SRC/NSF/A*STAR Forum on 2020 Semiconductor Memory Strategies: Processes, Devices, and Architectures

I. Introduction

In 1957, Richard Feynman asked in a lecture at Caltech if it might be possible to write the entire twenty-four volumes of the Encyclopedia Britannica on the head of a pen which he argued is about one 1.58 mm on a side. (The area of the head of the pin is about 2.5 mm^2.) Feynman envisioned that each printed dot would be scanned and stored and, even at that, he reasoned that there would be an adequate number of atoms available to accomplish this feat. Today, flash memory camera SD cards rated at 32 GB are commercially available and the SD card area is about 768mm^2. Assuming that the Encyclopedia Britannica could be stored in about 1 GB of memory, an SD chip area of about 25 mm^2 would be required to store the encyclopedia rather than 2.5 mm^2. Modern semiconductor technology is within an order of magnitude of affirmatively answering Feynman's question. Looking ahead to 2020, is it reasonable to consider the possibility of storing the U.S. Library of Congress (about ten Terabytes) in a one cm^3 volume while maintaining reasonable access time, retention times, and durability?

A group of technical leaders in the semiconductor memory field met in Singapore on October 20-21 2009 to consider what kinds of solid-state memory technologies might be achievable in the 2020 time frame. This report summarizes the discussions at the Forum. The charts used by each presenter can be found at: http://grc.src.org/member/event/E003676/default.asp. In Section II of this report the salient findings of the Forum are given in the form of an Executive Summary. Section III contains brief summaries for each of the presentations to accompany the charts found at the above URL. The labeling scheme used in Section III employs a 'K' as a prefix if the presentation was a keynote address and a 'P' if the brief presentation was given by a panelist.

II. Executive Summary

II.1 Minimum space-action metric for memory technologies

It is very encouraging that there are many options for continuing the rapid progress in memory technology, even as traditional memory devices and systems appear to be reaching physical limits for continued feature scaling. Each of the candidate technologies does, however, face substantial technical challenges if it is to emerge as a provider of ever more dense memory systems with fast Program/Erase (P/E) times, long retention periods, and high endurance.

The three essential components of a Memory Device are: 1) the *Storage Node* where data is stored and whose physics of operation differs across memory devices; 2) the *Sensor* that reads the state, and is typically an electrical device such as a transistor; and 3) the *Selector* that allows a given memory cell in an array to be addressed for reading or writing, and is a nonlinear element such as a transistor or diode. All three components impact the scaling limits of memory devices. Important properties of the memory device include: i) cell size/density; ii) operating time (e.g., write, read and retention times); iii) operating voltage and energy; and iv) endurance. All known memory technologies offer compromises across the density, speed and energy space. There is always interdependence between operating voltage, speed of operation (P/E), retention time, and cell dimensions.

Each candidate technology exhibits strengths and weaknesses and since they function differently and employ different technologies, it is desirable to utilize a metric that provides meaningful comparisons between the various contenders. At the Forum it was proposed that such a metric might be developed starting from fundamentals. If *action* is defined as the product of energy and time, then natural systems seem to evolve their states in such a way that *action* is minimized, i.e., the *least-action principle* of physics. In the context of memory devices, fast P/E times and minimum consumption of energy during switching are desirable attributes. The volume of the memory device drives the density of the memory system, i.e., the *space* it occupies is also an important memory element consideration. Thus, a memory metric should also include volume and hence the product of *space* and *action* (or more explicitly, the energy-space-time product) was selected as a performance indicator for memory candidates. Optimization studies for this metric, based on first-principles physical models, were carried out for several of the candidate memory technologies and the results are given in Table II.1 below.

	$N_{carriers}$	**$V_{storage}$, nm^3**	**E_{write}, J**	**t_{write}, ns**	**Space-Action, J·ns·nm^3**	**Biggest component**
DRAM	10^5	10^5	10^{-14}	1	~10^{-8}-10^{-9}	Storage node
Flash	10	10^3	10^{-16}	10^3	10^{-9}	Sensor (FET)
STT-MRAM	10^5	10^3	10^{-14}	1	10^{-13}	Selector (FET)
ReRAM	100	1	10^{-17}	1	10^{-14}	Selector (FET)
Comments		*V_{FET}~10^3 nm^3*			*with FET*	

Table II.1. The estimated minimum space-action metric for various memory technologies

One could argue that there are other properties of memory devices that are also important. For example, could endurance as well as long retention-time, be incorporated into the *space-action* metric? An indirect argument for endurance is that by minimizing the switching energies as implied by the space-action metric, there should be a positive impact on the number of cycles that a memory device can successfully perform. With respect to retention, this was included as a performance constraint during the minimization of the *space-action metric.*

II.2 Overview of Forum Discussions on Specific Memory Technologies

The six-transistor static random access memory (SRAM) is widely used as an embedded memory for many high performance applications and offers the fastest P/E times of all known or emerging memory technologies. There seems to be a general consensus that due to the gradual degradation of field effect transistors with scaling (leakage currents increasing, etc.) and increased failure rates, the SRAM may not survive many more scaling generations. However, adaptive and supply voltage control techniques along with special circuit techniques (with increased transistor count from six to eight of even ten transistors) are being used to manage some of these issues for SRAM technology, albeit at the cost of additional circuit complexity.

The dynamic random access memory (DRAM) has been the memory element of choice for processor level-two memory for many product generations. Like the SRAM, the DRAM is a volatile memory element that must continuously be refreshed due to charge leakage from the storage capacitor. As feature sizes continue to be scaled, the degraded cell transistor consumes more leakage current requiring the footprint of the capacitor be scaled accordingly and therefore innovations in stacked and trench capacitors are required to achieve the necessary capacitance for successful operation. The DRAM cell offers a footprint on the silicon that is about one order of magnitude smaller than that of the SRAM but it has somewhat slower P/E speeds than does the SRAM. The embedded DRAM (eDRAM) does require additional process steps to form the trench capacitor but this can be accomplished before the logic fabrication steps. Some companies, notably IBM, have incorporated eDRAM into processor designs. Research is underway for modified eDRAM structures that do not require a capacitor but rather utilize a double gate quantum-well structure incorporated into a single floating-body transistor. This is an extension of the floating body concept for memory cells wherein a single transistor is used and charge is stored in the quantum well. Such a structure provides about a 5x increase in memory density relative to an SRAM cell and power reduction due to decreased capacitance. However, these benefits come at a cost of more expensive SOI wafers.

The progress made in FLASH memory during the past decade has been remarkable and this memory technology is now ubiquitous across a wide range of consumer products. FLASH has even begun to make inroads as an alternative to magnetic disk storage for some highly portable computing applications. Nevertheless, current FLASH memory elements (floating-gate and charge-trapping) face several challenges as scaling continues beyond the ITRS 45 nm node. The root challenge stems from the difficulty in continuing to scale the gate oxide which manifests itself as a decrease in retention time for FLASH memory elements. The terminology *voltage-time dilemma* is often used to characterize this problem because of the conflict between high energy barriers required for retention and the need to offer higher P/E times for FLASH operations. It is believed that by improving electrostatics by use of nanocrystals and nanopores, incorporation of specially designed tunnel barriers, and by utilizing 3D stacks of thin germanium channels, FLASH scaling might be continued for few more generations. An alternative FLASH (SONOS) memory cell has been proposed that may circumvent some of the problems associated with FLASH scaling by the use of vertical silicon nanowires to form transistors with a wrap-around gate structure. Preliminary findings suggest that these structures could offer a lower silicon footprint with good performance metrics and avoid some of the electrostatic problems of conventional FLASH structures. Since this is a research project, many fabrication and process integration issues remain to be addressed. At this time the cost-effective integration of these novel flash technologies with CMOS remains a challenge.

One of the attractive features of magnetic random access memory (MRAM) technology is its compatibility for integration with CMOS Back End of the Line (BEOL) processes. Read times for MRAM devices appear to be intermediate between those of the SRAM and the DRAM and the cell size is similar to that of the eDRAM. In Spin Torque Transfer MRAM (STT MRAM) cell structures, current-induced domain wall movement is used to reduce write current and provide good memory stability with a smaller cell size. It appears that further improvements in STT MRAM technology are possible if the cell could be fabricated in a vertical orientation.

Ferroelectric materials have also been investigated for use in the creation of memory cells. Several types of cells have been proposed including capacitive, and the ferroelectric gate metal-insulator field effect transistor (MISFET). A major issue that has arisen with these classes of ferroelectric memories is their relatively short retention times due to interface defects.

