Nonvolatile CBRAM-crossbar-based 3-D-integrated hybrid memory for data retention

Wang, Yuhao; Yu, Hao; Zhang, Wei

2013

Wang, Y., Yu, H., & Zhang, W. (2014). Nonvolatile CBRAM-Crossbar-Based 3-D-Integrated Hybrid Memory for Data Retention. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 22(5), 957-970.

<https://hdl.handle.net/10356/102813>

<https://doi.org/10.1109/TVLSI.2013.2265754>

© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [DOI: <http://dx.doi.org/10.1109/TVLSI.2013.2265754>].

Downloaded on 26 Apr 2023 19:19:13 SGT

Non-volatile CBRAM-Crossbar based 3D-Integrated Hybrid Memory for Data Retention

Yuhao Wang, Hao Yu, *Member, IEEE* Wei Zhang, *Member, IEEE*

Abstract—This paper explores the design of 3D-integrated hybrid memory by conductive-bridge random-access-memory (CBRAM). Considering internal states, height and radius of conductive bridge, of one CBRAM device, accurate CBRAM device model is developed for CBRAM-crossbar based non-volatile memory (NVM) design with efficient estimation of area, access time and power. Based on this design platform, one 3D-integrated hybrid memory is designed by stacking one tier of CBRAM-crossbar with tiers of SRAM and DRAM, where the tier of CBRAM-crossbar is deployed for data retention during power gating of SRAM/DRAM tiers. One corresponding block-level data retention is developed to only write back dirty data from SRAM/DRAM to CBRAM-crossbar. When compared to phase-change random-access-memory (PCRAM) based system-level data retention, our design achieves 11x faster data-migration speed and 10x less data-migration power. When compared to ferroelectric random-access-memory (FeRAM) based bit-level data retention, our design also achieves 17x smaller area and 56x smaller power under the same data-migration speed.

Index Terms—Conductive-bridge RAM, non-volatile memory, data retention, 3D hybrid memory, power gating.

I. INTRODUCTION

Memory technologies have undergone a rapid advance from traditional CMOS SRAM/DRAM to nano-scale non-volatile memory (NVM) [1], [2], [3], [4], [5], [6], which can provide low power and high performance memory solutions for future computing systems. Characterized by different features the emerging NVMs may fit in different applications. For example, spin-torque-transfer based memory (STT-RAM) exhibits close-to-DRAM performance and virtually unlimited endurance, but requires large footprint and high driving current [1], [7], thus is considered to be cache-class memory. The phase-change-memory (PCRAM) has high integration density, 10~1000ns scale performance but limited endurance, thus is categorized as storage-class memory. The recent designs are also for hybrid memory structure [8] towards high integration density, fast accessing time, long endurance time, and low leakage power by leveraging strengths of different memory technologies. This paper shows one possible hybrid memory design by 3D-integrated SRAM/DRAM with CBRAM-crossbar.

From perspective of device technology, the recent conductive-bridge random-access-memory (CBRAM) has

been introduced as one emerging resistive random-access-memory (ReRAM) device [3], [9], [10], [11], [12], [13]. Based on the fast ion-migration process, the conductive filament of CBRAM is able to grow and shrink extremely fast under low supply voltage. As a result, compared to other types of NVMs, CBRAM has advantages in higher speed and lower power. Moreover, CBRAM fabrication process can easily result in extremely high density crossbar structure. The CBRAM is categorized as storage-class memory due to high density and limited endurance ($\sim 10^6$) [14]. However, accurate device model or description of non-volatile memory devices at nano-scale is different from the traditional devices. For example, to fully describe the state of CBRAM devices, one needs to include height and radius of conductive bridge to describe the state of CBRAM device such as CBRAM conductance. Obviously, the states of height and radius are different from the states of electrical currents and voltages. The approach in [15] introduces an equivalent circuit to describe such internal states, which is not scalable with geometries for large-scale designs. Therefore, one needs to develop a SPICE-like CBRAM design platform [16], [17], [18], [19] to accurately describe hybrid designs by CBRAM and CMOS devices. As such, one can extract circuit parameters for CBRAM-crossbar based memory such as access time, power and area similar to CACTI [20].

From perspective of system architecture, 3D die-stacking [21], [22], [23], [24], [25] is promising to integrate hybrid memory components with high density and low latency. One can design a hybrid memory system with each tier by different memory technology stacked by through-silicon-vias (TSVs) [21]. However, as leakage power is the primary concern of a memory system, one needs to have well-designed data-retention scheme such that power gating can be effectively deployed to reduce leakage power yet without degrading performance. Traditionally, the common approach for data retention of SRAM/DRAM [26] is to deploy a small retention voltage for all memory cells in sleep-mode, which still has non-negligible leakage power. Recently, the work in [27] applies the non-volatile PCRAM for *system-level* data retention of DRAM. DRAM layer and PCRAM layer are stacked in 3D fashion and connected by TSVs. Benefited from the increased number of vertical data paths, the bus bandwidth is significantly improved compared to 2D scenario. However, due to the asymmetric performance between DRAM and PCRAM, the data transfer frequency and bandwidth are limited by the low-write latency of PCRAM. Also, asynchronous handshaking protocol is required that incurs additional overhead. Another recent work in [28] has *bit-level* data retention by embedding one FeRAM device for each SRAM cell with bit-

Y. Wang and H. Yu are all with School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. W. Zhang is with School of Computer Engineering, Nanyang Technological University (NTU), Singapore. The preliminary work was published in IEEE/ACM ISLPED'12. This work is sponsored by Singapore MOE TIER-2 fund MOE2010-T2-2-037 (ARC 5/11) and A*STAR SERC-PSF fund 11201202015. Please address comments to haoyu@ntu.edu.sg.

wise data-transfer controllers. Although concurrent bit-level data migration achieves fast speed, the overhead is overwhelmed with additional bit-wise data-transfer controllers in SRAM cells, which degrades the SRAM performance during the normal active mode.

In this paper, we introduce a 3D hybrid memory architecture, in which CBRAM-crossbar is stacked as one tier for data retention of the other SRAM/DRAM tiers under power gating. The primary contributions of this paper are as follows. Firstly, one CBRAM-crossbar based NVM design platform is developed with accurate device-level modeling and system-level estimation for area, accessing time and power. Secondly, the CBRAM-crossbar is deployed for data retention of SRAM/DRAM under power gating in fine-grained *block-level*, in which intrinsic block-level incremental dirty-data write-back scheme is deployed for fast and power efficient data migration. Experiment results show that our 3D hybrid memory with CBRAM-crossbar achieves 11x faster migration time and 10x less migration power for hibernating transition compared to the system-level PCRAM-based data retention in [27], under the same amount of TSVs used. Compared to the bit-level FeRAM-based data retention in [28], our 3D hybrid memory system achieves 17x area reduction along with 2x SRAM performance improvement and 56x transition power reduction with similar transition speed, under the incremental dirty-data write-back scheme.

The rest of this paper is organized as follows. Section II presents the architecture of the proposed 3D hybrid memory with the use of CBRAM-crossbar for data retention. In section III and IV, CBRAM-crossbar based NVM design platform is developed from both device and circuit levels, respectively. In section V, block-level data retention with incremental dirty-data write-back is introduced based on CBRAM-crossbar. Finally, experiment results are shown in section VI with conclusion in section VII.

II. 3D HYBRID MEMORY WITH CBRAM-CROSSBAR FOR DATA-RETENTION

In this section, we introduce the proposed architecture for the 3D hybrid memory system based on CBRAM-crossbar with block-level data retention ability.

Fig. 1 illustrates the overall system architecture of the proposed 3D hybrid memory, which is composed of embedded DRAM, SRAM and CBRAM-crossbar located at three layers (or tiers) connected by TSVs. Similar to the common memory organization [20], the entire CBRAM-crossbar memory is broken into *banks*, where each bank can be accessed independently with dedicated data and address buses. Each bank is further broken into a $M \times N$ *mat* array, where each mat consists of one $m \times n$ CBRAM-crossbar. In addition, the terms of *sleep mode* and *active mode* are used to denote whether the system is power-gated or not. The active to sleep mode transition is called *hibernating transition*, and term *wakeup transition* is used vice versa. For better presentation, Table I summarizes the notations used throughout this paper.

