A NOVEL ARCHITECTURE FOR ADVANCED HIGH DENSITY DYNAMIC RANDOM ACCESS MEMORIES

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Electrical Engineering

in the

College of Graduate Studies

University of Idaho

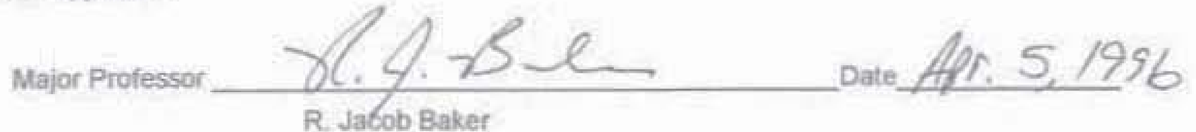
by
Brent Keeth

May 1996

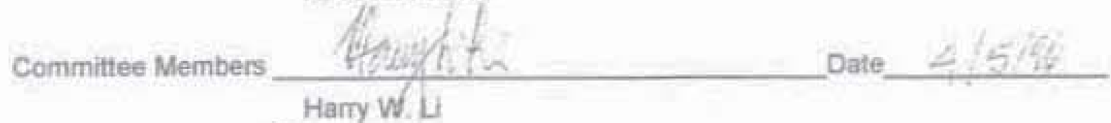
Major Professor: R. Jacob Baker, Ph.D.

AUTHORIZATION TO SUBMIT THESIS

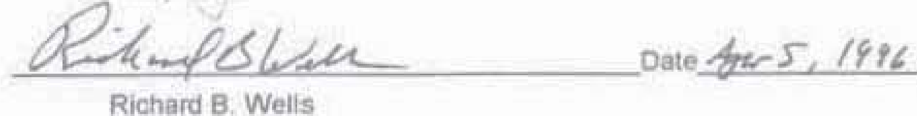
This thesis of Brent Keeth, submitted for the degree of Master of Science with a major in Electrical Engineering and titled "A NOVEL ARCHITECTURE FOR ADVANCED HIGH DENSITY DYNAMIC RANDOM ACCESS MEMORIES," has been reviewed in final form, as indicated by the signatures and dates given below. Permission is now granted to submit final copies to the College of Graduate Studies for approval.



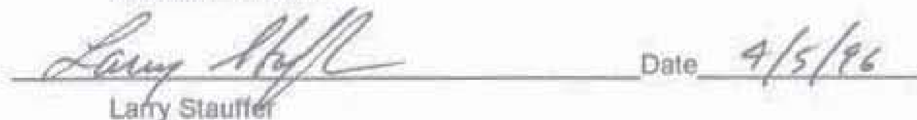
Major Professor ______ Date Apr. 5, 1996
R. Jacob Baker



Committee Members ______ Date 4/5/96
Harry W. Li



______ Date Apr 5, 1996
Richard B. Wells

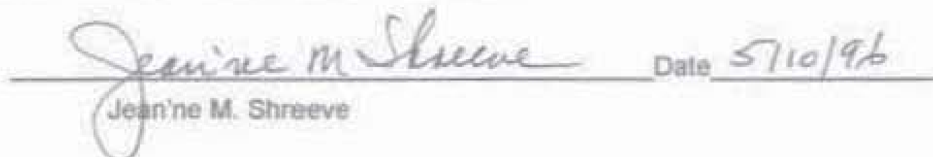


______ Date 4/5/96
Larry Stauffer

Department Administrator ______ Date 4-16-96
David Egolf

Discipline's College Dean ______ Date 4/16/96
Richard Jacobsen

Final Approval and Acceptance by the College of Graduate Studies



______ Date 5/10/96
Jean'ne M. Shreeve

ABSTRACT

Dynamic random access memory (DRAM) integrated circuits have been evolving for the past twenty-five years. From the early 1kilobit DRAMs, to the recently demonstrated 1 gigabit DRAMs, advances in semiconductor processing have fueled this evolution. Starting at the 64 kilobit generation, DRAM designs have consistently relied upon the folded digitline architecture. This architecture, based upon arrays utilizing eight square feature ($8F^2$) memory cells, achieves high signal-to-noise performance at the expense of larger die size. Alternative architectures, with smaller memory cells, are either too expensive, too complex, or otherwise fail to meet required signal-to-noise targets. This thesis describes a novel bi-level DRAM architecture developed to achieve significant reductions in die size while maintaining the signal-to-noise performance of the folded digitline architecture. The bi-level digitline architecture achieves die size reduction by building memory arrays with smaller six square feature ($6F^2$) memory cells in a form of open digitline layout. The memory arrays utilize stacked digitlines and vertical digitline twisting to otherwise achieve folded architecture operation and noise performance. This thesis is divided into five sections. Section one consisting of a review of basic DRAM construction and operation. Section two contains a detailed examination of essential DRAM building blocks used in array construction. Section three provides a detailed description of the traditional open digitline and folded digitline architectures. Section four introduces and examines the bilevel digitline architecture. All three architectures are employed in the theoretical construction of 32Mbit memory blocks, thus providing a valuable data point for die size comparison. Finally, the results of this architecture size comparison are presented and analyzed.

TABLE OF CONTENTS

AUTHORIZATION TO SUBMIT THESIS ii

ABSTRACT iii

TABLE OF CONTENTS iv

TABLE OF FIGURES v

TABLE OF TABLES vii

I. DRAM BASICS 1

Array Construction 1

Array Operation 3

II. DRAM ARRAY ELEMENTS 8

Memory Cells and Arrays 8

Sense Amplifier Elements 18

Row Decoder Elements 26

III. ARCHITECTURAL CHARACTERISTICS 34

Open Digitline Array Architecture 34

Folded Array Architecture 42

IV. ADVANCED BILEVEL DRAM ARCHITECTURE 48

Array Architecture Objectives 48

Bilevel Digitline Construction 49

Bilevel Digitline Array Architecture 53

Architectural Comparison 59

V. CONCLUSION 60

BIBLIOGRAPHY 61

TABLE OF FIGURES

Figure 1.1: 1T1C DRAM memory cell 1
Figure 1.2: Folded memory array schematic 2
Figure 1.3: Folded memory array layout 2
Figure 1.4: Simple folded array schematic 3
Figure 1.5: Cell access waveforms 4
Figure 1.6: Sense amplifier schematic 4
Figure 1.7: Sensing operation waveforms 5
Figure 1.8: Sense amplifier schematic with I/O devices 6
Figure 1.9: Write operation waveforms 7
Figure 2.1: Mbit pair layout 8
Figure 2.2: Layout to show array pitch 9
Figure 2.3: Layout to show $8F^2$ derivation 10
Figure 2.4: Folded digitline array schematic 11
Figure 2.5: Digitline twist schemes 12
Figure 2.6: Open digitline array schematic 13
Figure 2.7: Open digitline array layout 14
Figure 2.8: Buried capacitor cell process cross section 15
Figure 2.9a: Buried digitline mbit cell layout 16
Figure 2.9b: Buried digitline mbit process cross section 16
Figure 2.10: Trench capacitor mbit process cross section 17
Figure 2.11: Equilibration and bias circuit schematic 18
Figure 2.12: Equilibration and bias circuit layout 19
Figure 2.13: Schematic showing I/O transistors 21
Figure 2.14: Basic sense amplifier schematic 21
Figure 2.15: Standard sense amplifier block 23
Figure 2.16: Complex sense amplifier block 24
Figure 2.17: Minimized sense amplifier block 24
Figure 2.18: Single metal sense amplifier block 25
Figure 2.19: Waveforms for a read-modify-write DRAM cycle 26
Figure 2.20: Bootstrap wordline driver schematic 27
Figure 2.21: Bootstrap wordline driver operation 28
Figure 2.22: NOR wordline driver 29
Figure 2.23: CMOS wordline driver 30
Figure 2.24: Static decode tree 31

Figure 2.25: P&E decode tree.......31
Figure 2.26: Pass transistor decode tree schematic.......33
Figure 3.1: Open digitline architecture schematic.......35
Figure 3.2: Open digitline 32Mbit array block.......37
Figure 3.3: Single pitch open digitline architecture.......39
Figure 3.4: Open digitline architecture with dummy arrays.......40
Figure 3.5: Folded digitline array architecture schematic.......42
Figure 3.6: Folded digitline architecture 32Mbit array block.......45
Figure 4.1: Development of bilevel digitline architecture.......49
Figure 4.2: Digitline vertical twisting concept.......50
Figure 4.3: Bilevel digitline architecture schematic.......51
Figure 4.4: Vertical twisting schemes.......51
Figure 4.5: Plaid $6F^2$ mbit array.......52
Figure 4.6: Bilevel digitline array schematic.......54
Figure 4.7: Bilevel digitline architecture 32Mbit array block.......56

TABLE OF TABLES

Table 2.1: Predecoded address truth table 32
Table 3.1: 0.25μm design parameters 34
Table 3.2: Active current and power versus digitline length 36
Table 3.3: Wordline time constant versus wordline length 36
Table 3.4: Open digitline (local row decode)--32Mbit size calculations 38
Table 3.5: Open digitline (dummy arrays & global row decode)--32Mbit size calculations 41
Table 3.6: Open digitline (dummy arrays & hier row decode)--32Mbit size calculations 41
Table 3.7: Wordline time constant versus wordline length (folded) 44
Table 3.8: Folded digitline (local row decode)--32Mbit size calculations 46
Table 3.9: Folded digitline (global decode)--32Mbit size calculations 47
Table 3.10: Folded digitline (hier row decode)--32Mbit size calculations 47
Table 4.1: Active current and power versus bilevel digitline length 55
Table 4.2: Bilevel digitline (local row decode)--32Mbit size calculations 57
Table 4.3: Bilevel Digitline (Global Decode)--32Mbit Size Calculations 58
Table 4.4: Bilevel Digitline (Hier Row Decode)--32Mbit Size Calculations 58
Table 4.5: 32Mbit size calculations summary 59

I. DRAM Basics

This section contains a brief examination of basic dynamic random access memory (DRAM) array architecture and construction. The examination focuses upon developing a basic understanding of memory cell and array concepts. After developing this understanding, the analysis continues into basic array operations that include cell access, sensing, and refresh.

Array Construction

A modern DRAM memory cell or memory bit (mbit), as shown in figure 1.1, consists of one metal oxide semiconductor (MOS) transistor and one storage capacitor--accordingly referred to as a 1T1C cell. The mbit transistor operates as a switch, interposed between the mbit capacitor and the digitline. The mbit is capable of holding a single piece of binary information, as stored electric charge

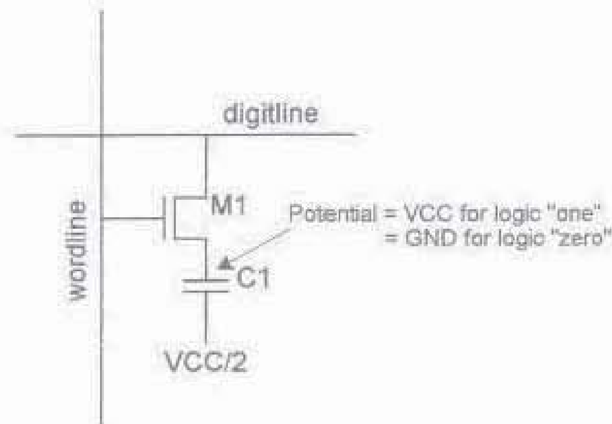


Figure 1.1: 1T1C DRAM memory cell

in the cell capacitor. Given a bias voltage of VCC/2 on the capacitor's common node [1], a logic one level is represented by +VCC/2 volts across the capacitor and a logic zero is represented by -VCC/2 volts across the capacitor. In either case, the amount of charge stored in the mbit capacitor is $Q = C \cdot VCC / 2$ coulombs, where C is the capacitance value in Farads.

The digitline, as depicted in figure 1.1, consists of a conductive line connected to a multitude of mbit transistors. Generally, either metal or silicided/polycided polysilicon forms the conductive line [2]. Due to the large quantity of attached mbits, its physical length, and proximity to other features, the digitline is very capacitive. For instance, a typical value for digitline capacitance on a 0.35um process might be around 300fF. Digitline capacitance is an important parameter since it dictates many other aspects of the design.

The mbit transistor's gate terminal connects to a wordline (rowline). The wordline, which connects to a multitude of mbits, consists of an extended segment of the same polysilicon used to form the transistor's gate. The wordline is physically orthogonal to the digitline. A memory array, shown in figure 1.2, is created by tiling a selected quantity of mbits together such that mbits along a given digitline do not share a common wordline and such that mbits along a common wordline do not share

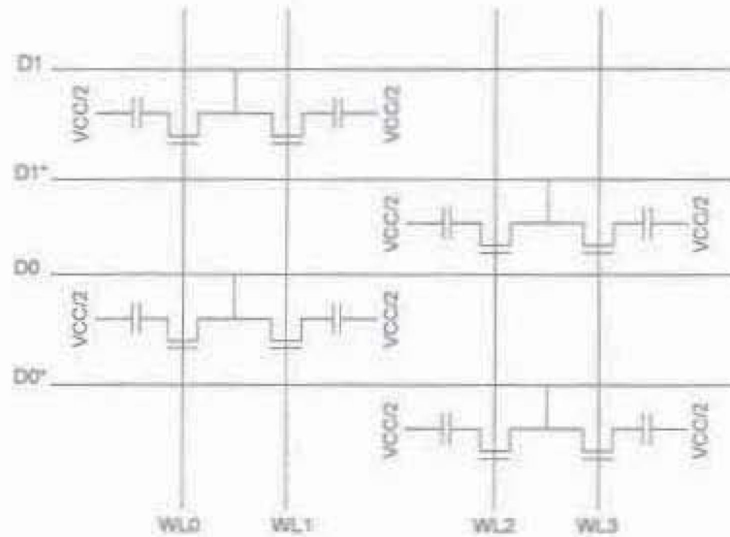


Figure 1.2: Folded memory array schematic

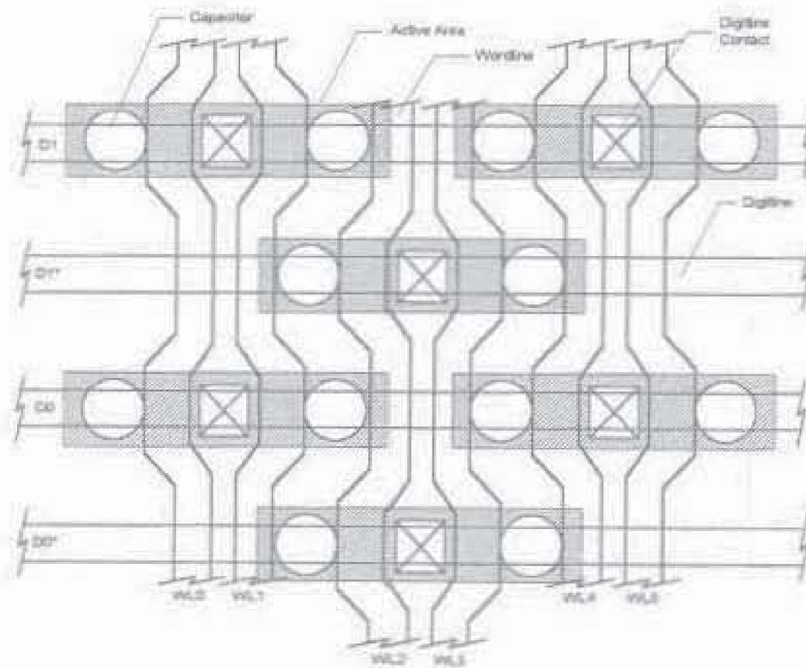


Figure 1.3: Folded memory array layout

a common digitline. Figure 1.3 contains an example of a memory array formed by tiling mbits. There are several features of this layout that need illumination. First, note that the mbits are in pairs to permit the sharing of a common contact to the digitline. This feature reduces the array size by eliminating unnecessary duplication. Second, note that any given wordline only forms (crosses) an mbit transistor on alternating digitlines. This feature allows the formation of digitline pairs and ensures that wordline activation enables transistors only on alternate digitlines. Digitline pairs are an inherent feature in folded digitline arrays, as depicted in figure 1.3. Further discussion of folded digitline architectures and their characteristics will be easier, though, after first introducing an alternative array structure called the open digitline architecture. A thorough understanding of both folded and open architectures is necessary to appreciate the characteristics and benefits of the bi-level digitline architecture--the subject of this thesis.

Array Operation

An understanding of basic DRAM operation, such as reading and writing, is necessary before the discussion of array architecture can continue. Assume that the capacitors in figure 1.4 have logic one levels (+VCC/2) stored on them. Consider each digitline pair or column as consisting of two adjacent digitlines. The digitlines, labeled D0 and D0* are initially equilibrated at VCC/2 volts [3]. All wordlines

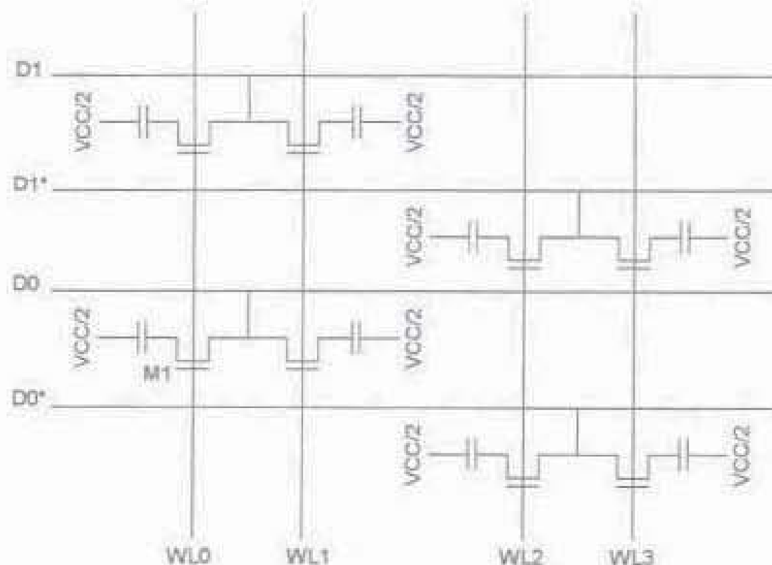


Figure 1.4: Simple folded array schematic

are initially at zero volts, which turns off the mbit transistors. To read mbit1, wordline WL0 transitions to a voltage that is at least one transistor V_{th} above VCC. This elevated wordline voltage level is referred to as VCCP or VPP. When the wordline voltage exceeds one V_{th} above the digitline voltage (VCC/2 in this example) and the mbit transistor turns on, the mbit capacitor will begin to discharge

onto the digitline. Essentially, reading or accessing a DRAM cell results in charge sharing between the mbit capacitor and the digitline capacitance. This sharing of charge causes the digitline voltage to either increase for a stored logic one or decrease for a stored logic zero. Ideally, the access will only modify the active digitline, leaving its complement digitline unaltered. In reality, the other digitline voltage will also change slightly, due to parasitic coupling between digitlines and overlap

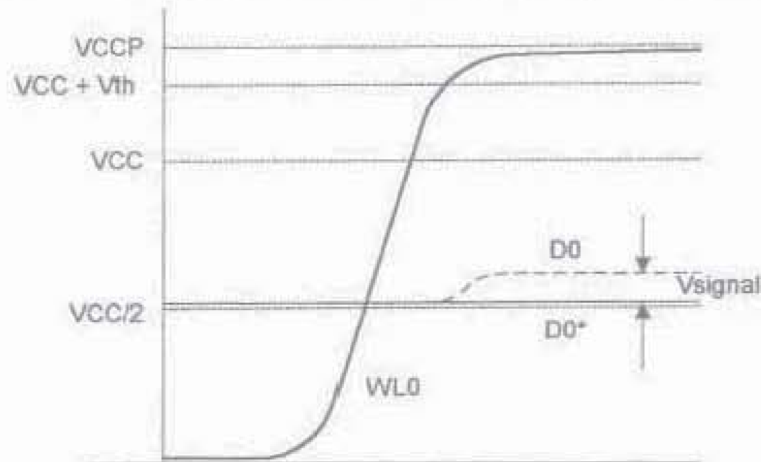


Figure 1.5: Cell access waveforms

capacitance between the firing wordline and the digitline. Regardless, a differential voltage develops between the two digitlines. The magnitude of this signal voltage is a function of the mbit capacitance (Cmbit), digitline capacitance (Cdigit), the mbit's stored voltage prior to the access (Vcell), and any noise terms. Accordingly, $V_{signal} = [(Vcell \cdot Cmbit) \div (Cdigit + Cmbit)] - V_{noise}$ volts. For a design in which Vcell=1.65V, Cmbit=40fF, Cdigit=300fF, and V_{noise}=0, this equation yields a V_{signal} of 194mV. Figure 1.5 contains typical waveforms for the cell access operation just described.

After the cell access is complete, the sensing operation can commence. The reason for forming a digitline pair will now become apparent. Figure 1.6 contains a schematic diagram for a simplified

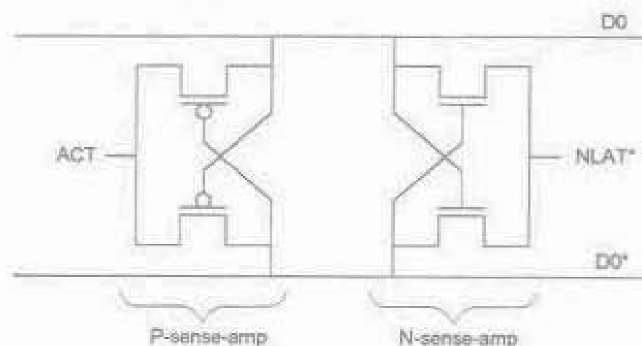


Figure 1.6: Sense amplifier schematic

sense amplifier circuit. Note that it consists of a cross-coupled PMOS pair and a cross-coupled PMOS pair. The NMOS pair or N-sense-amp common node is labeled NLAT* (for N-sense-amp LATch) in figure 1.6. Similarly, the P-sense-amp common node is labeled ACT (for ACTive pull-up). Initially, NLAT* is biased to VCC/2 and ACT is biased to VSS or ground. Since the digitline pair D0 and D0* are both initially at VCC/2 volts, the N-sense-amp transistors remain off due to zero Vgs potential. Similarly, both P-sense-amp transistors remain off due to their negative Vgs potential. As discussed in the preceding paragraph, a signal voltage develops between the digitline pair when the mbit access occurs. While one digitline contains charge from the cell access, the other digitline serves as a reference for the sensing operation. The sense amplifier firing generally occurs sequentially rather than concurrently [1], [4], [5]. The N-sense-amp fires first and the P-sense-amp second. The N-sense-amp is generally a better amplifier than the P-sense-amp because of the higher drive of NMOS transistors and better V_{th} matching. This provides for better sensing characteristics and lower probability of errors. Figure 1.7 contains waveforms for the sensing

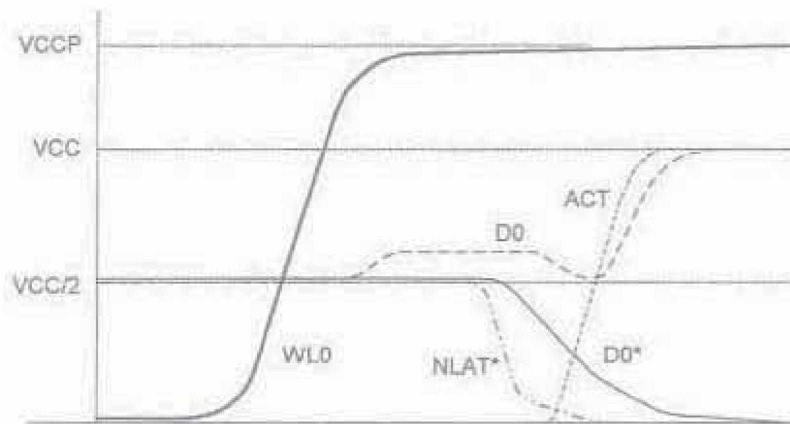


Figure 1.7: Sensing operation waveforms

operation. Dropping the NLAT* signal toward ground will fire the N-sense-amp. As the voltage between NLAT* and the digitlines approaches V_{th}, the NMOS transistor, whose gate connection is to the higher voltage digitline, will begin to conduct. Transistor conduction occurs first in the subthreshold region, progressing to the saturation region as the gate to source voltage exceeds V_{th}. Conduction results in the discharge of the low voltage digitline toward the NLAT* voltage. Ultimately, NLAT* will reach ground, bringing the digitline with it. Note that the other NMOS transistor will not conduct since its gate voltage derives from the low voltage digitline, which is discharging toward ground. In reality, parasitic coupling between the digitlines and limited subthreshold conduction by the second transistor will reduce the high digitline voltage.

