SA-ANT: Efficient Low-Bit Group-Wise Quantization for Large Language Models via Sign-Asymmetric Adaptive Numeric Type

Xinkuang Geng, Siting Liu, Hui Wang, Jie Han, Honglan Jiang

School of Integrated Circuits, Shanghai Jiao Tong University Shanghai, China

Shanghai Jiao Tong University

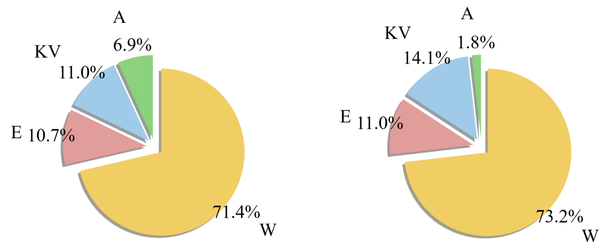
DATE 26

Introduction

- Weight Quantization of Large Language Models
  - Maps high-precision weights to low-bit discrete values [1]
- Group-Wise Quantization
  - Divides weights into groups with independent scaling factors
  - Improves accuracy through finer quantization granularity [1,3,4]

$$W_q = q(W_f, G) = \arg\min_{W_i} |W_i - \frac{W_f}{\Delta}|, \quad W_i \in G \qquad W_{qf} = \Delta \cdot W_q \approx W_f$$

- Adaptive Numeric Type
  - Dynamically selects the best quantization format for each group
  - Better fits diverse data distributions [2]



(a) Discriminative (b) Generative

Memory ratio of activations (A), KV cache (KV), embeddings (E), and weights (W) in Llama-3.1-8B under discriminative and generative workloads.

$$G^\star = \arg\min_{G_i} \|W_f - \Delta \cdot q(W_f, G_i)\|, \quad G_i \in G_S$$

Sub Numeric Type Adaptation

References

[1] J. Lin, J. Tang, H. Tang et al., "AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration," Proceedings of machine learning and systems, vol. 6, pp. 87–100, 2024.

[2] C. Guo, C. Zhang, J. Leng et al., "ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization," in 2022 55thIEEE/ACM International Symposium on Microarchitecture (MICRO),2022, pp. 1414–1433.

[3] W. Hu, H. Zhang, C. Guo et al., "M-ANT: Efficient Low-bit Group Quantization for LLMs via Mathematically Adaptive Numerical Type," in 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2025, pp. 1112–1126.

[4] Y. Chen, A. F. AbouElhamayed, X. Dai et al., "BitMoD: Bit-serial Mixture-of-Datatype LLM Acceleration," in 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2025,pp. 1082–1097.

Motivation

Why use smaller groups?

Finer quantization granularity

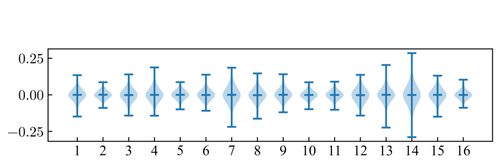
Enables **low-bit** quantization

What are the key challenges?

