Power Optimization in 3D NAND Flash: A Co-Design Study of Architecture, Circuits, and Algorithms

Ziqi Xie *

Arizona College of Technology, Hebei University of Technology, Tianjin, China

* Corresponding Author Email: xieziqiskrrr@outlook.com

Abstract. As data for storage and calculation continues to explode, 3D NAND flash is a key technology for storing information because of its high density. However, adding more layers of stacks and computer functions has led to the problem of power consumption, restricting further advancement. To solve this kind of challenge, in this paper, we creatively utilized a three-layer collaborative analysis framework at the architecture, circuit, and algorithm levels. The architecture level is like near-memory computing, but it gives me a little taste of what the hardware is doing. Then, have full link EE management and exact circuit control for the circuit to cut hardware cost. In terms of algorithms. Dynamic power—optimize based on intelligent prediction and global scheduling, research and study the internal relationship and principles of promoting and supporting the optimization method at different levels of research, create more collaborative rules in fewer lines, can really decrease hardware leakage, and can also respond to changes in regulations in real time. It supplies basic theoretical aid with specific systemic directions on low energy use, extremely thick 3D NAND memory storage systems, and the associated computation combined systems.

Keywords. 3D NAND flash, power consumption, collaborative analysis, near-memory computing.

1. Introduction

With the rapid growth of big data and artificial intelligence, the global demand for data storage is increasing rapidly. 3D NAND Flash, using vertical stacking technology to overcome the limitations of planar storage, has surpassed 400 layers and single-chip capacities up to 1Tb or more. It has become the primary storage form of SSDs and storage systems in data centers [1]. But if WL pitch is going down, parasitics are getting worse, and now we are also stacking them at higher densities, and things like RDI and PCI matter a huge tech more [2]. Data maintenance: To preserve the data, do data verification and data correction. This will become increasingly hard and require much effort [3]. Also, 3DNAND heads toward storage plus computing, not just on data storage alone. New apps want 3DNAND to calculate something for them, for example, the near memory computing and AI inference acceleration so having to send the computing and storage units around makes the power use problem even worse [4].

Power consumption is currently the main bottleneck for 3D NAND. Storage apps have used more power, making it harder to dissipate heat and keep the battery lasting as long, and the multichip parallel operations are no longer possible [5]. When that's what we're putting computation into it, and it uses too much power, it could just negate all the benefits of doing near-memory compute. So, it kills any chance of making near memory ever practical [6]. So, improving 3D NAND power efficiency is not just a single-point improvement anymore; now we need to do some re-architecting and redesign circuits and algorithms across all these layers [7].

Review this article: Optimizing the design at three levels: architecture, circuit, and algorithms. In contrast, an inductive summary of technical elements in each stage reflects the impact of interrelated multi-stage optimization methods for the overall power Optimization System. The study aims to clarify the primary sources and bottleneck mechanisms of power consumption in 3D NAND, covering core storage functional modules, hardware circuits, and emerging computing scenarios. It further organizes critical technological breakthroughs in architectural innovation, circuit refinement, and algorithmic adaptation, elucidating the power optimization principles of each technique. Finally, the research

summarizes the principal pathways for architecture-circuit-algorithm collaborative optimization, providing technical reference and trend guidance for the low-power design of 3D NAND.

2. Sources of Power Consumption

Before delving into specific power optimization techniques, it is essential to develop a clear understanding of the power composition in 3D NAND. As illustrated in Fig. 1, total power consumption primarily originates from several key components, including core array operations, peripheral circuits, and interface I/O. All optimization approaches discussed in this work aim to target these critical segments and effectively reduce power dissipation in each domain.

