# Array Architecture for a Nonvolatile 3-Dimensional Cross-Point Resistance-Change Memory

Elaine Ou and S. Simon Wong, *Fellow, IEEE*

***Abstract*—This work explores the design and capabilities of a three-dimensional cross-point array structure suitable for use with resistance-change non-volatile memory. The resistance-change cell serves as both the access element and the memory element, eliminating the need for individual selection devices. This work presents novel architecture and circuit techniques that minimize leakage current effects while maintaining a high effective bit density. A test chip fabricated in 0.18 $\mu$m CMOS technology verifies the architecture and circuit functionality. The performance of an 8 Gb memory chip built in 65 nm technology has been simulated. A random access time of 104 ns is achieved with a power dissipation of 61.2 mW. This makes the 3D cross-point memory competitive with NOR flash in terms of read time, and competitive with NAND flash in terms of area efficiency.**

***Index Terms*—3D, cross-point, flash, nonvolatile memory.**

## I. Introduction

IT IS anticipated that NAND flash may not scale below the $2\times$-nm process technology generation due to degradation in performance and reliability characteristics [1]. There is a wide range of emerging technologies under development to replace flash, and the most notable candidates include magnetoresistive random access memory (MRAM) [2], phase-change memory [3], [4] , and resistance-change memory [5]. Resistance-change metal-oxide materials have been shown to possess favorable characteristics that make them particularly suitable for a 3D memory architecture. They are compatible with modern CMOS processes and have demonstrated high-speed and low-power switching abilities [6], with sufficient retention times and endurance for many applications that would ordinarily use flash memory.

A typical resistance-change memory array requires a selection device such as a transistor or a diode for each memory element [7], [8] [9]. The MOSFET or bipolar junction transistor occupies a significant area and the cell size is normally 10-15 $F^2$, where $F$ is the minimum feature size available for the process technology. Memory cells using diode selection devices have been demonstrated with a cell size of about $4F^2$ [7]. The integration of selection polysilicon diodes in a 3D memory array has been commercially demonstrated [10], [11] . However, the high temperature processing requirements of polysilicon diode fabrication may not be compatible with current resistance-change metal-oxides. Furthermore, some metal-oxides require bipolar programming for proper operation [12]. Implementing this material in a memory array involves the use of two anti-parallel diodes per cell and results in a substantial increase in array area. For true scalability beyond 20 nm technology nodes, it is desirable to design a cross-point memory array that does not require selection devices for individual cells. Without the insertion of active cell-selection devices, the memory architecture should accommodate 3-dimensional stacking of memory layers. This will reduce the effective cell size to $4F^2/n$, where $n$ is the number of memory layers.

Manuscript received October 30, 2010; revised April 19, 2011; accepted April 20, 2011. Date of publication May 31, 2011; date of current version August 24, 2011. This paper was approved by Associate Editor Peter Gillingham.

E. Ou is with the Department of Electrical and Information Engineering, University of Sydney, New South Wales, Australia.

S. S. Wong is with the Center for Integrated Systems, Stanford University, Stanford, CA 94305-4070 USA (e-mail: wong@ee.stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2011.2148430

## II. Cross-Point Memory Architecture

A cross-point memory architecture without cell-selection devices will allow wordlines and bitlines to be laid out in the minimum metal pitch allowed by the technology process. However, the peripheral circuitry must be able to access the bitlines and wordlines in such a dense pitch.

The leakage current that arises as a result of unwanted memory cells being biased during a read or write operation is a major concern. For the read operation, this adds noise to the signal being sensed. For the write operation, this can create disturb conditions on unselected cells, or result in the incomplete programming of selected cells. The worst-case leakage condition occurs when the majority of the memory array is in the low-resistance state and a selected cell is in the high-resistance state. The unselected memory cells must also serve as rectifying devices, and are less effective at this when at a low resistance. The peripheral access circuitry must be designed to overcome this leakage current problem.

This paper addresses the design challenges of implementing the read operation of a cross-point memory array without individual cell-selection devices. This section describes the read strategies employed in a 2-dimensional cross-point memory array. The design procedure for the memory architecture is detailed and the read operation is explored at a circuit level. Section III describes a test chip that emulates the cross-point memory array and demonstrates the functionality of the read operation. Sections IV and V present the area and performance analysis of the read operation on a simulated 8 Gb 3-dimensional memory chip and compare the results to those of other nonvolatile memory products fabricated in similar process technologies.

### A. Memory Read Operation

A memory array is accessed wordline-by-wordline. During a read operation, the selected wordline is raised to $V_{\mathrm{READ}}$ and

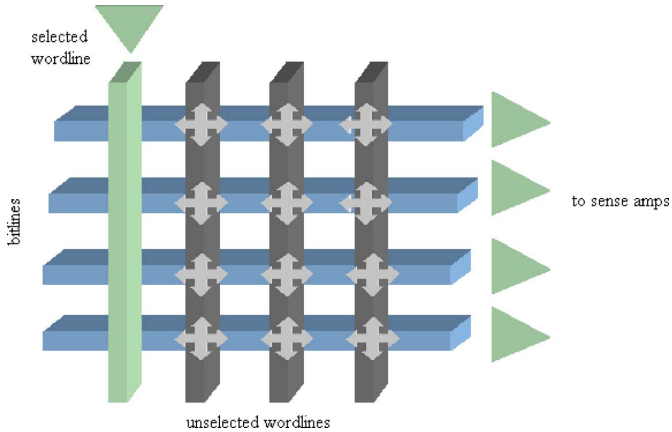



Fig. 1. A read operation performed on a cross-point array. The leftmost wordline is biased for reading, and the arrows indicate possible leakage current paths.

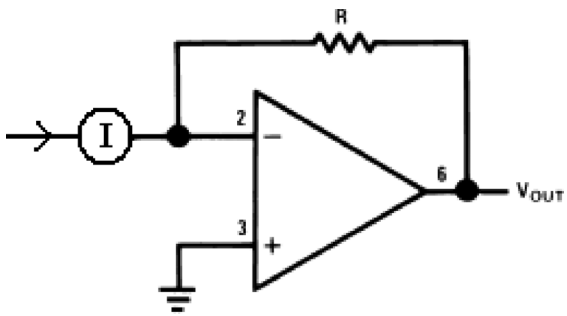



Fig. 2. Op-amp current-to-voltage converter.

a read current is driven in parallel through the bitlines. The unselected wordlines are terminated with high impedances while each bitline is connected to an individual sense amplifier at one end. In this manner, the sense amplifiers should provide the only current path to ground. Fig. 1 depicts the current leakage paths through the unselected wordlines. If the voltage difference between bitlines is minimized, then so is the leakage current. Hence, the sense amplifiers should maintain the bitlines at a nearly constant voltage regardless of the state of the memory cell. Current-sensing amplifiers provide significant reductions in bit-line voltage swing and sensing delays over voltage-sensing amplifiers [13].

An operational-amplifier can be implemented as a current-to-voltage converter, as shown in Fig. 2. The op-amp maintains the inverting input at virtual ground and the entire input current is directed across resistor $R$. The benefit of this design is that every bitline can be maintained at the same virtual ground, while the current differential between a high-resistance and a low-resistance state is seen as the voltage drop across resistor $R$. With this ideal sense amplifier design, the memory array will see no leakage current whatsoever. Unfortunately, in practice, the complexity level of an effective op-amp circuit renders it impossible to be fitted within the bitline pitch.

To address this challenge, a simpler alternative consisting of a diode-connected NMOS transistor followed by a current mirror is used to detect the read current from a bitline, as shown in Fig. 3. The mirrored current is compared to a bias current set

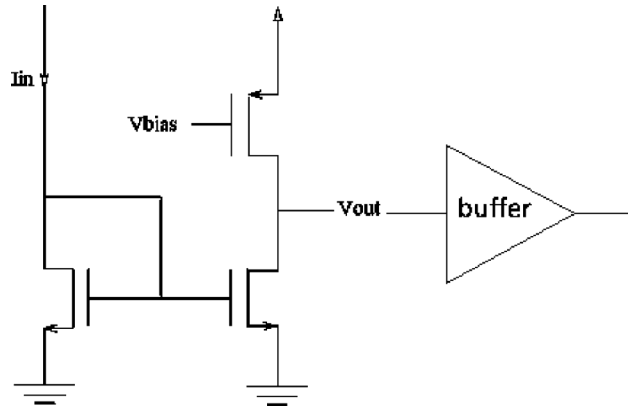



Fig. 3. Single gain stage current-sensing amplifier.

by the gate voltage $V_{bias}$ on the PMOS transistor in the reference branch, and this controls the voltage signal $V_{out}$, which is buffered and latched. The diode-connected NMOS must have a sufficiently low input resistance relative to the resistances of the memory cells to ensure that it provides enough conductance for sensing current. The NMOS transistor in the current-mirror branch must be sufficiently sized to detect the small amount of voltage swing.

