International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 7, Issue 2, PP 266-267, 2026
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V7I2P135

Original Article

FlashAttention and the Co-Evolution of Algorithms and Hardware: From IO-Awareness to Vector Optimization

Smitha Shivashankaraiah
Independent Researcher, USA.

Received On: 25/05/2026 ***Accepted On: 10/06/2026*** ***Published On: 25/06/2026***

Abstract: *FlashAttention has transformed transformer efficiency by solving the memory bottleneck of standard attention. However, its significance extends beyond a single algorithm. This paper argues that the FlashAttention family, from FA1 (2022) to VFA (2026), demonstrates a mandatory co-design loop between algorithms and hardware. Each generation did not simply improve performance; it solved the new bottleneck created by the previous hardware generation. FA1 solved HBM bandwidth. FA2 optimized parallelism for A100. FA3 introduced asynchrony for H100. FA4 targets Blackwell's asymmetric compute. VFA (April 2026) now solves the vector-unit bottleneck. This paper traces this evolution, synthesizes the pattern, and identify cross-layer depth communication as the next frontier, as emerging work such as MoDA (2026) begins to extend co-design beyond the single-layer boundary.*

Keywords: *Flashattention, Hardware-Algorithm Co-Design, Transformer, GPU Architecture, Attention Mechanism, IO-Awareness, Depth Attention.*

1. Introduction

Standard attention computes Attention(Q,K,V) = softmax($QK^T/\sqrt{d}$)V . The naive implementation materializes the N × N attention matrix in HBM (slow GPU memory). For a sequence length of 16K, this matrix exceeds 1GB and grows quadratically. The bottleneck is not arithmetic (FLOPs) but memory bandwidth: moving data between HBM and fast on-chip SRAM dominates runtime [1].

FlashAttention (FA1, 2022) solved this by introducing tiling and online softmax [1]. The full attention matrix never materializes; blocks are processed in fast SRAM, and partial results are combined incrementally. The result: 2–4× faster, 5–20× less memory.

But FA1 was not the end. It was the beginning. This paper traces the evolution from FA1 to FA2, FA3, FA4, and VFA (April 2026), and looks ahead to emerging work such as MoDA (2026) that extends this pattern beyond the single-layer boundary. The thesis is simple: each version solved the bottleneck that the previous hardware generation exposed. This is not coincidence; it is a pattern of hardware-algorithm co-design that the industry must embrace.

2. The Evolution: Four Generations

2.1. FA1 (2022): The HBM Bottleneck

- Hardware context: A100 GPUs (2020) had 40–80GB HBM2E with ~2 TB/s bandwidth, but SRAM was limited to ~20MB per SM. The gap between compute (TFLOPS) and memory bandwidth was growing.
- FA1's solution: Tiling + online softmax. Process attention in blocks that fit in SRAM. Never write the full N × N matrix to HBM.
- Result: 2–4× faster, 20× less peak memory on A100 [1]. IO-awareness was the breakthrough.

2.2. FA2 (2023): The Parallelism Bottleneck

- Hardware context: A100's 108 SMs could handle fine-grained parallelism. But FA1's split-K implementation caused synchronization overhead.
- FA2's solution: Better Warp scheduling, removing split-K overhead. Parallelize across sequence length, not just heads [2].
- Result: 2× faster than FA1, 70% of A100's peak TFLOPs [2].

2.3. FA3 (2024): The Asynchrony Bottleneck

- Hardware context: H100 GPUs (2023) introduced TMA (Tensor Memory Accelerator) for asynchronous memory copies and WGMMA (Warp Group Matrix Multiply-Accumulate) for faster matmul.
- FA3's solution: Asynchronous execution overlapped compute and memory transfers using H100's TMA. Warp-specialization hid memory latency [3].
- Result: 1.5–2× faster than FA2 on H100, reaching 75% peak utilization [3].

2.4. FA4 (2025): The Asymmetric Compute Bottleneck

- Hardware context: Blackwell GPUs (2025) have asymmetric compute: FP8/FP4 faster than

FP16/FP32. Memory bandwidth improved but not as dramatically as compute.

- FA4's solution: Leveraged lower-precision arithmetic (FP8) while maintaining accuracy through careful scaling. Balanced compute and memory with new tiling strategies.
- Result: Significant speedup over FA3 on Blackwell, with improved pipeline efficiency and FP8 throughput [5].

