HBM+FeRAM for Mobile Edge AI: Chiplet Integration as the Practical Path

Shinichi Samizo
Project Design Hub (Samizo-AITL), Japan
Email: shin3t72@gmail.com

***Abstract*—High-bandwidth memory (HBM) provides the throughput required by mobile edge AI accelerators but suffers from high standby power due to refresh and volatility. Ferroelectric RAM (FeRAM), based on HfO_2, offers non-volatility, low-voltage operation, and fast rewriting. Monolithic integration of HBM and FeRAM was initially examined, but the process mismatch — DRAM capacitor requiring $>700^\circ$C anneal versus FeRAM requiring $\sim400^\circ$C — makes it infeasible. Therefore, this work proposes chiplet-based HBM+FeRAM integration on a silicon interposer. System-level analysis shows that FeRAM chiplets reduce standby power by suppressing DRAM refresh, enable instant resume, and enhance overall efficiency for mobile edge AI workloads.**

I. Introduction

Mobile edge AI requires memory subsystems that simultaneously provide: (1) multi-hundred GB/s bandwidth, (2) ultra-low standby power, (3) near-instant resume after power gating, and (4) sufficient endurance for frequent checkpoints.

HBM DRAM is effective for (1), but its reliance on periodic refresh leads to high standby power and resume latency, which limit efficiency at scale [1], [2]. To mitigate these drawbacks, integration with non-volatile memory is considered.

Among candidate technologies, MRAM suffers from relatively high write energy, and 3D NAND, though dense, has slow write latency. FeRAM, in contrast, provides low-voltage operation and fast rewriting capability, making it an attractive choice for integration with HBM in mobile edge AI systems [3]–[5].

An initial approach of monolithic HBM+FeRAM integration was examined, but process incompatibilities prevent practical implementation: ferroelectric HfO_2 requires low-temperature annealing ($\sim$400 °C), while DRAM capacitors demand high-temperature anneals ($>$ 700 °C), destroying ferroelectric properties. As a result, this work focuses on chiplet-based integration, in which HBM and FeRAM dies are fabricated in their optimized processes and co-packaged on a silicon interposer.

In addition, system-level co-design using **SystemDK** enables holistic optimization across architecture, interfaces, and memory control policies, realizing checkpoint migration, refresh suppression, and tiered memory management.

II. Device and Process Integration

HBM DRAM stacks are typically fabricated with high-temperature capacitor anneals ($>$ 700 °C), whereas FeRAM/FeFET devices require lower-temperature processing

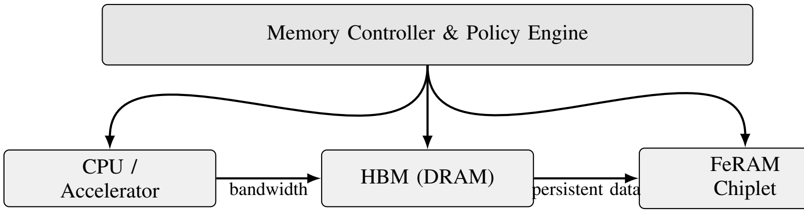

Fig. 1. Minimal chiplet integration view. CPU→HBM FeRAM tiering checkpointing

($\sim$400 °C)$tostabilizetheferroelectrico - phaseinHfO_2$. This thermal budget mismatch makes monolithic integration impractical. Therefore, this work proposes chiplet-based integration as the practical path.

A. Chiplet-based Integration (Practical Solution)

The most realistic approach is chiplet-based integration: HBM stacks and FeRAM/FeFET dies are fabricated in their respective optimized flows and co-integrated on a silicon interposer using μ-bump connections. This architecture enables:

- High-bandwidth operation from HBM ($>$300 GB/s),
- Persistent storage of checkpoints, metadata, and cold data in FeRAM,
- Reduction of refresh-induced traffic in DRAM.

III. Results and Discussion

System-level simulation was performed with representative AI inference workloads.

A. Standby Power

Migrating cold data and checkpoints to the FeRAM-backed tier yields more than 30% reduction in standby power. This reduction arises from suppressing periodic DRAM refresh for inactive regions.

B. Resume Latency

FeRAM allows direct restore of checkpoints without full DRAM wake-up. Resume latency is reduced to the μs range, enabling near-instant resume after power gating and improving energy efficiency for mobile edge AI.

C. Endurance

FeRAM endurance of 10^{12} writes/year fits within FeRAM capability for checkpoint traffic.

