ARIA: Additive ReRAM-based Integrity and Aging Monitoring for ICs

Thomas Schultz[1], Rashmi Jha[1], Brian Dupaix[2], Matt Casto[2]
[1]University of Cincinnati, Cincinnati, Ohio, United States, Email: schulttd@mail.uc.edu; jhari@mail.uc.edu
[2]Wright Patterson Air Force Base, Dayton, Ohio, United States

ABSTRACT This paper reports an approach for monitoring aging and integrity of CMOS circuits through additively manufactured Resistive Random-Access Memory (ReRAM) based test structures. MgO-based ReRAM devices demonstrated excellent temperature sensing and aging modalities with simultaneous storage of sensed temperature and age as a change in the resistive state. The Process Voltage Temperature (PVT) characteristics, aging, and temperature sensitivity of MgO-ReRAM devices were experimentally studied and modeled to capture resistance distributions and temperature-based modalities. This in-memory sensing feature of ReRAM was integrated with specially designed read circuitry using 180 nm CMOS technology, to produce a measurable change in spiking-frequency over the lifetime of the ReRAM under normal aging conditions with the underlying CMOS circuits. Large feature sizes were used so these circuits can be fabricated in-house in trusted foundry. Temporal changes in temperature of underlying CMOS circuit could be captured by instantaneous change in resistive state of ReRAM with local temperature fluctuations which translated to a change in read circuit output. The characteristics of this circuit is studied in detail using simulations. Due to additive integration of ReRAM and associated circuitry, this approach for aging and integrity monitoring (AIM) ensures large spatial and accurate temporal monitoring of underlying CMOS die with minimal loss of the functional chip area for these added security features. The passive, in-memory sensing, and non-volatile nature of ReRAM also ensures low-power consumption in these circuits. The devices resistance states and material composition are specific to every device preventing reverse engineering and tampering of the devices, thus making it an attractive approach for adding customized security and trust features in advanced CMOS nodes-based circuits.

INDEX TERMS Aging, Integrity, Monitoring, ReRAM

I. INTRODUCTION

The semiconductor industry has moved to global manufacturing as the complexity and expense of manufacturing Integrated Circuits (IC) components increases, specifically, beyond 45 nm Complementary Metal Oxide Semiconductor (CMOS) technology node. The extreme complexity of the industry provides a deeper explanation for this shift. The nonstop, consumer- and applications-driven demands for better power, performance, area, and cost (PPAC) of ICs requires a heavy investment in research and development (R&D) encompassing design, efficient and low-cost manufacturing, testing, assembling and packaging, and distribution. These demands have led to the consolidation of major semiconductor manufacturing facilities [1]. While this has benefitted fabless semiconductor design companies by providing access to advanced CMOS transistor technology nodes in a more cost-effective manner, it has also raised concerns about the integrity and security of these fabricated ICs. Globalization of the supply chain undoubtedly exposes ICs to various vulnerabilities. Particularly from IC fabrication standpoints, adversary can use compromised process technologies for IC fabrication. With scaling of feature sizes, maintaining process uniformity is becoming an increasingly challenging task. A slight variation in process parameters can cause significant impact on device performance, reliability, aging, and yield across the wafers.

To address this, foundries typically place dummy patterns on masks (a process referred to as dummification) to improve uniformity of processes, such as, Deep Reactive Ion Etching, Photolithography, Chemical Mechanical Polishing, Depositions, or Metallization's [2]. A slight modulation of features on mask can alter the uniformity of processes that can have significant impact on device and circuit performance. Additionally, dormant defects can be introduced in material stacks of devices or isolation layers that becomes active only under certain operational conditions (e.g. bias, temperature), thus, leading to device failure. Insertion of Trojan circuits during fabrication, or insertion of counterfeited dies during packaging are also possible.

Under these circumstances, we can assume an attack model where: (i) foundry uses compromised process technology leading to devices with compromised reliability not detected under normal burn-in tests, (ii) insertion of Trojans circuits during fabrication either during dummification by foundry or by malicious designer as part of IC design itself. An attacker, who is aware of these vulnerabilities, can leverage this knowledge and cause severe damages in several ways including: (i) accelerated aging of circuits during normal operations causing early failure of components, (ii) triggering of Trojan causing accelerated failure of circuits or leaking of information, (iii) modulation of circuit delays and power by triggering Trojans, (iv) altercation of data stored in on-chip memories or LUTs by intentional thermal variations, (v) alteration of parasitics, and (vi) isolation layer breakdown causing shorting of devices. These types of attacks become a concern as fabless design companies typically desire reliable and secure components with an expected lifetime of over 10 years.

Traditionally, IC designers have attempted age-aware designs and circuit failure predictions by capitalizing on the bias-temperature dependent changes in device switching characteristics over time. Previously employed methods for measuring reliability included individual device probing, ring oscillator (RO) frequency monitoring, and built-in self-test (BIST) structures. However, device probing increases measurement time significantly, lacks accuracy and requires an extensive measurement setup. To tackle this issue, various research works presented on-chip circuit reliability monitors for accurate measurement and statistical analysis [3,4]. The most common approach for IC integrity monitoring is a RO-based circuit. The output frequency of CMOS based RO circuit changes over time as the gate-oxide of CMOS transistors degrades due to Bias Temperature Instabilities (BTI) or hot carrier injection (HCI) under voltage and temperature stress. This provides sensing modalities by utilizing peripheral circuitry to measure the frequency changes [5]. However, there are inherent issues in this approach. For example, any portion of the RO circuit itself on an IC can be tampered with during the manufacturing process and their change in frequency over a 10-year lifespan is <10% which does not allow for a high level of accuracy when predicting the age of the circuit. Additionally, if one component of the RO fails then the entire circuit is unusable. RO circuits also consume significant static and active power which makes them unattractive for power constrained systems. To combat this several ROs are utilized in the circuit for aging with averages taken from the combined totals for age prediction.

Unlike a traditional aging monitors or BIST circuits that are more localized on-chip, when designing Aging and Integrity monitoring (AIM) circuits for ensuring trust and security features one has to also worry about the fact that during the run time of the circuit an adversary can attack any specific part of the circuit or any specific part of the circuit can fail at any time. This creates a need for having AIM circuitries with large spatial and temporal coverage of the CMOS IC design to ensure, at minimum, every crucial part of the circuit is monitored, as shown in Fig. 1(a). Typical IC devices such as the octal D-Flip Flop shown in Fig. 1(b) have thermal profiles which are not consistent across the entire chip and in this instance has a 5°C difference between the highest and lowest operating temperatures. This test was performed using a 5V bias and 1MHz clock with no other external inputs to map the base operating temperature of the IC. This indicates that some sections of the design with higher temperatures will fail sooner and thus must be monitored separately than the rest of the chip. Considering that each RO circuit will consume significant functional area on the chip as it needs to be integrated with the CMOS in the front end of the line (FEOL) process, it becomes expensive and impractical to deploy them for large spatial and temporal coverage [6]. This clearly leaves an area for improvement which can be addressed through arrays of Resistive Random-Access Memory (ReRAM) devices, as reported in this paper for the first time.

