# Emerging Non-Volatile Memory for Hybrid Computing System Project Description

## 1 Introduction

In modern computer architecture design, the storage of instruction and data follows a hierarchical arrangement called **memory hierarchy**, which takes advantage of the access locality and the performance-capacity trade-offs of diverse memory technologies. The importance of the memory hierarchy increases with the advances in microprocessor performance [**?**]. Figure 1 illustrates a typical memory hierarchy. The closer the memory is placed to microprocessor, the faster latency and higher bandwidth are required, with the penalty of the smaller capacity. Different memory technologies, such as SRAM, DRAM, and magnetic hard disk drives (HDD) are the common memory embodiments at the different levels in the memory hierarchy, respectively. With the improvements in speed, density, and cost of Flash memory, solid-state drives (SSD) have gained the momentum as the replacements of the traditional magnetic HDD (Figure 1).

Besides increasing leakage power dissipation, technology scaling also significantly degrades the reliability of SRAM and DRAM. In recent years, we have seen a lot of efforts have been made to address the research and development of some **emerging non-volatile memory (NVM) technologies**, e.g., *Phase-Change RAM (PCRAM)* and *Magnetic RAM (MRAM)*. By combing the speed of SRAM, the density of DRAM, and the non-volatility of Flash memory, these emerging memory technologies demonstrated a great potential to be the candidates of the future universal memories.

As such emerging memory technologies are getting mature, it is important for architecture designers to understand their pros and cons for better utilizing them to improve the performance/power/reliability of future computing systems. To be more specific, we are trying to answer the questions as follows.

- *How to model such emerging NVM technologies?*
- *What will be the impacts of such NVMs on the future memory hierarchy?*
- *What are the research challenges to overcome for such a new memory hierarchy?*
- *What will be the novel applications/architectures with emerging NVM technologies?*
- *How can graduate students be prepared for such future novel architectural innovations?*

To answer these questions, we propose a three-year project. The main objective of the proposal is to study the design implication of such emerging memory technologies for future computer architecture systems. The proposed program makes the following major contributions.

- **Task 1:** Developing **models** for emerging non-volatile memories (NVMs).
- **Task 2:** Proposing novel memory **architecture** to leverage emerging NVM technologies.
- **Task 3:** Exploring novel **applications** that leverage emerging NVM technologies.
- **Integrated educational plan:** Enhancing the core computer architecture courses with new NVM course modules.

Figure 1 illustrates the overview of the proposed project. The proposed work will initiate a novel research direction in high-performance system design and investigate the impact of emerging memory technologies on future computing systems. This work will support the deployment of modern microprocessor designs that use emerging nonvolatile memory technologies. The proposed research will provide a complementary perspective to the existing computing system research.

## 2 Background

In recent years, significant efforts and resources have been put on the researches and developments of emerging memory 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) and Magnetic RAM (MRAM), Resistive RAM (RRAM), and Memristor, have gained substantial attentions and are being actively pursued by industry [?]. In this section we will briefly describe the fundamentals of the two most promising emerging memory technologies to be investigated in our project, namely, the Magnetic RAM (MRAM) based on Spin-Torque Transfer RAM(STT-RAM), and the Phase-Change RAM (PCRAM).

**MRAM based on Spin-Torque Transfer RAM (STT-RAM) technology.** STT-RAM is a new type of Magnetic RAM (MRAM) [?,?,?,?,?], which features non-volatility, fast writing/reading speed (<10ns), high programming endurance ($>10^{15}$cycles) and zero standby power [?]. 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 [?, ?]. 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. Prototyping STT-RAM chips have been demonstrated recently by various companies and research groups [?, ?, ?, ?, ?, ?]. 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.

**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) [?]. 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. 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 [?]. PCRAM has shown to offer compatible integration with CMOS technology [?], fast speed [?], high endurance [?], and inherent scaling of the phase-change process at 22-nm technology node and beyond [?]. Compared to STT-RAM, PCRAM is even denser with an approximate cell area of $6 \sim 12F^2$ [?], where F is the feature size. In addition, phase change material has a key advantage of the excellent scalability within current CMOS fabrication methodology [?, ?, ?, ?, ?], with continuous density improvement [?, ?, ?]. Many PCRAM prototypes have been demonstrated in the past years by companies like Hitachi [?], Samsung [?], STMicroelectronics [?, ?], and Numonyx [?].

**Resistive RAM (RRAM).** RRAM can generally denote all the memory technologies that rely on the resistance change to store the data. Based on the storage mechanisms, RRAM materials can be cataloged as space-charge-limited-current (SCLC), filament, programmable-metallization-cell (PMC), Schottky contact and traps (SCT), *etc.* Among them, filament-based RRAM has been wided investigated because of the potentials on high-speed, high-endurance, and better scalability. The nsulating material between two electrodes can be made conducting through a hopping or tunneling conduction path after the application of a sufficiently high voltage, a process called electro-forming. The data storage could be achieved by break ("reset") or reconnect ("set") the conducting path. Such switching mechanism can in fact be explained with the fourth circuit

| | SRAM | DRAM | NAND Flash | PCRAM | MRAM (STT-RAM) |
|---|---|---|---|---|---|
| Data Retention | N | N | Y | Y | Y |
| Memory Cell Factor ($F^2$) | 50-120 | 6-10 | 2-5 | 6-12 | 4-20 |
| Read Time (ns) | 1 | 30 | 50 | 20-50 | 2-20 |
| Write /Erase Time (ns) | 1 | 50 | 106-$10^5$ | 50-120 | 2-20 |
| Number of Rewrites | $10^{16}$ | $10^{16}$ | $10^5$ | $10^{10}$ | $10^{15}$ |
| Power Read/Write | Low | Low | High | Low | Low |
| Power (Other than R/W) | Leakage Current | Refresh Power | None | None | None |

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

element, i.e., the memristor or the memory resistor [**?**, **?**, **?**]. Indeed, HP Labs plan to unveil RRAM prototype chips based on memristors with crossbar arrays soon.

