AN SRAM SYSTEM BASED ON A REDUCED-AREA FOUR-TRANSISTOR CMOS SRAM CELL

by

Stephan Joseph De Beer

Submitted in partial fulfillment of the requirements for the degree

Master of Engineering (Electronic Engineering)

in the

Faculty of Engineering

UNIVERSITY OF PRETORIA

July 2002

ABSTRACT

Keywords: SRAM systems, reduced-area SRAM cell, SRAM static noise margins, threshold voltage reference, constant transconductance biasing, low-impedance driver, analogue design, clamped bit line latching current sense amplifier, current transporting circuit, asynchronous control.

The traditional method of implementing SRAM in CMOS is via a six-transistor cell and five routing lines. If the number of transistors and the number of wires could be reduced, the packing density of the memory cells could be increased, and the area reduced.

This document describes the design of an SRAM system based on a new four-transistor SRAM cell. The primary design goal was to create a functional system, so that the relationship between reduced cell area and a potentially reduced system area could be investigated.

A new write method and associated array structure has been used, and the design of the system parameters was accomplished using static noise margin theory. The power dissipation and percentage reduction in cell area have been improved over previous designs.

The circuits to achieve the access to the cell have been designed and simulated. These include low-impedance driver circuits, that allow the power supply of the cell's devices to be individually modified to read and write the cell, and a current sense amplifier system to convert the output current to a digital voltage. These circuits allow complete and accurate control to be achieved, but a price is paid for the complexity in terms of layout area. The SRAM system emulates a standard SRAM, and could therefore be used to replace current SRAM implementations.

The design was simulated on a system level, and found to operate correctly. Although it is outperformed by its six-transistor cell counterpart in terms of power dissipation, speed and layout area, the groundwork for defining further research and improving the characteristics of further designs has been laid.

UITTREKSEL

Sleutelwoorde: Statiese lees skryf geheue (SLSG) stelsels, verminderde-area SLSG sel, SLSG statiese ruisgrense, drempelspanningsverwysing, konstante transkonduktansie-voorspanning, lae-impedansie drywer, analoog ontwerp, geklampte bislyn gegrendelde stroom-monster versterker, stroom oordrag netwerk, asinkrone beheer.

Die tradisionele implementering van statiese geheue in CMOS geskied deur middel van 'n ses-transistor geheuesel wat vyf elektriese verbindings het. Indien die aantal transistors en lyne verminder kan word, sou dit moontlik wees om die pakkingsdigtheid van die geheueselle te verhoog en die oppervlakte te verklein.

Hierdie dokument beskryf die ontwerp van 'n stelsel wat op 'n nuwe vier-transistor sel gebaseer is. Die primêre ontwerpsdoel was om 'n funksionele stelsel te skep sodat die verband tussen verminderde seloppervlakte en 'n potensieël verminderde stelseloppervlakte ondersoek kon word.

'n Nuwe metode om die sel te skryf binne 'n raamwerk van 'n matriks-struktuur is gebruik, en die ontwerp van stelselparameters is deur middel van statiese ruisgrensanalise gedoen. Die drywingdissipasie en die persentasie vermindering in seloppervlakte is verbeter in vergelyking met vorige ontwerpe.

Stroombane wat nodig is om die sel te beheer is ontwerp en gesimuleer. Dit sluit lae-impedansie drywers in, wat toelaat dat die toevoerspanning van die sel se nodes onafhanklik varieër kan word vir die doeleindes van lees en skryf. 'n Stroomsensor is ook ontwerp om die uitsetstroom van die sel na 'n digitale spanning te verander. Hierdie stroombane laat korrekte en volledige beheer toe, maar die prys word in terme van oppervlakte betaal. Na buite lyk die stelsel soos enige standaard statiese geheue stelsel, en kan dus gebruik word om huidige implementerings te vervang.

Die ontwerp is op stelselvlak gesimuleer en funksioneer korrek. Dit kompeteer egter nie met 'n ekwivalente ses-transistor stelsel in terme van drywingdissipasie, spoed en oppervlakte nie. Dit het egter die beginsels vir opvolgende navorsing en 'n volgende ontwerpsiterasie gedefinieër.

CONTENTS

1. Introduction 1

1.1 Summary of Related Work 2

1.2 Contributions of this Study 3

1.3 Dissertation Outline 4

2. Four-Transistor SRAM Cell 5

2.1 Introduction 5

2.2 Background 5

2.2.1 Six-Transistor SRAM Cell 5

2.2.2 Four-Transistor SRAM Cell 6

2.3 Problem Definition 11

2.4 Cell Operation 14

2.4.1 Cell in Retention Mode 14

2.4.2 The Read Cycle 15

2.4.3 The Write Cycle 17

2.4.4 Limitations of the Write Cycle 21

2.5 Alternative Write Cycle 24

2.6 Alternative Array Structure 26

2.7 Cell Design 28

2.7.1 Device Ratio 29

2.7.2 Noise Margins of Logic Circuits 30

2.7.3 Static-Noise Margin of the Four-Transistor SRAM Cell 31

2.7.4 Design of Voltage Deviations 33

2.7.5 Effects of Temperature 37

2.8 Transient Simulations 38

2.9 Experimental Verification 42

2.10 Six-Transistor SRAM Cell Comparison 48

2.10.1 Noise Margin of the Six-Transistor SRAM Cell 49

2.10.2 Six-Transistor SRAM Cell Design 51

2.10.3 Transient Simulations 53

2.11 Comparison between the Four- and Six-Transistor Cells 54

2.11.1 Speed 54

2.11.2 Power Dissipation 55

2.11.3 Cell Area 55

2.12 Conclusion 58

3. Source Driver Circuits 59

3.1 Introduction 59

3.2 Fundamental Principles 59

3.3 *D*/*O*-Line Driver Circuit 61

3.3.1 Overview 61

3.3.2 Line Capacitance 62

3.3.3 Currents 64

3.3.4 Switching Circuit 64

3.3.5 Voltage Reference Circuit 68

3.3.6 Low-Impedance Driver Circuit 74

3.3.7 Operational Amplifier 79

3.3.8 *D/O*-Line Driver Circuit Simulation 82

3.4 *RW*-Line Driver 85

3.4.1 Overview 85

3.4.2 Line Capacitance 86

3.4.3 Currents 86

3.4.4 Switching Circuit 87

3.4.5 Voltage Reference Circuit 89

3.4.6 Low-Impedance Driver Circuit 91

3.4.7 Operational Amplifier 92

3.4.8 *RW*-Line Driver Simulation 94

3.5 *CL*-Line Driver 96

3.5.1 Overview 96

3.5.2 Line Capacitance 96

3.5.3 Currents 97

3.5.4 Circuit Design 97

3.5.5 *CL*-Line Driver Simulation 98

3.6 Cell Control Simulation 99

3.7 Layouts 103

3.8 Conclusion 107

4. Current Sense Amplifier ... 108

4.1 Introduction ... 108

4.2 Sensing SRAM Cells ... 108

4.2.1 Voltage Sense Amplifier System ... 108

4.2.2 Current Sense Amplifier Systems ... 109

4.3 Current Sense Amplifier Design ... 116

4.3.1 Reference Current ... 117

4.3.2 Current Switching ... 118

4.3.3 Current Conveyor and Loads ... 121

4.3.4 Clamped Bit Line Sense Amplifier ... 123

4.4 Simulation ... 124

4.5 Layout ... 127

4.6 Conclusion ... 130

5. SRAM System Based on the Four-Transistor Cell ... 131

5.1 Introduction ... 131

5.2 SRAM Interface and Cycles ... 132

5.2.1 SRAM Control Signals ... 132

5.2.2 SRAM Read Operation ... 132

5.2.3 SRAM Write Operation ... 133

5.2.4 SRAM Timing ... 134

5.3 The SRAM System ... 135

5.3.1 Global System ... 135

5.3.2 Sense Amplifier 138

5.3.3 Memory Bank 139

5.3.4 Read Control 141

5.3.5 Write Control 145

5.4 Simulation 148

5.4.1 Timing Specifications 151

5.4.2 Power Dissipation 153

5.5 Layout 157

5.6 Comparison to a Six-Transistor System 160

5.7 Conclusion 162

6. Conclusion 164

6.1 What was Given? 164

6.2 What was the Aim? 164

6.3 What has been Accomplished ? 164

6.4 What can be Learned from This? 165

6.5 What is the Next Step? 166

References 168

Addendum A: C-Code for Noise Margin Analysis 172

A.1 Four-Transistor Cell Noise Margin Analysis 172

A.2 Six-Transistor Cell Noise Margin Analysis 177

Addendum B: Layout Legend 181

Addendum C: Circuit Diagrams 182

1. INTRODUCTION

Most digital data processing systems require some form of temporary data storage mechanism. As the size and speed of these systems increase, so does the requirement for larger memory space. This creates the need for very small area memory cells, so that silicon chip area, and therefore costs, can be kept as low as possible. Most advances in this field have taken place on the level of semiconductor processing technologies that were designed or adapted to create small memory cells. Very small densely packed memories can be created using dedicated processes (single-transistor dynamic RAM), or adding extra steps to standard processes (high ohmic load devices). Due to costs involved, these methods are only economically viable for the manufacturing of dedicated memory chips. Recently however, the need for high-density memories is coupled to the requirement that they be suitable for use in embedded systems, where processing circuits and memory circuits are manufactured on the same chip. Here dedicated processing technologies are usually too expensive, because they are not applied to the total chip area. Embedded memories therefore need to be based on a standard process (typically CMOS). In order to meet the requirements of smaller cell sizes, circuit topologies rather than processing technology need to be addressed and optimised.

The typical implementation for embedded temporary storage is the six-transistor SRAM cell. The memory is based on a cross-coupled inverter pair and two access transistors through which the cell can be read and written. The cell area could be reduced if it were possible to remove some of the devices and still retain satisfactory operation.

This dissertation presents a memory system utilising a smaller four-transistor SRAM cell where the access transistors have been omitted to save area [1], [2]. The system is implemented in a standard CMOS process, and is therefore usable in embedded applications. When compared to its six-transistor counterpart, the area per cell for equivalent performance is reduced. More complex peripheral circuits are however required to create a system that has the same external interface as a standard memory system.

The global aim of the work leading up to this dissertation was to create a functional system using the four-transistor SRAM cell, so that it could be investigated if the gain present in the reduced cell area could be transformed into an overall gain in system area. Other characteristics must also be investigated. The gain could then be used to add economical value to embedded SRAM systems.

1.1 SUMMARY OF RELATED WORK

There are several existing proposals in the field of low area SRAM systems, although the successful operation of some of them relies, in some form or another, on non-standard process technologies.

A good example is the four-transistor resistive load memory [3]. The structure is identical to the six-transistor cell except that the PMOS load devices are replaced by resistors. This implementation is well suited to early NMOS processes. A drawback of this system is undoubtedly the potentially high static current dissipation, but the absence of a second type of device in the memory array does produce a significant area advantage. A cell size of 7.4μm x 12.8μm in a 1.3μm process is reported [3]. This can be compared to a cell size of 15.0μm x 20.7μm in a 1.5μm process for the six-transistor cell [4]. A significant area advantage (69.5% reduction) can be seen, even though some area advantage will be inherent due to the better process used in [3].

A different approach is what is commonly termed a single-ended SRAM. A five-transistor cell, created by omitting one access transistor, is used [5]. The area advantage is present in the use of the single access transistor and bit line. The absence of the differential signals does however create some speed disadvantages.

More recently a transistorless architecture was proposed [6]. A tunnel switch diode (TSD), which is a stacking of p-type semiconductor, n-type semiconductor, insulator and metal, and has a thyristor-like current-voltage characteristic, can be used as a bistable element. By controlling the voltage across the device it may be placed in one of the two states. Reading is accomplished by monitoring the current at a nominal voltage. Very dense arrays can be manufactured using special processing steps, but the TSD memory array can be made to function in standard CMOS processes by increasing the cell size. Essentially, the minimum bit size is dependent on the minimum geometry widths allowable in a process.

A very recent publication describes a four-transistor SRAM cell where the PMOS load devices have been omitted [7]. The access transistors are PMOS. The leakage currents of the driver and access transistors are utilised to keep one of the nodes "high", by ensuring that the leakage through the PMOS from the bit line into the "high" node is higher than the leakage to ground through the NMOS. This has to be ensured for all conditions, including frequent writes, which tend to lower the average voltage on a bit line and cause the leakage into the node to decrease. If the ratio between the leakage into the node and the leakage from the node can be kept in the order of 100, the cell is adequately reliable. Problems in maintaining this ratio can occur at low temperature and require special circuits to ensure a sufficient off-state current ratio. A 35% reduction in cell size compared to a standard six-transistor cell implemented in the same process, is reported. This SRAM cell does however require an extra processing step, in that the threshold voltage of the cell NMOS-devices needs to be raised by about 0.3V [8]. This is necessary to create the required off-state current ratio.

1.2 CONTRIBUTIONS OF THIS STUDY

The research discussed in this dissertation aims to contribute to knowledge in the field of alternative static memory architectures, where the main criteria is reduced area. The viability of a novel memory architecture is evaluated by implementing a complete system that can be compared to standard six-transistor cell implementations. This allows the apparent gain in value to be verified and put to good use. Some analyses of the four-transistor SRAM cell operation are also presented, which aid to create better understanding of its operation. A different method of writing the cell, together with a new array structure, as well as a design method based on a noise margin analysis, is proposed.

1.3 DISSERTATION OUTLINE

Chapter 1 A brief introduction and perspective.

Chapter 2 A discussion of the operation of the four-transistor SRAM cell and an investigation and evaluation of possible array architectures.

Chapter 3 An outline of the design and simulation of the voltage references required for driving the word- and bit lines of the four-transistor SRAM system.

Chapter 4 A description of the design and simulation of the current sense amplifier required to read the data stored in the cell.

Chapter 5 An overview of the complete SRAM system together with some simulations, and a comparison to a six-transistor SRAM cell system.

Chapter 6 A concluding summary.

2. FOUR-TRANSISTOR SRAM CELL

2.1 INTRODUCTION

The foundation of the system designed in this dissertation is the four-transistor SRAM cell proposed by Seevinck [1] and evaluated by Joubert, Seevinck and Du Plessis [2]. In this chapter various aspects of this cell will be discussed. A design method based on noise margin analysis, by which the cell and other circuit parameters relating to it can be designed for any given CMOS process, is presented. To begin with, a brief outline of the basic operation of the six-transistor and proposed four-transistor cell, as discussed in [2], is given.

2.2 BACKGROUND

2.2.1 Six-Transistor SRAM Cell [9]

A standard six-transistor SRAM cell is shown in Figure 2.1. It consists of a bistable element in the form of a pair of cross-coupled inverters (*M1* - *M4*), and an access mechanism in the form of the two devices *M5* and *M6*.

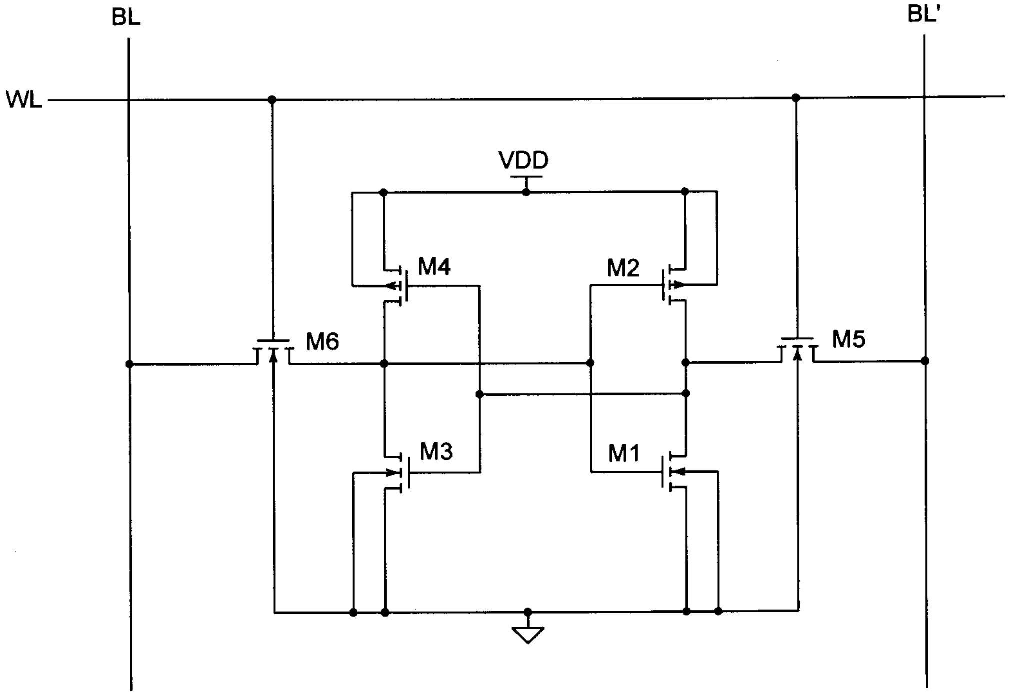


Figure 2.1 Standard six-transistor CMOS SRAM cell.

To read the state of the cell, the bit lines *BL* and *BL'* are precharged typically to close to *VDD*, and the word line *WL* is activated. This turns on the access transistors, and according to the state stored in the cell, one of the bit lines will be discharged. The differential voltage or the differential current on the bit lines may be sensed to determine the state of the cell.

When the cell needs to be written, one bit line is driven "high" and the other "low", depending on the value to be written to the cell, and the word line is activated. This forces the internal inverter nodes in the direction of the bit line voltages and the state on the bit lines is written to the cell.

A design issue for the six-transistor cell is the fact that the access devices may not be too strong, else the state of the cell may be modified during the initial read phase, where both bit lines are at a high potential. This constitutes a highly undesirable situation. To overcome this the ratio between the *W/L* of the driver transistor and that of the access transistor is typically designed to be in the order of 2 [10]. This requires either the access transistor to be rather long or the driver transistor to be rather wide, significantly increasing the cell area. The access devices also only provide access to the cell and do not contribute to its memory function, which resides purely in the cross-coupled inverter pair. The motivation behind the four-transistor SRAM cell is that if it were possible to devise a method of accessing the cross-coupled inverter pair other than using the access transistors, they could be omitted and the cell area reduced.

2.2.2 Four-Transistor SRAM Cell

Figure 2.2 depicts the four-transistor SRAM cell. As can be seen, the access transistors are no longer present. The cross-coupled inverter pair is retained, with one modification. The sources of the devices are no longer connected to the power supplies but are used as control nodes to achieve access to the cell.

In retention mode, that is when the cell is not being read or written, all these sources are still connected to the respective power supplies. The memory function of the cross-coupled inverter pair is therefore still present and unchanged. It is during the read and write operations that the voltages of the transistor sources are varied.

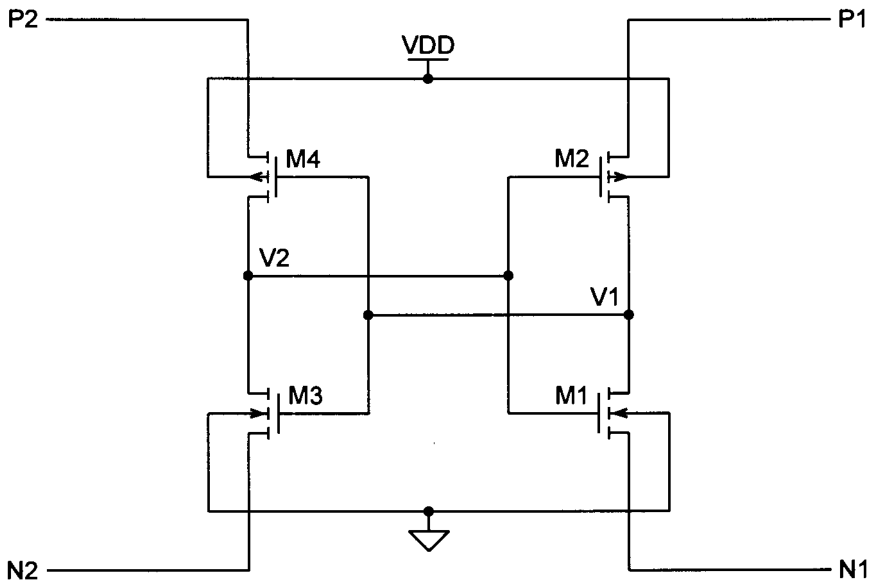


Figure 2.2 Four-transistor CMOS SRAM cell.

Reading the Four-Transistor SRAM Cell

The cell can be read by varying any of the four possible nodes (*N1*, *N2*, *P1*, *P2*) away from the supply voltage and beyond the threshold voltage of the devices, and then monitoring the current in the opposite inverter. For example, consider node *V1* to be "low" and therefore node *V2* to be "high". Devices *M1* and *M4* are turned on in the linear region and devices *M2* and *M3* are in cutoff. If the voltage at node *N1* is raised, the voltage of the node *V1* will track that of *N1* because *M1* is in a low-impedance mode. If the voltage deviation is larger than the threshold voltage of *M3*, then this device will be driven into saturation mode and therefore conduct a current. This current may be sensed either at node *N2* or *P2*. If however, node *V1* is "high" and node *V2* therefore "low", then *M1* is in cutoff. In this case, raising the voltage at node *N1* cannot turn *M1* on, so no conditions in the other parts of the circuit are changed. A current sensor attached to nodes *N2* or *P2* would therefore sense no current. The presence of a current is defined as one logic state and the absence of a current as the other state. As long as the voltage deviation is not large enough to force the tracking internal node *V1* beyond the trigger voltage of the other inverter, the state of the cell is not affected by the read.

Writing the Four-Transistor SRAM Cell

If an internal node *V1* or *V2* is driven beyond the trigger voltage of the opposite inverter, the state of the cell can be changed. The usual scheme of writing some cells in a large array is to apply the data to all cells and then to select which cells to write. The selection is done by reducing the supply voltage of the cells that need to be written. This can be done by either lowering both *P1* and *P2* or by raising both *N1* and *N2* equally. This reduction in power supply shifts the trigger voltage of the inverters. The data is applied to the other set of nodes, *N1* and *N2* or *P1* and *P2*, respectively. Depending on what logic state needs to be written to the cell, either one of the remaining nodes is deviated from the power supply.

For a more detailed description of the write operation, consider the following scheme. The power supply reduction is achieved by lowering nodes *P1* and *P2*. This lowers the trigger voltage of the cross-coupled inverter pair. The voltage of node *N1* is now raised. If the initial state of the cell is such that node *V1* is "low", then *M1* is in the linear region and the voltage of node *N1* will appear at node *V1*. If this voltage is larger than the reduced trigger voltage of the inverter *M3-M4*, the state of the cell will change. In the case where the initial state of *V1* is "high", the deviation of *N1* does not affect the cell because *M1* is in cutoff. Because the reduced trigger voltage requires a smaller deviation at node *N1* to create the necessary write condition, the reduction in power supply may be used to determine which cells are written. This will work as long as the deviation of node *N1* is not large enough to write cells with full power supply but is large enough to write those with reduced power supply.

Simplest Array Implementation

In order to use this scheme in a system an array of cells needs to be created, or at least emulated. The simplest way of creating an array of cells is by connecting the four-transistor cell as depicted in Figure 2.3.

The PMOS devices of several cells are connected at a common node named *OW*. The bulks are also connected to this node to minimise the bulk effect. This common line defines a single word. Several words are connected together via common *IR* and *IB* lines. This implementation requires the routing of only three signals, if the ground node is not routed. In a typical process with a low-impedance substrate, routing ground is not necessary, or at least not as part of every cell. Therefore two fewer lines are required than for a standard six-transistor SRAM cell, when comparing on a per cell basis. When implemented in a 1.2μm CMOS process a 37,3% shrink in size compared to a standard six-transistor cell is reported [2].

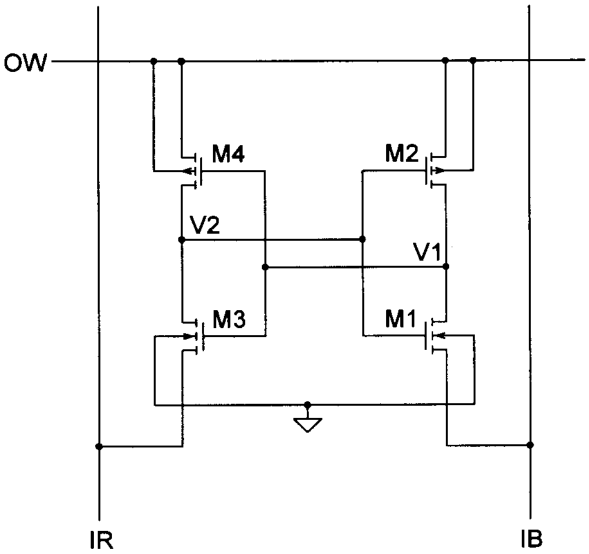


Figure 2.3 Smallest area four-transistor cell array implementation.

The cells can be written by lowering the voltage on the *OW* line and applying the data in the form of a raised voltage on either node *IR* or *IB*. Using this scheme, a number of cells can be written at the same time. In order to read any cell a single node needs to be deviated. In this scheme node *IR* can be raised. The current flowing in the *OW* line can then be monitored. This means that only one cell of a word may be read at a time, and that the equivalent bit of all other words in the array is also read. The reason that the current in the *IB* branch cannot be monitored is the fact that the current of these other cells being unintentionally read also flows into this node. These unwanted currents are a significant drawback because they have no purpose but do have the side effect of wasting power. Significant merit does however lie in the small cell size and this implementation may be very useful for serial memories where the output has to be supplied one bit at a time and the series read mode is therefore desired.

Advanced Array Implementation

In order to function similarly to a six-transistor SRAM cell array, it is important to devise an array configuration where it is possible to read and write a complete row of bits at once. This can be accomplished using a slightly more complicated scheme. The price paid is a larger cell area due to more signals needing to be routed. Figure 2.4 illustrates the configuration that can be used.

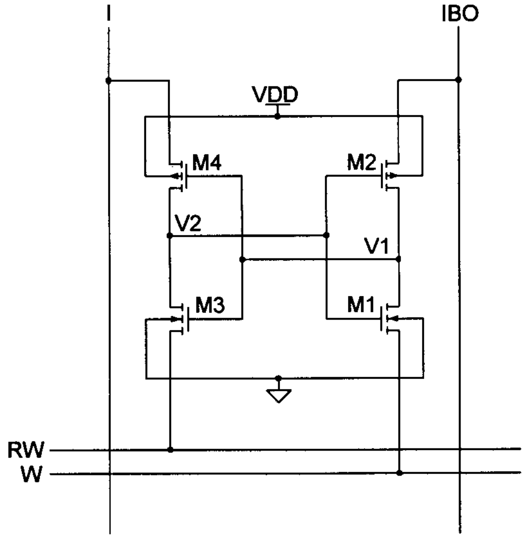


Figure 2.4 Advanced four-transistor cell array implementation.

The PMOS sources are connected vertically through the array and the NMOS sources horizontally. To write the cell, the data is applied to node *I* and *IBO*. Depending whether a "high" or a "low" needs to be written to the cell, one of these nodes is deviated from the power supply. The word to be written is selected by raising both *RW* and *W* together. This raises the trigger voltage of the inverters and allows the deviation of the PMOS node to switch the state of the cell if this is necessary. To read the cell, only node *RW* is deviated and the node *IBO* is monitored for the presence or absence of a current.

Here it can clearly be seen that it is necessary to route six lines in order to supply power and control signals to the cell. The power and ground node can be routed at less regular intervals because they only supply the bulk potential. This creates a cell with four lines, which is one line fewer than is required for the six-transistor cell. Due to the extra line, in comparison to the simplest array structure, the percentage shrink is reduced to 14.7% [2].

Lastly it is suggested by Joubert, Seevinck and Du Plessis [2] that the bulk effect present in some devices during writing and reading as well as the reduced supply voltages, will reduce the noise margin of the cell. High power dissipation is a further limitation of this array. The nodes *I* and *IBO* are connected across all words. This means that when one word is written all other words are being read. Depending on the state of the other cells a current will flow. If it is assumed that the probability of a "high" equals the probability of a "low" then one half of all cells in the array will conduct a wasted current while one specific word is being written. If, for example, a typical wasted write current of 80μA and an array size of 1024x32 bits is assumed, this amounts to a peak current of 1.31A. In the worst case scenario, where all cells in the array hold the same value and all bits of one word are written with the opposite value, double this current can be registered. This high wasted write current even for a relatively small array of cells could limit the usefulness of the proposed array structure as far as competitive power consumption specifications are concerned.

Apart from this, it has to be mentioned that area is still reduced in comparison to a six-transistor cell and that the current mode readout scheme as well as the small control voltage deviations should allow competitive read access times.

2.3 PROBLEM DEFINITION

In the light of the preceding discussion, various aspects of the design can now be defined. These can be grouped into two categories, those related to the design of the cell itself, and those related to the design of the SRAM system. Aspects of the cell which need to be addressed are:

- One of the design parameters of the cell itself are the device sizes. Here it is important to note that typically one device type will be chosen to be minimum size, so that the cell size can be kept minimum. Both NMOS devices and both PMOS devices should also be kept identical in size so that the operation and stability of the cross-coupled inverter pair are independent of its state. The parameter that requires further investigation is the device ratio, the ratio between the NMOS and the PMOS device sizes.

- The noise margin needs to be quantified and compared to the noise margin of the six-transistor cell.

- Further array configurations need to be investigated with the aim of eliminating, or at least reducing, the excessive power dissipation present during the write cycle.

- A design method for obtaining values for the required voltage deviations of the control lines to ensure successful cell operation, as well as stability, needs to be devised. Because larger voltage deviations imply larger currents, as well as smaller stability margins, this aspect of the design is strongly related to the power dissipation and the noise margin.

In order to create a complete system the following peripheral circuits are required:

- A current sense amplifier so that the output current can be sensed and converted to a digital voltage level. This sense amplifier has to be able to discriminate between a zero current state and a current being present.

- The sources of the transistors of the SRAM cell serve as the access points to control the cell. The control is accomplished by deviating certain source voltages away from the supply voltages. To achieve this, accurate voltage references combined with low output impedance driver circuits, need to be designed.

- In order to complete the system so that it functions just like a typical SRAM circuit at its outside ports, some control circuits including decoders and buffering systems are also required.

Figure 2.5 shows a block diagram of the complete SRAM system with the significant building blocks included. The control of the SRAM cell array is accomplished solely by the voltage reference and low-impedance driver circuits which control the source terminals of the transistors. Control circuits define what action is to take place and the decoded address input, as well as the data, define the state of the cells in the array. The current sense amplifier is connected to one of the device sources and therefore shares an interface to the SRAM array with the voltage reference circuits. Output drivers are present to provide sufficient driving power to charge and discharge the load capacitance without heavily loading the current sense amplifier.

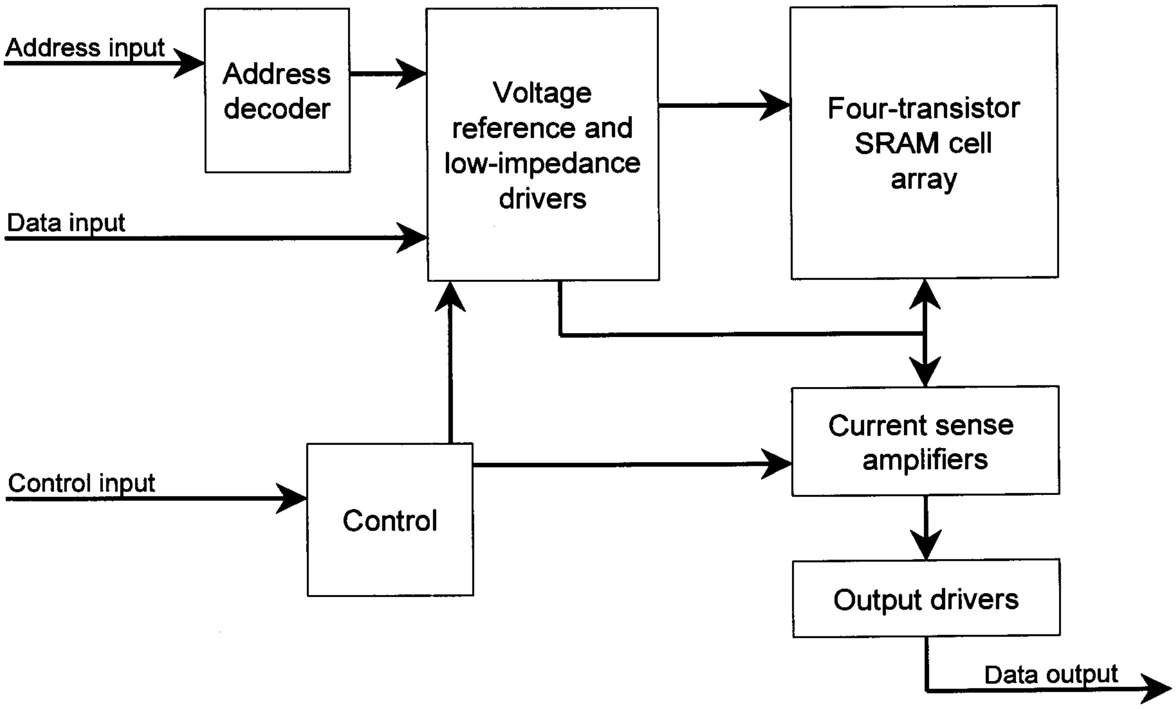


Figure 2.5 Block diagram of the SRAM system.

The word length of the RAM array was chosen to be 32 bits because this is representative of the word length of typical embedded digital systems. Furthermore, it was decided to design the system to contain 1024 words. This is not very large compared to benchmark systems [11], but most embedded memories do not have to be as large as dedicated systems. Another important aspect is that if a significant system area advantage is present in using the four-transistor cell, it should be observable at this memory size. Because some analog circuits are involved in the design, it was also decided to implement the design in a standard CMOS process suitable for both high-speed digital as well as analog design. The Austria Mikro Systeme (AMS)[1] processes were available, so the 0.6μm CMOS was chosen.

[1] Information available at: http://www.amsint.com

2.4 CELL OPERATION

Before the design parameters can be discussed, it is first necessary to describe the operation of the cell in greater detail. The aspects which need to be considered are the static operation, the read cycle and the write cycle. The discussions which follow, are all based on Figure 2.2

2.4.1 Cell in Retention Mode [12]

Retention conditions for the cell are deemed to be those conditions when no control signals are present and the cell holds its current value. This means that both NMOS sources (*N1* and *N2* in Figure 2.2) are connected to ground and both PMOS sources (*P1* and *P2*) to the power supply, *VDD*. The cell is therefore a standard cross-coupled inverter pair.

The network has only three possible operating points as can be seen from combining the voltage transfer curves of the two inverters, as shown in Figure 2.6. Because they are in a back-to-back configuration the operating points may be found by superimposing a true and mirrored transfer characteristic. Operating points are defined as those points where the voltage transfer characteristics intersect. If the loop gain around these points is smaller than unity then disturbances are weakened and therefore cannot upset the state of the system. Such a point is defined as a stable operating point. The cross-coupled inverter pair has two stable operating points, A and B. Each of these points is used to represent one digital state. A third operating point however exists at point C, but the loop gain around this point is larger than unity. Any disturbance such as noise or a device mismatch will therefore be amplified and the bias point moves to one of the stable operating points. Such a state is termed a metastable operating point.

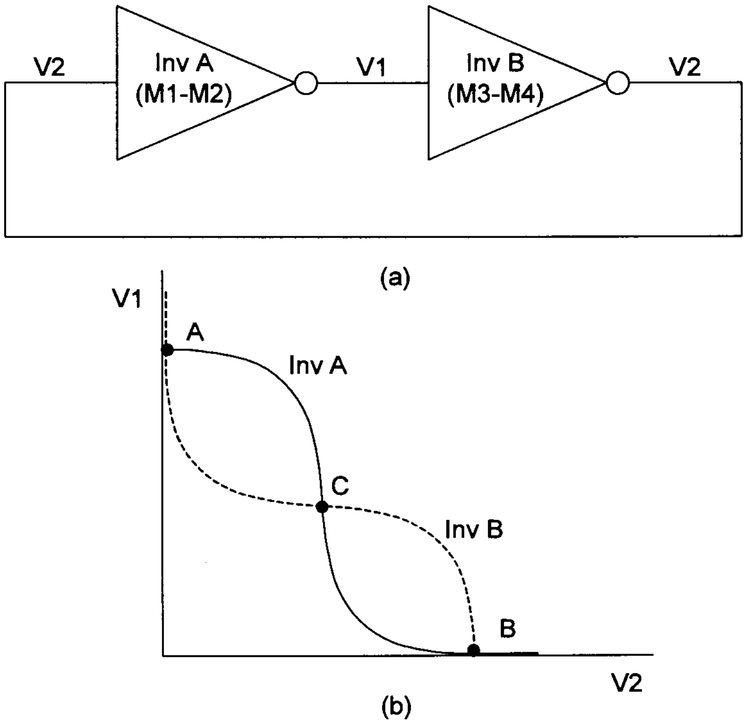


Figure 2.6 (a) A pair of cross-coupled inverters with (b) their voltage transfer characteristics showing the three operating points.

2.4.2 The Read Cycle

As has already been briefly discussed, when it is desired to read the cell, any one of the source nodes is deviated from the supply voltage. If the state of the cell is such that the device connected to the node where the deviation is applied is in cutoff, then nothing will happen in the circuit. But, referring to Figure 2.2, assume that node *V1* is "low", and that the deviation to initiate the read cycle is applied to node *N1*. Because node *V1* is "low" node *V2* will be "high". In a CMOS process at 5V supply, these node voltages will typically be 0V and 5V. Devices *M1* and *M2* therefore have gate-source voltages of 5V and 0V respectively. This implies that *M2* is in cutoff and no current can flow in the *M1-M2* branch. This places device *M1* in the linear operating region, defined by the equation

$$I_D = k'\frac{W}{L}\left[(V_{GS}-V_T)V_{DS}-\frac{1}{2}V_{DS}^{2}\right], \qquad (2.1)$$

where I_D is the drain current, V_{GS} and V_{DS} the gate-source and drain-source voltages respectively, k' the process transconductance parameter, V_T the threshold voltage and *W* and *L* the channel dimensions [13]. According to this equation, a zero current state at a high gate-source voltage implies a zero drain-source voltage. Any deviation in the source voltage of *M1* is therefore transferred directly to the node *V1* as long as the PMOS device *M2* remains in cutoff. The second inverter (*M3-M4*) is controlled by the voltage of the node *V1*. The NMOS device is initially in cutoff because its gate-source voltage is *V1*, and therefore zero. As this voltage is increased above the threshold voltage of *M3*, that device can start to conduct. It is biased in the saturation region because the drain-source voltage is much larger than the gate-source voltage. The current through this device is therefore given by

$$I_D = \frac{k'}{2}\frac{W}{L}\left(V_{GS} - V_T\right)^2, \tag{2.2}$$

if all secondary effects are ignored. In reality, the short channel effect, which is very dominant in sub-micron MOS devices, will tend to force the quadratic equation to a linear relationship [13]. The magnitude of the read current can therefore be controlled by varying the amount of voltage deviation. Two requirements are that the voltage deviation be larger than the threshold voltage, and smaller than the critical voltage which will cause the cross-coupled structure to trigger.

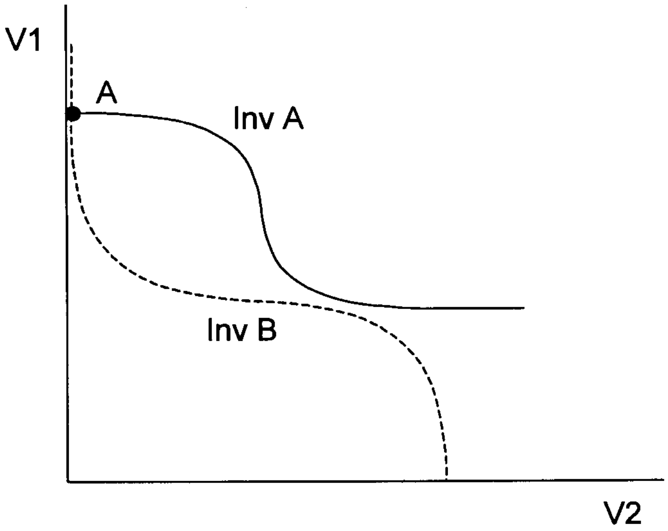


Figure 2.7 A cross-coupled inverter pair transfer curve when the ground node of inverter A is raised. Only one stable operating point exists.

Such a situation is illustrated in Figure 2.7 where the transfer characteristic of inverter A has been modified by raising the ground node. This modification is so large that the stable operating point where *V1* is "low" no longer exists, and the structure is forced to assume the other operating point, where *V1* is "high". A voltage between these two constraints ensures that one of the considerations in SRAM design, namely the non-destructive read condition, is satisfied [14]. This allows the content of one cell to be read without modifying its own content, or the content of any other cell in the array.

2.4.3 The Write Cycle

In order to force a cross-coupled inverter pair into a specific state, two conditions need to be satisfied [14]:

a. Static write condition: there has to exist one, and only one, stable operating point, which the circuit will assume when the static write condition is met. This deals only with the bias point of the circuit and does not include any transient effects.

b. Dynamic write condition: this condition determines the transient response the circuit undergoes while changing operating point during the write cycle. A slow write response is the result of a weak dynamic write condition.

A further requirement, as far as the system is concerned, is that the write to one cell may only modify the contents of that cell and not other cells in the system.

In order to change the stored value in the cell it has to be possible to force the circuit from state A to state B or vice-versa. This can be achieved by modifying the transfer characteristics so that the undesired point vanishes. At the same time it has to be assured that the desired operating point is still a stable point and that no other stable operating points exist. This is typically the writing method used in the six-transistor SRAM cell. By activating the access transistors the effective pull-up or pull-down strength of the inverters is modified. In one inverter the access device shunts the NMOS pull-down device and strengthens it, whereas in the second inverter the PMOS pull-up is shunted and strengthened.

A similar result may be achieved if the power supply of one cross-coupled inverter is reduced and the ground node of the other inverter is raised. One stable operating point will move closer to the metastable state until they become one point. If the changes are made larger still, only one point will remain as a single stable operating point and the circuit is forced to adopt that operating point.

A write to the four-transistor SRAM cell can be achieved by modifying the voltage transfer characteristics in such a way that only the single desired operating point exists. In the array of cells two types of modifications are applied, and only those cells affected by both are in a condition to change state. One of the modifications on its own, typically termed "half select", must not allow the cell to switch state. The data to be written to the cells is applied as a voltage deviation on one of two nodes and in one dimension through the array. The cells to be written are selected by changing their power supply. The applied data on its own cannot write cells. This is important because all cells in one dimension of the array are connected to the line to which the data is applied.

Consider once again the cell depicted in Figure 2.2, and assume that node *V1* is "low". It is now desired to write the other state, where *V1* is "high", to the cell. Firstly, the power supply to this cell is changed. This may be done by lowering the voltage on the PMOS-source or raising the voltage on the NMOS-sources. Assume that the PMOS-sources are used. This lowering of the supply voltage changes the output high voltage V_{OH} of both inverters, and also modifies their trigger voltages. The trigger voltage is defined as the point in the voltage transfer curve (VTC) where the input and output voltages are equal. At this point, both devices are in saturation because the V_{GS} of both is equal to their V_{DS}. An equation for the trigger voltage, ignoring all secondary effects, can therefore easily be derived by equating the device currents for the NMOS and PMOS in saturation [15] to obtain

$$V_{tr} = \frac{V_{Tn} + \sqrt{k'_p/k'_n}\left(V_{DD} - |V_{Tp}|\right)}{1 + \sqrt{k'_p/k'_n}}. \tag{2.3}$$

Lowering the supply voltage will therefore lower the trigger voltage as well. The VTC's of the cross-coupled inverter pair are modified as shown in Figure 2.8 (b).

The three possible operating points are still present. If the source voltage of the NMOS that is in cutoff, is modified, then no other conditions in the circuit are changed, so the state of the cell remains as it is. This situation is present if the cell is already in the state it needs to be written to. If this is not the case then the device connected to the raised node is in the linear region. For this explanation *M1* is linear and the voltage on node *N1* is raised. If this raised voltage is sufficiently close to the reduced trigger voltage of the opposite inverter, the cell can change state. This situation is best illustrated graphically. Raising the voltage of node *N1* modifies the transfer curve of only inverter A as is shown in Figure 2.8 (c). Only one operating point remains. At this point the output of inverter A is "high". This means that its PMOS device has been turned on. Therefore referring back to Figure 2.2, devices *M2* and *M3* are now turned on. This means the state of the cell has been flipped.

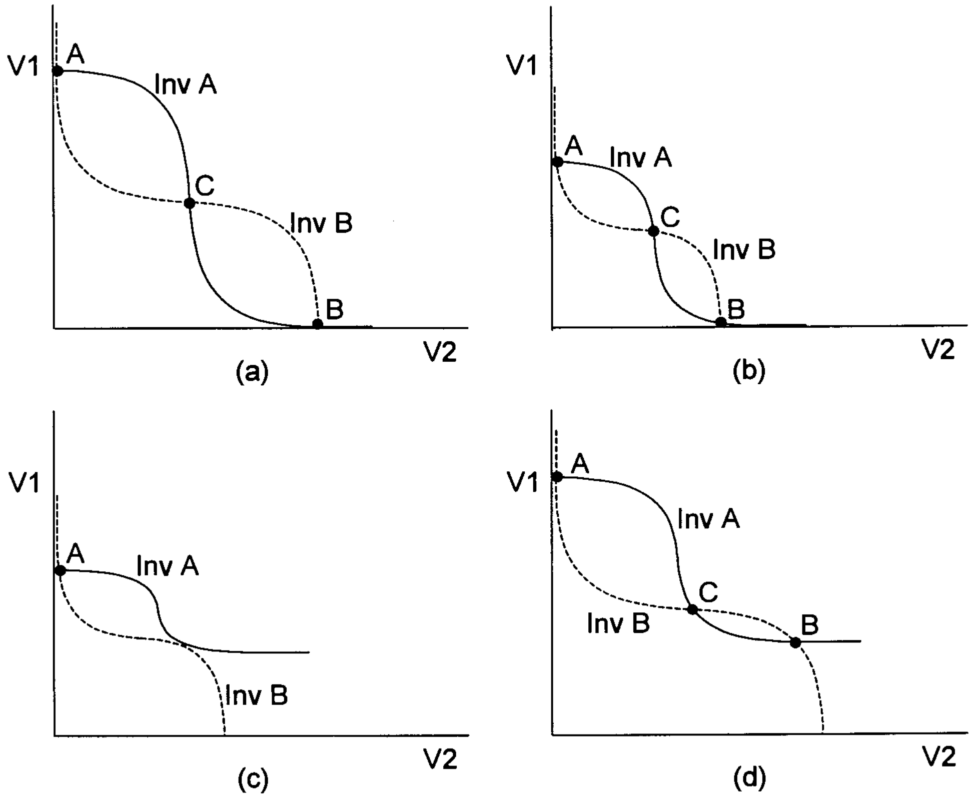


Figure 2.8 (a) The voltage transfer characteristic of a pair of cross-coupled inverters. (b) The supply voltage has been lowered and (c) one ground node has been raised to create a single operating point. (d) The ground node is raised without lowering the power supply and two stable operating points remain.

The array configuration dictates that the raising of one ground node is applied to all cross-coupled inverter pairs in the array. Because it is not desired to change their state, the voltage transfer characteristic under these conditions must still have the two stable operating points as is shown in Figure 2.8 (d).

Static write conditions are therefore satisfied by ensuring that the set of deviations applied creates only a single stable operating point. This means that for a given power supply reduction, there is a voltage deviation that has to be applied to node *N1* or *N2* depending on the value that needs to be written to the cell. There is however, a maximum allowable deviation to ensure that other cells in the array are not accidentally written.

Static write conditions deal only with the existence of a single stable bias point. They do not imply that a transient path to that point exists, and carry no information about how fast the switching takes place.

