Non-Volatile Memory Technologies: The Quest for Ever Lower Cost

Stefan Lai

BeingAMC, Inc.
Saratoga, CA 95070, USA

Abstract

Growth of flash memory business over last 20 years was driven by never ending reduction of memory cost through Moore's Law and innovations, and the quest for ever lower cost will continue for many years to come. This review begins with a brief summary of trend of flash memory cost reduction up to now. Then some of the improvement efforts on existing technologies reported by industry will be discussed. NAND flash will continue to be the cost reduction driver in next few years but will face increasing level of difficulties. Innovations will enable the trend to continue. For longer term, industry is developing new memory technologies that have promise to deliver ever lower cost. Some more mature new technology concepts will be discussed in this review. General direction is to go to multi-layer memories with multi-level cell capabilities. There are also alternative approaches like probe based storage. Enhancement from system level solution will help existing technologies as well as facilitating introduction of new technologies. It is expected that products from few new technologies will take off in coming years to enable continuation of cost reduction of non-volatile memories, meeting insatiable demand of existing devices for more memory capacity at lower cost, as well as creating new devices and new markets.

Introduction

Flash memory business grew from a small beginning in 1987 to over US$23.8B in revenue in 2007 (Figure 1). Products using flash memories like cell phones, music players, memory cards and USB drives are ubiquitous in everyday life of billions of people around the world. In 2007, more NAND memory bits (1.9 Exabyte or 1.9 x 10^{18} bytes) were shipped compared to DRAM bits over its entire history. This amazing growth is driven by Moore's Law cost reduction. Flash memory price dropped from about $80,000/Gigabyte in 1987 (for a 256Kb unit) to less than $1.50/Gigabyte late 2008 (for a 16Gb MLC unit). This price drop is more than 40% per year on average, ahead of Moore's Law learning curve of 30% per year. This relentless cost reduction was made possible by innovations along multiple fronts with lithography improvement being the essential driving force, but also included use of innovative self aligned technologies (1), introduction of NAND memory to reduce memory cell size, introduction of multi-level cell technology and wafer size increasing from 150 mm in 1987 to 300 mm in recent years. Reduction of memory price enabled creation of new markets, driving demand for more memory bits. It also stimulated continuous innovation of existing technologies as well as development of alternative memory technologies in anticipation of scaling challenges of existing flash memory technologies of NOR and NAND.

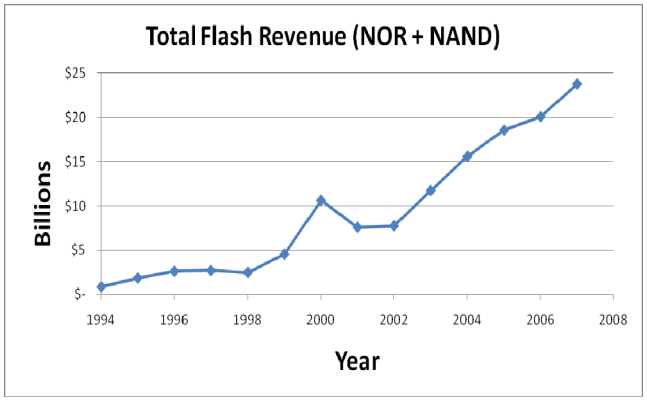

Figure 1: NOR + NAND Flash Revenue from 1994 to 2007 (Data from WSTS)

Continuation of Existing Technologies

Looking forward, existing flash memory technologies of NOR and NAND will continue to reduce in cost through scaling innovations in next couple generations. Key challenges to memory scaling originate from lithography and device characteristic (for a review of these topics, see reference 1 & 2). Looking at NAND memory as an example, with its regular cell layout consisting of straight lines and spaces for diffusion and polysilicon, it is possible to make use of optical enhancement techniques, enabling extension of conventional optical lithographic tools. One example is development of double patterning techniques (2, 3) which extend capability of existing lithographic tools to sub 30 nm nodes, at the expense of additional process steps and higher cost. Challenges to device characteristic are many, including coupling ratio, Vpass window, cell to cell interference and short channel effect of cell transistor (2). To address these problems, the most reported solution is to replace floating gate with a charge trapping layer in NAND flash memory. One such technology reported is TANOS (2, 3) (Figure 2) which can minimize cell to cell coupling while maintaining adequate coupling ratio. Floating trap approach also reduces effect of local oxide defects. To address short channel effect and Vpass window, new transistor structures are being

