CoolDRAM: An Energy-Efficient and Robust DRAM

Nezam Rohbani
School of Computer Science,
Institute for Research in
Fundamental Sciences (IPM)
Tehran, Iran

Mohammad Arman Soleimani
Department of Computer Engineering,
Sharif University of Technology
Tehran, Iran

Hamid Sarbazi-Azad
Department of Computer Engineering,
Sharif University of Technology,
and School of Computer Science, Institute for
Research in Fundamental Sciences (IPM)

***Abstract*—DRAM is the most mature and widely-utilized memory structure as main memory in computing systems. However, energy dissipation and latency of DRAM are two of the most serious limiting factors of this technology. All DRAM main operations are initiated by a Precharge phase, which is time-consuming and power-hungry. This work proposes a novel DRAM cell access scheme that entirely eliminates Precharge phase from DRAM read, write, and refresh operations, with a very slight modification in commodity DRAM structure. The proposed DRAM design, called CoolDRAM, operates using a single extra cell row as reference cells. CoolDRAM reduces energy dissipation by about 34% on average, with a negligible area overhead of about 0.4%. The robustness of CoolDRAM against process variation and environmental noises is 61× and 1.78× of the state-of-the-art, respectively, while maintaining the same power consumption and latency.**

***Index Terms*—DRAM, Power Consumption, Precharge, Sense Amplifier, Data Similarity**

I. Introduction

In a memory hierarchy, DRAM consumes the major portion of energy budget allocated to the memory subsystem [1], [9]. Tiny capacitors, arranged in two-dimensional memory arrays (MATs) are used as storage elements in this memory, which lose their charge over time and during accessing the cells. This characteristic of DRAM has made its cell access complex and energy-hungry [9].

Read, write, and refresh are three major DRAM operations which all of them, more or less, follow the same procedure [6], [17], [22]. Precharge and Activation are the two main micro-operations common to all DRAM operations [1], [9], [16]. These operations need charge and discharge cycles on a large number of bitlines with high parasitic capacitances, which makes them energy-hungry operations.

To reduce power/energy dissipation and latency of DRAM memories, many techniques are presented in the previous works. By utilizing multi-banking feature of DRAMs, some techniques put some cold banks in drowsy state and arrange data to maximize the number of cold banks [4], [6]. Some other techniques propose partitioning memory arrays to activate a smaller number of cells during memory access [7], [12]. DRAM refresh energy reduction is the goal for some other techniques, like selective refreshing and adaptive refresh rate by considering temperature, operating voltage, and application criticality [3], [5], [10], [11], [14]. Some previous works propose optimizations in sensing structures of DRAMs, like overlapping Precharge and Activation by Row-Buffer Decoupling (RBD) [19]. Data encoding to reduce DRAM bus activity [13], [15] and minimizing the number of Activations (row-hit boosting) for data access [20], [21], are the other techniques to reduce DRAM energy consumption. All of the above-mentioned power/energy consumption and delay reduction techniques need Precharge of bitlines to half-VDD (*VDD/2*) before Activating a row of cells. The only technique that eliminates the Precharge phase from DRAM cell access procedure is PF-DRAM [16]. However, this technique imposes significant memory structure modifications and area overhead.

979-8-3503-1175-4/23/$31.00 ©2023 IEEE

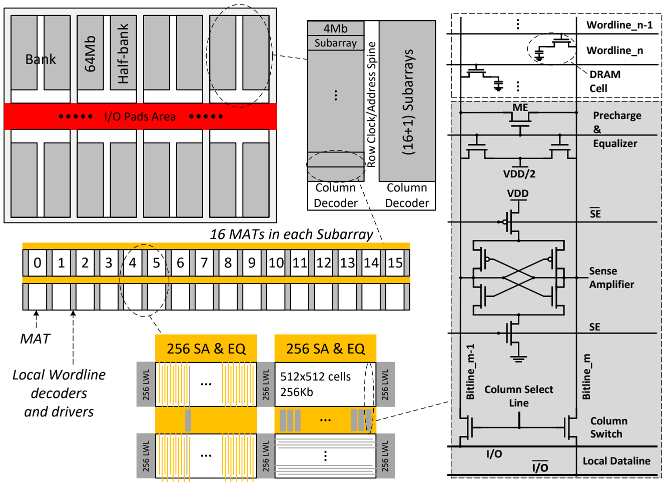

Fig. 1: DRAM chip general structure

This work proposes a DRAM memory design, called CoolDRAM, which not only entirely eliminates the need for the Precharge operation in DRAM during cell access, but also imposes a very slight modification to the commodity DRAM structure. By eliminating the Precharge phase, DRAM energy consumption is reduced by 34% on average. The area overhead of CoolDRAM is less than 0.4% compared with commodity DRAMs. Furthermore, CoolDRAM is 61× and 1.78× more robust against process variation and environmental noises, respectively, compared with the state-of-the-art.

