Resistive-RAM for Data Storage Applications

by

Siddharth Gaba

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in the University of Michigan
2014

Doctoral Committee

Associate Professor Wei Lu, Chair
Assistant Professor Emmanouil Kioupakis
Professor Jerome P. Lynch
Assistant Professor Zhengya Zhang
Associate Professor Zhaohui Zhong

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Non-volatile memories play an indispensable role in today's electronic systems and information driven society. Mainstream non-volatile memory technology, dominated by the floating gate transistor, has historically improved in density, performance and cost primarily by means of process scaling. This simple geometrical scaling now faces significant challenges due to fundamental constraints of electrostatics and reliability. Therefore novel non-transistor based memory paradigms are being widely explored by both industry and academia. Among the various contenders for next generation storage technology, resistive switching memory devices have got immense attention due to their high speed, multilevel capability, scalability, simple structure, low voltage operation and high endurance.

In this thesis, we present studies on resistance switching memory devices. Electrical and material characterization was carried out on a metal-insulator-metal device system and formation / annihilation of nanoscale filaments was shown to be the reason behind the resistance switching behavior. The metal/insulator/metal system was optimized to include an in-cell resistor which was shown to improve device endurance and reduce stuck-at-one faults. For highest density, the devices were arranged in a crossbar geometry and vertically integrated on CMOS decoders to demonstrate the feasibility of practical data storage applications.

Next, we show that these binary resistive switching devices can exhibit native stochastic nature of resistive switching. Even for a fixed voltage on the same device, the wait time associated with programming is not fixed but rather random and broadly distributed. However, the probability of switching can be predicted and controlled by the applied voltage and the pulse width used to program the device. These binary devices have been used to generate random bit streams with predicable bias ratios in time and space domains. The ability to produce random bit streams using binary resistive switching devices based on the native stochastic switching principle may potentially lead to novel non-von-Neumann, alternative computing paradigms.

Further, sub-1nA operating current devices showing pronounced rectifying behavior have been developed. This ultra-low current provides energy savings by minimizing programming erase and read currents. Despite having such low currents, excellent retention, on/off ratio and endurance have been demonstrated. Devices programmed with less than 1nA peak current pass 6 hour retention test at 85 °C and show no significant degradation after 10000 write/erase cycles and with good switching uniformity. Due to the partially written filament, the devices exhibit pronounced non-linear I-V and current rectification in the on-state at the low bias regime – both factors are very beneficial for array operation. Also, the filament shape can be modulated by controlling the compliance current to obtain multilevel storage.

Finally a scalable approach to simple 3D stacking is discussed. By implementation of a vertical sidewall-based architecture, the number of critical lithography steps can be reduced. A vertical device structure based on a W / WO_x / Pd material system is developed. The devices show well-defined incremental resistance switching behavior and good endurance. The devices can be programmed with less than 10% mismatch and no apparent crosstalk. This scalable architecture is well suited for development of analog memory and neuromorphic systems.

Chapter 1
Introduction

1.1 Data Explosion

Our society's ever growing demand for information and entertainment has led to an explosive growth in the amount of storage capacity in almost every electronic device we use – from laptops, phones and tablets used for personal communication, entertainment and business to large commercial data centers and server farms where a big chunk of the world's big data resides. Non-volatile memory is also extensively used in controller chips that have touched every aspect of modern society from programmable coffee makers, microwaves, toys to smart sensors in cars and buildings to automated tools used in manufacturing. Data retrieval times and instant access needs have seen intense increase of popularity and adoption of FLASH memory compared to the older and slower hard disk technology.

1.2 The NVM Workhorse: FLASH

Flash memory was developed by Toshiba in 1980s from electrically erasable programmable read-only memory (EEPROM). Flash memory can be written and read at the byte level (NOR Flash) or in in small blocks (NAND flash). Both types of flash – NAND and NOR- rely on the floating gate transistor to store information (Fig. 1-1). The floating gate transistor adds an extra gate and a tunneling oxide in addition to a normal MOS transistor. Programming is usually

done through FN tunneling or hot carrier injection and erasing of the cell is achieved by FN tunneling. The cell is nonvolatile in nature – charge trapped in the floating gate stays there for many years. It is this relatively simple structure and excellent data retention property that have made flash storage so popular in today's world.

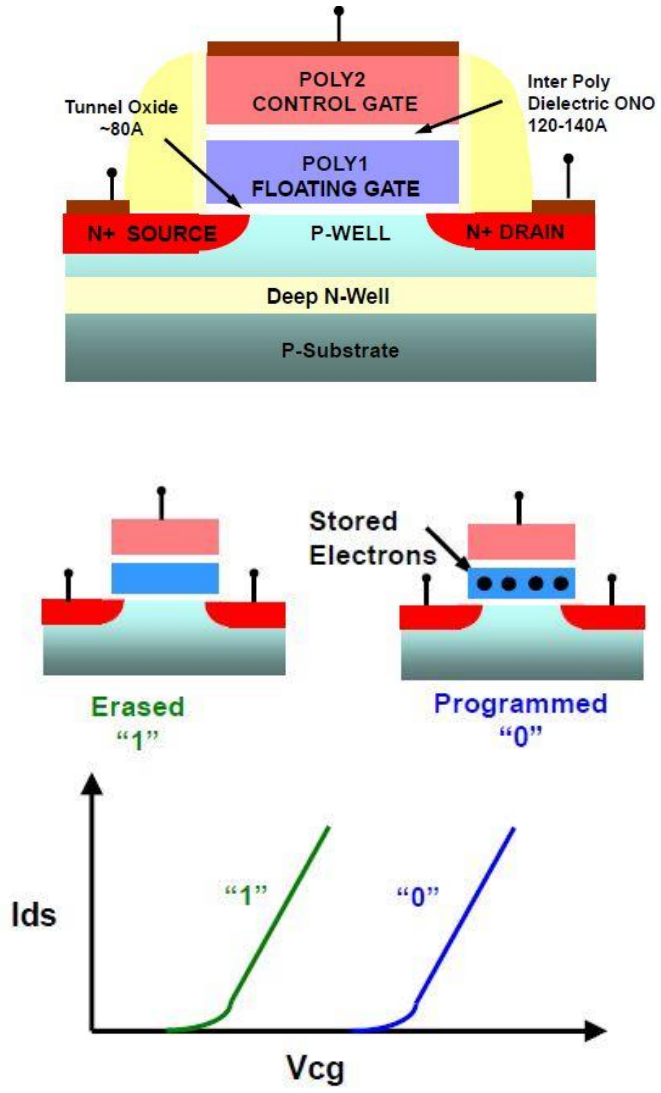


Figure 1-1: Flash memory works by adding (charging) or removing (uncharging) electrons to and from a floating gate. A bit's 1 or 0 state depends upon whether the floating gate is charged or not.

While both NOR and NAND are based on floating gate transistors they use different logical connection of these transistors to map data. NOR flash provides high-speed random access,

reading and writing data in specific memory locations. Thus, NOR is typically used to store code such as cell phones' operating systems and in BIOS chips. NAND flash reads and writes sequentially at high speed, handling data in small blocks called pages. This type of flash is typically used to store data in solid-state and USB flash drives, digital cameras, audio and video players, and TV set-top boxes. NAND flash reads faster than it writes, quickly transferring whole pages of data. Less expensive than NOR flash, NAND flash technology offers higher capacity for the same-size silicon and dominates today's flash market.

1.3 Semiconductor Flash Memory Scaling

Flash memory application has seen explosive growth in recent years and this trend is likely to continue because new and more demanding applications are constantly added partly due to the need for low power solid-state storage and partly due to rapidly declining prices. Conventional floating gate flash memories, no matter in NOR or NAND architecture, however, face steep challenges[1–4]. In addition to conventional CMOS scaling issues like short channel effects, loss of gate control and patterning issues, flash memory faces additional unique challenges.

For tight spacing rules, floating gate interference and the need for sufficient gate control (gate coupling ratio) have questioned the continued scaling of the floating gate device below 10nm. With scaling, crosstalk between the neighboring floating gates, word lines and even the adjacent channels becomes non-negligible and starts to cause data corruption.

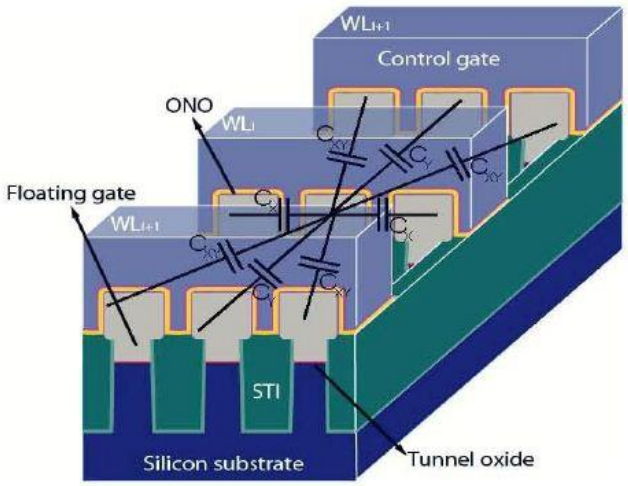


Figure 1-2: Diagram showing capacitive coupling in flash memory. Reproduced from [2]

Even though the channel dimensions scale, the dielectric thickness does not scale and is limited by the cell retention to about 50nm. The high voltage programming circuitry (~20V) needed for older generation also does not move towards lower voltages as the devices scale. Thus, running high voltages on metal lines with very small spacing raises endurance and reliability issues.

Further, as cell capacitance scales, less number for electrons are stored per unit shift in the threshold voltage. This causes higher impact and degradation from single electron events and increased 1/f noise. Channel electron trapping and de-trapping during read cause current to fluctuate resulting in increased random telegraph noise. Due to scaled channel lengths, channel dopant fluctuations increase the threshold voltage distributions. Continued degradation in threshold distribution width from cell scaling will eventually become unacceptable.

Fundamentally, the scalability of flash is based on charge which is quantized. With shrinking geometries, the number of electrons stored in the floating gate is fast reducing. At 20nm, there are less than 100 electrons present in the floating gate for a threshold of 1V. If 10% charge

loss during device lifetime is assumed to be within spec and retention specification is 10 years, this translates to 1 electron being lost per year.

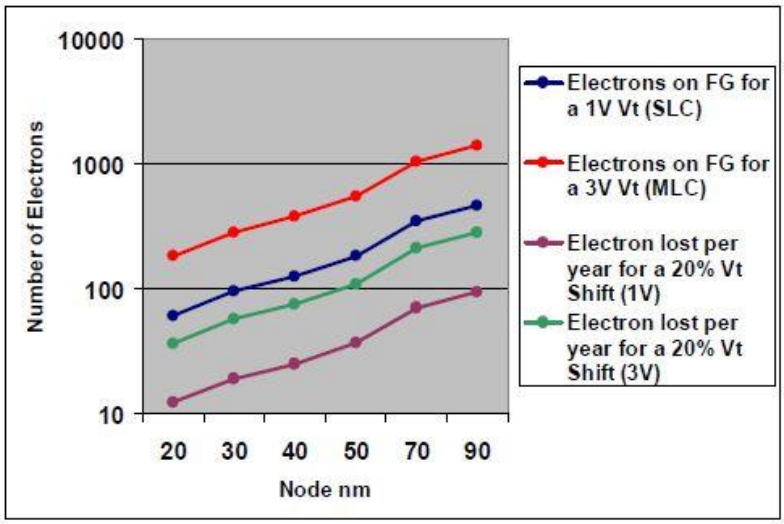


Figure 1-3: The number of electrons stored vs. lithographic dimension and electron loss requirement for retention. Reproduced from [2]

While recent development of 3D flash memory[5–9] will likely keep flash's dominance for the near future, such fundamental issues with flash scaling have forced both industry and academia to consider other non-charge based storage. Some of the more prominent emerging non-volatile memory devices are discussed on the next section.

1.4 Emerging Non-Volatile Memories

Due to this uncertain future of flash memory and the associated challenges, other non-volatile memory architectures are being explored. These include PCRAM, FeRAM and MRAM.

PCRAM or phase change random access memory (also known as phase change memory, ovonic unified memory or chalcogenide memory) relies on phase change of a material from an amorphous state to a crystalline state[10–12]. Each state has a distinct electrical conductivity which is sensed with a read process and interpreted as a 1 or a 0 for information storage purposes.

Phase change memory bit cells use localized joule heating to convert the switching material (generally a chalcogenide) from an amorphous phase (low conductivity or a binary 0) to a crystalline phase (high conductivity or a binary 1) or vice versa. In almost all prototype PCRAM devices, a chalcogenide alloy of germanium, antimony and tellurium (GeSbTe or GST) is used[13]. When heated to a high temperature (>600 °C), it loses its crystallinity and can be frozen in an amorphous glass-like state. By heating it again to a temperature above its crystallization point, but below its melting point, the material transforms into a low resistance conductive state.

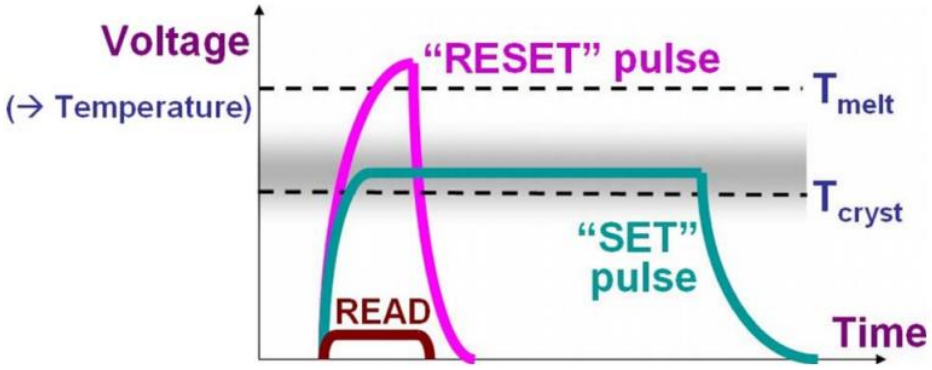


Figure 1-4: Programming of a PCM device involves application of electrical power through applied voltage, leading to internal temperature changes that either melt and then rapidly quench a volume of amorphous material during reset, or heat this volume at a lower temperature for sufficient time for recrystallization during set. A low voltage is used to sense the device resistance read so that the device state is not perturbed. Reproduced from [11].

Relatively high currents are needed for the crystalline to amorphous transition while a relatively long time (of the order of a 100ns) is used to fully convert the amorphous state to crystalline state. This continuous heating and quenching process results in a material expansion and contraction and causes significant thermal stress, which results in low device endurance[14]. Further, due to the thermal process involved, crosstalk between neighboring cells becomes an issue in large arrays.

FeRAM or ferroelectric RAM is similar in construction to a DRAM but uses a ferroelectric layer (typically lead zirconate titanate or PZT) instead of a dielectric layer in the capacitor to achieve non-volatility[15,16]. When an external electric field is applied across the ferroelectric layer, the dipoles tend to align themselves with the field direction, produced by small shifts in the positions of atoms and shifts in the distributions of electronic charge in the PZT crystal structure. After the charge is removed, the dipoles retain their polarization state. Binary "0"s and "1"s are stored as one of two possible electric polarizations in each data storage cell. It offers very low power consumption, very fast write performance and excellent endurance (often exceeding 1e16 cycles). However, a destructive read process makes its use as a replacement for flash less attractive. Further, CMOS compatibility, cost and density are the inhibiting factors which have prevented large scale adoption of FeRAM. While FeRAM devices from some manufacturers like Ramtron can now be found in niche areas like some electricity meters, air bag controllers, RAID disk controllers, RFID systems, FeRAM technology has failed to challenge flash as the medium of choice for large-scale data storage.

MRAM or magnetic RAM, in its simplest form, uses a magnetic tunneling junction to store information[17–19]. Two ferromagnetic plates, each of which can hold a magnetic field, are separated by an insulator. While one of the plates holds a permanent magnetic field, the field direction of the other can be switched. Due to the tunneling magnetic resistance effect, the electrical resistance of the cell changes due to the orientation of the fields in the two plates. By measuring the resulting current, the resistance inside any particular cell can be determined, and from this the polarity of the writable plate. Typically if the two plates have the same polarity this is considered to mean "1", while if the two plates are of opposite polarity the resistance will be higher and this translates to "0". Currently only one company (Everspin) produces a commercially

available 4Mbit part using an older 180nm process. MRAM still remains largely in the development phase. The large programming current and scalability issues (crosstalk issues when cell size scales) prevent it from being cost effective and challenging the well-established flash memory market.

Currently PCRAM, FeRAM and MRAM are in small volume production, but remain limited to niche applications. All the above have transitioned to the a commercial arena but have still failed to replace Flash

Resistive memory has recently emerged as another contender in the non-volatile memory race. The next section introduces this memory concept and discusses why it stands to challenge the flash dominance.

1.5 RRAM

A RRAM device (or a resistive random access memory device) is a two terminal device in which a switching medium (or matrix) is sandwiched between top and bottom electrodes (Fig.1.1). The resistance of the device is modulated by applying voltage or current signals to the electrode(s). Unlike PCRAM, there is no phase change involved in RRAM.

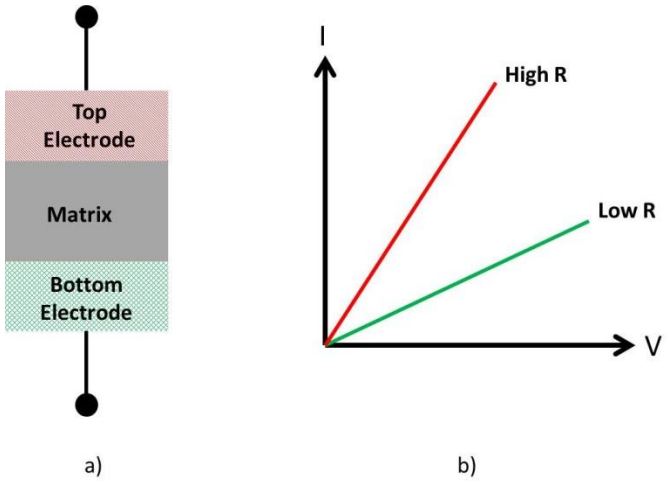


Figure 1-5: a) A RRAM cell has a very simple structure - matrix material sandwiched between two electrodes. b) The matrix resistance (high or low) is read out using a low voltage.

The basic idea is very simple and elegant: that a dielectric, which is normally insulating, can be made to conduct through a filament or conduction path formed after application of a sufficiently high voltage (or current). The conduction path formation can arise from different mechanisms, including defects, metal migration, etc. Once the filament is formed, it may be reset (broken, resulting in high resistance) or set (re-formed, resulting in lower resistance) by an appropriately applied voltage. This simple two terminal structure, among other things, is a reason for the popularity of RRAM devices among non-volatile memory researchers.

The first resistive switching effects were reported in the early 1960s[20]. Memories based on this concept never caught on due to concerns about scalability and stability and because of the emergence and rapid successful development of Si-based memories. Renewed interest in early 2000s[21–23] brought back the RRAM concept into limelight and it has been the focus of intense research in the last decade. The resistance switching effect has been observed in a broad range of materials including perovskites[24–27], binary metal oxides[28–31] and solid electrolytes[32–37].

While the exact classification based on mechanism is still debatable, these devices are often broadly sub-divided into two major categories based on the switching behavior: bipolar devices and unipolar devices. Bipolar devices need opposite voltage polarities for the two resistance transitions. For example, a device goes from high resistance (RESET) state to a low resistance (SET) state by application of a positive voltage higher than the threshold voltage. To RESET the device, a negative voltage must be applied. Since the devices show preference for switching polarities, the associated switching mechanism is electric field driven. Commonly used models and hypothesis include metal ion migration and redox processes [38–40] within a switching medium) and drift of oxygen vacancies[41,42]. Devices based on ion migration are often called conductive-bridge RAM (or CBRAM) while devices based on drift of oxygen vacancies are often simply called RRAM. This is different from unipolar devices where a single voltage polarity is enough to cause both transitions: SET to RESET and, also, RESET to SET. Because of lack of voltage polarity preference, Joule heating /thermal breakdown models are often used to describe such resistance switching behavior. RRAM has great potential as the memory of choice for next generation storage requirements. Non-volatility[39], sub-ns switching[43], high ON/OFF ratios[44], low switching voltages[39], (moderately) low current operation[45–48], capability to integrate multi-level characteristics[47,49,50] and high endurance have been demonstrated. A simple two terminal structure and absence of any select transistor has led to highly scaled bit cells often limited only by lithography. Further, RRAM does not suffer from any traditional scaling issues like short channel effects and loss of gate control. Since information is not stored as charge, there is no leakage or capacitive coupling issue like that in flash and DRAM. High performance devices have been shown with CMOS compatible materials and with minimum process

overheads[29, 41]. As a result, RRAM is considered as one of the most promising approaches for next generation memory needs, according to researchers in this field.

However, industry analysts and commercial companies have been a bit reserved in promoting RRAM primarily because the underlying mechanisms are not completely understood and the lack of systematic analysis of device performance including failure modes. Physics/device models still remain at a nascent stage. Most of the work done in the RRAM area has been experimental in nature and has thrown up a large number of materials as possible candidates for a commercially viable RRAM structure. In many studies, performance data comes from single isolated devices and not integrated large scale systems. Yield, cycle to cycle and device to device uniformity and array operation issues are also relatively unexplored areas.

This thesis attempts to explore some of these unknowns and endeavors to build a more holistic picture of RRAM devices and RRAM based systems.

1.6 Organization of the Thesis

In chapter 1, several essential topics pertaining to nonvolatile memory in general and resistive memory specifically have been discussed. In chapter 2, physical characterization studies (particularly in-situ transmission electron microscopy) are detailed. These studies explain the filamentary nature of resistive switching. Over-programming of the devices is shown to be linked to formation of robust filaments which can be prevented by incorporation of an in-cell resistor. This optimized structure is vertically integrated on a CMOS chip to show the feasibility of high density resistive memory and hybrid CMOS / RRAM systems.

In chapter 3, natural variations in switching parameters in resistive switching devices are shown to be due to the inherent stochastic nature of the switching process itself. Instead of forcing

the devices to behave in a deterministic way, the native non-determinism can be used to generate bit-streams. The use of these bit streams in stochastic computing are discussed.

The need for reduction of the operation current for resistive memories is described in Chapter 4. The retention-operating current tradeoff is discussed along with development of a novel device structure which can be programed and erased with less than 1nA. Despite such low currents, the device demonstrates excellent retention characteristics.

While all devices discussed in chapters 1-4 are 2D in nature, 3D scaling is discussed in chapter 5. A scalable device structure based on vertical sidewalls is developed. Mismatch between devices and properties like endurance and crosstalk are studied and characterized.

An implementation of a hybrid RRAM/CMOS architecture and future optimizations and studies are briefly mentioned in chapter 6.

References

[1] C.-Y. Lu, K.-Y. Hsieh, and R. Liu, "Future challenges of flash memory technologies," Microelectron. Eng., vol. 86, no. 3, pp. 283–286, Mar. 2009.

[2] K. Prall, "Scaling Non-Volatile Memory Below 30nm," 2007 22nd IEEE Non-Volatile Semicond. Mem. Work., pp. 5–10, 2007.

[3] A. Fazio, "Flash memory scaling," MRS Bull., vol. 29, no. 11, pp. 814–817, 2004.

[4] B. Govoreanu, D. P. Brunco, and J. Van Houdt, "Scaling down the interpoly dielectric for next generation Flash memory: Challenges and opportunities," Solid. State. Electron., vol. 49, no. 11, pp. 1841–1848, Nov. 2005.

[5] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and a. Nitayama, "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory," 2007 IEEE Symp. VLSI Technol., pp. 14–15, Jun. 2007.

[6] A. Hubert, E. Nowak, K. Tachi, C. Vizioz, C. Arvet, and J. Colonna, "A stacked SONOS technology , up to 4 levels and 6nm crystalline nanowires , with gate-all-around or independent gates (Φ -Flash), suitable for full 3D integration," Electron Devices Meet. (IEDM), 2009 IEEE Int. IEEE, pp. 637–640, 2009.

[7] S. Whang, K. Lee, D. Shin, B. Kim, M. Kim, J. Bin, J. Han, S. Kim, B. Lee, Y. Jung, S. Cho, C. Shin, H. Yoo, S. Choi, K. Hong, S. Aritome, S. Park, and S. Hong, "Novel 3-dimensional Dual Control-gate with Surrounding Floating-gate (DC-SF) NAND flash cell for 1Tb file storage application," 2010 Int. Electron Devices Meet., pp. 29.7.1–29.7.4, Dec. 2010.

[8] E.-S. Choi, H.-S. Yoo, H.-S. Joo, G.-S. Cho, S.-K. Park, and S.-K. Lee, "A Novel 3D Cell Array Architecture for Terra-Bit NAND Flash Memory," 2011 3rd IEEE Int. Mem. Work., pp. 1–4, May 2011.

[9] A. Katsumata, R., Kito, M., Fukuzumi, Y., Kido, M., Tanaka, H., Komori, Y., Ishiduki, M., Matsunami, J., Fujiwara, T., Nagata, Y. , Li Zhang, Iwata, Y., Kirisawa, R., Aochi, H., Nitayama, "Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices," 2009 Symp. VLSI Technol., pp. 136–7, 2009.

[10] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase Change Memory," Proc. IEEE, vol. 98, no. 12, pp. 2201–2227, Dec. 2010.

[11] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. a. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. S. Shenoy, "Phase change memory technology," J. Vac. Sci. Technol. B Microelectron. Nanom. Struct., vol. 28, no. 2, p. 223, 2010.

[12] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage.," Nat. Mater., vol. 6, no. 11, pp. 824–32, Nov. 2007.

[13] D. Lencer, M. Salinga, B. Grabowski, T. Hickel, J. Neugebauer, and M. Wuttig, "A map for phase-change materials.," Nat. Mater., vol. 7, no. 12, pp. 972–7, Dec. 2008.

[14] L. Goux, D. Tio Castro, G. a. M. Hurkx, J. G. Lisoni, R. Delhougne, D. J. Gravesteijn, K. Attenborough, and D. J. Wouters, "Degradation of the Reset Switching During Endurance Testing of a Phase-Change Line Cell," IEEE Trans. Electron Devices, vol. 56, no. 2, pp. 354–358, Feb. 2009.

[15] G. R. Fox, F. Chu, and T. Davenport, "Current and future ferroelectric nonvolatile memory technology," J. Vac. Sci. Technol. B Microelectron. Nanom. Struct., vol. 19, no. 5, p. 1967, 2001.

[16] S. L. Miller and P. J. McWhorter, “Physics of the ferroelectric nonvolatile memory field effect transistor,” J. Appl. Phys., vol. 72, no. 12, p. 5999, 1992.

[17] J. Slaughter and R. Dave, “Fundamentals of MRAM technology,” J. Supercond., vol. 15, no. 1, pp. 19–25, 2002.

[18] S. Tehrani, B. Engel, J. M. Slaughter, E. Chen, M. DeHerrera, M. Durlam, P. Naji, R. Whig, J. Janesky, and J. Calder, “Recent developments in magnetic tunnel junction MRAM,” IEEE Trans. Magn., vol. 36, no. 5, pp. 2752–2757, 2000.

[19] Y. Huai, “Spin-transfer torque MRAM (STT-MRAM): Challenges and prospects,” AAPPS Bull., vol. 18, no. 6, pp. 33–40, 2008.

[20] J. F. Gibbons, “Switching properties of thin NiO films,” Solid. State. Electron., vol. 7, no. 2, pp. 785–797, 1964.

[21] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, “Reproducible switching effect in thin oxide films for memory applications,” Appl. Phys. Lett., vol. 77, no. 1, p. 139, 2000.

[22] A. Shimaoka, K. Inoue, T. Naka, N. A. K. Sakiyama, Y. Wang, S. Q. Liu, N. J. Wu, and A. Ignatiev, “Novell Colossal Magnetoresistive Thin Film Nonvolatile Resistance Random Access Memory (RRAM),” IEDM, pp. 193–196, 2002.

[23] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found.,” Nature, vol. 453, no. 7191, pp. 80–3, May 2008.

[24] X. Chen, N. Wu, J. Strozier, and A. Ignatiev, “Spatially extended nature of resistive switching in perovskite oxide thin films,” Appl. Phys. Lett., vol. 89, no. 6, p. 063507, 2006.

[25] K. Szot, W. Speier, G. Bihlmayer, and R. Waser, “Switching the electrical resistance of individual dislocations in single-crystalline SrTiO3.,” Nat. Mater., vol. 5, no. 4, pp. 312–20, Apr. 2006.

[26] M. Hasan, R. Dong, H. J. Choi, D. S. Lee, D.-J. Seong, M. B. Pyun, and H. Hwang, “Uniform resistive switching with a thin reactive metal interface layer in metal-La0.7Ca0.3MnO3-metal heterostructures,” Appl. Phys. Lett., vol. 92, no. 20, p. 202102, 2008.

[27] A. Wang, “Bistable resistive switching of a sputter-deposited Cr-doped SrZrO3 memory film,” IEEE Electron Device Lett., vol. 26, no. 6, pp. 351–353, Jun. 2005.

[28] Z. Wei, Y. Kanzawa, and K. Arita, “Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism,” Meet. 2008. IEDM, 2008.

[29] J. J. Yang, M.-X. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Medeiros-Ribeiro, and R. S. Williams, "High switching endurance in TaO x memristive devices," Appl. Phys. Lett., vol. 97, no. 23, p. 232102, 2010.

[30] B. Govoreanu, G. S. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl, M. Jurczak, B.- Leuven, and K. U. Leuven, "10x10nm2 Hf / HfOx Crossbar Resistive RAM with Excellent Performance , Reliability and Low-Energy Operation," Electron Devices Meet. (IEDM), 2011 IEEE Int. IEEE, pp. 729–732, 2011.

[31] S. Yu, X. Guan, and H.-S. P. Wong, "Conduction mechanism of TiN/HfOx/Pt resistive switching memory: A trap-assisted-tunneling model," Appl. Phys. Lett., vol. 99, no. 6, p. 063507, 2011.

[32] D. Deleruyelle, M. Putero, T. Ouled-khachroum, M. Bocquet, M. Coulet, C. Calmes, C. Muller, E. Nationale, M. De Saint, C. D. M. De, P. G. Charpak, and D. Mimet, "Ge2Sb2Te5 layer used as solid electrolyte in Conductive-Bridge memory devices fabricated on flexible substrate," Solid State Electron., vol. 79, pp. 159–165, 2013.

[33] Q. Liu, J. Sun, H. Lv, S. Long, K. Yin, N. Wan, Y. Li, L. Sun, and M. Liu, "Real-time observation on dynamic growth/dissolution of conductive filaments in oxide-electrolyte-based ReRAM.," Adv. Mater., vol. 24, no. 14, pp. 1844–9, Apr. 2012.

[34] T. Fujii, M. Arita, Y. Takahashi, and I. Fujiwara, "In situ transmission electron microscopy analysis of conductive filament during solid electrolyte resistance switching," Appl. Phys. Lett., vol. 98, no. 21, p. 212104, 2011.

[35] S. Z. Rahaman, S. Maikap, H.-C. Chiu, C.-H. Lin, T.-Y. Wu, Y.-S. Chen, P.-J. Tzeng, F. Chen, M.-J. Kao, and M.-J. Tsai, "Bipolar Resistive Switching Memory Using Cu Metallic Filament in Ge[sub 0.4]Se[sub 0.6] Solid Electrolyte," Electrochem. Solid-State Lett., vol. 13, no. 5, p. H159, 2010.

[36] J. Jang, F. Pan, K. Braam, and V. Subramanian, "Resistance switching characteristics of solid electrolyte chalcogenide Ag(2)Se nanoparticles for flexible nonvolatile memory applications.," Adv. Mater., vol. 24, no. 26, pp. 3573–6, Jul. 2012.

[37] S.-J. Choi, G.-S. Park, K.-H. Kim, S. Cho, W.-Y. Yang, X.-S. Li, J.-H. Moon, K.-J. Lee, and K. Kim, "In situ observation of voltage-induced multilevel resistive switching in solid electrolyte memory.," Adv. Mater., vol. 23, no. 29, pp. 3272–7, Aug. 2011.

[38] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, "Device and SPICE modeling of RRAM devices.," Nanoscale, vol. 3, no. 9, pp. 3833–40, Sep. 2011.

[39] S. H. Jo and W. Lu, "CMOS compatible nanoscale nonvolatile resistance switching memory.," Nano Lett., vol. 8, no. 2, pp. 392–7, Feb. 2008.

[40] S. H. Jo, K.-H. Kim, and W. Lu, “Programmable resistance switching in nanoscale two-terminal devices.,” Nano Lett., vol. 9, no. 1, pp. 496–500, Jan. 2009.

[41] T. Chang, S.-H. Jo, K.-H. Kim, P. Sheridan, S. Gaba, and W. Lu, “Synaptic behaviors and modeling of a metal oxide memristive device,” Appl. Phys. A, vol. 102, no. 4, pp. 857–863, Feb. 2011.

[42] N. Raghavan, K. Pey, and X. Li, “Very low reset current for an RRAM device achieved in the oxygen-vacancy-controlled regime,” Electron Device Lett., vol. 32, no. 6, pp. 716–718, 2011.

[43] L. Goux, K. Sankaran, G. Kar, N. Jossart, K. Opsomer, R. Degraeve, G. Pourtois, G. Rignanese, and C. Detavernier, “Field-driven ultrafast sub-ns programming in W \ Al2O3 \ Ti \ CuTe-based 1T1R CBRAM system,” VLSI Technol. (VLSIT), 2012 Symp., pp. 69–70, 2012.

[44] S. H. Jo, K.-H. Kim, and W. Lu, “High-density crossbar arrays based on a Si memristive system.,” Nano Lett., vol. 9, no. 2, pp. 870–4, Feb. 2009.

[45] Y. Chen, H. Lee, P. Chen, W. Chen, K. Tsai, P. Gu, T. Wu, C. Tsai, S. Z. Rahaman, Y. Lin, F. Chen, M. Tsai, and T. Ku, “Novel Defects-Trapping TaOx / HfOx RRAM With Reliable Self-Compliance, High Nonlinearity, and Ultra-Low Current.,” IEEE Electron Device Lett., vol. 35, no. 2, pp. 202–204, 2014.

[46] C. Ho, C. Hsu, C. Chen, J. Liu, C. Wu, C. Huang, C. Hu, and F. Yang, “9nm half-pitch functional resistive memory cell with <1μA programming current using thermally oxidized sub-stoichiometric WOx film,” 2010 Int. Electron Devices Meet., pp. 19.1.1–19.1.4, Dec. 2010.

[47] K.-H. Kim, S. Hyun Jo, S. Gaba, and W. Lu, “Nanoscale resistive memory with intrinsic diode characteristics and long endurance,” Appl. Phys. Lett., vol. 96, no. 5, p. 053106, 2010.

[48] W. Kim, S. Il Park, Z. Zhang, Y. Yang-liauw, D. Sekar, H. P. Wong, and S. S. Wong, “Forming-Free Nitrogen-Doped AlOx RRAM with Sub- μA Programming Current,” VLSI Technol. (VLSIT), Symp., pp. 22–23, 2011.

[49] M. Wu, Y. Lin, and W. Jang, “Low-Power and Highly Reliable Multilevel Operation in 1T1R RRAM,” Electron Device Lett., vol. 32, no. 8, pp. 1026–1028, 2011.

[50] M. Cell, P. M. C. Memory, U. Russo, S. Member, D. Kamalanathan, S. Member, D. Ielmini, A. L. Lacaita, M. N. Kozicki, and A. Programmable, “Study of Multilevel Programming in Programmable,” Electron Devices, IEEE Trans., vol. 56, no. 5, pp. 1040–1047, 2009.

[51] W. C. Chien, Y. C. Chen, K. P. Chang, E. K. Lai, Y. D. Yao, P. Lin, J. Gong, S. C. Tsai, S. H. Hsieh, C. F. Chen, K. Y. Hsieh, R. Liu, and C. Lu, “Multi - Level Operation Of Fully CMOS Compatible WOX Rresistive Random Access Memory (RRAM),” Mem. Work. 2009. IMW’09. IEEE Int. IEEE, vol. 91, pp. 1–2, 2009.

Chapter 2

Amorphous Silicon Based Electrochemical Metallization Memories

2.1 Introduction

While resistive switching effects are attributed to formation / dissolution of conducting filaments[1–7], particularly for CBRAMs, the exact nature of these filaments and the dynamic filament growth processes have generated much debate[1–4]. In fact, understanding the filament growth is essential in optimization and the possible commercialization of these resistive switches. Physical characterization of the metallic filaments is carried out by scanning and transmission electron microscopy using planar metal-insulator-metal devices (Section 2.2). The filament growth was found to be dominated by cation transport through the dielectric. This planar MIM structure was further studied in actual crossbar devices and incorporation of an in-cell resistor is shown to improve device endurance (Section 2.3). Finally, device arrays are vertically integrated on a CMOS chip to demonstrate control of these nano-filaments for actual data storage (Section 2.4).