### B. Memory Cell Characterization

The memory array design depends strongly on the resistance-change memory cell characteristics. There are a number of metal oxide materials that offer reversible, voltage-induced, resistance changes. The resistance of the chosen material must be high enough to regulate leakage current in both the low-resistance and high-resistance states. The metal oxide should have low write-energy requirements, as high current or voltage requirements for programming would require extra peripheral circuitry and limit future scalability. In this work, the target device is a $HfO_2$-based resistance-change material [6]. The low-resistance state has a resistance of approximately 2 k$\Omega$ ($R_{\rm ON}$) and the high-resistance state has a resistance of over 1 M$\Omega$ ($R_{\rm OFF}$).

### C. Design Tradeoffs

The key variables to be determined in designing a read circuit architecture to support a memory array are: array dimensions, wordline driver size, and sense amplifier input impedance. The leakage current path between bitlines can be moderated if the input impedance of the sense amplifiers is sufficiently low. Alternately, the number of wordlines can be reduced, giving the read current fewer leakage paths. Finally, the wordline drivers can be made larger, resulting in a larger absolute current differential between the bitlines. The design methodology presented can be applied in the same manner when later optimizing the memory architecture for the write operation.

A cross-point memory array using a resistor model of the $HfO_2$-based RRAM device was created in HSPICE using a 0.18 $\mu$m process technology with a nominal voltage of 1.8 V. The array was simulated for accuracy while varying the number of wordlines, the sense amplifier NMOS sizes, and the wordline

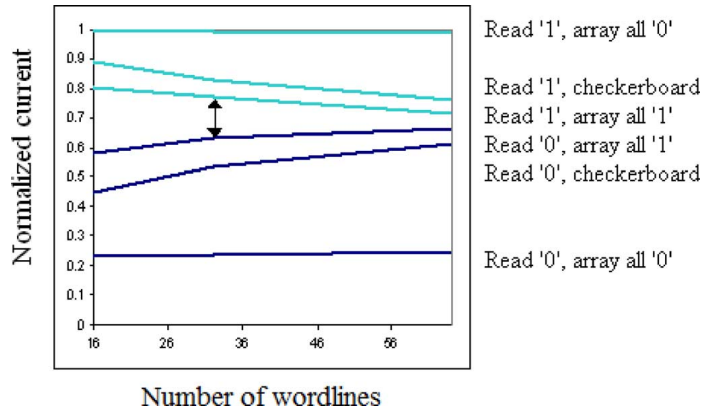



Fig. 4. Normalized current between the bitline and sense amplifier input under various array sizes and states. The upper lines represent the output when accessing a low-resistance cell, and the lower lines represent the output when accessing a high-resistance cell.

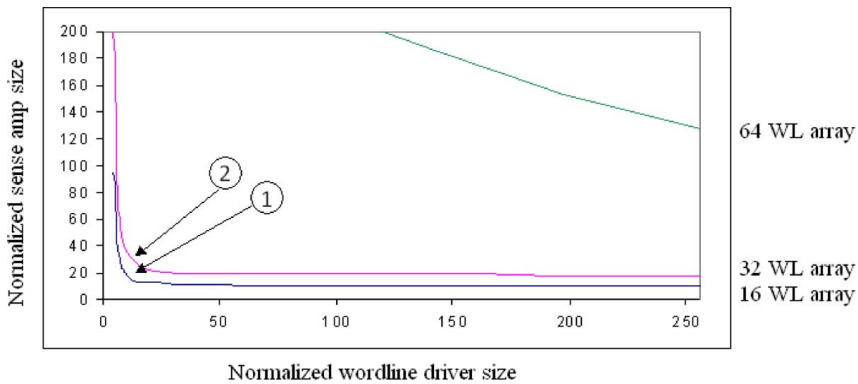



Fig. 5. Sense amplifier design tradeoffs. The optimum sense amplifier size for a 16-wordline array is marked (1), and the optimum sense amplifier size for a 32-wordline array is marked (2).

driver transistor sizes. Read accuracy was determined by accessing the leftmost wordline in a memory array and measuring the amount of current reaching the sense amplifier input from the bitlines. Fig. 4 illustrates the degradation of the differential between the sense current of a high- and low-resistance cell when the number of wordlines increases, and when most cells are in the low-resistance (logical "l") state. The current is normalized to the output of a low-resistance memory cell when the rest of the array is in a high-resistance state. In order to properly sense a signal, there should be a sufficient difference in the current between the high- and low-resistance states on the active wordline when the remainder of the array is in the low-resistance state, as this array configuration allows the most leakage current across unselected cells. It is determined that a 15% current differential at the sense amplifier input is sufficient for a minimum-size buffer to latch the sense amplifier output following a 4× current gain stage.

The results in Fig. 5 show the minimum transistor sizes that can be used during a worst-case read operation. The transistor sizes in the figure are normalized to the minimum transistor size for the process technology used in simulation. The rationale behind the sizing tradeoffs can be understood from Fig. 6, which is a simplified representation of the resistive network that occurs during a read operation in the worst-case array state. In order for a sense amplifier to be able to differentiate between the input currents $I_{\rm ON}$ and $I_{\rm OFF}$ in this scenario, the leakage current from the low-resistance bitline ($R_{\rm ON}$) to the high-resistance bitline ($R_{\rm OFF}$) must be less than the original bias current across the memory cell. The amount of current traversing from the $R_{\rm ON}$ bitline to the $R_{\rm OFF}$ bitline is equal to the difference between the bitline voltages ($\Delta V_{BL}$) divided by the effective resistance of all the memory cells on the bitline in parallel. As illustrated in Fig. 6, total resistance between a pair of bitlines through a wordline is $2 \cdot R_{\rm ON}/\#WL$, where $\#WL$ is the number of wordlines. Since each bitline has two neighboring bitlines (except for the first and last bitlines in an array), the worst-case effective resistance for leakage current is $R_{\rm ON}/\#WL$.

Then,

$$\frac{\Delta V_{BL} \cdot \#WL}{R_{\rm ON}} < \frac{V_{\rm READ}}{R_{\rm ON}}$$

$$\frac{V_{\rm READ}}{\Delta V_{BL}} > \#WL.$$

This gives the minimum requirements for a functional memory array.

For an array with a certain number of wordlines, $V_{\rm READ}$ can be increased and $\Delta V_{BL}$ decreased by increasing the width of the wordline drivers and sense amplifier transistors respectively. Increasing the wordline transistor size drives more current through the selected cells to account for the greater number of leakage paths. Increasing the size of the sense amplifier transistors provides a lower-impedance path relative to that of the unselected

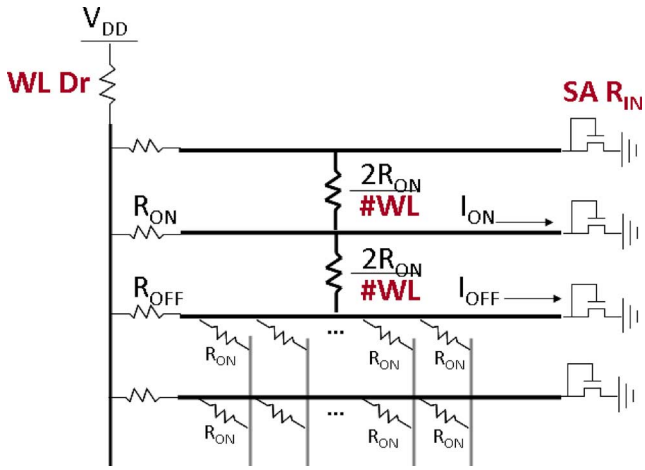



Fig. 6. Memory array design variables.

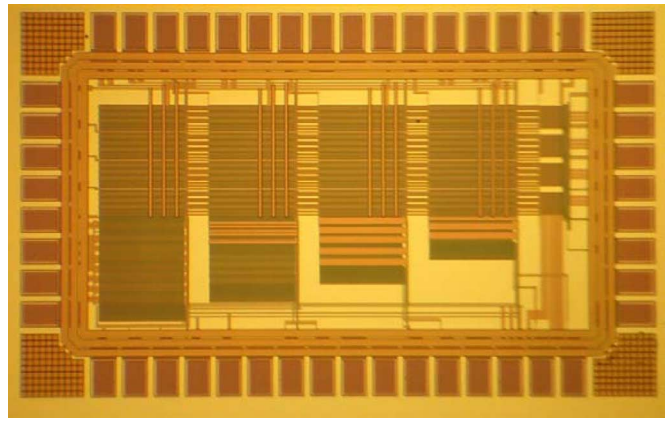


Fig. 7. Die micrograph of fabricated test chip with four sets of memory arrays.

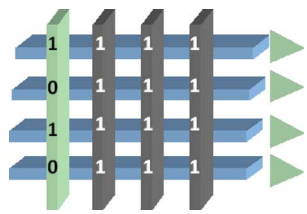



Fig. 8. Array configuration during a test read operation.

memory cells in parallel. This generates the L-shaped curve seen in Fig. 5. For a constant $V_{\mathrm{READ}}$, $\Delta V_{BL}$ must be halved to accommodate a doubling of the number of wordlines. In order to double the array size from 16 to 32 wordlines, either the wordline driver width, the sense amplifier transistor width, or both must be increased in order to reduce $\Delta V_{BL}$. When the array size is increased to 64 wordlines, the wordline drivers and sense amplifier widths must be increased in size again to the point where it is no longer practical for implementation.

