Project Description

1 Objective and Significance

The traditional memory technologies, e.g. SRAM, DRAM, and Flash memory, played a very important role in the modern computing system and portable multimedia device industries. However, the scaling of the traditional memories is facing significant challenges from process variations and device reliability at 22nm technology and below [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 design methodologies and circuit 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 NVMs:** The device models for emerging NVMs will be built to fill the gap between process development and circuit design. Memory design flow and optimization methodologies will be developed to facilitate the design space explorations.
- **Circuit techniques for emerging NVM technologies:** 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 developing NVM device models and design methodologis, inventing novel memory structures and circuit techniques, and study the design implications on future computing systems. The work will support the deployment of modern microprocessor and embedded system designs that utilize emerging NVM technologies. The proposed research will provide a complementary perspective to the existing computing system researches.

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 these emerging NVM technologies and summarize the previous related researches.

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 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

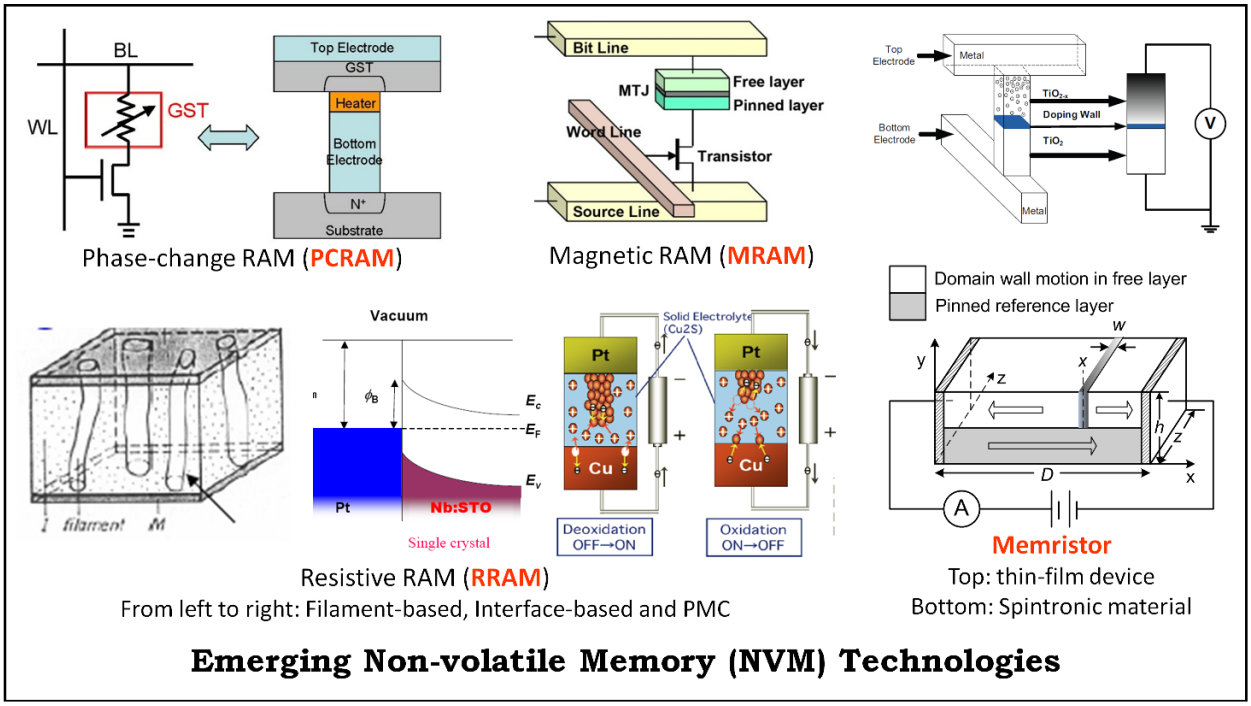

Figure 1: **Overview of Some Emerging Non-volatile Memory Technologies,** including Phase-Change RAM (PCRAM), Magnetic RAM (MRAM), resistive RAM (RRAM), and memristor.

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, fast speed, and good endurance [5–7]. Compared to MRAM, PCRAM is even denser with an approximate cell area of $6 \sim 12F^2$ [1], where F is technology feature size. In addition, phase change material has a key advantage of the excellent scalability within current CMOS fabrication methodology [6, 8–11], which enables the continuous density improvement [12–14].

