



SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models

Ziyi Guan^{1,2^}, Hantao Huang¹, Yupeng Su¹, Hong Huang¹, Ngai Wong² and Hao Yu¹

¹School of Microelectronics, Southern University of Science and Technology, Shenzhen

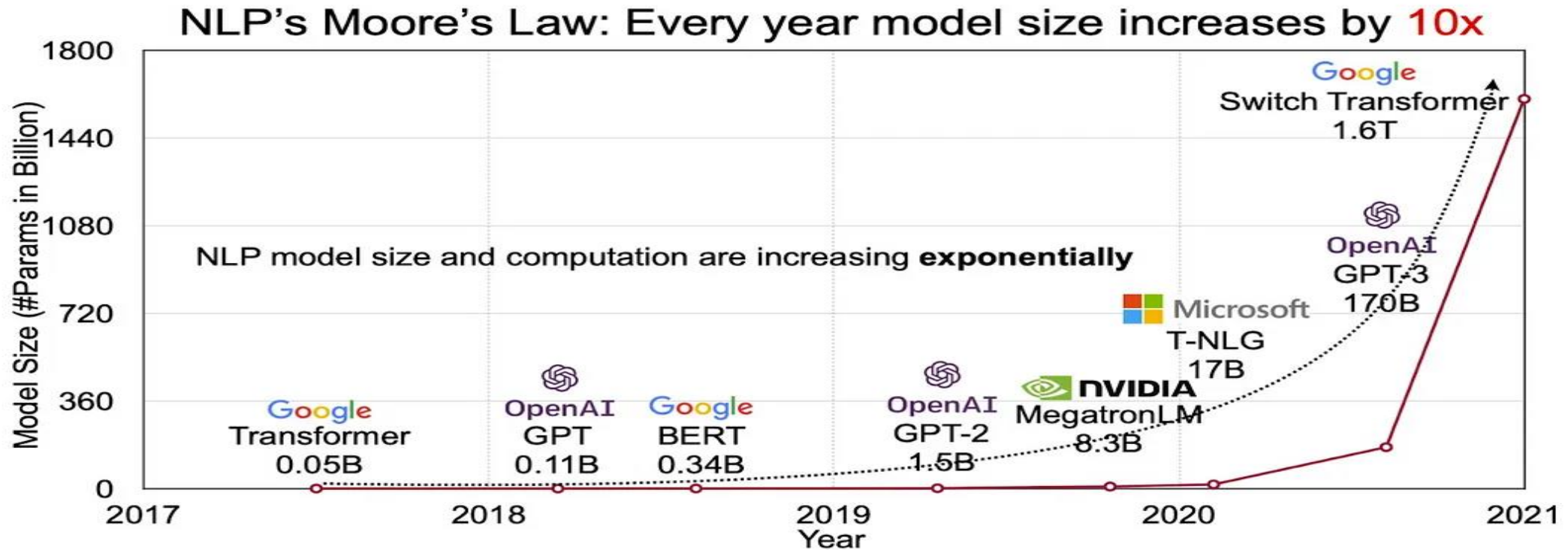
²Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