Two acceptable array configurations are marked (1) and (2) in Fig. 5. Design (1) is optimized for a 16-wordline array, with a wordline driver width of 12× the minimum transistor size, and a sense amplifier width of 10× the minimum transistor size. Design (2) is optimized for a 32-wordline array and has a wordline driver width of 12× and a sense amplifier width of 40×.

## III. Cross-Point Memory Array Measurement Results

A test chip was fabricated using 0.18 $\mu$m CMOS technology. The die micrograph is shown in Fig. 7. In order to evaluate the sense amplifier performance under different memory cell resistances, the memory cells were emulated using PMOS transistors at the cross-points. The gate voltage of these transistors was externally controlled to model a wide range of low-resistance $R_{\mathrm{ON}}$ and high-resistance $R_{\mathrm{OFF}}$ values. The test procedure involved setting a variable-size memory array to a fully low-resistance configuration by biasing the PMOS memory cell gate voltages to $V_{Ron}$. The cells along the leftmost wordline alternate between high- and low-resistance states (Fig. 8) by setting gate voltages to $V_{Roff}$ and $V_{Ron}$, respectively. This wordline would then be selected and biased, and the sense amplifier outputs would be latched and shifted out through the I/O pins. For each array test, $V_{Roff}$ begins at a low enough voltage that $R_{\mathrm{OFF}}$ is almost as low as $R_{\mathrm{ON}}$. If the outputs could not be successfully differentiated at the I/O pins, then $V_{Roff}$ was increased by 10 mV and the wordline was accessed again. If, even at the highest value of $V_{Roff}$, the outputs could not be detected, then $R_{\mathrm{ON}}$ was increased and the process was repeated. In this manner, we iteratively determined the lowest possible $R_{\mathrm{ON}}$ and $R_{\mathrm{OFF}}$ combinations that could still produce a distinguishable output through the latch.

Failures (in the form of high-resistance cells being detected as low-resistance cells, or vice versa) generally first occurred at the very bottom of the memory array, and would propagate upwards to the bit closest to the wordline driver. This is probably due to the voltage drop along the wordline providing a lower read voltage to the farther cells. As will be discussed in the following section, a memory array will have 32 bitlines after optimizing the tradeoff between array size and read current on the wordline. Thus, we designated a failure mode as the point when bit number 32 along the wordline first fails.

The results of the read operation tests on a 32-bitline by 32-wordline array are plotted in Figs. 9 and 10. These results show the range of $R_{\mathrm{ON}}$ and $R_{\mathrm{OFF}}$ resistance values that can be detected by each of the sense amplifier designs. The shaded areas indicate the $R_{\mathrm{ON}}$ and $R_{\mathrm{OFF}}$ combinations that failed, while the white areas indicate resistance combinations that yielded a differentiable output. As seen in Fig. 9, with a memory array of 32 wordlines, $R_{\mathrm{ON}}$ can be as low as 1746 Ω while $R_{\mathrm{OFF}}$ can be as low as 2400 Ω using the 10×-sized sense amplifier design that was optimized for 16 wordlines.

A larger operation region can be seen from the test results of the 40×-sized sense amplifier in Fig. 10. Here, for a 32-wordline array, $R_{\mathrm{ON}}$ can be as low as 1065 Ω with $R_{\mathrm{OFF}}$ as low as 2182 Ω and still be differentiable at the sense amplifier output. In practice, there will be deviations in the memory cell resistances due to process variations and operating conditions. The experimental results in Figs. 9 and 10 confirm that our sense amplifier design can cover the memory cell variations present in the target resistance-change device [6], as well as the mismatch in MOSFET threshold voltages.

The supply voltage could be reduced to 1.6 V while reading a 32-wordline array using the 10× sense amplifier design, and 1.4 V using the 40×-sized sense amplifier design. There was

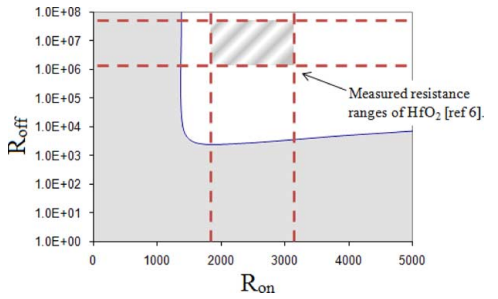



Fig. 9. Minimum detectable $R_{\rm ON}$ and $R_{\rm OFF}$ values in a 32-wordline array with 10×-sized sense amplifier. The shaded area indicates on/off resistance combinations resulting in failed read operations, while the white area indicates on/off resistance combinations resulting in successful read operations. Actual device values are shown in the overlaid box.

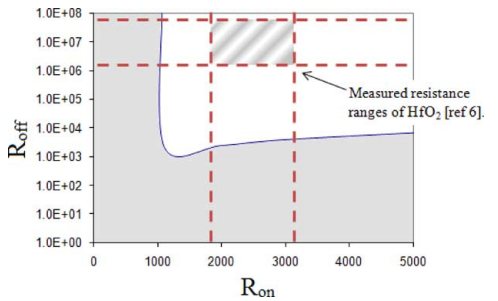



Fig. 10. Minimum detectable $R_{\rm ON}$ and $R_{\rm OFF}$ values in a 32-wordline array with 40×-sized sense amplifier.

no apparent benefit to increasing $V_{DD}$ to 2.0 V in either design. The value of $R_{\rm ON}$ must increase with a lower supply voltage because the amount of current generated from the memory cell will no longer be sufficient to withstand the effects of leakage across unselected cells. Ultimately, the sense amplifier completely fails because the latch is sized for a fixed range of sense amplifier output voltage. When the supply voltage becomes sufficiently low, then the output voltage falls outside the range that can be latched. The 40×-sized sense amplifier does not suffer from this problem because the original sense amplifier output range begins at a lower voltage due to the wider NMOS transistors and their lower effective resistance.

## IV. Three-Dimensional Memory Architecture

An array with 32 wordlines provides the most area-efficient solution for a cross-point array given the target memory cell characteristics presented in Section II. The number of bitlines does not affect the sense accuracy, but is limited by the amount of current that a wordline can drive during the write operation, where cells will be programmed in parallel. Based on a programming current of about 40 $\mu$A per cell, 32 bitlines will require a total programming current of approximately 1.2 mA, which is manageable. Thus, the chosen memory block size is 32 bitlines by 32 wordlines. There are no selection transistors within the block.

A single page architecture is shown in Fig. 11. The bitlines in each block are connected to the global bitlines through bitline-selection transistors. The wordlines in each block are connected to voltage buses through wordline driver transistors. There are 128 blocks in a row because this results in about a 500 Ω resistance along the global bitlines between the farthest memory cells and the sense amplifier inputs, keeping the voltage drop across the global bitline small enough to maintain the sense margin. There are also 128 blocks in a column. The limiting factor is the delay along the polysilicon line that runs vertically down the page and controls the block-selection transistors. The column decoders are located at the top of the page and control the bitline-selection transistors down a column of blocks. During a page access, an entire column of blocks is read in parallel by selecting a single wordline. The wordline decoders are located on the left side of the page and propagate control signals for the wordline drivers horizontally across the entire page. The horizontal routing is necessary to ensure that control signal lines do not interfere with wordlines when multiple layers of memory are used.

A popular approach in two-dimensional memory designs involves the arrangement of support circuitry between adjacent memory arrays [14]. A page is divided into two halves with 128 by 128 blocks on each half and a single set of sense amplifiers and latches in the middle. A read operation is performed in two cycles. In the first cycle, the odd-numbered blocks in a column are accessed on the left half of the page. The even-numbered blocks are accessed in the corresponding column on the right half of the page. In the second cycle, the even-numbered blocks are accessed in the same column on the left half of the page, and the odd-numbered blocks in the same column are accessed on the right half of the page. In these two cycles, an entire column of bits is read on both sides of the page. The column decoder provides an identical output for each half of the page. Because alternating blocks are accessed for each read cycle, it is possible to share wordline-driver transistors between vertically adjacent blocks.

With this architecture, selection transistors are required for every bitline and wordline of each block. The bitline-selection transistors connect the bitlines in each block to the global bitlines and serve to isolate the block and prevent additional leakage current from other blocks. These selection transistors need to fit in the minimum metal pitch. The wordline-selection transistors are used to drive the selected wordline to the desired voltages during read and write operations. In order to minimize area overhead, the wordline-selection transistors are shared between vertically adjacent blocks. Both wordline and bitline selection transistors are NMOS devices. This avoids the p-well spacing constraints involved if complementary devices are used. The body effect requires that a slightly wider transistor be used for driving a positive voltage into the wordline or bitline. A bitline only needs to accommodate up to 50 $\mu$A during a read operation. A minimum-size selection transistor may be used when fabricating the memory array in typical process technologies. When reduced to layout, the transistor will occupy a horizontal width of approximately 6$\lambda$. With staggered placement and routing, these selection transistors can be fit into a minimum bitline pitch as shown in Fig. 12. The resulting width is 32$\lambda$, and spans 8 wordline-widths after taking substrate contacts and diffusion spacing into account. As shown in Fig. 13, the wordline drivers can be staggered to fit into the minimum metal-width pitch. The wordline drivers for a 32-bitline block occupy a height of 24 bitlines.

