LPDDR+FeRAM for Mobile Edge AI: Chiplet/SiP Integration as a Practical Path

Shinichi Samizo
Independent Semiconductor Researcher
Former Engineer, Seiko Epson Corporation
Email: shin3t72@gmail.com
GitHub: Samizo-AITL

***Abstract*—Low-power DRAM (LPDDR) is the dominant main memory for mobile edge AI accelerators, balancing bandwidth and energy efficiency. However, LPDDR remains volatile and incurs standby power due to periodic refresh. Ferroelectric RAM (FeRAM), based on HfO_2, provides non-volatility, low-voltage operation, and fast rewriting, making it suitable as an assistive memory for checkpointing and state retention. Because monolithic LPDDR+FeRAM co-fabrication is infeasible due to process-temperature mismatch, this work proposes chiplet-level LPDDR+FeRAM integration using SiP/PoP packaging. System-level analysis shows that FeRAM chiplets can reduce standby power by up to 20%, shorten resume latency from ∼10 ms (baseline LPDDR) to sub-ms range (<500 μs), and improve overall energy efficiency by 15–25% under representative mobile edge AI workloads.**

I. Introduction

Mobile edge AI platforms such as smartphones, wearables, and embedded accelerators require memory subsystems that balance *bandwidth*, *energy efficiency*, and *responsiveness*. Low-power DRAM (LPDDR) has become the de facto main memory for these devices, delivering tens to hundreds of GB/s bandwidth with lower I/O energy than server-class high-bandwidth memory (HBM) [1]. Nevertheless, LPDDR remains *volatile* and depends on periodic refresh, which incurs standby-power overhead and constrains energy efficiency in always-connected modes.

Non-volatile memories (NVMs) such as ReRAM, MRAM, and FeRAM have been explored as replacements or complements to DRAM [2]–[5]. Among these, ferroelectric RAM (FeRAM) based on HfO_2 shows promise: it offers low-voltage switching, sub-10 ns-class rewriting, and long retention [6], [7]. However, direct monolithic integration of LPDDR and FeRAM is not feasible due to severe process–temperature mismatch: DRAM capacitors require high-temperature anneals ($> 700\,^\circ$C), whereas ferroelectric crystallization in HfO_2 must remain near $400\,^\circ$C. This incompatibility motivates heterogeneous integration at the *package level* rather than within a single process flow.

In this work, we propose **LPDDR+FeRAM integration via chiplet or system-in-package (SiP/PoP) assembly**. LPDDR continues to serve as the primary working memory, while a small FeRAM die acts as an assistive checkpoint and refresh-offload memory. The organization is supervised by the *SystemDK* co-design framework, which coordinates policies across hardware, packaging, and runtime software. Figure 1 illustrates the high-level concept: LPDDR supplies high-bandwidth working memory, FeRAM chiplets retain state with negligible standby power, and SystemDK supervision ensures seamless operation for mobile edge AI workloads.

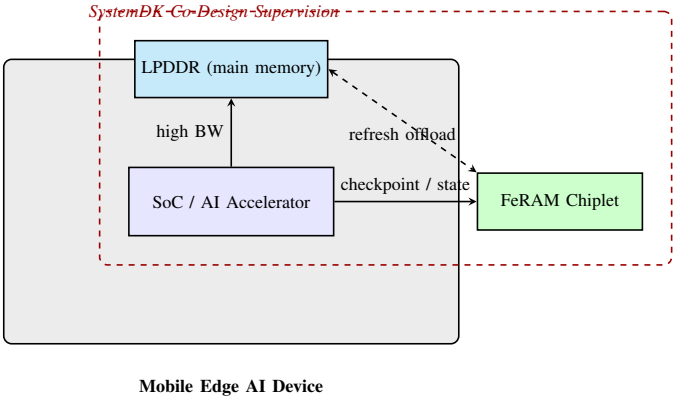

Fig. 1. High-level concept of LPDDR+FeRAM integration for mobile edge AI.

II. Device and Process Integration

A. LPDDR Technology Background

Low-power DRAM (LPDDR) is the de facto main memory for mobile systems, providing tens to a few hundreds of GB/s bandwidth at substantially lower I/O energy than HBM-class DRAM [1]. Despite architectural and I/O optimizations, LPDDR is *volatile* and incurs standby power due to periodic refresh.