^Presenter

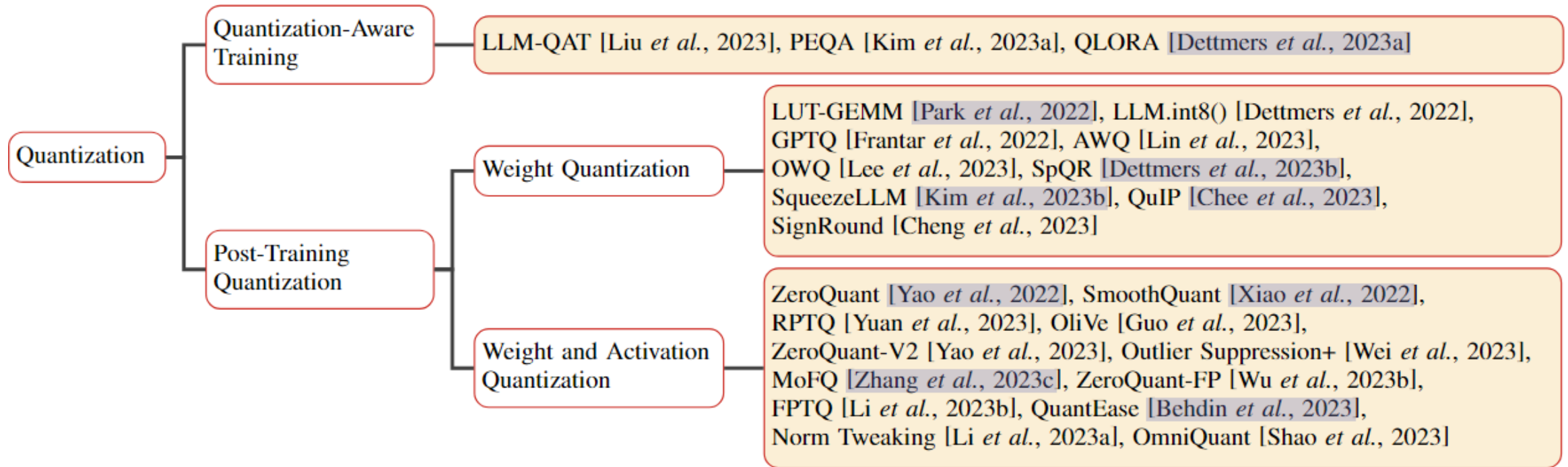


Surging Sizes in NLP Models

- Rapid Rise: NLP model sizes increase 10x annually
- This surge presents profound challenges for deployment on resource-constrained devices.



Related Works



Quantization-Aware Training (QAT): Quantization is integrated into the model's training process.

Post-Training Quantization (PTQ): Quantizing the parameters of a LLM after the training phase.

Related Works: OBQ GPTQ

Guide by Optimal Brain Quantization(OBQ):

Taylor Series of the error function(or loss function):

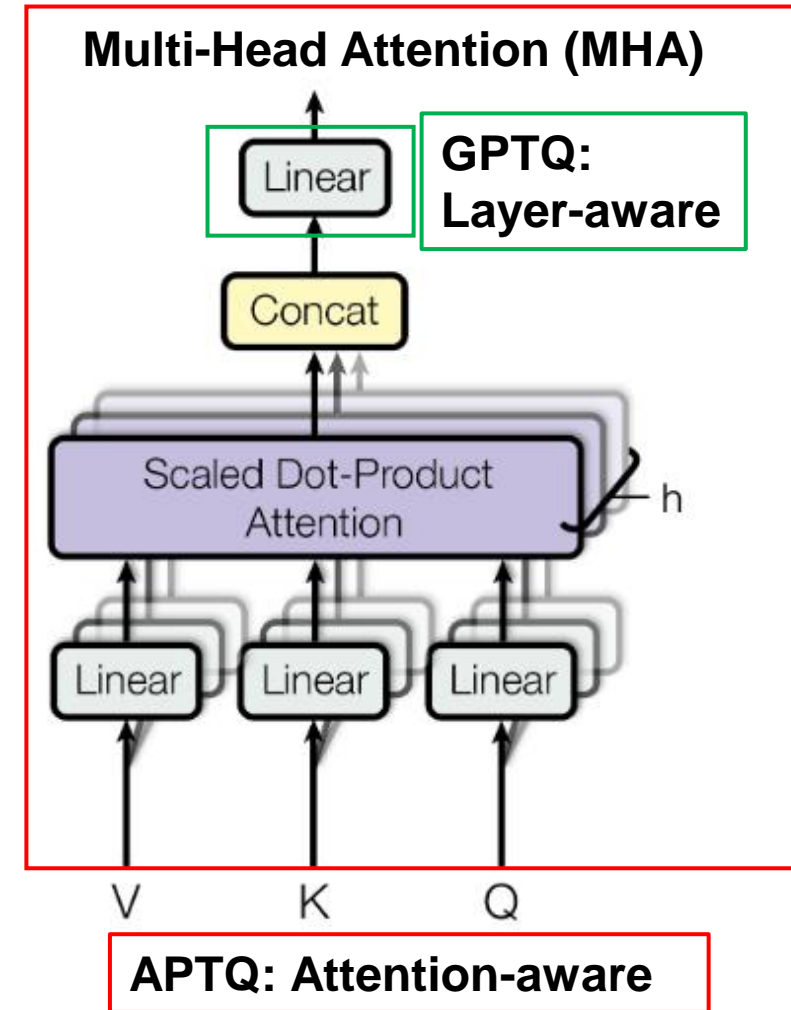
$$\delta E = \left(\frac{\partial E}{\partial \mathbf{w}} \right)^T \cdot \delta \mathbf{w} + \frac{1}{2} \delta \mathbf{w}^T \cdot \mathbf{H} \cdot \delta \mathbf{w} + O(\| \delta \mathbf{w} \|^3) \quad \text{Ignored}$$