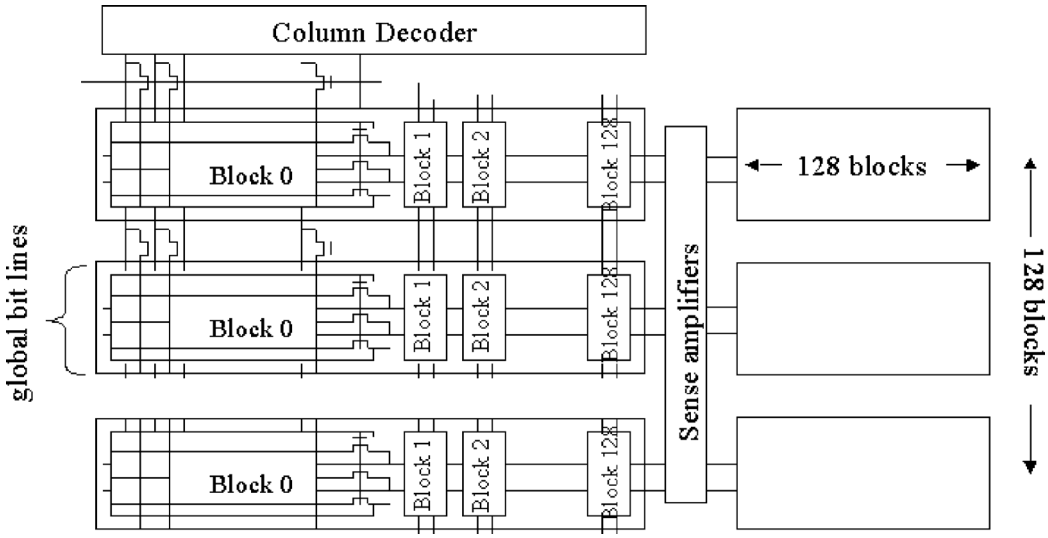



Fig. 11. Hybrid NAND/NOR architecture of cross-point memory. This shows a 32 Mb page consisting of 128 by 128 blocks. Each block has 32 wordlines and 32 bitlines.

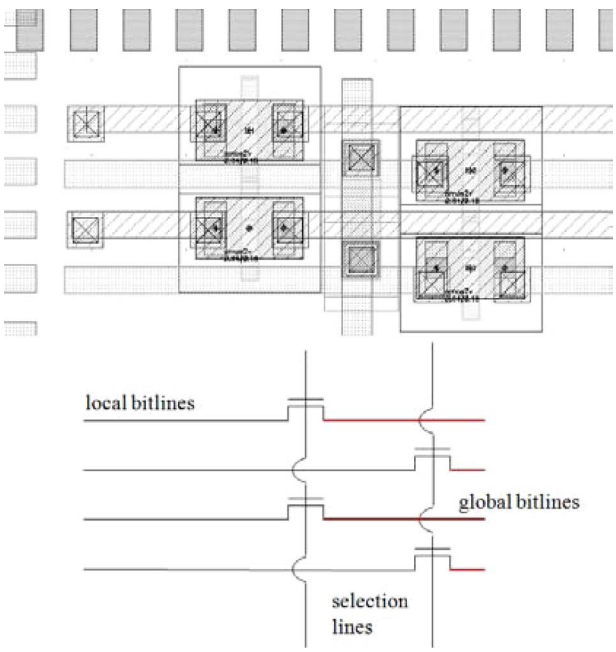



Fig. 12. Layout and schematic of four bitline-selection transistors in minimum bitline pitch.

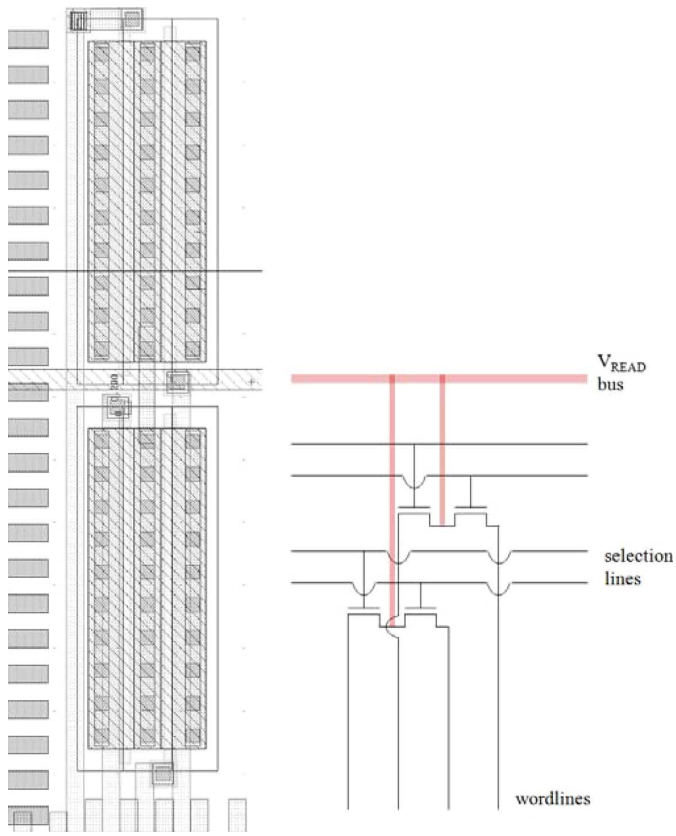



Fig. 13. Layout and schematic of wordline-selection transistors in minimum wordline pitch.

### *A. Layout Techniques and Array Efficiency*

Array efficiency is defined as the total area occupied by the memory cells divided by the total combined area of the memory and the peripheral circuitry. For a single layer of cross-point memory, it is possible to fold the access circuitry entirely underneath the memory array. The bitline-selection transistors span a width of 8 metal lines and the wordline drivers span a height of 24 metal lines, thus the total selection-transistor overhead will occupy the same area as a 32 wordline by 32 bitline block when packed as tightly as possible. Fig. 14 gives a three-dimensional perspective of the vertical vias and local interconnect involved with routing bitlines to the selection transistors. The same technique can be employed for the wordlines with an additional metal layer to bypass the bitline-selection transistors. To enable wordline-driver sharing, contacts and associated selection circuitry lie on alternating bitlines and wordlines. The corners of the blocks are already occupied by the bitline transistors, thus wordline drivers need to be folded under the adjacent blocks. The remaining space under the blocks is reserved for the bitline-selection transistors of neighboring blocks. The bitline and

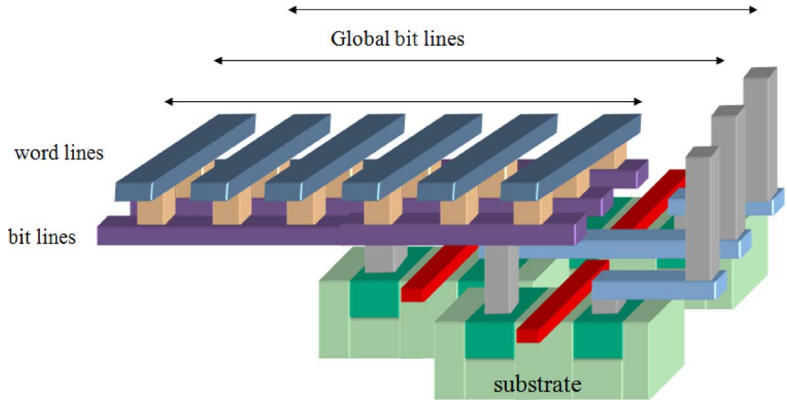



Fig. 14. Cross-sectional view of bitline-selection transistors made to fold under a cross-point memory array.