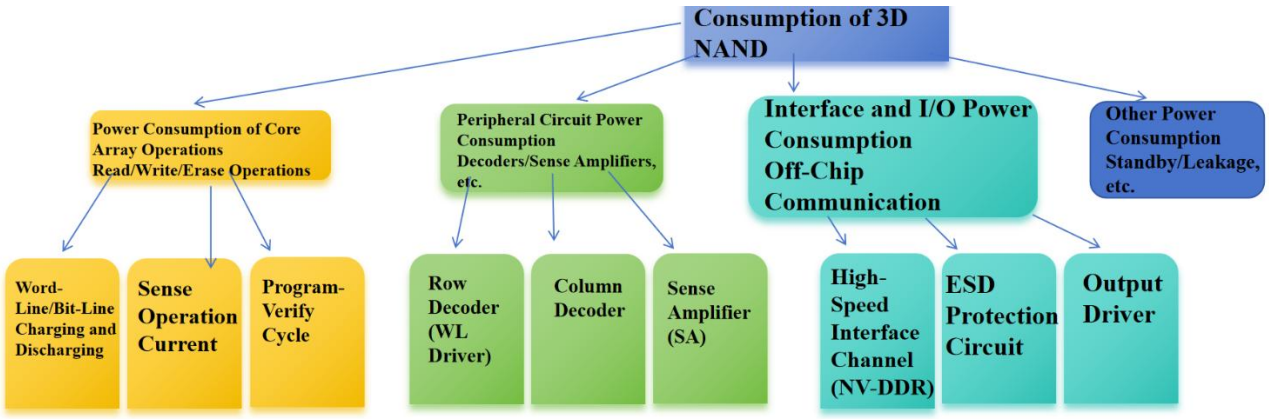

Figure 1. Overview of Power Consumption Sources in 3D NAND

3. Architectural-Level Power Optimization: Energy Efficiency Revolution from Storage to Computation

Architectural design serves as a high-level strategy for power optimization in 3D NAND. Restructuring data flow paths and integrating storage with computing functions reduces redundant operations and data movement, achieving significant power savings.

3.1. Architectural Optimization of Storage Core Modules

Review this article: Three-layer optimization at the architecture level, circuit level, and algorithm level, respectively, and an inductive summary of technical elements in each stage. Comparing these three stages will give the impact of the interrelated multi-stage optimization methods on the complete power optimization system [3]. This kind of design will store the LLR value of the suspected error bits in SRAM, thus reducing the memory access width and regularity of this kind of design. Power Save is achieved on average at 22.4%, for the first time, with the same error correction capability and area efficiency. This provides a 'selective data processing' perspective for ECC design.

3.2. Power Efficiency Innovation in Near-Memory Computing Architecture

Compute to (into) storage, looks like significant power efficiency breaks not in CPU/Soc but inside 3DNAND, take the compute and go directly to 3DNAND and not moving much data around, 3DNOR - Lue et al HB dCIM arch in 3DNOR, IMS in 3DFINN 3DNAND [4]. The Hb-CIM module will execute the multiplication operation for the matrix inside RAM, which uses 3D NOR's fast random access to reduce the I/O power. Based on a sensor, IMSA is a hardware accelerator approximating nearest neighbours. It works in parallel, and the data transmission happens at Terabytes/second (TBit/s), which means they're great for running an LLM or pulling big piles of data.

3.3. Architectural Acceleration for Bulk Operations

To solve the problem of performing bulk Bit-Wise operations for a Database, a Graph processing app, Park et al. have proposed Flash Cosmos, a new architecture for computing on the native 3D NAND array [8]. The architecture will use MWS to sense many words together: Intra-block MWS will sense bitwise AND, and Inter-block MWS will sense bitwise OR. The ESP needs to work correctly throughout these calculations. Experiments have shown that it has been increased between 3.5X to 32X and the energy consumption of 3.3x to 95x, respectively, compared to general processors that show in near memory computing schemes before, so significant power efficiencies can be gained for bulk operations with architectural optimization.

Architecture-level optimization comes from the development of in-memory dynamically programmable computing. So these architectures would be able to allocate compute hardware on demand and do smart pre-processing on data. Hence, there is much reduced energy consumption as a result of moving that data. But almost all current studies have been focused on individual patterns of computation. This is general and flexible, and most importantly, dynamic reconfiguration means extra time overhead and design overhead. Future work should explore lightweight hardware virtualization strategies to reduce reconfiguration costs and promote cross-layer co-design standards that harmonize architectural innovation with software compatibility. These efforts will help develop more universal and energy-efficient solutions for memory-computing integration.

As systematically compared in Table I, three primary technical directions in architectural-level power optimization are summarized and analyzed. The table categorizes representative research—including ECC decoder optimization, near-memory computing architectures, and bulk operation acceleration—across four dimensions: optimization direction, key innovation, identified limitations, and underlying mechanism. Clearly highlights each of these architectural strategies, taking advantage of huge ideas such as selecting data, computing locally, and reusing Hardware, etc., to achieve massive improvements in performance with respect to previous work, while at the same time cutting the power bill. And this structured comparison forms an intuitive, systematic reference for architectural pathways.