Shortly after the N-sense-amp fires, ACT will be driven towards VCC. This activates the P-sense-amp that operates in a complementary fashion to the N-sense-amp. With the low voltage digitline

approaching ground, a strong signal exists to drive the appropriate PMOS transistor into conduction. This conduction, again moving from subthreshold to saturation operation, will charge the high voltage digitline toward ACT, ultimately reaching VCC. Since the mbit transistor remains on during sensing, the mbit capacitor will charge to the NLAT* or ACT voltage level. The voltage, and hence charge, which the mbit capacitor held prior to accessing will restore a full level--VCC for a logic one and GND for a logic zero. It should be apparent now, why the minimum wordline voltage is V_{th} above VCC. If VCCP were anything less, the mbit transistor would turn off before the mbit capacitor attains a full VCC level.

A DRAM write operation is very similar to sensing and restore operations except that separate write driver circuits determine whether logic ones or zeros are placed into the cells. The write driver circuit is generally a tri-state inverter connected to the digitlines through a second pair of pass transistors as shown in figure 1.8. These pass transistors are referred to as I/O transistors. The gate terminals of the I/O transistors connect to a common CSEL (Column SELect) signal. The column address determines which CSEL signal activates and utlimately which pair (or multiple pairs) of digitlines route to the output pad or write driver. In most current DRAM designs, the write driver simply overdrives the sense amplifier pair, which remain active during the writing operation. The write

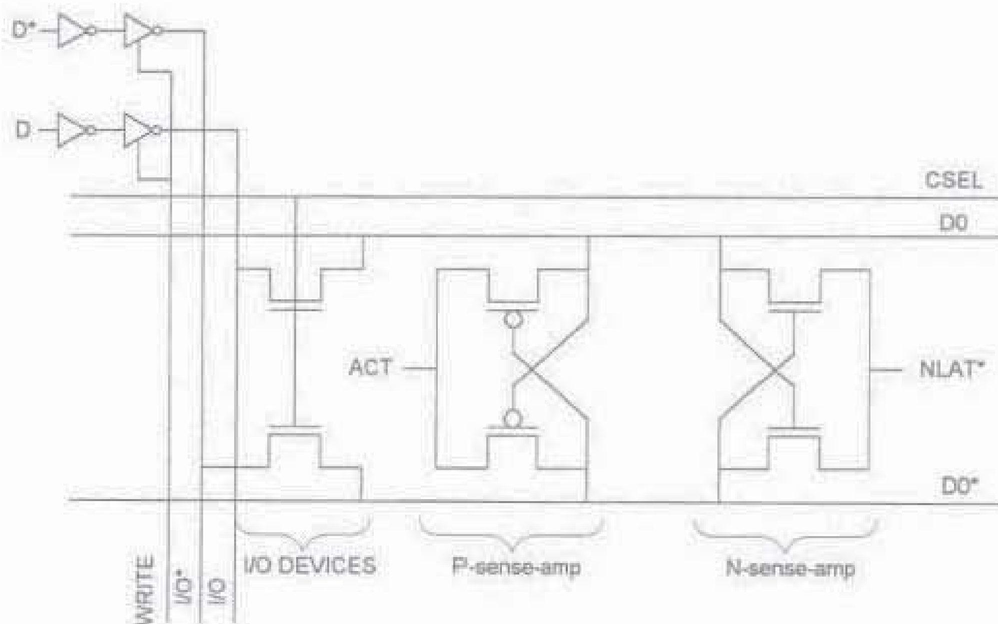


Figure 1.8: Sense amplifier schematic with I/O devices

operation needs to be long enough in duration to flip the sense amplifiers. After new data writes into the sense amplifiers, the amplifiers actually finish the write cycle by restoring the digitlines to full rail to rail voltages. Figure 1.9 contains an example of this, in which D0 is initially high after the sensing

operation, and low after the write operation. Each read cycle refreshes all of the mbit capacitors connected to the active wordline. A write operation, though, involves only two to four mbits within an array of mbits. This is because a single CSEL line generally connects to only four I/O transistor pairs. The remaining digitlines are accessible through additional CSEL lines corresponding to different column address locations.

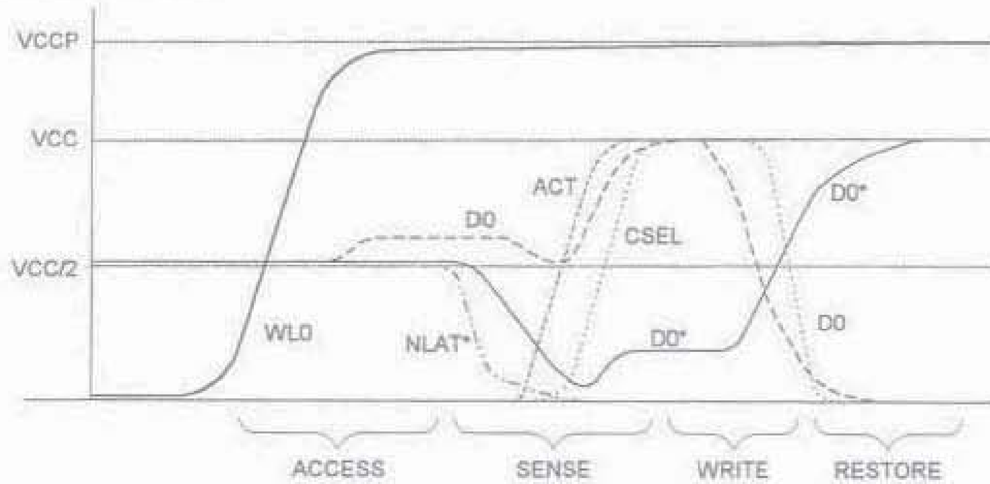


Figure 1.9: Write operation waveforms

II. DRAM Array Elements

Section-II begins a more detailed examination of standard DRAM array elements. This section is necessary to develop a clear understanding of fundamental DRAM elements and how they are used in memory block construction. A common point of reference is required before the analysis of competing array architectures can be considered. Included in this section is a detailed discussion of mbits, array configurations, sense amplifier elements and row decoder elements.

Memory Cells and Arrays

The primary advantage of DRAM, over other types of memory technology, is low cost. This advantage arises from the simplicity and scaling characteristics of its 1T1C memory cell [6]. Although the DRAM mbit encompasses simple concepts, its actual design and implementation are highly complex. Successful, cost effective DRAM designs require a tremendous amount of process technology.

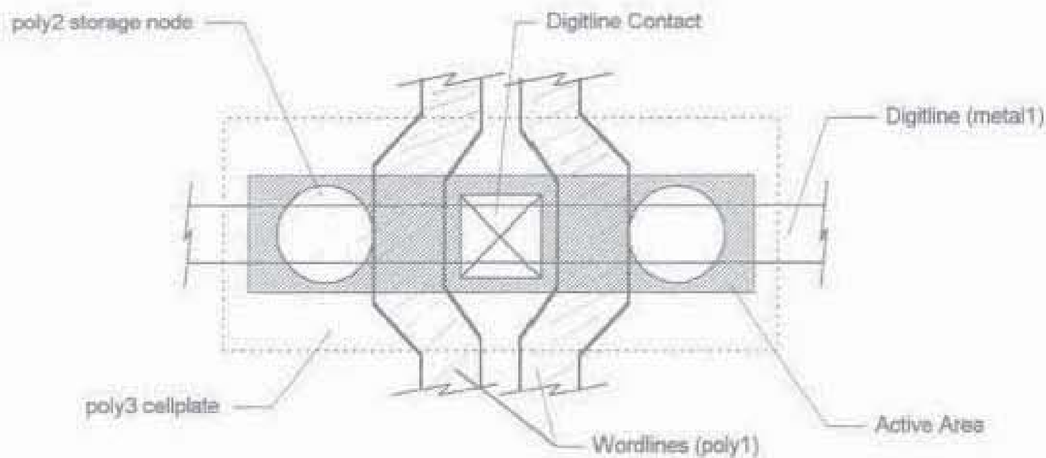


Figure 2.1: Mbit pair layout

A modern buried capacitor DRAM mbit pair appears in figure 2.1. DRAM mbits are constructed in pairs, to allow sharing of the digitline contact. Sharing a contact significantly reduces overall cell size. The mbits consist of an active area rectangle (in this case N+ active area), a pair of polysilicon wordlines, a single digitline contact, a metal or polysilicon digitline, and a pair of cell capacitors formed with oxide-nitride-oxide dielectric between two layers of polysilicon. For some processes, the wordline polysilicon is silicided to reduce the sheet resistance, permitting longer wordline segments without reducing speed. The mbit layout as shown in figure 2.1, is essentially under the control of process engineers, since every aspect of the mbit must meet stringent performance criteria.

A small array of mbits appears in figure 2.2. This figure is useful to illustrate several features of the mbit. First, note that the digitline pitch (width plus space) dictates the active area pitch and capacitor pitch. Process engineers adjust the active area width and the field oxide width to maximize transistor drive and minimize transistor to transistor leakage. The field oxide technology greatly impacts this balance. A thicker field oxide or a shallower junction depth will enable wider transistor active area. Second, the wordline pitch (width plus space) dictates the space available for the digitline contact, transistor length, active area space, field poly width, and capacitor length. Optimization of each of these features by process engineers is necessary to maximize capacitance, minimize leakage and maximize yield. Contact technology, subthreshold transistor characteristics, photolithography, etch and film technology will dictate the overall design.

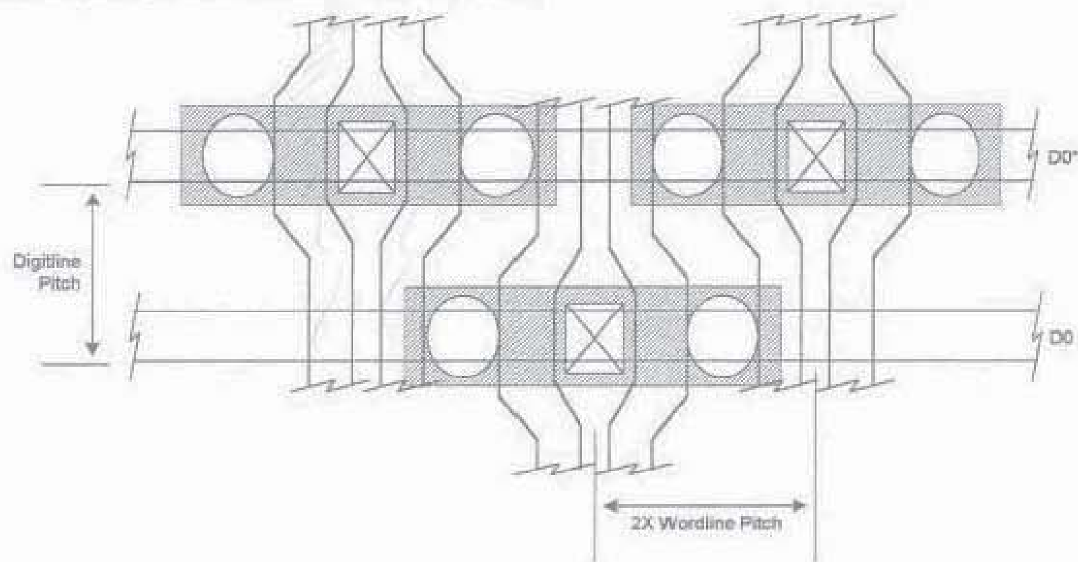


Figure 2.2: Layout to show array pitch

At this point in the discussion, it is appropriate to introduce the concept of feature size and how it relates to cell size. The mbit shown in figures 2.1 and 2.2 is by definition an eight square feature ($8F^2$) cell [7], [8]. The intended definition of feature in this case is minimum realizable process dimension, but in actual fact equates to a dimension that is half of the wordline (row) or digitline (column) pitch. A 0.25um process having wordline and digitline pitches of 0.6um yields an mbit size that is $8 \cdot (0.3um)^2 = 0.72um^2$. Explanation of the $8F^2$ designation is easier with the aid of figure 2.3. An imaginary box drawn around the mbit defines the cell's outer boundary. Along the x-axis, this box will include 1/2 digitline contact feature, 1 wordline feature, 1 capacitor feature, 1 field poly feature, and 1/2 poly space feature, which totals to 4 features. Along the y-axis, this box contains two 1/2 field oxide features and 1 active area feature, which totals to 2 features. The area of the mbit is therefore $4F \cdot 2F = 8F^2$. The folded array architecture, as shown in figure 2.2, always produces an

$8F^2$ mbit. This results from the fact that each wordline connects (forms a crosspoint) with an mbit transistor on every other digitline and must pass around mbit transistors as field poly on the remaining digitlines. The field poly in each mbit cell adds two square features to what would have been a $6F^2$ cell otherwise. Although the folded array yields a cell that is 25% larger than other array architectures, it also produces superior signal-to-noise performance, especially when combined with some form of digitline twisting [9]. Superior low noise performance made folded array architecture the architecture of choice since the 64kbit generation [1].

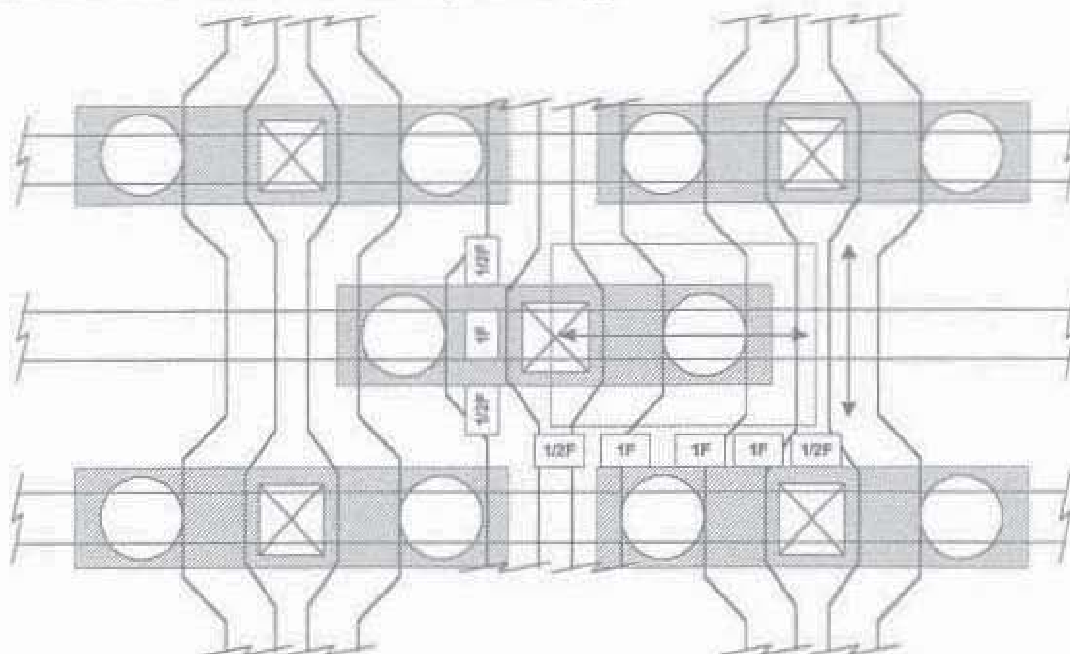


Figure 2.3: Layout to show $8F^2$ derivation

A folded array is schematically depicted in figure 2.4. Sense amplifier circuits placed at the edge of each array connect to both true and complement digitlines (D and D*) coming from a single array. Optional digitline pair twisting at one or more places can reduce and balance the coupling to adjacent digitline pairs and improve overall signal-to-noise characteristics [9]. Figure 2.5 serves as evidence of the variety of twisting schemes used throughout the DRAM industry [10]. Ideally, a twist scheme will equalize the coupling terms from each digitline to all other digitlines, both true and complement. If done properly, the noise terms will cancel or only produce common mode noise to which the differential sense amplifier is immune. Each digitline twist region consumes valuable silicon area. As a result, design engineers resort to the simplest and most efficient twisting scheme to get the job done. Since the coupling between adjacent metal lines is inversely proportional to their spacing, the signal-to-noise problem gets increasingly worse as DRAMs scale to smaller and smaller dimensions.

Evidence of this problem, is the industry trend to use more complex twisting schemes on succeeding generations [10], [11].

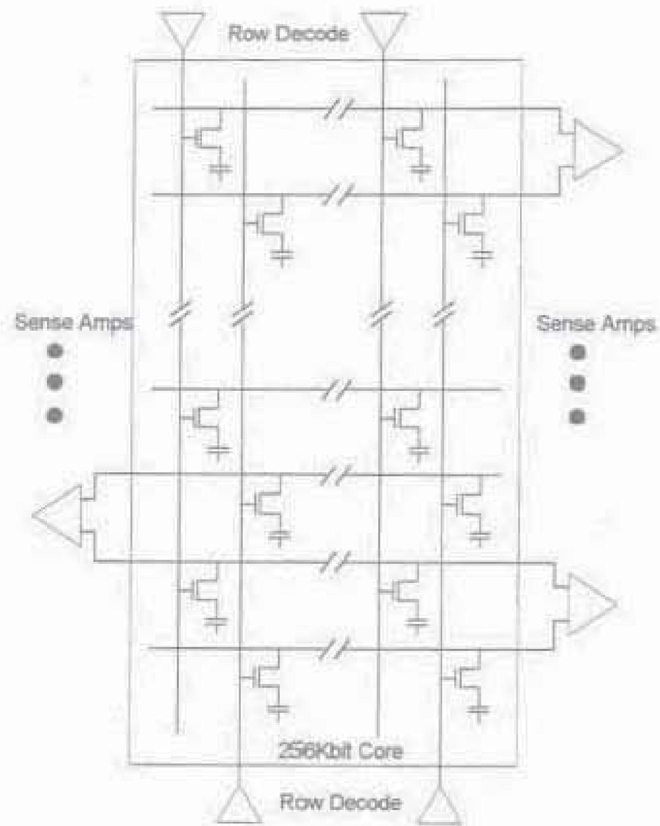


Figure 2.4: Folded digitline array schematic

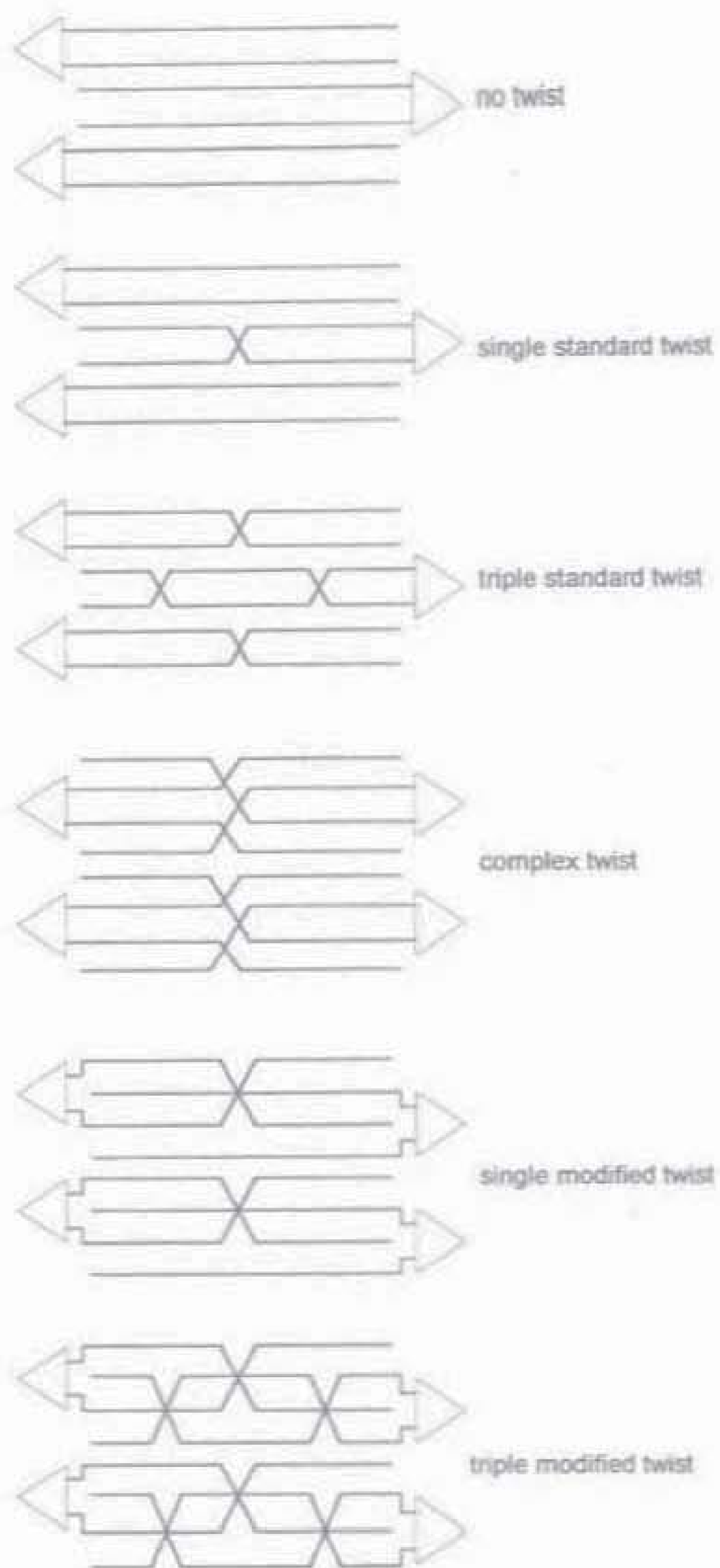


Figure 2.5: Digitline twist schemes

An alternative to the folded array architecture, popular prior to the 64kbit generation [6], was the open digitline architecture. Seen schematically in figure 2.6, this architecture also features the sense amplifier circuits between two sets of arrays [12]. Unlike the folded array, though, true and complement digitlines (D and D*) connected to each sense amplifier pair come from separate arrays [13]. This precludes the use of digitline twisting to improve signal-to-noise performance and identifies the prevalent reason why the industry switched to the folded array architecture. Also note that unlike

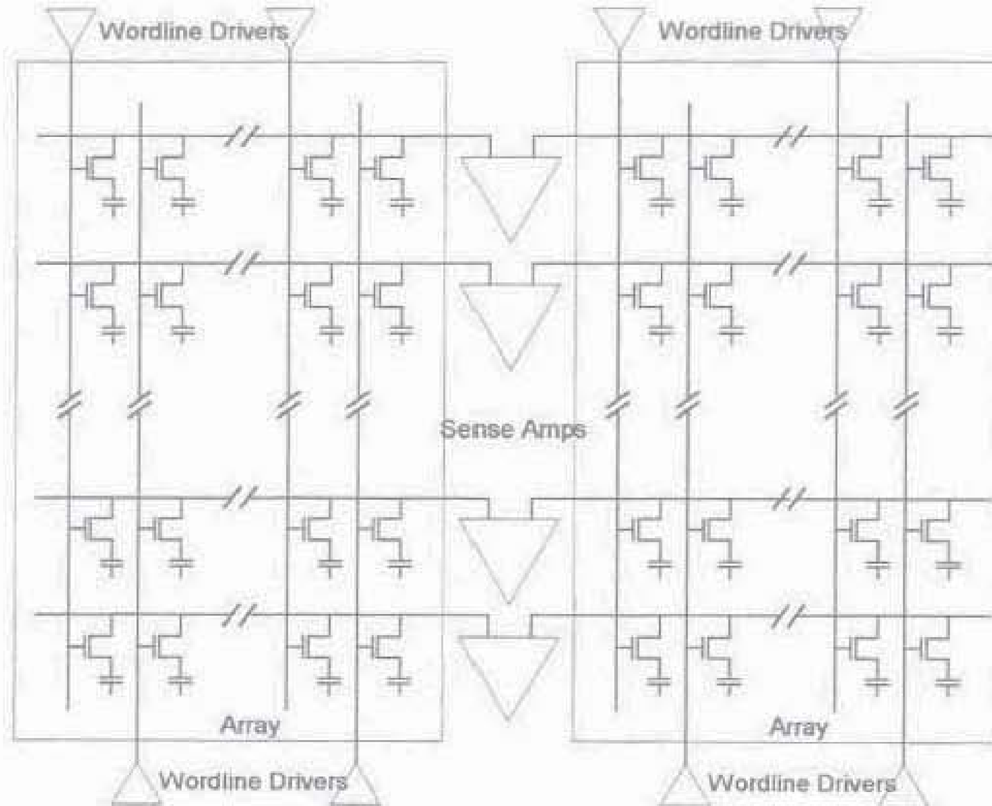


Figure 2.6: Open digitline array schematic

the folded array architecture each wordline in an open digitline architecture connects to mbit transistors on every digitline--crosspoint style arrays. This feature permits a 25% reduction in mbit size to only $6F^2$ since the wordlines do not have to pass alternate mbits as field poly. The layout for an array of standard $6F^2$ mbit pairs is shown in figure 2.7 [7]. A box is drawn around one of the mbits to shown the $6F^2$ cell boundary. Again, two mbits share a common digitline contact to improve layout efficiency. Unfortunately, most manufacturers have found that the signal-to-noise problems of open digitline architecture outweigh the benefits derived from reduced array size [12].