PCRAM device models have been developed from the point of views of reliability [15], low-frequency noise [16], and statistical analysis [17]. However, these models are dedicated to material and fabrication, which cannot be directly borrowed by circuit designers. Many PCRAM prototypes have been demonstrated in the past years by companies like Hitachi [18], Samsung [19], STMicroelectronics [20, 21], and Numonyx [22]. The maximum capacities achieved for single-level cell and multi-level cell are 1Gb and 256Mb [19, 22], respectively. However, to be more competitive to DRAM and Flash memory, PCRAM need further improve density and endurance.

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

STT-RAM is a new type of Magnetic RAM (MRAM) [1, 23–26], 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. 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 [27,28]. 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 [23, 24, 29–33]. Because of the fast access and soft-error resistance, MRAM potentially could be the next-generation on-chip cache or memory.

Continuous efforts on MRAM process development have been taken to improve yield [34], reduce power consumption [35], and increase density [36]. MRAM prototypes have been demonstrated recently by various companies and research groups [23, 27, 29, 37–39]. 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 been dedicated on developing STT-RAM models for circuit designers. A dynamic MTJ model with a more accurate (transient) description for MTJ resistance switching was proposed, which has been proved to reduce 20% pessimism in write time [40]. We also have analyzed the failure probability of STT-RAM cells due to parameter variations and developed a yield estimation model [41].

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 (multi-level cell, or MLC) resistance states of the resistive switch device. Resistive switching in transition metal oxides was discovered in thin NiO film decades ago [42]. From then, a large variety of metal-oxide materials have been verified to have resistive switching characteristics, including TiO_2 [43], NiO_x [44], Cr-doped $SrTiO_3$ [45], PCMO [46], and CMO [47] 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, RRAM devices can be divided into two types: unipolar switching and bipolar switching.

Programmable-metallization-cell (PMC) [48] is a promising bipolar switching technology. Its switching mechanism can be briefly 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 [49] 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 [50–52].

Memristor was predicted by Chua in 1971, based on the completeness of circuit theory [50]. Memristance is a function of charge, which depends upon the historic behavior of the current (or voltage) profile [52, 53]. 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 [51]. Afterwards, magnetic technology provided the other possible methods to build a memristive system [54, 55]. Due to its unique historic characteristic, memristor has very broad applications including nonvolatile memory, signal processing, control and learning system etc [56].

Many companies are working on technology developments and chip designs of RRAM and memristor-based memory, including Fujitsu, Sharp, HP lab, IMEC, Unity Semiconductor Corp., etc. [57]. The main efforts on RRAM research devote to material and process [43–47] and circuit design issues, such as power-supply monitoring [58] and timing control [59]. Unity Semiconductor Corp. has been processing 64Kb and 64Mb products and expects to demonstrate 64Gb RRAM chips in 2010 [60]. HP Labs also plan to unveil RRAM prototype chips based on memristor with crossbar arrays soon [61].

	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].

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) [62] and embedded PCRAM (Hitachi and STMicro) [18, 63] have been demonstrated, paving the way of integrating these NVMs to the traditional memory hierarchies. In addition, the emerging 3D integration technologies [64, 65] 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 non-volatile memories (NVMs), there are many critical technical issues to be solved. For example, how to integrate these new devices into the existing design flow? How to reduce the impacts of process variations? How to improve poor endurance and prolong life time? How to further increase memory capacity and throughput? etc. To answer these questions, we will start with the device modeling and analysis methodologies for the emerging NVM technologies; On top of it, novel circuit techniques will be proposed for each emerging memory technology based on its unique device characteristic. 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 optimization methodologies for memory design using emerging NVM technologies.

4.1 Task 1.1: Device Modeling for Emerging Memory Technologies

The emerging NVM technologies have introduced many new materials, e.g. 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 production cost, and reduce design margin, especially in the highly scaled technology with large process variations. For example, our previous work [40] 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 design 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 (such as temperature dependance on write pulse shape in PCRAM) need be considered; (2) the device parameter fluctuations induced by process variations, such as line-edge roughnesses, oxide thickness fluctuations, and random discrete dopants, need be analyzed; and (3) the models can provide enough accuracy with reasonable runtime. To be compatible to commercial EDA tools, e.g., Hspice from Synopsys [66] and Spectre from Cadence [67], 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.