To reduce the sneak-path power in crossbar structure [29], we propose to distribute the multi-bit data into different mats

Fig. 1. 3D hybrid memory system with CBRAM-crossbar based data retention

TABLE I
NOTATIONS

Notation	Descriptions
T_{Hi}/P_{Hi}	Hibernating transition time/power for bank i
T_{Wi}/P_{Wi}	Wakeup transition time/power for bank i
P_b/P_m	Bank/mat power
E_b/E_m	Bank/mat access energy
C_b/C_m	Memory bank/mat capacity
N_b/N_m	Number of memory banks/mats
W_d/W_a	Data/address bus width of each bank
W_D/W_A	Total data/address bus width
M/N	Number of row/column mats of each bank
B_C/B_D	Cache line/DRAM page size

of the same row concurrently, where each mat (i.e., one CBRAM-crossbar) accepts one bit of the written data at each cycle. To achieve this, the data address is decoded to find the exact same crossing point of each CBRAM-crossbar mat where one bit of the arriving data is to be kept. For read operation, the same address decoding is carried out and the bits from each mat are combined together as output. As such, during one hibernating transition, each SRAM/DRAM bank is associated with one dedicated CBRAM-crossbar bank of the same capacity for data retention. Therefore, power gating is employed at bank-level (i.e., block-level). Once the hibernating transition begins, all the data in the specified SRAM/DRAM bank must be copied or migrated to the corresponding CBRAM-crossbar memory bank through dedicated data and address buses implemented by TSVs in the vertical direction. On the other hand, all data must be migrated from CBRAM-crossbar back to the SRAM/DRAM during the wakeup transition.

However, the primary design challenges to develop the above-mentioned 3D hybrid memory with CBRAM-crossbar are from twofold. Firstly, there is no design platform for CBRAM device and CBRAM-crossbar circuit such that the delay, area and power can be estimated and optimized. Secondly, there is no memory controller developed and verified for the CBRAM-crossbar based data retention, which can perform efficient data migration for SRAM/DRAM. In the following, we show the development of one design platform for CBRAM device and CBRAM-crossbar circuit. Moreover,

we show one memory controller design for CBRAM-crossbar using an incremental block-level data retention.

III. CBRAM DEVICE MODELING

In this section, the CBRAM device physics is illustrated with discussion of its internal states, and the hybrid CMOS-CBRAM circuit simulation is discussed as well.

Fig. 2. (a) Working mechanism of CBRAM with the shape morphing of conductive filaments in several phases illustrated between ON state and OFF state; (b) cross section of CBRAM device with defined geometric variables

A. CBRAM Device with Internal State

The conventional electronic devices can be readily described by common state variables such as nodal voltage and branch current. Thus for any given circuit, its behavior can be fully decided by obtaining state variables based on circuit state equations under Kirchhoff's current and voltage laws. However, for the newly discovered non-volatile memory (NVM) devices at nano-scale, additional internal state variables such as doping-ratio for memristor, magnetization angle for spin-torque-transfer device, and crystalline fraction for phase-change device have to be considered to correctly describe their physics-based dynamical behaviors [29], [17].

As for CBRAM devices, also known as programmable metalization cell (PMC) [30], [31] or NanoBridge [32], [33], it is one emerging two-terminal non-volatile memory device that can be fabricated by depositing an active anode layer, a solid electrolyte layer, and an inert cathode layer, where the solid electrolyte layer is sandwiched by the other two

electrodes layers. A variety of electrode and solid electrolyte materials combinations have been reported in the literature [34], [35], [36], [37], [38]. CBRAM-based memory has also been successfully fabricated in chip-level [3].

The physical working mechanism of CBRAM device is illustrated in Fig. 2(a), which can be summarized as the shape-morphing of the conductive metal filament within ambient resistive solid electrolyte. When a positive bias-voltage is applied, the conductive filament, modeled as a cone with fixed base radius, first grows vertically by accumulating metal ions until the two electrodes are bridged together. Then the cone begins to morph into a cylinder, which will eventually set the CBRAM into low-resistance or *on* state. Similarly when a negative bias-voltage is applied, the cylinder-shaped conductive filament dissolves to a cone with same base area, disconnects the two electrodes, shrinks vertically and then turns the CBRAM in high-resistance or *off* state. From the defined geometric variables shown in Fig. 2(b), it can be observed that the variables h and r are capable to model the shape-morphing process, where detailed equations are given in [9].

In this paper, we find that the angle θ from the side of the cone to the bottom can determine the shape of filament, when assuming that a cylinder is simply a cone whose apex is at infinity. Thus the internal state variable s or $\tan\theta$, i.e. the height of cone in a projective perspective to base radius, is selected to determine the resistance of one CBRAM device. Hence s becomes a shape-determining state variable that a large value of s indicates the filament more cylinder-shaped while a small value for more cone-shaped. As such, the number of state variable for each device can be reduced from two to one, which will greatly improve the simulation speed as discussed later in section III-B.

As such, by modifying the equations from [9], the dynamic behavior of CBRAM device based on variable s can be modeled as

$$\frac{ds}{dt} = \begin{cases} \frac{k_1}{R} \cdot \sinh\left(\frac{k_2 \cdot V}{L + k_3 \cdot R \cdot s}\right) & s \leq \frac{L}{R} \\ \frac{s^2 \cdot k_1}{L} \cdot \sinh(k_4 \cdot V) & s > \frac{L}{R} \end{cases} \quad (1)$$

where k_1 , k_2 , k_3 and k_4 are constants, and their detailed definitions can be found in [9]. Additionally, V is the applied voltage on CBRAM device, L is the thickness of CBRAM and R is the constant base radius.

The equivalent resistance of CBRAM with respect to variable s can be calculated by

$$\begin{cases} R_{off} = \frac{\rho_{on} \cdot s \cdot R + \rho_{off}(L - s \cdot R)}{\pi \cdot R^2} & s \leq \frac{L}{R} \\ R_{on} = \frac{\rho_{on} \cdot L}{\pi \cdot R \cdot (R - L/s)} & s > \frac{L}{R} \end{cases} \quad (2)$$

where ρ_{on} and ρ_{off} are the conductive filament resistivity and non-conductive solid-electrolyte resistivity, respectively.

Two observations from equations (1) and (2) can be drawn for the fast write-speed of CBRAM. Firstly, there exists an exponential relation between the state changing-speed ds/dt

and the applied voltage V . In the literature, CBRAMs with less than 50ns write-latency have been demonstrated under 1.5V bias voltage [3], [11]. Secondly, the morphing steps between *off* state and intermediate state 1 as well as that between *on* state and intermediate state 2 are much more time consuming than that between intermediate states 1 and 2. Practically, by properly defining the intermediate states 1 and 2 and operating the devices within this range, the unnecessary time for morphing at two ends can be eliminated thus to significantly improve the write-speed while still to achieve acceptable *off/on* resistance ratio.

In addition, it can be observed that the switching speed indicated by Equation (1), and also R_{on} and R_{off} indicated by Equation (2), are all dependent only on the shape of internal conductive filament, rather than the external sizes of the upper and bottom contacts. As the filament size is significantly smaller than that of contacts, the CBRAM device exhibits a great potential of scalability. The scalability of CBRAM device has been confirmed by [14], within which it has been demonstrated that the threshold voltage, R_{on} and R_{off} are feature-size independent from um to nm regions.

B. Hybrid CMOS-CBRAM Simulation

Combining equations (1) to (2) together, we can reach the conclusion that the branch current of the CBRAM device can be decided by the conductive filament shape-determining variable s and also the applied voltage V . As such, the modified nodal analysis (MNA) that takes s as the additional new internal variable can be built and solved by extending the existing SPICE-like simulation approach developed in [17].

As discussed in [17], to perform transient analysis together with NVM devices, the following linearized circuit equation is established

$$\begin{bmatrix} G & E_i & E_l & S \\ -E_i & 0 & 0 & 0 \\ -E_l & 0 & 0 & 0 \\ K_v^g & 0 & 0 & K_s^g \end{bmatrix} \begin{bmatrix} \delta v_n \\ \delta j_i \\ \delta j_l \\ \delta s_m \end{bmatrix} + \begin{bmatrix} C & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & L_i & 0 \\ K_v^c & 0 & 0 & K_s^c \end{bmatrix} \begin{bmatrix} \delta \dot{v}_n \\ \delta \dot{j}_i \\ \delta \dot{j}_l \\ \delta \dot{s}_m \end{bmatrix} = -F(X_0, \dot{X}_0, t) \quad (3)$$

where E_i and E_l are incidence matrix to describe topological connections of voltage source and inductive elements, G , C , S the Jacobian terms of conductance, capacitance and also the newly introduced memory conductance of NVM and CMOS devices.