3. The Newest: VFA (April 2026) and the Vector Bottleneck

- Hardware context: On modern GPUs (H100, Blackwell), Tensor Cores handle matrix operations at peak throughput, but non-matrix operations such as the online softmax scaling factors run on slower scalar/vector units, creating a relative bottleneck [4].
- VFA's solution: Vector-Relieved FlashAttention (VFA) reduces vector operations by reordering computations. Instead of computing softmax separately, it fuses vector operations into matrix units where possible, avoiding scalar bottlenecks [4].
- Key insight: The vector bottleneck is not new. But hardware advances (Tensor Cores) made it visible. VFA proved that the pattern continues: new hardware reveals new bottlenecks, algorithms adapt.

4. The Co-Design Pattern

The FlashAttention family reveals a clear pattern:

Table 1: The FlashAttention hardware-algorithm co-design pattern across generations

Generation	Hardware	Bottleneck	Solution
FA1 (2022)	A100	HBM bandwidth	Tiling + online softmax
FA2 (2023)	A100	Parallelism overhead	Warp scheduling
FA3 (2024)	H100	Synchronous memory	Asynchrony (TMA)
FA4 (2025)	Blackwell	Asymmetric compute	Lower precision + tiling
VFA (2026)	H100/Blackwell	Vector unit ops	Vector-relieved reordering

The pattern is not accidental. Each version did not simply "improve performance." It solved the new bottleneck created by hardware evolution. This is hardware-algorithm co-design, a loop, not a one-time optimization. The FA series and VFA independently discovered that optimizing for today's hardware is insufficient. Algorithms must be designed to evolve with hardware.

5. Future Directions: The Depth Bottleneck

If the pattern holds, what is the next bottleneck?

Recent work suggests depth, specifically layer-to-layer communication. Standard attention optimizes within a single layer but discards the keys and values after each layer's computation. Later layers receive only the blended output, not the richer per-token information that was computed. Mixture-of-Depths Attention (MoDA, March 2026) addresses this by allowing each layer to attend to keys and values from preceding layers at the same token position [6]. Crucially, MoDA achieves 97.3% of FlashAttention-2's efficiency at 64K sequence length, a negligible 2.7% speed cost, while improving model accuracy through richer cross-layer information flow.

The prediction here: just as FA1 through VFA each solved the bottleneck exposed by the previous hardware generation, the next frontier is the depth dimension, specifically cross-layer redundancy that current attention ignores. The pattern of co-evolution continues.

6. Conclusion

FlashAttention is not just an algorithm. It is a case study in hardware-algorithm co-design. From FA1 to VFA, each generation solved the bottleneck that the previous hardware generation exposed. This pattern, from HBM bandwidth to parallelism to asynchrony to asymmetric compute to vector units, demonstrates that attention optimization is a moving target. Emerging work on depth-aware attention such as MoDA suggests the next frontier lies in cross-layer communication, extending the co-design pattern beyond sequence-level optimization. For engineers and researchers, the lesson is clear: do not optimize for today's hardware. Design algorithms that can co-evolve with hardware trends. The FlashAttention family provides the blueprint.

Conflicts of Interest

The author declares that there is no conflict of interest concerning the publication of this paper.

References

[1] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," NeurIPS, 2022.

[2] T. Dao, "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning," arXiv:2307.08691, 2023.

[3] T. Dao and Others, "FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low Precision," arXiv:2407.08608, 2024.

[4] Yupeng Sun, Yanzhao Li, Zhiqiang Zou, Bai Du, Zhiyuan Zhang, Hui Dong, Gaoyige Fan, and Hui Wang, "VFA: Relieving Vector Operations in Flash Attention with Global Maximum Pre-computation," arXiv:2604.12798, Huawei Technologies, 2026 (April).

[5] T. Zadouri, M. Hoehnerbach, J. Shah, T. Liu, V. Thakkar, and T. Dao, "FlashAttention-4: Algorithm and Kernel Pipelining Co-Design for Asymmetric Hardware Scaling," arXiv:2603.05451, 2026.

[6] B. Zeng et al., "Mixture-of-Depths Attention," arXiv:2603.15619, Shanghai Jiao Tong University, March 2026