Improving Neural Network Efficiency via Post-training Quantization with Adaptive Floating-Point

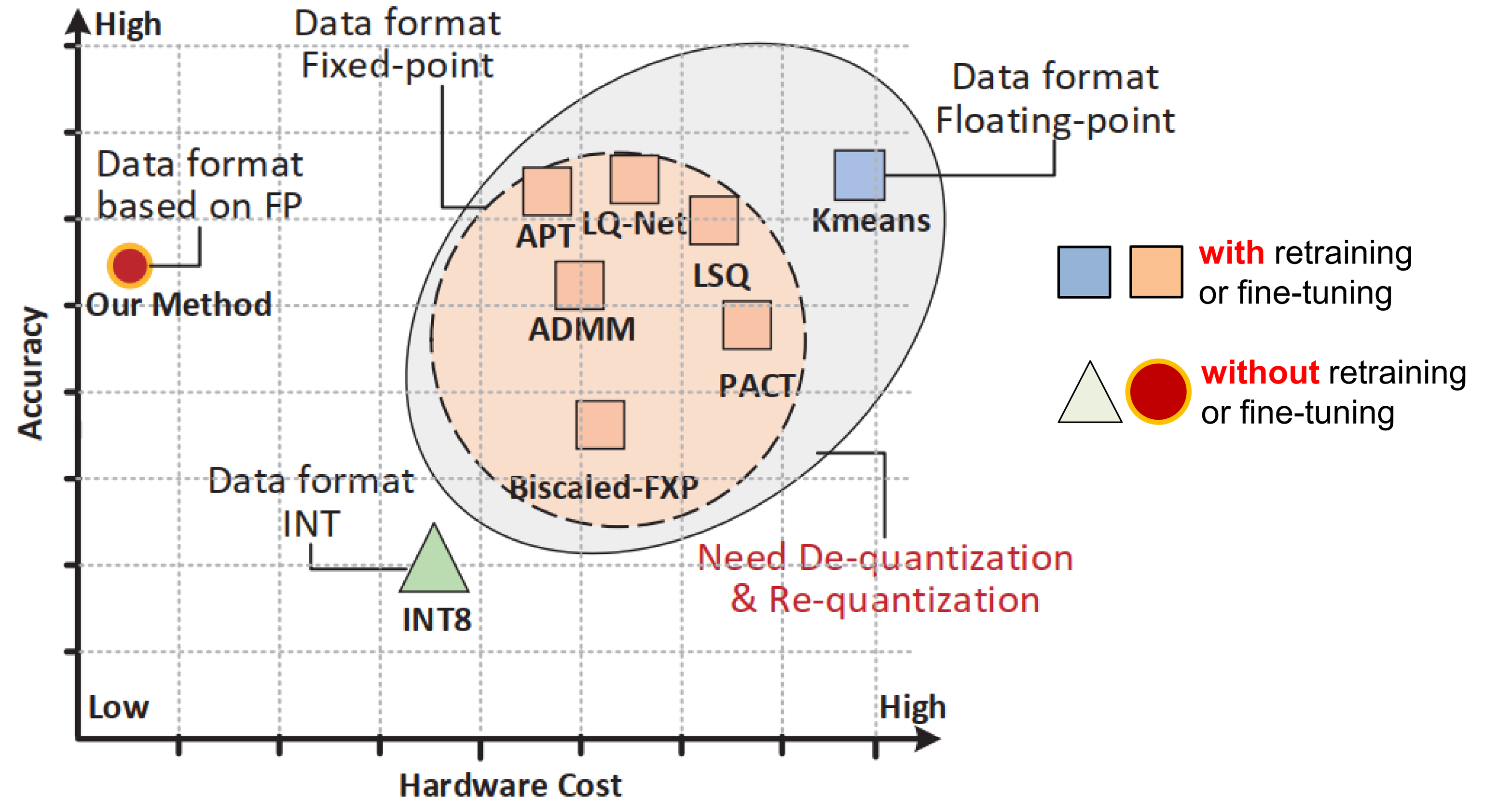
Fangxin Liu[1], Wenbo Zhao[1], Zhezhi He[1], Yanzhi Wang[2], Zongwu Wang[1], Changzhi Dai[3]
Xiaoyao Liang[1] and Li Jiang[1]
Shanghai Jiao Tong University, Northeastern University, DeepBlue Technology (Shanghai) Co., Ltd.