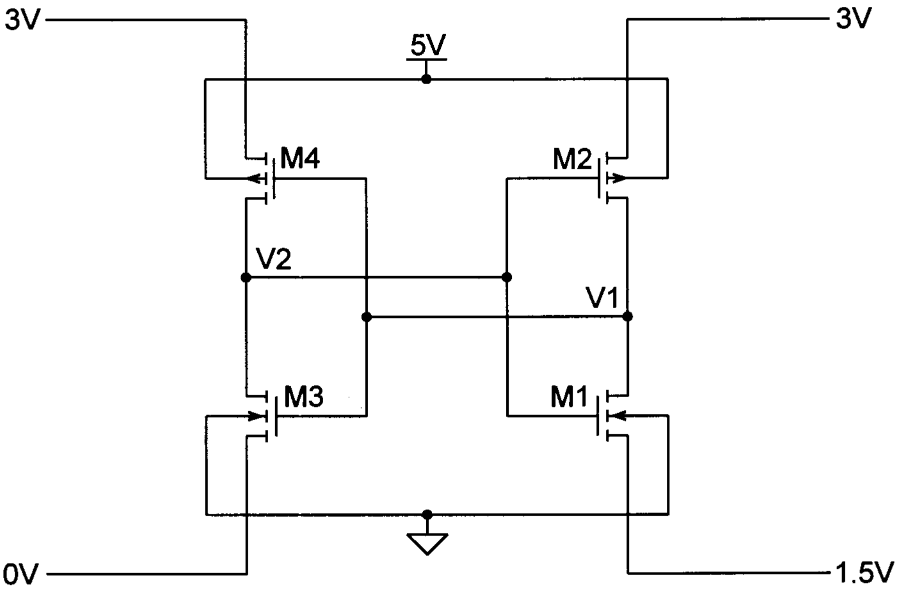


Figure 2.9 Four-transistor SRAM cell with write condition applied.

This situation can be illustrated by considering Figure 2.9. The transistors *M1* and *M4* are initially on. The voltage conditions applied to the nodes dictate that the state has to change. Due to the substrate effect present in all devices but *M3*, their threshold voltages are raised. The supply voltage of the *M1*-*M2* inverter is practically reduced to 1.5V with threshold voltages in the order of 1V each. This implies that the switching speed of this inverter is very slow due to only minimal sub-threshold conduction taking place. This situation is one where static write conditions are satisfied but the transient response is very slow because of the low supply voltages. The slow response is a result of needing to charge node *V1* to 3V and the required current having to be delivered through device *M2* which is barely on.

From this discussion it can be learned that one disadvantage of the four-transistor SRAM cell is the fact that it cannot operate at competitive speed for low supply voltages. In a standard CMOS process it seems that using a supply voltage of 5V is required to guarantee speed.

From the static and dynamic write conditions, maximum and minimum limits for the required voltage deviations can be defined. The design goal should be to use the minimum possible deviations in order to optimise the switching speed.

2.4.4 Limitations of the Write Cycle

Variations in Device Performance

The manufacturing process of an integrated circuit leads to variations in device quality. The manufacturers therefore typically supply a set of five simulation models. Because process variations are inevitable, the design has to cope with all process extremes in order to guarantee satisfactory operation. The following models are usually supplied:

- Typical mean (TM): All process variations are set to their average value.
- Worst case speed (WS): This model includes slow NMOS and slow PMOS devices. Typically this is brought about by high threshold voltage and low process transconductance factor. Currents are low and devices are therefore slow.
- Worst case power (WP): Process variables are set to obtain strong devices. Currents and speed are high due to high process transconductance factors

and low threshold voltages. The high currents bring with them high power dissipation but also fast response times.

- Worst case one (WO): This is a combination of a high quality NMOS and a low quality PMOS device. Speed and power dissipation are average but the relationship between the NMOS and the PMOS is distorted. Noise margin and stability problems may occur under these conditions.

- Worst case zero (WZ): This model is the opposite situation of the worst case one model. The effects on a circuit are however identical.

The extent of the effect that process variation can have on the devices can be shown graphically as in Figure 2.10. It shows a two dimensional plot of the simulated current through a saturated NMOS and PMOS device under equal bias conditions.

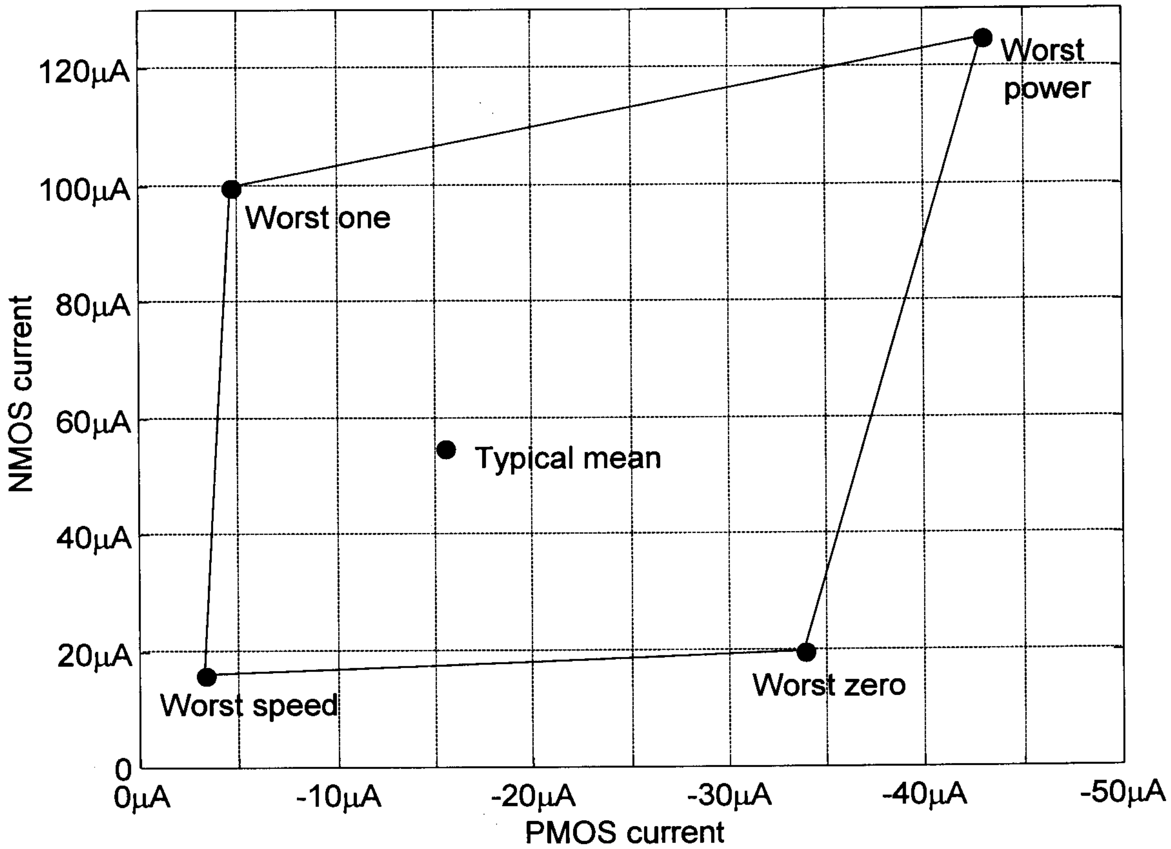


Figure 2.10 Drain current of saturated NMOS and PMOS devices at $|V_{GS}|$=1.3V and a *W/L* ratio of 1.4μm/0.6μm.

Simulation Issues

The four-transistor SRAM cell was simulated using the different models. For these simulations, values had to be assigned for the deviations to be applied. For the typical mean model it seemed that a good choice would be the same as was proposed by Joubert, Seevinck and Du Plessis [2]. That work proposes equal deviations of 1.5V. The cell could be verified as being operational. The time taken for the cell to switch its state after the control signals have been applied was simulated as 1.52ns. This indicates good dynamic write conditions. Problems were however experienced with other models.

- When using the worst case speed model the threshold voltages are too large (0.95V instead of the typical 0.8V). The 1.5V reduction in power supply from either side affecting the one inverter leaves a power supply headroom of 2V. At the high threshold voltages, which are further raised by the bulk effect present in both devices, the transient response becomes very poor. The deviations can be reduced, but this decreases the reliability of the write cycle, because the static write condition is weakened.

- When using the worst case zero model, the NMOS devices are weak and the PMOS devices strong. This raises the trigger voltage of the inverters. The 1.5V deviation applied to one of the NMOS source nodes is therefore not large enough to create reliable static write conditions. To obtain an operational cell under these conditions the power supply reduction via the PMOS source nodes had to be reduced and the NMOS source node deviation increased.

- Simulation with the worst case one model yielded a problem of a different sort. The 1.5V deviation applied to the NMOS source node to write data to the cell is also applied to all other cells in the array, and may therefore only create static write conditions if applied to a cell together with the power supply reduction. But due to the high quality NMOS devices combined with the poor quality PMOS devices, the trigger voltage of the inverters is reduced so far that a 1.5V deviation of one NMOS node creates static write conditions. All cells in the array are therefore written.

The cell could be designed to function more reliably by designing the voltage deviations to change as the process changes. The simulations prove that the system would then be just inside the reliable region of operation across all processes. A more trustworthy design would be one where the cell operates for a given set of deviations under all process conditions. Reliability can then be increased by designing the deviations to change slightly with process conditions.

A second issue is the power dissipation. It is desired to keep that deviation which represents the data, and therefore is applied to all cells, as low as possible. This reduces the wasted current flowing in those cells which are read during a write cycle. The 1.5V deviation applied is typically 0.6V above the threshold voltage. This means that wasted currents are typically as high as 80μA per cell.

2.5 ALTERNATIVE WRITE CYCLE

To increase the reliability of the write cycle, a different approach can be used. The limitation in the method described up to now is the low supply voltage present in one of the inverters, which is necessary so that static write conditions exist. Consider the scheme of writing the cells depicted in Figure 2.11(a). The power supply reduction is restricted to the PMOS node of the inverter which is opposite to the inverter where the NMOS source node is raised in response to the data. The advantage of this is that each inverter is only affected by a single power supply reduction. This allows the transistors to have larger gate-source voltages and restores good dynamic write conditions.

As far as static write conditions are concerned this configuration is very effective for creating a single stable operating point. Consider for example devices *M1* and *M4* are initially on. The applied source node deviations cause the trigger voltage of the *M1-M2* inverter to be raised and that of the *M3-M4* inverter to be lowered. This creates strong positive feedback towards the desired operating point. The previous method only lowered the trigger voltage of the *M3-M4* inverter, while leaving that of the other inverter unchanged. This is due to the fact that the lowering of the PMOS source node is cancelled by raising the NMOS source node.

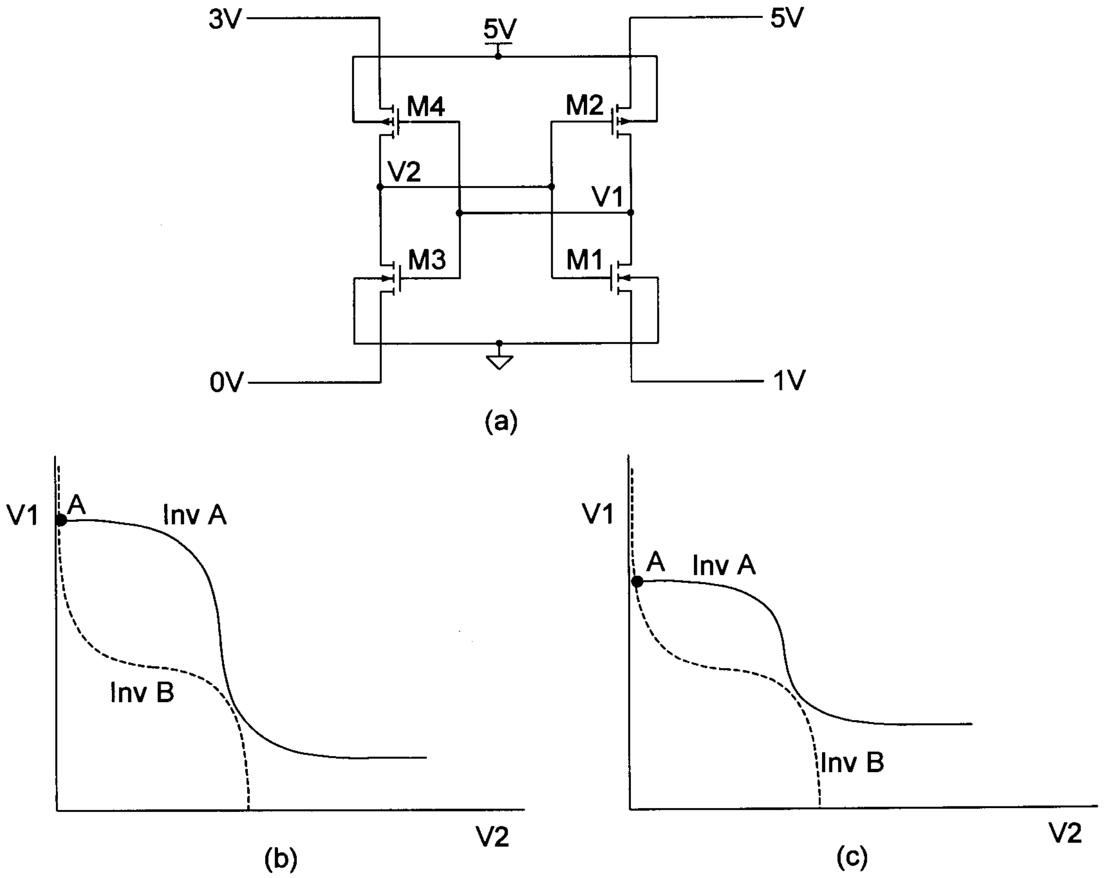


Figure 2.11 (a) Newly proposed scheme of writing the four-transistor SRAM cell. (b) The static write condition for this scheme compared to (c) the static write condition for the previously proposed scheme [2].

It can be seen from Figure 2.11(b) that static write conditions can be created while adequate power supplies to both inverters are maintained. This can be compared to the previously discussed scheme. The static write condition transfer curve is repeated in Figure 2.11(c) for comparison, where it is clear that inverter A is subjected to extremely low power supply. Further, it can be observed that a lower NMOS source node deviation is sufficient to create adequate static write conditions, because the trigger voltage of inverter A is not lowered by a power supply reduction. A single operating point is established at a lower source node deviation of *M1*, and this helps to achieve lower wasted currents during the write cycle thereby improving power dissipation.

A significant disadvantage is however also present. The deviations in both the PMOS and NMOS source nodes are data dependent. It is therefore no longer possible to select a complete row of cells and in one step write both binary values. A row can be selected and certain cells can be written to one binary value. The row may then again be selected using the other PMOS node and certain cells may be written to the other binary value. Alternatively a scheme could be devised to set all cells in a row to a known value and then use the proposed write method to set certain cells to the opposite binary value. Whichever scheme is used, the write cycle becomes a two-phase procedure, which will require more time to complete and more complex control mechanisms to implement.

Simulation of a cell using the different process conditions does however indicate that the cell is functionally operational without errors across all worst case models. This is achievable even if the deviations are kept constant. A set which works well is a PMOS source deviation of 1.8V and an NMOS source deviation of 1V.

The significance of the 1V NMOS source deviation is that the wasted power during the write cycle is reduced because the voltage which reads all other cells during a write, is reduced. According to equation (2.2) this reduces the current and therefore the power. In the typical mean case this current is reduced from 80μA to 20μA.

2.6 ALTERNATIVE ARRAY STRUCTURE

The newly proposed write mechanism has to be implemented within an array of cells. As mentioned above, the write cycle has to be structured as two separate sub-cycles. "Ones" and "zeros" can be written into the array in two separate cycles or the cells of one word can all be cleared and then selectively written with "ones". Clearing the cells can be accomplished by applying a large deviation on one node. This creates static write conditions quite easily. Whether to choose the NMOS or PMOS node depends on the design of the inverters. Typically it is desired to design all cell transistors minimum size. This allows the area of the cell to be minimised. In this case the trigger voltage of the inverters is in the region of 2V because the NMOS is a better device than the PMOS. It is therefore advantageous to use an NMOS source node to clear the cell. Because the trigger voltage of the inverters is closer to ground than it is to the power supply, static write conditions can be established at a smaller node voltage deviation. This means that the static write conditions are combined with a higher power supply to the inverters and therefore stronger dynamic write conditions.

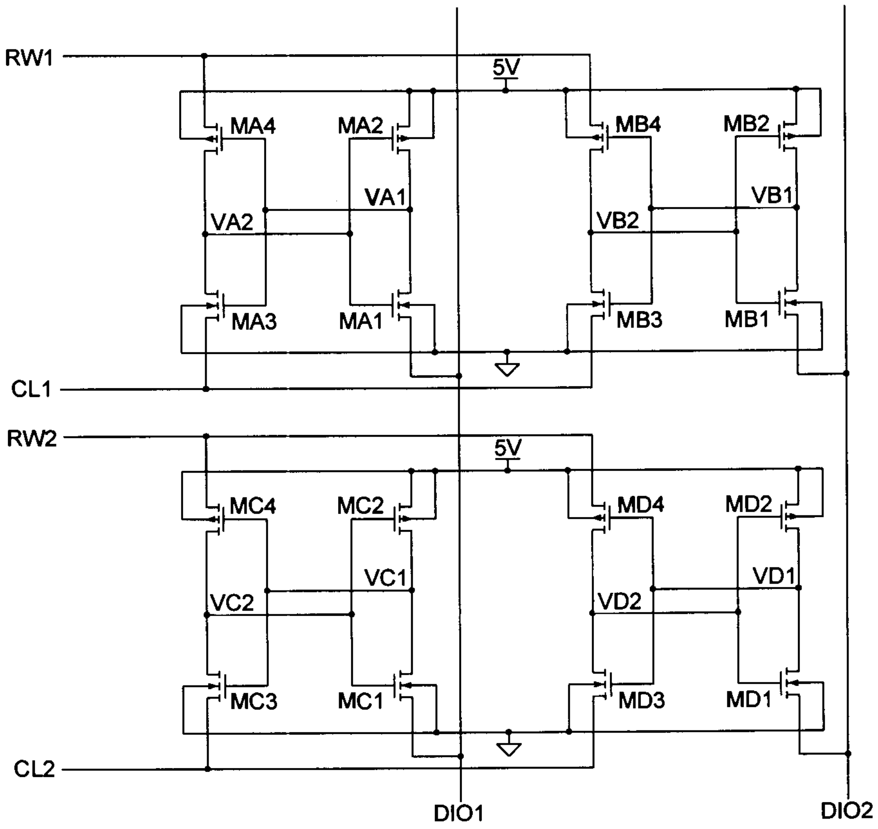


Figure 2.12 Proposed array structure incorporating the alternative write cycle.

The 2x2-array structure of Figure 2.12 shows how an array of cells can be implemented. A row of cells can be placed in a specific state by pulling the *CL*-line to the power supply. The cells are thereby forced into a state where *M1* and *M4* are on and thus *V1* is "low". This state is defined as a logic "zero". After this, certain cells may be placed in a logic "one" state by lowering the voltage on the *RW*-line and raising the voltage on specific *DIO*-lines. This complete procedure is required for writing a word. All cells are cleared and selected cells are then set.

Reading a word is accomplished by lowering the voltage on the *RW*-line. This causes a current to flow in the *DIO*-line if the cell connected to that line is in a logic "zero" state. If the cell is set no current will flow.

Compared to the array previously proposed [2] this implementation has several advantages:

- Functional operation is possible across all process deviations using a constant set of node deviations. This indicates greater reliability of the system.
- Five lines instead of six need to be routed, resulting in smaller cell size.
- The wasted power during the write cycle is significantly reduced by two mechanisms. Firstly, it is possible to use lower *D*/*O*-line voltages as already explained. This lowers the wasted current from 80μA to 20μA per cell. Secondly, under the assumption of equal probability data only half the *D*/*O*-lines will be activated and cause a wasted current in half the cells connected to them. One quarter of all cells in the array waste current instead of one half. Considering the 1024x32 array this amounts to a total wasted current of 163mA instead of 1.31A, a reduction of 87.5% when using the typical mean model. The worst case wasted current, that is when all cells are "zero" and one word is written to all "ones", decreases from 2.62A to 655mA. The percentage reduction here is 75%, once again assuming the typical mean simulation model is used.

The price paid for these advantages is the two cycle write procedure which requires more time and more complex control.

2.7 CELL DESIGN

In this section a design procedure that can be applied to design the four-transistor SRAM cell for any CMOS process is discussed. Two aspects require designing, namely the device ratio between the NMOS and PMOS device and the magnitude of the voltage deviations. The design of the latter is based on a noise margin analysis.

2.7.1 Device Ratio

One device is typically taken to be minimum size and the other is scaled to achieve the desired device ratio. Increasing the device ratio by an increase in the NMOS device strength will result in faster switching, because capacitance can be discharged faster. The trigger voltage of the inverters will be lowered and the cell size increased. Considering that lowering the trigger voltages of the inverters will ease the establishment of static write conditions if only a single NMOS source node is raised, this should be avoided. Larger cell size is also unwanted and the speed achieved from the cell is satisfactory, even for minimum size devices. Here it is important to note that the NMOS devices do not have to be strong to discharge large bit line capacitance because the cell is accessed differently. A good design choice is therefore to use minimum size transistors all round. The minimum allowable size is 0.8μmx0.6μm, requiring the so-called dog-bone layout shown in Figure 2.13(a). The design rules governing the process [16] dictate that a dog-bone transistor layout is larger in area than one which is sized to fit the minimum dimension of a diffusion contact, as in Figure 2.13(b). All cell transistors are therefore designed to be 1.4μmx0.6μm.

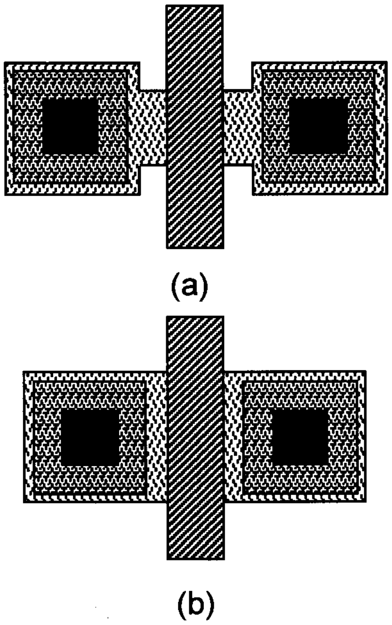


Figure 2.13 (a) Smallest device size transistor 0.8μmx0.6μm and (b) smallest area transistor 1.4μmx0.6μm.

2.7.2 Noise Margins of Logic Circuits [17]

Several types of noise may affect a logic circuit and there is a noise margin associated with each type of noise. The best case noise margin, sometimes called the typical noise margin, is defined to be the maximum noise magnitude that does not disturb the proper logic operation of an infinitely long chain of identical gates, when it is concentrated somewhere in a single gate. The worst case noise margin is the maximum noise amplitude that still guarantees proper operation when it is applied identically to each gate in an infinitely long chain of inverters. When considering the worst case noise margin of such a chain of inverters it has been proven that the chain may be replaced by a cross-coupled inverter pair for analysis purposes [17].

The following DC noise sources can be present in a logic circuit [18]:

- series-voltage noise: a series voltage exists between the gates,
- parallel-current noise: a current is present at the input and output of the gates,
- voltage-noise in the ground line
- voltage-noise in the power supply line.

These static noise sources are present all the time. Dynamic noise is present in short pulses. The noise amplitude may therefore be higher before incorrect operation results. The shorter the noise pulse, the higher the amplitude can be. The best method of obtaining these noise margins is by simulation [18].

Several methods exist to calculate the static noise margins. Most interest lies in obtaining the series-voltage noise margin, and it is typically referred to as the noise margin of a system. If the assumption is made that the output impedance of a gate is much smaller than the input impedance of the gate being driven, then the voltage transfer characteristic is invariant with loading. For CMOS this is typically the case due to the high input impedance of the MOS transistor gate terminal. To calculate the noise margin, the maximum square between the normal and mirrored transfer characteristic must be found. The length of the sides of that square represents the worst case noise margin.

2.7.3 Static-Noise Margin of the Four-Transistor SRAM Cell

The SRAM cell is a cross-coupled inverter pair and the noise margin may therefore be analysed in the same way as was proposed for an infinitely long chain of inverters. When referring to the noise-margin of the SRAM cell the series-voltage noise margin is implied. Typically only this noise margin is considered because it is the smallest of the four DC noise sources. Due to the low on-resistances of the MOS devices, high currents are required to upset the state of the cross-coupled inverter pair, and the parallel-current noise margin is very large. The power supply and ground noise is transmitted onto the internal nodes via the MOS devices operating in the linear region, and so only one internal node is affected at a time. The margins for these types of noise will therefore also be larger than the series-voltage noise margin which affects both internal nodes equally.

The series-voltage noise margin is found by superimposing the voltage transfer characteristics of the two inverters and finding the maximum square as shown in Figure 2.14(b).

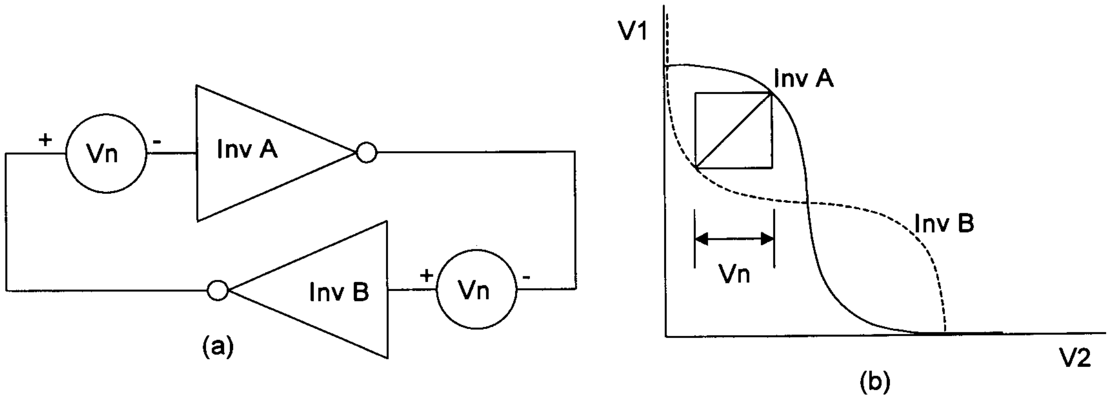


Figure 2.14 (a) Cross-coupled inverter with worst case series-voltage noise sources inserted and (b) the graphical representation of the worst case series-voltage noise margin.

A simple algorithm to find the maximum square is to define a new *u,v* coordinate system that is rotated 45° with respect to the original axes (Figure 2.15). The diagonal of the maximum square now lies parallel to the *v*-axis. The transfer function points are translated to the new coordinate system and the *v*-distance between the two curves is calculated as a function of *u*. The smaller of the maximum and minimum value of this distance is the length of the diagonal of the smaller maximum square. This, when translated back to the original coordinates (divide by square root of two) is the worst case static noise margin [10]. The transformation required to rotate the axes is defined by:

$$u = \frac{x-y}{\sqrt{2}} \tag{2.4}$$

and

$$v = \frac{x+y}{\sqrt{2}}. \tag{2.5}$$

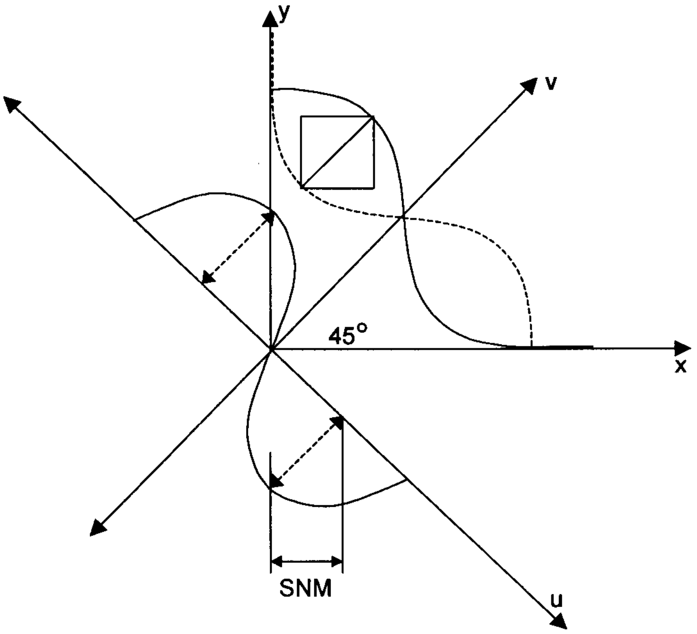


Figure 2.15 Static noise margin (SNM) estimation based on "maximum squares" in a 45° rotated coordinate system.

When any node of the four-transistor SRAM cell is deviated from the supply voltage a reduction in noise margin takes place. The two situations which need to be analysed are the reduction in noise margin when (a) a cell is being read and (b) a different cell in the array is being written. Further, it can be said that a zero noise margin implies that no external noise input is required to cause the cell to loose its current state. This is equivalent to static write conditions being present.

2.7.4 Design of Voltage Deviations

The algorithm presented above was used in a program (see addendum A.1 for the C-code) that calculates the noise margin from a set of inverter transfer curves. For the four-transistor SRAM cell, when node voltage deviations are applied, the two transfer characteristics differ. The program reads two sets of several transfer characteristics. In one set the PMOS node is lowered in steps and in the second set the NMOS node is raised in steps. The sets of transfer characteristics are generated using a circuit simulator and the models supplied by the manufacturer. One transfer characteristic of each set is used in the noise margin calculation algorithm. This therefore analyses the noise margin of the system of Figure 2.16. The deviation of the PMOS node is termed *Y* and that of the NMOS node on the opposite inverter *X*. This system caters for all noise margin degradation possibilities that can occur.

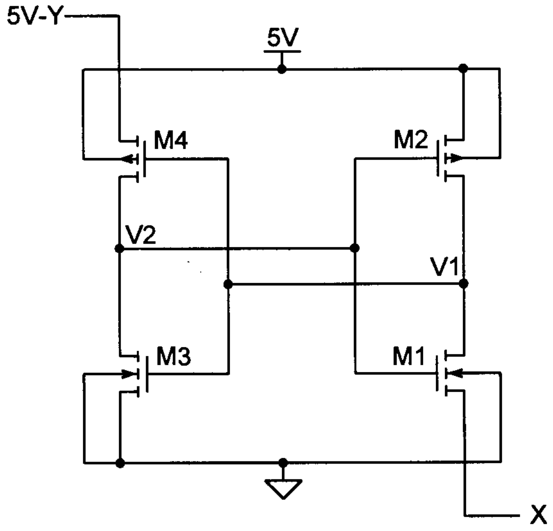


Figure 2.16 Static noise margin analysis system.

The program returns the noise margin as a function of *Y*, while *X* is zero, and the noise margin as a function of *X*, while *Y* is zero. These situations relate to the noise margin of a cell while being read, and that of a cell while another cell in the system is being written, respectively. A set of (*X*,*Y*) points where the static noise margin is zero is also returned. These points define the boundary that has to be crossed to achieve static write conditions.

The results generated are shown in Figures 2.17-19. In order to design the deviations it is required to consider all three plots together. Figure 2.17 is an indication of the noise margin as a function of *Y* across all simulation models while the node *X* is kept at zero volt. This is therefore an indication of the noise margin of the four-transistor cell while it is being read. Figure 2.18 shows the opposite situation where *Y* is kept zero and the noise margin of the cell as a function of *X* is plotted. This is interpreted as the noise margin of one cell while another is being written. The general method of design would be to choose *X* and *Y* such that the noise margins are equal.

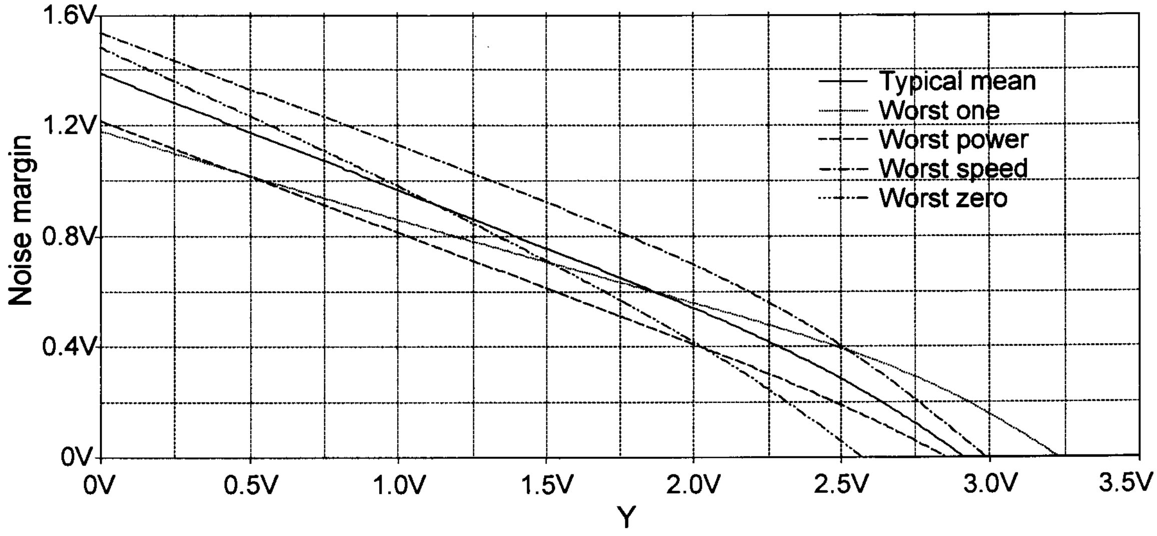


Figure 2.17 Noise margin plotted against *Y*-deviation for *X*=0 for the different simulation models.

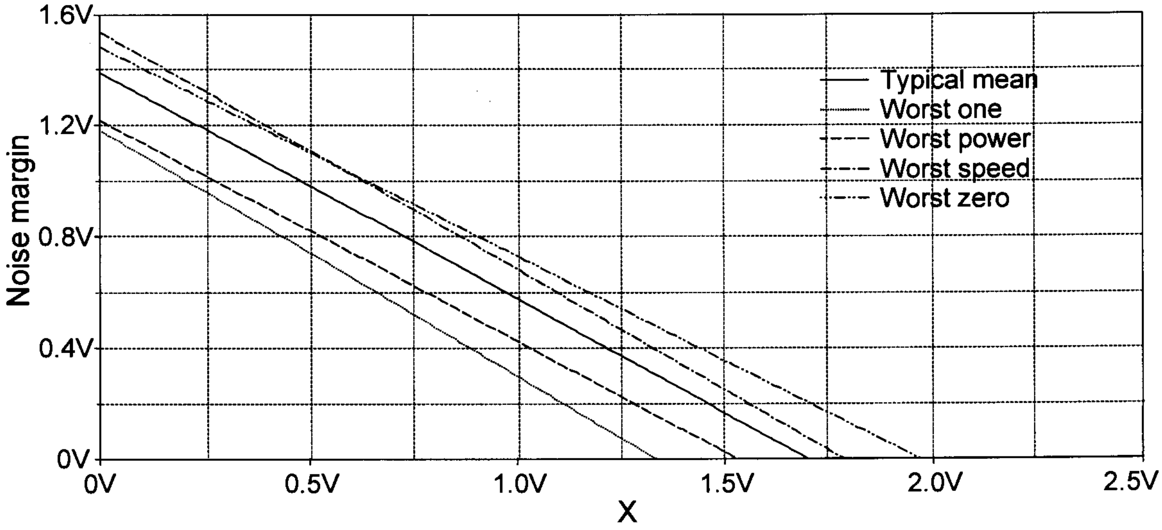


Figure 2.18 Noise margin plotted against *X*-deviation for *Y*=0 for the different simulation models.

A second constraint that has to be satisfied is that the selected *X*- and *Y*-deviations together have to create static write conditions. If the selected point is plotted on Figure 2.19 the point has to lie above the zero noise margin line. Designing the deviations therefore necessitates finding a set that yields large and equal noise margins as well as static write conditions. Selecting a point on the zero noise margin line will however not be sufficient, because it places the cell on the verge of being written. To ensure reliability in the write cycle a margin of safety is required, and the selected point should lie above the zero noise margin line, introducing a write safety margin.

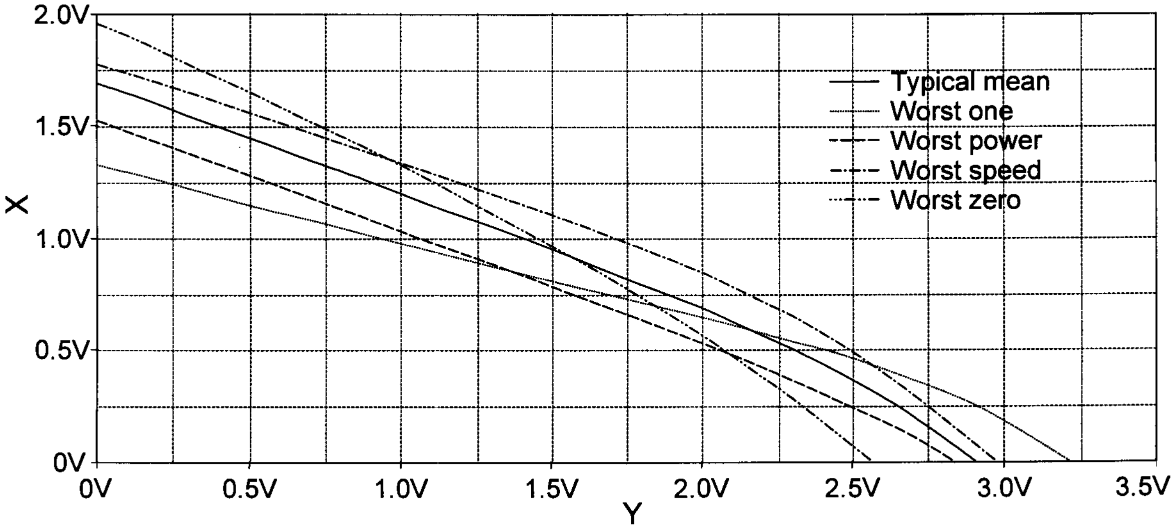


Figure 2.19 Zero noise margin trajectories for all simulation models. A point above the graphs implies that static write conditions are satisfied.

Using the three graphs the following deviation scheme was devised. The standard design point for the deviations is *X*=1V and *Y*=1.8V. This was selected because the static write conditions are achieved for all process conditions at a low *X*-deviation and an acceptable *Y*-deviation. Equal noise margins of 0.6V are achieved for the typical mean case. The selected point also lies at least 0.1V beyond any zero noise margin line, thereby introducing a write safety margin of 0.1V. Even though all margins change as the process conditions change, the chosen point guarantees operation across all conditions. It is however desirable to improve this situation. Referring to Figure 2.18 it is advantageous to decrease the *X*-deviation for the worst case power and worst case one situation, and increase it for the worst case speed and worst case zero situations. This is equivalent to scaling the *X*-deviation depending on the quality of the NMOS transistor. This decreases the spread on the noise margin and, importantly, counters the low noise margin of the worst case one situation.

Applying a scaling dependent on the PMOS device quality achieves similar results when considering the *Y*-deviation. This scheme also increases the write safety margin for the worst case speed model and reduces the excessive safety margin associated with the worst case power model.

The current flowing in the opposite inverter to the one where the specified single deviation is applied, is shown in Figures 2.20 and 2.21 for the NMOS and PMOS case. Figure 2.20 therefore illustrates the wasted write currents and Figure 2.21 the read currents. The current spread for a constant deviation is quite substantial as the process changes, and can be reduced by adapting the deviation voltages as discussed above. This is especially true for the *X*-deviation. A spread of 60μA can be reduced to 25μA by designing for a variation of 0.15V around 1V as the quality of the NMOS device changes. It can clearly be seen from Figure 2.20 that the wasted write current does increase significantly in a worst case power scenario. This can lead to excessively high power dissipation. Reducing the *X*-deviation in these situations will save power.

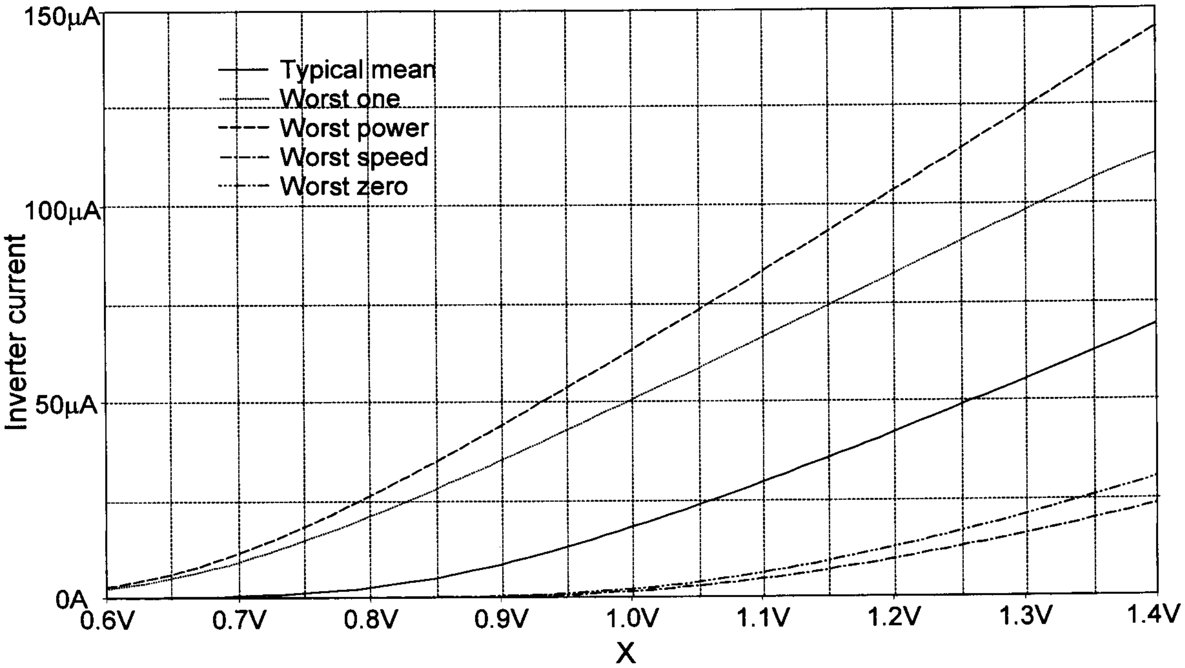


Figure 2.20 Simulated current flowing in the opposite inverter of the four-transistor SRAM cell across the five process models when a certain *X*-deviation is applied.

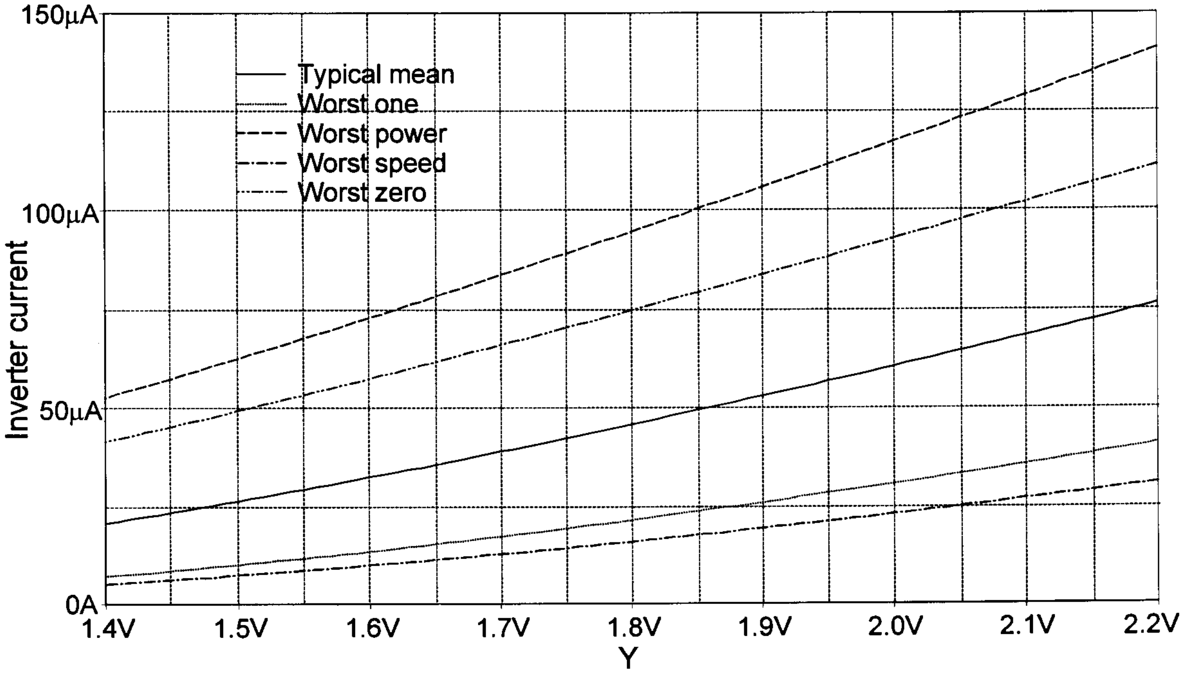


Figure 2.21 Simulated current flowing in the opposite inverter of the four-transistor SRAM cell across the five process models when a certain *Y*-deviation is applied.

Referring to Figure 2.21 the spread on the read current can also be reduced. This once again saves power but, more significantly, raises the minimum current that needs to be detected, while also lowering the maximum current. The higher minimum current, combined with the reduced spread, can potentially reduce the complexity of the current sense amplifier.

Typical process dependent variations of the *X*- and *Y*-deviations that still yield satisfactory safety margins on the static write conditions are 0.15V and 0.2V, respectively. The variation in the *X*-deviation compensates for quality variations of the NMOS and that of the *Y*-deviation compensates the PMOS device. These variations may therefore be generated using the device in question as a reference. If the deviations are generated using the threshold voltage of the respective device as a reference, a decrease in device quality which is largely due to an increase in the threshold voltage, will produce the correct change in the voltage deviation.

2.7.5 Effects of Temperature

Due to the potentially high power dissipation present during the write cycle, raised temperatures can be expected. As the temperature increases, the overall quality of the devices decreases. The following factors contribute to a variation in overall device quality as the temperature changes [9]:

- The effective carrier mobilities in the channel are decreased. This decreases the process transconductance parameters of the devices and they become weaker as temperature increases.

- The threshold voltages are reduced as temperature increases. For the given process the variations are -1.4mV/K and -1.9mV/K for the NMOS and PMOS device respectively [19].

Usually the first parameter is dominant and an overall performance degradation is observed with increasing temperature. The operation of the four-transistor SRAM depends only on the ratio of the process transconductance parameters of the two devices and both of them are affected equally. The variation in threshold voltages does influence the currents, as well as the zero noise margin points. The speed and power dissipation is also affected. At lower temperature higher currents are observed because of the higher mobility. The speed is reduced due to higher threshold voltages. Based on this there is another advantage to deriving the *X*- and *Y*-deviations from the threshold voltages. As previously mentioned the level of the deviation will track the threshold voltage. As far as the variation with temperature is concerned, as the threshold voltage changes, the deviations will track this change, therefore countering the effect of a change in threshold voltage. This allows operation over a wide temperature range.

2.8 TRANSIENT SIMULATIONS

To validate the results of the previous section a transient simulation is presented. One of four control procedures may be applied to the cell, namely

a. the *CL*-node raised to 5V (the cell is cleared),

b. the *RW*-node lowered by *Y* (the cell is being read),

c. the *D/O*-node raised by *X* (another cell in the array is being written) or

d. (b) and (c) are applied together (the cell is being written).

Each of these control operations may be applied irrespective of the state of the cell. It is therefore required to test each of these operations for each of the two cell states. A change of state of the cell may only take place if the state of the cell is "set" and operation (a) is applied or the state of the cell is "clear" and (d) is applied. Initially the cell is brought into a known state by activating the *CL* signal. The cell is in the "clear" state. The three operations which may not modify the contents of the cell are applied. The cell is then written and the state changes to "set". Again three operations that may not modify the contents are applied. Finally the cell is cleared again. This simulation is repeated using the different process models. The deviations are changed according to Table 2.1.

Table 2.1 Control voltages used for the different simulation models.

Deviation type	TM	WO	WP	WS	WZ
RW deviation (Y)	1.8V	2.0V	1.6V	2.0V	1.6V
DIO deviation (X)	1.0V	0.85V	0.85V	1.15V	1.15V
CL deviation	5V	5V	5V	5V	5V

The clear control voltage remains unchanged. A deviation of 5V is used not only with the objective that is quite simple to implement, but also that it can be generated without consuming static power. This source is only activated when the state of all cells needs to be made identical and it is only applied to those cells that need to be cleared and does not affect other cells. Noise margins are therefore not an issue and any control voltage that fulfils the static write conditions is adequate.

The important characteristics to be assessed are the correct functional operation, the read current, the wasted write current, the read access time, the write time and the clear time. The read access time is defined as the time difference between the 50% levels of the *RW*-signal and the output current pulse, whereas the write and clear times are taken as the time between a 50% level in the *DIO*-line or *CL*-line to the point where the voltages of the internal SRAM nodes are equal. A rise time of 1ns is used for all control signals. This was decided because 1ns is in the same time range as the response speed of the cell.

The simulation is also repeated at different temperatures. This part is however only performed to test the theory that the cell remains functional even if the temperature changes because the exact deviations of the control voltages with changing temperature are unknown.