Resistive random access memories (ReRAM) store information in the state of a resistor that can be set to controllable binary values. Usually, each element in a ReRAM array is connected in series with a diode that acts to ensure that the element does not cross-couple to other devices in the array. There are many types of ReRAM memory elements. In the fuse/anti-fuse ReRAM, conducting filaments are formed/dissolved by the application of a current of proper duration and amplitude across the element. Electron-effect based resistive memory elements change resistive states due to the movement of electrons into traps. In metal-ion or oxygen-ion ReRAMS, the migration of these heavier elements is used as the basic mechanism for resistive switching. The memristor, recently reported by Hewlett-Packard, is a member of the ion-migration class of resistive elements. Each of these classes of memories offers many desirable attributes for embedded non-volatile memory elements but each also faces technical challenges. For example, poor retention is characteristic of all classes with the exception of the fuse/anti-fuse of ReRAM devices. Several of these classes of ReRAM devices suffer from lack of scalability and poor uniformity.

One challenge faced by ReRAM devices is that due to their two-terminal structure, all control and sensing functions must be implemented via current pulses of specified sign, magnitude, and duration. This dependence on well defined current pulse characteristics may result in increased sensitivity due to fabrication variations across the chip. Moreover, the constraint of access to the devices through only two terminals may not provide adequate separation between the signal domains for P/E and memory state sensing. The phase change random access memory (PCRAM) avoids the problem of effecting all changes through the two terminals of the resistive element by employing an embedded heating element that is used to change the material phase of the PCRAM. Four distinct material phases have been observed for calcogenide materials offering the promise of a two-bit storage capability in a single device. The PCRAM memory element appears to be consistent with the BEOL processes and offers good stability, long-term storage, and speed of response. However, it will be sensitive to external temperatures, for example during solder reflow with packaging, it requires relatively high programming currents, and the PRAM is subject to resistance drift with time.

Although polymer memory systems were not the primary focus of the Forum, discussions at the Forum indicated that polymer materials can be used for a range of memory technologies, many of whose operational properties mimic those of semiconductor memories discussed above. Polymer memories hold the promise of low cost fabrication and a flexible substrate; however, to date they do not compare favorably with semiconductor memories in terms of speed of operation, density, and durability.

A challenge for embedded memory systems is compatibility with CMOS processing technologies. As an example, in System-on-a-Chip designs, there has been a trend to increase memory content at a rate that exceeds the rate of increase in logic content. In microprocessor applications, there is also an increased need for embedded memory systems to support the growing number of multi-core processors on a single chip and to provide much higher bandwidth for memory access in general.

The memory bandwidth needed to support multi-core architectures will soon be on the order of 1 Terabit per second, and, if current trends continue, operation of such memory systems will require the expenditure of more than one hundred watts. One approach to decreasing power consumption is to reduce the operating voltage for memory elements to the range of ~0.6 volts. This would reduce CV^2f losses and improve energy efficiency. However, sub one-volt operating levels are not compatible with many memory technologies. Through-silicon-via (TSV) technology is emerging as an option that could provide orders of magnitude increases in memory bandwidth while simultaneously decreasing memory power requirements. Alignment accuracies on the order of one micron and half pitches of five microns have been reported for TSV technology. A challenge for TSV technology, however, is to continue to scale alignment and pitch dimensions in such a way that they track scaling of logic chip dimensions.

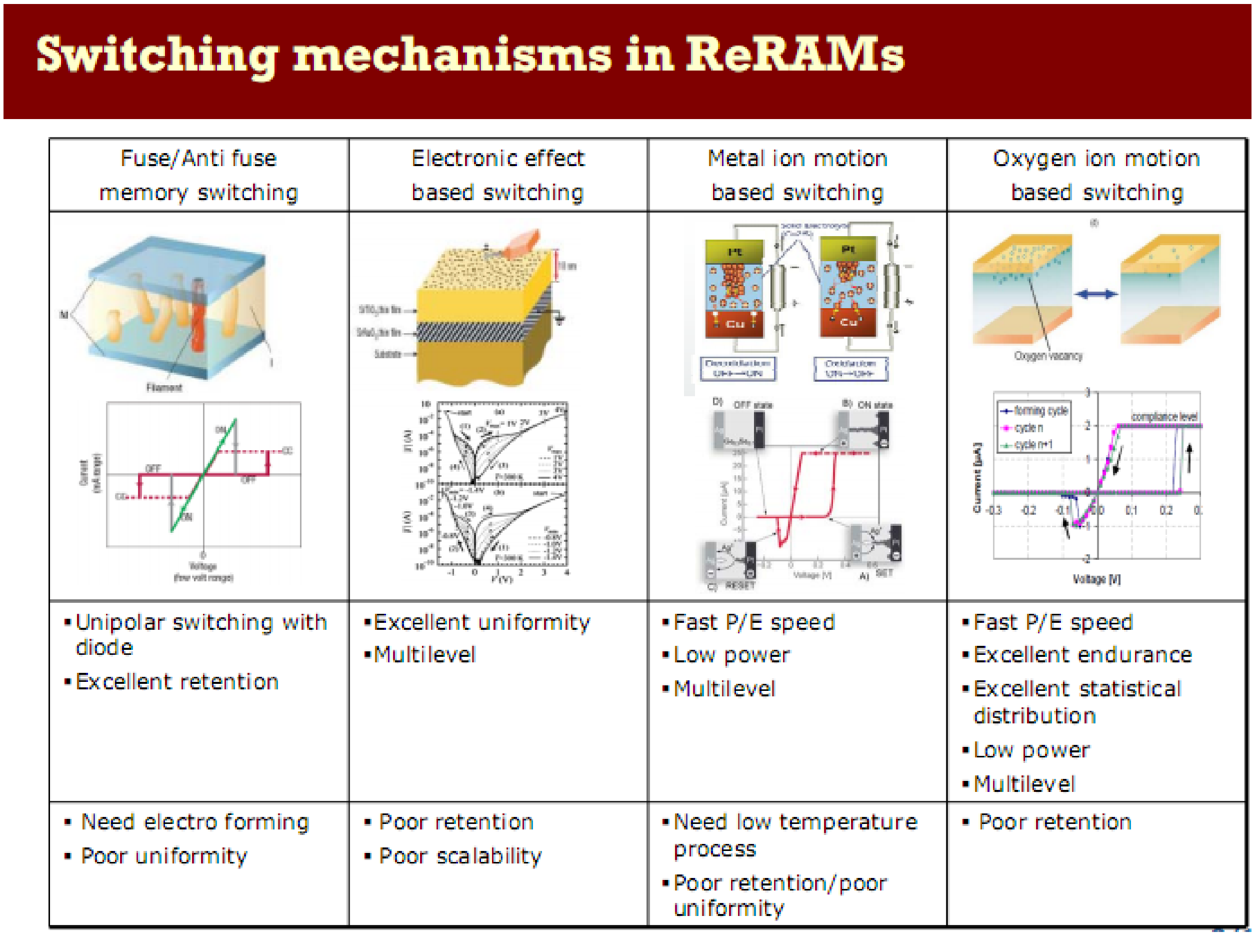
III. Presentation Summaries

Session 1: Novel Memory Devices

K1.0 In. K. Yoo, Samsung: "Perspectives on ReRAM"

Dr. Yoo began by taking a broad view of memory evolution from the points of view of *(i)* Society*, (ii)* Products and *(iii)* Technologies. Societies tend to operate in the domain of 'Contextualism' which uses the principle of inductive reasoning. If A implies B and A implies C, then B and C are probably related. {If a cloudy sky implies rain and a cloudy sky implies snow, then rain and snow are probably related.} 'Technologies tend to be based on Essentialism or deductive reasoning, i.e., if A implies B and B implies C, then A must imply C. {If a falling barometric pressure implies increased clouds and increased clouds imply rain, then falling pressure implies rain.} However, the development of products involves a different kind of thinking that Dr. Yoo calls "What-ism" or Abductive reasoning. If A implies C and B implies C, then A and B must be related in some way. {If very dense memories result in increased market share and memories with fast access times give increased market share, then fast and dense memories would likely result in increased market share.} Abductive reasoning results in the expansion of knowledge as it is ultimately a form of forecasting.