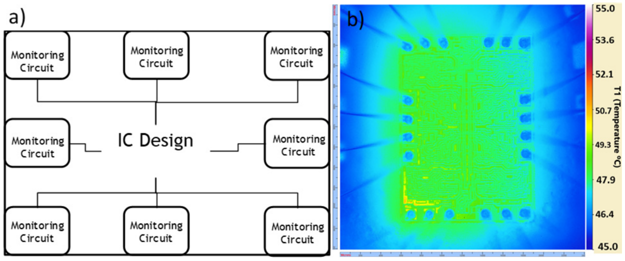

Fig. 1: Spatial coverage of an IC design with a) monitoring circuits taking large sections of functional area and b) temperature profile of an octal D-flip flop.

ReRAMs are two-terminal emerging non-volatile memory (NVM) devices consisting of a top electrode (TE), bottom electrode (BE) and a switching oxide (SO) between TE and BE. The resistance of the device can be toggled between a low resistance state (LRS) and a high resistance state (HRS) by changing SO properties by applying set and reset voltage biases. To set the ReRAM devices a positive voltage around 1-2V is applied to the TE while the BE is grounded and to reset the device a negative 1-2V is applied using the same configuration. The most common method for applying the voltage is a short pulse, but several researchers also sweep the voltage from 0-2V to understand the behavior of the devices across the voltage range. ReRAM devices have gathered tremendous research attention, primarily, for off-chip data storage, on-chip memory for System on Chip (SoC) applications due to low-switching voltages [7], reconfigurable elements for FPGAs [8], and in-memory computing architectures [9,10]. The crossbar integration of ReRAM devices with access devices provides opportunity to develop dense memory structures and compatibility with the CMOS process-flow. However, ReRAM devices currently suffer from device-to-device and cycle-to-cycle variabilities in switching states which is undesired for the above-mentioned NVM applications and much research is needed to address this issue.

While variabilities in ReRAM imposes issues for its applications as NVM, development of ReRAM devices and their applications has begun extending beyond an NVM structure. For example, the random variability in ReRAM resistive states has been explored for designing PUFs [11,12]. However, PUFs are usually desired to have a long lifetime and show minimal drift in Challenge Response Pair (CRP) tests over time under Process Voltage Temperature (PVT) variations, which remains to be proven for ReRAM. Interestingly, our observations on experimentally fabricated ReRAM devices have indicated accelerated aging and temperature-based changes in resistive states of the device. This provides us significant opportunity to further explore these modalities in ReRAM devices and specifically tailor these devices to use them as AIM elements for CMOS ICs. The objective of this paper is to present the concept of an additive ReRAM-based integrity and aging (ARIA) monitoring circuits for CMOS ICs. When compared to RO based monitors, ARIA monitors offer inherent benefits as it can be integrated on prefabricated CMOS dies using additive manufacturing techniques. The feature sizes of ARIA components can be relatively large ensuring such fabrication to be conveniently performed in trusted in-house foundries, thus, making these circuits secure against tampering. Additionally, ARIA monitors will not consume any FEOL functional area of CMOS ICs, therefore, designers can take advantages of high-density scaled transistor technologies in untrusted foundries while ensuring complete spatial and temporal coverage for aging, reliability, and integrity monitoring using ARIA monitors, fabricated in-house.

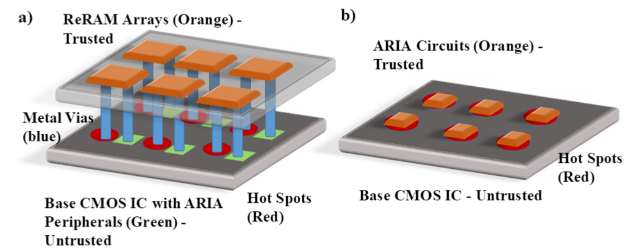
a)
ReRAM Arrays (Orange) - Trusted
Metal Vias (blue)
Base CMOS IC with ARIA Peripherals (Green) - Untrusted
Hot Spots (Red)
b)
ARIA Circuits (Orange) - Trusted
Hot Spots (Red)
Base CMOS IC - Untrusted

Fig. 2: Visual representation of a) on chip peripheral circuitry with metal via structures and b) single chip 3-D integration for additively manufactured ReRAM AIM circuit on top of silicon die.

Fig. 2(a) presents an approach to implement ARIA monitors where the signal to be sensed (e.g. temperature or current from underlying IC) can be fed to ARIA monitors by routing vias from underlying CMOS ICs to the top of the die. ARIA monitors can be additively integrated on this die aligned with these pre-routed sensing pads from CMOS ICs as well as partially integrated on the fabricated die. This allows the ARIA monitors to sense the signal of the desired underlying circuit. In a particular instance where temperature is used as sensing signal then a continuous stressing of ARIA monitors under this temperature will lead to the changes in ReRAM states that can be used to monitor the aging of ICs under normal operations. Also, if a Trojan is triggered in the underlying IC then it will cause a local heating and instantaneous increase in temperature. This is due to changes in power which are caused by the active Trojan and will also be reflected in the IC's thermal profile [13]. This will lead to an abrupt change in the resistive states of ReRAM in ARIA monitors. This feature can be utilized for integrity monitoring. One should note that unlike a usual diode- or transistor- based temperature sensor, ReRAM not only senses temperature but also stores it as change in resistive states, such allowing for in-senor storage of information which is very unique and novel feature explored in this work.

The second integration approach shown in Fig. 2(b) relies on individual ARIA monitors which have the full sensing capabilities of the previous design without having to utilize any of the original CMOS design space. This method requires through silicon vias and bonding of the two dies where access to each desired circuit to be monitored has a pad connection on the surface of the fabricated IC. The major benefit of this design is not having to customize the IC design to allow space for the monitoring circuit but rather be able to place the ARIA monitors wherever they are necessary.

The other benefits of this method include ability to integrate these components on CMOS dies using low thermal budget process that allows for additive integration of these components in trusted in-house foundries. With each CMOS design there are security metrics or design-specific areas of a circuit that needs to be monitored, thus there is no one solution that fits all needs. Customizable solutions for AIM through additive manufacturing, as proposed in this paper, will be cost effective for the designers and will monitor the aging and integrity of the portions of the circuit where it is placed.

The remainder of this paper is organized as: section-II discusses the additive fabrication of ReRAM as ARIA monitors and characterization of resistance states, section-III covers the temperature and aging modalities of ReRAM, its impact on resistances in LRS, HRS, and virgin resistance state (VRS), section-IV discusses the design of ARIA monitor circuit design and simulation results, and Section-V presents conclusions and prediction strategies for circuit monitoring using ReRAM based in these studies.