2.2 Physical characterization of metallic filaments

2.2.1 Device Fabrication

For transmission electron microscopy (TEM) studies, planar resistive switches were fabricated on low stress SiN_x membranes (Ted Pella). The SiN_x membranes have a small thickness of 15 nm and produce a low background under TEM. The active dielectric layer thickness was also kept low at ~15 nm to further improve the electron transparency of the devices. The electrodes were fabricated by electron beam lithography (Raith 150) followed by electron beam evaporation and lift-off processes. A two-step lithography process was employed. The 25-nm Pt electrode with a 5-nm Ti adhesion was fabricated first, followed by the Ag electrode alignment, patterning, and lift-off (thickness ~40 nm). Similar device structures were also fabricated on Si/SiO_2 substrates for scanning electron microscopy (SEM) studies. The a-Si was deposited using plasma-enhanced chemical vapor deposition (GSI UltraDep 2000) at 350-400 °C

2.2.2 Electron Microscopy Studies

The fabricated device was switched by applying a positive bias on the Ag top electrode (TE) while keeping the Pt bottom electrode (BE) at ground potential. After ~30secs the sharp increase in the current in the I-t graph marks the switching of the device to the ON state (Fig 2-1b). This device was transferred to a TEM and a metal filament was observed (Fig. 2-1a). The filament grew from the active electrode and had the thinnest part near the inert electrode interface. Another device was programmed with much lower current (Fig. 2-1c) to reveal an incomplete but stable filament (Fig. 2-1d). The crystal structure of the partially formed filament was studied by selected area electron diffraction (SAED). Because both the switching medium (a-Si) and the supporting SiN_x membrane underneath are amorphous in nature, they appear as diffusive haloes in the SAED pattern (Fig. 2-1e). SAED analysis confirms that the filament is composed of

elemental Ag (instead of some Ag compounds) with fcc structure. An EDS line profile suggests the filament is made of a chain of nano-particles rather than a solid Ag filament (Fig. 2-1f-g). Further, bright field STEM of a Ag nano-particle shows lattice fringes which were indexed to the Ag fcc structure (Fig. 2-1i). The filament is found to be composed of a series of Ag nanoparticles(Fig. 2-1h) separated by gaps of nanometer scale (ranging from 1 to 4 nm). These nanoscale gaps allow efficient electron tunneling so that the conducting filament can still offer appreciable conductance in the on-state, even though it is not a solid wire.

The filament growth can be explained by formation of Ag cations and transport of these cations inside the dielectric. The Ag cations are then reduced inside the dielectric forming Ag nano-particles (Fig. 2-1j). The reduction can happen inside the dielectric when the cations capture electrons injected by the cathode. So the shape of the filament is determined by the location where the cations are reduced. If the cation mobility is high, the cations would be reduced at the cathode and the filament would seem to grow from the inert electrode. This case indeed has been seen in material systems where sputtered SiO_2 is used as the dielectric[8]. Due to various defects in sputtered SiO_2, the cation mobility is enhanced and the filament formed has the broadest part near the inert electrode which is opposite to what is observed in the a-Si case. Detailed analysis of the electrochemical processes and cation transport dynamics during filament growth can be found in our recent publication [8].

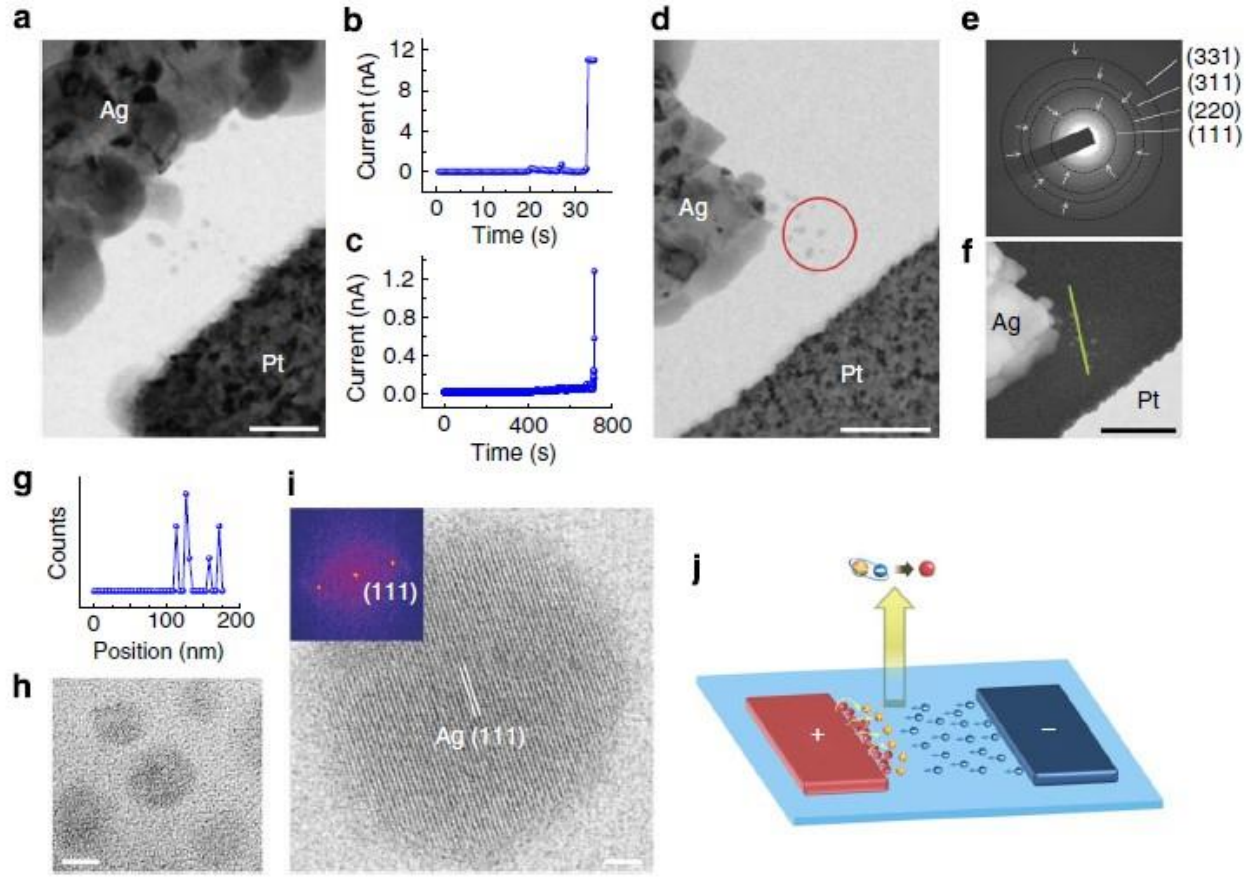


Figure 2-1: Observation of conducting filament dynamics in a-Si-based resistive memories. (a) TEM image of a complete conducting filament in a a-Si-based device. The thinnest region of the filament is at the dielectric/inert electrode interface (Scale bar, 50 nm). (b) The corresponding I–t curve for device in (a) during the forming process. The applied voltage was 8 V. (c) I–t curve for another device in (d) during the forming process. The applied voltage was 10 V. (d) TEM image of the device showing a partially formed filament. The filament apparently grew from the active electrode. The region indicated by the red circle was analyzed by SAED (Scale bar, 100 nm). (e) SAED pattern of the filament indexed with elemental Ag with fcc structure. (f) Z-contrast HAADF STEM image of the filament. Scale bar, 100 nm. The green line indicates the position for EDS line profile analysis. (g) Corresponding EDS line profile results showing the Ag signal intensity. (h) HRTEM image showing a group of nanoparticles inside the filament. Scale bar, 2 nm. (i) Bright-field STEM image of an Ag nanoparticle in the conducting filament. The lattice fringes were indexed to the Ag fcc structure. Scale bar, 2 nm. Inset: corresponding fast Fourier transformation results of the HRTEM image. (j) Schematic of the filament growth for a-Si-based resistive memories showing that the cations can be reduced inside the dielectric film by free electrons.

To further characterize the filament, multiple devices were programmed with different programming currents and an increase in filament size is observed with increasing programming current (Fig. 2-2). The filament increases both in width and length. Also, in hundreds of devices that were measured, no device had more than one complete filament. The formation of a single dominant filament is consistent with the hypothesis that the filament formation is a self-limiting

process, and, once a filament is completed, further filament growth will be suppressed owing to reduced electric field[9].

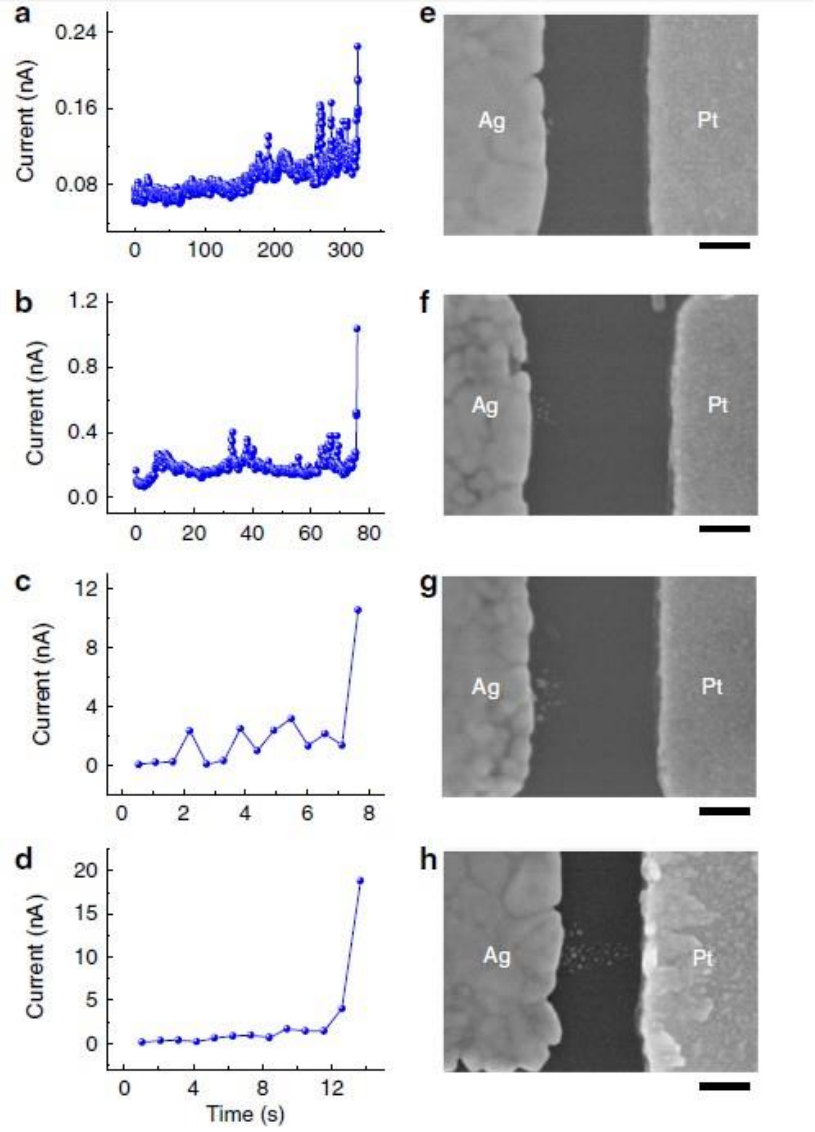


Figure 2-2: Controlling the filament size by limiting the programming current. (a–d) I–t curves with different programming currents during the forming process and (e–h) corresponding SEM images of the devices after forming, showing the correlation of the filament size with the programming current. Scale bar, 100 nm. Four different a-Si-based devices with similar geometry were used in this study. The applied voltages were (a) 18 V, (b) 26 V, (c) 22 V, and (d) 20 V, respectively.

Similar effects were also observed from a single device, where the application of a second programming process led to an extension of the existing filament towards the inert electrode, illustrating the possibility of step-by-step growth of the filament from the active electrode to the inert electrode in these devices (Fig. 2-3). These results indicate that the shape and size of the filament can be controlled during the programming stage and provide the foundation for multi-level storage within a single cell.

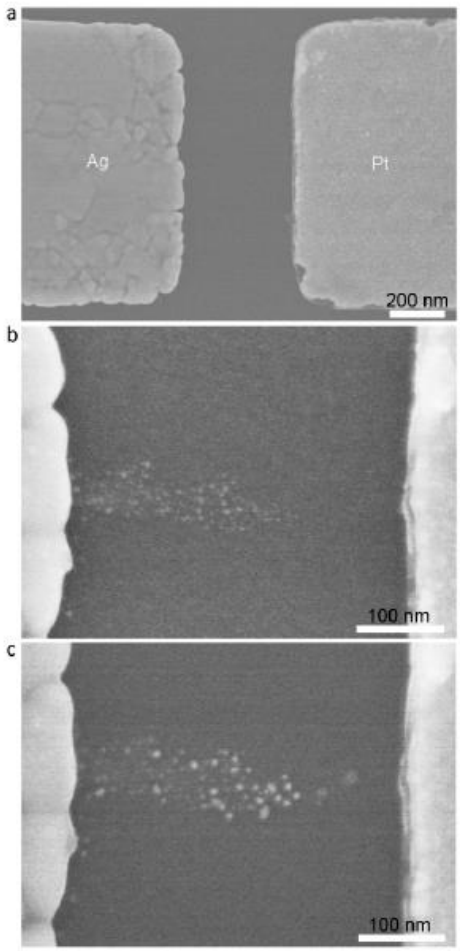


Figure 2-3: Step-by-step filament growth. (a) SEM image of the as-fabricated planar resistive switch based on a-Si. (b) Growth of an incomplete conducting filament after forming. (c) When a positive bias was applied again on the Ag electrode, the conductance of the device was increased for a second time. SEM imaging of the device shows an extension of the previous filament toward the Pt electrode

2.3 MIM vs. MIS

While the planar MIM devices described in the previous section are excellent for studying the filament growth and the dynamics of creation / annihilation of the conducting filament(s) upon application of SET and RESET voltages, crossbar structure is preferable from a density point of view. In addition to the density advantage, crossbar architecture based on two-terminal resistive switches has been the main focus among researchers due to its structural simplicity and large connectivity.

When actual crossbar devices were fabricated using the MIM structure, poor endurance was typically observed. The devices were very prone to getting stuck in the ON state and could not be cycled. In this section, we study the cause of the poor endurance of these MIM devices. The effect of transient currents on the endurance of the device was studied and presence of an in-cell resistor was shown to improve endurance significantly.

2.3.1 Device Fabrication

The memory devices studied were based on a cross-point structure, as illustrated in Fig. 2-4a, with a top electrode(100nm silver) and a bottom electrode (either 70nm p+ poly-silicon, resulting in an on-chip series-resistance of ~ 50 k-ohm, Case A or 70nm tungsten, Case B) sandwiching an insulating layer of amorphous silicon. All devices are fabricated on a silicon substrate with 100nm thermal oxide using electron-beam lithography and traditional etch /liftoff processes. A snapshot of the process flow for both kinds of devices is given in Fig. 2-4b. The area of the active cross-point is varied from 100nm x 100nm to 750nm x 750nm for all fabricated devices.

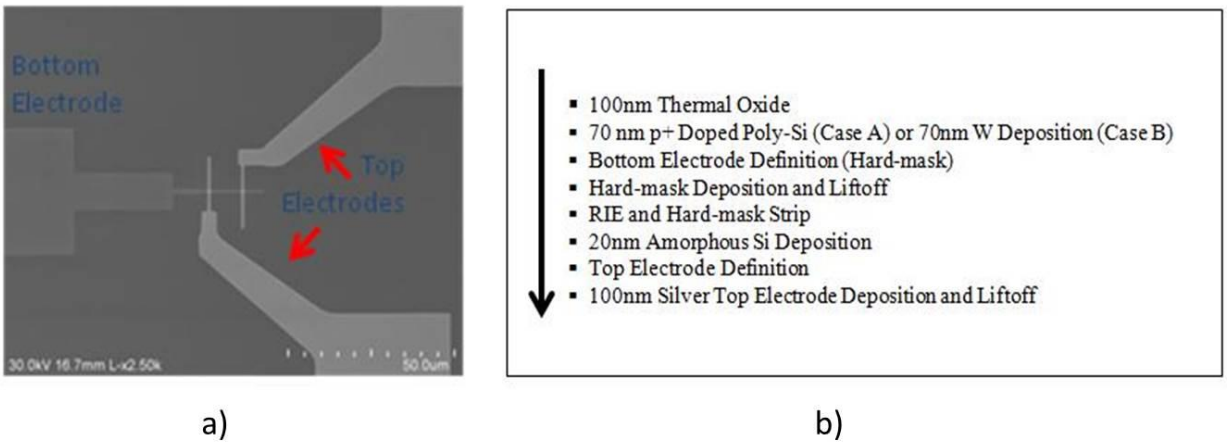


Figure 2-4: a) Scanning electron micrograph of a fabricated cell where two top electrodes share a common bottom electrode. b) An overview of the process flow for the two different devices. Except the bottom electrode material (p+ poly-Si vs. tungsten), all processing conditions were kept identical to allow a fair comparison.

2.3.2 Electrical Characterization

The electrical measurements were performed using a Keithley 4200 semiconductor characterization system or a customized LabVIEW-based measurement system in combination with a Desert Cryogenics TTP4 probe station. The bias voltage was always applied to the top electrode with the bottom electrode grounded during measurements. An external series resistor was used when measuring the tungsten bottom electrode devices but as the poly-BE already offers a series-resistance to the switching device, no external series resistor was used for the poly-silicon based devices.

The poly-BE based devices do not have appreciably high forming voltages (Fig. 2-5a and Fig. 2-5b) and show sharp switching characteristics (Fig. 2-5c).

.

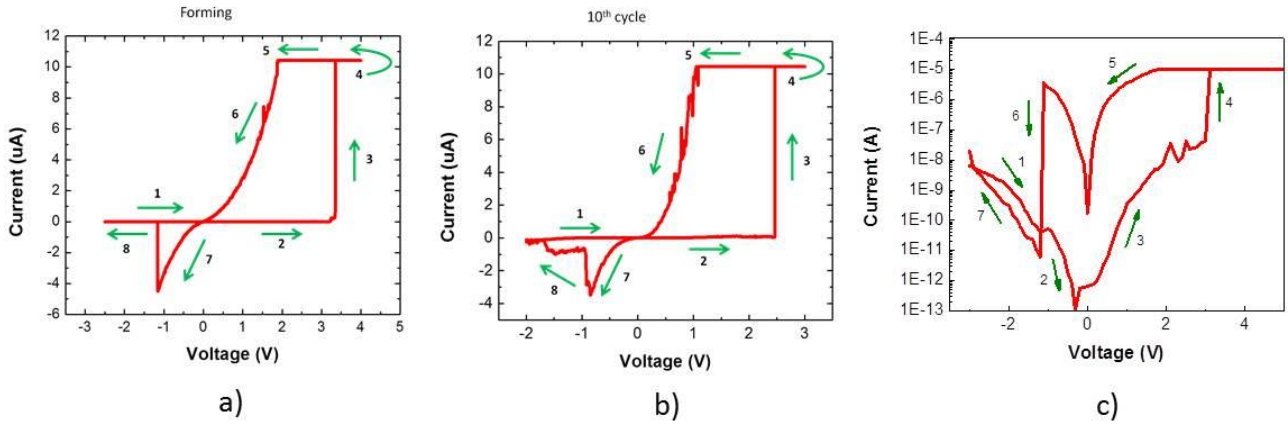


Figure 2-5: a) I-V sweep of a virgin device showing sharp switching at ~3.5V and non-linear I-V at on-state. b) Subsequent SET voltage is slightly lower than the initial forming voltage. Ron ~ 110 kohm. c) A typical log scale I-V curve demonstrating highly non-linear and sharp switching characteristics.

Devices exhibit high on-to-off current ratios (>10000) and are stable for more than 2 hours at room temperature (Fig. 2-6b). They can be cycled continuously and do not get stuck at the programmed or erased state (Fig. 2-6c).

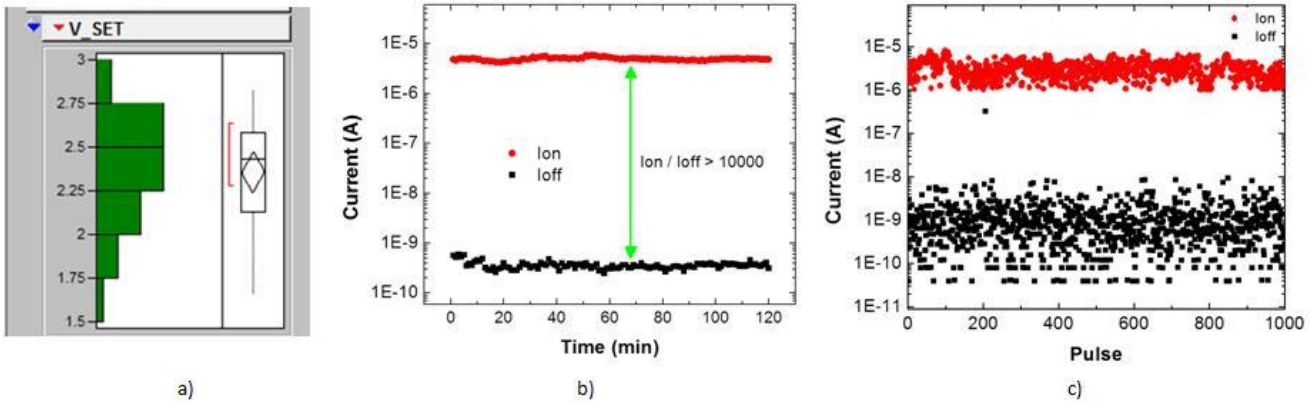


Figure 2-6 a) Distribution of VSET for 30 consecutive write cycles. b) Two hour retention data (red circles) and two hour read disturb data (black squares). Readout utilized a 1V-10ms pulse repeated every 60 seconds. Device was written with a 5V/400us pulse for the retention test and erased with a -4V/400us pulse for the read disturb test. c) Endurance Data. Devices can be cycled continuously and do not exhibit stuck-at-one (SA1) or stuck-at-zero (SA0) faults.

The metal-BE based devices show similar switching voltages (Fig. 2-7a) and retention behavior (Fig. 2-7b). However, in stark contrast with poly-silicon based devices, metal based devices have very limited endurance (Fig. 2-7c). Over 200 devices were tested and none yielded endurance numbers comparable to the polysilicon based devices.

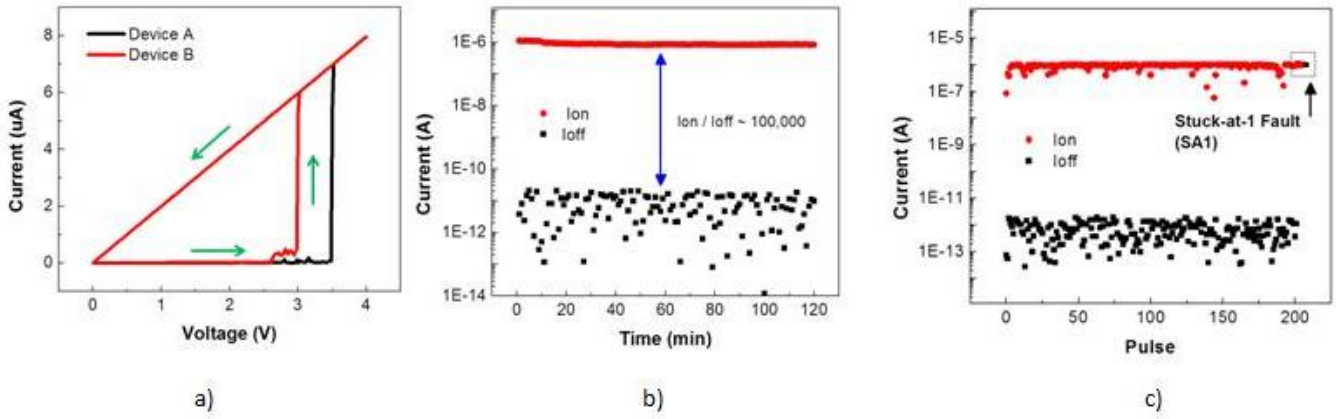


Figure 2-7: a) Switching curves from two devices. A 500k external series resistor was used. R_{on} ~ 5 kΩ. b) ON state retention (red) and OFF state read-disturb (black) data. A 1 MΩ series resistor was used in the setup and the device was read with a 1V-10ms pulse repeated every 60 seconds. c) Metal bottom electrode based devices have a significantly smaller endurance and tend to get stuck in the programmed state.

2.3.3 SPICE Simulation

Devices with poly-BE tend to show superior performance when compared to similar devices with metal-BE. This fundamental difference can be understood by modeling the two devices as shown in Fig. 2-8a and Fig. 2-8b. The 50 kΩ resistor depicts the resistance of the poly-BE in Case A and the external series resistor used in Case B. The 30 pF capacitor is used to model typical parasitic cabling capacitance from the measurement setup and is located between the external series resistor and the device-under-test (DUT) in Case B. SPICE model of the RRAM device is based on custom code as described in an earlier publication by us [9].

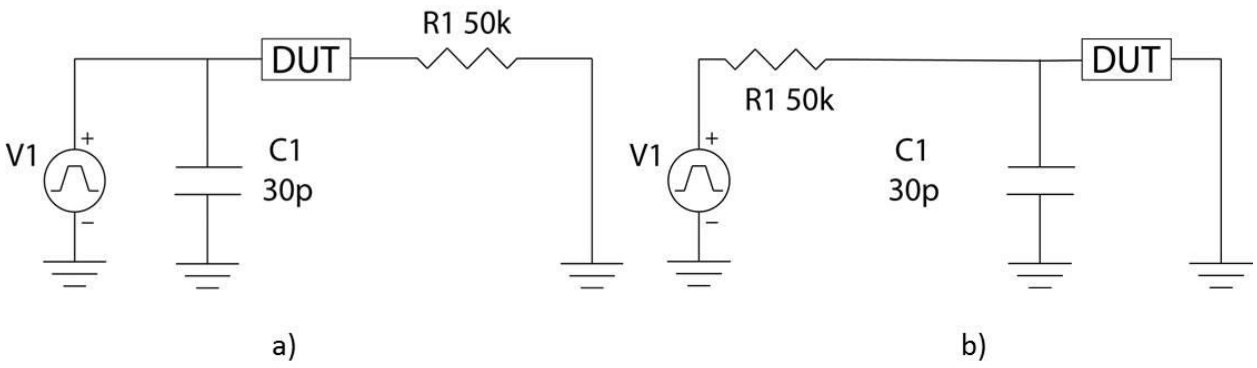


Figure 2-8: a) Case A: Equivalent circuit for a device based on poly-BE. b) Case B: Equivalent circuit for a device based on metal-BE and utilizing an external 50k series resistor.

First, let us consider the case of the poly-Si based devices (Fig. 2-9a, b). When the device-under-test (DUT) sees a high enough voltage; it begins to switch to a low resistance state. Even though the resistance of the DUT drops significantly, the presence of R1 (bottom electrode resistance due to poly-silicon) prevents the capacitor C1 from discharging through the DUT and suppresses any transient current during device switching. However, in the case of the metal-BE devices, as the DUT switches to a lower resistance (Fig. 2-9c), the voltage across C1 starts to drop and the capacitor discharges through the DUT causing significant transient effects. As shown in Fig. 2-9d, this momentary discharge can cause current spikes as high as tens of mAs even though the steady state DC current is never expected to go above 100μA(Fig. 2-9b) using an external current compliance. As a result, limiting the programming current using external resistors or current compliance is not efficient and the transient current causes the device to "over-program" and results in poor endurance.

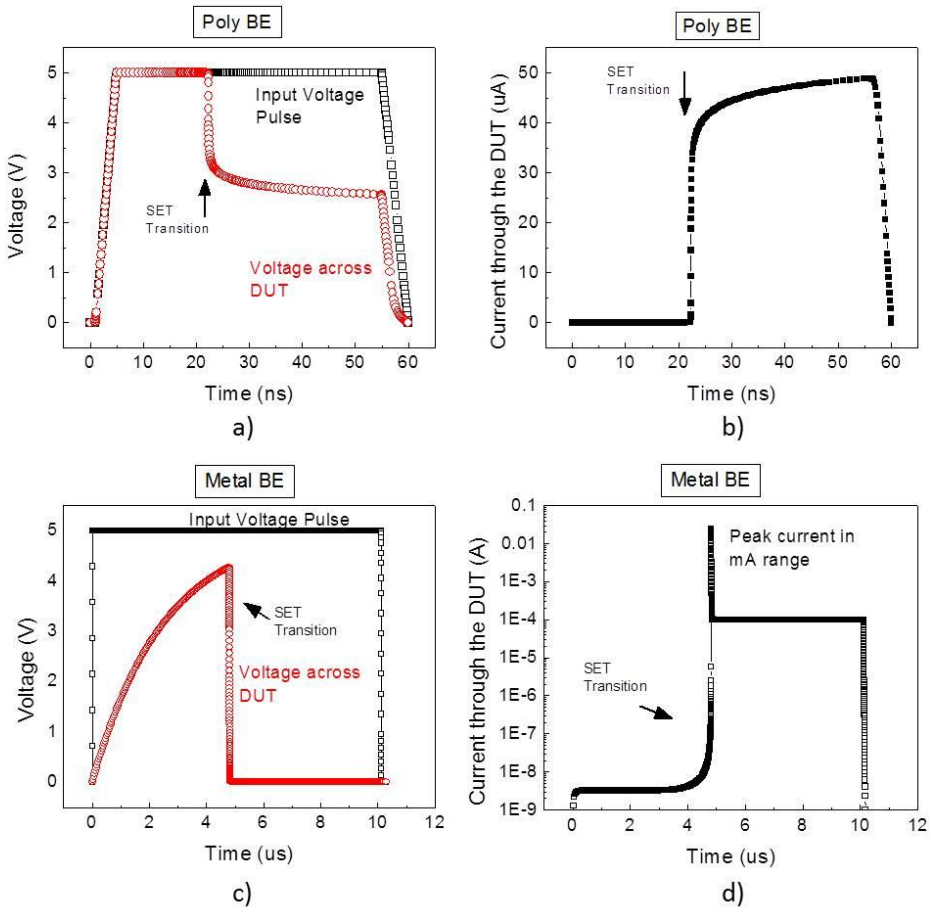


Figure 2-9: a) and b) Case A: The device switches in less than 20 ns as can be observed by monitoring the voltage across the DUT. The voltage across DUT settles at around half the voltage of the input pulse due to the series resistor effect. No sharp current transients are observed as the device switches. c) Case B: The device takes much longer to switch due to the RC delay. d) A sharp current discharge accompanies device switching.

In Case B, as the device switches, the accompanying current spike tends to over-program the device and drive it to a very low resistance state with a very robust and large filament (we typically observe Ron values of a few kilo-ohm for metal-BE devices). Erasing over-programmed cells is difficult and the associated erase tends to be thermal in nature. This has a limiting effect on the device endurance.

In strong contrast, in Case A, the poly-silicon devices tend to have higher R_{ON} values comparable to the on-chip series-resistance[9] due to feedback associated with voltage divider effect provided by the on-chip resistor during programming. The on-chip resistance limits the

filament growth and prevents the devices from shorting out. With finite R_{ON} values, the erase process is field based and not thermal in nature. This has a pronounced effect on the device endurance.

While these devices were tested with commercially available test equipment, the most optimized approach to test both single devices and arrays would be by using on-chip integrated decoders. This approach would have reduced RC delay and reduced parasitic capacitance issues.

As these MIS devices use high temperature poly-silicon, the bottom electrode process is not compatible with post-CMOS integration since the thermal budget is too high. To circumvent this problem, a low temperature SiGe process was developed to allow vertical integration of these MIS devices on CMOS IC with row and column decoders. This is described in the next section.

2.4 CMOS / RRAM Vertical Integration

2.4.1 Bottom Electrode Engineering

The MIS devices discussed in the previous section use high temperature poly-silicon and the bottom electrode process is not compatible with post-CMOS integration due to the high thermal budget. Various options were explored to achieve a low temperature CMOS compatible process.

Silicidation Approach

Metal silicides are used in commercial CMOS processes to contact the terminals of the transistors[10,11]. The silicide formation involves depositing a metal (usually Ti, Co or Ni) and a single or a sequence of moderate to high temperature anneals whereby the metal reacts with the underlying silicon forming MSi_x. The exact nature of the metal silicide formation and the kinetics are dependent on the metal used and the temperature regime and can often be predicted from phase diagrams. Finally the unreacted metal is removed by wet etching resulting in a self-aligned silicide layer.

Metal	1st RTP Temp. (°C)	Phase formed after 1st RTP	Wet selective etching	2nd RTP Temp. (°C)	Phase formed after 2nd RTP
Ti	650-730	C49 $TiSi_2$	H_2SO_4 / H_2O_2 + NH_4OH/H_2O_2	>850	C54 $TiSi_2$
Co	400-600	CoSi	H_2SO_4 / H_2O_2	>700	$CoSi_2$
Ni	300-350	Ni_2Si	H_2SO_4 /H_2O_2	400-550	NiSi

Table 2-1: Traditional metals used for silicide module in commercial CMOS devices.

While Ti and Co need high temperature anneals, Ni can be used at <450-500 °C (Table 2-1). However, test samples processed by us did not show any silicide formation. This was attributed to the native oxide which prevents reaction of the metal and the silicon below. The absence of multi-chamber systems, where wafers undergo Ar pre-sputter and metal deposition in one tool without ever breaking vacuum, makes the process window very small and results in large sample to sample variation.

Metal Induced Crystallization

Metal-induced crystallization (MIC) is a method by which amorphous silicon can be turned into polycrystalline silicon at relatively low temperatures[12–15]. In MIC, the a-Si film is capped with a metal, such as Al or Ni. The structure is then annealed, which causes the a-Si films to be transformed into polycrystalline silicon.

In a variant of this method, called metal-induced lateral crystallization (MILC), metal is only deposited on some area of the a-Si. Upon annealing, crystallization starts from the portion of

a-Si which is covered by metal and proceeds laterally[16–19]. Unlike MIC process, where metal contamination in the obtained poly-silicon is relatively high, the laterally crystallized silicon in MILC process contains very small amount of metal contamination. However, the crystallization speed is extremely slow (few um/hour was observed in test samples capped with evaporated Ni and annealed in a JetFirst RTP system in nitrogen ambient at 400 °C). Also, the metal removal process involved harsh acids and was deemed unfit for CMOS integration purposes as the I/O pads or routing in the upper metal levels would likely get damaged if strong acids / oxidizers were to be used.

Doped Silicon-Germanium SiGe has been a mainstay in the semiconductor industry for heterojunction bipolar transistors (HBTs) [20,21] and as a stressor in advanced CMOS technology nodes (45nm and below) [22]. While epitaxy is the process of choice for high mobility and low defect applications in CMOS, low temperature LPCVD deposited poly-SiGe is also a widely used MEMS structural layer [23–25]. Doped poly-SiGe was easily deposited in a First Nano EasyTube LPCVD system at the University of Michigan using a custom recipe at 425 °C using germane, silane, diborane and hydrogen (as a carrier/dilutant gas). The sheet resistance can be as low as 25×10^{-6} ohm·m.

Tests of standalone cells using SiGe as the bottom electrode show similar switching characteristics as the poly-Si BE devices (e.g. Fig. 2-5) and reasonable retention characteristics as shown in figure 2-10. Based on these studies, boron-doped poly-Si_xGe_{1-x} was identified as the bottom

electrode material for the integrated crossbar array. The ratio of Ge over Si was found to be 50% - 70% depending on SiH_4: GeH_4 ratio during SiGe film growth.

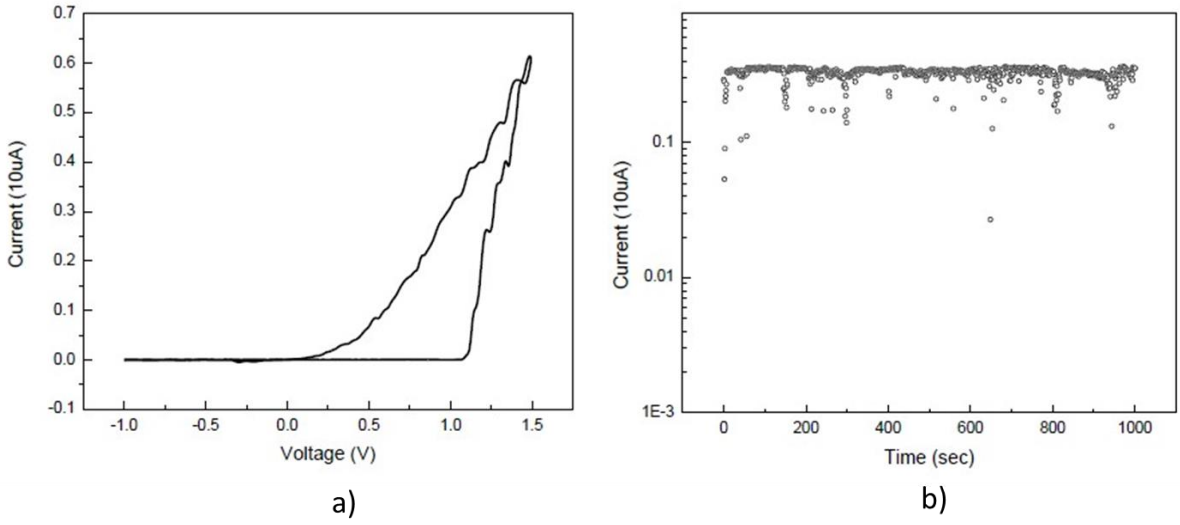


Figure 2-10: a) I-V switching curve from SiGe based RRAM showing intrinsically non-linear and rectifying I-V characteristics. b) Retention characteristics of the SiGe-based RRAM.

It is worth mentioning that the retention of the SiGe based devices was not as high as that for the poly-based devices and the SiGe based devices showed a more pronounced non-linearity and rectification. This can be attributed to a more weakly formed filament due to the different interface (SiGe vs. poly-Si). The topic of low operating current and retention is revisited in Chapter 4. Although, the SiGe based devices do not exhibit very high retention, they retain the state long enough for demonstration of array read/write operation and feasibility of the vertical integration concept.

Replacement of doped poly-silicon with low temperature metal nitrides TiN / TaN [26–29] is also being explored as discussed in Chapter 6.

