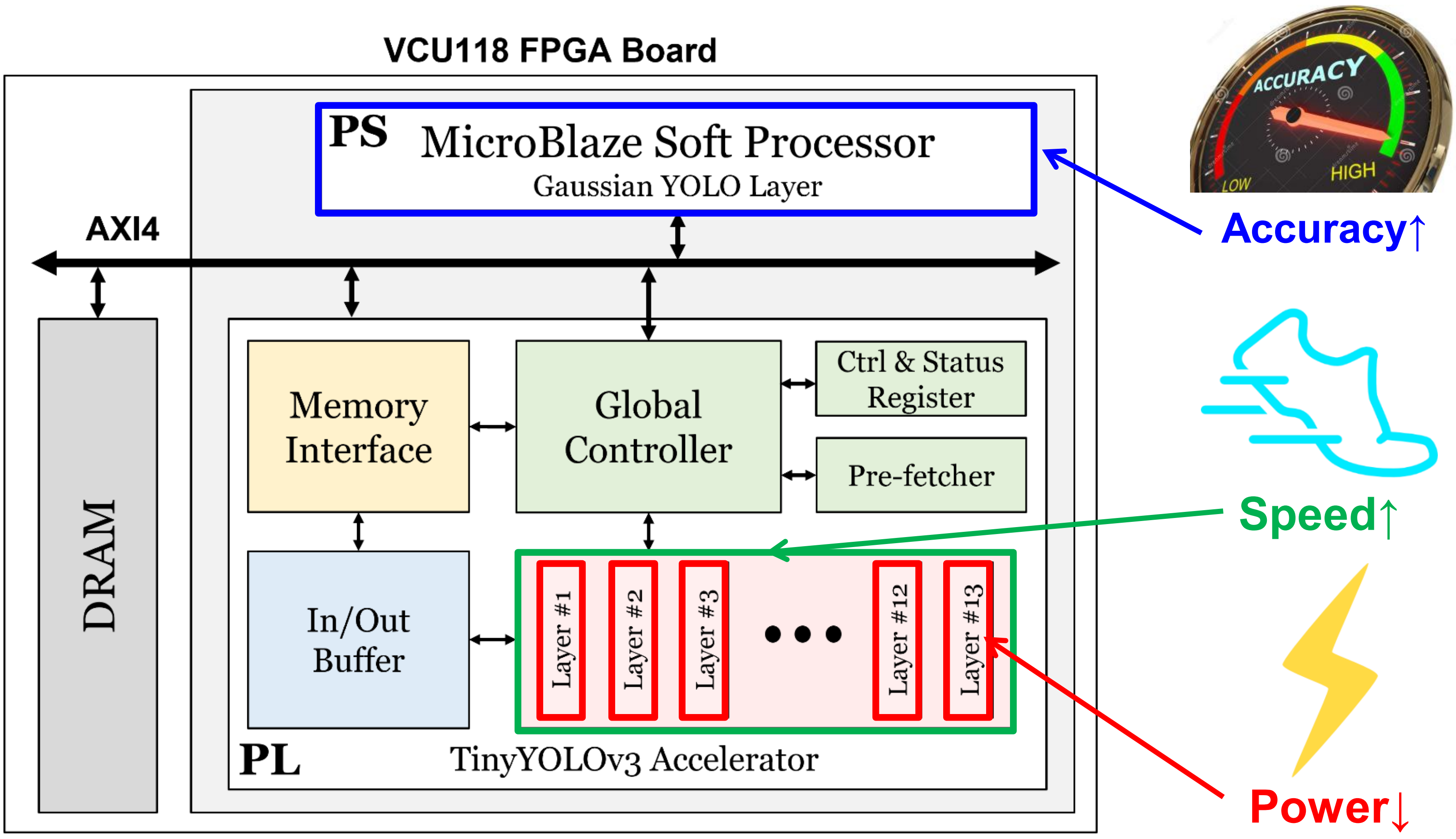
Implementation of a High-throughput and Accurate Gaussian-TinyYOLOv3 Hardware Accelerator

