DENSO
Crafting the Core

DENSO IT LAB

CVPR JUNE 3-7, 2026 DENVER COLORADO

Rethinking Asymmetric Quantization: Hidden Symmetry in Vision Model Weights

Masafumi Mori[1], Shinya Gongyo[2], Mitsuru Ambai[2]

[1]DENSO CORPORATION, [2]DENSO IT Laboratory

CVPR Poster: Day3 (Morning)

Background: PTQ for Efficient Vision Inference

- Many applications require both high accuracy and real-time inference on edge devices

AD/ADAS

Factory Automation

Smart Agriculture

...

- Efficient deployment of vision models has become essential

→ **Post-Training Quantization** (**PTQ)** enables efficient inference without retraining

- Recent PTQ methods commonly use **Asymmetric Quantization (AsymQ)** for weights, because it improves low-bit accuracy

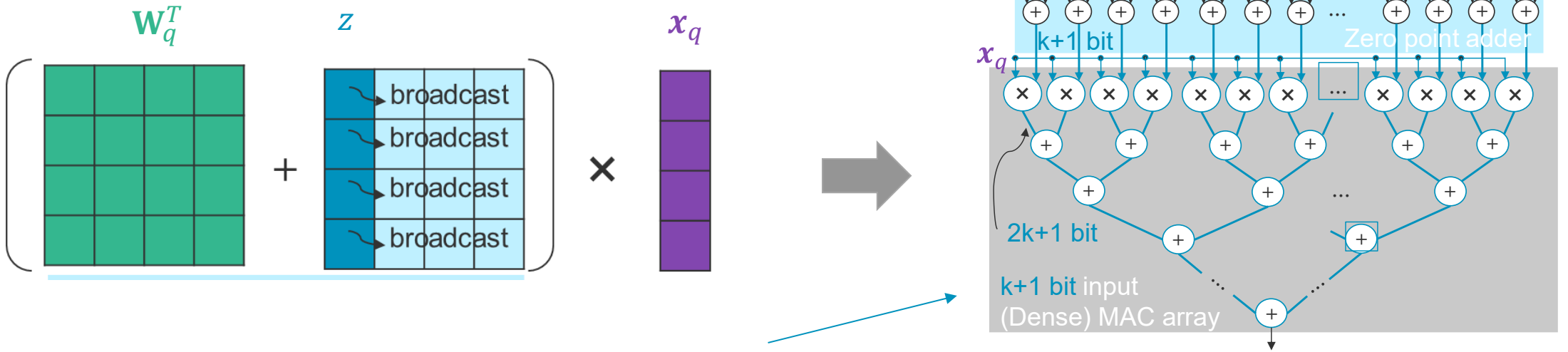
$$\hat{\mathrm{W}} \equiv \mathrm{Q}(\mathrm{W}, \Delta, \mathrm{z}, \mathrm{k}) = \Delta(\mathrm{W}_q - \mathrm{z})$$

$$\mathrm{W}_q \equiv \mathrm{clip}(\lceil \mathrm{W} \rfloor / \Delta + \mathrm{z}, -2^{k-1}, 2^{k-1} - 1)$$

PTQ is attracting increasing attention in both industry and academia

Motivation

- Asymmetric Quantization (AsymQ)
  - Low-bit symmetric W_q + zero-point z addition
  - Commonly adopted for better accuracy than SymQ ($z = 0$)