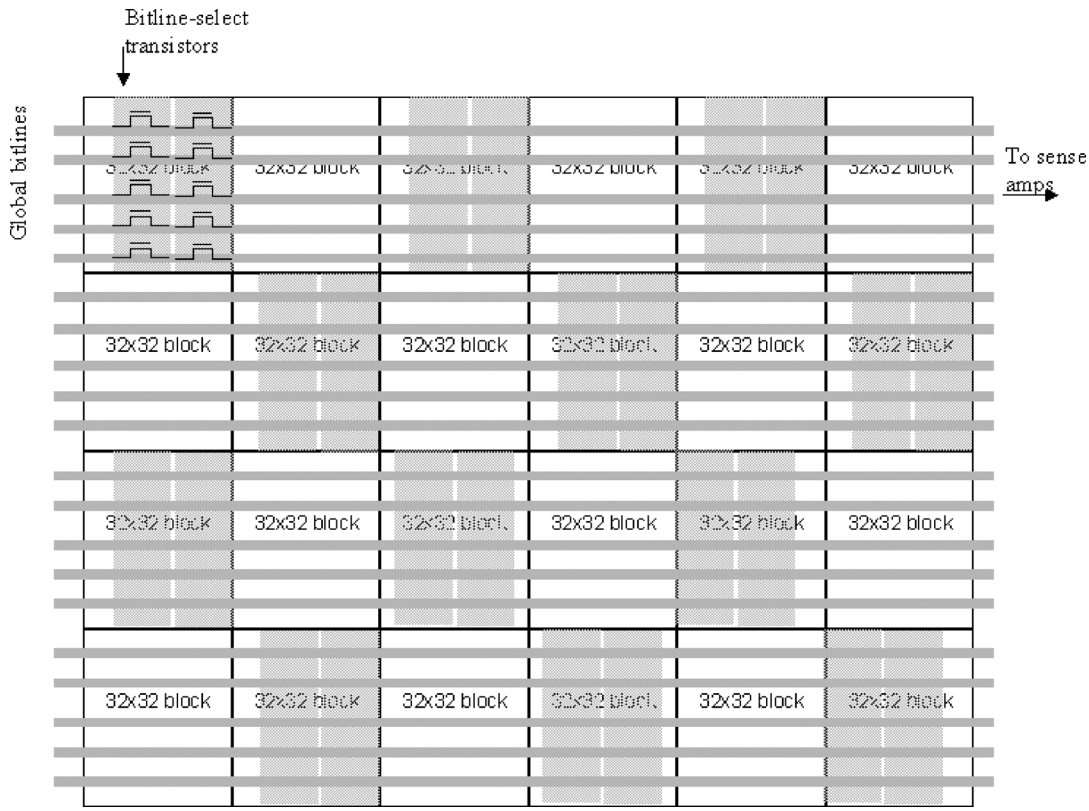



Fig. 15. Schematic of bitline-selection transistors made to fold under a cross-point memory array of 32-bit by 32-bit blocks connected to global bitlines.

wordline transistors are placed under adjacent memory blocks in a checkerboard pattern as shown in Figs. 15 and 16.

A page consisting of 32-bit by 32-bit memory blocks with the selection circuitry folded underneath the arrays as described above achieves an array efficiency of 91.3% (without taking into account the decoders or sense circuitry), as the only peripheral circuitry within the blocks that cannot lie beneath the memory array are the contacts to the global bitlines and wordline voltage buses.

### *B. Stacking Memory Layers*

A second memory array can be stacked on top of the first with little area overhead as illustrated in Fig. 17. In a single-layer layout, contacts to the wordline voltage buses occupy two peripheral edges of each memory array, and contacts to the global bitlines occupy one peripheral edge. The remaining peripheral edge can be used to route a second layer of bitlines through the selection devices to the global bitlines. Wordlines will be shared between two layers of bitlines. Only one layer of bitlines will be active at a time, so the wordline drivers will not drive any more current in a read operation than with the single-layer design. The additional bitlines should have no significant effect on the leakage current during the read or write operations as they are essentially ”floating” when unselected, with no path to ground. However, by layering bitlines between two sets of wordlines, the opportunities for current leakage effectively doubles in the worst-case array configuration. Thus, memory cell devices will

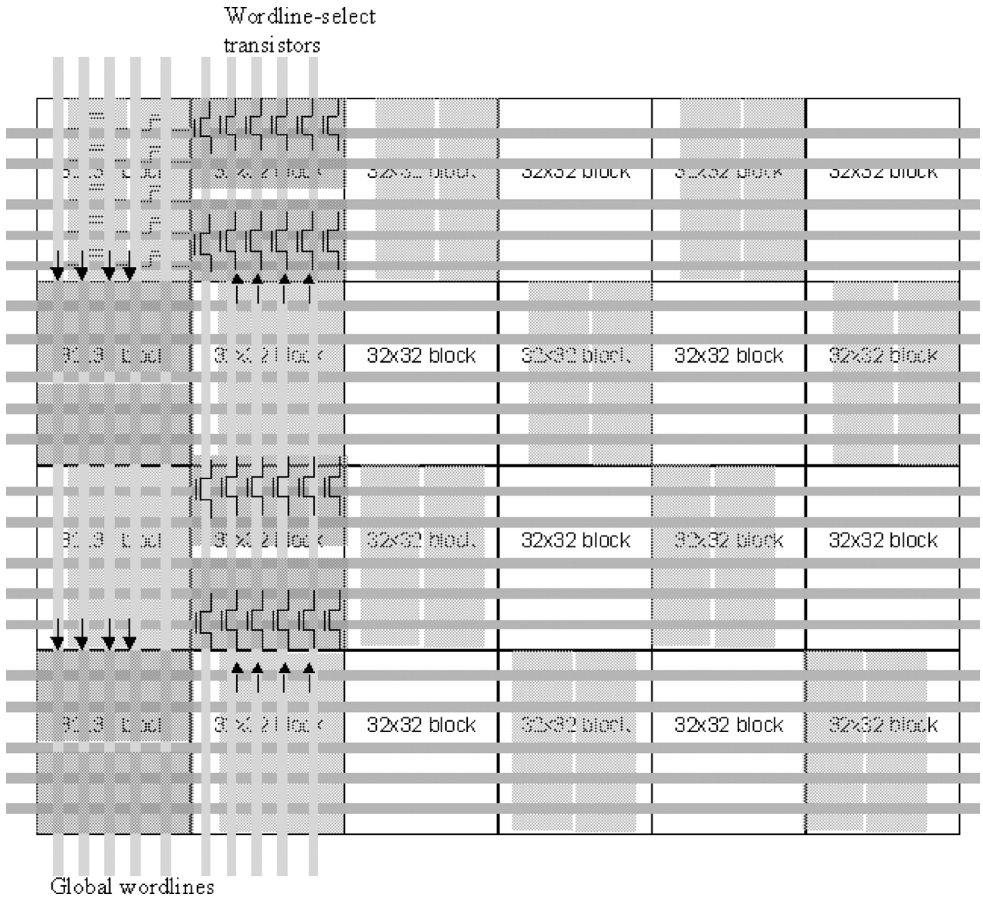



Fig. 16. Schematic of wordline drivers made to fold under a cross-point memory array made up of 32-bit by 32-bit blocks connected to global wordlines.

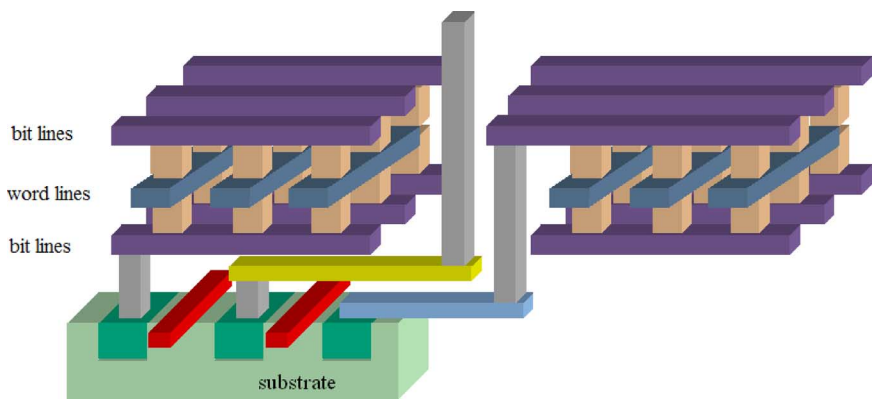



Fig. 17. Two-layer memory with shared wordline and bitline selection transistors.

need to fall within a tighter range of resistance values. The only additional area overhead is the space at the array edge where an additional set of bitline contacts must be inserted. The effective array efficiency is now 88.5% for two layers of memory cells in a 32 by 32 block.

With some additional decoder complexity, it is possible to build up to four memory layers by adding two sets of wordlines as shown in Fig. 18. The top and bottom wordline layers (1) and (3) can share a set of selection transistors, as they deliver current to different bitline layers. The middle wordline layer (2) is connected to driver transistors on the opposite side of the array. All wordline driver transistors are connected to a global bus that provides the appropriate voltage bias. The wordline-selection circuitry for the two-layer design occupies a height of 24 bitlines. After accounting for spacing and metal via requirements, an additional height of 10 bitlines will be necessary. This decreases the array efficiency to 71.7%. Now that all four edges of each memory array are surrounded by vias, it would be difficult to increase the number of memory layers beyond four without a significant decrease in array efficiency.

### C. 3D Memory Area Comparisons

In a 65 nm process technology, a 32-by-32-bit cross-point array (1 Kbit) has an area of 17.3 $\mu m^2$. Since our array archi-

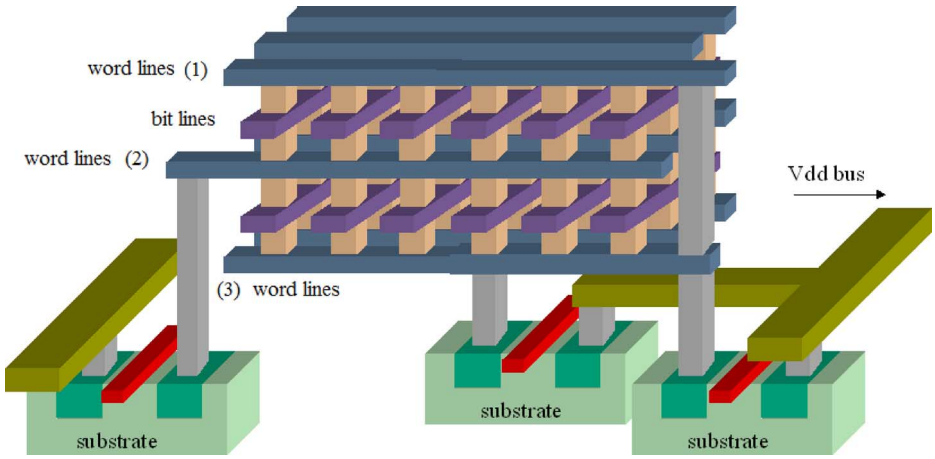



Fig. 18. Four-layer memory with wordline driver connections.