II. Proposed DRAM Design

A. Preliminaries and Motivation

Fig. 1 depicts a DRAM chip structure. A DRAM chip is organized in multiple banks that each bank can serve an operation in parallel with the others. Each bank is composed of two half-banks and each one contains subarrays. Each subarray is composed of multiple (generally 16) MATs, that share adjacent peripherals with each other. In a MAT the cells are arranged in a two-dimensional matrix, composed of rows (wordlines) and columns (bitlines). To access a cell in a MAT, all of the bitlines in the MAT are charged to half-VDD, using precharge and equalizer transistors, in the Precharge phase and then are floated. By Activating a wordline, each cell in the accessed row is connected to the corresponding bitline. Based on the stored charge in the memory cell, *0* or *VDD*, a tiny voltage perturbation (V_S), about ±90 mV, is generated on the corresponding bitline by charge sharing between the cell capacitor and the parasitic capacitor of the floated bitline. This voltage perturbation is detected by differential sense amplifiers connected to the bitlines. To cancel ambient and internal noises on long bitlines for a reliable cell access, differential sensing is mandatory. The sense amplifier compares the voltage of a bitline connected to the target cell with another float bitline in a pair. V_S can be calculated by Eq.(1). C_{BL} and C_{cell} are the parasitic capacitance of a bitline and memory cell, respectively.

$$V_S = \frac{V_{DD}}{2} . \frac{C_{cell}}{C_{BL} + C_{cell}} \tag{1}$$

The most popular sense amplifier in DRAMs is single-ended differential sense amplifier [9]. This sense amplifier is composed of two cross-coupled inverters in a positive feedback to compare bitlines voltage in each pair and detect the accessed cell value. In the Precharge phase, the sense amplifier is disconnected from VDD and GND power rails using SE and $\overline{SE}$ (sense amplifier enable) signals (Fig. 1). In this way, the internal nodes of the sense amplifier are precharged to half-VDD together with the bitlines in each pair.

After developing V_S on bitlines in the Activation phase, and consequently, on the sense amplifier internal nodes, the sense amplifier is activated. Since this initial condition on the sense amplifier is unstable, the sense amplifier node with lower voltage drops to *0*, and the other node rises to *VDD*, promptly.

During cell access (read/write/refresh), by activating one row, a large number of cells (e.g. 64 K cells in DDRx) are connected to the bitlines and their value are determined in the bitline pairs sense amplifiers. Because of the large parasitic capacitance of bitlines, a major portion of DRAM power dissipation is consumed for voltage toggling on the bitlines during Precharge and Activation phases [2], [9], [16].

The dissipated energy on a bitline pair during the Precharge phase can be calculated by Eq.(2). In which, Q_{pre} is the charge stored in the bitline during Precharge operation and β is a constant, determined by the width ratio of precharge transistors to equalizer transistor. By connecting bitlines in each pair to each other, through equalizer transistor (*ME* in Fig. 1), sharing the opposite charges on the bitlines aids the precharge transistors (and half-VDD power source) to precharge the bitlines to *VDD/2*, faster and more efficiently [9]. In commodity DRAM, β is about 0.54 [16].

$$E_{pre} = \beta . Q_{pre} . VDD = \beta . C_{BL} . \frac{VDD}{2} . VDD = \beta . C_{BL} . \frac{VDD}{2}^2 \tag{2}$$

During the Activation phase, the sense amplifier pulls the voltage of one bitline in each pair to *VDD* and pulls that of the other one to *0*, based on the voltage perturbation on the accessed bitline. The dissipated energy in this phase for commodity DRAM (E_{act}) can be calculated by Eq.(3). In which Q_{act} is the charge drawn from the power source during the Activation phase, that is equal to the electric charge needed to charge C_{BL} plus cell capacitor (C_{cell}) to *VDD*.

$$\begin{aligned} E_{act_C} &= Q_{act} . VDD \\ &= \begin{cases} (C_{BL} + C_{cell}) . (\frac{VDD}{2} - V_{S_C}) . VDD & , V_{S_C} > 0 \\ = (C_{BL} + C_{cell}) . (\frac{VDD}{2} - \frac{VDD}{2} . \frac{C_{cell}}{C_{BL} + C_{cell}}) . VDD & \\ & \\ C_{BL} . \frac{VDD}{2} . VDD & , V_{S_C} < 0 \end{cases} \\ &= C_{BL} . \frac{VDD}{2}^2 \end{aligned} \tag{3}$$

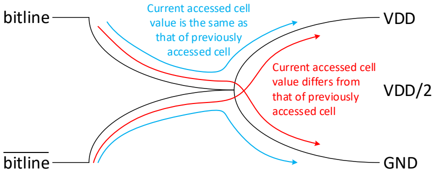
bitline
Current accessed cell value is the same as that of previously accessed cell
VDD
Current accessed cell value differs from that of previously accessed cell
VDD/2
$\overline{\text{bitline}}$
GND