developed. One example is a three dimensional cell transistor called Hemi-Cylindrical FET (HCFET) (2, 3) (Figure 2) where transistor channel length is extended physically above normal planar silicon surface through a half cylinder. Charge trapping TANOS layer wraps around silicon half cylinder to maintain close to uniform channel control. Through these innovations, memory transistor can be scaled down physically to sub 30 nm nodes (3). A bigger challenge for NAND comes from number of electrons stored in either

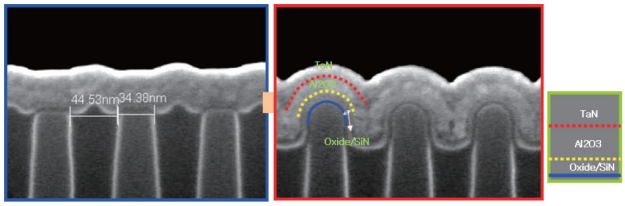

Figure 2: Cross section SEM of Planar and HCFET with TANOS (3)

floating gate or charge trapping layer. At sub 30 nm nodes, there are only few hundred electrons stored in charge storage layer and for Multi-Level-Cell (MLC) (4), only tens of electrons separate storage levels. With dielectric degradation mechanisms through cycling unchanged with scaling, the result is it takes less number of cycles to give significant charge leakage. While lower number of cycles may be adequate for consumer applications like music players or memory cards for cameras, it is stressing limits in enterprise applications like Solid State Disk (SSD) for servers.

Three Dimensional Memories

With traditional memory cell size scaling running into scaling limits, alternative approaches are being investigated. Most obvious approach is to add more memory layers in vertical dimension: 3D memories. One approach reported by Samsung (5) is single-crystal silicon layer on ILD stacking (Figure 3). NAND arrays are formed on single-crystal silicon on ILD as well as on bulk to double memory density without

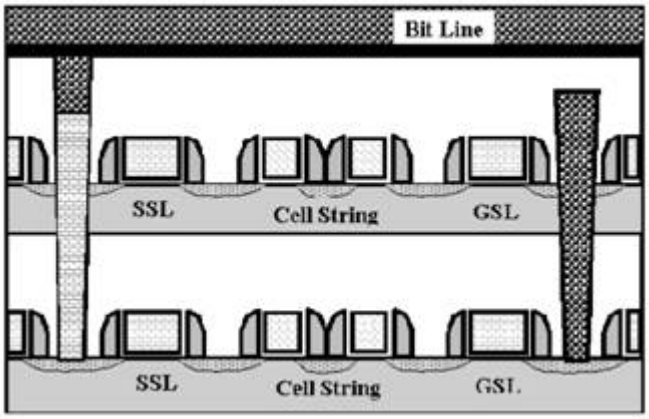

Figure 3: 3D stacked NAND Cells. Additional memory is fabricated on a single crystal silicon layer on ILD (5)

increasing chip size. Critical masking count is reduced by sharing bit line contacts as well as common source lines while base silicon process is shared between memory layers. This is different from combining two fully processed wafers which does not lower cost of wafer processing. One of the challenges is to fabricate high quality single crystal silicon layers on ILD. A similar approach was reported by Macronix (6) where thin film transistors (TFT) are used. Thin film transistors have polysilicon channels formed by annealing of LPCVD deposited amorphous silicon. Polysilicon is easier to prepare compared to single-crystal silicon on ILD but the transistors will have lower performance.