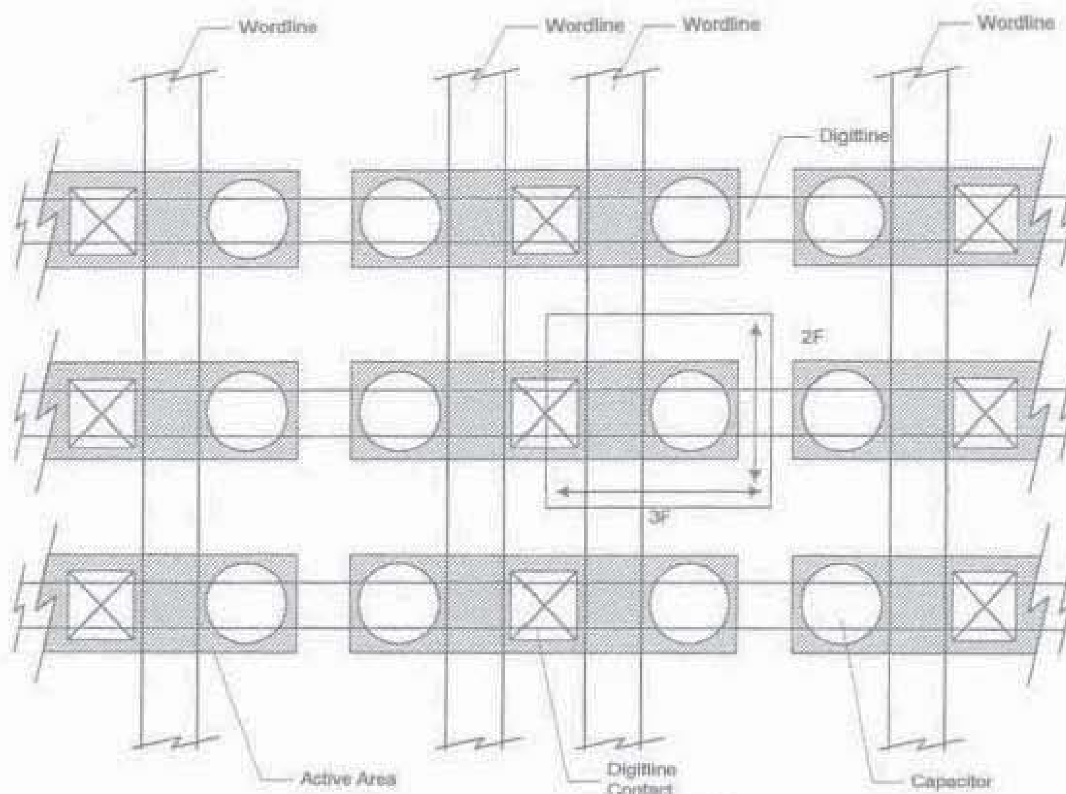


Figure 2.7: Open digitline array layout

Digitline capacitive components, contributed by each mbit, include junction capacitance, digitline to cellplate (poly3), digitline to wordline, digitline to digitline, digitline to substrate, and in some cases digitline to storage cell (poly2) capacitance. Each mbit connected to the digitline therefore adds a specific amount of capacitance to the digitline. Most modern DRAM designs have no more than 256 mbits connected to a digitline segment. Two factors dictate this quantity. First for a given cell size, as determined by row and column pitches, there is a maximum achievable storage capacitance without resorting to exotic processes or excessive cell height. For processes in which the digitline is above the storage capacitor (buried capacitor), contact technology will determine the maximum allowable cell height. This fixes the volume available (cell area multiplied by cell height) in which to build the storage capacitor. Second, as the digitline capacitance increases, the power associated with charging and discharging this capacitance during reading and writing operations increases. Any given wordline essentially accesses (crosses) all of the columns within a DRAM. For a 256Meg DRAM, each wordline crosses 16,384 columns. With a multiplier such as that, it is easy to appreciate why limits to digitline capacitance are necessary to keep power dissipation low.

Figure 2.8 contains a process cross section for the buried capacitor mbit depicted in figure 2.1. This type of mbit, employing a buried capacitor structure, places the digitline physically above the storage capacitor [14]. The digitline is constructed from either metal or polycide while the digitline contact is formed using metal or polysilicon plug technology. The mbit capacitor is formed with polysilicon (poly2) as the bottom plate, an oxide-nitride-oxide (ONO) dielectric, and a sheet of polysilicon (poly3) which forms the common node shared by all mbit capacitors. The capacitor shape can be simple, such as a rectangle, or complex, such as concentric cylinders or stacked discs. Exotic capacitor structures are the topic of many DRAM process papers [15], [16], [17]. The ONO dielectric undergoes optimization to achieve maximum capacitance with minimum leakage. It must also tolerate the maximum DRAM operating voltage without breakdown. For this reason, the cellplate (poly3) is normally biased at +VCC/2 volts. This will ensure that the dielectric will have no more than VCC/2 volts across it for either stored logic state, a logic one at +VCC/2 volts or a logic zero at -VCC/2 volts.

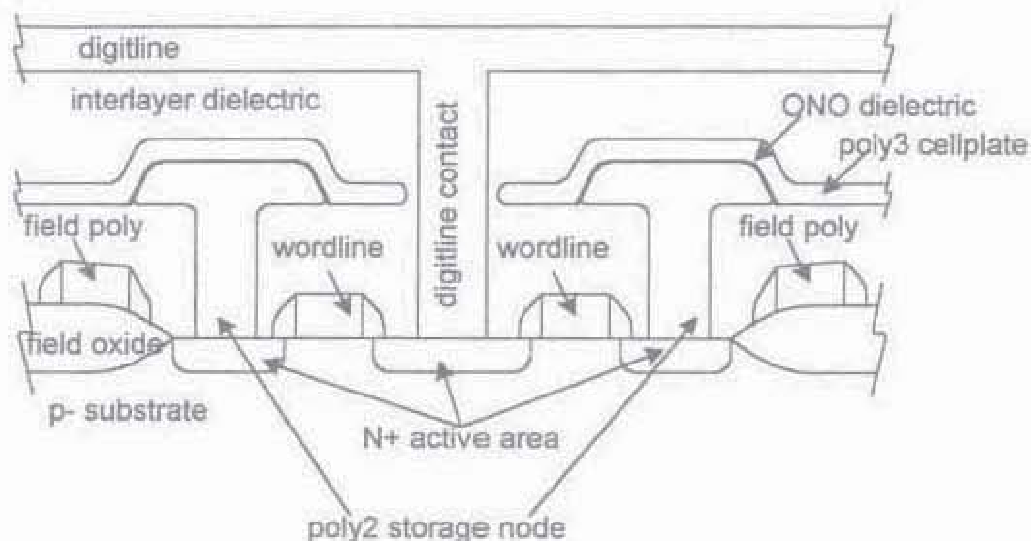


Figure 2.8: Buried capacitor cell process cross section

There are two other basic mbit configurations used in the DRAM industry. The first, shown in figure 2.9, is referred to as a buried digitline or capacitor over bitline cell [18], [19]. The digitline in this cell is almost always made from polysilicon rather than metal. As viewed from the top, the active area is normally bent or angled to accommodate the storage capacitor contact that must drop between digitlines. An advantage that the buried digitline cell has over the buried capacitor cell of figure 2.8 is that the digitline is physically very close to the silicon surface making digitline contacts much easier to produce. The angled active area, though, reduces the effective active area pitch, constraining the isolation process even further. Buried digitline cells also make formation of the

capacitor contact extremely difficult. Since the digitline is at or near minimum pitch for the process, insertion of a contact between digitlines is far from trivial.

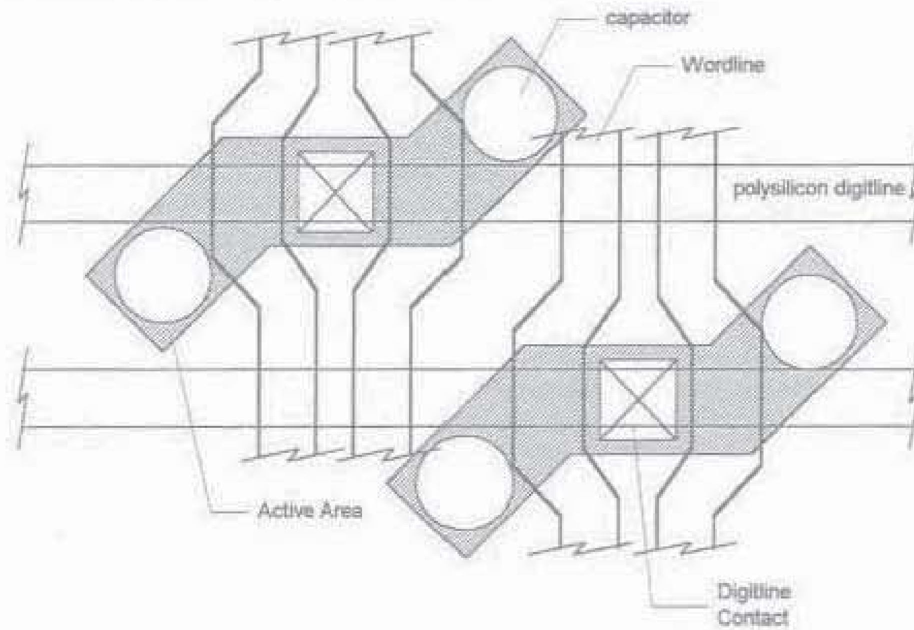


Figure 2.9a: Buried digitline mbit cell layout

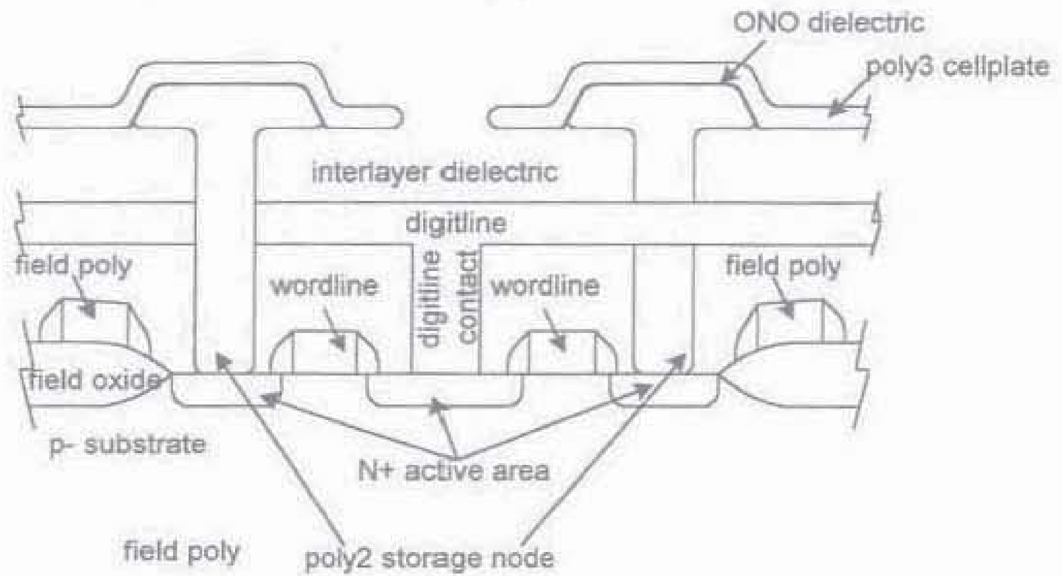


Figure 2.9b: Buried digitline mbit process cross section

Figure 2.10 contains a process cross-section of the third type of mbit used in the construction of DRAMs. This cell utilizes trench storage capacitors and is accordingly called a trench cell [16], [17]. Trench capacitors are formed in the silicon substrate, rather than above the substrate, after etching deep holes into the wafer. The storage node consists of a doped polysilicon plug deposited in the hole following growth or deposition of the capacitor dielectric. Contact between the storage node plug and the transistor drain is usually made through a poly strap. With most trench capacitor designs, substrate serves as the common node connection to the capacitors, preventing the use of +VCC/2 bias and thinner dielectric. The substrate is heavily doped around the capacitor to reduce the resistance and improve the capacitor's CV characteristics. A real advantage to trench cell is that the capacitance can be increased by merely etching a deeper hole into the substrate [20]. Furthermore, the capacitor does not add stack height to the design, which greatly simplifies contact technology. The disadvantage to trench capacitor technology resides in the difficulty associated with reliably building capacitors in deep silicon holes and in connecting the trench capacitor to the transistor drain terminal.

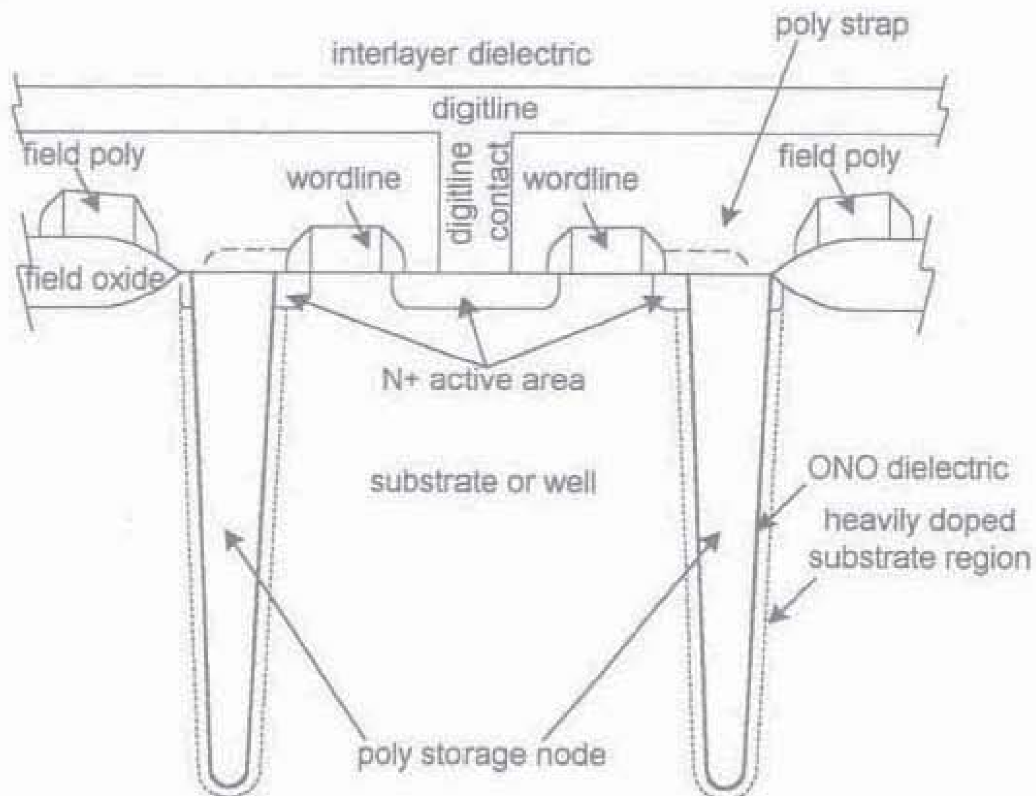


Figure 2.10: Trench capacitor mbit process cross section

Sense Amplifier Elements

The term sense amplifier actually refers to a collection of circuit elements that pitch up to the digitlines of a DRAM array. This collection most generally includes isolation transistors, devices for digitline equilibration and bias, one or more N-sense amplifiers, one or more P-sense amplifiers, and devices to connect selected digitlines to I/O signal lines. All of these circuits along with the wordline driver circuits, to be discussed later in this section, are called pitch cells. This designation comes from the requirement that the physical layout for these circuits is constrained by the digitline and wordline pitches of an array of mbits. For example, the sense amplifier layout for a specific digitline pair (column) generally consumes the space of four digitlines. This is commonly referred to as quarter-pitch or four-pitch, such that one sense amplifier exists for every four digitlines.

The first elements for review are the equilibration and bias circuits. From the earlier discussions on DRAM operation in section-I, the digitlines start at Vcc/2 volts prior to cell access and sensing [3]. It is vitally important to the sensing operation that both digitlines, which form a column pair, are at the same voltage before firing a wordline. Any offset voltage that appears between the pair, will directly

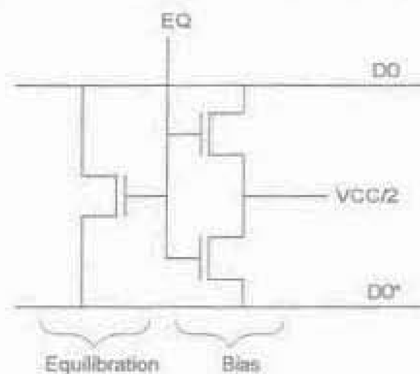


Figure 2.11: Equilibration and bias circuit schematic

reduce the effective signal voltage produced by the access operation [1]. Digitline equilibration is accomplished with one or more NMOS transistors connected between the digitlines. The higher drive strength of an NMOS device produces faster equilibration than a PMOS transistor of comparable size. An equilibration transistor, together with bias transistors, appears schematically in figure 2.11. The gate terminal is connected to a signal called EQ (EQuilibrate). EQ is held at Vcc whenever the external row address strobe (RAS*) is high, indicating an inactive or precharge state for the DRAM. When RAS* falls, EQ will transition low, turning off the equilibration transistor just prior to any wordline firing. Toward the end of each RAS cycle, EQ will again transition high and force the digitlines to re-equilibrate.

As shown in figure 2.11, Vcc/2 bias for the digitlines is provided by two additional NMOS transistors. The devices operate in conjunction with equilibration to ensure that the digitline pair remains at the prescribed voltage for sensing. Normally, a pair of digitlines that are at Vcc and ground respectively, will equilibrate to Vcc/2 volts [1]. The bias devices ensure that this occurs, and also guarantee that the digitlines remain at Vcc/2 despite leakage paths that might otherwise discharge them. NMOS transistors are again used because of their superior drive strength, but also to allow integration with the equilibration transistor. Most often, layout integrates the bias and equilibration transistors to reduce their effective size and since they share the same EQ control signal. Most modern DRAMs use Vcc/2 volt precharge since this reduces power consumption, improves sensing and reduces read time. An exception to Vcc/2 precharge exists in the IBM® 16Mbit PMOS mbit DRAM designs that equilibrate and bias the digitlines to Vcc [2]. Since the wordlines and digitlines are both at Vcc when the part is inactive, row to column shorts do not contribute to increased standby current. On the other hand, row to column shorts cause higher standby current in Vcc/2 precharge DRAMs since their wordlines are kept at ground in standby. A typical layout for the equilibration and bias circuit appears in figure 2.12.

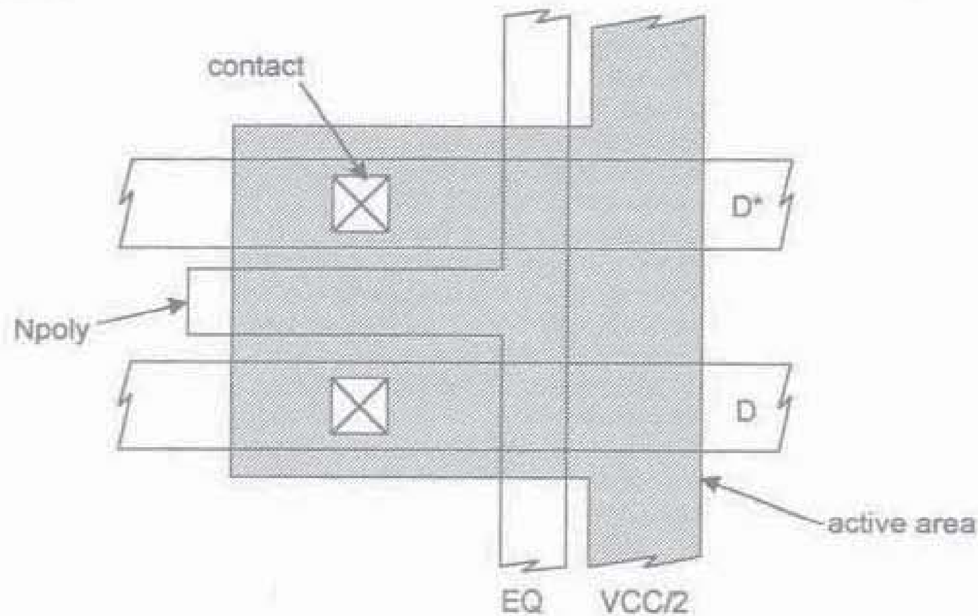


Figure 2.12: Equilibration and bias circuit layout

Isolation devices are important elements in sense amplifier circuits. Generally implemented as NMOS transistors, isolation transistors are placed between the array digitlines and specific sense amplifier components. As will be understood shortly, there are a multitude of possible configurations for the sense amplifier block. Isolation devices provide two functions. First, if the sense amps are positioned between and connected to two arrays, they allow one of the two arrays to be electrically

isolated. This isolation is necessary whenever a wordline fires high in one of the arrays. Isolation of the second array will reduce the total digitline capacitance connected to the sense amplifiers. This speeds read and write time, reduces power consumption, and extends refresh for the isolated array. Second, the isolation devices provide some resistance between the sense amplifier and the array digitlines. This resistance stabilizes the sense amplifiers and speeds up the sensing operation by minimizing the amount of capacitance directly connected to the sense amplifiers [5]. Capacitance of the sense nodes, between isolation transistors, is generally less than 15fF. This low capacitance permits the sense amplifiers to latch quickly. The restore operation slows, though, because of the isolation resistance, but this is less important than sensing and stability. Isolation transistors are physically located on both ends of the sense amplifier layout. For quarter pitch sense amplifiers, there is one isolation transistor for every two digitlines. Although this is twice the active area width and space of an array, it nevertheless establishes the minimum isolation used in the pitch cells.