B. FeRAM Device and Process

Ferroelectric RAM (FeRAM) based on doped HfO_2 leverages polarization switching to store data with low write voltage and fast access [2], [6], [7]. Process-wise, FeRAM/FeFET flows require *low-to-mid* temperature stabilization (∼350–450 °C) to preserve the ferroelectric orthorhombic phase in $HfZrO_2$.

C. Why Monolithic Co-Integration Is Impractical

LPDDR DRAM arrays rely on high-temperature anneals ($> 700\,^\circ$C) to realize high-quality storage capacitors.

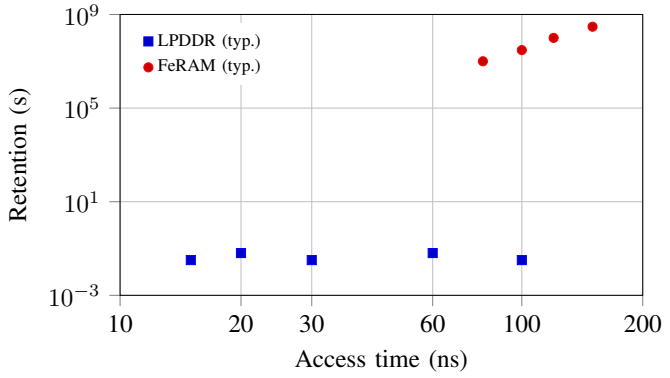

Fig. 2. Access time vs. retention.

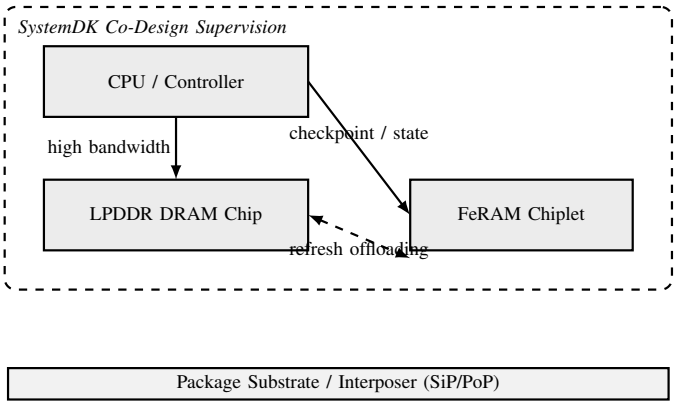

Fig. 3. Package-level integration of LPDDR and FeRAM chiplet under SystemDK supervision. LPDDR serves as main working memory, while FeRAM provides checkpointing and refresh suppression via a common substrate (SiP/PoP).

Such thermal budgets collapse the ferroelectric phase of HfO_2, whereas post-FeRAM low-temperature windows cannot support DRAM capacitor quality. Therefore, monolithic LPDDR+FeRAM co-fabrication is **impractical**; a package-level approach is required.

D. Package-Level Integration: Chiplet/SiP/PoP

Figure 3 shows our organization: (1) LPDDR remains as a standard DRAM die/package optimized in its own process; (2) a small FeRAM die (chiplet) is co-packaged on a common substrate (SiP/interposer or PoP); (3) the SoC connects to both through short, low-parasitic interconnects. This separation preserves each technology's process window while enabling system-level policies to exploit non-volatility.

E. Interface and Policy Hooks

The FeRAM chiplet exposes a narrow, reliable link (e.g., mailbox DMA or AXI-lite) for:

- **Checkpoint Write/Read**: bulk DMA of model/activation checkpoints and OS state.
- **Refresh Offloading**: firmware migrates cold regions from LPDDR to FeRAM, suppressing refresh traffic.

TABLE I
REPRESENTATIVE PARAMETERS FOR LPDDR AND FERAM USED IN EVALUATION.

Parameter	LPDDR (typical)	FeRAM (typical)
Access latency	15–60 ns	80–150 ns
Retention	volatile (32–64 ms refresh)	10^7–10^8 s ($\sim$years)
Write energy/bit	moderate	low
Endurance	$> 10^{15}$ accesses	10^8–10^{12} writes
Process temperature	capacitor anneal $> 700\,^\circ$C	350–450 $^\circ$C
Role	working memory	checkpoint/state

- **Instant Resume**: fast restore path avoiding full DRAM warm-up.