The new state variable vector $X = [v_n, j_i, j_l, s_m]^T$ contains nodal voltage, source current, inductor current, and internal state variable of NVM devices, respectively. $K_v^g, K_s^g, K_v^c, K_s^c$ are the additional Jacobian terms introduced from NVM device equations. As such, for CBRAM memory circuit designs, our approach avoids the use of equivalent circuit components, which is not scaled with dependence on geometries and not scalable for large-scale memory design. Moreover, the reduction of the number of CBRAM state variables from two

to only one will also greatly reduce the size of state matrix in (3).

Applying equations (1) and (2) to (3), we can obtain the required Jacobian terms as follows

$$G = \begin{bmatrix} g(s) & -g(s) \\ -g(s) & g(s) \end{bmatrix}, S = \begin{bmatrix} V \cdot \frac{dg(s)}{ds} \\ -V \cdot \frac{dg(s)}{ds} \end{bmatrix},$$

$$K_v^g = \begin{bmatrix} -\frac{df(V,s)}{dv} & \frac{df(V,s)}{dv} \end{bmatrix}, K_s^g = -\frac{df(V,s)}{ds},$$

$$K_v^c = 0, K_s^c = 1 \quad (4)$$

where $g(s) = \frac{1}{R}$ is the CBRAM conductance, and $f(V, s) = \frac{ds}{dt}$ as indicated in equation (1). As such, the new MNA for CBRAM is formulated and can be directly applied in transient analysis in a SPICE-like simulator.

IV. CBRAM-CROSSBAR MEMORY DESIGN

Based on the developed CBRAM device simulator, we further explore the design of CBRAM-crossbar based memory in this section.

Fig. 3. Two approaches for 3D stacked CBRAM-crossbar fabrication (a) crossbar structure integrated within the interconnect layers, and CBRAM devices fabricated at the bottom of vias (b) crossbar structure fabricated on the top of interconnect layers

A. CBRAM-crossbar Memory Unit

Fig. 5 illustrates the design for CBRAM-crossbar structure. Compared to the 1T-1R structure [39], the crossbar structure has two major advantages. Firstly, the extremely high integration density can be realized due to the small pitch-size of nano-bars [29]. Secondly, the crossbar structure can be stacked on top of the active transistor layer in a 3D fashion [32], [33], [40], which further reduces the area overhead and improves communication bandwidth. Fig. 3(a) shows the approach to fabricate the CBRAM-crossbar structure within the interconnect layers, where the CBRAM devices are deposited at the bottom of the copper vias [32]. Another approach of CBRAM-crossbar fabrication, shown in Fig. 3(b), is to stack the crossbar structure on top of the interconnect layer, which incurs the least modifications on the conventional CMOS process [33].

Fig. 4. Conventional crossbar read operation with incurred sneak-path issue

The conventional voltage-divider based readout design is shown in Fig. 4, where R_l follows $\frac{R_{off}}{R_l} = \frac{R_l}{R_{on}}$. As such, V_o will be logic-1 when target-cell is in low-resistance-state (LRS) and be logic-0 when in high-resistance-state (HRS). Note that all the unselected word-lines and bit-lines need to be floated to avoid sensing current branches before reaching R_l . This will incur sneak-path issue. When the target-cell is in HRS and is surrounded by other cells in LRS, there will be a significant leaked current flowing through the neighboring cells, which may lead to misinterpretation of the stored bit. The sneak-path issue can be addressed by adding a selection device for each CBRAM device [41], [42]. The selection device works like bidirectional-diode to ensure no current flowing through paths with more than one CBRAM device.

Instead of applying selection devices, alternative operations are proposed in this paper for the CBRAM-crossbar readout circuit to avoid the sneak-path issue. The *write* and *read* operations for CBRAM-crossbar are shown in Fig. 5(a) and 5(b), respectively. To write CBRAM-crossbar, the bias-voltage v_w needs to be applied on the designated cell through multiplexed voltage selector. As such, the corresponding word-line and bit-line are applied with $v_w/2$ and $-v_w/2$ voltages, respectively. By controlling the voltage level and polarity, one can determine the write-speed and change the on/off states of the cell. The cells that are half-selected will not change their states. To read CBRAM-crossbar, the bias-voltage v_r (normally $v_r < v_w$) is applied on the corresponding word-line. By measuring the current of designated column, the on/off state of target-cell can be detected. Note that because all unselected word-lines and bit-lines are grounded rather than

floated, the voltage-divider based readout scheme is incapable to function. As such, more sophisticated readout circuit design is required and is discussed later in this section.

Fig. 5. Crossbar operations and peripheral circuits

Fig. 5(c) further shows the peripheral circuit design for one CBRAM-crossbar. During write operation, the corresponding word-line and bit-line are applied with writing voltage through address decoding. The voltage selection is done by the row drivers and column drivers, which switch among different voltage supplies according to the command issued. During read operation, word-lines still need voltage selection while bit-lines are switched to current sensing circuit.

Fig. 6. Readout circuit design for CBRAM crossbar structured memory

To sense the current of designated column in the crossbar structure, readout circuit illustrated in Fig. 6 is proposed in this paper. The sensing procedure is done by two steps here. Firstly, a current mirror is deployed to amplify the current determined by the state of the target-cell. The bias current I_{bias} is applied to ensure that transistor M1 works in saturation region. The bias voltage V_{bias} has to be deliberately chosen according to I_{bias} to achieve a virtually grounded node A, i.e. a 0V column voltage required in Fig. 5(b). Otherwise, the current to detect will have branch at node A to other cells

at same column, which will weaken the current signal and thus lead to a degradation of sensing accuracy. Secondly, the amplified current signal is converted into voltage signal, and is further compared with the reference voltage to decide and output on/off state denoted by different logic levels.

Another issue in readout operation is the severe device resistance variations. It is reported that $5x \sim 100x R_{off}$ and $2x \sim 10x R_{on}$ variations can be expected based a variety of measurement data [43]. For one crossbar array, there exists a valid reference voltage in the readout circuit only when the minimum R_{off} of all CBRAM devices is greater than the maximum R_{on} of all devices, i.e.

$$\min(R_{off}) > \max(R_{on}) \quad (5)$$

which, however, may not hold true if variation is significant. A crossbar with device resistance variations that cannot meet above condition is called a failure. As such, the sense amplifier is not able to guarantee successful readout for every operation. In the experiment section (Section VI.B), we will show the failure rate for different crossbar array sizes, and possible solutions to address such issue when failure rate is high.

B. CBRAM-Crossbar Performance Modeling

Based on the circuit design for CBRAM-crossbar, here we propose its system-level delay, power and area models like the CMOS memory design tool CACTI [20]. We further verify such models by the transistor-level SPICE-level simulator developed in the previous section.

1) *Delay Model*: Generally, the time needed for write operation in CBRAM-crossbar is the signal propagation delay on wires (i.e. word-line and bit-line), denoted by D_{wire}^w , plus the device switching time $D_{switching}$; for read operation, it is composed of wire delay D_{wire}^r and sensing delay $D_{sensing}$. Therefore we have

$$D_{write} = D_{wire}^w + D_{switching} \quad (6)$$

$$D_{read} = D_{wire}^r + D_{sensing}. \quad (7)$$

Different from DRAM/SRAM cell, the CBRAM device has asymmetrical write/read-delays. Since the write operation requires physical change of CBRAM cell, i.e. the shape-morphing of the conductive filament as illustrated in Fig. 2, it usually takes much longer time than simply detecting the resistance of CBRAM cell in the read operation. The CBRAM switching time $D_{switching}$ can be obtained from the proposed CMOS-CBRAM simulator, where high accuracy can be achieved thanks to the use of physical model introduced in section III-A.

Due to the existence of leakage current in crossbar structure, CBRAM cells at different positions suffer from different amount of reduced applied voltages introduced by IR-drop along the word-line and bit-line. Since the switching time is very sensitive to even the slightest of voltage deviation from the expected value, CBRAM cells at different positions of the crossbar array exhibit different values of switching time. Therefore, the exponential relation between $D_{switching}$ and the applied voltage v_w , represented as a lookup table, is built into CACTI.

Fig. 7. RC-delay model of one CBRAM-crossbar for: (a) read operation; (b) write operation

Different from SRAM/DRAM sensing schemes where the subtle voltage/current swing or capacitor driving signal needs to be detected, the current signal in Fig. 6 is amplified and driven by power source thus the sensing can be performed really fast. Note that the sensing delay can also be obtained by performing SPICE-level simulation. In the following, we focus on the modeling and calculation of wire delay D_{wire} for read/write operations.