Fig. 2: Bitlines voltage activity in a bitline pair during consecutive Activations

The total dissipated energy for a precharge-activation cycle on a bitline pair, in the case of a row conflict, can be calculated using Eq.(4). This value is constant regardless of the accessed cell value, '0' or '1'.

$$E_{acc_C} = E_{pre} + E_{act_C} = \frac{1}{2}(1 + \beta) . C_{BL} . VDD^2 \tag{4}$$

During the execution of real-world workloads, a very high data similarity (about 90%) exists between the stored data in DRAM main memory [16], [18]. This is because of large number of '0's stored in DRAM memory, narrow-width values, locality of references, data dependency, high-frequently used instructions, and low hamming distance of op-codes. Regarding the very high similarity of stored data in DRAM, it is expected that a large portion of bitlines stabilizes to the same previous value of former Activation. Thus, a useless discharge and charge on each bitline pair (equal to E_{acc_C}) is dissipated with no change on a large number of bitlines states. In other words, during reading the same previous value on a bitline pair, one bitline is discharged to half-VDD (during Precharge phase) and again is charged to the same previous value (in Activation phase), and the other bitline is charged to the half-VDD (during Precharge phase) and is discharged to the *0* (during Activation phase) again. This is while, if the bitlines were not precharged to half-VDD, no voltage swing occurred on bitlines during reading the same value on a bitline pair (reading '0' after '0' or reading '1' after '1'), see blue arrows in Fig. 2.

B. *CoolDRAM*

Precharging bitlines to half-VDD is a power-hungry and time-consuming operation in DRAM cell access, however, differential sense amplifiers need equal voltage on both bitlines in a pair as an *initial point*. The technique presented in PF-DRAM [16], charges or discharges both bitlines in each pair equally to *VDD* or to *0*, during the Activation phase (does not drive bitlines in a pair to opposite values like commodity DRAMs). PF-DRAM utilizes *VDD* or *0* as an initial point on bitlines in each pair. In this condition, if the initial value on bitlines is equal to the accessed cell value, no voltage swing occurs on that bitline pair and thus, overall power dissipation reduces considerably [16]. However, the main drawback of PF-DRAM is that it utilizes an unbalanced sense amplifier to detect equal voltage on bitlines, which makes it vulnerable to transistors' threshold voltage drift. Furthermore, the DRAM vendors are very reluctant to apply any modification to the DRAM structure, especially the sense amplifier which needs to be very accurately engineered to fit inside the narrow space between DRAM bitlines [9].

In this work we present CoolDRAM which not only entirely eliminates the Precharge phase from the DRAM cell access

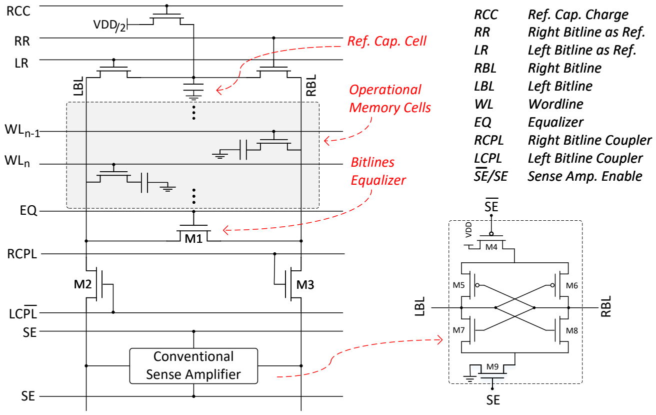

Fig. 3: Proposed CoolDRAM structure

sequence, but also uses exactly the same sense amplifier structure of commodity DRAMs.

Fig. 3 shows the general design of CoolDRAM. CoolDRAM imposes the minimum modifications to the commodity DRAM structure. Its sense amplifier is exactly the same as the sense amplifier utilized in commodity DRAMs and the memory cell array remains untouched. The precharge circuit in CoolDRAM is slightly modified for updating bitlines voltage compared with commodity DRAM. *M1* connects two bitlines in each pair to each other to balance their voltage before Activation phase, just like commodity DRAM memories (*ME* in Fig. 1). *M2* and *M3* are utilized to couple and decouple bitlines from the sense amplifier.

The only added part of CoolDRAM to commodity DRAM is a dual-contact reference cell row to the memory array (shown at top of Fig. 3). The reference cell can be connected to Right Bitline (RBL) or Left Bitline (LBL) in each pair, by activating Right Bitline as Reference (RR) or Left Bitline as Reference (LR) signals, respectively. This cell develops a voltage perturbation on the reference bitline for proper operation of the sense amplifier. Before connecting this cell to a bitline, it is charged to half-VDD by activating Reference Capacitor Charge (RCC) signal, while RR and LR are deactivated. This signal is activated by the activating power source (VPP) [1] of DRAM, 2.5 V in DDR4, to avoid voltage drop on the n-type control transistor. The charge of the reference cell can be performed in parallel with restoring data in accessed cell, thus, no timing overhead is added to the cell access sequence.