**Summary.** Figure 1 illustrates the comparison of two emerging memory technologies – PCRAM and MRAM (STT-RAM) – against the traditional main-stream SRAM, DRAM, and NAND-based Flash memory [**?**]. Note that both CMOS-compatible embedded MRAM (NEC) [**?**] and embedded PCRAM (Hitachi and STMicro) [**?**, **?**] have been demonstrated, paving the way of integrating these NVMs to the traditional memory hierarchies. In addition, the emerging 3D integration technologies [**?**, **?**] 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 Research Overview

As such emerging memory technologies are getting mature, it is important for architecture designers to study the design implication of such emerging memory technologies for future computer architecture systems. In this project, we start with Task 1 to study the modeling and analysis methodologies for emerging non-volatile memories (NVMs); In Task 2, we investigate the new memory hierarchy design with NVM technology and study architectural design techniques to enable such memory architecture; Finally in Task 3, we explore novel applications and architectures that are enabled by the unique features of emerging NVM technologies. 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: MRAM/PCRAM Modeling

To help the architectural level and system-level design of the SRAM-based or DRAM-based cache and memory, various modeling tools have been developed during the last decade. For example, CACTI [**?**, **?**, **?**, **?**] and DRAMsim [**?**] have become widely used in the computer architecture community to estimate the speed, power, and area parameters of SRAM and DRAM caches and main memory. Similarly, to explore new design opportunities that these emerging memory technologies can bring to the designers at architecture and system levels, it is imperative to have a high-level model for caches and memories built with emerging NVMs, such as MRAM/PCRAM. The model needs to provide the extraction of all important parameters, including access latency, dynamic access power, leakage power, die area, and I/O bandwidth *etc.*, to facilitate architecture and system-level analysis and to bridge the gap between the abundant research activities at process and device levels and the lack of a high-level cache and memory model for emerging NVMs.

### 4.1 Task 1-A: Device Modeling and Circuit Analysis

Currently, most of researchers working on circuit, architecture and system levels are using highly-simplified characteristics of the emerging devices, due to the lack of knowledge on material physics. 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, the data storage element MTJ at a certain resistance state is usually modeled as a constant resistor by ignoring the dependency of the MTJ resistance on the magnitude of the read/write current driven by the NMOS selection transistor in an MRAM cell. Our previous work [**?**] 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 model the emerging NVM devices and to build a design environment that can be seamlessly integrated with the existing CMOS logic design flow.

Figure **??** illustrates the proposed scope of device modeling and circuit analysis methodology for the emerging NVMs. In Stage I, we will develop the dedicated device models, which will be based on physical mechanism and corroborated by device measurements. The sources and impacts of process variations will be also analyzed and integrated into the dedicated device models. The models will be implemented with SPICE-compatible languages, such as Verilog-A or C language. On top of it, the simplified behavior models will be extracted. High-level languages, *i.e.* VHDL/Verilog or C will be used. In Stage II, we will build an emerging memory design flow, which can realize the creation and optimization of novel hierarchal memory array structure and peripheral circuitry. The accuracy of the corresponding device model will determine the credibility of the design, such as critical timing/power simulation and corner analysis. Therefore, the dedicated device model will be used in this step. High-level synthesis and function verification will also be an important part in Stage II. 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 [**?**] for more thoroughly analysis, *i.e.*, the critical path timing at design corners. In Stage III, we will build IP's (Intelligence Properties) for emerging NVM technologies with the aid of the proposed design flow in Stage II. The IP's will provide the extracted parameters of memory array cell including area, dynamic and leakage power, access latency, *etc.*, the recommendable memory array structures and the corresponding trade-offs, as well as the optimized peripheral circuitry design, *i.e.*, sense amplifier and write drivers. Those IP's will be used in the researches at architectural and system levels.

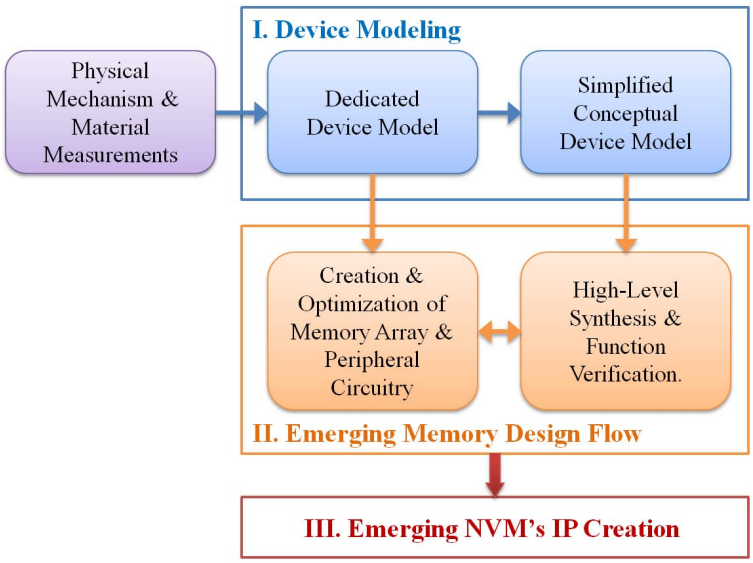


*Figure 2:* The proposed scope of device modeling and circuit analysis methodology for the emerging NVMs.

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.

**Preliminary results:** 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 [**?**] by test-chip design and fabrication at Seagate. We are also one of the first researchers to propose spintronic memristor structures [**?**], which was interviewed by IEEE Spectrum [**?**]. The corresponding compact model and corner analysis [**?**] have also been developed. In this project, we will further extend this methodology to other emerging NVMs, such as PCRAM.

### 4.2 Task 1-B: Architectural Modeling

Based on the device/circuit-level modeling and analysis methodologies described in Task 1-A, we will develop a PCRAM/MRAM simulator, which can be easily integrated with architecture simulators including Simplescalar-based single core simulator [**?**, **?**], and multi-core simulators such as M5 [**?**],GEMS [**?**] or PTLsim [**?**].