II. Additive Approach for ReRAM AIM

The concept of additive manufacturing relies on taking a silicon die from a foundry, such as, TSMC or Global Foundries, to additively manufacture components in house or at a trusted foundry. The benefit of additive manufacturing is the lower density of structures and the simple design of a ReRAM device structure. In order to fabricate the ReRAM devices on top of a silicon die, just three masks will be required, one for patterning the bottom electrodes, one for patterning the isolation layer, and one for patterning the top electrodes. Once the die is received from the manufacturer, a chemical mechanical polish (CMP) will be required to remove the passivation layer to gain access to the metal vias at the top layer. Once this is done the first deposition of the bottom electrode consisting of a metal such as Ru, TiN, Pt, Cu and Au is done and patterned to create a larger sized pad for the deposition of the isolation layer. Separating the two metal layers, top electrode and bottom electrode is crucial to prevent shorting of the device and also allows a hole to be etched for the deposition of the oxide. This buffer layer typically consisting of Low-Temperature SiN or SiO_2, is patterned and etched. The transition metal oxide (TMO) layer is then deposited which can be tailored to the desired lifetime by changing oxides such as HfO_2, TaO_x, MgO or SiO_2 and defect-engineering via intrinsic or extrinsic doping in these oxides. Once the oxide is deposited, a top electrode is then deposited and patterned to create a contact to access the metal layers of the CMOS. This process can be done using a low density design on top of high density areas with only the need of a metal via coming up to the bottom electrode contact point as shown in Fig. 3. The other benefit of this process is that it can be accomplished below 200°C which will not impact the underlying CMOS layers. The ideal additive design will also use long channel mosfets (~180 nm) that can be fabricated in house to design all of the peripheral circuitry on top of the CMOS structure as well, but this step was not part of this research.

The ReRAM devices studied in this work were fabricated as a Ruthenium (Ru)/Magnesium Oxide (MgO)/Titanium (Ti)/Tungsten (W) stack. The layer

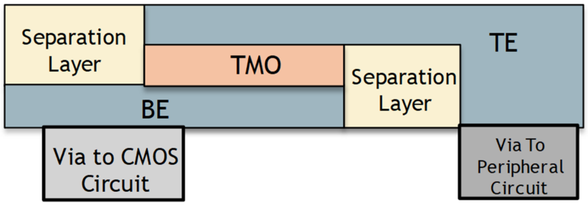

Fig. 3: Diagram of fabrication structure for additively manufactured ReRAM on top of silicon die with contact vias to the top and bottom electrode where the BE via connects to the circuit to be monitored and the TE via goes to the peripheral circuitry required to monitor the device state.

thicknesses were deposited as follows: 40 nm thick Ru BE. Then a 6 nm of MgO TMO layer was deposited by reactive sputtering of Mg from Mg target in a mixed 12 SCCM Ar and 5 SCCM O_2 at 2mTorr pressure environment at room temperature. Once this was done, the interface layer of 10 nm Ti and TE of 70 nm W was deposited. The wafer was patterned using photolithography and the TE was etched to define the device areas. Devices of size 30µm x 30 µm squares were tested in this work for all experimental results. All measurements were performed using Keithley 4200 Semiconductor Characterization System fitted with 4225 Pulse Modulation Unit by probing devices in a Cascade probe station. Temperature based measurements were performed by probing devices on a Cascade Mircrotech SummR 12000AP. This data will be presented and explained in secion III.

III. Temperature Sensitivity and Aging Model for ReRAM

The proposed ARIA circuits capitalizes on exploting the temperature sensing and aging modalities in ReRAMs, which is not well-understood. In this section we present our experimental observations and models to capture these dynamics in MgO-ReRAM. Switching characterstics and temperature modalities have been previously studied in Mg doped HfO_2 [14], HfO_2 multi-state characteristics [15], and variabilities in MgO based ReRAM [16]. To the best of our knowledge, this is the first attempt to systematically report and model these parameters in ReRAM devices. These measurements can have implications not just for ARIA circuits but also for the other applications of ReRAMs. Schematically shown in VRS (i.e. right after fabrication without any voltage stress) in Fig. 4(a). MgO as switching oxide was investigated in this work because of the unique defect dynamics in MgO owing to an interplay between Mg^{2+}, V_o^{2+} as a function of process parameters and operating conditions which could provide an increased temperature and aging sensitivity, as desired by ARIA circuits. Ti was used as an interfacial layer (IL) due to its relatively low Gibbs free energy of oxide formation that can help in gettering excessive oxygen from MgO. During the deposition of MgO/Ti in ReRAM, the initial defect concentration in the oxide is typically governed by the material stacks and process parameters that creates an unknown distributions of devices in the VRS. To better understand the variability in the VRS, 100 individual devices of 30µm x 30µm were tested using a 0.2V read bias and their resistance distribution is shown in Fig. 5(a). The distribution of the resistances is based on a mean of 18MΩ and a std. deviation of 4MΩ. This variance is somewhat extensive when compared to the typical 6% tolerance in 180 nm transistors [17,18]. Thus, if ReRAM in VRS is used in ARIA circuits then slight changes in the resistances will not be detected by the monitoring circuits as it would be within the tolerance bounds for VRS. Therefore, other states in ReRAM needs to be investigated for this purpose.

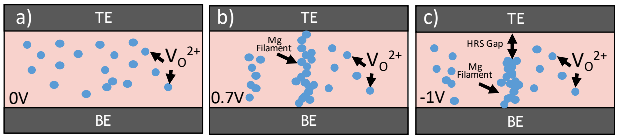

Fig.4: Schematic representation of ReRAM oxygen vacancies and filament formation mechanics where the voltages represent the applied bias from TE with BE grounded in (a) VRS, (b) LRS, and (c) HRS.

Next, to obtain other states, a one-time electroforming process was necessary to observe the desired switching characteristics in this device. Due to the manual nature of the tests only a limited number of the devices (45) were chosen to undergo the swtiching process from LRS to HRS. The electroforming process causes a soft breakdown of oxide by forming a defect-assisted filament consisting of V_o^{2+} or metal rich-regions, such as Mg^{2+}, schematically shown in Fig. 4(b). Electroforming process was performed by a 2V voltage-pulse across the device for a 100ns period. Fig. 5(b) shows the distribution of LRS after electroforming where the mean is 220Ω with a std. deviation of 30Ω. The device can be brought back to HRS by applying a reset voltage that causes retraction of filament, possibly due to local oxidation, shown in Fig. 4(c). HRS is obtained by applying a -1.5V voltage-pulse for a 100ns period to the devices already in LRS. The resultant resistances of each device were measured by applying a 0.2V read pulse for 1 µs. The HRS distribution across multiple devices is shown in Fig. 5(c). This distribution has a mean of 18KΩ and a std. deviation of 3.5KΩ. The tolerance of HRS is better than VRS by 2% with a single switching cycle, but can be further improved by taking devices whose resistance lies outside one std. deviation and cycle the devices again to decrease the std. deviation allowing the ARIA circuit to have a better resolution. Fig. 5(d) shows a typical DC switching cycle for these devices. A low set (0.7V) and reset (-1.0V) was observed which attests that these devices can be programmed conviniently using nominal power supplies on-chip.

Next, temperature sensitivity (S_T) of these states were measured. For this test, independent 30µm x 30µm devices were used for measuring VRS, HRS, and LRS. To achieve the LRS state a device was set by applying a 1.2V pulse with a 100µA compliance current to prevent breakdown of the oxide.