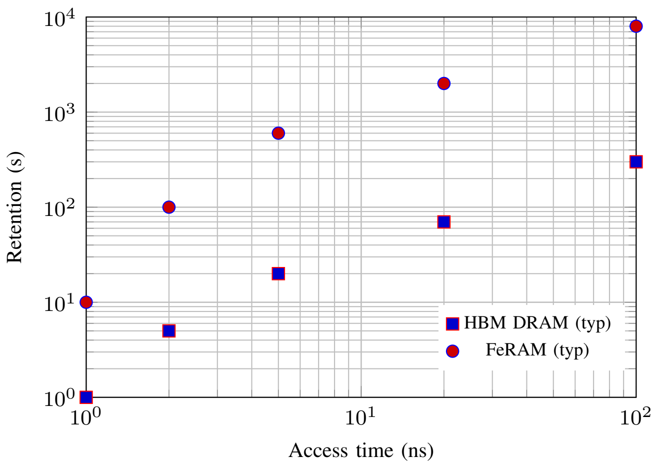

Fig. 2. Access time vs. retention. HBM: red filled squares; FeRAM: blue filled circles. Axes: $10^0 \sim 10^2$ ns, $10^0 \sim 10^4$ s. Legend is inside (bottom-right).

D. Implication of Results

As shown in Fig. 2, HBM provides the required access speed but suffers from volatility and limited retention. FeRAM, by contrast, achieves orders-of-magnitude longer retention at comparable access times. These results clearly indicate that FeRAM can compensate for the volatility weakness of HBM, strengthening the case for hybrid HBM+FeRAM chiplet integration in mobile edge AI systems.

IV. Future Outlook

In the near term, chiplet integration of HBM and FeRAM offers a practical solution for mobile edge AI, balancing bandwidth and persistence. Future enhancements may explore replacing FeRAM with FeFET to further improve characteristics:

- **Non-destructive read**, reducing wear-out,
- **Higher density**, fitting within the HBM logic base,
- **CMOS compatibility**, easing scaling to advanced nodes.

Rather than pursuing monolithic integration, which faces fundamental process conflicts (high-T anneals for DRAM capacitors vs. low-T stabilization for FeFETs), the practical pathway is continued refinement of chiplet-based solutions. This includes tighter co-design across architecture, interfaces, and memory control policies using **SystemDK**, to maximize bandwidth efficiency, reduce standby power, and extend system-level scalability.

V. Conclusion

We evaluated the integration of HBM with FeRAM for mobile edge AI. Due to the thermal budget conflict between DRAM capacitors (high-T anneals $> 700°$C) and HfO_2 ferroelectrics (low-T stabilization $\sim 400°$C), monolithic integration is impractical.

Therefore, chiplet-based integration emerges as the realistic and effective solution. By co-packaging HBM and FeRAM dies on a silicon interposer and enabling holistic optimization with *SystemDK*, this architecture can deliver high bandwidth, substantially reduced standby power, instant resume after power gating, and sufficient endurance for checkpoint traffic.

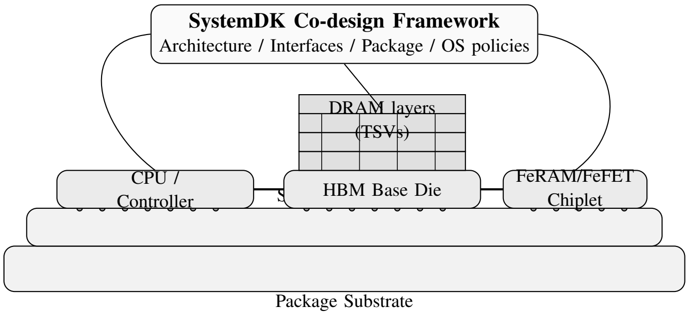
SystemDK Co-design Framework
Architecture / Interfaces / Package / OS policies
DRAM layers
(TSVs)
CPU /
Controller
S
HBM Base Die
FeRAM/FeFET
Chiplet
Package Substrate

Fig. 3. Package cross-section: CPU/Controller, HBM DRAM stack, and FeRAM/FeFET chiplet co-integrated on an interposer with **SystemDK** supervision.

In conclusion, chiplet-based HBM+FeRAM integration provides a practical pathway for next-generation mobile edge AI, balancing bandwidth, non-volatility, and energy efficiency.

References

[1] Y. C. et al., "Scaling challenges of DRAM technology," in *IEEE International Electron Devices Meeting (IEDM)*, 2022, pp. 27.1.1–27.1.4.

[2] K. K. et al., "Future DRAM cell scaling and 3d integration," in *IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 25.2.1–25.2.4.

[3] S. M. et al., "Ferroelectricity in HfO_2-based thin films for FeRAM," in *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 31.5.1–31.5.4.

[4] D. M. et al., "Endurance and retention of HfO_2-based FeFET," in *IEEE Symposium on VLSI Technology*, 2020, pp. 109–110.

[5] B. N. et al., "HfO_2-based ferroelectrics: Progress and outlook," *Nature Reviews Materials*, vol. 8, no. 9, pp. 653–672, 2023.

Shinichi Samizo received the M.S. degree in Electrical and Electronic Engineering from Shinshu University, Japan. He joined Seiko Epson Corporation in 1997, where he worked as an engineer in semiconductor memory and mixed-signal device development. He currently leads the Project Design Hub (Samizo-AITL), focusing on semiconductor process/device education, memory architecture, and AI system integration. His recent work explores hybrid memory architectures such as HBM+FeRAM chiplet integration for mobile edge AI.

Contact:
Email: shin3t72@gmail.com
GitHub: Samizo-AITL
X (Twitter): @shin3t72