Note that tools such as CACTI [**?**, **?**, **?**, **?**] and DRAMsim [**?**] have been widely used in the computer architecture community to estimate the speed, power, and area parameters of the traditional caches and main memory. However, these existing tools were initiated and built based on the cache and memory modelings of SRAM/DRAM. The architectural modeling for PCRAM/MRAM raises unique research issues and challenges on building such simulators. First, some circuitry modules in PCRAM/MRAM have different requirements from those originally designed for SRAM/DRAM. For example, the existing sense amplifier model in CACTI [**?**, **?**, **?**, **?**] and DRAMsim [**?**] is voltage-mode sensing, while PCRAM data reading usually uses a current-mode sense amplifier. Second, due to the unique device mechanisms, the models of PCRAM/MRAM need specialized circuits to properly handle their operations. We can still take PCRAM as an example. The specific pulse shapes are required to heat up GST material quickly and to cool it down gradually during the RESET and especially SET operations. Hence, a model of the slow quench pulse shaper need to be created. Finally, the most obvious and important difference between PCRAM/MRAM and SRAM/DRAM is their distinct memory cell structure. PCRAM and MRAM typically use a simple "1T1R" (one-transistor-one-resistor) or "1D1R" (one-diode-one-resistor) structure, while SRAM and DRAM cell has a conventional "6T" structure and "1T1C" (one-transistor-one-capacitor) structure, respectively. The difference of cell structures directly leads to different cell sizes and array structures.

In addition, where to place these NVM memories in the traditional memory hierarchy also influences the modeling methodologies. For example, the emerging NVMs could be used as a replacement for on-chip cache or for off-chip DIMM (dual in-line memory module). Obviously, the performance/power of on-chip cache and off-chip DIMM would be quite different: When a NVM is integrated with logics on the same die, there is no off-chip pin limitation so that the interface between NVM and logic can be re-designed to provide a much higher bandwidth. Furthermore, off-chip memory is not affected by the thermal profile of the microprocessor core while the on-chip cache is affected by the heat dissipation from the hot cores. While higher on-chip temperature has a negative impact on SRAM/DRAM memory, it actually has a positive influence on PCRAM because the heat can facilitate the write operations of PCRAM cell. The performance estimation of PCRAM becomes much more complicated in such a case. Moreover, building an accurate PCRAM/MRAM simulator needs close collaborations with the industry (see collaboration letters from HP, IBM, IMEC, and Seagate) to understand physics and circuit details, as well as architectural level requirements such as the interface/interconnect with the multi-core CPUs.

**Preliminary Result and Collaborations:** The PSU PI Xie has developed a stacked SRAM cache simulator called 3DCacti [**?**, **?**], which has been widely downloaded and used by other re-

searchers. 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 [**?**, **?**]. Xie also collaborated with Dr. Norm Jouppi from HP Labs, developed a preliminary version of PCRAM simulator [**?**]. We will extend our tools to support architectural exploration in Task 2, especially for hybrid memory systems with an emphasis on multi-core architecture (for example, interface design and coherency modeling) and with Non-Uniform Cache Architecture(NUCA) model (for large memory). Dr. Norm Jouppi from HP Labs, with his expertise in memory architecture modeling, will keep a close collaboration with us for the development of the architectural models for NVMs (see supporting letter from Dr. Jouppi), and we will integrate our models to HP Labs' CACTI tool [**?**], which is an integrated cache and memory model that is widely used in computer architecture community for design space exploration.

### 4.3 Task C-2: Explore novel circuit techniques for NVM

The advent of novel devices have introduced a number of new design issues. In general, the requirements for all the memories are similar, *e.g.* fast speed, high density, affordable yield, low power, *etc.* However, depending on the applications and process development stage, the primary concerns for different technologies could be quite different. Furthermore, the different technologies may need different solutions for the same issue, which mainly rely on the specific device characteristics. Our task is to investigate the corresponding circuit design issues and explore novel circuit techniques for the emerging NVMs.

#### 4.3.1 Task C-2.1: STT-RAM – process variation-tolerant design

In STT-RAM design, there are two main constrains from MTJ device characteristics. (i) The difference between two resistance states of MTJ ($R_H$ & $R_L$) is fairly small: $\Delta$R=$R_H - R_L \approx 1000\Omega$ at 45nm technology node [**?**]. (ii) The large MTJ resistance variation $\sigma_R$ is another challenge. For example, MTJ resistance increases exponentially with the thickness of oxide barrier between two magnetic layers. It was reported in [**?**] that MTJ resistance increases by 8% when the thickness of oxide barrier changes from $14\mathring{A}$ to $14.1\mathring{A}$. Moreover, the MTJ resistance variation will be aggravated by the further reduction of oxide barrier thickness in scaled technologies. Besides oxide barrier thickness, MTJ resistance is also significantly affected by the large MTJ geometry variations.

The small $\Delta$R and the large $\sigma_R$ could lead to the false detection of the stored value. For example, in a conventional voltage sensing scheme, read current $I_R$ is sent to the STT-RAM cell and generates bitline (BL) voltage $V_{BL,L} = I_R \cdot (R_L + R_{TR})$ or $V_{BL,H} = I_R \cdot (R_H + R_{TR})$, when the MTJ is at the low resistance state and the high resistance state, respectively [**?**]. Here, $R_{TR}$ is the resistance of NMOS transistor. By comparing the BL voltage to a reference voltage $V_{REF}$ between $V_{BL,L}$ and $V_{BL,H}$, the MTJ resistance state can be readout. Usually a $V_{REF}$ is shared by multiple STT-RAM bits, hence it needs to satisfy: $\max(V_{BL,L}) < V_{REF} < \min(V_{BL,H})$. Here, $\max(V_{BL,L})$ and $\min(V_{BL,H})$ denote the maximal $V_{BL,L}$ and the minimal $V_{BL,H}$ generated by all involved STT-RAM bits, respectively. Unfortunately, $\max(V_{BL,L}) < \min(V_{BL,H})$ ***may not be always true when the bit-to-bit variation of MTJ resistance is large.*** New read-out scheme is required to overcome the STT-RAM processor variation and improve chip yield.

**A nondestructive self-reference read scheme.** We propose a nondestructive self-reference methodology to reduce read failures in STT-RAM design. The basic idea 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 *destructive* self-reference schemes were used in toggle-mode MRAM design [**?**, **?**]. We also successfully utilized it in STT-RAM design [**?**, **?**]. We call these schemes "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 introduces reliability issue.

Different from previous self-reference schemes, we propose a ***non-destructive self-reference*** approach based on the special R-I characteristic of MgO-based MTJ's. As we can see in Figure 2, the MTJ current dependence of the high and the low resistance states are quite different: the current roll-off slope of high resistance is much steeper than that of low resistance. 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 $dR = R1 - R2$. Obviously $dR_H$ when the MTJ is at the high resistance state should be pretty big, while $dR_L$ when the MTJ is at the low resistance state is close to '0'.