Juntae Park, Subin Ki, and Hyun Kim

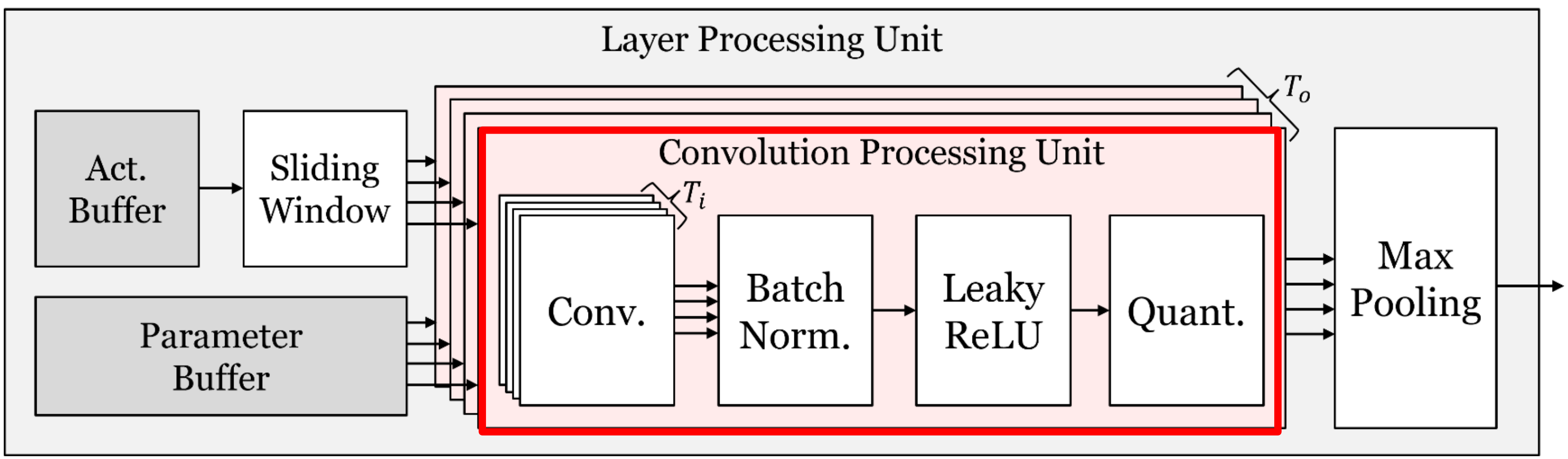
Seoul National University of Science and Technology

