1 Objective and Significance

The traditional memory technologies, e.g. SRAM, DRAM, and Flash memory, played a very important role in the development of modern computing system and portable multimedia device industries. However, the further scaling at 32nm technology node and below is facing significant technical difficulties, such as large process variations, high leakage power consumption, increased capacitive coupling between adjacent cells, and the device endurance and retention issues [1, 2].

In recent years, significant efforts and resources have been put on the researches and developments of **emerging non-volatile memory (NVM) technologies** that combine attractive features such as scalability, fast read/write, negligible leakage, and non-volatility. Multiple promising candidates, such as Phase-Change RAM (PCRAM), Magnetic RAM (MRAM), Resistive RAM (RRAM), and Memristor, have gained substantial attentions and are being actively pursued by industry [1, 3].

The main objective of this 3-year project is to investigate modeling and design techniques for emerging NVMs in order to enable the massive production and to accelerate the commercialization of these emerging memory technologies. The proposed program makes the following major contributions.

- **Design methodologies for emerging NVM memories:** The device models for emerging NVMs will be developed to fill the gap between process development and circuit design. Memory array design flow and optimization methodologies will be developed to facilitate the design space explorations
- **Circuit techniques for emerging NVM memories:** Various circuit techniques will be proposed to improve the reliability (including lifetime improvement and variation mitigation), yield, and density.
- **Integrated educational plan:** The educational plan will enhance the existing standard curricula by integrating new course modules on emerging NVMs to complement and upgrade the core device and circuit design courses, and bring the awareness of emerging memory technologies into the circuit design and computer architecture community through tutorials and workshops.

The proposed work will initiate a novel research direction in memory design by integrating NVM devices into the standard memory design flow, inventing novel array structure and circuit techniques, and investigating the impact to future computing system. The work will support the deployment of modern microprocessor and embedded system design that use emerging NVM technologies. The proposed research will provide a complementary perspective to the existing computing system research.

2 Background and Related Work

Figure 1 illustrates the fundamentals of the most promising emerging memory technologies to be investigated in our project, namely, the Phase-Change RAM (PCRAM), the Magnetic RAM (MRAM) based on Spin-Torque Transfer RAM(STT-RAM), the resistive RAM (RRAM), and the memristor. In this section, we will briefly describe the physical mechanisms of the emerging NVM devices. The research and development related to this proposal will also be described.

2.1 Phase-Change RAM (PCRAM)

PCRAM technology is based on a chalcogenide alloy (typically, Ge_2–Sb_2–Te_5, GST) material, which is similar to those commonly used in optical storage means (compact discs and digital versatile

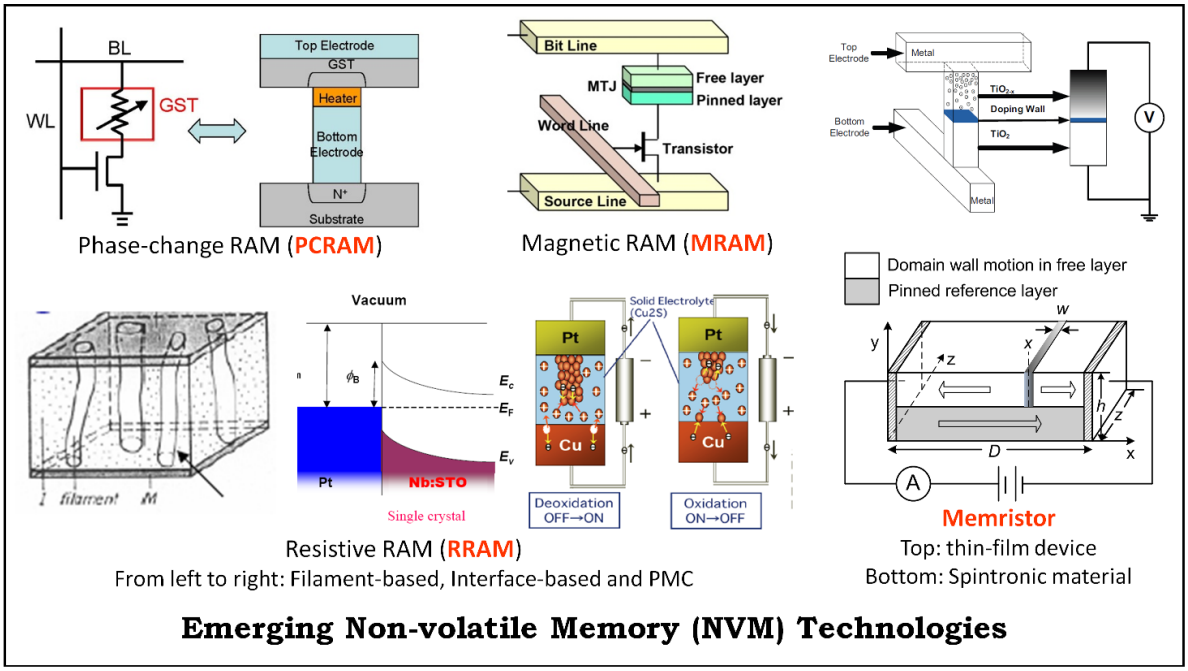

Figure 1: **Overview of Some Emerging Non-volatile Memory Technologies,** including Phase-Change RAM (PCRAM), Magnetic RAM (MRAM), resistive RAM (RRAM), and memristor.

discs) [4]. The data storage capability is achieved from the resistance differences between an amorphous (high-resistance) and a crystalline (low-resistance) phase of the chalcogenide-based material as shown in Figure 1. In SET operation, the phase change material is crystallized by applying an electrical pulse that heats a significant portion of the cell above its crystallization temperature. In RESET operation, a larger electrical current is applied and then abruptly cut off in order to melt and then quench the material, leaving it in the amorphous state [3].

PCRAM has shown to offer compatible integration with CMOS technology [5], fast speed [6], high endurance [7], and inherent scaling of the phase-change process at 22-nm technology node and beyond [8]. Compared to STT-RAM, PCRAM is even denser with an approximate cell area of $6 \sim 12F^2$ [1], where F is the feature size. In addition, phase change material has a key advantage of the excellent scalability within current CMOS fabrication methodology [6, 9–12], with continuous density improvement [13–15].

Although many device models were built from reliability [16], low-frequency noise [17], statistical analysis [18] point of views, they were mainly dedicated to process and device, which cannot be directly borrowed by circuit design and computer community. Many PCRAM prototypes have been demonstrated in the past years by companies like Hitachi [19], Samsung [20], STMicroelectronics [21, 22], and Numonyx [23]. The maximum capacities achieved are 1Gb and 256Mb for single level cell (SLC) [23] and multi-level cell (MLC) [20], respectively. However, to be more competitive to the existing DRAM and Flash memory, PCRAM need further improvement on density and endurance. In this project, we will address this issue from circuit design point of view.

2.2 MRAM based on Spin-Torque Transfer RAM (STT-RAM)

STT-RAM is a new type of Magnetic RAM (MRAM) [1, 24–27], which features non-volatility, fast writing/reading speed (<10ns), high programming endurance (>10^{15}cycles) and zero standby power [1]. The storage capability or programmability of MRAM arises from magnetic tunneling junction (MTJ), in which a thin tunneling dielectric, e.g., MgO., is sandwiched by two ferromagnetic layers, as shown in Figure 1. One ferromagnetic layer ("pinned layer") is designed to have

its magnetization pinned, while the magnetization of the other layer ("free layer") can be flipped by a write event. An MTJ has a low (high) resistance if the magnetizations of the free layer and the pinned layer are parallel (anti-parallel). In first-generation MRAM design, the magnetization of free layer is changed by the current-induced magnetic field [28, 29]. In STT-RAM, a new write mechanism called "polarization-current-induced magnetization switching" is introduced – the magnetization of free layer is flipped by the electrical current directly. Because the current required to switch an MTJ resistance state is proportional to the MTJ cell area, STT-RAM is believed to have a better scaling property than the first-generation MRAM [24, 25, 30–34].

Continuous efforts on process development have been taken on yield improvement [35], write power reduction [36], and high density [37]. Prototyping STT-RAM chips have been demonstrated recently by various companies and research groups [24, 28, 30, 38–40]. Commercial MRAM products have been launched by companies like Everspin (which is a spin-off from Freescale to expedite the technology commercialization in 2008) and NEC.

We have proposed a dynamic MTJ model with more accurate (transient) description for MTJ resistance switching [41]. Compared to highly conceptual fixed resistance used in traditional STT-RAM design flow, the dynamic model can help to reduce 20% pessimism in write time at TSMC $0.13\mu m$. The failure probability of STT-RAM cells due to parameter variations was considered and discussed in [42]. A model was proposed to predict memory yield and design optimization to minimize memory failures. MRAM potentially could be next-generation on-chip cache or memory due to its fast access and soft-error resistance. We will work toward this direction and look for new solutions and more applications to fast this procedure.

2.3 Resistive RAM (RRAM) and Memristor

In an R-RAM cell, the data is stored as two (single-level cell, or SLC) or more resistance states (multi-level cell, or MLC) of the resistive switch device (RSD). Resistive switching in transition metal oxides was discovered in thin NiO film decades ago [43]. From then, a large variety of metal-oxide materials have been verified to have resistive switching characteristics, including TiO_2 [44], NiO_x [45], Cr-doped $SrTiO_3$ [46], PCMO [47], and CMO [48] etc. Based on the storage mechanisms, RRAM materials can be cataloged as filament-based, interface-based, programmable-metallization-cell (PMC), etc. Based on the electrical property of resistive switching, RSDs can be divided into two categories: unipolar or bipolar.

Programmable-metallization-cell (PMC) [49] is a promising bipolar switching technology. Its switching mechanism can be explained as forming or breaking the small metallic "nanowire" by moving the metal ions between two sold metal electrodes. Filament-based RRAM is a typical example of unipolar switching [50] that has been widely investigated. The insulating material between two electrodes can be made conducting through a hopping or tunneling conduction path after the application of a sufficiently high voltage. The data storage could be achieved by breaking (RESET) or reconnecting (SET) the conducting path. Such switching mechanism can in fact be explained with the fourth circuit element, the **memristor** [51–53].