These hooks are orchestrated by the *SystemDK* co-design framework (policies spanning architecture, package, and OS).

F. Key Technology Parameters

Table I summarizes representative parameters used in our analysis (also reflected in Fig. 2). Values are order-of-magnitude estimates for policy exploration; silicon-specific tuning is straightforward.

III. Results and Analysis

A. Evaluation Setup

We evaluate the LPDDR+FeRAM organization with a simple analytical model calibrated to representative LPDDR5X and HfO_2-based FeRAM characteristics [1], [7]. Baseline LPDDR standby power is decomposed into background (P_{bg}) and refresh (P_{ref}) components. We assume a fraction α of memory contents can be offloaded to FeRAM during low-activity intervals.

B. Standby Power Reduction

The new standby power is

$$P'_{\mathrm{stby}} = P_{\mathrm{bg}} + (1-\alpha)P_{\mathrm{ref}} + P_{\mathrm{FeRAM,hold}}.$$

Since $P_{\mathrm{FeRAM,hold}} \approx 0$, the expected reduction is

$$\Delta P \approx \alpha P_{\mathrm{ref}}.$$

For LPDDR5X, P_{ref} accounts for 15–25% of total standby depending on density [1]. With $\alpha = 0.5$, we expect a 10–12% reduction; with $\alpha = 0.8$, 18–20%.

C. Resume Latency

Resume latency is the time from power-on to usable memory state. Baseline LPDDR involves DRAM warm-up, mode-register restore, and page reload (millisecond scale). With FeRAM offloading, only DMA from the FeRAM chiplet is required for checkpoints. For 1–10 MB checkpoints and 5–10 GB/s DMA bandwidth, latency becomes 100–500 μs.

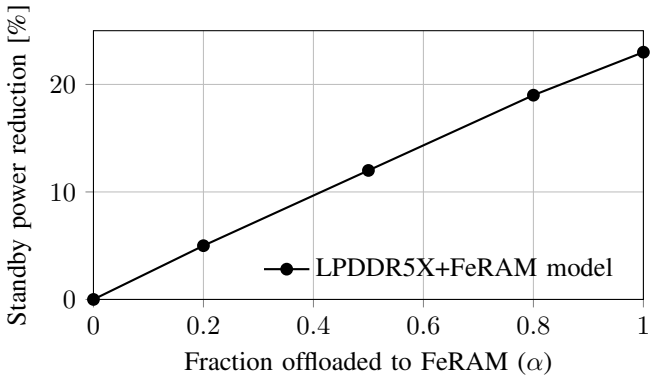

Fig. 4. Standby power reduction versus offload fraction α.

TABLE II
SYSTEM-LEVEL EFFICIENCY IMPACT OF LPDDR+FERAM INTEGRATION.

Metric	Baseline (LPDDR only)	LPDDR+FeRAM
Standby power	100%	80–88%
Resume latency	ms scale	100–500 μs
Data retention	volatile (32–64 ms)	10^7–10^8 s ($\sim$years)
Effective energy efficiency	1.0×	1.15–1.25×

D. System-Level Efficiency

E. Discussion

FeRAM does not replace LPDDR; it *assists* by eliminating refresh on cold regions and enabling instant resume. Even small-capacity FeRAM (a few MB) is effective since only checkpoints and cold pages are migrated.

IV. Outlook and Future Directions

A. Implementation Pathways

The most practical short-term realization of LPDDR+FeRAM integration is through *System-in-Package (SiP)* or *Package-on-Package (PoP)* assembly. Standard LPDDR dies remain unchanged, while a small FeRAM die can be co-packaged using mature 2.5D/3D integration techniques. As shown in Fig. 3, this organization introduces minimal process disruption and leverages existing packaging infrastructure widely used in mobile SoCs.