Another 3D approach is reported by Toshiba (7, 8) in their Bit-Cost Scalable (BiCS) flash memory (Figure 4). A NAND string is fabricated vertically with multi-layer electroplates as control gates of SONOS type memory on a vertical column of

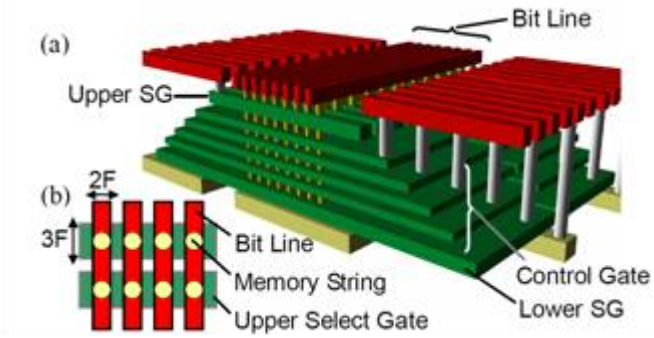

Figure 4: Birds-eye as well as top-down view of BiCS Flash memory array (7, 8)

polysilicon depletion mode transistors. Top and bottom electroplates act as select gates of NAND string. In this structure, number of critical lithographic steps remains constant because whole stack of control gates is punched in a single step. Compared to multi-layer NAND approach which requires critical lithography on each layer, BiCS is lower in cost (7).

Cross Point Memories

NAND 3D memories are complex structures with memory transistors subjected to traditional physical and electrical scaling limits. To be a cost effective replacement to existing technologies, new technologies must be simple in structure and be scalable down to sub 10 nm nodes. For ultimate lowest cost and scalable NV memories, industry has generally converged in general approach of a cross point memory (Figure 5). It is the simplest possible device defined by lithography. The structure is a two terminal device that is located at cross point of a conductor line in X-direction at the bottom and a conductor line in Y-direction at the top. X and Y conductor lines are at minimum pitch giving a cell area of 4 λ^2, where λ is minimum half pitch defined by limit of lithography capability. The two terminal device has two

elements: a memory storage element for memory function and a switch element to isolate memory element so that it will not be disturbed during normal device operation of other

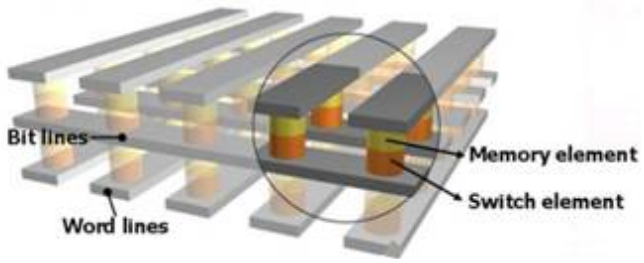

Figure 5: Generalized cross-point structure with one bit of array consists of a memory element and a switch element (9)

devices in memory array. The two terminal device may be defined simply by cross point lithography or it may require an extra masking step. Key point of this cell structure is it scales directly with lithography without complex layout structure or transistor size limit in DRAM, NOR and NAND memories. Ultimately, smallest memory device is determined by memory and switch material limitations at dimensions down to cluster of atoms.

With a simple cross point memory, it is possible for memory layers to be fabricated on top of each other to achieve a three dimensional memory (Figure 5). For such memory layer stacking, it is important that process temperature required for successful fabrication of additional memory layers stays within process limit. For example, if copper lines are to be used for conductors, then memory and switch elements should be fabricated at temperatures lower than about 400° C. And if higher temperature processing is required, other conducting material like tungsten will be necessary. Another consideration is bit cost reduction is not directly proportional to number of memory layers. Each memory layer still requires minimum dimension definition and thus more expensive lithographic steps. Base silicon cost is one element shared. Also, yield tends to go down with more memory layers. Typically, a 4 layer memory may give 50% to 70% per bit cost reduction.

Memory Technology Candidates

Memory mechanism reported for 3D two terminal device typically involves change in device resistance in response to application of electric field or passage of electric current. Examples of resistance change memory element are phase change memory (PCM), resistive memory (RRAM) and programmable metallization cell (PMC). Examples of switch element are polysilicon P/N diodes, nanowire P/N diodes and metal oxide P/N diodes.