Memristor was predicted by Chua in 1971 [51], based on the completeness of circuit theory. Memristance (M) is a function of charge (q), which depends upon the historic behavior of the current (or voltage) profile [53, 54]. In 2008, the researchers at HP reported the first real device of a memristor in a solid-state thin film two-terminal device by moving the doping front along the device as shown in Figure 1 [52]. Afterwards, magnetic technology provides the other possible methods to build a memristive system [55, 56]. Due to its unique historic characteristic, memristor has very broad application including nonvolatile memory, signal processing, control and learning system etc [57].

Many companies are working on RRAM technology and chip design, including Fujitsu, Sharp,

	SRAM	DRAM	NAND Flash	PC-RAM	STT-RAM	R-RAM & Memristor
Data Retention	N	N	Y	Y	Y	Y
Memory Cell Factor (F^2)	50-120	6-10	2-5	6-12	4-20	<1
Read Time (ns)	1	30	50	20-50	2-20	<50
Write /Erase Time (ns)	1	50	10^6-10^5	50-120	2-20	<100
Number of Rewrites	10^{16}	10^{16}	10^5	10^{10}	10^{15}	10^{15}
Power Read/Write	Low	Low	High	Low	Low	Low
Power (Other than R/W)	Leakage Current	Refresh Power	None	None	None	None

Figure 2: The comparison of various memory technologies [1].

HP lab, Unity Semiconductor Corp., Adesto Technology Inc. (a spin-off from AMD), etc. And in Europe, the research institute IMEC is doing independent research on RRAMs with its partners Samsung Electronics Co. Ltd., Hynix Semiconductor inc., Elpida Inc. and Micron Technology Inc [58]. The main efforts on RRAM research devote to material and devices [44–48]. Many circuit design issues have also been addressed, such as power-supply voltage and current monitoring [59], timing control [60], etc. Unity has been processing 64Kb and 64Mb products and expects to demonstrate 64Gb in 2010 [61]. HP Labs also plan to unveil RRAM prototype chips based on memristor with crossbar arrays soon [62].

Summary Figure 2 illustrates the comparison of emerging memory technologies – PCRAM, MRAM (STT-RAM), RRAM and Memristor – against the traditional main-stream SRAM, DRAM, and NAND-based Flash memory [1]. Note that both CMOS-compatible embedded MRAM (NEC) [63] and embedded PCRAM (Hitachi and STMicro) [19, 64] have been demonstrated, paving the way of integrating these NVMs to the traditional memory hierarchies. In addition, the emerging 3D integration technologies [65, 66] enables cost-effective integration of these NVMs with CMOS logic circuits. With all the NVM technology advances in recent years, it is anticipated that the emerging NVM technologies will break important ground and move closer to market in the near future ("Non-volatile memory goes commercial", EEtimes, 12/02/2009).

3 Proposed Research

To enable the massive production and commercialization of the emerging memory technologies, there are many critical technical issues to be solved. For example, how to introduce the novel devices into the existing design flow? How to minimize the process variation impacts? How to relieve the effect of the poor endurance and improve life time? In this project, we start with the modeling and analysis methodologies for emerging non-volatile memories (NVMs); Next, novel circuitry schemes will be proposed for each emerging NVMs based on their physical characteristics or issues; Our proposed research takes a holistic design perspective with close collaboration between two PIs with complementary expertise, aiming at accelerating the adoption of emerging NVMs for future computer architecture design.

4 Task 1: Design Methodologies for Emerging Memory Technologies

This proposed task focuses on device modeling and design flow and optimization methodologies for memory design using emerging memory technologies.

4.1 Task 1.1: Device Modeling for Emerging Memory Technology

The emerging NVM technologies have introduced many new materials, i.e. phase change material Ge_2–Sb_2–Te_5 in PCRAM and magnetic tunneling junction (MTJ) in MRAM. Due to the lack of

knowledge on physical mechanisms behind, most of researches on circuit, architecture and system levels nowadays are based on highly-simplified characteristics of the emerging devices. This methodology can cause a large design overhead, increase the production cost, and reduce the design margin, especially in the highly scaled technology with large process variations. For example, our previous work [41] showed that after adopting a dynamic MTJ model that can take into account the time-varying electrical inputs in MRAM design flow, the design pessimism can be dramatically minimized and the memory array area can be reduced by more than 40%. Therefore, one of the important tasks of our proposal is to build device models of emerging NVM technologies. Both dedicated device model and simplified behavioral model will be developed to adapt the requirements at different levels.

The dedicated device models, which will be based on physical mechanisms and corroborated by device measurements, need to satisfy three criteria: (1) both static characteristics (i.e., I-V relationship and high/low resistances) and the dynamic behaviors need be considered; (2) the device parameter fluctuations induced by process variations, such as line-edge roughnesses, oxide thickness fluctuations, and random discrete dopants, will be analyzed; and (3) the models should provide enough accuracy with reasonable runtime. To be compatible to commercial EDA tools, i.e., HSPICE from Synopsys [67] and Spectre from Cadence [68], Verilog-A or C language will be utilized to implement these models. The dedicated model will be used for memory array optimization and timing/power analysis.

On top of the dedicated device model, critical timing and function related parameters, i.e. read/write access time and critical switching current, will be extracted and fed into the simplified behavior models. High-level languages, i.e. VHDL/Verilog or C will be used. The simplified conceptual model is expected to provide sufficient accuracy and can be easily integrated in the commercial EDA tools and design methodologies such as *Primetime* and *Timemill* from Synopsys [67] for more thoroughly analysis, *i.e.*, the critical path timing at design corners.

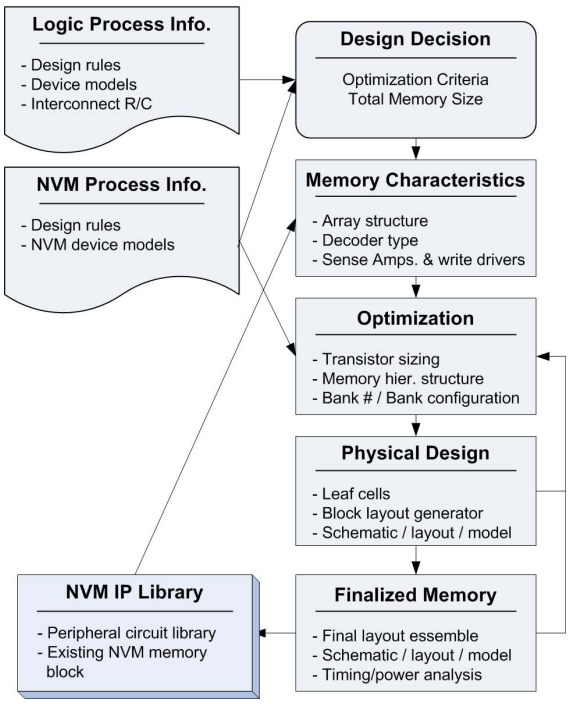

Figure 3: The proposed design methodology for the emerging NVMs.

4.2 Task 1.2: Memory Circuit Design Flow

Another important task of our proposal is to build a design environment that can be seamlessly integrated the emerging NVMs with the existing CMOS logic design flow. Our goal is to use generic memory cells to generate the memory specified by the user. The basic generation methodology and flow is illustrated in Figure 3.

The flow will start working technology specification. Note that the NVM technology could be completely separated from logic process. A high level NVM device model could be used to make initial design decision. The next step is to define the specific characteristics of the desired memory. These characteristics include memory size, array structure, sensing scheme, decoders, etc. Some circuitries can be imported from NVM IP (Intelligence Properties) library to reduce design cycle. Moreover, the constraints for the given application, if any (i.e. energy, delay, area, and noise margin), need to be defined in this stage. Optimization

is the key step within the whole design flow so that he dedicated NVM device model have to be used. It could take several iterations in order to meet the design specifications. The last two steps deal with the actual memory generation. The design will be checked at the end of each step. Re-optimization may be necessary if the specification could not be satisfied.

An IP library for emerging NVM technologies will be built with the aid of the proposed design flow. The library will provide a completed set of IP's, including single non-volatile memory cell, peripheral circuit design, and even the whole NMV design based on general design specifications. Those IP's will be used in the researches at architectural and system levels.

The whole methodology and the corresponding outcomes, including device models, memory design flow, and IP's, will be distributed to the architecture and system design community. Our project will build a channel and provide a friendly interface among material development, device fabrication and architecture design.

4.3 Task 1.3: Energy/Performance/Reliability Design Space Exploration

The physical characters of a NVM cell is mainly depends on the material characters and fabrication process. However, the circuit design of the memory array, such as the access device of the memory cell, the operational voltage, and the pheriphral circuitry, can also impact the operation conditions of the cell, such as current and power consumption. In this task, we propose to explore the design space of the NVM memory array, and study the energy/performance/reliability tradeoffs for memory design with such emerging memory technologies.