Proposed Architecture



Proposed Architecture

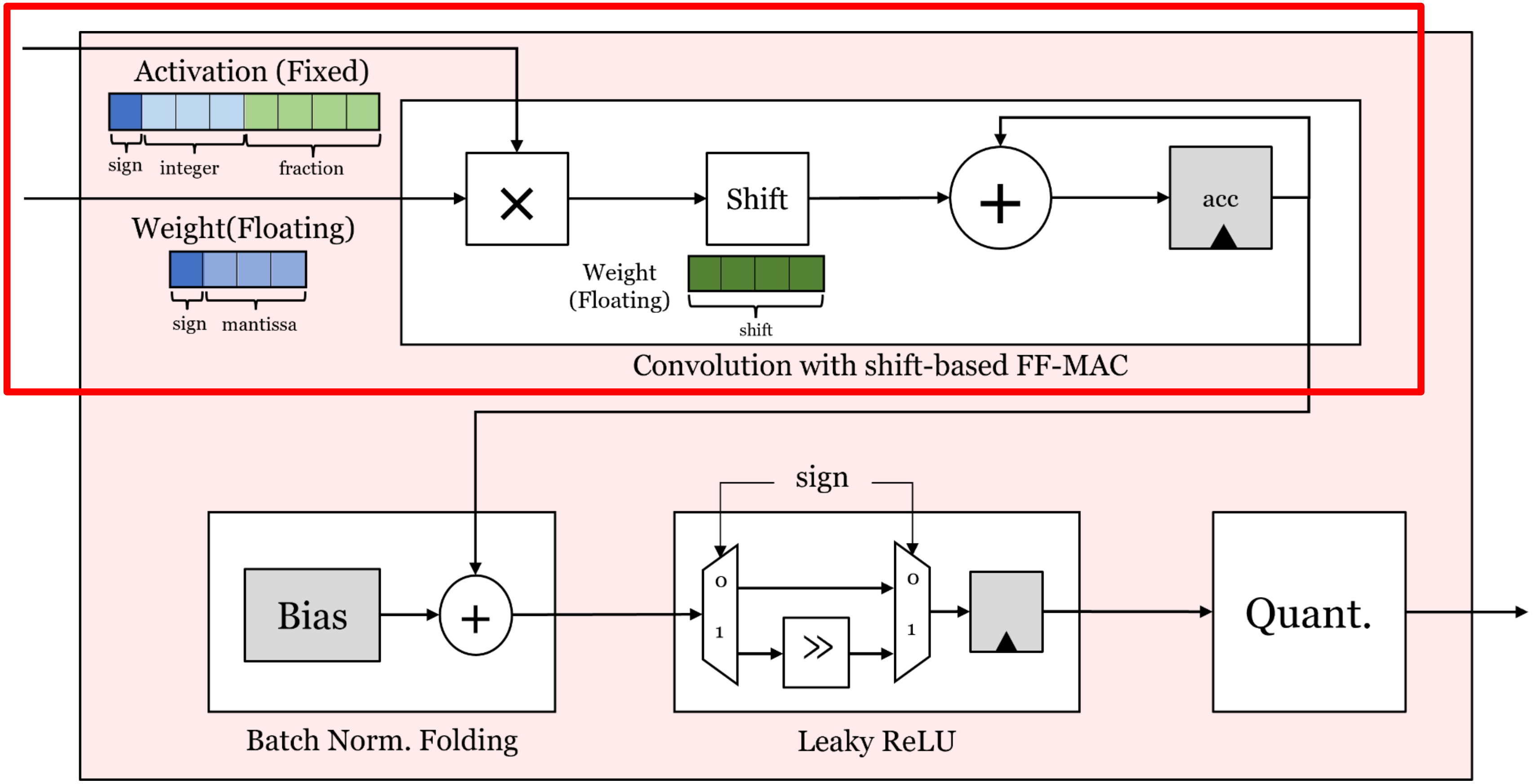
- **Layer processing unit**
  - It performs all the required operations for each layer.
  - It consists of the following modules, and they are imple mented **differently according to the layers**.



Proposed Architecture

Convolution processing unit

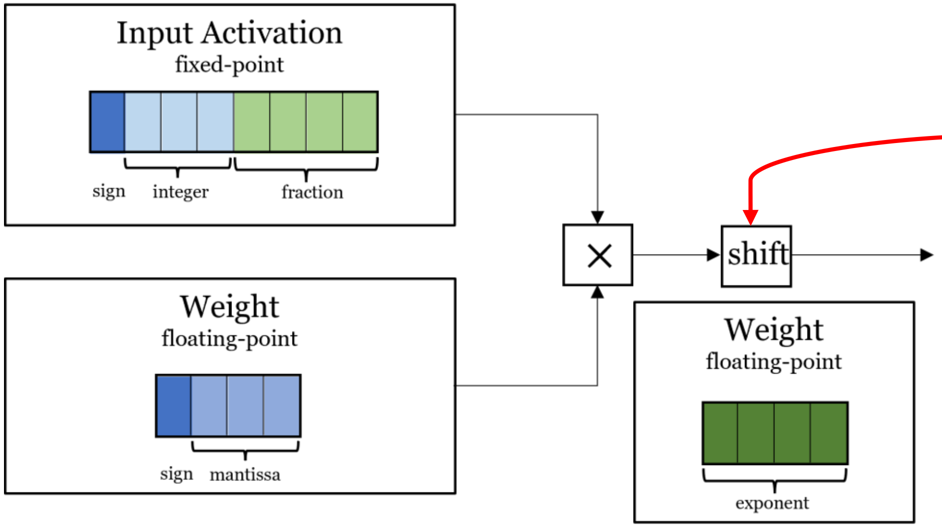
- It receives 8-bit fixed-point activation and 8-bit floating-point weight.