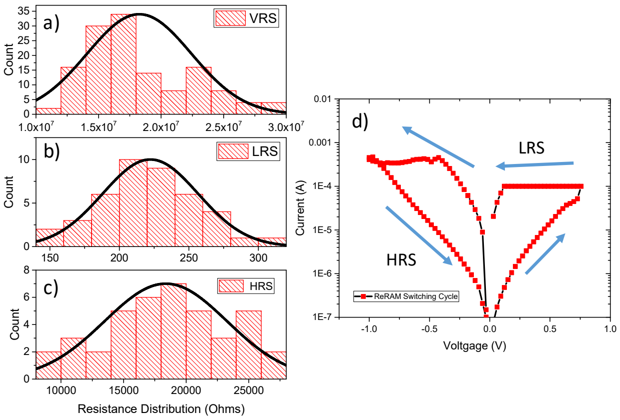

Fig. 5: Variability in ReRAM devices showing the Distribution of 30µm x 30µm ReRAM devices in a) 100 devices in VRS (b) 45 in LRS (c) 45 in HRS with a normal distribution fitted to the data to be used in simulation and (d) a typical DC switching cycle with 0.7V set and -1.0V reset.

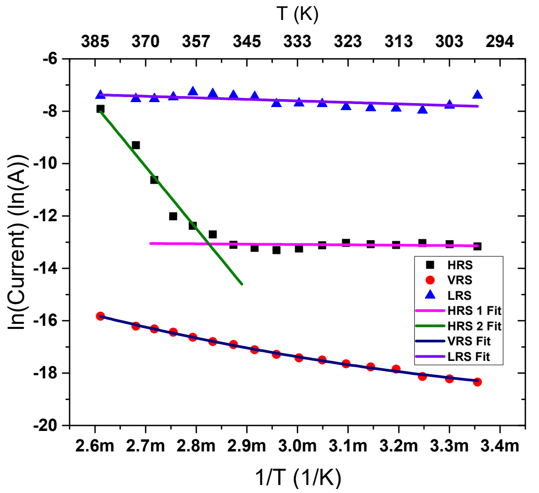

Fig. 6: Test results from Temperature testing over 300K to 385K in LRS, HRS, and VRS read with a 0.2V bias.

To achieve the HRS, separate devices were set using the previously described method then reset by applying a -1V pulse. This created three distinct states where the LRS, HRS and VRS that had no dependence on each other to prevent biasing of the data. Measurements were then taken from 7 devices in each resistance state (i.e. VRS, HRS, LRS) as temperature was increased from 300K to 385K. The current (I) through devices in different states were read by applying a 0.2V read pulse on each device at a given temperature. The temperature was increased linearly with several reads of each device taken at 5K intervals. The ln(I) vs. 1/T plot for VRS, HRS, and LRS is shown in Fig.6. Clearly, each state showed different sensitivity to the temperature as evident from the slope of ln (I) vs. 1/T. A linear fit to this curve was obtained for each state indicating mechanism of conduction to be governed by Frenkel-Poole (FP), given by equation (1) below.

$$\ln(I) = \ln(q\mu N_c AE) - \frac{q\left(\Phi_T - \sqrt{{qE}/{\pi\varepsilon_i\varepsilon_0}}\right)}{kT} \tag{1}$$

where, Φ_T is the trap energy level, ε_0 is the permittivity in vacuum, and ε_i is the dielectric constant, q is unit electronic charge, μ is electron mobility, N_c is the initial density of states, E is the electric field, k is the Boltzmann's constant, T is temperature in K [19]. The slope $(\Phi_T - \sqrt{{qE}/{\pi\varepsilon_i\varepsilon_0}})$ of the curve is indicative of temperature sensitivity $\left(S_T = {d(lnI)}/{d({1}/{T})}\right)$ and was extracted from Fig. 6 and reported in Table-I for different states. The FP curve fit for VRS was consistent across the temperature range. LRS showed very low S_T which is not desired by ARIA circuit. It should be noted that ohmic conduction through oxide also has similar dependence on T as FP, therefore, the transport mechanism ambiguity in LRS can be resolve by I vs. V fitting, indicating ohmic conduction in LRS, as reported by previous work [19]. Most interesting features were observed in HRS which is inarguably the most intriguing state of the device due to complex defect dynamics. HRS showed two distinct slopes above and below around 360K as evident from two different linear curve fits in these regions. Below 360K HRS current showed minimal dependence on temperature indicating mechanism to be governed by either temperature independent process (such as tunneling), or existence of shallow traps, or trap-assisted tunneling (TAT) that has similar dependence on temperature as FP [19]. Above 360K, slope is much higher indicating higher S_T, particularly desired for integrity sensing, discussed later in section IV.

Table-I: S_T in different states of MgO-ReRAM device

State	kS_T (eV)
VRS	0.28
HRS-1	0.01
HRS-2	2.01
LRS	0.05

To understand the aging characteristics of these states Accelerated Lifetime tests were performed. The time to failure (TTF) of an oxide at a given temperature is given by equation (2) below as,

$$TTF = Ae^{{E_a}/{kT}}e^{\beta V_s} \tag{2}$$

$$or, \ln(TTF) = \ln(A) + \beta V_s + {E_a}/{kT} \tag{3}$$

where, Ea is activation energy, A is a constant , k is Boltzmann's constant, and T is temperature in K, β is a constant, and V_s is stress voltage [20,21]. The TTF of several devices were tested at varying temperatures by keeping devices at a constant V_s of 0.2V to determine how long it would take to fail. The device states used to test were VRS and HRS and the criteria for determining failure was when the device state becomes more conductive than LRS. Devices in LRS were already very conductive and did not show any further breakdown, therefore, they were eliminated from this test. Shown in Fig. 7 is the natural log of the TTF as a function of T. The E_a was calculated from the slope as 0.76eV for HRS and 0.72eV for VRS. Once the E_a is known experimentally, the

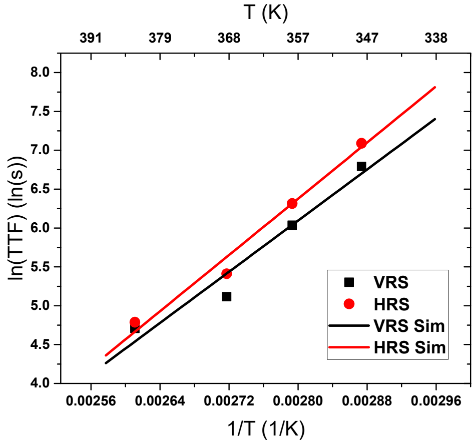

Fig. 7: Natural log of the time to failure vs Temperature of RRAM devices in HRS and VRS.

overall acceleration factor ($AF_{T,V}$) for aging at a given V_s can be written as:

$$AF_{T,V} = e^{E_a/k\left(\frac{1}{T_0}-\frac{1}{T_S}\right)} \tag{4}$$

where, T_0 is the chip operating temperature, T_s is the stress temperature.