GPTQ: Layer-Wise Layer-aware Quantization.

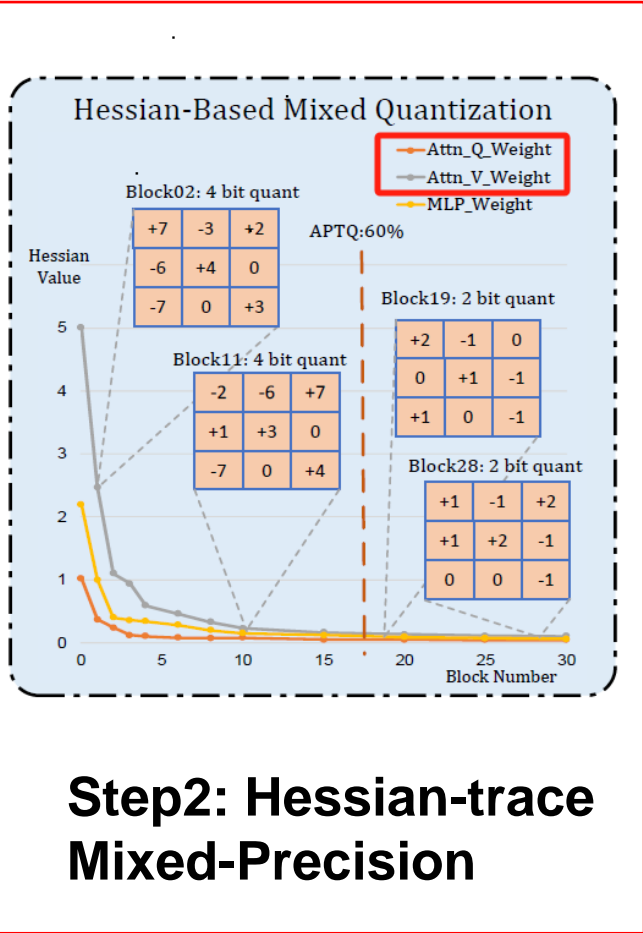
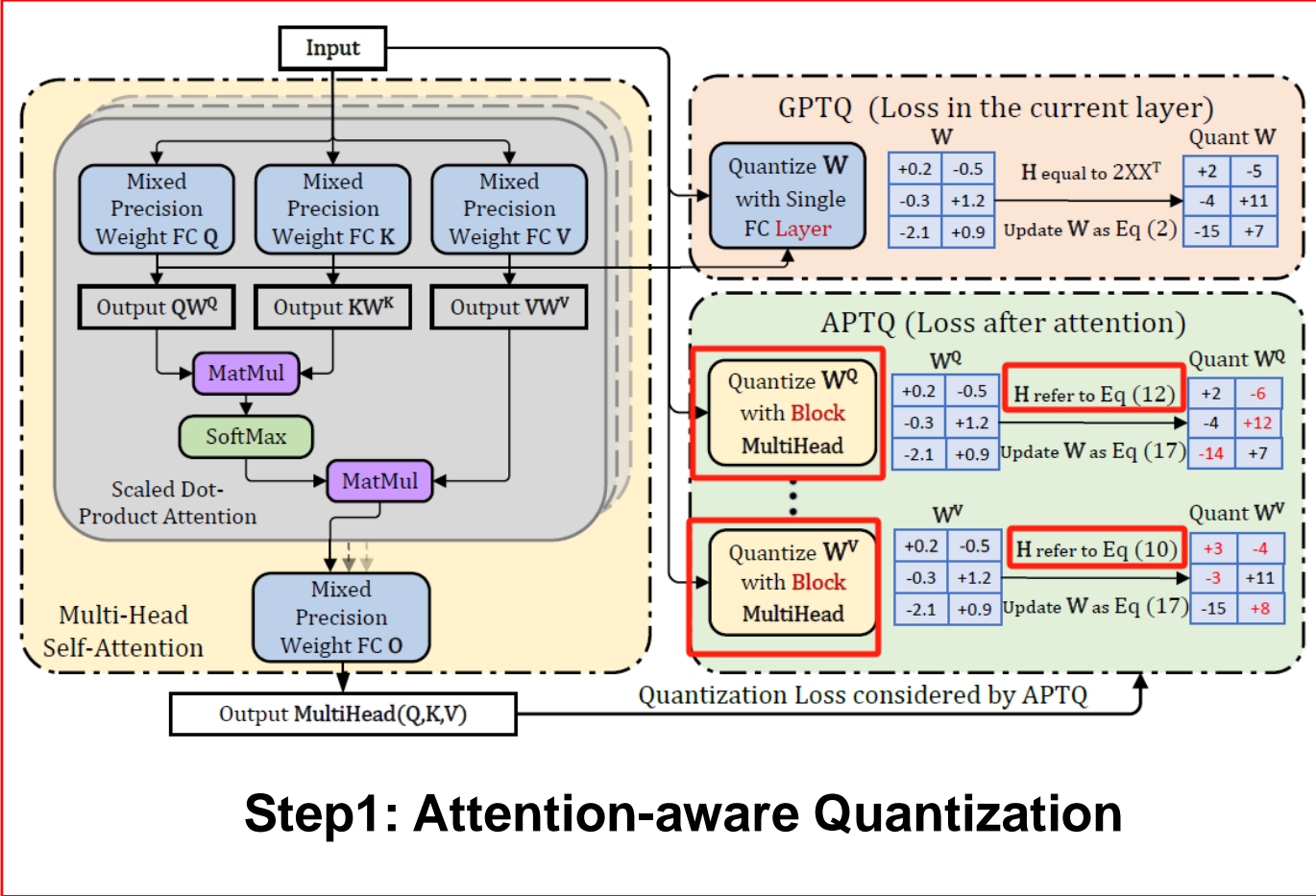
$$E = \| \mathbf{W}\mathbf{X} - \widehat{\mathbf{W}}\mathbf{X} \|_2^2 \quad \longrightarrow \quad \text{Local Optima}$$