2.4.2 CMOS Circuit Operation

The CMOS circuits needed for the vertical integration prototype demonstration were designed by our collaborators at HRL Laboratories. The chip has multiple decoders each controlled by 8 I/O signals. To access a particular bit the corresponding address code is input to the row decoder which also connects the DATA A signal to the target row through pass transistors. All unselected rows are connected to DATA B. A similar configuration is used for the column electrodes so when the correct address combination is input to the decoders the desired programming or read voltage (supplied to DATA A) can be applied across the selected bit, while all other unselected bits will be connected either with predefined protective voltages, ground, or left floating through DATA B. As a result, the integrated system allowed us to program 1600 cells (the 40×40 array) randomly with only 2 data inputs and 5 address inputs at each side, instead of having to supply 40×2 data inputs simultaneously for the case without CMOS decoder circuitry.

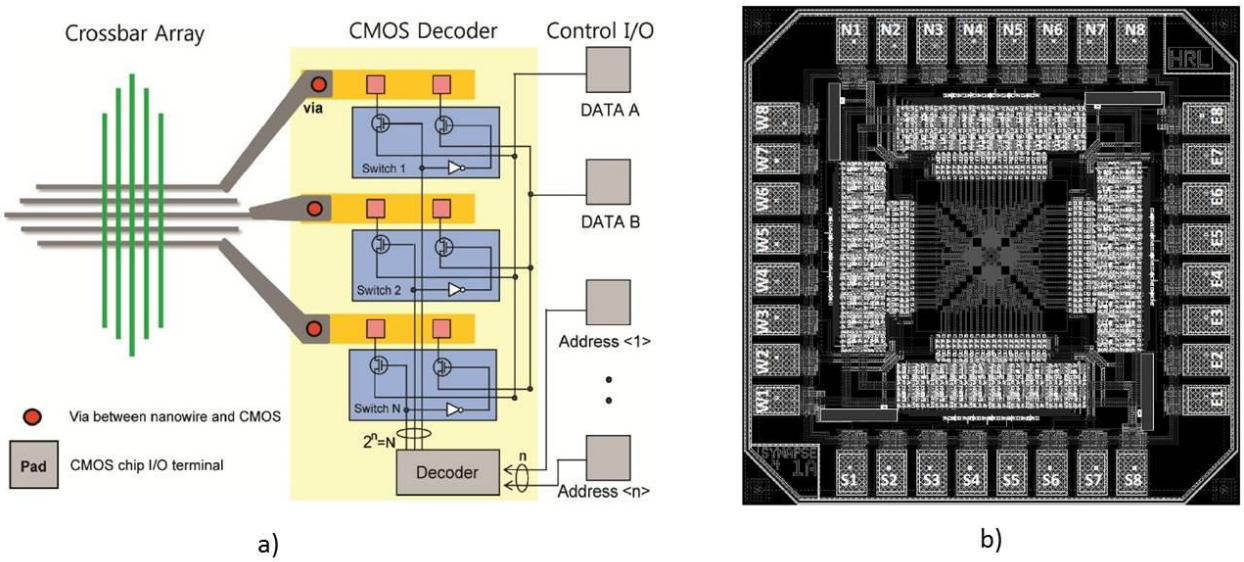


Figure 2-11: a) Schematic of the program/read schemes. Each column or row in the crossbar array is connected to one of the two external signal pads (DATA A for signal applied to the selected column/row, DATA B for signal connected to the unselected column/row) through CMOS decoder circuits controlled by address I/O pads. b) Die image of the CMOS decoder circuit.

This custom design was taped out using the IBM 7RF (0.18 micron) standard CMOS process. The wafers ran the standard front-end-of-line (FEOL) and middle-end-of-line (MEOL) processes but skipped all processes after M5 in the back-end-of-line (BEOL). Instead of the standard polyimide process, the wafers were protected with standard photoresist and shipped to University of Michigan for subsequent processing. The actual device integration is described in the next section.

2.4.3 Fabrication Process Flow

The eight inch wafers from HRL Laboratories were diced into small samples for easier handling using a dicing saw and the protective photoresist removed using standard solvents.

A blanket W film was sputter deposited followed by 15nm SiGe deposition in a LPCVD system. The switching matrix (a-Si) was deposited at 350°C using PECVD process. The stack comprising of the bottom electrodes and the switching matrix was patterned by using electron beam lithography (EBL) and a hard mask process in combination with standard reactive ion etching (RIE). After stripping the hard mask, sidewall spacers were formed by deposition and etch back of PECVD silicon oxide. The surface was planarized using spin on glass (SOG). The SOG was etched back to the expose the top of the a-Si surface. The top electrodes (Ag/Pd) were patterned using EBL and standard liftoff process. Table 2-1 summarizes all the fabrication steps. A scanning electron micrograph of the completed device is shown in Fig 2-4.

Seq.	Step	Description / Recipe / Tool	Comments
1	Dicing	ADT 7100 Dicing Saw	Front side wad protected using PR.
2	Strip PR	Remover PG (Microchem), 30 min.	
3	W Dep.	15nm, Enerjet Sputter, 7mT, 1A	DC Sputtering
4	a-Si Seed	GSI PECVD, Few nm	a-Si seed improves quality of the SiGe deposited in next step.
5	SiGe Dep.	400C, 15nm, FirstNano EZ3000 LPCVD, Silane/Germane/ Diborane	
6	EBL	PMMA A2, 50nm CD, 150nm Pitch	
7	Ni Dep.	Cooke electron beam evaporator	Hardmask
8	Ni Liftoff	Acetone, No Agitation	
9	Stack Etch	RIE LAM 9400 HDP	
10	Spacer Oxide Dep.	GSI PECVD, Few nm	
11	Spacer Etchback	LAM 9400 HDP RIE	Spacer oxide prevents W from reacting with SOG
12	SOG Coating	700B SOG, Filmtronics, ~300nm after cure.	
13	SOG Etchback	LAM 9400 HDP RIE	Inline SEM was used to determine end point.
14	EBL	Lithography to open W vias	
15	Via Etch	LAM 9400 HDP RIE	
16	EBL	Pattern top electrodes, 50nm CD, 150nm Pitch	
17	Ag/Pd Evaporation	Cooke electron beam evaporator	
18	Liftoff	Acetone, No Agitation	Agitation or heating is avoided due to poor adhesion of silver

Table 2-2: Fabrication process flow for the vertically integrated devices.

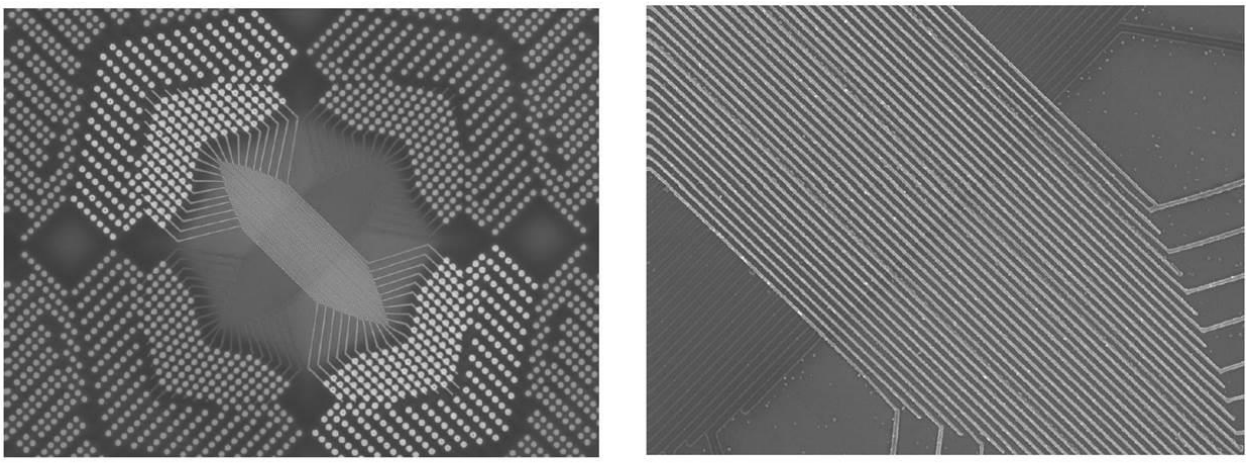


Figure 2-12: The complete device structure of the integrated crossbar array. (a) SEM image of the 40x40 crossbar array along with the CMOS vias. (b)High magnification image of the crossbar array. The density of the crossbar memory is 10 Gbits/cm2 with 100 nm pitch.

2.4.4 Electrical Data

2.4.4a DC Characteristics

Figure 2-13a shows the I-V switching characteristics of a single device after the vertical integration. Device fabrication did not affect functioning of the below CMOS as all decoders and pass transistors were verified using functional tests. All programming and read signals can be passed through the CMOS circuit to the crossbar array as designed.

In addition, Fig. 2-13b shows that very similar switching curves can be obtained from devices in the fabricated crossbar array with a narrow threshold voltage distribution. Tight distribution of the switching characteristics is a prerequisite to the operation of resistive memories at large scale to avoid accidental program/erase process when applying protective or read voltages.

Also noteworthy that the cells maintain an intrinsic current-rectifying behavior as shown in Fig. 2-13a, b such that the current at reverse bias is pronouncedly suppressed compared to the current at forward bias, consistent with earlier reports on similar stand-alone cells [5]. Even though

the current through the device is suppressed at relatively small reverse bias, the device still remains in the on-state and only become erased with large (e.g. < −1.5 V) negative voltages. The intrinsic current-rectifying characteristic is a key reason that the array studied here can operate without having an external transistor or diode at each crosspoint.

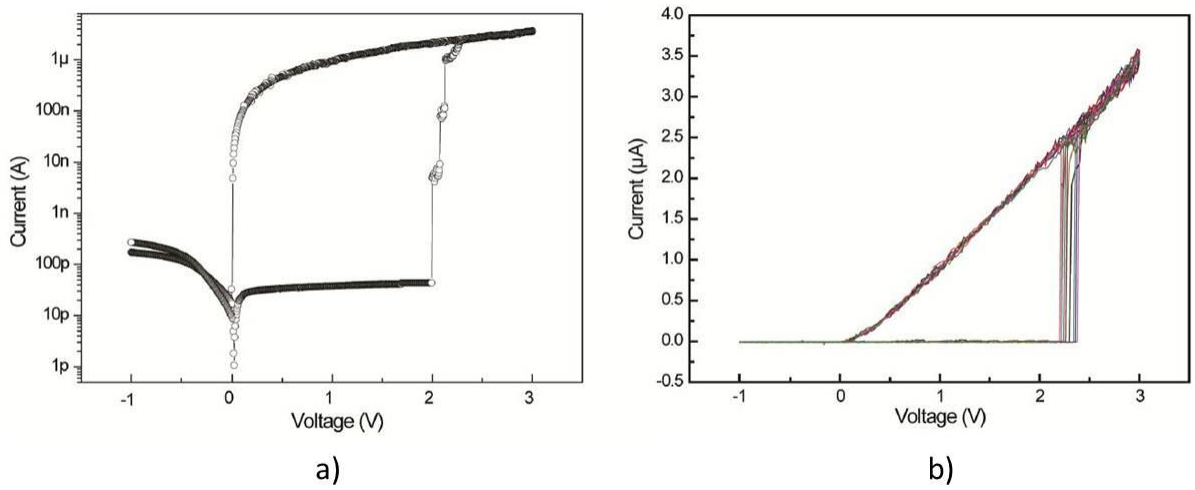


Figure 2-13: a) I-V switching curve from SiGe based RRAM showing intrinsically rectifying I-V characteristics. b) I-V switching characteristics from 10 different cells in the crossbar array.

2.4.4b Array Operation

To demonstrate the data storage capability of the integrated array, a binary bitmap image with 1600 pixels (40×40) that represents the University of Michigan logo was used (Fig. 2-14a). A dark pixel represents a "0" or OFF state of the device while a bright pixel represents a "1" or ON state of the device. The image was then programmed into the 40×40 integrated array and read out.

For writing '1', a 3.5 V, 100 µs pulse was applied across the selected cell through the CMOS decoder circuit using the protocol discussed above, while the other unselected electrodes were connected to a protective voltage with amplitude equaling half of the programming voltage to minimize disturbance of unselected cells. The same approach was also used for writing '0' into

a cell with a –1.75 V, 100 μs erase pulse. The programming/erasing was carried out based only on the input pattern and regardless of the current state of the memory cells, and a single programming/erase pulse was sufficient for a given cell. Once all data were programmed in an array, the information in the array was then read out one cell at a time by applying a 1 V, 500 μs read pulse across the target cell, while grounding all unselected electrodes through the CMOS decoder. To minimize cell wear out, the 40×40 array was divided into 25 8×8 sub-arrays and each sub-array was programmed as a whole followed by readout. The 40×40 pixel bitmap image was reconstructed by stitching results from the 25 8×8 sub-arrays together (Fig. 2-14b).

The resultant reconstructed image after reading the states of the devices is compared to the original image and shows good resemblance. The worst case ON and OFF cell states are separated by at least 50x (Fig. 2-15, upper panel). To further illustrate the functionality of the integrated crossbar array, a complementary image of the original logo was stored into the same array (Fig. 2-14c). Again, the reconstructed image closely resembles the "inverted" logo (Fig. 2-14d). Here too, the ON and OFF states are separated by at least 20x (Fig. 2-15, lower panel).

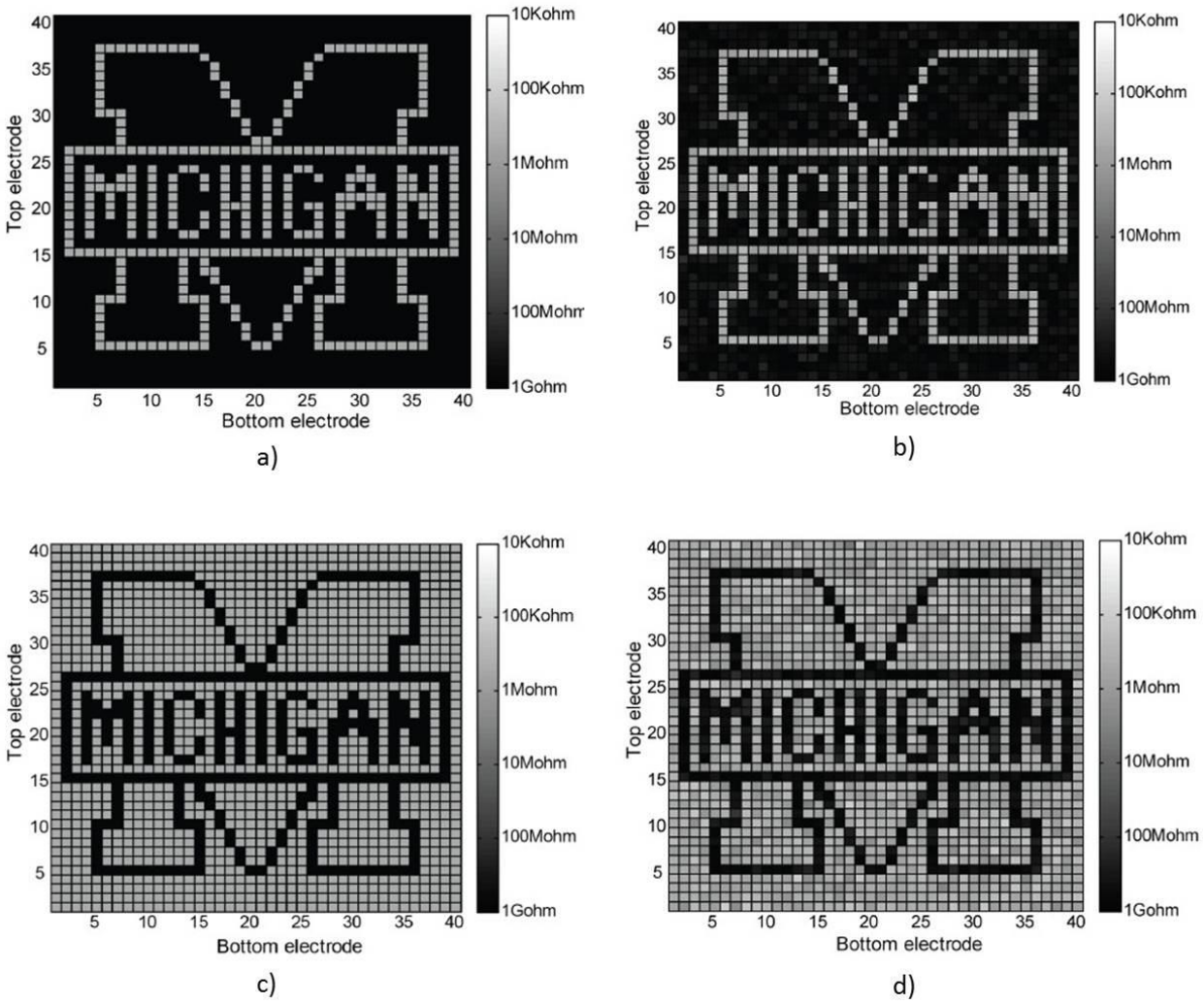


Figure 2-14: The original 40×40 bitmap image representing the UM logo with more number of 0s than 1s (a) and more 1s than 0s (c). The reconstructed bitmap images (b and d) after storing and retrieving data in the 40×40 crossbar array for each case above.

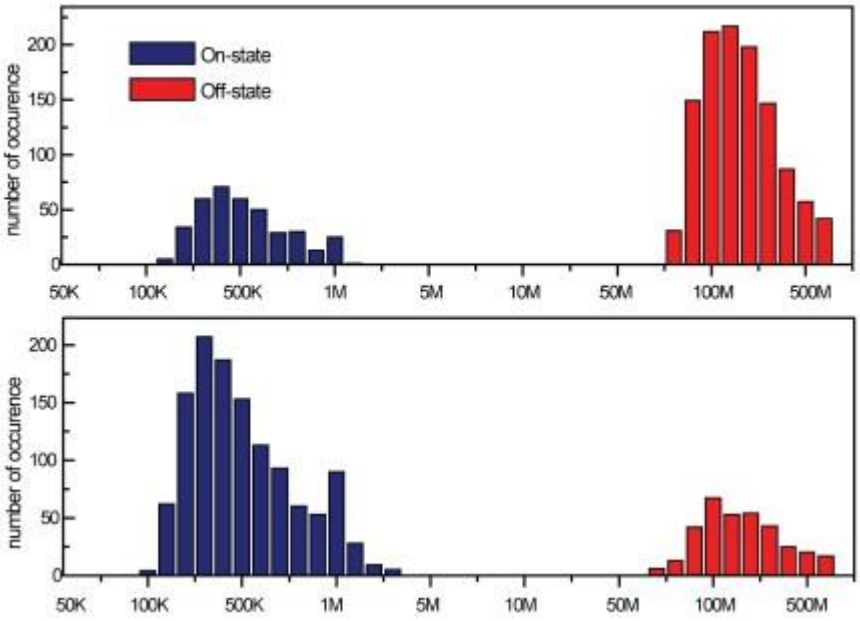


Figure 2-15: Histograms of the on- and off-state resistances for the data in Fig. 2-14b and Fig. 2-14d, respectively.

2.5 Conclusion

In conclusion, physical characterization of the resistive switching devices using transmission electron microscopy and scanning electron microscopy proved that the switching is caused by nanoscale filaments and the growth of these filaments is strongly affected by cation transport through the dielectric film[8]. The narrowest region of the filament was found to be near the dielectric/inert-electrode interface in these devices, suggesting that this region deserves particular attention for continued device optimization.

We have further demonstrated the need to include on-chip resistors for reliably programming RRAM cells. The on-chip resistor protects the actual device from high current discharge events and enhances the endurance of the device[30].

Finally, devices were integrated on top of a CMOS IC to show feasibility of vertical integration and hybrid CMOS / RRAM systems[31]. Integrated hybrid crossbar/CMOS systems has

been fabricated and successfully characterized. Crossbar arrays were programmed without incorporation of transistors or diodes at each cross-point by leveraging the intrinsic rectification of the devices. Binary bitmap images were successfully stored and retrieved with read margin > 20. This demonstration suggests that it is possible to build high density storage systems based purely on passive crossbar arrays using resistive switching devices. CMOS integration also opens pathway for novel integrated systems combining functionality of CMOS and high density and non-volatile memory provided by resistive switches.

References

[1] D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X.-S. Li, G.-S. Park, B. Lee, S. Han, M. Kim, and C. S. Hwang, "Atomic structure of conducting nanofilaments in TiO2 resistive switching memory.," *Nat. Nanotechnol.*, vol. 5, no. 2, pp. 148–53, Feb. 2010.

[2] Y. C. Yang, F. Pan, Q. Liu, M. Liu, and F. Zeng, "Fully room-temperature-fabricated nonvolatile resistive memory for ultrafast and high-density memory application.," *Nano Lett.*, vol. 9, no. 4, pp. 1636–43, Apr. 2009.

[3] T. Sakamoto, K. Lister, N. Banno, T. Hasegawa, K. Terabe, and M. Aono, "Electronic transport in Ta2O5 resistive switch," *Appl. Phys. Lett.*, vol. 91, no. 9, p. 092110, 2007.

[4] G.-S. Park, X.-S. Li, D.-C. Kim, R.-J. Jung, M.-J. Lee, and S. Seo, "Observation of electric-field induced Ni filament channels in polycrystalline NiOx film," *Appl. Phys. Lett.*, vol. 91, no. 22, p. 222103, 2007.

[5] K.-H. Kim, S. Hyun Jo, S. Gaba, and W. Lu, "Nanoscale resistive memory with intrinsic diode characteristics and long endurance," *Appl. Phys. Lett.*, vol. 96, no. 5, p. 053106, 2010.

[6] S. H. Jo, K.-H. Kim, and W. Lu, "Programmable resistance switching in nanoscale two-terminal devices.," *Nano Lett.*, vol. 9, no. 1, pp. 496–500, Jan. 2009.

[7] S. H. Jo and W. Lu, "CMOS compatible nanoscale nonvolatile resistance switching memory.," *Nano Lett.*, vol. 8, no. 2, pp. 392–7, Feb. 2008.

[8] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, "Observation of conducting filament growth in nanoscale resistive memories.," *Nat. Commun.*, vol. 3, p. 732, Jan. 2012.

[9] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, "Device and SPICE modeling of RRAM devices.," *Nanoscale*, vol. 3, no. 9, pp. 3833–40, Sep. 2011.

[10] C. M. Osburn, Q. F. Wang, M. Kellam, C. Canovai, P. L. Smith, G. E. McGuire, Z. G. Xiao, and G. a. Rozgonyi, "Incorporation of metal silicides and refractory metals in VLSI technology," *Appl. Surf. Sci.*, vol. 53, pp. 291–312, Nov. 1991.

[11] S.-L. Zhang and M. Östling, "Metal Silicides in CMOS Technology: Past, Present, and Future Trends," *Crit. Rev. Solid State Mater. Sci.*, vol. 28, no. 1, pp. 1–129, Nov. 2003.

[12] S. Joo, "Understanding of Metal-Induced Lateral Crystallization Mechanism-A Low Temperature Crystallization Phenomenon," *Electron. Mater. Lett.*, vol. 1, no. 1, pp. 7–10, 2005.

[13] Z. Jin, G. a. Bhat, M. Yeung, H. S. Kwok, and M. Wong, "Nickel induced crystallization of amorphous silicon thin films," *J. Appl. Phys.*, vol. 84, no. 1, p. 194, 1998.

[14] S. Y. Yoon, K. H. Kim, C. O. Kim, J. Y. Oh, and J. Jang, "Low temperature metal induced crystallization of amorphous silicon using a Ni solution," *J. Appl. Phys.*, vol. 82, no. 11, p. 5865, 1997.

[15] G. Radnoczi, a. Robertsson, H. T. G. Hentzell, S. F. Gong, and M. -a. Hasan, "Al induced crystallization of a-Si," *J. Appl. Phys.*, vol. 69, no. 9, p. 6394, 1991.

[16] J. Joo, "Low-temperature polysilicon deposition by ionized magnetron sputtering," *J. Vac. Sci. Technol. A Vacuum, Surfaces, Film.*, vol. 18, no. 4, p. 2006, 2000.

[17] Y. Yoon, M. Kim, G. Kim, and S. Joo, "Metal-induced lateral crystallization of a-Si thin films by Ni-Co alloys and the electrical properties of poly-Si TFTs," *IEEE Electron Device Lett.*, vol. 24, no. 10, pp. 649–651, Oct. 2003.

[18] M. Miyasaka, K. Makihira, T. Asano, E. Polychroniadis, and J. Stoemenos, "In situ observation of nickel metal-induced lateral crystallization of amorphous silicon thin films," *Appl. Phys. Lett.*, vol. 80, no. 6, p. 944, 2002.

[19] I. Hong, T. Hsu, S. Yen, F. Lin, M. Huang, and C. Chen, "Nickel Induced Lateral Crystallization of Amorphous Silicon Thin Film Studied by SPESM," pp. 270–272.

[20] I. Z. Mitrovic, O. Buiu, S. Hall, D. M. Bagnall, and P. Ashburn, "Review of SiGe HBTs on SOI," *Solid. State. Electron.*, vol. 49, no. 9, pp. 1556–1567, Sep. 2005.

[21] J. D. Cressler and S. Member, "SiGe HBT Technology : A New Contender for Si-Based RF and Microwave Circuit Applications," vol. 46, no. 5, pp. 572–589, 1998.

[22] M. L. Lee, E. a. Fitzgerald, M. T. Bulsara, M. T. Currie, and A. Lochtefeld, “Strained Si, SiGe, and Ge channels for high-mobility metal-oxide-semiconductor field-effect transistors,” *J. Appl. Phys.*, vol. 97, no. 1, p. 011101, 2005.

[23] A. Witvrouw, M. Gromova, and A. Mehta, “Poly-SiGe, a superb material for MEMS,” *MRS Proc.*, vol. 782, pp. A2.1.1–12, 2004.

[24] S. Sedky, A. Witvrouw, and K. Baert, “Poly SiGe, a promising material for MEMS monolithic integration with the driving electronics,” *Sensors Actuators A Phys.*, vol. 97–98, pp. 503–511, Apr. 2002.

[25] A. Witvrouw, R. Van Hoof, and G. Bryce, “(Invited) SiGe MEMS Technology: A Platform Technology Enabling Different Demonstrators,” *ECS Trans.*, vol. 33, no. 6, pp. 799–812, 2010.

[26] N. D. Cuong, D.-J. Kim, B.-D. Kang, and S.-G. Yoon, “Effects of Nitrogen Concentration on Structural and Electrical Properties of Titanium Nitride for Thin-Film Resistor Applications,” *Electrochem. Solid-State Lett.*, vol. 9, no. 9, p. G279, 2006.

[27] N. D. Cuong, D.-J. Kim, B.-D. Kang, C. S. Kim, K.-M. Yu, and S.-G. Yoon, “Characterization of Tantalum Nitride Thin Films Deposited on SiO[sub 2]/Si Substrates Using dc Magnetron Sputtering for Thin Film Resistors,” *J. Electrochem. Soc.*, vol. 153, no. 2, p. G164, 2006.

[28] T. Riekkinen, J. Molarius, and T. Laurila, “Reactive sputter deposition and properties of Ta x N thin films,” *Microelectron. Eng.*, vol. 64, pp. 289–297, 2002.

[29] A. Malmros, M. Südow, K. Andersson, and N. Rorsman, “TiN thin film resistors for monolithic microwave integrated circuits,” *J. Vac. Sci. Technol. B Microelectron. Nanom. Struct.*, vol. 28, no. 5, p. 912, 2010.

[30] S. Gaba, S. Choi, P. Sheridan, T. Chang, Y. Yang, and W. Lu, “Improvement of RRAM Device Performance Through On-Chip Resistors,” *MRS Proc.*, vol. 1430, pp. mrss12–1430–e09–09, Jun. 2012.

[31] K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, “A functional hybrid memristor crossbar-array/CMS system for data storage and neuromorphic applications.,” *Nano Lett.*, vol. 12, no. 1, pp. 389–95, Jan. 2012.

Chapter 3
Stochastic RRAM Devices for Computing Applications

3.1 Introduction

In the previous chapter we discussed that resistance switching is associated with the formation and rupture of a single dominant nanoscale filament. Filament formation involves oxidation, ion transport and reduction processes, all of which are thermodynamically driven [1, 2] and require overcoming specific activation energies. Typically one of the processes is rate-limiting so that switching is associated with thermal activation over a dominant energy barrier and is thus probabilistic in nature, if a dominant filament is involved[3].

From a device perspective this probabilistic nature manifests itself as variations in switching parameters. Switching parameters like threshold voltage, on-current etc. show cycle to cycle variation even for a single device. It is important to distinguish this "temporal" variation from spatial variation. While spatial variations (i.e. device-to-device variations) are related to line edge roughness and film thickness irregularity and can be corrected to some extent through process control and variation aware design, temporal variation in switching parameters is tied to the intrinsic probabilistic nature of the resistance switching.

To confirm this probabilistic nature of switching from an electrical measurement perspective, we fabricated and tested semiconductor / insulator / metal (poly-Si / a-Si / Ag) crossbar devices. Devices based on cation (Ag) migration inside a solid state electrolyte (a-Si) are used in this study, but the results can be generalized to other cation[1] or anion[4] (e.g. oxygen vacancy in oxides) based RRAM systems as well.

In this chapter, the stochastic nature of device switching has been confirmed. Even though individual switching events cannot be predicted, the overall probability of switching is predictable. This unique predictability of device switching can be used in stochastic computing architectures.

3.2 Device Fabrication

Experiments were conducted on stand-alone two-terminal cross-point devices (Fig. 3.1(a) (b)). We started the device fabrication by depositing 60nm boron-doped poly-silicon (sheet resistance ~ 1000 ohm / square) on a Si / SiO_2 substrate using a low pressure chemical vapor deposition (LPCVD) furnace. Poly-silicon bottom electrodes were patterned by using electron beam lithography and reactive ion etching (RIE) methods. Photolithography and liftoff were used to pattern a thin etch stop layer (ESL) on top of the poly-silicon bottom electrode pads. Next, a 35nm insulating film of amorphous silicon was deposited in a plasma enhanced chemical vapor deposition system. Silver top electrodes (capped with palladium) were patterned using electron beam lithography and conventional liftoff method. Subsequently, bottom electrode contacts were opened using photo-lithography and RIE. The ESL prevents over-etch into the poly-silicon during the a-Si etch. Finally, large gold contact pads were formed to ensure good electrical connectivity and to provide a suitable surface for probing / wire bonding. Detailed process flow is given in Table 3-1.

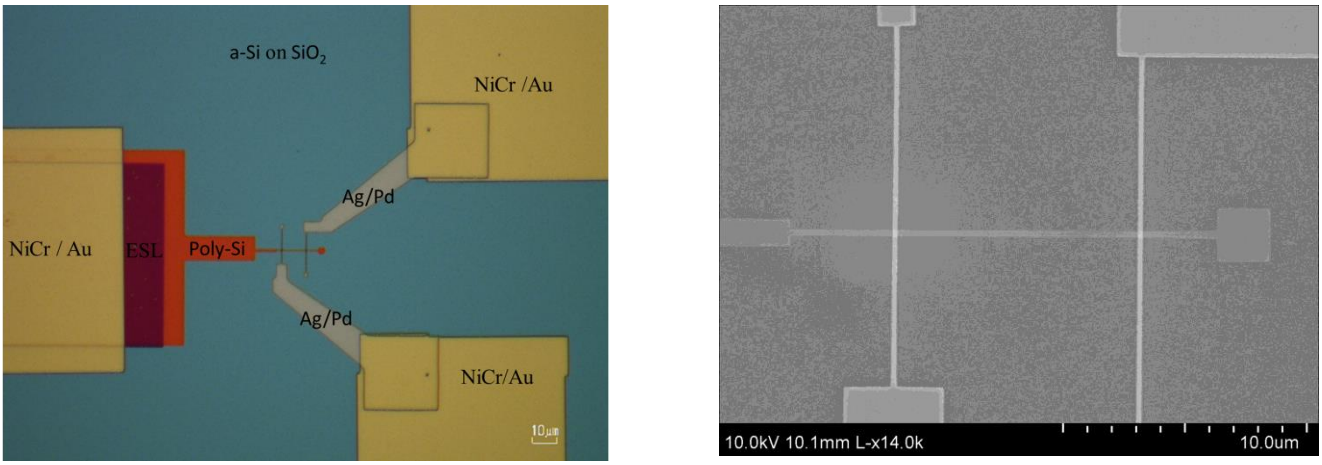


Figure 3-1: (a) An optical micrograph of the fabricated device. A resistive switch is formed at each location where the Ag top electrode crosses over the poly-Si bottom electrode. (b) Scanning electron micrograph of the crosspoint device structure.

Module	**Layer**	**Sub-Step**	**Comment**
Poly-Si Deposition		p-doped poly-Si deposition	LPCVD using SiH_4 / BH_3, 60nm.
	1	Photolithography	
		Metal evaporation	Adhesion Layer (NiCr, Ti or Cr) + Au
		Liftoff in acetone	
	2	Electron-beam lithography	200nm minimum feature size
		Ni evaporation	40nm Ni is used as a hard mask.
Poly-Si BE		Liftoff in acetone	
		poly-Si etch	SF_6 / C_4F_8 RIE chemistry
		Ni strip	1:1 HCl : DI Water
	3	Photolithography	
Etch Stop Layer (ESL)		Metal evaporation	
		Liftoff in acetone	
a-Si Deposition		Pre-clean	Dilute HF dip to remove native oxide
		a-Si depostion	380 °C PECVD, 35nm
	4	Electron-beam lithography	200nm minimum feature size
Ag Top Electrodes		Ag/Pd evaporation	40nm Ag capped with 20nm Pd
		Liftoff in acetone	
	5	Photolithography	
Bottom Electrode Pads		RIE	Remove a-Si on top of bottom electrode pads
		Strip photoresist mask	Dry resist strip + Wet clean
	6	Photolithography	
Pads		Metal evaporation	Adhesion Layer (NiCr, Ti or Cr) + Au
		Liftoff in acetone	

Table 3-1: Fabrication flow used to fabricate cross-point devices based on the Ag / a-Si / poly-Si sandwich structure.

3.3 Stochastic Nature of Devices

The fabricated devices were electrically probed in a Lakeshore PS-100 Tabletop Cryogenic Probe Station. Electrical data was collected using a National Instruments data acquisition system (NI USB-6259 BNC) in conjunction with a DL 1211 current preamplifier from DL Industries. Custom code written in MATLAB was used to generate and collect the signals.

Unless otherwise mentioned, electrical bias was applied to the top electrode while the bottom electrode was grounded. All measurements were carried out at room temperature at atmospheric pressure. Binary resistive switching was reliably observed in the devices, as shown in Fig. 3-2. The device exhibited standard switching hysteresis with clearly defined ON and OFF states and an apparent threshold voltage ~5V. The change from OFF to ON state is explained by the formation of a dominant filament as discussed in Ch. 2 and in previous work [5,6].

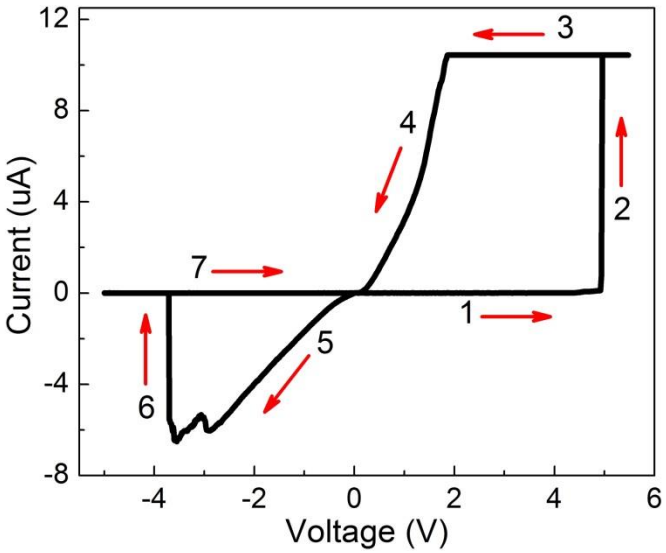


Figure 3-2 DC switching curve of the fabricated device.

To study the inherent temporal variation switching behavior, electrical bias lower than the threshold voltage was applied to slow down the switching process and allow accurate recording of the switching time. For example, in Fig. 3-3, the device was initially reset to the OFF state and a constant voltage of 3.5V was applied to the device at t = 0. The current through the device was then monitored continuously. After an initial wait time, the sharp increase in current marked the transition of the device to the ON state. This wait time measured the time elapsed between the application of the voltage and the switching of the device[7].

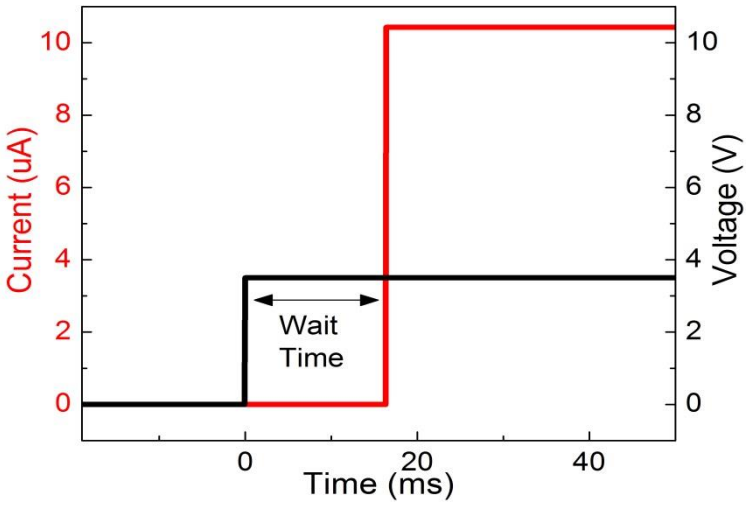


Figure 3-3 Example of a wait time measurement.