Ideally, for ARIA circuits as age monitors, it is important to be able to measure and model degradation in HRS and VRS at different T_0 over time before device fails. This is because T_0 will be different for different IC to be monitored or different regions of the IC to be monitored. Degradtion in HRS and VRS over time can be attributed to the generation of additional traps in MgO (such as Mg^{2+} or V_o^{2+}) under continuous thermal stress of T_0, generated by operation of underlying IC (Fig. 1 and Fig. 2). The trap density, D_t, generated over time can be given as Eq. 5 below,

$$D_t = \alpha * t^m \tag{5}$$

where α is the approximated initial concentration, m is the logarithmic defect generation rate and t is the stress time [22]. Based on this equation, $I(t)_{T,V}$, i.e. current measured through VRS or HRS ReRAM stressed with constant T and V over t before device fails can be empirically written as:

$$I(t)_{T,V} = I(0) + \gamma * I(0) * t^m \tag{6}$$

where, I(0) is the initial current through unstressed device at temperature T, and γ and m are fitting parameters that need to be extracted experimentally. Note, once m is known for a given T and V, then it can be calculated for other T and V by modifying TTF using $AF_{T,V}$, shown in equation (4). To extract γ and m experimentally, I was read through device in HRS over time under constant temperature of 368K (95°C) at stress bias of 0.2V, shown in Fig. 8(a). This temperature was chosen as an appropriate temperature from which several data points could be taken during the lifetime otherwise the device would fail too quickly at higher temperatures and too slowly at lower temperatures. From this plot, γ and m values were extracted as 3.75E-8 and 4.03, respectively. From these fitting parameters, equation (6) was then used to simulate I vs. Time in HRS at higher and lower temperatures showing how the I profile of the device changes at different T over its lifespan, shown in Fig. 8(b). Note, the m values, calculated using equation (6), and used for these simulations are shown in Table-II, while I(0) values at different T is taken from Fig. 6 for HRS. Clearly, the HRS degrades faster over t at higher stress temperatures and degradation in HRS current over time is an apparent indicator of aging over its operational lifetime. Upper current limit was set at 5mA, when device was declared as completely failed.

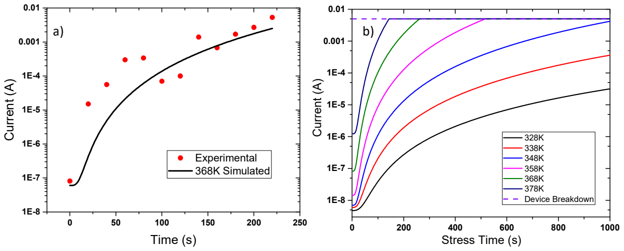

Fig. 8: Oxide degradation due to high temperature and a 0.2V stress where (a) increase in current over time experimentally observed at 368K and simulated curve using Eq. (6) and (b) extrapolating simulation of increasing current at other temperatures based on Eq. (4) and Eq. (6) with a dashed line showing the maximum current flowing through the ReRAM device indicating a complete oxide breakdown.

Table-II: Logarithmic defect-generation rate in MgO-ReRAM at different temperatures with 0.2V stress

Temperature (K)	Log Defect Generation Rate
338	3.26
348	3.58
358	3.87
368	4.03
378	4.55

By using this approach, Fig. 9 shows a simulation of degradation of resistance in VRS and HRS at constant temperature of 328K over its lifetime. It should be noted that 328K was chosen to understand feasibility of using these devices to monitor the normal operating temperature and age of a MIPS processor which typically heats up to this temperature [23]. From Resistance vs. Lifetime, it can be seen the resistance changes rapidly in the beginning indicating that there is significant degradation of the oxide during the beginning of its life. Then, resistance changes slow down to saturate towards the end of its life and eventually degrades below LRS. This clearly, indicates potential for using these devices for age monitoring.

The devices can also be used for integrity monitoring by capitalizing on S_T of these devices, particularly, the HRS, shown in Fig. 6. High S_T in HRS indicates the potential for sensing the temperature of a circuit at high resolution. For example, when a Trojan is activated in this processor the local area will heat upwards of a few degrees K which will cause an

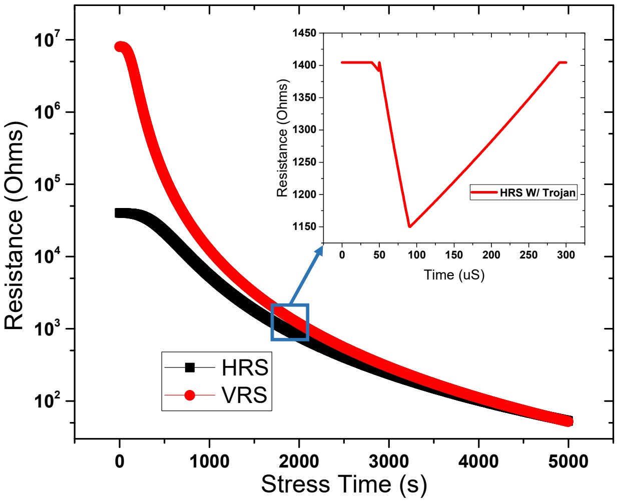

Fig. 9: Extrapolation of HRS and VRS degradation over lifetime given starting condition with 328K operating temperature. Inset indicates a drop in relatively instantaneous resistance as Trojan turns-on leading to local temperature increase from 328K to 332K.

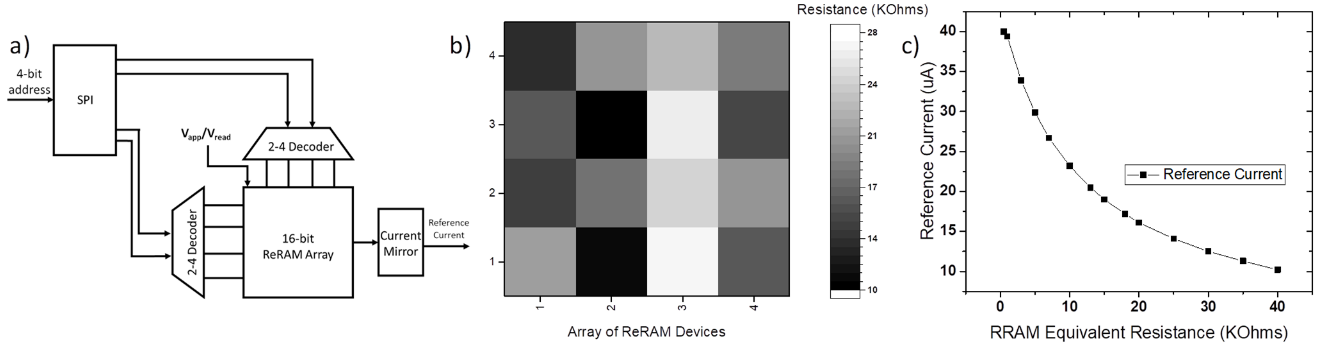

Fig. 10: Circuit diagram for the ReRAM control logic circuit and current mirror output using 180nm technology where a) shows the control logic for a 16-bit array of RRAM devices used to create a reference current for a current mirror using L=180nm and W=360nm, b) Resistance weight distribution of the ReRAM devices in HRS and c) Current output vs. ReRAM resistance of the current mirror circuit.

instantaneous change in resistance of device based on its S_T. Based on Fig. 6, it should be noted in order to detect a Trojan the thermal impact of the Trojan circuit must cause a minimum of a 2K temperature increase to be detected by the circuit otherwise it will not be detected as the temperature difference will be indistinguishable and within normal read variability. The result of this instantaneous change is captured in the inset of Fig. 9. During this small time frame the ReRAM device will change resistance state but will return to its normal state after the Trojan becomes inactive and the normal aging of the devices continues. One should note that ARIA ReRAMs do not require high write endurance cycles. All that is needed is one-time programming into HRS. However, when used in ARIA circuits, its it important to minimize die-to-die and inter-die variations in resistance values in various states. Due to stochastic nature of switching, HRS helps in averaging out the impact of process-induced Die-To-Die and Inter-Die resistance variations that can be evident in VRS. Therefore, if one samples multiple ReRAMs in HRS randomly across wafers then it can be expected to have a similar Gaussian distribution. HRS also provides opportunity to program the intial state which can have two implicaitons: (i) tailoring the starting point allows the intial state of the devices to have approximately the same resistance and age in the same manner eliminating the random distribution and (ii) makes it extremely difficult to replicate the aging of the specific monitor or reverse engineer the design and materials used to develop it. Another interesting point to be noted is that typically variability in ReRAMs increases as devices are scaled down below 100 nm which is critical when they are used for high-density NVM applications. However, ARIA circuits does not need ultra-small ReRAMs. Ideally, relatively larger feature sizes are desired so these can be fabricated at low-density in trusted foundries on top of CMOS dies. The larger feature sizes helps in mitigating the variability issues in ReRAMs. Next section presents peripheral circuit design of ARIA monitors that can leverage these temperatures and aging modalities of ReRAM for predicting age and integrity of the underlying IC.

