HAP: Accelerating DNNs with Resolution-Preserved Quantization by Harnessing Adaptive-Precision

Erjing Luo*, Xinkuang Geng†, Honglan Jiang†, Leibo Liu‡, Jie Han*

*Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada

†Department of Micro-Nano Electronics, Shanghai Jiao Tong University, Shanghai, China

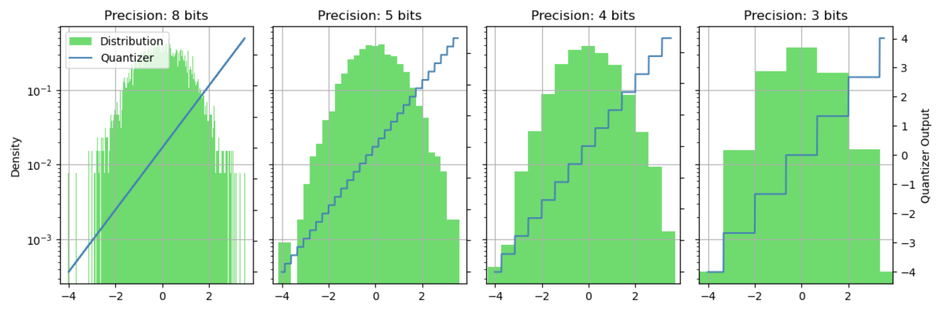
‡School of Integrated Circuits, Tsinghua University, Beijing, China

Email: {erjing1, jhan8}@ualberta.ca

Background & Motivations

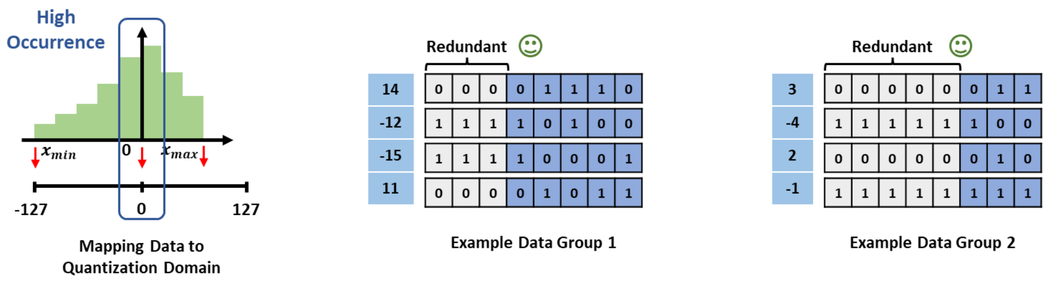
Post-Training Quantization (PTQ) & Limitation

- PTQ has been widely used to accelerate Deep Neural Networks (DNNs) [1-6].
- However, PTQ faces fundamental limits in low-precision domain, as the quantization resolution (i.e., quantizer capacity) decrease exponentially as precision reduces.



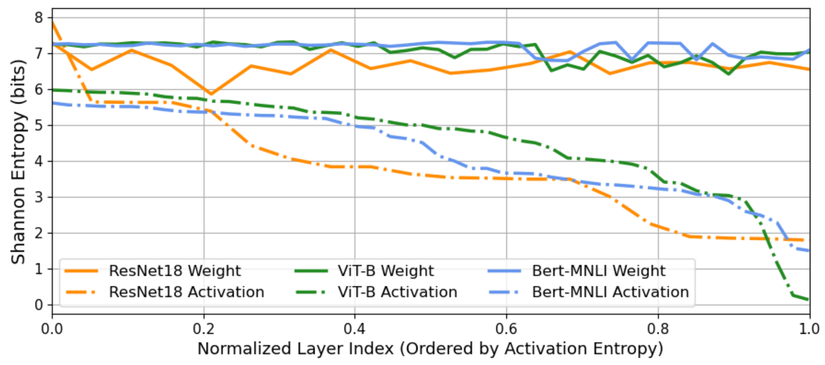
Adaptive-Precision Quantization (APQ)

- **Key idea**: encoding bit-width < quantization precision
- Values with smaller magnitude require less encoding precision
- APQ adjusts bit-width for localized data without sacrificing resolution [8-10].

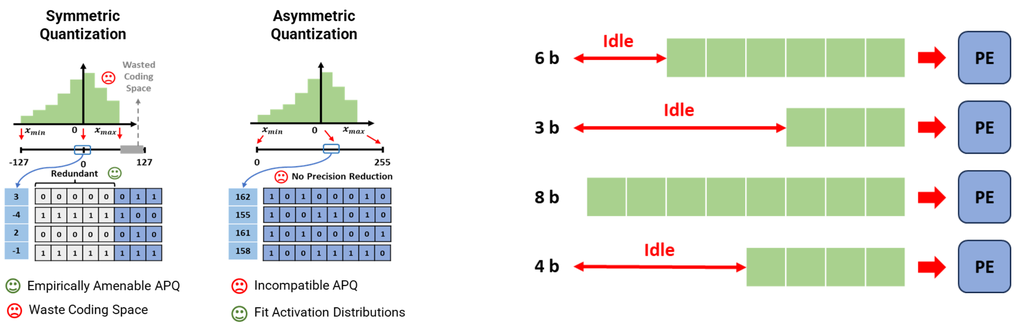


APQ is Especially Suitable for the Activations in DNNs

- Activations are quantized in coarser tensor-level while weights are typically quantized in finer channel-level [1-3].
- In terms of accuracy, the PTQ for activations is less robust than for weights [4-6].
- From an information theory perspective, activations exhibit more exploitable redundancy in their binary representations.