It appears that there are several application areas for ReRAM (Resistive Random Access Memories); two terminal devices whose resistance valve can be changed by the appropriate current pulse sequence. Some of these applications include one-time programmable memory for Digital Rights Management, high capacity high speed memories for digital cameras, tunable resistors for analog applications and for some (oxide) materials transparent electronic memories may be possible. There are several different switching mechanisms in ReRAMS including *i)* Fuse/anti-fuse switching, *ii)* electron effect based switching, *iii)*metal ion motion switching and oxygen ion motion-based switching. Each of these categories of devices appears to have major strengths and weaknesses. Data retention and uniformity of operation are weaknesses of almost all devices in these categories; however, oxygen ion motion appears to offer more advantages relative to the other ReRAM categories. See Figure III.1 below from Dr. Yoo's presentation:


Filament
Pt
Cu
OFF state
ON state
RESET
SET
Oxygen vacancy
forming cycle
cycle n
cycle n+1
compliance level
Current [µA]
Voltage [V]

Switching mechanisms in ReRAMs

Fuse/Anti fuse memory switching	Electronic effect based switching	Metal ion motion based switching	Oxygen ion motion based switching
▪Unipolar switching with diode ▪Excellent retention	▪Excellent uniformity ▪Multilevel	▪Fast P/E speed ▪Low power ▪Multilevel	▪Fast P/E speed ▪Excellent endurance ▪Excellent statistical distribution ▪Low power ▪Multilevel
▪ Need electro forming ▪ Poor uniformity	▪ Poor retention ▪ Poor scalability	▪Need low temperature process ▪Poor retention/poor uniformity	▪ Poor retention

Figure III.1. Properties of Four classes of ReRAMs

The presentation then focused on application of a percolation cluster model for resistance switching in which fractal cluster dimensions play an important role.

P1.1 D. Strukov, University of California at Santa Barbara: "Memristor as a New Memory Element"

Hewlett Packard Company has developed a memristor based on the use of TiO2 materials that effects resistance change due to oxygen vacancy migration. (In memristors, the voltage across the device is dependent on the rate of change of the device current.) Dr. Strukov then described bulk and interface models for a memristive element in which barrier height is modulated to oxygen ion movement. A general conclusion of the HP research group is that a strong nonlinearity in ionic transport is required for high retention by the memristor.

A variety of CMOL (CMOS Molecular) architecture were described that could be utilized to develop memristor memory arrays. There are, however, substantial tradeoffs with both device resistance and switching speeds. For example, low memristor resistance is usually accompanied by electromigration issues, less dense arrays and smaller readout margins. Conversely, higher device resistance is associated with higher volatility, slower response and larger switching voltages. Lower switching speeds usually signify increased volatility while faster switching speeds are usually accompanied by lower endurance and less repeatability when in operation. The presentation concluded with a discussion of work needed to optimize circuit architectures that would embody work in device characterization, comprehension of design constraints and consideration of fabrication requirements.

P1.2 M. G. Erotsum/K. Saraswat, Stanford University: "Capacitorless Double Gate Quantum Well Single Transistor DRAM."

In a 1-transistor DRAM, the cell senses holes accumulated in the floating body as the threshold voltage changes; i.e., in these devices, I_D=f (V_g, V_{th}). The 1T DRAM size is $4F^2$ and requires conventional CMOS materials and processes. Scalability issues include lithography constraints and sensing widnow insuffficiency. This talk describes an extension of the 1T DRAM concept that retains most of its advantages, and some extra advantages such as the introduction of an extra 'storage pocket' and the ability to tune spatial hole distribution, but does require introduction of new materials. In particular, a vertical double gate capacitorless-single transistor DRAM has been demonstrated and characterized. By reverse biasing the back gate, memory operation was obtained for scaled, fully-depleted devices.

A novel 1T-quantum well structure has been proposed that provides the opportunity to engineer the spatial hole distribution within the device body. It turns out that the spatial location of the quantum well is an important control parameter whose effect on V_{th} increases rapidly as body thickness decreases. The use of SiGe in the fabrication of the quantum well provides improvement in erased-state degradation as well as easier fabrication.

P1.3 Patrick Lo Guo-Qiang, IME: "NV-Memory Elements with Gate All Around Transistors"

A technology platform has been developed at the Institute of Microelectronics for vertical silicon transistors with a wrap-around gate that appears to be a viable candidate for sub-22nm ITRS nodes. It can be shown that electrical gate oxide thickness for the gate-around vertical nanowire transistor will scale less rapidly than the physical oxide thickness. In addition, the electric field is greatly increased relative to the planar double gate transistor. Another benefit is that the silicon footprint of the vertical nanowire is equivalent (in feature size) to the planar transistor. This could greatly reduce circuit size and estimates are that the vertical transistor will be ~6 times faster and consume ~3 times less power than a comparable planar transistor.

The vertical nanowire transistor can be used to construct SONOS-class flash memories whose electrostatics are sufficient to improve (P/E) performance by increasing threshold voltage swing over time. Decreasing nanowire diameters from 8nm to 5 nm has been shown to further improve P/E performance. Looking ahead, in an effort to develop high-density and high performance memory devices, plans are to arrange multiple cells on each vertical pillar and to further enhance P/E fields. It appears that these memory elements will have good retention and endurance properties.

P1.4 Edwin Kan, Cornell University: "My (Optimistic) 3 Cents to Flash Scaling"

The list of flash memory scaling challenges is quite long and includes inability to scale tunnel oxide to give 10 year retention, stress induced leakage currents, lack of cost effective 30 channel stacking, high P/E voltage requirements, insufficient cycle life times, etc. Dr. Kan pointed out that there is a rich set of material choices for each of the components of flash memory. He also indicated that the key for effective scaling is the ability to design and control electrostatics in the gate stack, and that nano-crystals can be used to modify gate stack electrostatics. For example, if C_{60} is used in the gate stack, then enhanced flat band shift can be obtained and resonant tunneling through C_{60} can significantly increase the ratio of retention time to P/E times. If a bi-layer tunnel oxide (e.g. S_iO_2 and HfO_2) is used in conjunction with metal nanocrystals, Fowler Nordheim tunneling occurs during P/E and direct tunneling during retention.

Further voltage and density scaling can be enabled by improved electrostatics via the use of nanocrystals and nanopores, by specifically designed tunnel barriers, and by planar 3D stacks of Ultra Thin Body Chemical Mechanical Polished Ge Channels.

P1.5 H. S. Philip Wong, Stanford University: "Memory Cell Selection Device Enabling High Density Memory"

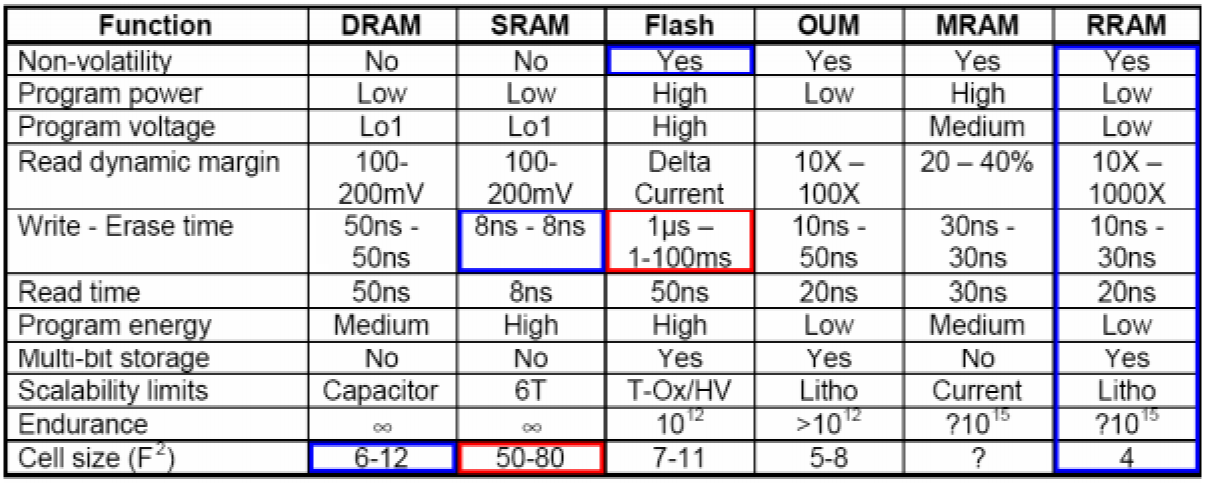
Professor Wong focused first on the Cross-Point Memory Cell Selection Device memory arrays. The function of the selection device is to provide a margin for read signal to noise, to insure that memory cells can be written and to control power dissipation. If one considers an array of vertical nanowire memory cells, then it is reasonable to consider integrated selection diodes with each memory cell. The requirements for the diode include; the diode must be scalable to $4F^{2-}$, it must operate with a high current density, sustain a large (10^4) I_{ON}/I_{OFF} rates and it must be fabricated with a low temperature process that is consistent with CMOS processing.

