Hybrid SLC/MLC ReRAM for Quantized KV Cache in Transformer-based LLM

Shota Suzuki[1, *], Naoko Misawa[1], Chihiro Matsui[1], and Ken Takeuchi[1, 2]

[1] Department of Electrical Engineering and Information Systems, The University of Tokyo, Japan
[2] Systems Design Lab., Graduate School of Engineering, The University of Tokyo, Japan
*E-mail: shota.suzuki@co-design.t.u-tokyo.ac.jp

Abstract

Quantizing KV Cache is a remarkable technology to reduce the footprint of memories during LLM inference. Multi-level cell (MLC) ReRAM realizes smaller memory areas than single-level cell (SLC) ReRAM, but it causes false detection of cell state due to read variation from random telegraph noise. This work proposes a strategy to store quantized KV cache in a hybrid SLC/MLC ReRAM to reduce memory area while maintaining LLM inference performance. The strategy splits elements into MSB and LSB, then stores MSB in SLC ReRAM, LSB in MLC ReRAM. This work achieves 25% fewer cells than the conventional method to keep perplexity at an acceptable level.

1. Introduction

Transformer-based large language models (LLMs) [1, 2] excel in natural language processing. During inference, KV Cache minimizes computational overhead by storing the previous key and value for each token. As new tokens are processed and attention scores are generated, KV Cache continuously expands (Fig. 1), consequently increasing its DRAM footprint (Fig. 2(a)). This substantial KV Cache size necessitates significant data transfer between GPUs and memories, driving the need for size reduction strategies and high-capacity memories [3]. Quantization is one technique to shrink KV Cache, while maintaining LLM perplexity at lower bit-widths [4] and decreasing DRAM usage (Fig. 2(b)). ReRAM is an emerging non-volatile memory for storage-class applications. ReRAM can achieve multiple conductance states, offering higher density and energy efficiency [5, 6, 7]. However, some data stored in multi-level cell (MLC) ReRAM can be susceptible to readout errors due to current fluctuations caused by random telegraph noise (RTN) [8]. Single-level cell (SLC) ReRAM stores less data but avoids these errors. This work investigates the utility of ReRAM as storage for quantized KV Cache. A hybrid SLC/MLC ReRAM architecture (Fig. 2(c)) is proposed to reduce memory area while mitigating MLC ReRAM's current fluctuation impact on LLM inference.

2. Quantized KV Cache with Hybrid SLC/MLC ReRAM

ReRAM Device and Read Variation

Fig. 3 illustrates a TaO_X-based ReRAM structure and switching mechanism. ReRAM possesses two basic states: a high resistance state (HRS) and a low resistance state (LRS), switched by appropriate Set/Reset voltages [7]. Fig. 4 depicts cell current distributions for a 2 bits/cell MLC ReRAM at 0.2 V read voltage [8]. RTN causes current fluctuations, leading to values outside the expected range. A confusion matrix (Fig. 5) summarizes occurrences of false detections over 10^3 readings in MLC ReRAM [8]. This matrix helps estimate the read variation's impact on LLM perplexity. When utilized as SLC, states S0 and S3 are chosen for logical "0" and "1", respectively. SLC ReRAM has no false detection due to the wide memory window between S0 and S3.

Quantization of KV Cache and Storing Data

Fig. 6 shows KV cache quantization and storage. Quantization occurs when the sequence length becomes a multiple of a predefined residual length, preventing LLM perplexity degradation [4]. Each KV Cache element is 4-bit quantized using quanto [9], which is a simple and fast tool. A quantized KV Cache consists of one tensor and two parameters. Quantized tensors (X_Q) are uint4 data and share the original KV Cache tensor's (X) channel dimension length. Scale (S) defines the quantization step size, and zero point (Z) is another parameter used to map X_Q elements within the 0 to $2^4 - 1$ range. Quanto calculates S and Z from X's min/max element. All bits of each component are stored in either SLC or MLC ReRAM, based on bit significance (Fig. 7). X_Q is split into its upper 2 bits (MSB) and lower 2 bits (LSB), stored in SLC and MLC ReRAM, respectively. Z is int8, but its values range from 0 to $2^4 - 1$, so its lower 2 bits (LSB) go to MLC ReRAM. S is a 16-bit floating-point number; its fractional part, less critical than the sign and exponent, is saved in MLC ReRAM.

3. Evaluation and Result

The proposed method's perplexity is evaluated on Llama3.1-8B [1] and Phi-4 [2] using the PG-19 test dataset [10] (Table I). Fig. 8 presents perplexity for each condition. The proposal's perplexity (magenta) closely matches the conventional (black or red) for both LLMs up to input sequence lengths of 2,000. At a length of 5,000, it stays around 7.0, still acceptable for LLM tasks [4]. Ablation studies investigate each component's (X_Q, S, Z) vulnerability. Storing all bits of each element solely in MLC ReRAM leads to catastrophic perplexity (orange, brown, or gold). This underscores the necessity of the SLC/MLC ReRAM combination to prevent perplexity spikes. When stored in hybrid ReRAM, X_Q (blue) and Z (cyan) demonstrate greater susceptibility to MLC ReRAM's read variation than S (green). The proposed method reduces memory cell counts by 25 % compared to conventional 4-bit quantized KV Cache implementations (Table II). These results indicate that the proposed approach effectively reduces memory area while maintaining low perplexity.

