Efficient Multidimensional Data Mapping for Low-Power AI Tensor Processing

Tatsuya Ishizaki

Utsunomiya Kyowa University, Utsunomiya, Tochigi, 320-0811, Japan

(E-mail: capestone0214@gmail.com, t-ishizaki@kyowa-u.ac.jp)

Keywords : AI, Artificial Intelligence, Tensor, Image Processing, Memory, Power Consumption

Abstract

Efficient management of multidimensional data is critical for high-performance AI tensor and image processing. This paper explores the application of a patented data mapping technique (US7907473, JP5133073), originally designed for semiconductor memory, to improve memory access locality in AI workloads. By efficiently mapping multidimensional tensors according to their coordinates, the approach reduces unnecessary memory activations, potentially lowering power consumption and enhancing computational efficiency. We present a conceptual framework for integrating this method into AI tensor operations, analyze its potential benefits in terms of memory access patterns, and discuss implications for low-power AI system design.

1. Introduction

Since the early 2020s, generative AI has begun to gain widespread attention and adoption, contributing to a new phase in the advancement of artificial intelligence. Recent developments in large language models (LLMs) have contributed to more natural forms of interaction between artificial intelligence systems and human users. Language generation in LLMs relies on large-scale tensor operations and substantial memory access. Gholami et al. (2024) highlighted the importance of memory bandwidth in generative AI systems, suggesting that improving memory performance has become an essential requirement in AI development.

In practical systems, memory controllers determine the physical placement of tensor data in semiconductor memory based on instructions issued by GPUs. Generative AI systems built upon LLMs process large-scale tensor data, where reductions in power consumption and improvements in computational latency are

increasingly important. Accordingly, optimizing how tensor data are written to and read from semiconductor memory, particularly in terms of reducing energy usage and minimizing access latency, represents a critical design consideration.

This paper suggests that the data storage method claimed in U.S. Patent No. 7,907,473 (Ishizaki et al., 2011) and Japanese Patent No. 5,133,073 (Ishizaki et al., 2013) may contribute to reductions in power consumption and improvements in data processing speed when applied to tensor data management in semiconductor memory.

2. Background

Ghose et al. (2018) investigated accurate methods for analyzing DRAM power consumption, noting that DRAM can account for nearly half of total system power. This section describes memory write and read operations using DRAM as a representative architecture. However, the scope of the referenced patents is not restricted to DRAM and extends to broader classes of semiconductor memory systems.

2.1. General Memory Architecture

In conventional DRAM architectures, data write operations are carried out by specifying X and Y addresses that uniquely identify a target memory cell. The X address activates a corresponding word line, whereas the Y address selects the associated bit line (digit line). If a burst length is specified, data are written sequentially from the initial address, with the Y address incremented along the same activated word line.

In DRAM, data write and read operations require not only peripheral circuit control but also activation of the word line connected to the target memory cells. The activation of each word line incurs a certain energy cost. Consequently, overall power consumption increases proportionally to the number of activated word lines.

In general, tensor processing involving multi-dimensional data may require activation of multiple word lines, depending on the data layout in memory. Therefore, in generative AI systems, power consumption during tensor processing is expected to increase as the number of activated word lines increases. Reducing the number of activated word lines during tensor operations may improve energy efficiency, which is an important consideration in large-scale generative AI systems.

2.2. Patented Data Mapping

U.S. Patent No. 7,907,473 (Ishizaki et al., 2011) and Japanese Patent No. 5,133,073 (Ishizaki et al., 2013) propose an address mapping scheme intended to reduce power consumption in semiconductor memory during multi-dimensional data processing. The patents describe converting external X and Y addresses into internal addresses such that multi-dimensional data can be stored within a single word line. Specifically, the internal row address (CAX), which drives the word line and its associated circuitry, is constructed from common bits of the multi-dimensional data address, while the internal column address (CAY), which drives the digit lines and associated circuitry, is formed from the remaining bits. The patents indicate that the address conversion can be implemented either by internal memory circuitry or by external processing logic.

As stated in Ishizaki et al. (2011, 2013), claims 1, 2 and 3 of these patents describe a method for storing multidimensional data in memory, where word lines are selected based on common addresses, and the remaining addresses are used to select bit lines. This approach allows multidimensional data to be stored within a single word line.

3. Integration of Patented Data Mapping into Tensor Processing

LLMs and other generative AI models are known to process tensor data. Tensor data consist of multi-dimensional bits and are stored in memory as needed during computation. The method of storing this data depends on the memory controller and is generally not disclosed outside of the hardware design; however, it is assumed that efficient storage mechanisms are employed.

The patents discussed in this paper propose a technique for linearizing multi-dimensional data and storing it in a single memory word line. For tensor data used in generative AI, the common bits of the multi-dimensional address can be used to select the word line, while the remaining bits are used for digit line selection. This approach allows the tensor data to be stored in a single or minimal number of word lines. By doing so, it potentially reduces the number of activated word lines during tensor processing, contributing to more energy-efficient operations in generative AI systems.