Input/output (I/O) transistors allow data to be read from or written to specific digitline pairs. A single I/O transistor connects to each sense node as shown in figure 2.13. The outputs of each I/O transistor are connected to I/O signal pairs. Commonly, there are two pairs of I/O signal lines permitting four I/O transistors to share a single column select control signal. DRAM designs employing two or more metal layers run the column select lines across the arrays using either metal2 or metal3. Each column select activates four I/O transistors on both sides of an array, permitting the connection of four digitline pairs (columns) to peripheral data path circuits. The I/O transistors are carefully sized to ensure that the I/O bias voltage or remnant voltage on the I/O lines does not introduce instability into the sense amplifiers. Although designs vary significantly as to the numerical ratio, I/O transistors are two to eight times smaller than the N-sense amplifier transistors. This relationship is referred to as beta ratio. A beta ratio between five and eight is common, although proper selection can only be verified with silicon, since simulations fail to adequately predict sense amplifier instability.

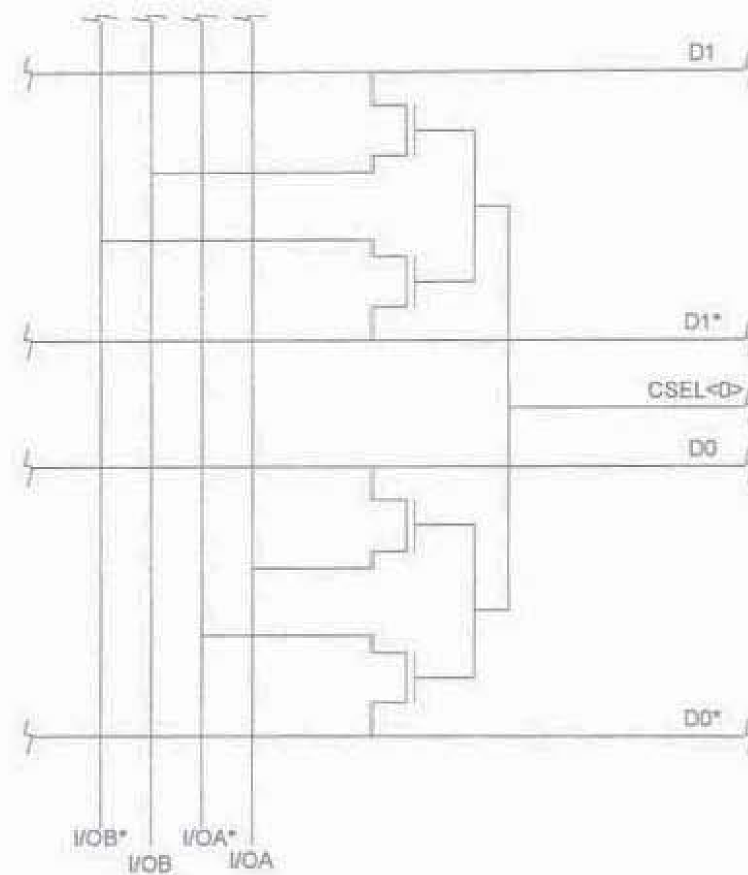


Figure 2.13: Schematic showing I/O transistors

The fundamental elements of any sense amplifier block are the N-sense amplifier and the P-sense amplifier. These amplifiers, as previously discussed, work together to detect the access signal voltage and drive the digitlines, accordingly to Vcc and ground. The N-sense amplifier, depicted in figure 2.14, consists of cross-coupled NMOS transistors. The N-sense amplifier drives the low

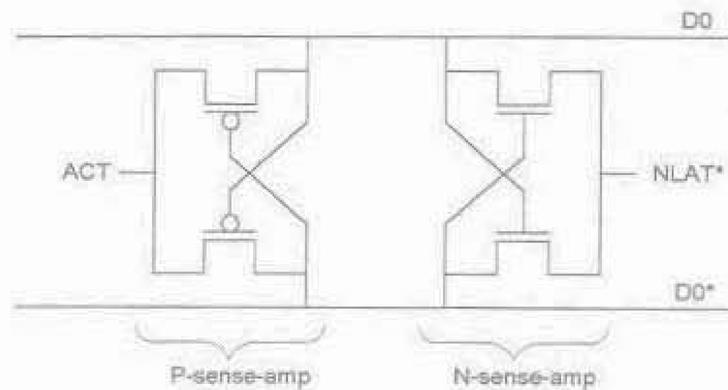


Figure 2.14: Basic sense amplifier schematic

potential digitline to ground. Similarly, the P-sense amplifier, consists of cross-coupled PMOS transistors and drives the high potential digitline to Vcc. The sense amplifiers are carefully designed to guarantee correct detection and amplification of the small signal voltage produced during cell access (less than 200mV) [1]. Matching of transistor Vth, transconductance, and junction capacitance within close tolerances helps ensure reliable sense amplifier operation. Ultimately, the layout dictates the overall balance and performance of the sense amplifier block. As a result, a tremendous amount of time is spent ensuring that the sense amplifier layout is optimum. Symmetry and exact duplication of elements are critical to a successful design including balanced coupling to all sources of noise, such as I/O lines and latch signals (NLAT* and ACT). Balance is especially critical for layout residing inside the isolation transistors, since the sense node capacitance is very low, making it more sensitive to noise and circuit imbalances.

While the majority of DRAM designs latch the digitlines to Vcc and ground, a growing number of designs are beginning to reduce these levels. Various technical papers report improved refresh times and lower power dissipation through reductions in latch voltages [21], [22]. At first, this appears contradictory, since writing a smaller charge into the memory cell should produce lower refresh time. The benefits derive from maintaining lower drain to source voltages (Vds) and negative gate to source voltages (Vgs) across non-accessed mbit transistors. Lower Vds and negative Vgs translate to substantially lower subthreshold leakage and longer refresh, despite the smaller stored charge. Most designs that implement reduced latch voltages, generally raise the N-sense amplifier latch voltage without lowering the P-sense amplifier latch voltage. Designated as boosted sense ground designs, they write data into each mbit using full Vcc for a logic one and boosted ground for a logic zero. The sense ground level is generally a few hundred millivolts above true ground. In standard DRAMs which drive digitlines fully to ground, the Vgs of non-accessed mbits becomes zero when the digitlines are latched. This results in high subthreshold leakage for a stored one level, since full Vcc exists across the mbit transistor while the Vgs is held to zero. Stored zero levels do not suffer from prolonged subthreshold leakage since any amount of cell leakage produces a negative Vgs for the transistor. The net effect is that a stored one level leaks away much faster than a stored zero level. One's level retention, therefore, establishes the maximum refresh period for most DRAM designs. Boosted sense ground extends refresh, by reducing subthreshold leakage for stored ones. This is accomplished by guaranteeing negative gate to source bias on non-accessed mbit transistors. The benefit of extended refresh from these designs is somewhat diminished, though, by the added complexity of generating boosted ground levels and the problem of digitlines that no longer equilibrate at Vcc/2 volts.

The rate at which the sense amplifiers are activated has been the subject of some debate. A variety of designs utilize multistage circuits to control the rate at which NLAT* fires. Especially prevalent with boosted sense ground designs are two stage circuits that initially drive NLAT* quickly toward true ground, to speed sensing, and then bring NLAT* to the boosted ground level to reduce cell leakage. An alternative to this approach using two stage drivers, first drives NLAT* slowly toward ground to limit current and digitline disturbances. Following this phase is a second phase in which NLAT* drives strongly toward ground to complete the sensing operation. The second phase usually occurs in conjunction with ACT activation. Although these two designs have contrary operation, they each meet specific performance objectives--trading off noise and speed.

Figure 2.15 shows a sense amplifier block commonly utilized in double or triple metal designs. It features two P-sense amplifiers placed outside the isolation transistors, a pair of EQ/Bias devices, a single N-sense amplifier, and a single I/O transistor for each digitline. This design is quarter pitch (one sense amplifier for every four digitlines), as are figures 2.15 and 2.16, since only half of the sense amplifiers required for an array are on either side. Placement of the P-sense amplifiers outside the isolation devices is necessary since a full one level (Vcc) cannot pass through NMOS ISO transistors whose gate terminals are driven to Vcc. EQ/Bias transistors are also placed outside the ISO devices to permit continued equilibration of digitlines in the isolated arrays. The I/O transistor gate terminals for four adjacent digitlines connect to a common CSEL signal. Each of the four I/O transistors being tied to a separate I/O bus. This sense amplifier, although simple to implement, is somewhat larger than other designs due to the presence of two P-sense amplifiers.

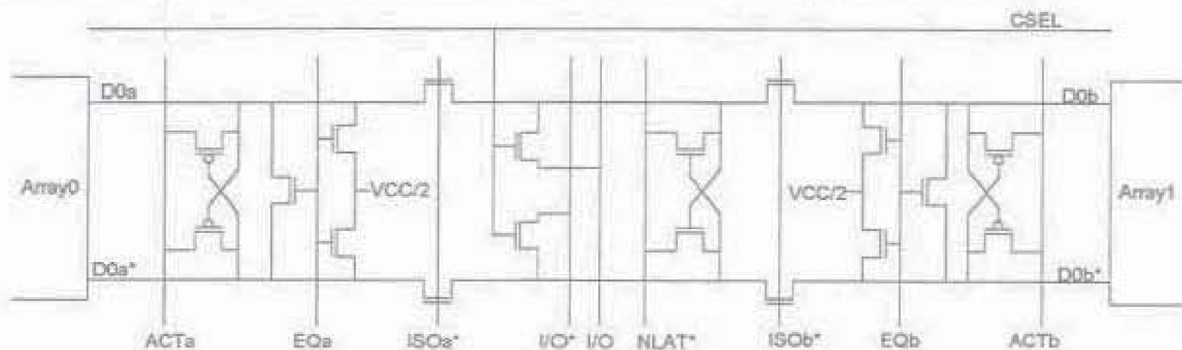


Figure 2.15: Standard sense amplifier block

Figure 2.16 shows a second, more complicated style of sense amplifier block. This design employs a single P-sense amplifier and three sets of N-sense amplifiers. In this design the P-sense amplifier is between the isolation transistors. Writing a full one level to the mbits requires that the isolation devices be either NMOS depletion, PMOS enhancement, or NMOS enhancement mode with boosted gate drive. The use of three N-sense amplifiers suggests the use of PMOS isolation

transistors, which prevent the writing of full zero levels unless the N-sense amplifiers exist outside the isolation devices. In this design the use of three N-sense amplifiers guarantees faster sensing and higher stability than a similar design using only two N-sense amplifiers. The inside N-sense amplifier fires prior to the outside N-sense amplifiers to ensure proper sensing. This design using three N-sense amplifiers will not yield a minimum layout, but this objective is sacrificed for higher performance.

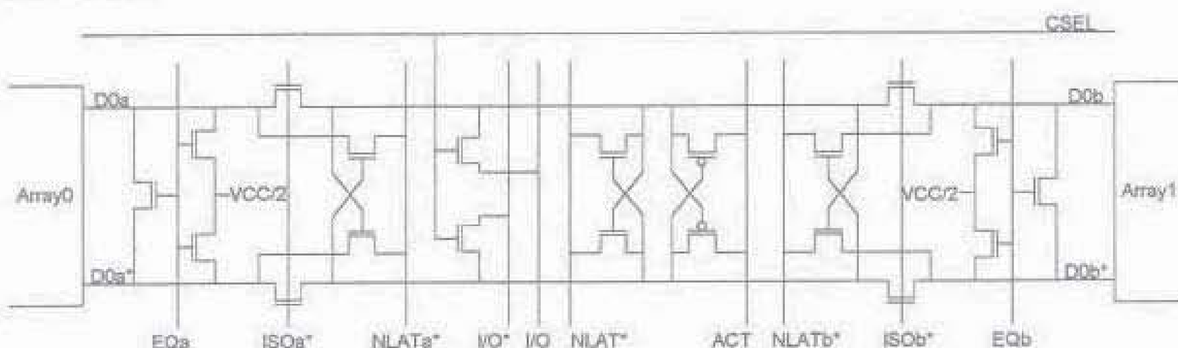


Figure 2.16: Complex sense amplifier block

The sense amplifier block shown in figure 2.17 is at or near the minimum configuration. This design features single N-sense and P-sense amplifiers placed between the isolation transistors. To write full logic levels requires that the isolation transistors be either depletion mode devices or that the gate voltages be boosted above Vcc by at least one Vth. This design still includes a pair of EQ/Bias circuits to maintain equilibration on isolated arrays. Only a few designs have ever tried to operate with a single EQ/Bias circuit residing within the isolation devices. Most designers consider floating digitlines a risky proposition since cell leakage increases and sensing operations degrade as the digitlines drifted away from the Vcc/2 precharge level. Future DRAM designs implemented on SOI (Silicon On Insulator) could permit the use of single EQ/Bias circuits, though, since this technology has negligible junction leakage [23], [24], [25].

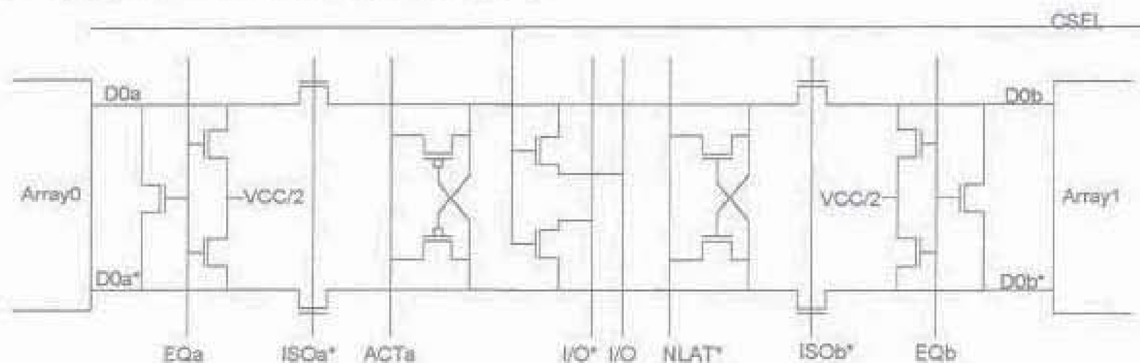


Figure 2.17: Minimized sense amplifier block

A sense amplifier design for use on a single metal DRAM appears in figure 2.18. Prevalent on 1Mb and 4Mb designs, single metal processes conceded to mult-imetal processes at the 16Mb generation. Unlike the double metal sense amplifiers shown in figures 2.15 through 2.17, single metal sense amps lay out at half pitch--one amplifier for every two array digitlines. This type of layout is extremely difficult, placing tight constraints on process design margins. Because of the unavailability of metal2, column select signals cannot route across the memory arrays. As a result, column select signal generation requires a localized rather than a global column decoding circuit. Figure 2.18 depicts how the N-sense amplifier and P-sense amplifiers are located on opposite ends of the mbit arrays in single metal designs. The sharing of sense amplifiers by two adjacent arrays is especially beneficial for single metal designs. Figure 2.18 illustrates this concept by showing how two arrays and their associated P-sense amplifiers share a single N-sense amplifier. In the depicted case, where I/O devices exist on only one end, the rightmost P-sense amplifier activates only when the right array is being accessed. The left P-sense amplifier, though, always activates, regardless of which array is being accessed, since all read and write operations must pass through the left P-sense amplifier to reach the I/O devices.

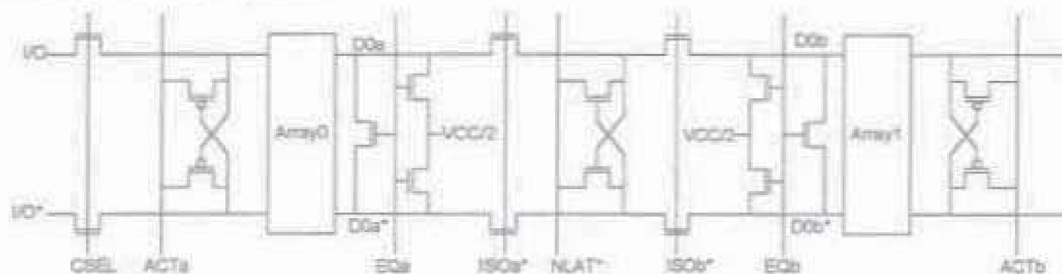


Figure 2.18: Single metal sense amplifier block

A set of operating signal waveforms appears in figure 2.19 for the sense amplifier shown in figure 2.15. The waveforms depict a read modify write cycle (late write) in which a read operation precedes a write operation during a single cell access. In this example, a one level is read from the cell, indicated by D0* rising above D0 during the cell access. A one level always appears as +Vcc/2 in the mbit cell, regardless of whether the cell connects to a true or complement digitline. The actual data state as viewed from the DRAM's data terminal (DQ) is a function of the data topology and the presence of data scrambling. Data scrambling or topo scrambling is implemented through additional datapath circuits to ensure that the stored mbit data level and DQ logic level are in agreement. An mbit one level (+Vcc/2) corresponds to a logic one at the DQ and an mbit zero level (-Vcc/2) corresponds to a logic zero at the DQ terminal. The late write timing sequence shown in figure 2.19 depicts a wordline firing in array1. Prior to the wordline firing, ISOA* must transition low to isolate array0 from the sense amplifiers. EQB also goes low and disables the EQ/Bias transistors connected to array1. The wordline fires high, accessing an mbit, which dumps its charge onto D0*. NLAT*,

initially at Vcc/2, drives low to begin the sensing operation by firing the N-sense amplifier to pull D0 toward ground. ACT fires, moving from ground to Vcc, which activates the P-sense amplifier, driving D0* toward Vcc. CSEL0 rises to Vcc, which enables the I/O transistors and connects the peripheral data circuits to the memory cells. The I/O lines are biased at a voltage close to Vcc. The I/O line bias causes D0 to rise slightly while the column is active. Following this read operation, write drivers in the periphery activate and drive the I/O lines to opposite data states (in this example). This new data propagates through the I/O devices and writing over the previous data held by the sense amplifiers. After the sense amplifiers latch the new data, the write drivers and the I/O devices shut down, allowing the sense amplifiers to restore the digitlines to full levels. The wordline transitions low to shut off the mbit transistor after the cell restoration completes. Finally, EQB and ISOA* fire high to equilibrate the digitlines back to Vcc/2 and to reconnect array0 to the sense amplifiers. While timing for each of these events needs to be minimized for optimum DRAM performance, timing cannot be pushed too far as to eliminate all timing margins. Margins are necessary to ensure proper device operation over the expected range of process variations and over a wide range of operating conditions.

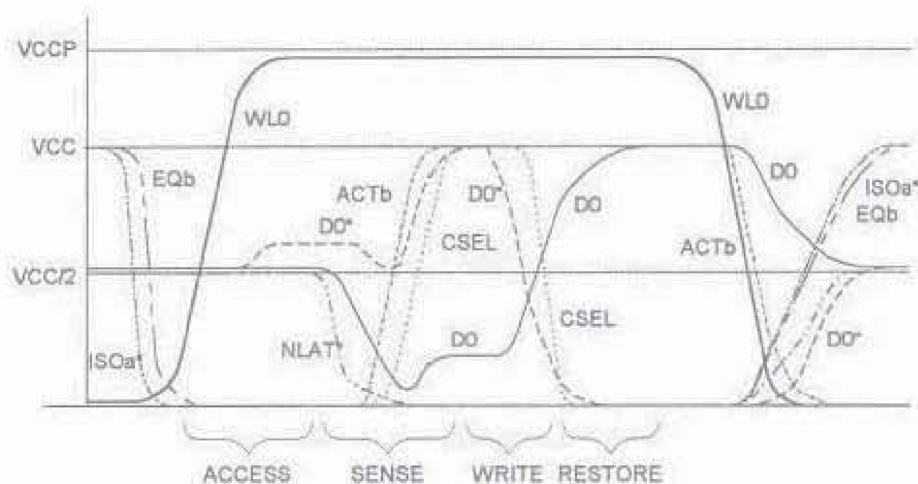


Figure 2.19: Waveforms for a read-modify-write DRAM cycle

Row Decoder Elements

Row decode circuits are similar to sense amplifier circuits in that they also pitch up to mbit arrays and have a variety of implementations. A row decode block consists of two basic elements, a wordline driver and an address decoder tree. There are three basic configurations for wordline driver circuits that include the nor driver, the inverter (CMOS) driver, and the bootstrap driver. Additionally, the drivers and associated decode trees can either be configured as local row decodes for each array

section or as global row decodes which drive a multitude of array sections. Global row decodes connect to multiple arrays through metal wordline straps. The straps are stitched to the polysilicon wordlines at specific intervals dictated by the polysilicon resistance and the desired RC wordline time constant. Most processes that strap wordlines with metal do not silicide the polysilicon, although doing so would reduce the number of stitch regions required. Strapping wordlines and using global row decoders obviously reduces die size--in some cases very dramatically. The penalty to strapping is that it requires an additional metal layer and that this layer is at minimum array pitch. This puts a tremendous burden on process technologists in which three conductors are at minimum pitch--wordlines, digitlines, and wordline straps. Distributed row decoders, on the other hand, do not require metal straps, but do require additional die size [26]. It is highly advantageous to reduce the polysilicon resistance in order to stretch the wordline length and reduce the number of needed row decodes especially on large DRAMs such as the 1gigabit.

The bootstrap wordline driver shown in figure 2.20 is built exclusively from NMOS transistors [26], producing the smallest layout for the three types of driver circuits. As the name denotes, this driver relies on bootstrapping principles to bias the output transistor's gate terminal. This bias voltage must be high enough to allow the NMOS transistor to drive the wordline to the boosted wordline voltage

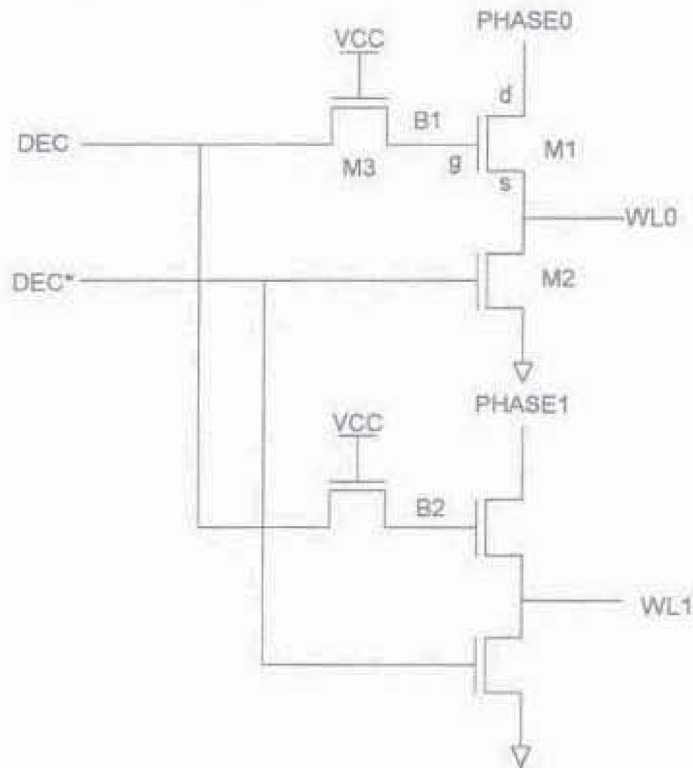


Figure 2.20: Bootstrap wordline driver schematic

VCCP. Operation of the bootstrap driver will be described with the help of figure 2.21. Initially the

wordline driver is off and the wordline and phase terminals are at ground. Transistor M2 holds the wordline at ground since the decoder output signal DEC* is at Vcc. The gate of pass transistor M3 is fixed at Vcc. The signals DEC and DEC* are generated by a separate decode circuit. DEC and DEC*, as a complement pair, represent the first of two terms necessary to decode the correct wordline. PHASE represents the second term. If DEC rises to Vcc and DEC* drops to ground, as determined by the decoder, the boot node labeled B1 will rise to Vcc-Vtn volts and transistor M2 will turn off. Transistor M1 continues to hold the wordline to ground since PHASE0 is still at ground. Once B1 reaches Vcc-Vtn, the PHASE signal fires to the boosted wordline voltage VCCP. As a result of gate to drain and gate to source capacitance of M1, the gate of M1 will boot to an elevated voltage, Vboot. This voltage is determined by the parasitic capacitance of node B1, Cgs1, Cgd1, VCCP and the initial voltage at B1, Vcc-Vtn. Accordingly, $Vboot \cong [(VCCP \cdot Cgd1) \div (Cgs1 + Cgd1 + CB1)] + (Vcc - Vtn)$. In conjunction with the wordline voltage rising from ground to VCCP, the gate to source capacitance of M1 provides a secondary boost to the boot node. The secondary boost helps to ensure that the boot voltage is adequate to drive the wordline to a full VCCP level.