Professor Wong described a memory cell due by Samsung that meets many of these requirements and that can be fabricated using conventional processes. In another example, it was shown that the use of Ti between Si and TiN can dramatically reduce diode contact resistance. The Stanford group has utilized germanium (Ge) nanowires for diode fabrication with the result that programming current is reduced and nanowire synthesis can be achieved at temperatures <350^0C. Another promising example of the Ge diode used in conjunction with phase change memory was discussed.

Session II: Prospective Materials for Memory Applications

K2.0 Professor Tseng, Tseung-Yuen, National Chiao Tung Univesity: "Materials for Future Memories"

An ideal non-volatile memory has low operating voltage, low power consumption, long retention time, non-destructive readout, a simple structure, low cost, etc. Professor Tseung-Yuen began by introducing two major classes of emerging non-volatile memory elements: the capacitive type and the resistive type. The capacitive type includes the Ferroelectric RAM and the resistive type includes a variety of resistive RAMs (e.g., the Atomic Switch, Fuse/Anti-fuse RAM, etc). Candidates considered in this light are capacitive FeRAM, the ferroelectric-gate FET RAM, the STT-MRAM, PCM (Phase-Change memory), the Floating Body DRAM, and the resistive RAM (ReRAM). Figure III.2 gives an overview of the parameters for a variety of memory technologies.



Function	DRAM	SRAM	Flash	OUM	MRAM	RRAM
Non-volatility	No	No	Yes	Yes	Yes	Yes
Program power	Low	Low	High	Low	High	Low
Program voltage	Lo1	Lo1	High		Medium	Low
Read dynamic margin	100-200mV	100-200mV	Delta Current	10X – 100X	20 – 40%	10X – 1000X
Write - Erase time	50ns - 50ns	8ns - 8ns	1μs – 1-100ms	10ns - 50ns	30ns - 30ns	10ns - 30ns
Read time	50ns	8ns	50ns	20ns	30ns	20ns
Program energy	Medium	High	High	Low	Medium	Low
Multi-bit storage	No	No	Yes	Yes	No	Yes
Scalability limits	Capacitor	6T	T-Ox/HV	Litho	Current	Litho
Endurance	∞	∞	10^{12}	$>10^{12}$	$?10^{15}$	$?10^{15}$
Cell size (F^2)	6-12	50-80	7-11	5-8	?	4

2002 IEDM Zhuang et.al. Sharp

Figure III.2. Properties of several classes of semiconductor memories

Professor Tseung-Yuen first contrasted the categories of materials (PZT and SBT) for FeRAM memory elements and briefly characterized the advantages/disadvantages of both material systems. FeRAM memory elements can either be of the capacitive type or of ferroelectric gate FET type which includes the Metal-Insulator-Ferroelectric Field Effect Transistor (MFISFET) type and the Metal-Ferroelectric-Metal-Insulator Field Effect Transistor (MFMIFET) type. He showed several examples of industrial FeRAM prototypes and introduced some of their operating properties. Some of the materials issues associated with capacitive FeRAM are:

- Protection of ferroelectric material from hydrogen damage
- Maintenance of high remnant polarization and reliability
- Oxidation resistant plug between plug and bottom electrode
- Effect of CMOS thermal budget and contamination.

A limiting factor for capacitive FeRAM scaling is the 2D capacitor area; this might be addressed by moving to 3D capacitor structures and resolving electrode coverage uniformity with ferroelectric issues.

The other type of FeRAM device is the Ferroelectric Gate Field Effect Transistors where the ferroelectric material is in the gate stack. Prof. Tseung-Yuen reviewed the Ferroelectric gate FET operating principles and provided data on retention time for a range of materials. A major issue with this class of memory is the relatively short retention times (~few to 30 days). The problem arises from the introduction of defects and hence traps at interfaces during processing. Solutions are based on the introduction of insulating buffer layers between the ferroelectric material and silicon.

Next, ReRAM materials were discussed. There are a number of switching mechanisms for ReRAMs including filamentary, interface-controlled, phase change and solid-state electrolytes for ion migration. The ReRAM cell typically includes a diode select element and a resistor as the memory element fabricated at the BEOL; although a transistor-resistor configuration is also sometimes used. The latter configuration has a cell size of $6F^2$. There is a wide array of candidate materials for implementing the ReRAM. The thin film properties of the ferroelectric materials are influenced by many factors including film composition and dopants, crystalline structure, microstructure, surface morphology, etc. Illustrations of some of these effects were presented.

In summary, the 3D ferroelectric capacitor structures, ferroelectric materials with higher P_r, lower crystallization temperatures, and formation of thin films are needed for high density ferroelectric memory. Complex high-k metal oxides are among possible materials that can be used in future non-volatile memory NVM applications as the insulating buffer layer in Metal-Ferroelectric-Insulator Silicon Field Effect Transistor memory elements. In these structures, high k materials with high crystallization temperature and low leakage currents are needed for the buffer insulation layer in order to improve retention times.

In the case of ReRAMS, composition, non-stochiometry, dopants, crystallization and thickness of resistive switching films all relate to the definition of switching characteristics.

P2.1 L. P. Shi and T. C. Chong, A*STAR: "Phase Change Materials and the PCRAM"

The principles of operation of a PCRAM, based on the change of material phase, were briefly reviewed. The class of chalcogenide-based phase change materials was introduced and the required properties of these materials were connected to the write/read/erase/stability/endurance performance requirements of the PCRAM. Going beyond available materials, it was suggested that super lattice materials could be engineered to reduce required currents and increase operating speeds. The important factors relating to the ultimate scalability of the PCRAM at the nanoscale are the ratio of atoms at the interface to atoms in the volume and the mean free path of electrons in the phase change material. It was estimated that the scaling limit for the PCRAM is about 5 nm. The PCRAM has many of the attributes of a universal memory element combining non-volatility, high speed (~20-100 ns), high density ($5\text{-}7F^2$) and high endurance ($>10^{12}$ cycles).

P2.2 Hero Akinaga and Hisashi-Shima, AIST: "Materials for ReRAM"

In the cell area/cell clock frequency space of competing non-volatile memories, the ReRAM requires a cell area of about 5 F^2 and operates at a clock frequency of 1 GHZ; both parameters are near the best offered by any of the competing technologies. It is believed that its CMOS compatibility and scalability should make the ReRAM an attractive option for future non-volatile memories. The cycle of setting and resetting a ReRAM involves ion migration in the filament to obtain low and high resistance states. Below 20 nm features, electrochemical interactions, not filamentary connections, brings about non-volatile resistance switching. However, the authors argue that these two mechanisms are complementary and do not conflict with one another. The Gibbs Free Energy Diagram provides the guiding principle for selection of the proper combination of electrode and oxide layer.

It was argued that the current status of ReRAM technology includes:

- CMOS Compatibility
- No physical scaling limits
- Ultra-fast operation (~10 ns)
- Large ON/OFF ratios (~10^3)
- Reasonably low power operation (~I mA)
- Reasonably high endurance (~10^6-10^9 R/W)
- Reasonably good retention at elevated temperatures (~150 °C)

Key challenges include:

- Synthesis of a highly reproducible oxide interface
- Decrease operating current for higher endurance
- Realization of zero current, resistive switch.

P2.3 N. Ishiwata, NEC: "MRAM: Materials and Devices – Current – induced Domain Wall Motion High Speed MRAM"

In comparison with other memory technologies, the Spin Transfer Torque MRAM is non-volatile, has unlimited endurance (>10^{15} cycles), very fast access time (~10 ns) and can be used for working memory applications. In the access frequency/cell size space, the MRAM is at about 200 MHZ with a cell size of 12 F^2. A typical Domain Wall Motion STT-MRAM cell has two transistors and a Magnetic Tunneling Junction. The key issue this structure addresses is the reduction of write current which needs to be less than 0.2 mA. Spin-Transfer Torque Switching is thought to be a promising approach to lowering the write current. One way to achieve spin transfer torque switching is through current-induced domain wall motion. Generally, domain wall motion has many positive characteristics and is applicable to two transistor/one magnetic tunneling junction structures. Simulations have shown that materials with perpendicular magnetic anisotropy (PMA) provide a smaller cell size and better memory stability. CoNi is judged to be best material because of smaller current density accompanied by a large pinning field and high velocity. Dr. Ishiwata has shown that using CoNi, the write current is less than 0.2 mA for device widths less than 100 nm and a write time of 2 ns can be obtained. A 4 Kb memory has been demonstrated for this material system.