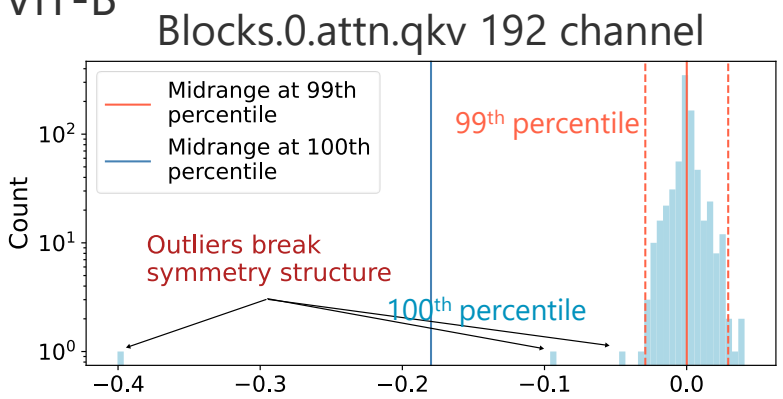


Zero-point addition increases multiplier bit-width→ higher hardware cost

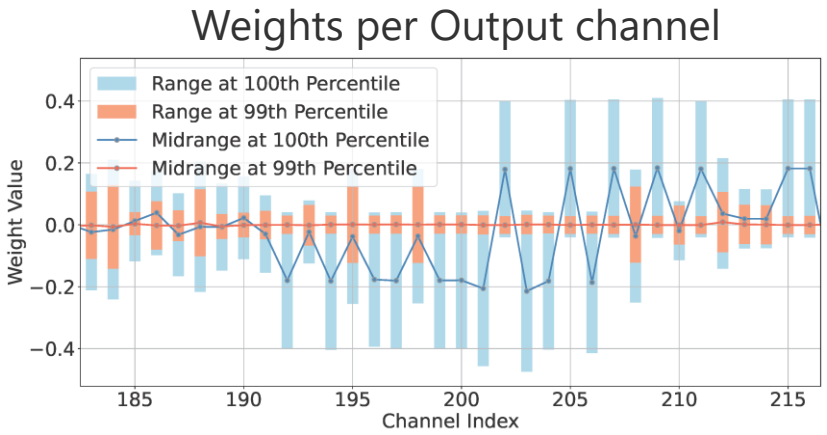
Do we really need AsymQ?

Observation

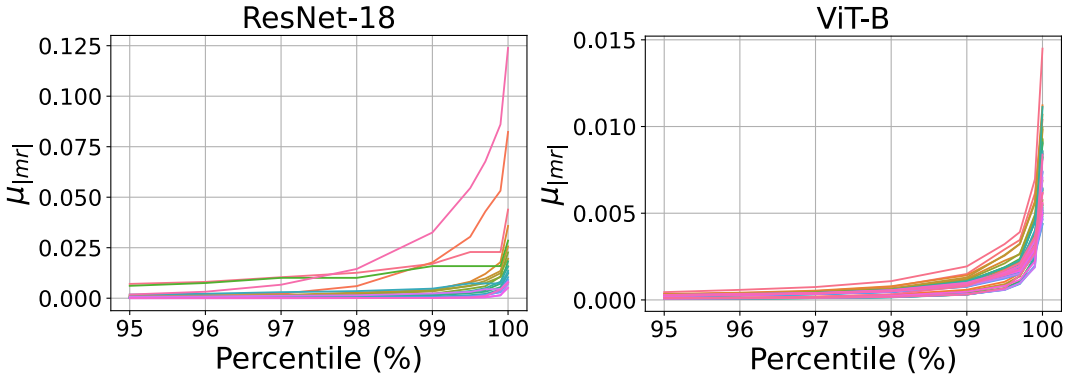
- We analyze the distribution of network weights
  - Weights of ViT-B