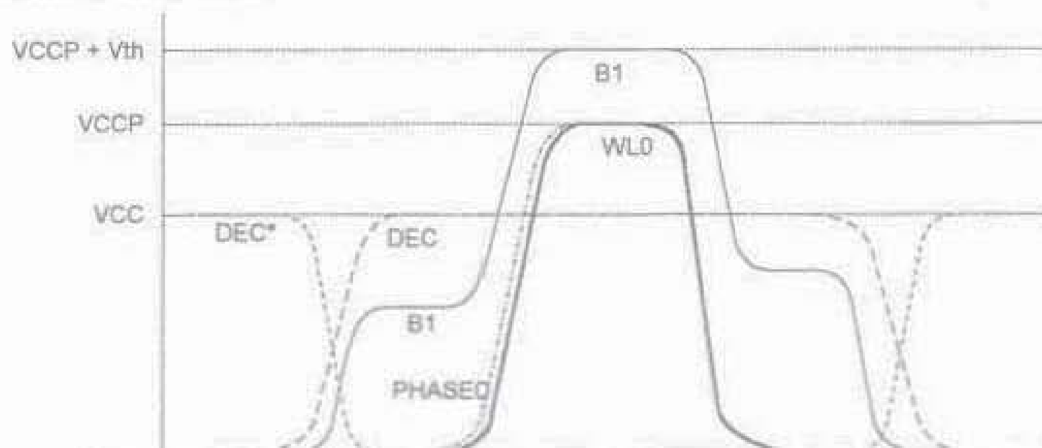


Figure 2.21: Bootstrap wordline driver operation

The bootstrap driver is turned off by first driving the PHASE0 signal to ground. M1 remains on, since node B1 cannot drop below Vcc-Vtn, substantially discharging the wordline toward ground. This is followed by the address decoder turning off, bringing DEC to ground and DEC* to Vcc. With DEC* at Vcc, transistor M2 turns on and fully clamps the wordline to ground. A voltage level translator is required for the PHASE0 signal since it operates between ground and the boosted voltage VCCP. For a global row decode configuration this is not much of a burden. For a local row decode configuration the level translators can be very difficult to implement. Generally, these translators are placed in array gaps which exist at the intersection of sense amplifier and row decode blocks, or they

are distributed throughout the row decode block itself. The translators require both PMOS and NMOS transistors and must be capable of driving large capacitive loads. Layout of the translators is very difficult, especially since the overall layout must be as small as possible.

The second type of wordline driver, shown in figure 2.22, is called a NOR driver. It is similar to the bootstrap driver in that two decode terms drive the output transistor from separate terminals. The NOR driver uses a PMOS transistor for M1 and does not rely on bootstrapping to derive the gate bias. Rather, the gate is driven by a voltage translator that converts DEC* from Vcc to VCCP voltage levels. This translation is necessary to ensure that M1 remains off for nonselected wordlines since the PHASE signal that is common to multiple drivers is driven to VCCP. To fire a specific wordline, DEC* must be low and the appropriate PHASE must fire high. Generally, there are four to eight PHASE signals per row decoder block. The NOR driver requires separate level translators for each

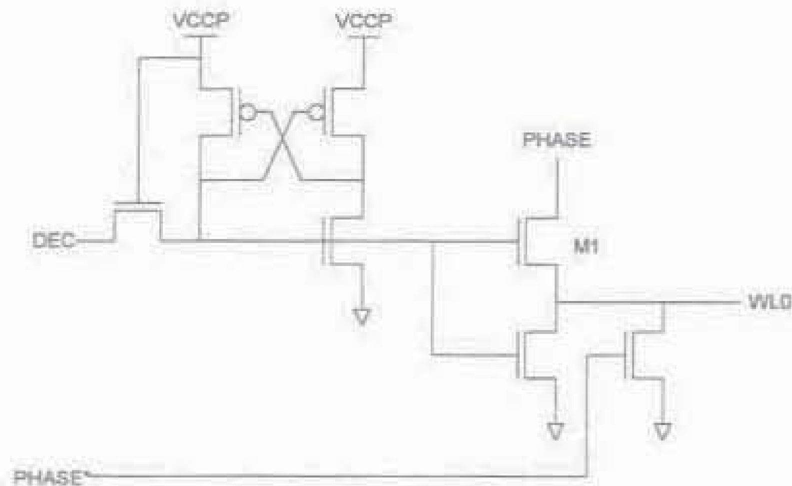


Figure 2.22: NOR wordline driver

PHASE and DEC* signal. The bootstrap driver by comparison requires translators only for the PHASE signals, far few than the NOR driver.

The final wordline driver configuration seen in figure 2.23 lacks a specific name, but is sometimes referred to as a CMOS inverter driver or CMOS driver. Unlike the first two drivers, the output transistor, M1, has its source terminal permanently connected to VCCP [27]. This driver, therefore, requires a voltage translator for each and every wordline. Both decode terms DEC and PHASE* combine to drive the output stage through the translator. The advantage of this driver, other than simple operation, is low power consumption. The CMOS driver conserves power because the translators drive only the small capacitance associated with a single wordline driver. The PHASE translators of both the bootstrap and NOR drivers must drive considerable junction capacitance. Disadvantages to the CMOS driver are layout complexity and high standby current. Subthreshold

leakage in transistor M1 and the high drain to source bias (VCCP) cause high standby current for this design. The bootstrap and NOR drivers do not suffer from high standby current since VCCP is not applied to the output transistors when the drivers are disabled. For DRAM designs having thousands of wordline drivers, the cumulative leakage current can easily exceed the standby current budget. Only careful design of transistor M1 will prevent this from occurring.

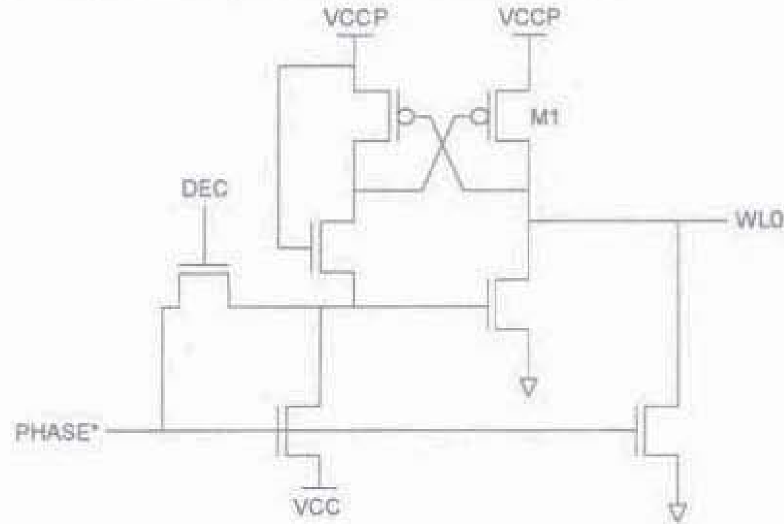


Figure 2.23: CMOS wordline driver

Address decode trees are the final element of the row decode block to be discussed. Decode trees are constructed from all types of logic--static, dynamic, pass gate, or a combination thereof. Regardless of what type of logic that an address decoder is implemented with, the layout must completely reside beneath the row address signal lines to constitute an efficient, minimal design. In other words, the metal address tracks dictate the die area available for the decoder. For DRAM designs that utilize global row decode schemes, the penalty for inefficient design may be insignificant, but for distributed local row decode schemes, the die area penalty can be significant. As with mbits and sense amplifiers, any time invested in row decode optimization is well spent.

The simplest type of address decode tree utilizes static CMOS logic. Shown in figure 2.24, a simple tree can be designed using 2-input NAND gates. While easy to design schematically, static logic address trees are not popular, since they waste silicon and are difficult to layout efficiently. Static logic requires two transistors for each address term, one NMOS and one PMOS--producing a large layout for designs with many address terms. Furthermore, static gates must be cascaded to accumulate address terms, adding gate delays with each level. For these and other reasons, static logic gates are not utilized in row decode address trees for today's state of the art DRAM.

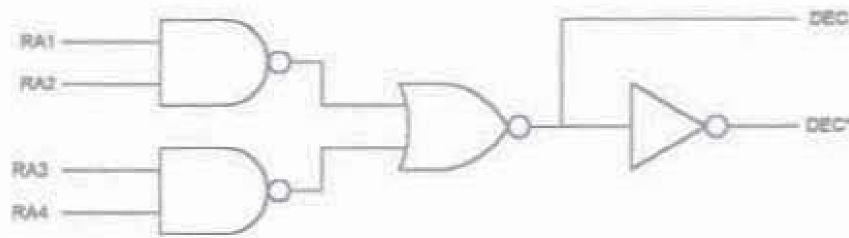


Figure 2.24: Static decode tree

The second type of address tree utilizes dynamic logic, the most prevalent being precharge and evaluate (P&E) logic. Used by the majority of DRAM manufacturers, P&E address trees come in a variety of forms, although the differences between one and another can be subtle [26], [27]. Figure 2.25 shows a simplified schematic for one version of a P&E address tree designed for use with

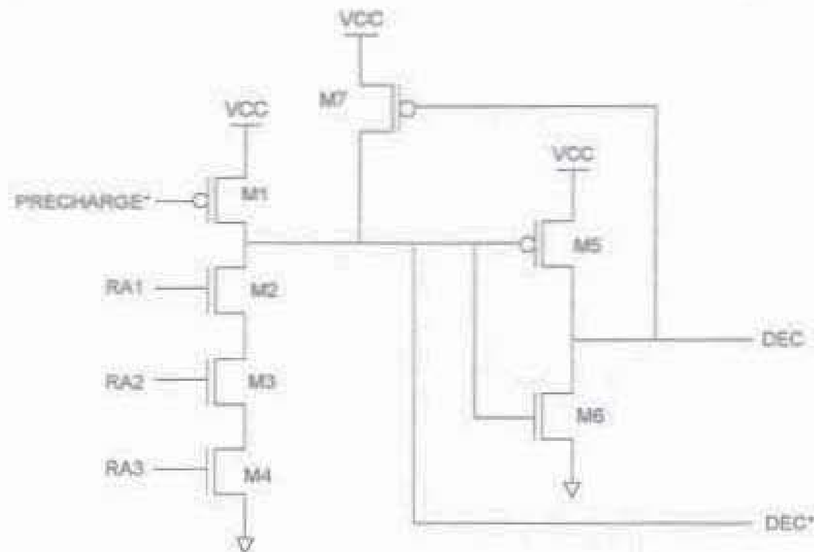


Figure 2.25: P&E decode tree

bootstrapped wordline drivers. P&E address tree circuits feature one or more PMOS precharge transistors and a cascade of enable NMOS transistors M2-M4. This P&E design uses half of the transistors required by the static address tree of figure 2.24. As a result, the layout of the P&E tree is much smaller than the static tree fitting more easily under the address lines. The PRECHARGE* signal, under control of the RAS chain logic, drives the precharge transistor. PRECHARGE* and transistor M1 ensure that DEC* precharges high, disabling the wordline driver and preparing the tree for row address activation. M7 is a small PMOS transistor driven by the DEC inverter (M5 and M6). Together, M7 and the inverter form a latch ensuring that DEC* remains high for all rows not selected by the row addresses. At the beginning of any RAS cycle, PRECHARGE* is low and the row addresses are all disabled (low). After RAS* falls, PRECHARGE* transitions high to turn off M1. Following this, the row addresses are enabled. If RA1 through RA3 all go high, then M2-M4 will turn

on, overpowering M7 and driving DEC* to ground and DEC to Vcc. The output of each tree segment normally drives four bootstrapped wordline drivers--each connected to a separate PHASE signal. For an array with 256 wordlines there will be 64 such decode trees.

The row address lines shown as RA1-RA3 can be either true and complement address lines or predecoded address lines. Predecoded address lines are formed by logically combining (AND) addresses as shown in table 2.1. Advantages to predecoded addresses include lower power, since

Table 2.1: Predecoded address truth table

RA0	RA1	PR01(n)
0	0	0
1	0	1
0	1	2
1	1	3

fewer signals make transitions during address changes, and higher efficiency, since only three transistors are necessary to decode six addresses for the circuit of figure 2.25. Predecoding is especially beneficial for redundancy circuits. Predecoded addresses are used throughout most DRAM designs today.

The final type of address tree, shown in figure 2.26, utilizes pass transistor logic. Pass transistor address trees are similar to P&E trees in many ways. Both designs utilize PMOS precharge transistors and NMOS address enable transistors. Unlike P&E logic, though, the NMOS cascade does not terminate at ground. Rather, the cascade of M2-M4 goes to a PHASE* signal that is high during precharge and low during the evaluation. The address signals operate the same as in the P&E tree, high to select and low to deselect. The pass transistor tree is shown integrated into a CMOS wordline driver. This is necessary since they are generally used together and their operation is very complementary. The cross-coupled PMOS transistors of the CMOS level translator provide a latch necessary to keep the final interstage node biased at Vcc. The latch has a weak pull-up, easily overpowered by the cascaded NMOS enable transistors. A pass transistor address tree is not used with bootstrapped wordline drivers, since the PHASE signal feeds into the address tree logic rather than into the driver as required by the bootstrap driver.

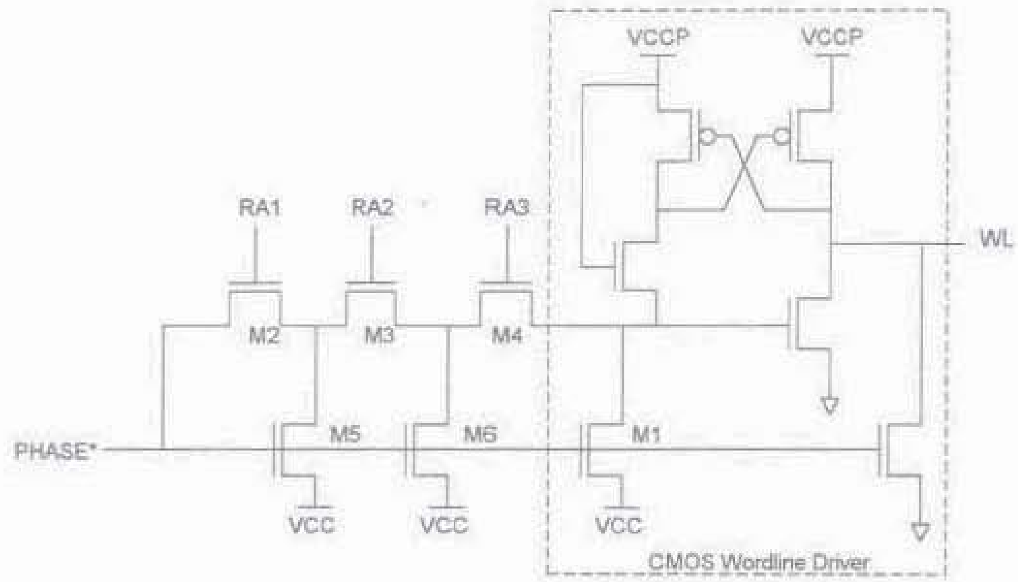


Figure 2.26: Pass transistor decode tree schematic

III. Architectural Characteristics

This section provides a detailed description of the two most prevalent array architectures under consideration for future large scale DRAMs--the aforementioned open digitline and folded digitline architectures. To provide a viable point for comparison, each architecture will be employed in the theoretical construction of 32Mbit memory blocks for use in a 256Mbit DRAM. Design parameters and layout rules from a typical 0.25μm DRAM process provide the needed dimensions and constrainsts for the analysis. Some of these parameters are shown in table 3.1. Examination of DRAM architectures in the light of a real world design problem permits a more objective and

Table 3.1: 0.25μm design parameters

Parameter	Value
Digitline width Wdl	0.3 μm
Digitline pitch Pdl	0.6μm
Wordline width Wwl	0.3 μm
Wordline pitch for $8F^2$ mbit Pwl8	0.6μm
Wordline pitch for $6F^2$ mbit Pwl6	0.9μm
Cell capacitance Cc	30fF
Digitline capacitance per mbit Cdm	0.8fF
Wordline capactivance per $8F^2$ mbit Cw8	0.6fF
Wordline capactivance per $6F^2$ mbit Cw6	0.5fF
Wordline sheet resistance Rs	6Ω/sq

unbiased comparison. An added benefit to this approach is that the strengths and weaknesses of either architecture should become readily apparent.

Open Digitline Array Architecture

The open digitline array architecture was the prevalent architecture prior to the 64Kbit DRAM. A modern embodiment of this architecture as shown in figure 3.1 [8], [14], is constructed with multiple crosspoint array cores separated by strips of sense amplifier blocks in one axis and either row decode blocks or wordline stitching regions in the other axis. Each 128Kbit array core is built using $6F^2$ mbit cell pairs. There are a total of 131,072 (2^{17}) functionally addressable mbits arranged in 264 rows and 524 digitlines. The 264 rows consist of 256 actual wordlines, 4 redundant wordlines and 4 dummy wordlines. The 524 digitlines consist of 512 actual digitlines, 8 redundant digitlines, and 4 dummy digitlines. Photolithography problems usually occur at the edge of large repetitive structures such as mbit arrays. These problems produce malformed or nonuniform structures, rendering the

edge cells useless. Inclusion of dummy mbits, wordlines and digitlines on each array edge ensures that these problems only occur on dummy cells, leaving live cells unaffected. Although dummy structures make each array core larger, they also significantly improve device yield making them necessary items on all DRAM designs.

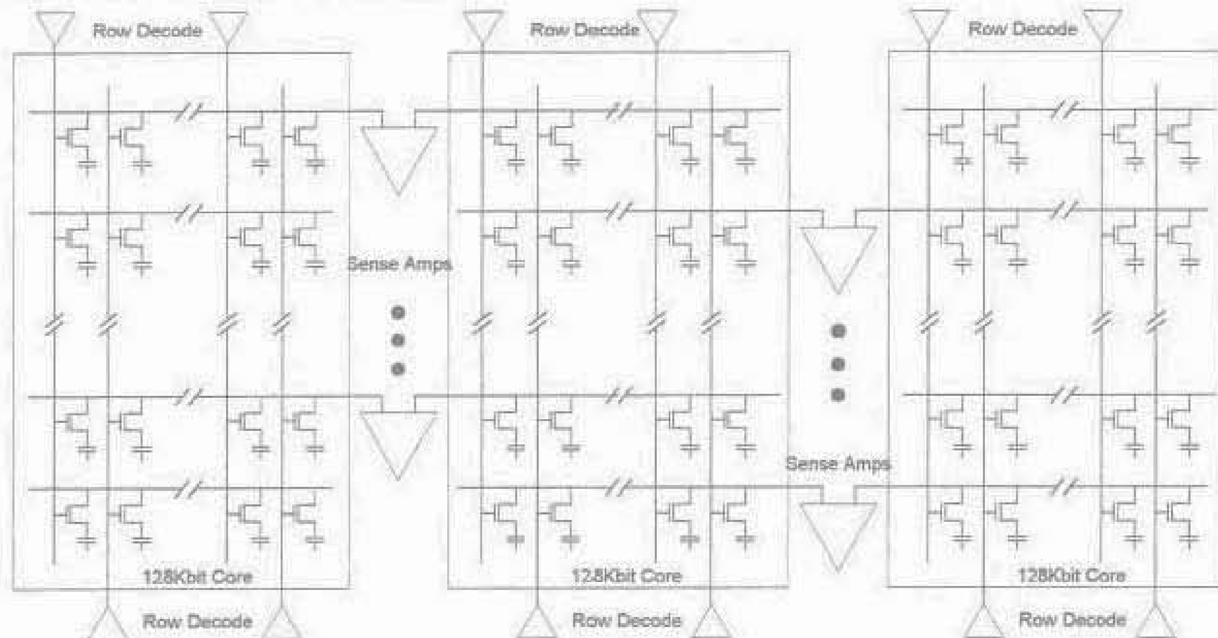


Figure 3.1: Open digitline architecture schematic

Array core size, as measured in the number of mbits, is restricted by two factors--a desire to keep the quantity of mbits binary and practical limits on wordline and digitline length. The need for a binary quantity of mbits in each array core derives from the binary nature of DRAM addressing. Given N row addresses and M column addresses for a given part, there are a total of 2^{N+M} addressable mbits. Address decoding is greatly simplified within a DRAM if array address boundaries are derived directly from address bits. Since the addressing is binary, the boundaries naturally become binary. Therefore, the size of each array core must necessarily have 2^X addressable rows and 2^Y addressable digitlines. The resulting array core size is 2^{X+Y} mbits, which is of course a binary number. The second set of factors limiting array core size are practical limits on digitline and wordline length. From earlier discussions in section-II, the digitline capacitance is limited by two factors. First, the ratio of cell capacitance to digitline capacitance must fall within a specified range to ensure reliable sensing. Second, operating current and power for the DRAM is in large part determined by the current required to charge and discharge the digitlines during each active cycle. Power considerations restrict digitline length for the 256Mbit generation to approximately 128 mbit pairs (256 rows), with each mbit connection adding capacitance to the digitline. The power dissipated during a read or refresh operation is proportional to the digitline capacitance (Cd), the supply voltage

(Vcc), the number of active columns (N), and the refresh period (P). Accordingly, the power dissipated is given as $Pd = Vccx \cdot (N \cdot Vcc \cdot (Cd + Cc)) \div (2 \cdot P)$ watts. On a 256Mbit DRAM in 8K refresh there are 32,768 (2^{15}) active columns during each read, write, or refresh operation. Active array current and power dissipation for a 256Mbit DRAM are given in table 3.2 for a 90nS refresh

Table 3.2: Active current and power versus digitline length

Digitline Length	Digitline Capacitance	Active Current	Power Dissipation
128 mbits	102fF	60mA	199mW
256 mbits	205fF	121mA	398mW
512 mbits	410fF	241mA	795mW

period (-5 timing) at various digitline lengths. The budget for active array current is limited to 200mA for this 256Mbit design. To meet this budget, the digitline cannot exceed a length of 256 mbits.

Wordline length, as described in section-II, is limited by the maximum allowable RC time constant of the wordline. To ensure acceptable access time for the 256Mbit DRAM, the wordline time constant

Table 3.3: Wordline time constant versus wordline length

Wordline Length	Rwl	Cwl	Time Constant
128 mbits	1,536ohms	64fF	0.098nS
256 mbits	3,072ohms	128fF	0.39nS
512 mbits	6,144ohms	256fF	1.57nS
1024 mbits	12,288ohms	512fF	6.29nS

should be kept below four nanoseconds. For a wordline connected to N mbits, the total resistance and capacitance using 0.3μm polysilicon are $Rwl = Rs \cdot N \cdot Pwl \div 0.3\mu m$ ohms and $Cwl = Cw6 \cdot N$ Farads respectively. Table 3.3 contains the effective wordline time constants for various wordline lengths. As shown in the table, the wordline length cannot exceed 512 mbits (512 digitlines) if the wordline time constant is to remain under four nanoseconds.

The open digitline architecture does not support digitline twisting since the true and complement digitlines which constitute a column are in separate array cores. Therefore, no silicon area is consumed for twist regions. The 32Mbit array block requires a total 256 128Kbit array cores in its construction. Each 32Mbit block represents an address space comprising a total of 4096 rows and 8192 columns. A practical configuration for the 32Mbit block is depicted in figure 3.2. In this figure,

the 256 array cores appear in a 16 by 16 arrangement. The by 16 arrangement produces 2Mbit sections consisting of 256 wordlines and 8192 digitlines (4096 columns). A total of 16 2Mbit sections are required to form the complete 32Mbit block. Sense amplifier strips are positioned vertically between each 2Mbit section. Row decode strips or wordline stitching strips are positioned horizontally between each array core.