P2.4 C-X Zhu, National University of Singapore: "Polymer Electronic Memories: Materials, Devices, and Mechanisms"

Polymer memories can be either of capacitive type, transistor type or resistor type. In polymer ReRAMs, data storage is based on electrical bi-stability of materials arising from changes in intrinsic properties. The polymer ReRAM structure is relatively straightforward and would, in practice, contain a diode to avoid parasitic currents. ReRAM of the filament class can be fabricated using either carbon-rich or metallic materials. A variety of alternative ReRAM structures have been reported including, space charges and traps using poly (N-vinylcarbazole)- carbon nanotube composite films. Polymer ReRAMS have also been realized based on conformation change effects, from charge transfer effects (for SRAM memories), and from polymer fuse effects. Other polymer ReRAM demonstrations have utilized ionic conduction and tunneling.

P2.5 Pierre Fazan, Innovative Silicon: "ZRAM Floating Body Memory – Materials, Devices and Processes"

The Floating Body Memory (FBM) exploits isolated (SOI) transistors floating body (FB) effects, where the floating body is used as a storage node. The presence of carriers (holes) in the FB defines one memory state whereas, the emission of carriers (holes) by junction forward biasing providing electrons in the FB defines the second memory state. State identification is obtained by V_{th} changes and current sensing. In 2007, a MOS transistor in conjunction with an intrinsic bipolar junction transistor induced by the floating body was shown to demonstrate increased margin, longer retention time, higher read current, less variability and increased scalability. Applications include stand-alone DRAM replacement (4 to $6F^2$ cell) or embedded DRAM or SRAM replacement (35 to 45 F^2 cell).

The roadmap for stand-alone memory calls for the realization of 3D devices on bulk silicon and, for embedded memory planar devices on SOI. Materials challenges for stand-alone memories include: 3D devices, engineered substrates, low-resistive gates, contacts and lines, interface quality, tighter pitch contacts and low resistance interconnect. Challenges for embedded applications include engineered substrates, partially and fully depleted SOI, ultra-thin silicon and ultra thin buried oxide (BOX), interface quality and defect reduction. SPICE and TCAD models for the FBM are also needed. Process integration challenges for stand-alone memories include: 3D integration, optimum doping for the MOSFET and the intrinsic BJT, self aligned contacts and tight pitch interconnect. In the embedded case, ultra-thin Si device integration, optimum doping, and elevated source/drains are process integration challenges.

Session III: Memory Architectures

K3.0 Paul Franzon, North Carolina State University: "Architecture for Extremely Sealed Memories"

It appears that the memory bandwidth requirement for multi core architectures is fast approaching 1 Tbps. Graphics, networking and computing systems will soon have similar bandwidth requirements. An important challenge is to provide this access bandwidth with greatly reduced power since estimates are that power requirements would be of the order of several hundred watts. This power is consumed by cells at about 81 fJ per bit, by row decoders, sense amps and relatively inefficient charge pumps to supply ~2 V to the core, to command, address, data pipeline and 'assist' circuits and to Input/Output circuits. If memory circuits operated at 0.6 V, there would be a reduced need for charge pumps and a net core power reduction.

Dr. Franzon pointed out that DRAM failures today are almost all due to packaging and that future scaled processors could well spend 80% of their time checkpointing if soft error rates are not improved. With respect to the percent of memory area used for cells, most DRAMs have a fill factor of 30% to 40%. From the perspective of cell size, it appears that ReRAMS may offer the best opportunity because F, the interconnect half-pitch, can be small.

Opportunities for memory technologies arise in several areas. The trend toward 3D integrated circuits fabricated with TSVs could provide for increased memory bandwidth. There are architectural and processing tradeoffs that arise from the choice of coarse versus fine pitch TSVs. In a synthetic radar application, an FFT processor realized using 3D TSV technologies resulted in a 65% power reduction and an 800% increase in memory bandwidth relative to the planar counterpart. A detailed analysis of the reductions in metrics such as area, wire length, power, etc. for the FFT 3D TSV processor was provided in the presentation. An important ancillary design benefit is that floor planning, TSV placement, and partitioning are made easier in 3D. It should be pointed out that TSV technology sometimes is accompanied with a small area penalty relative to planar technology. Some of the 3D IC issues include cost for low volumes using 12" equipment, Known Good Die, and thermal management issues.

Looking ahead to exascale computing, extreme stacking will be needed to manage energy and bandwidth requirements. The 3D IC RAM will create architectural opportunities including separation of the memory array structure from the architectural specification, co-optimization of floor-plan, logic, and memory, and potential removal of the need for L2 cache. There is a need for fast compact memory to support neuromorphic computing. For example, a highly programmable analog FET based on a nano-crystal metal floating gate exhibits a high density of states, is reliable, and has good retention properties.

P3.1 John Barth, IBM: "eDRAM to the Rescue"

The integration of logic with DRAM using a deep trench process should be logic friendly since the trench capacitor could be fabricated first before the logic processes. Indeed, if SOI is used, the buried oxide can be used to simplify the process and reduce parasitics, and the pass transistor can be scaled for higher performance. In the Itanium®- 2 processor, 9 Mb of on-die L3 Cache was provided using eDRAM technology. In one application, it was found that the use of eDRAM rather than SRAM for a CPU resulted in a chip that was 43% smaller, but 2x slower. It appears that at the 45 nm node, eDRAM and SRAM total latencies are approximately the same when memory size approaches 64 Mb. The eDRAM is being used to advantage in IBM processor systems; recently the eDRAM was integrated with a processor in SOI technology at the 45 nm node.

P3.2 Wei Wang, SUNY-Albany: "Nano-ReRAM for Novel FPGA Architectures"

The focus of this presentation was on emerging memories for FPGA applications; e.g., CNT, Graphene, MRAM, and ReRAM. For example, cross bar switches can be implemented by a one-transistor, one-resistor (1T1R) circuit as demonstrated by Spansion in 2005. ReRAMs can provide 4x to 6x density increases and power improvements for FPGA circuits relative to the SRAM. In addition, routing resource utilization can be reduced by 2x to 3x using 1T1R memory elements, again relative to SRAM memories. Comparisons of 2D CMOS with 2D rFPGA {ReRAM memory elements} shows an area improvement of 2x-3x, a static power improvement of 10x and delay, dynamic power reductions of ~20%.

P3.3 Kevin Zhang, Intel "Low Power and High Performance SRAMS in Nano-Scale CMOS Technology"

The SRAM continues to be very important for processor applications. At the 32 nm node, the bit cell area is ~0.2 μm^2. There are, however, a number of challenges with continued SRAM scaling including the problem of transistor mismatch at the cell level. In addition, the R/W window is shrinking with conventional scaling and correction of this trend by adding transistors extracts a steep overhead penalty.

By creating a differential voltage between the word line and the cell supply voltage by multiple supply voltages, the R/W window can be expanded, and a reduction in leakage/stand-by power can be achieved. Also, dynamic 'sleep' management of supply voltage to reduce power consumption of inactive cells is a viable technique to manage array power consumption. Indeed, it appears that adaptive design of SRAM cells along with process innovations will enable continued use of this workhorse memory technology.

P3.4 Jimmy Zhu, Carnegie-Mellon University: "Low Power Crossbar MRAM with Scalability"

This presentation considers a diode/MRAM as a memory crossbar element. The estimated switching energy for spin torque transfer MRAM devices is in the range from 0.1 pJ to 1 pJ. An alternative approach is to use FeRh whose magnetization state can be altered by heating. This memory device is estimated to have a switching energy of .01 pJ and switching speed on the order of a few ps. Since a diode would be included in the memory cell, the need for a transistor is removed. The memory cell footprint is on the order of 4 F^2 and, moreover, the cell is expected to be stable even below 10 nm. An areal density of 1.5 Tb/in^2 is projected.

P3.5 Leon Chua, University of California at Berkeley: "The Memristor as a Memory Element"

Professor Chua, who argued for the existence of the memory resistor almost four decades ago based on a device voltage-current-charge-flux taxonomy, gave a brief overview of memristors and their possible applications. He argued that memristors are realized in the structure of the brain including in synapses and axons. He gave an illustration of the use of memristor to obtain a déjà vu response to excitations. (A déjà vu response is learning to recognize and ignore benign and boring stimuli.) Associative memories have also been demonstrated with memristive neural networks. Professor Chua concluded by noting that memristive circuits have the potential to create a new market for intelligent hardware capable of adaptive intervention.