Of three memory technology candidates, PCM (10) (Figure 6) is most mature as high density array of up to 512 Mb has been reported (11). Chalcogenide memory material is switched to either amorphous state (high resistance) or crystalline state (low resistance) by Joule heating and quenching. Temperature is controlled by magnitude and

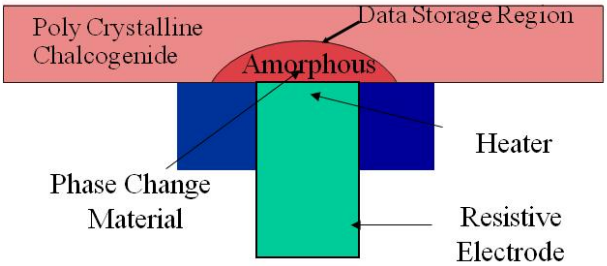

Figure 6: PCM memory element. With passage of switch current from electrode to Chalcogenide, region around the contact heats up and can be quenched into high resistance amorphouse phase (10)

timing of current through the device. More than 10^{10} write cycle capability was demonstrated on single cells (10). High cycle count is of interest for some RAM like applications. Switch current is relatively high (~ 600 μA) for current technology node (~ 90 nm) which limits write bandwidth to lower than exiting NAND products. Switch current is expected to be reduced with future finer litho nodes as well as improvement in cell design and material. Write performance can further be improved with new alloy that has faster switch time.

The alternative RRAM technology is gaining more interest (12, 13). RRAM consists of either simple or complex oxides that can be switched between different resistance states by applying suitable voltages across the structure. Examples of simple oxides are Cu_xO, NiO, TiO_x, ZrO_x and HfO_x (14). Examples of complex oxides are $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO), (Nb, Cr)-Doped (Ba, Sr) TiO3 or $SrZrO_3$ (14). There is no consensus on exact switching mechanism which is an area of intense research. In many cases, an initial forming process is required for device to function properly as a memory. One

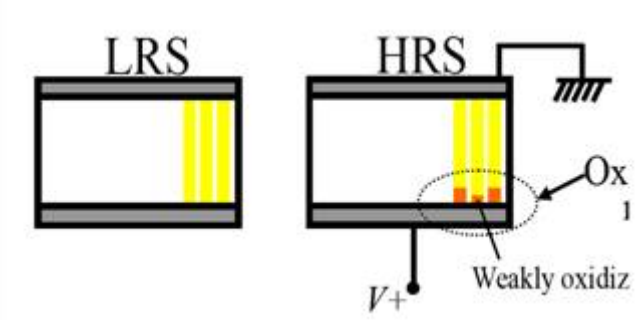

Figure 7: Model of RRAM switching showing conditions of the filaments in Low Resistance State and High Resistance State (12)

model (12) postulates that forming process creates filament conducting paths (Figure 7) giving low resistance states (LRS). Under bias, oxygen vacancy migration at interface

changes barrier at interface, giving high resistance states (HRS). The process is reversible under different bias conditions, which can be uni-polar or bi-polar. Typically, reported cycling capability is limited to tens of thousands of cycles. To date, publication on high density Mb level products has not been reported.

Programmable metallization cell (PMC) (Figure 8) (15) relies on transport of mobile metal ions in a solid electrolyte under electric field to form a conduction bridge between two electrodes. One electrode is reactive to be source of metal and the other electrode is inert. Bridge is broken with reversed voltage. Switching process is fast and requires very low applied voltage. One example of PMC system is Ag doped Ge_xS_{1-x} and a 2Mb test array has been demonstrated (15). Forward switching voltage is about 300 mV whereas reverse switching voltage is about 80 mV. On/Off resistance ratios

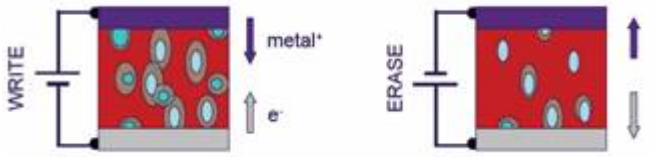

Figure 8: Model of switching of the PMC cell, showing the metal bridging on the on state and no metal contact during the off sate (15)

are as high as 10^6. Switching time is in nSec range and over 10^6 write/erase cycles have been demonstrated. For multi-layer memory, low switching voltage is a concern as voltage coupling from adjacent lines may disturb individual cells.