Table 1. Architectural-Level Power Optimization

Optimization Direction	Core Innovation	Drawbacks and Limitations	Optimization Mechanism
ECC Decoder Optimization [3]	Hamming product codes & Chase-Pyndiah algorithm	Error location sensitivity; Extra control logic	Selective storage of error-suspected bit LLRs to reduce SRAM accesses
Near-Memory Computing Architecture [4]	HB dCIM (GEMV) & IMS (Search)	Limited flexibility; SW compatibility challenges	Compute Localization: Eliminates Data Movement to CPU/DRAM
Bulk Operation Acceleration [8]	Multi-Wordline Sensing (MWS)	Structure-dependent; Limited to bit-wise operations	Native NAND array logic (AND/OR)

4. Circuit-Level Power Optimization: Precise Control of Hardware Energy Efficiency

Circuit design serves as the hardware foundation for power optimization in 3D NAND. By suppressing noise interference, improving the interface transmission, and improving driving efficiency, it can avoid loss and energy that can be used again when working with hardware equipment. As we have more and more stacked layers and even higher and faster interface speeds, parasitics and signal problems are just becoming more and more complicated at the circuit level.

4.1. Noise Suppression and Reliability Circuit Design

In high-density 3D NAND, due to the stacking up of the wordlines, the parasitic parameters increase, the threshold voltage shifts, and CSL noise is also higher during the read process. And these are the kinds of things that we need extra calibration steps for—and much power. Kim et al. introduced a multi-directional noise reduction circuit: the Offset Canceling Sense Latch (OCSL) decreased the effective offset by 81%, with no repetition for sensing actions; Q-IFR handles noise from nearby wordlines in sets, reducing the bit error rate by 11.2%; the CSL noise tracking method modifies the reference voltage with fluctuations. These methods enhance reliability while avoiding redundant power consumption [2].

4.2. Power Optimization for High-Speed InterfacesPower Optimization for High-Speed Interfaces

The increase in interface speed to 5.6 Gb/s has intensified the trade-off between signal integrity and power consumption. Park et al. proposed a power-isolated low tap-terminated (PI-LTT) interface, which employs an N/N transmission driver structure to reduce channel power consumption by 38% to 82%. A read duty cycle adjuster (RDCA) compensates for impedance nonlinearity. At the same time, an external power-assisted core driver (EPACD) technique introduces a low supply voltage (VPPL), reducing program and read core power consumption by 9.4% and 6.2%, respectively, and also decreasing operational latency [1].

4.3. Energy Efficiency Improvement in I/O Peripheral Devices

Focusing on the area and dynamic power consumption of I/O driver circuits, Spessot et al. introduced a thermally stable low-voltage high-k metal gate (LV HKMG) platform. The HKMG Gate First approach reduces equivalent oxide thickness (EOT) to 1.4-1.5 nm, decreasing driver area by 39% and dynamic power consumption by 8%. The replacement metal gate (RMG) scheme further scales EOT down to 1.1 nm, reduces area by 58%, and lowers power consumption by 15%, making it well-suited for low-power high-speed interfaces [9].

4.4. Innovative Calibration Circuit Design for Multi-Load Scenarios

ZQ calibration becomes error-prone in multi-chip packaging environments when capacitive loading is increased. Lee et al. proposed a reference-voltage-loop ZQ calibration technique that shifts the reference node from the high-load ZQ pin to a low-load internal node within the chip. The method rapidly generates calibration codes using a 2-bit per cycle successive approximation register (SAR) algorithm. Experimental results show that under a 4.7 μF capacitive load, voltage error is controlled within 13 mV. At the same time, power consumption remains as low as 8.2mW, significantly improving calibration energy efficiency and stability in multi-load scenarios [10].

Circuit-level optimization should aim for full-link energy efficiency co-management, leveraging asynchronous circuit design and adaptive biasing to achieve dynamic power control. However, asynchronous design faces compatibility issues with conventional EDA toolflows, and intelligent sensing modules may introduce new bottlenecks in accuracy and power overhead. Furthermore, many techniques rely on specific advanced process nodes. Future efforts should focus on developing EDA toolchains that support asynchronous design, creating low-overhead and highly robust sensing and control algorithms, and embedding energy-optimized IP in early-stage process node definitions to facilitate process-circuit co-optimization.