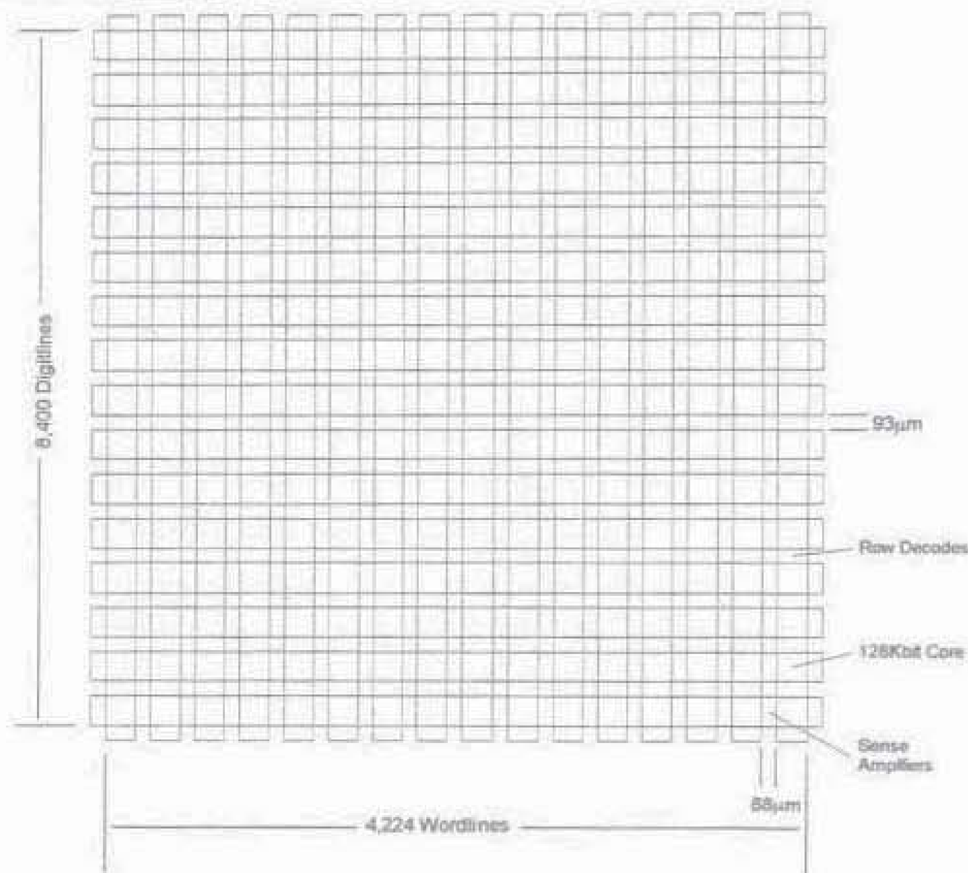


Figure 3.2: Open digitline 32Mbit array block

Layout was generated for the various 32Mbit elements depicted in figure 3.2 using commercial IC layout software. This layout is necessary to obtain reasonable estimates for pitch cell size. These size estimates, in turn, allow overall dimensions of the 32Mbit memory block to be calculated. The results of these estimates appear in figure 3.2. Essentially, the overall height of the 32Mbit block can be found by summing the height of the row decode blocks (or stitch regions) together with the product of wordline pitch and total number of wordlines. Accordingly,

$Height32 = (Tr \cdot Hldec) + (Tdl \cdot Pdl)$ microns, where Tldec is the number of local row decoders, Hldec is the height of each decoder, Tdl is the number of wordlines including redundant and dummy lines, and Pdl is the digitline pitch. Similarly, the width of the 32Mbit block is found by summing the total width of the sense amplifier blocks together with the product of the wordline pitch and the number of

wordlines. This bit of math yields $Width32 = (Tsa \cdot Wamp) + (Twl \cdot Pwl6)$ microns, where Tsa is the number of sense amplifier strips, Wamp is the width of the sense amplifiers, Tdl is the total number of wordlines including redundant and dummy lines, and Pwl6 is the wordline pitch for the $6F^2$ mbit . Table 3.4 contains calculation results for the 32Mbit block shown in figure 3.2. While overall size is the best measure of architectural efficiency, a second popular metric is array efficiency. Array efficiency is determined by dividing the area consumed by functionally addressable mbits by the total die area. To simplify the analysis in this thesis, peripheral circuits will be ignored in the array efficiency calculation. Rather, the array efficiency calculation will only consider the 32Mbit memory

Table 3.4: Open digitline (local row decode)--32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	88μm
Number of local decode strips	Tldec	17
Height of local decode strips	Hldec	93μm
Number of digitlines	Tdl	8,400
Number of wordlines	Twl	4,224
Height of 32Mbit block	Height32	6,621μm
Width of 32Mbit block	Width32	5,298μm
Area of 32Mbit block	Area32	35,078,058μm^2

block ignoring all other factors. With this simplification, the array efficiency for a 32Mbit block is given as $Efficiency = (100 \cdot 2^{25} \cdot Pdl \cdot Pwl6) \div (Area32)$ percent, where 2^{25} is the number of addressable mbits in each 32Mbit block. The open digitline architecture yields a calculated array efficiency of 51.7 percent.

Unfortunately, the ideal open digitline architecture presented in figure 3.2 cannot be realized in practice. There are serious implications to using an open digitline architecture. The first problem is that the sense amplifiers must connect to two separate array cores. While this requirement is easily understood, the implications are far from trivial. Memory block edges present a real problem since the sense amplifiers cannot be placed outside of arrays as depicted in figure 3.2--they must always be between arrays [12]. There are two methods of solving this problem. First, design the sense amplifiers such that the sense amplifier block contains a set of sense amplifiers for every digitline in the array. This solution, seen in figure 3.3, eliminates the need to have sense amplifiers on both sides of an array core since all of the digitlines connect to a single sense amplifier block. Not only does this solution eliminate the edge problem, but it also reduces the 32Mbit block size since there

are only eight sense amplifier strips instead of the seventeen shown in figure 3.2. Unfortunately, it is next to impossible to layout sense amplifiers in this fashion [28]. Even single metal sense amplifier layout, considered the tightest layout in the industry, can only achieve one sense amplifier for every two digitlines.

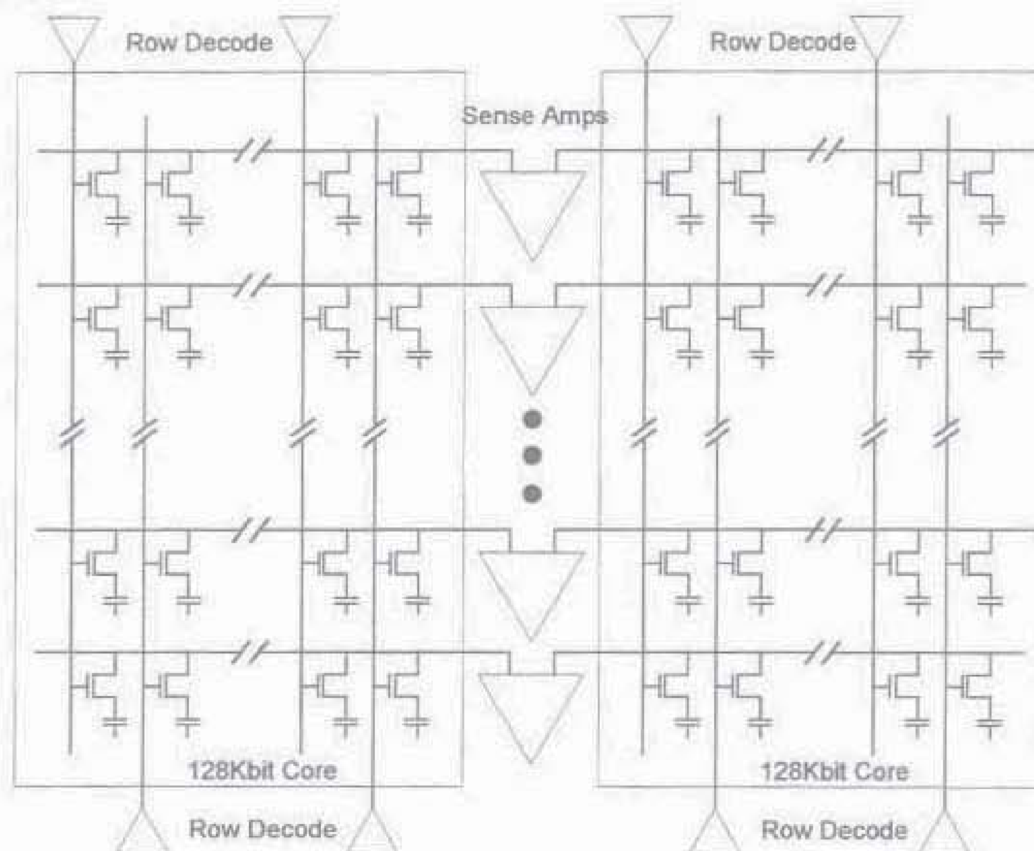


Figure 3.3: Single pitch open digitline architecture

A second approach to solving the array edge problem in open digitline architectures is to maintain the configuration shown in figure 3.2, but include some form of reference digitline for the edge sense amplifiers. The reference digitline can assume any form, as long as it accurately models the capacitance and behavior of a true digitline. Obviously the best type of reference digitline is a true digitline. Therefore, with this approach, additional dummy array cores are added to both edges of the 32Mbit memory block as shown in figure 3.4. The dummy array cores only need half as many wordlines as a true array core since only half of the digitlines are connected to any single sense amplifier strip. The unconnected digitlines are used to double the effective length of the reference digitlines. While this approach solves the array edge problem, it results in lower array efficiency and a larger 32Mbit memory block. Dummy arrays solve the array edge problem inherent in open digitline architecture, but they necessitate sense amplifier layouts that are on the edge of impossible.

The problem of sense amplifier layout is all the worse because of the need to route global column select lines through the layout. For all intents and purposes the sense amplifier layout cannot be completed without the presence of an additional conductor such as a third metal or without time multiplexed sensing. Therefore, for the open digitline architecture to be successful an additional metal must be added to the DRAM process.

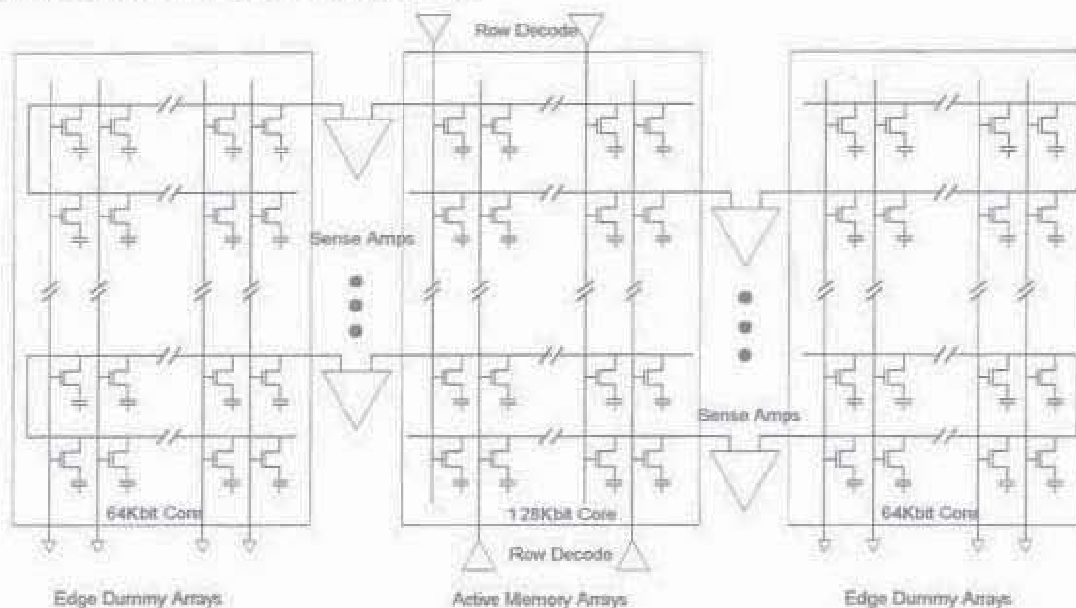


Figure 3.4: Open digitline architecture with dummy arrays

The presence of metal3 not only enables the sense amplifier layout, but also permits the use of a either a full or hierarchical global row decoding scheme. A full global row decoding scheme using wordline stitching places great demands upon metal and contact/via technologies, but represents the most efficient use of the additional metal. Hierarchical row decoding using bootstrap wordline drivers is slightly less efficient, but relaxes process requirements significantly [26]. For a balanced perspective, both approaches, global and hierarchical, were analyzed. The results of this analysis for the open digitline architecture are summarized in tables 3.5 and 3.6 respectivley. Array efficiency for global and hierarchical row decoding calculated to 60.5 percent and 55.9 percent respectively for the 32Mbit memory blocks based upon data from these tables.

Table 3.5: Open digitline (dummy arrays & global row decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	88μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	200μm
Number of stitch regions	Nst	17
Heigh of stitch regions	Hst	10μm
Number of digitlines	Tdl	8,400
Number of wordlines	Twl	4,488
Height of 32Mbit block	Height32	5,410μm
Width of 32Mbit block	Width32	5,535μm
Area of 32Mbit block	Area32	29,944,350μm^2

Table 3.6: Open digitline (dummy arrays & hier row decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	88μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	190μm
Number of hier decode strips	Thdec	17
Height of hier decode strips	Hhdec	37μm
Number of digitlines	Tdl	8,400
Number of wordlines	Twl	4,488
Height of 32Mbit block	Height32	5,859μm
Width of 32Mbit block	Width32	5,535μm
Area of 32Mbit block	Area32	32,429,565μm^2

Folded Array Architecture

The folded array architecture depicted in figure 3.5 is the standard architecture used in today's modern DRAM designs. The folded architecture is constructed with multiple array cores separated by strips of sense amplifiers and either row decode blocks or wordline stitching regions. Unlike the open digitline architecture, which uses $6F^2$ mbit cell pairs, the folded array core uses $8F^2$ mbit cell pairs [7]. Modern array cores include 262,144 (2^{18}) functionally addressable mbits arranged in 532 rows and 1044 digitlines. The 532 rows consists of 512 actual wordlines, 4 redundant wordlines and 16 dummy wordlines. Each row (wordline) connects to mbit transistors on alternating digitlines. The 1044 digitlines consist of 1024 actual digitlines (512 columns), 16 redundant digitlines (8 columns) and 4 dummy digitlines. As discussed earlier, photolithography limitations necessitate the use of the dummy wordlines and digitlines. These photo problems are very pronounced at the edges of large repetitive structures such as the array core.

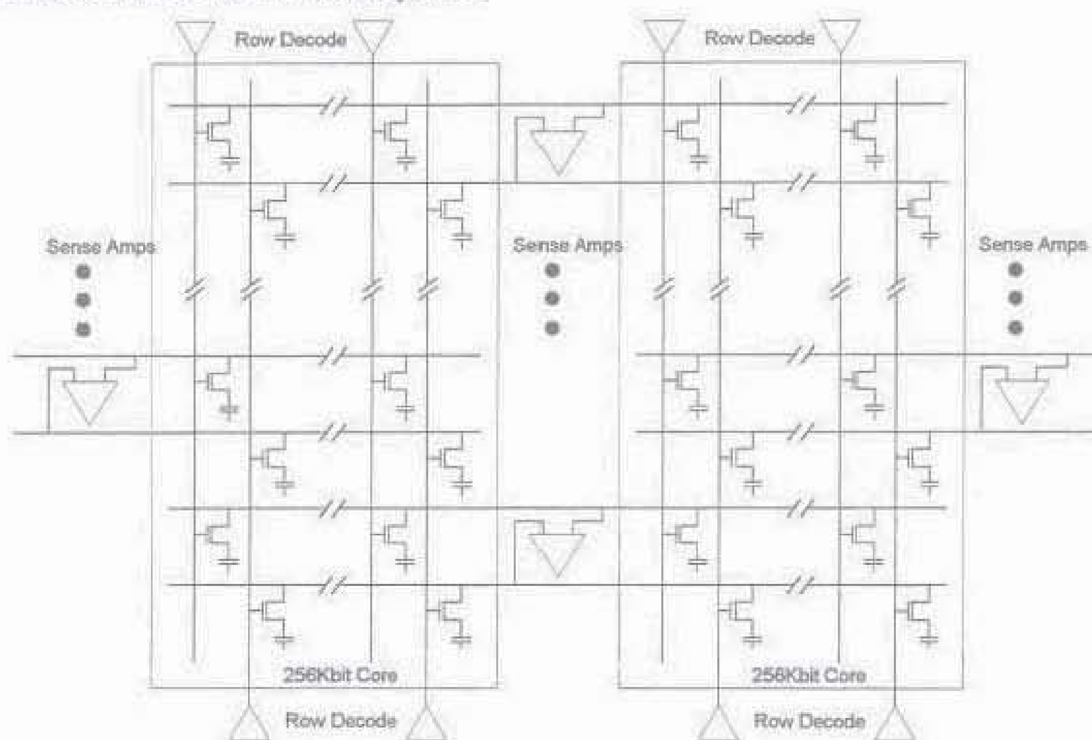


Figure 3.5: Folded digitline array architecture schematic

Sense amplifier blocks are placed on both sides of each array core. The sense amplifiers within each block are laid out at quarter pitch--one sense amplifier for every four digitlines. Each sense amplifier connects through isolation devices to columns (digitline pairs) from both adjacent array cores. Odd columns connect on one side of the core and even columns connect on the opposite side. Each sense amplifier block is therefore connected to only odd or even columns, never connecting to both odd and even columns within the same block. Connecting to both odd and even columns requires a

half pitch sense amplifier layout--one sense amplifier for every two digitlines. While half pitch layout is possible with certain DRAM processes, the bulk of production DRAM designs remain quarter pitch due to ease of layout. The analysis presented in this section is accordingly based upon quarter pitch design practices.

Location of row decode blocks for the array core depends upon the number of available metal layers. For one and two metal processes, local row decode blocks are located at the top and bottom edges of the core. Three and four metal processes support the use of global row decodes. Global row decodes require only stitch regions or local wordline drivers at the top and bottom edges of the core [11]. Stitch regions consume much less silicon area than local row decodes, substantially increasing array efficiency for the DRAM. The array core also includes digitline twist regions that run parallel to the wordlines. These regions provide the die area required for digitline twisting. Depending upon the particular twisting scheme selected for a design (see section-II), the array core will need between one and three twist regions. For the sake of analysis, a triple twist is assumed, since it offers the best overall noise performance and is the choice of DRAM manufacturers on advanced large scale applications [29]. Each twist region constitutes a break in the array structure necessitating the inclusion of dummy wordlines. For this reason there are 16 dummy wordlines (2 for each array edge) in the folded array core rather than 4 dummy wordlines as in the open digitline architecture.

The array core for folded digitline architectures contains more mbits than is possible for open digitline architectures. Larger core size is an inherent feature of folded architectures arising from the very nature of the architecture. Folded architectures get their name from the fact that a folded array core results from folding two open digitline array cores one on top of the other. The digitlines and wordlines from each folded core are spread apart (double pitch) to allow room for the other folded core. After folding, each constituent cores remains intact and independent, except for mbit changes ($8F^2$ conversion) that are necessary in the folded architecture. The array core size doubles since the total number of digitlines and wordlines doubles in the folding process. It does not quadruple as one might suspect, because the two constituent folded cores remain independent--the wordlines from one folded core do not connect to mbits in the other folded core. Digitline pairing (column formation) is a natural outgrowth of the folding process since each wordline only connects to mbits on alternating digitlines. The existence of digitline pairs (columns) is the one characteristic of folded digitline architectures that produces superior signal-to-noise performance. Furthermore, the digitlines that form a column are physically adjacent to one another. This feature permits various digitline twisting schemes to be used, as discussed in section-II, which further improves signal-to-noise.

Similar to the open digitline architecture, digitline length for the folded digitline architecture is again limited by power dissipation and minimum cell to digitline capacitance ratio. For the 256Mbit generation, digitlines are restricted from having connection to more than 256 cells (128 mbit pairs). The analysis to arrive at this quantity is similar to that for the open digitline architecture. Refer back to Table 3.2 to view the calculated results of power dissipation versus digitline length for a 256Mbit DRAM in 8K refresh. Wordline length is again limited by the maximum allowable RC time constant of the wordline. Contrary to an open digitline architecture in which each wordline connects to mbits on each digitline, the wordlines in a folded digitline architecture only connect to mbits on alternating digitlines. Therefore, a wordline can cross 1024 digitlines while only connecting to 512 mbit

Table 3.7: Wordline time constant versus wordline length (folded)

Wordline Length	Rwl	Cwl	Time Constant
128 mbits	3.072ohms	77fF	0.24nS
256 mbits	6,144ohms	154fF	0.95nS
512 mbits	12,288ohms	307fF	3.77nS
1024 mbits	24,576ohms	614fF	15.09nS

transistors. The wordlines will have twice the overall resistance, but only slightly more capacitance since the wordlines run over field oxide on alternating digitlines. Table 3.7 contains the effective wordline time constants for various wordline lengths for a folded array core. For a wordline connected to N mbits, the total resistance and capacitance using 0.3μm polysilicon are $Rwl = 2 \cdot Rs \cdot N \cdot Pwl \div 0.3\mu m$ ohms and $Cwl = Cw8 \cdot N$ farads respectively. As shown in table 3.7, the wordline length cannot exceed 512 mbits (1024 digitlines) for the wordline time constant to remain under four nanoseconds. Although the wordline connects to only 512 mbits it is two times longer (1024 digitlines) than wordlines in open digitline array cores. The folded digitline architecture therefore requires half as many row decode blocks or wordline stitching regions as the open digitline architecture.

A diagram of a 32Mbit array block using folded digitline architecture is shown in figure 3.6. This block requires a total of 128 256Kbit array cores. In this figure, the 128 array cores are arranged in an 8 row and 16 column configuration. The by 8 row arrangement produces 2Mbit sections consisting of 256 wordlines and 8192 digitlines (4096 columns). A total of 16 2Mbit sections form the complete 32Mbit array block. Sense amplifier strips are positioned vertically between each 2Mbit section, as was done in the open digitline architecture. Again, row decode blocks or wordline stitching regions are positioned horizontally between the array cores.