As mentioned before, a memory element by itself is not sufficient to give a working memory array. A switch is required to isolate memory element in an array. Development work on switch element is not as widely reported as memory element. Ideal switch has requirement of being able to pass very high current in conducting state so that memory switching is not compromised and read current is not reduced. It also has requirement of very low leakage current in reverse path such that many memory cells can be connected in parallel in a high density array. A simple switch is silicon PN diode. However, to get high quality silicon PN

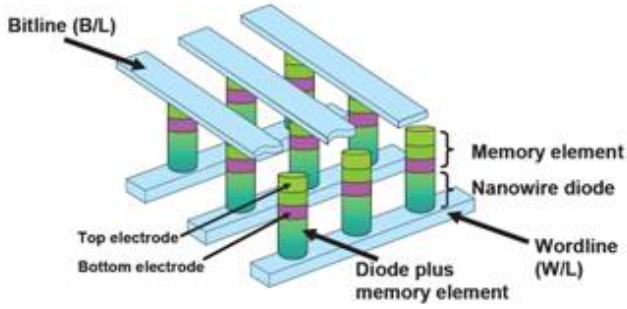

Figure 9: Schematic of PCM memory arrays with nanowire diodes as memory cell selection devices (16)

diode, relatively high process temperature in 700° C range is required, which is in conflict with typical back end aluminum or copper processes. In most cases, development of switch focuses on innovative material that can be fabricated at relatively low temperature. One example of switch reported for PCM is nanowire diode (Figure 9) (16). In this case, phosphorous doped germanium nanowires (GeNW) are grown by vapor-liquid-solid (VLS) technique with in-situ doping. Functional PCM cell with integrated nanowire diode was demonstrated with well behaved switching behavior.

Use of oxide diode as switch element is reported for RRAM. In this demonstration (9), p-CuO_x/n-$InZnO_x$ heterojunction thin film was fabricated with Ti-doped NiO as memory element. A 2 stack memory was fabricated with all processes at room temperature allowing for compatibility with current

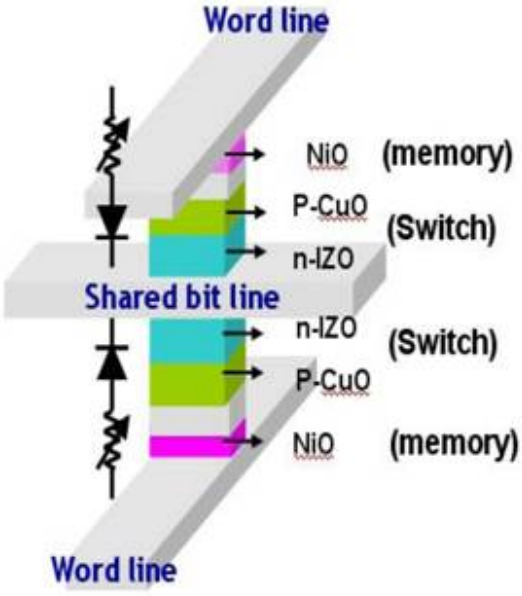

Figure 10: Schematic diagram of a 2-stack 1D-1R memory cell with upper layers reversed to share the bit line (9)

CMOS technologies. The diode can carry over $10^4 A/cm^2$ in forward direction and operating voltage for diode and storage element is less than 3V.