5. Algorithm-Level Power Optimization: Strategic Adaptation for Energy Efficiency Gains

Algorithmic strategies dynamically adapt to the reliability characteristics of 3D NAND and application requirements to reduce unnecessary hardware operations and redundant energy consumption. As a flexible means for power improvement, algorithm-level optimizations must collaborate deeply with architectural and circuit technologies to maximize overall system efficiency.

5.1. Power Management for Open-Block Read Operations, During open-block read operations, the high conduction

Current of unprogrammed wordlines significantly increases the integrated current (ICC), particularly under low temperatures, where it can be 50% higher than in closed blocks, leading to voltage drops and additional power consumption. Chen et al. proposed a block status detection method based on the current valley during WL ramp, which dynamically adjusts the bit-line voltage (VBL) and read voltage (Vread) to suppress current overdrive [5]. This strategy substantially reduces peak and average power consumption during open-block reads while mitigating read disturbance, providing a new approach for power management in memory systems.

5.2. Reliability Algorithms in Computing Scenarios

In 3D NAND-based computing-in-memory (nvCIM), threshold voltage shifts caused by device aging and read disturbance introduce computational errors and increase calibration overhead. Hsu et al. proposed a block-level self-calibration algorithm that extracts the offset ratio during the calibration phase and dynamically compensates computation results during inference. A neutral block is utilized for real-time aging monitoring, enabling online calibration [6]. This approach maintains accuracy even under 10G read stress, avoiding the energy consumption associated with repeated computations and effectively balancing reliability and power consumption.

5.3. Energy Efficiency Optimization Through Differentiated Error Correction

The choice of error correction code (ECC) scheme significantly impacts read power consumption: low-code-rate LDPC offers strong error correction but high decoding power, while high-code-rate LDPC may lead to frequent read retries. Song et al. proposed a differentiated ECC (DECC) strategy that dynamically selects ECC schemes based on data access patterns. Hot data employs low-code-rate LDPC (code rate 0.8) to reduce retries with strong error correction capability. In contrast, other data uses high-code-rate LDPC (code rate 0.9) to control redundancy overhead. Combined with hot data identification and background re-encoding mechanisms, this approach achieves a 41.7% improvement in IOPS and a 32% reduction in energy consumption under Zipf-like workloads, demonstrating the energy efficiency value of algorithmic adaptation [7].

Algorithm-level optimization aims to shift from passive adaptation to active control, leveraging lightweight AI agents and global energy efficiency schedulers to improve system-wide power. However, AI agents' training and inference overhead may offset the energy savings achieved, and global scheduling faces challenges in acquiring and integrating large amounts of telemetry data. It is advisable to adhere to low-overhead design principles, employ unsupervised learning to reduce dependency on labeled data, and define standardized cross-layer state interfaces. These steps can provide efficient and reliable data support for algorithms, ensuring the practical implementation of intelligent power management.

As systematically summarized in Table II, algorithm-level power optimization strategies are categorized across three dimensions: optimization scenario, core challenge, and algorithmic strategy with corresponding effectiveness. The table examines three typical high-power consumption scenarios—open-block reading, computational reliability, and data retrieval—detailing their primary challenges, algorithmic solutions, and actual energy efficiency improvements. It highlights how algorithms are a flexible strategy to dynamically address diverse power optimization needs, achieving

significant system-level power reduction with minimal hardware investment. This effectively outlines the application context and outcomes of algorithm-level optimization.

Table 2. Summary of Algorithm-Level Optimization Scenarios and Strategies

Optimization Scenario	Core Challenge	Algorithm Strategy
Open-Block Read [5]	High current in unprogrammed wordlines, high peak power	Dynamically adjust VBL/Vread via status detection
Computational Reliability [6]	Computation errors from aging/read disturbance	Online self-calibration with aging monitoring
Data Read [7]	ECC strength vs. decoding power trade-off	Dynamic ECC selection by data hotness

6. Conclusion

This study systematically investigates the power optimization challenges in 3D NAND Flash, presenting for the first time a tri-layer collaborative analysis spanning architecture, circuits, and algorithms to dissect the bottlenecks of power consumption and elucidate the mechanisms of cross-layer synergistic optimization. The main contributions of this work include the following three aspects:

This research moves beyond conventional localized optimization approaches at the architectural level by proposing a dynamically reconfigurable "compute-defined storage" paradigm. It demonstrates that by establishing a heterogeneity-aware memory-computing integration architecture and intelligent data pre-processing mechanisms, data flow paths can be restructured at the source, thereby significantly reducing data migration—a major contributor to power consumption. This provides a systematic solution for achieving high energy efficiency in memory-computation integrated systems.