Midrange is defined as (max + min) / 2



  - CNN and ViT

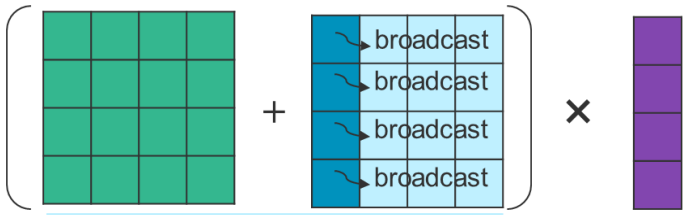
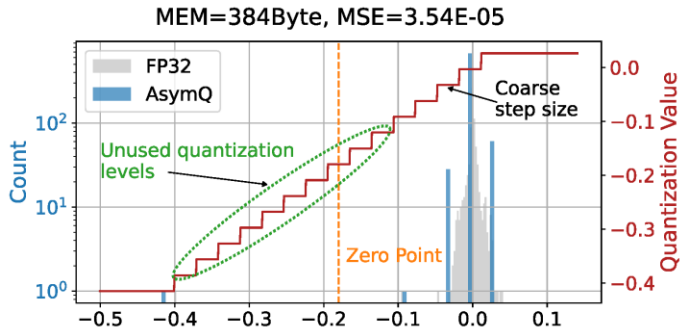
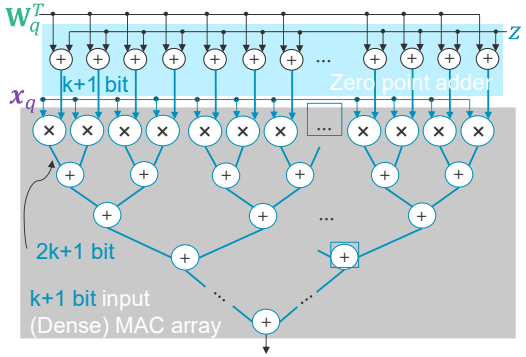
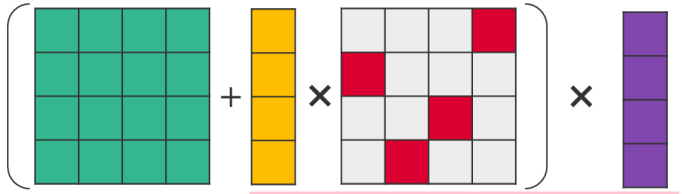
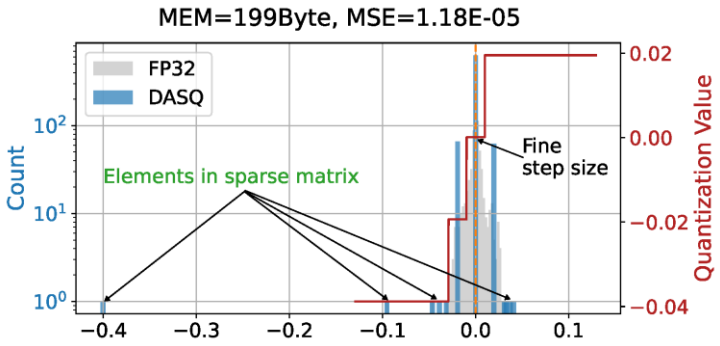
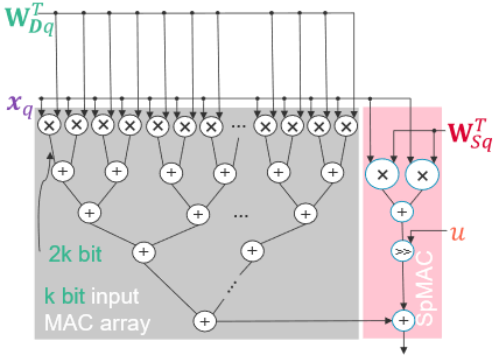


Weight become **symmetric** across CNN and Vision Transformer

$\mu_{|mr|}$ represents the mean of the absolute midrange averaged over output channels in each layer.