Proposed Architecture

◈ Hardware-friendly shift-based floating-fixed MAC and quantization

- We **unify the shift direction** for each layer by adjusting the **AS** value.
- **It only decreases the average accuracy by 0.05%.**



$$\boldsymbol{shift} = \boldsymbol{AS} - ((IS + M) - (E - B))$$

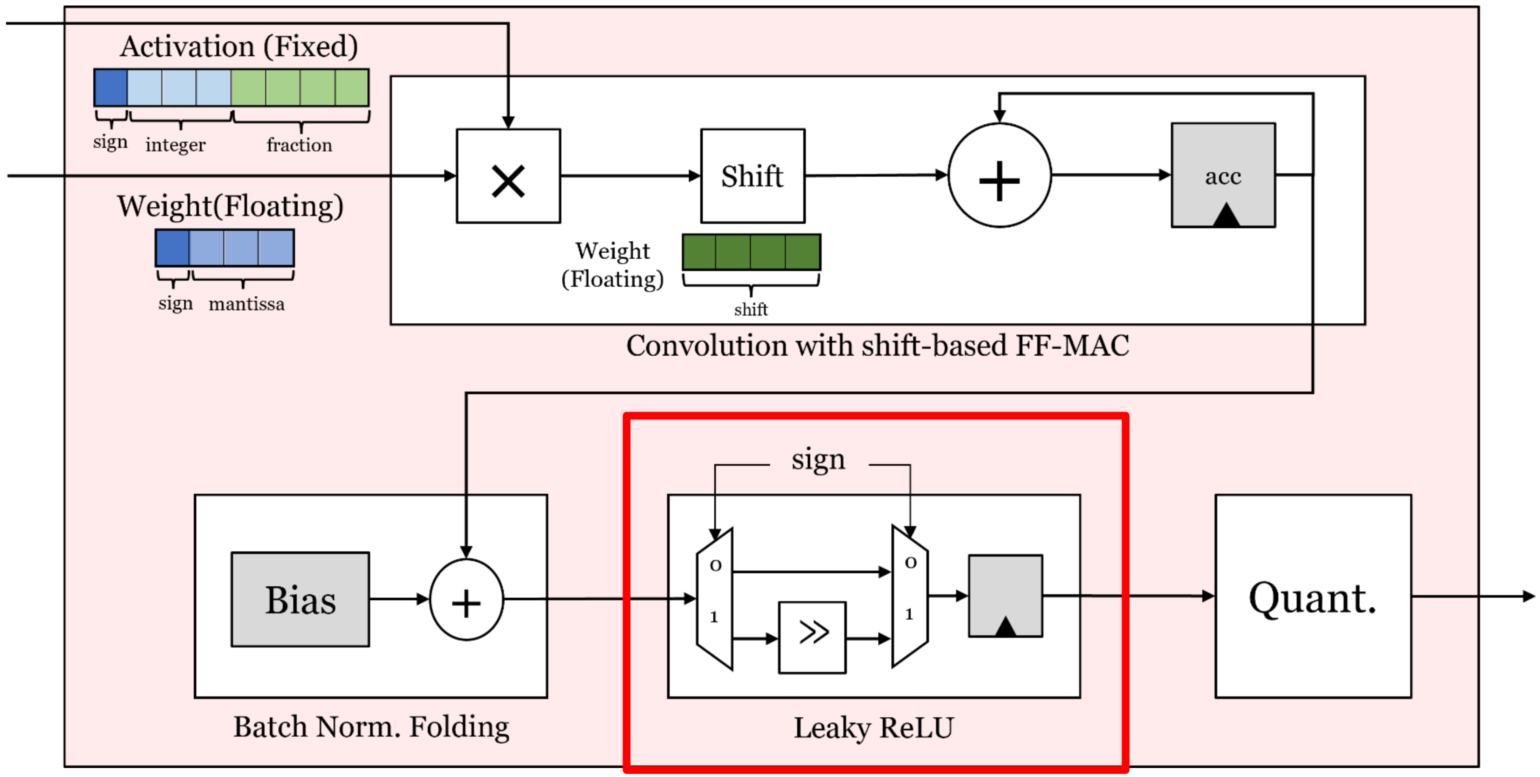
- **AS: Accumulate Scale**
- **IS: Input activation Scale**
- **M: # of mantissa bits**
- **E: Exponent of the weight**
- **B: Bias of the weight**