Once the device was verified to be in the ON state, the voltage bias was turned off and the device was reset to the OFF state by applying a negative voltage pulse. This process was then repeated one hundred times at each bias condition – 2.5V, 3V, 3.5V, 4V and 4.5V – to analyze the temporal variations in the switching behavior.

The wait time before switching shows apparent stochastic behavior. For example, Fig. 3-4a shows the histogram for the wait times associated with an applied voltage of 2.5 V. As can be seen, even for a given voltage applied to the same device, the wait time is not fixed but rather

shows a large distribution. Even for one device, the wait time is broadly distributed and in principle can only be predicted in terms of statistics while the individual switching events occur randomly.

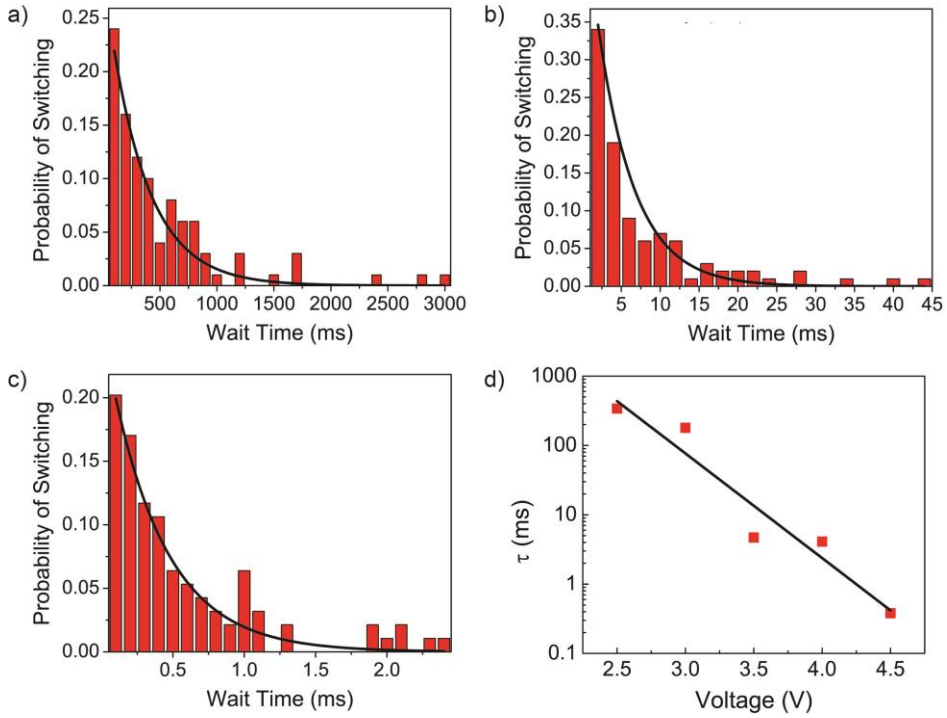


Figure 3-4 : Stochastic wait time distribution. (a–c) Distributions of wait times applied voltages of 2.5 V (a), 3.5 V (b) and 4.5 V (c). Solid lines: fitting to the Poisson distribution of eqn. (1) using s as the only fitting parameter. τ= 340 ms, 4.7 ms and 0.38 ms for (a)–(c), respectively. (d) Dependence of s on the programming voltage. Solid squares were obtained from fitting of the wait time distributions while the solid line is an exponential fit.

Mathematically, if only one dominant energy barrier limits the switching process, the wait time is expected to follow a Poisson distribution and the probability that a switching event occurs within Δt at a given time t is given by

$$P(t) = \frac{\Delta t}{\tau}\exp(-\frac{t}{\tau}) \qquad \text{(Equation 3-1)}$$

where τ is the characteristic wait time[3]. Fig. 3-4 shows that excellent fit of the wait times to the Poisson distribution in equation 3-1 can be obtained with just one fitting parameter (τ), in agreement with the hypothesis of thermal activation over a dominant energy barrier during

filament formation. The Poisson distribution of the wait time, with the standard deviation equaling the mean, further verifies that the switching is random and stochastic in nature.

Further, the characteristic wait time τ is strongly voltage dependent. As can be seen from Fig. 3-4, the wait time distribution at different programming voltages preserves the Poisson nature but the characteristic wait time decreases significantly as the bias voltage is increased. With an increase in 2 V in the applied voltage, the characteristic time drops exponentially by almost three orders of magnitude (Fig. 3-4d). The strong voltage dependence of (average) switching time is expected within the filament formation picture since the energy barriers for both the oxidation and the ion transport processes are field-dependent (the reduction processes of the ions are not thought to be the rate-limiting process)[1,2]and the effective barrier height is reduced upon the application of the bias voltage[3,8,9].

Additionally, the probability of switching is related only to the voltage amplitude and the total time the voltage is applied, i.e. the switching probability is cumulative. This provides an analog component (e.g. the cumulative time during which the programming voltage is applied) to these binary devices if they are used as synapses in neuromorphic systems[10]. For example, the time variable may be discretized such that the probability of switching can in turn be written as a function of the number of (short) pulses of a fixed voltage, regardless of the gap between the pulses. This has the advantage of allowing a look-up table to be used to determine the number of pulses needed to achieve a given probability, rather than a more complex timing circuitry.

As a proof of concept, we apply a string of pulses, each of which having 2.5 V amplitude and 100 ms duration, and count the number of pulses it takes for the device to switch. As expected, the distribution for the number of pulses needed for switching the device also follows a Poisson

distribution and can be fitted with just one fitting parameter τ. The time constant τ obtained by counting the number of discrete pulses agrees well with the time constant obtained from applying continuous voltage biases (Fig. 3-5a). Significantly, almost identical τ values were obtained when increasing the gap between the applied pulses from 10 ms (Fig. 3-5a) to 500 ms (Fig. 3-5b) while keeping the pulse width constant, proving the cumulative effect of the programming pulses on switching.

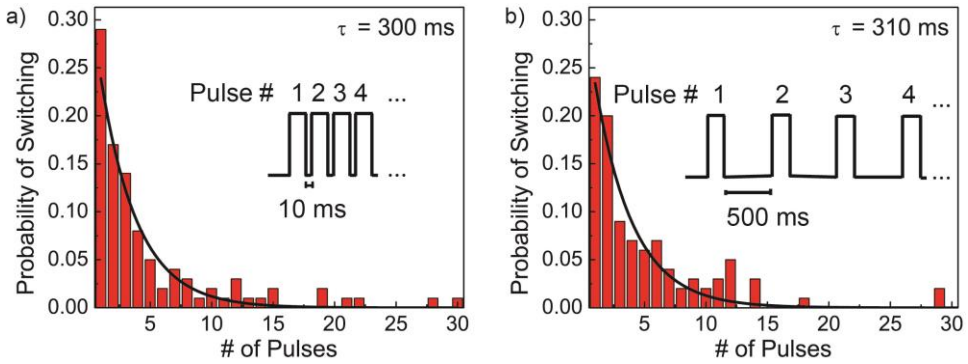


Figure 3-5: Switching probability when subjected to a series of short pulses The average switching time can be calculated by measuring the number of pulses, regardless of the gap between the pulses. The pulse width was kept constant at 100 ms while the gap was changed from 10 ms (a) to 500 ms (b). The voltage amplitude was fixed at 2.5 V for all pulses.

3.4 Random But Predictable

First, we note that the switching probability can be calculated by integrating the Poisson distribution in equation 3-1, which leads to

$$C(t) = 1 - \exp(-\frac{t}{\tau}) \qquad \text{(Equation 3-2)}$$

For an applied voltage of 2.5 V, the prediction based on equation 3-2 is shown in Fig. 3-6a as the solid line. The switching probabilities can also be obtained by calculating the cumulative probability distribution function from data in Fig. 3-4a, shown as the squares in Fig. 3-6a. Again good agreements can be obtained, in agreement with the model.

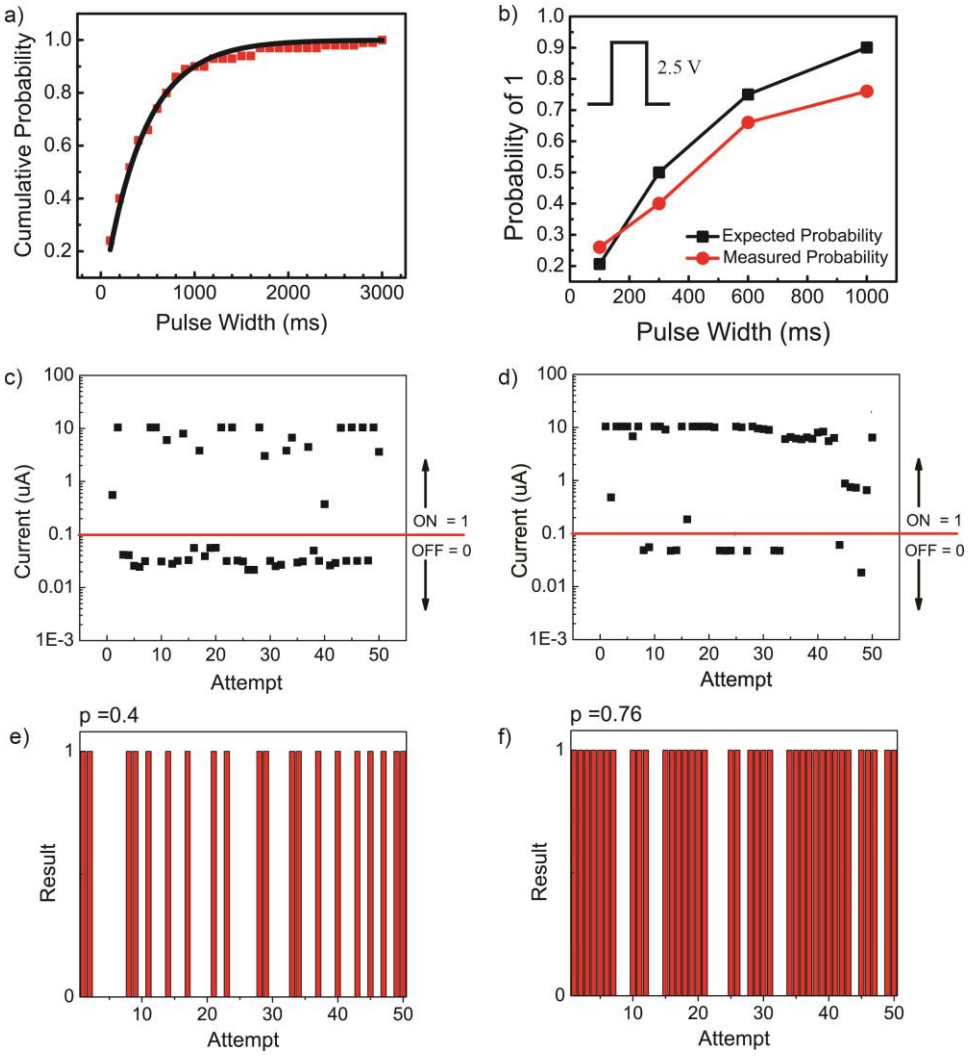


Figure 3-6 : Probability of switching within a single pulse. (a) Pulse width dependence. The solid line shows values predicted from equation 3-2, squares were obtained from measuring the cumulative probability obtained from Fig. 3-5a. (b) Expected and measured probability using a single 2.5 V pulse. (c–d) Device current measured after repeated application of a single 2.5 V, 300 ms (c) and 1000 ms (d) pulse. The device was reset after each measurement. (e–f) Corresponding bitstreams of (e) p= 0.4 and (f) p= 0.76 corresponding to (c) and (d).

The significance of equation 3-2 is that we can now predict the switching probability at a given programming voltage and pulse width, even though each switching event is random. We verify these predictions by applying a programming pulse at fixed amplitude (e.g. 2.5 V) and pulse width (e.g. 300 ms) and measuring whether the device was switched to the ON state during the pulse or not. The device currents measured after the application of the programming pulse are shown in Fig. 3-6c for fifty such attempts. In this experiment, after each programming pulse the

device was reset to the same original OFF state so that each programming pulse starts with the same initial condition. The device switches to the ON state twenty times out of the fifty trials or attempts. This number closely matches that predicted from the theoretical curve given by equation 3-2. Obviously, longer pulses result in a higher probability of the device switching during the programming pulse, as predicted by equation 3-2. Results for 2.5 V/1000 ms pulses are shown in Fig. 3-6d; the device switches thirty eight times compared to twenty times in Fig. 3-6c. Results from other pulse widths are shown in Fig. 3-6b.

This ability to predict the probability of switching in stochastic memristive switching events can be used for novel computing schemes. In one such example, the measured current can be digitized – for example, each value above 0.1 μA can be assigned as "1" while each value below 0.1 μA can be assigned as "0". Thus, here we essentially obtained a stream of bits (a bitstream) which is fifty bits long in time and having 20 1s randomly distributed in the bitstream (Fig. 3-6e), corresponding to a bias ratio, p = 0.4. Here the bias ratio of the bitstream is defined as the ratio of the number of 1s to the total stream length. The bias ratio in this random bitstream can be controlled by the pulse width and the voltage used. For example, the bias ratio changes to p= 0.76 if the pulse width is increased to 1000 ms (Fig. 3-6f). It is also important to note here that we can control only the bias ratio of the bitstream (i.e. the overall number of 1s), while the locations of individual 1s are random. The randomness is an essential requirement for applications such as bitstream based stochastic computing, where the correlation between the locations of 1s can lead to systematic computing errors[11]. Use of this "predictable randomness" property of these devices is further discussed in the next section.

3.5 Application: Stochastic Computing

Stochastic computing was first proposed in the 1960s[12,13] but like many other concepts proposed ahead of their time it never got commercial attention due to the exponential growth of digital computers. In stochastic computing, analog values are represented as probabilities in bitstreams. For example, a bitstream B containing 25 percent 1s and 75 percent 0s denotes the number b= 0.25. The length or the structure of B need not be fixed. For example {1,0,0,0}, {0,0,1,0} and {0,0,0,0,1,0,1,0} all are possible representations of b= 0.25. This value depends only on the ratio of the total number of 1s to the total length of the bitstream and not on the value of a particular position in the bitstream.

The stochastic bit stream representation is error resilient since all the bits in the bit stream are LSBs. Therefore, a few bit flips in any given stochastic number will not significantly alter any given result. This stochastic representation has an inherent advantage over the binary radix representation when it comes to noise tolerance, since a bit flip only causes an error of 1/n where n is the length of the bitstream regardless of where the error occurs, while in the binary radix representation a bit flip can cause an error of 2^{n-1} if it occurs at the most significant bit[14].

Another advantage of stochastic computing is that relatively simple circuits can be used for tasks which otherwise are computationally more intensive in binary radix representation. Since the order of the bits in a stochastic bit stream has no meaning, unlike binary encoded numbers, bit level parallelization becomes straightforward.

For example, the multiplication of two numbers x and y can be achieved by simply performing a bit-wise AND operation on the two bitstreams representing x and y[14] (Fig. 3-7). To understand

how this works, consider bit streams *a* and *b* with corresponding probabilities P_a and P_b. The probability of a bit at a given position in both stream *a* and *b* are both 1 is equal to $P_a \times P_b$.

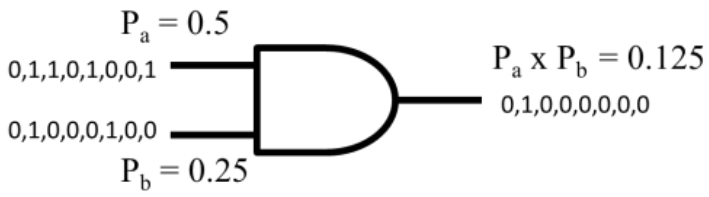


Figure 3-7 : Stochastic multiplication using a logic AND gate.

This above operation assumes that x and y are independent bitstreams that are completely uncorrelated. Any correlation degrades the accuracy of stochastic computing. For example, if we multiply two identical bitstreams using the AND gate, the product will be x, instead of x^2. The independence assumption requires the bitstreams to be randomized. The randomization is done by stochastic number generators (SNGs)[15] as shown in Fig. 3-8. Adding SNGs in systems significantly increases overheads, sometimes as high as 80% of the total resource usage[16]. Further, SNGs need to be inserted along multiple intermediate stages to mitigate correlations introduced by reconvergent signals. The extensive deployment of SNGs easily surpasses core logic as the dominant cost in stochastic computing. This added overhead of SNGs has been a major roadblock in the widespread adoption of stochastic computing.

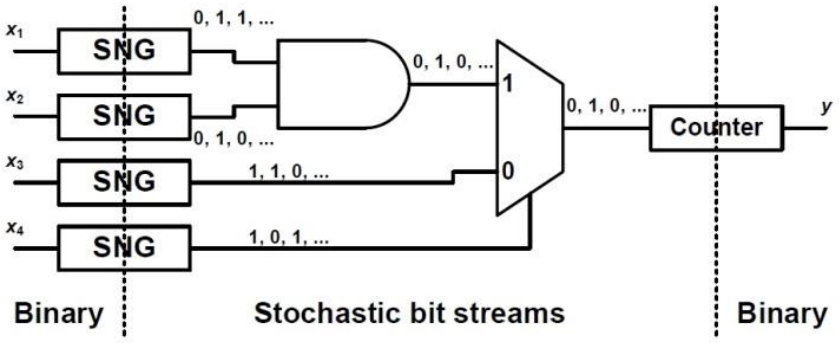


Figure 3-8: Stochastic implementation of logic function y= x1x2x3 + x3 (1- x4)

Implementing SNG in CMOS is expensive because CMOS inherently lacks true randomness, and the randomness needed by SNGs must be created with psuedo-random number generators like linear-feedback shift registers (LFSRs).

However, given that RRAMs have inherent randomness that can be captured very easily without complicated circuitry; RRAMs are a potentially ideal implementation of SNGs because of their low cost and high integration density. The problem of generating independent random bit streams is solved by simply programming multiple memristive devices with identical voltage pulses.

Using RRAMs to produce uncorrelated bit streams with nearly identical biases has been experimentally shown in this section. Here four different devices were used to obtain twenty bit long bit streams (Fig. 3-9). While the bias for a given pulse height and width can be predicted from equation 3-2, the exact position of the 1s and 0s in the bitstream is completely random. Thus, the devices are used to produce streams of bits which are uncorrelated but the total number of 1s to the string length in each case is almost the same (0.6 in this case). In other words, four different independent bitstreams representing the number 0.6 have been generated without resorting to expensive SNGs.

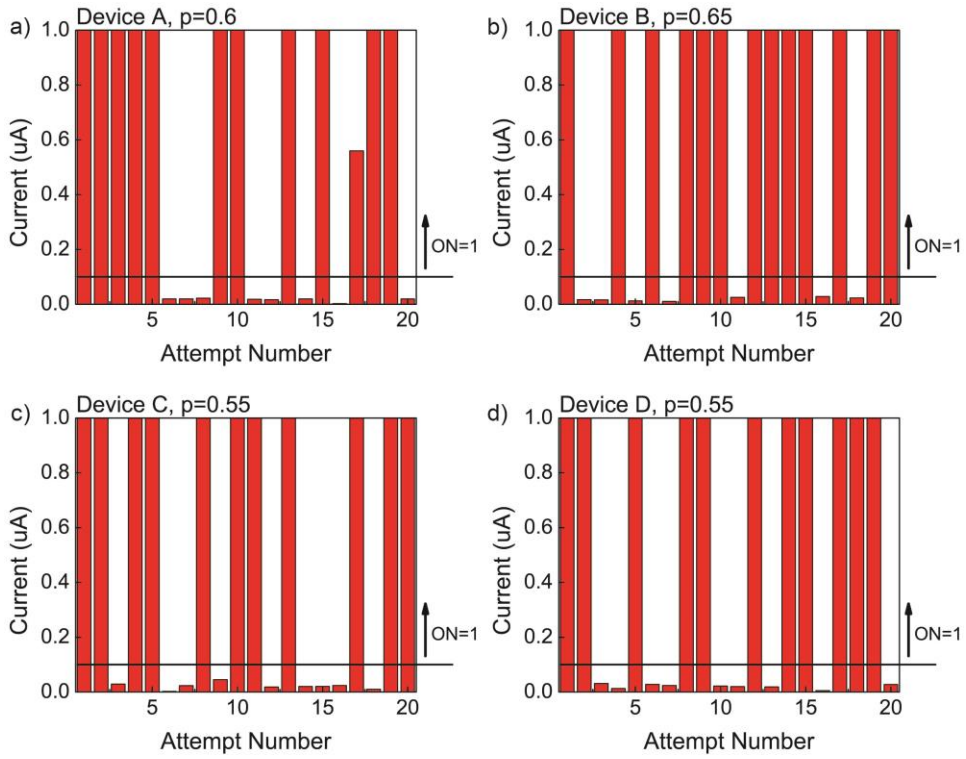


Figure 3-9 : Stochastic switching in different devices (a–d) obtained from a second fabrication run. The devices were measured with twenty 2.75 V, 1000 ms pulses.

In Fig. 3-9 we see that each bit in the bitstream was generated at a different time instance. In other words, the random bits were produced in serial fashion one after the other. This is commonly termed as "stochastic bitstream in time". For more efficient computing, bitstreams in the space domain, i.e. bits generated in parallel or "stochastic bitstreams in space" are also required, which leads to parallel processing of the bitstreams.

Stochastic bitstreams in space can also be obtained by using memristive devices in parallel in an array form. Bits produced in different physical locations (cells) were generated by programming an array of 4 devices (A, B, C, and D) with a single pulse in parallel as shown in

Fig. 3-10 Here the state of each device, 1 or 0, represents the bits of the bitstream in space and was read out individually after the programming pulse.

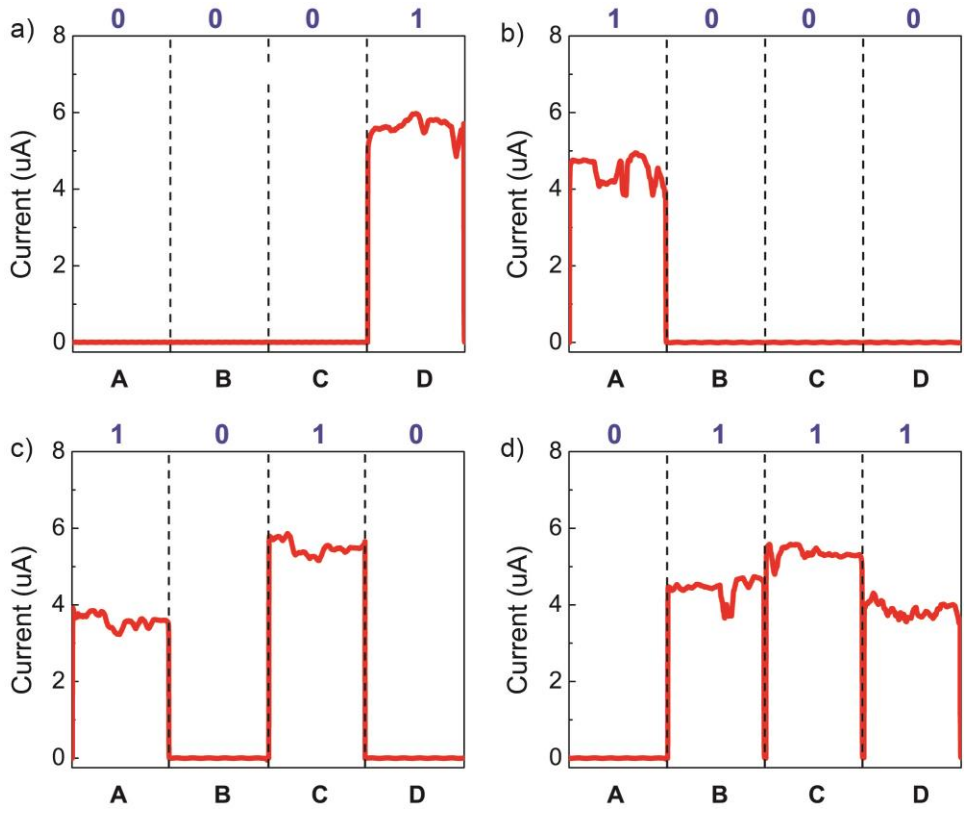


Figure 3-10 : Stochastic programming of a device array. 4 representative combinations {0001} (a), {1000} (b), {1010} (c), and {0111} (d) are shown here out of a total of $2^4 = 16$ combinations. The different combinations were obtained in the same array under identical pulses. The devices were programmed in parallel using a single 2.75 V, 1000 ms pulse and their states were measured using 2 V, 100 ms pulses individually after the programming pulse. The array was reset after each measurement. The measurement setup is further discussed in Appendix 3.

A question can then be raised as to why the device array did not behave as a single, larger device simply having four times the area. In the latter case, since only a dominant filament exists we should observe only one device in the ON state in the 4-device array instead. The answer can be obtained after a careful examination of the filament formation process. During programming, after the first filament bridges the electrodes the voltage drop across the device is subsequently

reduced, an effect caused by the voltage divider formed with the external circuit components (i.e. series resistance)[3,7]. The reduction of bias voltage in turn slows the growth of additional filaments and results in a single, dominant filament. Indeed, if the 4 devices share the same series resistor, we observed only one device in the ON state after each programming pulse, i.e. the device array simply behaves as a single, larger device. However, if each device has its own local, series resistor, as shown in the inset of Fig. 3-11, the completion of the dominant filament in one device does not cause the decrease of voltage seen by other devices; thus multiple devices can be switched.

A lot of scientific research recently has been focused on using stochastic computing methodologies in areas like image processing[16] and artificial neural networks[17]. All these applications fit the stochastic operating paradigm where errors can be tolerated. Further, in applications where the required accuracy changes with time, stochastic processors allow trading off accuracy for energy consumption by simply using longer or shorter bit streams. By using a longer bit stream, accuracy of the system increases at the cost of increased time and energy required to produce these bit streams. Compared to SNGs in binary systems, RRAMs allow more efficient production of bit streams and thus are suitable for integration into stochastic systems for current and future compute-intensive but error tolerant probabilistic applications.

3.6 Conclusion

In conclusion, we show that binary resistive switching devices (RRAMs) can exhibit the native stochastic nature of resistive switching. Even for a fixed voltage on the same device, the wait time is not fixed and is random and broadly distributed. However, the probability of switching can be predicted and controlled by the applied voltage and the pulse width used to program the device.

The RRAM devices have been used to generate random bitstreams with predicable bias ratios in time and space domains. The ability to produce random bitstreams using binary memristive devices based on the native stochastic switching principle may potentially lead to novel non-von-Neumann, alternative computing paradigms.

References

[1] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories—fundamentals, applications, prospects," *Nanotechnology*, vol. 22, no. 25, p. 254003, Jul. 2011.

[2] W. Lu, D. S. Jeong, M. Kozicki, and R. Waser, "Electrochemical metallization cells—blending nanoionics into nanoelectronics?," *MRS Bull.*, vol. 37, no. 02, pp. 124–130, Feb. 2012.

[3] S. H. Jo, K.-H. Kim, and W. Lu, "Programmable resistance switching in nanoscale two-terminal devices.," *Nano Lett.*, vol. 9, no. 1, pp. 496–500, Jan. 2009.

[4] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing.," *Nat. Nanotechnol.*, vol. 8, no. 1, pp. 13–24, Jan. 2013.

[5] S. H. Jo and W. Lu, "CMOS compatible nanoscale nonvolatile resistance switching memory.," *Nano Lett.*, vol. 8, no. 2, pp. 392–7, Feb. 2008.

[6] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, "Observation of conducting filament growth in nanoscale resistive memories.," *Nat. Commun.*, vol. 3, p. 732, Jan. 2012.

[7] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, "Device and SPICE modeling of RRAM devices.," *Nanoscale*, vol. 3, no. 9, pp. 3833–40, Sep. 2011.

[8] H. Schroeder, V. V. Zhirnov, R. K. Cavin, and R. Waser, "Voltage-time dilemma of pure electronic mechanisms in resistive switching memory cells," *J. Appl. Phys.*, vol. 107, no. 5, p. 054517, 2010.

[9] D. B. Strukov and R. S. Williams, "Exponential ionic drift: fast switching and low volatility of thin-film memristors," *Appl. Phys. A*, vol. 94, no. 3, pp. 515–519, Nov. 2008.

[10] K. Likharev, A. Mayr, I. Muckra, and Ö. Türel, "CrossNets: High-Performance Neuromorphic Architectures for CMOL Circuits," *Ann. N. Y. Acad. Sci.*, vol. 1006, no. 1, pp. 146–163, Dec. 2003.

[11] A. Alaghi and J. P. Hayes, “Survey of Stochastic Computing,” *ACM Trans. Embed. Comput. Syst.*, vol. 12, no. 2s, pp. 1–19, May 2013.

[12] B. Gaines, “Stochastic computing systems,” *Adv. Inf. Syst. Sci.*, pp. 37–172, 1969.

[13] W. J. Poppelbaum, C. Afuso, and J. W. Esch, “Stochastic computing elements and systems,” *Proc. Novemb. 14-16, 1967, fall Jt. Comput. Conf. - AFIPS ’67*, p. 635, 1967.

[14] W. Qian, J. Backes, and M. Riedel, “The synthesis of stochastic circuits for nanoscale computation,” *Int. J. Nanotechnol. Mol. Comput.*, vol. 1, no. 4, pp. 39–57, 2009.

[15] X. Li, W. Qian, M. D. Riedel, K. Bazargan, and D. J. Lilja, “A reconfigurable stochastic architecture for highly reliable computing,” *Proc. 19th ACM Gt. Lakes Symp. VLSI - GLSVLSI ’09*, p. 315, 2009.

[16] W. Qian, S. Member, and X. Li, “An Architecture for Fault-Tolerant Computation with Stochastic Logic,” *Comput. IEEE Trans.*, vol. 60, no. 1, pp. 93–105, 2011.

[17] B. D. Brown and H. C. Card, “Stochastic neural computation. I. Computational elements,” *IEEE Trans. Comput.*, vol. 50, no. 9, pp. 891–905, 2001.

Appendix 3: Experimental Setup

For automatic testing of multiple cells in parallel, a custom circuit was built using analog switches (ADG1412 from Analog Devices) and a microcontroller (Atmega32u4). The switches were closed during the application of the programming pulse so the 4 devices can be programmed in parallel. During read the switches were sequentially closed so the device states can be read out individually.

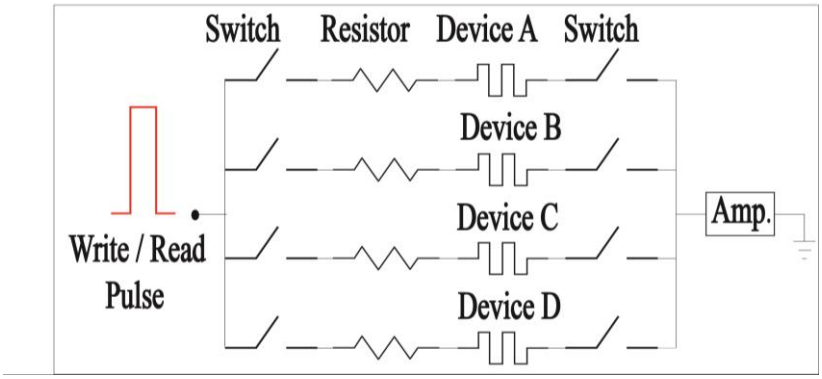


Figure 3-11 Experimental setup used to generate bitstreams that are stochastic in space.

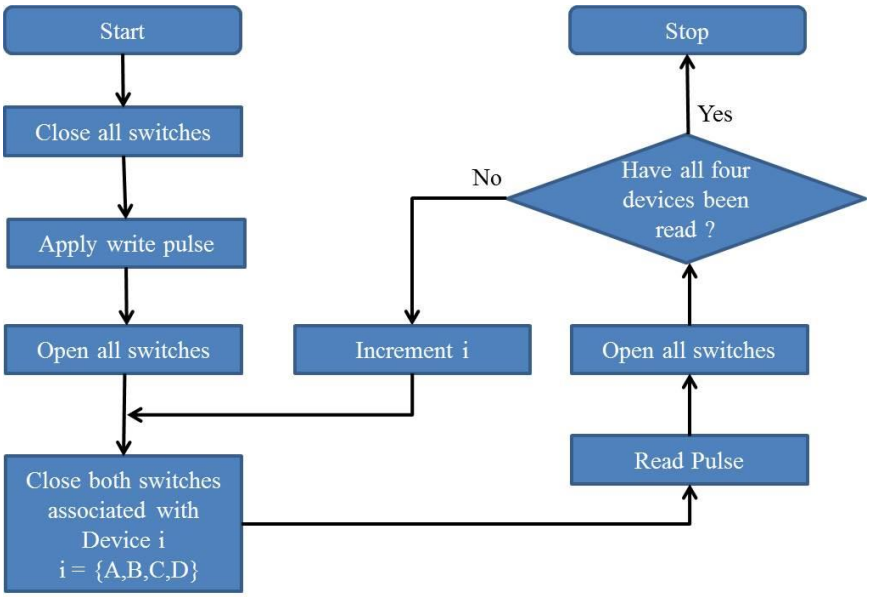


Table 3-2: Flowchart representing the parallel write and serial read of four devices. The analog switches are controlled by a microcontroller which is programmed using MATLAB before initiating the measurement routine.

Chapter 4

Ultralow Sub-1nA Operating Current Resistive Memory

4.1 Introduction

Resistive switching devices are a prime candidate for ultra-high density non-volatile memory. High speed operation [1], superior scalability [2] and feasibility of 3D integration (Chapter 5) have been demonstrated. However, most reported RRAM devices require high programming currents, which complicates the array design and the driving circuitry. Even with advanced CMOS generations with high drive current capability (>1000μA/μm), transistor sizes need to be increased well beyond the minimum sizes allowed by lithography to be able to drive sufficient current through these (high current) resistive switches. . Lowering RRAM operating current also opens up the path for hybrid systems using novel devices with relatively low drive currents - carbon nanotube transistors[3–5], graphene transistors[6], nanowire transistors[7,8] etc.

Low operating current is also beneficial for large array operation. In order to investigate the influence of the operating current, a simple structure with one word line was studied (Fig. 4-1).

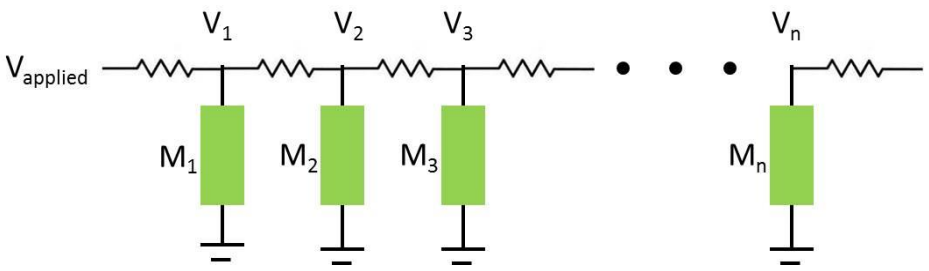


Figure 4-1: Voltage is dropped due to the line resistance. All cells are assumed in high resistance state.

All cells sharing the bit line are assumed to be undergoing the writing process. The number of cells, sharing the same word line, for which the voltage loss is within 10% of the programming supply voltage is calculated as a function of the operating current and the line resistance (Fig. 4-2).

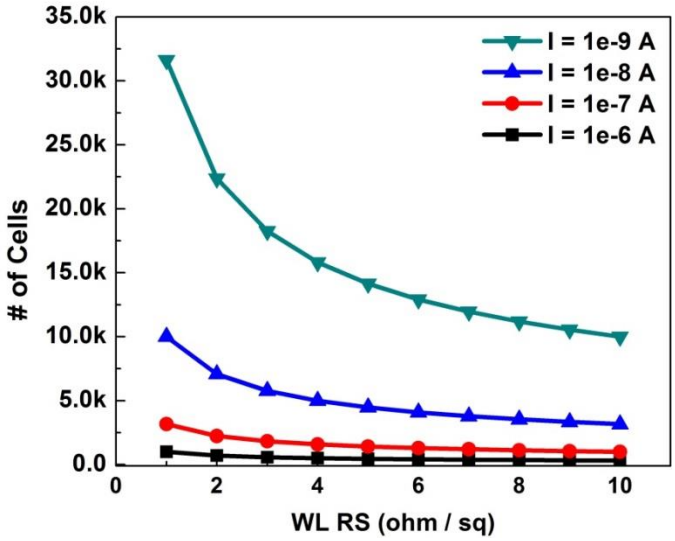


Figure 4-2: Number of resistive switching cells along a word line as a function of the operating current and line resistance.

As expected, the number of cells for which the voltage loss is within 10% drops sharply from >30000 (1nA operating current) to less than 100 (1µA operating current) for sheet resistance of 1 ohm/square. Thus to achieve high density, it becomes essential to decrease the operating current.

While there is intense interest in reducing the operating current in resistive memory devices, the inherent tradeoff of programming current vs. retention makes such scaling non-trivial[9]. If low current is used to program a device, the resultant filament is very weakly formed and tends to dissolve away relatively fast resulting in volatile memory or nonvolatile memory with very short retention times[9]. Scaling of operating currents through material design and development has seen considerable progress in the last few years. Many material systems are being studied to find reliable low current non-volatile resistive switches. Since most films involved are very thin (<10-20nm) with interface defects playing a major role in the formation and stabilization of the filament, theoretical modeling of such systems is still at a nascent stage and most published data is experimental in nature. Sub-μA operating currents have been demonstrated recently (Table 4-1). Sub-nA current operation along with appreciably good retention has been shown for the first time.