Challenges in Applying APQ to Activations



Existing APQ methods are limited to symmetric quantization (SQ), failing to generalize to asymmetric quantization (AQ). However, AQ is more suitable for activations [11-12].

The variable precision in activations leads to computational workload imbalance in parallel architectures, reducing performance gains.

Reference & Acknowledgement

Reference

[1] Yvinec, Edouard, et al. "Spiq: Data-free per-channel static input quantization." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[2] Moon, Jaehyeon, et al. "Instance-aware group quantization for vision transformers. " Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[3] Guo, Cong, et al. "SQuant: On-the-Fly Data-Free Quantization via Diagonal Hessian Approximation. " Proceedings of International Conference on Learning Representations. 2022

[4] Liu, Yijiang, et al. "Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Xiao, Guangxuan, et al. "Smoothquant: Accurate and efficient post-training quantization for large language models." International conference on machine learning. PMLR, 2023.

[6] Lin, Ji, et al. "Awq: Activation-aware weight quantization for on-device llm compression and acceleration." Proceedings of machine learning and systems 6 (2024): 87-100.

[7] Judd, Patrick, et al. "Stripes: Bit-serial deep neural network computing." 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2016.

[8] Lascorz, Alberto Delmás, et al. "Shapeshifter: Enabling fine-grain data width adaptation in deep learning." Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. 2019.

[9] Shi, Man, et al. "Bitwave: Exploiting column-based bit-level sparsity for deep learning acceleration." 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2024.

[10] Chen, Yuzong, et al. "BBS: Bi-directional bit-level sparsity for deep learning acceleration." 2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2024.

[11] Geng, Xinkuang, et al. "QUQ: quadruplet uniform quantization for efficient vision transformer inference." Proceedings of the 61st ACM/IEEE Design Automation Conference. 2024.

[12] Kam, Dongyun, et al. "Panacea: Novel DNN Accelerator using Accuracy-Preserving Asymmetric Quantization and Energy-Saving Bit-Slice Sparsity." 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2025.

Acknowledgement

The work at the University of Alberta was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (Project Numbers: RES0048688, RES0051374 and RES0054326) and Alberta Innovates (Project Number: RES0053965).

News: The full version of this work has been accepted to DAC 2026. Please see the paper for more details.

Erjing Luo, Xinkuang Geng, Honglan Jiang, Leibo Liu, and Jie Han. "HAP: Efficient Quantization Harnessing Adaptive Precision for DNN Hardware Acceleration." 63rd ACM/IEEE Design Automation Conference (DAC '26), July 26--29, 2026, Long Beach, CA, USA

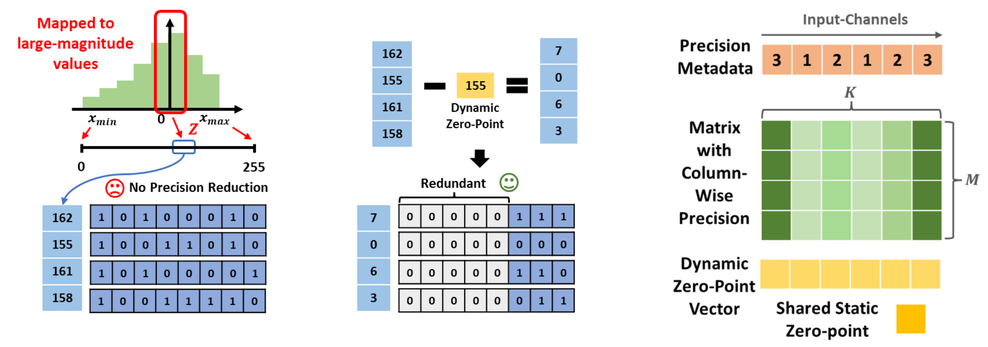
HAP Methodologies

HAP: A Quantization and Accelerator Design to Accelerate DNN Inference

- PTQ for activations: 8-bit AQ + APQ
- PTQ for weights: channel-level dual-precision (4-/8-bit) quantization
- Accelerator: bit-serial computing + dynamic precision matching

