Agile-DRAM: Agile Trade-Offs in Memory Capacity, Latency, and Energy for Data Centers

Jaeyoon Lee, Wonyeong Jung, Dongwhee Kim, Daero Kim, Junseung Lee, and Jungrae Kim

Abstract

Data centers frequently face significant memory under-utilization due to factors such as infrastructure overprovisioning, inefficient workload scheduling, and limited server configurations. This paper introduces Agile-DRAM, a novel DRAM architecture that addresses this issue by flexibly converting the under-utilized memory capacity into enhanced latency performance and reduced power consumption. Through minor modifications to the conventional DRAM architecture, Agile- DRAM supports multiple operational modes: low-latency, lowpower, and the default max-capacity mode. Notably, Agile-DRAM facilitates agile transitions between these modes in response to workload fluctuations in data centers at runtime. Evaluation results demonstrate that the low-latency mode can boost singlecore execution speed by up to 25.8% and diminish energy usage by up to 22.4%. Similarly, the low-power mode can reduce DRAM standby and self-refresh power by 31.6% and 85.7%, respectively.

Introduction

Challenges

DRAM high access latency: Despite doubling in density, DRAM has seen only a 16.7% reduction in access latency over two decades.
DRAM high refresh overhead: Power consumption for refreshes increases with capacity, sometimes comprising half of total power use.
Data center memory under-utilization: Data centers face memory under-utilization due to over-provisioning and inefficient workload scheduling, resulting in significant excess capacity and increased costs.

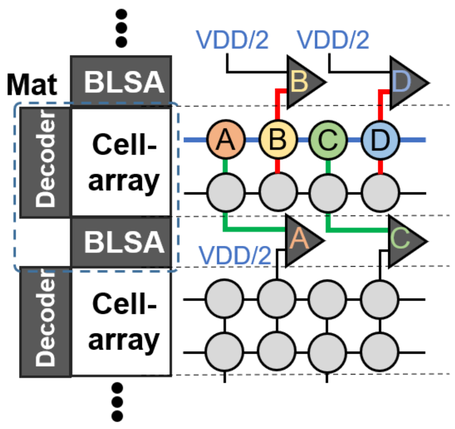
Goal

1. Reduce access latency to enhance system performance.
2. Lower refresh overheads for better energy efficiency and cost reduction.
3. Maintain current DRAM chip size while implementing improvements.
4. Optimize under-utilized memory capacity to boost performance efficiently.
5. Implement agile mode switching for flexible performance and energy management.

Key Idea

Agile-DRAM introduces a ‘mirrored mat' structure centralizing bitline sense amplifiers for efficiency, and supports multiple modes—max-capacity, low-latency, and low-power modes. Also enables agile mode switching, optimizing DRAM performance and energy consumption.

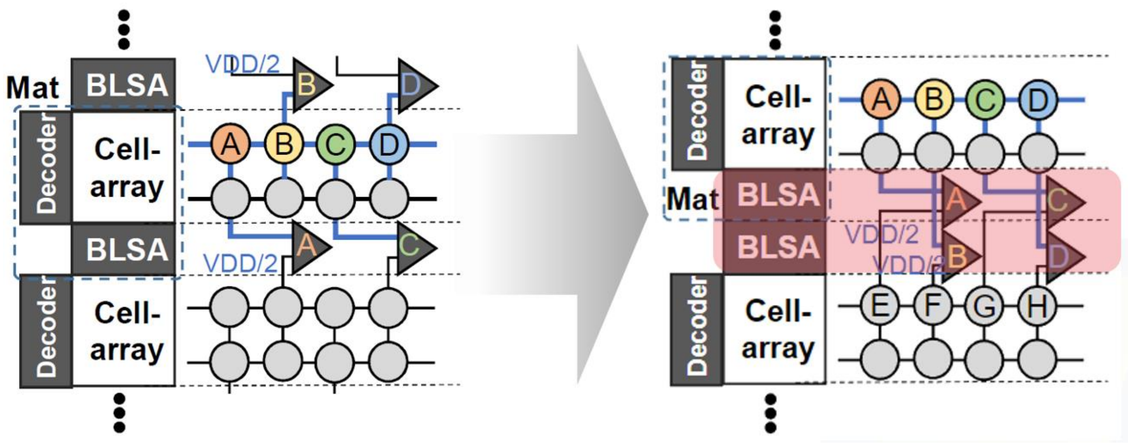
Conventional Structure



The traditional DRAM features an ‘open-bitline structure', where bitlines are divided and connected to separate upper and lower bitline Sense Amplifiers, streamlining signal detection and amplification for compact and efficient memory organization.