With the aid of the dedicated device models, the critical timing and function related parameters, such as 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 language, 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 [66] for more thorough analysis, e.g. the critical path timing at design corners.

4.2 Task 1.2: Emerging NVM Design Flow

Another important task of our proposal is to build a design environment that can be seamlessly integrated the emerging NVMs into the existing CMOS 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 starts with technology specification. Note that logic process and NVM technology could come from different foundries or even at the different technology nodes. Hence, their compatibility and the corresponding impact on design need be considered. 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 our 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 step. Optimization is the key step within the whole design flow so that the dedicated NVM device model will be used. It could take several iterations 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.

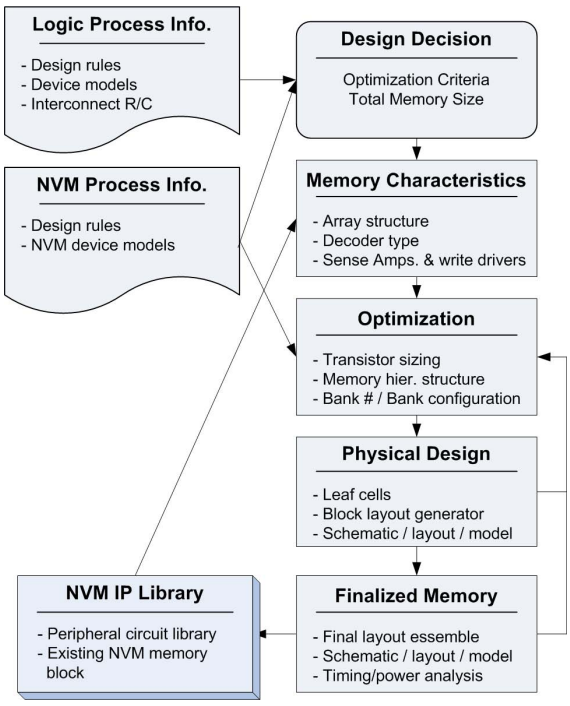

Figure 3: The proposed design flow for emerging NVMs.

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 generic memory cells, peripheral circuitry blocks, and even the completed NMV designs 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 peripheral 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 the 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.

4.4 Preliminary Result and Collaborations:

The NYU-Poly 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 [40] by test-chip design and fabrication at Seagate. We are also one of the first researchers to propose spintronic memristor structures [55], which was interviewed by IEEE Spectrum [68]. The corresponding compact model and corner analysis [56] have also been developed. In this project, we

[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.

will further extend this methodology to other emerging NVMs, such as PCRAM and RRAM. The PSU PI Xie has developed a stacked SRAM cache simulator called 3DCacti [69,70], which has been widely downloaded and used by other researchers.

The PIs have collaborated together when the PI Li was in Seagate, to develop a preliminary version of MRAM simulator for cache stacking [65, 71]. Xie also collaborated with Dr. Norm Jouppi from HP Labs, developed a preliminary version of PCRAM simulator [72]. 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 the memories are similar, i.e. high density, fast speed, low power, affordable yield and reliability, etc. On the other hand, each NVM technology faces different process integration difficulties, owns unique device characteristic, and targets on different memory market. Therefore, the primary concerns and the optimal solutions of different NVM chip designs are different. Our proposal 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 emerging 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 them from massive production and commercialization. Write endurance is usually measured by 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. Based on a survey of PCRAM device and circuit prototypes published within the last five years, the best reported write endurance is 10^9 [73]. Currently, the best test result of STT-RAM write endurance is $< 4 \times 10^{12}$ programming cycles [30]. 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. The gap mainly comes from the immature 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 tricky. 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}, e.g. $t_{wr} \propto -I_{wr}$ 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 by using 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 chip is shipped 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, for example, to adjust the bit-redundancy or to select ECC algorithm. The granularity of the self-contained local control scheme depends on each specific NVM technology and the targeted application.

Circuit techniques to enable wear leveling. Architectural level techniques have been proposed to mitigate the endurance problems in PCRAM, such as *Read-before-Write* or *White Cancellation* [73,74]. 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 [73, 74].

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. The success of MLC in NAND flash memory has been explored in PCRAM [4,11], STT-RAM [36], and RRAM [75]. MLC 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.