L. Lai, N. Suda, and V. Chandra, "Deep convolutional neural network inference with floating-point weights and fixed-point activations," arXiv preprint arXiv:1703.03073, 2017.

Proposed Architecture

Approximated shift-based Leaky ReLU

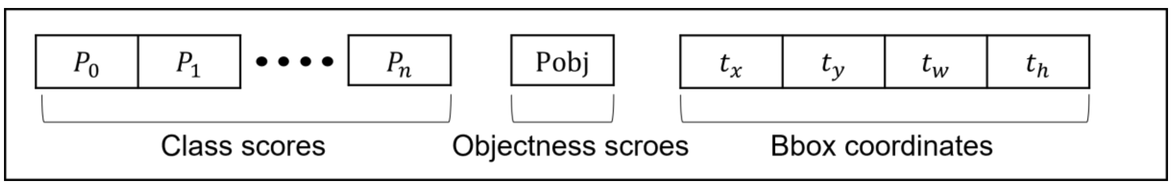
- Instead of the commonly used alpha value of **0.01**, we utilized an approximate value, **0.09375**, which is a combination of $(2^{-4}+2^{-5})$.



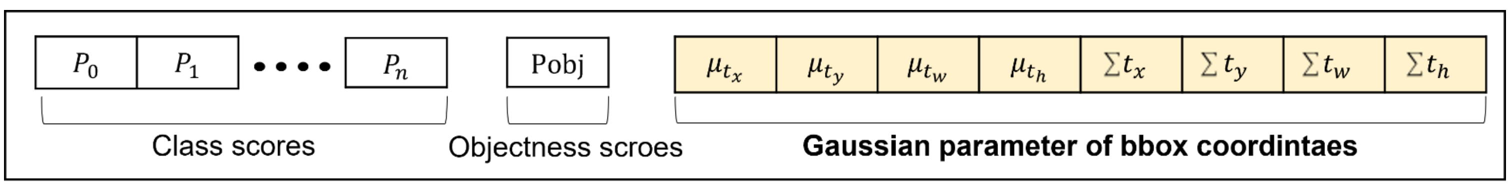
Proposed Architecture

◆ Gaussian YOLO layer

- It utilizes the uncertainty of each bbox coordinate to improve accuracy.
- It **increases** **accuracy** by **0.8% ~ 1.2%** with **negligible amount of increased FLOPs**