IV. ARIA PERIPHERAL CIRCUIT DESIGN AND SIMULATION

Utilizing the discussed HRS, aging, and temperature-based changes in ReRAM, an ARIA circuit was designed to monitor the age and temperature of neighboring areas. The circuit needed to be capable of analyzing the state of the ReRAM device and produce a digital output which could be used by the CPU to determine the temperature or age of the circuit. Several circuit designs would suffice for this application, but the circuit was designed to be capable of being entirely fabricated on top of the CMOS die. The ARIA circuit designed in Fig. 10(a) includes: a 4-bit addressable array of 16 ReRAM devices connected in a crossbar, used as the input to the current mirror (CM) to create a reference current. The array configuration is

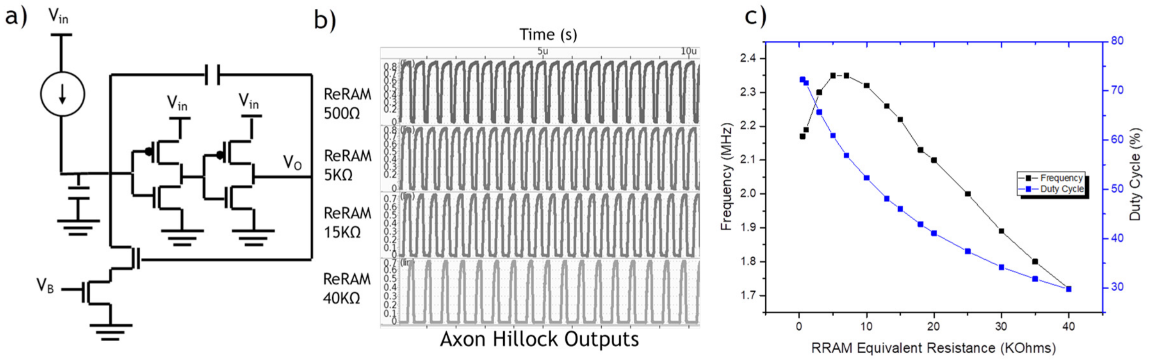

Fig. 11: The second section of the ARIA circuit includes the a) Axon Hillock Circuit (AHC) which current source is the current supplied from the current mirror b) The raw AHC output simulation data of four equivalent resistance states: 40KΩ, 15KΩ, 5KΩ, 500Ω and c) the frequency values and duty cycle of the oscillating signal as the ReRAM resistance changes.

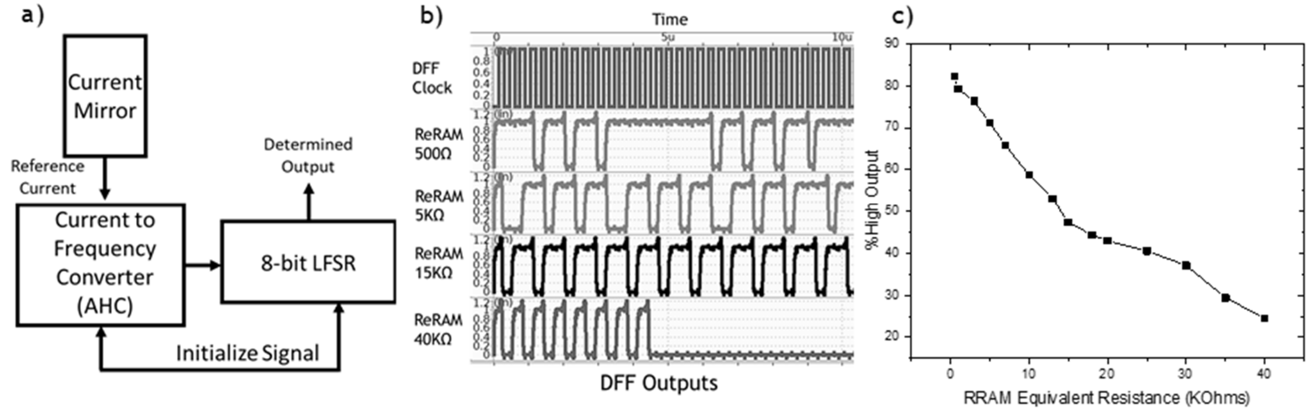

Fig. 12: The third section of the ARIA circuit which includes a) a series of DFFs creating an LFSR to sample the output of the AHC and produce a digital output signal b) the raw simulation data of the DFFs output for four equivalent resistance states: 40KΩ, 15KΩ, 5KΩ, 500Ω and c) the Average output of the LFSR is given as a percentage of 1's or high output as the ReRAM device changes resistance.

used to create multiple references for a single ReRAM device in HRS allowing the system to average out the impact of PVT and aging variabilities in each individual ReRAM device. The 0.2V read and 1.8V write signals are connected to access transistors controlled by the serial-parallel interface which requires a one-time write to place the devices into the initial state and then a constant read bias for the remainder of their lifetime. The current mirror is comprised of two transistors designed with L=180nm and W=360nm to handle the current range produced by the RRAM devices. Fig. 10(b) shows the resulting resistance values of the ReRAM devices from the one-time programming in the array to the input of the CM circuit. This way if a single device fails it will not stop the circuit from functioning, though adjustments need to be made to compensate for single device failure and the eventual failure of all the devices. Fig. 10(c) indicates the reference current (Iref) vs. equivalent resistance based on the aging or instantaneous temperature changes of the devices. If the equivalent resistance is equal to two ReRAM devices, 40KΩ, the output current is 10µA and once these devices degrade down to 500Ω the output current increases to 40µA.

Next, to re-use some of the previous design concepts based on RO-based AIM, the objective was to convert the change in Iref as an array of voltage pulses (0/1) with frequencies and duty cycles proportional to the Iref. The current to spike converter, such as Axon Hillock Circuit (AHC), was identified as suitable for this purpose [24]. The AHC circuit using 180 nm CMOS technology was design in Cadence, shown in Fig. 11(a), with Iref from CM circuit as input. The transistor sizes for the dual inverter are all W=L=180nm, the transistors which control the reset mechanism are W=180nm and L=360nm to allow for faster resetting of the AHC, and the capacitors are 4pF and 2pF. The bias voltage for the NMOS transistor VB was set to 0.7V as to control the resetting of the AHC. The input voltage is set to 1V to keep a low voltage supply and reduce the power consumption of the circuit. The design of the transistors and capacitors were based on the stable range of current produced by the current mirror to operate the AHC.