Non-Gaussian distributions

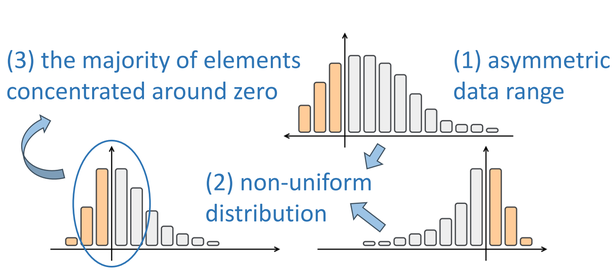
Large variation across groups

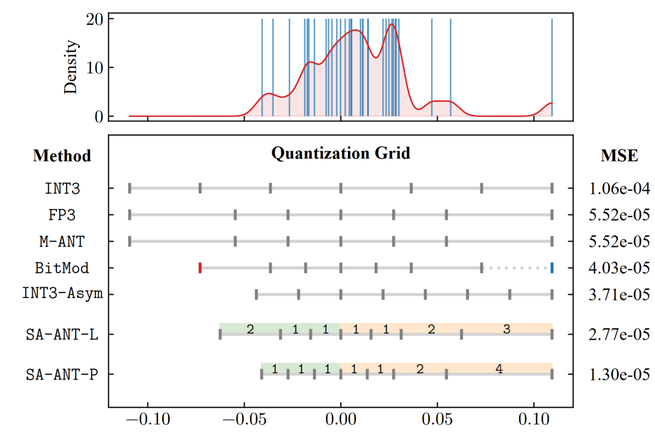
- **INT3** fails to capture the asymmetry and non-uniformity of the data → high MSE
- **FP3** provides finer resolution for values near zero → better matches non-uniformity → reduced MSE
- **M-ANT** [3] provides 16 strictly symmetric sub numeric types
- **BitMod** [4] exploits the redundant zero encoding → introduces asymmetry → reduced MSE
- **INT3-Asym** [1] employs a zero point to offset the INT3 grid → better matches asymmetry → reduced MSE



(a) 2048 elements per channel (b) 32 elements per group

Data distributions of 16 sampled channels (a) and 16 sampled groups (b) from Llama-3.2-1B.





Data distribution, quantization grids and errors of different numeric types for one group from Llama-3.2-1B. Blue vertical lines denote the elements within this group.

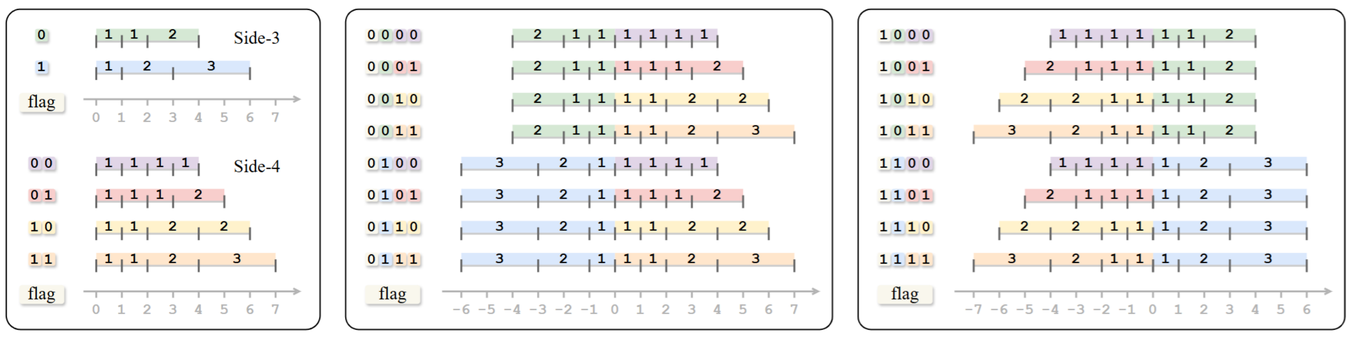
Sign-Asymmetric Adaptive Numeric Type

- Small-group data follows a **skewed normal distribution**
- SA-ANT Quantization Grid Construction
  - Predefine half-grids separately for positive and negative sides
  - Combine the half-grids to form asymmetric adaptive quantization grids

Match Well

Non-decreasing sequences employed in SA-ANT and their corresponding INT targets

Variant	Side	Spacing of the half grids				Max	Target
SA-ANT-L	*T*	1,1,2	1,2,3			7	INT4
	Q	1,1,1,1	1,1,1,2	1,1,2,2	1,1,2,3		
SA-ANT-P	*T*	1,1,1	1,1,2	1,2,2	1,2,4	13	INT5
	Q	1,1,1,1	1,1,1,2	1,1,2,2	1,1,2,4		
		1,2,2,2	1,2,2,4	1,2,4,4	1,4,4,4		



