

NITRO: 3D NAND Flash-Based In-Storage LLM Computing with Enhanced Activation Dataflow

Sanghun Shin

Dept. of Electronic Engineering
Sogang University
Seoul, Republic of Korea
sanghun@sogang.ac.kr

Gisan Ji

Dept. of Electronic Engineering
Sogang University
Seoul, Republic of Korea
gisANJI@sogang.ac.kr

Sungju Ryu*

Dept. of System Semiconductor Engineering
Sogang University
Seoul, Republic of Korea
sungju@sogang.ac.kr

Abstract—In-storage computing (ISC) has emerged as a next-generation memory architecture to relieve the data movement bottleneck between host processors and memory systems. While recent NAND flash-based processing-in-memory works leverage the high density of 3D NAND flash for deep neural networks, they primarily focus on optimizing computation inside the NAND array. Consequently, these approaches often fail to address the critical latency overhead associated with managing intermediate activation data. To overcome such a limitation, we propose a heterogeneous NAND flash-based ISC architecture with enhanced activation buffering. By buffering intermediate values in a DRAM subsystem rather than programming them into the NAND flash array, our approach effectively mitigates the high programming latency penalties. We also introduce a distributed dataflow scheme that maximizes computational parallelism through optimized plane- and bank-level data mapping. The results show that our proposed architecture achieves performance improvements, reducing inference latency by up to 86% compared to the baseline.

Index Terms—3D NAND flash memory, in-storage computing, large language models, DRAM buffering, pSLC buffering.

I. INTRODUCTION

The exponential scaling of Transformer-based large language models (LLMs) has introduced a severe memory bottleneck, often referred to as the *memory wall*. To address this challenge, in-storage computing (ISC) has emerged as a promising solution by integrating computational units directly inside the memory devices. Among various platforms, 3D NAND flash memory is an attractive candidate for LLM storage and accelerator due to its exceptional storage density, non-volatility, and low cost-per-bit. While prior NAND flash-based processing-in-memory (NAND-PIM) works have successfully demonstrated in-situ matrix-vector multiplication (MVM) [1]–[3], they primarily focus on optimizing the MVM kernel itself, overlooking the management of intermediate activation data. In Transformers, writing intermediate results (e.g., the *Query* matrix from Q_{proj} operation) back to conventional triple-level cell (TLC) NAND flash before subsequent computations incurs severe latency. These delays cause pipeline stalls that negate the performance benefits of in-storage processing.

To overcome this fundamental limitation, we propose NITRO, a 3D NAND flash-based architecture optimized for LLM inference. NITRO introduces a heterogeneous PIM strategy that

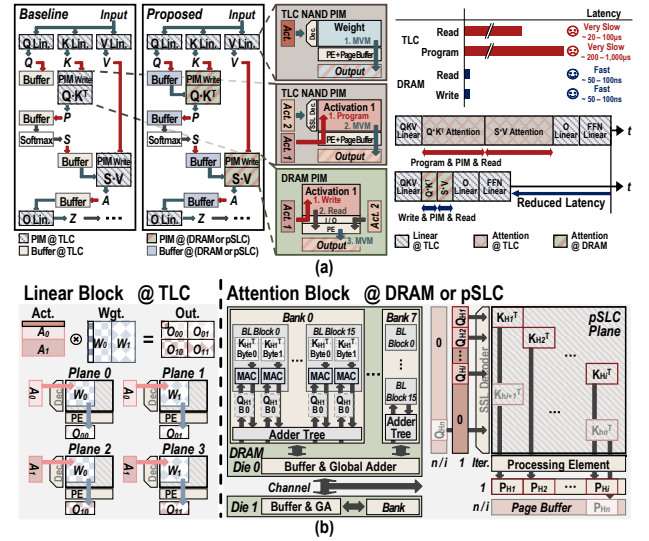


Fig. 1. Overview of the proposed NITRO architecture. (a) Comparison of activation handling between the baseline and our design. (b) Our distributed dataflow to increase throughput on linear and attention computations.

allocates computational tasks based on the LLM’s data access patterns. Specifically, weight-intensive linear computations are assigned to high-density TLC NAND-PIM arrays, while activation-intensive attention computations are offloaded to a high-speed DRAM-PIM module embedded inside the SSD. Furthermore, NITRO includes a distributed dataflow mechanism designed to maximize throughput by mapping tensor operations across the parallel planes of the NAND Flash dies. By combining computational offloading with optimized data parallelism, NITRO effectively reduces the bottlenecks caused by intermediate data management.