The current values supplied by the current mirror circuit were used as the input of the AHC. The raw simulation data of four equivalent resistance states: 40KΩ, 15KΩ, 5KΩ, 500Ω are shown in Fig. 11(b). From the raw output of the AHC, the frequency of the spikes generated, and the duty cycle were calculated over a 10µs period to allow for a significant number of spikes to occur. Several other resistance ranges were tested in the AHC to show the entire spectrum of resistance values. Fig. 11(c) displays the frequency values and duty cycle of the signal as the ReRAM begins to degrade in resistance and the current supplied by the current mirror increases. The frequency hovers around the 2.25MHz range until the resistance of the ReRAM increases beyond 15KΩ. As the resistance of the ReRAM increases further the frequency drops down to 1.75MHz at 40KΩ. While the frequency of the AHC does not change much over the full resistance range of the ReRAM devices, the duty cycle of the output does. The duty cycle shows a change from 75% at 500Ω to only 30% at 40KΩ.

Knowing the operating frequency of the output of the AHC, the pulse train can be sampled through linear feedback shift register (LFSR) to produce a definitive output for changes in the frequency. The final output of the circuit can be generated through a series of Flip Flops creating the aforementioned LFSR shown at the output of the AHC in Fig. 12(a) for a digitized serial output value. The clock frequency of the LFSR is set to the output of the AHC to monitor the frequency changes over a given time period. Spectre simulations of the DFFs for four equivalent resistance states: 40KΩ, 15KΩ, 5KΩ, 500Ω are shown in Fig. 12(b). The average of the output state of the DFFs were calculated over a 20µs length of time and converted to a percentage of 1's which is the percentage of time the DFFs output 1V over that time interval. The results of the output of the DFFs is shown in Fig. 12(c) where the percentage of high or 1 value outputs decreases linearly as the resistance increases. When the resistance of the ReRAM array is at the minimal value, the outputs of the DFFs are around 85% ones and when the resistance is at the maximum value the output changes to only 23% ones. This output shows a linear change in the % of high output which provides a high-quality step metric for differentiating between subtle resistance changes. This creates a high-resolution monitor for the aging and temperature of a circuit. The outputs of the DFFs would

Table-III: Comparing ARIA circuit with other monitoring systems metrics

Metric	**Dynamic NBTI Sensing [5]**	**Silicon Odometer [4]**	**NBTI/HCI models [3]**	**Built-in BTI Monitor [25]**	**Proposed ARIA Circuit**	**Scaled ARIA Circuit**
Voltage	2V	1.2V	1.1V	1V	0.2V	0.2V
Evaluation time	100µs	2µs	400µs	500µs	50µs	1µs
Transistor Node	45nm	130nm	45nm	45nm	180nm	22nm
Area	77.3µm^2	34,980µm^2	148µm^2	2,400µm^2	10,000µm^2	~40µm^2
Power	10-100nW/Sensor	Unreported	20-30µW	Unreported	10-150µW	1-10µW
Lifetime	< 2000s*	Unreported	Unreported	5 years**	1.5hrs*	Unknown

*Lifetime determined by constant biasing and evaluation of a single device

**Lifetime determined by only biasing and evaluating once a week

allow for a LUT to be used for monitoring the age or temperature of the circuit but would need to start sampling on the rising edge of the AHC due to the slight changes in frequency from the change in ReRAM resistances. A comparator could also be used to differentiate between the number of high outputs within a string of a given length. Monitoring circuitry should consume minimal power, use low voltage, and be low cost. The ARIA circuit metrics compared to other solutions is presented in Table-III below.

V. Conclusions and Future Work

In conclusions, our studies indicated that ReRAM devices can be additively manufactured on an IC and used to monitor the age and temperature of the underlying circuit. This AIM circuit uses low power, frequency, and voltage. Fabrication strategies for the process of additively manufacturing the ReRAM devices and their integration with the underlying CMOS architecture was discussed. The number of ReRAM devices which can be used to monitor the system will depend on how long the system is expected to be used. Since each ReRAM device has a lifespan dictated by Eq. 6 and there are 16 devices in the array, the desired number of arrays which can be used for monitoring one section can be calculated. By utilizing a counter and changing to another array once one has been completely used the circuit size will increase but it will provide a longer aging monitor. The thermal profile of the SoC may want to be studied through simulation or aging to determine the best areas to place the ReRAM system to provide an accurate age and integrity monitor. Experimental studies in combination with simulations showed resistance distributions in LRS, HRS, and VRS of devices. These studies also included data for changes in the temperature impacting the resistance in the three different states. Accelerated aging of ReRAM devices at temperatures of 338K-378K were performed and the time to failure of the devices were measured. This was used to develop model to predict the lifespan of the devices at different operating temperatures and long-term degradation due to aging. Devices had a manageable variability across HRS and temperature changes. Additionally, the impact of these variabilities in HRS of ReRAM on AIM circuit operation was mitigated through using an array of multiple devices whose states could be evenly distributed, thus, reducing the impact of variability. The HRS was found to be much more susceptible to changing states due to increases in local temperature and aging which was measured as change in current using the current mirror circuit appropriately. While HRS is used to determine the temperature and integrity of the IC, if the underlying IC is designed to operate at high temperatures at all times the device will not be able to detect any difference as it will already be close to LRS. This limits the temperature range in which this device and ultimately system will be able to operate and thus the device would need to be modified to handle higher temperatures. A Spectre simulation was performed on a ReRAM array resistance divider network fed into a current mirror and AHC to monitor the change in frequency and duty cycle of the output. DFFs were used to sample these outputs and create a usable data stream for a comparator to use to determine the temperature or age of the underlying circuit. The simulation showed a significant change in output and a linear correlation as the ReRAM devices changed resistance based on temperature or aging.

We observed that MgO-ReRAM, experimentally studied in this work, showed appropriate sensitivity to temperature variations and aging. Typically, ReRAMs are optimized to mitigate these effects when used as NVM for high-density data storage, however, our studies indicate that it is important to look at ReRAM devices beyond just NVM structures. The capability of these devices to change resistance over temperature and time allows them to be utilized in new designs. Additionally, these changes are stored as change in resistive states that provides in-memory sensing feature. MgO-ReRAM in HRS, however, showed a short lifespan of just 1.5 hrs. for complete degradation at 328K, which needs to be improved so circuit can operate as monitors for entire lifetime (~10 years) for CMOS ICs. One method of improving lifetime would be modifying interface layers or improving the quality of the oxide, but other methods are designed to create pre-existing filament areas to be less initially destructive to the oxide during the forming process [26]. Approaches to improve this can include decreasing of E_a and defect generation rates, increasing the HRS: LRS ratio, and integrating sequential arrays to be activated once the lifetime of one end which will be explored in our future work. Currently ARIA is designed to operate in the MHz range which should be acceptable for age monitoring. However, for integrity monitoring, the clock frequency of ARIA needs to be comparable to a processor clock frequency to detect anomalous behavior which may only occur during a single instruction cycle. The next steps will also include fabricating the proposed

ARIA circuit on CMOS ICs to validate the additive manufacturing process and overall performance of the IC.