Due to the intrinsic non-volatile characteristics of these emerging memory technology, naturally the read and write behaviors are asymmetric in terms of performance and energy. For example, the write-operation of PCRAM/MRAM requires a large current to be applied for a period of time so that the state of the storage junction is flipped; while the read-operation is realized by applying a small voltage to the cell and sensing the current across the cell. We propose to analyze the constrained conditions of he design of NVM memory array, and study the sizing of the transistors as well as the operational voltages to investigate the tradeoff of the distinguishability [1], energy consumption, lifetime and speed. For example, both the width of NMOS access device and word-line voltage can obviously affect the distinguishability. A higher word-line voltage or a larger NMOS device is more desirable to obtain a high distinguishability. However, the power consumption issues and area consideration always require a low voltage and small device size. Another example is on the lifetime/energy/performance tradeoff. For example, the lifetime of PCRAM is represented by the cycling endurance, which is a function of pulse energy applied for the memory cell during the RESET writing. The reason for higher energy pulse induced cycle lifetime degradation is that the RESET resistance can be saturated when the writing current is higher than a critical level. This "over programming" phenomena can result in lager amorphous volume, and then degrade the PCRAMs lifetime. Consequently, a large write current can help improve the performance, but affect both lifetime and energy. Consequently, depending on the application, we plan to investigate two optimization strategies: (1) *Energy-driven optimization.* For low power application, such as mobile computing platform, energy consumption may be the most important design goals. We can optimize the word-line and bit-line voltage as well as the transistor sizing of the access NMOS device for memory array to achieve minimal energy consumption while satisfy constrains on lifetime, performance, and area. (2) *Performance-driven optimization.* For high performance application, We can perform the optimization to achieve the best read/write performance while satisfy constrains on lifetime, energy, and area. Such optimization strategies can also be extended to lifetime-driven optimization and density-driven optimization.

[1] During the read operation, the ratio of high resistance to low resistance in the storage junction reflects the distinguishability between logic 1 and 0

4.4 Preliminary Result and Collaborations:

The PI Li has built a combined magnetic and circuit design analysis and optimization methodology for MRAM, which has been proved to improve design efficiency significantly [41] by test-chip design and fabrication at Seagate. We are also one of the first researchers to propose spintronic memristor structures [56], which was interviewed by IEEE Spectrum [69]. The corresponding compact model and corner analysis [57] have also been developed. In this project, we will further extend this methodology to other emerging NVMs, such as PCRAM.

The PSU PI Xie has developed a stacked SRAM cache simulator called 3DCacti [70, 71], which has been widely downloaded and used by other researchers. The PI and co-PI have collaborated together when the PI Li was in Seagate, to develop a preliminary version of MRAM simulator for cache stacking [66, 72]. Xie also collaborated with Dr. Norm Jouppi from HP Labs, developed a preliminary version of PCRAM simulator [73]. We will extend our existing toolsets to support architectural exploration.

5 Task 2: Circuit Techniques to Improve Reliability, Yield and Density

The advent of novel materials and devices have created many opportunities in circuit design. On one hand, the general requirements to all type of memories are similar – high density, fast speed, low power, affordable yield and reliability, etc. On the other hand, every NVM technology faces different processor integration difficulties, owns unique device characteristic, and targets on different market. Therefore, the primary concerns and the optimal solutions of different NVM chip designs are various. Our task here is to investigate the common design issues and to exploit distinctive circuit techniques for each individual emerging NVM technology. More specifically, we will focus on three main concerns in NVM design – reliability, yield, and density.

5.1 Task 2.1: Reliability Improvement

Reliability is an important parameter in NVMs, which is usually evaluated by data retention and write endurance. While data retention is not a big issue for the emerging NVMs (see Figure 2), write endurance becomes one of the biggest obstacles that prevent from massive production and commercialization. Write endurance is usually measured as the number of writes performed before the cell cannot be programmed reliably. SRAM and DRAM both have endurance of about 10^{16} programming cycles [1], which are sufficient for use even in high-performance processors.

For different emerging NVMs, the physical mechanisms to cause endurance issues are different. In PCRAM, writing is a primary wear mechanism: when injecting current into phase change material, thermal expansion and contraction degrades the electrode-storage contact [74]. Based on a survey of PCRAM device and circuit prototypes published within the last five years, the best reported write endurance for PCRAM is 10^9 [74]. Theoretically MRAM should be able to be programmed $> 10^{15}$ times [1] since its magnetic stack is similar to the one used in hard disk drive. Currently, the best test result of STT-RAM is $< 4 \times 10^{12}$ programming cycles [31] due to the particles and pin-holes introduced in process integration. In parallel to improving material and process development, circuit design techniques can help out in many ways.

Self-contained local control scheme. In general, the damage on NVM material has an ***exponential*** relationship with the current/energy applied on it. And it is an accumulative procedure of total time period. Hence, the most effective approach to improve write endurance is to reduce the write current (I_{wr}) and write operation period (t_{wr}).

For example, one possible solution is smoothing I_{wr} shape during write operations and avoiding overshot on NVM materials. Accordingly, how to design a write driver to provide a sleek but fast ramp-up curve is the tricky part. Another interesting alternative could be lowering the voltage

on memory device to meet only the minimal required current. Obviously, an accurate self-timing control scheme is necessary, which can stop providing writing current to memory cells once detecting successful programming operations. On one hand, we observe that a longer t_{wr} is needed when a smaller I_{wr} is provided. t_{wr} could be very sensitive to I_{wr}, for example, $t_{wr} \propto -I_{wr}$ in MRAM. On the other hand, the process variations at nano-scale technology node, including variations of both CMOS devices and memory elements, make it very hard to control I_{wr} precisely.

Hence, we propose to add a self-contained local control scheme. The scheme is ***self-contained*** because it is mainly composed of a number of memory cells with the same emerging NVM device. These cells are divided into three functional groups used for configuration, detection and control, respectively. The initial configuration should be programmed at testing stage before the chip shipping out. During write operations, the detection cells are also programmed and the degradation extent of these cells can be used to predict the status of memory cells in the main array. The prediction result will be fed into control schemes periodically to adjust I_{wr} and t_{wr} on the fly. When needed, the control signals can even be used at system level, i.e. the bit-redundancy and ECC algorithm. The granularity of the self-contained local control scheme depends on each specific NVM technology and application requirement.

Circuit techniques to improve endurance. Architectural level techniques have been proposed to mitigate the endurance problems in PCRAM, such as *Read-before-Write* or *White Cancellation* [74, 75]. In this task, we plan to investigate circuit-level techniques to enable wear leveling for NVM memories. Specifically, we plan to study two wear leveling techniques: 1) *Bit-line Shifting:* Usually, write-operations in applications demonstrate an extremely uneven distribution in a memory block. Consequently, a few hot memory cells are worn out much faster than other cells, making the entire memory block useless even though most of the cells are still functional. Based on such observation, we plan to design a bit-line shifter to spread out the writes over all memory cells in a memory block. In our design, each memory block has its own shift offset register (SOR) and shift interval counter (SIC). An SOR stores the shift offset of its cache block, and the bit-line shifter refers to it to determine how many bits to shift the incoming data before store it to the cache block. An SIC records the number of writes performed to its cache block, and when it reaches to a predefined threshold, we update the corresponding SOR value. Hence, we can achieve the balance between the wear-leveling performance and the additional bit changes caused by changing shift offsets. 2) *Word-line Remapping:* Similar to Bit-line shifter, we can also use a word-line remapper to spread out writes over all cache blocks. The word-line remapper consists of an adder and a word-line SOR to keep the current word-line shift offset, through which a word-line decoder logic can determine what word-line should be enabled. After changing the value of the word-line SOR, we should invalidate the contents in the cache because the word-line mapping is changed. Therefore, the word-line remapping period should be long enough (e.g., a second) to avoid excessive remapping overhead. These two circuit-level modification to the memory array will be evaluated on the effectiveness of the endurance improvement, as well as the performance/area/energy overhead, and will be compared against architectural level techniques such as those proposed in [74, 75].

Multi-level cell (MLC) write endurance. Multi-level cell (MLC) can effectively improve the integration density of memory by storing more than one bit information in a single memory device: n bits are represented by 2^n states of a storage device. MLC technology has achieved significant commercial success in NAND flash memory [76] and it has been explored in PCRAM [4, 12], STT-RAM [37], and RRAM [77]. It can effectively improve the integration density of memory. However, the write endurance is degraded significantly due to the smaller resistance gap between two adjacent states. In the project, we will seek optimal solutions to improve MLC write endurance by considering write patterns of both physical mechanism and system requirement.

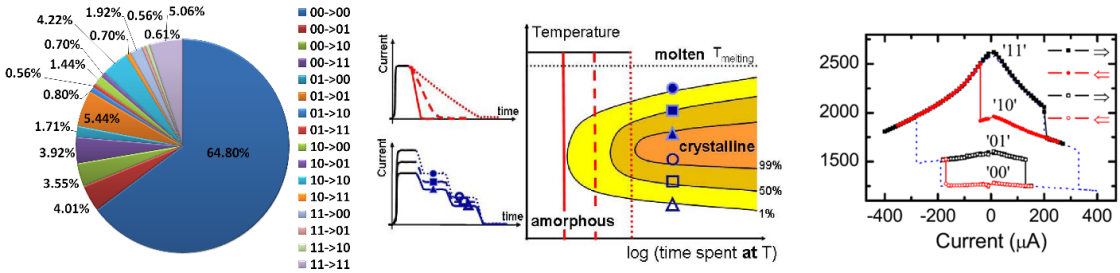

Figure 4: MLC write patterns. Left: The transition distribution between the different values of MLC MRAM bit; Middle: PCRAM – Schematic plot of time-temperature-transformation-chart [13]; Right: MRAM R-I sweep curve [37].

Let's use a 2-bit MLC as an example. Each memory cell can represent four logic states, namely $L00$, $L01$, $L10$, and $L11$. Figure 4 shows the transition distribution between the different logic values in an in-order microarchitecture. We noticed that most of transitions occur between the same values, and hence, there is no need to change resistance state at all. The observation is aloes true for most of embedded applications. Therefore, "write-after-read" scheme, which conducts only the necessary transitions based on the values of the new data being written and the original data stored in the MLC bit, could be the most efficient way for energy saving and lifetime improvement. However, the extra read in write operation introduce performance overhead, can it be absorbed or minimized? Can we detect the data in memory cell at the same time as cell is programmed and terminate the writing earlier? These questions will be discussed in the project.