In CoolDRAM, during the Activation phase, both bitlines are charged/discharged to *VDD*/*0* and not opposite values. The equal voltage on bitlines during Activation phase eliminates the need for precharging bitlines to half-VDD as the initial point and next cell access can be started right after the previous cell access.

In the first step of cell access, the sense amplifier internal nodes are floated by gating the sense amplifier from the power rails, VDD and GND, by deactivating *M4* and *M9* using $\overline{SE}$ and *SE*, the same as commodity DRAMs (see Fig. 3. In this phase, *M2* and *M3* are activated using Left Bitline Coupler (LCPL) and Right Bitline Coupler (RCPL) signals. *M1* connects LBL and RBL till right before activating the target wordline, to keep the voltage of bitlines in each pair as close as possible.

By activating the target row, the corresponding memory cell is connected to one of the bitlines in a pair. Connecting the cell to LBL or RBL depends on activating even or odd wordlines. Together with activating the target wordline, one of the RR or LR signals is activated to connect the reference cell to the opposite bitline in that pair.

During cell access in this design, two scenarios are possible: 1- the charges of target cell and floated bitlines are the same and 2- the charges of target cell and floated bitlines differs.

In the first scenario, the voltage of connected bitline to the target cell remains unchanged (*VDD* or *0*). This is while the voltage perturbation amplitude on the bitline connected to the reference cell is equal to the commodity DRAMs ($V_S = \frac{V_{DD}}{2} \cdot \frac{C_{ref\,Cell}}{C_{BL}+C_{ref\,Cell}}$). If the bitlines initial voltage is *VDD*, the voltage of bitline connected to the reference cell drops to $VDD - V_S$ and if its initial voltage is *0*, its voltage rises to V_S.

In the second scenario, a voltage perturbation of equal to $2V_S$ is developed on the bitline connected to the target cell ($VDD.\frac{C_{ref\,Cell}}{C_{BL}+C_{ref\,Cell}}$). Thus, the voltage of the bitline corresponding to the target cell will be $VDD - 2V_S$ or $2V_S$ at the end of charge sharing phase, while the voltage change on the reference bitline will be $VDD - V_S$ or V_S. In both scenarios and with all of possible bitlines and target cells initial charges, the voltage difference between bitlines is equal to V_S.

After charge sharing, bitlines are decoupled from the sense amplifier by deactivating *M2* and *M3*, and the sense amplifier is activated, by turning *M4* and *M9* on. Based on the initial condition on the sense amplifier internal nodes, the node with higher voltage rises to *VDD* and the other node drops to *0*; the same operation as the Stabilization phase in commodity DRAMs. Decoupling transistors are exploited in the previous works as well, to decrease tRCD and power dissipation during sense amplifier stabilization [19]. By decoupling bitlines with high parasitic capacitance during stabilization of the sense amplifier, the sense amplifier stabilizes faster and more power-efficient [19].

After stabilization of the sense amplifier, *M1* together with one of *M2* or *M3* are activated to update bitlines and restore the accessed cell. If the memory cell connected to LBL is accessed, *M2* and *M1* are activated, and if the cell connected to RBL is accessed, *M3* and *M1* are activated to update bitlines and cell restoration.

All of the control signals are triggered by VPP to prevent voltage drop on the n-type transistors, like the commodity DRAMs. Since a large number of cell accesses lead to reading the same value on bitlines [8], [16], thus, the bitlines voltage swing in a large portion of Activations is limited to charge and discharge of the reference bitline in each pair by V_S. A bitline pair voltage flips only when the stored charge in the target cell differs from the previously accessed cell through that bitline pair.

The column decoder selects the target column(s) right after the stabilization of the sense amplifier to transfer the read value to the global sense amplifier to be sent to the I/O drivers. Bitlines are updated in parallel with the required time for accessing the target column (tCL).

During the restoration of the accessed cell, the reference cell is disconnected from the bitline and is charged/discharged to half-VDD by activating the RCC signal, to prepare the memory array for the next Activation. Thus, right after the previous Activation, the next row Activation can be issued in CoolDRAM with no need to perform a Precharge operation on bitlines. Both open-row and closed-row policies are applicable in CoolDRAM.