< Prediction box of original YOLO >



< Prediction box of Gaussian YOLO >

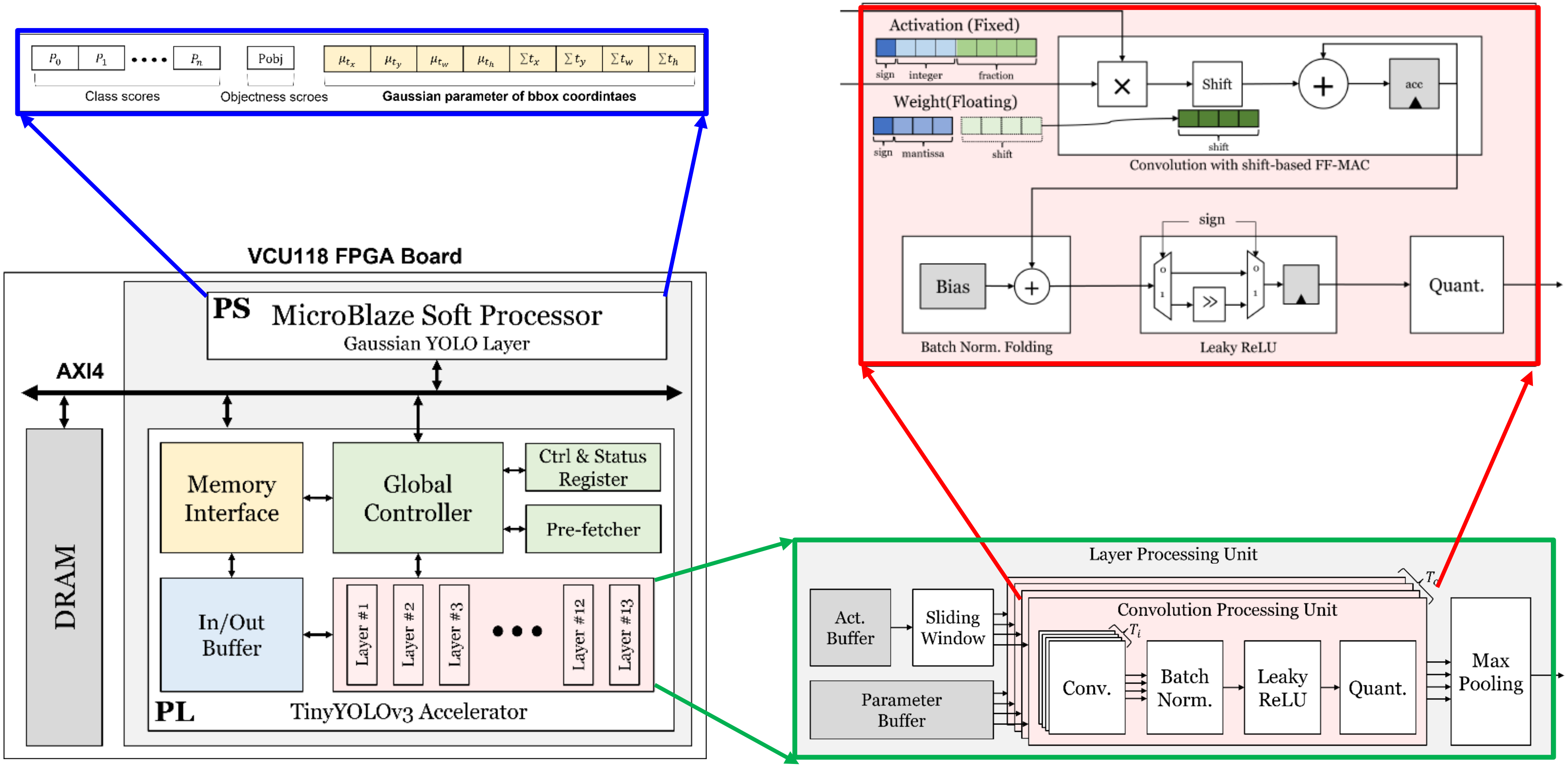
Choi, Jiwoong, et al. "Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving." *Proceedings of the IEEE/CVF International conference on computer vision*. 2019.