Session IV: Technological Platforms for Future Memories

K4.0 Masao Fukuma, NEC: "Challenges of the Integration of New Memory into CMOS Platforms"

Both memory capacity and memory to logic bandwidth requirements are increasing for SoC applications at a faster rate than logic performance requirements. This growing need is compounded by the fact that all conventional memory technologies have fundamental scaling limits which will severely impact their ability to meet these needs. In particular SRAM suffers from large leakage currents, small operating margins and complicated circuits. DRAM is challenged with the required storage capacitor size and the difficulty in obtaining low leakage transistors. FLASH suffers from the non-scalability of dielectrics and poor reliability. To meet these growing memory needs, it is clear that only embedded memory technologies can simultaneously provide higher bandwidth, lower power, reduced pin count and board area and lower noise generation. The relative performance of eSRAM, eFLASH, eDRAM, DRAM and FLASH are summarized in Figure III.3.

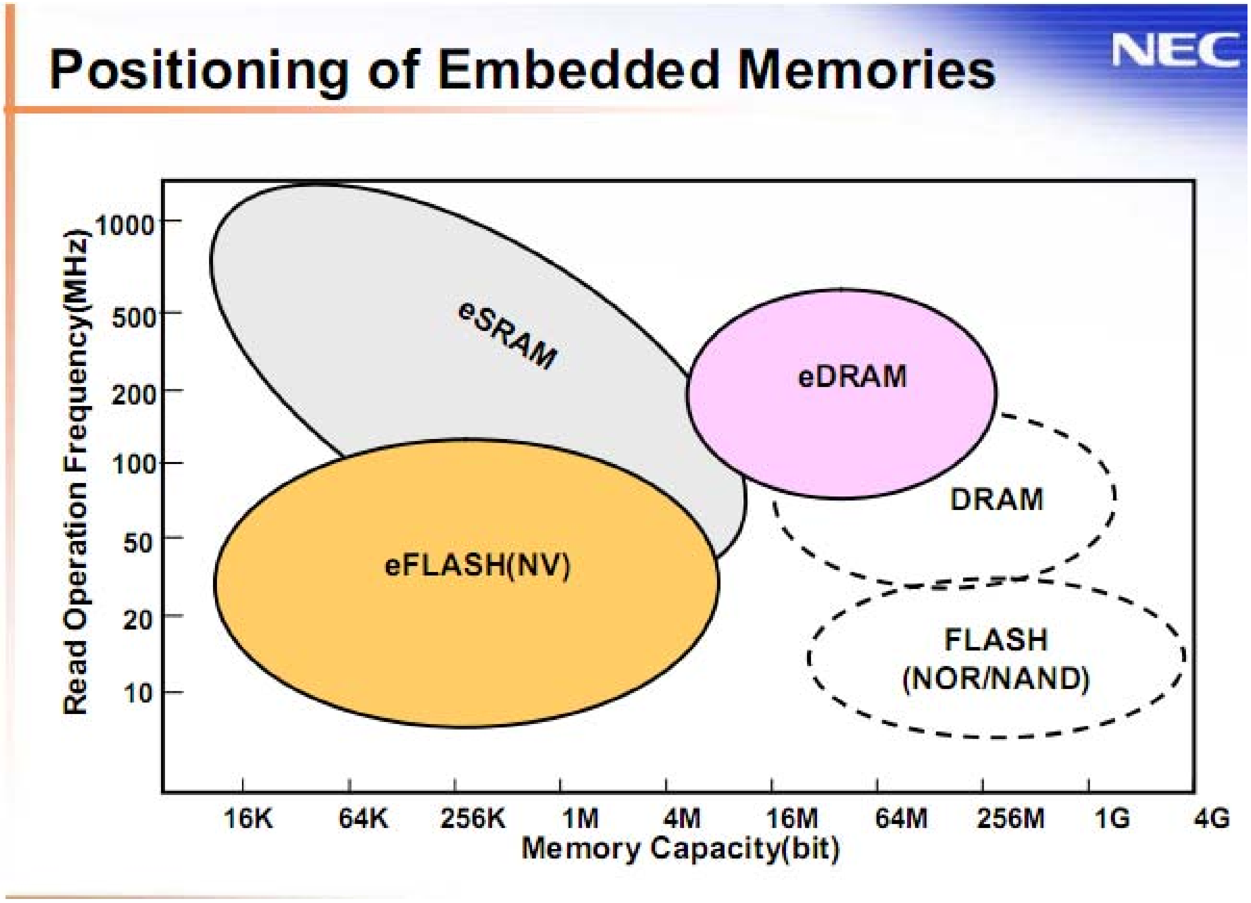


Figure III.3. The performance space for several semiconductor memory technologies

Dr. Fukuma pointed out that the integration of memory in the BEOL allows customization of BEOL for memory integration while leaving the Front End of the Line (FEOL) as a standard CMOS process. There are many new candidates for Back End of Line (BEOL) memory integration including RAM, FeRAM, PCRAM, Nanobridge, etc., which have been reported in the literature. Three added process steps are required for MRAM but there is no effect on logic circuits and the MRAM cells can be inserted at any level above the first metal layer, M1. There have also been reports of the integration of Filamentary ReRAM and FeRAM cells at the top metal layer. Nanobridge cells have also been demonstrated as a nonvolatile electrolyte switch embedded in Cu interconnect.

After the 22 nm node, CMOS processes may undergo significant changes including the use of different device structures, e.g., nanowire transistors, new materials and new processes. It is believed that design complexity and variability will increase and that reliability will decrease. The philosophy of manufacturing and design will probably follow the trajectory from 'Never produce defects' to 'Never let defects escape' to 'Never let defects lead to failure.' The mitigation to prevent failure from defects will likely require a change in self-checking programmable logic cells that self-detect faults in real-time and automatically correct them. Of course, continued scaling must be accomplished in the face of rapidly rising mask costs for lithography and stringent smoothness requirements for the mirror in the Extreme Ultra Violet Lithography (EUVL) tool. Some of these effects can be reduced by using regular architectures with dynamic reconfiguration of programmable hardware. There has been progress at NEC toward the design of power efficient lithography friendly circuits that have virtual/autonomic capability.

P4.1 T. Hamamoto, Toshiba: "Floating Body RAM (FBRAM) : Overview and Future Challenges"

In contrast to the DRAM cell, that stores charge in a capacitor embedded in a bulk silicon substrate, the FBRAM stores charge in a floating body, all on a SOI substrate. Target applications include high performance as a replacement for eSRAM and high density as a replacement for eDRAM. Dr. Fukuma indicated that most performance metrics for FBRAM compare favorably with these competitors. Over the past several years, there has been a steady evolution of FBRAM technologies with improvement, for example, in scalability, single cell operation and autonomous refresh. When scaled to support large signals, the change in threshold voltage should be maintained at about 0,4V, maximum device electric fields should be on the order of 0.7 MV/cm and the signal-to-noise ratio should be on the order of 18.

Dr. Hamamoto examined the structures and performance characteristics of scaled fully depleted Floating Body Ram Cells (FD-FBRAM) as they are scaled from 65nm to 32 nm. He concluded that the FD-FBRAM is scalable without performance sacrifices to the 32 nm node maintaining signal margin, retention time, and signal-to-ratio.

P4.2 D. Rinerson, Unity: "Storage Class Memory"

The memory element of the ReRAM class is composed of a select device and a memory cell that utilizes a variable conductivity metal oxide. The claimed attributes include 0.5 F^2 memory cell size, 4x the density of current generation NAND flash, and a fast write speed. The .5 F^2 memory footprint is achieved by fabricating four physical memory layers and storing 2 bits/cell. The principle of operation is based on oxygen ion exchange between the insulating and conductive metal oxide under the influence of an electric field. Unity has achieved successful data pattern reads and writes on the passive cross-point memory array. Like with the MRAM, it is envisioned that the fabrication for these memory elements would be done at the BEOL, leaving the CMOS FEOL unaffected. Unity plans an aggressive product integration schedule with 64 Gb memories as early as 2010.