For conventional 1T-1R structure, word-line delay can be calculated by distributed RC-line delay and the bit-line delay can be estimated by Seevinck model [44]. However, although the same approach has been applied to estimate the wire delay of crossbar structure in [45], it lacks accuracy due to the following reasons. Firstly, since there is no transistor in a crossbar, the word-line and bit-line delay are symmetric. In other words, the word-line and bit-line delay have to be modeled in the same manner. Moreover, the leakage current of cells along word-line and bit-line is a phenomenon specific to crossbar structure, which is not considered by conventional RC-line delay model. The leakage current will weaken the driving ability of row and column drivers, and hence a longer delay can be predicted compared to the conventional RC-line delay.

Fig. 7(a) and Fig. 7(b) illustrate the crossbar delay model for read/write operation, with leakage path of cell i along the word-line and bit-line modeled as parallel R_{li} , whose value is its corresponding CBRAM on/off resistance. For the read operation, the bit-line is virtually grounded as illustrated in Fig. 5, thus only the word-line delay for the propagation of v_r contributes to the wire delay. For the write operation, both word-line driving voltage $v_w/2$ and bit-line $-v_w/2$ will propagate to target cell, and the total wire delay is determined by the slower one. Therefore, for a CBRAM-crossbar with m rows and n columns, the worst case wire delay for both read operation and write operation can be calculated respectively by

$$D_{wire}^r = \alpha RCn^2 \quad (8)$$

$$D_{wire}^w = \max(\alpha RCn^2, \alpha RCm^2) \quad (9)$$

where R and C are parasitic resistor and capacitor in unit length similar as the distributed RC-line model. Note that when the CBRAM device is scaled down, the R and C will be reduced accordingly, which produces a more smaller wire delay. In addition, the location-dependent switching speed

issue incurred by the IR-drop along the word-line and bit-line will be relieved as well.

Compared to the conventional RC-line delay expression, a fitting parameter α has been added to approximate the expected longer delay due to the CBRAM-crossbar structure, such as the effect introduced by R_{li} in Fig. 7. Practically, α can be obtained by fitting with a few samples obtained by simulating entire CBRAM-crossbar in different sizes using the developed SPICE-like simulator.

2) *Power Model*: The energy per write-access for CBRAM-crossbar is composed of two parts: the energy consumed to switch the target cell state, and the energy dissipated along the word-line and bit-line. Consider a CBRAM-crossbar with m rows and n columns, the write-access energy E_{write} can be calculated as:

$$E_{write} = E_{switching} + E_{static}^w + E_{dynamic}^w \quad (10)$$

where $E_{switching}$ is the energy for changing the target CBRAM cell state, and needs to be obtained through the developed SPICE-like simulator since its dynamics is hard to be approximated using resistor; $E_{dynamic}^w$ is the energy for charging parasitic capacitors along word-line and bit-line; and E_{static}^w is Joule heat dissipated on the half-selected CBRAM cells along word-line and bit-line, and can be modeled as resistors since their resistance values remain constant during write operation.

The E_{static}^w and $E_{dynamic}^w$ can be calculated by

$$E_{static}^w = \left(k \cdot \frac{v_w^2}{4 \cdot R_{on}} + l \cdot \frac{v_w^2}{4 \cdot R_{off}} \right) \cdot D_{write} \quad (11)$$

$$E_{dynamic}^w = \frac{1}{8} C \cdot v_w^2 \cdot (m + n - 2) \quad (12)$$

where v_w is the write-voltage, R_{on} and R_{off} are the on/off state resistance of CBRAM, D_{write} is the crossbar write-delay, C the distributed unit capacitance of crossbar wire, k and l the numbers of CBRAM cells in on-state and off-state respectively along the path, following $k + l = m + n - 2$.

Similarly, the read-access energy E_{read}^r for crossbar can be calculated by

$$E_{read} = E_{static}^r + E_{dynamic}^r \quad (13)$$

Note that for read operation, all the bit-lines are virtually grounded and only cells in the target row consume power as shown in Fig. 5, thus we have

$$E_{static}^r = \left(k \cdot \frac{v_r^2}{R_{on}} + l \cdot \frac{v_r^2}{R_{off}} \right) \cdot D_{read} \quad (14)$$

$$E_{dynamic}^r = \frac{1}{2} C \cdot v_r^2 \cdot n \quad (15)$$

where v_r is the read-voltage, and D_{read} is the crossbar read-delay. Moreover, k and l are the numbers of CBRAM cells in on-state and off-state respectively in the target row, where $k + l = n$. Note that the scalability of the CBRAM device is beneficial to dynamic power reduction. When scaled down, the smaller pitch size will reduce the wire capacitance with the power reduction at the word-line and bit-line.

3) *Area Model*: The area model is consisted of two parts: A_c for the pure area crossbar structure and A_p for corresponding CMOS peripheral circuits. Utilizing the 3D integration technique shown in Fig. 3, the crossbar is stacked over the active layer where its peripheral circuits are located. As such, the total area becomes

$$A = \max(A_p, A_c). \quad (16)$$

For a CBRAM-crossbar with M rows and N columns, its area can be calculated by

$$A_c = M \cdot N \cdot L_{pitch}^2 \quad (17)$$

where L_{pitch} is the nano-bar pitch size, determined by the technology node of the CBRAM device. Therefore, at advanced technology node, extremely small CBRAM-crossbar area can be achieved due to the scalability of the CBRAM device. Besides the reduced wire delay inside the CBRAM-crossbar, the addressing delay outside the CBRAM-crossbar can be greatly reduced as well, which together will lead to significant memory access latency reduction. Note that for the peripheral circuits the similar area model can be developed based on [20].

C. Design Space Exploration

In this part, based on the developed CBRAM-crossbar models, we further show how to optimize the 3D hybrid memory system by performing design space exploration under different optimization objectives.

Given design freedoms of parameters shown in Table I, the total capacity of SRAM, DRAM or CBRAM is calculated by

$$C_M = N_b \cdot C_b = N_b \cdot N_m \cdot C_m = N_b \cdot M \cdot N \cdot m \cdot n \quad (18)$$

where $M \cdot N$ is the dimension of mat array for each bank, and $m \cdot n$ is the dimension of crossbar array for each mat.

As indicated by equations (21) and (18), different combinations of bank number, mat array dimension and crossbar array dimension need to be explored for the optimal performance. For example, based on equation (21) it is desired to maximize the data-bus width W_d and CBRAM working frequency f_b . There are several design constraints to be considered as illustrated below one by one as below:

Data bandwidth constraint: as mentioned in section II, since each CBRAM-crossbar supports only one-bit-write for each cycle, we need to design the multi-bit data to be distributed into different crossbars at the same row. Consequently, the data-bus width W_d must equal to N , which is the number of crossbars (i.e., mats) that each row contains.

Transition power constraint: during the hibernating and wakeup transition, the concurrent data migration will induce high transition power

$$\sum P_{bi} = \sum f_b W_{di} E_{mi} \leq P \quad (19)$$

where W_{di} and E_{mi} is the bus-width and energy consumed of bank i . The transition power needs to be limited by reliability concerns for hot-spots of thermal and supply-current.

TSV density constraint: since the TSVs for data transmission occupy certain area, diameter in 5um for example [46], they lead to larger footprint as well as difficulties in placement and routing. As a result, the average TSV density must be limited which leads to an upper-bound on data bandwidth

$$\frac{W_D + W_A}{A} \leq D \quad (20)$$

where A is the chip area.

Fig. 8. Design space exploration of 3D hybrid memory with CBRAM-crossbar

For our 3D hybrid memory system, one can perform design space exploration based on design parameters such as bank capacity C_b , data-migration bus bandwidth W_d and CBRAM-crossbar write-frequency f_b . Fig. 8 shows the performance-objective with different tradeoffs bounded by design constraints, which can be summarized as follows:

- For speed optimization i.e., minimizing hibernating/wakeup transition time, large bandwidth is desirable. As such, small C_b , large W_d and high f_b are preferred. However, it is limited by TSV density and power constraints.
- For power optimization i.e., minimizing transition power, small C_b and W_d architecture working in a low f_b is favorable, which is mainly limited by TSV density and speed constraints.
- For memory performance optimization i.e., minimizing memory SRAM/DRAM performance degradation due to high TSV density, small W_d and large C_b help reduce TSV density, which in turn alleviates memory performance degradation, mainly limited by speed constraint.

For example, regions I, II, III in Fig. 8 are corresponding to the limiting design constraints of TSV density, power and speed, respectively.

V. CBRAM BLOCK-LEVEL INCREMENTAL DATA RETENTION

In this section, the proposed *block-level* data retention is discussed in details for the 3D hybrid memory architecture with CBRAM-crossbar.