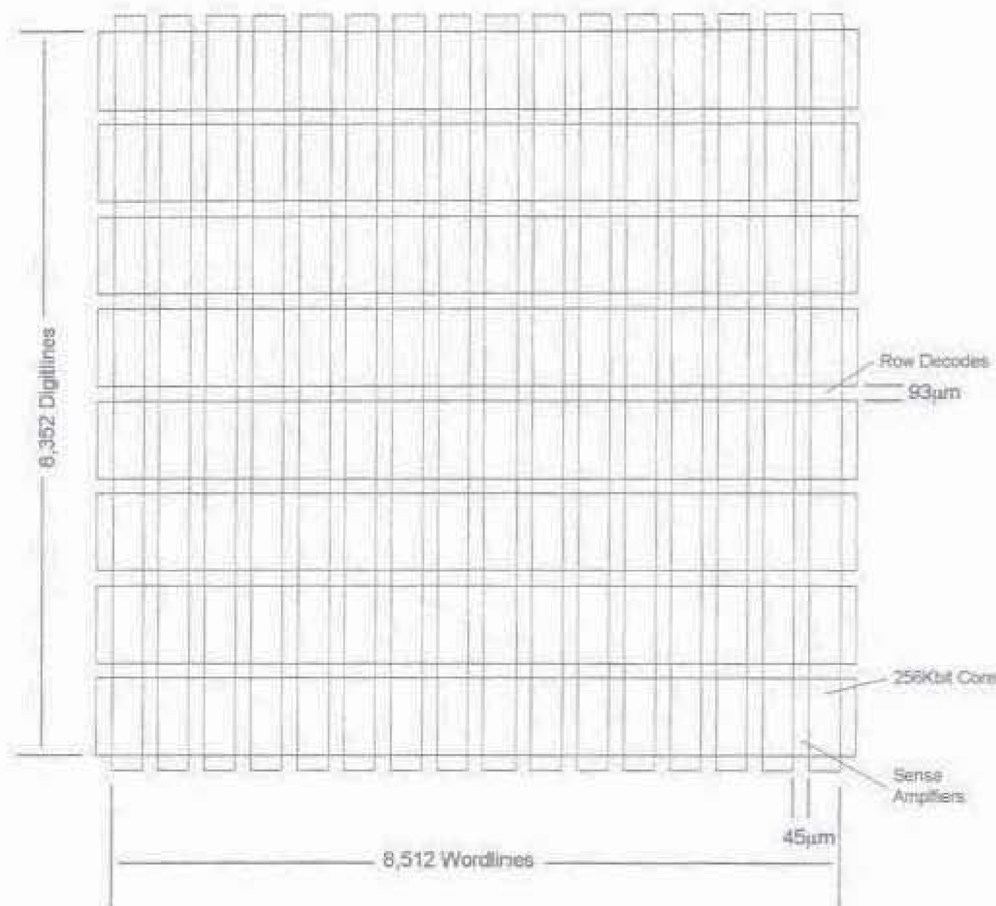


Figure 3.6: Folded digitline architecture 32Mbit array block

The 32Mbit array block shown in figure 3.6 includes size estimates for the various pitch cells. Layout was generated where necessary to arrive at the size estimates. Overall size for the folded digitline 32Mbit block can be found by again summing the dimensions for each component. Accordingly, $Height32 = (Tr \cdot Hrdec) + (Tdl \cdot Pdl)$ microns, where Tr is the number of row decoders, Hrdec is the height of each decoder, Tdl is the number of digitlines including redundant and dummy, and Pdl is the digitline pitch. Similarly, $Width32 = (Tsa \cdot Wamp) + (Twl \cdot Pwl8) + (Ttwist \cdot Wtwist)$ microns, where Tsa is the number of sense amplifier strips, Wamp is the width of the sense amplifiers, Twl is the total number of wordlines including redundant and dummy, Pwl8 is the wordline pitch for the $8F^2$ mbit, Ttwist is the total number of twist regions, and Wtwist is the width of the twist regions. Table 3.8 shows the calculated results for the 32Mbit block shown in figure 3.6. This table assumes the use of a double metal process necessitating the use of local row decoder blocks. Note that table 3.8 contains approximately twice as many wordlines as table 3.4 for the open digitline architecture. The reason for this is that each wordline in the folded array only connects to mbit on alternating digitlines,

whereas each wordline in the open array connects to mbits on every digitline. A folded digitline design therefore needs twice as many wordlines as a comparable open digitline design.

Table 3.8: Folded digitline (local row decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	45μm
Number of local decode strips	Tldec	9
Height of local decode strips	Hldec	93μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	8,512
Number of twist regions	Ttwist	48
Width of twist regions	Wtwist	6μm
Height of 32Mbit block	Height32	6,592μm
Width of 32Mbit block	Width32	6,160μm
Area of 32Mbit block	Area32	40,606,720μm^2

Array efficiency for the 32Mbit memory block from figure 3.6 is again found by dividing the area consumed by functionally addressable mbits by the total die area. For the simplified analysis presented in this thesis, the peripheral circuits are ignored. Array efficiency for the 32Mbit block is therefore given as $Efficiency = (100 \cdot 2^{25} \cdot Pdl \cdot 2 \cdot Pwl8) \div (Area32)$ percent, which yields 59.5 percent for the folded array design example.

The addition of metal3 to the DRAM process enables the use of either a global or hierarchical row decoding scheme--similar to the open digitline analysis. While global row decoding and stitched wordlines achieve the smallest die size, they also place greater demands upon the fabrication process. For a balanced perspective, both approaches were analyzed for the folded digitline architecture. The results of this analysis are presented in tables 3.9 and 3.10. Array efficiency for the 32Mbit memory blocks using global and hierarchical row decoding calculated to 74.0 percent and 70.9 percent respectively.

Table 3.9: Folded digitline (global decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	45μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	200μm
Number of stitch regions	Nst	9
Height of stitch regions	Hst	10μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	8,512
Number of twist regions	Ttwist	48
Width of twist regions	Wtwist	6μm
Height of 32Mbit block	Height32	5,301μm
Width of 32Mbit block	Width32	6,160μm
Area of 32Mbit block	Area32	32,654,160μm^2

Table 3.10: Folded digitline (hier row decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	17
Width of sense amplifiers	Wamp	45μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	190μm
Number of hier decode strips	Nhdec	9
Height of hier decode strips	Hhec	37μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	8,512
Number of twist regions	Ttwist	48
Width of twist regions	Wtwist	6μm
Height of 32Mbit block	Height32	5,534μm
Width of 32Mbit block	Width32	6,160μm
Area of 32Mbit block	Area32	34,089,440μm^2

IV. Advanced Bilevel DRAM Architecture

This section introduces a novel advanced architecture for use on future large scale DRAMs. First, technical objectives for the proposed architecture are discussed. Second, the concept for an advanced array architecture capable of meeting these objectives is developed and physically described. Third, a 32Mbit memory block is conceptually constructed with this new architecture for use in a 256Mbit DRAM. Finally, the results achieved with the new architecture are compared to those obtained for the open digitline and folded digitline architectures from section-III.

Array Architecture Objectives

The open digitline and folded digitline architectures both have distinct advantages and disadvantages. While open digitline architectures achieve smaller array layouts by virtue of using smaller $6F^2$ mbit cells they also suffer from poor noise performance. A relaxed wordline pitch which stems from the $6F^2$ mbit simplifies the task of wordline driver layout. Sense amplifier layout, though, is difficult because the array configuration is inherently half pitch--one sense amplifier for every two digitlines. Folded digitline architectures, on the other hand, have superior signal to noise [10], at the expense of larger, less efficient array layout. Good signal to noise performance stems from the adjacency of true and complement digitlines and the capability to twist these digitline pairs. Sense amplifier layout is simplified because the array configuration is quarter pitch--one sense amplifier for every four digitlines. Wordline driver layout is more difficult since the wordline pitch is effectively reduced in folded architectures.

The main objective of the new array architecture is to combine the advantages both folded and open digitline architectures while avoiding their respective disadvantages. To meet this objective, the architecture needs to include the following features and characteristics:

- Open digitline mbit configuration
- Small $6F^2$ mbit
- Small efficient array layout
- Folded digitline sense amplifier configuration
- Adjacent true and complement digitlines
- Twisted digitline pairs
- Relaxed wordline pitch
- High signal to noise

An underlying goal of the new architecture is to reduce overall die size beyond that obtainable from either the folded or open digitline architectures. A second, yet equally important goal, is to achieve signal to noise performance that meets or approaches that of the folded digitline architecture.

Bilevel Digitline Construction

The bilevel digitline architecture resulted from 256Mbit DRAM design and research activities carried out at Micron Technology, Inc. in Boise, Idaho. The bilevel digitline architecture is an innovation that evolved from a comparative analysis of open and folded digitline architectures. The analysis served as design catalyst, ultimately leading to the creation of a new DRAM array configuration--one that allows the use of $6F^2$ mbits in an otherwise folded digitline array configuration.

$6F^2$ memory cells are a byproduct of crosspoint style (open digitline) array blocks. Crosspoint style array blocks require that every wordline connect to mbit transistors on every digitline, precluding the formation of digitline pairs. Yet, digitline pairs (columns) remain an essential element in folded digitline type operation. Digitline pairs and digitline twisting are important features that provide for good signal to noise performance. The bilevel digitline architecture solves the crosspoint and digitline pair dilemma through vertical integration. Essentially, two open digitline crosspoint array sections are

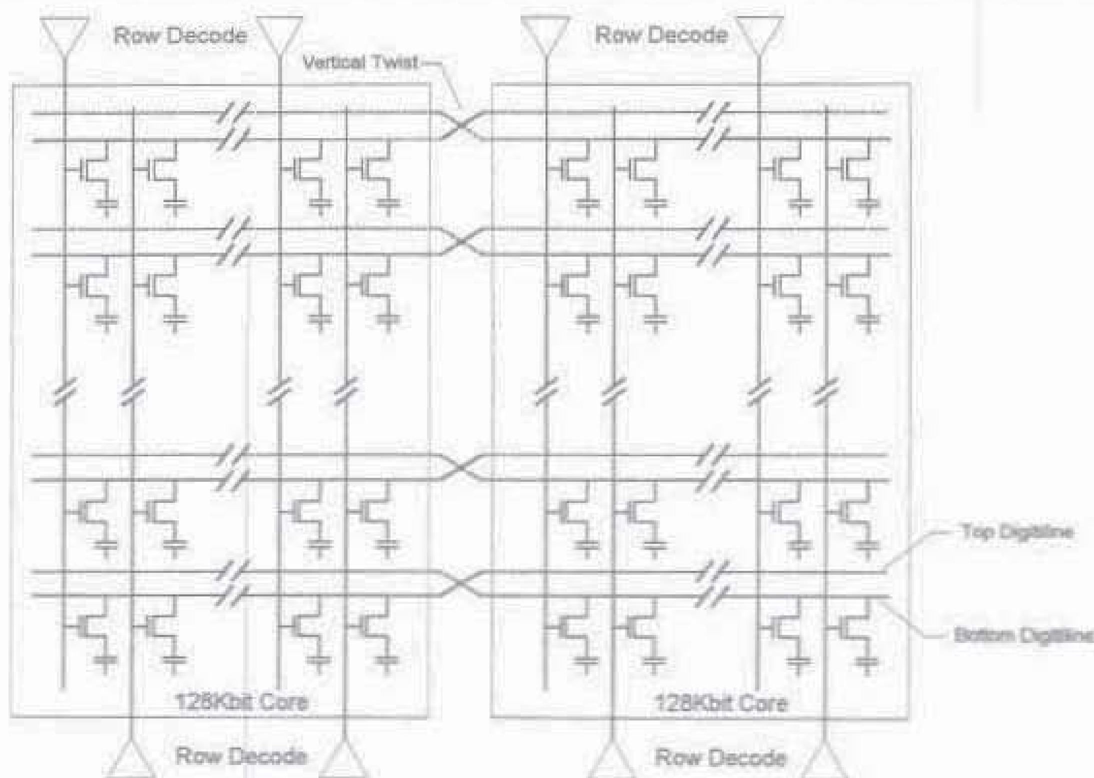


Figure 4.1: Development of bilevel digitline architecture

placed side by side as seen in figure 4.1. Digitlines in one array section are designated as true digitlines while digitlines from the second array section are designated as complement digitlines. An additional conductor is added to the DRAM process to complete formation of the digitline pairs. The added conductor allows digitlines from each array section to route across the other array section--both true and complement digitlines being vertically aligned. At the juncture between each section,

the true and complement signals are vertically twisted. This twisting allows the true digitline to connect to mbits in one array section and the complement digitline to connect to mbits in the other array section. The twisting concept is illustrated in figure 4.2.

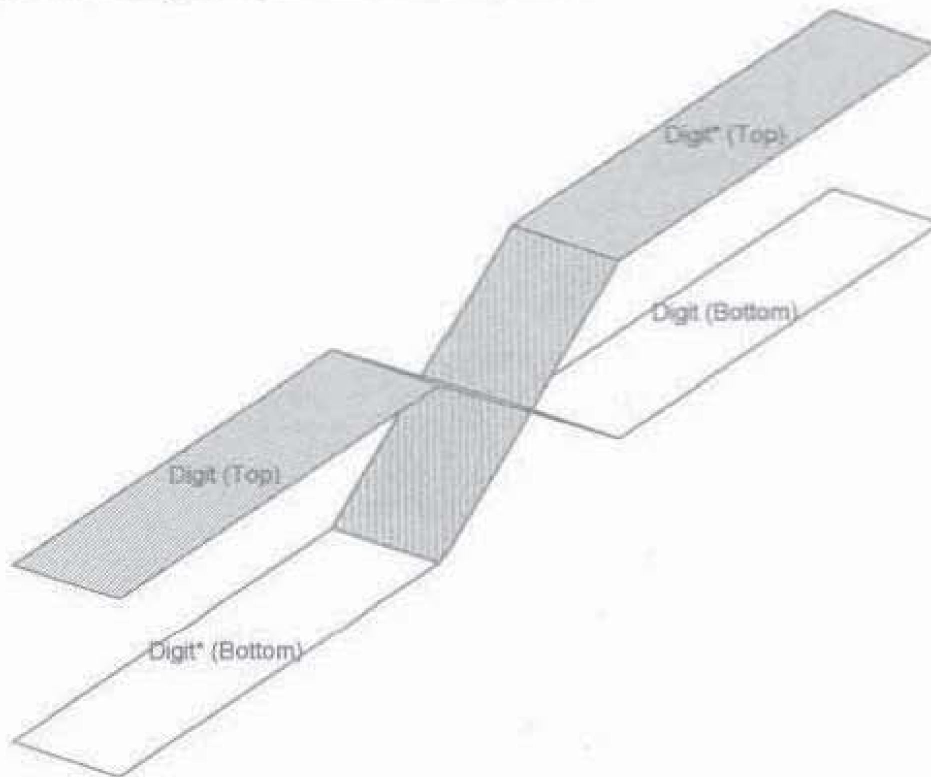


Figure 4.2: Digitline vertical twisting concept

To improve signal to noise characteristics of this design the single twist region is replaced by three twist regions as illustrated in figure 4.3. An added benefit to multiple twist regions is that only half of the digitline pairs actually twist within each region, thus making room in the twist region for each twist to occur. The twist regions are equally spaced at the 25%, 50%, and 75% marks in the overall array. Assuming that even digitline pairs twist at the 50% mark, then odd digitlines twist at the 25% and 75% marks. Each component of a digitline pair, true and complement, spends half of its overall length on the bottom conductor connecting to mbits and half of its length on the top conductor. This characteristic balances the capacitance and the number of mbits associated with each digitline. Furthermore, the triple twisting scheme guarantees that the noise terms are balanced for each digitline, producing excellent signal to noise performance.

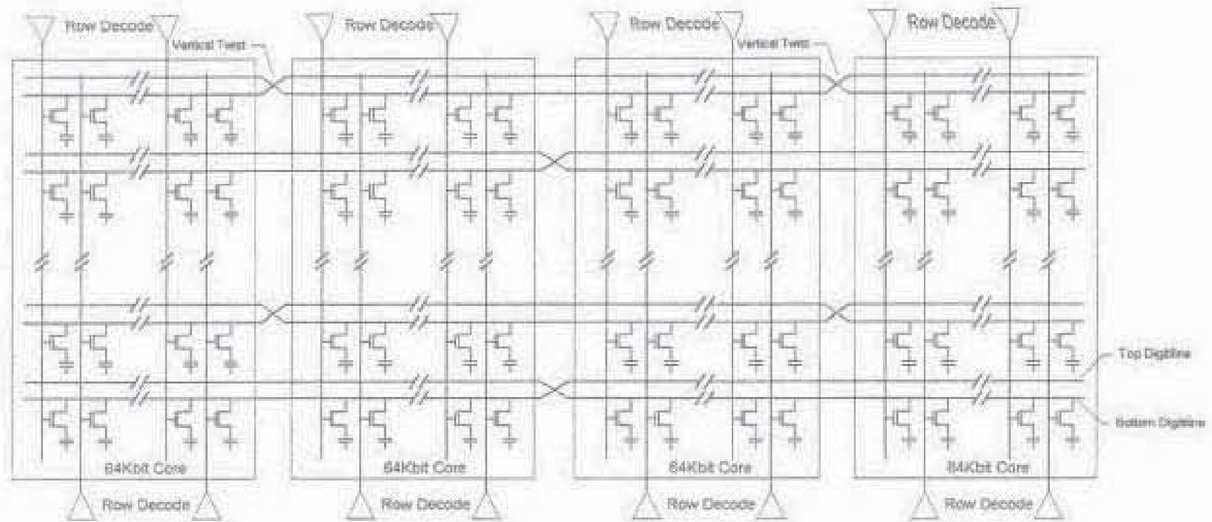


Figure 4.3: Bilevel digitline architecture schematic

A variety of vertical twisting schemes are possible with the bilevel digitline architecture. As shown in figure 4.4, each scheme utilizes conductive layers already present in the DRAM process to complete the twist. Vertical twisting is simplified since only half of the digitlines are involved in a given twist region. The final selection of a twisting scheme is based upon available process technology, yield factors and die size consideration.

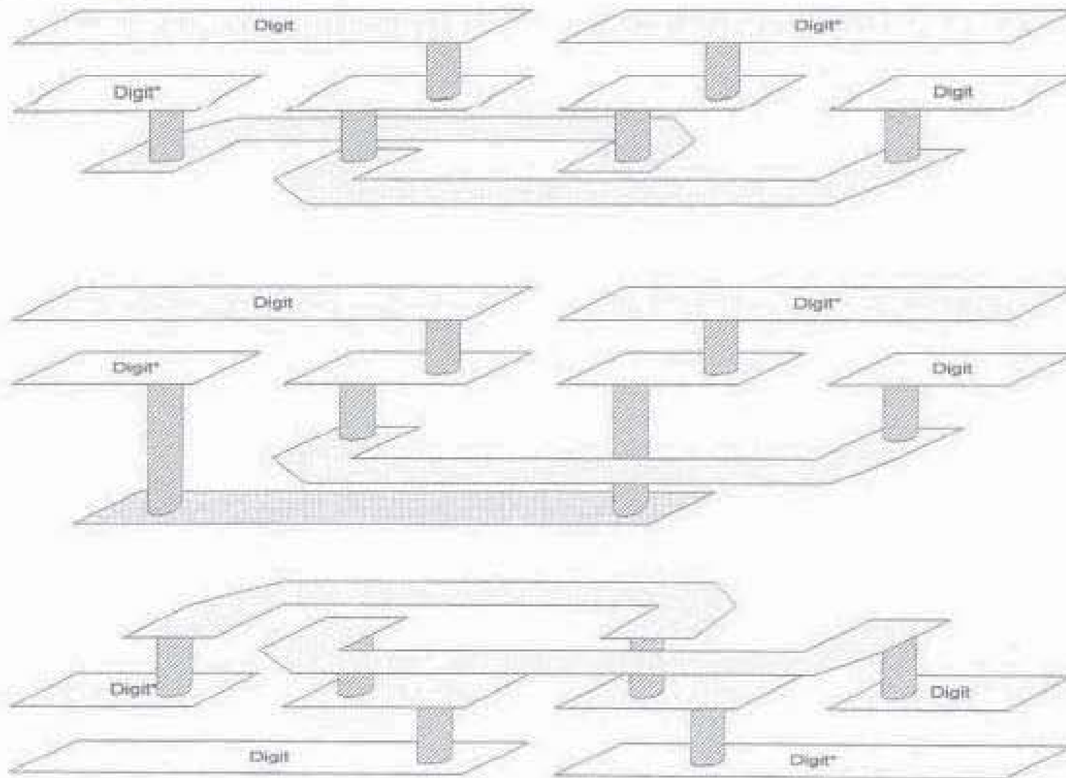


Figure 4.4: Vertical twisting schemes

To further advance the bilevel digitline architecture concept, its $6F^2$ mbit was modified to improve yield. Shown in arrayed form in figure 4.5, the 'plaid' mbit is constructed using long parallel strips of active area vertically separated by traditional field oxide isolation. Wordlines run perpendicular to the active area in straight strips of polysilicon. Plaid mbits are again constructed in pairs that share a common contact to the digitline. Isolation gates (transistors) formed with additional polysilicon strips provide horizontal isolation between mbits. Isolation is obtained from these gates by permanently connecting the isolation gate polysilicon to either a ground or negative potential. The use of isolation gates in this mbit design eliminates one and two dimensional encroachment problems associated with normal isolation processes. Furthermore, many photolithography problems are eliminated from the DRAM process as a result of the straight, simple design of both the active area and polysilicon in the mbit. The 'plaid' designation for this mbit is derived from the similarity between an array of mbits and tartan fabric--very apparent in a color array plot.

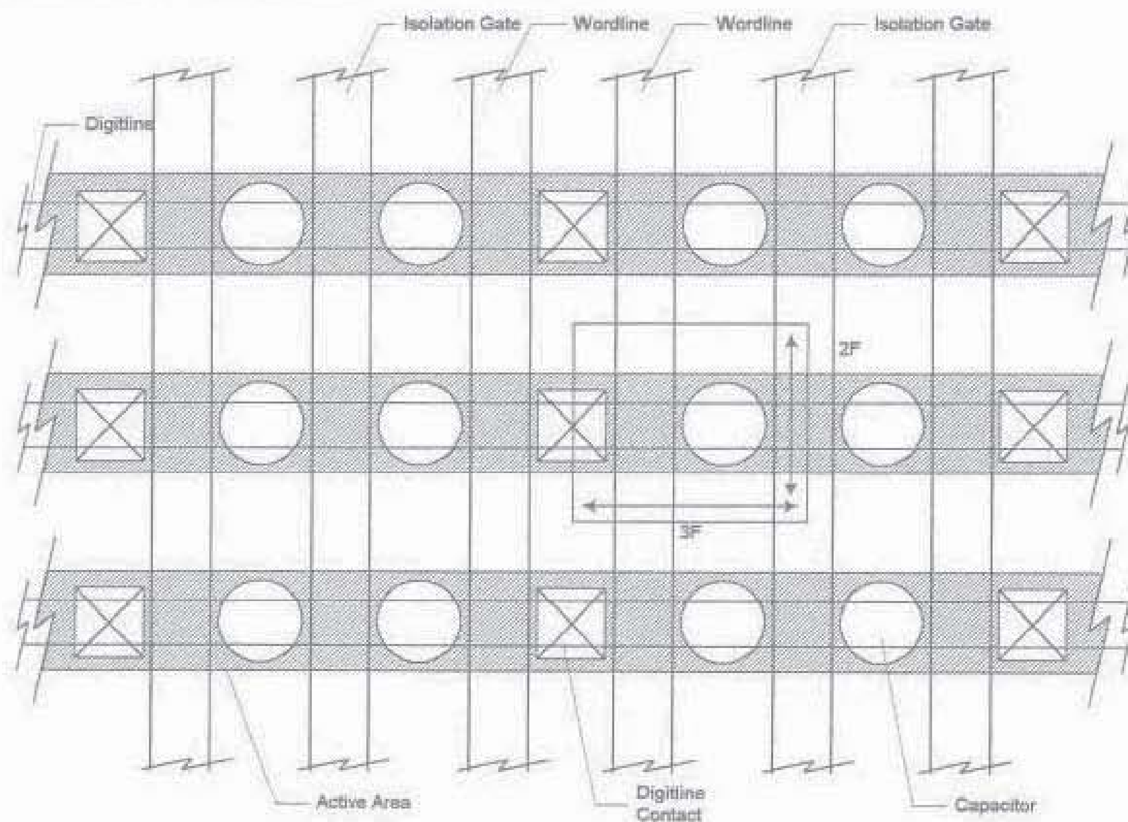


Figure 4.5: Plaid $6F^2$ mbit array

In the bilevel and folded digitline architectures both true and complement digitlines exist in the same array core. Accordingly, the sense amplifier block needs only one sense amplifier for every two digitline pair. For the folded digitline architecture, this yields one sense amplifier for every four

metal1 digitlines--quarter pitch. The bilevel digitline architecture that uses vertical digitline stacking, needs one sense amplifier for every two metal1 digitlines--half pitch. Sense amplifier layout is therefore more difficult for bilevel than folded designs. The triple metal DRAM process needed for bilevel architectures concurrently enables and simplifies sense amplifier layout. Metal1 is used for lower level digitlines and local routing within the sense amplifiers and row decodes. Metal2 is available for upper level digitlines and column select signal routing through the sense amplifiers. Metal3 can therefore be used for column select routing across the arrays and control and power routing through the sense amplifiers. The function of metal2 and metal3 can easily be swapped in the sense amplifier block depending upon layout preferences and design objectives.