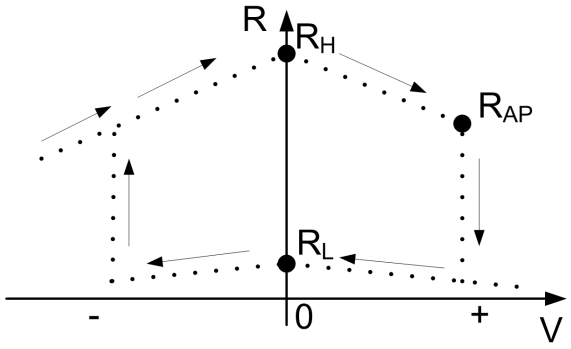


*Figure 3:* The static R-I curve of MgO-based MTJ.

**Preliminary results:** The scheme of our proposed nondestructive self-reference scheme is shown in Figure 3. A switch transistor SLT1 is connected to BL as well as the corresponding voltage storage element C1. The other switch transistor SLT2 is connected to a voltage divider. The top connect point of C1 ($V_{BL1}$) and the output of the voltage divider ($V_{BL2O}$) are connected to the two inputs of a voltage sense amplifier, respectively. The voltage divider is used to eliminate the impact of difference between two read currents $I_{R1}$ and $I_{R2}$. The operation of our proposed nondestructive self-reference scheme includes three steps: (1) First read: A read current $I_{R1}$ is applied and incurs the corresponding $V_{BL1}$, which is stored in C1. (2) Second read: Another read current $I_{R2}$ is applied and incurs $V_{BL2}$. $V_{BL2}$ goes through voltage divider and generate $V_{BL2O}$. (3) Sensing: $V_{BL1}$ and $V_{BL2O}$ are compared by the voltage sense amplifier. If $V_{BL1}$ is significantly larger than $V_{BL2O}$, the original value of STT-RAM bit is ??1?? (high resistance state). Otherwise, the original value of STT-RAM bit is ??0?? (low resistance state).

Compared to the conventional self-reference scheme [**?**, **?**, **?**], the proposed scheme can provide much faster read speed by eliminating two write steps (erase and write-back), which make it possible to utilize STT-RAM on-chip as well as satisfy the performance requirement. The reliability of STT-RAM is improved too by eliminating the unnecessary disturbances (writing) during read operations. However, to fully realize this approach, there are still some circuit issues to be solved. For example, the sensing margin of the proposed scheme is bigger than the one of conventional STT-RAM read scheme, but smaller the one of the "destructive" self-reference. So what type of sense-amplifier design is the best with the consideration on both accuracy and speed requirement? How to reduce the impact on overall area of peripheral circuitry? Will the traditional SRAM-liked

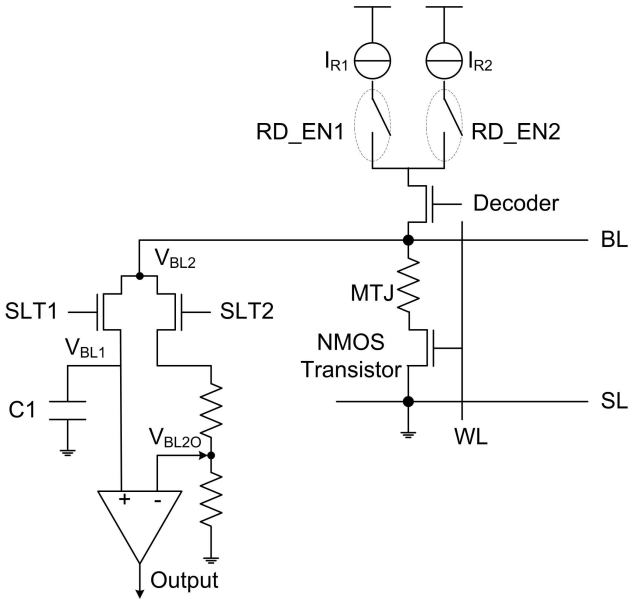


*Figure 4:* Nondestructive self-reference sensing scheme for STT-RAM.

memory array structure is good enough or some
novel structure could be more beneficial? Furthermore, how will process variations, including both MTJ and CMOS processes, affect the proposed read scheme? In this project, we will address these circuit issues from both device and circuit point of views and explore the solutions.

#### 4.3.2 Task C-2.2: PCRAM – endurance enhancement

PCRAM exploits the large resistance contrast between the amorphous and crystalline states in phase-change materials [**?**]. The resistance difference could be four or five orders sometime. Given such a large resistance contrast, the difference in read current is read current is more than sufficient for binary storage and even MLC (multi-level cell) operation [**?**].

The most stringent requirement to prevent PCRAM from production is endurance. While READ endurance is not likely a problem, the best reported WRITE endurance for PCRAM is only $10^9$ based on a survey of PCRAM device and circuit prototypes published within the last five years [**?**]. Besides the improvement on material, we could also help out in circuit design level. From circuit design point of view, we could also help out in many ways and this will be our objective in the proposal.

One possible solution is smoothing the driving current during write operations and avoiding the overshot on phase change materials. Here, how to design a write driver to provide a sleek but fast ramp-up curve is the tricky part. We should also reduce memory access time, especially for SET/RESET operations, because endurance failure is directly related to the duration of energy applied on phase change material. Obviously an accurate self-timing control scheme is required, which can stop providing current to memory cells once detecting successful SET/RESET operations. Another interesting method could be lowering the voltage on phase change material to meet only the minimal SET/RESET requirements because the failure increases exponentially when energy increases. Furthermore, for some applications that non-volatility is not a requirement (i.e. directly replace SRAM with PCRAM), we can trade data retention with endurance by further reducing the energy pulse. Of course, the statically or dynamically fixing by using redundancy and ECC will keep useful. However, will the more complex ECC algorithms or bit-level redundancy be needed? We will investigate it in the proposal.