At the circuit level, this study demonstrates that power optimization should evolve from individual module improvements to collaborative energy efficiency management across the entire link. The paper confirms that adopting an event-driven asynchronous circuit architecture and real-time sensing-based intelligent biasing techniques enables precise reduction of inherent hardware energy consumption and supports dynamic adaptation. This provides a robust hardware foundation for addressing the complex physical challenges arising from high stacking layers.

From the algorithmic perspective, this research highlights the role of algorithms as global energy efficiency managers. The design of lightweight AI agents and global energy-aware schedulers allows algorithms to transition from passive adaptation to active prediction and guidance, intelligently coordinating hardware resources to achieve an optimal balance among power consumption, performance, and reliability at the system level.

This study is significant because it clearly demonstrates that power optimization in 3D NAND must rely on multi-level collaborative innovation. The synergistic optimization pathways and technical system summarized in this research provide direct guidance for improving the energy efficiency of existing storage devices. Furthermore, they offer a practical and theoretically grounded low-power design paradigm for 3D NAND technology as it advances toward higher density and deeper integration with computing applications.

References

[1] S. S. Park, J. D. Lyu, M. Kim, et al., "A 28Gb/mm² 4XX-Layer 1Tb 3b/cell WF-Bonding 3D-NAND Flash with 5.6Gb/s/pin IOs," in Proc. 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, IEEE, 2025, pp. 504-505.

[2] M. Kim, S. W. Yun, J. Park, et al., "A 1Tb 3b/Cell 8th-Generation 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4Gb/s Interface," in Proc. 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, IEEE, 2022, pp. 136-137.

[3] S. Kim, B. Ahn, B. Jun, et al., "Low Power Decoder Architecture of Product Code for Storage Controller," in Proc. 2022 19th International SoC Design Conference (ISOCC), Gangneung-si, Korea, IEEE, 2022, pp. 324-325.

[4] H. T. Lue, C. H. Hung, K. C. Wang, et al., "Prospects of Computing In or Near Flash Memories," in Proc. 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, IEEE, 2024, Art. no. 10873502.

[5] A. B. Chen, D. I. Moon, Z. Wan, et al., "On the Challenges of Open-Block Reads in 3D NAND," in Proc. 2025 IEEE International Memory Workshop (IMW), Monterey, CA, USA, IEEE, 2025, Art. no. 11026944.

[6] P. K. Hsu, P. Y. Du, C. Lo, et al., "An Approach of 3D NAND Flash Based Nonvolatile Computing-In-Memory (nvCIM) Accelerator for Deep Neural Networks (DNNs) with Calibration and Read Disturb Analysis," in Proc. 2020 IEEE International Memory Workshop (IMW), Dresden, Germany, IEEE, 2020, Art. no. 9108116.

[7] Y. Song, Y. Lv, L. Shi, "DECC: Differential ECC for Read Performance Optimization on High-Density NAND Flash Memory," in Proc. 2023 28th Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan, IEEE, 2023, pp. 104-109.

[8] J. Park, R. Azizi, G. F. Oliveira, et al., "Flash-Cosmos: In-Flash Bulk Bitwise Operations Using Inherent Computation Capability of NAND Flash Memory," in Proc. 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO), Chicago, IL, USA, IEEE, 2022, pp. 937-955.

[9] A. Spessot, S. M. Salahuddin, R. Escobar, et al., "Thermally Stable Packaged Aware LV HKMG Platforms Benchmark to Enable Low Power I/O for Next 3D NAND Generations," in Proc. 2022 IEEE International Memory Workshop (IMW), Dresden, Germany, IEEE, 2022, Art. no. 9779308.

[10] J. H. Lee, J. E. Park, D. H. Shin, et al., "A Reference Voltage Loop Operation Based ZQ Calibration Technique for Multi-Load High-Capacity NAND Flash Memory Interface," IEEE Access, vol. 13, pp. 95563-95573, 2025.