Figure 2.22 shows the control signals *RW*, *DIO*, *CL* for the typical mean case. The results of the simulation are shown in Figure 2.23. The simulation results clearly indicate the state of the two internal nodes of the SRAM cell, *V1* and *V2*. The two inverter currents are also shown. The wasted write current, the read current as well as the peak currents that flow while the state of the cell is changing, can be seen. The state of the cell changes at only the correct times, so the SRAM cell is operational. This holds for all five process models using different control voltages. The cell is operational at a junction temperature in the range from -55°C to +125°C. The simulation results are summarised in Table 2.2.

Table 2.2 Simulated specifications for the four-transistor SRAM cell.

Model type	Read current (μA)	Wasted write current (μA)	Read access time (ps)	Write time (ps)	Clear time (ps)
Typical mean	44.9	17.7	390	859	143
Worst case one	28.1	27.6	440	967	184
Worst case power	77.0	34.0	327	724	136
Worst case speed	21.7	6.8	589	1340	370
Worst case zero	61.6	9.0	342	764	263

Apart from the fact that the cell is operational independent of process and temperature, it can also be seen that the current specifications do not vary as drastically as can be expected from Figures 2.20 and 2.21. The read current is at least 20μA, which does not require an extremely sensitive current sense amplifier. The wasted write currents are low, considering what the initial estimates amounted to. The access times will be compared to those of the six-transistor SRAM cell later in this chapter.

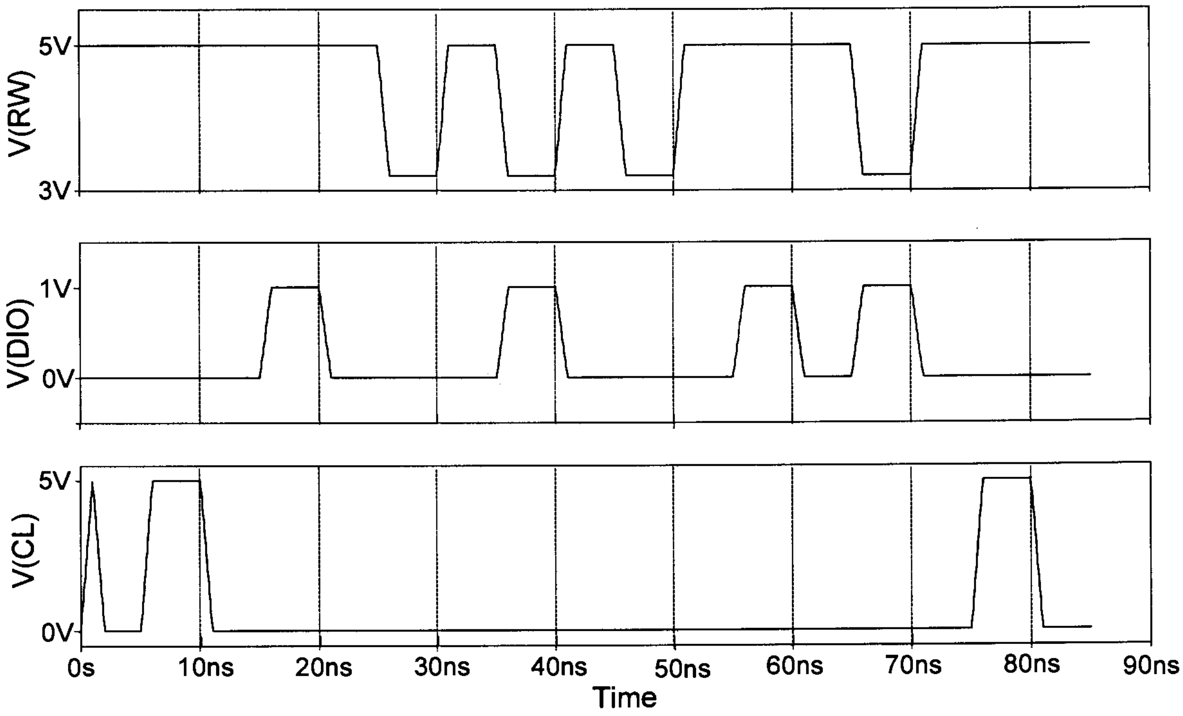


Figure 2.22 Control signals applied to the four-transistor SRAM cell for the typical mean case.

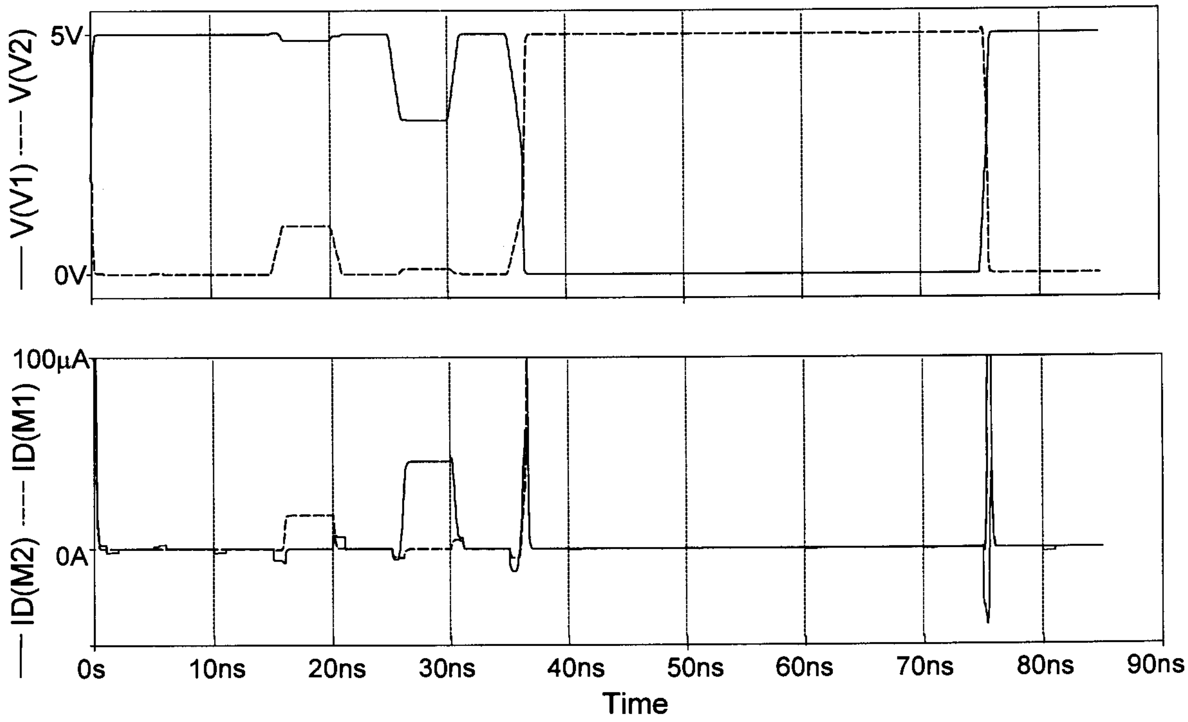


Figure 2.23 Response of the four-transistor SRAM cell using the typical mean model.

2.9 EXPERIMENTAL VERIFICATION

The proposed array structure combined with the proposed scheme of accessing the cell was verified experimentally. A 2x2 array of cells manufactured in the AMS 0.6μm CMOS process was tested. This array was initially manufactured to suit the access scheme proposed by Joubert, Seevinck and Du Plessis [2]. The equivalent PMOS sources are connected in the horizontal array dimension and the NMOS sources in the vertical dimension. This means it was not possible to use the NMOS node for clearing the cells. As previously mentioned the NMOS source was chosen because of the speed advantages. The measurement equipment, as well as the peripheral circuits, operate at speeds in the microsecond range, so this speed advantage is not significant. Two measurement set-ups were therefore used, one of them demonstrates the functional operation of an array and the other verifies that cells may be cleared using the NMOS source. The first setup uses the unused PMOS source to clear the cells. In order to use digital input signals to control the cell some interface circuits were constructed to perform the following tasks:

- NMOS source driver: to convert a logic "high" input signal to an adjustable deviation from 0V and a logic "low" to 0V,
- PMOS source driver: to convert a logic "high" input signal to an adjustable deviation from 5V and a logic "low" to 5V,
- current-to-voltage converter: to sense a current of at least 20μA and convert it to a measurable voltage swing.

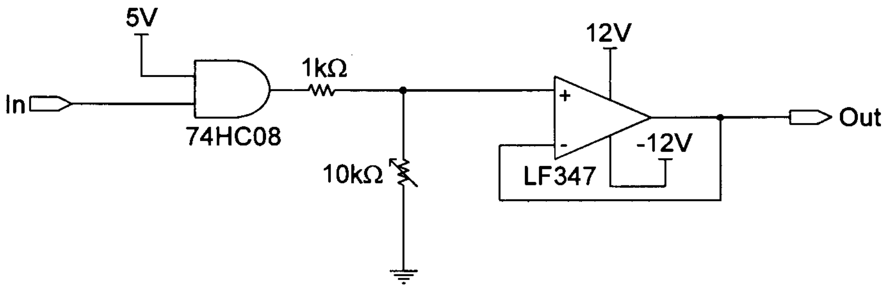


Figure 2.24 Discrete NMOS source driver circuit.

The NMOS source driver circuit in Figure 2.24 uses the CMOS input gate to buffer the logic input signal to a signal with rail-to-rail swing. The amplitude of this signal can be adjusted with the voltage divider and is then buffered to the output through the voltage buffer.

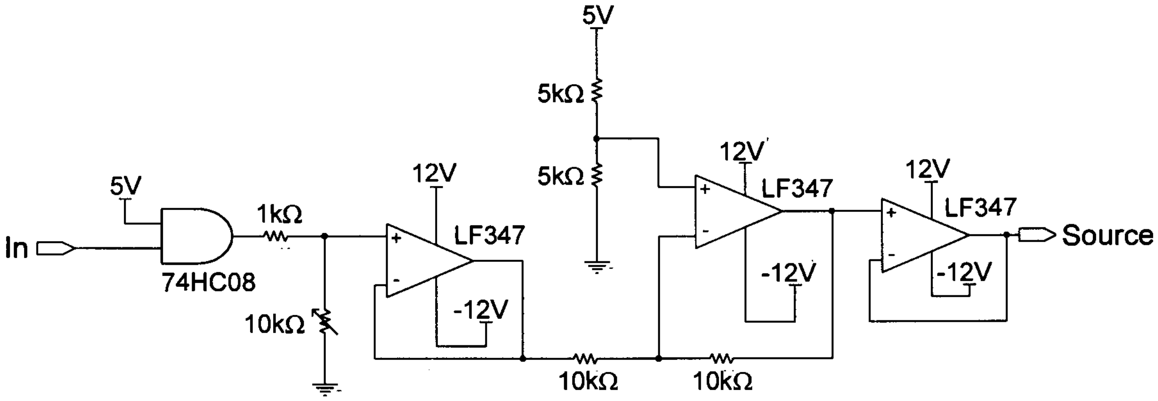


Figure 2.25 Discrete PMOS source driver circuit.

For the PMOS source driver shown in Figure 2.25 the input signal is once again buffered and the amplitude adjusted to the desired level by the adjustable voltage divider circuit. The signal is fed into a differential amplifier with unity gain through a voltage follower. The amplitude-adjusted input signal is subtracted from 5V and buffered through a unity-gain voltage follower to the output.

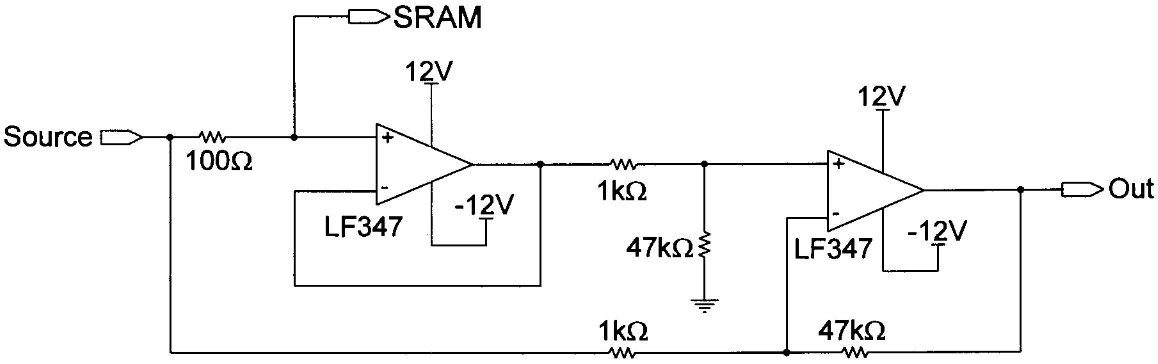


Figure 2.26 Discrete current-to-voltage converter.

The current-to-voltage converter circuit of Figure 2.26 is attached between the NMOS source driver and the SRAM. The current out of the SRAM flows through the 100Ω resistor. The side of the converter attached to the RAM is buffered so that the current required by the differential amplifier does not influence the current through the sense resistor. The resistor value is chosen as 100Ω because this gives rise to a voltage drop of 5mV at 50μA. This voltage drop is large enough to sense but not large enough to influence the operation of the SRAM array. A differential amplifier with a gain in the region of 50 amplifies the differential signal across the resistor to a detectable level.

To drive the SRAM, a word generator capable of generating a sequence of 32 words that are 8 bits wide was used. The 2x2 array requires 6 bits (2 *RW*, 2 *CL* and 2 *D/O*). As already mentioned the clear of the cell is accomplished using the free PMOS source nodes. Each control word has to be isolated from the next by a word containing only "zeros". This allows 16 actions to be performed. Both words are initially cleared by activating both *CL*-lines simultaneously. This procedure is verified by reading the words in succession by activating the respective *RW*-lines. After reading both words, the word read first is read again so that it may be verified that reading the words did not affect their contents. Next the first word is written with data '10' by activating the corresponding *RW*-line and *D/O*-line. The write procedure is verified by reading the word (activating *RW*-line). To verify that writing and reading did not modify the other word it is also read and the first word is read once more. The second word is written with data '01' and an identical verification procedure is used. In the final cycle one word is cleared and the effect on the array verified.

The two plots in Figure 2.27 were captured from the oscilloscope and show that the SRAM array operates correctly. Except for the *CL*-signals, the signals indicated in these plots are identical to those of Figure 2.12. The current-to-voltage converter is connected to the *D/O*-lines. A pulse on the current-to-voltage converter output indicates a current is flowing. The presence of a current during a read cycle is an indication that devices *M2* and *M3* are on (see Figure 2.12), and is therefore an indication that the state of the cell is a logic "zero". The spikes present on the output are a result of unequal delays to the differential amplifier of the current-to-voltage converter. One signal path is directly connected to the differential amplifier and the other is buffered. This causes unequal delays if the common mode voltage of the two nodes of the resistor is changed. The spike is can also be observed when simulating the circuit shown in Figure 2.26.

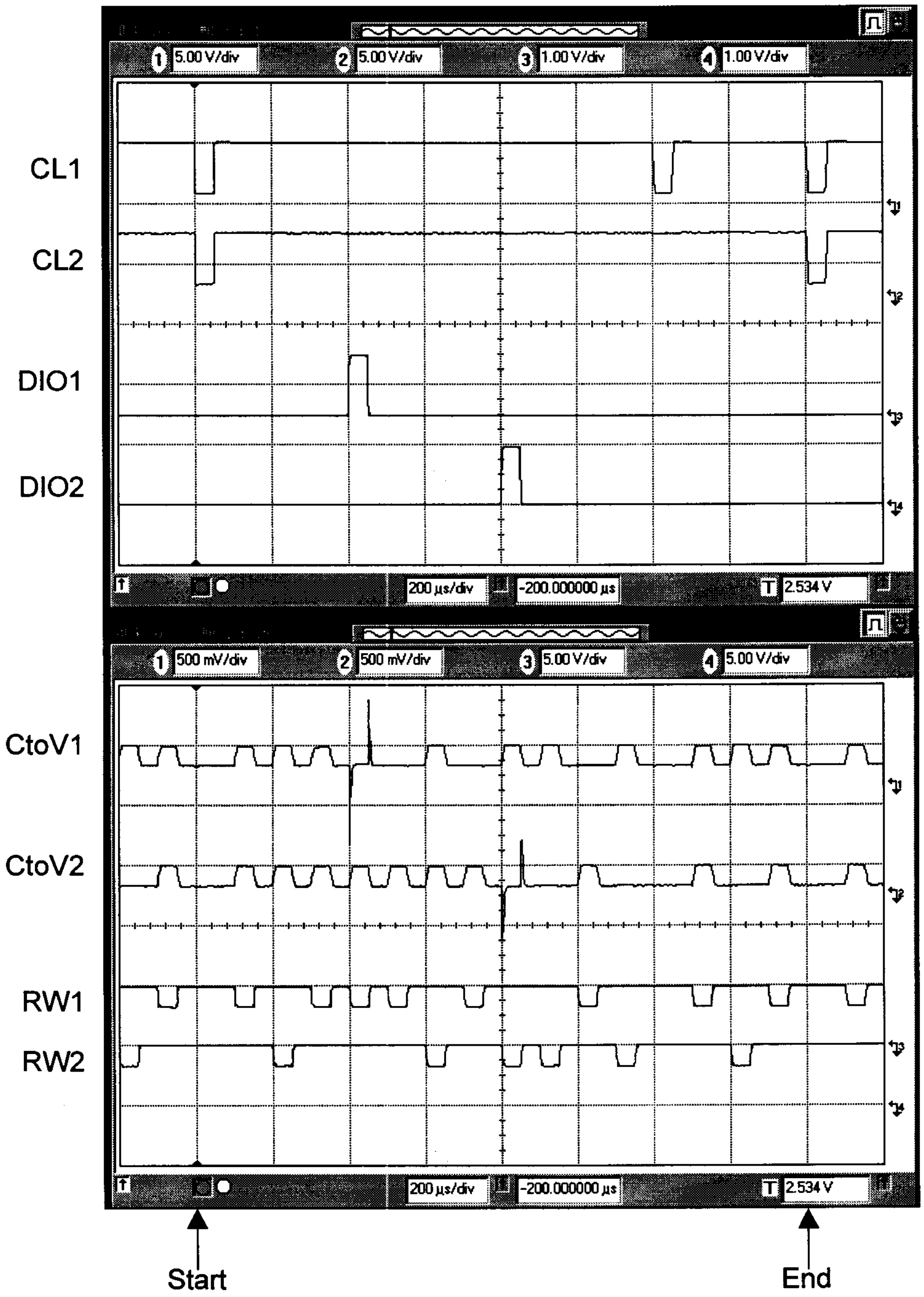


Figure 2.27 Experimental results for the 2x2 SRAM array showing the four described procedures between the "Start" and "End" indicators.

The plots also show the deviations used. The *D/O*-lines operate at a deviation of 1V and the *RW*-lines at 1.8V. For the clear signals the maximum deviation possible with the peripheral circuits was used (typically in the order of 4.5V). In order to verify that pulling an NMOS source node very high in voltage can also clear the cell, a second experiment was performed.

Exactly the same sequence as described above is used. Instead of tying the alternative NMOS source of the second bit of the words to ground it is connected to another NMOS source driver circuit (called *CLN*). Instead of activating the *CL1* line in the last cycle to clear the first word, the *CLN*-line is activated to a very high voltage (4.5V). This clears the second bits of both words and does not affect anything else in the array. The two clear lines (*CL1* and *CL2* in Figure 2.12) are connected together. The plots of this sequence are shown in Figure 2.28.

From the circuit diagrams of the discrete interface circuits it can be seen that the voltage levels are adjustable. This allowed some ranges of the deviations to be determined. A specific deviation level was adjusted in a certain direction until incorrect operation resulted (typically a certain bit not being written or cleared anymore or a certain bit being written or cleared when it was not supposed to be). Four chips were measured and the data averaged to obtain the results given in Table 2.3.

Table 2.3 Measured maximum and minimum deviation data.

Minimum required deviation on a PMOS node to flip the cell	2.72V
Maximum allowable deviation on a PMOS node not to flip the cell	2.65V
Minimum required deviation on an NMOS node to flip the cell	1.47V
Maximum allowable deviation on an NMOS node not to flip the cell	1.45V
Minimum required deviation on an NMOS node required to write the cell if a standard deviation is applied on the opposite PMOS node	0.46V

The minimum deviation of a PMOS node required to write the cell together with a standard deviation on the opposite NMOS node could not be measured, because the low *RW*-line deviation then makes it impossible to read the cell to verify what happened.

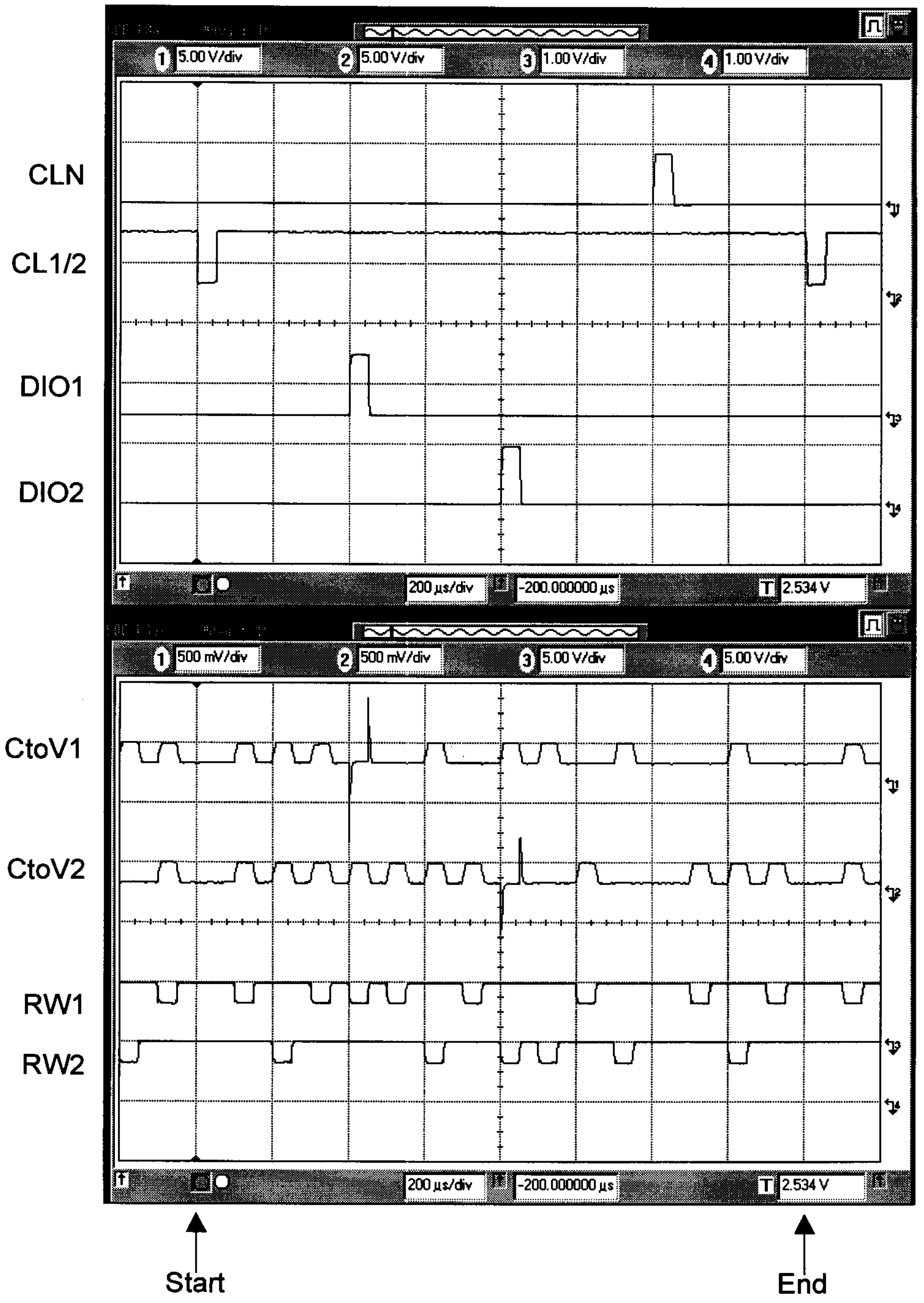


Figure 2.28 Experimental results for the sequence that tests clearing the cell via the NMOS source.

In order to compare the measured data to the simulated data the approximate location of the process on Figure 2.10 was measured. This was done by ensuring the cell is in a known state and deviating an NMOS and a PMOS node in such a way that the state does not change, but that a current flows in the opposite inverter. This current was measured and plotted against the gate-source voltage in similar fashion to Figures 2.20 and 2.21. This allowed the device quality of the measured chips to be defined relative to the five simulation models provided by the manufacturer. The measured NMOS characteristic was found to coincide with that simulated using the worst case zero model and that of the PMOS lies between the typical mean and the worst case one model. This indicates that the quality of both device types on the manufactured chip is poor. If the measured point were to be plotted on Figure 2.10 it would lie at the point 20μA/10μA (NMOS current / PMOS current), therefore closest to the worst case speed point.

Considering the deviation ranges measured against the theoretic ranges the overshoot present in the response of the operational amplifiers used needs to be considered. The flip of the cell when a single NMOS node is raised takes place around a deviation of 1.46V. This is lower than the 1.8V calculated using noise margin analysis (see Figure 2.19). The same is valid for the situation when a PMOS node is used. The flip takes place at a deviation of 2.7V instead of the expected 3.0V. Simulations of the discrete op-amp circuits together with the array confirm that there is approximately 0.25V overshoot present. The overshoot peak is in the region of 100ns wide, which is more than 50 times the width required by the cell (assuming a write time less than 2ns). The cell can therefore easily respond to the peak overshoot value. This falsifies the measured deviation ranges slightly. When adding the overshoot to the deviation, the experimental results agree well with the theory.

2.10 SIX-TRANSISTOR SRAM CELL COMPARISON

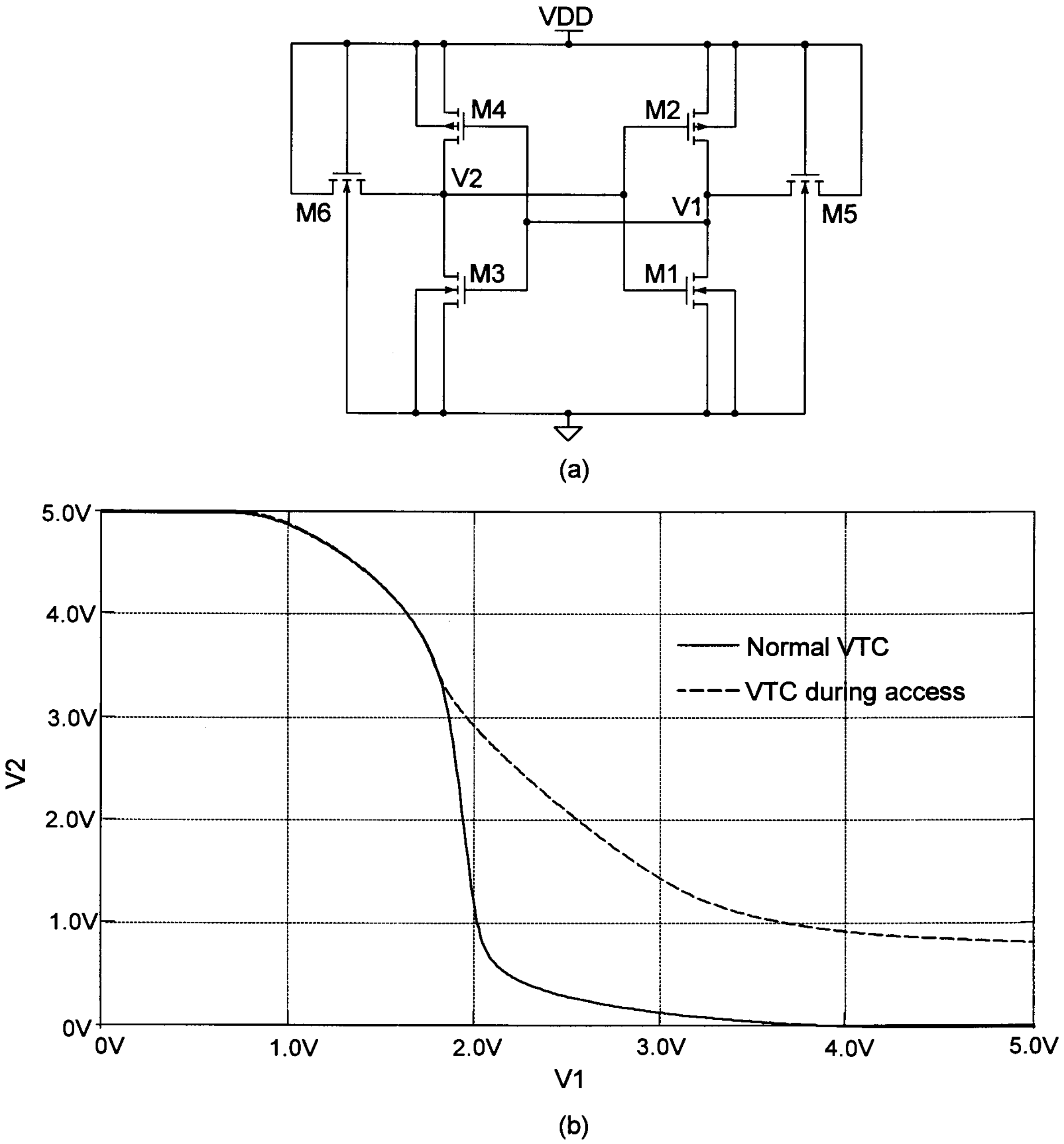
To end this discussion on the four-transistor SRAM cell, it needs to be compared to the six-transistor SRAM cell. Here it is important that as many design parameters as possible are equal for both cells. This allows a comparison of the cell areas to be based on two systems that have equivalent performance characteristics. It was decided to design the six-transistor SRAM cell to have the same noise margin as the four-transistor cell, because this is an important factor on which the design of the latter was based. The six-transistor cell was designed to have a typical noise margin in the order of 0.6V and an absolute worst case noise margin of at least 0.43V. These are the noise margins of the four-transistor cell given the following conditions:

- typical noise margin: 0.6V for a typical process and NMOS and PMOS source node deviations of 1V and 1.8V respectively,
- smallest noise margin: 0.43V for the worst case one model and NMOS and PMOS source node deviations of 0.85V and 2.0V respectively.

2.10.1 Noise Margin of the Six-Transistor SRAM Cell [10]

When considering the six-transistor cell, the noise margin under retention conditions is simply the noise margin of the unmodified cross-coupled inverter pair. For a cross-coupled inverter pair with unity device ratio, this noise margin is 1.39V given typical conditions, as can be seen from Figures 2.17 and 2.18. This value can be found by reading off the noise margin associated with *X*=0 and *Y*=0, because these are the values of the deviations during data retention. It is once again during the access that the noise margin drops. For the six-transistor cell only the read access needs to be considered. The write cycle does not affect any cells but the ones intended.

Just before the access transistors are turned on to initiate the read of the data in the cell, both bit lines are charged to an equal potential which is typically also close to *VDD*. Therefore, when the access devices are turned on, one of them shunts the pull-up device and the other weakens the pull-down device. For example, in Figure 2.29(a) the initial conditions of node *V1* and *V2* are "low" and "high" respectively. When turned on, the device *M6* shunts *M4* by assisting to pull node *V2* "high", and the device *M5* weakens *M1* by pulling node *V1* "high" against the action of *M1*. This modifies the voltage transfer characteristic as is shown in Figure 2.29(b).



(a)

(b)

Figure 2.29 (a) Six-transistor SRAM cell during initial read access and (b) the effect this has on the voltage transfer characteristic.

When the NMOS device is in cutoff the VTC is not modified but once the inverter PMOS enters cutoff and the NMOS the linear region, the diode connected access transistor causes current to flow. The weaker this device is the greater will be the voltage drop across it and the less deterioration in the noise margin will be present. This illustrates the theory that the access transistors are typically made weaker than the driver transistors to preserve noise margin [10]. What needs to be equal noise margins during the read access. This is done by designing the cell ratio, the ratio between the device sizes of the NMOS driver transistor and the access transistor. The noise margin calculation algorithm is utilised to plot the noise margin from a set of transfer characteristics as a function of the cell ratio. A set of inverter characteristics similar to the one shown in Figure 2.29(b), with varying cell ratio, is used as an input. The C-code for this program is given in addendum A.2 of this dissertation. Figure 2.31 shows the results obtained for the different simulation models.

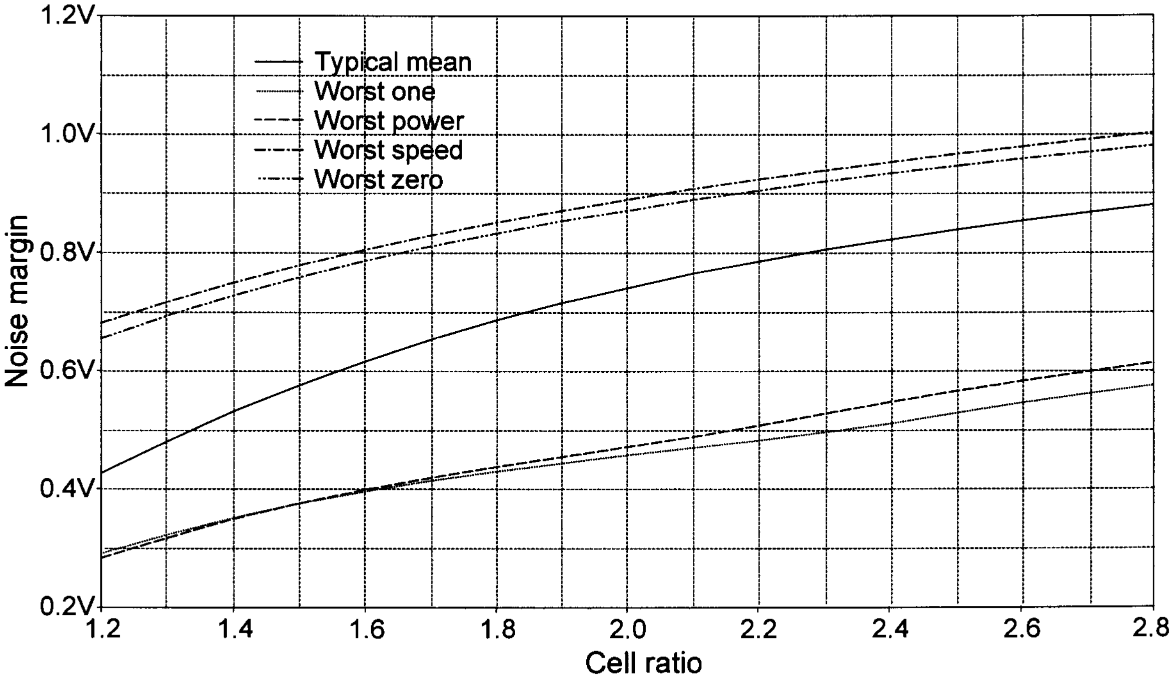


Figure 2.31 Static noise margin of a six-transistor SRAM cell during read access as a function of cell ratio for different simulation models.

As can be seen, a cell ratio of 1.55 guarantees the 0.6V static noise margin for the typical mean process. This however means that the noise margin of the cell for worst case power and worst case one conditions is very low. To raise these noise margins to the 0.43V level the cell ratio has to be increased to 1.8. This correlates well with the typical choice of around 2 [10].

The inverter devices have dimensions of 1.4μmx0.6μm. This means the access devices require dimensions of 1.4μmx1.1μm. This guarantees equal noise margins to the four-transistor cell, as well as static and dynamic write conditions and completes the design of the six-transistor SRAM cell.

2.10.3 Transient Simulations

Dynamic write conditions can be tested via simulation. The cell is initialised in a defined state and written to the other state by pulling the corresponding bit lines "low" and "high". The access transistors are activated and the internal state of the cell is observed. In order to simulate the read cycle specifications the access transistors are activated after the nodes have been precharged. The output can be a differential current or a differential voltage. Therefore bit line capacitance has to be added. A typical value is 0.5pF. Rise and fall times of all signals are once again taken as 1ns.

The cell is operational across all process variations. The military specification temperature range was also simulated with similar results as for the four-transistor cell, namely that the functional operation is not affected, but the cell does tend to become slower as the temperature increases. This is an indication of the fact that the degradation in mobility has more influence on the operation of the six-transistor cell than the decrease in threshold voltage.

Table 2.4 shows the simulated characteristics. The write time is considered as the time difference between the 50% level of the word line control signal and the time where the internal cell voltages are equal. Two read access times are specified because two methods of sensing the cell exist. The differential voltage sense time is taken as the time between the 50% level of the word line and a differential voltage of 1V. This value is chosen because it should allow good sensing given a large differential mode voltage as well as a common mode voltage adequately distant from the power supply. The differential current flowing into the cell as one bit line is discharged can also be sensed. This current is initially constant because the access transistor is in saturation. The current mode read access time is the time difference between the 50% levels of the word line input and the differential current output.

Table 2.4 Simulated specifications for the six-transistor SRAM cell.

Model type	Write time (ps)	Voltage mode read access time (ns)	Current mode read access time (ps)
Typical mean	435	1.64	155
Worst case one	290	1.35	135
Worst case power	302	1.14	140
Worst case speed	489	2.25	170
Worst case zero	489	1.85	166

2.11 COMPARISON BETWEEN THE FOUR- AND SIX-TRANSISTOR CELLS

2.11.1 Speed

Due to the high bit line capacitance the voltage mode access times of the six-transistor cell are high. On a more comparable level the current mode access times are substantially faster than for the four-transistor cell, as are the write times. The fact that the write times are longer than the read times is identical to the four-transistor cell. Here it has to be mentioned that the current mode read access times do compare well to the clear times of the four-transistor cell. As a whole the six-transistor cell does seem faster. This will have to be further investigated in the system environment rather than on a stand-alone cell basis.

The slower operation of the four-transistor cell is due to the fact that the control voltage deviations are small. This creates a small difference between the gate-source voltage and the threshold voltage of the devices (over-voltage) and causes smaller currents. It also has to be mentioned that the supply voltage reduction present in the four-transistor cell also slows down the circuit. Some of this speed loss may however be made up when implementing a system because the smaller control voltage deviations take place faster if rise and fall rates stay constant.

2.11.2 Power Dissipation

The six-transistor cell does not suffer from high internal currents. The power dissipation of the cell itself is restricted to the switching currents. Significant current does however flow when the bit line is discharged during reading.

The four-transistor cell has similar switching currents and smaller read currents. But as already discussed the wasted write currents will definitely penalise this SRAM configuration in terms of power dissipation, especially due to the fact that these currents do not serve any purpose. High currents may also occur in the six-transistor SRAM system when bit line voltages need to be changed. These currents do however serve the purpose of bringing the bit lines to the correct voltage required for operation of the system.

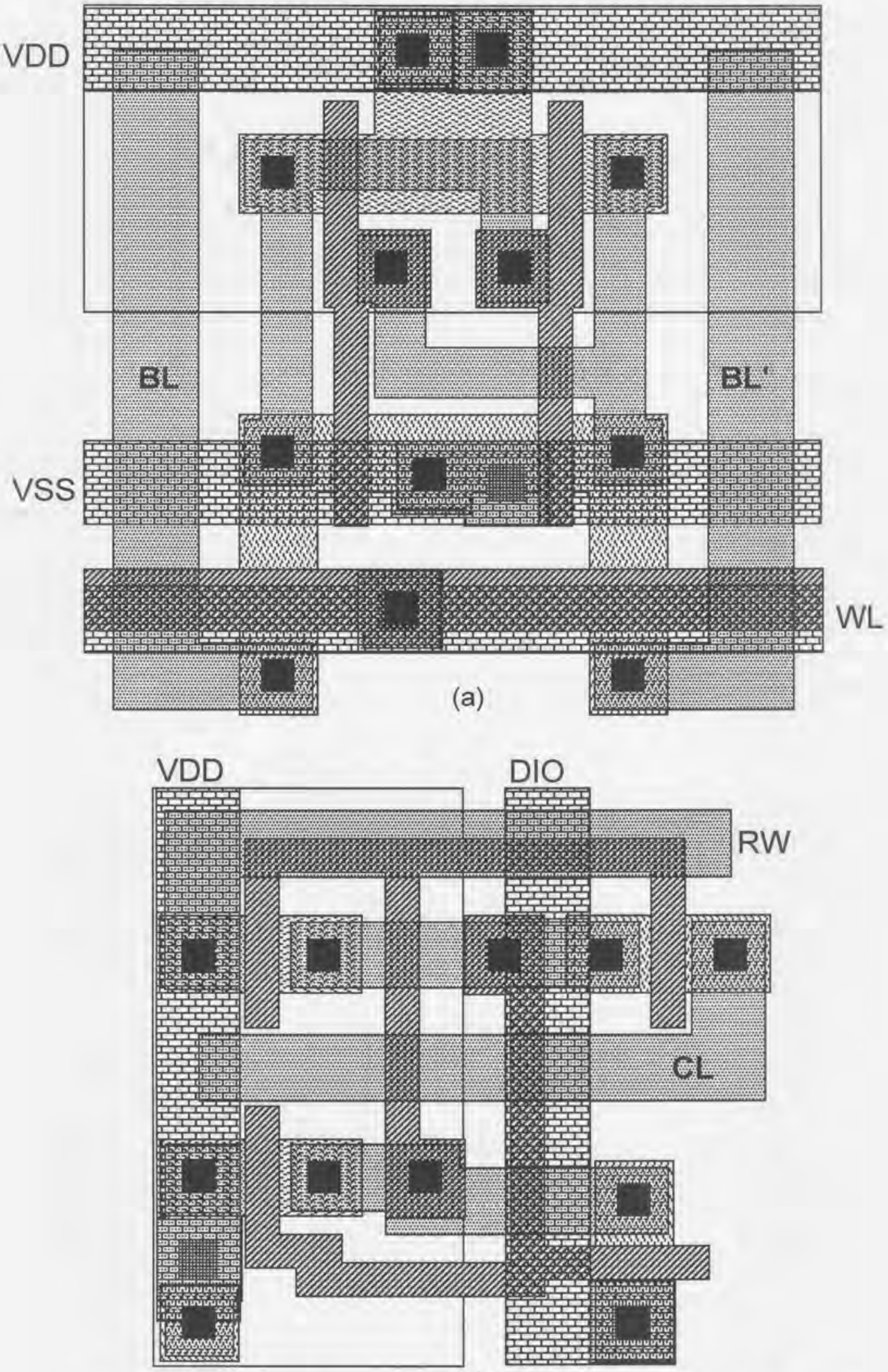
2.11.3 Cell Area

The advantage of the four-transistor SRAM cell lies in the fact that the access transistors are omitted. This allows a smaller cell area. A layout for each cell is shown in Figure 2.32, while Table 2.5 summarises the characteristics of the layouts. The following constraints were applied to both layouts:

- Nodes and lines common to adjacent cells may be shared. The NMOS-substrate contacts may be placed at regular intervals throughout the array, but this distance is large so it is neglected when considering the cell size. Therefore the VSS line for the four-transistor cell need not be routed.

- Those signals routed across the array (common to all bits of a word) are routed in a 1.2μm metal and those signals routed vertically in the array (common to a specific bit of all words) must be routed in a 1.5μm metal. The latter signals travel longer distances and have higher capacitance associated, so a wider track was chosen.

- For obvious reasons an array of cells must adhere to all geometric and electric constraints.



(a)

(b)

Figure 2.32 Layouts of the (a) six- and (b) four-transistor SRAM cells (Legend in addendum B).

The layout of the six-transistor cell shares the *VDD* line together with a butted N-Well substrate contact. One diffusion of each access transistor is also shared. A line routed in metal layer 2 straps the poly-silicon word line. For the four-transistor cell the external nodes *CL*, *RW* and *DIO*, as well as the *VDD* line and substrate contact, are shared among adjacent cells. The N-Well substrate contact is included as part of each cell because the process design rules require small spacing between them so that a contact is required every four cells. Having a dedicated contact channel would require more space than including one as part of every cell. The layouts clearly indicate that sharing internal diffusion (*VDD* and *VSS*) area is not possible with the four-transistor cell as it is for the six transistor cell. Therefore it is possible to share every external node.

Table 2.5 Comparison between the layouts for the two SRAM cells.

Characteristic	Four-transistor cell	Six-transistor cell
Cell dimensions (H x W)	9.6μm x 9.4μm	11.2μm x 13μm
Cell area	90.24μm^2	145.6μm^2
256x32 Array dimensions (H x W)	2457μm x 338μm	2867μm x 416μm

The reduction in cell area associated with the four-transistor SRAM cell is 38.02%. This is a significant improvement over the 14.7% reduction achieved using the initially proposed array structure [2]. Given that one line fewer needs to be routed compared to the initially proposed array structure, this is less than should be expected. The comparison made in [2] however, uses a six-transistor SRAM cell that has a significantly higher noise margin compared to the four-transistor cell due to a far greater cell ratio (16.7). This results in larger layout for the six-transistor cell, and an overestimation of the reduction in area. The comparison for this dissertation has been based on two cells with equal characteristics. Considering the layout in Figure 2.32(b) it is also evident that there is sufficient area left to route an extra line in the vertical dimension if it were required. When comparing array sizes, it can be seen that a reduction in height and width has been achieved by using the four-transistor SRAM cell.

2.12 CONCLUSION

This section covered all aspects of the four-transistor SRAM cell. Initially the current published knowledge about the cell was analysed. Some problems with operation were identified and a new array structure, which is based on a new method of writing data to the cell, was proposed. The noise margins and reliability of the cell were analysed and the voltage deviations were designed by making use of the results thereof. A six-transistor cell was designed for an identical noise margin. Regarding performance, the latter is faster and consumes less power but it is larger. An acceptable reduction in cell area was achieved. Some performance characteristics (mean values) of the new array structure and cell are:

- 38% reduction in area compared to the six-transistor cell,
- sub-nanosecond read, write and clear times,
- 0.6V noise margin at 5V power supply (compared to 0.43V at 1.8V power supply for the loadless four-transistor SRAM cell [8] using a low threshold voltage process together with a high threshold voltage option on the NMOS),
- suitable for a standard 5V CMOS process with no extra processing steps,
- 87.5% reduction in wasted power compared to previous design [2],
- one line fewer to route compared to previous proposal [2].

3. SOURCE DRIVER CIRCUITS

3.1 INTRODUCTION

The previous chapter dealt with aspects of the four-transistor cell itself. A set of voltage deviations, as well as a scheme of using these deviations to access the cell, was devised. These results may be used for designing the voltage references and the low-impedance drivers.

Three circuits are required, one for each of the three control nodes of the four-transistor SRAM cell. The important aspects of the design of these drivers, as well as simulation results, are given in this chapter. The designs necessitated several choices to be made. These are also explained.

3.2 FUNDAMENTAL PRINCIPLES

The three driver circuits have two characteristics in common:

a. They need to present a low impedance in the off-state. The off-state of a driver circuit is defined as that state when the node which is being driven, is connected to the power supply. Depending on the type of driver, it may need to supply the read current, wasted write current, transient switch current or different combinations of these. This current causes a voltage drop across the internal resistance of the driver. The maximum allowable voltage drop and the magnitude of these currents together define the maximum allowable output impedance in the off-state. Because a large voltage drop across the internal output resistance of the driver circuit will cause the noise margins to degrade, it was decided to limit this voltage to 0.1V. This is 10% of the smallest voltage deviation and is small enough not to cause significant noise margin problems, but also large enough that very low output resistance is not required.

b. When a driver circuit is activated, only transient currents need to be supplied. This is due to the fact that the inverter branch to which a voltage deviation is being applied, always has one transistor in cutoff. This transistor may be the one on whose source the deviation is applied. In this case no variables in the SRAM cell change, and no static current can reach the driver circuit. If the transistor whose source is being driven, is in the on-state, the voltage deviation is transmitted to the internal SRAM node. This will cause a static current to flow in the other inverter, but the complementary device in the same branch as the one being driven remains in cutoff and no static current is possible. Only if the state of the cross-coupled inverter pair changes will a short current peak flow in both inverters. This transient switch current will have to be supplied by the driving circuit. The other transient currents that need to be supplied or sunk, are those required to charge and discharge the capacitance associated with a node, while a voltage change takes place. Because no static current flows while a deviation is applied on a specific source node, the driver circuit sees that node as a pure capacitance. During the design of the on-state of the circuits, the source nodes can therefore be modelled as a capacitance as far as switching behaviour is concerned.

From these characteristics it can be derived that the drivers are best implemented by switching the source nodes between two defined voltages. This allows a large device to connect to the respective power supply in order to create the low output impedance in the off-state. A second switching transistor is used to connect the node to a predefined voltage when it needs to be deviated from the standard power supply. This part of the circuit may have a larger output impedance because no large static currents are present. The output impedance here is determined by the rate at which the capacitance must be charged.