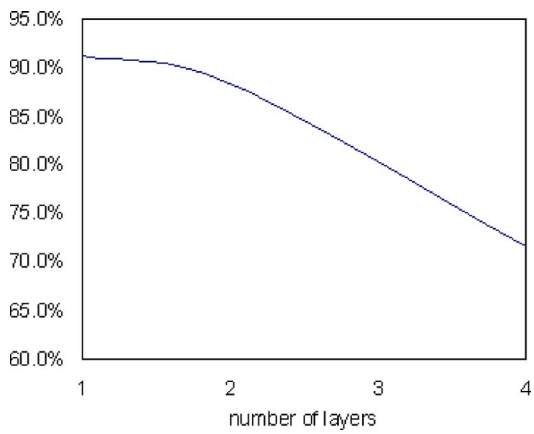



Fig. 19. Array efficiency vs. number of vertical memory layers, assuming 32-by-32-bit arrays.

tecture allows wordline and bitline selection transistors to be placed under the memory array, as shown in Figs. 15 and 16, we assume a conservative estimate of 50% memory area efficiency. Then, approximately 4 Gbits will fit on a 139 $\mathrm{mm}^2$ die with the single-layer cross-point memory design, with an overall bit-efficiency of 28.9 $\mathrm{Mbits}/\mathrm{mm}^2$. Assuming the peripheral circuitry (decoders, buffers, sense amplifiers, etc.) occupies the same area for a multi-layer memory as for a single-layer memory, and that the only difference in total layout area is the decrease in array efficiencies shown in Fig. 19, then a two-layer cross-point memory will have a die efficiency of 48.5%, but allow 8 Gbits to fit on a 144 $\mathrm{mm}^2$ die. A four-layer cross-point memory will have a die efficiency of 39.3%, but an 8 Gbit capacity will occupy a die area of only 89 $\mathrm{mm}^2$.

Table I lists some production NAND flash memory devices fabricated in similar process technologies and their area efficiencies to provide a basis of comparison to the multi-layer cross-point memories presented in this work. Even with a conservative efficiency estimate, the four-layer cross-point memory will achieve a far better bit-efficiency than a multi-level NAND flash that stores two bits per cell.

TABLE I
NAND FLASH MEMORY AREA EFFICIENCY COMPARISONS

| | memory size | die size ($mm^2$) | bit-efficiency (Mbits/$mm^2$) | die efficiency |
|---|---|---|---|---|
| 65nm SLC NAND [15] | 4 Gb | 131 | 31.3 | 54% |
| 65nm SLC NAND [16] | 4 Gb | 137 | 29.2 | 60.4% |
| 50nm SLC NAND [17] | 8 Gb | 170 | 47.2 | 65% |
| 63nm MLC [18] | 8 Gb | 133 | 61.6 | 70% |
| 65nm 1-layer X-point | 4 Gb | 139 | 29.5 | 50% |
| 65nm 2-layer X-point | 8 Gb | 144 | 56.9 | 48.5% |
| 65nm 4-layer X-point | 8 Gb | 89 | 92.4 | 39.3% |

## V. PERFORMANCE ANALYSIS

An 8 Gb cross-point memory array built in four layers was modeled and simulated in HSPICE using device parameters for a typical 65 nm CMOS process with a nominal operating voltage of 1.2 V. The simulation includes all the peripheral circuits, address buffers, decoders, sense amplifiers, and output drivers.

### A. 8 Gb Memory Architecture

A general block diagram for the 8 Gb memory is shown in Fig. 20. Four layers of 32-bit by 32-bit blocks are tiled into a page, with 128 rows and 256 columns of blocks per page. The column decoders control the bitline-selection devices and are located above the page. The wordline decoders control the wordline-selection devices and are located to the sides of the page. The sense amplifiers used in this model are the 40×-sized sense amplifiers described in Section II. A page stores 32 Mb per layer, so a four-layer page has a memory capacity of 128 Mb. The memory utilizes a single-core architecture with 64 four-layer pages. A design with 32 pages in two vertical columns was chosen for simulation as this layout resembles most NAND flash architectures.

### B. Decoder Architecture

A complete read operation reads two columns of bits, one from each half of a page. This takes two cycles, as a read cycle

TABLE II
NONVOLATILE MEMORY PERFORMANCE COMPARISONS.

| | tR_access | tR_sequential | $I_{avg}/I_{max}$ | $P_{avg}/P_{max}$ |
|---|---|---|---|---|
| 1Gb NOR Flash [20] | 100 ns | 25 ns | 21 mA/24 mA | 35.7 mW/40.8 mW |
| 8Gb NAND Flash [17] | 25 $\mu$s | 30 ns | 15 mA/30 mA | 40.5 mW/81 mW |
| 1Gb 3-D OTP [21] | 140 $\mu$s | 100 ns | 20 mA/30 mA | 54 mW/81 mW |
| 4-layer 8Gb Cross-point | 104 ns | 201 ps | 27 mA/51 mA | 32.4 mW/61.2 mW |

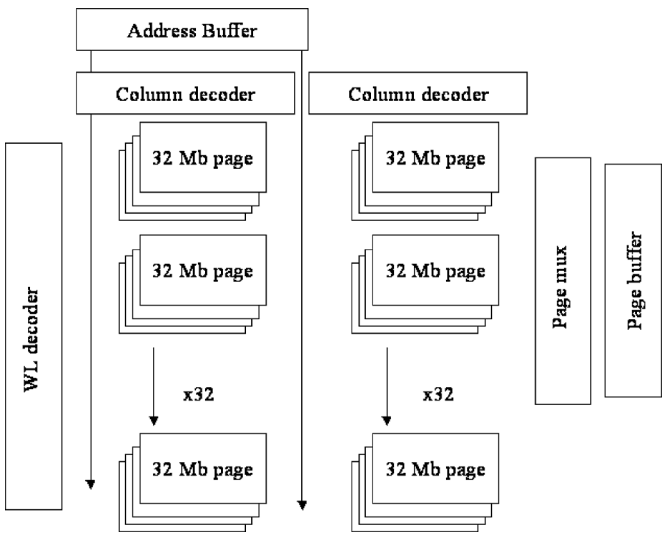



Fig. 20. Four-layer cross-point memory implemented in an 8 Gb architecture.

accesses alternating rows from each half. Each read cycle senses 4 Kb of data. A 21-bit address is required to access a column of 4096 bits in a single read cycle. Every page receives a 5-bit wordline address to select one of 32 wordlines. The wordline-selection signal propagates across the page. Each page also receives an 8-bit column address and a 2-bit layer selection code. The column decoder output controls the bitline-selection transistors. The wordline address is decoded using a two-stage decoder and the column address enters a three-stage decoder. The page outputs ultimately pass through a multiplexer controlled by a 6-bit, two-stage page decoder.

### C. Critical Path

The slowest-case read operation occurs when trying to access one of the bottom-most pages in the two columns shown in Fig. 20. In this case, the address bits must be propagated from the address buffer all the way across the die. The critical path follows the address bits to the wordline decoder, which must drive the capacitance of a metal line the width of a page as well as 256 wordline drivers. The local wordlines and bitlines charge within picoseconds, as they are only 32 bits in length. The remainder of the critical path involves charging the global bitlines, sensing and latching the bitline signal, and propagating the data to the output buffer.

All decoders and buffers are built using minimum-size transistors, with the exception of the wordline drivers and the ouput buffer, which must drive a 10 pF load. In order to achieve a fast burst speed, the output buffer is optimized for a fan-out of 4 (FO4). An eight-stage buffer is used to drive a 10 pF output load. The interface for the 8 Gb memory will have 8 I/O pads. The 21 address bits will be stored in an address buffer in three address cycles. For these simulations, each single read cycle is considered independently addressed. Each read cycle stores 4 kb of data in the page buffer, shifted out 8 bits at a time.

### D. Simulation Methodology

Parasitic interconnect delay is the greatest contributor to read latency in the cross-point memory array. Long bitlines and wordlines are fabricated in the minimum metal pitch, leading to high parasitic resistance and capacitance. In order to accurately model parasitic capacitance, each wire and memory element must be treated as a three-dimensional structure in metal or polysilicon, interacting with all of the surrounding wires and the ground plane. Parasitic capacitances in the multi-layer cross-point memory arrays were modeled and extracted using Ansoft Q3D Extractor [19].

### E. Waveform Analysis and Timing Diagrams

Fig. 21 shows the modeled waveforms of a single access performed on the 8 Gb four-layer memory architecture. A worst-case access was simulated, reading a bit from the left-most column of the bottom-most page in the memory structure.