P4.3 James J. Q. Lu, RPI: "3D Memory Integration Platforms - Novel Opportunities for Large-Scale, Heterogeneous, High-Bandwidth/Low-Power Memories"

Professor Lu began by comparing the differences between capacity and latency for different segments of the memory hierarchy area circa 2015, and he raised the possibility of 3D memory. He argued, based on analyses of Kerry Bernstein, formerly of IBM, that 3D extends the transfer of performance from device to the core level. There are several 3D integration approaches, e.g., 3D packaging, die on wafer assembly, transistor and BEOL wafer-to-wafer bonding for what he calls 3D Hyper-Integration. A 3D chip stack could include from bottom-up, processors, memory and A/D's, I/O's, sensors, MEMS, etc. Chips speed increases would be on the order $N^{3/2}$, where N is the total number of devices, and power reduction would go as $1/N^{1/2}$.

He described 3D platforms based on dielectric and metal bonding, and hybrid metal/dielectric bonding. As has been indicated, there are many potential benefits from 3D structures. There are, however, many technology challenges including thermal and power management, yield, test methodologies and the associated design and equipment infrastructure. There has been significant progress in 3D technologies alignment accuracies on the order of 1 micron, bonding at less than 400^0 C, thinning technologies, and in inter-wafer interconnect (i.e., metal bond via and through strata via (TSV)). The Known Good Die problem with both processor and memory has been ameliorated with redundancy, error correction coding (ECC), design-for-testability (DFT), built-in self-test (BIST), and simplified/optimized processing for each functional component on a separate wafer.

Basic research opportunities remain across the spectrum from materials and processing to CAD tools. Compact models are needed for prediction of density, speed and energy requirements of the 3D memory stack. Tools are needed to predict the impact of 3D integrated memories on microprocessor performance and on the optimum and new physical architecture for 3D memory systems.

P4.4 Luigi Capodieci, Global Foundries: "Extreme Patterning Solutions"

At the limits of geometric scaling, variability as a percentage of feature size invariably increases since atoms, k_BT/q and photons don't scale. Variability is due to intrinsic process variability, environmental effects and physical limits. Variability can either be systematic or random and can occur at length scales ranging from the transistor, across die, across wafer, wafer-to-wafer, or lot-to-lot. The cost of lithography processing tools is increasing at about 3X every five years; e.g., a scanner for the 45 nm node costs approximately $40 million.

Historically, due to wafer size increases and scanner throughput increases, the lithography cost per transistor declined rapidly through the 65 nm node. The cost per transistor has flattened at the 45 nm and 32 nm nodes and is projected to increase for the first time with the introduction of EUV. It does not appear that EUV will be available for the 32 nm and 22 nm nodes and the industry is currently relying on Resolution Enhancement Techniques (RET) to continue scaling.

RET techniques include: regular layout fabrics and templates, non-rectangular transistor modeling, DFM aware routing, optical proximity corrections, process variability-aware design techniques. Another approach is that of source optimization based on a library of layout features subject to different symmetry source illuminations. In addition, resist material complexity increases as lithographic feature sizes decrease.

Extreme Resolution Enhancement Techniques (ERET), can meet 22 nm needs, but with much greater cost increases. The simplest ERET approaches are various versions of double exposure/ double patterning. EUV lithography unfortunately won't be ready for 22 nm introduction. Nano imprint methods have problems with defectivity and overlay throughput while directed self assembly is still in the demonstration phase. Maskless lithography seems destined for low volume applications.

Session 5: Physical Limits of Memory Elements

K5.0 Ralph Cavin and Victor Zhirnov, SRC: "Nanoscale Memory Devices: Space-Time-Energy Tradeoffs"

There are three generic operations common to all memory arrays. A mechanism must be provided for selecting the memory cell to be addressed, there must exist a capability to sense the state of the cell, and of course the cell must be able to store the data that is written to it. The implementation of each of these basic functions establishes the scaling properties of the memory cell. In this paper a universal metric is proposed for use in evaluating a wide range of memory cells. Its origins lie with the Least Action Principle of Physics which states that physical systems evolve in such a way as to minimize the product of energy and time. Of course operational energies and times are important in memory systems but the size of the memory cell is also critical for high density memory systems. Thus, this presentation defined a Space-Action metric to be used in the comparison of various types of memory systems.

The operational properties of the various classes of memory elements considered in this presentation are derived from Boltzman and Heisenberg fundamentals and the use of energy barriers represented in terms of the various state variables upon which memory operation is based. In particular, the presentation estimated the Space Action metric for charge-based memory devices such as DRAM, SRAM, Floating Gate Memory and SONOS.

In all cases, the write-time for the memory cell was used as the 'time' term in the Space-Action metric. One of the dilemmas in electron-based memories is that high energy barriers (high voltages) are usually needed for long retention times but these do not allow fast injection of charge during the 'write' event. In the case of flash and DRAM memories, it was estimated that the optimum Space-Action metric is approximately 10^{-9}J-ns-nm^{3}.

Next a spin torque transfer MRAM was analyzed and the results gave a Space-Action metric of about 10^{-11}J-ns-nm^3, which is two orders of magnitude better than its electron-based counterparts. Finally, the atomic relay was analyzed as a prototypical ReRAM element and an impressive Space-Action metric of 10^{-17}J-ns-nm^3 was obtained. In the STT RAM and ReRAM analyses, the constraints due to the sensor were not included in the calculations and thus it is reasonable to argue that the Space-Action metrics would degrade somewhat. Nevertheless, the FeRAM may remain significantly superior in the Space-Action metric. Future work will focus on refining the estimates offered in the presentation and modifying the approach to reflect cell endurance.

P5.1 Rich Liu, Macronix: "Scaling Limitations of Flash Memory"

In this presentation, the fundamentals of Floating Gate NAND operation were reviewed. Both programming and erasure are achieved by Fowler-Nordheim tunneling. The Gate Coupling Rates, defined as the ratio of the capacitance between the control gate and the floating gate to the total floating gate capacitance, must be greater than 0.6 so that most of the gate voltage drop is across the tunnel oxide. In SONOS devices, electrons are trapped in SeN and de-trapping is very slow. Hole tunneling must be used to erase the cell contents, but this is very slow and the rate of erasure can only be increased by using thin tunnel oxides. Unfortunately this results in poor retention. One solution is to use bandgap engineered tunnel dielectrics to allow both fast hole erasing and good data retention. A P-poly gate is used instead of an N-poly gate to reduce gate injection.

The ultimate scaling limit for both floating gate devices and charge trapping devices is the small number of storage electrons available. (~15 electrons in a 10 nm floating gate device) The only known solution is 3D layering which does not solve physics and physical limitations but rather bypasses these limits using relatively large devices (~40 nm). Only charge trapping devices (not floating gate) are thought to be amenable for use in 3D memory structures.

P5.2 Nam Sung Kim, University of Wisconsin: "SRAM Scaling Limit: Its Circuit and Architecture Limitations"

Supply voltage and frequency scaling for logic will require that SRAMS to operate at lower voltages and this will exacerbate SRAM failure rates. Process variations that increase with feature down-scaling will also increase SRAM failure rate. As an example, it is projected that SRAM failure probabilities will be six or seven orders of magnitude greater as features are scaled from the 45 nm node to the 16 nm node corresponding to a reduction from a power supply voltage from 1.0 volts to 0.6 volts. Some of the approaches to addressing these challenges include the use of dynamic/adaptive techniques for a 6-T SRAM and the combination of SRAM cell sizing and error correction coding. It is believed that the dynamic/adaptive techniques can provide for an order of magnitude failure rate reduction for the 6-T SRAM with only small additional overhead. But the trend of using ancillary techniques to improve SRAM reliability probably has limited usefulness. Architectural relief derived from smaller cache sizes is possible if used in conjunction with lower supply voltages and lower operating frequencies since the gap between operating frequencies and memory bandwidth is reduced.

P5.3 Han Guchang; A*STAR: "Scaling Limits of MRAM"

MRAM structures include toggle, STT, racetrack and others. The focus of this presentation was on STT MRAM. Scaling challenges in performing the read operation for a conventional MRAM arise from the need for an ultra-thin barrier (~1 nm) to achieve a resistance and area product of 3 $\Omega\mu m^2$. In addition maintenance of uniform resistance over the wafer is needed. The STT MRAM offers a simpler structure and hence lower cost; a smaller area, and the half-selection issue of conventional MRAM's does not arise. However, the STT MRAM requires a high aspect ratio to provide high stability and this decreases memory density. If the perpendicular anisotropy STT MRAM is used, several advantages accrue including scalability, switching current that is independent of shape, high stability and density and small cell-to-cell interactions.