It was already mentioned that the SRAM system comprises 1024 words of 32 bits each. The *RW*-line drivers, as well as the *CL*-line drivers, therefore need to drive 32 cells each. This represents a manageable capacitance and current. The *DIO*-line drivers, however, need to drive 1024 cells each. This is 32 times more than the other two driver circuits. Apart from this, the *RW*-driver and the *CL*-driver never drive more than one word at a time. This is not so for the *DIO*-driver because the number of *DIO*-lines deviated during a single write cycle depends on the data values being written. It was therefore decided to split the memory array into four independent banks of 256 words each. This lowers the bit line capacitance, as well as the currents required for charging and discharging the *D/O*-lines, by a factor four and eases the design of the driver circuits. Splitting an array into banks is a common method of increasing the speed by reducing the capacitance [20]. To further decrease the load, it was decided to also implement four independent driver circuits, each of which are therefore only attached to eight *D/O*-lines.

A positive spin-off of dividing the array into banks, is the fact that no more than one quarter of all cells have their *D/O*-lines activated, and can potentially waste current while a word is being written. This implies that the power dissipation due to wasted write currents is also reduced further by a factor four.

3.3 *D/O*-LINE DRIVER CIRCUIT

3.3.1 Overview

A block diagram of the driver circuit is given in Figure 3.1. There are essentially three parts to the circuit, the voltage reference, the low-impedance driver circuit and the switching circuit. The first part generates the required deviation. This is specified as 1V for a typical mean process, and it should vary approximately 0.15V in either direction as the quality of the NMOS devices changes.

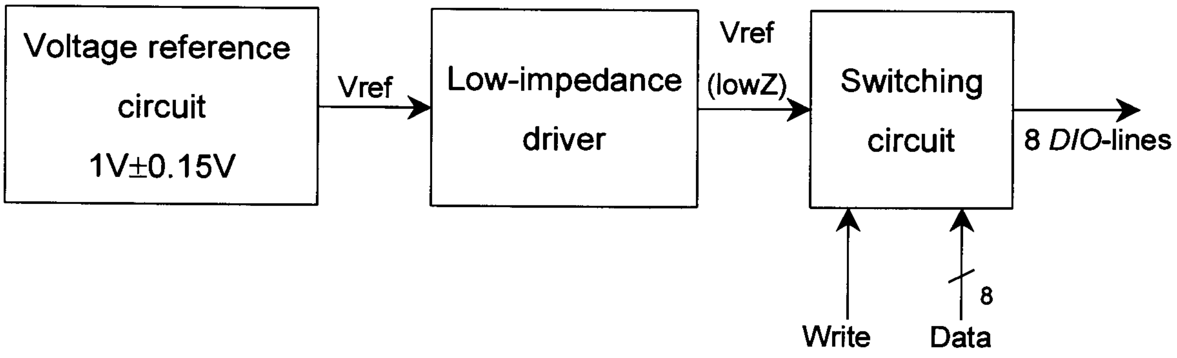


Figure 3.1 Functional block diagram of the *D/O*-line driver circuit.

The low-impedance driver circuit assures that the capacitance associated with the *D/O*-lines can be switched between 0V and 1V at the required speed. Rise and fall times of the signal should approach the read and write times of the cell, but any specification of less than 5ns was deemed to be sufficient for a first iteration. Here it needs to be mentioned that very short rise and fall times require high charge and discharge currents and therefore large driver circuits. This aspect compromises the area advantage present in using the small four-transistor cell. The output of this driver circuit is a buffered version of the reference voltage. The buffer circuit therefore has to be process independent, so that the carefully tuned process dependent reference voltage is not changed.

The switching circuit selectively connects various *D*/*O*-lines to the low-impedance driver. This connection must be established if a "one" needs to be written to a specific bit of the word being addressed. Therefore the *D*/*O*-line must be driven to the reference voltage if the write strobe is activated and the data input on a specific bit is "high". This switching circuit contains the pull-down device which ensures that a specific *D*/*O*-line is always connected to ground via a very low impedance, unless it is being deviated.

3.3.2 Line Capacitance

Before the circuit can be designed, some characteristics of the load which the *D*/*O*-line presents to the driver circuit, need to be investigated. The maximum total capacitance associated with a single cell is dependent on the state of the cell, and that state is data dependent. The capacitance associated with a specific source node is dependent on whether the transistor at that node is on or off. If the device is off, the node capacitance is that of the source-bulk pn-junction. If the device is on, the two drain-bulk capacitances of the NMOS and PMOS devices, as well as the gate input capacitance of the other inverter, need to be added to this. A worst case design has to be followed to ensure that the system is functional even under worst case data conditions, so it has to be assumed that all cells connected to a specific line present their worst case loading. The capacitances can be calculated from the device dimensions and the process data [19]. To calculate the gate capacitance the device dimensions and the gate capacitance per unit area are used. The drain-bulk and source-bulk capacitances are calculated using the equation [13]

$$Cj = \frac{C_{j0}}{\sqrt[M]{1 - \frac{V_B}{\psi_0}}}. \tag{3.1}$$

For the equation C_{j0} is the zero bias junction capacitance, V_B the bias voltage of the junction, ψ_0 the built in potential and M the grading coefficient. The last mentioned variable is an indication of the abruptness of changes in the impurity concentrations at the junction. This equation is used to calculate the capacitance of the drain and source diffusions by applying it to the area as well as the side wall. These two parts of the diffusion have different parameters. The capacitance is dependent on bias conditions. The smaller the reverse bias voltage, the larger the junction capacitance is. Because the potential of nodes varies it is best to use the largest possible capacitance, namely that at V_B = 0, and equation (3.1) reduces to C_{j0}. Using the dimensions of the devices and assuming typical process data, the total capacitance per cell is calculated to be 14.4fF. Considering that there are 256 cells connected to one line, the total capacitance per line amounts to 3.68pF.

Considering that for long lines the metal interconnect capacitance can become quite significant, it has to be added to the capacitance of the line. The length of the *DIO*-line on metal layer 2 is 2.5mm at a width of 1.5μm. The area and fringe capacitances per unit dimension are 0.032fF/μm^2 and 0.05fF/μm respectively. The line therefore contributes 370fF, making the total capacitance 4.05pF.

All *DIO*-lines of one bank together present 129.6pF to the driver circuit. This is the reason it was decided to split the driver circuit into four independent circuits, each of which drives eight *DIO*-lines. The total capacitance is thereby reduced to 32.4pF per circuit. Assuming a rise and fall rate of 1V/ns, this equates to a transient current peak of 32.4mA. At this point the advantages associated with reducing the peak current are evident, given the fact that it was reduced by a factor 16. High peak currents require wide tracks that waste area and add capacitance.

3.3.3 Currents

In the off-state each line switching circuit must be able to cope with the read current of one cell (45µA), as well as the transient current peak present when the cell is cleared (200µA). There is never more than one word accessed at once so the static and transient currents to cope with in the off-state are small. In the on-state the circuit also has to sink the transient current peak that flows in the inverter while the cell is switching state (200µA).

The read current, as well as the transient current peaks, are very small compared to the current peaks required to charge and discharge the capacitance. The situation on which the design has to focus is therefore the charging and discharging of the large capacitance. The small operating currents will have little or no effect because of the high capacitance.

3.3.4 Switching Circuit

The basic schematic for the switching circuit is shown in Figure 3.2. It consists of two wide devices, one of which is connected to the reference voltage and the other to ground. One of them is always on, and this allows the *D/O*-line to be switched between the two voltages.

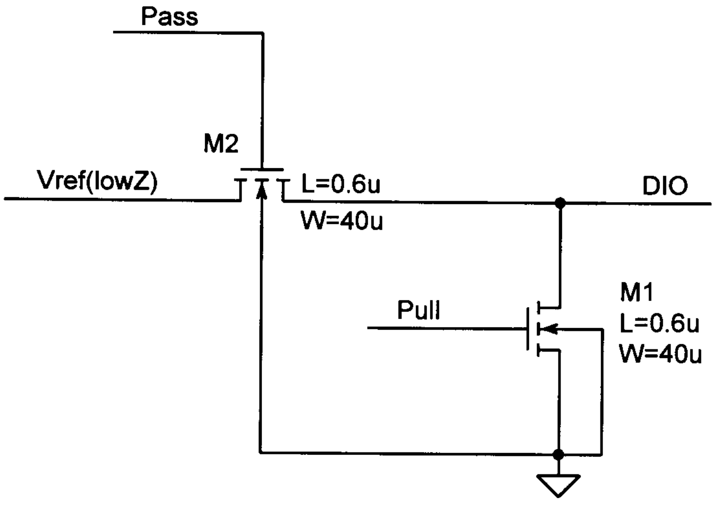


Figure 3.2 Circuit to switch the *D/O*-line between ground and the reference voltage.

To design the device dimensions their resistance has to be calculated. While the capacitance is discharging through the pull-down device *M1*, the circuit can be modelled as an RC-circuit. The drain-source voltage of the device is the voltage of the *D/O*-line so it is always in the linear region, and its operation is therefore described by equation (2.1). As V_{DS} increases the rate of increase of I_D will decrease, which is equivalent to the resistance of the device increasing. The highest resistance is therefore associated with the highest *D/O*-line voltage. To achieve a discharge time of 1ns, the time constant of the RC discharge has to be 0.2ns. The resistance of the linear transistor should therefore not be more than 70Ω. Using equation (2.1) at a V_{DS} of 1V implies that the width has to be 33μm. This does however not include the short channel effect and other secondary effects, which can be quite dominant in a sub-micron circuit. By means of simulation, which includes secondary effects, it was decided to use W=40μm. From Figure 3.3 it can be seen that W=40μm results in the required 70Ω device resistance.

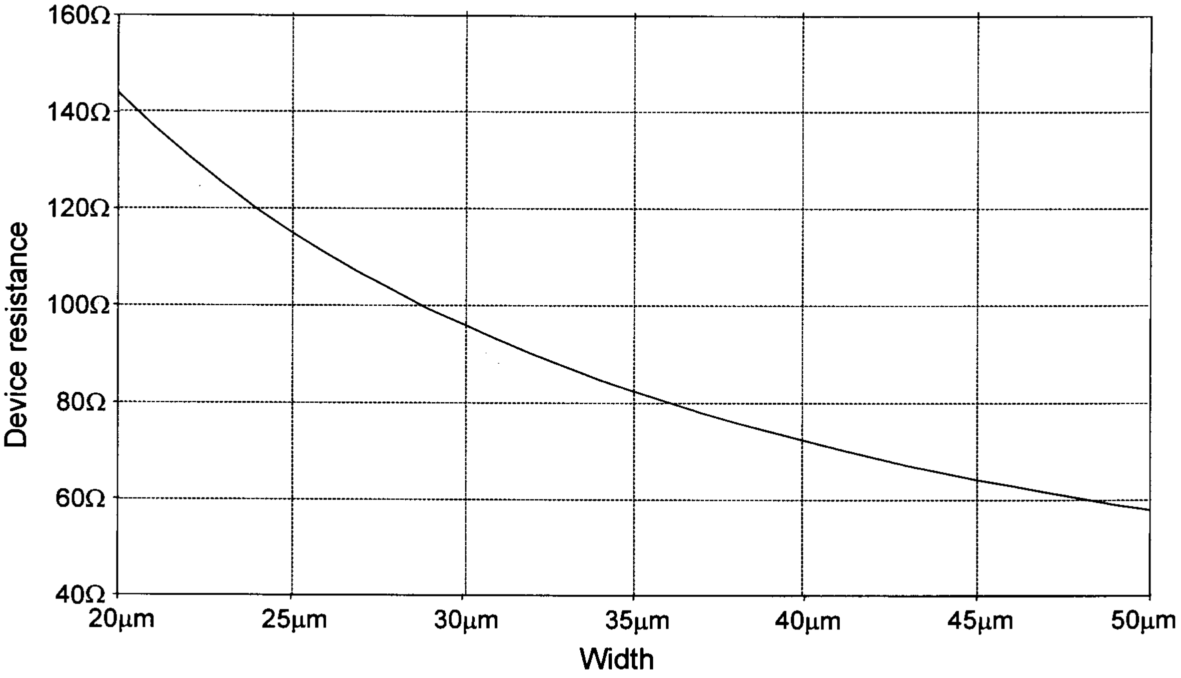


Figure 3.3 Resistance of the pull-down device as a function of the device width. The gate length is 0.6μm.

The pass device *M2* in Figure 3.2 follows the same design equation except that the resistance for low drain-source voltages is slightly higher due to the bulk-effect.

This is however still lower than the maximum resistance so the same width as the pull-down device is used. The size of this device is also less critical because, as will be seen later, the charging rate is limited by the output resistance of the low-impedance driver circuit.

The circuit of Figure 3.2 requires two control signals, where one is the inverse of the other. During the switching a problem can exist if both *M1* and *M2* are on at the same time, namely that there is a low resistance path from the reference voltage to ground. This creates an unnecessary load on the low-impedance driver in the form of the so-called short-circuit current which also wastes power, especially due to the large widths of the pull-down and pass devices. It is therefore not ideal to generate one control signal by inverting the other. A control circuit is required, where the falling edge of one control signal is the trigger to allow the other control signal to rise, as shown in Figure 3.4. This will limit the time where both devices are turned at the same time. Some conduction to ground will always occur but it is greatly reduced.

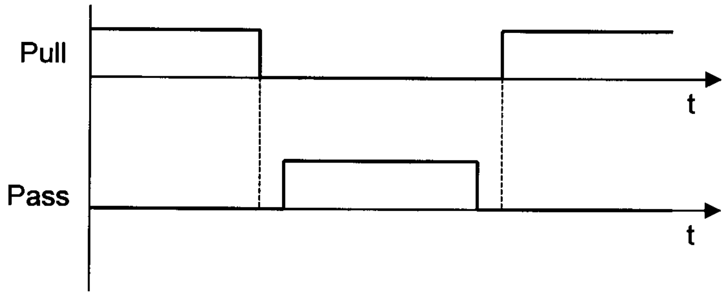


Figure 3.4 Timing pattern for the circuit in Figure 3.2 to limit short-circuit current and load on the driver circuit.

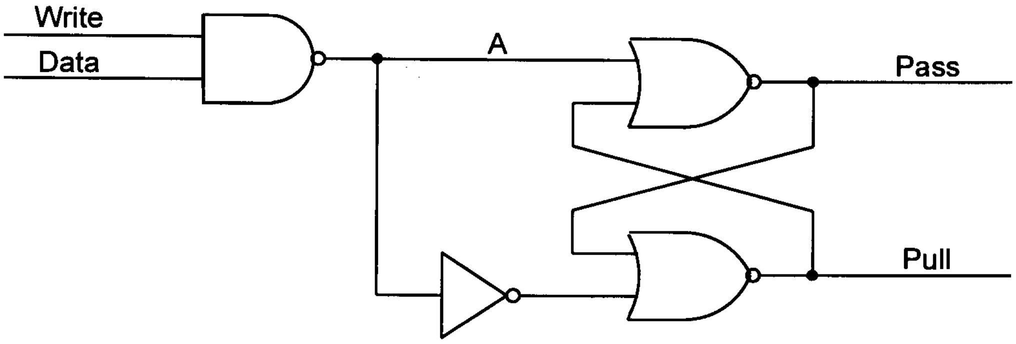


Figure 3.5 Control circuit to activate the switches of Figure 3.2 ensuring that short-circuit current is low.

The circuit of Figure 3.5 can be used, shown here on gate level for clarity. A transistor level circuit showing the device sizes is given as part of the full circuit diagrams in addendum C.

Node *A* will go "low" when the condition to activate the *DIO*-line is true, namely a write is taking place and the data bit is "high". The NOR-gates form a latch and the inverter ensures it is always being set or reset. The *Pull* signal will go "low" in response to *A* going "low" and the *Pass* signal will be activated via the feedback loop and therefore only rises in response to *Pull* going "low". During deactivation node *A* rises and forces node *Pass* "low". Now the *Pull* signal can only be changed via the feedback loop and will therefore change only in response to *Pass* changing. This circuit therefore cannot have *Pass* and *Pull* "high" at the same time, and this is what is desired.

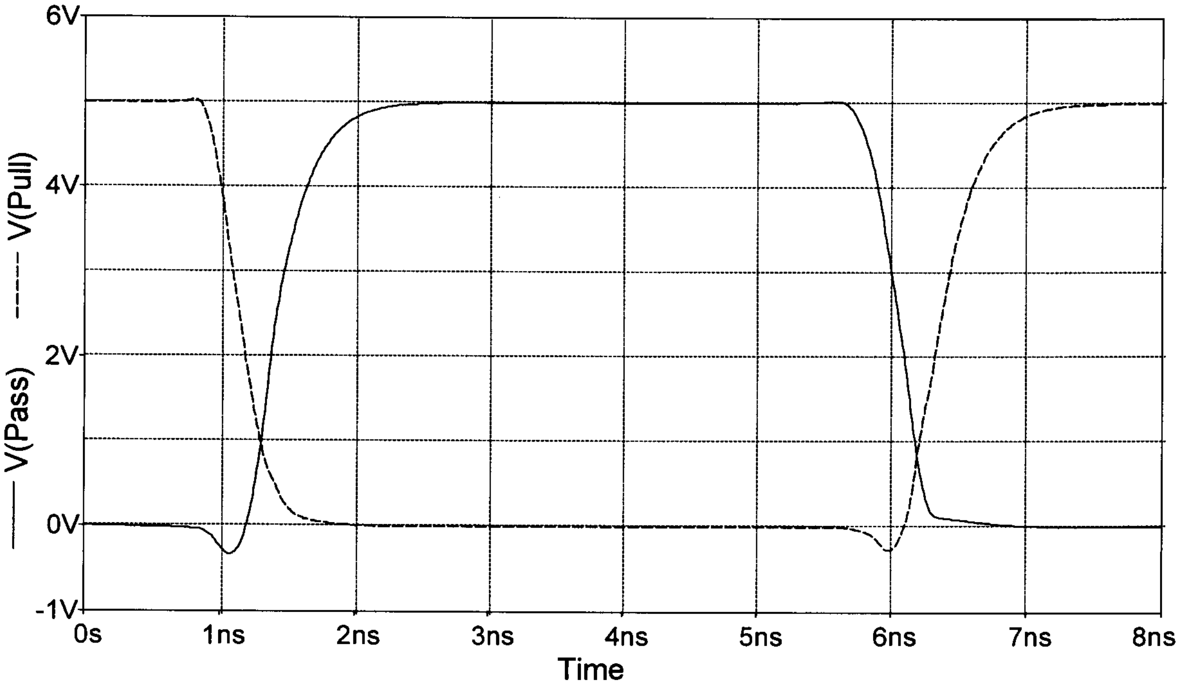


Figure 3.6 Simulation showing the correct delays between the *Pull* and the *Pass* signal.

It can be seen from the simulation results shown in Figure 3.6, that the correct sequence of the signals has been achieved. Figure 3.7 shows the simulated peak short-circuit current of the circuit in Figure 3.2 when inverter control signals and non-overlapping control signals are used. The reduction in peak current from approximately 4.0mA to 0.75mA can clearly be seen, and this minimises the loading on the low-impedance driver, as well as saving power. The time difference between the two peaks is due to longer delays in the non-overlapping control circuit.

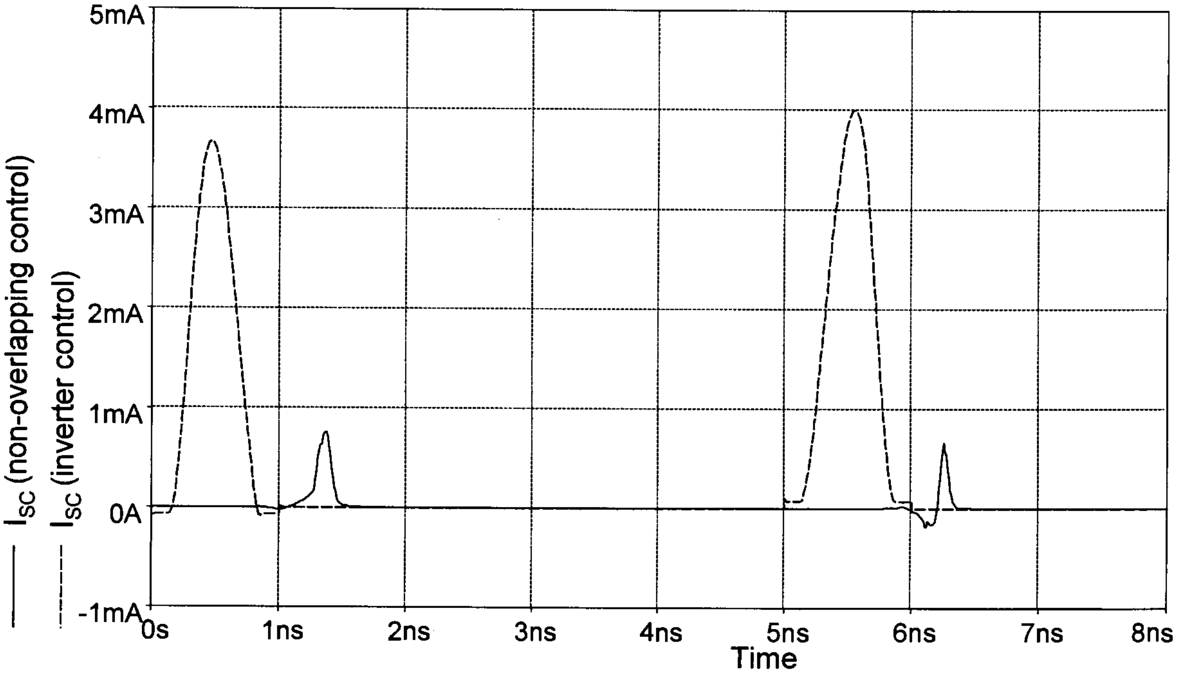


Figure 3.7 Simulation showing the difference in peak short-circuit current between the non-overlapping *Pull* and *Pass* signal control and inverter control.

3.3.5 Voltage Reference Circuit

Circuit Topology

The voltage reference circuit has to generate a voltage that is dependent on the quality of the NMOS device. The output voltage is specified as 1V for a typical quality NMOS, 0.85V for high NMOS quality and 1.15V for poor quality NMOS devices. The threshold voltage, V_T, for the process is specified as 0.6V, 0.72V and 0.84V for high, typical and poor quality devices. This indicates that the variation is 0.12V in either direction, which is very close to the specified 0.15V variation in voltage deviation. A circuit with the output of V_T+0.15V would therefore produce an output voltage very close to the specification.

According to Gray and Meyer [13], a threshold voltage reference can be implemented by steering a sufficiently low constant current through a diode-connected device with a sufficiently large *W/L* ratio. This device will then operate at a V_{GS} that is very close to its threshold voltage. Two aspects that have to be considered are the fact that 0.15V has to be added to the threshold voltage and that a process independent current has to be generated.

Consider the topology shown in Figure 3.8. All three devices carry an identical current *I*. The devices *M1* and *M2* have the same gate voltage, so their gate-source voltages differ by the voltage drop across the resistor *R*. The devices are saturated and the current-voltage relationship is described by equation (2.2).

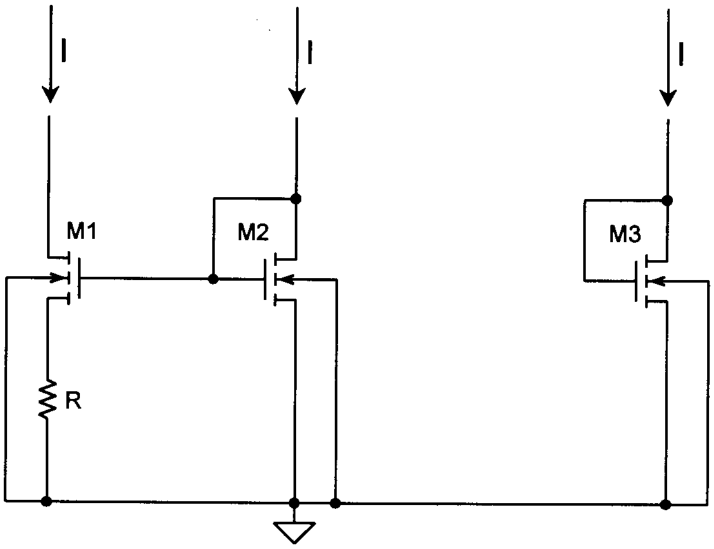


Figure 3.8 Basic topology of the voltage reference circuit.

Equating the gate voltages, and ignoring the bulk effect, as well as all other secondary effects, leads to

$$\sqrt{\frac{2I}{k'W/L_1}} + V_T + IR = \sqrt{\frac{2I}{k'W/L_2}} + V_T. \tag{3.2}$$

The threshold voltages can be cancelled. If the special case where $W/L_1 = 4W/L_2$ is considered, then 3.2 reduces to

$$\sqrt{2k'W/L_2\, I} = \frac{1}{R} = g_{mM2}. \tag{3.3}$$

The small signal transconductance of *M2*, g_m, is therefore independent of all parameters except the resistance *R*, and consequently the *M1-M2* configuration is typically referred to as a constant transconductance bias circuit. This is very useful for stabilising the performance of analog integrated circuits [21]. Because the device dimensions are fixed, the product *Ik'* is a constant. Device *M3* is also saturated and the gate-source voltage is given by

$$V_{GS-M3} = V_T + \sqrt{\frac{2I}{k' W/L_3}}\,. \tag{3.4}$$

Substituting *I* from equation (3.3) into this relationship produces

$$V_{GS-M3} = V_T + \frac{1}{Rk' \sqrt{W/L_3 \, W/L_2}}\,. \tag{3.5}$$

$V_{GS\text{-}M3}$ is in the form V_T*+constant*, and can be used as the required reference voltage output. The value that is added to the threshold voltage is however not invariant to process conditions. This introduces some error, but is not critical. As the NMOS quality increases due to the transconductance parameter increasing and the threshold voltage decreasing, these two effects tend to work together in equation (3.5). This will cause a larger variation of the reference voltage from the typical value of 1V, but the threshold voltage varies by 0.12V and 0.15V is required. The effect of *k'* can be limited by choosing large *W/L* ratios and resistance *R*. During the SRAM cell analysis it was shown that the aim should be to design a stable reference voltage. A maximum allowable variation as process conditions change was specified, but not to create reliable operation but rather to optimise circuit performance. As long as the range remains below the maximum specification, no serious situations arise.

Circuit Design

The complete circuit diagram for the voltage reference circuit is depicted in Figure 3.9. Some important aspects of the circuit are:

- In order to make the current in all three branches as equal as possible cascodes are used. This raises the output impedance of the current mirrors and improves their accuracy [13].

- To further improve accuracy the transistors are not chosen minimum length. Minimum length transistors exhibit a large variation in their characteristics, mostly due to the effective length of the gate. By choosing the length greater than the minimum, the percentage variation can be reduced. By making the gate longer, the impact of the short-channel effect is also limited. The overall behaviour of the devices becomes more predictable. The standard device size was chosen to be 20μmx1.2μm. The large width is required due to the stacking of four devices, each of which has to be in saturation for the circuit to operate. This requires the over-voltages to be small, and large device width at low current levels ensures this.

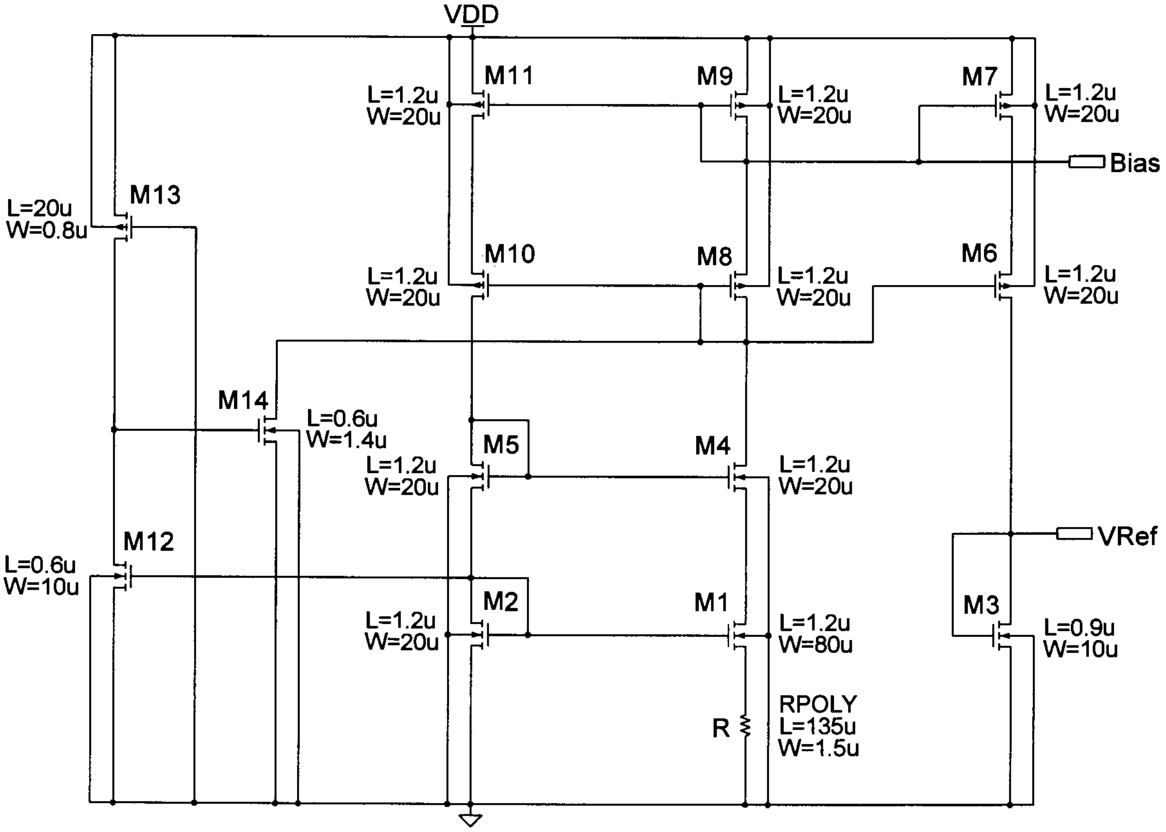


Figure 3.9 Constant transconductance bias circuit and reference voltage generator.

- The node named *Bias* is used to bias the low-impedance driver circuit. The device *M2* is biased at a constant g_m. The current through this device is *I*, the reference current. Assume this current is mirrored to another transistor using a mirroring ratio *P*. The *W/L* ratio of this new device is *Q* times larger than that of *M2*, where it is preferred that the length of the transistors remains constant so that the accuracy of the mirroring ratio is not compromised. Then the small signal transconductance of the new transistor is given by

$$g_m = g_{mM2} PQ \,. \tag{3.6}$$

  This implies that the transconductance of *M2* can be mirrored with any ratio to any other device in the circuit. Stable transconductances are important in ensuring stable circuit performance as the process conditions change.

- The 4/1 ratio between the *W/L* ratios of *M1* and *M2* can be seen. This was an assumption in deriving equation (3.5).

- The resistance is implemented using a poly-silicon resistor. This was done because this type of resistor has the smallest temperature coefficient, varies least as the process conditions change, and is not voltage dependent, as is the case for the well resistors. By designing the poly to be wider than minimum width, the tolerance can be reduced as well. This is more difficult to do using a well resistor, because the variation in effective width is larger. The sheet resistance is low (33Ω/square), so the resistor has to be quite long. The currents in the bias network should be low to limit static power dissipation, and a bias current of 25μA was chosen. From equation (3.3) it follows that the resistance should be in the region of 3kΩ. This translates to a length of 135μm if the width is taken as 1.5μm.

- The topology of the circuit is what is known as self-biased. It has a stable state where currents flow, but also a stable zero current state. To ensure that the circuit cannot start up in this zero current state, a start-up circuit is required [21]. This part of the circuit senses the gate voltage of *M2*. If this voltage is zero the bias circuit is in the zero current state and device *M12* will also be in cutoff. *M13* pulls the gate of *M14* high and turns it on. This pulls the drain node of *M14*, which is at *VDD* in the zero current state, down and turns on devices *M6* and *M8* to force the bias network out of the zero current state. Once this occurs, the gate voltage of *M2* increases and turns on *M12* which in turn turns off *M14*. The start-up circuit now has no influence on the bias network and only consumes the minimal current flowing in the *M12-M13* branch.

Simulation

The bias point simulation data is shown in Table 3.1 for different processes. Three simulation models, highest (WS), typical (TM) and lowest (WP) resistance, are supplied for the poly-silicon resistor. This results in 15 possible combinations that should be simulated. Because the reference voltage is critical, all these simulations were performed.

Table 3.1 Simulated reference voltage across the different process corners.

Transistor model Resistor model	TM	WP	WS	WO	WZ
TM	0.999V	0.845V	1.16V	0.855V	1.14V
WP	1.037V	0.875V	1.21V	0.887V	1.19V
WS	0.972V	0.822V	1.12V	0.831V	1.10V

It can clearly be seen that the desired objective of making the reference voltage dependent on the quality of the NMOS device, has been achieved. If the typical mean resistor model is used, the specified variation of 0.15V in either direction of 1V, is present. As can also be seen from equation (3.5), as the value of the resistance increases, the deviation voltage will decrease. This effect can clearly be observed, and does cause some deviations to be outside the specified range. This is once again not considered to be too serious, because the 0.15V specification was set as a guideline. Using Figure 2.19 it can be verified that static write conditions do still exist for those values outside the specified range. Bias currents for the complete bias circuit depend largely on the value of the resistance. A lower resistance causes a large increase in bias current. The complete circuit consumes on average 80μA for a typical resistance. This values goes to 60μA for the highest resistance and up to 100μA for worst case power resistance model. This is considered to be acceptable.

3.3.6 Low-Impedance Driver Circuit

Circuit Topology

The function of the low-impedance driver circuit is to buffer the voltage reference adequately, so that the *D*/*O*-lines may be driven at the required speed, without loading the reference voltage circuit. It has already been mentioned that the capacitance of a single *D*/*O*-line is 4.05pF, and that the total load for one *D*/*O*-line driver circuit is 32.4pF. To charge this total capacitance with the same time constant as the discharge cycle, namely 0.2ns, requires the driver circuit to have an output impedance of no more than 6.1Ω. Another aspect is that the output voltage has to be an accurate replica of the input voltage. To achieve these specifications a circuit topology like the one depicted in Figure 3.10 can be used.

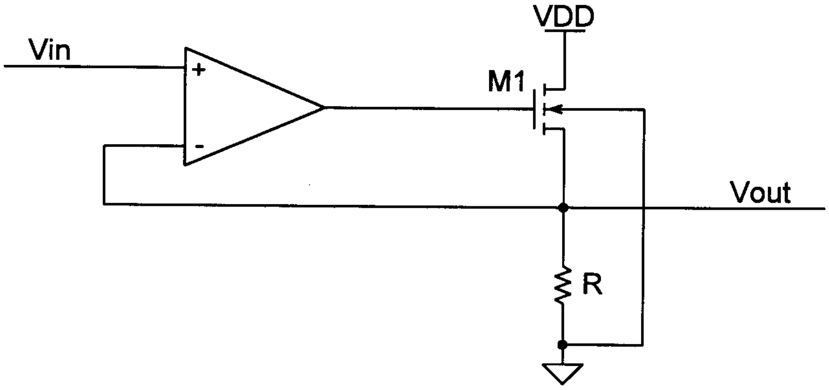


Figure 3.10 Proposed circuit topology for the low-impedance driver.

The negative feedback loop sets the output voltage equal to the input voltage, as long as the required gate voltage to *M1* can be delivered. This circuit consumes static power, because a current has to flow through the resistance in order to create the required output voltage. The difference between this output voltage and the maximum output voltage of the operational amplifier is the maximum possible gate-source voltage that can be applied to the transistor. This, together with the device dimensions and the current through the resistor, determine the maximum output current the circuit can deliver. The lowest output impedance is the inverse of the transconductance of the transistor in maximum current state. For a large transistor, this circuit can achieve very low output impedance, but only for sourcing current. If current needs to be sunk, the output voltage will rise and the transistor will turn off. The current can thus only be sunk via the resistor.

This circuit operates on the principle of negative feedback. Upon close inspection it can be seen that two negative feedback loops exist.

a. The operational amplifier forms a negative feedback loop. The input voltage is constant. Any change in the output voltage of the circuit, typically brought about by a change in load conditions causes the output of the operational amplifier to respond in such a way to oppose the change. For example if the output voltage rises due to a lower current requirement, the output voltage of the operational amplifier falls in response to a negative differential input voltage. This causes the V_{GS} of the transistor to decrease with the result that the current decreases. This decrease in output current satisfies the lower current requirement.

b. In order to identify the second negative feedback loop, consider that the output voltage of the operational amplifier is constant, at least over the time period being considered. If the output voltage of the circuit drops due to the sudden addition of uncharged capacitance, as is the case when a *D/O*-line is connected to its output, the gate-source voltage of the transistor will increase. This in turn increases the current supplied, and speeds up the rate of charging. The current reduces back to the static level once the capacitance is charged to the reference voltage level. The higher the voltage change due to adding uncharged capacitance, the larger will be the increase in the charging current.

The first feedback loop has significant delay, mostly due to the operational amplifier. The op-amp also has finite slew rates that slows down the response to a quickly changing input. This slow response typically causes overshoot, which is not desirable. Overshoot causes higher wasted write currents, lower noise margins and, if large enough, can even cause unintentional writes, as the *D/O*-line voltage becomes too high. The overshoot can be overcome by applying adequate frequency compensation, but this in turn slows down the charging rate. The second feedback loop however is only limited in speed by the inherent cutoff frequency of the device. This can be significantly higher than that of the operational amplifier.

The required speed dictates that the specifications for the operational amplifier that satisfies the requirements, are not realistic. An op-amp used to charge or discharge a capacitor within a certain given time, T_{ch}, should have a minimum unity-gain bandwidth of [22]

$$\omega_0 \geq \frac{15}{T_{ch}}. \tag{3.7}$$

In order to charge the gate capacitance of the driver transistor *M1*, in 2ns, requires a unity-gain bandwidth of 1.2GHz. There is some additional delay from the op-amp to the output so an even higher bandwidth is required. Combined with this, a high slew rate in the order of at least 1V/ns is required, so that the output can change fast enough. This set of specifications is unrealistic given the application.

It is far more advantageous to use the second negative feedback loop that is inherently fast. The response of this loop is immediate because the change is applied directly to the device that is responsible for countering it. The strength of the loop can be adjusted by adjusting the *W/L* ratio of the transistor. A weaker device requires a higher voltage change on the output node for the same change in current. The operational amplifier is used merely to present a high impedance to the reference circuit and to bias the transistor *M1* correctly. Depending on the quality of *M1*, its gate voltage needs to be set to supply the correct static current. Once this is set up, the second negative feedback loop is employed to charge the load capacitance. If the op-amp has a very slow response in comparison to the rest of the circuit, its output voltage does not change much as the load is changed.

Driver Design

The proposed configuration does however have the disadvantage that static current flows. Because the resistor is the only method of discharging the output node, the resistance has to be fairly small and therefore carries a high current. It would be advantageous to turn off this current when the driver is not being used. The turn-on has to be fast though, and when the large current is turned off, the output voltage of the op-amp has to be kept at the correct bias level. If this is not done, the purposeful slow response of the op-amp renders the driver circuit useless for a long time, because the output voltage will be incorrect. The circuit given in Figure 3.11 can fulfil the set requirements.

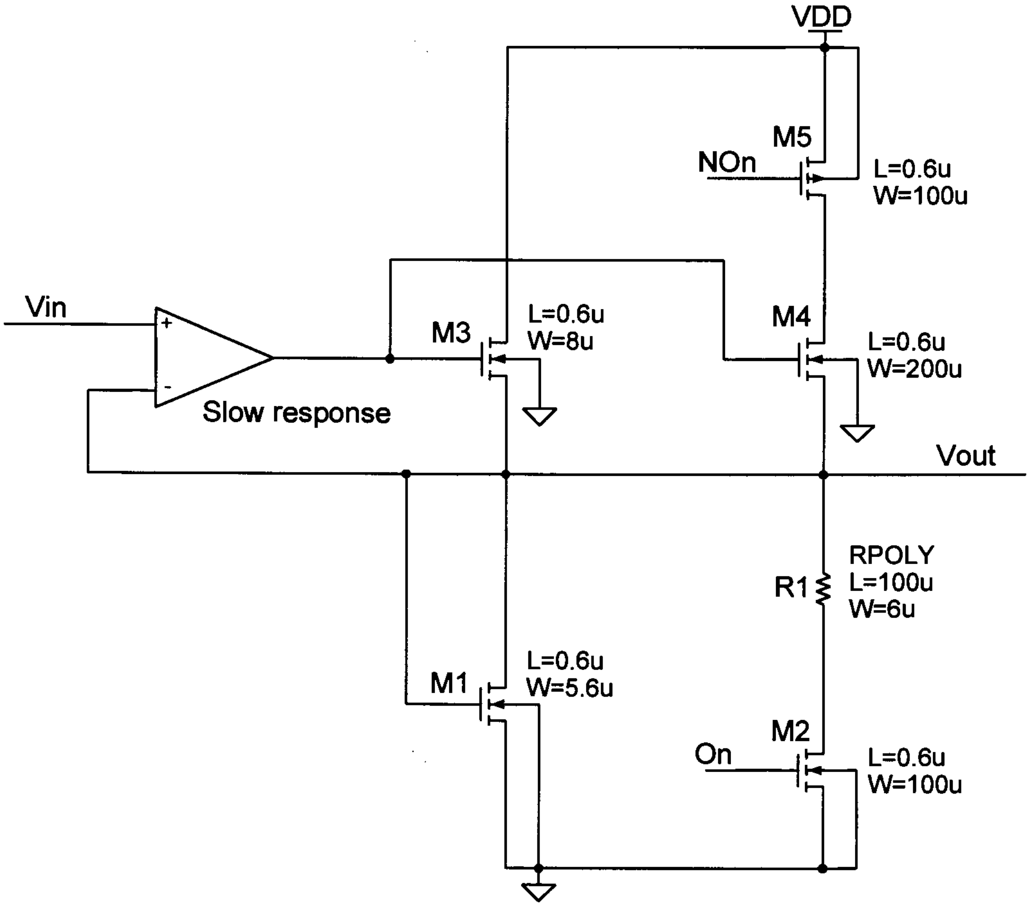


Figure 3.11 Low power, low-impedance driver circuit.

The basic principle behind the circuit is to have two negative feedback loops for the operational amplifier, one of which can be turned off to save power. The other, using only a small bias current, is always on. Both configurations are designed to present themselves as being identical to the op-amp. The main feedback loop may be turned off via *M2* and *M5*. Device *M2* stops current draining out of the output node through the resistor and *M5* stops any current flowing into the node through *M4*. These two devices are controlled by the *On* signal and its inverse *NOn*. When *On* is "high" the main feedback loop is turned on. The feedback network made up of devices *M1* and *M3* is always on. Its function is to keep the gate voltage of *M4* at the level that is required to deliver the correct current through the resistor to maintain the output voltage of the circuit equal to the input voltage.

To design the circuit, the size of the resistor is considered first. It has to be able to sink all transient and static currents associated with the eight *D/O*-lines, without the voltage drop over it exceeding the minimum reference voltage of 0.8V. The peak transient current when the cell changes state is 200μA, so the peak current that has to be sunk by the resistor at 0.8V is therefore 1.6mA. This translates to a maximum resistance of 667Ω. The maximum allowable poly current density is 0.45mA/μm and the sheet resistance is 33Ω/square. The resistor needs to be 100μm long and 6μm wide.

The dimensions of the driver transistor *M4* are chosen large enough so that the peak charging current of 32.4mA can be supplied at moderate gate-source voltages. The device has to be quite powerful because the speed of the feedback system depends on it. A small deviation in the gate-source voltage has to cause a large change in the current. The switch devices *M2* and *M5* operate in the linear region. Their dimensions are chosen in such a way that the resistance at peak currents is insignificant as far as operation of the circuit is concerned.

The alternate feedback network is based on a small current permanently flowing in device *M1*. This transistor is diode-connected and will always be on, because its gate-source voltage is equal to the reference voltage which is 0.15V larger than the threshold voltage. The transistor *M3* is designed to supply the current to *M1* at an identical gate-source voltage to that required by *M4* to supply enough current to the resistor to create the correct voltage drop. This principle is based on ratios between devices, rather than on absolute device strength, so it is independent of process conditions, as long as all devices in question are well matched. The op-amp also aids in this, by allowing the circuit to adapt to the process and temperature conditions.

3.3.7 Operational Amplifier [22]

The specifications for the op-amp are not very stringent. The speed does not have to be high, nor does the slew rate. More importantly, the input offset voltage should be low so that the output voltage does not differ too much from the input voltage. This also means the gain should not be too low, typically 60dB. To ensure stability the op-amp has to be adequately compensated, so the phase margin has to be high (greater than 60°). The response times of the circuit must be slow, so a maximum unity-gain bandwidth of 10MHz seems reasonable, because this is about 10 times slower than the typical cycle frequency. Normal circuit operation takes place without affecting the op-amp. The output load is a pure capacitance in the order of 600fF, so an output driver stage is not required. Figure 3.12 shows the basic circuit diagram of the two-stage compensated op-amp.

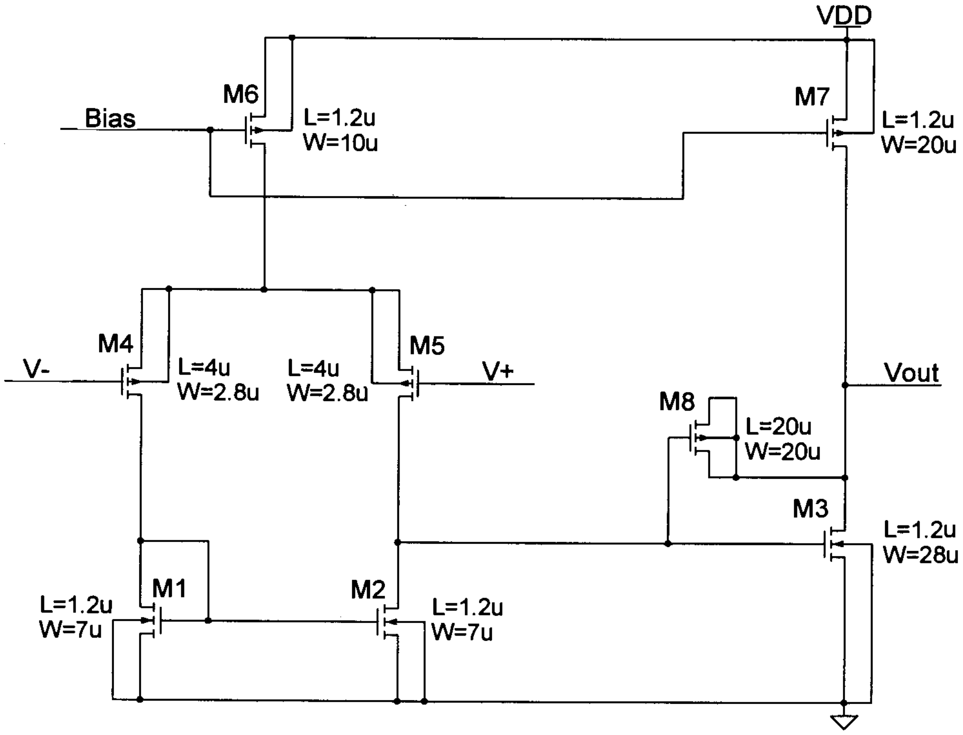


Figure 3.12 Two stage compensated op-amp.

Once again, for the purpose of good matching, the minimum transistor length is chosen as 1.2μm. All matched transistors must also have the same length, because this aids in reducing the input offset voltage [23]. The layout has to be possible in a standard CMOS process, so the compensation capacitor has to be implemented using a transistor gate capacitance. To aid in stabilising the circuit the value of this capacitance is taken double the usual estimate of C_C=C_{Load}. To achieve the low gain-bandwidth product the input stage is biased at very low currents. The bias voltage is generated as part of the reference voltage and represents a constant transconductance bias. The 25μA are scaled down to 12.5μA for the differential input pair, to lower the transconductance. The unity-gain frequency is

$$\omega_0 = \frac{g_{mi}}{C_c}, \tag{3.8}$$

where g_{mi} is the transconductance of the input devices. Using this equation the *W/L* ratio of the input devices can be found to be 0.6. The sizes are calculated based on the fact that the input transistors need to be made up of two parallel devices. This allows common-centroid layout to be used to reduce the offset voltage [13]. The width is therefore chosen as double the minimum width and the length correspondingly calculated. In order to satisfy the phase margin constraint the second pole has to lie at a frequency of least

$$\omega_{p2} = \frac{g_{m3}}{C_{Load}} = 3\omega_0, \tag{3.9}$$

The constraint ensures a phase margin of 60°, but it should be higher to avoid all overshoot in the circuit. The small load capacitance compared to the higher compensation capacitor and the low gain of the input stage make this a simple constraint to achieve. Any width above 3μm for the device *M3* in Figure 3.12 ensures it is satisfied.