Max Current (nA)	Device Structure	Critical Dimension(nm)	Retention (sec)	Endurance (# of cycles)	Reference
1000	TiN/ HfOx / TaOx / Ta / TiN	3000	1e4 @ 85C	>1000 (DC)	[10]
1000	TiON/ WOx / W / TiN	9	1e6	<1000(DC)	[11]
250	Ag / a-Si / Poly-Si	100	5e4 @ 100C	>1e8	[12]
50	Al / N-AlOx / Al	1000	1e6 @ 125C	>1e5	[13]
0.5	Cu / Al_2O_3 / Poly-Si	2000	2.1e4@ 85C	>10000	This work
0.01	Cu / SiO_2 /Ir (or Pt)	>100 000	No Data	No Data	[14]

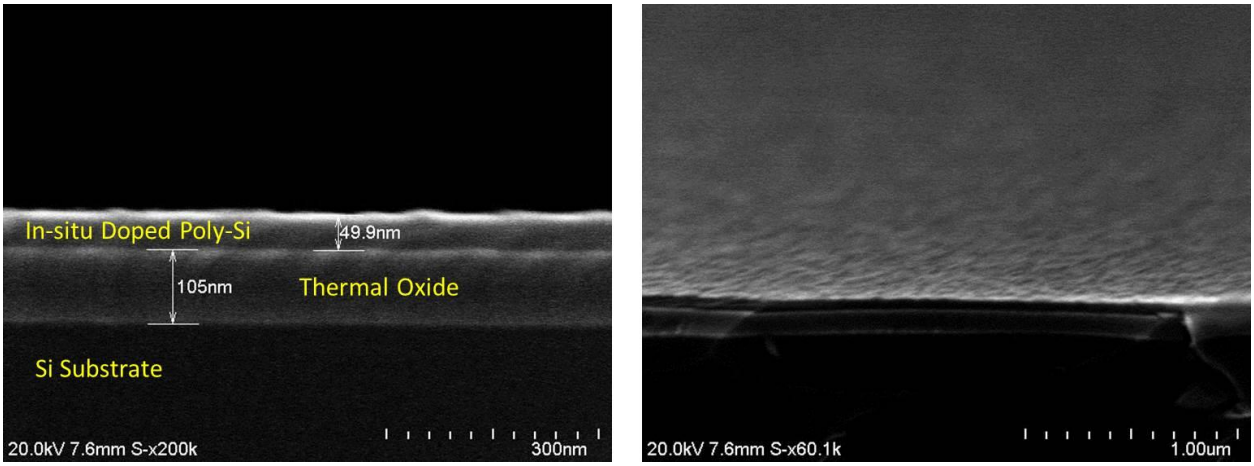
Table 4-1: Comparison of the Cu/Al_2O_3/Poly-Si devices with other recent low current systems.

4.2 Device Structure / Fabrication

Experiments were conducted on two-terminal cross-point devices with a structure similar to the devices described in Chapter 3.

Device fabrication begins with a blanket deposition of 50 nm boron-doped polysilicon (with a sheet resistance of 1000 ohm per square) on a Si/SiO_2 substrate with 100 nm SiO_2 using a Tempress TS 6604 low pressure chemical vapor deposition furnace (Standard SC-1 and SC-2

cleans were performed just before loading the wafers into the furnace to prevent contamination issues). Poly-silicon was deposited at 570 °C at a pressure of 360mT using silane (78 sccm) and boron trichloride (12 sccm). Test wafers were cleaved and analyzed in a scanning electron microscope to determine the deposition rate, actual thickness (Fig. 4-1a) and surface roughness (Fig4-1b).



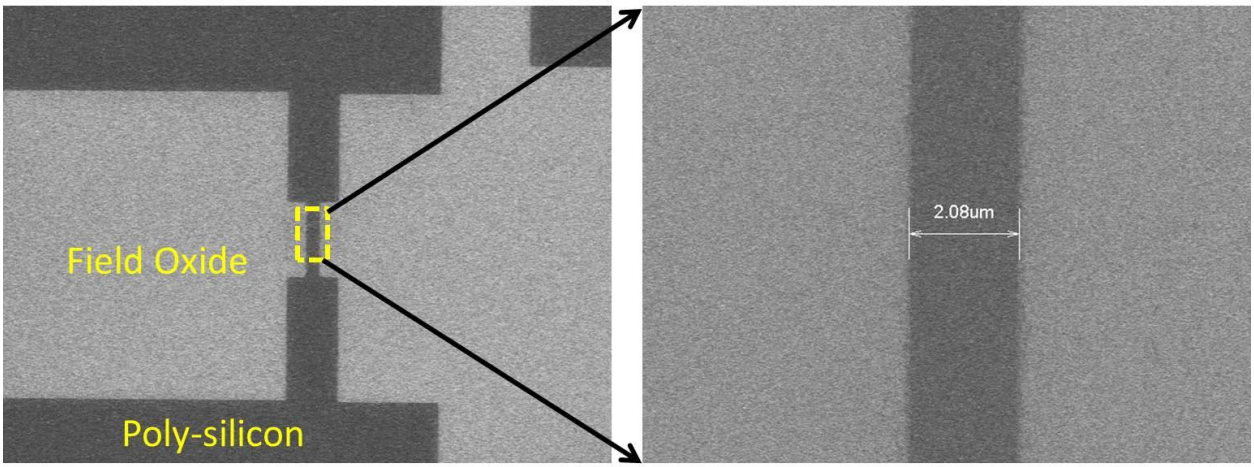
a) b)

Figure 4-3: a) X-SEM of the deposited poly-silicon film. 9 minutes deposition time gives a film thickness ~50nm. b) Tilt view (45 degrees) of the poly-silicon film confirms relatively flat surface.

Next, standard photolithography and liftoff techniques were used to pattern alignment marks on the samples. An exclusive step for alignment marks can often be skipped when using metal bottom electrodes (for example in the tungsten devices described in Chapter 6) and the lithography alignment marks can be patterned along with the bottom electrodes since even a thin metal layer provides enough contrast for later alignment steps. However, 50nm polysilicon on top of thermal oxide provides very little contrast. Thus an extra mask for alignment marks was added in the process flow to minimize yield loss due to poor alignment.

Polysilicon bottom electrodes are then patterned by photolithography and reactive ion etching (RIE). A CF_4/O_2 chemistry (40 sccm CF_4, 2 sccm O_2, 80W, 100mT, Plasmatherm-790) was

carefully optimized to minimize sidewall roughness. End point was confirmed by multiple collaborative techniques - measuring the field oxide using an ellipsometer (confirms all poly-silicon on top of oxide has been removed), a multi-meter (an open circuit confirms all the doped poly has been etched away) and by SEM (confirms no etch residue and clean sidewalls, Fig. 4-2ab). The resist mask is stripped using oxygen plasma and standard solvents. Extended oxygen plasma cleans are avoided to minimize oxidation of polysilicon.



a) b)

Figure 4-4: a) Scanning electron micrograph of the etched poly-silicon bottom electrodes. b) A 2um as-designed line becomes slightly larger in dimension due to lithography/etch process bias.

Next, the samples are dipped in a dilute HF solution (1:20 HF: DI) to remove native oxide on the poly-silicon. The HF dip is timed (30sec), to remove around 100Å of silicon dioxide (calibrated on a blanket SiO_2 film). This is found to be sufficient to remove the native oxide, which tends to self-saturate below 50Å[14,15], and prevent excessive loss of field oxide and undercut of the poly-silicon bottom electrodes.

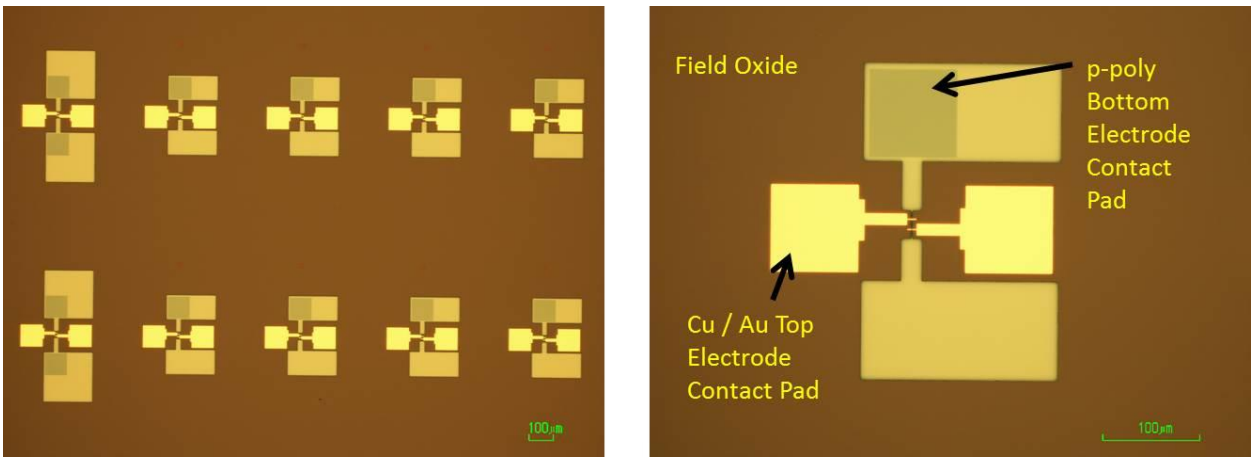
Immediately after the dilute-HF dip, the samples are placed in an atomic layer deposition reactor (Oxford OpAL) and 10nm Al_2O_3 is deposited using standard precursors (tri-methyl-

aluminum and water[16,17] at 150 °C. The chamber was preconditioned before loading the samples to prevent flaking from the chamber walls.

Top electrodes are fabricated using photo-lithography and electron-beam evaporation / liftoff techniques. The copper electrodes (800Å) are capped with gold (400Å) to prevent accidental corrosion of copper.

Finally, the Al_2O_3 on the bottom electrode contact pads is removed using photolithography and a timed wet etch (dilute-HF, 1:20 HF: DI, 45 seconds). For wire bonding, gold contact pads can be deposited using electron beam-evaporation and liftoff. An optical micrograph of the completed devices is shown in Fig. 4-3.

The devices were measured in a Lakeshore Cryogenic Probe Station using a Keithley 4200 Semiconductor Characterization System (SCS). The 4200-SCS was found to be better for providing very low current compliance as compared to the 1211 amplifier used for measuring the Ag-based devices (Chapter 3) and the WO_X based devices (Chapter 6). For all measurements, bias was applied to the Cu top electrode while the polysilicon bottom electrode was grounded.



a) b)

Figure 4-4: a) Optical micrograph of the devices after opening the pads for the bottom electrodes. The first device in each row has two contacts for the bottom electrode and serves as a test structure to allow measurement of line resistance. b) Higher magnification view of each device. Each poly-silicon bottom electrode is shared between two copper top electrodes.

4.3 Electrical Results

4.3.1 Low Current Operation

By using a metal/insulator/ semiconductor structure (instead of a metal/insulator/metal) in conjunction with very low current compliance, sub-nA current operation is demonstrated. Writing, erasing and reading are all accomplished within 0.5nA (Fig. 4-6). The sharp increase (decrease) in the current - the vertical jumps in the log I-V curve (Fig. 4-6b) - corresponds to the formation (rupture) of the filament during set (reset).

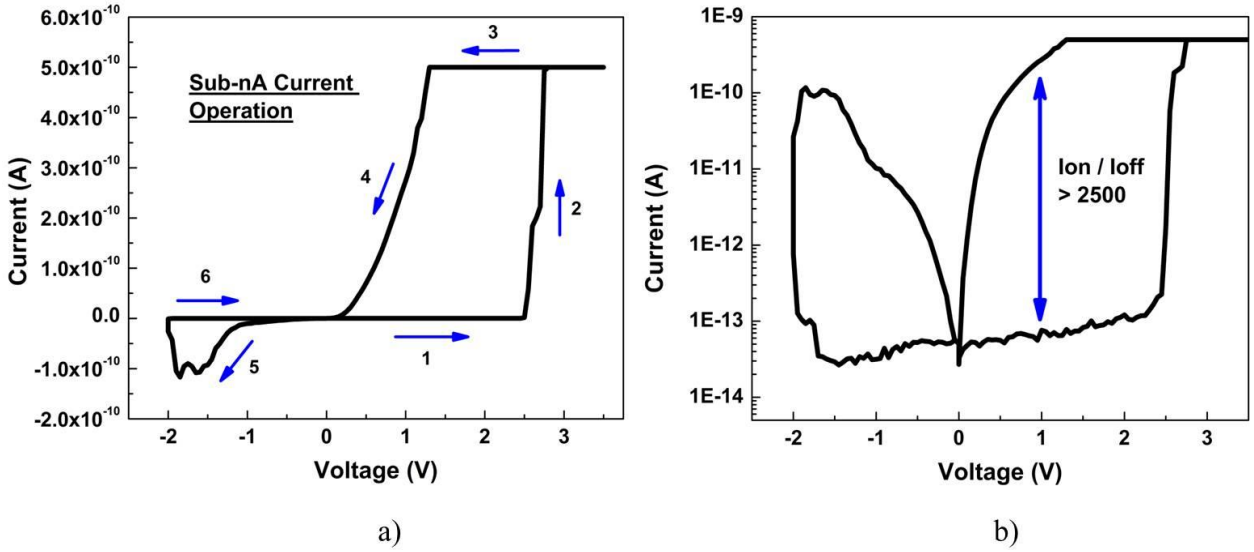


Figure 4-6: Linear (a) and log (b) scale I-V curve showing sub-1nA current operation. Despite the low current, large (> 2500) Ion-Ioff ratio is obtained.

Particularly, the polysilicon bottom electrode serves as an in-cell resistor to effectively limit potential voltage overshoot during filament growth to prevent the formation of thick filaments (Fig. 4-7), as discussed earlier in Section 2.3

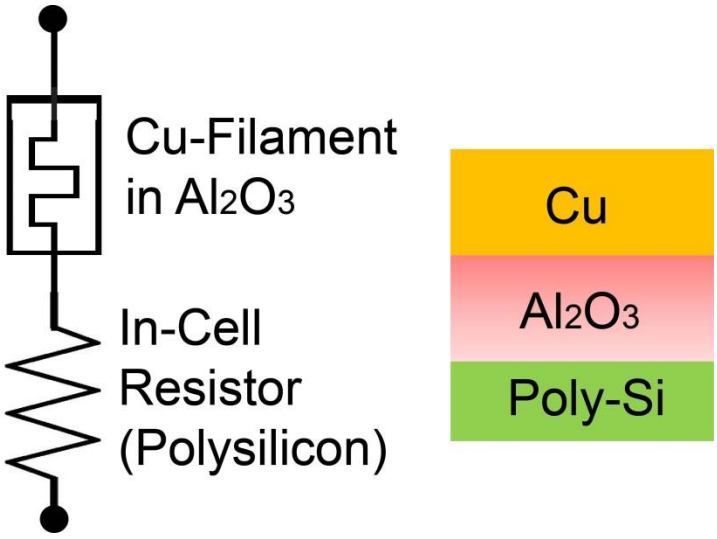


Figure 4-7: Device structure schematic. The polysilicon BE effectively acts as an in-cell resistor and prevents overshoot during writing

The effect of the polysilicon bottom electrode in-cell resistor was confirmed by fabricating devices with 10nm Al_2O_3 but with metal (e.g. W) bottom electrodes. All devices with W BEs shorted out (Fig. 4-8) after the forming operation even when very low current compliance (0.5nA) was used. This is consistent with previous reports suggesting reduction in voltage/current overshoot during forming / SET process is required to prevent over-programming of the device[18,19].

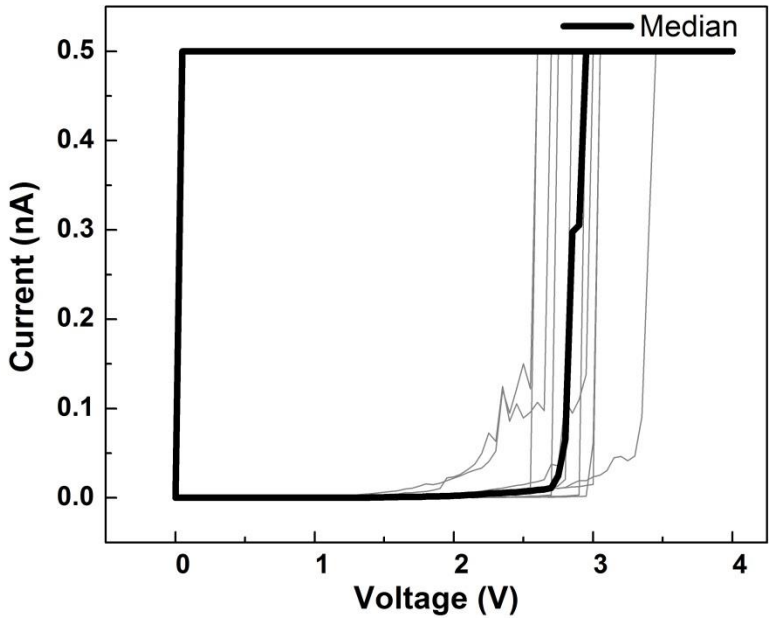


Figure 4-8: Forming curves of 10 different control devices (W / 10nm Al_2O_3 / Cu). The devices cannot be cycled even with very low current compliance.

If the Al_2O_3 thickness increases beyond 15nm, stuck-at-1 (SA1) faults are seen. With increased dielectric layer thicknesses, the voltage needed for forming increases. With increasing voltages, the overshoot in the current (when the device forms) also increases and this tends to make the formed filament too robust[21]. Due to this robust filament, the device yield drops and most devices get stuck in the written state right after forming. For example, devices with 20nm Al_2O_3

have a relatively high but uniform forming voltage ~13V (Fig. 4-9) and get SA1 right after forming. 10nm Al_2O_3 devices, on the other hand, show a non-linear IV after forming and can be cycled well. The fact that the forming voltage of the 20nm device is much more than double that of the 10nm device can be explained by the incorporation of the copper in the dielectric during the fabrication stage prior to forming.

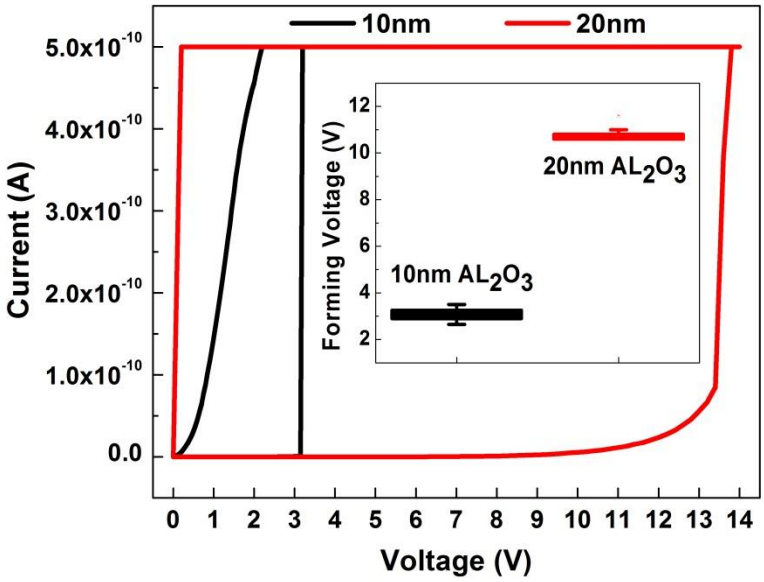


Figure 4-9: Comparison of forming curves of devices with different thickness. Forming voltages show tight distribution (inset). The 20nm Al_2O_3 devices require significantly higher forming voltage and typically cannot be cycled.

On the other extreme, devices fabricated with thin Al_2O_3 are relatively leakier with higher off-state currents (Fig. 4-10). For example, the off-state current (at 2 V) for 8.5nm Al_2O_3 devices is higher than the off state current for the 10nm devices at the same voltage. Also the devices with thinner ALD layer do not show sharp switching and have writing voltages that are very similar to the forming voltage. Thus, thinner insulator layers are a possible solution for achieving forming free devices.

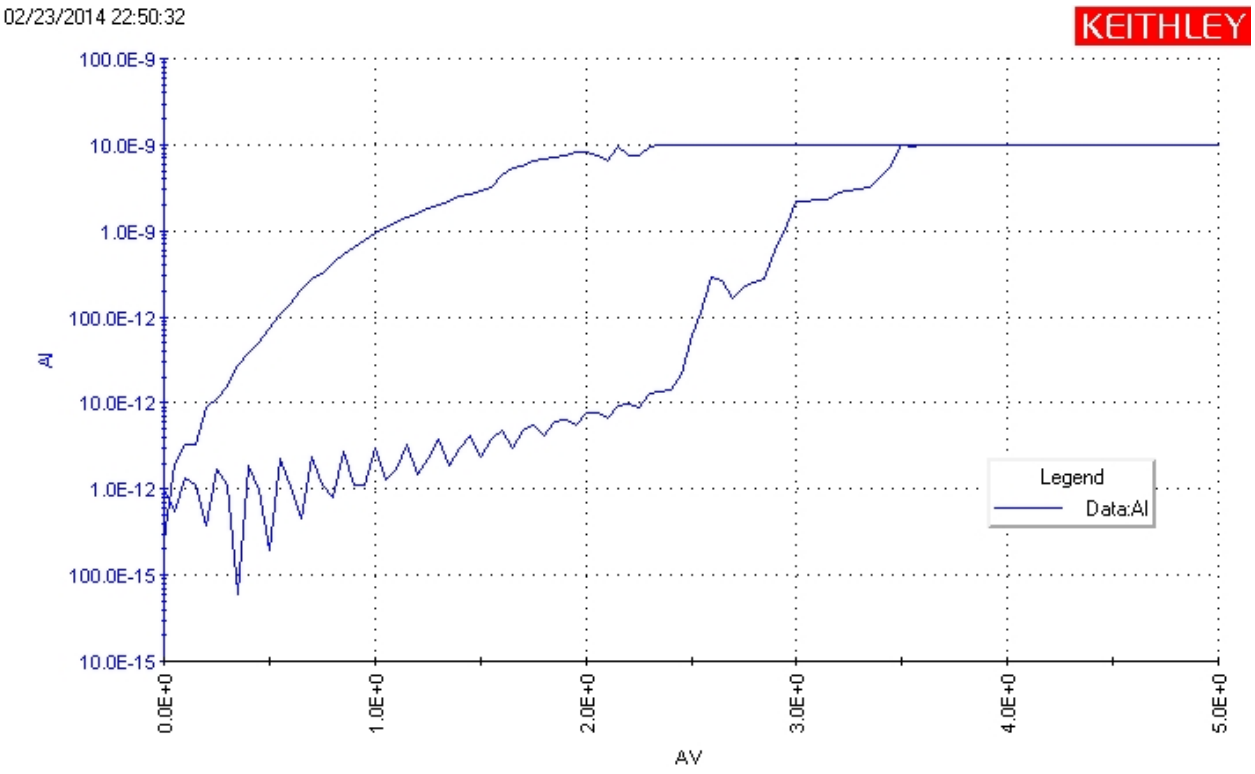


Figure 4-10: Current-voltage curve of a Cu / 8.5nm AL_2O_3 / Poly-Si virgin device.

As the off state current scales with device area[22], making the device smaller (by using electron beam lithography, nano-imprint lithography etc.) would suppress the off state current even further and allow the devices to have even higher I_{on}/I_{off} ratio for practical array operations.

The low current compliance and the in-cell resistor prevent formation of thick solid filaments which would otherwise result in high on-state current and linear I-V characteristics. It has been shown that the dilemma between programming current and retention is due to the spontaneous diffusion of the filament material when a very weak filament is formed at low programming current[9] . With the low current compliance and the in-cell resistor to prevent voltage overshoot, we expect to be able to control the filament growth process. Specifically, instead of the formation of a very weak filament, we target an incomplete filament which has a

solid base but leaves a gap between the filament tip and the bottom electrode, schematically shown in the Fig. 4-11. This idea is based on the filament growth concept verified by SEM and TEM studies discussed in Section 2.2

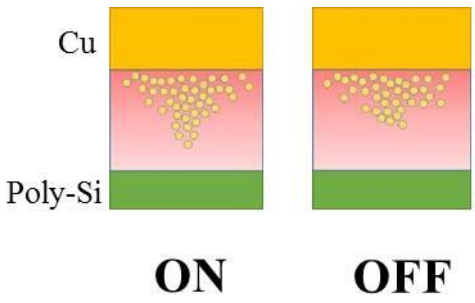


Figure 4-11: Since the copper filament does not completely bridge the two electrodes in the ON state, very low operating current can be obtained.

In this case, low current is obtained since the filament does not completely bridge the two electrodes; while good retention can still be maintained as the filament does not have a very weak tip that leads to retention loss. The concept of controlling filament growth has been recently demonstrated by our group and verified through in-situ TEM studies[23,24], as discussed in section 2.2.

The hypothesis of the incomplete filament formation was supported by analyzing the measured I-V curve in the low-resistance state (LRS). The LRS I-V can be well fitted with a SCLC (space-charge-limited-conduction) model (Fig. 4-12), consistent with electron transport through a thin ALD Al_2O_3 layer[25].

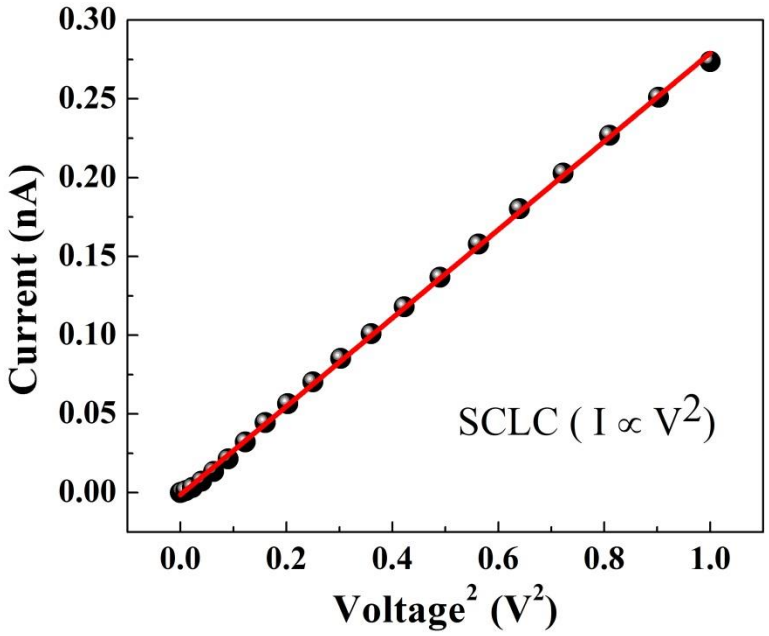


Figure 4-12: Linear fit for on-state current vs. voltage^2 indicating SCLC conduction mechanism.

The Mott-Gurney expression for space charge limited current through a thin material of thickness d and dielectric constant ε is given by

$$J = \frac{9\varepsilon\mu V^2}{8d^3} \qquad (1)$$

where V is the applied voltage. By assuming a gap d between the filament tip and the BE to be ~2-3nm[26], μ(electron mobility in Al_2O_3)=7e-9 m^2/V-s[25], ε_r = 4[25], the effective electrode area responsible for LRS conduction can be calculated (from (1)) to be 8-26nm^2, suggesting the presence of a dominant filament with an effective tip diameter of 3-5nm. The estimated filament and gap size are consistent with the observed filament shape/characteristics from experiments targeted at visualizing the actual filament[23,24,27] and support the concept of having a partially formed filament in the LRS to maintain low programming current and retention (note on/off > 3000 can still be obtained as shown in Fig. 4-6).

4.3.2 Retention and Endurance

Despite the ultra-low programming current, the devices show good retention behavior (Fig. 4-13a), since the current is not limited by a very thin (thus unstable) filament but rather by the gap between the filament tip and the bottom electrode. An incomplete but robust filament results in stable read current even at high temperature (Fig. 4-13a). Excellent endurance can be obtained in both operation modes - 10000 pulse cycles (Fig. 4-13b) and >100 DC cycles (Fig. 4-13c). Cycle-to-cycle uniformity of SET voltage (Fig. 4-13d) also shows the reliability of the controlled filament growth process even under such ultra-low set and reset current conditions.

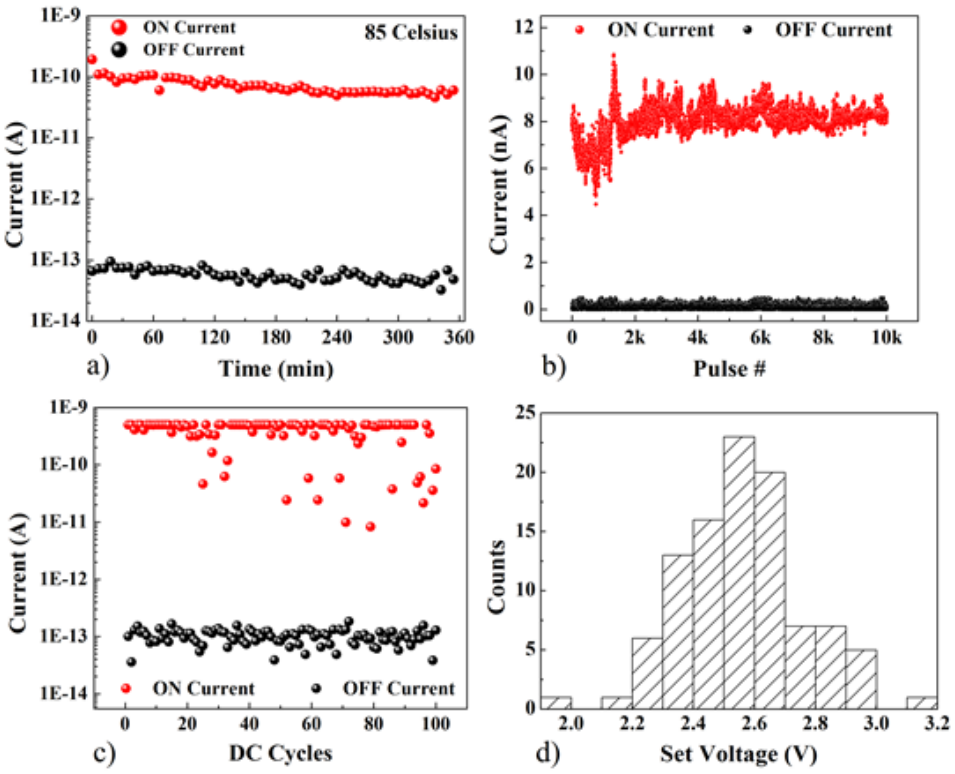


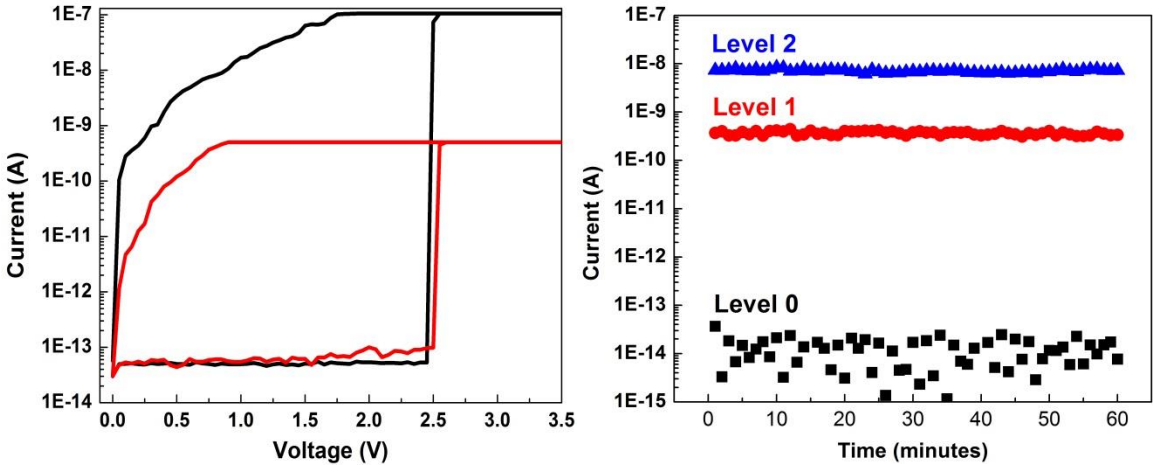
Figure 4-13: a) Elevated temperature retention test. Read pulse (1V/10ms) was repeated every 6 minutes. Large read window is maintained after 6 hours at 85°C. b) 10 000 cycle pulse data indicating robust endurance. Write pulse: 5V/5ms, erase pulse: -2.5V/4ms. , read pulse: 1V, 10ms c) Endurance data from 100 DC cycles. ON current and OFF current were read at 1V. d) Distribution of the SET voltage taken from 100 consecutive DC sweeps. Mean SET voltage is 2.53V with standard deviation of 0.2V.

4.3.3 Multilevel Operation

Multilevel capability is widely studied in memory devices to allow scaling of the cost/bit metric. A cell with multi-level capability – or a multi-level cell (MLC) is capable of storing more than one bit of information. The benefit of multi-level cell storage is that storage capacity may be increased without a corresponding increase in process complexity.

MLC capability is a widely researched topic in RRAM. Different current levels can be obtained by changing the external series resistor[12,27], current compliance[28,29], write (erase) pulse width / height[30–33]. Due to the high ON resistance of our devices, the series resistors needed to obtain multilevel memory need to have very high values (due to the voltage divider effect). This can be detrimental since the added series resistors will increase the RC delay seen by the programming circuitry. Modifying the programming pulse height / width to program the device to different states adds algorithmic and design overheads and makes the control circuitry more complex. One of the simplest and cheapest ways to obtain multilevel memory is to control the compliance current while programming the device. The compliance current can be controlled easily by changing the gate potential of the transistor associated with an RRAM device in an already ubiquitous 1T-1R structure[29,34–36].

In this work, the current level can be easily controlled by controlling the shape of the filament. This is achieved by changing the current compliance while programming the device (Fig. 4-14a). Stable read current can be obtained from these states with a well-defined read margin exceeding 10 (Fig. 4-14b). Programming with higher compliance currents decreases endurance as the filament tends to become too robust and devices tend to get SA1. On the other extreme, devices tend to not retain their state if programmed below 50pA.

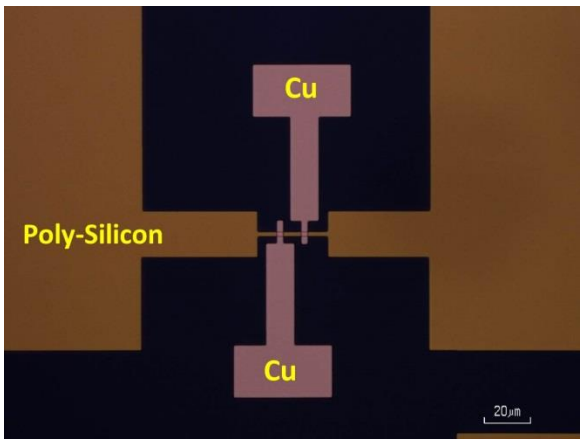


a) b)

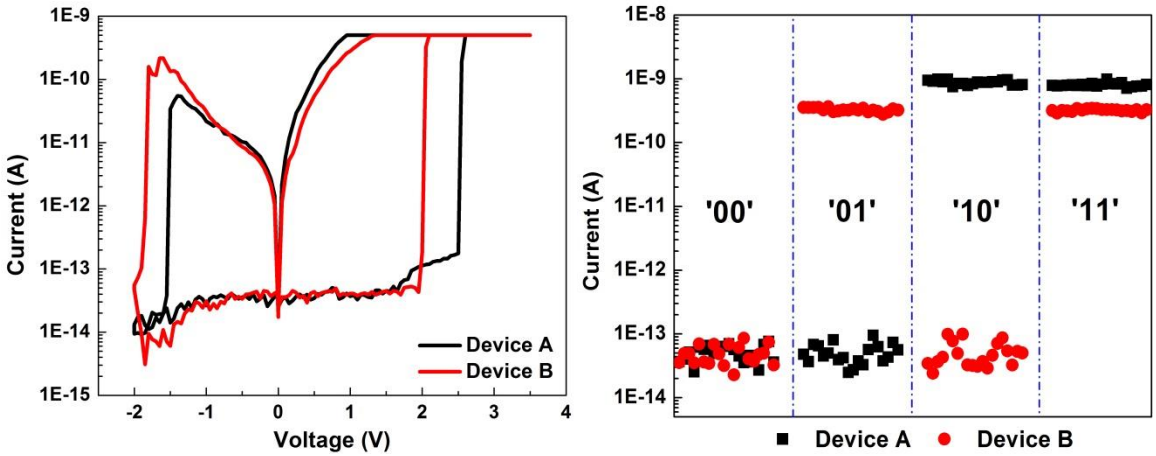
Figure 4-14: Multilevel cell (MLC) capability achieved by controlling compliance current during programming (a). Different states exhibit stable read window (b). Device state was read with 1V pulse repeated every 1 minute

4.3.4 Array Operation

2 x 1 arrays were fabricated. (Fig. 4-15a). The devices show well matched I-V curves (Fig. 4-15b) and were used for storing various bit patterns – 00, 01, 10, 11 (Fig. 4-15cc). To avoid interference between the two cells, voltage bias is applied on the copper top electrodes while the polysilicon bottom electrode is held at ground potential using the Keithley 4200 SMU. No crosstalk is observed in this small array even though the two top electrodes share a common bottom electrode.



a)



b) c)

Figure 4-15: (a) 2x1 arrays fabricated with two Cu top electrodes sharing a poly-silicon bottom electrode (Scale bar 20um). (b) Matched I-V curves for one such 2x1 array. (c) Devices exhibit stable read currents for different combinations of device states – reset / reset, reset/set, set/reset and set/set. The devices were written with DC sweeps.