Wordline pitch is effectively relaxed for the plaid $6F^2$ mbit used in the bilevel digitline architecture. The mbit is still built using the minimum process feature size of 0.3μm. The relaxed wordline pitch stems from structural differences between a folded digitline mbit and an open digitline or plaid mbit. There are essentially four wordlines running across each folded digitline mbit pair compared to two wordlines that run across each open digitline or plaid mbit pair. Although the plaid mbit is 25% shorter than a folded mbit (3 features versus 4 features), it also has half as many wordlines effectively reducing the wordline pitch. This relaxed wordline pitch makes layout much easier for the wordline drivers and address decode tree. In fact, both odd and even wordlines can be driven from the same row decoder block, thus eliminating half of the row decoder strips in a given array block. This is an important consideration since the tight wordline pitch for folded digitline designs necessitates separate odd and even row decode strips.

Bilevel Digitline Array Architecture

The bilevel digitline array architecture depicted in figure 4.6 is the proposed architecture for tomorrow's large scale DRAM designs. The bilevel architecture is constructed with multiple array cores separated by strips of sense amplifiers and either row decode blocks or wordline stitching regions. Wordline stitching requires a four metal process while row decode blocks can be implemented in a three metal process. The array cores include 262,144 (2^{25}) functionally addressable plaid $6F^2$ mbits arranged in 532 rows and 524 bilevel digitline pairs. The 532 rows consist of 512 actual wordlines, 4 redundant wordlines, and 16 dummy wordlines. There are also 267 isolation gates in each array due to the use of plaid mbits, but since they are accounted for in the wordline pitch, they can be ignored. The 524 bilevel digitline pairs consist of 512 actual digitline pairs, 8 redundant digitline pairs, and 4 dummy digitline pairs. The term digitline pair is used in describing the array core structure since pairing is a natural product of the bilevel architecture. Each digitline pair consists of one digitline on metal1 and a vertically aligned complementary digitline on metal2.

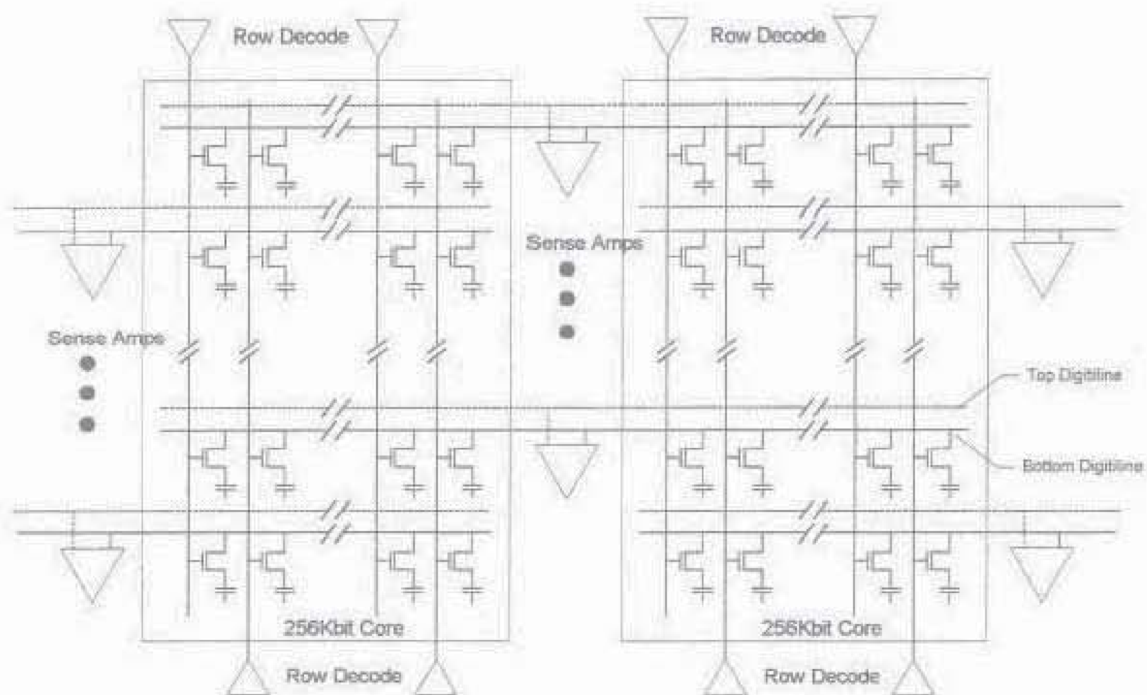


Figure 4.6: Bilevel digitline array schematic

Sense amplifier blocks are placed on both sides of each array core. The sense amplifiers within each block are laid out at half pitch–one sense amplifier for every two metal1 digitlines. Each sense amplifier connects through isolation devices to columns (digitline pairs) from two adjacent array cores. Similar to the folded digitline architecture, odd columns connect on one side of the array core and even columns connect on the other side. Each sense amplifier block is then exclusively connected to either odd or even columns, never both.

Unlike a folded digitline architecture that needs a local row decode block connected to both sides of an array core, the bilevel digitline architecture only needs a local row decode block connected to one side of each core. As stated earlier, the relaxed wordline pitch allows both odd and even rows to be driven from the same local row decoder block. This feature helps make the bilevel digitline architecture more efficient than alternative architectures. A four metal DRAM process allows local row decodes to be replaced by either stitch regions or local wordline drivers. Either approach could substantially reduce die size. The array core also includes the three twist regions that are necessary for the bilevel digitline architecture. The twist region is somewhat larger than that used in the folded digitline architecture, due to the complexity of twisting digitlines vertically. The twist regions again constitute a break in the array structure necessitating the inclusion of dummy wordlines.

As with the open digitline and folded digitline architecture, the bilevel digitline length is limited by power dissipation and minimum cell to digitline capacitance ratio. In the 256Mbit generation the digitlines are again restricted from having connection to more than 256 mbits (128 mbit pairs). The analysis to arrive at this quantity is the same as that for the open digitline architecture, except that the overall digitline capacitance is higher since the digitline runs equal lengths in metal2 and metal1. The capacitance added by the metal2 component is small compared to the metal1 component since metal2 does not connect to mbit transistors. Overall, the digitline capacitance increases by about 25 percent compared to an open digitline. The power dissipated during a read or refresh operation is proportional to the digitline capacitance (Cd), the supply voltage (Vcc), the number of active columns (N), and the refresh period (P) and is given as $Pd = Vccx \cdot (N \cdot Vcc \cdot (Cd + Cc)) \div (2 \cdot P)$ watts. On a

Table 4.1: Active current and power versus bilevel digitline length

Digitline Length	Digitline Capacitance	Active Current	Power Dissipation
128 mbits	128fF	75mA	249mW
256 mbits	256fF	151mA	498mW
512 mbits	513fF	301mA	994mW

256Mbit DRAM in 8K refresh there are 32,768 (2^{15}) active columns during each read, write, or refresh operation. Active array current and power dissipation for a 256Mbit DRAM are given in table 4.1 for a 90nS refresh period (-5 timing) at various digitline lengths. The budget for active array current is limited to 200mA for this 256Mbit design. To meet this budget, the digitline cannot exceed a length of 256 mbits.

Wordline length is again limited by the maximum allowable RC time constant of the wordline. The calculation for bilevel digitline is identical to that performed for open digitline due the similarity of array core design. These results are given in table 3.3 in section-II. Accordingly, the wordline length cannot exceed 512 mbits (512 bilevel digitline pairs) if the wordline time constant is to remain under the required four nanosecond limit.

Layout of various bilevel elements was generated to obtain reasonable estimates of pitch cell size. These size estimates allow overall dimensions for a 32Mbit array block to be calculated. The diagram for a 32Mbit array block using the bilevel digitline architecture is shown in figure 4.7. This block requires a total of 128 256Kbit array cores. The 128 array cores are arranged in 16 rows and 8 columns. Each 4Mbit vertical section consists of 512 wordlines and 8192 bilevel digitline pairs (8192 columns). A total of eight 4Mbit strips are required to form the complete 32Mbit block. Sense amplifier blocks are positioned vertically between each 4Mbit section. Row decode strips are

positioned horizontally between every array core. There are only a total of eight row decode strips needed for the sixteen array cores since each row decode contains wordline drivers for both odd and even rows.

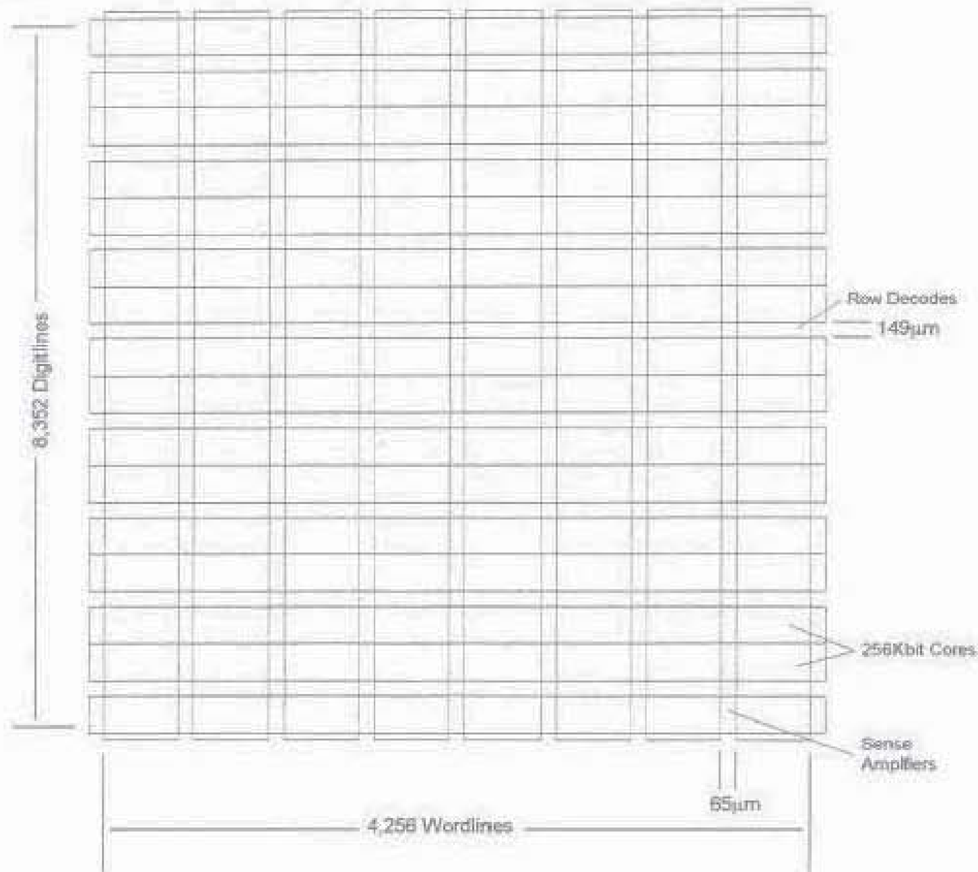


Figure 4.7: Bilevel digitline architecture 32Mbit array block

The 32Mbit array block shown in figure 4.8 includes pitch cell layout estimates. Overall size for the 32Mbit block is found by summing the dimensions for each component. As before, $Height32 = (Tr \cdot Hrdec) + (Tdl \cdot Pdl)$ microns where Tr is the number of bilevel row decoders, Hrdec is the height of each decoder, Tdl is the number of bilevel digitline pairs including redundant and dummy, and Pdl is the digitline pitch. Also, $Width32 = (Tsa \cdot Wamp) + (Twl \cdot Pwl6) + (Ttwist \cdot Wtwist)$ microns, where Tsa is the number of sense amplifier strips, Wamp is the width of the sense amplifiers, Twl is the total number of wordlines including redundant and dummy, Pwl6 is the wordline pitch for the plaid $6F^2$ mbit, Ttwist is the total number of twist regions, and Wtwist is the width of the twist regions. Table 4.2 shows the calculated results for the bilevel 32Mbit block shown in figure 4.7. A triple metal process is assumed in these calculations since it requires the use of local row

decoders. Array efficiency for the bilevel digitline 32Mbit array block is given as

$Efficiency = (100 \cdot 2^{25} \cdot Pdl \cdot 2 \cdot Pwl6) \div (Area32)$ percent, which yields 63.1 percent for this design example.

Table 4.2: Bilevel digitline (local row decode)–32Mbit size calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	9
Width of sense amplifiers	Wamp	65μm
Number of local decode strips	Tldec	8
Height of local decode strips	Hldec	149μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	4,256
Number of twist regions	Ttwist	24
Width of twist regions	Wtwist	9μm
Height of 32Mbit block	Height32	6,203μm
Width of 32Mbit block	Width32	4,632μm
Area of 32Mbit block	Area32	28,732,296μm^2

With metal4 added to the bilevel DRAM process, the local row decoder scheme can be replaced by a global or hierarchical row decoder scheme. The addition of a fourth metal to the DRAM process places even greater demands upon process technologists. Regardless, an analysis of 32Mbit array block size was performed assuming the availability of metal4. The results of the analysis are shown in tables 4.3 and 4.4 for the global and hierarchical row decode schemes. Array efficiency for the 32Mbit memory block using global and hierarchical row decoding calculates to 74.5 percent and 72.5 percent respectively.

Table 4.3: Bilevel Digitline (Global Decode)–32Mbit Size Calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	9
Width of sense amplifiers	Wamp	65μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	200μm
Number of stitch regions	Nst	4
Height of stitch regions	Hst	10μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	4,256
Number of twist regions	Ttwist	24
Width of twist regions	Wtwist	9μm
Height of 32Mbit block	Height32	5,251μm
Width of 32Mbit block	Width32	4,632μm
Area of 32Mbit block	Area32	24,322,632μm^2

Table 4.4: Bilevel Digitline (Hier Row Decode)–32Mbit Size Calculations

Description	Parameter	Size
Number of sense amplifier strips	Tsa	9
Width of sense amplifiers	Wamp	65μm
Number of global decode strips	Tgdec	1
Height of global decode strips	Hgdec	190μm
Number of hier decode strips	Nhdec	4
Height of hier decode strips	Hhec	48μm
Number of digitlines	Tdl	8,352
Number of wordlines	Twl	4,256
Number of twist regions	Ttwist	24
Width of twist regions	Wtwist	9μm
Height of 32Mbit block	Height32	5,393μm
Width of 32Mbit block	Width32	4,632μm
Area of 32Mbit block	Area32	24,980,376μm^2

Architectural Comparison

Although a straight comparison of DRAM architectures might appear simple, in actual fact it is a very complicated problem. Profit remains the critical test of architectural efficiency and is the true basis for comparison. This in turn requires accurate yield and cost estimates for each alternative. Without these estimates and a thorough understanding of process capabilities, conclusions are elusive and the exercise remains academic. The data necessary to perform the analysis and render a decision also varies from manufacturer to manufacturer. Accordingly, a conclusive comparison of the various array architectures is beyond the scope of this thesis. Rather, the architectures will be compared in light of the available data. To better facilitate a comparison, the 32Mbit array block size data from section-III and section-IV is summarized in table 4.5 for the open digitline, folded digitline, and bilevel digitline architectures.

Table 4.5: 32Mbit size calculations summary

Architecture	Row Decode	Metals	32Mbit Area (μm^2)	Efficiency (%)
Open Digit	Global	3	29,944,350	60.5
Open Digit	Hier	3	32,429,565	55.9
Folded Digit	Local	2	40,606,720	59.5
Folded Digit	Global	3	32,654,160	74.0
Folded Digit	Hier	3	34,089,440	70.9
Bilevel Digit	Local	3	28,732,296	63.1
Bilevel Digit	Global	4	24,322,632	74.5
Bilevel Digit	Hier	4	24,980,376	72.5

From table 4.5 it can be concluded that overall die size (32Mbit Area) is a better metric for comparison than array efficiency. For instance, the triple metal folded digitline design using hierarchical row decodes has an area of 34,089,440μm^2 and an efficiency of 70.9%. The triple metal bilevel digitline design with local row decodes has an efficiency of only 63.1%, but an overall area of 28,732,296 μm^2. Array efficiency for the folded digitline is higher, but this is misleading, since the folded digitline yields a die that is 18.6% larger for the same number of conductors. Table 4.5 also illustrates that the bilevel digitline architecture always yields that smallest die area, regardless of the configuration. The smallest folded digitline design at 32,654,160μm^2 and the smallest open digitline design at 29,944,350μm^2 are still larger than the largest bilevel digitline design at 28,732,296μm^2. Also apparent is that the bilevel and open digitline architectures both need at least three conductors in their construction. The folded digitline architecture still has a viable design option using only two conductors. The penalty to two conductors is of course a much larger die size--a full 41% larger than the triple metal bilevel digitline design.

V. Conclusion

This thesis described a novel bilevel digitline architecture for use on advanced DRAM designs. The bilevel digitline architecture achieves significant reductions in die size while maintaining the high signal to noise performance of traditional folded digitline architectures. The bilevel digitline uses vertically stacked digitline pairs connected to arrays of $6F^2$ memory cells. Vertical digitline twisting at three array locations ensures balanced noise cancellation and equalizes the quantity of memory cells contacting each digitline. DRAM die size reduction results primarily from the use of smaller $6F^2$ memory cells in crosspoint style arrays and secondarily from efficient pitch cell utilization. Overall, the bilevel digitline approach presented in this thesis combines the best characteristics of both folded and open digitline architectures into an efficient new DRAM architecture. Future work will need to include the design and construction of prototype DRAMs employing the bilevel digitline architecture. Prototypes will enable a thorough examination of noise and yield issues with working silicon, thus permitting an unbiased comparison of the bilevel digitline architecture to the traditional folded digitline architecture.

BIBLIOGRAPHY

[1] R. Kraus, et.al., "Optimized Sensing Scheme of DRAM's," in IEEE Journal of Solid-State Circuits, VOL. 24, NO. 4, pp. 895-899, August 1989.

[2] E. Adler, et.al., "The Evolution of IBM CMOS DRAM Technology," in IBM Journal of Research and Development, VOL. 39, NO. 1/2, pp. 167-188, March 1995.

[3] N. C. Lu, et.al., "Half-Vdd Bit-Line Sensing Scheme in CMOS DRAM's," in IEEE Journal of Solid-State Circuits, VOL. SC-19, NO. 4, pp. 451, August 1984.

[4] R. Sarpeshkar, et.al., "Mismatch Sensitiviy of a Simultaneously Latched CMOS Sense Amplifier," in IEEE Journal of Solid-State Circuits, VOL. 26, NO. 10, pp. 1413-1422, October 1991.

[5] R. Kraus, "Analysis and Reduction of Sense-Amplifier Offset," in IEEE Journal of Solid-State Circuits, VOL. 24, NO. 4, pp. 1028-1033, August 1989.

[6] K. Itoh, "Trends in Megabit DRAM Circuit Design," in IEEE Journal of Solid-State Circuits, VOL. 25, NO. 3, pp. 778-791, June 1990.

[7] D. Takashima, et.al., "Open/Folded Bit-Line Arrangement for Ultra-High-Density DRAM's," in IEEE Journal of Solid-State Circuits, VOL. 29, NO. 4, pp. 539-542, April 1994.

[8] H. Hidaka, et.al., "A Divided/Shared Bit-Line Sensing Scheme for ULSI DRAM Cores," in IEEE Journal of Solid-State Circuits, VOL. 26, NO. 4, pp. 473-477, April 1991.

[9] M. Aoki, et.al., "A 60-ns 16-Mbit CMOS DRAM with a Transposed Data-Line Structure," in IEEE Journal of Solid-State Circuits, VOL. 23, NO. 5, pp. 1113-1119, October 1988.

[10] T. Yoshihara, et.al., "A Twisted Bitline Technique For Multi-Mb DRAMs," in IEEE International Solid-State Circuits Conference, pp. 238-239, 1988.

[11] Y. Oowaki, et.al., "A 33-ns 64Mb DRAM," in IEEE Journal of Solid-State Circuits, VOL. 26, NO. 11, pp. 1498-1505, November 1991.

[12] M. Inoue, et.al., "A 16Mb DRAM with An Open Bit-Line Architecture," in 1988 IEEE ISSCC Digest of Techical Papers, pp. 246-247.

[13] Y. Kubota, et.al., "Alternatively-Activated Open Bitline Technique for High Density DRAM's," in IEICE Trans. Electron., VOL. E75-C, NO. 10, pp. 1259-1266, October 1992.

[14] T. Hamada, et.al., "A Split-Level Diagonal Bit-Line (SLDB) Stacked Capacitor Cell for 256Mb Drams," in 1992 IEDM Technical Digest, pp. 799-802.

[15] T. Mohihara, et.al., "Disk-Shaped Capacitor Cell for 256Mb Dynamic Random-Acess Memory," in Japan Journal of Applied Physics, VOL. 33, Part 1, NO. 8, pp. 4570-4575, August 1994.

[16] J.H. Ahn, et.al., "Micro Villus Patterning (MVP) Technology for 256Mb DRAM Stack Cell," in 1992 Symposium on VLSI Tech. Digest of Technical Papers, pp. 12-13.

[17] K. Sagara, et.al., "Recessed Memory Array Technology for a Double Cylindrical Stacked Capacitor Cell of 256M DRAM," in IEICE Trans. Electron., VOL. E75-C, NO. 11, pp. 1313-1322, November 1992.

[18] S. Ohya, "Semiconductor Memory Device Having Stacked-Type Capacitor of Large Capacitance," United States Patent, Number 5,298,775, March 29, 1994.

[19] M. Sakao, et.al., "A Capacitor-Over-Bit-Line (COB) Cell with Hemispherical -Grain Storage Node for 64Mb DRAMs," in 1990 IEDM Technical Digest, pp. 655-658.

[20] G. Bronner, et.al., "A Fully Planarized 0.25mm CMOS Technology for 256Mbit DRAM and Beyond," in 1995 Symposium on VLSI Tech. Digest of Technical Papers, pp. 15-16.

[21] M. Asakura, et.al., "A 34ns 256Mb DRAM with Boosted Sense-Ground Scheme," in 1994 IEEE ISSCC Digest of Techical Papers, pp. 140-141.

[22] T. Ooishi, et.al., "An Automatic Temperature Compensation of Internal Sense Ground for Sub-Quarter Micron DRAMs," in 1994 Symposium on VLSI Circuits Digest of Technical Papers, pp. 77-78.

[23] T. Nishihara, et.al., "A Buried Capacitor Cell with Bonded SOI for 256-Mbit and 1-Gbit DRAMs," in Solid State Technology, pp. 89-94, June 1994.

[24] F. Morishita, et.al., "Leakage Mechanism due to Floating Body and Countermeasure on Dynamic Retention Mode of SOI-DRAM," in 1995 Symposium on VLSI Tech. Digest of Technical Papers, pp. 141-142.

[25] H. Komiya, et.al., "Future Technological and Economic Prospects for VLSI," in IEICE Trans. Electron., VOL. E76-C, NO. 11, pp. 1555-1563, November 1993.

[26] K. Noda, et.al., "A Boosted Dual Word-line Decoding Scheme for 256Mb DRAMs," in 1992 Symposium on VLSI Circuits Digest of Technical Papers, pp. 112-113.

[27] H. Shichijo, "DRAM Technology Trend and Prospect," in 1991 VLSITSA Digest of Technical Papers, pp. 349-353.

[28] M. Inoue, et.al., "A 16-Mbit DRAM with a Relaxed Sense-Amplifier-Pitch Open-Bit-Line Architecture," in IEEE Journal of Solid-State Circuits, VOL. 23, NO. 5, pp. 1104-1112, Oct. 1988.

[29] Y. Nakagome, et.al., "The Impact of Data-Line Interference Noise on DRAM Scaling," in IEEE Journal of Solid-State Circuits, VOL. 23, NO. 5, pp. 1120-1127, October 1988.