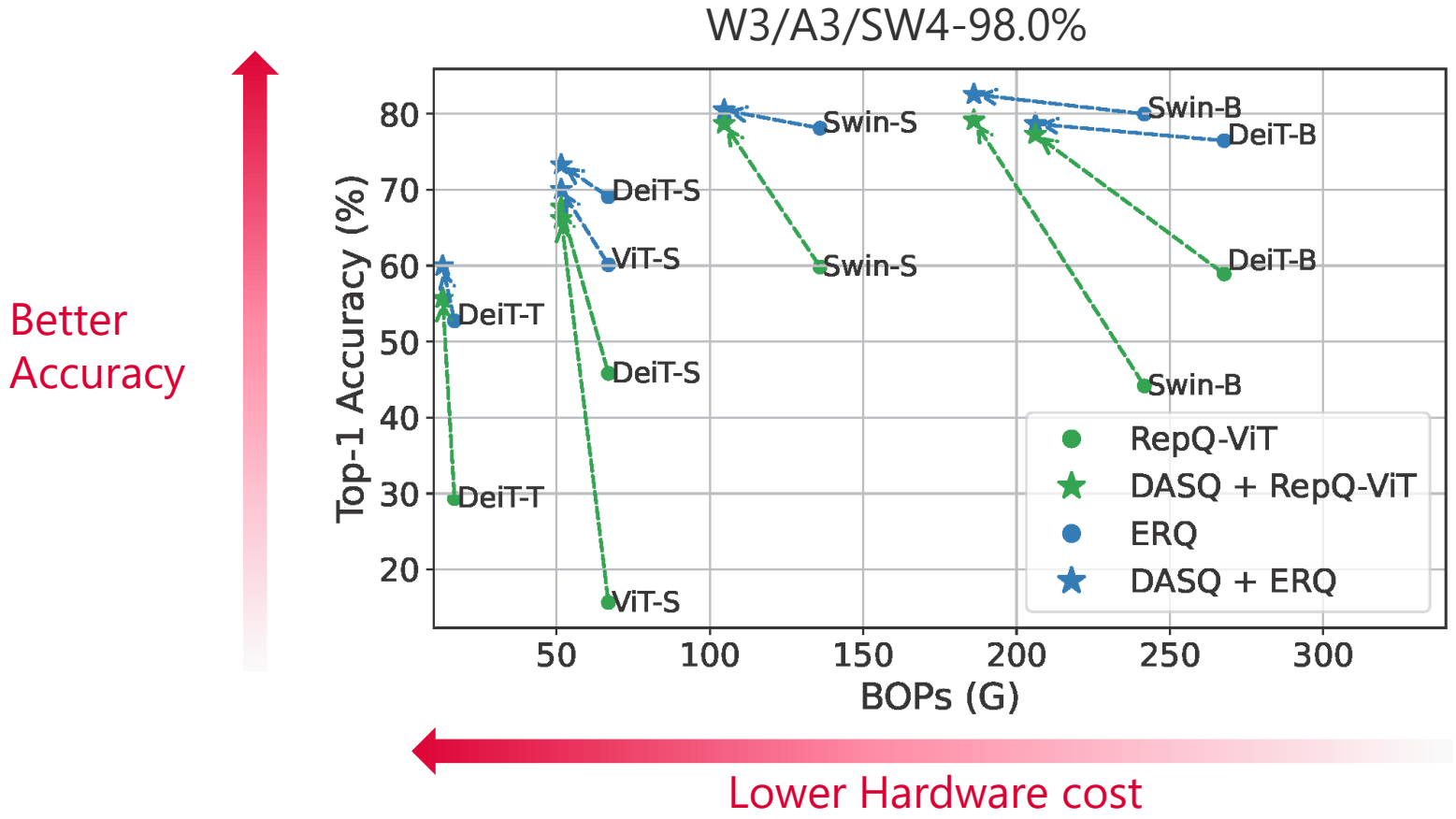
Hidden Symmetry in Vision Model weights

Method: DASQ (Dense and Additive Sparse Quantization)

	Weight Representation	Quantization Efficiency	Hardware Efficiency
AsymQ (Baseline)	$\mathbf{W}_q^T$ z x_q  Zero point Add Asymmetry handled by **zero-point**	 **Larger Quantization Error**	 **Larger Multiplier tree**
DASQ (Ours)	$\mathbf{W}_{Dq}^T$ $1/2^u$ $\mathbf{W}_{Sq}^T$ x_q  Sparse MatMul Asymmetry handled by **sparse outliers**	 **Smaller Quantization Error**	 **Smaller Multiplier tree**

Experimental Results

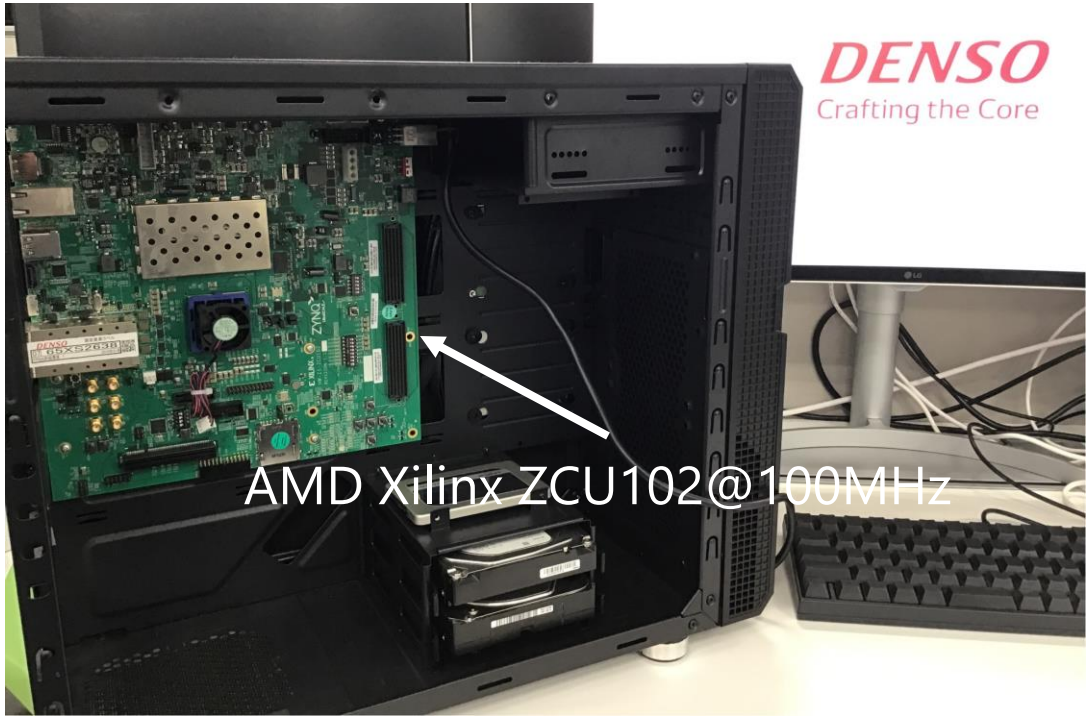
- ImageNet Evaluation
  - Weight Quantization has been changed **AsymQ** to **DASQ**