The main contributions of CoolDRAM are as follows:

- Very simple modification is applied to DRAM arrays to entirely eliminate Precharge phase from DRAM cell access procedure. In the only previous work [16] that eliminates Precharge phase, the sense amplifier is modified to detect different voltages on bitlines.
- Imbalanced sense amplifier in the previous work [16] is very complex to be tuned in nano-scale technology, with high process variation rate. This is while CoolDRAM uses the same straightforward balanced sense amplifier structure of commodity DRAMs which is more straightforward to fabricate.

III. Experimental Setup

Circuit-level simulations are conducted using Synopsys HSPICE using 14 nm Multi-Gate technology model. DRAM memory MAT size is considered 512×512 cells [9]. Operational DRAM cells and reference cell access transistors' dimensions are considered the smallest in the technology. Coupling transistors and equalizer transistor width-to-length ratio is set at 2-to-1 and the sense amplifier transistors dimensions are considered 10-to-1.

The voltage of activating power source is 2.5 V (the same as commodity DRAMs). The main VDD, used for sense amplifiers, peripheral units, and charging DRAM cells is considered 1.2 V. Half-VDD power source (0.6 V) is generated by the internal voltage regulator of DRAM chips.

IV. Evaluation

A. CoolDRAM Functionality

Bitlines' parasitic capacitance and resistance are accurately modeled by the Lumped-element model by segmenting the bitlines into 3 elements. Regarding the distributed parasitic resistance and capacitance on bitlines, the physical location of accessed cell may generate different V_S on the sense amplifier nodes. The highest V_S is generated by the DRAM cells which are physically located closer to the sense amplifier and the weakest V_S is generated by the furthest cell from the sense amplifier. In our evaluations we considered the worst-case scenario of cell access which is the access to the furthest memory cell from the sense amplifier in the memory array. The reference cell location is fixed, thus the developed signal by connecting the reference cell to the bitline is predictable and experiences fewer fluctuations.

Fig. 4 shows all possible signal activities on the reference cell, accessed cell, bitlines, and sense amplifier during cell access through the LBL (as an example). The reference cell is charged to 0.6 V before Activation. By connecting the reference cell to RBL, by sharing its charge with the RBL parasitic capacitance, its voltage drops or rises to about 1.11 V or 0.09 V, respectively. This voltage perturbation is almost half of the voltage perturbation generated by the target cell charge. See the voltage perturbation on LBL and RBL during the time period between 1 ns and 3 ns.

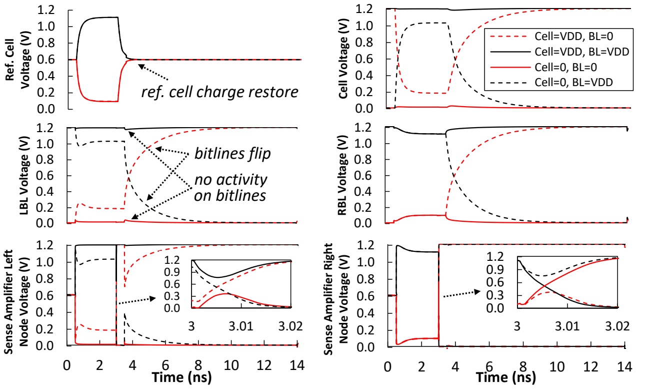

Fig. 4: Voltage waveform of nodes in a bitline pair, showing the functionality of CoolDRAM

TABLE I: Simulation parameters

Technology	14 nm Multi-Gate Predictive Technology Model
Operating Voltage	Chip power supply (VDD) = 1.2 V Activating power supply (VPP) = 2.5 V
MAT size	512×512 cell
C_{BL}, C_{cell}	144 fF, 24 fF
Transistors W/L	SA 10/1, M1∼M3 2/1, access transistors 1/1
Process Variation	σ = 2.5%∼15%, m = V_{th}, Cap., Res.

Two different conditions can be observed on the LBL after connecting the target cell to this bitline. If the stored charge on the LBL is the same as the stored charge in the cell, black and red solid lines in Fig. 4, no voltage perturbation occurs on LBL. In the other case, if the stored charge in the cell differs from the charge on the LBL, the voltage perturbation is about 180 mV. In both cases, the amplitude of voltage difference between LBL and RBL is about 90 mV.

Sense amplifier internal nodes follow LBL and RBL voltages during the time period 0.5 ns to 3 ns. By decoupling the sense amplifier from the bitlines at 3 ns and activating the sense amplifier, the sense amplifier's node with higher voltage rises to *VDD* and the other node drops to *0*. This is performed in about 20 ps because of the low parasitic capacitance of sense amplifier internal nodes. After stabilization of the sense amplifier, by activating *M2* (the target cell is connected to LBL in this example), together with *M1* (bitlines equalizing transistor), the LBL, RBL, and accessed cell voltages are restored or updated. It is worth mentioning that LBL and RBL voltages flip only if the recently accessed cell value differs from the previously accessed cell value. Otherwise, LBL and connected cell voltages remain untouched and voltage perturbation on the RBL is recovered to *VDD* or *0*. Together with updating bitlines and accessed cell voltage, the reference cell voltage is charged to 0.6 V by triggering the RCC signal.