APTQ: Layer-Wise **Attention-aware** Quantization.

$$E = \| \text{MHA}(\mathbf{W}, \mathbf{X}) - \text{MHA}(\widehat{\mathbf{W}}, \mathbf{X}) \|_2^2 \quad \longrightarrow \quad \text{Global Optima}$$



APTQ Overview



Hessian-Attention-based Quantization

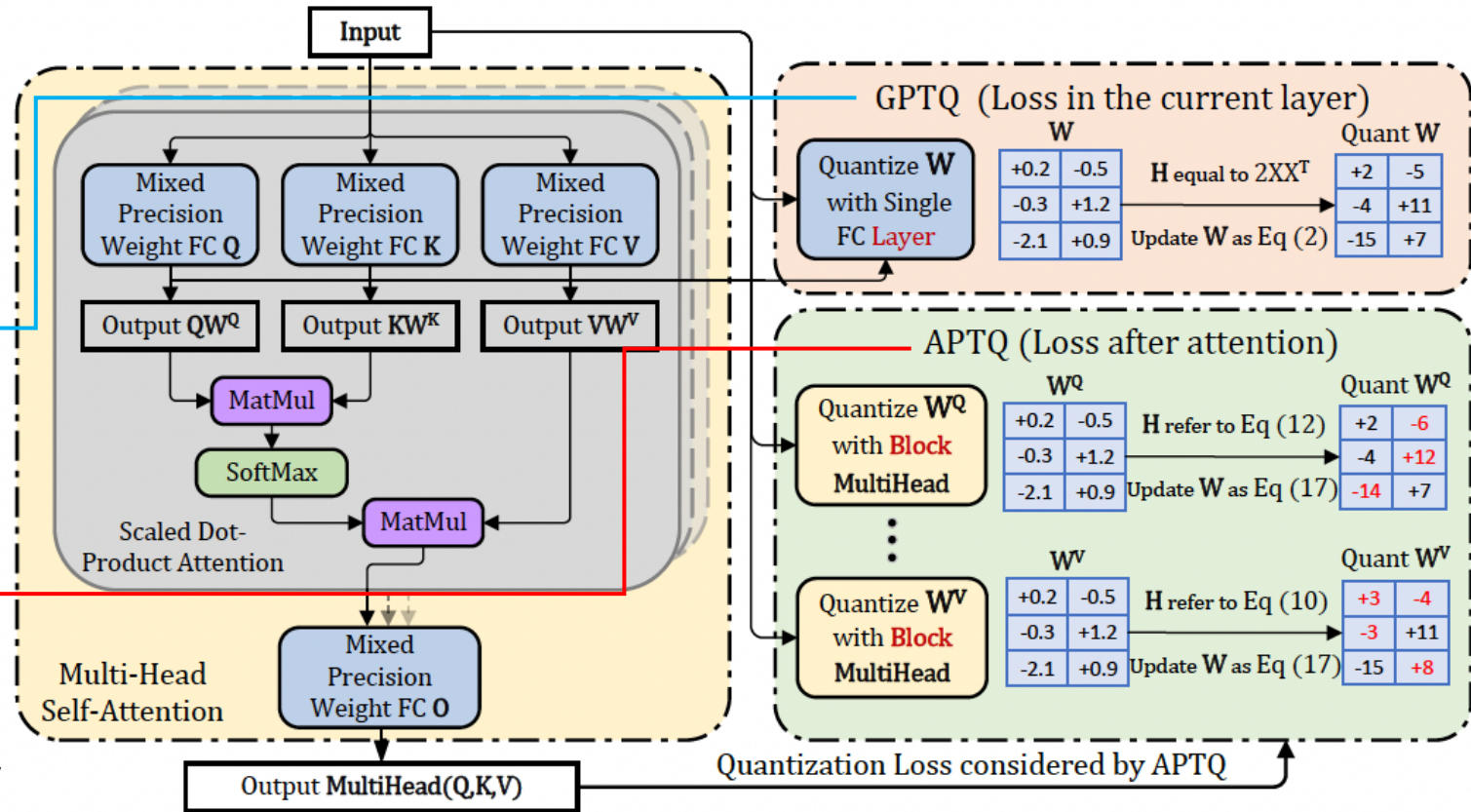
GPTQ Weight updating formula:

$$H_F = 2X_F X_F^T$$

APTQ Weight updating formula:

$$H_{\hat{w}} = 2 \cdot [F'(\hat{w}) \cdot F'(\hat{w})^T]$$

The weights are updated according to the gradient of the whole Attention layer

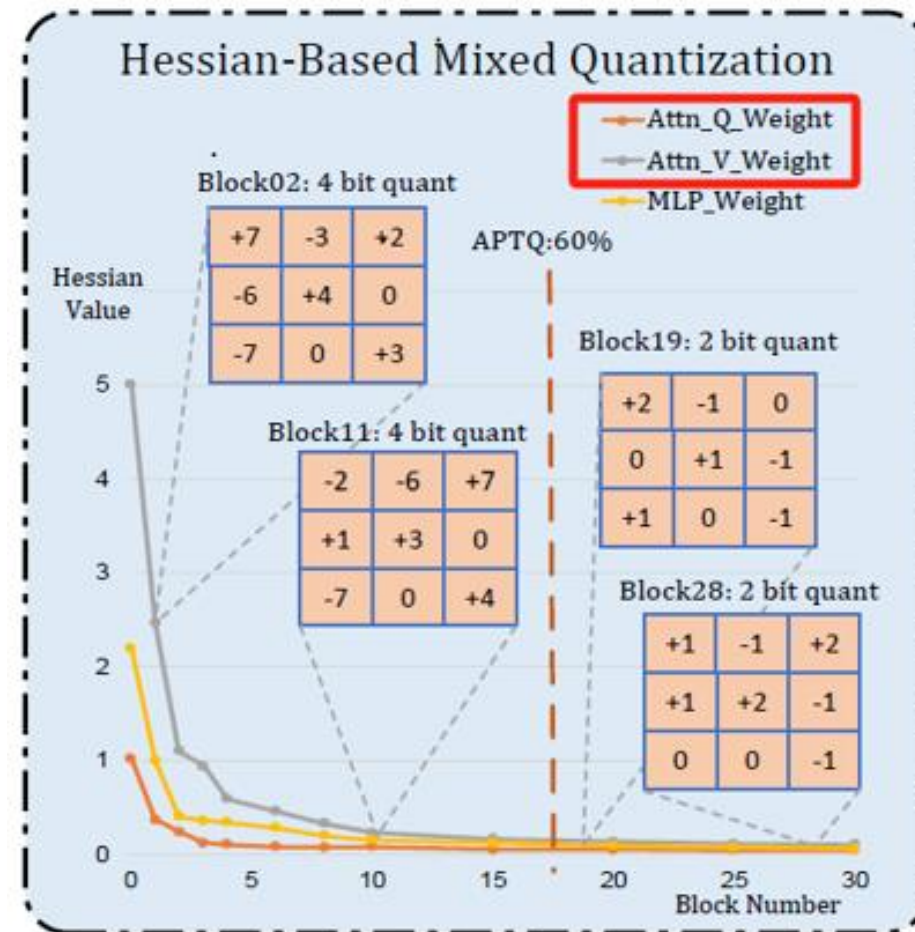


Hessian-Trace-based Mixed-Precision Quantization

Mixed-precision metric:
average Hessian trace values

$$\text{Average bits} = 4 \times R + 2 \times (1 - R)$$

where R denotes the proportion of weights quantized at 4 bits
(In this figure, $R = 0.6$)



APTQ Algorithm

Input: Pre-trained model weights W , blocksize B , Hessian matrix H , quantization function quant , Layer names layerName , Ratio of 4-bit in 2/4 mixed-precision R .

- 1: Initialize quantized weight matrix $Q \leftarrow 0_{d_{\text{row}} \times d_{\text{col}}}$.
 - 2: Initialize block quantization error matrix $E \leftarrow 0_{d_{\text{row}} \times B}$.
 - 3: **Step 1: 4-bit Hessian-Attention-Based Quantization**
 - 4: **for** $i = 0, B, 2B, \dots$ **do**
 - 5: **for** $j = i, \dots, i + B - 1$ **do**
 - 6: **if** "self_attn.k_proj" in layerName **then**
 - 7: $H_{\hat{W}}^K = 2 \left[\frac{\partial F}{\partial W^K} \cdot \frac{\partial F}{\partial W^K}^T \right]$ from Equation (13)
 - 8: $Q_{:,j}^K \leftarrow \text{quant}(W_{:,j}^K)$
 - 9: $E_{:,j-i}^K \leftarrow (W_{:,j}^K - Q_{:,j}^K) / [H_{\hat{W}}^{-1}]_{jj}^K$ based on Equation (16)
 - 10: $W_{:,j:(i+B)}^K \leftarrow W_{:,j:(i+B)}^K - E_{:,j-i}^K \cdot (H_{\hat{W}}^{-1})_{:,j:(i+B)}^K$ based on Equation (17)
 - 11: For self_attn.Q, V, and O projection layers, similar updates are applied
 - 12: Compute the average Hessian trace for each layer in block $i : (i + B)$.
 - 13: **end if**
 - 14: **end for**
 - 15: **end for**
 - 16: **Step 2: Hessian-trace-based Mixed-Precision Quantization**
 - 17: Calculate Hessian trace values for each layer, and order them from highest to lowest, starting with the previously established 4-bit quantization.
 - 18: Determine the layers for mixed-precision quantization based on the computed Hessian trace values and R .
 - 19: **for** each selected layer **do**
 - 20: Calibrate the bit allocation in line with each layer's Hessian trace sensitivity and R .
 - 21: Implement 2/4 bit mixed-precision quantization
 - 22: **end for**
- Output:** The resulting quantized model weights Q are characterized by scale, zero-point, and quantization error.