#### 4.3.3 Task C-2.3: RRAM – density improvement

RRAM is expected to replace NAND Flash memory as main storage in near future [**?**]. Hence, increasing memory density becomes the ultimate goal in RRAM design. Usually an NMOS transistor is used as selection device in a random access memory cell (*e.g.* DRAM and MRAM) by connecting it in series with the data storage element. Such a cell structure needs three sets of terminals – word line (WL), bit line (BL) and source line (SL). The routing requirement and design rules determine that the minimal cell size is $12F^2$ [**?**], which is too big to be tolerant in RRAM design. Here, $F$ represents the technology feature size. Theoretically, the smallest memory cell is $4F^2$, which has only two terminals – one is horizonal and another is vertical. The storage element is built at the cross-point of two metal wires. Hence, this is called cross-point structure. Moreover, the cross-point structure can 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. Hence, cross-point structure is widely investigated in RRAM design.

From design point of view, RRAM technologies can be divided into two operation types: unipolar switching and bipolar switching. Unipolar operation executes the programming/erasing by using short and long pulse, or by using high and low voltage with the same voltage polarity. Usually a diode is served as selection device (1D1R). The data in bipolar switching RRAM can be changed

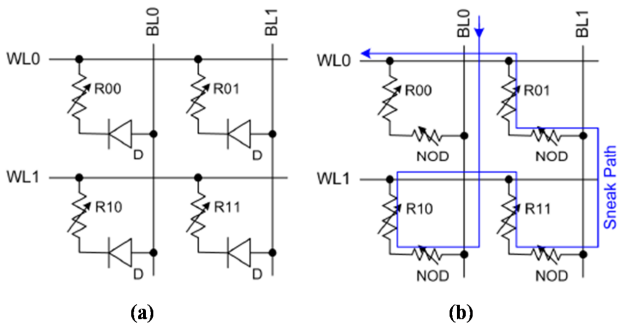


*Figure 5:* RRAM memory cell scheme. (a) 1D1R; (b) 1NOD-1R.

by short voltage/current pulses with opposite voltage polarity. For such memory structures, non-ohmic device (NOD) [**?**] is used to provide two-direction driving current as well as support process integration of cross-point structure. We call it as 1NOD-1R (See Figure 4).

However, 1D1R and 1NOD-1R cell structures are facing on some design difficulties due to process limitation. 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 4(b). The sneak current can introduce disturbance on unintended cells during read, write and erase operations. Therefore, diode (P-N or Schottky) is more favorable as a selective element for RRAM array and intra-die stacking. However, it is extremely difficult to achieve the high quality diode with large $I_{on}/I_{off}$ ratio (large forward current $I_{on}$ and extremely small reverse current $I_{off}$) by using temperature limited BEOL (back end of line) process ($< 400°C$) [**?**].

**Using bipolar PMC as the selective element.** We propose to bipolar resistive switching devices as the selection device. Programmable-metallization-cell (PMC) could be a good candidate. PMC [**?**] is a promising bipolar RRAM technology, which is composed of two solid metal electrodes – relatively, one is inert and the other is electrochemically active. Between the two electrodes locates a thin electrolyte film. When a negative bias is applied to the inert electrode in programming operation (SET), metal ions in the electrolyte together with those flew from the positive active electrode can be reduced by the inert electrode. As a result, the metal ions form a small metallic "nanowire" between the two electrodes, which produces a low resistance. In erasing operation (RESET), a positive bias is applied on the inert electrode. Metal ions migrate back into the electrolyte and eventually to the negatively-charged active electrode. The "nanowire" is broken and the resistance increase back. The I-V curve is illustrated in Figure 5(a). A higher voltage is required in RESET operation ($V_r$)than the one in SET operation ($V_s$).

**Preliminary results:** Figure 5(b) demonstrates a double cell configuration built by stacking two PMC's back to back with a barrier layer in between. In this structure, 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. This cell structure has been successfully demonstrated in process [**?**]. Since PMC technology itself is compatible to CMOS process, the integration of the proposed structure is friendly to CMOS technology too.

The operations can be explained by using the corresponding I-V curves of the two PMC devices shown in Figure 5(c). Here, the top and bottom I-V curves are for the memory and selection devices, respectively. We can see that when the voltage drop cross the RRAM cell exceed $V_r$, the

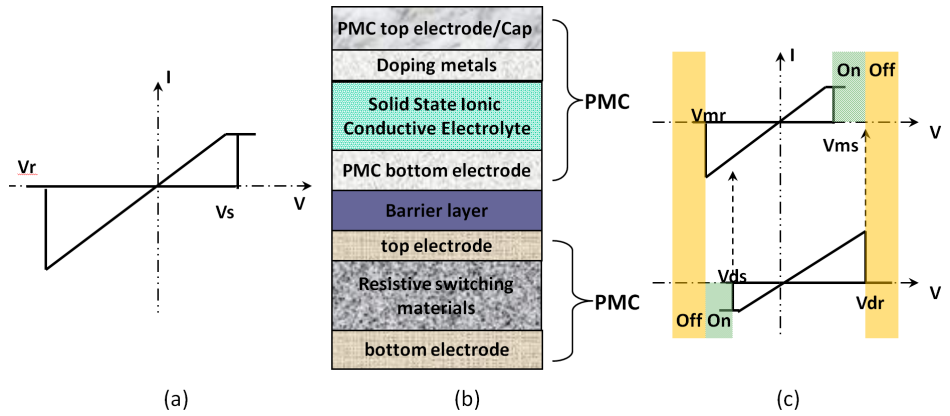


*Figure 6:* (a) I-V curve of PMC; (b) The proposed double cell configuration based on PMC technology; (c) I-V curves of the two back-to-back PMC devices shown in (b).

RRAM cell is "off" because either PMC1 or PMC2 are in high resistance state.

- SET operation: First, a negative bias in between $-V_{ds}$ and $-V_{mr}$ is applied to turn on the switch; Then a positive voltage higher than $V_{ms}$ but smaller than $V_{dr}$is used to program memory. After data is successfully written to the cell, further increase voltage to higher than $V_{dr}$ in order to turn off the switch.
- RESET operation: Apply a negative bias lower than $-V_{mr}$, it will first turn on the switch and than reset the memory cell. In such a situation, there is no way to turn off the switch. But there is no leakage path either due to the high resistance of the memory.
- READ operation: Like SET and RESET, the switch needs to be turned on first by applying a negative bias in between $-V_{ds}$ and $-V_{mr}$. Then a small current can be used to read memory cell resistance state. At the end of the read operation, the switch is required to turned off or keep it on when memory is in low or high resistance state, respectively.