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY · Northeastern University · 先进计算机体系结构实验室 Advanced Computer Architecture Laboratory · DeepBlue Technology · 2021 ICCV OCTOBER 11-17 VIRTUAL

Motivation

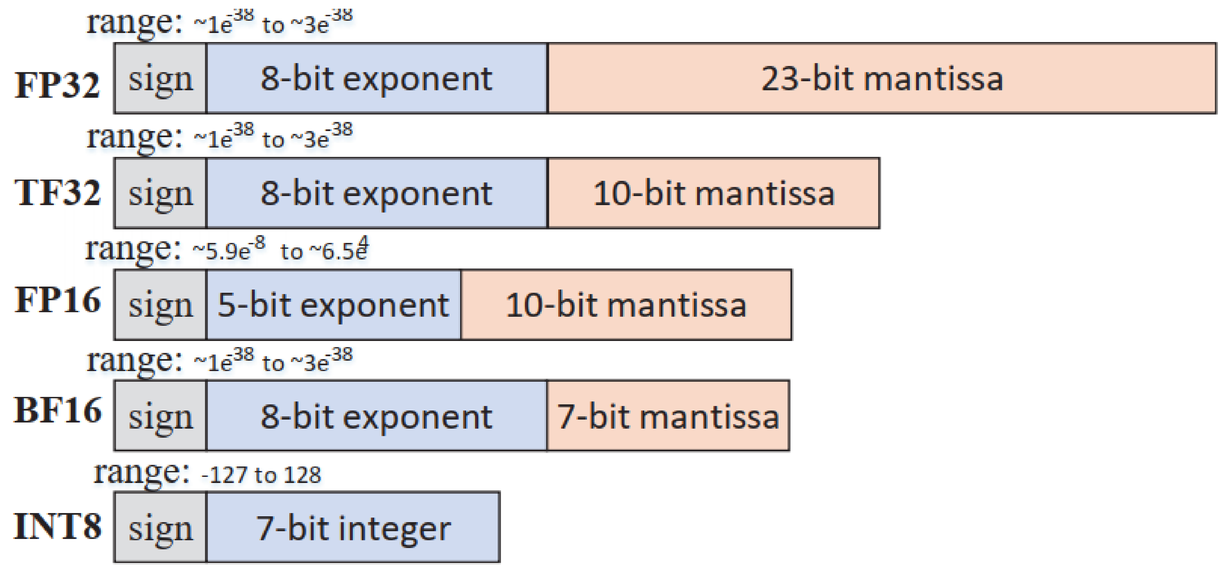
- Existing quantization methods can be generally separated into **non-uniform** methods, and **uniform** methods.
- Since many users are incapable of retraining DNN due to the **lack of computing-resource or retraining data**, quantization **without retraining** becomes the most popular compression method in many real-world scenarios.
- **Low latency** is critical for real-time interactions, while **low energy** consumption can help companies reduce cost in data-centers and improve the endurance of edge devices.