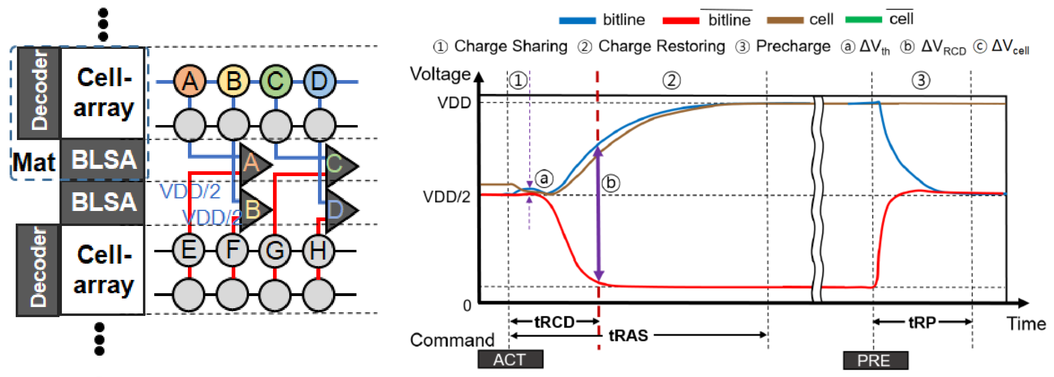
Proposed Method

Agile-DRAM Structure

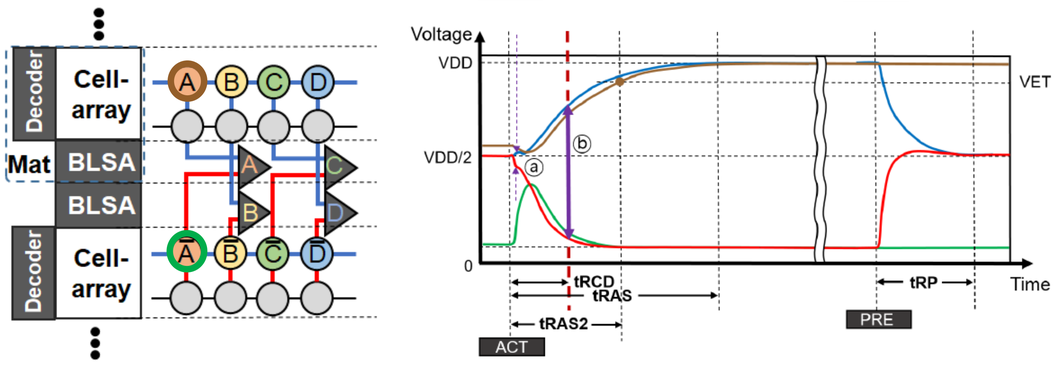


The ‘mirrored mat' represents a leap in DRAM architecture, centralizing bitline sense amplifiers to streamline connections and enhance performance, laying the groundwork for advanced, efficient memory operations within our Agile-DRAM framework.

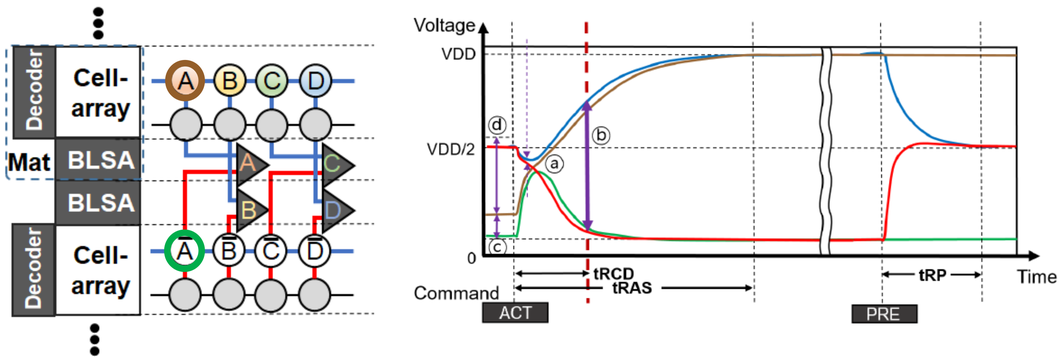
Max-Capacity (MC)



Low-Latency (LL)