Due to the very low transconductance of the first stage, that of the second stage has to be adequately high to achieve the set gain specification. The overall gain of the amplifier is

$$A_v = \frac{g_{m5}}{g_{o5}+g_{o2}} \frac{g_{m3}}{g_{o3}+g_{o7}}. \qquad (3.10)$$

To aid in achieving the higher transconductance the bias current of the second stage is doubled. No data is presented by the manufacturer on the output transconductance of the devices. A bias point simulation of a device in saturation revealed it to be in the order of 10μS. Using the specification of 60dB the minimum width of *M3* can be calculated to be 28μm.

To avoid a systematic offset voltage the scaling of the current mirror devices of the input stage has to be

$$\frac{W/L_1}{W/L_3} = \frac{W/L_2}{W/L_3} = \frac{1}{2}\frac{W/L_6}{W/L_7}. \qquad (3.12)$$

This fixes the width of *M1* and *M2* to 7μm.

The characteristics of this op-amp can be seen in Figure 3.13. It depicts the simulation results of a frequency response simulation that has been repeated across all fifteen combinations of simulation models. The characteristics are well matched and within specifications for all possible process conditions.

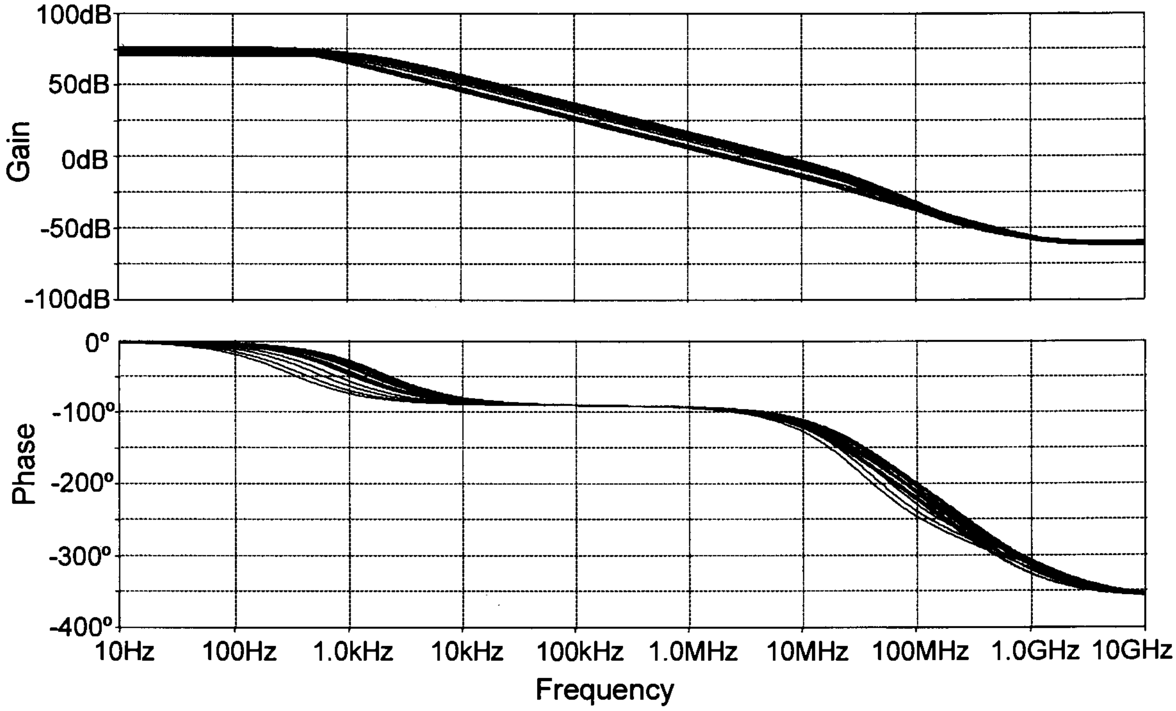


Figure 3.13 Op-amp gain and phase response for different process conditions.

The typical mean performance figures are:

• unity-gain frequency	6.2MHz
• open-loop gain	72dB
• phase margin	75°.

3.3.8 *D*/*O*-Line Driver Circuit Simulation

So that the complete driver circuit may be simulated, the sub-circuits discussed up to now are placed together as shown in Figure 3.1. Eight switch circuits, each with their own control circuit, are connected to one bias network, operational amplifier and driver circuit. Each switch circuit is loaded with the calculated capacitance of 4.05pF. The drain-bulk and source-bulk capacitances of all devices are included in the simulation, by ensuring that the area and perimeter values for the drain and source of each transistor are included in the netlist.

The responses of the circuit to the following situations has to be simulated:

- the maximum capacitance is switched, that is all *D*/*O*-lines are switched and their capacitance is maximum,
- the minimum capacitance is switched, that is a single *D*/*O*-line is activated with minimum capacitance (only a single source-bulk capacitance per cell),
- the circuit is switched on with no capacitance connected, that is when a write occurs, but no bits in the group of eight need to be set.

In the last scenario it is important to test the effect on the operational amplifier. The average output voltage of the operational amplifier should remain constant. Any adverse deviations in the output voltage are not desired, because these take a long time to disappear, given the intentional slow speed of the op-amp.

The response of the circuits to these various situations has to be tested for short pulses (10ns), as well as long pulses (50ns). The time between the pulses also has to be varied (10ns and 50ns). A simulation has been set up that tests the three situations using the sequence of 10ns on, 10ns off, 10ns on, 50ns off, 50ns on and 10ns off. The results are shown in Figures 3.14, 315 and 3.16.

Figure 3.14 shows the voltage of a single *D*/*O*-line when the driver circuit is loaded with a maximum capacitance. The simulation was performed for all different model combinations. It can clearly be seen that the rise time is virtually independent of the process conditions. There is some delay present which is mostly due to the delay in the peripherals circuits. The range of the pulse amplitudes is within limits to ensure correct operation of the SRAM cells.

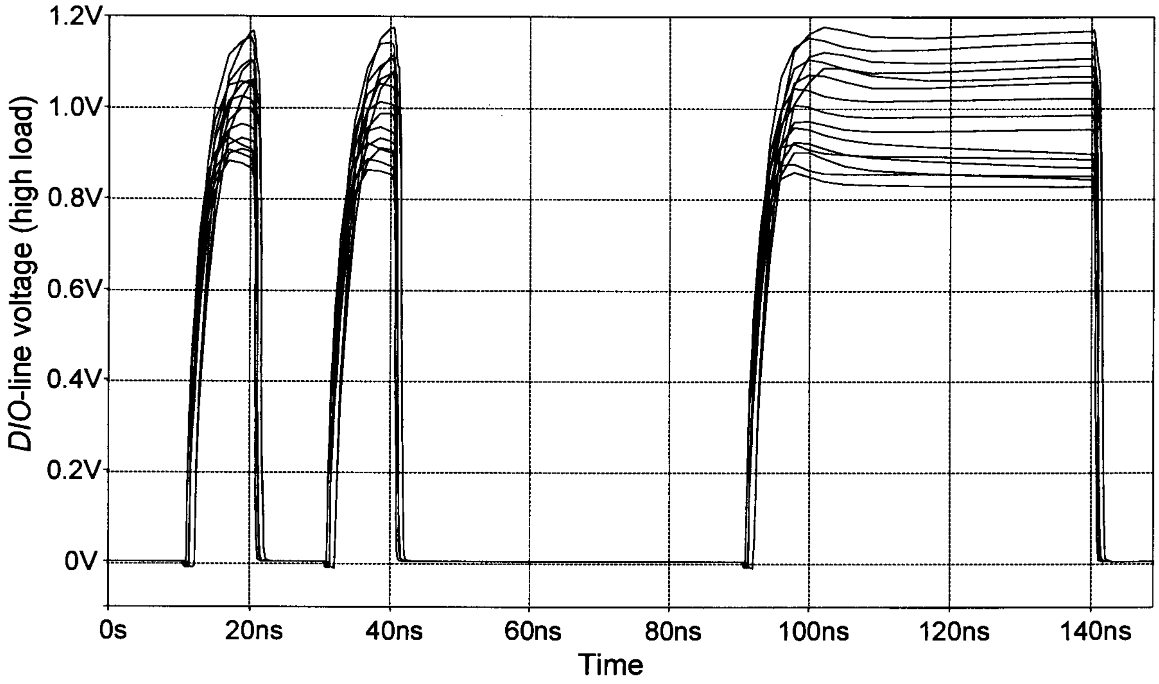


Figure 3.14 *D*/*O*-line voltage for the maximum load condition across all process models.

In the case of minimum load conditions (Figure 3.15) some overshoot is present in the characteristic. In those cases where the quality of the NMOS is high (lower deviation voltage), this is not serious because the SRAM cells were shown to be able to operate at higher deviations without error. As the quality of the NMOS devices decreases, the deviation increases, but it can be seen that the overshoot that occurs is no longer higher than the value of the reference voltage, because the average amplitude of the pulse is actually lower than the reference voltage. This once again is not critical, because the cells do operate correctly at a deviation of 1V, irrespective of process conditions. The rise times are decreased for the low load condition but still seem to be quite independent of process conditions. The range of the deviations is also slightly increased, but still within acceptable limits when verified with the results of the noise margin analysis of the SRAM cell.

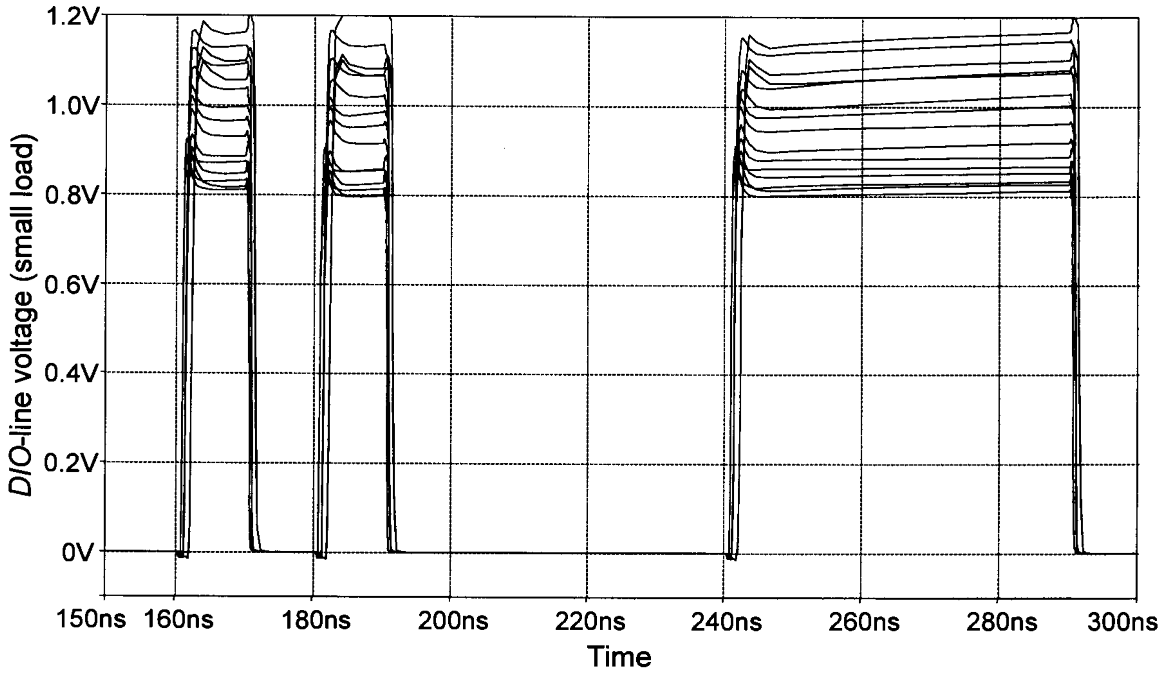


Figure 3.15 *D/O*-line voltage for the minimum load condition across all process models.

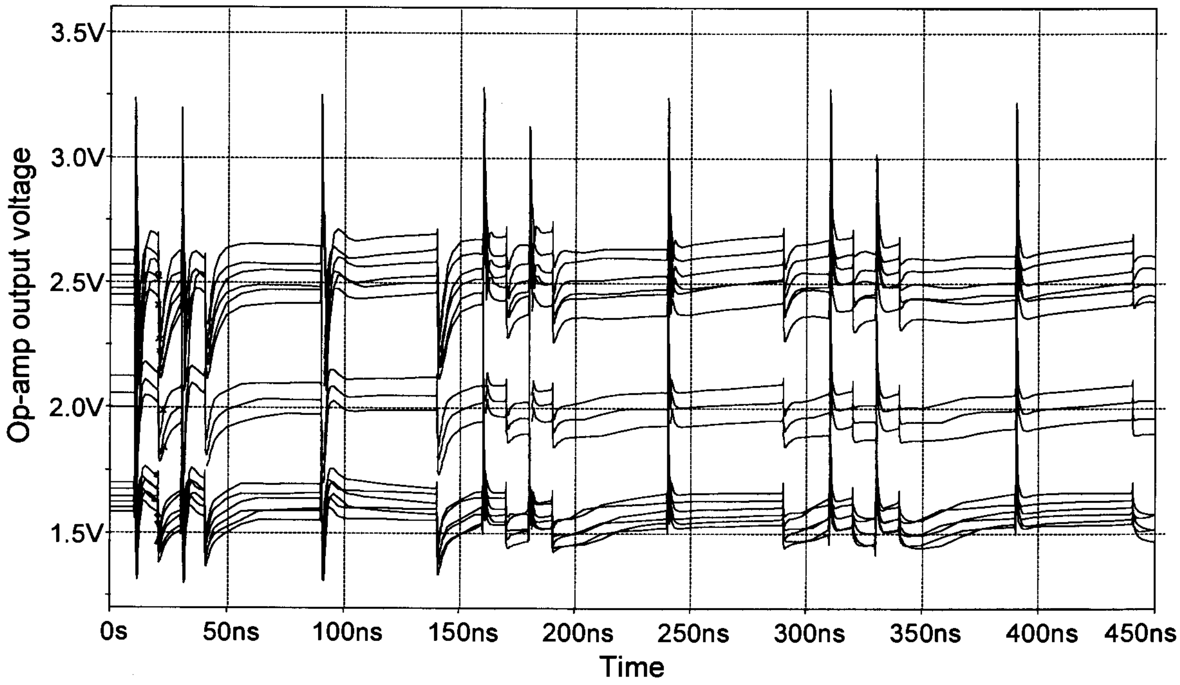


Figure 3.16 Simulation of the operational amplifier output voltage for different processes and load conditions.

The final test is whether different conditions can affect the output voltage of the op-amp in such a way that incorrect voltage levels will occur. This is shown in Figure 3.16, where the output voltage of the operational amplifier is shown for the full length of the simulation (all three load situations). It can clearly be seen that turning the circuit on and off does cause certain responses to occur, but none of these affect the average voltage. This simulation proves that the scheme of using the operational amplifier to adjust only for varying conditions can produce the required results. The load can be adequately charged using the inherent negative feedback loop of the driver transistor.

3.4 *RW*-LINE DRIVER

3.4.1 Overview

This driver circuit is similar to the *D/O*-line driver, as can be seen by comparing its functional block diagram (Figure 3.17) to that of the *D/O*-line driver (Figure 3.1).

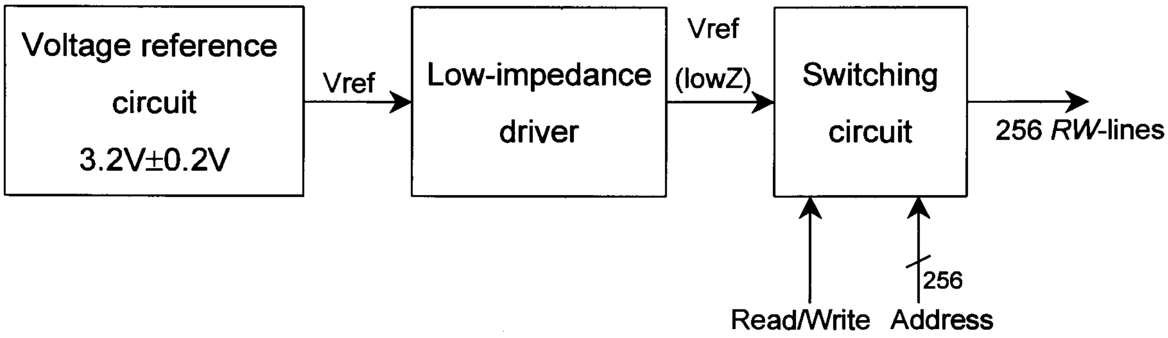


Figure 3.17 Functional block diagram of the *RW*-line driver circuit.

Basically the circuit has to perform an identical function to the *D/O*-line driver, but some small differences are present. The basic analog circuits are identical except that they are mirrored, that is the operation is with respect to 5V rather than ground. The design procedures and equations derived for the previously discussed circuit are valid here without change. Therefore only the differences will be discussed. The final circuit simulations are given to show that the circuit operates correctly.

3.4.2 Line Capacitance

Different to the *DIO*-lines, the capacitance associated with a single *RW*-line is small. The capacitance contributed by one cell is the drain-bulk and source-bulk capacitance of one PMOS device, the drain-bulk capacitance of an NMOS device and the gate capacitance of one NMOS and one PMOS transistor. If the cell is in the opposite state the capacitance is reduced to the source-bulk diffusion capacitance of one PMOS device. These two values are 14.27fF and 3.17fF per cell respectively, and add up to an *RW*-line capacitance of between 456fF and 101fF. Added to this is the capacitance associated with the metal routing, 52fF in this case. The routing capacitance is low because the lines are short, and the total switched capacitance is also smaller in comparison to the previous circuit, especially given the fact that only one *RW*-line is activated at a time.

3.4.3 Currents

The currents that have to be sourced by the driver circuit are slightly different. The wasted write currents flow in the inverter opposite to the one where the *DIO*-voltage deviation is applied, and therefore need to be supplied by the *RW*-line driver in its off-state. Because there are 32 cells connected to one switching circuit, each potentially requiring 20μA of wasted write current, the total current that needs to be supplied at a low voltage drop across the driver, is 640μA. This situation is present when another word in the array is being written. Here it is very important that the voltage drop over the internal resistance of the pull-up device is small, because one NMOS source node (*DIO*) is deviated. If the voltage of the *RW*-node drops too much, static write conditions could occur and the cell could unintentionally be written. The voltage drop should be strictly limited to below 0.05V. In the off-state, the transient switching currents when cells are being cleared, also need to be supplied. The peak current is 32 x 200μA, a 6.4mA peak current. For this the voltage drop over the pull-up may be quite large because the cells are in the process of being cleared. Static write conditions are present and the voltage drop in the *RW*-line will not affect this.

The capacitance associated with a single *RW*-line is small, so the charge and discharge currents are also small. When the driver is in the on-state, all 32 cells connected to it can potentially be written. This once again causes transient currents. These currents should not modify the *RW*-line deviation too much, because this voltage may be connected to some cells that must not be written. If the voltage deviation increases too much, their noise margin will degrade to the point where writing might accidentally happen.

3.4.4 Switching Circuit

The circuit diagram of the *RW*-line driver switching circuit is shown in Figure 3.18. Topology wise the circuit is identical to the *D/O*-line driver, so further explanation of the operation is not required. A certain *RW*-line switch is turned on when the corresponding *Select* and the *ReadWrite* signal are "high". The *RW*-driver is required for reading and writing, so it is controlled by a signal that is active for either a read or a write cycle. The *Select*-lines are the output of an address decoder that selects the word to be accessed.

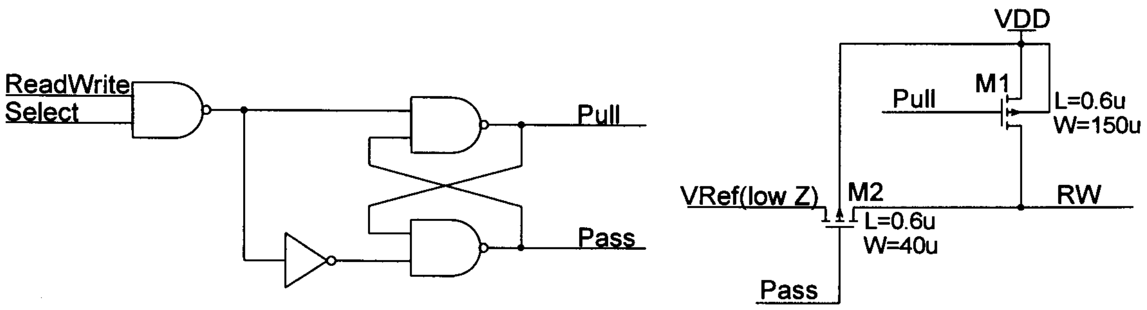


Figure 3.18 Switching circuit for the *RW*-line driver including the control latch.

Once again a latch is used to prevent both *Pass* and *Pull* from being "low" at the same time. This allows the short-circuit current between the power supply and the low-impedance reference voltage to be significantly reduced. The NAND-gate driving the *Pull* node is designed to have more driving strength, to compensate for the larger load capacitance.

The sizing of the pass and pull-up devices is derived using an identical procedure to that used for the *D/O*-line driver. Three factors need to be considered when determining the specification for the resistance of the pull-up device. Firstly, there is the pull-up time constant. Because the capacitance is low the resistance may be quite high. To achieve a time constant of 0.2ns the resistance has to be smaller than 400Ω. Secondly, the wasted write current of 640μA per row of cells may not cause a voltage drop of more than 0.05V, which leads to a resistance specification of no more than 78Ω. This is valid for the typical mean process. In the worst case one situation the resistance of the PMOS devices is high, but the wasted write current is large, because it is determined from the NMOS devices. The total current per row is 800μA. The maximum resistance is therefore 62.5Ω. Figure 3.19 is a plot of the resistance of a PMOS device at a drain-source voltage of 0.05V as a function of the device width. It can clearly be seen that the required width is 150μm.

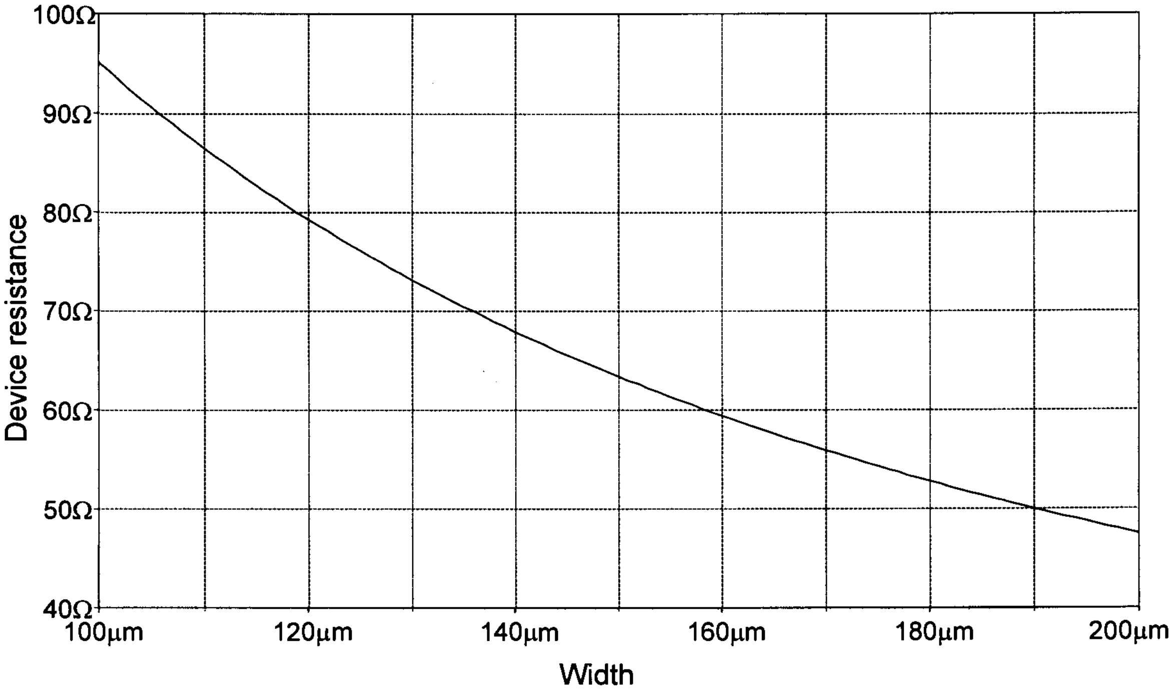


Figure 3.19 PMOS device resistance at V_{DS}=0.05V as a function of the gate width. The gate length is 0.6μm.

For the pass transistor the maximum resistance is determined by the charging time. The drain-bulk diffusion of the wide pull-up device does add significant loading, so that the total node capacitance is increased to 800fF, which means the charging resistance has to decrease to 250Ω. The device resistance at a drain-source voltage of 1.8V needs to be considered. This is shown in Figure 3.20, and it seems the 40μm device width is a choice that should guarantee satisfactory performance.

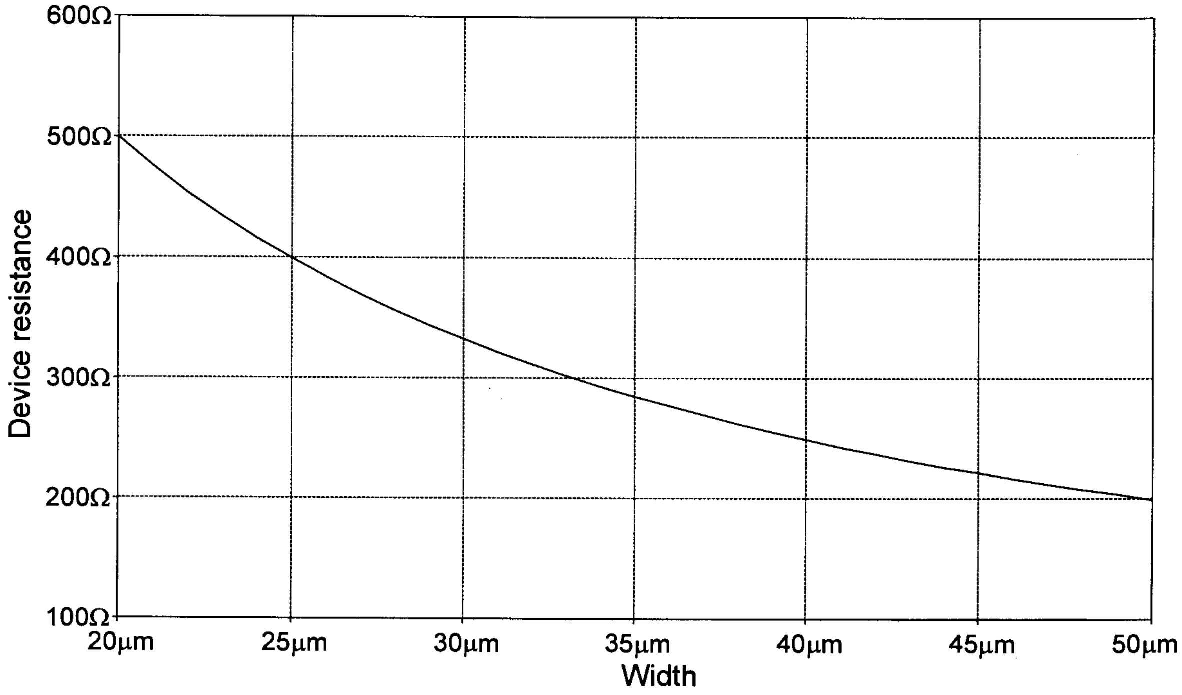


Figure 3.20 PMOS device resistance at V_{DS}=1.8V as a function of the gate width. The gate length is 0.6μm.

3.4.5 Voltage Reference Circuit

The deviation scheme devised via the noise margin analysis of the four-transistor SRAM cell requires the deviation to be 1.8V±0.2V. The deviation has to increase as the quality of the PMOS devices drops and decrease as the quality increases. The minimum, typical and maximum threshold voltage of the PMOS devices is specified to be 0.68V, 0.8V and 0.9V respectively [19]. In this case the over-voltage required for a single device to transform the reference current to the required deviation is too large. Comparing the required deviation to the threshold voltage shows that two threshold voltages can fit into the deviation and in that case the over-voltage is very small. The PMOS devices can be placed in unique wells which can be connected to the source, so the bulk effect can be overcome. The devices are designed to have a large *W/L* ratio, so that the over-voltage is small and the final deviation is very close to two threshold voltages. The bias current of the branch with the reference devices is also reduced. The simulated reference voltages for different process conditions are given in Table 3.2. The reference voltage is equal to the deviation subtracted from 5V. A weaker PMOS device therefore creates a lower reference voltage.

The rest of the circuit is very similar to the *D/O*-line driver reference circuit. The current per branch was also designed to be 25μA, and cascodes help increase the output impedance of the current mirror devices. This improves the matching and therefore the accuracy of the constant transconductance bias circuit. An identical start-up circuit is used to prevent the zero current state. This circuit is designed to consume minimal static power, without the load device becoming too long, requiring large chip area. The current mirror bias voltage, *Bias*, is an output that is used in biasing the operational amplifier.

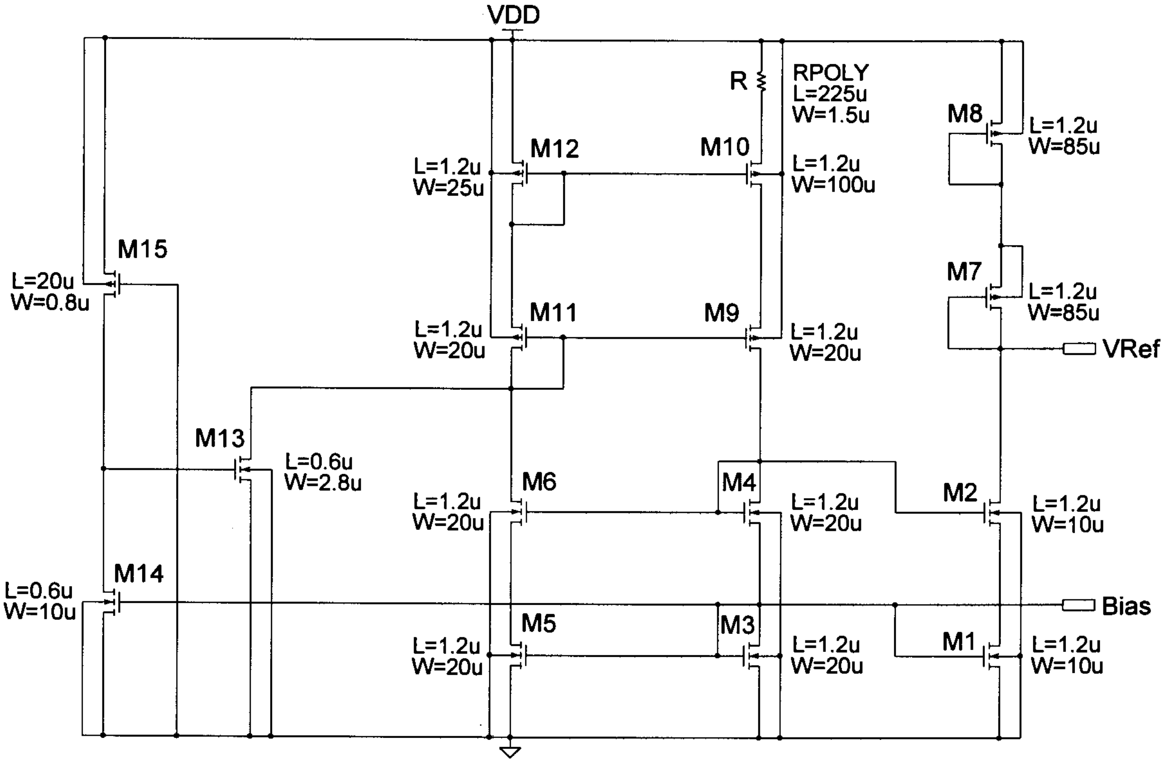


Figure 3.21 Reference voltage generator for the *RW*-line driver.

Table 3.2 Simulated reference voltage for the *RW*-line driver reference circuit across the different process corners.

Transistor model Resistor model	TM	WP	WS	WO	WZ
TM	3.20V	3.44V	2.90V	2.93V	3.42V
WP	3.14V	3.40V	2.84V	2.87V	3.38V
WS	3.24V	3.47V	2.96V	2.98V	3.46V

3.4.6 Low-Impedance Driver Circuit

A choice that needs to be made is the number of driver circuits to use. The capacitance that needs to be switched at any time is never more than one *RW*-line, equalling 800fF. The capacitance associated with the output node of the low-impedance driver circuit is high. Every switch circuit connected to this line adds 63.4fF capacitance by means of the 40μm wide device *M2* in Figure 3.18. A maximum of 256 switch circuits can be connected to this line. This makes the capacitance very large (16.2pF), but this is advantageous to the operation of the circuit, because it creates a situation where the switched capacitance is more than an order of magnitude smaller than the precharged capacitance. The capacitance of the driver output is permanently kept at the correct voltage by the op-amp feedback network. Activating an *RW*-line will only cause a small change in output voltage due to the charge being shared among the large and correctly charged output capacitance and the small capacitance of the *RW*-line. The only observation required here is that a smaller voltage change will occur on the output node of the driver when an *RW*-line is connected. The main driver transistor *M3* in Figure 3.22 therefore has to be sufficiently strong to quickly recharge the node to the correct potential.

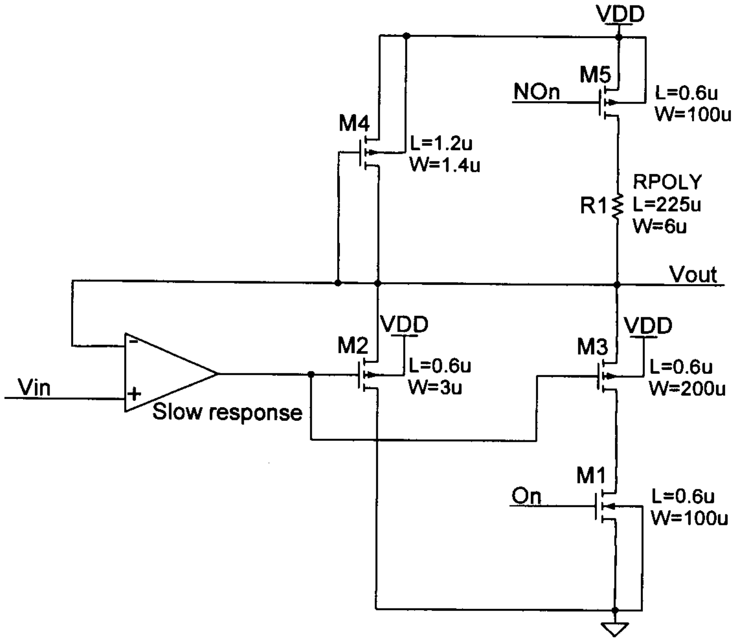


Figure 3.22 Low-impedance driver for the *RW*-line driver circuit.

3.4.7 Operational Amplifier

As far as the operational amplifier is concerned, the set specifications, as well as the circuit configuration and design procedure, are identical. The circuit is however mirrored with respect to the power supplies for two reasons. The NMOS input devices help to accommodate input voltages that are closer to the power supply than they are to ground, and the NMOS current sources make it possible to use the constant PMOS-transconductance bias network. The circuit diagram is given in Figure 3.23.

The specifications for the operational amplifier when simulated using a typical mean process model are:

- unity-gain frequency	9.3MHz
- open-loop gain	75dB
- phase margin	70°.

The frequency response simulations of Figure 3.24 show that the amplifier characteristics do not vary significantly across the full range of process variations.

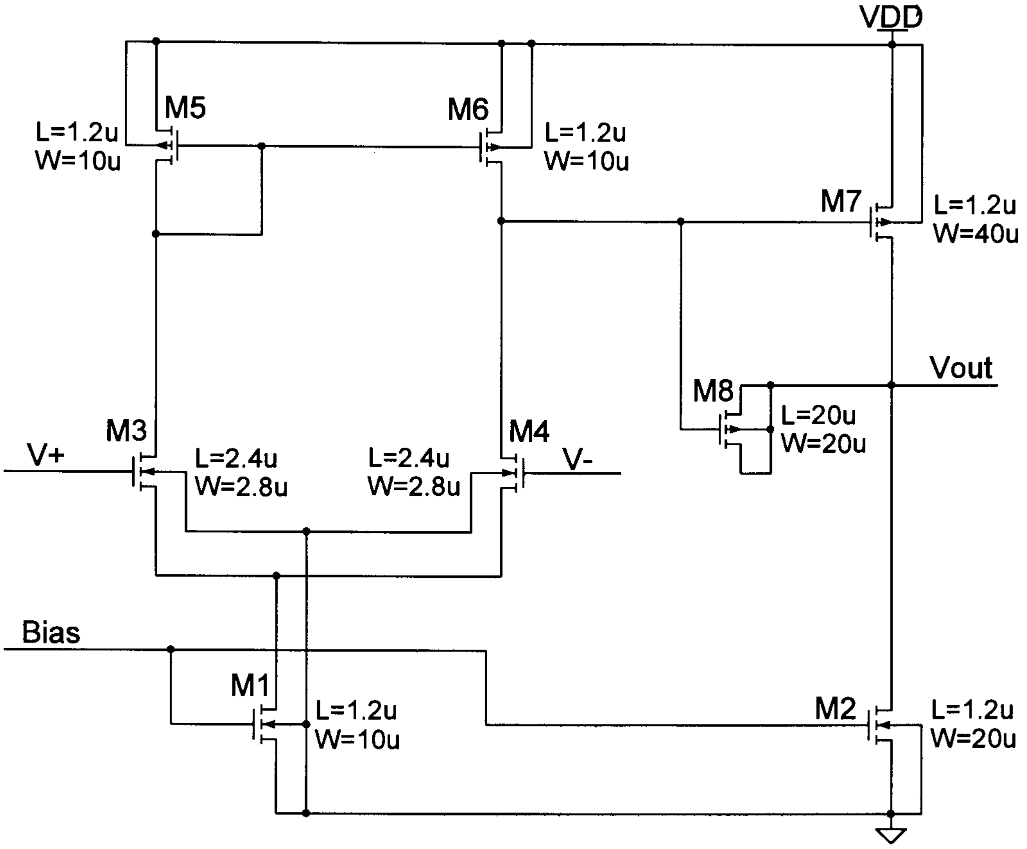


Figure 3.23 Operational amplifier circuit diagram for the *RW*-line driver circuit.

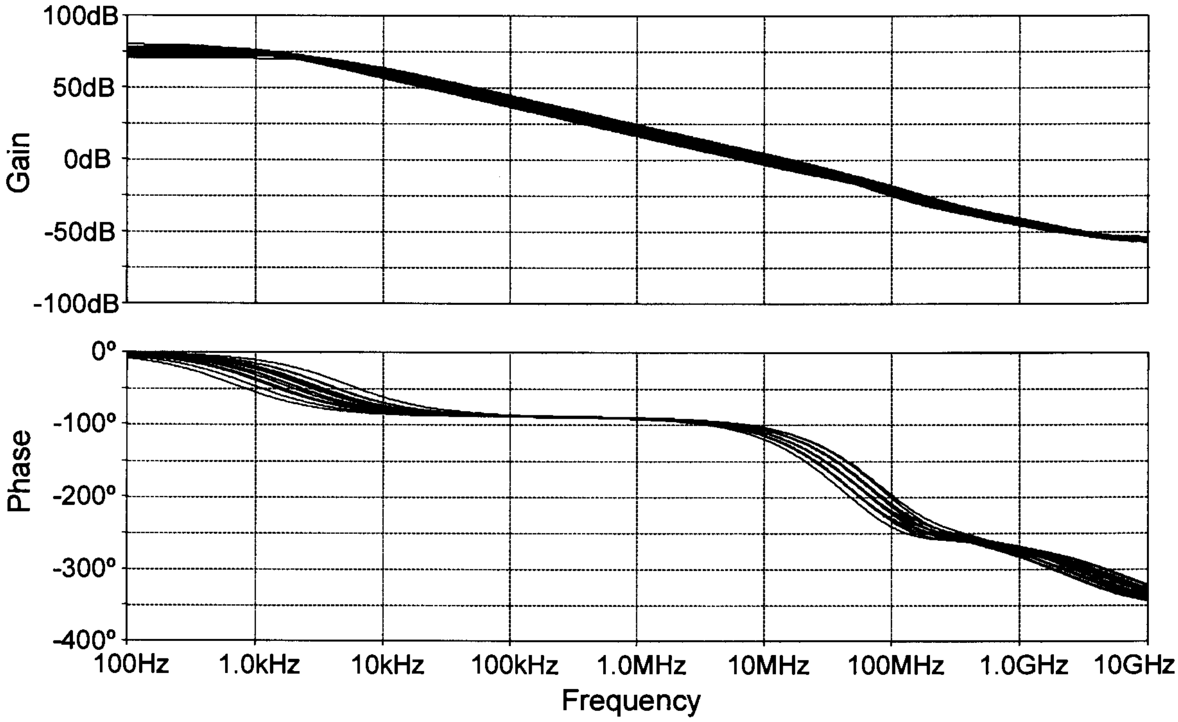


Figure 3.24 Frequency response of the *RW*-line driver operational amplifier.

3.4.8 *RW*-Line Driver Simulation

To test the correct operation of the driver circuit all components are placed together. The 16.2pF capacitance on the output node is added together with a switch circuit, including its control circuit. The load consists of 32 cells, and because this is relatively small, it was decided to use this as the load instead of the capacitance used during the design. This improves the accuracy of the simulation. The cells are all initialised in the "zero" state. This implies that the devices connected to the *RW*-line are on and capacitance conditions are maximum.

The conditions that need to be tested are the maximum and minimum load capacitance, for short and long pulses, and the spacing between the pulses also has to be varied. The times given in the following simulation description can be referred to Figures 3.25 and 3.26 which show the first and the second half of the simulation respectively.

Two short pulses (10ns), with a short delay between them (10ns) are initially applied. A pause of 50ns is added between 40ns and 90ns, and a 50ns long pulse is tested thereafter. In the pause time the *D/O*-lines of all cells are activated, to allow the effect of the wasted write current to be analysed. This happens in the time range 50ns to 60ns. The wasted write current is allowed to flow by activating the *D/O*-line. This causes a voltage drop that should be less than 0.05V to occur in the *RW*-line voltage.

After the long pulse from 90ns to 140ns all cells are written (150ns to 160ns), with the aim of testing the effect of this on the circuit. Because all cells are now cleared, the load conditions are minimum, and the initial sequence of two short pulses and one long pulse is repeated. This time it makes no sense to activate the *D/O*-lines in the 50ns pause time from 200ns to 250ns because the devices connected to this line are all turned off. Finally the cells are all cleared by activating the *CL*-line at 310ns, to verify the effect of the transient current peak when the cells are cleared.

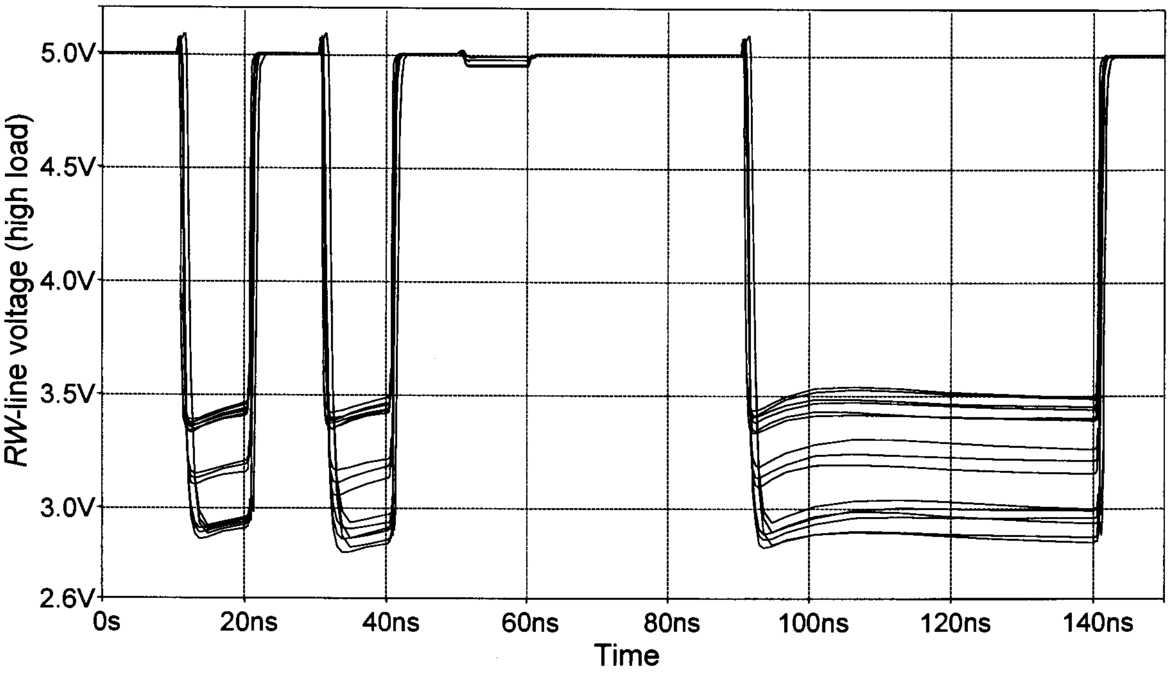


Figure 3.25 *RW*-line voltage for the maximum load condition across all process models.

Some overshoot is evident in the circuit response. This is even more so when the minimum load condition is present (Figure 3.26), although there is not much of a difference between the two cases, given the small difference in capacitance. When considering the operation, this overshoot is most problematic where the deviation is inherently high due to the poor PMOS quality. In these cases the noise margin does however remain above 0.6V (see Figure 2.17). The three groups of deviations can clearly be seen. A simulation model either has a high, typical or low quality PMOS device. The groups were less evident for the *D/O*-line driver. Due to the higher voltage deviation present in the *RW*-line driver, the spread is larger and the groups more clearly defined. Figure 3.25 also shows that the wasted write current does cause a small voltage drop from 50ns to 60ns, but this is less than the specified 0.05V.

When all cells are being written (time range 150ns to 160ns), the current spikes flowing do not seriously affect the *RW*-line voltage, as can be seen by comparing the three short pulses in Figure 3.26. The final glitch in the voltage (at 310ns) is caused by the transient peak currents that flow when all cells are being cleared by raising the *CL*-line voltage. As already mentioned the amplitude of this glitch can be quite large, as long as it is below the threshold voltage of the PMOS devices.

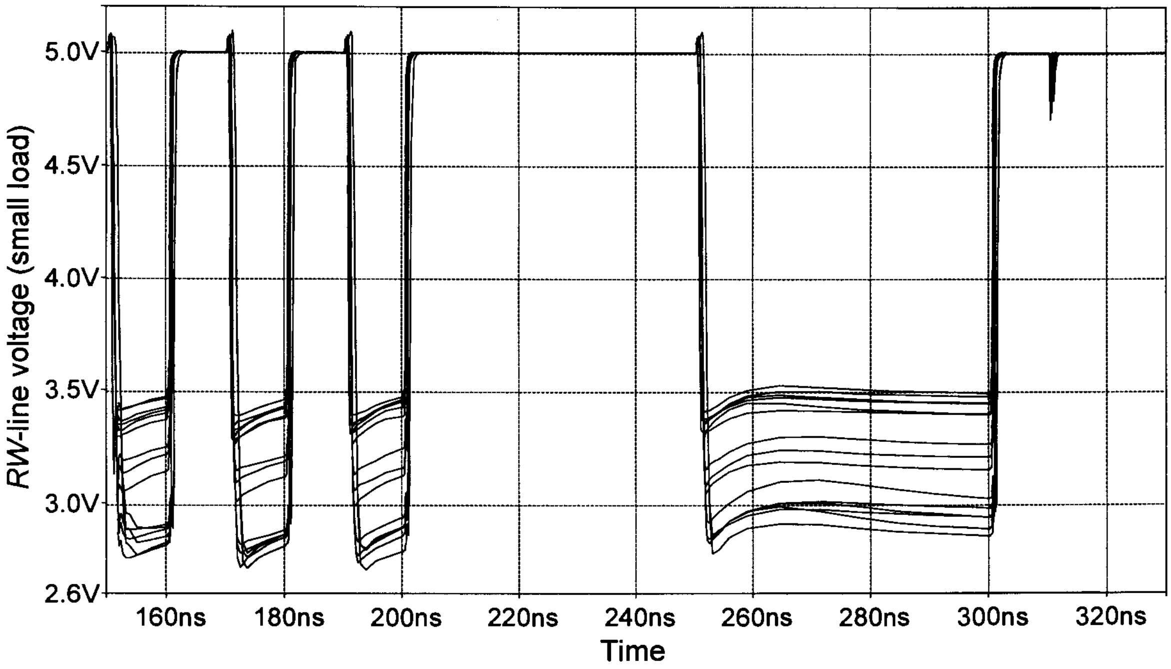


Figure 3.26 *RW*-line voltage for the minimum load condition across all process models.