Hardware cost versus prediction accuracy



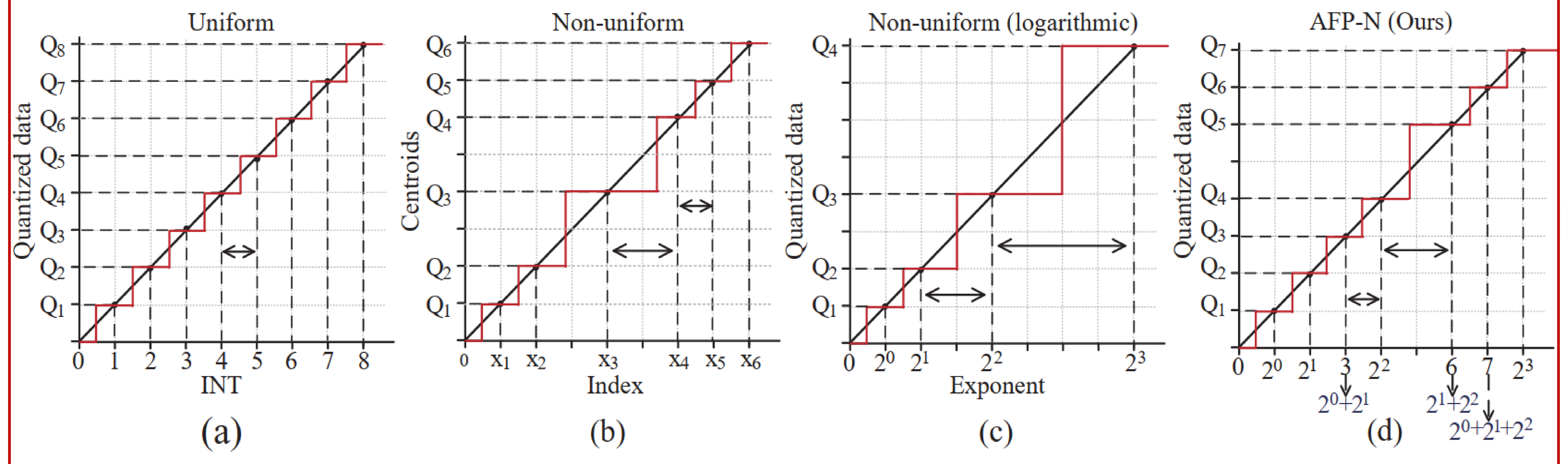
- Most **dynamic quantization** methods have to perform the dequantization and re-quantization process to rescale parameters with the aim of ensuring accuracy.
- As the trade-off of prior quantization methods in terms of data format precision and hardware efficiency, we develop a floating-point representation variant, named **Adaptive Floating-Point (AFP)**.

Comparison numeric format with INT8, FP32/16, BFP16 and TF32



Adaptive Floating-Point Quantization

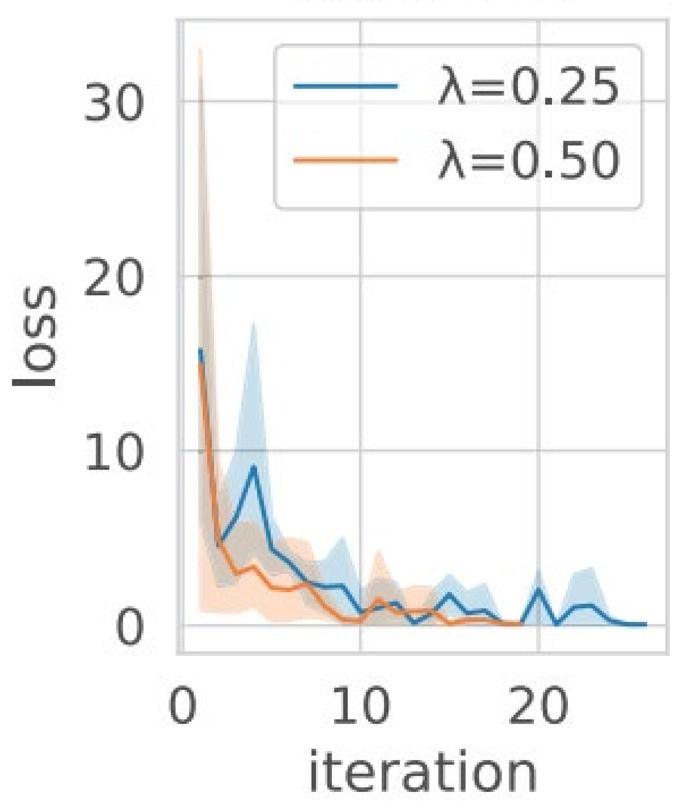
- AFP owns varying bitwidth for exponent and mantissa parts (n_{exp} and n_{man}), where the bit-width are chosen w.r.t the target application.
- In contrast to the fixed bias term adopted by the FP32 (i.e., k = 127), we make such a bias term a tunable as well.