Proposed Architecture

◆ Design summary

Accuracy ↑

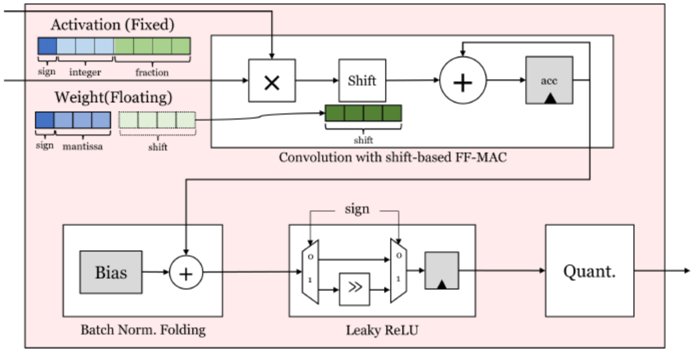
Power ↓



Speed↑

Experimental Results

◈ Comparison of **hardware resources** among **different MAC operators**



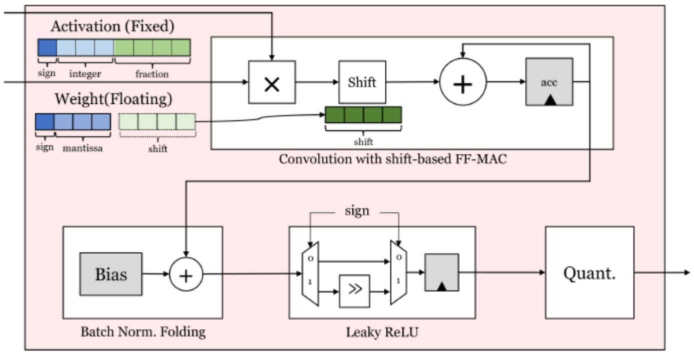
***[1]: 8b-Floating-Fixed MAC without unified shift direction**

Methods	LUT	FF	CARRY8	DSP
32b-Fixed-Fixed MAC	38670	5120	2270	4
8b-Fixed-Fixed MAC	21157	4905	2050	0
8b-Floating-Fixed MAC [1]	18746	4273	1912	0
8b-Proposed	**12381**	**3800**	**440**	**0**
	↓ 67.9%	**↓ 25.8%**	**↓ 80.6%**	

[1] L. Lai, N. Suda, and V. Chandra, "Deep convolutional neural network inference with floating-point weights and fixed-point activations," arXiv preprint arXiv:1703.03073, 2017.

Experimental Results

◈ Comparison of **hardware resources** among **different MAC operators**



***[1]: 8b-Floating-Fixed MAC without unified shift direction**

Methods	LUT	FF	CARRY8	DSP
32b-Fixed-Fixed MAC	38670	5120	2270	4
8b-Fixed-Fixed MAC	21157	4905	2050	0
8b-Floating-Fixed MAC [1]	18746	4273	1912	0
8b-Proposed	**12381**	**3800**	**440**	**0**
	↓ 33.9%	↓ 11.1%	↓ 76.9%	

[1] L. Lai, N. Suda, and V. Chandra, "Deep convolutional neural network inference with floating-point weights and fixed-point activations," arXiv preprint arXiv:1703.03073, 2017.

Experimental Results

◈ Evaluation of **overall** **accuracy**

***[1]: 8b-Floating-Fixed MAC without unified shift direction**

Dataset	**Model**	**mAP (%)** Baseline (32-b)	Quant. [1] (8-b)	**Proposed Quant. (8-b)**	Drop [1]-**Proposed**	
COCO 2014	TinyYOLOv3	33.1	32.81	32.79	0.02	
	Gaussian TinyYOLOv3	34.38	34.07	**34.01**	0.06	**↑ 0.91%**
VOC2007	TinyYOLOv3	68.54	68.22	68.18	0.04	
	Gaussian TinyYOLOv3	69.37	69.07	**69.02**	0.05	**↑ 0.48%**

[1] L. Lai, N. Suda, and V. Chandra, "Deep convolutional neural network inference with floating-point weights and fixed-point activations," arXiv preprint arXiv:1703.03073, 2017.