New and Different Challenges

In cross point memory structures, one is trading manufacturing of complex transistor structures, which have a long history of development and well known processes of optimization, to manufacturing of simple memory structures, which rely on very exact material properties to control switch function as well as memory function. For memory mechanisms, there are many candidates and only three of most reported ones are discussed here. Within the three different types of memories, there are almost infinite numbers of combination of material choices. This is something new to

silicon memory industry. Requirement for fast development and evaluation of multitude of material have stimulated new concepts of material characterization and optimization (17). New processes may be required to deposit, etch and clean of new materials in very small dimensions and typically difficult topography. One example is requirement to deposit switch and memory materials in small openings less than 20 nm in diameter a few years from now. Electrical testing is another new challenge. Two terminal device switches in single nSec to hundreds of nSec. With very small dimensions, probe capacitance can be much larger than device capacitance. Special hardware is required to accurately test dynamic of switching. In an array, operation of memory relies on schemes like 1/2 voltage or 1/3 voltage inhibit for the array to function properly with no disturb. On/off ratio of 2 terminal element determines size of memory "tile" (minimum array block) and amount of array overhead required for support circuits. Decoders and sensing circuits have to support higher currents compared to conventional memories. The challenges are multiplied when memory layers are stacked on top of each other. The challenges presented opportunities for innovations and if one is successful to overcome the challenges, one is rewarded with a simple memory array that can be scaled continuously for many generations.

System Enhancement

One of the paradigm shift that enabled NAND technology to ship in high volume is use of external controller to overcome some intrinsic limitations of the technology. Controller enhances data integrity through error detection and error correction, and manages cycling by distributing data through wear leveling. But memory control and sensing on chip level are still the same approach of threshold detection that has been in use since the 60's. These traditional ways to read and write memory devices are now being challenged by a different system approach to manage the information.

The new system approach is a fundamentally different way to approach the problem of reconciling information on cell level. Instead of viewing signal detection as a discrete event, it is important to view all data as having a signal to noise ratio. All storage media have an inherit capacity limit based on bit error rate. There is no such thing as *perfect data*. This new approach is a multi-faceted, multi-staged method of reconciling data through established techniques from modern "communications theory". This new system approach will be necessary for the success of future high density, multi-level memory solutions. Also, such approach may change the process of silicon technology learning from physical silicon to system level, which may allow for faster as well as less expensive technology development.

One example of this approach is the work reported by Storage Genetics Inc. (SGI) of Longmont, Colorado (18). In an experiment on a NAND device on a leading technology node, the company has recently demonstrated at least a 2x capacity gain and an order of magnitude improvement in endurance without adding significant system costs. In Figure 11, top histogram showed raw data of 16 levels while lower histogram showed how sigma of one level is improved by the SGI processor and compensations engines, with sigma value

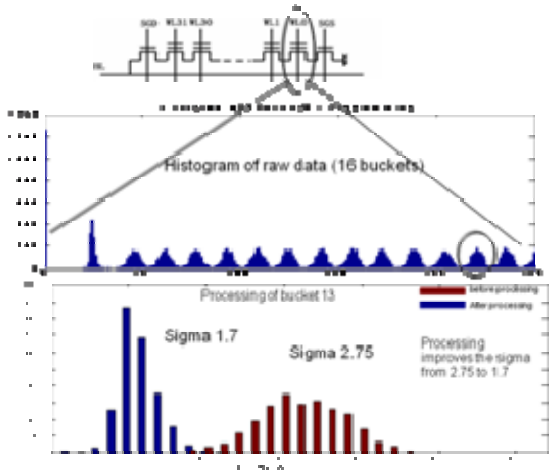

Figure 11: Top graph showed raw data of 16 levels for 4 bit/cell NAND device. The bottom showed the histogram of how one level is improved by the SGI processor and compensation engines (18)

going down from 2.75 to 1.7. This approach will prove to have a major impact on NVM cost and reliability in the future multi-bit per cell and multi-layer memories.

Alternative Approach

Cross point memory is limited by lithography, which is becoming very expensive with latest generations of immersion tools, adding to cost of manufacturing. One of the more innovative alternative technology reported is probe based memory, with pioneer work known as Millipede developed by IBM (19). In this case, instead of addressing a memory bit by X and Y conducting lines defined by lithography, scanning microprobes are used to make contact to memory media. Movement of microprobes, as well as other elements, is enabled by Micro Electro Mechanical Systems (MEMS) technology to give high degree of precision. Horizontal as well as vertical movement in nanometer or finer steps has been reported. Memory density is not determined by lithography but instead by ability to position probes in nanometer dimension as well as size of probe tip. High read and write bandwidth is made possible by use of multiple probes in parallel. With MEMS technology, manufacturing steps are based on same high volume Cleanroom processing used in standard high volume silicon processing. No advanced lithography is required and older generation fabs can be used to produce such memories at very low cost. Memory density improvement comes from improvement of control of nanometer scale movement as well as smaller tip geometries. Figure 12 showed latest implementation by Nanochip Inc. where three chips bonded

together at wafer level give a working high density memory (20). They are planning to introduce high density products in tens of GByte range at lower cost compared to NAND in next