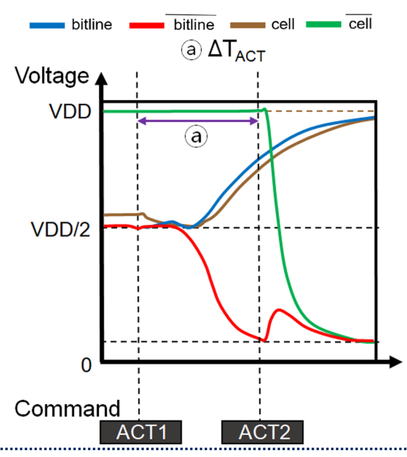


Low-Power (LP)



Agile-DRAM innovates with a max-capacity mode (MC) for full storage use, a low-latency mode (LL) for rapid data sensing through paired cell activation with differential charge encoding, and a low-power mode (LP) enhancing energy efficiency by extending refresh intervals and maintaining reliable sensing with lower cell voltage.

Agile mode Switching



Agile-DRAM introduces agile mode switching, a critical feature enabling dynamic adjustments between max-capacity, low-latency, and low-power modes to optimize memory utilization. Utilizing a dual-activation mechanism with a specific time delay (ΔT_{ACT}) ensures seamless transitions without data integrity loss, offering non-disruptive, efficient memory management for varying workloads in dynamic server environments.

Results

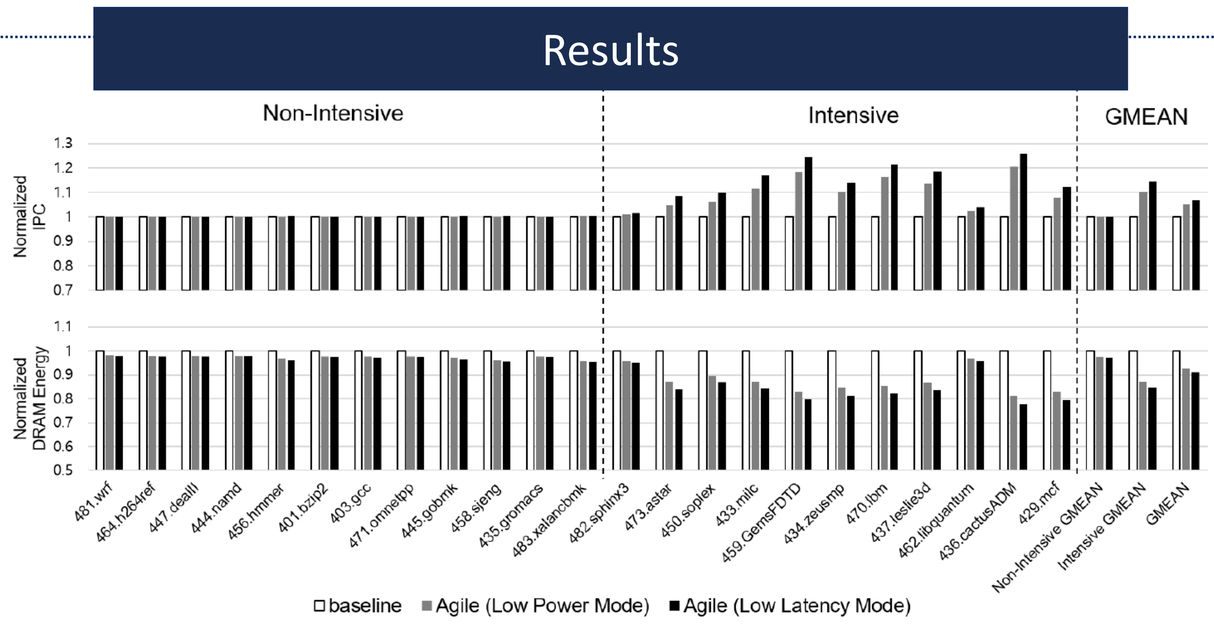


Figure compares IPC / Energy of the baseline open-bitline architecture and other modes of Agile-DRAM. Overall, the LL mode can speed up single-core execution by 6.9% using 23 SPEC CPU 2006 benchmarks. Also, LL mode reduces DRAM energy consumption by 9.0% on average.

Conclusion

Agile-DRAM presents an innovative DRAM architecture addressing the challenge of memory under-utilization in data centers by facilitating transitions between max-capacity, low-latency, and low-power modes. This dynamic adaptability enhances system performance and energy efficiency without service disruption or significant area overhead, offering a cost-effective solution for both data center operators and DRAM vendors to optimize memory use and system operations.

The 30th IEEE International Symposium on High-Performance Computer Architecture’2024 HPCA