Compared to diode or NOD, PMC based switch has two advantages – bipolar switching and large $I_{on}/I_{off}$ ratio. Hence, the proposed scheme could be used in bipolar switching RRAM design with minimized sneak current. Although we have investigate the feasibility based on theoretical analysis, there are still a lot of unsolved issues. For example, how to control timing and applied voltage? What kind of peripheral circuitry floorplan will be optimal for the proposed RRAM design? And again, how will process variations affect the proposed RRAM scheme? In this project, we will address these circuit issues from both device and circuit point of views and explore the solutions.

## 5 Related Work

In recently years, there have been active efforts on emerging NVM technologies. However, most of efforts were at process and device levels. Relatively, the architecture and system level analysis is less due to the lack of a high-level cache and memory model for emerging NVMs.

**PCRAM.** Compared to STT-RAM, PCM is even denser with an approximate cell area of $6 \sim 12F^2$ [**?**], where F is the feature size. In addition, phase change material has a key advantage of the excellent scalability within current CMOS fabrication methodology [**?**, **?**, **?**, **?**, **?**]. Continue density improvement is the most important task for PCRAM process development. 2 and 4-bit MLC PCRAM material and the correponding write strategies were demonstrated by Nirshl et al. [**?**].

New process integration techniques, such as ultra-small lithography independent contact area [?] and unified 7.5nm dash-type confined cell [?], could also help enhance density. Reliability is the major challenge in PCRAM process, which severely limits its applications. Researches on different angles have been done: Lacaita et al. discussed projected its impact on scaling [?], Shih et al described the mechanisms of retention loss in GST material [?], and Lavizzari et al presented the impact of transient effects [?]. Accordingly, many device models were built from reliability [?], low-frequency noise [?], statistical analysis [?] point of views. Those models mainly were dedicated to process and device, which cannot be borrowed by computer community.

Many PCRAM prototypes have been demonstrated in the past years. In 2007, a 1.5V 512KB embedded PCRAM in a 0.13$\mu$m CMOS by Hitachi [?] and a 512b diode-switch PCRAM in a 90nm CMOS by Samsung [?]. A year later, a 256Mb MLC PCRAM in a 90nm technology by STMicroelectronics [?]. Accordingly, a multi-level programming algorithm was developed and embedded into the chip, demonstrating 2b/cell feasibility. Very recently, a 45nm 1Gb 1.8V single-level cell (SLC) PCRAM was designed with 85ns random-access time and 9MB/s program throughput was demonstrated by Numonyx [?], and A 90nm 4Mb embedded PCRAM with 1.2V 12ns read access time and 1MB/s write throughput was made by STMicroelectronics [?]. In addition, The peripheral circuit design for high density diode-switch PCRAM was discussed in ISQED 2009 [?]. Diode-switch PCRAM was demonstrated in VLSI Symposium 2007 [?]

Discussions on endurance limitation were recently brought up to off-ship memories. Lee et al. proposed techniques to reduce the write accesses to PCM-based main memory to improve its endurance [?]. One of the techniques utilizes the dirty bits in L2 at the word granularity to check if a word has been updated since it was last fetched on-chip. A technique along a similar line for PCM memory at bit level was proposed by Zhou et al. [?]. In this technique, a memory row is first read out, then compared with the new data, and finally written back for those changed values. This technique can significantly save performance and energy by avoiding pre-write operations. These technologies, however, all incurred significant access performance degradation and cannot be directly used in on-chip cache structure.

**MRAM.** MRAM features non-volatility, fast access speed, zero standby power and high programming endurance [?, ?]. The research on MRAM material and device mainly devoted on spin-torque based MTJ due to its better performance, higher density, and better scalability compared to the conventional MRAM [?, ?, ?, ?]. Certainly, the yield improvement is an important topic in nanoscale devices. Miura et al. presented a SPRAM with synthetic ferrimagnetic free layer, which has high immunity to read disturbance and sufficient margin between read and write currents [?]. The MTJ structure with synthetic ferrimagnetic free layer can achieve a lower critical current density without degrading the thermal stability [?]. Very recently, a 2-bit MLC (Multi-level cell) MTJ device was reported in [?] for further density enhancement. Two-digit information – 00, 01, 10, and 11, are represented by four MTJ resistance states. The transitions between different MTJ resistance states can be realized by passing the spin-polarized currents with different amplitudes and/or directions.

A 4Kb STT-RAM using tailored MTJ design was fabricated by Sony in 0.18$\mu m$ technology in 2005 [?]. The test chip demonstrated that STT-RAM is a prominent candidate for the next generation memory because of its high speed, low power and high scalability. In 2007, Kawahara et al. prototyped a larger 2Mb STT-RAM in 0.2$\mu m$ technology [?]. This chip improves memory access latency by featuring an array scheme with bit-by-bit bidirectional current write and a prallelizing-direction current read. Recently, a even larger capacity – 32Mb MRAM prototype in 90nm technology was demonstrated by NEC [?]. A cell structure with 2 transistors and 1 magnetic tunneling junction (2T1MTJ) was adopted to improve access time to 12ns. Besides the SRAM-like array [?, ?, ?], other memory structures are also investigated by using MRAM/STT-RAM technol-

ogy. In [**?**], Wang et. al. described a CAM structure based on conventional MRAM technology. In [**?**], Wu et. al. propsoed a novel STT-RAM read scheme with high sensing margin and illustrates a new CAM design. The possibility of applying STT-RAM in reconfigurable logic block for 3D-stacked reconfigurable spin processor was investigated [**?**].

A write disturbance fault (WDF) model for conventional MRAM was proposed by Su et al. [**?**]. The fault affects the data stored in MRAM cells due to excessive magnetic filed during a write operation. This should not be a problem to STT-RAM since it uses spin-polarized current to flip data. We have proposed a dynamic MTJ model with more accurate (transient) description for MTJ resistance switching [**?**]. 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 [**?**]. A model was proposed to predict memory yield and design optimization to minimize memory failures.