3.5 *CL*-LINE DRIVER

3.5.1 Overview

This driver circuit, when activated, has to pull the *CL*-line up to *VDD*. This causes the cells to be forced into a certain state. The 5V deviation was chosen for simplicity, because noise margins and static write conditions are not an issue. Any deviation that creates static and dynamic write conditions is sufficient. The simplest to implement is to pull the node up to *VDD* and otherwise connect it to ground via a low impedance. Essentially the *CL*-driver is therefore an inverter.

3.5.2 Line Capacitance

To determine what transistor sizing is required, the worst case capacitance associated with the *CL*-line has to be calculated. The worst case capacitance contributed per cell is the same as for the *DIO*-line (14.4fF). The *CL*-line thus has an associated capacitance of 500fF, if the capacitance of the metal routing is included.

3.5.3 Currents

The currents which flow in the *CL*-node of the four transistor SRAM cell are the transient currents (200µA) while the state switches, and the wasted write current of 20µA each. The wasted write current may once again not create a significant voltage drop across the pull-down device, so that the noise margin of the cell is not degraded. The 0.05V specification used for the *RW*-line driver is set here as well. The transient current flowing when cells are written falls under the same constraints as the transient clear currents that affect the *RW*-driver circuit. As long as the voltage drop caused by them is below the threshold voltage, they hardly affect the cells.

3.5.4 Circuit Design

The parameters that need to be designed are the widths of the PMOS pull-up and the NMOS pull-down. The pull-down has to allow 640µA at 0.05V voltage drop and 6.4mA at no more than 0.5V voltage drop. The resistances required for both cases are 78Ω. Similar simulations as those of Figure 3.19 and 3.20 were repeated for the NMOS and the required width was found to be 40µm. The pull-up device has to charge the 500fF capacitance with a time constant of 0.2ns, as is used throughout the design. This implies a resistance of 400Ω. This resistance has to be applicable over the complete voltage range, but this is not possible for a MOS device. Initially the resistance is high, and drops as the drain-source voltage decreases. The appropriate width is chosen based on the simulated rise time and found to be equal to 40µm. Once again it is desired not to turn on both devices simultaneously in an effort to save power. An identical scheme to the previous two driver circuits is used to implement this. The circuit diagram is shown in Figure 3.27.

The driver circuit of a certain row of cells is activated if the *Clear* signal and the *Select* line of that row are "high". The *Select* line is the identical signal that is also part of the activation scheme of the *RW*-line driver switching circuits. Inverters A and B together have an identical delay to inverter C, so that the delays between the latch circuit and the pull-up and pull-down devices are identical. This prevents distortion of the correct shift between the activation signals of *M1* and *M2* created by the latch circuit.

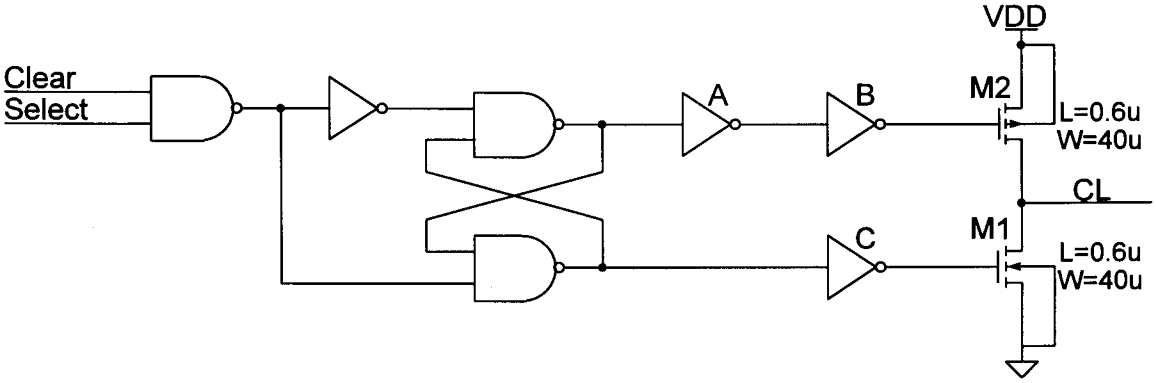


Figure 3.27. Circuit diagram of the *CL*-line driver.

3.5.5 *CL*-Line Driver Simulation

To simulate the operation of the driver circuit, a row of 32 cells is attached to it. The aspects that need to be simulated are that the cells can be cleared effectively, that the wasted write currents do not cause a significant voltage drop over the pull-down device *M1*, and that writing the cell does not cause a voltage drop higher than the threshold voltage. This needs to be done for the five process corners. The cells are initialised in the "clear" state. The *DIO*-lines are then activated (5ns to 15ns) and after this the state is changed to "set" at 22ns. This simulates the last two aspects respectively. Finally the cells are cleared (45ns to 55ns).

The simulation results of Figure 3.28 show the voltage of the *CL*-line for the described simulation run. It can clearly be seen that the wasted write currents only cause a very small voltage drop. The highest drop is for the worst case power and was measured as 46mV. This is within specification. The maximum voltage drop during the write cycle is 0.5V which is lower than the threshold voltage. The clear signal activation and levels are satisfactory, although the delay for the worst case speed is substantially more than for all other process models. The rate of change of the signal is however almost identical to all other simulation runs, so it can be concluded that the longer delay is caused in the control circuit, rather than by an insufficient pull-up or pull-down device strength. Therefore the delay is difficult to overcome, unless some other circuit parameter, typically power dissipation, is compromised.

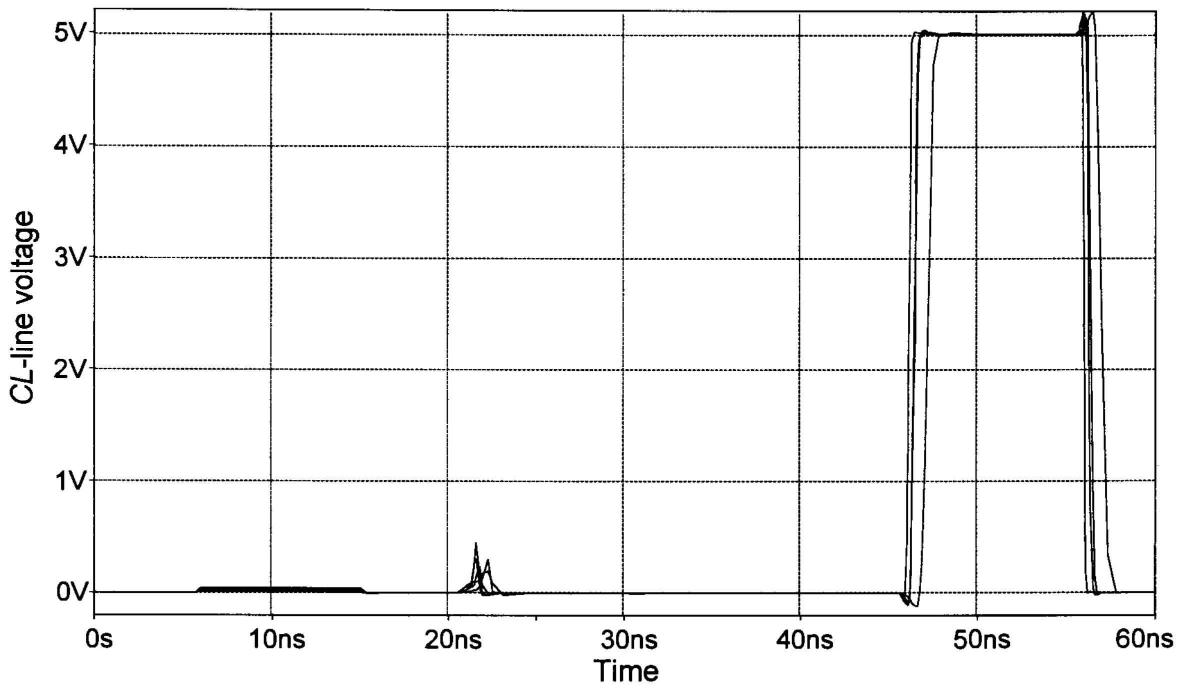


Figure 3.28 Simulation of the *CL*-line voltage for the different process models.

3.6 CELL CONTROL SIMULATION

At this point it is required to test all line drivers together. Correct methods require that a 256x32-bit array be simulated. A netlist of this circuit contains more than 32k devices in the cell array. This requires extensive simulation time and yields only slightly more information than simulating a single cell does. Therefore the identical simulation to that of Section 2.8 is performed using the designed driver circuits to drive the SRAM-nodes. The rest of the array is added as capacitance to the simulation. To verify that errors do not occur, independent of the load conditions, the simulation is repeated for minimum and maximum load conditions.

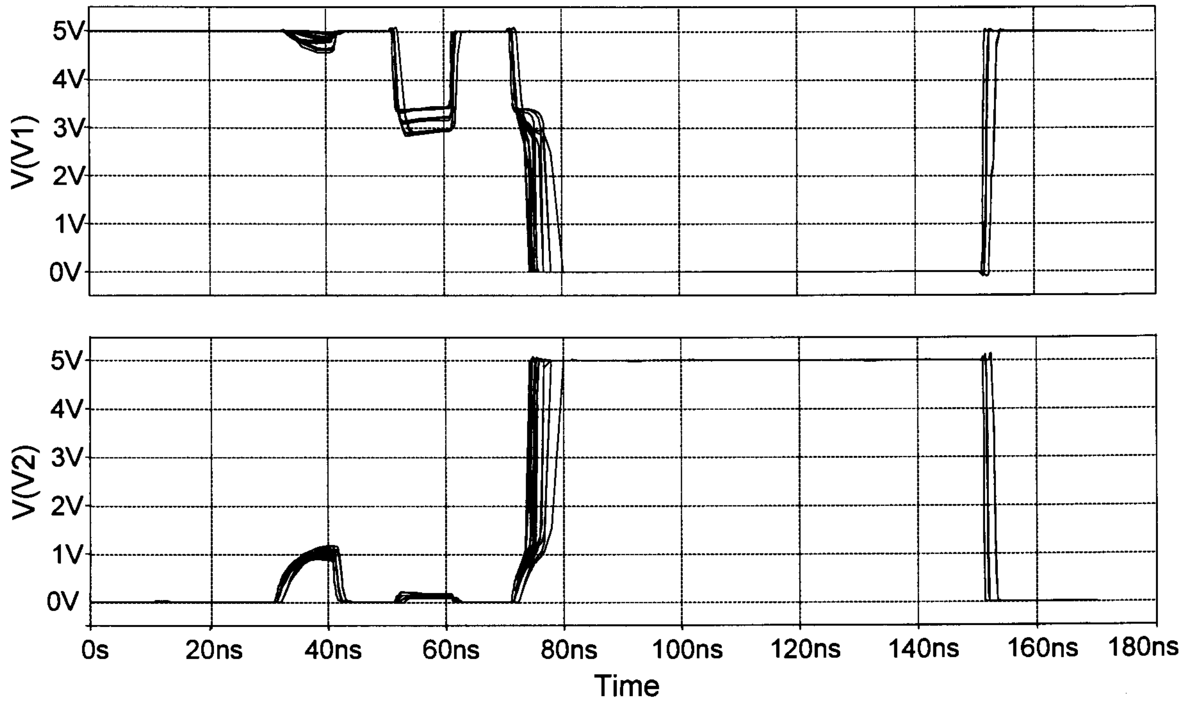


Figure 3.29 Voltages of SRAM internal nodes for different process conditions when the loading of the drivers is maximum.

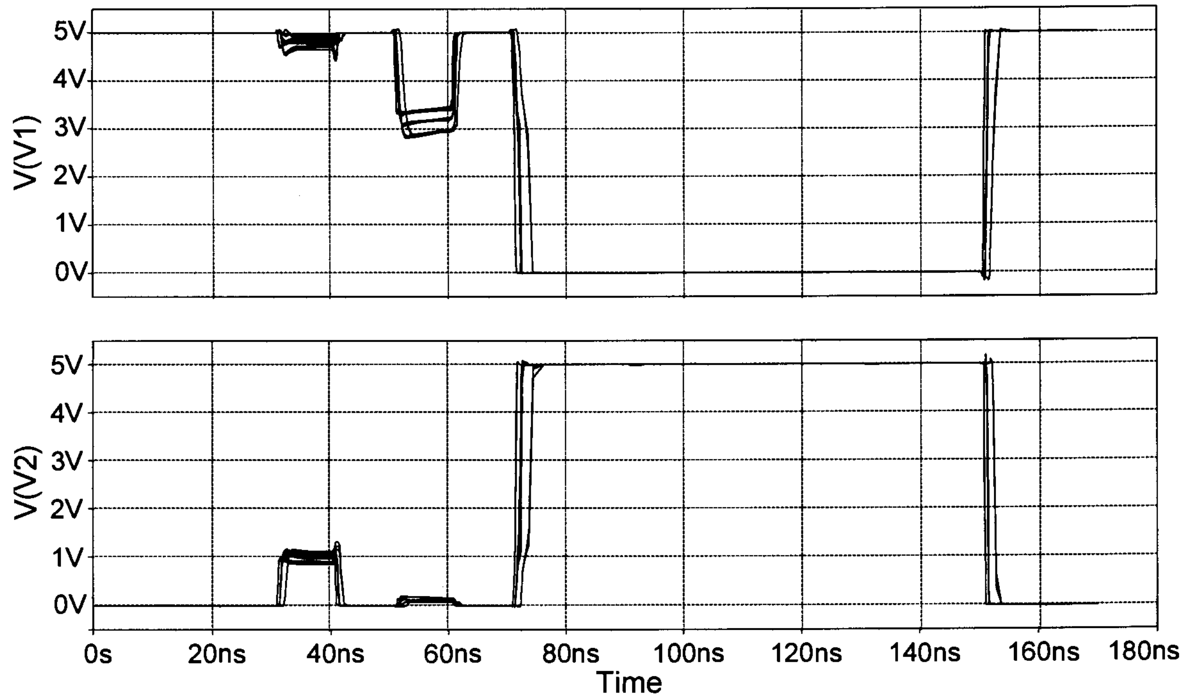


Figure 3.30 Voltages of SRAM internal nodes for different process conditions when the loading of the drivers is minimum.

Table 3.3 Simulated minimum, typical and maximum specifications for the four-transistor SRAM cell together with maximum load peripheral circuits.

Specification type	Read current (μA)	Wasted write current (μA)	Read access time (ns)	Write time (ns)	Clear time (ns)
Minimum	27.3	4.8	1.27	3.53	1.07
Typical	52.5	20.2	1.99	4.92	1.65
Maximum	80.1	48.5	3.48	8.20	2.86

Table 3.4 Simulated minimum, typical and maximum specifications for the four-transistor SRAM cell together with minimum load peripheral circuits.

Specification type	Read current (μA)	Wasted write current (μA)	Read access time (ns)	Write time (ns)	Clear time (ns)
Minimum	28.5	2.50	1.20	1.43	0.98
Typical	54.8	15.97	1.83	2.16	1.49
Maximum	83.9	43.0	3.16	3.88	2.66

Figures 3.29 and 3.30, as well as Tables 3.3 and 3.4 show the results of the simulations. Comparing the figures to Figure 2.23 it can be seen that the cell operates correctly for all process variations. Comparing the parameters given in the tables to those of the cell alone, shown in Table 2.2, it can be seen that the timing specifications and the spread on the currents have increased. This is due to the less than ideal control voltages applied. The timing specifications have increased due to the delay of the peripheral circuits. The timing data show some signs of being slow for the maximum load situation, especially the write access time. This is mostly due to long delays in the *D*/*O*-line driver circuit. Comparing Figures 3.14 and 3.15 to Figures 3.25 and 3.26, shows that the *D*/*O*-line driver response deteriorates more than the *RW*-line driver response as the load conditions worsen, mostly because the *D*/*O*-line driver drives a high switched capacitance and the *RW*-line driver not. In order to establish how much of the simulated delay lies in the peripheral circuits, the delays from the node voltage of the SRAM cell to the desired effect are given in Table 3.5 for comparison. The

times given in the table were measured from the simulation results of the minimum load simulation. This gives the best estimate of the delay in the peripheral circuits, because the delay due to high capacitance has been removed.

Table 3.5 Delay specifications from applied control signal on the SRAM nodes to the required response.

Specification type	Read access time (ps)	Write time (ps)	Clear time (ps)
Minimum	176	367	60.3
Typical	355	634	117
Maximum	716	1490	333

These specifications relate well to those of Table 2.2, meaning that the driver circuits, as well as their control, add delay to the system. Herewith it is proven that the delay of the SRAM itself is not increased. If this were so it would mean that for instance dynamic write conditions are weakened and could be an indication of poor control circuits. It does seem as if the voltage sources are capable of correctly applying the required control signals.

Table 3.6 shows the maximum and minimum read and wasted write currents compared to those that would be present if the process adaptive voltage generators were not used. Here it can clearly be seen that although the spread is no longer as ideal as it is in Table 2.2, a definitive advantage can be drawn from using the described voltage generators. A clear reduction in the range can be observed. Especially the maximum value which is the one that potentially causes highest power dissipation, is reduced by 20% and 33% for the read current and the wasted write current respectively. For those manufactured systems close to the worst case power specification, this implies a power saving of the stated percentages in comparison to a system where fixed voltage control is used.

Table 3.6 Comparison between the currents when adaptive voltage control and fixed voltage control are used.

Control mechanism	Read current (μA)			Wasted write current (μA)		
	Minimum	Typical	Maximum	Minimum	Typical	Maximum
Fixed voltage	19.3	53.9	106.2	5.0	17.7	63.0
Adaptive voltage	28.5	54.8	83.9	2.50	15.97	43.0

3.7 LAYOUTS

Layouts for the driver circuits were created, so that it may be verified how much area they require in relation to the array of cells. The switching circuits should typically fit into the pitch of the SRAM array. This was not possible for the *DIO*-line driver switches, or the *CL*-line drivers, or the *RW*-line driver switch circuits. The reason is that the pitch of the cell is too small to accommodate the circuits. It was therefore decided to fit the circuits into double the pitch and to place two next to each other to drive the cell rows. The low-impedance driver and op-amp of the *DIO*-line driver were designed to fit the vertical pitch of eight cells.

Wherever matching between devices is required, common-centroid layout was used and the orientation of devices was kept identical. This has been shown to reduce the offset voltage of differential pairs [24]. The third metal layer was used to ease routing, but no high resistive poly or poly capacitor modules were used. The circuit can therefore be manufactured using only a standard CMOS core module.

The source driver circuits are a combination of sensitive analog circuits operating with small currents (op-amp input stage and bias networks) and circuits that switch large currents (low-impedance driver circuits). This can cause interference if the substrate is not isolated adequately. For this purpose, the large transistors that switch large currents or charge large capacitance were adequately surrounded by substrate contacts. If the geometry allowed it, guard rings were used. The aim was to create the lowest possible resistance in the bulk, to prevent voltage spikes. The same was done for all sensitive analog components. The power supply tracks were also made wide to prevent high resistance building up and causing a voltage drop at high currents.

Figures 3.31 to 3.37 show the layouts of the driver circuit building blocks. The legend is given in addendum B.

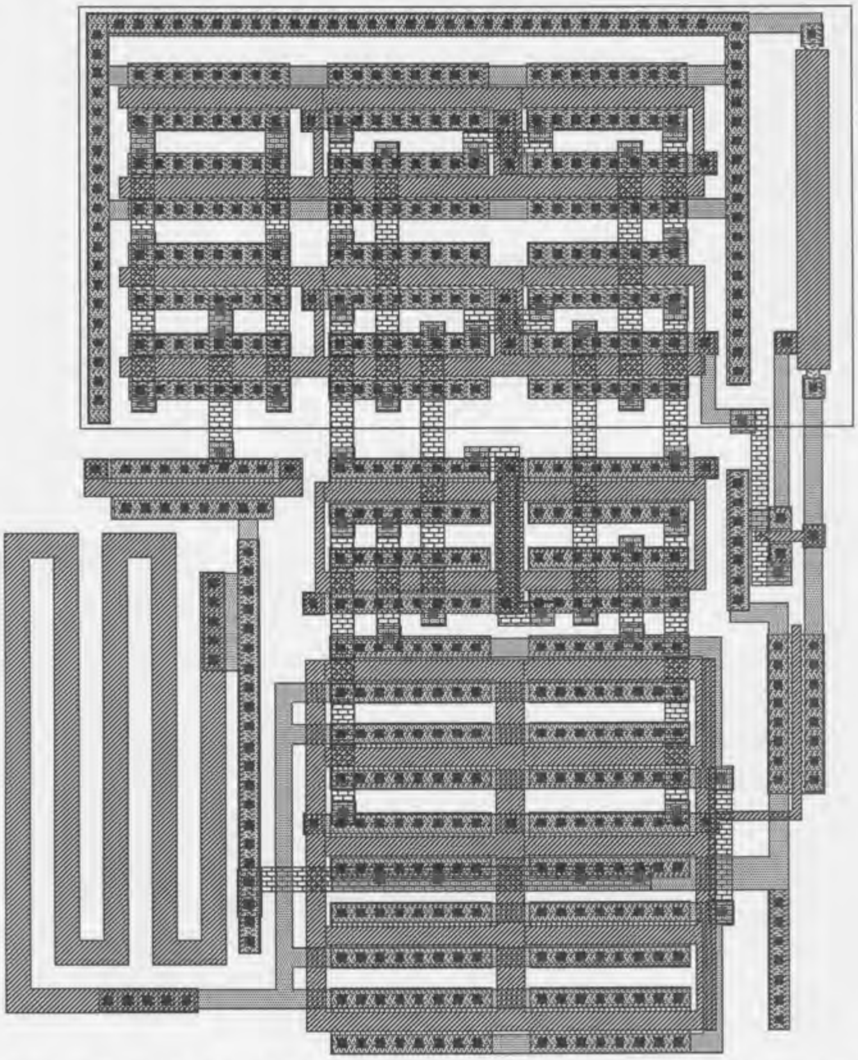


Figure 3.31 Layout of the *D/O*-line driver reference voltage and bias network.

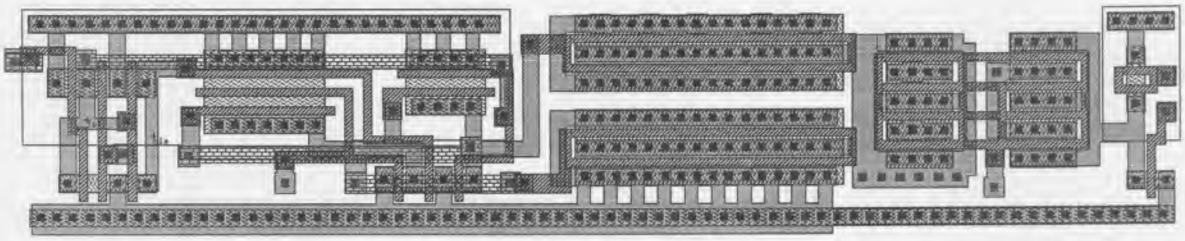


Figure 3.32 Layout of the *D/O*-line driver switch circuit with some peripherals added that are discussed in Chapter 4.

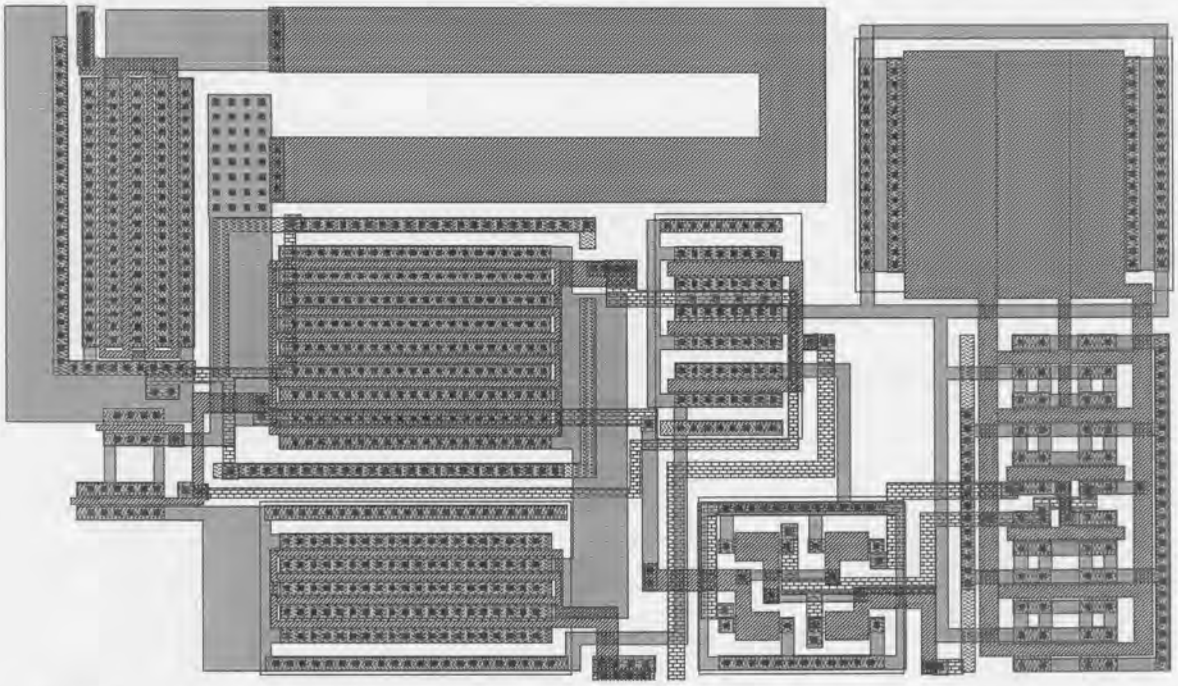


Figure 3.33 Layout of the *D*/*O*-line driver op-amp and low-impedance driver circuit.

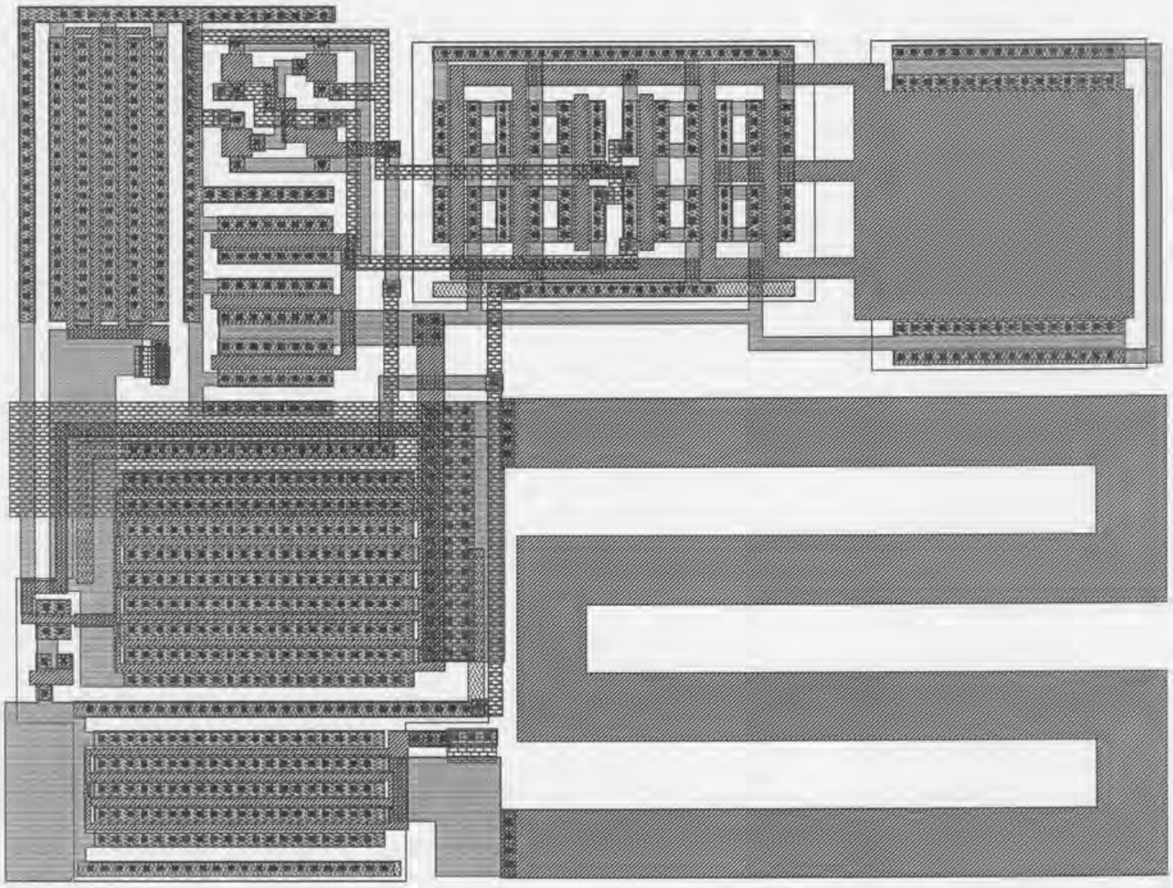


Figure 3.34 Layout of the *RW*-line driver op-amp and low-impedance driver circuit.

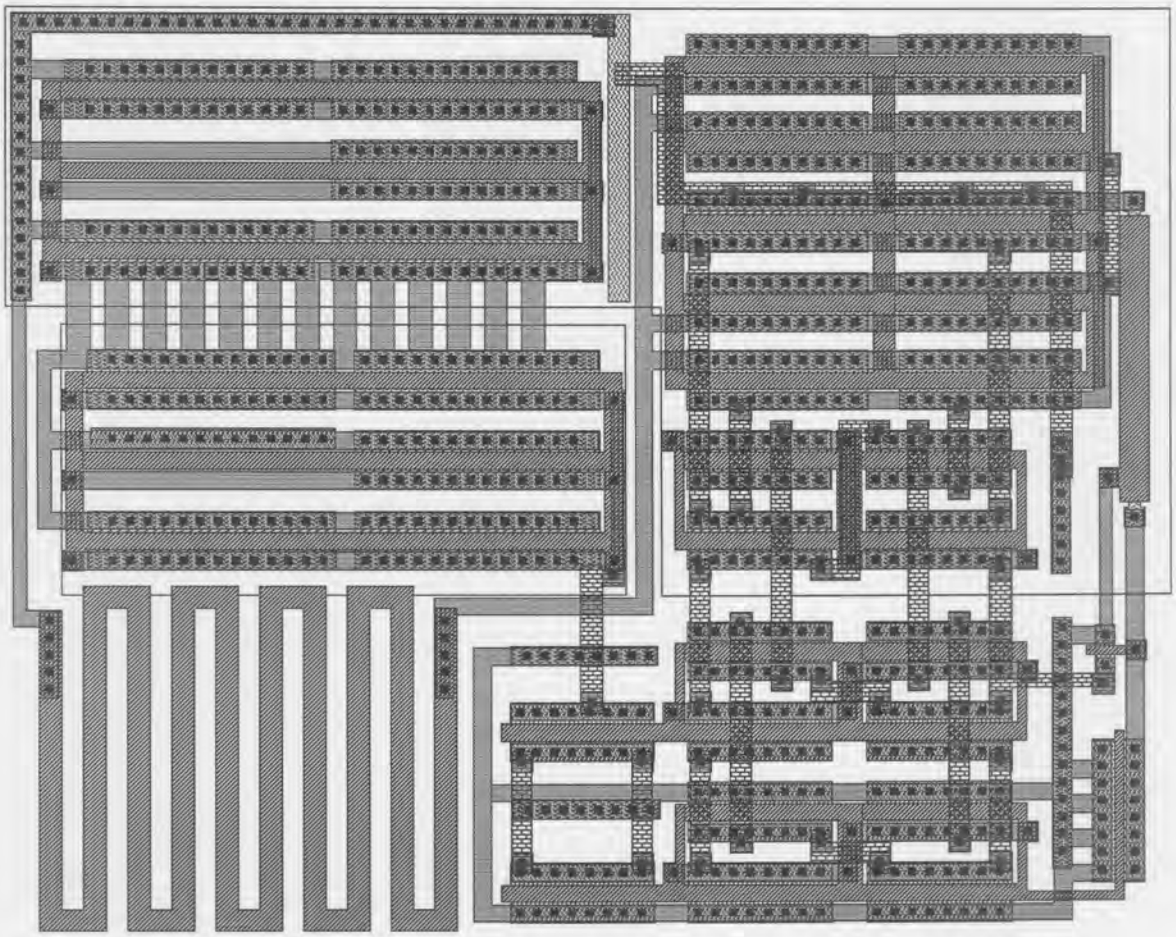


Figure 3.35 Layout of the *RW*-line driver reference voltage and bias network.

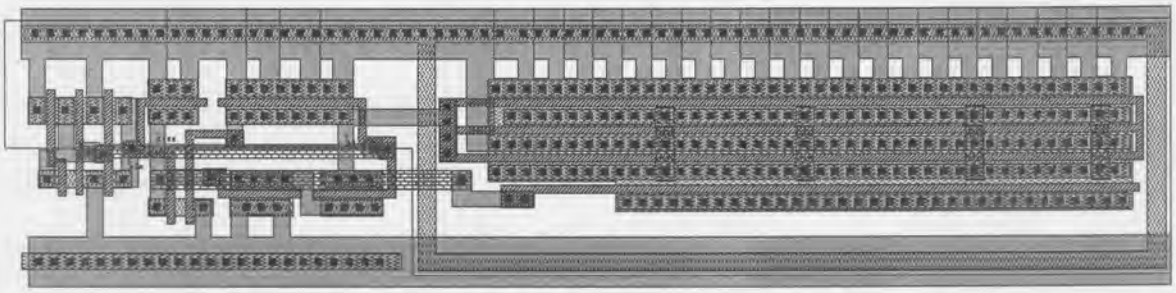


Figure 3.36 Layout of the *RW*-line driver switch circuit.

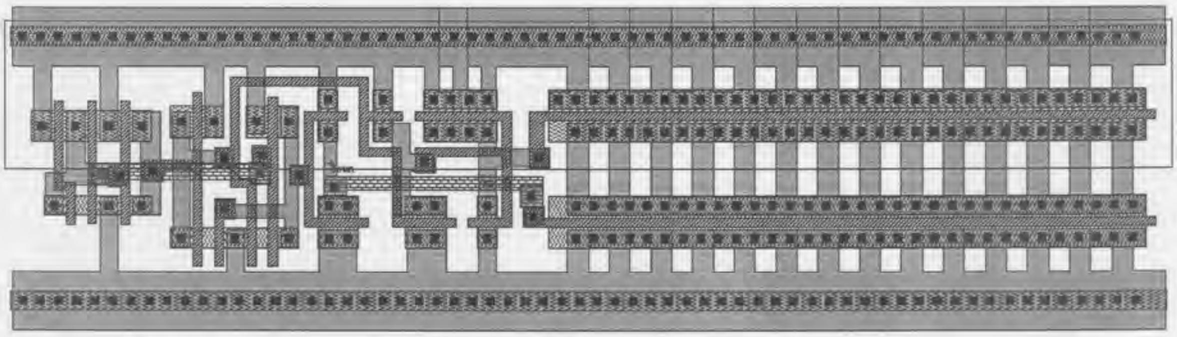


Figure 3.37 Layout of the *CL*-line driver circuit.

3.8 CONCLUSION

This chapter discussed the design procedures of the driver circuits required to correctly control the four-transistor SRAM using the methods proposed in Chapter 2. The three driver circuits for the *DIO*-, *RW*- and *CL*-line were designed and simulated. To prove that all circuits can operate together to form a system, a complete simulation was performed. The results indicate successful operation. The specifications of the system that were extracted from the simulation results indicate that the design goals set during the discussion of the SRAM cell were met. The driver circuits add significant delay, but this is considered to be inevitable. The *DIO*-line driver seems to be the worst as far as delay is concerned, especially when the loading is worst case. An effort to increase the speed of the system should initially focus on this circuit.

The current spread as process conditions change, and especially the maximum currents during the read and write cycle, have been reduced effectively by designing voltage reference circuits that adapt to the process conditions. These voltages are buffered to drive the capacitance associated with the SRAM array. To achieve this effectively, two feedback loops have been used, where the inherently fast loop keeps the output voltage constant as the load conditions change and the slower loop is used to ensure that the circuit performance is mostly independent of process conditions.

4. CURRENT SENSE AMPLIFIER

4.1 INTRODUCTION

The control of the source nodes of the four-transistor SRAM cell was discussed in the previous chapter. The circuits initiate the read and write actions using digital signals as inputs. The read output is a current flowing in the *D/O*-line to ground. A sense amplifier is required to transform this current into a digital voltage signal. The initial design choices of the SRAM cell dictate that the absence of a current indicates a "one" is being read, whereas the presence of a current indicates a "zero" is stored in the cell.

In this chapter the design and simulation of the current sense amplifier is discussed. The function and operation of different sensing structures are evaluated to find the sensing system best suited for the four-transistor SRAM cell. The chosen circuit is designed and simulated.

4.2 SENSING SRAM CELLS

4.2.1 Voltage Sense Amplifier System

Early SRAM systems used voltage-mode sensors [5], [25]. When the access transistors of a six-transistor SRAM cell are activated, one of the precharged bit lines is discharged into the cell via the access transistor. The other bit line remains at the precharged voltage level. This means that a differential voltage is created across the bit lines.

The differential bit line voltage is commonly sensed using a differential amplifier system, like the one shown in Figure 4.1 [25]. Two differential amplifiers with the input signals applied in the opposite configuration are used. This is a differential input to differential output amplifier. The output signal from the first stage is then used as input to the second stage amplifier, a differential input to single-ended output configuration. The high gain of the two cascaded stages ensures that even a small differential voltage on the bit lines causes a logic level voltage swing on the output node.

This amplifier system is simple to bias and its operation is independent of the bit line common-mode voltage. As the number of cells connected to one bit line increases, the length and the capacitance of the bit lines also increases. Combined with the resistance of the bit line this causes increased sensing delays.

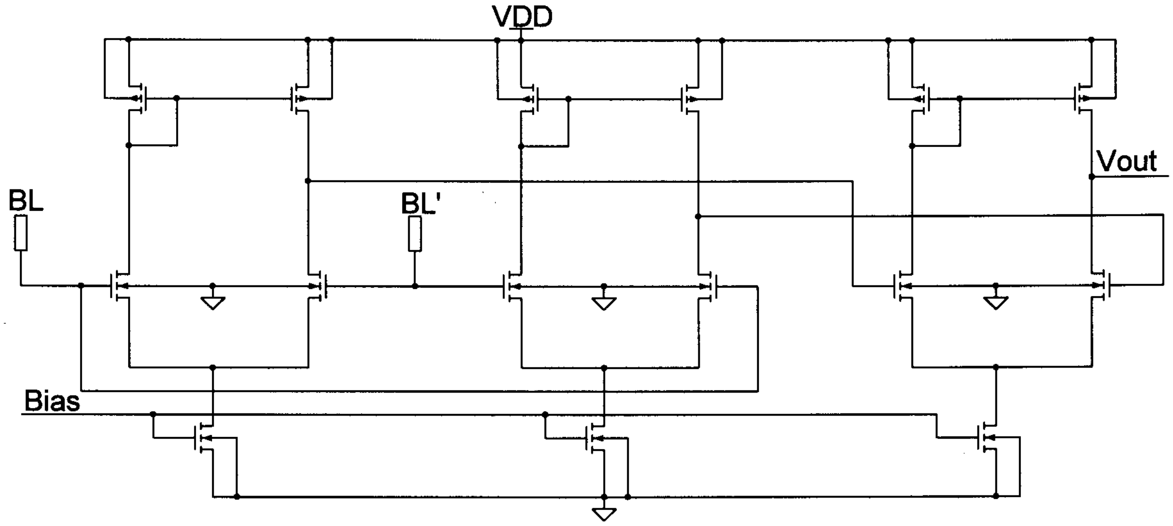


Figure 4.1 Voltage sense amplifier.

The voltage sense amplifier senses a differential voltage and can be slow due to high bit line capacitance. This makes it useful for sensing small arrays of six-transistor SRAM cells. The four-transistor cell has a current-mode output rendering voltage-mode sensing circuits unusable. A current sense amplifier is required.

4.2.2 Current Sense Amplifier Systems

To enable a current to be sensed, as well as overcoming the increasing delays, current-mode methods were developed, where a differential current, rather than a differential voltage is sensed. The reasoning behind this is the fact that the delay through an RC-tree network for voltage-mode signals is about 20 times longer than the delay for current-mode signals [26]. Voltage-mode signals typically require an infinite terminating impedance, but current signals require a short circuit. A current-mode sensor should therefore present as low an impedance as possible. If the input impedance is too high, a voltage swing is required during operation, the speed of which is dependent on the associated capacitance. The aim of current-mode techniques is to minimise or totally avoid this swing.

During access in the six-transistor SRAM cell, one access transistor conducts a current and the other not. This creates a differential current into the sense amplifier, as shown in Figure 4.2. The voltages of the two bit lines are equal because the current-mode sensing circuit has zero differential input impedance, so the currents I_a and I_b are equal. The sensing circuit requires a bias current, *i*, and the accessed cell draws a current *I* on one side and no current on the other. This presents the sensing circuit with the differential current *I* at the input nodes.

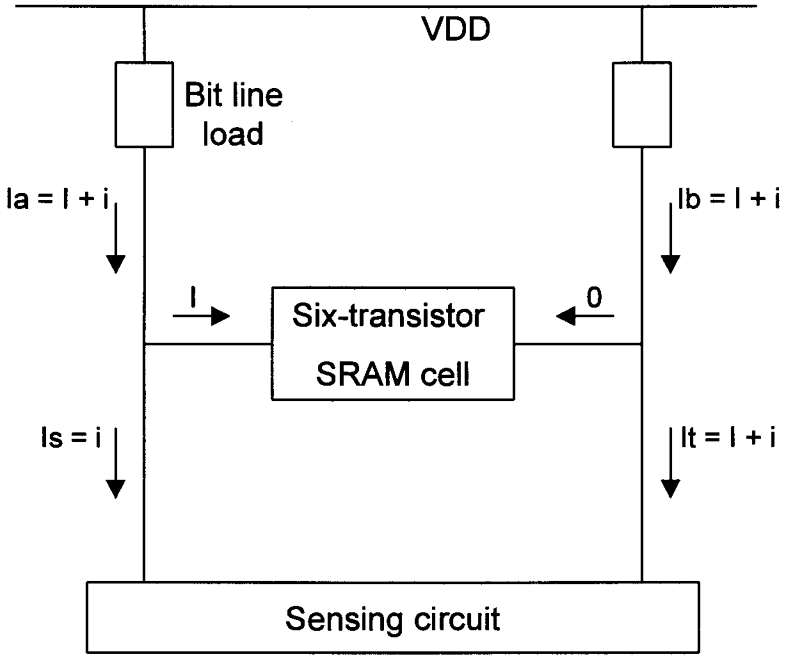


Figure 4.2 Current-mode sensing theory for an SRAM cell.

Current Sensing Circuit

In order to implement a current-mode sensing scheme a circuit is required that converts a differential current to a differential voltage, typically by guiding the currents through a load device. Such a circuit is shown in Figure 4.3, where the differential current is guided through two diode-connected devices, *M1* and *M2*. This creates a differential voltage between node *Va* and *Vb* which is sensed by a differential voltage amplifier. The output stage in the form of an inverter ensures the signal is amplified to logic voltage levels.

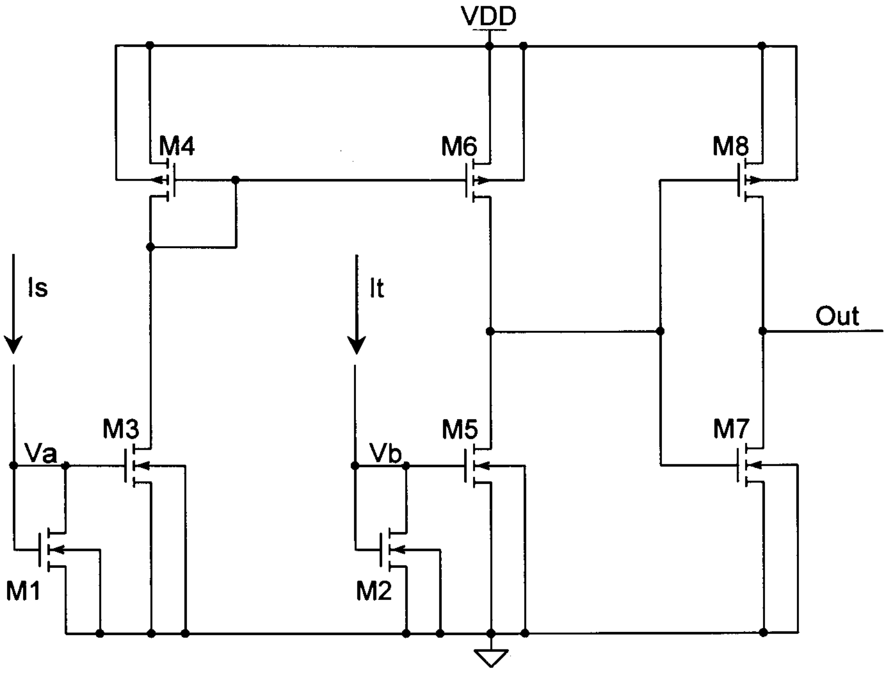


Figure 4.3 Differential current sense circuit.

To operate, this topology relies on a differential voltage to be present on the input nodes, and this implies the differential input impedance is not low. The voltages of the bit lines still have to be changed, and this is slow due to the large associated capacitance.

Current Conveyor Circuit

To overcome this problem a current transport circuit with ideally zero input impedance is required. This is essentially a differential current buffer. A low impedance is presented to limit the voltage swing of the bit lines. The transport circuit decouples the differential voltage swing required for sensing from the bit lines, without modifying the differential current. This allows the current-mode signal to be converted to voltage-mode while presenting a very low differential input impedance to the bit line pair. A simple circuit to implement this transport function is the current conveyor [26] shown in Figure 4.4.

The operation of the circuit is as follows. Assume the circuit is implemented using four equally sized devices, as is typically the case. The gate-source voltages of *M1* and *M2* are equal to *V1* because they are both saturated and carry the same current. This is also valid for *M3* and *M4*, where the gate-source voltage is *V2*. The gate voltages of *M2* and *M4* are fixed at ground, but may also be fixed at any other constant bias voltage. From this it follows that the voltages of the input lines are *V1+V2*, and *V2+V1* respectively. They are therefore equal, regardless of the current distribution, and this is seen by the driving circuit as a zero differential-mode input impedance. Some non-ideal behaviour is present because of the body effect which causes the threshold voltages of all devices to differ. This will cause the gate-source voltage of for example *M1* and *M2* to differ and the voltages of the input nodes are no longer equal. When both branch currents are equal, the input lines have the same voltage, but if one branch carries a higher current than the other, its voltage will tend to rise. This means a positive differential input impedance is present. This non-ideal behaviour actually aids in preventing a negative input impedance which could result in latching behaviour [26].

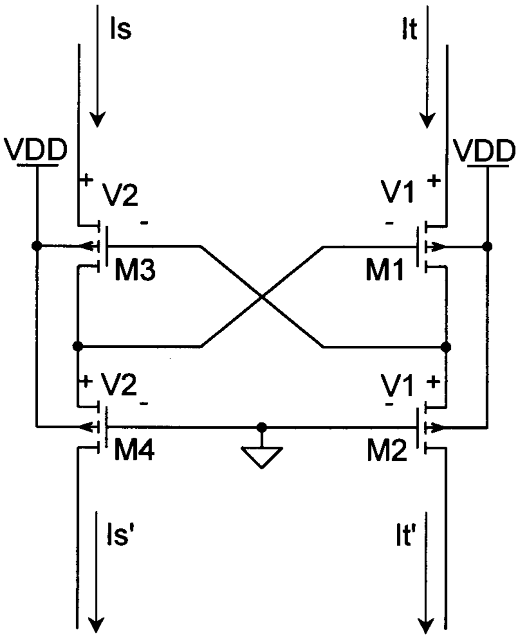


Figure 4.4 Differential input, differential output current conveyor circuit.

The input currents are passed to the output nodes without change, where they may be converted to voltages across load devices without affecting the voltages of the input nodes. As long as the load voltage drops are small enough to keep *M2* and *M4* biased in the saturation region, a differential current can be correctly sensed. Placing Figure 4.4 between the bit lines and the input of the sensor of Figure 4.3 creates a very effective current sense amplifier with sensing delay independent of the bit line capacitance [26]. The sensing delay only increases for very high bit line capacitance values.