Furthermore, let's name the four resistance states of a 2-bit MLC are $R00$, $R01$, $R10$, and $R11$ from low to high. We noticed that switching to different resistance state in an MLC need follow specific sequence and/or demand different write current as shown in Figure 4. For example, the multiple resistances in PCRAM are achieved by different size and shape of the amorphous region at the top of the pillar-heater within the phase change material. Hence, the target resistance strongly depends on temperature and time during write operations. An MRAM MLC has two free layers whose magnetization directions can be switched separately. Therefore, two-step writing – a large current switching followed by a low current one – is required.

Corresponding to the four resistance states, an MLC cell has total of $4! = 24$ encoding schemes for its four logic states. As we stated above that the breakdown probability of a NVM cell has an exponential relationship with the current amplitude through it, the damage to memory material has different weight when writing different data. Properly selecting the encoding scheme of logic vs. physical states based on the transition distributions can further improve the write endurance and lifetime of MLC NVM technologies.

5.2 Task 2.2: Yield Enhancement

Higher defect rates and low yield are brought by the continuous shrinking of devices and the unlimited demand on higher densities. As technology enters into nanometer scale, device parameter fluctuations induced by process variations, such as line-edge roughnesses (LERs) and oxide thickness fluctuations (OTFs) have become critical issues [78]. Emerging non-volatile memories, which are among the densest circuits in systems, are greatly impacted by the large process variations. For example, MTJ resistance in MRAM increases exponentially with the thickness of oxide barrier between two magnetic layers. It was reported in [79] that MTJ resistance increases by 8% when the thickness of oxide barrier changes from $14\mathring{A}$ to $14.1\mathring{A}$. In the program, we propose to overcome the impact of process variations and to enhance yield with the aid of the unique device characteristics of NVMs.

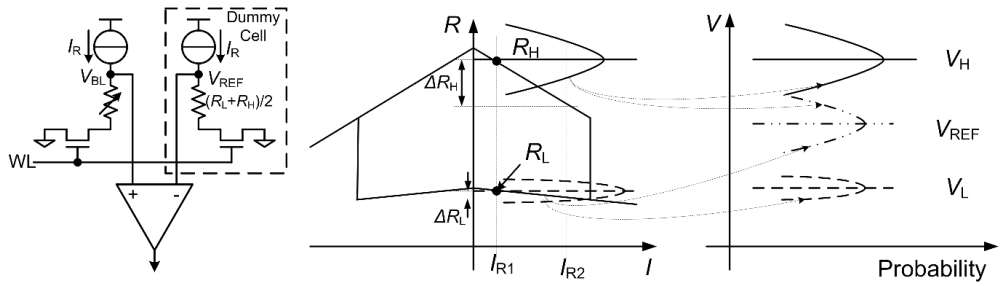

Figure 5: Left: Conventional read-out scheme of MRAM; Middle: R-I characteristic of MgO-based MTJ; Right: MTJ resistance distribution incurred read failure [42].

Non-destructive self-reference technology in MRAM. Like most of the emerging NVM technologies, MRAM uses device resistance as the data storage media. Figure 5 shows a conventional voltage sensing scheme, which compares the bit line voltage V_{BL} generated by the selected memory cell with a reference signal V_{REF} produced by the dummy cell. And a dummy cell is shared by multiple memory cells to reduce overhead. Ideally the resistance of the dummy cell should be set in the middle of the high and low resistance states (R_H and R_L). In reality, process variation incurs the resistance distribution of MTJ in memory cells as well as the dummy cells. When the resistance variation σ_R is large, the tails of R_H or/and R_L could be overlapped with R_{dummy} and lead to the false detection of the stored value as illustrated in Figure 5. We called it as ***Read Failure***.

Read failure is a severe problem in STT-RAM design for two main constraints: (1) The difference between two resistance states of MTJ is fairly small: ΔR=$R_H - R_L \approx 1000\Omega$ at 45nm technology node [42]; and (2) the MTJ resistance variation σ_R is relatively high because it is extremely difficult to control oxide barrier thickness within a small range of variation, i.e. $0.5\mathring{A}$ [80]. Besides the regular yield improvement techniques, such as redundant column/row and ECC (Error Correction Code), a self-reference read-out scheme could be another effective way to fix read-failure problem.

The basic idea of a self-reference reading is to compare the stored data in a memory cell with a reference value written to the same cell. By limiting the comparison within one single STT-RAM cell, the impact of bit-to-bit variation of MTJ resistance can be avoided. Previously some self-reference schemes were used in toggle-mode MRAM design [25,80]. We also successfully utilized it in STT-RAM design [81]. These schemes are all "destructive" because the original value in memory cell is wiped out when writing the reference value into MTJ, and has to be recovered at the end of the read operation. Obviously it prolongs read latency and aggravate reliability issue.

In this project, we will work on a ***non-destructive self-reference*** methodology, which does not need disturb the original data during read operations. The approach comes from the special R-I characteristic of MgO-based MTJ. As we can see in Figure 5, the MTJ current dependence of R_H and R_L are quite different: the current roll-off slope of R_H is much steeper than that of R_L. Therefore, we can sample the stored value of an MTJ twice by using two read currents I_{R1} and I_{R2} and compare the resistance difference $\Delta R = R1 - R2$. Obviously ΔR_H is pretty big, while ΔR_L is close to '0'. There are some uncertainties to realize this approach. For example, how much is the sensing margin in the new read-out scheme after considering process variations? What type of sensing circuitry is more optimal? Will a new sense-amplifier (SA) design be necessary? How does it impact memory array structure? How much yield improvement can be achieved with the new scheme? Will this scheme be still valid when technology further scales down? In this proposal, we will investigate these issues and exploring the solutions. Our target is to minimize the effect of process variation and to improve read speed.

Resistance drift. Resistance drift has been observed in both PCRAM and memristor-based memory. In PCRAM, especially multi-level memory, the amorphous phase (and other phases obtained by incomplete phase transition) is metastable and can experience structural relaxation [82], which results in resistance drift over the time. For a memristor-based memory, if the read operation cannot provide zero flux, the resistance could "drift" to one direction continuously due to the accumulative effect of the input flux [83]. Resistance drift can increase resistance variation σ_R and hence, spread out the resistance distribution. Memory access patterns (i.e. the resistance state stored in the cell, read/write access frequency and interval, etc) strongly impact the resistance drift. On top of the process variations, the resistance drifts make the design margin even smaller, which aggravates read failure and further hurt chip yield.

The resistance drift in memristor-based RRAM design is mainly impacted by read access frequency. hence, Ho et. al. proposed a refreshing scheme [83]: after a particular number of reading operation, a refresh operation is needed to eliminate the effect of read pulse mismatch. Our proposal is to flip the read current direction whenever accessing the memristor. The hard part of this proposal is to decide the granularity. The most effective way obviously is to implement it at block level. However, it will cause performance overhead in order to get the information of previous read current direction. On the contrary, a coarse granularity could reduce the overhead, but it won't help in the worst situation.

The resistance drift in multi-level PCRAM has different behavior pattern from memristor-based RRAM. It is determined only by the state stored in the cell and interval to previous write, no matter if the chip is powered up or not. Xu and Zhang [84] proposed to solve the resistance drift in PCRAM with a complex ECC, which can minimize the error rate to a sufficient low level during its whole lifetime. The drawback of the scheme is the large overhead on performance and circuit complexity. Instead, we propose to build a self-adjustable sensing scheme, in which the reference current/voltage can be self-adjusted to follow the trend of resistance drift. Phase change material or memristor-based material could be used to monitor the device degradation. Similar to our proposal for memristor-based RRAM, properly selecting the granularity and cover the worst-case situation is the key issue to be solved.

5.3 Task 2.3: High Density

Increasing memory density is an ultimate goal in memory design. In the past, technology scaling is always the biggest driving force to reduce single cell size. Process development plays an important role as well. For example, to continue scalability, the charge storage materials of NAND Flash have gone through several generations: from standard double polysilicon gate, to SONOS, to bandgap engineered SONOS, and to TaNOS [85]. Circuit design techniques also plan an important role to boost memory density, i.e. word-line overdrive scheme can help reduce select transistor size in memory cell [42]. In the project, we propose to introduce novel devices into emerging NVM design and investigate the corresponding design techniques.

Double cell configuration in bipolar switching RRAM. In RRAM design, crossbar structure is widely investigated. In each cell, only two terminals are needed – one is horizonal (WL) and another is vertical (BL). The storage element is built at the cross-point of two metal wires. A diode could be used as selective element in unipolar switching RRAM. When RRAM material is bipolar switching, a non-ohmic device (NOD) [86] is needed to provide two-direction driving current and to support process integration of cross-point structure. We call it as 1NOD-1R as shown in Figure 6. Data access is supported by properly controlling the voltages applied on WL's and BL's. Theoretically, crossbar structure has the smallest memory cell area $4F^2$. Here, F represents the technology feature size.

Due to process limitation, 1NOD-1R cell structures are facing some design difficulties. Concep-

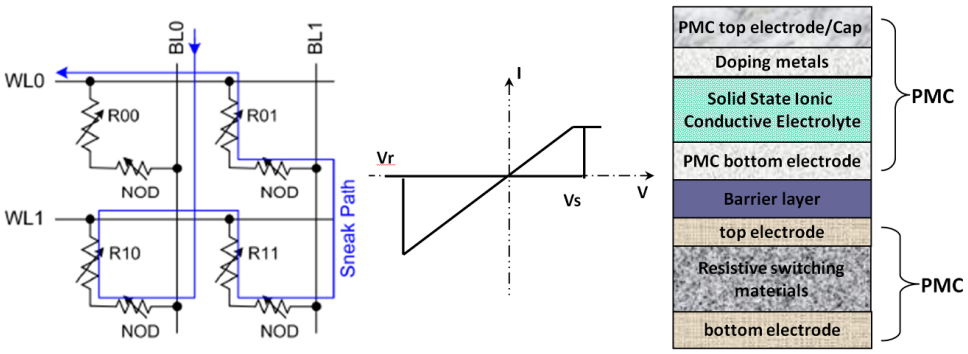

Figure 6: Left: 1NOD-1R and sneak path; Middle: R-I characteristic of PMC; Right: double cell configuration by stacking two cell back to back with a barrier layer in between.

tually, NOD can be understood as two parallel connected diodes. Ideally, it turns on only when the voltage drop between the two terminals exceeds its threshold. However, the I-V characteristic curve of real device could be quite different. This results in sneak path which has three or more cells in series as shown in Figure 6. The sneak current can introduce disturbance on unintended cells during read, write and erase operations.