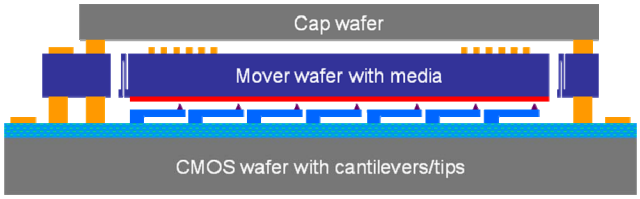

Figure 12: Cross section schematic of Nanochip Memory Module (20)

couple of years.

Conclusions

Desire for more memory bits has been growing constantly with digitization of every day sound, pictures, videos and other information. It was made possible by ever lower cost of NV memories enabled by Moore's Law and innovations. With transistor based memory running into physical as well as reliability scaling limits, a new class of simple, two terminal, cross point, three dimensional memories will emerge to take NV memories through the last leg of NV memory scaling. The transition is not easy and extremely challenging, but industry has overcome big challenges before. I am confident that they are ready for this challenge now.

Acknowledgement

The author wish to acknowledge contribution of many people in industry who have made flash memories such exciting technologies and businesses over the years, and pioneers who continue to innovate to create an even more exciting future for NV memories. Personally, he would like to thank his colleagues at Intel over many years who together as a team, they have changed the world.