II. PROPOSED NITRO ARCHITECTURE

A. Activation DRAM Buffering

A critical challenge in accelerating LLM inference in NAND flash is the high latency associated with programming intermediate activation data (Fig. 1a). This bottleneck is particularly severe in self-attention computations, which require multiplications between dynamically generated intermediate results. In conventional TLC NAND-PIM architectures, the pipeline is forced to stall for hundreds of microseconds while programming intermediate matrices (e.g., *Key* matrix) into the NAND flash cells before subsequent operations can proceed.

*Corresponding Author

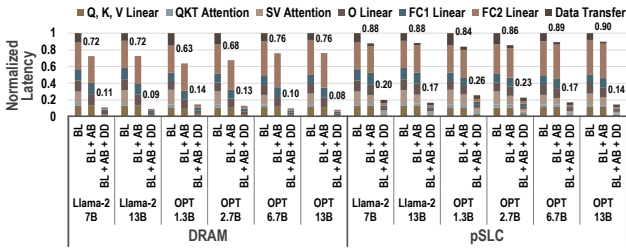


Fig. 2. Normalized inference latency of linear layer at TLC and attention layer at DRAM/pSLC with baseline as 3D-FPIM. BL: Baseline. AB: Activation buffering. DD: Distributed dataflow.

NITRO mitigates this bottleneck by partitioning computational tasks based on their data access patterns. Linear computations are assigned to the high-density TLC NAND-PIM, which involves large read-only weight matrices. This leverages the massive and cost-effective storage capacity of NAND flash. Conversely, self-attention computations with dynamic intermediate activations are offloaded to a high-speed DRAM-PIM subsystem integrated inside the SSD. By bypassing slow NAND flash programming steps for dynamic data, NITRO eliminates the associated pipeline stalls and reduces execution latency. However, in some cases, SSDs are configured without DRAM for cost reasons. NITRO can alternatively utilize NAND blocks operating in pseudo-SLC (pSLC) mode, which offers lower latency and higher endurance compared to standard TLC.

B. Distributed Dataflow

Mapping large tensors linearly across pages without considering the hardware architecture often leads to resource underutilization. To address this, NITRO uses plane-level parallelism for linear layer computations, as illustrated in Fig. 1b (left). The weight matrix (W) is vertically partitioned into sub-matrices, which are then distributed across multiple NAND planes. Concurrently, the input activation matrix (A) is broadcast to planes containing the corresponding weight segments. These arrangements enable the simultaneous execution of MVM operations across all planes. The resulting partial outputs are then concatenated to form the final result, maximizing NAND-PIM resources utilization and enhancing overall throughput.

For self-attention computations inside the DRAM-PIM, NITRO leverages bank-level parallelism by assigning different attention heads to separate DRAM banks (Fig. 1b (center)). In DRAM-less configurations that use pSLC-mode NAND-PIM as a substitute for DRAM, key matrices (K_{H1}, \dots, K_{Hn}) are distributed across different bit-line (BL) groups via a diagonal placement scheme (Fig. 1b (right)). This layout enables input query matrices (Q_{H1}, \dots, Q_{Hn}) to be calculated with multiple key matrices in a single MVM iteration, allowing parallelized multi-head attention computations directly inside the NAND array structure.

III. EVALUATION RESULTS

To evaluate the latency and energy consumption of the proposed NITRO architecture, we designed an in-house simulator based on the NVSIM [4] and 3D-FPIM [2] using the parameters derived from commercial TLC NAND flash [5]. The digital processing elements inside the NAND-PIM were designed using Verilog HDL and synthesized in a 28nm node