Experimental Results

◆ Performance Comparison with **previous works**

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	Proposed
Year	2020	2022	2020	2021	2021	2021	2022	2023
Model	TinyYOLOv2	TinyYOLOv2	TinyYOLOv3	TinyYOLOv3	TinyYOLOv3	TinyYOLOv3	TinyYOLOv3	TInyYOLOv3
Platform	Xilinx XCZU9EG	Xilinx XC7Z045	Xilinx XC7Z020	Xilinx XCKU040	Xilinx XC7Z020	Xilinx XCVU9P	Intel 10AX115	Xilinx XCVU9P
Freq.(MHz)	300	200	100	143	100	200	200	150
OCM(KB)	1,105	508.5	185	984	120	-	6,095	9,166
DSPs	609	448	160	839	208	2693	1122	96
LUTs(k)	95.1	99.4	25.9	139	33.4	17.7	146.1(Altera ALMs)	132
FFs(k)	90.6	98.9	46.7	-	-	145.7		39.5
Image Size	416×416	416×416	416×416	416×416	416×416	416×416	416×416	416×416
Precision	16b	16b	16b	16b	16b	-	32b	8b
Accuracy(%)	-	-	30.9	-	30.8	-	33.1	**34.01**
FPS	16.1	-	1.88	32.4	14.7	32.1	36.3	**62.9**
Throughput (GOPS)	102	138.8	10.45	180	-	166.4	202	**351.1**
Power(W)	-	-	3.36	3.87	-	-	-	5.52
Power Eff. (GOPS/W)	-	-	3.11	46.51	-	-	-	**63.61**

1.4x ~ 20.5x

S. Ki, J. Park and H. Kim, "Dedicated FPGA Implementation of the Gaussian TinyYOLOv3 Accelerator," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 10, pp. 3882-3886, Oct. 2023.

[1] S. Zhang, J. Cao, Q. Zhang, Q. Zhang, Y. Zhang, and Y. Wang, "An fpgabased reconfigurable cnn accelerator for yolo," in 2020 IEEE International Conference on Electronics Technology, pp. 74–78, 2020.

[2] H. Hongmin, L. Xueming, Q. Yadong, H. Xianghong, and X. Xiaoming, "An efficient parallel architecture for convolutional neural networks accelerator on fpgas," in 6th Int. Conf. High Performance Compilation, Computing and Communications, pp. 66–71, 2022.

[3] Z. Yu and C.-S. Bouganis, "A parameterisable fpga-tailored architecture for yolov3-tiny," in Proceedings of 16th International Symposium of Applied Reconfigurable Computing, pp. 330–344, 2020.

[4] D. Pestana, P. Miranda, J. Lopes, R. Duarte, M. Véstias, H. Neto, and J. De Sousa, "A full featured configurable accelerator for object detection with yolo," IEEE Access, vol. 9, pp. 75864–75877, 2021.

[5] P. Miranda, D. Pestana, J. Lopes, R. Duarte, M. Véstias, H. C. Neto, and J. T. de Sousa, "Configurable hardware core for iot object detection," Future Internet, vol. 13, no. 11, p. 280, 2021.

[6] M. Sharma, R. Rahul, S. Madhusudan, S. Deepu, and D. S. Sumam, "Hardware accelerator for object detection using tiny yolo-v3," in IEEE 18th India Council International Conference, pp. 1–6, 2021.

[7] V. Herrmann, J. Knapheide, F. Steinert, and B. Stabernack, "A yolo v3-tiny fpga architecture using a reconfigurable hardware accelerator for real-time region of interest detection," in 2022 25th Euromicro Conference on Digital System Design (DSD), pp. 84–92, IEEE, 2022.

Demonstration

Thank You !