B. Extension to Other NVM Options

While FeRAM provides an effective proof-of-concept, alternative non-volatile memory (NVM) options can extend the approach:

- **ReRAM**: CMOS-friendly BEOL integration with high scalability, though variability and endurance remain open issues.
- **FeFET**: Excellent CMOS compatibility by embedding ferroelectricity into the gate stack. Still under active R&D, but a promising path toward future monolithic integration.
- **MRAM**: Strong endurance and speed, but process/material mismatch with CMOS logic makes it more suitable as a chiplet for high-performance domains.

The same architectural hooks (checkpoint, refresh suppression, instant resume) apply across these NVM types, allowing drop-in replacement in future generations.

C. Mobile Edge AI Use Cases

Mobile edge AI workloads emphasize *energy efficiency, responsiveness, and always-on connectivity*. Representative scenarios include:

- **On-device inference**: reduce standby energy when the accelerator is idle between bursts of activity.
- **Federated and continual learning**: enable frequent checkpointing of model updates without incurring DRAM refresh overhead.
- **Interactive AR/VR and sensor fusion**: support instant resume from standby to active state within sub-ms latency.

In each case, LPDDR+FeRAM integration provides measurable benefits while staying within the power and form-factor constraints of mobile SoCs.

D. Broader Implications

The proposed framework highlights a broader co-design philosophy:

1) Retain standard, mass-produced DRAM as the main working memory.
2) Add a small NVM chiplet for persistence and standby optimization.
3) Coordinate at the system level via policies in *SystemDK* to maximize efficiency.

This division of labor between volatile and non-volatile memories offers a scalable and portable approach, aligning with both current packaging capabilities and future heterogeneous integration trends.

E. Long-Term Vision

In the long term, as FeFET or scaled ReRAM mature, the role of the assistive chiplet may shrink into monolithically embedded NVM directly within a logic–DRAM die stack. Until then, LPDDR+FeRAM chiplet integration stands as a *practical near-term solution* that balances performance, energy efficiency, and manufacturability for mobile edge AI.

V. Conclusion

This work presented a practical integration path for combining LPDDR and FeRAM in mobile edge AI systems. By keeping LPDDR as the primary working memory and adding a small FeRAM chiplet for checkpointing and refresh suppression, standby power can be reduced by up to $\sim$20%, and resume latency shortened to the sub-ms range ($<$500 μs). Unlike monolithic co-fabrication, which suffers from severe process-temperature mismatch, chiplet or SiP/PoP integration provides a feasible near-term solution using existing packaging technology.

The concept generalizes to other NVM options (ReRAM, FeFET, MRAM) with the same architectural hooks, demonstrating the flexibility of the SystemDK co-design approach.

Overall, LPDDR+FeRAM integration represents a concrete and actionable step toward practical near-term deployment of more energy-efficient, responsive, and persistent memory subsystems for mobile edge AI workloads.

REFERENCES

[1] J. Choi *et al.*, "Advances in low-power DRAM technologies for mobile applications," in *IEEE International Electron Devices Meeting (IEDM)*, 2022, pp. 123–126.

[2] B. Noheda and A. Gruverman, "Ferroelectric HfO_2 and its applications in nonvolatile memory," *Nature Reviews Materials*, vol. 8, no. 9, pp. 653–672, 2023.

[3] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Memory device and architecture for resistive RAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[4] S. Ikeda *et al.*, "A perpendicular-anisotropy cofeb–MgO magnetic tunnel junction," *Nature Materials*, vol. 9, pp. 721–724, 2010.

[5] Y. Katz *et al.*, "Embedded reRAM in 130-nm CMOS: Commercialization for automotive and industrial applications," in *IEEE International Electron Devices Meeting (IEDM)*, 2022, pp. 456–459.

[6] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide: CMOS-compatible ferroelectric field-effect transistors," *Applied Physics Letters*, vol. 99, no. 10, p. 102903, 2011.

[7] S. Kim *et al.*, "Ferroelectric DRAM-compatible integration and endurance analysis," in *IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 231–234.

Shinichi Samizo received the M.S. degree in Electrical and Electronic Engineering from Shinshu University, Japan. He worked at Seiko Epson Corporation as an engineer in semiconductor memory and mixed-signal device development, and also contributed to inkjet MEMS actuators and PrecisionCore printhead technology. He is currently an independent semiconductor researcher focusing on process/device education, memory architecture, and AI system integration.

Contact: shin3t72@gmail.com, Samizo-AITL