B. CoolDRAM Robustness

1) Process Variation: The Gaussian distribution is used to model process variation effect on transistors threshold voltages (V_{th}), as well as cells and bitlines capacitance, and resistance drift. The mean value (m) is the nominal parameter value with a standard deviation of σ (see Table I). In our evaluation, we considered Systematic Process Variation as one of the most important process variation effects in nano-scale chip fabrication. Under this variation, nearby devices experience

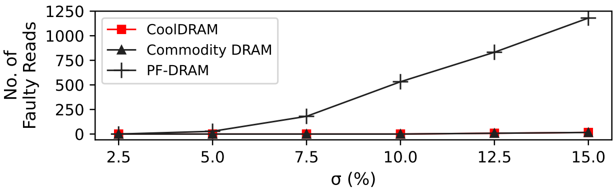

Fig. 5: Number of faulty reads due to process variation over 20 K simulation rounds

the same drift in their nominal parameters. To evaluate this, Monte Carlo simulation is employed.

Fig. 5 shows the number of faulty operations caused by process variation over 20K simulation rounds. The results show that CoolDRAM and commodity DRAM are highly resilient to process variation, with no faulty reads observed even when increasing σ to 10% and only 8 and 16 faults are observed by increasing σ to 12.5% and 15%, respectively. Since PF-DRAM operation is based on an imbalanced sense amplifier, a slight drift in its sense amplifier transistors' threshold voltage may lead to a faulty cell access operation. The study's findings indicate that the fault rate for PF-DRAM significantly increases to 6.1% at σ of 15%, whereas CoolDRAM's fault rate is less than 0.1%.

2) Noise Immunity: There are two types of noise that can impact DRAM cell access: ambient noise and internal noise. Bitlines are particularly vulnerable to ambient noise due to their lengthy structure, which acts like an antenna and absorbs electromagnetic noise during the charge sharing phase. Internal noise in DRAM is dependent on the clearance between the bitlines and wordlines (space between them) and can be categorized into several types, including Wordline Drive Noise, Power-Supply Voltage Bounce, and Bitline-to-Bitline and Bitline-to-Wordline coupling noises.

Ambient Noise: The Monte Carlo simulation with 20K iterations is utilized to detect the number of faulty reads during cell access. Ambient noise on the bitlines is simulated using a current source which simulates a spike noise by Eq.(II). Here, Q and T are the injected charge and duration, and t and Per are simulation time and noise injection moment, respectively.

$$I_{noise} = (\frac{Q}{T})\sqrt{SGN(1+SGN(t-Per)) \times \frac{t-Per}{T}} \times \exp(-\frac{t-Per}{T}) \tag{5}$$

Table II presents the number of faulty read operations observed during 20 K simulation rounds in the presence of ambient noise. The results demonstrate that PF-DRAM exhibits the highest fault rate. With an ambient noise charge of 2.5 fC, 3 faulty reads are observed in PF-DRAM. This value increases significantly to 3350 faulty reads with 10 fC ambient noise charge, which is 1.76× and 1.78× of that of commodity DRAM and CoolDRAM, respectively. The slightly higher robustness of CoolDRAM compared with commodity DRAM is because of higher capacitance on the reference bitline by connecting the reference cell to it during cell access.

Fig. 6 illustrates the output voltage of the sense amplifier before it reaches its final stabilization point in the presence of various noise levels on the bitlines during a read operation of a '1' on the accessed cell (2000 samples are presented). In a correct read, the sense amplifier should be stabilized to *VDD*, thus the cases with the sense amplifier output voltage under *0* V lead to a faulty read. As Fig. 6 shows, PF-DRAM sense amplifier output voltage is much more affected by the ambient noise compared with the CoolDRAM, with the variance of 5.8 mV compared with 2.8 mV, respectively.

TABLE II: Number of faulty read operations at the presence of ambient noise in 20 K simulation rounds

Average Charge	CoolDRAM	PF-DRAM	Commodity DRAM
0C	0	0	0
2.5fC	0	3	0
5fC	40	648	48
7.5fC	694	2060	722
10fC	1873	3350	1902

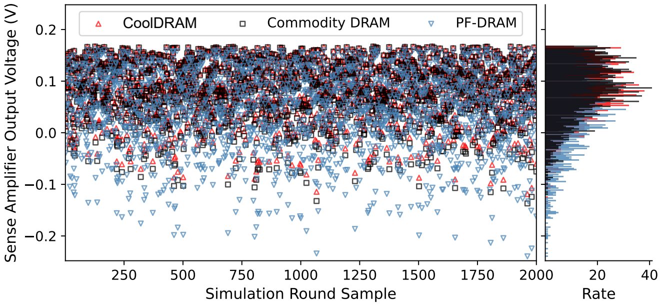

Fig. 6: Noise effect on sense amplifier output voltage (2K Samples are shown)