4. Discussion

Vaswani et al. (2017) presented the Transformer, which incorporates the attention mechanism, for generative AI. We consider a four-dimensional tensor defined by the dimensions (b, l, d, h), where b represents the batch size, l the sequence length, d the

embedding dimension, and h the number of attention heads. The following discussion focuses on how tensor data across these dimensions can be written to and read from memory in a manner that reduces word-line power consumption.

To analyze memory activation behavior, we introduce a simplified model of the attention computation stage. Specifically, we assume that the fundamental feature processing unit within each attention head corresponds to a N-dimensional vector indexed by (b, l, h) with N defined as d/h. Under this assumption, we map (b, l, h) to the word-line selection address cax, and the feature dimension N to the bit-line selection address cay. This mapping allows each feature vector to be written or read with a single (or minimal) word-line activation per computational unit.

- The word-line selection address (CAX) can be assigned b, l and h.
- The bit-line selection address (CAY) can be assigned N.

In contrast, if the address mapping is reversed, with N assigned to the word-line selection address and (b, l, h) to the bit-line selection address, reading a single N-dimensional feature vector would require N word-line activations. Therefore, in this simplified model, the proposed mapping can theoretically reduce word-line activation counts, and hence word-line-associated energy consumption, by up to a factor of 1/N under an idealized model.

An alternative address translation example is as follows. Consider data with two-dimensional addresses, where each word stores data corresponding to h and N. If b and l are treated as common addresses, this mapping also allows each feature vector to be written or read with a single (or minimal) word-line activation per computational unit.

- The word-line selection address (CAX) can be assigned b and l,
- The bit-line selection address (CAY) can be assigned h and N.

It should be noted that the amount of data that can be stored within a single word line depends on the underlying memory architecture, and a full tensor feature vector cannot necessarily be mapped onto a single word line in practical implementations. The 1/N reduction is only a rough estimate. Actual energy savings depend on the memory design and could be smaller. Our approach may help alleviate the memory wall bottleneck by reducing the cumulative latency overhead associated with repeated row activation and precharge operations.

The above discussion is based on a simplified analytical model. As Wang et al. (2020) explored to uncover address mapping, practical address translation remains unknown. In order to perform efficient read and write operations at the memory cell level, appropriate address translation mechanisms must be designed while taking the physical memory organization into account.

5. Conclusion

This paper examined a method for efficiently writing and reading tensor data in generative AI, including LLM-based models. In general, memories such as DRAM consist of multiple word lines, bit lines, storage cells, and driving circuits. Writing and reading data at the cell level require selecting both the word line and the bit line connected to the target cell. Consequently, the more word lines are activated, the higher the energy consumption. In generative AI, where reducing power consumption is critical, an increase in the number of activated word lines is a significant concern.

In this paper, we proposed a method based on the aforementioned patents to write and read tensor data using a single or minimal number of word lines. Specifically, the approach assigns the common bits of the tensor data to the word-line selection address (CAX) and the non-common bits to the bit-line selection address (CAY) for the data intended to be grouped together.

This method potentially enables more efficient writing and reading of tensor data for generative AI using a single or minimal number of word lines, which in turn may contribute to reduced energy consumption.

Overall, this paper represents an initial investigation highlighting the potential usefulness of the patented address translation for storing tensor data in memory. Future research could explore more concrete implementations and evaluate the associated reductions in energy consumption, which may contribute to the advancement of AI.

Conflict of Interest

The author is a co-inventor of the patents referenced in this paper.

References

Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M. W., & Keutzer, K. (2024). AI and memory wall. *IEEE Micro, 44*(3), 33-39.

Ghose, S., Yaglikçi, A. G., Gupta, R., Lee, D., Kudrolli, K., Liu, W. X., Hassan, H., Chang, K.

K., Chatterjee, N., Agrawal, A., O' Connor, M., & Mutlu, O. (2018). What your DRAM power models are not telling you: Lessons from a detailed experimental study. *Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2*(3), Article 38, 1-41.

Ishizaki, T., Nakamura, H., Kurokawa, T., & Ushikoshi, K. (2011). Semiconductor memory device and data storage method including address conversion circuit to convert coordinate information of data into onedimensional information to amplifier (U.S. Patent No. 7,907,473). U.S. Patent and Trademark Office. https://patents.google.com/patent/US7907473B2/en

Ishizaki, T., Nakamura, H., Kurokawa, T., & Ushikoshi, K. (2013). Semiconductor memory device and data storage method including address conversion circuit to convert coordinate information of data into one-dimensional information to amplifier (Japan Patent No. 5,133,073). Japan Patent Office. https://patents.google.com/patent/JP5133073B2/ja

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30.

Wang, M., Zhang, Z., Cheng, Y., & Nepal, S. (2020). DRAMDig: A knowledge-assisted tool to uncover DRAM address mapping. In *2020 57th ACM/IEEE Design Automation Conference (DAC)* (pp. 1-6). IEEE.