References

(1) S. Lai, "Flash Memories: Successes and challenges", *IBM Journal of Research and Development*, vol. 52, No. 4/5, pp. 529-535, July/September 2008.
(2) K. Kim, "Future memory technology: challenges and opportunities", *International Symposium on VLSI-TSA Proc of Tech Program*, pp. 5-9, 2008.
(3) D. Kwak, J. Park, K. Kim, Y. Yim, S. Ahn, Y. Park, J. Kim, W. Jeong, J. Kim, M. Park, B. Yoo, S. Song, H. Kim, J. Sim, S. Kwon, B. Hwang, H. Park, S. Kim, Y. Lee, H. Shin, N. Yim, K. Lee, M. Kim, Y. Lee, J. Park, S. Park, J. Jung and K. Kim, "Integration Technology of 30nm Generation Multi-Level NAND Flash for 64Gb NAND Flash Memory", *Symposium on VLSI Technology Digest of Tech Papers*, pp. 12-13, 2007.
(4) K. Prall, "Scaling Non-Volatile Memory below 30 nm", *Proceedings Non-Volatile Semiconductor Memory Workshop*, pp. 5-10, 2007.
(5) S. Jung, J. Jang, W. Cho, H. Cho, J. Jeong, Y. Chang, J. Kim, Y. Rah, Y. Son, J. Park, M. Song, K. Kim, J. Lim and K. Kim, "Three Dimensionally Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30 nm Node", *IEDM Technical Digest*, pp. 37-40, 2006.
(6) E. Lai, H. Lue, Y. Hsiao, J. Hsieh, C. Lu, S. Wang, L. Yang, T. Yang, K. Chen, J. Gong, K. Hsieh, R. Liu and C. Lu, "A Multi-Layer Stackable Thin-Film Transistor (TFT) NAND-Type Flash Memory", *IEDM Technical Digest*, pp. 41-44, 2006.
(7) H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi and A. Nitayama, "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory ", *Symposium on VLSI Technology Digest of Tech Papers*, pp. 14-15, 2007.
(8) Y. Fukuzumi, R. Katsumata, M. Kito, M. Kido, M. Sato, H. Tanaka, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi and A. Nitayama, "Optimal Integration and Characteristics of Vertical Array Devices for Ultra-High Density, Bit-Cost Scalable Flash Memory", *IEDM Technical Digest*, pp. 449-452, 2007.
(9) M. Lee, Y. Park, B. Kang, S. Ahn, C. Lee, K. Kim, W. Xianyu, G. Stefanovich, J. Lee, S. Chung, Y. Kim, C. Lee, J. Park, I. Baek and I Yoo, "2-stack 1D-1R Cross-point Structure with Oxide Diodes as Switch Elements for High Density Resistance RAM Applications", *IEDM Technical Digest*, pp. 771-774, 2007.
(10) S. Lai, "Current status of the phase change memory and its future," *IEDM Technical Digest*, pp. 255 - 258, 2003.
(11) J. Oh, J. Park, Y. Lim, H. Lim, Y. Oh, J. Kim, J. Shin, J. Park, Y. Song, K. Ryoo, D. Lim, S. Park, J. Kim, J. Kim, J. Yu, F. Yeung, C. Jeong, J. Kong, D. Kang, G. Koh, G. Jeong, H. Jeong and K. Kim, "Full Integration of Highly Manufacturable 512 Mb PRAM based on 90 nm Technology", *IEDM Technical Digest,* pp. 49-52, 2006.
(12) K. Kinoshita, T. Tamura, H. Aso, H. Noshiro, C. Yoshida, M. Aoki, Y. Sugiyama and H. Tanaka, "New Model Proposed for Switching Mechanisms of ReRAM", *Proceedings Non-Volatile Semiconductor Memory Workshop* pp. 84-85, 2006.
(13) S. Karg, G. Meijer, J. Bednorz, C. Rettner, A. Schrott, E. Joseph, C. Lam, M. Janousch, U. Staub, F. La Mattina, S. Alvarado, D. Widmer, R. Stutz, U. Drechsler and D. Caimi, "Transition-metal-oxide-based resistance-change memories", *IBM Journal of Research and Development*, Vol. 52 No. 4/5, pp. 481-492, July/September 2008.
(14) G. Burr, B. Kurdi, J. Scott, C. Lam, K. Gopalakrishnan and R. Shenoy, "Overview of candidate device technologies for storage-class memory", *IBM Journal of Research and Development,* Vol. 52 No. 4/5, pp. 449-464, July/September 2008.
(15) R. Symanczky, R. Dittrich, J. Keller, M. Kund, G. Muller and B. Ruf, "Conductive Bridging Memory Development from Single Cells to 2Mbit Memory Arrays", *Proceedings Non-Volatile Memory Technology Symposium*, pp. 70-74, 2007.
(16) Y. Zhang, S. Kim, J. McVittie, H. Jagannatghan, J. Ratchford, C. Chidsey, Y. Nishi and H. Wong, "An Integrated Phase Change Memory Cell with Ge Nanowire Diode for Cross-Point Memory", *Symposium on VLSI Technology Digest of Tech Papers*, pp. 98-99, 2007.
(17) Kathy Klotz-Guest, "Intermolecular and Elpida Launch R&D Collaboration For Next-Generation Memory Technology", *Press Release Intermolecular Inc.*, July 15, 2008.
(18) Ken Eldridge, Storage Genetics Inc., *Private communication*, September 2008.
(19) A. Pantazi, A. Sebastian, T. Antonakopoulos, P. Bachtold, A, Bonaccio, J. Bonan, G. Cherubini, M. Despont, R. DiPietro, U. Drechsler, U. Durig, B. Gotsmann, W. Haberle, C. Hagleitner, J. Hedrick, D. Jubin, A. Knoll, M. Lantz, J. Pentarakis, H. Pozidis, R. Pratt, H. Rothuizen, R. Stutz, M. Varsamou, D. Wiesmann and E. Eleftheriou, " Probe-based ultrahigh-density storage technology", IBM *Journal of Research and Development*, Vol. 52 No. 4/5, pp. 493-511, July/September 2008,
(20) G. Knight, "1 TeraByte per Chip?", *Presentation at INSIC Annual Meeting Symposium*, San Diego, August 2008.