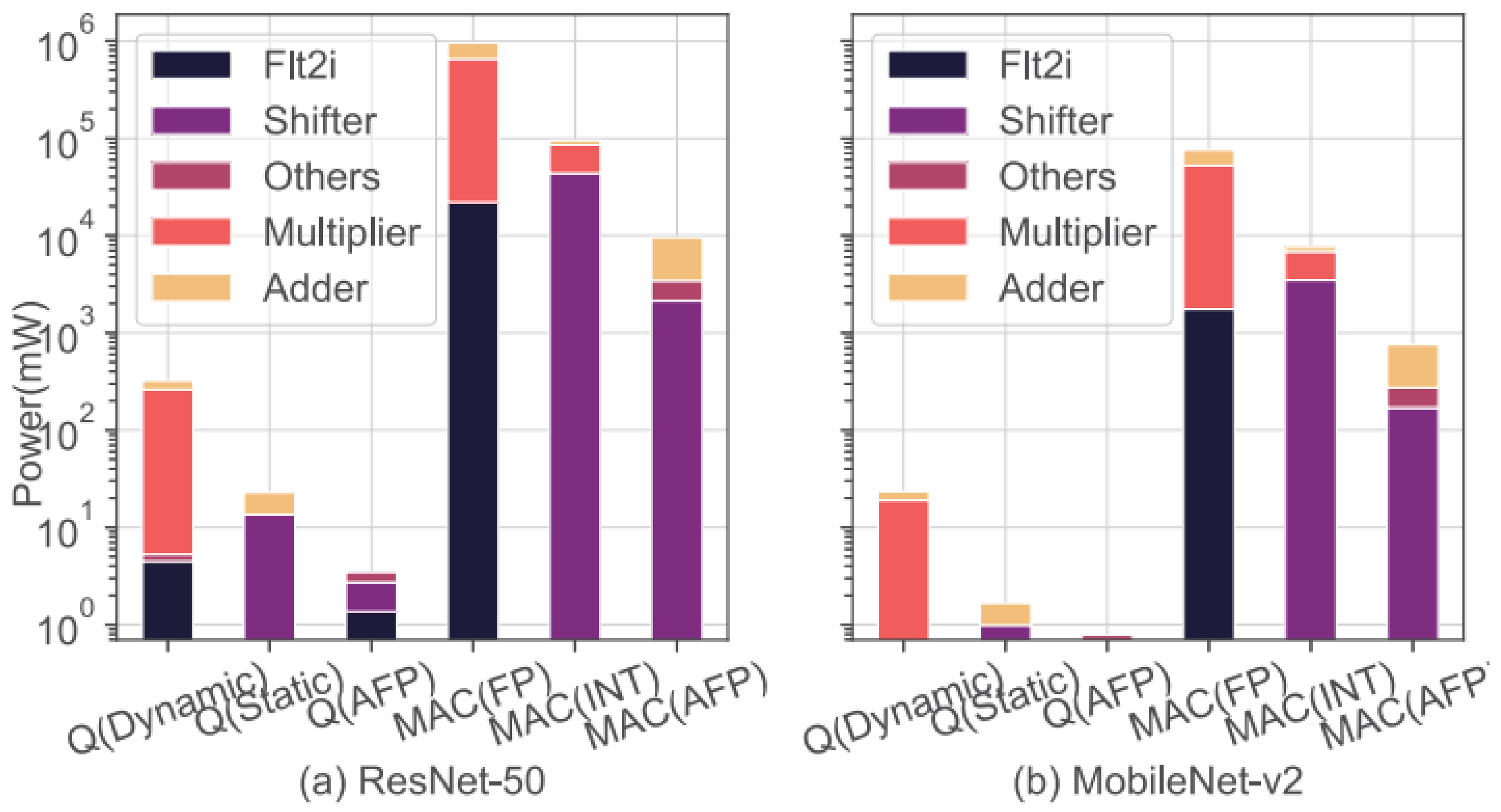
Layer-wise Quantization with AFP

- **Determine bias:** The bias is chosen to allow the maximum value of quantized weights and the maximum value of weights to be consistent, and the range of quantization can cover as much of the distribution of weights as possible.
- **Determine the bit-width of exponent:** The bit-width of exponent should be determined to enable that the range of exponent part can adequately cover the distribution of the weights.
- **Determine the bit-width of mantissa**: The mantissa is a component of a finite floating-point number, with the radix point immediately following the first digit.