4. Conclusions

This work proposes a strategy to store quantized KV Cache in a hybrid SLC/MLC ReRAM architecture. This approach reduces memory area while maintaining performance during LLM inference. The proposed method achieves a 25 % reduction in memory cells and keeps perplexity below 7.5.

Acknowledgements The authors thank S. Yoneda, S. Ito, and S. Awamura of NTCJ. This paper is based on results obtained from a project, JPN23015, subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

References [1] A. Grattafiori *et al.*, *arXiv*, 2024, doi: 10.48550/arXiv.2407.21783. [2] M. Abdin *et al.*, *arXiv*, 2024, doi: 10.48550/arXiv.2412.08905. [3] H. Liu *et al.*, *arXiv*, 2024, doi: 10.48550/arXiv.2412.19442. [4] Z. Liu *et al.*, *ICML*, 2024, pp. 32332-32344. [5] R. Mochida *et al.*, *VLSI Tech.*, 2018, pp. 175-176. [6] M. Morimoto *et al.*, *SSDM*, 2024, pp. 121-122. [7] T. Ninomiya *et al.*, *JJAP*, vol. 53, no. 11R, pp. 114-201, 2013. [8] K. Yamauchi *et al.*, *IRPS*, 2025, pp. 10C.4-1-10C.4-6. [9] "optimum-quanto", *GitHub repository*, [Online]. Available: https://github.com/huggingface/optimum-quanto, Accessed: Jul. 11, 2025. [10] J.W. Rae *et al.*, *arXiv*, 2019, doi: 10.48550/arXiv.1911.05570.

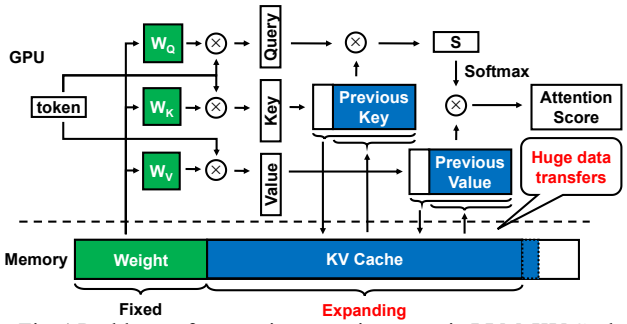

Fig. 1 Problems of generating attention score in LLM. KV Cache expands as new token is processed. Large KV Cache introduces significant data transfer between GPU and memory.

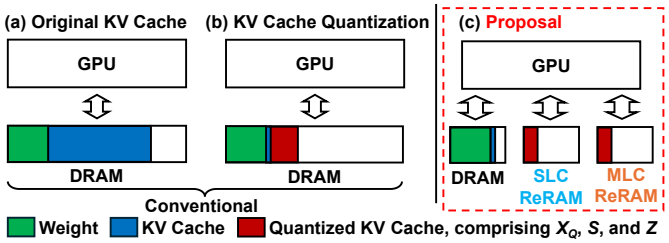

Fig. 2 Comparison of conventional and proposal. (a) Original KV Cache, occupying DRAM. (b) KV Cache quantization, compressing size. Some KV Cache remains temporarily in DRAM to preserve LLM performance (c) Proposal, storing quantized KV in hybrid SLC/MLC ReRAM to realize higher capacity while maintaining LLM inference performance.

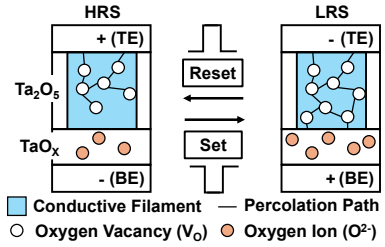

Fig. 3 Structure and switching mechanism of TaOx ReRAM between HRS and LRS by applying Set/Reset voltage [7].

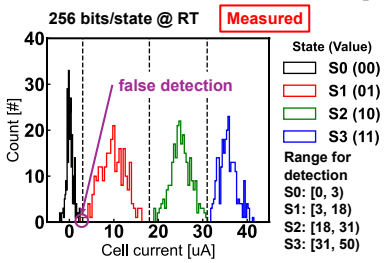

Fig. 4 Distribution of cell current on MLC ReRAM at 0.2 V read voltage [8]. Some cell states are detected incorrectly due to fluctuation.

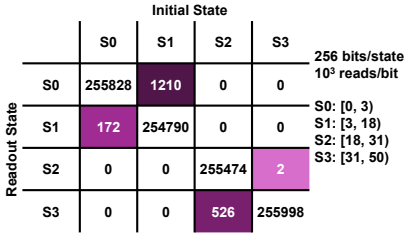

Fig. 5 Confusion matrix of initial state vs. readout state in MLC ReRAM at 0.2 V read voltage over 10^3 read cycles [8].