At architecture level, there are several recent efforts in using STT-RAM as an on-chip last level cache. Desikan et. al. conducted an architectural evaluation of on-chip conventional MRAM cache in a single micro-processor [**?**]. Dong et al. developed a delay and energy model for MRAM-based cache and conducted a detailed comparison between the cache with SRAM and STT-RAM technologies in terms of area, performance and energy in the context of 3D stacking [**?**]. Sun et al. extended the application of STT-RAM based cache to Chip Multiprocessor (CMP) and proposed new techniques to improve latency and to reduce write energy [**?**].

## 6 Education Plan

An academic job is not only to create knowledge but also to disseminate that knowledge to others both at the graduate and undergraduate levels. Education is an integral part of the PI's career development plan and is a supportive and inseparable part of the PI's research. It is the PI's belief that research and academic activities should be inextricably linked in a healthy academic environment. The PI envisions that students play an important role in the research program. Successful teaching and other forms of interaction with students will attract them to the research area, excite them to learn more about the ongoing research, and eventually contribute to the PI's research program. The interaction between academia and industry is also very important. The following sections present the objectives of the proposed educational plan.

### 6.1 Course Development and Teaching

- **Course module development**

  As computer engineering educators, we should not only preserve the historical domain of our discipline, but also expand it. Current standard curricula on VLSI design/computer architecture/embedded system design are still mainly oriented to *deterministic design paradigm*, giving none or little emphasis on the non-deterministic behaviors introduced by technology scaling or novel device. In stead of introducing a new course, **the major goal of the PI is to develop and disseminate course modules that complement or upgrade existing core courses, by introducing new development and challenges in the forthcoming probabilistic design paradigm**. These class modules (overheads, labs, and notes) on process variations and probabilistic design techniques, will easily upgrade and complement a variety of courses, including embedded system designs, computer architecture, VLSI circuits and systems, and design automation tools/methodologies. The course modules and class projects related to PVT variations will be developed to provide students with hands-on experience on statistical timing analysis and statistical optimization techniques. Some of the modules will be incorporated into undergraduate

engineering courses and others will be more suitable for graduate-level courses:

- *Undergraduate courses.* The PI plans to revamp a senior-level course (CSE 477: VLSI Digital Circuits) by introducing new course modules, such as circuit-level process variation, new labs on SPICE simulation of process variation, and gate-level statistical timing analysis tools. The PI also plans to incorporate variation-aware micro-architecture design concepts, such as the Razor architecture [**?**], into a senior-level computer architecture course (CSE 431: Computer Organization and Design).
- *Graduate courses.* The PI introduced a new graduate level course (CSE 598C: Design of Reliable Power Efficient Systems) in his first semester on the faculty at Penn State. The new course already included the reliability issues caused by temperature variation and power supply noise variation. Another graduate-level course that the PI co-teaches is a research seminar course (CSE597D: Embedded System Design). The PI plans to develop and incorporate new modules on process variation as well as the new research outcomes from the CAREER research program for these courses, and make them available to other academic institutions through the WWW or CDs, so that they can be easily integrated into a variety of courses. In fact, some of the modules from the PI's course have been adopted by University of Minnesota (Prof. Antonia Zhai) and University of Connecticut (Prof. Yunsi Fei). The PI believes that the new modules will continue to benefit his colleagues at Penn State and other institutions.

- **Fostering Interaction with Industry**

  In teaching computer engineering courses, the PI believes in the clear need to present real-life example products and applications, so that students can understand the significance of what they are learning. For instance, in the past, the PI infused his industrial design experience into the classroom and discussed the real problems he encountered when working as a designer. In this education plan, the PI plans to invite his industry contacts, who are the experts in process variation (Dr. Kerry Bernstein from IBM and Dr. Tanay Karnik from Intel), to give guest lectures on how complex chips are designed and what biggest challenges the industry faces in the probabilistic design paradigm.
- **Bringing Research into the Classroom**.

  The PI views research and teaching as complementary and promotes this view in all the classes he teaches. For example, some projects from his graduate course involve both theory and implementation on new topics, and have led to conference quality papers [**?**, **?**, **?**]. The PI will propose related research topics as possible projects and the tools developed via the research plan will also be used in the classroom.

  The semiconductor industry is a fast growing and swiftly changing area. The PI believes that it is extremely important to keep the advanced-level graduate courses up-to-date with latest research papers and novel ideas. When the PI was at IBM, the company had a tradition of giving a review seminar to all employees after a major conference (such as DAC or ISSCC). The PI plans to borrow this idea and develop a graduate seminar course, in which the students will review the latest major design automation conferences and real-time embedded system conferences, such as DAC, ICCAD, and RTSS. The PI will guide the students to review the most significant papers from each conference, encouraging students to explore their interesting topics and discover potential research topics.
- **Striving for Teaching Excellence**

  The PI has worked very hard to improve his teaching effectiveness rapidly. He has consulted senior professors with numerous questions about teaching and signed up for the teaching seminars organized by the Penn State's graduate school. These efforts have been successful; for the first 3 years of teaching at Penn State, the PI received an average 6.03/7 rating for his teaching

evaluation, which was higher than the departmental and college average (5.17/7 and 5.28/7, respectively). The PI will keep using all resource available at PennState to improve his teaching. In particular, he will keep participating all activities provided by PennState's Schreyer Institute for Teaching Excellence (www.schreyerinstitute.psu.edu), including teaching luncheon, seminars, and workshops.

### 6.2 Multidisciplinary/Multi-institution/International Education Collaboration

The PI plans to investigate *collaborative teaching experiments* that can then be adopted by other universities. In fact, The PI has **already** made a plan with Carnegie Mellon University and University of Pittsburgh: The PI will work with *Prof. Rob Rutenbar* from CMU and *Prof. Alex Jones* from UPitt to organize a graduate-level course on design automation tools and algorithms in Fall 2006 (CSE 578: CAD Tools). This course will span design automation flow from high-level synthesis to physical synthesis. The course will be simultaneously offered at Penn State, CMU, and University of Pittsburgh through online course delivery system (WebEx). Lectures will originate from the different schools based on the topics to be covered.

Based on the experience and assessment from this design automation course, the PI plans to organize another graduate-level course on probabilistic design flow for MPSoC embedded system design, which complements the CAREER research plan. The course will involve Princeton University (Prof. Wayne Wolf on conventional embedded MPSoC design), North Carolina State University (Prof. Frank Mueller on embedded software and compiler design), and Northwestern University (Prof. Hai Zhou on statistical timing analysis and optimization). The PI will incorporate the latest research outcomes from this project. Students at PSU/Princeton/Northwestern/NCSU 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 by 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.