The data is migrated between memory blocks (i.e., banks) sequentially through dedicated 3D TSV buses. Compared to the *bit-level* data retention scheme [28], where each SRAM memory cell is associated with one neighboring FeRAM

cell with cell-wise controllers, our block-level approach can achieve much smaller area overhead since the data migration controller is shared by all memory cells of the same block. In addition, our block-level approach will not degrade the SRAM performance since no change is made inside the SRAM memory cell. Furthermore, [27] has a system-level data retention by updating check-points. Because the use of PCRAM limits the frequency and amount of data that are retained, the system in [27] can only keep system check-points between relatively longer time intervals, which is insufficient when more fine-grained data retention is required. Based on the CBRAM-crossbar structure, this section has introduced one block-level data retention, which includes the block-level memory controller with two operations: dirty-bit set-up and incremental write-back.

A. Dirty Bit Set-up

The target for data retention is to synchronize the data of CBRAM-crossbar with corresponding SRAM/DRAM contents at block-level. For any SRAM/DRAM bank i , the time needed to copy the data to CBRAM-crossbar is decided by

$$T_{Hi} = \frac{M_b}{f_b \cdot W_d} \quad (21)$$

where f_b is the write-frequency of CBRAM-crossbar memory limited by its latency, W_d is the bank-level data bandwidth, and M_b is the amount of data to be migrated, which equals to the bank capacity C_b in the brute-force approach.

Clearly, reducing M_b directly reduces the T_{Hi} . Considering the common law of locality [47], the system tends to access relatively local-memory regions during a given period of time. As such, between two successive power gating stages, only part of the content in CBRAM-crossbar and SRAM/DRAM becomes unsynchronized, which is denoted as *dirty data* throughout this paper. By only incrementally writing dirty data to CBRAM-crossbar, significant amount of migration data and power can be saved.

As shown in Fig. 1, to keep the dirty-data status information, an extra CBRAM-crossbar called *dirty-data pool* is embedded to each CBRAM-crossbar bank, where each bit in the pool, referred as *dirty bit*, indicates the dirty status of a few continuous bytes of data in SRAM/DRAM, referred as *dirty-data group*. Empirically, we design the group granularity G_d as the cache line size B_C for SRAM or the page size B_D for DRAM.

The dirty-bit set-up occurs simultaneously each time when the content of SRAM/DRAM is changed during active mode. As shown in Fig. 9, each CBRAM-crossbar bank *listens* to all memory write operations issues to its corresponding SRAM/DRAM bank to update the dirty pool during active mode. Once the SRAM/DRAM write-action is detected, the corresponding data group becomes dirty.

As such, the corresponding bit in the dirty pool needs to be SET. The dirty pool size C_p is decided by

$$C_p = \frac{C_b}{G_d} = p \cdot q \quad (22)$$

Fig. 9. Circuit diagram for dirty bit set-up at active mode

where p and q is the CBRAM-crossbar dimension of dirty pool. Therefore, the designated dirty bit position can be located by decoding the first $\log(p)$ and the following $\log(q)$ bits of the physical memory write-address, respectively.

B. Incremental Write-back

Fig. 10. Circuit diagram for dirty-data write-back at hibernating transition

The flow to write back dirty SRAM/DRAM data to CBRAM-crossbar during hibernating transition is illustrated in Fig. 10. Specifically, one address counter is used to check the status of all dirty bits in the dirty-data pool. Once the dirty bit in SET state is detected, the corresponding data group needs to be copied to the CBRAM-crossbar. Due to the limited data-bus bandwidth W_d , the group data of size G_d needs to be written back to CBRAM-crossbar in several cycles. As such, the address counter generates the memory address of the next

piece of data to be copied from SRAM/DRAM by adding W_d -offset each cycle, and the read-signal to DRAM and write-signal to CBRAM-crossbar are issued for data migration. Once finished, the corresponding dirty bit is RESET. During wakeup transition, all data will be copied from CBRAM-crossbar back to SRAM/DRAM with similar hardware supports.

Here we will discuss how the use of CBRAM-crossbar is able to instinctively facilitate the dirty-data write-back without additional design efforts. As discussed above, we can see that intensive bit operations are required for dirty flags update and read out. For the conventional for 1T-1R structured memory, accesses are done in the unit of byte or word. Therefore to read a bit, the byte or word which contains the target bit is first read out, then OR, AND or SHIFT instruction is further applied to obtain one bit; and similarly to write a bit, the byte or word is first read out, then merged with the bit by OR or AND operation; and finally the bit-modified byte or word can be written back.

As such, the bit operations completed in multiple cycles may not be able to meet the real-time dirty flag update requirement. Also, significant power overhead may be incurred. Therefore, without additional design efforts, the byte-addressable or word-addressable conventional memory is not suitable for dirty-data write-back where intensive bit operations are required. On the other hand, as discussed in section IV-A, bit operation is the instinctive way that CBRAM-crossbar is operated. Byte or word operations are achieved by multiple identical CBRAM-crossbar units to work in parallel. In other words, both bit operations required by dirty-data pool and word operations can co-exist by using identical CBRAM-crossbar units naturally. Additionally from the physical design point of view, each CBRAM-crossbar block at top tier in Fig. 1 needs to have smaller size compared to their counterpart memory block in other tiers. This requirement is to ensure one vertical data path between the pairs, and can be achieved since the CBRAM-crossbar has very high density.

VI. EXPERIMENT RESULTS

This section shows the simulation experiment evaluation of the 3D hybrid memory from twofold. Firstly, the evaluation and optimization of the stacked CBRAM-crossbar memory is discussed in section VI-B. Based on the optimized CBRAM-crossbar memory design, the comparison between block-level data retention with the other schemes is discussed in section VI-C.

A. Experiment Settings

To evaluate the CBRAM-crossbar memory performance, the overall CBRAM-crossbar platform is implemented by C++ on Linux. All presented results are computed on an Intel Core i5 with 2.6GHz clock frequency and 4GB of RAM.

Firstly, one CMOS-CBRAM SPICE-like simulator is developed with CBRAM models. The model parameters k_1 , k_2 , k_3 and k_4 are based on [9] with 1.8V voltage supply. $2M\Omega$ and $1G\Omega$ are assumed as thresholds for *on* and *off* states of CBRAM, respectively. A pitch size of 100nm is assumed for the crossbar structure, with $50nm \times 50nm$ cross-section area

of nano-wire made by copper. Multiple supply voltages are used, where 65nm technology node with 1.2V is assumed for CMOS logic and 1.8V for CBRAM-crossbar operations.

Next, CBRAM-crossbar the power, delay and area models proposed in IV-B are verified by the SPICE-like simulator. Such verified models are further integrated into the memory design tool CACTI [20] to perform the system-level evaluation for the proposed 3D hybrid memory system. Therefore, by deploying the extended CACTI for CBRAM-crossbar, one can obtain the memory block performance data, which is later utilized in the data-retention behavioral model.

Finally, the system under investigation is then composed of one 2MB level-2 SRAM cache, one 64MB embedded DRAM and one 66MB CBRAM for data retention, which are vertically connected by TSVs with $50\mu\text{m}$ length. The TSV power and delay models in [48] are integrated into the extended CACTI. Here we adjust the design parameters C_b , W_d and f_b under constraints of the maximal 10W transition power, maximal $60/\text{mm}^2$ TSV density and maximal 3ms transition speed.

B. Stacked CBRAM-Crossbar Memory

We first discuss the advantage of CBRAM-crossbar memory, and then show the design space exploration to obtain the optimal architecture-level parameters.

1) *CBRAM-crossbar Modeling*: One recently developed SPICE-level simulator for NVM devices [16], [17] has been extended with CBRAM device based on section III-A and III-B. The CBRAM device model parameters are extracted from [9]. Besides, 100nm L_{pitch} , 1.8V v_w , 0.9V v_r , and 6.4fF distributed C are utilized as settings.

Fig. 11. CBRAM model validation against the published measurement data [3], [31]

Fig. 11 shows the validation of CBRAM device model illustrated in section III-A against the published measurement data [3], [31]. It can be seen that the exponential relation between applied voltage and switching time is valid. The transient analysis for CBRAM switching is conducted in the proposed SPICE-like simulator with results shown in Fig. 12. The CBRAM is initialized at off-state, and 1.8V voltage is applied to set the device (off to on). The internal state variable s grows from 0.2 to around 4, indicating that the conductive filament first grows its height as a cone, reaches the top

Fig. 12. Transient response of CBRAM device Set and Reset when v_w amplitude of 1.8V is applied

electrode, and then transforms towards cylinder. This is also confirmed by the change of resistance from $\text{G}\Omega$ scale to $\text{M}\Omega$ scale. The CBRAM has been successfully set after around 5ns, then a -1.8V voltage is applied to reset the device, where a reverse process can be observed.