Here, we propose to build a double cell configuration in bipolar switching RRAM, i.e. PMC [49], by stacking two cells back to back with a barrier layer in between. Figure 6 illustrates such a double cell with PMC material as example: The PMC with active electrode on top is used as data storage (RRAM), while the main function of the PMC with active electrode at bottom is the selective device. The functionality of the proposed structure is guaranteed by the asymmetric I-V characteristic of PMC – a higher voltage is required in RESET operation (V_r) than the one in SET operation (V_s). A large I_{on}/I_{off} ratio (large turn-on current I_{on} and extremely small turn-off current I_{off}) due to the big different between the high and low resistance states minimizes the disturbance from sneak current. The integration of the proposed structure is friendly to CMOS technology since RRAM technology itself is compatible to CMOS process.

To make this structure feasible, there are still a lot of unsolved issues. For example, what is the proper timing sequence during read and write operations? What is the optimal peripheral circuitry floorplan for the proposed RRAM design? What is the impact of process variations? Furthermore, can we expand this structure to other RRAM and memristor-based memory design? In this project, we will address these circuit issues from both device and circuit point of views and explore the solutions.

Peripheral circuitry density improvement. RRAM crossbar structure can also grow in third dimension, which is called intra-die stacking. The memory storage cell is located in between any two adjacent metal layers which are used as interconnects. Within the same die size, the multiple memory layers further improve the memory density. If we still use the bottom layer as the logic layer and connect upper layers through via holes, the peripheral circuit area will increase several times, and hence decrease the high-density effect as a result of the intra-die stacking structure. Therefore, a 3-D stacking peripheral circuitry become necessary in order to increase array efficacy.

Previously, Song et. al. showed GIZO thin film transistors (TFTs) can be stacked vertically and might be used in intra-die memory [87]. However, their work still remained at material level. In this project, we will further explore it in device and circuit design level. For example, the initial proposal was to use TFTs as word-line (WL) selection transistor only. It does not reduce the number of via holes from bottom layer to memory layer because WL's has the identical number as the WL drivers. Our plan is expanding the usage of TFTs to some simple logic functions. For example, we can use it to implement the last stage of WL decoder. In this case, only signals from

pre-decoding schemes need to be fed into each memory layer. Hence, the functional blocks on the logic layer and the via holes between layers can be dramatically reduced. Many design techniques will be involved: How to simplify the circuit with NMOS-type only logic? Can this design fit into the small pitch of crossbar array? How to distribute the power supplies in the stacking logic structure? Depending on the process development status of TFT technology, we even expect to use to implement more complex functionalities and further improve the density of peripheral circuitry.

5.4 Preliminary Results and Collaborations:

The PI Li has worked on memory design for years, from traditional SRAM to current emerging NVMs. Her research on low-power SRAM design [88–90]has been used in Intel processor. In Seagate, she has led a design team on MRAM and RRAM test-chip design. Many circuit design techniques for NVM design have be invented, including the destructive self-reference scheme in STT-RAM design [81], feasibility of the non-destructive self-reference scheme [91], memory array density improve with new array structure [92, 93] or peripheral circuit [94, 95], yield improvement by using hybrid ECC scheme [96] or defective bit [97], etc.

Yuan: Will you add anything here?

6 Broader Impacts, Outreach, and Education

Research Impact and Technical Merit: Memory design is one of the key components in modern VLSI ICs and microprocessors. The importance of the memory hierarchy increases with the advances in performance of the microprocessors [1]. A key *transformative aspect* of the proposed research is that the success of the project will result in innovations in the modern microprocessor design, potentially leading to better performance, higher energy-efficient, and more reliable computer systems.

Collaborations and Partnership: It is naturally important to have industry support and guidance for this research. The NYU-Poly PI Li has been with industry for 5 years before joining academia. She has a strong connection with Memory Product Group at Seagate, where she did research and led a design team on nonvolatile memories. The PSU PI Xie worked for IBM Microelectronics division before joining academia, and has built a good relationship with IBM research. In the past 6 years as a faculty member, Xie has close collaborations with industry partners. The proposed research has intrigued our industry partners, and the project will be carried out with close collaboration with partners in several companies, including IBM, Intel, HP, IMEC, Qualcomm, and Seagate. The investigators anticipate that the techniques and tools developed in this project will be used in both classroom projects and academic/industrial research. We will closely work with our industry partners to transfer research results into commercial designs. The proposed technology is of immense interest for companies.

Outreach and Knowledge Dissemination: As part of outreach efforts, the PIs will actively disseminate results to a wide audience and to different professional communities. The NYU-Poly PI Li believes that the communication between academia and industry is very important. In NANOARCH 2009, she organized a panel on Emerging Technologies, which brought industrial voices into emerging NVM research. The PSU PI Xie has delivered over 30 invited talks in the past at IEEE Chapters, universities, and companies. He has been a tutorial speaker at several forums, offering tutorials on 3D ICs in MICRO 2006, ISCA 2008, GLSVLSI 2008, and MICRO 2009 [98]. Penn State is part of the University-Industry-Government partnership called The Technology Collaborative (TTC) that focuses on research, training and education issues related with system design. The PI from Penn State has been actively involved with their education programs and have offered courses to the local industry in the past through TTC. We will use this forum to

disseminate findings of the proposed research to industry practitioners, who in turn can facilitate technology transition and incorporate research breakthroughs in real systems.

Women and Minority Student Recruiting Activities While this research program will make contributions in educating all students to be well prepared for designing future computer systems, it will make additional efforts to promote diversity. Being a woman faculty herself, the NYU-Poly PI Li plans to actively recruit and mentor women and minority students. The PSU PI has an impressive record of graduate student advising, especially those from underrepresented groups, having graduated several women and minority graduate students. The PIs will continue to attract underrepresented students by getting their current graduate students from underrepresented communities to present their research at minority undergraduate institutions and to serve as role models. The PIs have been working with women and minority recruiting programs in both universities, i.e., the Multicultural Education and Programs at NYU-Poly and the WISER (Women in Science and Engineering Research) and MURE (Minority Undergraduate Research Experience) programs at PSU.

Integration with Education: This project will involve graduate and undergraduate students in all aspects of the research. The PIs, as in the past, will actively integrate the research results from this project into the graduate and undergraduate curricula, especially related to computer architecture. The NYU-Poly PI teaches a graduate-level course EL5473 (Introduction to VLSI), and this project will allow the PIs to integrate additional practical material to make the class more appealing for engineering students. A graduate-level course on advanced topics in computer architecture will be developed at NYU-Poly in collaboration with colleagues who are experts in architecture and circuit design. Undergraduate students will be especially targeted and encouraged to pursue graduate studies. Support for undergraduate researchers will also be sought from NSF REU supplements and by involving the outstanding students from the Schreyers Honors program at Penn State. Beyond involving students in all aspects of research, the PIs will develop new courses on different aspects of advanced computer architecture and VLSI, to train the next generation work-force. In addition, the PIs plans to organize workshops and tutorials at major conferences to support other faculty to adapt new teaching and research material in their curricula. Class notes, slides, and laboratory manuals related to the new courses developed will be made publicly available. The PIs will educate industrial practitioners and use this grant to disseminate findings to industry practitioners, who in turn can facilitate technology transition and incorporate research breakthroughs in real systems.

Collaborative Teaching Experiments: A graduate-level course on emerging non-volatile memories will be simultaneously offered at Penn State and NYU-Poly (in a *virtual classroom*) through an online course delivery system (WebEx). Lectures will originate from both schools based on the topics to be covered. The PIs will incorporate the latest research outcomes from this project. Students at PSU and NYU-Poly will also experiment with the tools developed as a part of this research. This multi-institution education plan will not only provide a unique opportunity for students to learn from experts in other universities/areas but also promote collaborations among students in different schools through working together on course projects. Such remote collaboration is a critical skill in today's global economy, where many companies have offices throughout the world.

7 Project Management and Industry Collaborations

The research team poses complementary skills required for the project. The PIs are well qualified for the proposed research with significant prior experience in various areas. The PI Prof. Li has 5-years industrial experience related to device modeling and circuit design with focus on emerging

non-volatile memories, and just recently joined NYU-Poly as an assistant professor. The co-PI Prof. Xie's expertise span areas of VLSI and architecture, with extensive experience in architectures with emerging technologies, such as 3D architecture. The PIs will work in close coordination on different parts of this project. The integration of all these research components and tool will be a coordinated effort by all the investigators.

The project is a three-year effort involving multiple PhD students. Li will lead the effort in the first year with 2 PhD student from NYU with 1 PhD student from PSU on device modeling and build memory design flow in Task 1. From the second year, both PIs will work together with 1 PhD from each institute. While Xie continue exploring energy/performance/reliability design space in non-volatile memories, Li will focus on circuit techniques to improve NVM reliability with the aid of design flow built in first year. The yield and density enhancements will be the emphasis in the final year, which will be led by Xie. Detailed project milestones are given in Figure 7.

The PIs have a well-established collaboration in the past years, when the PI was still in Seagate, and published preliminary results on NVM architectures in DAC 2008 and HPCA 2009 [66, 72]. The existing collaboration and preliminary results will allow rapid ramp-up for the proposed research. The two teams will coordinate with each other via weekly teleconferences and regular mutual visits (with only 4-hour driving between two institutes).