The address signal takes over 100 ns to fully propagate to the decoder input of the farthest page from the address buffer. However, the wordline selection transistors can latch the address by 30 ns. The global wordline is charged by 35 ns, and the local wordline can drive sufficent current for sensing to the local bitlines by 45 ns. The global bitline is charged at 50 ns and the sense amplifier output can be latched by 66 ns. The data from the page buffer can be propagated through the I/O driving a 10 pF load by 70 ns, but the page buffer output lines are not stabilized until 104 ns. After a stable signal is latched by the page buffer at 104 ns, the page buffer data can be shifted out and the next read cycle can commence. Only 8 bits are available for the I/O, while a single read cycle accesses 4096 bits. Thus, it takes 512 cycles to fully output the 4096 bits. An eight-stage output buffer optimized for a fan-out of 4 can shift out each set of data in 173 ps, but it could be clocked as slowly as 201 ps so that the data is shifted out within a read cycle of 104 ns.

### F. Capacitive Noise Considerations

Due to the close proximity of adjacent wordlines and bitlines, capacitive coupling is a concern in the design. The crosstalk capacitance extracted from the 3-D model is 0.615 fF between adjacent bitlines or wordlines. The vertical capacitance between the metal lines is only 0.0892 fF. Fig. 22 shows the transient voltage induced on adjacent wordlines and bitlines in the array

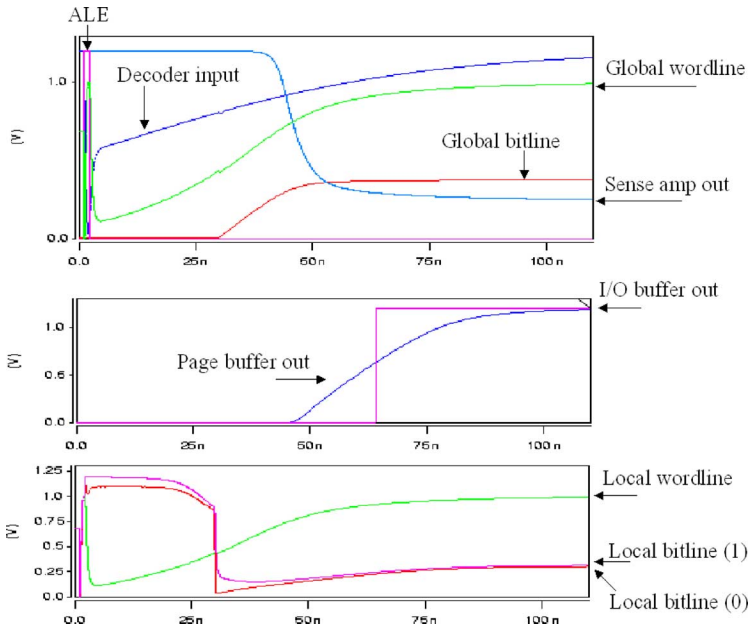



Fig. 21. Simulation waveforms for 8 Gb memory read access.

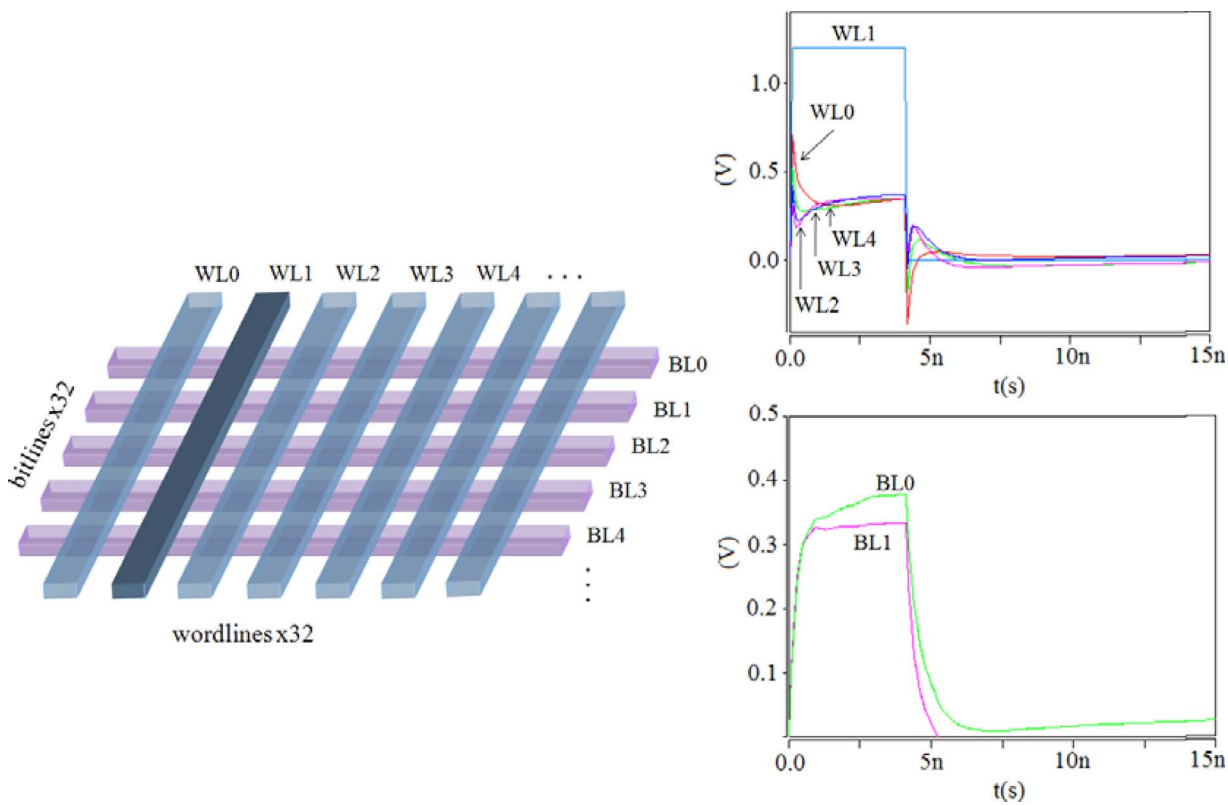



Fig. 22. Transient effects of capacitive coupling on adjacent wordlines and bitlines in a 32-by-32-bit array. WL1 is raised to 1.2 V for a read operation.

as wordline WL1 is raised to 1.2 V during a read operation. The unselected wordlines are left floating. The resistance-change memory cells along the selected WL1 alternate between ON and OFF states. The immediately-adjacent wordlines WL0 and WL2 see a peak coupling voltage of 0.65 V, and then settle to between 0.3 and 0.4 V. The other wordlines also settle to this range. This is also in the same range as the bitline voltages, BL0 with an ON-state cell and BL1 with an OFF-state cell. Thus, the unselected wordlines do not induce any additional leakage current and there are no significant coupling effects on the bitlines.

### G. Performance Comparisons With Commercial Products

Table II shows a comparison of the performance of various types of nonvolatile memory currently in production. The 3D cross-point memory has an access time (tR_access) almost as fast as that of a lower-capacity NOR flash, and a burst read time

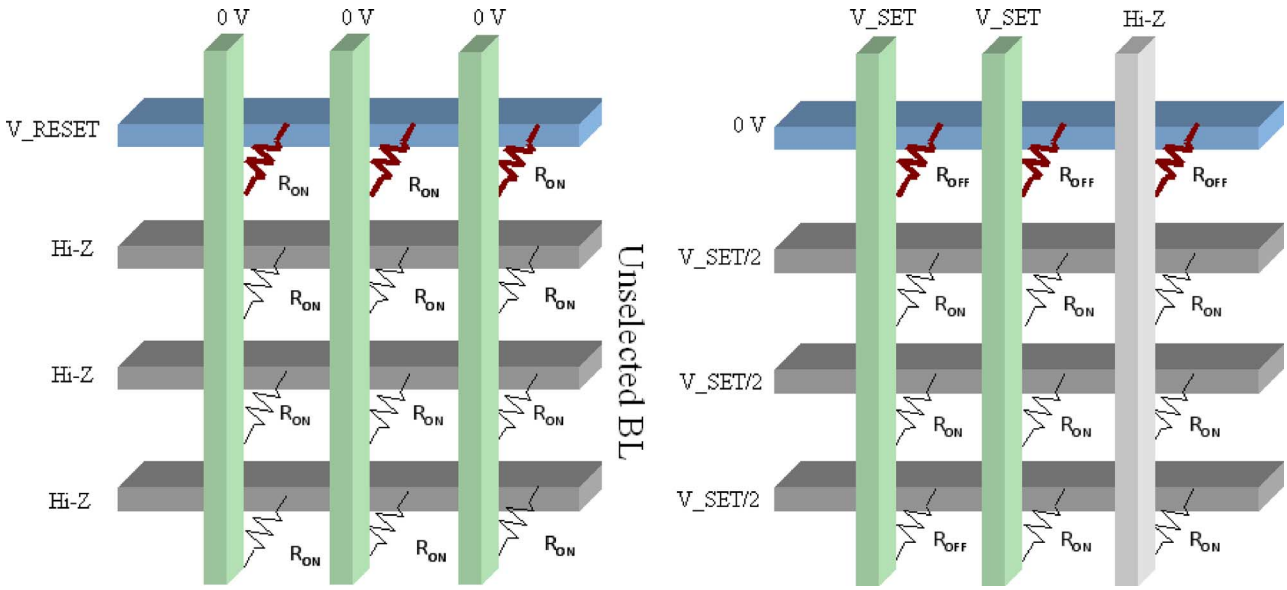



Fig. 23. A RESET (left) and a SET (right) operation performed on a cross-point array.