Fig. 13. Monte Carlo simulation of 100×100 crossbar-array with device resistance variations

The Monte Carlo simulation with 5000 times is conducted in order to investigate the device resistance variations impact on readout operation, which is shown in Fig. 13. In each experiment, a 100×100 crossbar-array is generated, where R_{off} and R_{on} of all CBRAM devices have normal distribution with μ of $1\text{G}\Omega$ and $2\text{M}\Omega$, and deviation δ that follows the variation in [43], i.e. 10x for R_{on} and 100x for R_{off} . The $\max(R_{on})$ and $\min(R_{off})$ are calculated and then used as the x-axis and y-axis values for each point. The whole domain is divided into two regions, pass region and fail region, separated by the dashed line $\max(R_{on}) = \min(R_{off})$. It can be observed that there are 9 failed cases, which leads to a 0.18% failure rate for the 100×100 crossbar size. In other words, 9 out of 5000 crossbars in this size do not have the ideal reference voltage value.

The Monte Carlo simulations for crossbar sizes starting at 50×50 to 600×600 with a step of 50 are conducted with results shown in Table II. It can be observed that the failure

rate increases as the crossbar size enlarges. For large crossbar arrays where the failure rate is not negligible, robust readout schemes namely the self-reference [49] and runtime ECC/ECP [50] can be applied to increase the readout reliability. Thus it is favorable to limit the crossbar size within a few hundreds by a few hundreds.

TABLE II
DEVICE RESISTANCE VARIATION CAUSED SENSE AMPLIFIER FAILURE RATE FOR DIFFERENT CROSSBAR-ARRAY SIZES

Array size	50^2	100^2	150^2	200^2	250^2	300^2
Failure rate	0.05%	0.18%	0.85%	0.95%	2.65%	3.20%
Array size	350^2	400^2	450^2	500^2	550^2	600^2
Failure rate	4.30%	5.50%	5.55%	7.65%	10.85%	13.40%

Fig. 14. Verification of proposed CBRAM-crossbar specific wire delay model against simulation results, with fitting parameter $\alpha = 1.2$

Fig. 14 shows the verification of the proposed crossbar specific delay model against accurate simulation results obtained through the developed SPICE-like simulator. It can be observed that the proposed model with fitting parameter $\alpha = 1.2$ is able to predict the crossbar delay well. As expected in section IV-B1, the crossbar wire delay, calculated by $1.2RCn^2$ therefore, is more than twice longer compared to the conventional distributed RC-line delay calculated by $0.5RCn^2$. In other words, an error of more than 50% will be incurred if the conventional distributed RC-line delay model is used instead.

The verification of the power model against simulation results is shown in Fig. 15. The simulation results are obtained by simulating $n \times n$ square CBRAM-crossbars with different n values, where n is used to denote crossbar size. It can be observed that the proposed model for read/write operations is able to capture the trend with minor error when crossbar size increases.

2) *Stacked CBRAM-crossbar Memory*: Table III evaluates and compares the performance of the stacked CBRAM-crossbar memory with the other memory technologies for the same capacity of 16MB. The data for SRAM and DRAM are generated by CACTI with default settings, and PCRAM data is extracted from PCRAMsim [39]. As shown in Table III, mainly due to the fast device-level accessing speed by CBRAM device as well as the high density of crossbar structure, the CBRAM-crossbar performance, especially the

Fig. 15. Verification of proposed CBRAM-crossbar specific power model for read/write operations against simulation results

TABLE III
PERFORMANCE COMPARISON OF 16MB SRAM, DRAM, PCRAM, AND CBRAM MEMORIES

Feature	SRAM	DRAM	PCRAM	CBRAM
area (mm^2)	27.4	4.19	3.77	2.33
read latency (ns)	3.43	2.25	2.54	3.90
write latency (ns)	3.43	2.55	RESET 42.54 SET 102.54	8.01
read energy (nJ)	0.83	0.61	0.73	1.8
write energy (nJ)	0.75	0.61	RESET 6.66 SET 2.28	2.0

TABLE IV
OPTIMIZED PERFORMANCE FOR DIFFERENT DESIGN OBJECTIVES

Objective	Speed	Power	MP
C_b	256KB	256KB	4MB
W_d (bits)	16	8	128
P_H/P_W (W)	2.3/9.0	1.8/1.8	4.3/7.1
T_H/T_W (ms)	0.5/1.5	3/3	1.9/3
TSV density ($/\text{mm}^2$)	60	48	26

accessing latency for read/write operations, is already close to DRAM, which shows its potential for the future application as the main memory. Moreover, compared to PCRAM, the CBRAM-crossbar shows 9x faster write-latency, 1.6x smaller area, 4.5x less write-energy per access, and 1.5x slower read-latency on average. Only a slightly slower read-latency is observed for the CBRAM-crossbar.

3) *Design Space Exploration and Optimization*: As discussed in section IV-C, when employing CBRAM-crossbar for hybrid memory design, one needs to decide the optimal architecture-level parameters for performance objectives under certain constraints. As one example, we show the procedure of transition time (i.e., speed) optimization. Fig. 16 shows the trend of TSV density and P_H/P_W with C_b/W_d . Among the available combinations, the one with C_b of 256KB and W_d of 16-bits is chosen since it results in largest W_D and power margin when increasing frequency and when satisfying the constraints defined in IV-C. As such, when working at its maximal allowed frequency, 0.5ms T_W and 1.5ms T_h can be achieved.

Table IV demonstrates the optimal results for different performance objectives. An optimal 0.5ms hibernating transition

Fig. 16. (a) TSV density and (b) mode transition power under different architecture-level parameters

time and 1.5ms wakeup transition time is achieved for the speed optimization. For transition power optimization, 5x less wakeup transition power and 1.6x less hibernating transition power are achieved with 6x wakeup transition time and 2x hibernating transition time penalties. For memory performance (MP in Table IV) optimization, the lowest TSV density is obtained as the minimized memory performance degradation in normal active mode. The design exploration leaves designer the freedom to choose different design parameters based on system requirements. The configuration with speed optimized is used in the following experiments.

C. 3D Hybrid Memory with Data Retention

In this part, we evaluate the proposed block-level data retention with the use of incremental write-back. We also present the comparison with system-level and bit-level data retention respectively.

Fig. 17. Hibernating power and time reduction by incremental dirty-data write-back

1) *Dirty Data Write-back Strategy*: In order to evaluate the dirty-data write-back strategy, a set of benchmark programs are selected from SPEC2006 suite and are run in gem5 simulator [51], where memory-access traces are generated. As the advantage of the dirty-data write-back strategy may

depend on the memory access patterns of executed programs, the benchmarks with different memory access characteristics are picked in general. For examples, *mcf* and *lbn* have high cache-miss rates while *h264* and *namd* have low cache-miss rates; *perlbench* and *gcc* have intensive store-instruction intensive while *astar* and *namd* have low store-instruction [52]. For each benchmark, its dirty-data flags are updated according to the memory-access trace. Then, the dirty-data write-back strategy is deployed to evaluate the power saving and speedup during data migration. Fig. 17 compares the hibernating transition power and time with and without using dirty-data write-back strategy. Averagely, system with dirty-data write-back strategy achieves 5x reduction in power and 1.5x reduction in time during the hibernating transition.

2) *Data-retention Performance Comparison*: Next, we compare the performance of data retention and leakage power in sleep mode for the following data retention schemes:

- *STD*: standard SRAM/DRAM without power gating.
- *DPG*: data-retentive power gating (DPG) of both SRAM [53] and DRAM [54] with reduced supply voltage.
- *PCRAM*: system-level data retention by PCRAM [27] which is used to keep the entire SRAM/DRAM contents.
- *FeRAM*: bit-level data retention by FeRAM [28].
- *CBRAM*: our proposed block-level data retention with incremental dirty-data write-back by CBRAM.