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 can be extended to 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

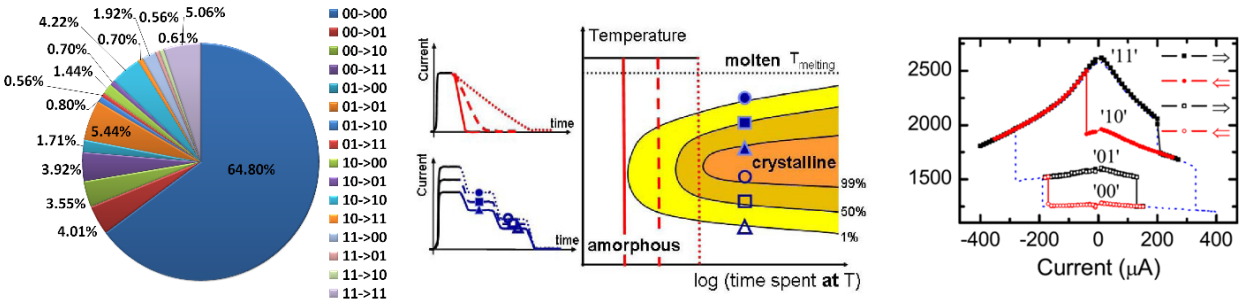

Figure 4: MLC write patterns. Left: The transition distribution between the different values of MLC MRAM bit; Middle: PCRAM – Schematic plot of time-temperature-transformation-chart [12]; Right: MRAM R-I sweep curve [36].

improvement. However, the extra read in write operation results in performance overhead. Is this overhead severe? 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, 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, i.e. 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 [76]. 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 [77] 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.

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. 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

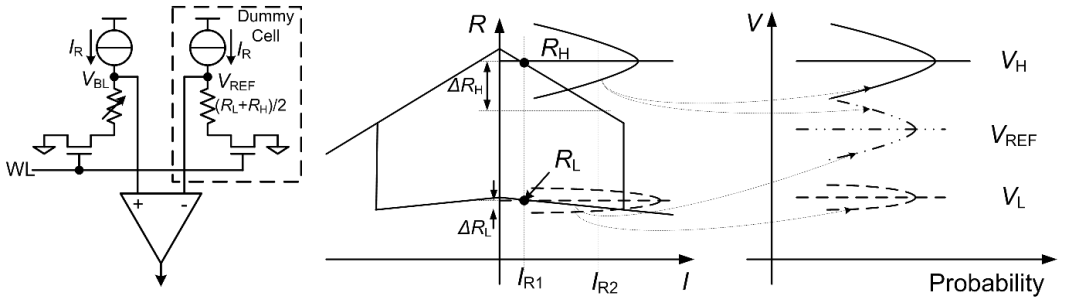

Figure 5: Left: Conventional read-out scheme of MRAM; Middle: R-I characteristic of MgO-based MTJ; Right: MTJ resistance distribution incurred read failure [41].

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 [41]; 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. $0.5\mathring{A}$ [78]. 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 [24, 78]. We also successfully utilized it in STT-RAM design [79]. 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 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'.

To approve the feasibility of this approach and implement it, more questions need to answered. For example, how much is the sensing margin in the new read-out scheme after considering process variations? Is it tolerable within current sensing scheme? 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 minimizing the impacts of process variations as well as improving read performance.

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 [80], 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 [81]. Resistance drift can increase resistance variation σ_R and hence, spread out the resistance distribution. Memory access patterns (e.g. 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 memory design is mainly impacted by read access frequency. Ho et. al. proposed a refreshing scheme: after a particular number of reading operation, a refresh operation is needed to eliminate the effect of read pulse mismatch [81]. We propose to flip the read current direction whenever accessing the memristor. The difficult part of this approach is to determine the granularity. Obviously a fine granularity is more effective to overcome the resistance drift in memristor. However, it also results in high performance and area overhead to record/achieve all the required information, i.e. previous read current directions. On the contrary, a coarse granularity could reduce the overhead, but it may help much in the worst situation.

The resistance drift in multi-level PCRAM has different behavior pattern from memristor-based memory. 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 accessed. Xu and Zhang [82] 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 memory, properly selecting granularity and covering the worst-case situation will be the key issues to be investigated in the project.