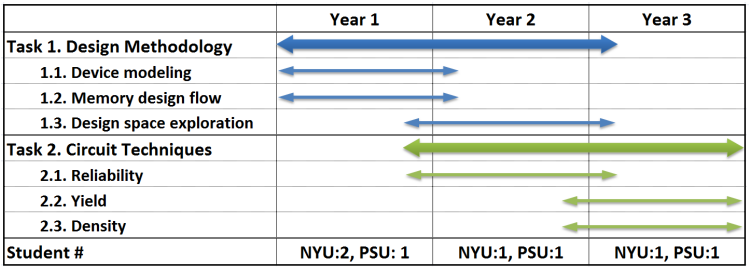

Figure 7: Project Management.

Industry Collaborations. By leveraging both PI's past industry experience and successful collaborations with companies, the project will be carried out in close collaboration with industrial partners from IBM, HP, Intel, Qualcomm, Seagate, ITRI, as well as with a partner from IMEC in Belgium. The industrial collaborators will play important roles in the proposed project by enabling the acquisition of realistic data, discussion of the practicality of ideas, placement of students in internships and permanent positions, and eventually the transfer of the technologies. By working closely with researchers in industry, the PIs will be able to ensure that the proposed methodologies and tools are practical and have a real impact on industry.

8 Results from Prior NSF Support

Hai (Helen) Li recently just joined NYU-Poly as an assistant professor, after 5-years industrial experience in Qualcomm, Intel, and Seagate. She doesn't have any NSF grant yet.

Yuan Xie: The most related prior NSF grant is CCF-0903432 (ADAM: Architecture and Design Automation for 3D Multi-core Systems; 08/2009-07/2012; $480K). This project aims at developing architectural design techniques and design automation tools for future 3D multi-core architectures. Xie actively collaborates with industry in 3D IC design research (IBM, Qualcomm, Honda, and Seagate). He has published extensively in the 3D IC design and 3D architecture areas, covering various aspects, including 3D architecture [66,70,72,99–105] and 3D EDA tools [65,70,71,106–110].

One of the benefits for 3D integration technologies is the capability of enabling cost-effective heterogeneous integration, which makes it much more practical to integrate emerging NVM with CMOS logic circuits. Consequently, the research plan described in this proposal will complement and be synergistic with the ongoing project.

The PIs have also submitted another proposal titled "Collaborative Research:SHF:Small:Modeling, Architecture and Application for Emerging Memory Technologies" to NSF-CISE-CCF-SHF pro-

gram recently (December 2009), with a focused scope of computer architecture research. The research topics work proposed in this proposal is circuit-oriented, and will complement and be synergistic with the other pending proposal.

REFERENCES CITED

References

[1] International Technology Roadmap for Semiconductor, 2007. http://www.itrs.net/.

[2] K. Kinam and J. Gitae. Memory technologies for sub-40nm node. In *IEEE International Electron Devices Meeting (IEDM)*, pages 27–30, 2007.

[3] G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, and R. S. Shenoty. Overview of candidate device technologies for storage-class memory. *IBM Journal Research and Device*, 52(4/5):449–464, 2008.

[4] Ferdinando Bedeschi, Rich Fackenthal, Claudio Resta, Enzo Michele Donzand Meenatchi Jagasivamani, Egidio Cassiodoro Buda, Fabio Pellizzer, David W. Chow, Alessandro Cabrini, Giacomo Matteo Angelo Calvi, Roberto Faravelli, Andrea Fantini, Guido Torelli, Duane Mills, Roberto Gastaldi, and Giulio Casagrande. A bipolar-selected phase change memory featuring multi-level cell storage. *IEEE Journal of Solid-State Circuits*, 44(1):217–227, 2009.

[5] J. H. Oh, J. H. Park, Y. S. Lim, H. S. Lim, Y. T. Oh, J. S. Kim, J. M. Shin, and et al. Full integration of highly manufacturable 512Mb PRAM based on 90nm technology. In *Proceedings of the IEEE International Electron Devices Meeting*, pages 2.6.1–2.6.4, 2006.

[6] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, , and R. Bez. Scaling analysis of phase-change memory technology. In *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, pages 29.6.1–29.6.4, 2003.

[7] S. Lai. Current status of the phase change memory and its future. In *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, pages 10.1.1–10.1.4, 2003.

[8] Y. C. Chen, C. T. Rettner, S. Raoux, G. W. Burr, S. H. Chen, R. M. Shelby, M. Salinga, and et al. Ultra-thin phase-change bridge memory device using GeSb. In *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, pages 30.3.1–30.3.4, 2006.

[9] S. L. Cho et al. Highly scalable on-axis confined cell structure for high density PRAM beyond 256Mb. In *Symposium on VLSI Technology Digest of Technical Papers*, pages 96–97, 2005.

[10] S. Kim and H.-S. P. Wong. Generalized phase change memory scaling rule analysis. In *Non-Volatile Semiconductor Memory Workshop*, 2006.

[11] S. Lai and T. Lowrey. OUM – A 180nm nonvolatile memory cell element technology for standalone and embedded applications. In *IEEE International Electron Devices Meeting (IEDM)*, pages 36.5.1–36.5.4, 2001.

[12] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C.Chen, R. M. Shelby, M. Salinga, and et al. Phase-change random access memory: A scalable technology. *IBM Journal Research and Device*, 52(4/5), 2008.

[13] T. Nirschl andJ. B. Philipp, T D. Happ, G. W Burrt, B. Rajendrant, M.-H. Lee, A. Schrottt, M. YangT, M. Breitwischt, C.-F. Chen, E. JosephT M Lamorey, R. Chee, S.-H. Chen, S. Zaidi, S. Raoux, Y.C. Chen, Y. Zhu, R.Bergmann, H.-L. Lunge, and C. Lamf. Write strategies for 2 and 4-bit multi-level phase-change memory. In *Proceedings of the IEEE International Electron Device Meeting Technology (IEDM)*, pages 461–464, 2007.

[14] W. S. Chen, C. Lee, D. S. Chao, Y. C. Chen, F. Chen, C. W. Chen, R. Yen, M. J. Chen, W. H. Wang, T. C. Hsiao, J. T. Yeh, S. H. Chiou, M. Y. Liu, T. C. Wang, L. L. Chein, C. Huang, N. T. Shih, L. S. Tu, D. Huang, T. H. Yu, M. J. Kao, and M. J. Tsai. A novel cross-spacer phase change memory with ultra-small lithography independent contact area. In *IEEE International Electron Devices Meeting (IEDM)*, pages 319–322, 2007.

[15] D. H. Im, J. I. Lee, S. L. Cho, H. G. An, D. H. Kim, I. S. Kim, H. Park, D. H. Ahn, H. Horii, S. O. Park, U. I. Chung, and J. T. Moon. A unified 7.5nm dash-type confined cell for high performance PRAM device. In *IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2008.

[16] D. Ielmini, S. Lavizzari, D. Sharma, and A. L. Lacaita. Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation. In *IEEE International Electron Devices Meeting (IEDM)*, pages 939–942, 2007.

[17] P. Fantini, G. Betti Beneventi, A. Calderoni, L. Larcher, P. Pavan, and F. Pellizzer. Characterization and modelling of low-frequency noise in PCM devices. In *IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2008.

[18] D. Mantegazza, D. Ielmini, E. Varesi, A. Pirovano, and A. L. Lacaita. Statistical analysis and modeling of programming and retention in PCM arrays. In *IEEE International Electron Devices Meeting (IEDM)*, pages 311–314, 2007.

[19] S. Hanzawa, N. Kitai, K. Osada, A. Kotabe, Y. Matsui, N. Matsuzaki, N. Takaura, M. Moniwa, and T. Kawahara. A 512KB embedded PRAM with 416KBs write throughput at 100μA cell write current. In *IEEE International Solid-State Circuits Conference (ISSCC)*, page 26.2, 2007.

[20] K-J. Lee, B. Cho, W-Y. Cho, S. Kang, B-G. Choi, H-R. Oh, C-S. Lee, H-J. Kim, J-M. Park, Q. Wang, M-H. Park, Y-H. Ro, J-Y. Choi, K-S. Kim, Y-R. Kim, W-R. Chung, H-K. Cho, K-W. Lim, C-H. Choi, I-C. Shin, D-E. Kim, K-S. Yu, C-K. Kwak, and C-H. Kim. A 90nm 1.8V 512Mb diode-switch PRAM with 266MB/s read throughput. In *IEEE International Solid-State Circuits Conference (ISSCC)*, page 26.1, 2007.

[21] F. Bedeschi, R. Fackenthal, C. Resta, E. Donze, M. Jagasivamani, E. Buda, F. Pellizzer, D. Chow, A. Fantini, A. Calibrini, G. Calvi, R. Faravelli, G. Torelli, D. Mills, R. Gastaldi, and G. Casagrande. A multi-level-cell bipolar-selected phase-change memory. In *IEEE International Solid-State Circuits Conference (ISSCC)*, page 23.5, 2008.

[22] G. De Sandre, L. Bettini, A. Pirola, L. Marmonier, M. Pasotti, M. Borghi, P. Mattavelli, P. Zuliani, L. Scotti, G. Mastracchio, F. Bedeschi, R. Gastaldi, and R. Bez. A 90nm 4Mb embedded phase-change memory with 1.2V 12ns read access time and 1MB/s write throughput. In *IEEE International Solid-State Circuits Conference (ISSCC)*, page 14.7, 2010.

[23] C. Villa, D. Mills, G. Barkley, H. Giduturi, S. Schippers, and D. Vimercati. A 45nm 1Gb 1.8V phase-change memory. In *IEEE International Solid-State Circuits Conference (ISSCC)*, page 14.8, 2010.

[24] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano. A novel nonvolatile mem-

ory with spin torque transfer magnetization switching: Spin-RAM. In *Proceeding of IEEE International Electron Device Meeting (IEDM)*, pages 459–462, 2005.