4.4 Passive Crossbar Arrays

The incomplete filament allows the device to exhibit non-linear I-V at LRS (e.g. Fig. 4-6). Additionally, the device shows self-rectification. As shown in Fig. 4-16, in the ON state, the current at 1V is much higher than the current at -1V. This rectification ratio can exceed 25 typically.

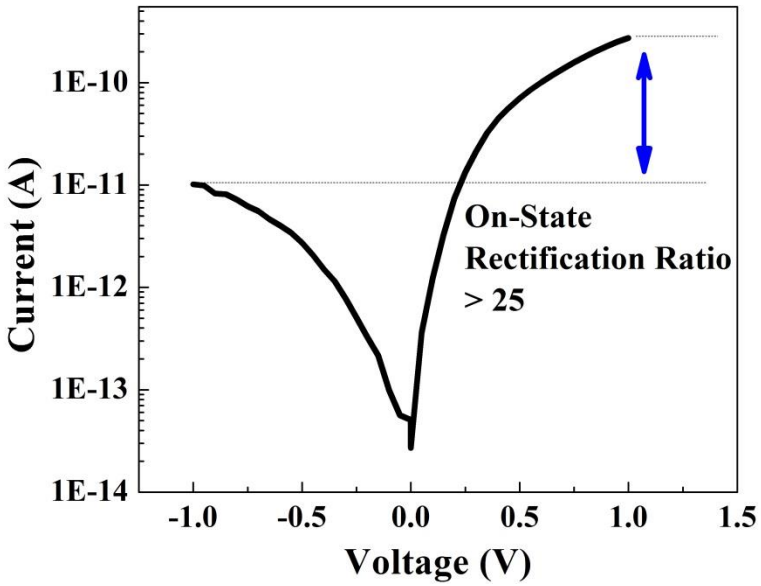


Figure 4-16: Intrinsic rectifying behavior in the device on-state.

The self-rectification behavior and the non-linear I-V characteristics are beneficial to improve the read margin in passive RRAM crossbar arrays [37]. Combined with the ultra-low operating current, we expect good performance from high density arrays based on these devices. Indeed, numerical simulation for N*N square arrays (Fig. 4-17a) confirms improved read margin in arrays due to device non-linearity and intrinsic rectification. The read current loss is negligible in a 1M-bit array (1024*1024) and is less than 10% within a 16M-bit array (4096*4096). Further, the low current negates the detrimental effects of line resistance of the word/bit lines. Normally, increased line resistance causes increased voltage drops on the selected word line and bit line and also raises or lowers the potential on unselected word lines which can lead to inaccurate read current. Due to the low current, the voltage drops are minimized and no significant read current degradation is observed for 512*512 arrays with line resistance as high as 1000 ohm/square(Fig. 4-17b).

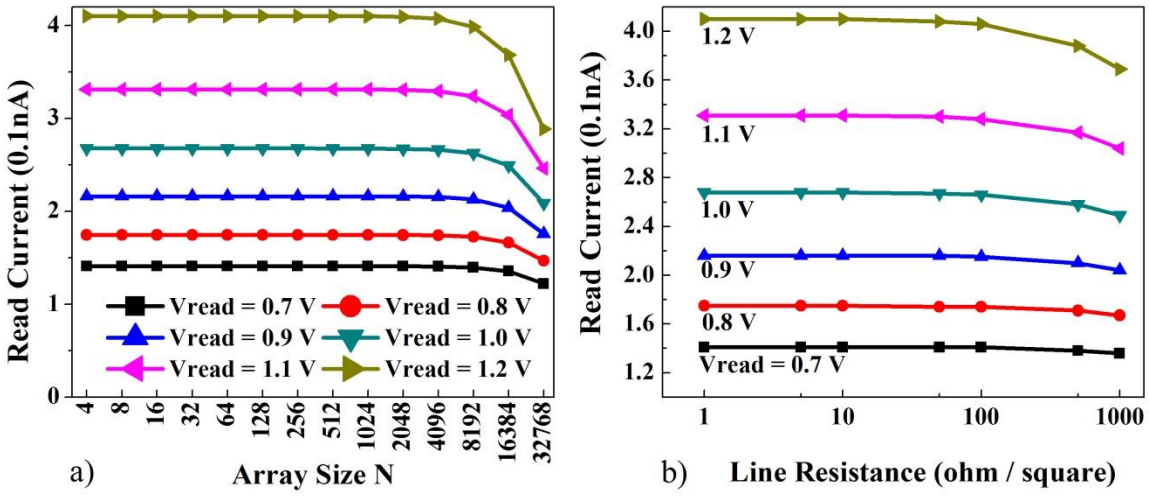


Figure 4-17: Read margin for square arrays with N rows and N columns. Line resistance of 100 ohm/sq. is assumed. Grounding scheme is utilized for array simulation - unselected word-lines and bit- lines are held at ground potential. The selected word-line is biased biased at V_{READ} while the selected bit-line is grounded. Worst case scenario with the target cell located at the farthest corner and all unselected cells are in low resistance state. b) Read current degradation as a function of word / bit line resistance for N = 512.

4.5 Conclusion

Sub-nA operation of RRAM devices with intrinsic current rectification has been demonstrated. The polysilicon in-cell resistor and low current compliance prevent over programming of the device and leads to controlled tuning of the filament geometry. Such low current operation coupled with excellent retention and non-linear self-rectifying behavior is promising for array operation and for integration with novel nanowire / nanotube based systems.

References

[1] L. Goux, K. Sankaran, G. Kar, N. Jossart, K. Opsomer, R. Degraeve, G. Pourtois, G. Rignanese, and C. Detavernier, “Field-driven ultrafast sub-ns programming in W \ Al2O3 \ Ti \ CuTe-based 1T1R CBRAM system,” VLSI Technol. (VLSIT), 2012 Symp., pp. 69–70, 2012.

[2] B. Govoreanu, G. S. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl, M. Jurczak, B.- Leuven, and K. U. Leuven, “10x10nm2 Hf / HfOx Crossbar Resistive RAM with

Excellent Performance , Reliability and Low-Energy Operation," Electron Devices Meet. (IEDM), 2011 IEEE Int. IEEE, pp. 729–732, 2011.

[3] A. D. Franklin and Z. Chen, "Length scaling of carbon nanotube transistors.," Nat. Nanotechnol., vol. 5, no. 12, pp. 858–62, Dec. 2010.

[4] J.-P. Colinge, C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A.-M. Kelleher, B. McCarthy, and R. Murphy, "Nanowire transistors without junctions.," Nat. Nanotechnol., vol. 5, no. 3, pp. 225–9, Mar. 2010.

[5] M. M. Shulaker, G. Hills, N. Patil, H. Wei, H.-Y. Chen, H.-S. P. Wong, and S. Mitra, "Carbon nanotube computer.," Nature, vol. 501, no. 7468, pp. 526–30, Sep. 2013.

[6] F. Schwierz, "Graphene transistors.," Nat. Nanotechnol., vol. 5, no. 7, pp. 487–96, Jul. 2010.

[7] W. Lu, P. Xie, and C. M. Lieber, "Nanowire Transistor Performance Limits and Applications," IEEE Trans. Electron Devices, vol. 55, no. 11, pp. 2859–2876, Nov. 2008.

[8] J. Yao, H. Yan, S. Das, J. F. Klemic, J. C. Ellenbogen, and C. M. Lieber, "Nanowire nanocomputer as a finite-state machine.," Proc. Natl. Acad. Sci. U. S. A., vol. 111, no. 7, pp. 2431–5, Feb. 2014.

[9] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current," 2013 IEEE Int. Electron Devices Meet., pp. 10.1.1–10.1.4, Dec. 2013.

[10] Y. Chen, H. Lee, P. Chen, W. Chen, K. Tsai, P. Gu, T. Wu, C. Tsai, S. Z. Rahaman, Y. Lin, F. Chen, M. Tsai, and T. Ku, "Novel Defects-Trapping TaOx / HfOx RRAM With Reliable Self-Compliance, High Nonlinearity, and Ultra-Low Current.," IEEE Electron Device Lett., vol. 35, no. 2, pp. 202–204, 2014.

[11] C. Ho, C. Hsu, C. Chen, J. Liu, C. Wu, C. Huang, C. Hu, and F. Yang, "9nm half-pitch functional resistive memory cell with <1μA programming current using thermally oxidized sub-stoichiometric WOx film," 2010 Int. Electron Devices Meet., pp. 19.1.1–19.1.4, Dec. 2010.

[12] K.-H. Kim, S. Hyun Jo, S. Gaba, and W. Lu, "Nanoscale resistive memory with intrinsic diode characteristics and long endurance," Appl. Phys. Lett., vol. 96, no. 5, p. 053106, 2010.

[13] W. Kim, S. Il Park, Z. Zhang, Y. Yang-liauw, D. Sekar, H. P. Wong, and S. S. Wong, "Forming-Free Nitrogen-Doped AlOx RRAM with Sub- μA Programming Current," VLSI Technol. (VLSIT), Symp., pp. 22–23, 2011.

[14] C. Schindler, M. Weides, M. N. Kozicki, and R. Waser, “Ultra-low current resistive memory based on Cu-SiO2,” Silicon Nanoelectron. Work. 2008,IEEE., pp. 1–2, 2008.

[15] M. Morita, “Native Oxide Films and Chemical Oxide Films,” in Ultraclean Surface Processing of Silicon Wafers, T. Hattori, Ed. Springer Berlin Heidelberg, 1998, pp. 543–558.

[16] H. Kahn, C. Deeb, I. Chasiotis, and a. H. Heuer, “Anodic oxidation during MEMS processing of silicon and polysilicon: native oxides can be thicker than you think,” J. Microelectromechanical Syst., vol. 14, no. 5, pp. 914–923, Oct. 2005.

[17] M. Groner, J. Elam, F. Fabreguette, and S. George, “Electrical characterization of thin Al 2 O 3 films grown by atomic layer deposition on silicon and various metal substrates,” Thin Solid Films, vol. 413, pp. 186–197, 2002.

[18] S. M. George, “Atomic layer deposition: an overview.,” Chem. Rev., vol. 110, no. 1, pp. 111–31, Jan. 2010.

[19] F. Xiong, M.-H. Bae, Y. Dai, A. D. Liao, A. Behnam, E. a Carrion, S. Hong, D. Ielmini, and E. Pop, “Self-aligned nanotube-nanowire phase change memory.,” Nano Lett., vol. 13, no. 2, pp. 464–9, Feb. 2013.

[20] S. Gaba, S. Choi, P. Sheridan, T. Chang, Y. Yang, and W. Lu, “Improvement of RRAM Device Performance Through On-Chip Resistors,” MRS Proc., vol. 1430, pp. mrss12–1430–e09–09, Jun. 2012.

[21] S. Tirano, L. Perniola, J. Buckley, J. Cluzel, V. Jousseaume, C. Muller, D. Deleruyelle, B. De Salvo, and G. Reimbold, “Accurate analysis of parasitic current overshoot during forming operation in RRAMs,” Microelectron. Eng., vol. 88, no. 7, pp. 1129–1132, Jul. 2011.

[22] S. H. Jo and W. Lu, “CMOS compatible nanoscale nonvolatile resistance switching memory.,” Nano Lett., vol. 8, no. 2, pp. 392–7, Feb. 2008.

[23] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, “Observation of conducting filament growth in nanoscale resistive memories.,” Nat. Commun., vol. 3, p. 732, Jan. 2012.

[24] Y. Yang, P. Gao, L. Li, X. Pan, S. Tappertzhofen, S. Choi, R. Waser, I. Valov, and W. D. Lu, “Electrochemical dynamics of nanoscale metallic inclusions in dielectrics.,” Nat. Commun., vol. 5, no. May, p. 4232, Jan. 2014.

[25] H. Spahr, J. Reinker, T. Bülow, D. Nanova, H.-H. Johannes, and W. Kowalsky, “Regimes of leakage current in ALD-processed Al2O3 thin-film layers,” J. Phys. D. Appl. Phys., vol. 46, no. 15, p. 155302, Apr. 2013.

[26] S. H. Jo, K.-H. Kim, and W. Lu, “High-density crossbar arrays based on a Si memristive system.,” Nano Lett., vol. 9, no. 2, pp. 870–4, Feb. 2009.

[27] U. Celano, L. Goux, A. Belmonte, K. Opsomer, A. Franquet, A. Schulze, C. Detavernier, O. Richard, H. Bender, M. Jurczak, and W. Vandervorst, “Three-dimensional observation of the conductive filament in nanoscaled resistive memory devices.,” Nano Lett., vol. 14, no. 5, pp. 2401–6, May 2014.

[28] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, “Device and SPICE modeling of RRAM devices.,” Nanoscale, vol. 3, no. 9, pp. 3833–40, Sep. 2011.

[29] S. Sheu, P. Chiang, and W. Lin, “A 5ns fast write multi-level non-volatile 1 k bits rram memory with advance write scheme,” VLSI Circuits, 2009 Symp., pp. 82–83, 2009.

[30] C.-H. Hsu, Y.-S. Fan, and P.-T. Liu, “Multilevel resistive switching memory with amorphous InGaZnO-based thin film,” Appl. Phys. Lett., vol. 102, no. 6, p. 062905, 2013.

[31] W. C. Chien, Y. C. Chen, K. P. Chang, E. K. Lai, Y. D. Yao, P. Lin, J. Gong, S. C. Tsai, S. H. Hsieh, C. F. Chen, K. Y. Hsieh, R. Liu, and C. Lu, “Multi - Level Operation Of Fully CMOS Compatible WOX Rresistive Random Access Memory (RRAM),” Mem. Work. 2009. IMW’09. IEEE Int. IEEE, vol. 91, pp. 1–2, 2009.

[32] S. Lee, Y. Kim, and M. Chang, “Multi-level switching of triple-layered TaOx RRAM with excellent reliability for storage class memory,” VLSI Technol. (VLSIT), 2012 Symp., pp. 71–72, 2012.

[33] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, “High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm.,” Nanotechnology, vol. 23, no. 7, p. 075201, Feb. 2012.

[34] Y. Tseng and C. Huang, “High density and ultra small cell size of contact ReRAM (CR-RAM) in 90nm CMOS logic technology and circuits,” IEDM Tech. Dig. IEEE Int. Electron Devices Meet. 2009., pp. 5.6.1–5.6.4, 2009.

[35] M. Wu, Y. Lin, and W. Jang, “Low-Power and Highly Reliable Multilevel Operation in 1T1R RRAM,” Electron Device Lett., vol. 32, no. 8, pp. 1026–1028, 2011.

[36] S. Koveshnikov and K. Matthews, “Real-time study of switching kinetics in integrated 1T/HfO x 1R RRAM: Intrinsic tunability of set/reset voltage and trade-off with switchingtime,” IEDM Tech. Dig. IEEE Int. Electron Devices Meet. 2012, pp. 486–488, 2012.

[37] J. Zhou, K. Kim, and W. Lu, “Crossbar RRAM Arrays: Selector Device RequirementsDuring Read Operation,” IEEE Trans. Electron Devices, vol. 61, no. 5, pp. 1369–1376, 2014.

Chapter 5

3D Vertical Dual-Layer Oxide RRAM for Vertical Memory

5.1 Introduction

To maintain functional scaling of integrated circuits, vertical scaling, that aims at enhancing the device performance or functionality through expansion in the vertical direction, is now being widely researched for future memory and logic applications. In particular, one advantage of RRAMs based on the simple two-terminal structure [1-3] is their compatibility with vertical scaling. Generally, two different approaches are being investigated for vertical scaling in resistive memory devices: traditional 3D cross-point RRAM[4] using stacked cross-point arrays in a layer-by-layer fashion; and vertical RRAM structures[5-6] based on devices formed at the sidewall between a vertical electrode and a lateral electrode with the capability to form multiple layers simultaneously.

In this chapter, the two approaches are first compared. We discuss the advantages and disadvantages of both approaches and argue that the vertical 3D structure is more cost-effective. This advantage in terms of less number of lithography steps is demonstrated in a prototype vertical RRAM structure using a W/WO_X/Pd RRAM system. Gradual analog switching behavior thus obtained is shown to be useful in analog memory and neuromorphic systems.

5.2 3D Integration

3D integration technology is being actively researched as a possible solution to continue the scaling trajectory predicted by the Moore's law [7–9]. 3D integration solves important performance limitations correlated with simply shrinking of CMOS device sizes. 3D integration promises larger functionality packed into smaller form factors while improving performance and reducing costs.

3D integration includes many technologies like wire-bonding [10], wafer-on-wafer integration [11-12], die-on-wafer integration [13], monolithic integration [9] etc. Among all these available technologies, monolithic integration is the one which yields the highest density of devices[9]. In monolithic approaches, devices are processed sequentially starting from the bottommost layer. A second layer of devices is then built after depositing an isolation layer. This process of isolation and device formation continues to make a multi-layer structure. No space penalties associated with TSVs and alignment apply to monolithic 3D integration. Additionally, more interconnects between the different device layers can be fabricated in a more local and distributed fashion to improve bandwidth between the different layers. Because of these advantages, memory chip vendors have looked at monolithic integration to increase bit density and reduce the cost/bit metric. While 3D FLASH memory has already transitioned to the commercial markets, 3D RRAM has received broad of attention from the research community as a possible replacement for 3D FLASH.

As 3D FLASH memory scales, the cost of lithography steps increases drastically. The number of critical masks spiral with increasing number of layers and has been shown to be economically unviable (Fig. 5-1a). The fabrication cost is limited by the critical mask count and the cost of the lithography tools. 8 layers are possible without resorting to currently immature EUV

pattering techniques (Fig5-1b). Since a RRAM cell is inherently smaller and simpler (and hence cheaper to fabricate) than FLASH; it is being looked at as a possible replacement for FLASH in the near future. As such, it becomes increasingly important to study the vertical scaling of RRAM to understand the issues and possible roadblocks.

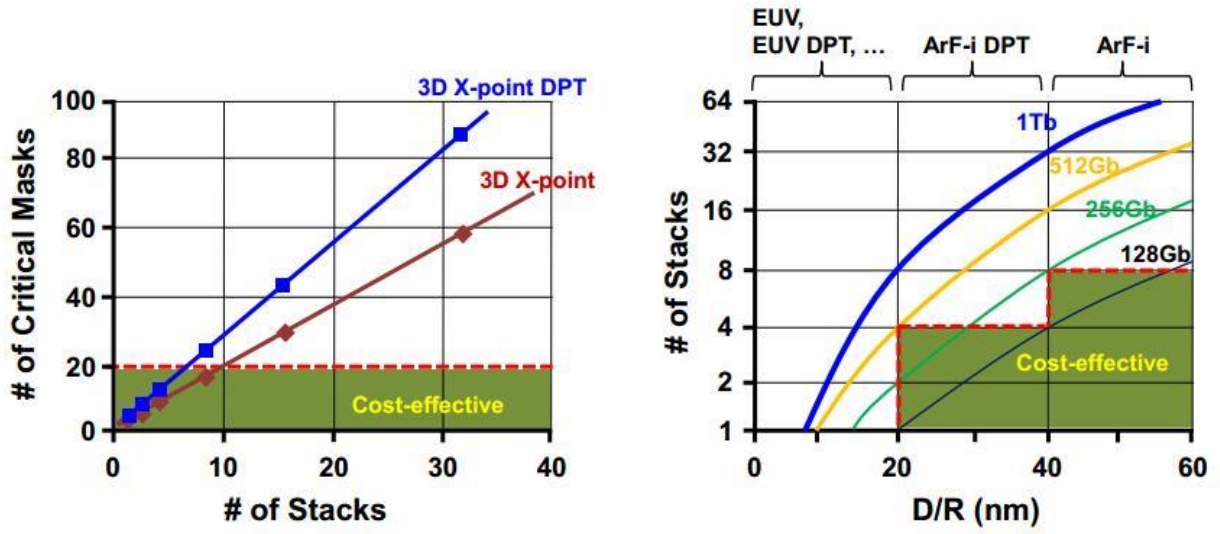


Figure 5-1 : Minimum number of critical masks necessary for 3D X-point FLASH arrays as a function of stacked layer number, (b) Memory density as a function of design rule and the number of stacked memory layers composed of $4F^2$ sized 2bit MLC cells at the condition of chip area x cell efficiency = $100mm^2$. Reproduced from [5].

5.3 3D Monolithic Integration: Traditional Crosspoint vs. Vertical Sidewall Structure

Two different approaches have been proposed for monolithic vertical scaling of RRAM devices. The first – the traditional crosspoint approach – involves fabricating the first layer of crosspoint devices and then depositing an interlayer dielectric. Then the crosspoint fabrication is repeated to yield the second layer. One layer is formed each time as shown in Fig 5-1a. The vertical 3D RRAM structure takes a different approach whereby multi-layer horizontal electrodes are fabricated first and then the vertical electrodes are fabricated yielding cross-points on the sidewalls

of the horizontal electrodes rather than the top of the horizontal electrodes as is the case in the traditional 3D structure.

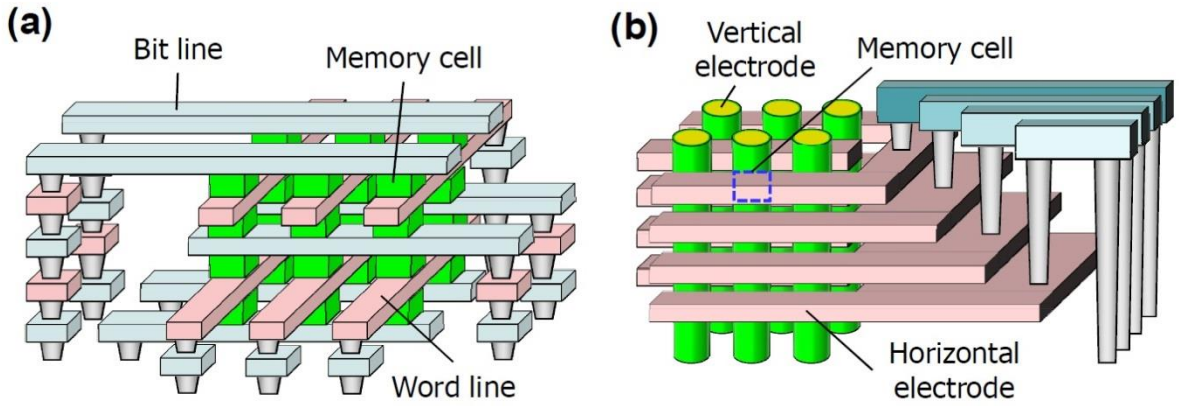


Figure 5-2: Schematics of the traditional crosspoint structure (a) and vertical 3D structure (b). Reproduced from [5]

The vertical 3D RRAM structure or sidewall type of structure has several advantages over the traditional crosspoint-type of device structure. In a traditional cross point device, the active area dimensions are completely defined by lithography; while in sidewall devices at least one active device dimension is not critically dependent on lithography. The top (vertical) electrode is still defined by lithography in both conventional crosspoint devices and in sidewall devices. In the latter case, however, the active area dimension is determined by the thickness of a deposited film which can be precisely controlled to the atomic level, as opposed to the lithographically defined dimension of the bottom electrode. Additionally, since deposition thicknesses can be controlled to a much better extent than defined by lithography, device to device variation can be improved. Finally, for vertical RRAM structures the number of critical lithography steps remains fairly constant as the number of layers increases, implying significant improvements in cost savings compared with stacked cross-point approaches. In the next section, we detail the fabrication of a

prototype vertical 3D sidewall device. Another vertical 3D sidewall device with digital switching behavior is being developed and is described in Chapter 6.

5.4 Device Fabrication

The dual layer vertical devices were fabricated on a Si/SiO_2 substrate with 100nm of thermal oxide. The first horizontal electrode layer of tungsten (40nm) was deposited at room temperature by DC sputtering in a Kurt J Lesker LAB 18 system. Although many different metal oxides have been studied as candidates for memristive devices[14], tungsten-based materials were chosen for this demonstration due to the ubiquitous use of W in commercial CMOS processes and the rich knowledge of this material. Silicon dioxide (60nm) serving as the inter layer dielectric, was then deposited at 200°C in a plasma enhanced chemical vapor deposition system (GSI PECVD). Ellipsometric and X-SEM methods were utilized to control all thicknesses to within +/- 10% of nominal values. The tungsten and silicon dioxide depositions were then repeated to form the dual-layer horizontal electrode. Next, photolithography and reactive ion etching (RIE) were used to pattern the film stack (Fig 5-3).

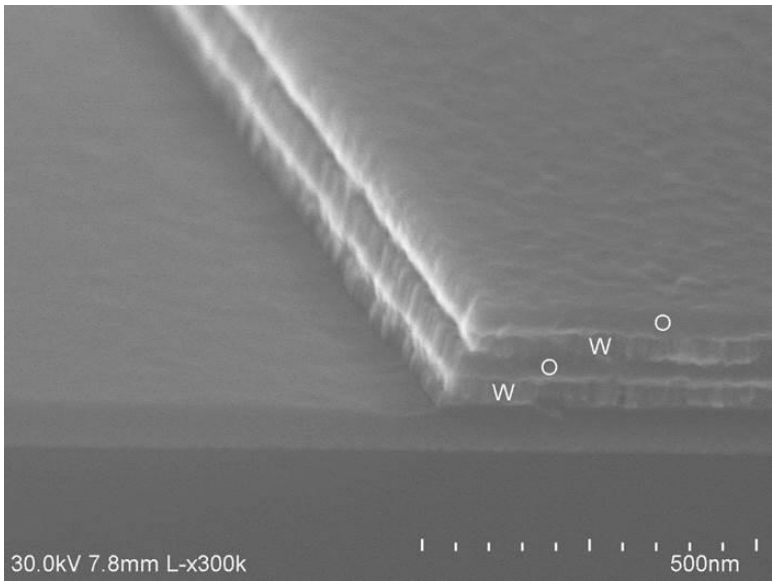


Figure 5-3: Scanning electron micrograph after the bottom electrode stack etch.

To form the tungsten oxide (WO_X) switching layer, the sample was annealed in an oxygen rich ambient at 375°C at atmospheric pressure for 60 seconds in a JetFirst 150 RTP system. The exposed sidewalls were oxidized to form WO_X while the remaining bulk of the tungsten, which was covered by the PECVD silicon dioxide, served as the horizontal electrodes. Afterwards, the vertical Pd electrodes were patterned through photolithography, e-beam evaporation and liftoff techniques to complete the $Pd/WO_X/W$ device structure at each sidewall junction (Fig 5-4).

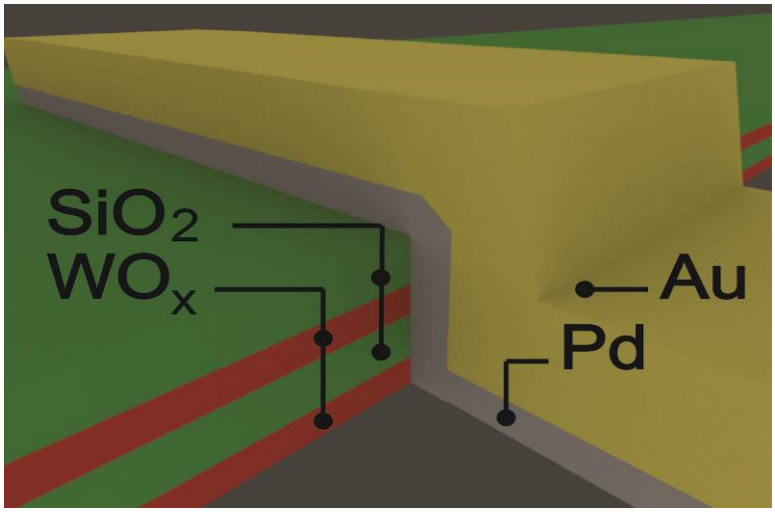


Figure 5-4: Schematic showing the dual layer device structure

To improve sidewall coverage of the top electrode, the sample was placed at an angle of ~45 degrees to the normal incident direction during electron beam evaporation of the vertical electrode material (Pd 600 Å / Au 2400 Å). Finally, photolithography and RIE were used to open contact pads to the two horizontal tungsten electrode layers, followed by gold pad deposition (Fig 5-5).

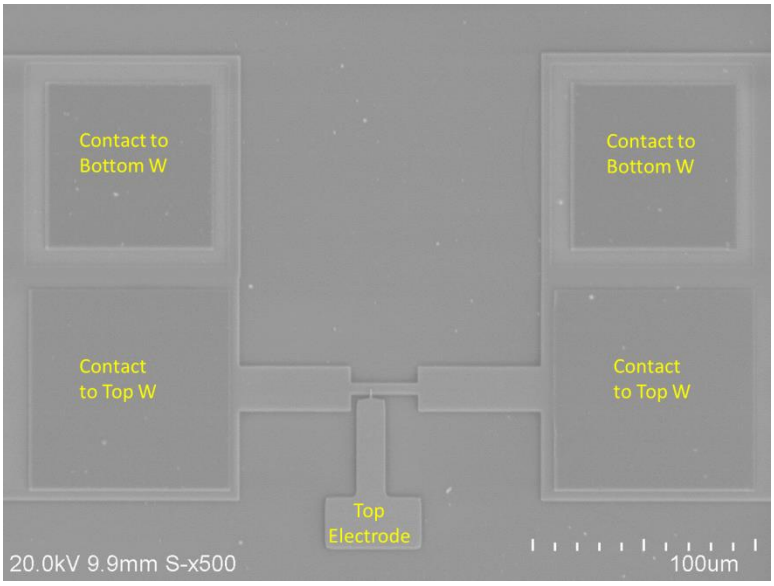


Figure 5-5: Contacts to the bottom W and top W layers were opened using lithography and RIE.

Throughout the process the peak temperature was limited to 375°C to maintain CMOS compatibility. Cross-sectional SEM images of a completed device are shown in Fig. 5-6.

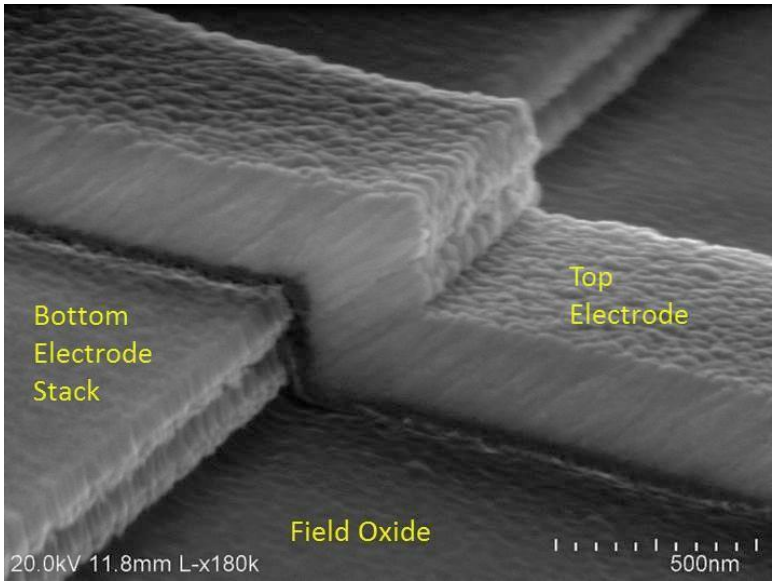


Figure 5-6: Scanning electron micrograph of the completed device.

5.5 Measurement Setup

The fabricated devices were electrically probed in a Lakeshore PS-100 Tabletop Cryogenic Probe Station. Electrical data was collected using a National Instruments data acquisition system (NI USB-6259 BNC) in conjunction with a DL 1211 current preamplifier from DL Industries. Custom code written in MATLAB / LabVIEW was used to generate and measure the signals. The shared Pd vertical electrode is held at ground potential while the electrical bias is applied to the W horizontal electrodes as shown in Fig 5-7. Electrical bias can be selectively applied to either the top W layer or the bottom W by utilizing analog switches S1 and S2 (Analog Devices ADG1412).

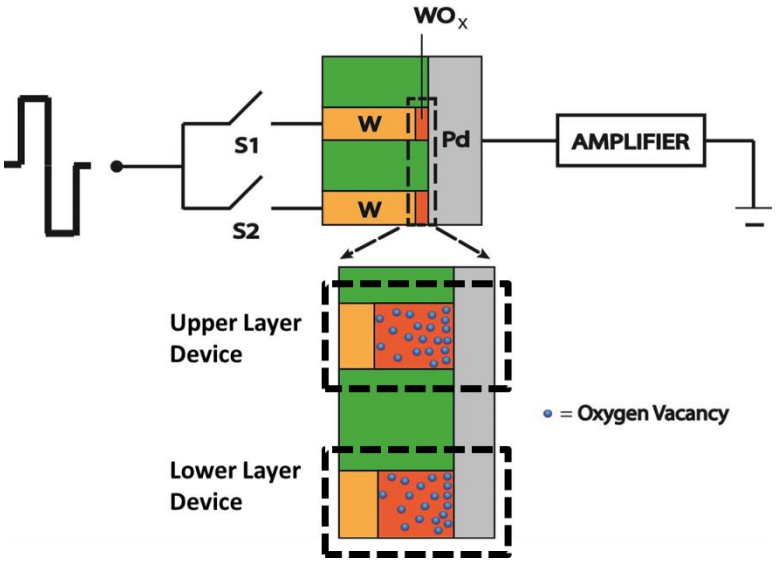


Figure 5-7: Experimental setup for electrical characterization of the dual layer device

Contrary to the CBRAM type of devices discussed in Chapters 2-5, which require an active electrode (e.g. Ag or Cu) to inject metal cation into the switching layer (a-Si or Al2O3, which does not directly participate in the redox process), the device discussed here has inert electrodes (Pd and W) while the switching happens internally inside the WOx layer by redistributing the oxygen vacancies (V_Os) inside the film, as schematically shown in Fig 5-7. More detailed discussion on the device operation and modeling can be found in the next section.

5.6 Results and Discussions

5.6.1 Bipolar Operation

The devices were tested in DC operation mode using the custom-built test circuitry described earlier (Fig. 5-7). Fig. 5-8 shows the I-V characteristics obtained from both the upper device and the lower device. A typical resistive switching behavior can be observed with well-defined hysteresis.

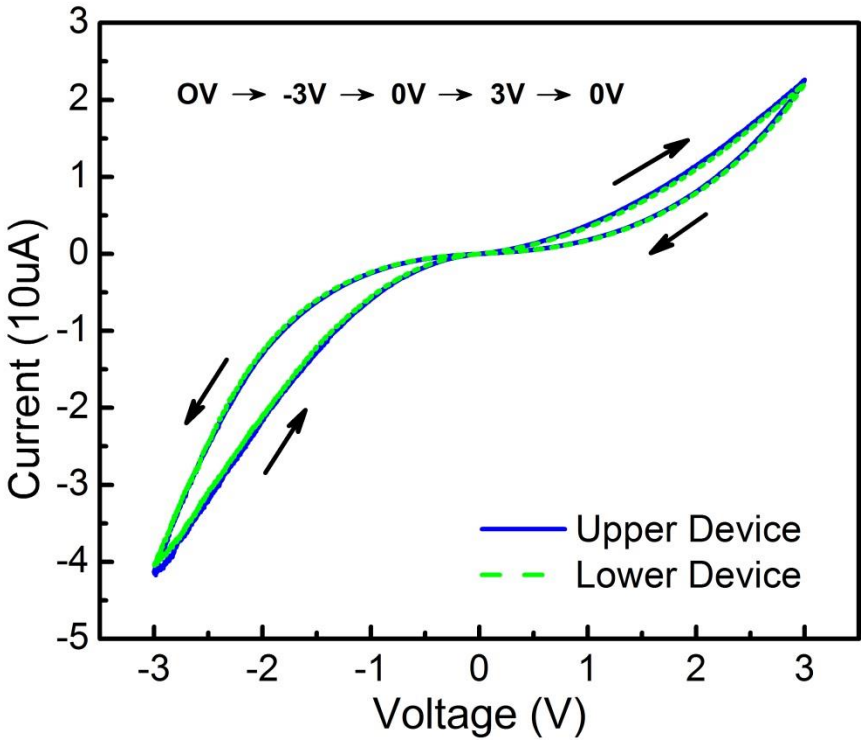


Figure 5-8: I-V plot for each device in the dual layer structure.

The resistive switching observed in these WO_X devices is bipolar - the voltage polarities for increasing/decreasing device resistance are opposite, in contrast to unipolar devices where the resistance can be increased/decreased with single-polarity signals by adjusting current compliance levels. The bipolar nature and the hysteresis in the I-V curves are attributed to the migration of oxygen vacancies in the non-stoichiometric WO_X matrix [15] . During oxidation process, there are more oxygen vacancies generated near the outer surface (i.e. near the vertical Pd electrode side (Fig. 5-7)). These oxygen vacancies act as dopants and modulate the local conductivity of the WOx matrix. Due to the excess dopants, a nearly ohmic contact is created at the WOx/vertical electrode interface while a Schottky contact is created at the WOx/W horizontal electrode interface. In this configuration, the total device resistance is dominated by the V_O poor region near the horizontal W electrode. Applying a negative voltage to the W electrode drives the migration of the positively charged oxygen vacancies towards the W electrode. As sufficient oxygen vacancies move close to

the W electrode, the Schottky junction width is reduced to allow efficient tunneling through it and improves the overall device conductance. Conversely, applying a positive voltage to the W electrode drives the oxygen vacancies away towards the Pd electrode and thus makes the device less conductive.

This device behavior is similar to results obtained from 2D horizontal devices[15]. While horizontal devices rely on oxidation of a pristine as-deposited W interface to generate the WO_X matrix with an oxygen vacancy concentration gradient, these vertical devices are obtained by oxidation of an etched sidewall. This vertical device concept suggests that high quality WO_X materials can still be obtained at the electrode sidewall reliably.