Internal Noise: Since bitlines in commodity DRAM, with folded structure, swing in opposite directions during the Precharge and Activation phases, the crosstalk effect between even and odd bitlines and wordlines cancels each other out pretty well. However, in open bitline structure DRAMs, Bitline-to-Wordline noise is more significant. In CoolDRAM, bitlines voltage swing reduces considerably in normal system operation, because of data similarity and special data access scheme. Nevertheless, in the worst-case, i.e. a Trojan which arranges and accesses data in DRAM in a way that in any access a swing on all of the bitlines occurs, can increase Bitline-to-Wordline noise. However, since row buffer access and restoration can be performed in parallel in CoolDRAM, by decreasing the steep of signals on the bitlines, this noise source can be controlled.

C. System-Level Evaluation

The energy consumption reduction during CoolDRAM access is dependent on both the row-hit rate and bit-flip probability in consecutive accesses to the same bitline. To determine these parameters, we selected 16 random workloads from the SPEC CPU2017 benchmark suite and used the gem5 full system simulator on an X86 64-bit processor with single-, 2-, 4-, and 16-cores, two levels of 32 KB and 2 MB of caches. Our system-level evaluation considers the main memory as a DDR4-2400 with a capacity of 8 GB and an 8 KB row buffer size, with open-row policy. The Micron power model [2] is used to calculate the energy dissipation of different components of DRAM.

The results show removing Precharge phase from cell access (read/write/refresh) reduces energy consumption by 28%, 29%, 31%, and 34%, on average for single-, 2-, 4-, and 16-cores, respectively. Increasing the number of parallel executing workloads also improves the efficiency of CoolDRAM. Fig. 7 depicts the Activation energy reduction of CoolDRAM compared to commodity DRAM. The required energy for charge-discharge cycles of bitlines, one of the main power-

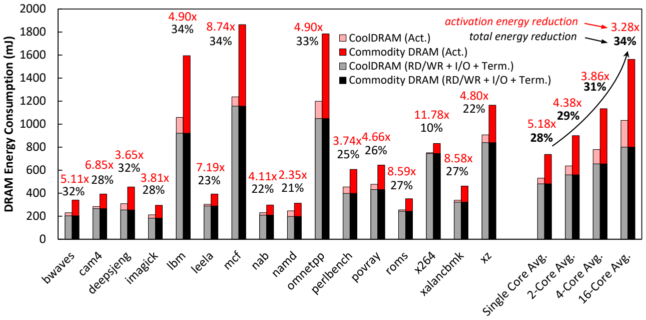

Fig. 7: Comparison of energy consumption between CoolDRAM and commodity DRAM for executing one billion instructions from each workload in the SPEC CPU2017 benchmark suite

hungry operations in DRAM [2], is reduced by an average of 5.18×, 4.38×, 3.86×, and 3.28× for single-, 2-, 4-, and 16-cores, respectively. The energy dissipation reduction for *x264* and *namd* are 11.78× (highest) and 2.35× (lowest), respectively. This difference in reduction is attributed to x264 having lower locality of references to DRAM and a higher bit-flip probability during execution compared to namd. The peripheral energy consumption for Read and write, I/O, and I/O termination power are the same for commodity DRAM and CoolDRAM. As both CoolDRAM and PF-DRAM eliminate the precharge phase, their energy dissipation during SPEC CPU2017 execution is very close, with a difference of less than 1% due to the sense amplifier structure. Hence, the results for PF-DRAM are not presented here.

D. Area Overhead

Sense amplifiers occupy nearly 8% of the entire commodity DRAM memory chip area [16], [19]. The proposed CoolDRAM memory structure keeps sense amplifiers of DRAM untouched. The area of bitlines coupling transistors and bitlines voltage equalizer transistor is exactly the same as the required area for the precharge circuit in commodity DRAMs [9]. The only additional circuit in CoolDRAM, as compared to commodity DRAM, is one row of dual-contact reference cells and one half-VDD charge transistor per bitline pair. The occupied area by the added circuit in CoolDRAM is less than two rows in the memory array. Thus, the area overhead imposed by the added circuit to DRAM MATs with 512×512 cells is less than 0.4%. It is worth mentioning that the two access transistors to the dual-contact cell and the charging transistor are made using the smallest possible DRAM transistors in the technology, the same as DRAM cell access transistors. The area overhead of CoolDRAM is approximately 22× less than that of PF-DRAM (the area overhead of PF-DRAM is about 8.8%).