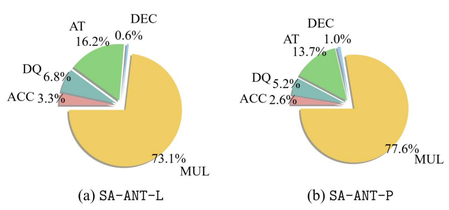
(a) Half grids of Side-3 & Side-4 (b) Side-3 for negative & Side-4 for positive (c) Side-4 for negative & Side-3 for positive

Overview of the construction of 16 quantization grids (sub numeric types) in **SA-ANT-L**

- **Combination of half grids** $G_{SA} = \bigcup_{i=1}^{n_T} \bigcup_{j=1}^{n_Q} \{-T_i \cup Q_j, -Q_j \cup T_i\}$
- **Variants:** **SA-ANT-L** : 16 sub-numeric types ⟶ Lightweight **SA-ANT-P** : 64 sub-numeric types ⟶ High-precision

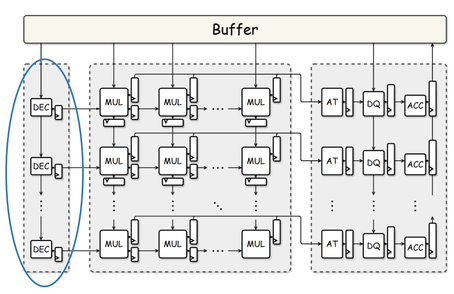
Accelerator Design

- **Low Decoder Overhead**
  - The decoder introduces only 0.6% / 1% of the compute-core area.
- **INT-Compatible Computation**
  - Supports standard low-bit INT operations (Mul / Acc / Dequant).
  - **SA-ANT-L** → INT4
  - **SA-ANT-P** → INT5



(a) SA-ANT-L (b) SA-ANT-P

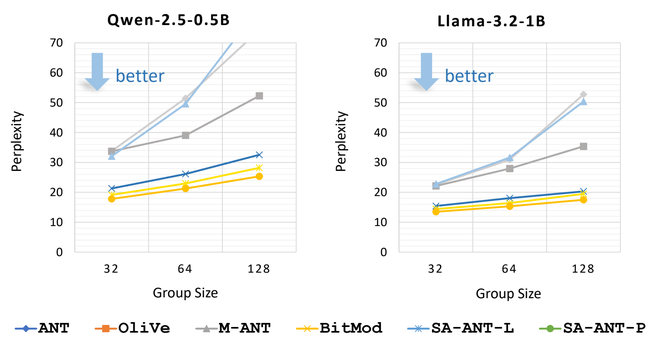
Area breakdown of the compute core



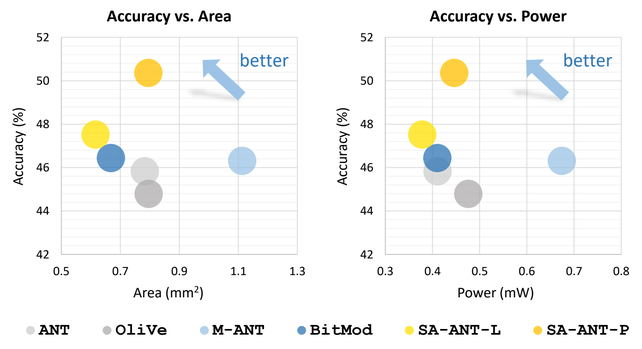
Microarchitecture of the accelerator

Experiments

- Across different models and group sizes, SA-ANT consistently achieves lower perplexity than existing methods [2,3,4], , indicating better language modeling performance.
- SA-ANT not only improves accuracy but also reduces area and power compared to existing methods [2,3,4].



Perplexity across different group sizes



Average accuracy vs. area & power of the compute core