Search time

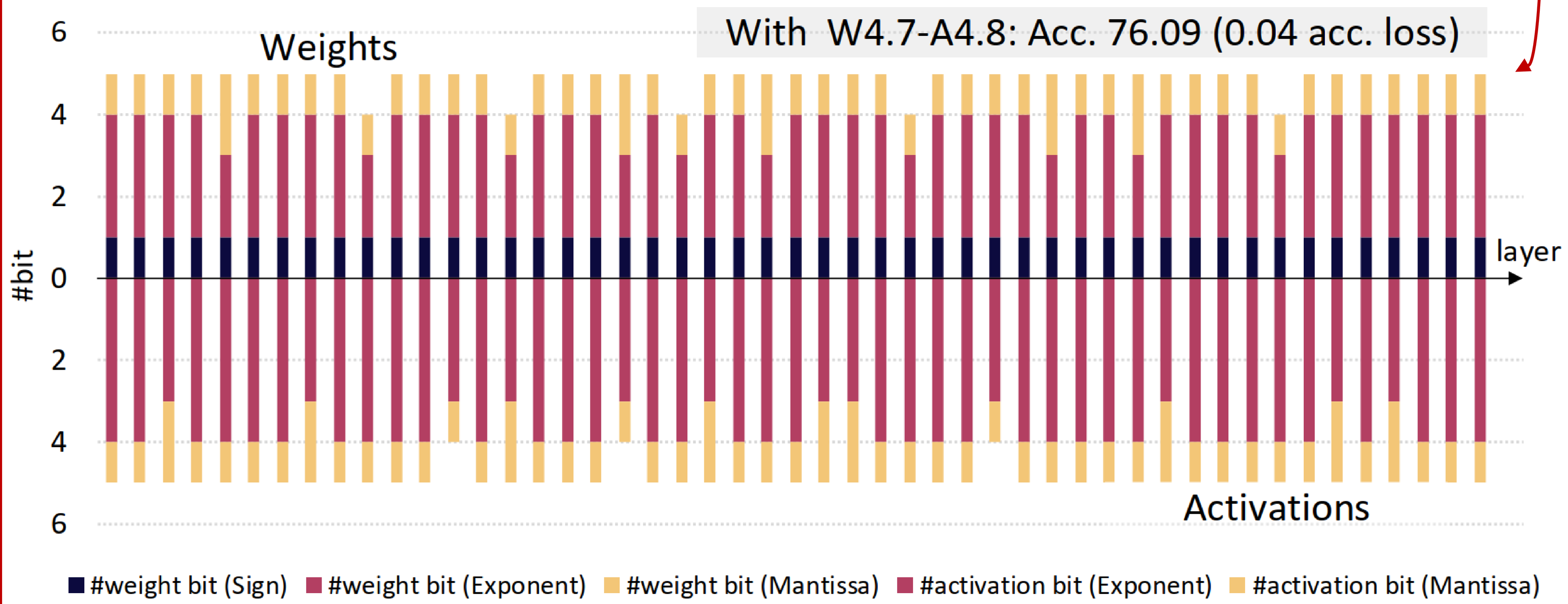


Hardware efficiency



Results on ImageNet

Quan. scheme	Bit width	First layer	Last layer	Acc. Top-1(%)	Acc. loss Top-1(%)	Quan. type	No retrain	Data format
ResNet-50								
Full precision	32	32	32	76.13	-	-	-	-
INT8 [12]	8	8	8	74.9	-1.5	Uniform	✓	INT
V-Q [23]	7	7	7	75.89	-0.27	Uniform	×	FP
Biscaled-FxP [13]	6	6	6	70.46	-5.67	Non-uni.	✓	INT
ADMM [31]	6	6	6	75.93	-0.2	Non-uni.	×	FP
INQ [35]	5	32	32	74.81	-1.59	Non-uni.	×	FP
Focused-C. [34]	5	5	5	75.86	-1.54	Non-uni.	×	FP
APT [17]	4	32	32	75.95	-0.18	Non-uni.	×	FP
UNIQ [2]	4	4	4	74.84	-1.29	Non-uni.	×	FP
this work(dynamic)	**4.8**	**5**	**5**	**76.09**	**-0.04**	**Non-uni.**	✓	**FP**
this work(dynamic)	**3.9**	**4**	**4**	**75.27**	**-0.86**	**Non-uni.**	✓	**FP**
this work (static)	**4.8**	**5**	**5**	**76.00**	**-0.13**	**Non-uni.**	✓	**FP**
this work (static)	**3.9**	**4**	**4**	**75.11**	**-1.02**	**Non-uni.**	✓	**FP**



Hardware cost with the sweet-spot