Experiment

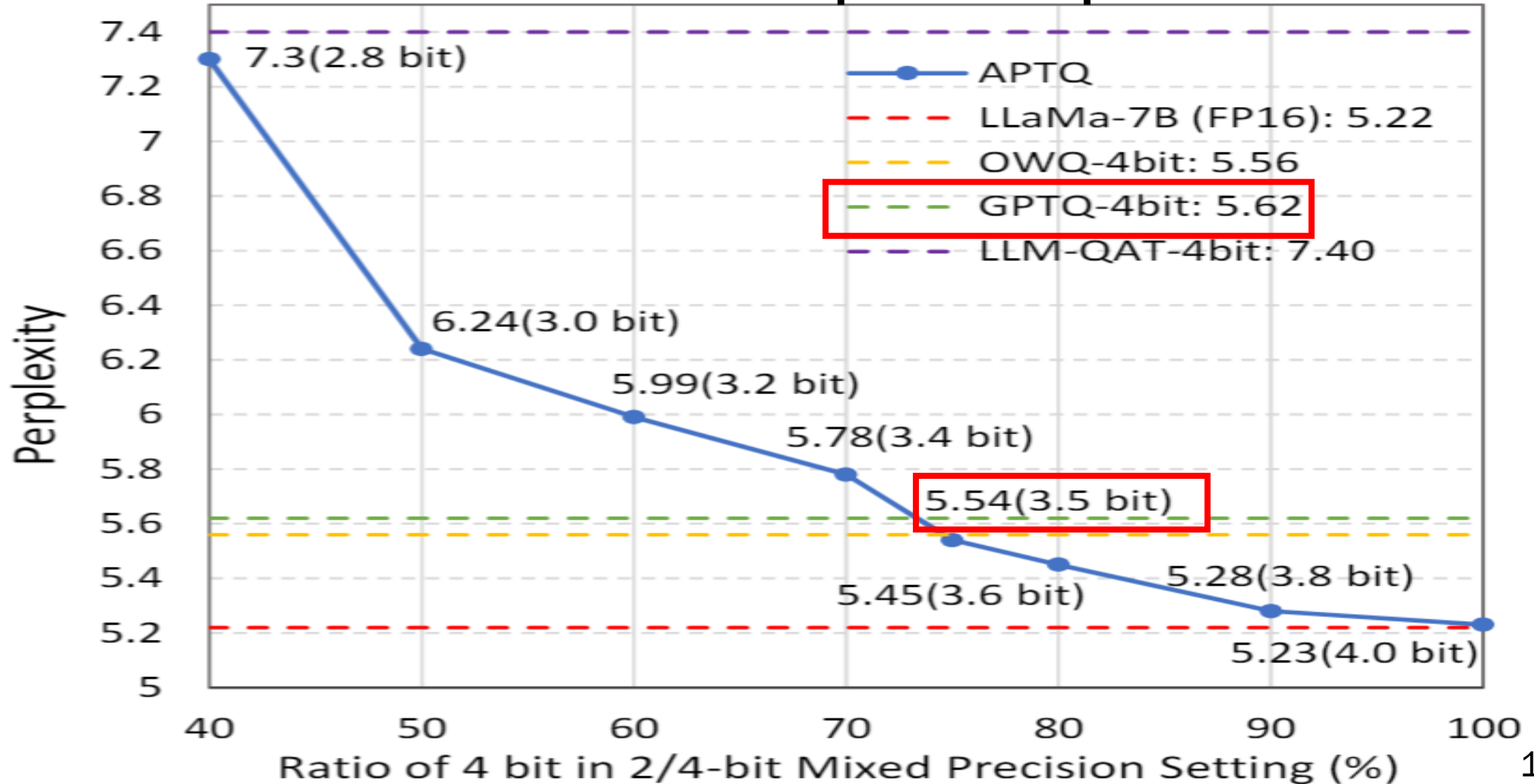
Better PPL under low-bit weight-only quantization

Table 1: Comparison of Perplexity of Quantized LLaMa Models on C4 and WikiText-2 Datasets.