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 [83]. Circuit design techniques play an important role to boost memory density too, e.g. word-line overdrive scheme used to reduce select transistor size in memory cell [41]. 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) [84] 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. Conceptually, 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 [48], by stacking two cells back to back with a barrier layer in between. Figure 6 illustrates such a double

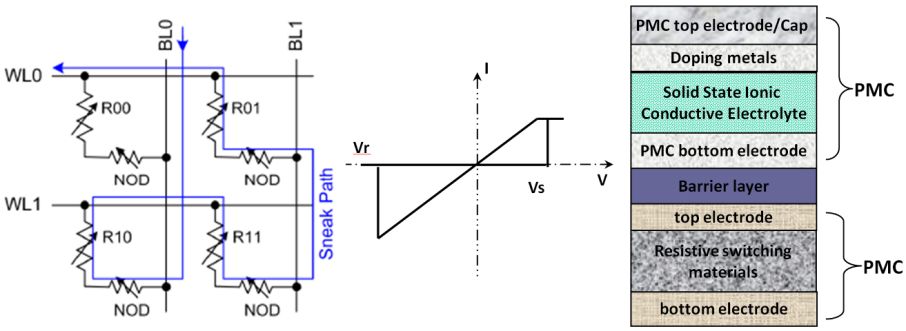

Figure 6: Left: 1NOD-1R and sneak path; Middle: R-I characteristic of PMC; Right: double cell configuration by stacking two cell back to back with a barrier layer in between.

cell by using 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. However, to make this structure feasible to production, there are still many device and design details to be discussed. For example, to guarantee the correct functionality, the timing sequence need be adjusted. Consequently, the peripheral circuit design and floorplan need to be re-designed. The impacts of process variations need to be considered too. Furthermore, will we be able to expand this structure to other RRAM and memristor-based memory design? What kind of device characteristic is favorable in such a structure? In this project, we will address these issues and explore solutions.

3-D stacking peripheral circuitry. 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 memory capacity. If we still use the bottom layer as logic layer and connect upper memory layers through via holes, the area of peripheral circuit will increase significantly, and hence cut down the high-density effect introduced by 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 [85]. 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 upper memory layers because WL's has the identical number as the inputs of WL drivers. Our plan is expanding the usage of TFTs to some simple logic functions, e.g. the last stage of WL decoder. In such a 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 connecting different layers can be dramatically reduced.

Design of such a hybrid circuit will be very challenging but interesting because new device can bring more constraints, which again motivate the invention of new design techniques. Also, many details need be discussed in the approach. For example, how to arrange TFT layout to fit into the small pitch of crossbar array? And how to build power supply network in 3-D stacking peripheral 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 NYU-Poly PI Li has worked on memory design for years, from traditional SRAM to current emerging NVMs. Her research on low-power SRAM design [86–88]has been used in Intel processor. In Seagate, she has led a design team on MRAM and RRAM test-chip design and proposed many circuit techniques to improve reliability [89,90], yield [79,91] and density [92–95] of emerging NVMs. The PSU PI Xie has rich experience in SRAM-based cache design [69, 70, 96–98], PCRAM-based cache design [99], and MRAM-based cache design [65, 71] (in collaboration with the NYU-Poly PI Li) .

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 [100]. 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-Poly and 1 PhD student from PSU on device modeling and NVM 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.

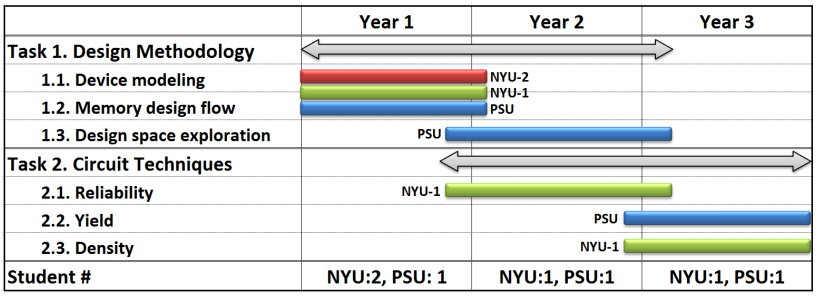

Figure 7: Project Management.

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 [65,71]. 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).

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 [65, 69, 71, 97, 98, 101–105] and 3D EDA tools [64, 69, 70, 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 program 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.