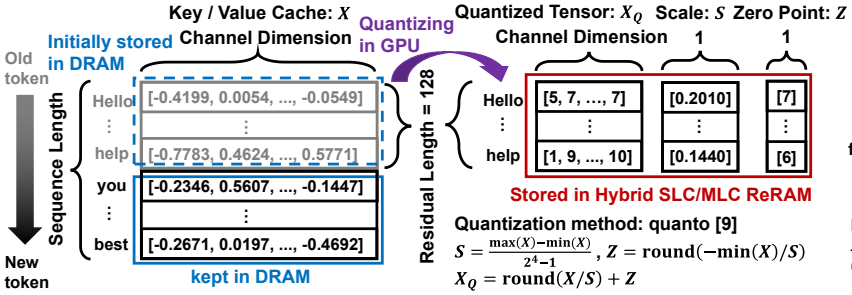

Fig. 6 Quantization of KV Cache and proposed hybrid SLC/MLC ReRAM. Sequence length is number of processed tokens, and each token is a vector of channel dimension length. KV Cache is initially stored in DRAM. Residual length defines number of temporal tokens kept in DRAM to prevent perplexity degradation and save computation time. When sequence length is multiple of residual length, KV Cache is quantized per token using [9]. Quantized KV Cache comprises quantized tensor X_Q, scale S, and zero point Z. S is quantization step, and Z is offset to shift X_Q in range of 0 to $2^4 - 1$. All components are stored in hybrid SLC/MLC ReRAM.

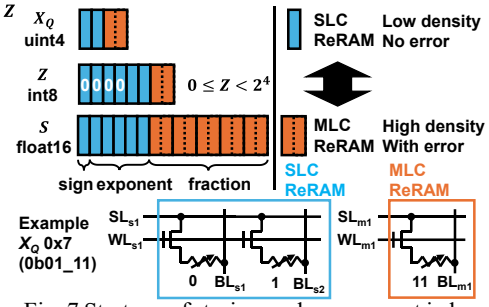

Fig. 7 Strategy of storing each component in hybrid SLC/MLC ReRAM. Each data is separated into MSB and LSB, and MSB is stored in SLC ReRAM, whereas LSB is stored in MLC ReRAM. X_Q's LSB is lower 2 bits. Z has values between 0 and $2^4 - 1$; LSB is lower 2 bits. S is floating-point number; LSB is fractional part.

Table I Evaluation Method

Model	Llama3.1-8B [1], Phi-4 [2]
Dataset	PG-19 test* [10]
Index	Perplexity
Quantization bits	4 bits
Quantization Method	quanto [9]

* 1 of 100 samples is used due to time constraints

Table II Number of necessary cells per token

	Cells per token * Channel Dimension of X or X_Q ** dimension of S, *** dimension of Z
Conventional KV Cache	2048 cells (DRAM) = (16 bits / 1 bit/cell) × 128*
Conventional 4Q-KV	536 cells (DRAM) = (4 bits / 1 bit/cell) × 128* + (16 bits / 1 bit/cell) × 1** + (8 bits / 1 bit/cell) × 1***
Proposal	402 cells (Hybrid SLC/MLC ReRAM) = (2 bits / 1 bit/cell + 2 bits / 2 bit/cell) × 128* + (6 bits / 1 bit/cell + 10 bits / 2 bit/cell) × 1** + (6 bits / 1 bit/cell + 2 bits / 2 bit/cell) × 1***

- 25%

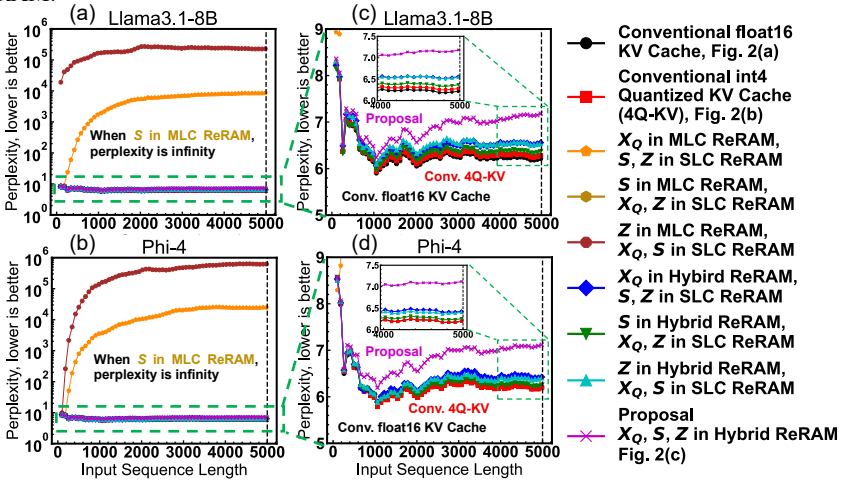

Fig. 8 Perplexity of (a) Llama3.1-8B and (b) Phi-4 for each KV Cache, plotted every 80 inputs. Detailed comparison of conventional and proposal in (c) Llama3.1-8B and (d) Phi-4. Proposal prevents perplexity degradation for long input sequence lengths.