Method	Avg bit	C4	Wikitext-2
LLaMa-7B	16	5.22	5.68
GPTQ	4.0	5.62	8.14
OWQ	4.01	5.56	7.15
LLM-QAT	4.0	7.40	10.90
PB-LLM-20%	3.4	20.61	17.19
APTQ	4.0	5.23	6.45
APTQ-75%	3.5	5.54	6.54
APTQ-50%	3.0	6.24	6.76

Experiment

Better PPL under mixed-precision quantization



Experiment

Better zero-shot accuracy under mixed-precision quantization

Table 2: Zero-shot accuracy of quantized LLaMa models on common sense reasoning tasks.

Model	LLaMa-7B							LLaMa-13B						
Method	Avg bit	PIQA	Hellaswag	Arc-E	Arc-C	WinoGrande	$\overline{Acc\%} \uparrow$	PIQA	Hellaswag	Arc-E	Arc-C	WinoGrande	$\overline{Acc\%} \uparrow$	
FP16	16	79.2	76.2	72.8	44.7	69.9	68.56	80.3	79.0	74.8	47.9	72.7	70.94	
RTN [12]	4.0	77.3	72.7	68.8	43.1	66.9	65.76	79.1	76.8	72.6	46.5	70.5	69.10	
SmoothQuant [18]	4.0	76.4	68.1	67.3	39.6	66.0	63.48	77.9	74.2	76.3	45.5	69.7	68.72	
FPQ [11]	4.0	77.8	75.0	72.4	41.7	69.0	66.60	79.4	77.7	72.8	47.3	71.5	69.74	
LLM-QAT [12]	4.0	78.3	74.0	70.0	41.7	69.0	66.60	79.4	77.7	72.8	47.3	71.5	69.74	
GPTQ [6]	4.0	76.0	69.4	66.9	43.0	66.7	64.40	79.8	77.7	73.2	45.9	72.6	69.84	
PB-LLM 30% [16]	4.1	78	74.3	69.0	42.3	69.7	66.66	-	-	-	-	-	-	
PB-LLM 10% [16]	2.7	67.8	68.1	58.7	39.6	67.4	60.32	-	-	-	-	-	-	
APTQ	4.0	78.6	75.7	72.4	44.4	69.3	68.08	79.9	78.8	73.9	47.0	72.1	70.34	
APTQ-90%	3.8	78.8	75.9	73.6	43.5	69.4	68.24	79.4	78.8	73.8	47.8	72.6	70.48	
APTQ-80%	3.6	78.0	75.3	70.2	43.7	69.5	67.34	79.5	78.2	72.8	46.5	72.6	69.92	
APTQ-75%	3.5	77.5	74.5	68.7	44.2	70.2	67.02	79.3	77.6	71.8	46.1	73.2	69.60	
APTQ-70%	3.4	77.6	73.4	66.9	41.3	68.9	65.62	78.3	77.5	71.4	46.3	72.5	69.20	
APTQ-60%	3.2	76.8	72.1	63.1	39.3	69.5	64.16	78.6	74.2	69.5	44.2	69.5	67.20	
APTQ-50%	3.0	74.5	68.3	57.9	36.4	65.3	60.48	74.4	71.2	64.1	41.0	68.0	63.74	

Conclusion

- APTQ integrates **attention-based gradients** with Hessian optimization, significantly enhancing quantization precision.
- APTQ uses a **novel Hessian trace-driven mixed-precision** scheme to optimize performance by adjusting bitwidths based on layer sensitivity.
- APTQ achieves near **full-precision** results at 4-bit quantization and demonstrates **state-of-the-art (SOTA)** zero-shot performance compared to other methods in experiments on LLaMa models.



**THE CHIPS
TO SYSTEMS
CONFERENCE**

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

Thank you for your attention