(tR_sequential) that is much faster than that of NAND flash. However, the 8 Gb cross-point memory architecture described in this chapter does not implement any sort of error-correction or redundancy, which adds some amount of overhead to the access times of flash memory. The power requirements for the read operation are also compared in Table II, and are similar to that of other designs. In the worst-case read operation, the current draw for propagating an 8-bit address to the farthest page is 802 $\mu$A, the current draw for reading and latching 4096 bits in parallel is 27.3 mA, and the current draw for buffering the 4096 bits to the output is 22.8 mA. The total power consumption is 61.2 mW on the critical path.

## VI. Memory Write Operation

Programming resistance-change memory cells is accomplished by applying either a SET voltage ($V_{SET}$) or a RESET voltage ($V_{RESET}$). "SET" is defined as the transition of cells from a high-resistance state to a low-resistance state, while "RESET" brings the cells back to a high-resistance state from a low-resistance state. Fig. 23 shows a representation of the RESET and SET operations. The entire memory array is RESET bitline-by-bitline before the SET voltage is applied to those cells that need to be programmed. This enables us to predict the current requirements for the SET operation to avoid over-programming (bringing memory cells to too low a resistance). Over-programming is not a concern during RESET because the higher resistance will automatically limit the current flow. In the RESET operation, $V_{RESET}$ is applied sequentially to one bitline at a time. All wordlines in the array are pulled to ground (0 V), and every cell in this row of bits becomes RESET. The unselected bitlines are terminated with high-impedances so that the only current path is from the selected bitline to the wordlines. It is expected that the unselected bitlines will drift to some voltage slightly higher than 0 V. Leakage current is not a significant concern in the RESET operation because every cell in the array will eventually be brought to the high-resistance state. In the SET operation, 0 V is applied sequentially to one bitline at a time, while $V_{SET}$ is applied only to selected wordlines. The unselected wordlines are terminated with a high-impedance and are expected to drift to some voltage between ground and $V_{SET}/2$. Unselected bitlines are biased at $V_{SET}/2$. The unselected cells see a bias of at most $V_{SET}/2$. While this allows some leakage current in the array, it is necessary to prevent unselected bitlines from drifting to too low a voltage and causing an inadvertent SET. The capacitive coupling effects during SET and RESET need to be carefully simulated to avoid accidental programming.

Although some resistance-change memory cells can be set to intermediate low-resistance states by modulating the SET current, a SET voltage close to $V_{SET}$ is still required [6]. The current requirements during the write operation may further limit the size of the memory block. This will be a subject for future investigation.

## VII. Conclusion

A novel sense amplifier and memory architecture for resistance-change memory has been designed and experimentally verified in this work. The cross-point memory array does not require individual access transistors, thus the memory architecture can be integrated into a 3-dimensional stacked structure simply by layering the arrays. The resulting multi-layer memory array has an expected bit-density exceeding that of single- and multi-level NAND flash. It should be noted that memory cell device characteristics become increasingly stringent with multiple memory layers, as vertical connections between memory array layers allow for more opportunities for leakage current. Test chip measurement results show that the sense amplifier design described in this work can indeed tolerate increased leakage current from connecting multiple layers of memory when memory cell variations are within target device specifications.

Simulation results also show that an 8 Gb memory architecture can be accessed with power and speed that are competitive with those of NOR flash. The greatest benefit of a resistance-change cross-point memory over NAND flash is its scala-

bility. A four-layer cross-point memory can have a significantly greater bit-density than NAND flash memories fabricated in the same technology node. This provides a great advantage in the cost-per-bit scaling for future nonvolatile memory.

## REFERENCES


[1] ITRS Roadmap [Online]. Available: http://www.itrs.net/

[2] R. E. Scheuerlein, "Magneto-resistive IC memory limitations and architecture implications," in *1998 Proc. 7th Biennial IEEE Int. Nonvolatile Memory Technology Conf.*, Jun. 1998, pp. 47–50.

[3] S. Lai and T. Lowrey, "OUM – A 180 nm nonvolatile memory cell element technology for stand alone and embedded applications ," in *Int. Electron Devices Meeting, IEDM 2001 Tech. Dig.*, Dec. 2001, pp. 36.5.1–36.5.4.

[4] M. Gill, T. Lowrey, and J. Park, "Ovonic unified memory – A high performance nonvolatile memory technology for stand-alone memory and embedded applications," in *IEEE ISSCC Dig. Tech. Papers*, 2002, pp. 202, 459.

[5] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible switching effect in thin oxide films for memory applications," *IEEE Electron Device Lett.*, vol. 77, no. 1, pp. 139–141, 2000.

[6] H. Y. Lee *et al.*, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust $HfO_2$ based RRAM," in *IEEE Int. Electron Devices Meeting, IEDM 2008*, Dec. 2008, pp. 1–4.

[7] K.-J. Lee *et al.*, "A 90nm 1.8V 512Mb diode-switch PRAM with 266MB/s read throughput," in *IEEE ISSCC Dig. Tech. Papers*, 2007, pp. 472, 616.

[8] F. Bedeschi *et al.*, "A multi-level-cell bipolar selected phase change memory," in *IEEE ISSCC Dig. Tech. Papers*, 2008, pp. 428, 625.

[9] S. Kang *et al.*, "A 0.1-$\mu$m 1.8-V 256-Mb phase-change random access memory (PRAM) with 66-MHz synchronous burst-read operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 210–218, Jan. 2007.

[10] K. Kim *et al.*, "Multilevel programmable oxide diode for cross-point memory by electrical-pulse-induced resistance change," *IEEE Electron Device Lett.*, vol. 30, no. 10, pp. 1036–1038, Oct. 2009.

[11] M. Johnson *et al.*, "512-Mb PROM with a three-dimensional array of diode/antifuse memory cells," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1920–1928, Nov. 2003.

[12] M. Rozenberg *et al.*, "Mechanism for bipolar resistive switching in transition-metal oxides," *Phys. Rev. B*, vol. 81, no. 11, 2010.

[13] T. Blalock and R. Jaeger, "A high-speed clamped bit-line current-mode sense amplifier," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 542–548, Apr. 1991.

[14] A. Mohsen *et al.*, "The design and performance of CMOS 256K bit DRAM devices," *IEEE J. Solid-State Circuits*, vol. 19, no. 5, pp. 610–618, Oct. 1984.

[15] "K9XXG08UXA: 512Mx8 bit/1Gx8 bit NAND flash memory," Samsung Electronics, Datasheet, 2006.

[16] "4-Gbit NAND built at 65 nm," *EE Times*, Jul. 2006.

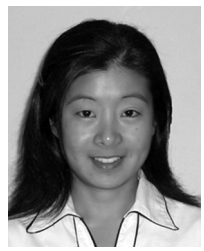
[17] D. Nobunaga *et al.*, "A 50nm 8Gb NAND flash memory with 100MB/s program throughput and 200MB/s DDR interface ," in *IEEE ISSCC Dig. Tech. Papers*, 2008, pp. 426, 625.

[18] D.-S. Byeon *et al.*, "An 8 Gb multi-level NAND flash memory with 63 nm STI CMOS process technology ," in *IEEE ISSCC Dig. Tech. Papers*, 2005, vol. 1, pp. 46–47.

[19] Ansoft Q3D Extractor. [Online]. Available: http://www.ansoft.com

[20] Numonyx Axcell M29EW, Numonyx, Datasheet, 2009.

[21] SanDisk 3-D OTP Memory, SanDisk, Datasheet, Document Number DS034, 2006.

**Elaine Ou** received the B.S. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, in 2003, the S.M. degree in computer science from Harvard University, Cambridge, MA, in 2005, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2010. Her doctoral research under Prof. S. Simon Wong investigated sense amplifiers and 3-dimensional cross-point array architectures for nonvolatile resistance-change memory.

She is currently a Visiting Lecturer in the Department of Electrical and Information Engineering at the University of Sydney, New South Wales, Australia. Her current research concentrates on low-latency hardware for financial network interfaces.

**S. Simon Wong** (M'83–SM'91–F'99) received the Bachelor degrees in electrical engineering and mechanical engineering from the University of Minnesota, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley.

His industrial experience includes semiconductor memory design at National Semiconductor (1978–1980) and semiconductor technology development at Hewlett Packard Labs (1980–1985). He was an Assistant Professor at Cornell University (1985–1988). Since 1988, he has been with Stanford University where he is now Professor of Electrical Engineering. His current research concentrates on understanding and overcoming the factors that limit performance in devices, interconnections, on-chip components and packages. He is on the board of Pericom Semiconductor and the advisory board of Atheros Communications.