[25] Hiroaki Tanizaki, Takaharu Tsuji, Jun Otani, and et al. A high-density and high-speed 1T-4MTJ MRAM with Voltage Offset Self-Reference Sensing Scheme. In *IEEE Asian Solid-State Circuits Conference*, pages 303–306, 2006.

[26] W. Zhao, E. Belhaire, Q. Mistral, C. Chappert, V. Javerliac, B. Dieny, and E. Nicolle. Macro-model of spin-transfer torque based magnetic tunnel junction device for hybrid magnetic-CMOS design. In *IEEE International Behavioral Modeling and Simulation Workshop*, pages 40–43, 2006.

[27] T. M. Maffitt, J. K. DeBrosse, J. A. Gabric, E. T. Gow, M. C. Lamorey, J. S. Parenteau, D. R. Willmott, M. A. Wood, and W. J. Gallagher. Design considerations for MRAM. *IBM Journal of Research and Development*, 2006.

[28] M. Motoyoshi, I. Yamamura, W. Ohtsuka, M. Shouji, H. Yamagishi, M. Nakamura, H. Yamada, K. Tai, T. Kikutani, T. Sagara, K. Moriyama, H. Mori, C. Fukamoto, M. Watanabe, R. Hachino, H. Kano, K. Bessho, H. Narisawa, M. Hosomi, and N. Okazaki. A study for 0.18μm high-density MRAM. In *IEEE VLSI Symposium on Technology*, pages 22–23, 2004.

[29] Y.K. Ha, J.E. Lee, H.-J. Kim, J.S. Bae, S.C. Oh, K.T. Nam, S.O. Park, N.I. Lee, H.K. Kang, U.-I. Chung, and J.T. Moon. MRAM with novel shaped cell using synthetic anti-ferromagnetic free layer. In *VLSI Symposium on Technology*, pages 24–25, 2004.

[30] T. Kawahara et al. 2Mb spin-transfer torque ram (SPRAM) with bit-by-bit bidirectional current write and parallelizing-direction current read. In *Proc. IEEE International Solid-State Circuits Conference, Tech. Dig*, pages 480–617, 2007.

[31] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai. Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory. *Journal of Physics: Condensed matter*, 19(16):165209, 2007.

[32] S. Salahuddin, D. Datta, P. Srivastava, and S. Datta. Quantum transport simulation of tunneling based spin torque transfer (STT) devices: Design trade offs and torque efficiency. In *IEEE International Electron Devices Meeting (IEDM)*, pages 121–124, 2007.

[33] R. Beach, T. Min, C. Horng, Q. Chen, P. Sherman, S. Le, S. Young, K. Yang, H. Yu, X. Lu, W. Kula, R. Xiao T. Zhong, A. Zhong, G. Liu, J. Kan, J. Yuan, J. Chen, R. Tong, J. Chien, T. Torng, D. Tang, P. Wang, M. Chen, S. Assefa, M. Qazi, J. DeBrosse, M. Gaidis, S. Kanakasabapathy, Y. Lu, J. Nowak, E. O'Sullivan, T. Maffitt, J. Z. Sun, and W. J. Gallagher. A statistical study of magnetic tunnel junctions for high-density spin torque transfer-MRAM (STT-MRAM). In *IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2008.

[34] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando. Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM. In *IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, 2008.

[35] K. Miura, T. Kawahara, R. Takemura, J. Hayakawa, S. Ikeda, R. Sasaki, H. Takahashi, H. Matsuoka, and H. Ohno. A novel SPRAM (SPin-transfer torque RAM) with a synthetic ferrimagnetic free layer for higher immunity to read disturbance and reducing write-current dispersion. In *IEEE VLSI Symposium on Technology*, pages 234–235, 2007.

[36] M. Durlam, P. J. Naji, A. Omair, M. DeHerrera, J. Calder, J. M. Slaughter, B. N. Engel, N. D. Rizzo, G. Grynkewich, B. Butcher, C. Tracy, K. Smith, K. W. Kyler, J. J. Ren, J. A. Molla, W. A. Feil, R. G. Williams, and S. Tehrani. A 1-Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects. *IEEE Journal of Solid-State Circuits*, 38(5):769–773, 2003.

[37] X. Lou, Z. Gao, D. V. Dimitrov, and M. X. Tang. Demonstration of multilevel cell spin transfer switching in MgO magnetic tunnel junctions. *Applied Physics Letter*, 93:242502, 2008.

[38] R. Nebashi, N. Sakimura, H. Honjo, S. Saito, Y. Ito, S. Miura, Y. Kato, K. Mori, Y. Ozaki, Y. Kobayashi, N. Ohshima, K. Kinoshita, T. Suzuki, K. Nagahara, N. Ishiwata, K. Suemitsu, S. Fukami, H. Hada, T. Sugibayashi, and N. Kasai. A 90nm 12ns 32Mb 2T1MTJ MRAM. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 462–463, 2009.

[39] T. W. Andre, J. J. Nahas, C. K. Subramanian, B. J. Garni, H. S. Lin, A. Omair, and Jr. W. L. Martino. A 4-Mb 0.18-μm 1T1MTJ toggle MRAM with balanced three input sensing scheme and locally mirrored unidirectional write drivers. *IEEE Jour. Of Solid-State Circuits*, 40(1):301–309, 2005.

[40] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Y. M. Lee, R. Sasaki, Y. Goto, K. Ito, T. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno. 2 Mb SPRAM (SPin-transfer torque RAM) with bit-by-bit bi-directional current write and parallelizing-direction current read. *IEEE Jour. of Solid-State Circuits*, 43(1):109–120, 2008.

[41] Y. Chen, X. Wang, H. Li, H. Liu, and D. Dimitrov. Design margin exploration of spin-torque transfer RAM (SPRAM). In *International Symposium on Quality Electronic Design*, pages 684–690, 2008.

[42] H. Li and Y. Chen. An overview of nonvolatile memory technology and the implication for tools and architectures. In *Design, Automation and Test in Europe Conference and Exhibition*, pages 731–736, 2009.

[43] J. F. Gibbons and W. E. Beadle. Switching properties of thin NiO films. *Solid State Electronics*, 7:785–797, 1964.

[44] M. Fujimoto, H. Koyama, M. Konagai, Y. Hosoi, K. Ishihara, S. Ohnishi, and N. Awaya. TiO_2 anatase nanolayer on TiN thin film exhibiting high-speed bipolar resistive switching. *Applied Physics Letter*, 89(22):223509, 2006.

[45] R. Jung, M.-J. Lee, S. Seo, D. C. Kim, G.-S. Park, K. Kim, S. Ahn, Y. Park, I.-K. Yoo, J.-S. Kim, and B. H. Park. Decrease in switching voltage fluctuation of Pt/NiO_x/Pt structure by process control. *Applied Physics Letter*, 91(2):022112, 2007.

[46] M. Janousch, G. I. Meijer, U. Staub, B. Delley, S. F. Karg, and B. P. Andreasson. Role of oxygen vacancies in Cr-doped $SrTiO_3$ for resistance-change memory. *Adv. Mater.*, 19(7):2232–2235, 2007.

[47] S. Q. Liu, N. J. Wu, and A. Ignatiev. Electric-pulse-induced reversible resistance change effect in magnetoresistive films. *Applied Physics Letter*, 76(19):2749, 2000.

[48] S. T. Hsu and T. Li. Resistance random access memory switching mechanism. *Journal of Applied Physics*, 101(2):024517, 2007.

[49] M.N. Kozicki, M. Balakrishnan, C. Gopalan, C. Ratnakumar, and Mitkova. Programmable metallization cell memory based on Ag-Ge-S and Cu-Ge-S solid electrolytes. In *Non-Volatile Memory Technology Symposium*, pages 83–89, 2005.

[50] I. H. Inoue, S. Yasuda, H. Akinaga, and H. Takagi. Nonpolar resistance switching of metal/binary-transition-metal oxides/metal sandwiches: Homogeneous/inhomogeneous transition of current distribution. *Physical Review B*, 77(3):035105, 2008.

[51] L. O. Chua. Memristor – the missing circuit element. *IEEE Trans. Circuit Theory*, CT-18(5):507–519, 1971.

[52] J. M. Tour and T. He. The fourth element. *Nature*, 453(7191):42–43, 2008.

[53] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. StanleyWilliams. The missing memristor found. *Nature*, 453:80–83, 2008.

[54] L. O. Chua. Memristive devices and systems. *Proc. IEEE*, 64:209–223, 1976.

[55] Yu. V. Pershin and M. Di Ventra. Spin memristive systems: Spin memory effects in semiconductor spintronics. *Phys. Rev. B, Condens. Matter*, 78(11):113309, 2008.

[56] X. Wang et al. Spin memristor through spin-torque-induced magnetization motion. *IEEE Electron Device Lett.*, 30(3):294–297, 2009.

[57] Y. Chen and X. Wang. Compact modeling and corner analysis of spintronic memristor. In *IEEE/ACM International Symposium on Nanoscale Architectures 2009 (Nanoarch09)*, pages 7–12, 2009.

[58] R. C. Johnson. Superlattices enable small, fast, low-power rram.

[59] Rabindranath Balasubramanian and Gregory Bakker. Programmable system on a chip for power-supply voltage and current monitoring and control, 2009. US Patent pending, application number 12/350,419.

[60] Jongtae Kwak. Delay line circuit, 2009. US Patent US 2009/0243689 A1.

[61] Mark LaPedus. Unity rolls 'storageclass' memory technology, 2009. http://www.eetimes.eu/217500737.

[62] R. Colin Johnson. Memristors ready for prime time, 2008. http://www.eetimes.com/showArticle.jhtml?articleID=208803176.

[63] R. Hoding. NEC develops 32 Megabit MRAM for embedded SoCs, 2009. EETimes, Feb. 12.

[64] F. Pellizzer, A. Pirovano, F. Ottogalli, M. Magistretti, M. Scaravaggi, and et al. Novel μTrench Phase-Change Memory Cell for Embedded and Stand-Alone Non-Volatile Memory Applications. In *IEEE Symposium on VLSI Technology 2004*, pages 18–19, 2004.

