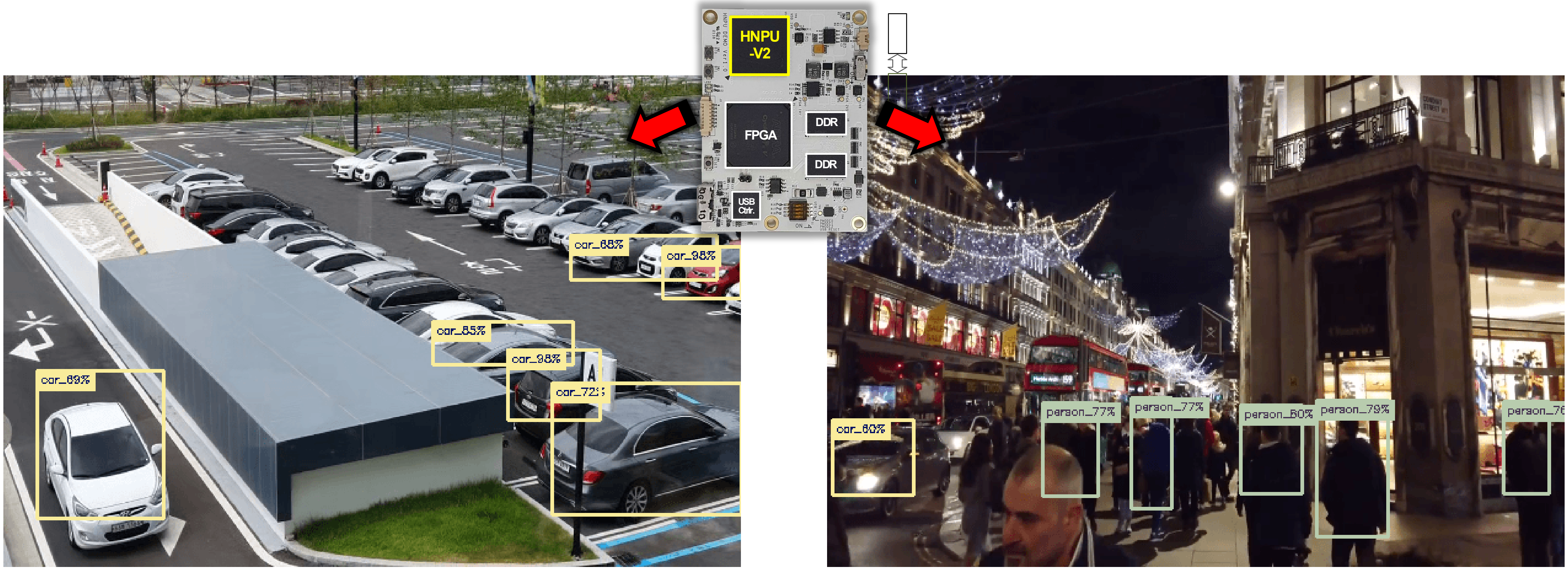
75% ▲

44% ▼

**Total number of detections in a single image*

Robust Object Detection w/ DNN Training

- HNPU-V2: Online DNN Tuning for Accuracy Compensation
- Automatic Accuracy Recovery from Unexpected Situations



Conclusion

- **intrinsic-TRNG**
  - On-chip random number generation for DNN training functionality
  - Stochastic rounding → Low-precision DNN training
- **Versatile Sparsity Exploitation Core with PCRU**
  - Input-slice skipping w/ workload balancing
  - Pruning-aware online DNN tuning by supporting channel removal
- **Multi-Learning-Task-Allocation w/ LT-flag based RAN Control**
  - Enable BU for low-latency online DNN tuning

HNPU-V2: A 0.95 mJ/frame DNN Training Processor for 46.6 FPS Real-time Environmental Adaptation

Thank You!

- **Questions? Feel Free to Contact Me!**
  - E-mail: hdh4797@kaist.ac.kr
  - LinkedIn: https://www.linkedin.com/in/donghyeon-han-90b439170
  - Personal Web-site:
    - https://hdh4797.wixsite.com/dhan
    - https://www.youtube.com/channel/UC1JOzBOZtHnWPEgP2QVdRQQ/
  - Zoom Meeting: https://zoom.us/j/6238458176?pwd=QldmbnhDOWNFdU9wcDhIKzdDN2ZiZz09 (Password: Donghyeon)