The limits of STT MRAM scaling may be determined by the maximum current that can be provided by CMOS estimated to be ~600 µA/um. The writing current must be smaller than 600 µA/µm for a 1T 1MTJ STT MRAM. For 10 year retention and 64 GB capacity, the required energy barrier is about 85 k_BT and it is estimated that a write current of 30 µA would be required. A challenge is to reduce the write current while maintaining the energy barrier.

P5.4 Shunri Oda, Tokyo Institute of Technology: "Performance Projections for Nanomechanical Memory"

A novel NEMS memory element is described that operates by physically bonding the floating gate between two stable states. Since the floating gate is suspended in an air gap above the channel, there is no charge tunneling through a gate oxide, the write/erase operations can be high speed, and the configuration is compatible with a CMOS process. By deflecting the floating gate, there is a measurable shift in the threshold voltage for the transistor that can be used to reflect memory state. The required fabrication processes to achieve the novel floating gate transistor were included in the presentation. Charge on the floating gate is stored in silicon nanodots embedded in the gate material.

The NEMS floating gate transistor was shown to maintain a current On/Off ratio of ~10^5-10^6 under scaling to gate lengths of about 100 nm. Also with scaling, the switching time decreases to about 15 ns at a gate length of 100 nm, and the switching voltage also decreases to about 2.5 V at 100 nm. The current best estimate of energy consumption for the device with gate length of 100 nm is about 0.2 f J. The memory window collapses at ~ 50 nm gate length foreclosing further scaling.

P5.5 Pinaki Mazumder, NSF/University of Michigan: "Physics of Double and Single Barrier Tunneling and Their Impacts on Semiconductor Memories"

This presentation focused on the physical modeling and analyses of tunnel barrier systems in the context of memory applications. Professor Mazumder has contributed to the analysis of a wide variety of devices including the TSRAM (Tunneling SRAM), the QMOS Sense Amplifier, CMOS radiation hardened and noise resistant domino logic circuits, nanoscale ionic transport, analog memristors, brain-like computing and many others. He has also authored three books on memory test, fault tolerance and reliability.

Session 6 – Needs and Models for Collaborative Research

K6.1 Steven Hillenius, SRC: The Unprecedented Power of Collaborative Research

In this keynote presentation, a review was presented of the evolution of the research landscape since 1950 from research conducted principally within individual companies to the more collaborative model that is widely used today. It was noted that the Semiconductor Research Corporation (SRC) was formed in 1982 in the midst of a severe recession in the United States; perhaps even more severe than the present recession. At its outset, SRC defined a set of ambitious research goals that defined its research agenda and laid the foundation for what eventually became the International Technology Roadmap for Semiconductors (ITRS). SRC, financially and technically supported by its members, operates a virtual laboratory with research in over one hundred universities worldwide. This structure provides considerable flexibility to begin and end new research programs as industry needs change. Over its twenty-seven year history, SRC has gained experience on how to operate a highly successful consortium at the interface between industry, government, and universities.

SRC studies have shown that in the semiconductor industry, the average latency between the first publication of a research paper and the introduction of the research results into products is on the order of twelve years. SRC influential paper (100 citations are more) citations show a citation trajectory over time that is markedly different for technology-related papers than for design-related papers. After a technology is introduced into manufacturing, citations for related technology papers usually begin to decline whereas the citation rate for influential design papers is relatively flat over time, even after introduction of the ideas into practice. Dr. Hillenius indicated that there is a strong need for such a consortium to focus pre-competitive research on memory technologies; particularly at a time when foreseeable limits exist for several classical memory technologies and with the increased emphasis on embedded memories. In his view, the SRC model for collaboration would work well to meet this research need.

P6.2 Gurtej Sandhu, Micron: Research priorities/Needs

From the perspective of a memory company, there are too many options for future memories and too little time and money to identify the optimum choices. The socket-level replacement of existing memories with emerging memories seems to be unlikely and the need is for application specific memory solutions. Research focus should be directed to those nodes at the end of ITRS. There are barriers to collaboration in emerging memory research relating to standards, intellectual property, and the current competitive landscape. These barriers are counter-balanced by needs for resource/expertise sharing, standard performance characterization methodologies and consolidate success metrics for the research. Collaborative research must embrace fast-paced learning, materials screening to develop device-level data at the scales of interest, and it must offer a comprehensive approach to address all aspects of memory performance. The memory eco-system is quite broad including manufacturing, assembly, packaging and test, circuit design and architecture, process module integration, and the equipment and material suppliers. Collaborative research must comprehend the needs of each of these eco-system members.

P6.3 Ajith Amerasekera, TI: Role of Memory in More than Moore

The driver for exponential growth is system performance and industry will take the path of least resistance to increase performance. As the application space broadens for integrated circuits, process technologies are entering into unchartered territory that is no longer driven by simple scaling paradigms. One can classify the various phases of integrated circuit applications over time as first *centralized*, then *interactive* and eventually *immersion*. Today, we are at the cusp of the transition from *interactive* to *immersion* integrated circuit technologies. According to Professor Jan Rabaey (University of California at Berkeley), the application space in the future can be viewed as having three interrelated levels; the infrastructural core will continue to be driven by Moore's Law gains while the next two levels, mobile applications and the sensory swarm are much more likely to fall within the purview of More than Moore. There are a wide variety of future application drivers including smart buildings, personal health technology, energy-efficient transportation, energy generation and management, and many more. In this context, Dr. Ameraskera believes that the demand for memory will only increase over time. The drive to an 'intelligent ambient' will increase the demand for non-volatile memory since there is a super-linear increase in the amount of data. Speed of memory operation will become increasingly important by the 2020's. The advent of autonomous systems with novel battery technology and energy harvesting will require circuits that can operate with unreliable power sources. The role to be played by 3D systems is yet to be defined. One should not forget about hard disk drive technology and indeed, how much additional scaling can this technology offer?

P6.4 Zoran Krivokapic, Global Foundries: Foundry Perspective on Embedded Memory

Multi-chip modules, 3D integration, and embedded are the three forms of delivery of memory to a system application. In each case, performance, cost, and form factor must be considered. In the case of the SRAM, at what node will random dopant fluctuations require fully depleted devices? How perfect a dielectric can be fabricated to minimize random telegraph noise in a small area? SRAMs can be based on tunneling devices and this could potentially offer area savings. However, the valley current of the Negative Differential Resistance (NDR) is critical to achieve low power operation for these devices. With respect to DRAMs, cost is a major limitation for high performance solutions. Research for a novel dynamic memory replacement is important. In the embedded class of memories, flash is used to achieve smaller memory sizes but the large operating voltages must be reduced. Possible improvements may be derived from nanocrystal and SONOS forms of flash memory. Spin torque transfer MRAMs is an excellent candidate for embedded memory since its power scales inversely with geometrical scaling. Penta-layer STT and perpendicular magnetization devices could offer significant improvement and offer further research guidance. Also, by using current flowing perpendicular to the ferromagnet domain wall, its velocity can be increased at smaller current densities (faster switching). At the system level, memory-based architectures for computing will require large amounts of memory.

P6.5 Dim Lee Kwong, IME: Government –University-Industry Collaborative Models: Singapore Perspective

This presentation depicted the Singapore national organizations related to technology and then focused on the Ministry of Trade and Industry to which A*STAR, an R&D funding body, reports. In Singapore, there are currently 40 IC design companies, 14 wafer fabs, and 20 assembly and test companies. The total worldwide market share of Singapore companies is about 12%. The 2008 projected industry financial contribution of the semiconductor sector is S$37B. The Institute of Microelectronics (IME) research thrusts are: i) Moving up the value chain by providing integrated micro-system products and services, ii) enhancing More Moore capabilities while driving innovations in More than Moore through nanotechnology, integrated optoelectronics and green electronics, iii) setting up a medical electronics R&D hub, and iv) enabling new growth areas e.g., e-health. To support these thrusts, IME has established capabilities in design/modeling, fabrication, characterization, packaging, circuits and interfaces. All IME programs are designed to involve strategic partners from industry, medical centers, hospitals, and universities.

At the present time, more than 80% of IME programs are of the More-than-Moore variety (innovative applications of VLSI technology) and IME is making continued investments in its infrastructure which includes 200 mm CMOS and MEMS processing, 300 mm through-silicon-via (TSV) and elective CMOS, silicon photonics multi-project wafer facility, and several consortia addressing specific topics, e.g. an analog centre of excellence in Singapore. The international recognition for the quality and relevance of IME programs is continuing to grow as for IME staff members are recognized for their papers at top-ranked professional meetings.