5.6.2 Current Non-linearity

In the WOx device[15], the current can be modeled as a sum of two components: 1) Schottky current, formed between WOx and W bottom electrode in the V_O poor region, 2) tunneling current between WOx and W bottom electrode in the more conductive V_O rich region after programming. If we assume the normalized area of the V_O rich region formed by oxygen vacancies is w, then

$$I = w\gamma sinh(\delta V) + (1 - w)\alpha[1 - exp(-\beta V)] \qquad (1)$$

, where I is the current, V is the applied voltage, $\alpha, \beta, \gamma, \delta$ are all device-specific parameters. The first term is the tunneling current, indicating the current through the conduction region and the second term is the current through Schottky contact region.

Therefore, if positive voltage is applied to device, $w\gamma sinh(\delta V)$ dominates the current and nonlinearity comes from sinh function. If negative voltage is applied, $w\gamma sinh(\delta V) +$

$(1-w)\alpha[1-exp(-\beta V)] \approx w\gamma sinh(\delta V) - (1-w)\alpha exp(-\beta V)$, which also shows strong nonlinearity.

From an application perspective, this non-linearity is extremely useful for array operation since it alleviates the sneak path problem [16–18] as discussed earlier in Section 2.4

5.6.3 Device Matching

Very similar I-V curves can be obtained from both devices in different stacks along the same vertical electrode (Fig. 5-8) indicating close matching between the two devices in the dual layer approach. The close matching is further demonstrated by comparing the current though each device while applying five consecutive negative DC set cycles (0V to -3V) followed by five consecutive positive DC reset cycles (0V to +3V) to each device (Fig. 5-9a). Each device demonstrates an incremental change in conductance as expected from an analog memristive device – gradual increase on applying a negative voltage and gradual decrease on applying a positive voltage, and very similar behavior can be observed in both devices (Fig. 5-9b).

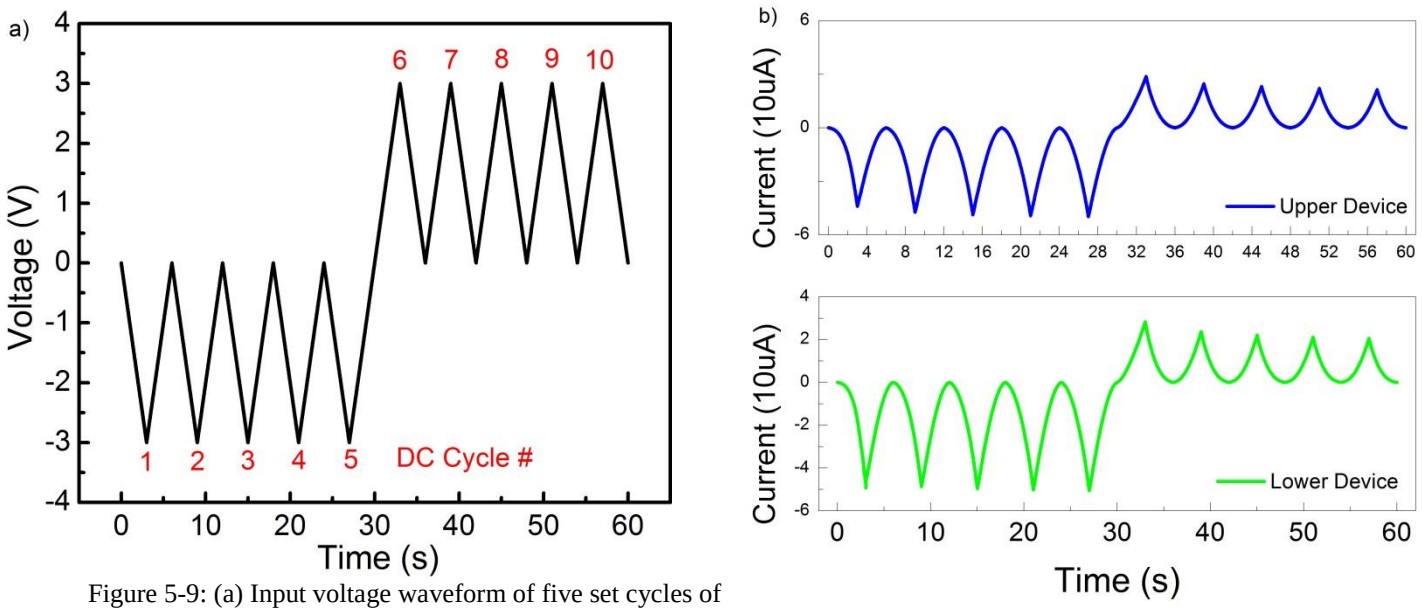


Figure 5-9: (a) Input voltage waveform of five set cycles of 0 to -3V followed by five reset cycles of 0 to +3V at 1V/s. (b) Current output for the upper device(upper panel) and the lower device (lower panel).

Fig. 5-10 plots the device current measured at the maximum programming voltage of -3V during the 5 consecutive set cycles, demonstrating <10% mismatch between the upper layer and lower layer devices. Matching response to pulse stimuli is discussed in section 5.6.5.

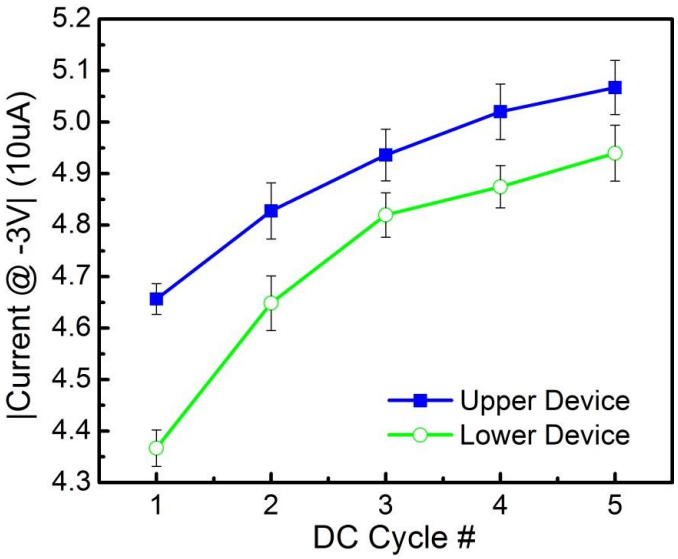


Figure 5-10: Incremental change of the maximum current during DC programming for both devices. The error bars were obtained from five different DC sweep measurements.

5.6.4 Pulse Operation

While DC operation gives more insight into the physics behind device operation, pulse operation is more desirable from an integrated circuit perspective. The fabricated dual layer devices were tested using pulses as shown in Fig 5-11. Switching polarity is the same as that in DC operation. A negative voltage pulse (applied to the horizontal W electrode) increases the conductance of the device while a positive voltage pulse decreases the conductance of the device. In other words, a negative voltage pulse writes the device while a positive voltage pulse erases the device. A low voltage read is used to read the state of the device. The voltage is kept low in order not to disturb the state of the device.

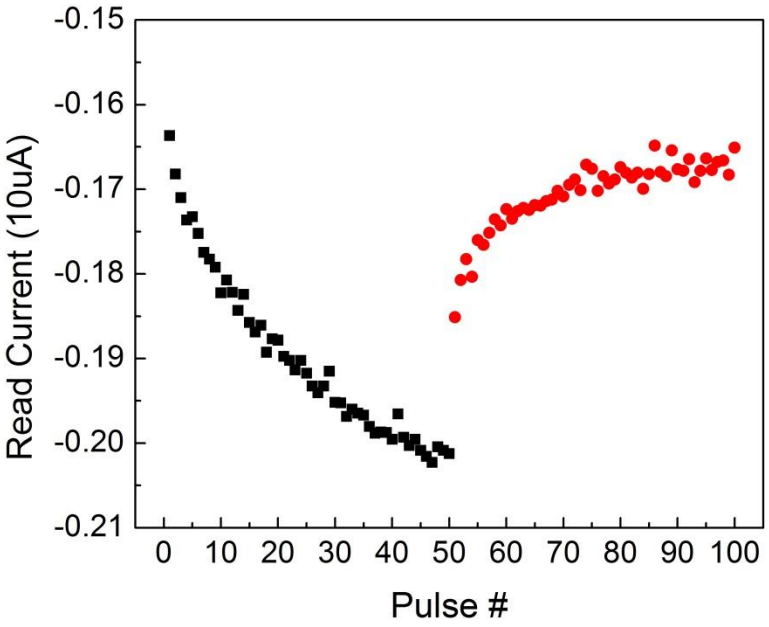


Figure 5-11 : Write/erase voltages were applied to the device in the form of 50 write pulses followed by 50 erase pulses. Read current after each voltage pulse is measured and plotted. Each write pulse was -3V/400us and each erase pulse was +3V/400us. The read voltage pulse was -0.8V/10ms. Upper device in the dual layer stack was used for this test.

This incremental analog increase / decrease in current is remarkably similar to potentiation /depression observed in biological synapses (Fig. 5-12). In a biological synapse, the activities of the pre-neuron and post neuron (e.g. frequency and timing of the action potentials or spikes) result in either an increase of synaptic weight (potentiation) or decreases of synaptic weight (depression). The ability to modulate the conductance of analog RRAM devices with voltage pulses allows emulation of synaptic functions with these nanoscale devices. Specifically, in biological synapses, the weight change is believed to be mediated by certain ions (e.g. Ca2+) whose concentration is in turn modulated by the spikes, while in analog RRAMs the conductance change can be argued to follow similar ionic dynamics (e.g. oxygen vacancies) modulated by voltage pulses.

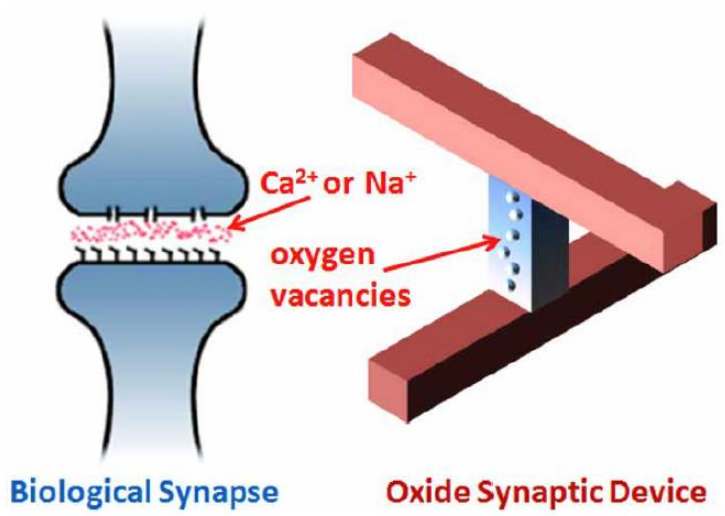


Figure 5-12: A representation showing the similarity between a biological synapse and a RRAM device. Taken from [19]

Analog resistive switching devices have been widely proposed to emulate synaptic functions in neuromorphic circuits [20–24]. In particular, large connectivity can be obtained when these devices are organized in a crossbar structure [25] that mimics the network structure of

biological systems [20]. The ability to extend the network to the vertical direction in the 3D vertical structure demonstrated in this chapter will further improve the connectivity and allow large scale network developments. The two devices operate in parallel and thus can be conceptually seen as two independent devices sharing common bottom electrode, but with separate top electrodes. Alternatively, the vertically stacked memristive device array can effectively act as a single synapse to mitigate the limited resolution and the stochastic switching characteristics seen in single devices.

5.6.5 Endurance

Endurance is a very important metric in resistive switching devices. For large scale integration and use in real-life systems, devices should not degrade with prolonged operation. We applied 10K cycles of potentiation-depression pulses and measured the read current after each potentiation (depression) pulse. Each cycle consist of 50 potentiation pulses (each followed by a read pulse) and 50 depression pulses (each followed by a read). The measured read currents are plotted in Fig. 5-13. No apparent degradation is observed even after 10K cycles.

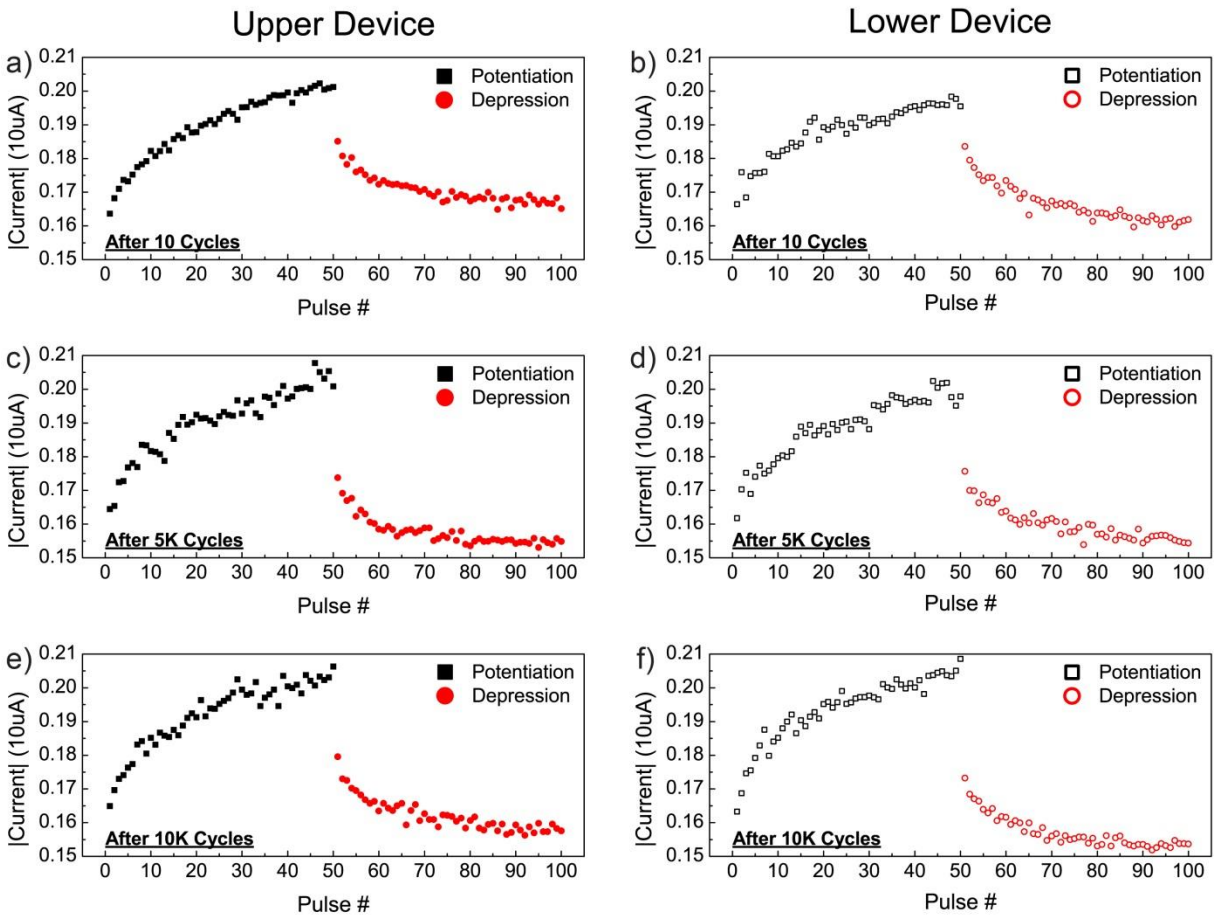


Figure 5-13: Current measured after consecutive potentiation (-3V/400us) and depression pulses (3V/400us) for the upper device (a) and lower device (b). Each cycle comprises of 50 potentiation pulses and 50 depression pulses. The read voltage was -0.8V. The device performance remains unchanged after 5000 cycles (c, d) and 10000 cycles (e, f).

5.6.6 Crosstalk

Since the two devices in the stack share a common electrode, it is important to measure if there is crosstalk between the two cells. Crosstalk may be detrimental when designing large arrays for real-life applications.

We show that each device in the dual layer stack can be programmed and read independently without any crosstalk with the non-addressed device, even though the devices are directly on top of each other and share the same vertical electrode. The two devices operate in

parallel and thus can be conceptually seen as two independent devices sharing common bottom electrode, but with separate top electrodes.

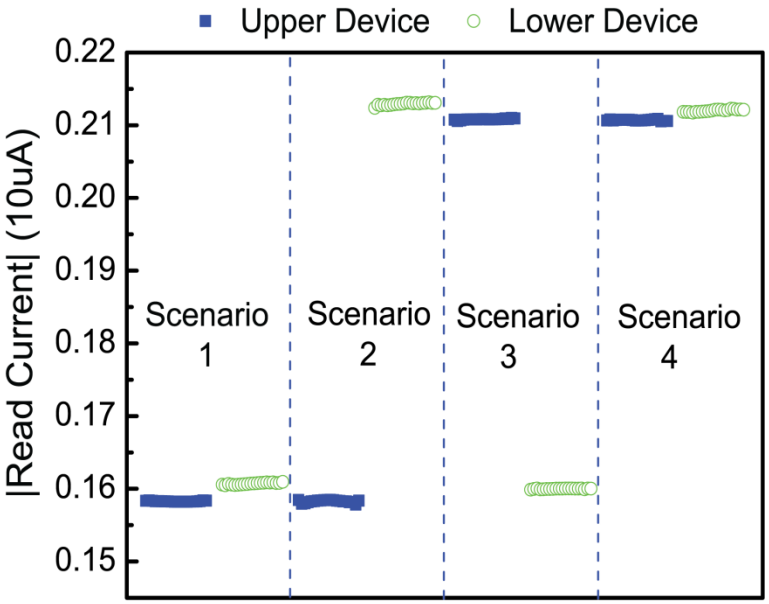


Figure 5-14: Independent programing/read of devices in the dual-layer structure. The read currents obtained during 20 consecutive read pulses were plotted for the four scenarios after the upper/lower device has been reset/reset, reset/set, set/reset, and set/set.

Results from four different programming scenarios are shown in Fig. 5-14 showing independent control of the devices. For example, in the fourth scenario, the upper device was programmed with a -3V/400us pulse train (same as the condition used in Fig. 3). Switch S1 was then opened and the lower device was programmed by closing switch S2. The current level attained by the lower device in this case was then compared to that in scenario 2, where the upper device has not been programmed, and almost identical results were obtained demonstrating independent programming of each device.

5.7 Conclusion

In summary, CMOS compatible, dual-layer vertical tungsten oxide resistive switching devices were demonstrated. The devices show well-defined incremental resistance switching

behavior and good endurance exceeding 10,000 potentiation/depression cycles. The devices can be programmed with less than 10% mismatch and no apparent crosstalk. This scalable architecture is well suited for development of analog memory and neuromorphic systems.

References

[1] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges," Adv. Mater., vol. 21, no. 25–26, pp. 2632–2663, Jul. 2009.

[2] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories—fundamentals, applications, prospects," Nanotechnology, vol. 22, no. 25, p. 254003, Jul. 2011.

[3] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found.," Nature, vol. 453, no. 7191, pp. 80–3, May 2008.

[4] I. Baek, D. Kim, and M. Lee, "Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application," Electron Devices Meet. 2005. IEDM Tech. Dig. IEEE Int., pp. 750–753, 2005.

[5] I. G. Baek, C. J. Park, H. Ju, D. J. Seong, H. S. Ahn, J. H. Kim, M. K. Yang, S. H. Song, E. M. Kim, S. O. Park, C. H. Park, C. W. Song, G. T. Jeong, S. Choi, H. K. Kang, and C. Chung, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," 2011 Int. Electron Devices Meet., pp. 31.8.1–31.8.4, Dec. 2011.

[6] S. Yu, H.-Y. Chen, B. Gao, J. Kang, and H.-S. P. Wong, "HfOx-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture.," ACS Nano, vol. 7, no. 3, pp. 2320–5, Mar. 2013.

[7] D. Choudhury, "3D integration technologies for emerging microsystems," Microw. Symp. Dig. (MTT), 2010 IEEE MTT-S Int., pp. 1–4, 2010.

[8] E. Beyne, "3D System Integration Technologies," 2006 Int. Symp. VLSI Technol. Syst. Appl., pp. 1–9, Apr. 2006.

[9] M. Vinet, P. Batude, C. Tabone, B. Previtali, C. LeRoyer, a. Pouydebasque, L. Clavelier, a. Valentian, O. Thomas, S. Michaud, L. Sanchez, L. Baud, a. Roman, V. Carron, F. Nemouchi, V. Mazzocchi, H. Grampeix, a. Amara, S. Deleonibus, and O. Faynot, "3D monolithic integration: Technological challenges and electrical results," Microelectron. Eng., vol. 88, no. 4, pp. 331–335, Apr. 2011.

[10] M. Karnezos, “3-D Packaging : Where All Technologies Come Together,” Electron. Manuf. Technol. Symp. 2004. IEEE/CPMT/SEMI 29th Int. IEEE, pp. 64–67, 2004.

[11] J. Burns, B. Aull, and C. Chen, “A wafer-scale 3-D circuit integration technology,” Electron Devices, IEEE Trans., vol. 53, no. 10, pp. 2507–2516, 2006.

[12] C.-T. Ko and K.-N. Chen, “Wafer-level bonding/stacking technology for 3D integration,” Microelectron. Reliab., vol. 50, no. 4, pp. 481–488, Apr. 2010.

[13] G. Katti, A. Mercha, J. Van Olmen, C. Huyghebaert, A. Jourdain, M. Stucchi, M. Rakowski, I. Debusschere, P. Soussan, W. Dehaene, K. De Meyer, Y. Travaly, E. Beyne, S. Biesemans, and B. Swinnen, “3D stacked ICs using Cu TSVs and Die to Wafer Hybrid Collective bonding Front End CMOS devices Through Silicon Vias (TSVs),” Electron Devices Meet. (IEDM), 2009 IEEE Int. IEEE, pp. 14.4.1–14.4.4, 2009.

[14] J. J. Yang, D. B. Strukov, and D. R. Stewart, “Memristive devices for computing.,” Nat. Nanotechnol., vol. 8, no. 1, pp. 13–24, Jan. 2013.

[15] T. Chang, S.-H. Jo, K.-H. Kim, P. Sheridan, S. Gaba, and W. Lu, “Synaptic behaviors and modeling of a metal oxide memristive device,” Appl. Phys. A, vol. 102, no. 4, pp. 857–863, Feb. 2011.

[16] K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, “A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications.,” Nano Lett., vol. 12, no. 1, pp. 389–95, Jan. 2012.

[17] M. A. Zidan, H. A. H. Fahmy, M. M. Hussain, and K. N. Salama, “Memristor-based memory: The sneak paths problem and solutions,” Microelectronics J., vol. 44, no. 2, pp. 176–183, Feb. 2013.

[18] S. C. Puthentheradam, D. K. Schroder, and M. N. Kozicki, “Inherent diode isolation in programmable metallization cell resistive memory elements,” Appl. Phys. A, vol. 102, no. 4, pp. 817–826, Jan. 2011.

[19] S. Yu, B. Gao, Z. Fang, and H. Yu, “Stochastic learning in oxide binary synaptic device for neuromorphic computing,” Front. Neurosci., vol. 7, no. October, pp. 1–9, 2013.

[20] G. S. Snider, “Spike-Timing-Dependent Learning in Memristive Nanodevices,” Nanoscale Archit. 2008. NANOARCH 2008. IEEE Int. Symp. on. IEEE, pp. 85–92, 2008.

[21] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, “Nanoscale memristor device as synapse in neuromorphic systems.,” Nano Lett., vol. 10, no. 4, pp. 1297–301, Apr. 2010.

[22] T. Chang, S.-H. Jo, and W. Lu, “Short-term memory to long-term memory transition in a nanoscale memristor.,” ACS Nano, vol. 5, no. 9, pp. 7669–76, Sep. 2011.

[23] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, “Short-term plasticity and long-term potentiation mimicked in single inorganic synapses.,” Nat. Mater., vol. 10, no. 8, pp. 591–5, Aug. 2011.

[24] D. Kuzum, R. Jeyasingh, B. Lee, and H. Wong, “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing,” Nano Lett., vol. 12, no. 5, pp. 2179–2186, 2011.

[25] S. H. Jo, K.-H. Kim, and W Lu, "High-density crossbar arrays based on a-Si memristive system," Nano Lett., vol. 9, no. 2, pp 870-4, Feb. 2009.

Chapter 6
Future Work

6.1 Integration with CMOS Circuits and RRAM Arrays – A Hybrid Approach

Resistive switches described in previous chapters are all passive in nature and do not have any ability to amplify signals. However, they provide very high density that can provide functions such as routing, data storage and information processing (e.g. stateful logic[1] and neuromorphic systems [2,3]) which is unmatched by CMOS. Thus, to take advantage of the high density of these resistive switches and the tremendous functionality of CMOS circuits, a hybrid CMOS/Resistive-switch architecture can be can be realized on the same lines as the proposed hybrid CMOS / nanoelectronics (or CMOL) architecture[4–6]. In this hybrid architecture the interface between the CMOS and the resistive switching devices is provided by regularly distributed pins mounted on top of the CMOS circuits and directly under the resistive switches (Fig 7-1) rather than on the periphery (Chapter 2).

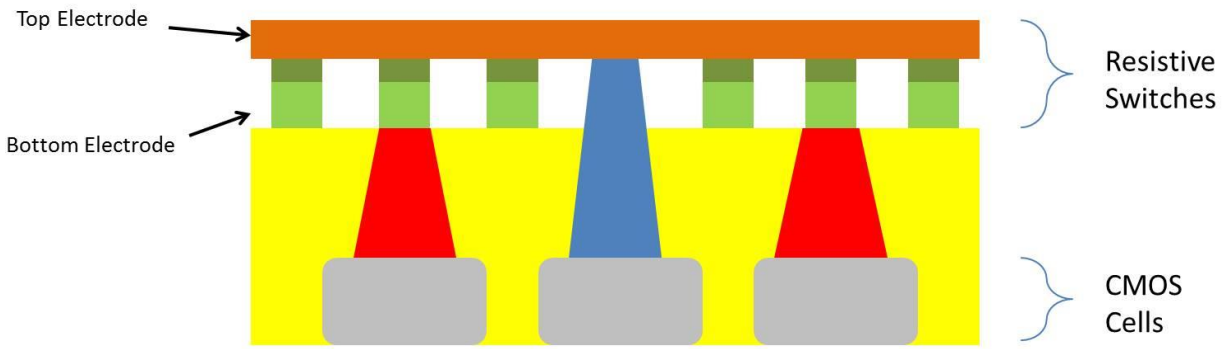


Figure 6-1: Representation of connection of CMOS cells and resistive switches using a CMOL approach.

In the general CMOL approach, the pins are of two types (shown as red and blue in Fig. 6-1), with the blue pins contacting the top electrode of the resistive switches and the red pins contacting the bottom electrodes. The pin array is turned, relative to the resistive switch crossbar, by an angle α where

$$\alpha = \arcsin (F_{NANO} / \beta F_{CMOS})$$

where, F_{NANO} and F_{CMOS} are the pitch of the resistive switch crossbar and the CMOS pin array respectively and β is a numerical factor (typically well above 1) which depends on CMOL cell complexity.

Since identical CMOS and resistive switch units are distributed uniformly, precise alignment is not necessary in the CMOL configuration. In this vertical configuration 2N CMOS cells can be used to control N^2 switches in a N×N crossbar array. This approach thus effectively addresses the pitch mismatch problem since all resistive switches can be accessed through the CMOS circuitry even though $F_{NANO} << F_{CMOS}$ typically.

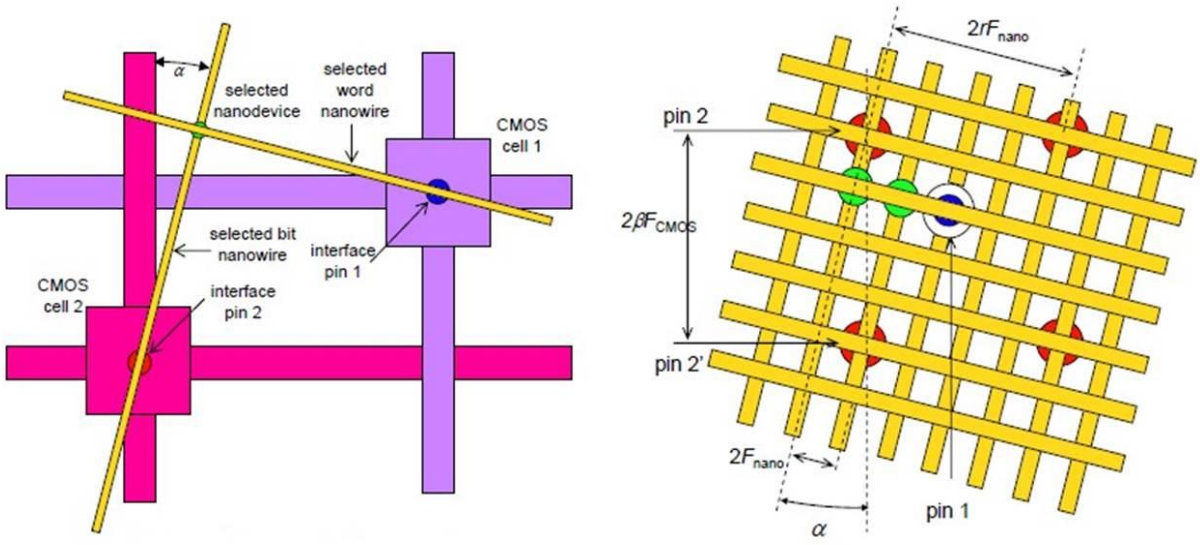


Figure 6-2: Each nano-device or resistive switch can be addressed by a red pin and a blue pin (a). By tilting the crossbars by angle α, each resistive switch can be uniquely addressed by a red pin and a blue pin, without the need for perfect alignment. Reproduced from [4].

Apart from providing the mating capability of two device technologies with very different minimum pitches, the CMOL architecture also provides improved fault tolerance. The CMOS cell connections can be reconfigured around the defective resistive switches, thus improving functional yield despite having high failure occurrences in the resistive switches.

Further, a large standalone crossbar can be broken up into smaller segments accessible through different CMOS cells in the CMOL architecture. Thus, the line resistance issue and sneak path issues are alleviated and array read margins are not affected by increasing array size.

To experimentally demonstrate this hybrid concept, a test chip with such a CMOL interface was designed by our collaborators at University of California, Santa Barbara. The red and blue pins are located in the middle of the chip as shown in the block diagram (Fig. 6-3).

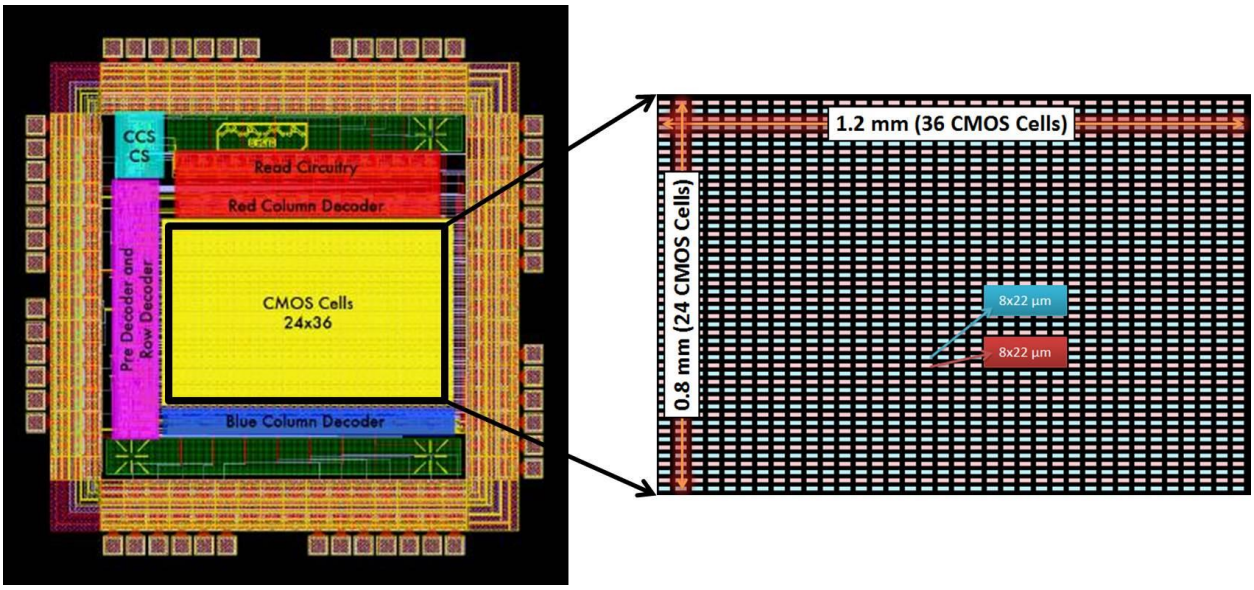


Figure 6-3: Block diagram of the test chip (a). The red and blue pins are designed using the top metal (M3) and form an array in the center of the test chip (b)

The CMOS chip design can be broken down into basic building blocks. The designed chip contains row/ column decoders to address the red and blue pins. Shift registers are also included to take serial address inputs. Current mirror circuits are used to generate reference current signals to distinguish between a stored "1" or "0" – the current through a RRAM cell is compared to a reference current to decide if the cell is in the ON state or OFF state.

This test chip was taped out using AMI's 0.5um 5V process and fabricated at ON Semiconductor. Since the 2mm x 2mm x 260µm chips are too small for handling at UM's Lurie Nanofabrication Facility, they were mounted on a larger piece of silicon using a "wafer level integration" scheme developed at University of California, Santa Barbara[7].

The chip was integrated using two silicon substrates – the carrier and the holder (Fig. 6-4).An oxide coated wafer was covered with negative resist (AZ5214). A hole with the same size as the chip was patterned by using the actual chip as a photo-mask. The pattern was transferred to the oxide layer using a RIE process. This oxide pattern was used to make a through-wafer hole by

using a standard DRIE process. At the end of the DRIE process, the oxide was stripped using BHF. The chip was placed upside down in this cavity by using a handle wafer. Another carrier coated with benzocyclobutene (BCB, Cyclotene 4024-40, Dow Chemical) was bonded onto the holder wafer holding the chip.

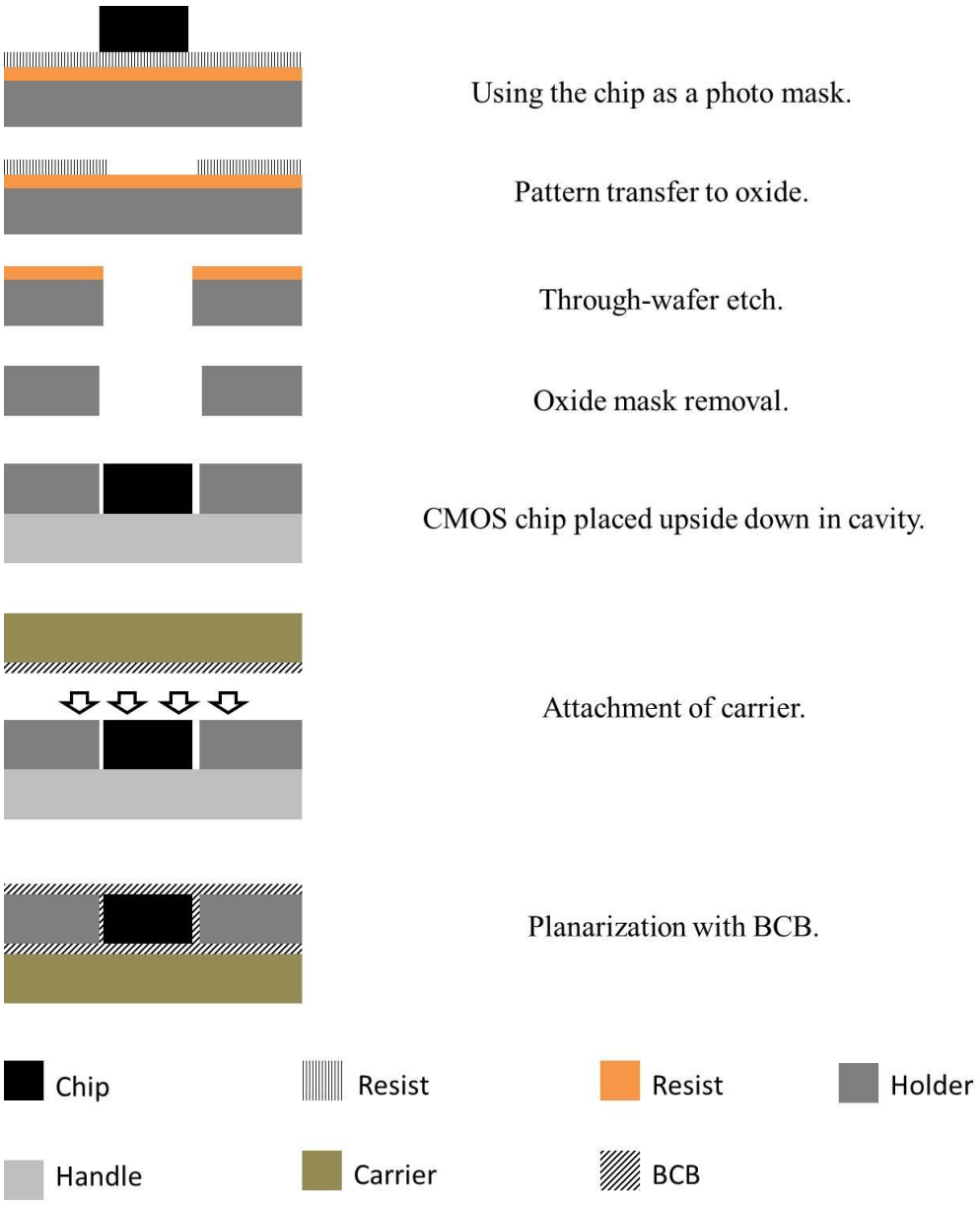
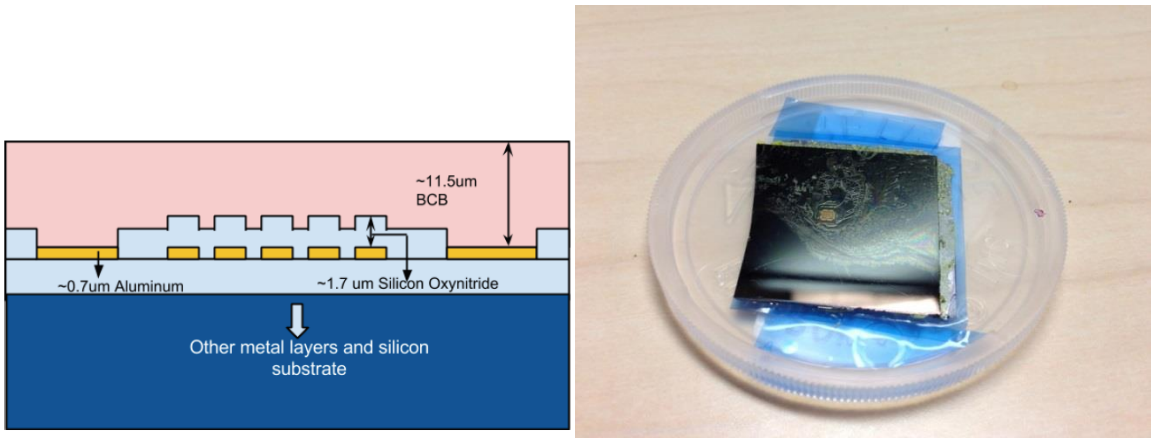


Figure 6-4: Wafer level integration scheme used to integrate the 2mm x 2mm chip onto a larger piece of silicon.