ACKNOWLEDGEMENT

The authors gratefully acknowledge NSF for the Award# CNS-1556301, DAGSI Fellowship which funds our research.

REFERENCES

[1] International Trade Administration, “2016 ITA Semiconductors and Semiconductor Manufacturing Equipment Top Markets Report”, US Department of Commerce, 2016

[2] Tseng-Chin Luo, Mango C.-T. Chao, Philip A. Fisher, and Chun-Ren Kuo, “A Novel Design Flow for Dummy Fill Using Boolean Mask Operations”, IEEE Transactions on Semiconductor Manufacturing, Vol. 25, No. 3, August 2012.

[3] Kim, Kyung Ki, Wei Wang, and Ken Choi. “On-chip aging sensor circuits for reliable nanometer MOSFET digital circuits.” IEEE Transactions on Circuits and Systems II: Express Briefs 57, no. 10 (2010): 798-802.

[4] Singh, Prashant, Eric Karl, Dennis Sylvester, and David Blaauw. "Dynamic nbti management using a 45 nm multi-degradation sensor." *IEEE Transactions on Circuits and Systems I: Regular Papers* 58, no. 9 (2011): 2026-2037.

[5] Kim, Tae-Hyoung, Randy Persaud, and Chris H. Kim. "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits." *IEEE Journal of Solid-State Circuits* 43, no. 4 (2008): 874-880.

[6] Tajik, Shahin, Julian Fietkau, Heiko Lohrke, Jean-Pierre Seifert, and Christian Boit. "PUFMon: Security monitoring of FPGAs using physically unclonable functions." In On-Line Testing and Robust System Design (IOLTS), 2017 IEEE 23rd International Symposium on, pp. 186-191. IEEE, 2017.

[7] Cong and Xiao, “FPGA-RPI: A Novel FPGA Architecture with RRAM-Based Programmable Interconnects”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems (Volume: 22, Issue: 4, April 2014), pages 864-877.

[8] Nassif, Sani R. "Modeling and analysis of manufacturing variations." In *Proceedings of the IEEE 2001 Custom Integrated Circuits Conference (Cat. No. 01CH37169)*, pp. 223-228. IEEE, 2001.

[9] Liu, Tz-yi, Tian Hong Yan, Roy Scheuerlein, Yingchang Chen, Jeffrey KoonYee Lee, Gopinath Balakrishnan, Gordon Yee et al. "A 130.7-mm^2 2-Layer 32-Gb ReRAM Memory Device in 24-nm Technology." *IEEE Journal of Solid-State Circuits* 49, no. 1 (2013): 140-153.

[10] Zha, Yue, and Jing Li. "IMEC: A Fully Morphable In-Memory Computing Fabric Enabled by Resistive Crossbar." *IEEE Computer Architecture Letters* 16, no. 2 (2017): 123-126.

[11] Kim, Jeeson, Taimur Ahmed, Hussein Nili, Jiawei Yang, Doo Seok Jeong, Paul Beckett, Sharath Sriram, Damith C. Ranasinghe, and Omid Kavehei. "A physical unclonable function with redox-based nanoionic resistive memory." *IEEE Transactions on Information Forensics and Security* 13, no. 2 (2018): 437-448.

[12] Yoshimoto, Y., Y. Katoh, S. Ogasahara, Z. Wei, and K. Kouno. "A ReRAM-based physically unclonable function with bit error rate< 0.5% after 10 years at 125° C for 40nm embedded application." In *VLSI Technology, 2016 IEEE Symposium on*, pp. 1-2. IEEE, 2016.

[13] Forte, Domenic, Chongxi Bao, and Ankur Srivastava. "Temperature tracking: An innovative run-time approach for hardware Trojan detection." In *Proceedings of the International Conference on Computer-Aided Design*, pp. 532-539. IEEE Press, 2013.

[14] Long, Branden Michael, Saptarshi Mandal, Joseph Livecchi, and Rashmi Jha. "Effects of Mg-Doping on HfO_2 Based ReRAM Device Switching Characteristics." *IEEE Electron Device Letters* 34, no. 10 (2013): 1247-1249.

[15] Chen, Wenbo, Wenchao Lu, Branden Long, Yibo Li, David Gilmer, Gennadi Bersuker, Swarup Bhunia, and Rashmi Jha. "Switching characteristics of W/Zr/HfO2/TiN ReRAM devices for multi-level cell non-volatile memory applications." *Semiconductor Science and Technology* 30, no. 7 (2015): 075002.

[16] Schultz, Thomas, and Rashmi Jha. "Understanding vulnerabilities in ReRAM devices for trust in semiconductor designs." In *Aerospace and Electronics Conference (NAECON), 2017 IEEE National*, pp. 338-342. IEEE, 2017.

[17] Onabajo, Marvin, and Jose Silva-Martinez. “*Analog circuit design for process variation-resilient systems-on-a-chip*.” Springer Science & Business Media, 2012.

[18] Borkar, Shekhar, Tanay Karnik, Siva Narendra, James Tschanz, Ali Keshavarzi, and Vivek De. "Parameter variations and impact on circuits and microarchitecture." In *Proceedings 2003. Design Automation Conference (IEEE Cat. No. 03CH37451)*, pp. 338-342. IEEE, 2003.

[19] Chiu, Fu-Chien, “A Review on Conduction Mechanisms in Dielectric Films”, Advances in Materials Science and Engineering Volume 2014, Article ID 578168, 18 pages http://dx.doi.org/10.1155/2014/578168.

[20] Toshiba Memory Corporation, “Toshiba Memory Corporation Reliability Handbood.” Ver.2, July 2018, pp. 116-120.

[21] Micron Technology Inc., “Uprating Semiconductors for High-Temperature Applications.” Technical Note, Rev. F 2004, pp. 3-6.

[22] O'Connor, Robert, Greg Hughes, and Thomas Kauerauf. "Time-Dependent Dielectric Breakdown and Stress-Induced Leakage Current Characteristics of 0.7-nm-EOT HfO_2 pFETs." *IEEE Transactions on Device and Materials Reliability* 11, no. 2 (2011): 290-294.

[23] Kangqiao Hu, Abdullah Nazma Nowroz, Sherief Reda and Farinaz Koushanfar, “High-Sensitivity Hardware Trojan Detection Using Multimodal Characterization”, Proceedings of the Conference on Design, Automation and Test in Europe, pp. 1271-1276. EDA Consortium, 2013.

[24] Wang, Xiaofei, John Keane, Tony Tae-Hyoung Kim, Pulkit Jain, Qianying Tang, and Chris H. Kim. "Silicon odometers: Compact in situ aging sensors for robust system design." IEEE micro 34, no. 6 (2014): 74-85.

[25] Lu, Pong-Fei, and Keith A. Jenkins. "A built-in BTI monitor for long-term data collection in IBM microprocessors." In 2013 IEEE International Reliability Physics Symposium (IRPS), pp. 4A-1. IEEE, 2013.

[26] Retamal, José Ramón Durán, Chin-Hsiang Ho, Kun-Tong Tsai, Jr-jian Ke, and Jr-Hau He. "Self-organized Al nanotip electrodes for achieving ultralow-power and error-free memory." *IEEE Transactions on Electron Devices* 66, no. 2 (2019): 938-943.