TABLE V
DATA-RETENTION PERFORMANCE COMPARISON FOR DIFFERENT LEAKAGE REDUCTION SCHEMES

Scheme	P_s^L/P_d^L (mW)	T_H/T_W (ms)	P_H/P_W (W/MB)
STD	209/220	NA	NA
DPG	21/22	1e-4	0
PCRAM	0/0	3.7/1.3	0.072/0.1
FeRAM	0/22	1	0.75
CBRAM	0/0	0.33/1.5	0.007/0.14

Table V shows the full comparison results in terms of sleep-mode leakage power for SRAM and DRAM (P_s^L/P_d^L), hibernating and wakeup transition time (T_H/T_W), and hibernating and wakeup transition power (P_H/P_W). The leakage power for SRAM/DRAM under standard scheme is generated by CACTI at 65nm technology node. Leakage power under DPG scheme is calculated by the leakage reduction factors reported in [53], [54] for both SRAM and DRAM. Block-level CBRAM based data-retention performance is derived from our platform, combining the architecture optimization results in Section VI-B3 and power and time reduction results in Section VI-C1. PCRAM based scheme performance estimation is based on [27], with PCRAM memory performance obtained from PCRAMsim [39]. For FeRAM based scheme, bit-to-bit data retention performance is extracted from [28]. Because all data is migrated concurrently in this scheme, the data retention power performance can be estimated by multiplication and speed performance is the same with bit-to-bit performance.

Due to the use of NVM for data retention, PCRAM, FeRAM and CBRAM based memory systems all outperform the STD and DPG schemes in terms of leakage power reduction. Moreover, our proposed 3D hybrid CBRAM-crossbar memory system allows a shut-down of both SRAM and DRAM during the power gating. As a result, compared to

the block-level PCRAM based data retention, we achieve 11x faster hibernating transition time and 10x smaller hibernating transition power with the same number of TSVs. As shown in Table III, the hibernating transition performance improvement comes from the advantageous CBRAM-crossbar memory performance and also the block-level incremental dirty-data write-back strategy. However, its the wakeup transition time and power is slightly inferior to PCRAM based scheme. This is mainly because the incremental dirty-data write-back strategy does not apply to the wakeup transition. In addition, the proposed CBRAM system also outperforms the FeRAM by 107x/5.4x hibernating/wakeup transition power saving, and around 3x faster hibernating transition time.

TABLE VI
CACHE PERFORMANCE COMPARISON BETWEEN BLOCK-LEVEL DATA RETENTION (CBRAM) AND BIT-LEVEL DATA RETENTION (FeRAM) IN ACTIVE MODE

Cache capacity	Feature	Block-level CBRAM	Bit-level FeRAM
128KB	access time (ns)	1	1.8 (-80%)
	access energy (nJ)	0.13	0.21 (-62%)
	area (mm^2)	0.82	7.6 (-827%)
512KB	access time (ns)	1.5	2.8 (-87%)
	access energy (nJ)	0.33	0.6 (-82%)
	area (mm^2)	3.8	28.6 (-642%)
2MB	access time (ns)	2.7	4.3 (-59%)
	access energy (nJ)	0.6	1.3 (-117%)
	area (mm^2)	14.2	113 (-696%)

Further illustrated in Table VI, when compared with the block-level CBRAM based data retention, the bit-level FeRAM based data retention induces the overwhelming cache performance degradation in normal active mode due to bit-wise controllers embedded in the SRAM. For bit-level FeRAM based data retention scheme, the area overhead for retaining one bit data is 6.1um^2 , which is only 0.36um^2 in our CBRAM system, estimated by the our developed CBRAM-crossbar based CACTI. Due to such a large area overhead, the FeRAM based system shows an 84% SRAM access time degradation, 74% SRAM access energy overhead on average. The performance of bit-level FeRAM based cache is derived from CACTI by replacing SRAM cell with FeRAM-SRAM cell pair. Therefore, our CBRAM-crossbar based data retention outperforms not only the system-level PCRAM based data retention but also the bit-level FeRAM based data retention.

VII. CONCLUSION

In this paper, the accurate CBRAM device model is developed for SPICE-like simulation, which provides area, access time and power model for the optimization of CBRAM-crossbar based memory design. As such, the design of 3D hybrid memory is explored based on the CBRAM-crossbar memory for the block-level data retention of SRAM/DRAM with the incremental dirty-data write-back.

Experiment results show that the proposed hybrid memory achieves 11x faster migration time and 10x less migration power for hibernating transition, when compared to the system-level 3D PCRAM-based data retention in [27] with the same amount of TSVs used. Moreover, the proposed hybrid memory also achieves 17x area reduction along with

2x SRAM performance improvement, average 56x transition power reduction respectively with similar transition time, when compared to the bit-level FeRAM-based data retention in [28].

REFERENCES

- [1] K. Tsuchida and et.al., "A 64Mb MRAM with clamped-reference and adequate-reference schemes," in *Proc. Int. Solid State Circuit Conf. (ISSCC)*. IEEE, 2010, pp. 258–259.
- [2] D. Schinke and et.al., "Computing with novel floating-gate devices," *Computer*, vol. 44, no. 2, pp. 29–36, 2011.
- [3] S. Dietrich and et.al., "A nonvolatile 2-mbit cbram memory core featuring advanced read and program control," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 42, no. 4, pp. 839–845, 2007.
- [4] Y. Wang and H. Yu, "Ultralow-power memory-based big-data computing platform by nonvolatile domain-wall nanowire devices," in *Proceedings of the 2013 ACM/IEEE international symposium on Low power electronics and design*. ACM, 2013.
- [5] B. Lee and et.al., "Phase-change technology and the future of main memory," *Micro*, vol. 30, no. 1, pp. 143–143, 2010.
- [6] Y. Wang and H. Yu, "Design exploration of ultra-low power non-volatile memory based on topological insulator," in *Nanoscale Architectures (NANOARCH), 2012 IEEE/ACM International Symposium on*, 2012, pp. 30–35.
- [7] D. Halupka and et.al., "Negative-resistance read and write schemes for STT-MRAM in 0.13 um CMOS," in *Proc. Int. Solid State Circuit Conf. (ISSCC)*. IEEE, 2010, pp. 256–257.
- [8] X. Wu and et.al., "Hybrid cache architecture with disparate memory technologies," in *Intl. Symp. Computer Architecture (ISCA)*, vol. 37, no. 3. IEEE, 2009, pp. 34–45.
- [9] S. Yu and H. Wong, "Compact modeling of conducting-bridge random-access memory (cbram)," *IEEE Trans. on Electron Devices*, vol. 58, no. 5, pp. 1352–1360, 2011.
- [10] S. Sheu and et.al., "A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability," in *Proc. Int. Solid State Circuit Conf. (ISSCC)*. IEEE, 2011, pp. 200–202.
- [11] C. Gopalan and et.al., "Demonstration of Conductive Bridging Random Access Memory (CBRAM) in Logic CMOS Process," in *Proc. Memory Workshop (IMW)*, 2010, pp. 1–4.
- [12] J. Jameson and et.al., "Quantized Conductance in $\text{Ag/GeS}_2/\text{W}$ Conductive-Bridge Memory Cells," *IEEE Electron Device Lett.*, vol. 33, no. 2, pp. 257–259, 2012.
- [13] Y. Wang and et.al., "Design of low power 3d hybrid memory by non-volatile cbram-crossbar with block-level data-retention," in *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*. ACM, 2012, pp. 197–202.
- [14] M. Kund and et.al., "Conductive bridging ram (cbram): An emerging non-volatile memory technology scalable to sub 20nm," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*. IEEE, 2005, pp. 754–757.
- [15] N. Gilbert, C. Gopalan, and M. Kozicki, "A macro model of programmable metallization cell devices," *Solid-state electronics*, vol. 49, no. 11, pp. 1813–1819, 2005.
- [16] W. Fei, H. Yu, W. Zhang, and K. Yeo, "Design exploration of hybrid cmos and memristor circuit by new modified nodal analysis," *IEEE Trans. on Very Large Scale Integration Systems*, no. 20, pp. 1012–1025, 2012.
- [17] Y. Shang, W. Fei, and H. Yu, "Analysis and modeling of internal state variables for dynamic effects of nonvolatile memory devices," *IEEE Trans. on Circuits and Systems I*, no. 59, pp. 1906–1918, 2012.
- [18] Y. Wang, W. Fei, and H. Yu, "Spice simulator for hybrid cmos memristor circuit and system," in *Cellular Nanoscale Networks and Their Applications (CNNA), 2012 13th International Workshop on*. IEEE, 2012, pp. 1–6.
- [19] Y. Wang, Y. Shang, and H. Yu, "Design of single saw-tooth-pulse based stt-ram readout circuit by nvm spice," in *IEEE Non-Volatile Memory Technology Symposium (NVMTS)*. IEEE, 2012.
- [20] Cacti: An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model. [Online]. Available: <http://www.hpl.hp.com/research/cacti/>
- [21] G. Loh, "3d-stacked memory architectures for multi-core processors," in *Proc. Int. Symp. Computer Architecture (ISCA)*, vol. 36, no. 3. IEEE, 2008, pp. 453–464.