The PI also believes that the success of universities in the United States stems from their willingness to leverage the best talent around the world. The PI plans to foster the connections between PennState and other top universities in the world. In fact, the PI will spend 7 weeks during the summer of 2006 visiting top universities in Asia. He will spend 4 weeks at Tsinghua University in Beijing to teach a short course on embedded system design. He will then spend 2 weeks and 1 week at National Taiwan University and Hong Kong University of Science and Technology, respectively, to give seminars and explore research collaborations with these universities. He also plans to conduct an international teaching experiment with Tsinghua University by offering his graduate level course on embedded system designs to Tsinghua University's graduate curriculum.

### 6.3 Outreach and Broader Impact

**Tutorials and Workshops:** The PI also believes that it is important to share research findings with experts in related areas. Such activities can spawn inter-disciplinary and inter-university research collaborations. The PI has a good track record of presenting tutorials on his latest research outcome, together with other experts in the field. For example, **he has presented tutorials** in ASPLOS 2004 [**?**], ASPDAC 2005 [**?**], ISCA 2005 [**?**], ASICON 2005 [**?**], and will present in MICRO 2006 [**?**]. From past offerings of tutorials, the PI has found these tutorials to serve as a spark plug for drawing more researchers to start working on a particular research area by creating awareness of the problem's importance and by introducing tools to facilitate in solving the problem. The PI

will continue this tradition in this program, and prepare **tutorials for the future embedded system conferences such as EMSOFT, ISSS+CODES, RTSS, or CASES**, based on the new outcomes from this research. The PI has served as the tutorial chair for **EMSOFT 2005** and helped organized the conference. He also served in the **CASES 2006** program committee. In the future, the PI plans to organize a **workshop on probabilistic embedded system design**, since a workshop is a great aid to attract other researchers to exchange ideas in this important area.

**Industrial Courses:** University and industry interactions are very important and benefit each other. On the one hand, the class materials in university education need to be infused with the latest developments in the industry; on the other hand, professional engineers need continuing education and training to develop new specialized skills and keep up with the rapid changes in the industry. Penn State is part of the University-Industry-Government partnership called *The Technology Collaborative (TTC)* (www.techcollaborative.org), which focuses on research, training, and education issues related with system design. The PI is actively involved with their education programs and has offered courses to the local industry in the past through TTC. Based on his graduate level course (Design of Reliable Power Efficient Systems), the PI has developed a two-day short course on designing reliable power efficient systems for industrial engineers. The course was delivered twice (in Jan 2004 and May 2004) to industrial engineers in companies like IBM, Seagate, and ADC, via an online course delivery system (Webex) as well as local attendance. He was also invited by Synopsys Inc. to give a five-day short course to industrial engineers on designing reliable power efficient circuits. The PI plans to maintain a close relationship with industry and disseminate findings of the proposed research to industry practitioners, who in turn can incorporate these into real Embedded SoC designs.

### 6.4 Student Advising

- **Advice Webpage**. Currently the PI has four Ph.D. advisees and co-advisees, including one female student. The PI has created an advice webpage (www.cse.psu.edu/ỹuanxie/advice.htm), which is a collection of more than 100 useful links categorized as follows: (1) Ph.D. dissertation/research advice; (2) presentation advice; (3) technical writing advice; (4) technical reviewing advice; (5) Job hunting advice; and (6) English learning advice. These advice links were collected by the PI when he was a graduate student at Princeton University and was extremely helpful for the smooth completion of his Ph.D. study. The PI's advisees have found them very useful when they start their graduate study. The webpage has also benefited other faculty members' advisees and the PI has received much positive feedback.
- **Advising Under-represented Groups and Undergraduate Research**
  The Computer Science and Engineering disciplines have exhibited a growing gender gap [**?**, **?**]. Working close together with his mentor Dr. Mary Jane Irwin, who is very active in CRA-W and ACM-W, the PI has strived for the promotion of women and minority in engineering. Currently he has one female Ph.D. student, who has been involved in the preliminary work proposed in this program [**?**], and will be supported by this program if it is funded. The PI has encouraged her to participate PennState Women in Engineering Program (WEP), to attend CRA-W (Women in Computing Research Association)'s computer architecture summer school in 2006, and to attend DAC 2006 with ACM-W scholarship. He plans to recruit two minority undergraduate students through the Penn State WISER (Women in Science and Engineering Research) and MURE (Minority Undergraduate Research Experience) programs, with funding provided by Penn State. These two students can help develop software for this program. PennState's WEP program also organizes Engineering Camp for Girls and Girl Scout Saturdays every year. The PI plans to participate and design small projects for these students, to promote computer engineering among

high school female students.

### 6.5 Outreach and Broader Impact

**Tutorials and Workshops:** 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. She also gave a tutorial in Tsinghua University in 2008. Li will continue her effort, and prepare tutorials for the future device and circuit conferences such as DATE, ISQED, or DAC based on the new outcomes from this research.

## 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 students from NYU working with 1 PhD student from PSU on the circuit and architectural modeling in Task 1. In the second year, Xie will lead the effort in Task 2, with 2 PhD students from PSU and 1 student from NYU, to study architectural techniques using NVM technologies. In the final year, both PIs will work together with 1 PhD from each institute to study novel applications that leverage NVM technologies. Detailed project milestones are given in Figure 6.

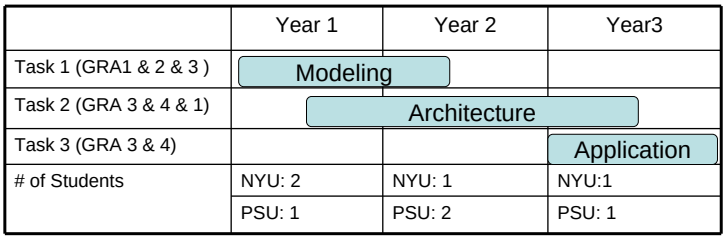


*Figure 7:* Project Management (The first student in each task would be the lead).

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 [**?**, **?**]. 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 (see attached supporting letters). 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 [?,?,?,?,?,?,?,?,?,?] and 3D EDA tools [?,?,?,?,?,?,?,?,?].

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.

# REFERENCES CITED