The final surface was planarized using another layer of BCB (Fig. 6-5a). The final integrated platform (Fig. 6-5b) was shipped to UM for RRAM fabrication.



a) b)

Figure 6-5: a) A schematic showing the CMOS pins buried under a thick film of BCB after the completion of the wafer level integration process. b) Photograph of the CMOS IC mounted onto a larger piece of silicon. The round plastic wafer carrier has a diameter of 2 inches, for size reference.

6.1.1 Device Fabrication

WOx test devices were fabricated on the CMOS chip by using standard micro-fabrication techniques. To make electrical contact to the CMOS chip, the BCB was etched using a high power C_4F_8/O_2 based recipe. The BCB etching was found to be very sensitive to the etching chemistry. If the ratio of C_4F_8 to O_2 is too low, the silicon content in the film gradually increases and the film cracks due to high stress buildup. Visual inspection and inline SEM check were used to determine the etch end point (Fig. 6-6).

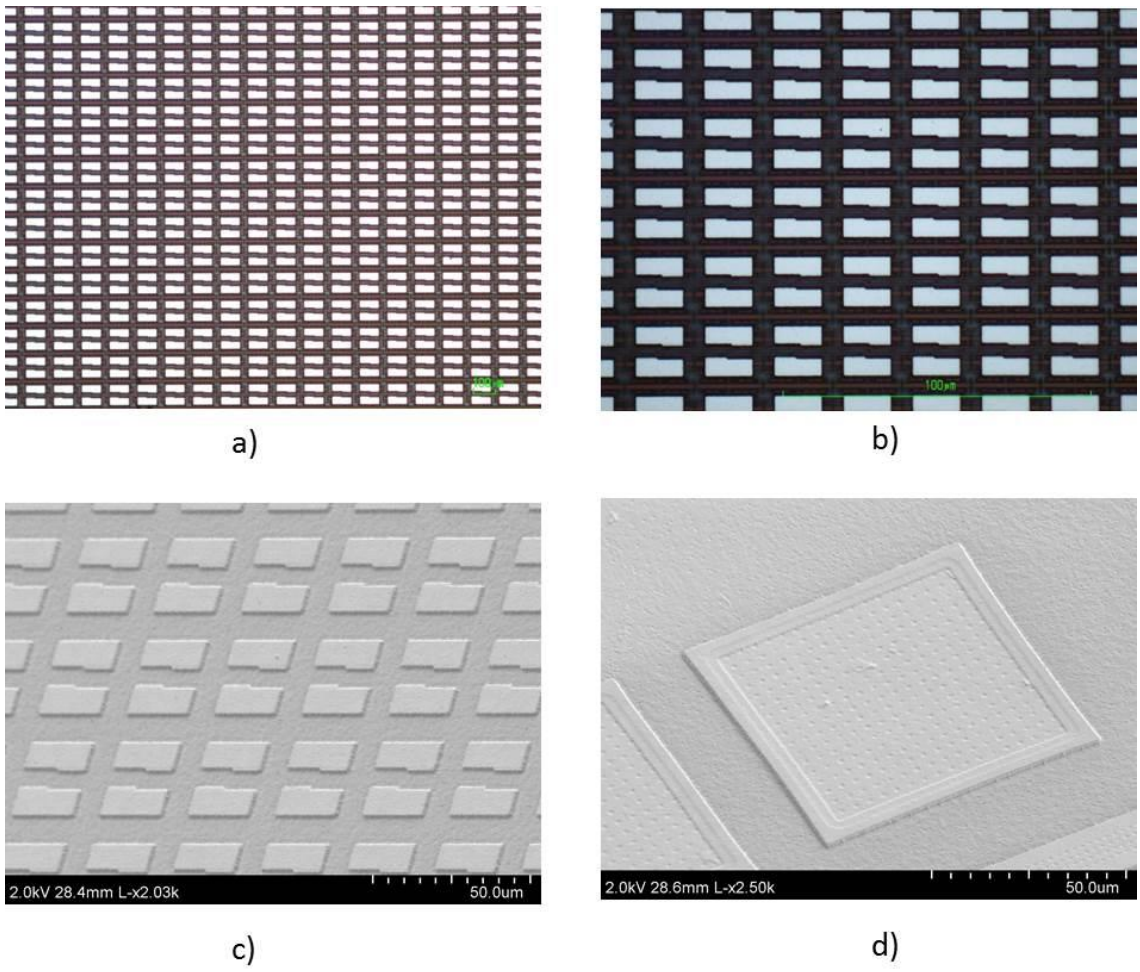
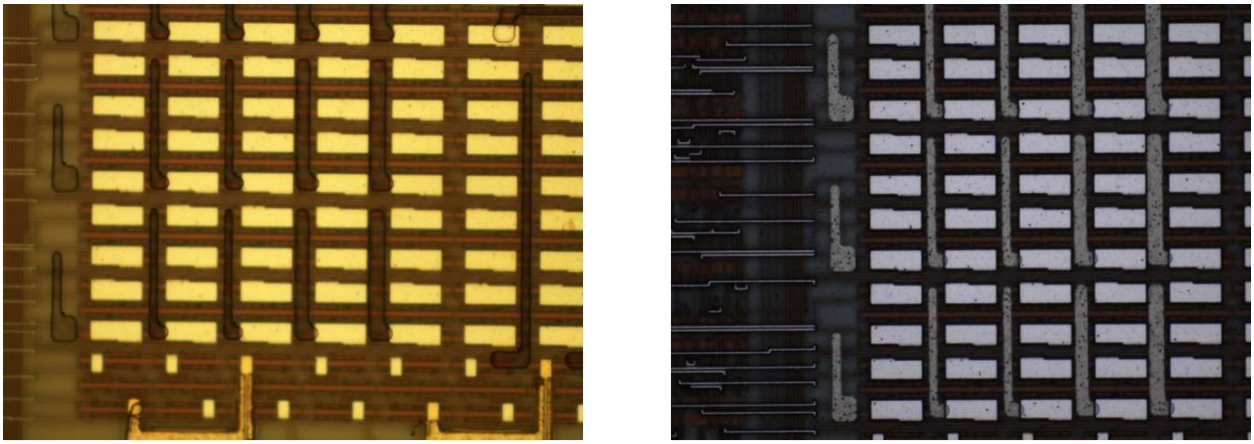


Figure 6-6: Low (a) and high magnification (b) optical images of the CMOS pins after etching away the BCB film. Scanning electron micrographs taken at 35 degree tilt show that the CMOS pins(c) and the I/O pads (d) are completely exposed.

To prevent accidental shorting between the fabricated RRAM devices and the routing on the top metal layer, a layer 500nm PECVD oxide layer was used as an inter-layer dielectric (ILD). Next the bottom electrodes were defined using photolithography (MA/BA-6 contact aligner, CD = 3µm), DC Sputtering (Lab 18 K.J Lesker system, 60nm W, 300W power and 2.5mT sputter pressure with 3e-6 Torr base pressure) and liftoff processes (Fig. 6-7).



a) Post Develop b) Post W Liftoff

Figure 6-7: Optical micrographs of the CMOS chip after develop step and after the W liftoff step. Each CMOL pin is 8µm x 22µm for size reference.

The active layer (WOx) is formed by oxidizing the bottom electrodes for 1 minute at 350°C and atmospheric pressure in a Jetfirst 150 RTO system. The non-stoichiometric oxide has a oxygen vacancy gradient as described in Chapter 5. It is important to note that peak temperature during post-CMOS processing has to be kept below 400°C to prevent damage to the CMOS BEOL metallization. The top electrodes (Pd/Au) are then formed using photolithography and liftoff techniques (Fig. 6-8).

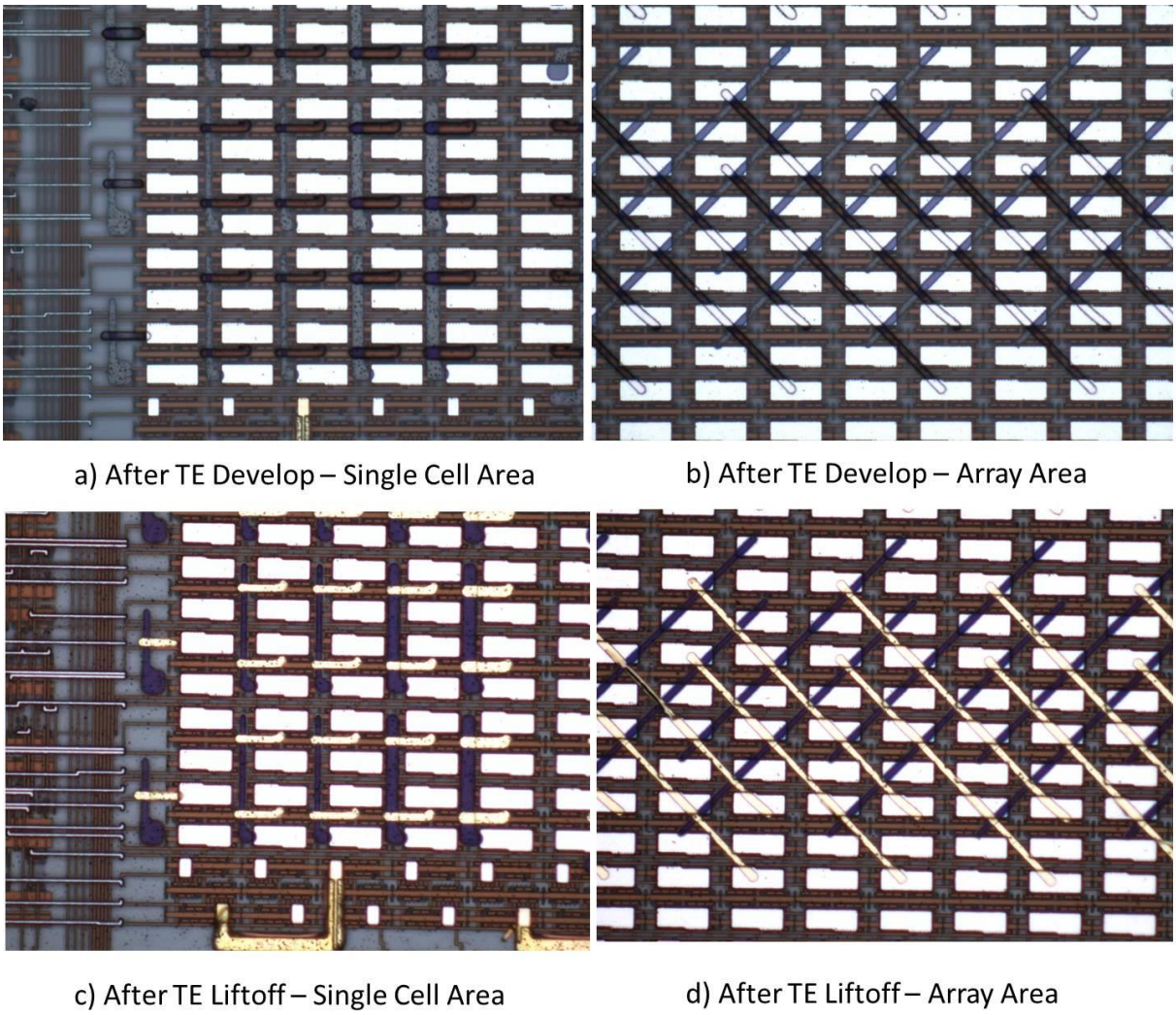


Figure 6-8: Optical micrographs of the CMOS chip after develop step (a,b) and after the Pd liftoff step (c,d). Each CMOL pin is 8µm x 22µm for size reference

Next the WOx formed on the bottom electrodes is etched away using the top electrode as a mask. After the RIE step, WOx exists only at the cross points – between the top electrode and bottom electrode. The RIE step needs to be timed very well to leave the W BE undamaged. Under-etch tends to leave WOx on the W bottom electrodes and results in bad electrical contact later in the process. Too much over-etch, on the other hand, gouges into the W and results in increased line resistance.

The process of connecting the RRAM device to the CMOL pins starts by defining vias in the PECVD oxide using photolithography and RIE as shown in Fig. 6-9. The I/O pads which are covered with PECVD oxide are also opened in this mask/step. Finally the vias are filled with metal (Fig. 6-10) to connect the RRAM devices to the CMOL pins below.

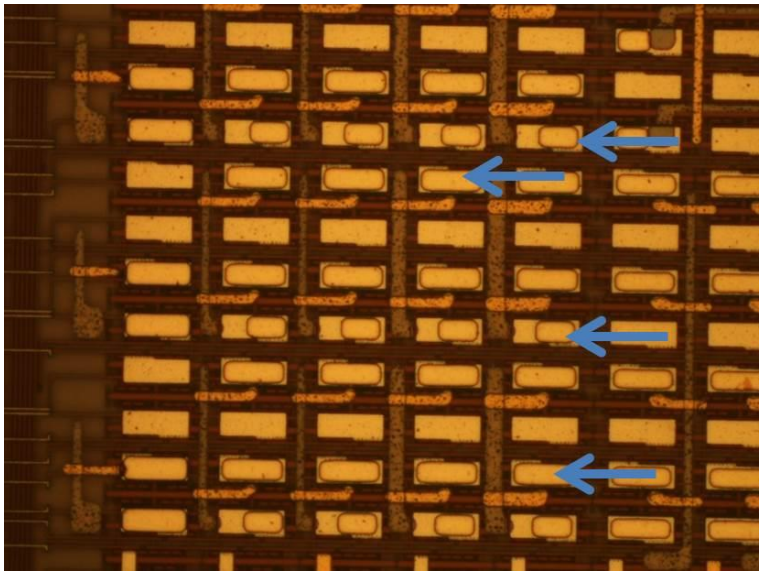


Figure 6-9: Optical micrograph of the vias etched in the PECVD ILD to access the CMOL pins.

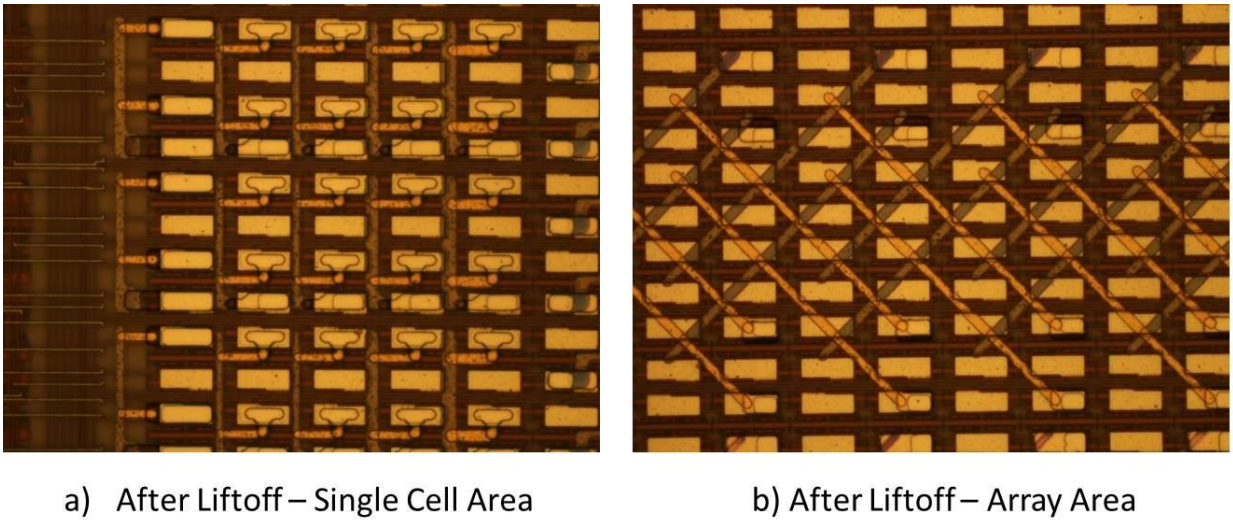


Figure 6-10: Optical scope images of the final metal liftoff step to connect the RRAM devices to the CMOL pins.

A summary of the process flow is included below in Table 6-1.

Sequence	Step	Process
1	Etch down BCB / oxynitride to expose CMOL Pins and I/O Pads	Reactive Ion Etching (RIE)
2	Cover I/O Pads with gold	Photolithography, Ebeam Evaporation, Liftoff
3	PECVD Oxide	Plasma Enhanced Chemical Vapor Deposition(PECVD)
4	Bottom Electrodes	Photlithography, DC Sputtering, Liftoff
5	Oxidize W	Rapid Thermal Oxidation (RTO)
6	Top Electrode	Photolithography, Ebeam Evaporation, Liftoff
7	Etch WOx	RIE
8	Open Vias on CMOL Pins and I/O Pads	Photolithography, RIE
9	Metal routing to connect top and bottom elecrodes to red/blue pins.	Photlithography, DC Sputtering, Liftoff

Table 6-1: Brief description of the process flow for the hybrid CMOS-RRAM test chip.

6.1.2 Electrical Testing

BCB etching end point and good contact to the CMOL pins was confirmed by fabricating (lithography > DC sputtering > liftoff) a small metal line connecting two special CMOL pins – these CMOL pins are different by design and instead of being connected to the CMOS circuitry below, they are connected to the large I/O pads. The advantage of such a test structure lies in the fact that this structure can be tested without powering on the CMOS chip.

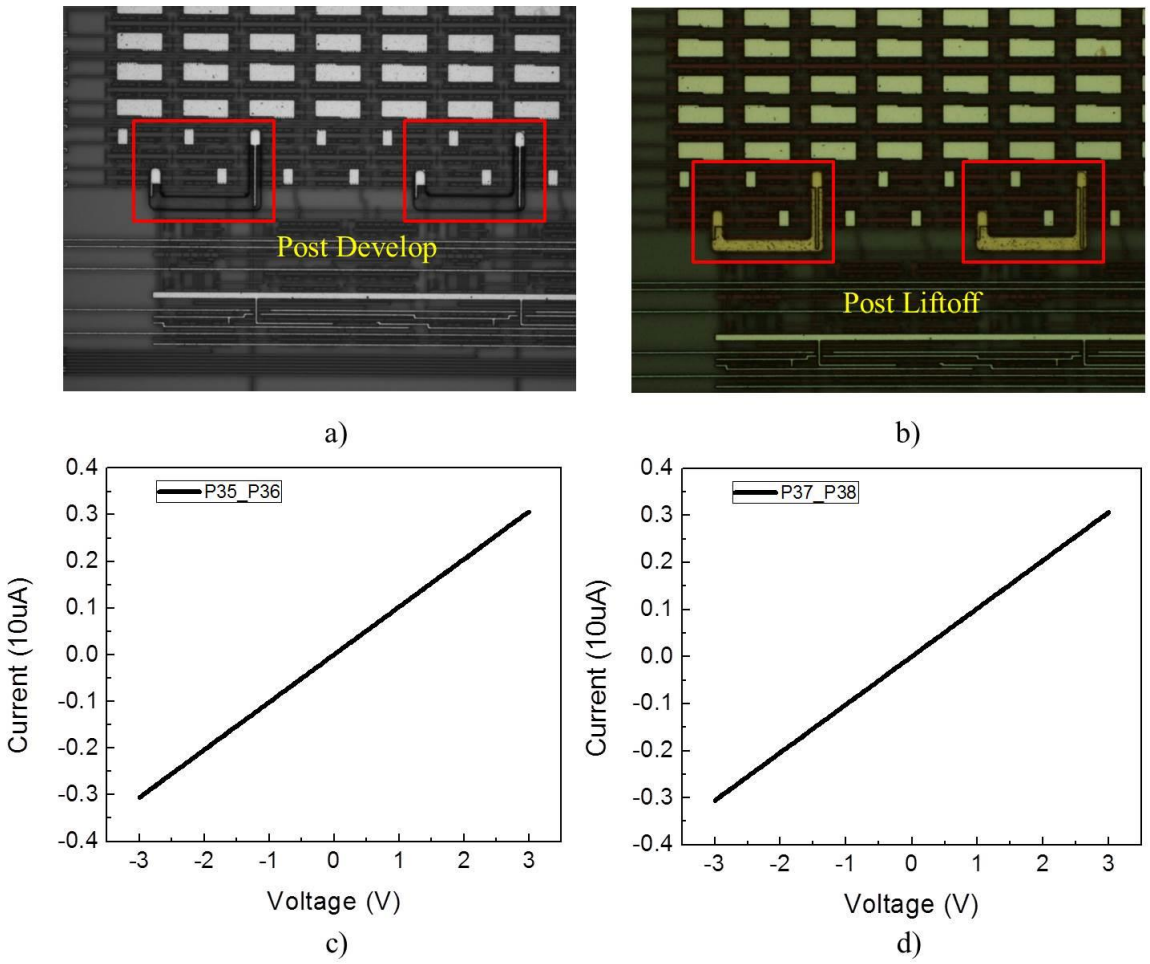


Figure 6-11: Optical image of the etch end point test structure – post develop (a) and post liftoff (b). Good contact to the CMOL pins is indicated by the linear I-V(c and d). Two test structures (I/O pad 35/36 and I/O pad 37/38) show similar results.). A 1M series resistor was used to prevent any damage to the metal lines due to Joule heating.

The hybrid CMOS/RRAM (or the CMOL prototype) system must be powered up to be able to test the integrated devices. Standalone devices connected directly to the I/O pads (similar to the etch test structures) do not work as intended due to the protection ESD diodes connected to each I/O pad. Instead of observing the gradual analog I-V characteristics seen from WOx devices (Chapter 5), we only measure the current through the back-to-back ESD diodes. Since the WO_X devices have much higher resistance than the ESD diodes, almost all of the current flow through the low resistance ESD diodes (Fig. 6-12).

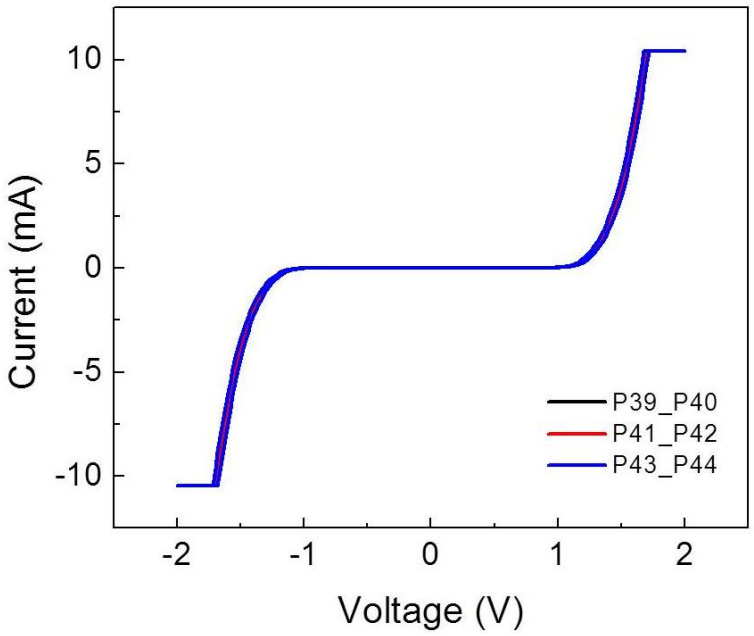


Figure 6-12: Current thorough ESD diode (instead of the current through the test devices) is measured by probing the I/O pads with test devices due to the shunt connection of the ESD diodes.

To test the integrated chip, an integrated test platform has been developed by University of California, Santa Barbara and the chip is being currently tested at UCSB using this integrated platform. In the meanwhile, another revision / tape-out is being prepared. The major changes from the first iteration are as follows:

Larger chip size: Since 2mm x 2mm chip size is too small to handle, the wafer level integration process and BCB need to be used. This adds extra processing steps and fabrication time. A simpler approach to increase the chip size to 4mm x 4mm or 5mmx5mm is being considered.

Thinner final chip: Because of the wafer level processing, the final chip is almost 1mm thick and is incompatible with UM's GCA auto-stepper (CD ~1µm, alignment tolerance ~500nm). Because of this limitation, the MA/BA-6 aligner (CD ~3-4µm, alignment tolerance ~2-3µm) has been used. Using thinner wafers for the wafer level integration process (250µm wafers instead of the 550µm)

would allow the final chip to be used with UM's auto-stepper and would therefore reduce overlay errors.

Signal Routing: In the current design, there is metal routing in the same layer as the CMOL pins (Fig.6-13). This increases the chances of accidental shorting while fabricating the RRAM devices. The possibility of moving the routing to a metal lower below the actual CMOL pins is being explored.

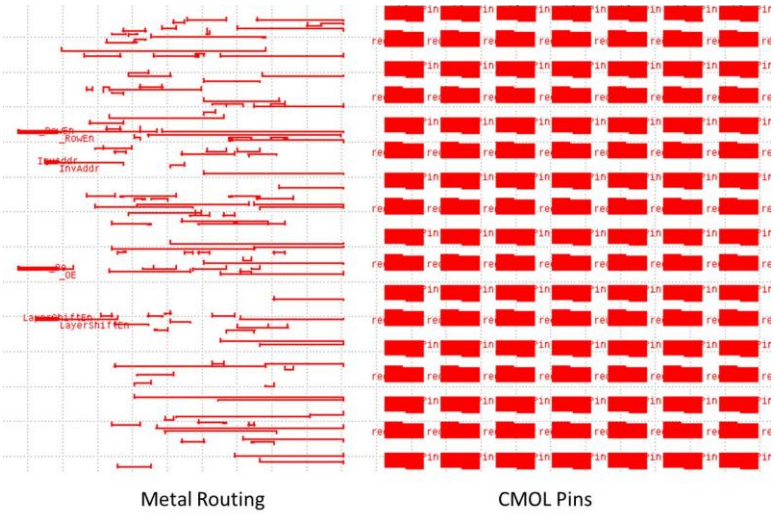


Figure 6-13: Top metal layer is used for defining both the CMOL pins and also for signal routing.

The final goal of the project is to fabricate resistive switching devices on top of this CMOS IC to demonstrate the CMOL concept. The device fabrication is being approached as a multi-tiered task based on increasing complexity, as described below.

1. Fabrication of shorts to test metal contact and to verify functioning of the CMOS circuitry.
2. Single isolated resistive switching devices connected to the CMOL interface.
3. Two layers of crossbars with CMOL interface (Fig 6-14) .
4. Rotated crossbar with CMOL interface.

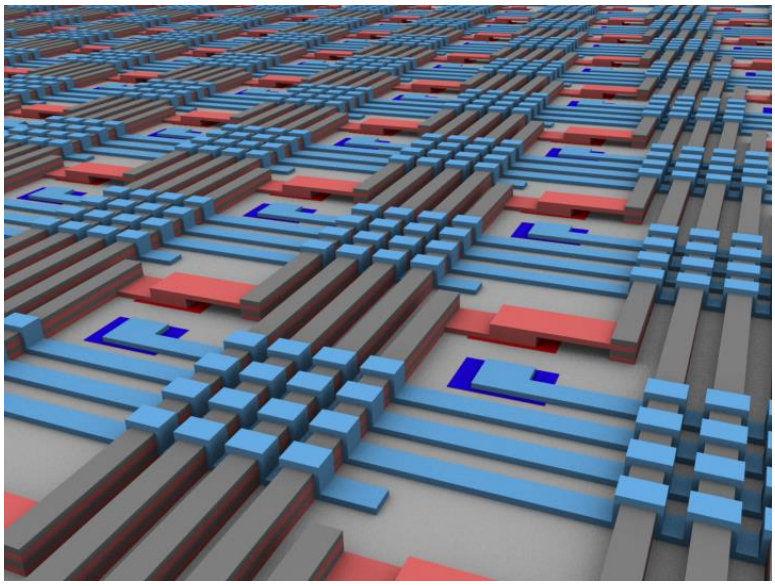


Figure 6-14: Representation of a dual layer crossbar vertically integrated on top of the CMOS pins.

6.2 Achieving self-compliance in devices

The Cu/Al_2O_3/poly-Si devices described in Chapter 4 are very interesting from a device research/optimization perspective but rely heavily on accurate external current compliance, which may not be very easy to implement in a large array system. A 1T-1R structure, where a transistor is placed at each crosspoint [8–11] has the capability to limit the current through the resistive switch but comes with an area penalty. A more device oriented approach would involve inserting a barrier layer between the bottom electrode and the Al2O3 matrix. The barrier layer needs to be chosen such that the metallic filament does not extend into this layer and acts like a filament-stop layer to prevent the filament from shorting out the two electrodes. The barrier layer thus limits the programming current and prevents the formation of very rubust filaments that lead to SA1 failures. On the other hand, the barrier layer also needs to be thin enough so that when the filament is formed in the switching matrix (i.e. device having been programmed into the ON state), sufficient

current can pass through (e.g. via tunneling) during read. Inclusion of such a layer may remove the need for external compliance while programming these devices.

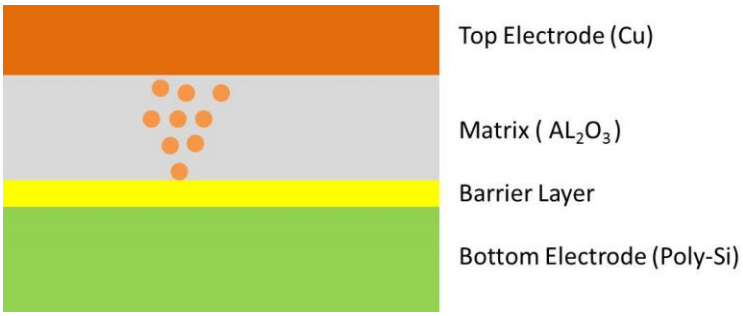


Figure 6-15: Insertion of a barrier layer to prevent over-programming of the device and to achieve self-compliance.

6.3 CMOS compatible ultra-low current devices

The current generation of devices uses poly-silicon as the bottom electrode material. The polysilicon is deposited at 570 °C making the process thermal budget too high for vertical integration with CMOS circuits. For CMOS compatibility, the process temperature needs to be limited to 350-400°C. There are various options to lower the thermal budget. These include the following:

- **Low temperature poly-silicon deposition and dopant activation anneals (<400C)** : Methods to deposit polysilicon at low temperatures are being widely studied, especially by manufacturers of LCD displays [12,13]. While great advances have been made in the deposition and annealing, grain size control and dopant activation are still issues.

- **Replace the polysilicon with metal resistors**: This approach would need very long winding metal lines to obtain sufficiently high resistance values due to the resistivity of most metals. This would affect device density adversely and also increase capacitance drastically. While decreased device density is still not an issue for prototype device level research, increased capacitance would increase the voltage overshoot while programming

devices [14,15] and most likely over-program the devices and result in poor endurance and very high likelihood of stuck-at-1 faults.

- **Replace the polysilicon with metal nitrides like TiN and TaN**: While pure metals have low resistivity numbers, metal nitrides like TiN and TaN can be engineered to have high resistivity[16–19], comparable to the resistivity of the poly-silicon used in our devices. This route is being explored currently using reactive sputtering by varying the nitrogen/argon flow rates and the process parameters (sputter pressure/ power etc.). A baseline TiN recipe (N_2: Ar ~ 6%, 300W RF, 3.5mT, 1200ohm/square) has been established while a TaN recipe is being optimized.

References

[1] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, "'Memristive' switches enable 'stateful' logic operations via material implication.," *Nature*, vol. 464, no. 7290, pp. 873–6, Apr. 2010.

[2] D. Kuzum, R. Jeyasingh, B. Lee, and H. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano Lett.*, vol. 12, no. 5, pp. 2179–2186, 2011.

[3] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems.," *Nano Lett.*, vol. 10, no. 4, pp. 1297–301, Apr. 2010.

[4] D. B. Strukov and K. K. Likharev, "CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices," *Nanotechnology*, vol. 16, no. 6, pp. 888–900, Jun. 2005.

[5] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, G. S. Snider, G. Medeiros-Ribeiro, and R. S. Williams, "Memristor-CMOS hybrid integrated circuits for reconfigurable logic.," *Nano Lett.*, vol. 9, no. 10, pp. 3640–5, Oct. 2009.

[6] D. B. Strukov and K. K. Likharev, "Prospects for terabit-scale nanoelectronic memories," *Nanotechnology*, vol. 16, no. 1, pp. 137–148, Jan. 2005.

[7] A. Uddin, K. Milaninia, C.-H. Chen, and L. Theogarajan, “Wafer Scale Integration of CMOS Chips for Biomedical Applications via Self-Aligned Masking.,” *IEEE Trans. Compon. Packaging. Manuf. Technol.*, vol. 1, no. 12, pp. 1996–2004, Dec. 2011.

[8] S. Sheu, P. Chiang, and W. Lin, “A 5ns fast write multi-level non-volatile 1 k bits rram memory with advance write scheme,” *VLSI Circuits, 2009 Symp.*, pp. 82–83, 2009.

[9] Y. Tseng and C. Huang, “High density and ultra small cell size of contact ReRAM (CR-RAM) in 90nm CMOS logic technology and circuits,” *IEDM Tech. Dig. IEEE Int. Electron Devices Meet. 2009.*, pp. 5.6.1–5.6.4, 2009.

[10] M. Wu, Y. Lin, and W. Jang, “Low-Power and Highly Reliable Multilevel Operation in 1T1R RRAM,” *Electron Device Lett.*, vol. 32, no. 8, pp. 1026–1028, 2011.

[11] S. Koveshnikov and K. Matthews, “Real-time study of switching kinetics in integrated 1T/HfO x 1R RRAM: Intrinsic tunability of set/reset voltage and trade-off with switching time,” *IEDM Tech. Dig. IEEE Int. Electron Devices Meet. 2012*, pp. 486–488, 2012.

[12] J. . Rath, “Low temperature polycrystalline silicon: a review on deposition, physical properties and solar cell applications,” *Sol. Energy Mater. Sol. Cells*, vol. 76, no. 4, pp. 431–487, Apr. 2003.

[13] J. Joo, “Low-temperature polysilicon deposition by ionized magnetron sputtering,” *J. Vac. Sci. Technol. A Vacuum, Surfaces, Film.*, vol. 18, no. 4, p. 2006, 2000.

[14] S. Tirano, L. Perniola, J. Buckley, J. Cluzel, V. Jousseaume, C. Muller, D. Deleruyelle, B. De Salvo, and G. Reimbold, “Accurate analysis of parasitic current overshoot during forming operation in RRAMs,” *Microelectron. Eng.*, vol. 88, no. 7, pp. 1129–1132, Jul. 2011.

[15] Y. Chen, H. Lee, and P. Chen, “Robust High-Resistance State and Improved Endurance of Resistive Memory by Suppression of Current Overshoot,” *IEEE Electron Device Lett.*, vol. 32, no. 11, pp. 1585–1587, 2011.

[16] N. D. Cuong, D.-J. Kim, B.-D. Kang, and S.-G. Yoon, “Effects of Nitrogen Concentration on Structural and Electrical Properties of Titanium Nitride for Thin-Film Resistor Applications,” *Electrochem. Solid-State Lett.*, vol. 9, no. 9, p. G279, 2006.

[17] N. D. Cuong, D.-J. Kim, B.-D. Kang, C. S. Kim, K.-M. Yu, and S.-G. Yoon, “Characterization of Tantalum Nitride Thin Films Deposited on SiO[sub 2]/Si Substrates Using dc Magnetron Sputtering for Thin Film Resistors,” *J. Electrochem. Soc.*, vol. 153, no. 2, p. G164, 2006.

[18] T. Riekkinen, J. Molarius, and T. Laurila, “Reactive sputter deposition and properties of Ta x N thin films,” *Microelectron. Eng.*, vol. 64, pp. 289–297, 2002.

[19] A. Malmros, M. Südow, K. Andersson, and N. Rorsman, “TiN thin film resistors for monolithic microwave integrated circuits,” *J. Vac. Sci. Technol. B Microelectron. Nanom. Struct.*, vol. 28, no. 5, p. 912, 2010.