V. Conclusions

The precharge phase in DRAM cell access is a power-hungry and time-consuming operation. This work proposes a DRAM structure, called CoolDRAM, in which cell access operation is performed without the need for Precharging bitlines. The area overhead of CoolDRAM is less than 0.4%, and its robustness against noise and process variation almost remained untouched compared with commodity DRAM. Meanwhile, CoolDRAM's energy consumption is reduced by an average of 34% compared to commodity DRAM.

References

[1] "Ddr4 sdram," https://www.micron.com/-/media/client/global/documents/products/data-sheet/dram/ddr4/8gb_ddr4_sdram.pdf, 2015.

[2] "Micron technical note: Calculating memory power for ddr4 sdram," https://www.micron.com/-/media/client/global/documents/products/technical-note/dram/tn4007_ddr4_power_calculation.pdf, 2017.

[3] A. Agrawal, A. Ansari, and J. Torrellas, "Mosaic: Exploiting the spatial locality of process variation to reduce refresh energy in on-chip edram modules," in *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2014, pp. 84–95.

[4] A. M. Amin and Z. A. Chishti, "Rank-aware cache replacement and write buffering to improve dram energy efficiency," in *2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*. IEEE, 2010, pp. 383–388.

[5] I. Bhati, Z. Chishti, S.-L. Lu, and B. Jacob, "Flexible auto-refresh: Enabling scalable and energy-efficient dram refresh reductions," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, 2015, pp. 235–246.

[6] K. K. Chang, A. G. Yağlıkçı, S. Ghose, A. Agrawal, N. Chatterjee, A. Kashyap, D. Lee, M. O'Connor, H. Hassan, and O. Mutlu, "Understanding reduced-voltage operation in modern dram devices: Experimental characterization, analysis, and mechanisms," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 1, pp. 1–42, 2017.

[7] N. Chatterjee, M. O'Connor, D. Lee, D. R. Johnson, S. W. Keckler, M. Rhu, and W. J. Dally, "Architecting an energy-efficient dram system for gpus," in *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2017, pp. 73–84.

[8] E. Cooper-Balis and B. Jacob, "Fine-grained activation for power reduction in dram," *IEEE Micro*, vol. 30, no. 3, pp. 34–47, 2010.

[9] K. Itoh, *VLSI memory chip design*. Springer Science & Business Media, 2013, vol. 5.

[10] S. Khan, D. Lee, and O. Mutlu, "Parbor: An efficient system-level technique to detect data-dependent failures in dram," in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2016, pp. 239–250.

[11] S. Khan, C. Wilkerson, D. Lee, A. R. Alameldeen, and O. Mutlu, "A case for memory content-based detection and mitigation of data-dependent failures in dram," *IEEE Computer Architecture Letters*, vol. 16, 2016.

[12] S.-L. Lu, Y.-C. Lin, and C.-L. Yang, "Improving dram latency with dynamic asymmetric subarray," in *48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2015, pp. 255–266.

[13] S. Mittal and J. S. Vetter, "A survey of architectural approaches for data compression in cache and main memory systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, pp. 1524–1536, 2015.

[14] M. Patel, J. S. Kim, and O. Mutlu, "The reach profiler (reaper) enabling the mitigation of dram retention failures via profiling at aggressive conditions," *ACM SIGARCH Computer Architecture News*, vol. 45, 2017.

[15] G. Pekhimenko, V. Seshadri, O. Mutlu, M. A. Kozuch, P. B. Gibbons, and T. C. Mowry, "Base-delta-immediate compression: Practical data compression for on-chip caches," in *International Conference on Parallel Architectures and Compilation Techniques (PACT)*. IEEE, 2012.

[16] N. Rohbani, S. Darabi, and H. Sarbazi-Azad, "Pf-dram: a precharge-free dram structure," in *ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 126–138.

[17] N. Rohbani, M. A. Soleimani, and H. Sarbazi-Azad, "Pipf-dram: processing in precharge-free dram," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 1075–1080.

[18] H. Seol, W. Shin, J. Jang, J. Choi, J. Suh, and L.-S. Kim, "Energy efficient data encoding in dram channels exploiting data value similarity," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, 2016.

[19] O. Seongil, Y. H. Son, N. S. Kim, and J. H. Ahn, "Row-buffer decoupling: A case for low-latency dram microarchitecture," in *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. IEEE, 2014, pp. 337–348.

[20] S. Srikanth, L. Subramanian, S. Subramoney, T. M. Conte, and H. Wang, "Tackling memory access latency through dram row management," in *Proceedings of the International Symposium on Memory Systems*, 2018.

[21] X. Tao, Q. Zeng, and J.-K. Peir, "Hot row identification of dram memory in a multicore system," in *Proceedings of the International Conference on High Performance Compilation, Computing and Communications*, 2017, pp. 71–75.

[22] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. P. Jouppi, "Rethinking dram design and organization for energy-constrained multi-cores," in *Proceedings of the 37th annual international symposium on Computer architecture*, 2010, pp. 175–186.