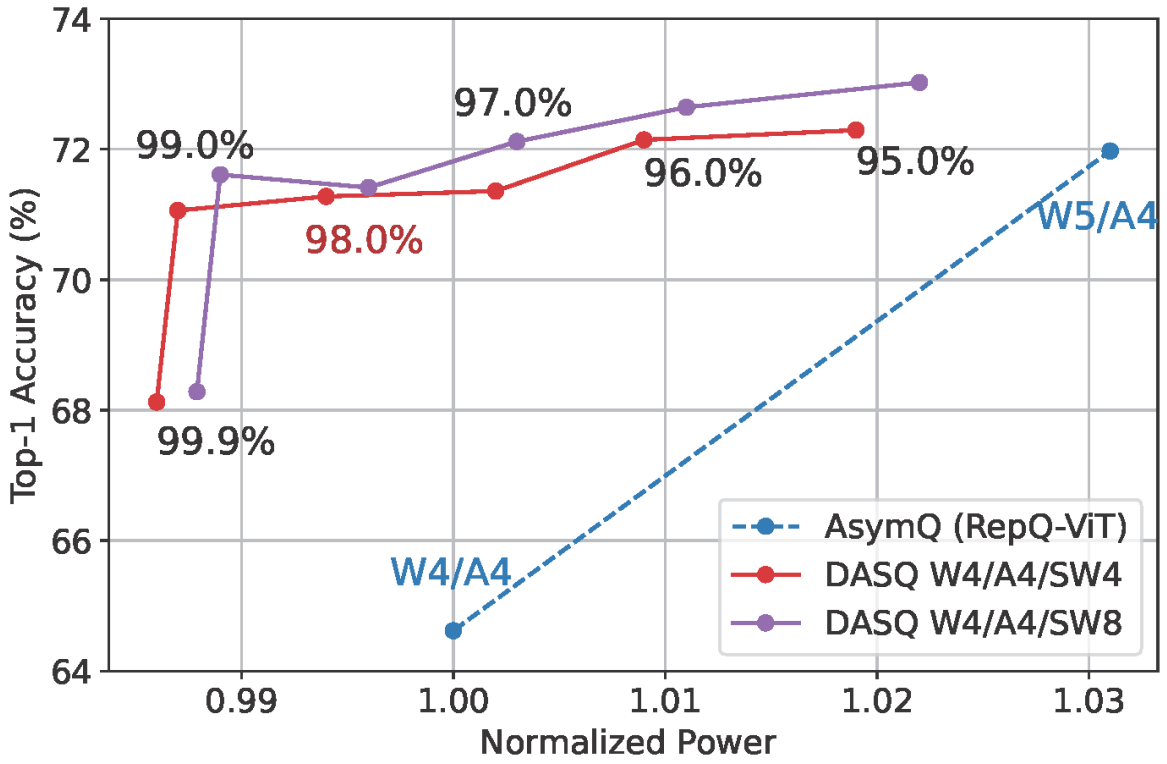
DASQ achieves higher accuracy with lower hardware cost

FPGA Evaluation

FPGA Deployment



Accuracy-Power Tradeoff: ViT-S



DASQ outperforms AsymQ on FPGA

Conclusion

- Rethinking Asymmetric Quantization (AsymQ) for Vision Model Weights
  - ✓ Observation: Hidden symmetry in vision model weights

    Vision model weights become symmetric after removing sparse outliers
  - ✓ Method: DASQ

    Asymmetry is represented by sparse outliers, not zero-point additions
  - ✓ Result

    Higher accuracy and better hardware efficiency than AsymQ

DASQ enables accurate and hardware-friendly inference

CVPR JUNE 3-7, 2026 DENVER COLORADO

Thank you

CVPR Poster: Day3 (Morning)

DENSO

Crafting the Core

DENSO IT LAB