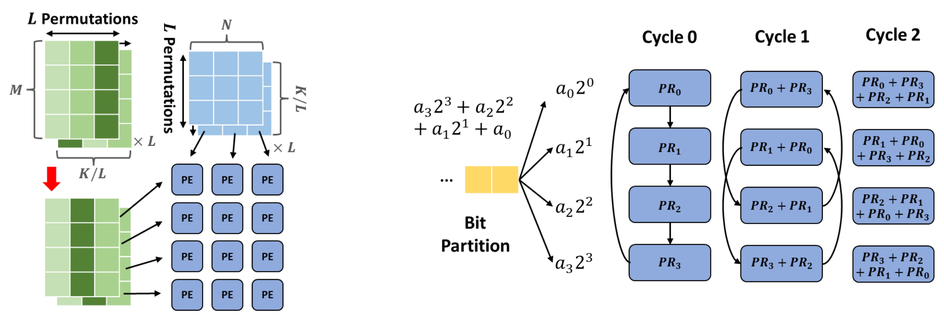
Dynamic Asymmetric Re-Quantization (DAR)

- Introducing a dynamic zero-point to generalize APQ in the scenario of AQ
- Intra-channel grouping strategy to exploit value locality for lower precision



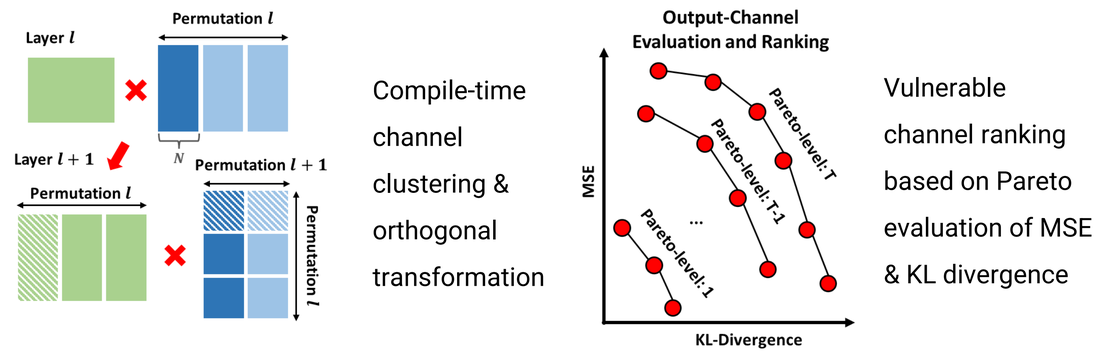
Dataflow for DAR

- Bit-serial output-stationary processing array
- Reordering outer-products to match activation precisions
- Reuse architecture for the dynamic zero-point term



Dual-Precision Vulnerable Channel Protection for Weights

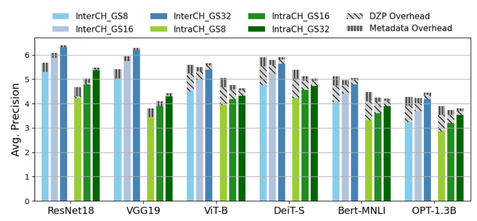
- Targeting 4-bit PTQ while remaining a fraction of channels in 8-bits



Experimental Results

Setup

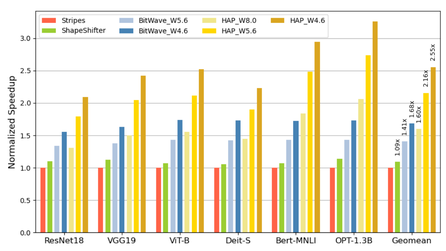
- **DNNs & Datasets:** CNNs (ResNet18, VGG19) and ViTs (ViT-B, Deit-S) on ImageNet-1K, Bert-Base on GELU-MNLI, and OPT-1.3B on WikiText-2
- **Quantization:** SQuant [3], NoisyQuant [4], SmoothQuant [5]
- **Accelerators:** Stripes [7], ShapeShifter [8], and BitWave [9]
- **Methodology:** cycle-accurate simulator + ISO-compute comparison (512 8-bit multipliers)



Average precision for DAR, under different settings of grouping Strategies (intra-channel or inter-channel) and group sizes.

Top-1 Accuracy (%) & Perplexity (for OPT-1.3B)

Prec (W/A)	8/8	4/8	8/4	5/5	4.6/4~	5.6/4~
CNNs	SQuant [3]				HAP+SQuant	
ResNet18↑	69.65	68.49	63.41	68.09	68.84	**69.21**
VGG19↑	74.16	73.61	69.22	72.95	73.91	**74.09**
Vision Trans.	NoisyQuant [4]				HAP	
ViT-B↑	84.31	81.62	72.42	79.04	82.98	**83.47**
Deit-S↑	79.24	76.50	48.32	61.05	78.26	**78.84**
NLP Models	SmoothQuant [5]				HAP	
Bert-MNLI↑	84.62	83.82	73.92	75.31	84.65	**84.74**
OPT-1.3B↓	16.77	230.58	1243.69	55.14	25.85	**18.84**



Performance comparisons (normalized to Stripes [7]).

Conclusion

- **HAP**: a resolution-preserved quantization and accelerator solution for DNN acceleration
- A novel **adaptive precision-quantization** method for **activations**
- A bit-serial accelerator that balances computational workloads via **outer-product reordering**
- A **dual-precision weight quantization** providing flexibility for accuracy-performance trade-off
- Superior accuracy at 4- and 5 bits across typical DNNs and benchmarks
- 2.55x speedup on average over existing accelerators