- [22] N. Minas and et.al., "3d integration: Circuit design, test, and reliability challenges," in *Intl. On-Line Testing Symp. (IOLTS)*. IEEE, 2010, pp. 217–217.
- [23] T. Thorolfsson, N. Moezzi-Madani, and P. Franzon, "Reconfigurable five-layer three-dimensional integrated memory-on-logic synthetic aperture radar processor," *Computers & Digital Techniques, IET*, vol. 5, no. 3, pp. 198–204, 2011.
- [24] Y. Wang and et.al., "Design exploration of 3d stacked non-volatile memory by conductive bridge based crossbar," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*. IEEE, 2012, pp. 1–6.
- [25] X. Huang and et.al., "A nanoelectromechanical-switch-based thermal management for 3-d integrated many-core memory-processor system," *Nanotechnology, IEEE Transactions on*, vol. 11, no. 3, pp. 588–600, 2012.
- [26] C. Hua, T. Cheng, and W. Hwang, "Distributed data-retention power gating techniques for column and row co-controlled embedded sram," in *Proc. Intl. Workshop on Memory Technology, Design, and Testing*. IEEE, 2005, pp. 129–134.
- [27] J. Xie, X. Dong, and Y. Xie, "3d memory stacking for fast checkpointing/restore applications," in *Intl. 3D Systems Integration Conf. (3DIC)*. IEEE, 2010, pp. 1–6.
- [28] M. Koga and et.al., "First Prototype of a Genuine Power-Gatable Reconfigurable Logic Chip with FeRAM Cells," in *Proc. Intl. Conf. on Field Programmable Logic and Applications (FPL)*. IEEE, 2010, pp. 298–303.
- [29] W. Lu and et.al., "A functional hybrid memristor crossbar-array/cmos system for data storage and neuromorphic applications," *Nano Lett.*, no. 12, p. 389, 2012.
- [30] M. Kozicki and et.al., "Programmable metallization cell memory based on ag-ge-s and cu-ge-s solid electrolytes," in *Non-Volatile Memory Technology Symposium, 2005*. IEEE, 2005, pp. 7–pp.
- [31] U. Russo and et.al., "Study of multilevel programming in programmable metallization cell (pmc) memory," *Electron Devices, IEEE Transactions on*, vol. 56, no. 5, pp. 1040–1047, 2009.
- [32] S. Kaeriyama and et.al., "A nonvolatile programmable solid-electrolyte nanometer switch," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 1, pp. 168–176, 2005.
- [33] M. Tada and et.al., "Nonvolatile crossbar switch using tio_x/tasioy solid electrolyte," *IEEE transactions on electron devices*, vol. 57, no. 8, pp. 1987–1995, 2010.
- [34] K. Tsunoda and et.al., "Bipolar resistive switching in polycrystalline tio₂ films," *Applied physics letters*, vol. 90, no. 11, pp. 113 501–113 501, 2007.
- [35] T. Sakamoto and et.al., "Electronic transport in ta₂o₅ resistive switch," *Applied Physics Letters*, vol. 91, no. 9, pp. 092 110–092 110, 2007.
- [36] C. Schindler and et.al., "Bipolar and unipolar resistive switching in cu-doped sio₂," *Electron Devices, IEEE Transactions on*, vol. 54, no. 10, pp. 2762–2768, 2007.
- [37] M. Haemori and et.al., "Impact of cu electrode on switching behavior in a cu/hfo₂/pt structure and resultant cu ion diffusion," *Applied Physics Express*, vol. 2, no. 6, p. 1401, 2009.
- [38] W. Guan and et.al., "Nonpolar nonvolatile resistive switching in cu doped zro₂," *Electron Device Letters, IEEE*, vol. 29, no. 5, pp. 434–437, 2008.
- [39] X. Dong, N. Jouppi, and Y. Xie, "Pcrasim: System-level performance, energy, and area modeling for phase-change ram," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*. IEEE, 2009, pp. 269–275.
- [40] M. Ben-Jamaa and et.al., "Silicon nanowire arrays and crossbars: Top-down fabrication techniques and circuit applications," *Science of Advanced Materials*, vol. 3, no. 3, pp. 466–476, 2011.
- [41] W. Park and et.al., "A pt/tio₂/ti schottky-type selection diode for alleviating the sneak current in resistance switching memory arrays," *Nanotechnology*, vol. 21, no. 19, p. 195201, 2010.
- [42] K. Gopalakrishnan and et.al., "Highly-scalable novel access device based on mixed ionic electronic conduction (miec) materials for high density phase change memory (pcm) arrays," in *VLSI Technology (VLSIT), 2010 Symposium on*. IEEE, 2010, pp. 205–206.
- [43] A. Chen and M. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," in *Reliability Physics Symposium (IRPS), 2011 IEEE International*. IEEE, 2011, pp. MY–7.
- [44] E. Seevinck, P. van Beers, and H. Ontrop, "Current-mode techniques for high-speed vlsi circuits with application to current sense amplifier for cmos sram's," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 26, no. 4, pp. 525–536, 1991.
- [45] C. Xu, X. Dong, N. Jouppi, and Y. Xie, "Design implications of memristor-based ram cross-point structures," in *Proc. Design Automation and Test in Europe (DATE)*. IEEE, 2011, pp. 1–6.
- [46] International technology roadmap of semiconductor. [Online]. Available: <http://www.itrs.net>
- [47] J. Hennessy and D. Patterson, *Computer architecture: a quantitative approach*. Morgan Kaufmann Pub, 2011.
- [48] Y. Shang and et.al., "Thermal-reliable 3d clock-tree synthesis considering nonlinear electrical-thermal-coupled tsv," in *Proc. Asia and South Pacific Design Automation Conf. (ASPDAC)*, 2013.
- [49] G. Jeong and et.al., "A 0.24- μ m 2.0-v 1t1mtj 16-kb nonvolatile magnetoresistance ram with self-reference sensing scheme," *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 11, pp. 1906–1910, 2003.
- [50] S. Schechter and et.al., "Use ecp, not ecc, for hard failures in resistive memories," in *ACM SIGARCH Computer Architecture News*, vol. 38, no. 3. ACM, 2010, pp. 141–152.
- [51] N. Binkert, B. Beckmann, G. Black, S. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. Hower, T. Krishna, S. Sardashti et al., "The gem5 simulator," *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, 2011.
- [52] S. Bird and et.al., "Performance characterization of spec cpu benchmarks on intel core microarchitecture based processor," in *SPEC Benchmark Workshop*, 2007.
- [53] H. Qin and et.al., "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. Int. Symp. on Quality Electronic Design (ISQED)*. IEEE, 2004, pp. 55–60.
- [54] T. Nagai and et.al., "A 65nm low-power embedded DRAM with extended data-retention sleep mode," in *Proc. Int. Solid State Circuit Conf. (ISSCC)*. IEEE, 2006, pp. 567–576.

Yuhao Wang (S'12) received the B.S. degree in microelectronics engineering from Xi'an Jiaotong University, Xi'an, China in 2011. He is currently working towards Ph.D. degree with School of Electrical and Electronic Engineering at Nanyang Technological University. His research interests include emerging non-volatile memory and 3D-IC technology for exa-scale computing platform.

Hao Yu (M'06) received the B.S. degree from Fudan University, China; and Ph. D degree from the Electrical Engineering Department, University of California, Los Angeles (UCLA), USA with major in integrated circuit and embedded computing. He was a Senior Research Staff at Berkeley Design Automation (BDA). Since October 2009, he has been an Assistant Professor at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His primary research interest is in CMOS emerging technologies such as 3D-IC and RF-IC at nano-tera scale. He has 94 top-tier peer-reviewed publications, 2 books, Best Paper Award from the ACM-TODAES'10, Best Paper Award nominations in DAC'06, ICCAD'06, and ASP-DAC'12, Best Student Paper Competition Finalist (advisor) in SiRF'13, and Inventor Award from Semiconductor Research Cooperation (SRC) in 2008. He is associate editor and technical program committee member for a number of IEEE/ACM journals and conferences.

Wei Zhang (M'05) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively; and the Ph.D. degree in computer engineering from Princeton University, Princeton, NJ, in 2009. She is currently an Assistant Professor at the School of Computer Engineering, Nanyang Technological University, Singapore. Her research interests include embedded systems, reconfigurable computing, nanotechnology, electronic design automation, and network-on-chip. Dr. Zhang received a Princeton Research Prize and a Best Paper Award in 2007 and 2009, respectively.