[65] Y. Xie, G. Loh, B. Black, and K. Bernstein. Design space exploration for 3D architectures. *ACM Journal of Emerging Technologies in Compuing Systems*, 2(2):65–103, 2006.

[66] Xiangyu Dong, Xiaoxia Wu, Guangyu Sun, Yuan Xie, Hai Li, and Yiran Chen. Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement. In *45th ACM/IEEE Design Automation Conference (DAC)*, pages 554–559, 2008.

[67] http://www.synopsys.com.

[68] http://www.cadence.com.

[69] Spintronic memristors, March 2009. http://www.spectrum.ieee.org/semiconductors/devices/spintronic-memristors/0.

[70] Yuh-Fang Tsai, Yuan Xie, N. Vijaykrishnan, and M. J. Irwin. Three-dimensional cache design exploration using 3DCacti. In *International Conference on Computer Design (ICCD)*, pages 519–524, 2005.

[71] Yuh-Fang Tsai, Feng Wang, Yuan Xie, N. Vijaykrishnan, and M. J. Irwin. Design space exploration for 3-D cache. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(4):444–455, 2008.

[72] Guangyu Sun, Xiangyu Dong, Yuan Xie, Jian Li, and Yiran Chen. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In *IEEE 15th International Symposium on High Performance Computer Architecture, 2009.*, pages 239–249, 2009.

[73] Xiangyu Dong, Norm Jouppi, and Yuan Xie. PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM. In *Proceedings of International Conference on Computer-Aided Design (ICCAD)*, pages 269–275, 2009.

[74] Benjamin Lee, Engin Ipek, Onur Mutlu, and Doug Burger. Architecting phase change memory as a scalable DRAM alternative. In *The 36th International Symposium on Computer Architecture (ISCA)*, 2009.

[75] Moinuddin K. Qureshi, Viji Srinivasan, and Jude A. Rivers. Scalable High Performance Main Memory System Using Phase-Change Memory Technology. In *36th International Symposium on Computer Architecture (ISCA)*, 2009.

[76] Jong-Ho Park, Sung-Hoi Hur, Joon-Hee Leex, Jin-Taek Park, Jong-Sun Sel, Jong-Won Kim, Sang-Bin Song, Jung-Young Lee, Ji-Hwon Lee, Suk-Joon Son, Yong-Seok Kim, Min-Cheol Park, Soo-Jin Chai, Jung-Dal Choi, U. In Chung, Joo-Tae Moon, Kyeong-Tae Kim, Kinam Kim, and Byung-Il Ryu. 8Gb MLC (multi-level cell) NAND flash memory using 63nm process technology. In *IEEE International Electron Devices Meeting (IEDM)*, pages 876–876, 2004.

[77] I.G. Baek, D.C. Kim, M.J. Lee, H.J. Kim, E.K. Yim, M.S. Lee, J.E. Lee, S.E. Ahn, S. Seo, J.H. Lee, J.C. Park, Y.K. Cha, S.O. Park, H.S. Kim, I.K. Yoo, U.In Chung, J.T. Moon, and B.I. Ryu. Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application. In *IEEE International Electron Devices Meeting (IEDM)*, pages 750–753, 2005.

[78] A. Asenov, S. Kaya, and A.R. Brown. Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness. *IEEE Transactions on Electron Devices*, 50(5):1254–1260, 2003.

[79] S. Tehrani et al. Recent developments in magnetic tunnel junction MRAM. In *IEEE Trans. Magn.*, volume 36, pages 2752–2757, 2000.

[80] Gitae Jeong, Wooyoung Cho, S. Ahn, Hongsik Jeong, Gwanhyeob Koh, Youngnam Hwang, and K. Kim. A 0.24μm 2.0-V 1T1MTJ 16-kb nonvolatile magnetoresistance RAM with self-reference sensing scheme. *IEEE Jour. of Solid-State Circuits*, 38(11):1906–1910, 2003.

[81] H. Li, Y. Chen, H. Liu, K. Kim, and H. Huang. Spin-transfer torque memory self-reference read scheme. US Patent pending, application number 12/147,723.

[82] A. Pirovano, A. L. Lacaita, S. A. Kostylev, A. Benvenuti, and R. Bez. Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials. *IEEE Transactions on Electron Devices*, 51:714–719, 2004.

[83] Y. Ho, G. M. Huang, and P. Li. Nonvolatile memristor memory: device characteristics and design implications. In *IEEE/ACM 2009 International Conference on Computer-Aided Design (ICCAD)*, pages 482–490, 2009.

[84] W. Xu and T. Zhang. Using time-aware memory sensing to address resistance drift issue in multi-level phase change memory. In *IEEE International Symposium on Quality Electronic Design (ISQED)*, 2010.

[85] Chih-Yuan Lu, Kuang-Yeu Hsieh, and Rich Liu. Future challenges of flash memory technologies. *Microelectronic Engineering*, 86(3):283–286, 2009.

[86] Man F. Yan. Non-ohmic device using TiO_2. US Patent number 4430255.

[87] Ihun Song, Sunil Kim, Huaxiang Yin, Chang Jung Kim, Jaechul Park, Sangwook Kim, Hyuk Soon Choi, Eunha Lee, and Youngsoo Park. Short channel characteristics of Gallium-Indium-Zinc-Oxide thin film transistors for three-dimensional stacking memory. *IEEE Electron Device Letters*, 29:549–552, 2008.

[88] Agarwal, H. Li, and K. Roy. Drg-cache: A data retention gated-ground cache for low power. In *39th Design Automation Conference (DAC)*, pages 473–478, 2002.

[89] Agarwal, H. Li, and K. Roy. A single-vt low-leakage gated-ground cache for deep submicron. *IEEE Jour. Of Solid-State Circuits*, 35(2):319–328, 2003.

[90] S. Bhunia, H. Li, and K. Roy. A high performance iddq testable cache for scaled cmos technologies. In *IEEE Proceedings of the 11th Asian Test Symposium (ATS'02)*, pages 157–162, 2002.

[91] Y. Chen, H. Li, H. Liu, R. Wang, and D. Dimitrov. Spin-transfer torque memory non-destructive self-reference read. US Patent pending, application number 112/147,727.

[92] H. Li, Y. Chen, H. Liu, and X. Wang. Static source line in stat-ram. US Patent pending, application number 12/242,331.

[93] Y. Chen, H. Li, H. Liu, Y. Lu, and Y. Li. Transmission gate-based spin-transfer torque memory unit. US Patent pending, application number 112/170,549.

[94] H. Li, Y. Chen, H. Liu, and H. Huang. Non-volatile resistive sense memory on-chip cache. US Patent pending, application number 12/250,027.

[95] H. Li, Y. Chen, H. Liu, H. Huang, and R. Wang. Write current compensation using word line boosting circuitry. US Patent pending, application number 12/426,098.

[96] Y. Chen, H. Li, H. Liu, Y. Lu, and S. Xue. Data devices including multiple error correction codes and methods of utilizing. US Patent pending, application number 112/198,516.

[97] H. Li, Y. Chen, D. Setiadi, H. Liu, and B. Lee. Defective bit scheme for multi-layer integrated memory device. US Patent pending, application number 12/502,194.

[98] http://www.cse.psu.edu/∼yuanxie/.

[99] B. Vaidyanathan, Yu Wang, and Yuan Xie. Cost-aware lifetime yield analysis of heterogeneous 3D on-chip cache. In *IEEE International Workshop on Memory Technology, Design, and Testing*, pages 65–70, 2009.

[100] S. Sridharan, M. DeBole, Guangyu Sun, Yuan Xie, and V. Narayanan. A criticality-driven microarchitectural three dimensional (3d) floorplanner. In *Asia and South Pacific Design Automation Conference*, pages 763–768, 2009.

[101] Dongkook Park, S. Eachempati, R. Das, A. K. Mishra, Y. Xie, N. Vijaykrishnan, and C. R. Das. MIRA: A multi-layered on-chip interconnect router architecture. In *35th International Symposium on Computer Architecture (ISCA)*, pages 251–261, 2008.

[102] Xiaoxia Wu, Jian Li, Lixin Zhang, Evan Speight, Ram Rajamony, and Yuan Xie. Hybrid cache architecture with disparate memory technologies. In *International Conference on Computer Architecture (ISCA)*, pages 34–45, 2009.

[103] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan, and M. Kandemir. Design and management of 3D chip multiprocessors using network-in-memory. In *International Symposium on Computer Architecture (ISCA'06)*, 2006.

[104] Gabriel H. Loh, Yuan Xie, and Bryan Black. Processor design in 3D die-stacking technologies. *IEEE Micro*, 27(3):31–48, 2007.

[105] Y. Xie, G. Loh, B. Black, and K. Bernstein. Tutorial: 3D integration for microarchitecture. In *The 39th Annual IEEE/ACM International Symposium on Microarchitecture*, 2006.

[106] Xiangyu Dong and Yuan Xie. System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs). In *Asia and South Pacific Design Automation Conference (ASP-DAC 2009)*, pages 234–241, 2009.

[107] Xiaoxia Wu, Paul Frankstein, and Yuan Xie. Scan chain design for three-dimentional(3D) ICs. In *International Conference on Computer Design*, 2007.

[108] Xiaoxia Wu, Yibo Chen, Yuan Xie, and Krish Chakrabarty. Test-access mechanism optimization for core-based three-dimensional SOCs. In *International Conference on Computer Design*, 2008.

[109] W. L. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan, and M. J. Irwin. Interconnect and thermal-aware floorplanning for 3D microprocessors. In *International Symposium on Quality Electronic Device*, pages 98–104, 2006.

[110] O. Ozturk, Feng Wang, M. Kandemir, and Yuan Xie. Optimal topology exploration for application-specific 3D architectures. In *Asia and South Pacific Design Automation Conference*, page 6, 2006.