Latched Sense Amplifier

The circuit configuration discussed up to now is capable of sensing the large differential current present in the six-transistor SRAM cell. A simulation of the six-transistor SRAM cell of Chapter 2 indicates the differential current to be in the region of 250μA. The results of Chapter 2 also indicate the sense current of the four-transistor cell to be 45μA which would result in a maximum differential current of 22.5μA. This is an order of magnitude smaller than the differential read current of the six-transistor cell. Sensing with the circuit of Figure 4.3 would be slow because the smaller current takes longer to modify the voltage across the load devices.

A more sensitive sensor is required. A good choice is the latched sense amplifier that is normally used for sensing dynamic RAM systems where the differential read currents are in the order of 0.3μA [28].

The speed of a latched sense amplifier is based on the fact that positive feedback is used to amplify very small signals. This is typically achieved by using a cross-coupled inverter pair. The circuit is forced to assume a metastable state by equalising the voltages on the two inverter nodes. Typically a single device or a transmission gate can be used for doing this. The differential current signal is then applied to either the PMOS source nodes or the NMOS source nodes, causing a very small shift in the operating points of the two inverters. Due to the very high gain and the positive feedback, this shift is quickly amplified to digital voltage levels.

Another advantage of using a latched amplifier is the fact that the sensed value is latched in the sensing structure. No additional latch is required, and this saves area and reduces delays.

A sensing system where the current conveyor is used in conjunction with a latched sense amplifier is shown in Figure 4.5 [29].

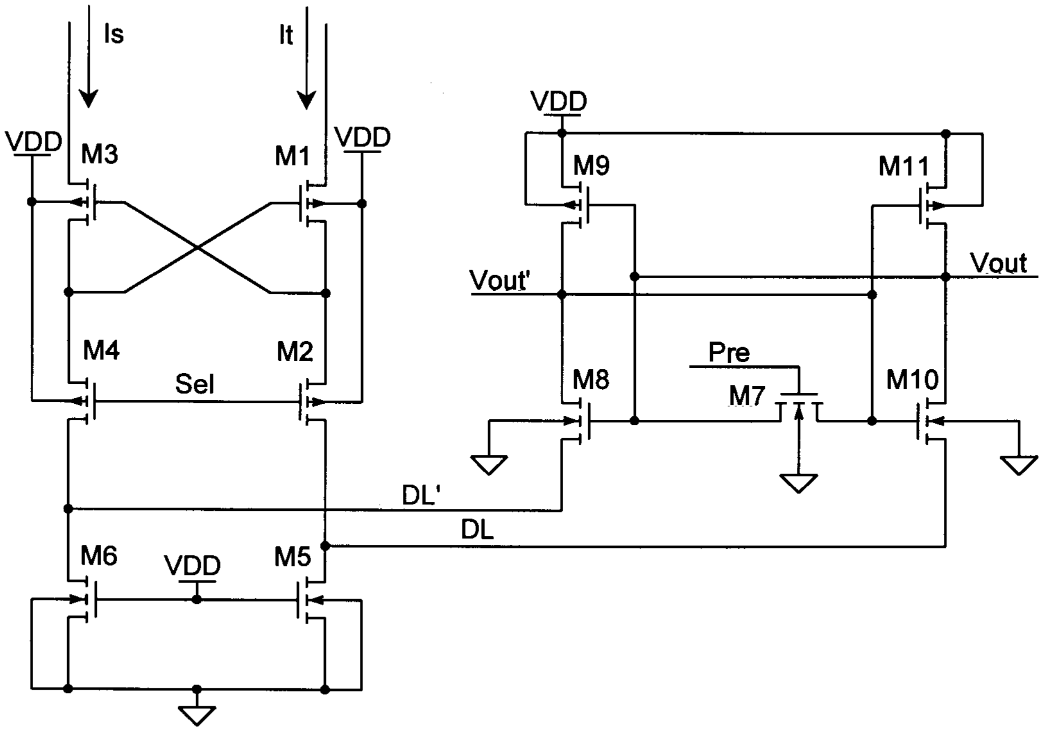


Figure 4.5 Current-mode sensing system using a cross-coupled inverter pair as a sense amplifier.

The current conveyor transports the differential current. The cross-coupled inverter pair is placed in the metastable state by activating the *Pre* line. This turns on *M7* and equalises the input voltages of the two inverters. After a predetermined delay, *M7* is turned off again. The differential current causes a slight differential voltage to occur in the data lines due to the small resistance to ground of the linear devices *M5* and *M6*. This differential voltage causes a slight imbalance in the gate-source voltages of the NMOS devices of the cross-coupled latch. As a result of this, one of the devices will become stronger and tend to pull the associated internal node "low", while the other device will become weaker and the internal node associated with it will to be pulled "high". This causes an imbalance in the latch, and due to the high small signal gain and the positive feedback, the outputs are driven to logic output levels. They remain there until the latch amplifier is reset for the next sensing cycle by activating the *Pre* signal. For example, if the input current *It* is larger than the input current *Is*, then the voltage of *DL* will be slightly higher than that of *DL'*, and node *Vout* will be pulled "high" and *Vout'* "low". The sensing delay is independent of the bit line capacitance and almost invariant to the data line capacitance. This is due to the fact that the voltage changes required on the data lines to create a fast response can be very small, given the high gain of a cross-coupled inverter pair in the metastable condition.

The gate nodes of devices *M2* and *M4* in Figure 4.5 are shown connected to a select line, *Sel*. This line can be used to activate the current conveyor, and allows several current conveyors connected to the same data lines, *DL* and *DL'*, to be individually activated, allowing current-mode multiplexing. This is very useful when using a single sense amplifier to sense currents from different banks of the memory system.

Clamped Bit Line Sense Amplifier

The response time of the cross-coupled inverter pair is dependent on its small signal gain-bandwidth product [30]. The fastest response time is thus obtained by maximising the gain-bandwidth product. The first parameter to optimise is the device ratio, the ratio between the NMOS and PMOS device size. The highest gain-bandwidth product is achieved if this ratio is 1 [30]. A second method is to slightly modify the current sense amplifier by adding another equalisation device to the data lines. The clamped bit line sense amplifier in Figure 4.6 is suggested to have a gain-bandwidth product an order of magnitude higher than the conventional sense amplifier of Figure 4.5 [31].

This equalisation device is activated using the same signal *Pre* as the cross-coupled inverter pair equalisation device. When active, it conducts the difference between the currents applied to the sense amplifier, and holds the data lines at the same potential. To begin sensing, both equalisation devices are turned off and the cross-coupled inverter pair once again acts as a high gain positive feedback amplifier. The data lines which usually have significant associated capacitance, have been held at the same potential before sensing, but the equalisation device enforcing this has now been removed. The difference current therefore has to be supplied via a different path, and the only one available is the cross-coupled inverter pair. Sourcing this additional current causes a voltage differential to occur across the output nodes of the cross-coupled inverter pair, and this is amplified to a stable state.

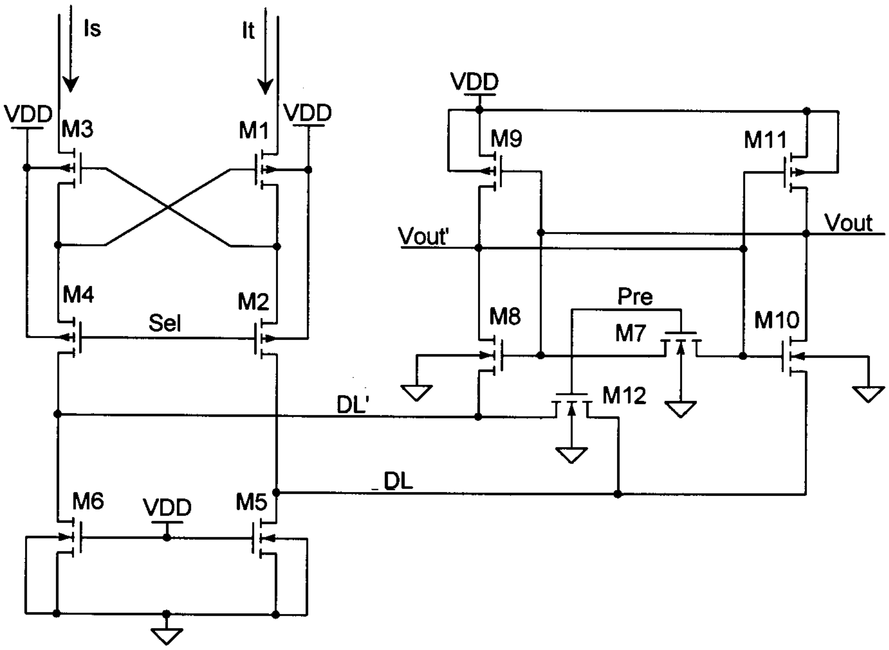


Figure 4.6 Clamped bit line current-mode sense amplifier.

A small signal analysis presented in [31] concludes that by adding the data line clamping device a zero in the region of the first pole is added to the small signal gain of the conventional sense amplifier. This increases the gain-bandwidth product and results in higher sensing speeds. The clamping device removes the need to slightly change the voltage of the high capacitance data lines and this results in improved speed. The speed is now limited by the internal capacitances of the cross-coupled inverter pair and these are an order of magnitude smaller than the data line capacitances.

4.3 CURRENT SENSE AMPLIFIER DESIGN

In view of the discussion of the previous section, it was decided to implement the current sense amplifier using the clamped bit line sense amplifier together with the current conveyor. The latter is used to ensure low input impedance and to decouple the sensing action from the source nodes of the SRAM cells. If the current conveyor were not used, the currents that flow in the sensing latch, especially while it is switching, would affect the SRAM NMOS source node directly. This situation should be avoided because it could cause voltage and current spikes that degrade the noise margin of the SRAM cells. The current conveyor allows sensing, while protecting the SRAM cell array from any transients that occur in the process.

4.3.1 Reference Current

All current sensing circuits discussed in the previous section require a differential current. As with voltage sensing, current sensing techniques seem to be based on differential signals. Their value is usually more process independent and they usually have a speed advantage over single-ended sensing techniques. The four-transistor SRAM cell generates a single-ended current, and the information lies in the fact that a current is present or absent. In order to sense this, a reference current for comparison is required. If this reference current is half the magnitude of the expected cell current, a dual polarity differential current signal is created, as can be seen in Figure 4.7. This is the structure of the signal required by the sense amplifier.

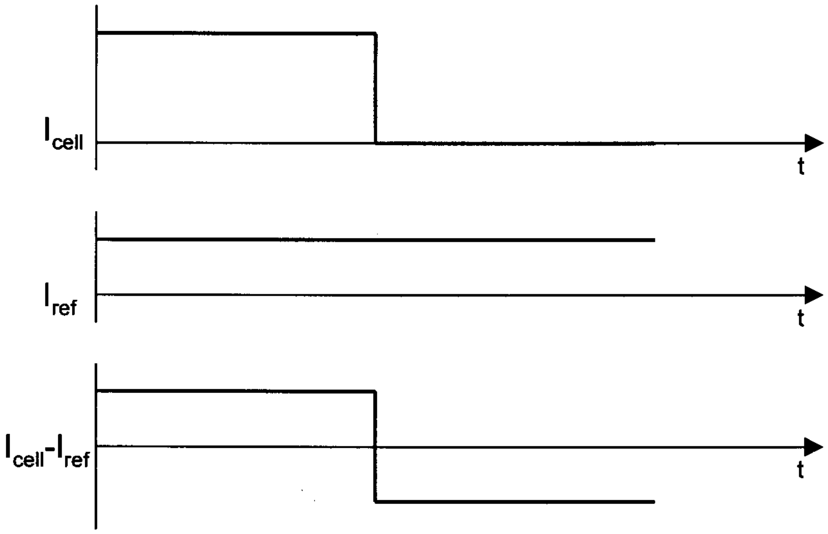


Figure 4.7 Generating a differential current by using a constant reference current of half the cell current amplitude.

The reference current should be generated using the same circuits that are used to generate the cell current. This allows the reference current to change in response to environment and process conditions, just like the cell current does.

The scheme of reading the cell is considered in Figure 4.8(a). It can be seen that the voltage deviation of the *RW*-line is applied to the input terminal of the opposite inverter without much of a change, if the appropriate devices are on. When this happens, the current is dependent only on the PMOS device in saturation (*M2*) because the NMOS device (*M1*) is linear and therefore acts only as a low resistance. As far as the magnitude of the current is concerned, the two circuits depicted in Figure 4.8 are almost equivalent if the transistor sizing is equal.

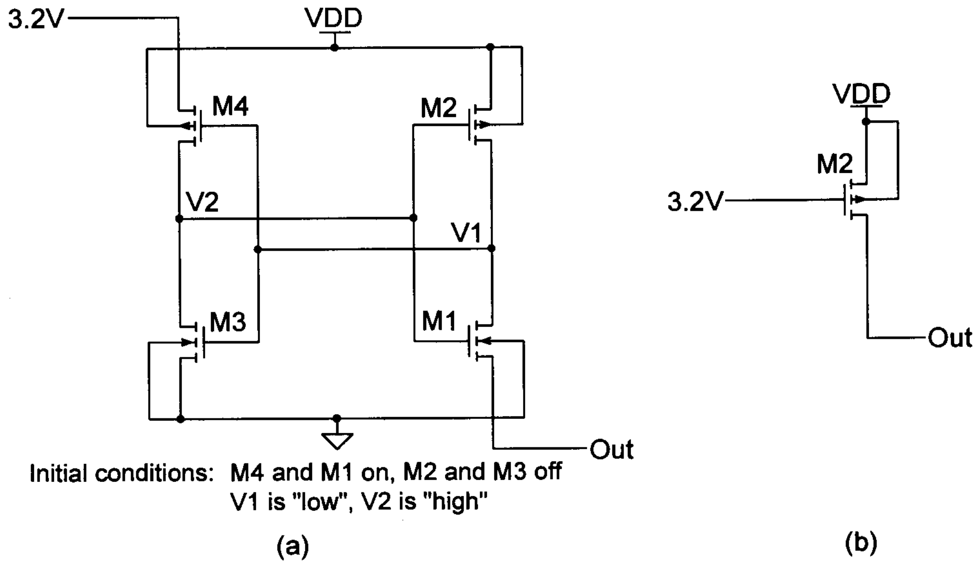

3.2V
VDD
M4
M2
V2
V1
M3
M1
Out
Initial conditions: M4 and M1 on, M2 and M3 off
V1 is "low", V2 is "high"
(a)
VDD
M2
3.2V
Out
(b)

Figure 4.8 Two virtually equivalent circuits for generating read current. *Out* is the output node where the current flows and is connected to a potential very close to ground.

When applying equation (2.2) to the circuit of Figure 4.8(b) it is evident that the magnitude of the current can be scaled by scaling the device size. The required reference current has to be half the magnitude, which implies the *W/L* ratio has to be half of that of the cell devices. The width cannot be halved, as this would result in a device width that is smaller than minimum. The length should therefore be doubled, but this causes the short-channel effect to become less prominent and the device characteristics will change. The best method is therefore to use two minimum area transistors in series.

4.3.2 Current Switching

The cell current, as well as the reference current, need to be steered through the current conveyor. There are four memory banks in the system, but only one word is accessed at any time. The four banks need to be multiplexed into the current sense amplifier, so that only a single sensor per bit is required instead of one sensor per bit per bank. The overhead required to do this multiplexing should be small. Using a single sensor together with a current multiplexing circuit reduces delays and saves area.

Current-mode multiplexing is very simple, because it only entails switching a specific current into a node. This is done with a single pass device biased in the linear region. This device should be wide enough that the voltage drop is not too large. Assume a voltage deviation is applied to the *RW*-node of the cells. If the read current now causes a voltage deviation on the *D/O*-node, this decreases the noise margins further. Some voltage drop over the current-conveyor pull-down device is however required to ensure that the current conveyor can operate properly. The *D/O*-line switch pull-down device width of 40μm is too large, because the voltage drop is so small, it causes the current conveyor to have a low differential current gain. The switching circuit of the *D/O*-line driver therefore needs to be modified as is depicted in Figure 4.9 to turn off the pull-down device *M1* during reading and activate a second pass transistor *M3*. This transistor connects to the current conveyor and is also the device used for the current-mode multiplexing. An appropriate pull-down will be supplied in the current conveyor circuit.

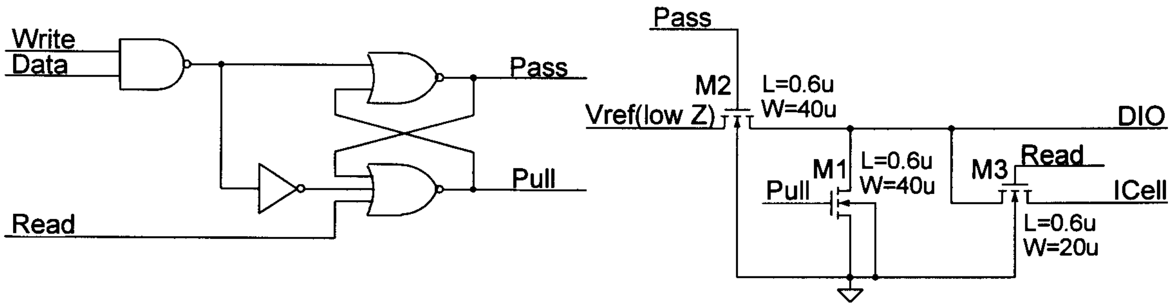


Figure 4.9 Modified *D/O*-line driver to allow the pull-down *M1* to be disconnected and the output current to be guided to the *ICell* node.

Several of these switching circuits, one from each bank, can be connected to a single current conveyor and pull-down load and only one is activated at a time.

The dimension of the current switching device was chosen large so that the voltage drop across it when the read current flows, is insignificant. The typical read current is 45μA and a device of 20μm width has a resistance of 140Ω (Figure 3.3). This produces a voltage drop of 6mV which is an order of magnitude smaller than the voltage deviation that can be allowed.

A similar multiplexing structure is used for the reference currents. It was decided to generate the reference currents at each memory bank, because this equalises the delays between the reference currents and the cell currents and also improves the matching between the two. The complete reference generator for one bit is shown in Figure 4.10.

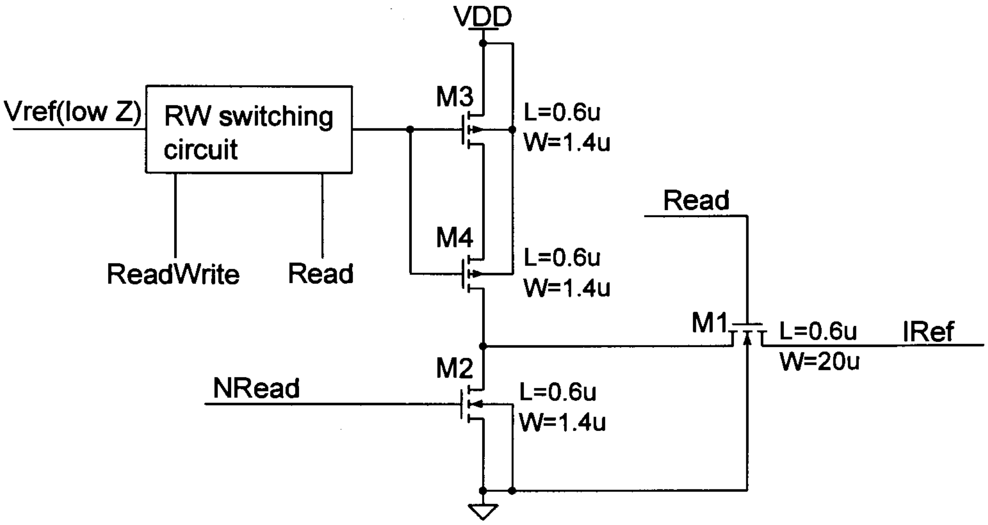


Figure 4.10 Reference current generator circuit.

A switching circuit like that used for the *RW*-lines is used to drive the gate nodes of the two current source devices, *M3* and *M4*. The switch circuit is connected to the same *RW*-line low-impedance driver as the cells, to ensure identical conditions between the cells and the reference current generator. The control signals used are the *ReadWrite* and the *Read* signal. The latter prevents the activation of the reference currents during the write cycle. During the read an identically dimensioned pass device to the pass device for reading in the *D/O*-line driver (*M3* in Figure 4.9), is activated. The *IRef* nodes are connected to the other side of the current conveyor. A pull-down device *M2* is turned on when the reference current is not required. This was done to pull the voltage down to ground, just like it is done with the *D*/*O*-line when the cells are not being read. This once again aids in establishing almost identical operating conditions.

4.3.3 Current Conveyor and Loads

The design of the current conveyor is based on a compromise between differential current gain and low load resistances. High load resistances mean a higher gain can be achieved [26]. On the other hand a high load in the case of the four-transistor SRAM cell implies a larger voltage deviation on the *D*/*O*-line and this is not desired because it degrades noise margins. The load resistances are therefore designed to have a voltage drop no higher than 0.1V. The current conveyor circuit used is shown in Figure 4.11.

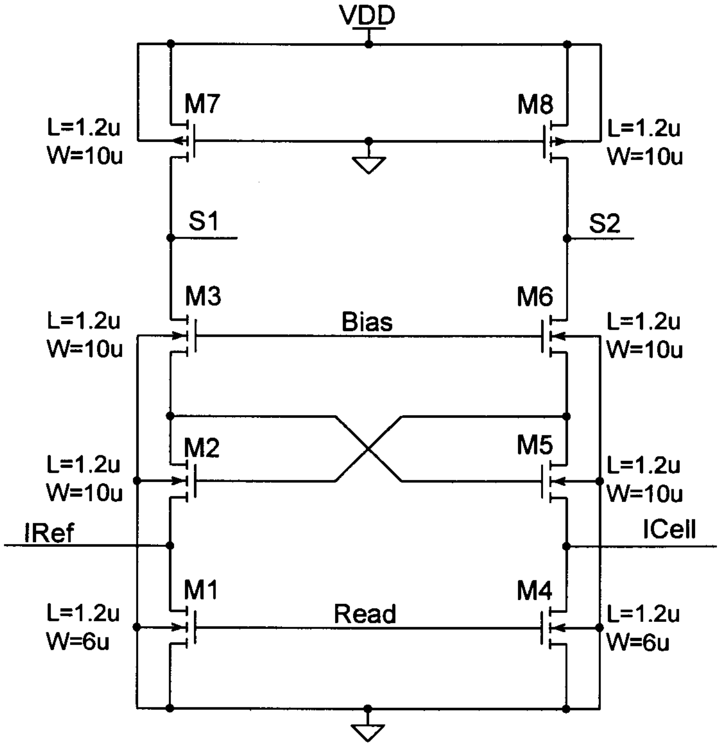


Figure 4.11 Current conveyor circuit for the sensing system.

It can be seen that the bias current of the current conveyor also flows through the load devices, *M1* and *M4*. The current conveyor has to be biased with at least the current magnitude that has to be sensed. This is required to prevent a zero current state in one of the branches. The bias current was therefore chosen as 50μA. A maximum voltage drop of 0.1V across the load devices is allowed at a total current of 100µA. This implies a device resistance of 1kΩ and a device size of 6µmx1.2µm. A larger length is used because the operation of the circuit relies on the matching between the load devices.

For good matching the four devices making up the conveyor are chosen equal size [26]. The gates of *M3* and *M6* have to be biased at a fixed voltage to allow a bias current to flow. It was decided to turn off this bias current by deactivating the load devices, rather than turning off the current conveyor. This allows quicker start-up times. When the conveyor is required during the read cycle, the *Read* signal goes "high" and places the load devices in the linear region. The reference current generators of each bank are tied to the *IRef* node and the respective *DIO*-lines to the *ICell* node. The difference current is conveyed to the sensing nodes *S1* and *S2*. The clamped bit line current sense amplifier is attached here.

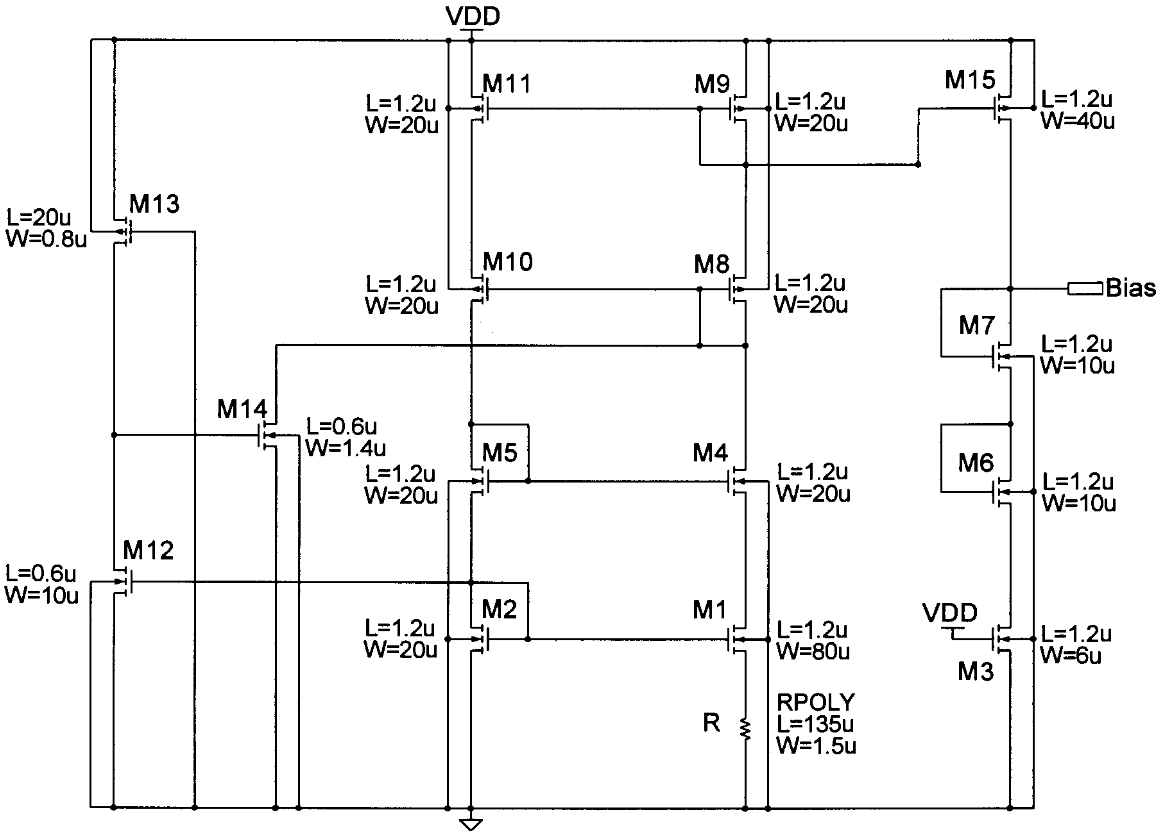


Figure 4.12 Bias network for the current conveyor.

The load devices for these two lines are two devices biased in the linear region (*M7* and *M8* in Figure 4.11). The width is chosen large enough so that the currents required by the cross-coupled inverter pair in its metastable state do not cause a large voltage drop. The two devices are also not minimum length, because good matching is required for reliable operation.

The performance of the current conveyor is dependent on the small signal transconductance of devices *M2*, *M3*, *M5* and *M6*. An identical bias network to the one used for the *DIO*-line driver is therefore required, so that these devices may be biased at a constant transconductance, irrespective of process conditions. The bias voltage is generated by using an identical stack of devices to the current conveyor *M1-M3* or *M4-M6* device stack. The bias network is given in Figure 4.12.

4.3.4 Clamped Bit Line Sense Amplifier

The clamped bit line sense amplifier (CBLSA) is given in Figure 4.13. The sensing node *S1* and *S2* are connected to the output nodes of the current conveyor.

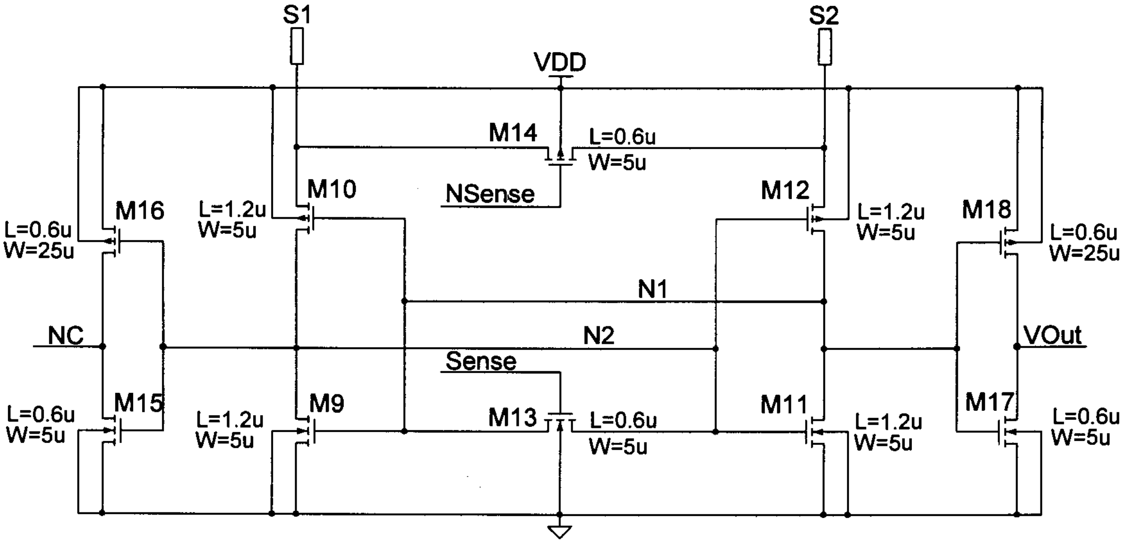


Figure 4.13 Clamped bit line current sense amplifier.

The equalisation device *M13* is chosen strong enough to drive the latch into its metastable state and device *M14* is designed to be able to conduct the differential current between the bit lines at a low voltage drop. The latter device has to be a PMOS because the potential of the nodes it has to equalise is high. The signals *Sense* and *NSense* are inverses of each other. The dimensions of the devices of the cross-coupled inverter pair are chosen equal, to fit with the findings in [30] that this maximises the gain-bandwidth product. If these devices are chosen wide, the response is fast but the current flowing in the metastable state is high. A device width that limits the current to 250μA per branch was chosen, and this still yields fast response times. Only one side of the cross-coupled inverter pair needs to be used, but an output inverter is placed on both to present each side with approximately equal loading. The output inverters are designed to have a high trigger voltage so that their devices do not turn on and conduct current when the cross-coupled structure is in the metastable state. This prevents disturbances on the sense amplifier nodes *N1* and *N2*.

4.4 SIMULATION

The complete sense amplifier system was simulated. For this simulation the *RW*-driver circuit was used to generate the required voltage for reading the cell and generating the reference current. The aspects simulated were whether a "one" and a "zero" could be sensed reliably, as well as how fast the sensing can be accomplished. There are two aspects to this, namely the delay from the *Read* signal until the differential current appears at the sense amplifier. Only at this point may the equalisation signal (*Sense* and its inverse *NSense*) be deactivated. The second specification is the time taken from this deactivation until the valid data appears on the output node.

The sense amplifier is also a cross-coupled latch, and therefore has to be initialised at the start of the simulation. It is initialised so that the data output is "zero". First a cell that is in the "one" state, that means no read current will flow, is read, and then a cell in the "zero" state, where a read current does flow. A simulation is set up where a read cycle is characterised by the *Read* and *Sense* signals being active for 5ns. The *Sense* signal is then deactivated and the *Read* is kept "high" for another 5ns. The sensing is performed in the time ranges 10ns-20ns and 30ns-40ns.

Figure 4.14 shows the output of the current sense amplifier and the internal node voltages of the sensing latch *N1* and *N2*. The differential current carried by the current conveyor (the drain current of *M3* and *M6* in Figure 4.11) and the current flowing through the bit line clamp device (*M14* in Figure 4.13) are shown in Figure 4.15.

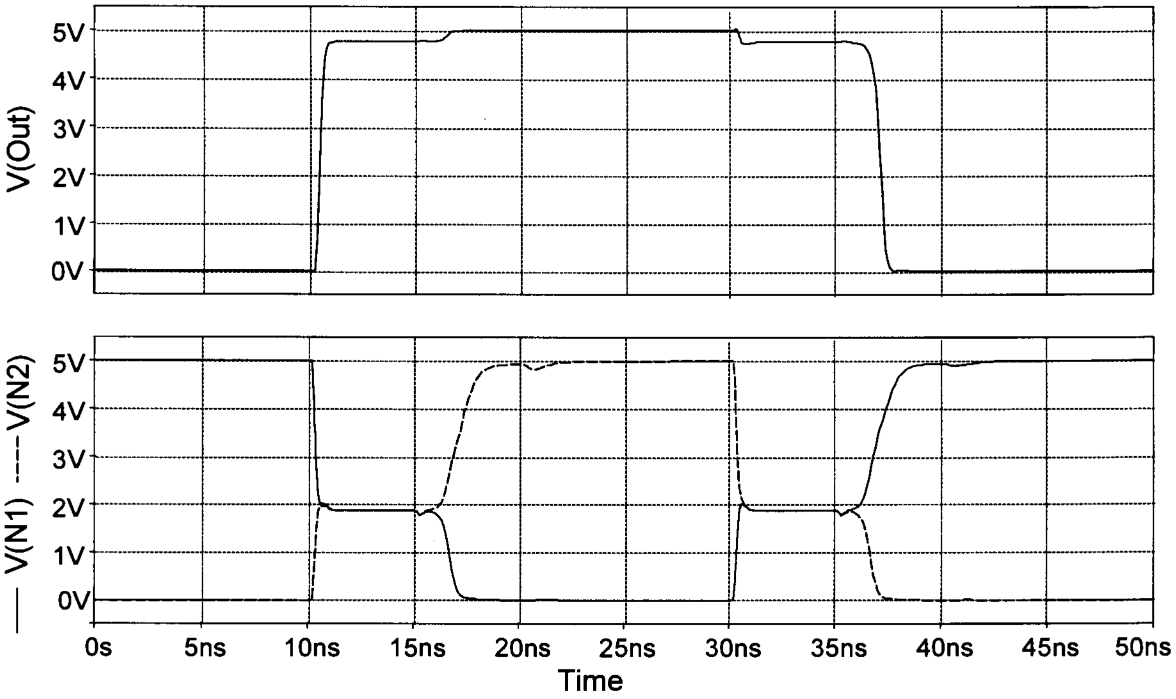


Figure 4.14 Simulated output voltage and internal voltages of the sense amplifier. The metastable condition can clearly be seen.

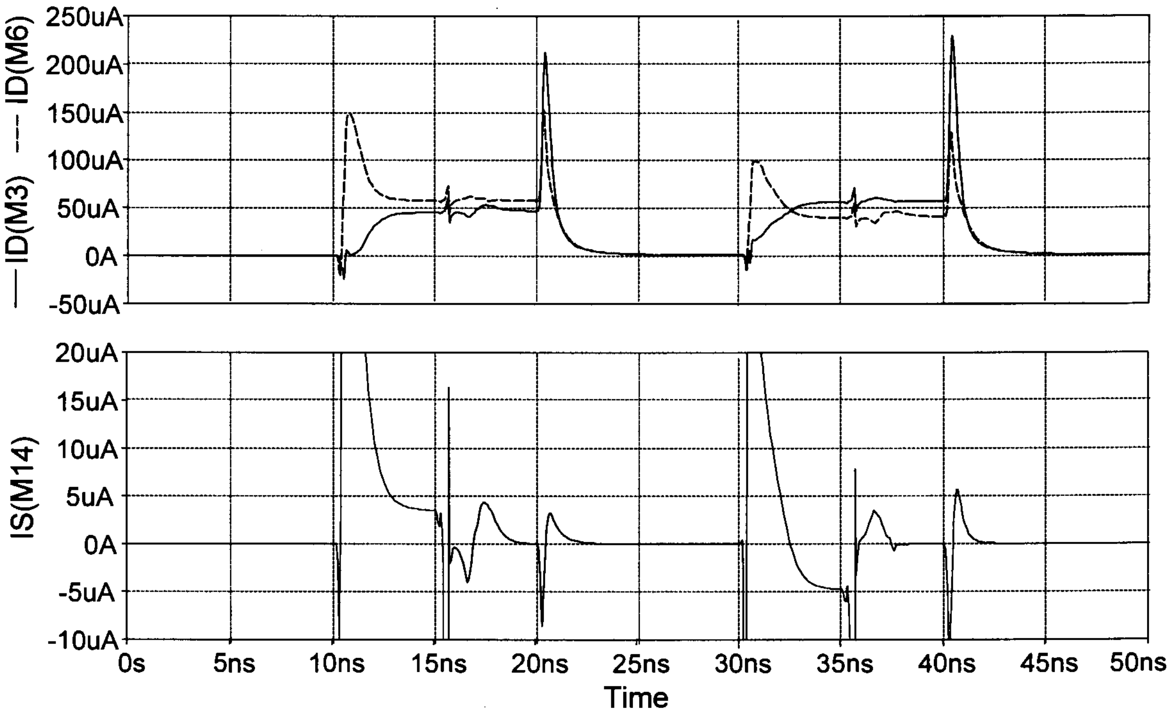


Figure 4.15 Simulated current conveyor currents and bit line clamp device current. The device indices refer to Figures 4.11 (top) and 4.13 (bottom).

The output waveform can clearly be seen to sense a "one" at 12ns and then a "zero" at 37ns. The forced metastable condition of the cross-coupled sensing latch, where the internal voltages are equal, is evident when sensing starts. Once the differential current is established (shown in Figure 4.15) the latch is released from the metastable state and assumes the correct value. The currents carried by the current conveyor are shown. The average value, when flowing, is equal to 50μA, the bias value. The differential current is the difference between the two. The bit line clamp device, *M14*, can be seen to carry some of this differential current.

This simulation yields correct data output values for all process conditions. The best, typical and worst simulated delays are given in Table 4.1.

Table 4.1 Best, typical and worst time specifications for the complete current sense amplifier system.

Specification type	*Read* to valid differential current (ns)	*Sense* release to valid output (ns)	"Zero" sense time (ns)	"One" sense time (ns)
Minimum	1.77	1.58	3.35	0.478
Typical	2.57	2.15	4.72	0.607
Maximum	4.15	3.30	7.45	0.836

The first two specifications are only given for the case where a "zero" is being sensed, because they are zero for the case where a "one" is sensed. If a one is being sensed the differential current will be positive all the time, because no current is being delivered from the cell. The latter takes longer to arrive at the current conveyor and this means a delay is present before the negative differential current exists, when a "zero" is being sensed. As soon as the latch enters a metastable state the output value goes "high", due to the way the output inverter was designed. When the *Sense* node is released, the value is already "high" and the delay is zero. The time to sense a "zero" is therefore the sum of the two delays making up the sensing process. The time to sense a "one" is only the time for the output to go "high" once the sensing has been initiated, and is therefore small. The times were simulated using a load capacitance of 100fF.

One aspect that needs to be investigated is how the sensing delay is modified if the capacitance of the input node to the current conveyor changes. According to previous published works, [26] [27] [28] [29], sensing delay is invariant as this capacitance changes. Figure 4.16 shows the delay as a function of the capacitance. It can be seen that an increase in the delay does occur. This is due to the fact that the voltage of the input nodes changes when not being used. The fastest turn on times for the current conveyor were achieved if the pull-downs are disconnected, but this lets the input nodes float, and they assume higher voltages than during normal operation. At larger capacitance, this high voltage takes longer to discharge.

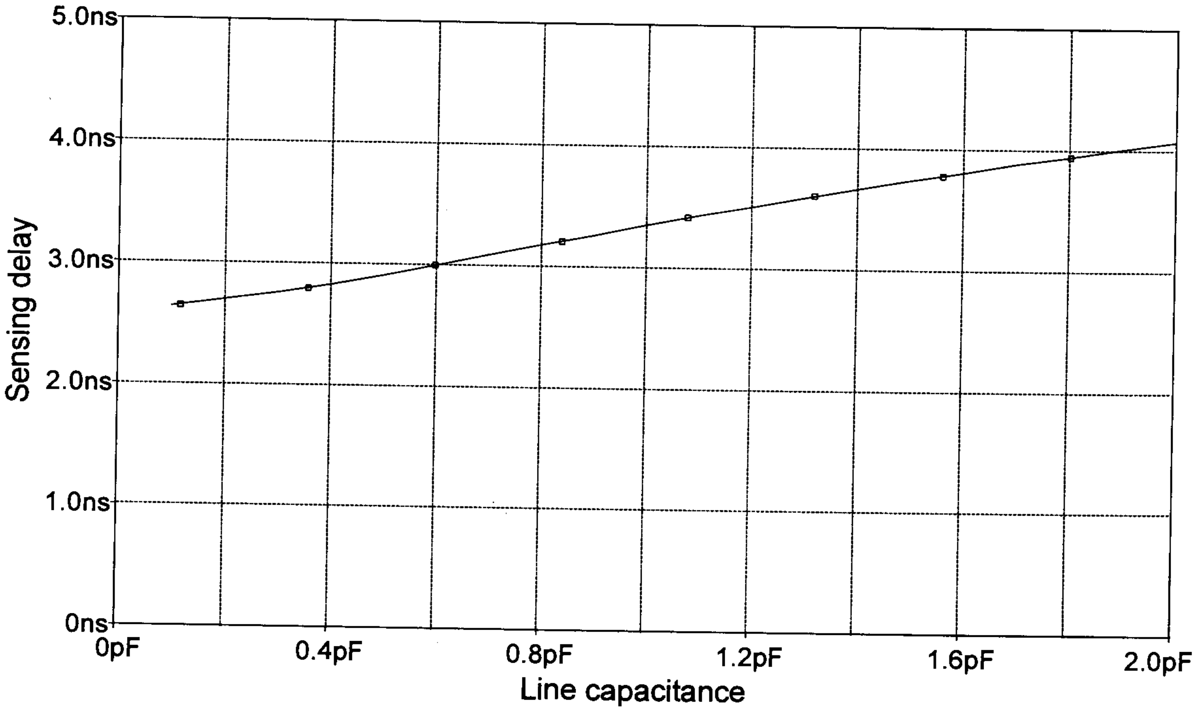


Figure 4.16 Read to valid differential current as a function of the capacitance of the input lines to the current conveyor.

4.5 LAYOUT

The layout of the current sense amplifier is shown in Figure 4.17. The reference current generator is part of the layout of the *D*/*O*-line driver switching circuit. This is shown in Figure 3.33 which also has the modifications to the switch circuit incorporated.

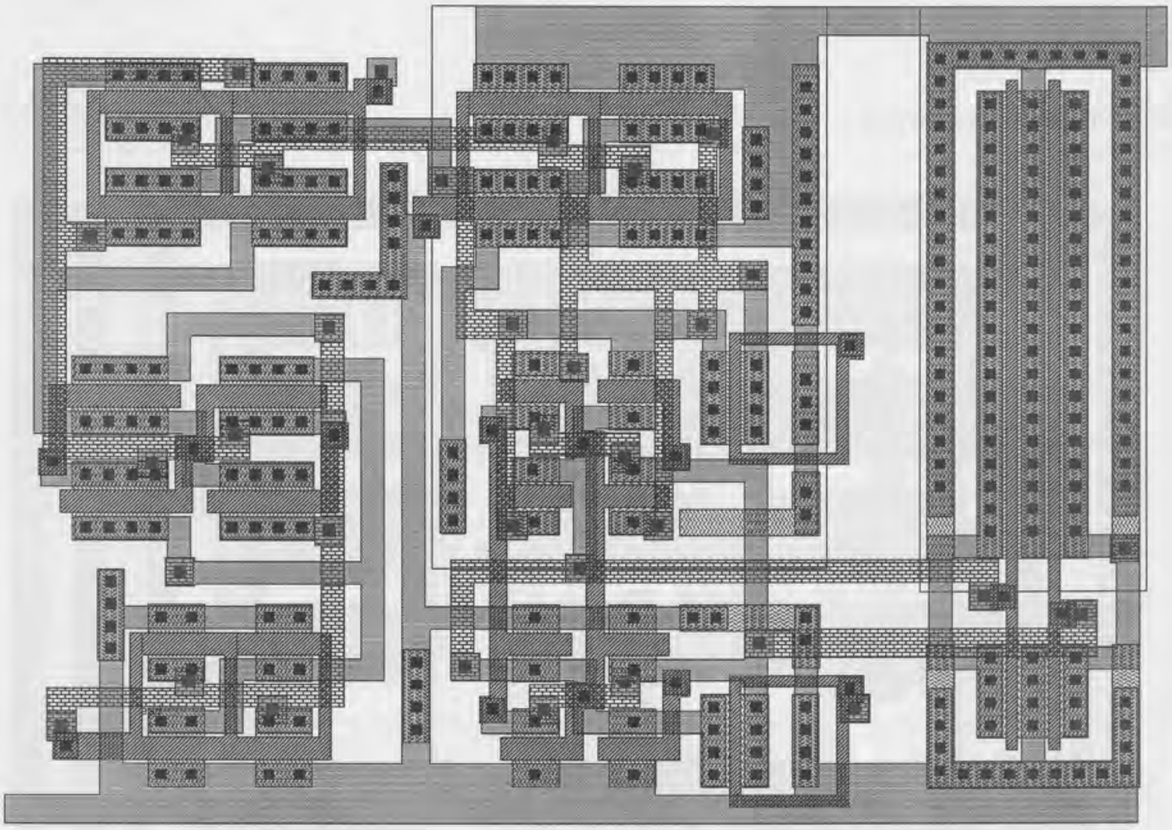


Figure 4.17 Layout of a single current conveyor and clamped bit line current sense amplifier. The devices on the right are the noisy and thus heavily guarded output inverters.

The sense amplifier circuit is very sensitive to noise. This is due to the fact that it is placed into a metastable state and gently nudged in a certain direction. If strong noise sources or other disturbances exist, they may have a dominant effect on the cross-coupled latch and force it into the wrong state. Adequate guarding of the sensitive structures, as well as the noisy circuits, is therefore required. The output inverters are well isolated from the sensitive cross-coupled inverter pair. Where possible, a separate well was used to isolate the devices. Common centroid layout is used to ensure better matching between devices that need to be identical. This is quite important to ensure correct operation of the sense amplifier.

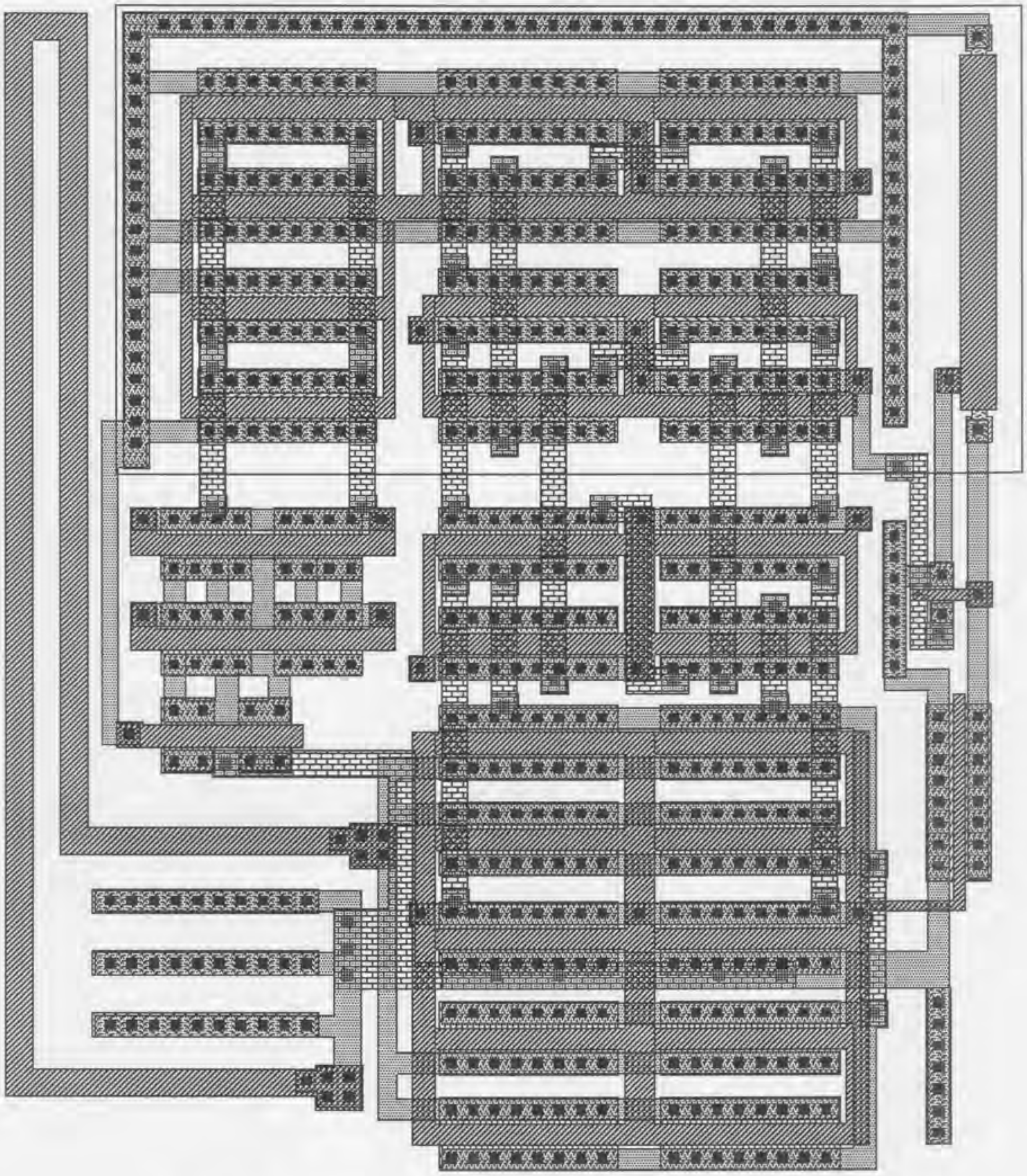


Figure 4.18 Layout of the bias network of the current conveyor.