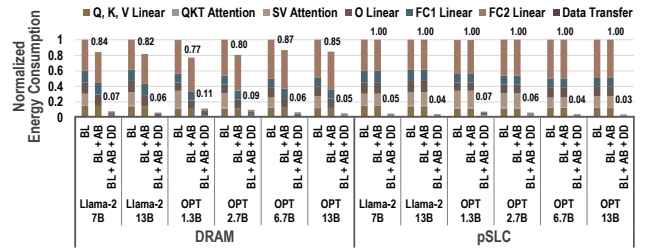


Fig. 3. Normalized inference energy consumption of linear layer at TLC, attention layer at DRAM/pSLC with baseline as 3D-FPIM. BL: Baseline. AB: Activation buffering. DD: Distributed dataflow.

using the Synopsys design compiler. These results were then scaled to a recent 7nm node following the methodology in [6]. For the DRAM-PIM subsystem, the computing behavior and parameters were based on McDRAM [7] and Micron's LPDDR4 specifications [8]. To show the effectiveness of our approach, we compared it with the previous analog NAND-PIM baseline, 3D-FPIM [2], using LLaMa-2 [9] and OPT [10].

As illustrated in Fig. 2, NITRO achieves an average latency reduction of 86% compared to the 3D-FPIM baseline. This speedup is achieved by offloading the attention computations to the high-speed DRAM-PIM and distributed dataflow scheme. This combination effectively reduces the pipeline stalls caused by slow TLC NAND programming while maximizing parallelism for large-scale matrix operations. Furthermore, shifting the frequent and energy-intensive activation writes from NAND flash to low-power DRAM, NITRO reduces average energy consumption by 93% (Fig. 3). In terms of plane area, the proposed NAND-PIM subsystem occupies approximately 46.7mm². We used 16-bit SAR-ADCs for high-accuracy computation, which occupy approximately 38% of the total plane area.

IV. CONCLUSION

This paper presented NITRO, a high-performance NAND flash-based ISC architecture designed to accelerate LLM inference. While conventional NAND flash-only PIM approaches encounter a performance degradation due to the high latency of programming intermediate activation data, NITRO addresses these limitations with a heterogeneous PIM strategy. By selectively offloading attention computations to a high-speed DRAM-PIM subsystem and executing linear computations inside the high-density NAND array, NITRO achieves an optimal balance between storage density and processing speed. Furthermore, NITRO leverages a distributed dataflow scheme to exploit internal parallelism across NAND flash planes and DRAM banks, thereby maximizing computational throughput.

Experimental results show that our proposed architecture reduces inference latency by up to 86% compared to the previous analog NAND-PIM study.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-02263669, Development of server-level DRAM-stacked PIM solution for accelerating ultra-large AI models, 100%). The EDA tool was supported by the IC Design Education Center(IDECE), Korea.

REFERENCES

- [1] H.-T. Lue, P.-K. Hsu, M.-L. Wei, T.-H. Yeh, P.-Y. Du, W.-C. Chen, K.-C. Wang, and C.-Y. Lu, "Optimal design methods to transform 3d nand flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvcim) accelerator for deep-learning neural networks (dnn)," in *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 38–1.
- [2] H. Lee, M. Kim, D. Min, J. Kim, J. Back, H. Yoo, J.-H. Lee, and J. Kim, "3d-fpim: An extreme energy-efficient dnn acceleration system using 3d nand flash-based in-situ pim unit," in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022, pp. 1359–1376.
- [3] M. Kang, H. Kim, H. Shin, J. Sim, K. Kim, and L.-S. Kim, "S-flash: A nand flash-based deep neural network accelerator exploiting bit-level sparsity," *IEEE Transactions on Computers*, vol. 71, no. 6, pp. 1291–1304, 2021.
- [4] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.
- [5] *Gen2 256Gb TLC 3D NAND Flash*, YMTC Corporation, 2019, client Rev. 0.2. [Online]. Available: <https://www.ymtc.com>
- [6] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of cmos device performance from 180 nm to 7 nm," *Integration*, vol. 58, pp. 74–81, 2017.
- [7] H. Shin, D. Kim, E. Park, S. Park, Y. Park, and S. Yoo, "Mcdram: Low latency and energy-efficient matrix computations in dram," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2613–2622, 2018.
- [8] *LPDDR4X/LPDDR4 SDRAM*, Micron Corporation, 2022, rev. D 3/2023 EN. [Online]. Available: <https://www.micron.com/>
- [9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
- [10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "OPT: Open Pre-Trained Transformer Language Models," *arXiv preprint arXiv:2205.01068*, 2022.