Figure 4.18 shows the layout of the bias circuit. The devices responsible for generating the bias voltage (*M3*, *M6* and *M7* in Figure 4.12) need to have performance characteristics that are as identical as possible to those of the current conveyor which have been split into two parallel devices of half the width in order to create the common centroid geometry. The equivalent devices in the bias network (middle left in Figure 4.18) have therefore also been split the same way.

4.6 CONCLUSION

The design of the current sense amplifier was discussed in this chapter. Various techniques in the field of current sensing have been integrated into this design. The four-transistor SRAM cell is sensed by comparing the current out of the cell to a reference current generated under identical conditions. The comparison is carried out via a current conveyor circuit coupled to a clamped bit line current sense amplifier. The current conveyor was used to isolate the sensing circuit from the SRAM cells, yet maintain the differential current. The use of positive feedback in the sense amplifier helped speed up sensing and also aided in obtaining the required sensitivity to sense small differential currents. The circuit is however used to sense differential currents an order of magnitude larger than what it is capable of sensing. This makes accurate control and isolating the circuit from noise sources less critical, but still necessary.

The interface circuit to the SRAM cells is compact. This is important because it is repeated often, therefore allowing area to be saved.

The sense amplifier circuit is operational across all process conditions and the simulated delays are not too dependent on the capacitance associated with the input nodes to the sensing system. This is advantageous because these nodes connect the equivalent bits of all four memory banks. The capacitance associated with them can vary from bit to bit and can be large because of the long distances the node is routed. Sensing delays are however dependent on process conditions and typically double as the process changes from best speed to worst speed conditions. Given that a large portion of this delay is associated with the charging and discharging of capacitances, the increase as device strength weakens, is inevitable.

5. SRAM SYSTEM BASED ON THE FOUR-TRANSISTOR CELL

5.1 INTRODUCTION

The four-transistor SRAM cell, as well as all circuits required for a purely digital interface to an array of cells, have been discussed. The source node driver circuits apply the correct control voltages in response to digital input signals, and the current sense amplifiers sense the output currents and convert them to digital voltages. The following step is to integrate all the components to form a complete system. The circuits designed up to now require numerous control signals that need to be generated in a predetermined sequence. The write cycle for example is composed of two sub-cycles that need to be sequenced correctly in response to one external input. This requires several control, buffer and other digital interface circuits.

To begin with, the general global characteristics of an SRAM system interface are listed. An overview of the complete system is given and explained. Some important sub-circuits are subsequently discussed in greater detail. Simulation results, as well as specifications (timing, power, area) of the system, are given. The chapter concludes with a comparison between the four-transistor SRAM cell system and a similar six-transistor SRAM cell system.

Throughout this chapter most circuits will be displayed on gate level for the sake of clarity. The exact circuit structure and device sizing are not considered to be important for understanding the operation of the circuits. Complete circuit diagrams on transistor level may be found in addendum C. Where it is of importance, the drive strength of a gate relative to an inverter with transistor widths 2μm and 5μm for the NMOS and PMOS respectively, is shown by means of a factor inside the symbol of the gate.

5.2 SRAM INTERFACE AND CYCLES

5.2.1 SRAM Control Signals [32]

Apart from the obvious address and data inputs, three other control signals are required to create a standard SRAM interface:

a. The write enable (*WE*) signal determines whether data is written to a selected cell or read from it.

b. The chip enable (*CE*) signal facilitates selection and activation of the SRAM system. It is also the signal used to select a single chip if a large memory is created by combining several chips. This allows the data and address lines to be shared. Timing of the SRAM can be derived from the edges of the chip enable signal, in which case it initiates the read and write cycles.

c. The output enable (*OE*) signal removes the output drivers from high-impedance mode. This is an extra signal required to ensure that a selected chip can only write the data held in its output latches on the external data bus when instructed to do so.

5.2.2 SRAM Read Operation [32]

Referring to Figure 5.1, which depicts the typical SRAM read operation, the following steps are required in sequence to read a word from the SRAM:

a. The address of the word to be read is placed on the address bus of the memory.

b. The *WE* signal is deactivated to signal to the SRAM system the word has to be read.

c. The *CE* signal is activated to start the read operation. The completion of the read cycle can be indicated by the RAM system or can be set by a maximum timing specification. After completion, the *CE* must be deactivated and the next memory cycle can begin.

d. The *OE* signal may be activated at the start, during or upon completion of the read operation. Typically the last option could be used to prevent unnecessary voltage transitions on the data bus. Considering that the data read from the array is latched, *OE* may be activated at any time before the next read operation.

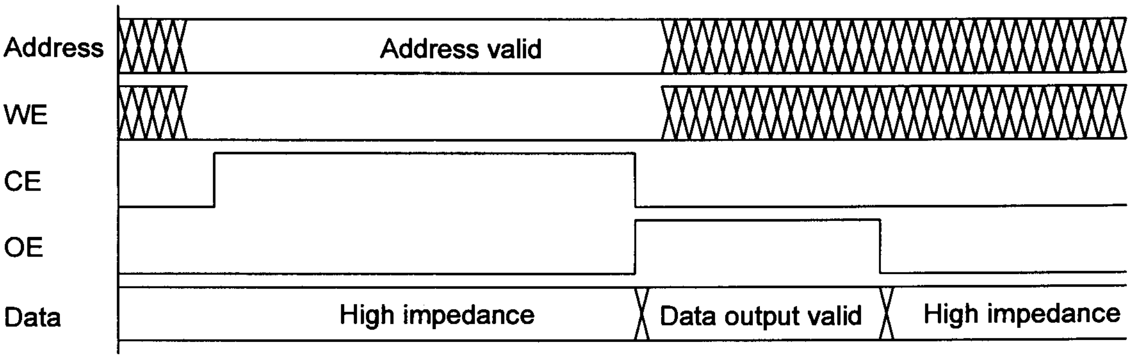


Figure 5.1 SRAM read cycle.

5.2.3 SRAM Write Operation [32]

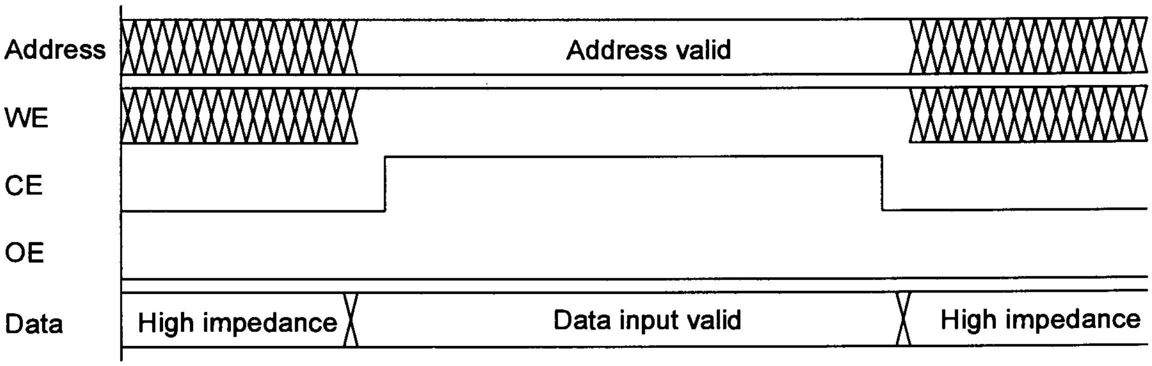


Figure 5.2 SRAM write cycle.

The typical write operation is shown in Figure 5.2. To perform the write operation the following sequence of external signals should be used:

a. The address of the word to be written, as well as the data to be written to that word, are placed on the address and data buses respectively.

b. The *WE* signal is driven "high" to indicate to the SRAM system that a write operation is to follow.

c. The *CE* signal is then activated to indicate the start of the write operation, and is deactivated again after completion. Once again the completion may be signalled by the SRAM system or be derived from maximum timing specifications.

d. The *OE* signal is kept "low" so that the output drivers are deactivated to prevent bus contention.

5.2.4 SRAM Timing [33]

The timing of the SRAM system may be accomplished asynchronously or synchronously. In the latter case, the control signals are applied and the operations of the system are carried out as a sequence of events timed to the edges of one or more clocks. The speed is determined by the number of clock cycles required for an operation and the clock period. An asynchronous RAM derives all timing from the edges of the control signals. Typically the address lines are used, but this requires the use of address change detection. The simpler method to use, is the edge of the *CE* signal. The completion of the cycle is assumed after a specified time has elapsed, or the control circuit of the RAM may supply a signal indicating completion. The latter allows the RAM to be used constantly at maximum cycle speed.

For the four-transistor SRAM system design, it was decided to use asynchronous timing derived from the edge of the *CE* signal with an output signal indicating completion. The main reason behind this choice is the multiple control steps with vastly different cycle lengths. Examples are the write cycle which is a combination of a clear and a write, or the read cycle where the current sense amplifiers need to be forced into a metastable state and then released once the differential current is applied to their inputs.

Several timing specifications exist that can be used to evaluate the performance of a specific SRAM:

- Read access time: This is the main specification and usually refers to the time difference between the read cycle initiation (*CE* activation in this case) and valid output data, assuming that the *OE* signal is active. This specification is load dependent.

- Read cycle time: The previous specification does not include the time difference required between valid address input and the activation of the read cycle. Adding this time to the read access time results in the minimum time between successive read cycles, the read cycle time.

- Write access time: This is the time difference between the write cycle initiation signal (*CE* activation in this case) and the completion of the write cycle.

- Write cycle time: The time difference required between the valid address and data and the write cycle initiation signal is added to the write access time to obtain the write cycle time. This is the minimum time between successive writes.

- Cycle time: The maximum of the read cycle time and the write cycle time.

5.3 THE SRAM SYSTEM

The discussion of the SRAM system is divided into two sections. First the top level is discussed, where the interaction between the four memory banks and the common peripherals is described. This is followed by considering the operation of one memory bank.

5.3.1 Global System

A block diagram of the complete memory system showing all top level building blocks and the interconnections between them, is depicted in Figure 5.3.

The system contains 32768 bits of memory grouped into 1024 words of 32 bits each, and further split up into four banks of 256 words each. This has already been discussed and motivated as a method of reducing the bit line capacitance and the wasted write currents. The result is increased speed and lower power dissipation.

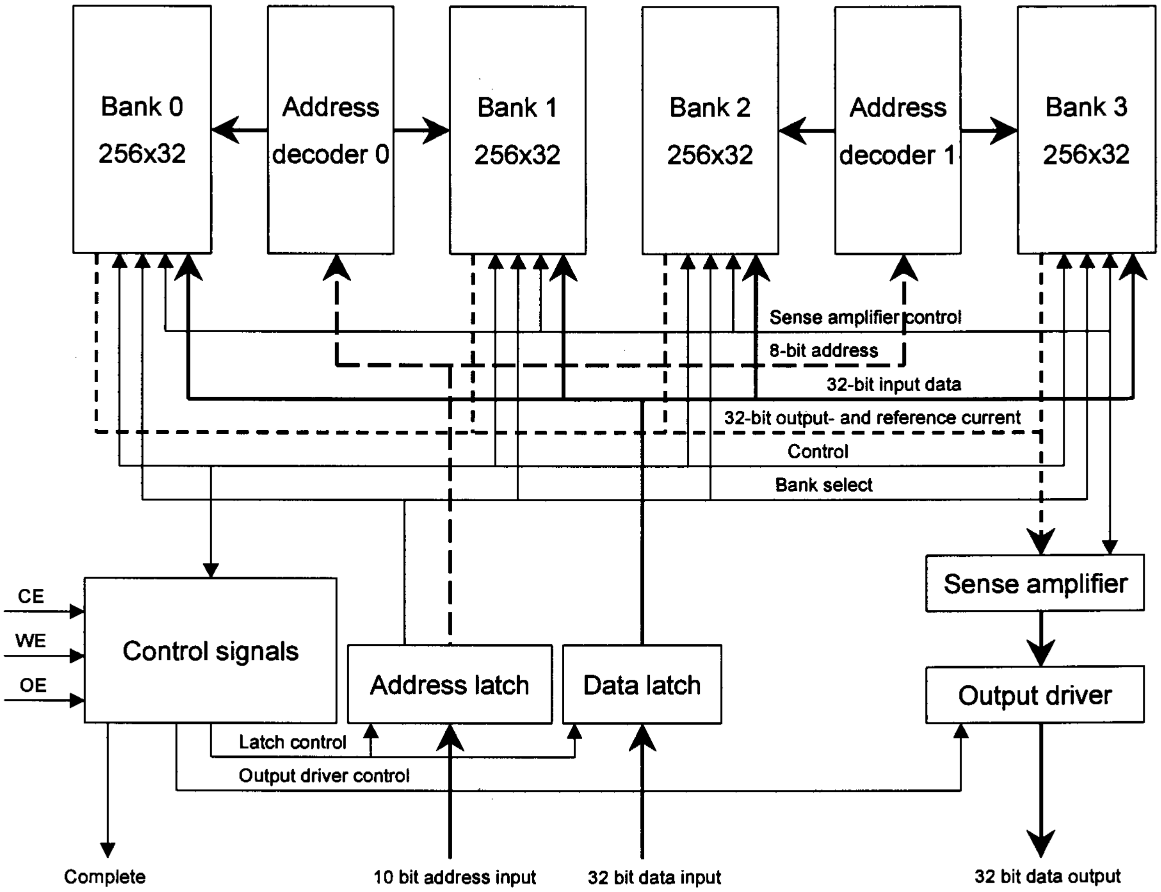


Figure 5.3 Top level block diagram of the SRAM system.

There are 1024 addressable words which implies a 10 bit wide address bus, and each bank contains 256 words meaning the address decoder is an 8-bit input to 256 line output decoder. The two most significant bits of the address serve as bank selection bits. From the block diagram of Figure 5.3 it is clear that two signals emerge from the address latch. One is the 8-bit wide address that points to a word within each bank. The other signal is the 2-bit bank select, which selects one of the four banks, thereby selecting one of the four addressed words. Two banks share one address decoder. Although this seems to be wasteful, it means that only a 16-bit wide address bus, the true and complement of each address bit, needs to be routed. This compares to routing 256 decoded address lines if a single address decoder is used. The area required for implementing two complete decoders is far less than the area required to route the decoded address to all words in each memory bank. Each output of the address decoder is implemented via an 8-input

AND-gate that is synthesised using two four-input NAND-gates feeding a two-input NOR-gate.

One of the features of the source node driver circuits discussed in Chapter 3 is the fact that the large currents they require to operate can all be turned off when the circuits are not needed. The bank select scheme also ensures that only those drivers in the currently addressed bank are turned on when needed, thereby saving power.

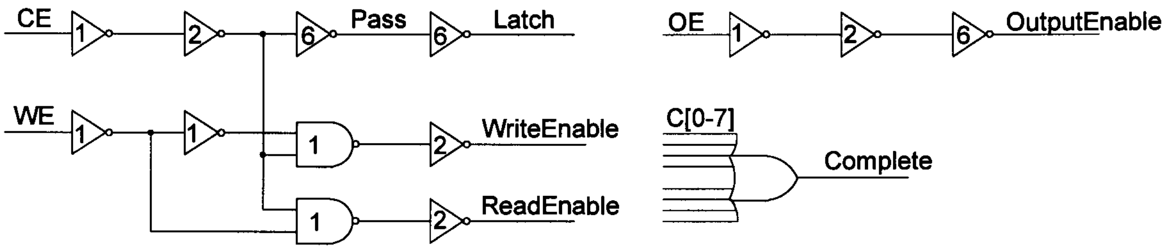


Figure 5.4 Circuit diagram of the control signals block in Figure 5.3.

The address input as well as the data input are guarded with transparent latches. Once the *CE* control signal is activated, these latches are locked, and hold their value until the *CE* is disabled again. This is done for two reasons:

a. While the SRAM is busy with a write or a read cycle the data and address input may not change. The reasons for this are to prevent data in the array from being corrupted through two words being accessed in one cycle, or the data changing during a write cycle. The latches ensure that the data and address inputs cannot change after the cycle has been started until it is complete.

b. Once a cycle has been initiated, the address and data are stored within the SRAM as long as they are required. This means the address and data bus may be used for addressing other peripheral components.

The latches are based on two inverters that may be placed in either a pass or a feedback configuration by means of two transmission gates. The control is achieved by using the *CE* signal and its inverse as is shown in Figure 5.4. The

Latch signal places the latches into feedback mode while the *Pass* signal makes them transparent.

The sense amplifier drives the output data bus, which can be the same bus as the input data bus, via a tri-state output driver. This driver is controlled by the *OE* signal. As shown in Figure 5.4, the *CE* and the *WE* signal are combined to create read and write enable signals which initiate and control all actions taking place in the memory banks. There are eight completion signals, one for each action (read and write) of each memory bank. These are combined in an 8-input OR-gate and presented as a system output to indicate completion of a cycle. The *ReadEnable*, *WriteEnable* and the *Complete* signals are the global control signals passed between each bank and the global control circuit. All internal control signals required by each bank are derived from the global signals within the bank. This prevents skew between control signals that is usually a result of travelling long distances and having different associated load capacitances.

5.3.2 Sense Amplifier

The sense amplifier is a common module. The current outputs of each bank are connected together onto the sense amplifier input. Only one of the four can be activated at any time. This achieves current-mode multiplexing and allows a single sense amplifier to be used for sensing all four banks. Figure 5.5 is a diagram of the sense amplifier circuit.

It accepts the 32 cell current inputs and 32 reference currents. Each bank generates its own reference currents to minimise skew between the cell current and reference current arrival at the sense amplifier. This also allows the reference current to be generated using the same *RW*-line voltage as the cell read current. The sense amplifier receives no control from the global control unit, but is controlled by the currently active bank. This is to ensure that the signal to release the cross-coupled latches from the metastable state is in sync with the signals that supply the currents. The memory banks each supply the *Equalise* signal, as well as the *NRead* signal. The first is required to force the metastable condition and clamp the bit lines, whereas the second is an active low indication that a read cycle is being executed. This signal turns on the loads for the current conveyor, and thereby allows bias currents to flow. The signals from each bank are active low and the NAND configuration therefore performs a logic OR function.

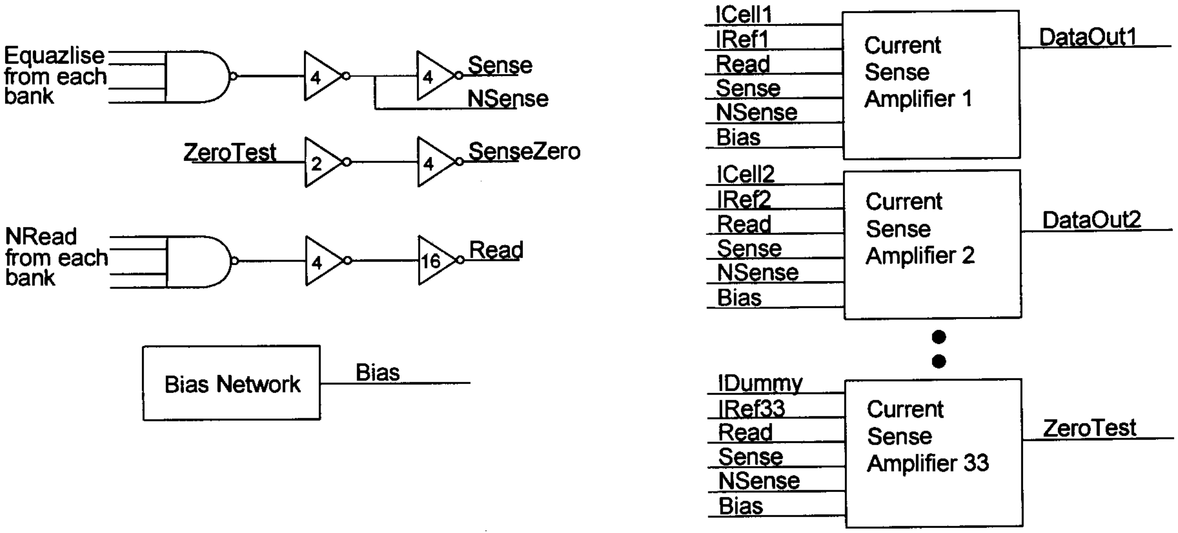


Figure 5.5 Complete current sense amplifier system.

For timing purposes, an extra 33rd sense amplifier that is always supplied with a reference current and a dummy cell current, *IDummy*, is incorporated into the sense amplifier. This specific circuit will always sense a "zero" and is used to indicate to the read control circuit of the currently selected bank that the sensing procedure is complete. The output inverter of this sense amplifier returns a logic "high" when the latch is in the metastable state. Once the "zero" is sensed, this output will therefore change and indicate completion of the sensing via the *SenseZero* signal, which is routed to the read control circuitry of every bank.

5.3.3 Memory Bank

The block diagram of one memory bank is shown in Figure 5.6.

The following list mentions relevant aspects of blocks discussed elsewhere in this document and therefore require no further explanation:

- the array of 256 words with 32 bits each,

- the *RW*-line driver system comprising the switching circuits, low-impedance driver, op-amp and bias network that drives the *RW*-nodes of all words,

- the *CL*-line driver to clear a row of cells,

- the four *DIO*-line driver systems, each comprising the switch circuits, op-amp, low-impedance driver and the bias network, that program the data into the words and also contain the access mechanism to guide the read currents to the sense amplifiers,

- the reference current generator which generates 33 reference currents and a dummy cell current for timing purposes.

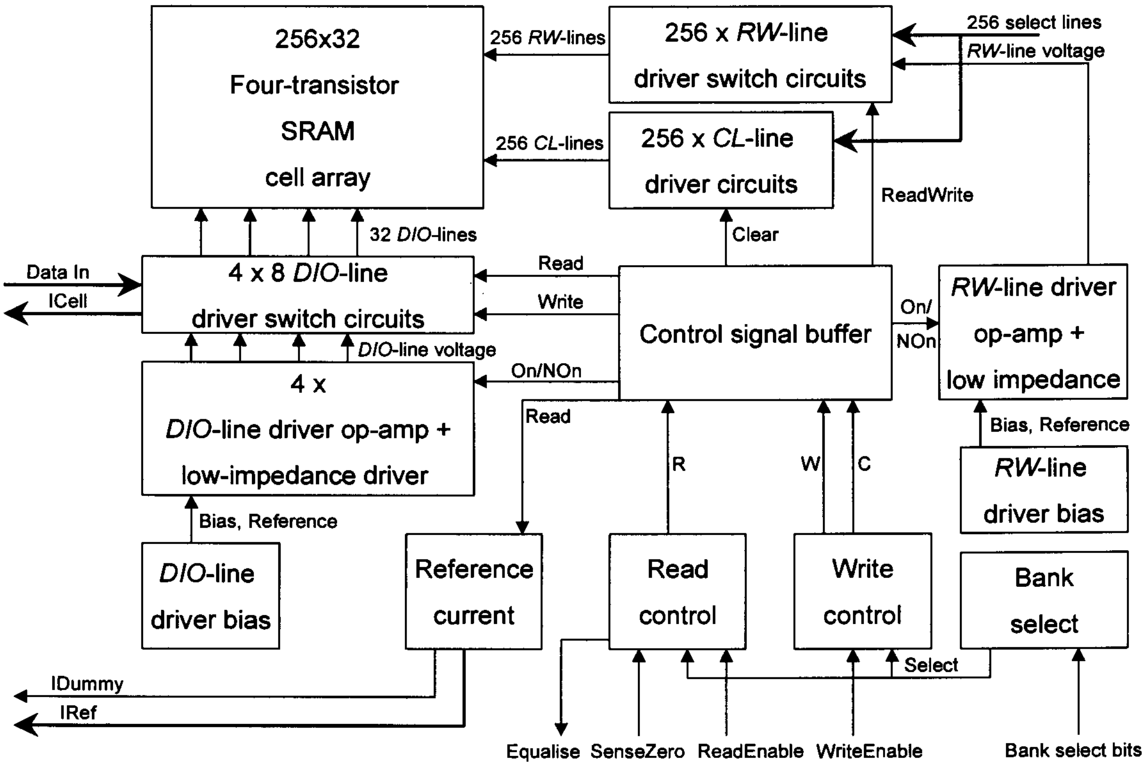


Figure 5.6 Block diagram of a single memory bank.

What therefore remains to be discussed are only those blocks dedicated to control.

- The bank select block decodes the two most significant address lines to decide if the specific memory bank is being addressed or not. If the bank is not being addressed, all functionality is deactivated by masking the *ReadEnable* and *WriteEnable* signals. This prevents any memory cycles from being initiated in the specific bank.

- The control signal buffer system accepts three inputs from the control logic and creates buffered versions of these for the various peripheral devices. These include the *On* and *NOn* signals for the low-impedance drivers and the *Read*, *Write*, *ReadWrite* and *Clear* signals (and their inverses where required) for the reference current generator and driver circuits. The three inputs are signals indicating the various cycles of the memory, the read (*R*), write (*W*) and clear (*C*). The operation of the control circuits is explained in the following two sections.

5.3.4 Read Control

The read control circuitry is a circuit to sequence a predefined chain of events. Consider that the required address has been decoded to one of the 256 *Select* lines, and that a specific memory bank is selected via the bank select circuitry. The read cycle is initiated on a falling edge of the masked version of the *ReadEnable* signal. The *R* signal has to be activated to turn on the *RW*-line driver, as well as the reference current generator. At the same time, the *Equalise* signal has to be activated to force the current sense amplifier latch into a metastable state to start the sensing operation. As soon as sufficient time has elapsed for the differential current to be present at the current sense amplifier, the *Equalise* signal has to be deactivated again. This time is measured by the circuit given in Figure 5.7. The sense amplifier then senses the currents and the *SenseZero* signal will be activated. This indicates the end of the read cycle, and the *Read* signal is deactivated. To inform the system containing the memory of completion of the read cycle, the *Complete* signal is activated.

Equalisation Cycle Timing

The circuit of Figure 5.7 is used to time the deactivation of the *Equalise* signal. It operates on the principle that the fast current-mode signals are present at the sense amplifier shortly after the *RW*-line voltage is sufficiently low to allow adequate currents to flow. It is this condition that is tested by the circuit. The *RW*-line voltage is applied to the gate of *M2* and, in an identical fashion to the reference current generator, will cause a current to flow. The current passes through the diode-connected load device *M1* and causes a voltage drop, *Vx*, to occur. This voltage drop triggers the inverter chain to pull its output node low. This happens at a low voltage drop across the load device because the trigger voltage of the first inverter has been designed to be close to ground. The second and third inverters have weak devices to add delay, so that the differential current signal may establish properly at the sense amplifier input nodes, before the clamping devices are turned off. The device *M3* is present to pull down node *Vx* when the circuit is not being used and is hence activated by the active low read strobe. This feature is required because the load device cannot pull *Vx* lower than its threshold voltage.

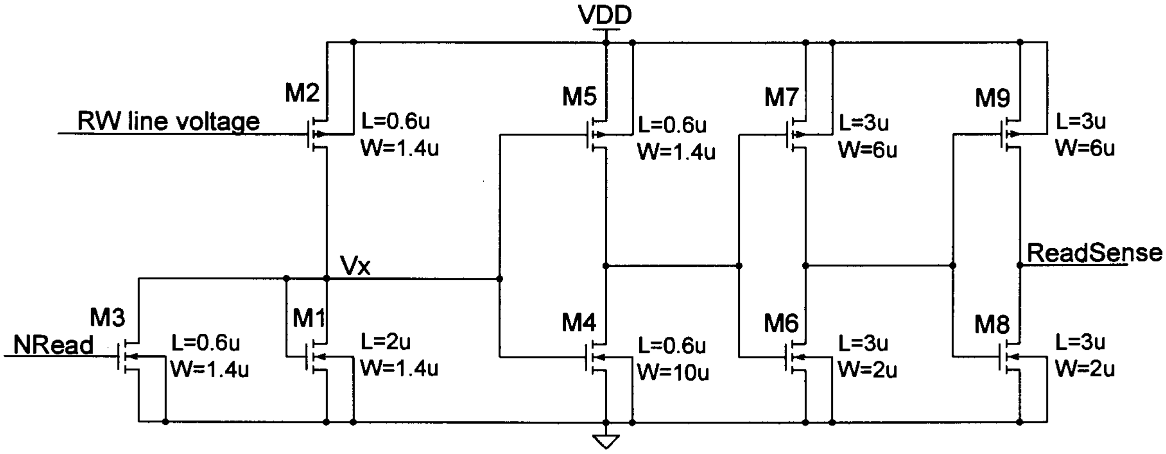


Figure 5.7 Circuit to sense completion of the equalisation cycle

Timing Circuit Structure

The timing control circuit contains a latch for each control signal that is required. The latches may be set or reset depending on certain events. These events are typically indicated by a falling edge on a particular signal. In order to use that signal to set or clear a latch, the edge needs to be converted to a pulse. This has to be done because the condition that set the latch may still be valid at the time the latch has to be reset. To reset the latch however, the signal that set it may no longer be present. This problem is solved by converting the edge to a pulse that has expired by the time the reset pulse is applied to the latch.

Edge to Pulse Converter

The circuit shown in Figure 5.8 can be used to convert a falling edge to a positive pulse, and consists of a NOR gate and three inverters. If the signal indicating a particular event has a rising edge, an inverter can be placed before the edge to pulse converter circuit.

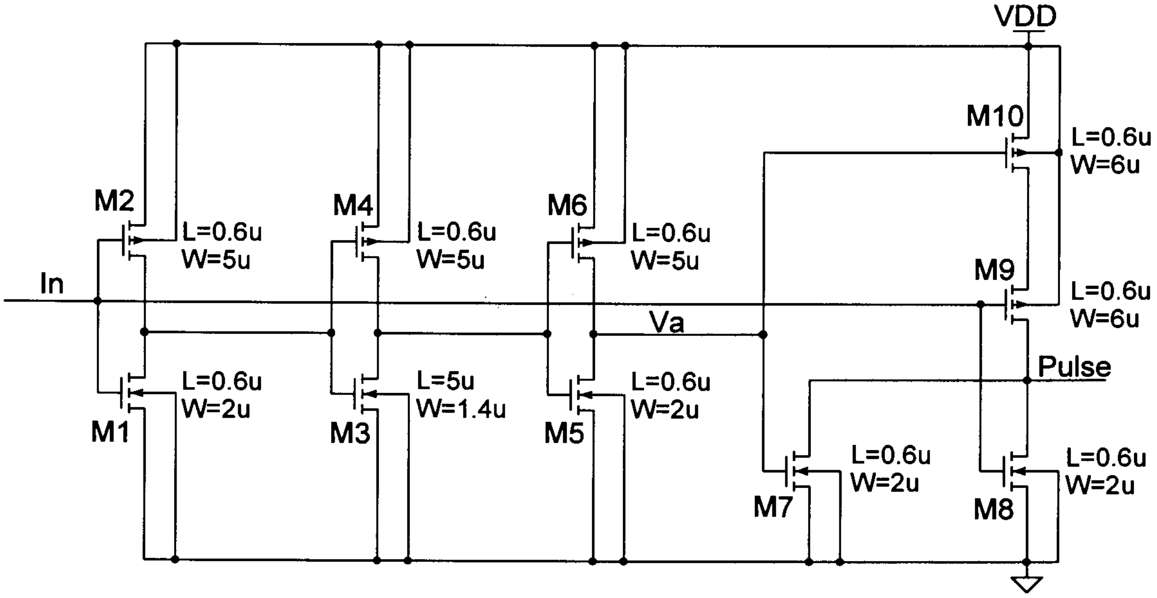


Figure 5.8 Falling edge to positive pulse converter.

From the device scaling it is evident that a falling edge on the input to the string of inverters will be transmitted slowly. This introduces delay between the falling edge of the *In* input to the NOR gate and the rising edge of the *Va* input, and creates the condition where both inputs are "low" for a short time. In this time the output of the gate will be "high". The opposite path through the inverter chain has normally sized devices so the reset of the system is fast and it may be used to respond correctly to a falling *In*-edge almost immediately after a rising *In*-edge. The width of the pulse is defined by the delay through the inverter chain. The circuit of Figure 5.8 returns a pulse width of 1.4ns. This is sufficient to set or reset an SR-latch. The simulation of Figure 5.9 also indicates that the voltage at node *Va* is slow to rise but falls quickly as the input signal rises, thereby rapidly resetting the converter.

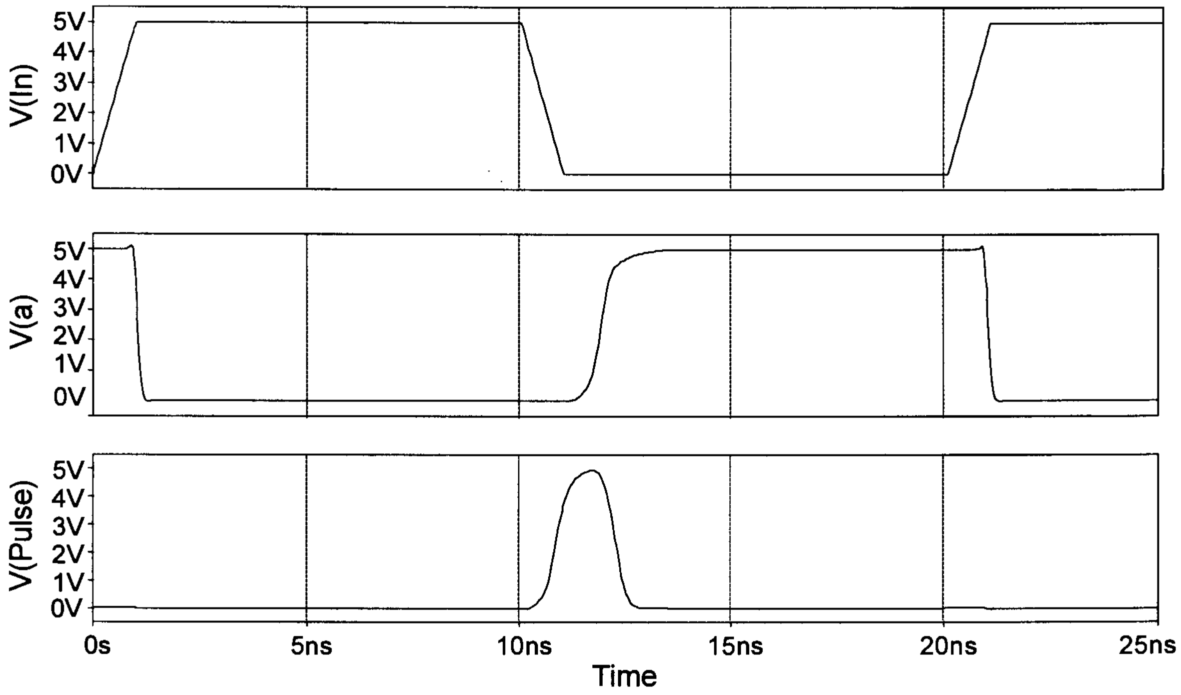


Figure 5.9 Simulation of the falling edge to pulse converter.

Read Control Circuit

The read control circuitry is given in Figure 5.10. It is based on the principle that a latch is set and reset by different events. The latches used are implemented using a cross-coupled NOR-gate structure. The set and reset signals are therefore active high. Some latches must be reset by two separate events, requiring two reset signals, and consequently one of the NOR-gates in these latches is a three-input gate.

The read cycle is initialised by a falling edge on the masked read enable signal, *RDEn*. The falling edge sets the *R* and *Equalise* signals by means of the edge to pulse converter. The circuit remains in this state until the falling edge on the *ReadSense* signal indicates stable currents, and resets the *Equalise* signal, allowing the sense amplifiers to sense the differential current and latch the digital voltage level. The sense amplifier then returns a falling edge on the *SenseZero* line as explained previously. In response to this, the *R* line is deactivated and the *Complete* signal is set. The latter is deactivated upon deactivation of the *RDEn* signal. The cycle may also be interrupted at any time by deactivating the *RDEn* signal, and all the latches are reset.

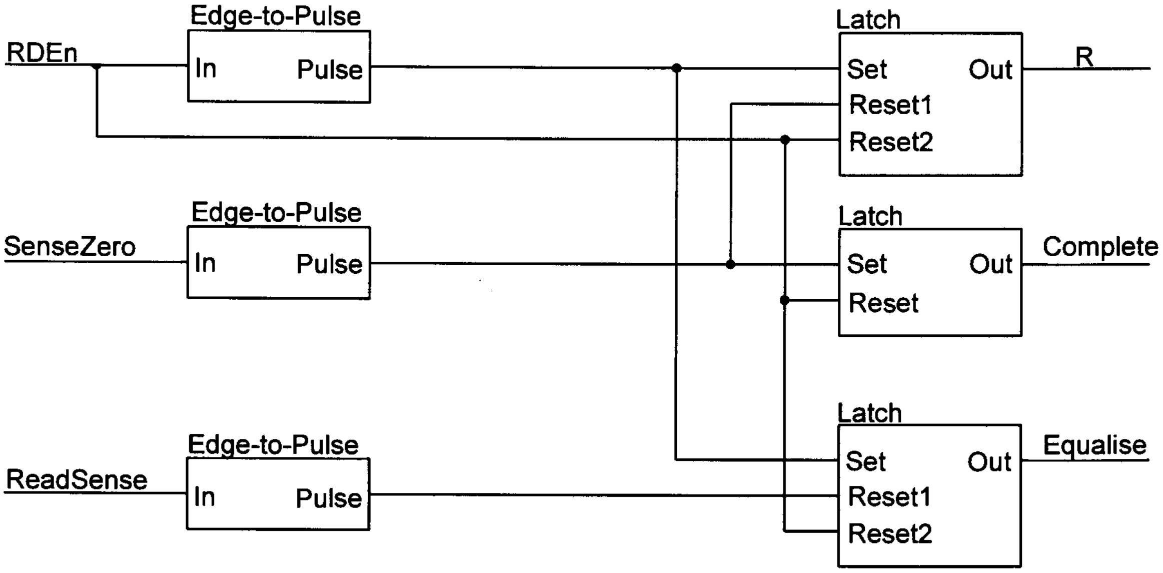


Figure 5.10 Read control and timing generator circuit.

5.3.5 Write Control

A structure based on the same principles as that for the read control circuitry is used for the write control. Once again consider that a specific word is being addressed, a specific bank selected and the data to be written to that word present at the input to the *DIO*-line driver. The first step is to clear the addressed word by activating the *C* signal. Once all bits in the word are cleared, the *C* signal must be deactivated and the *W* signal set. This activates the *RW*-line driver as well as the *DIO*-line driver, in order to write the specified data to the selected word. Upon completion of this step the *W* signal must be deactivated and the end of the cycle indicated by activating the *Complete* signal.

Timing Cells

The control circuitry requires notification of the cell clear and cell write completion. This is generated by using four dummy SRAM cells, one associated with each

D/O-line driver circuit. Each *D/O*-line driver circuit has to be timed, because the loading, and thus the delay, may differ. The action that needs to be timed is applied to the dummy cells in the same fashion as to the array cells. The strength of the switching circuits has been adapted to fit the smaller load presented by these cells. The fact that an operation has been completed in all dummy cells is therefore indicative of the fact that it is most likely also completed in the array cells. To sense completion of an operation in a dummy cell the internal nodes are sensed via inverters, *M5-M6* and *M7-M8* in Figure 5.11.

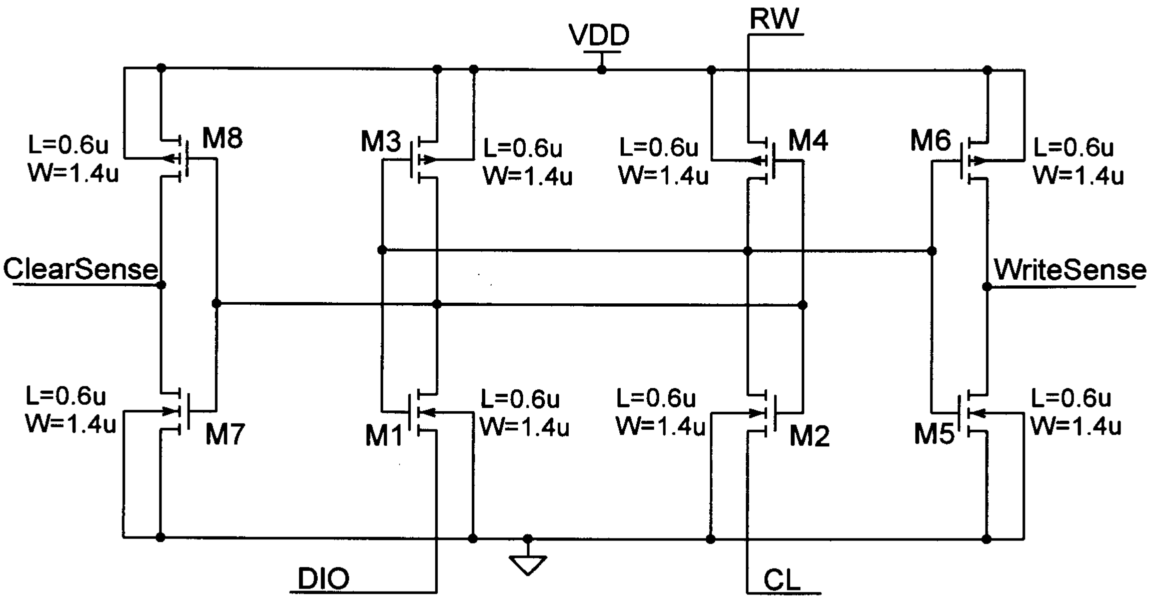


Figure 5.11 Dummy four-transistor SRAM cell with internal nodes sensed to time the write cycle.

The write cycle always consists of a cell clear followed by a cell write and each of these operations is applied to the dummy cells. They are therefore always in the correct state to time the next operation. To ensure a correct state at start-up or after a cycle has been interrupted, the *D/O*-line driver switching circuit for the dummy cells is modified to pull the *D/O*-line high when a write cycle is not currently taking place. Figure 5.12 shows this modified circuit, compared to Figure 4.8 for the array cells. With this modification the dummy cells are in a "set" state and ready to be cleared when a new write cycle starts.

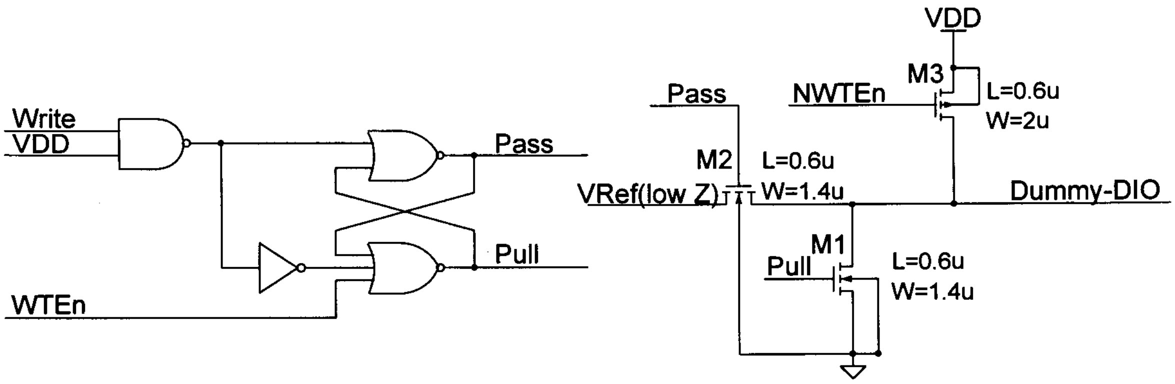


Figure 5.12 Modified *D/O*-line driver to ensure the dummy cells are in the correct initial state.

The *WTEn* is an active low version of the *WriteEnable* that has been masked by the bank selection bits and *NWTEn* is its inverse. This signal is active for the duration of the write cycle. This means that when the bank is not in a write cycle, the PMOS pull-up *M3* is turned on and pulls the *D/O*-line of the dummy cell "high", thereby setting the cell. The *WTEn* signal applied to the latch ensures that while the pull-up device is on the pull-down device *M1* is off. The scaling of the pull-down and pass devices, *M1* and *M2*, as is evident in Figure 5.12, is to adapt to the low line capacitance.

Write Control Circuit

Figure 5.13 depicts the write control circuit. When all four dummy cells are in the correct state, the active high *ClearSense* and *WriteSense* signals are combined via NAND-gates to the required falling edge.

The write cycle is initiated using a falling edge on the *WTEn* signal. This sets the *C* signal and the word is cleared. The dummy cells are also cleared and the *C* line is reset in response to the falling edge of the *ClearSense* signals on completion. The *ClearSense* pulse also sets the *W* line to start writing the cells. When all dummy cells are written, the pulse in response to the *WriteSense* signals all being true, clears the *W* line and activates the *Complete* signal. The latter is reset by deactivating the *WTEn* line. This typically happens via an external circuit in response to the *Complete* signal being activated. Each of the three latches can be cleared at any time by deactivating the *WTEn* signal, which also allows a cycle to be interrupted and the control circuit to be reset to the initial state at any time.

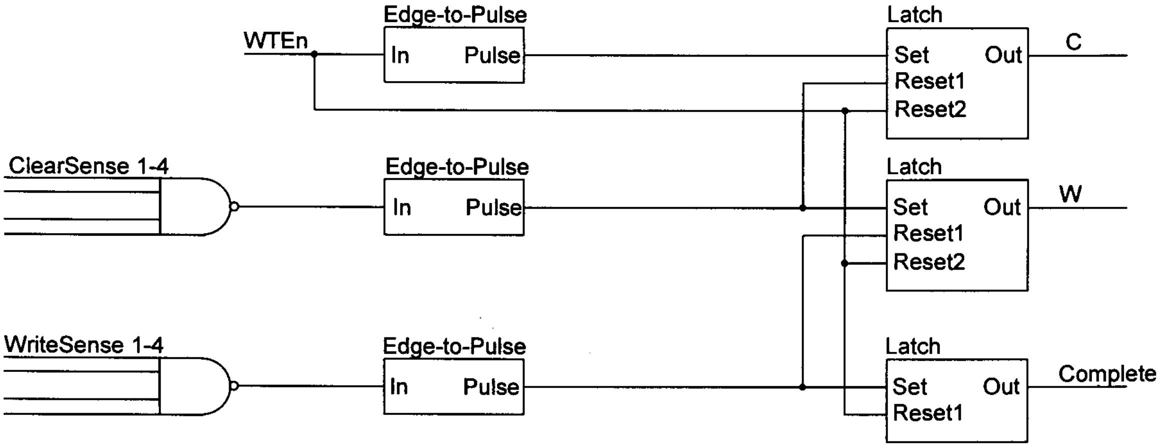


Figure 5.13 Write control signal generator circuit.

5.4 SIMULATION

A full scale transistor level simulation of the SRAM system was not possible with the available design tools. The complete system contains in excess of 170000 devices. The results obtained by performing such a simulation do not justify the effort involved. In order to obtain results within a reasonable amount of time the scale of the simulation should be limited to no more than 2000 devices. The following aspects need to be simulated, preferably across all process conditions:

- correct functional operation,
- an estimate of the power dissipation
- and an estimate of the access and cycle times.

Considering that the most important aspect to simulate is the functional operation, the first step can be to omit three memory banks. The banks are connected in parallel except for the bank select scheme, and those that have been left out can be modelled as capacitance on the shared lines. Furthermore, as far as a simulation is concerned, if it is possible to read one word without modifying one other word, all words can be read without modifying any other word in the array.

This is so because all circuits are 100% identical in a simulation. The same is valid for the write. What therefore needs to be tested is that one word can be written and read without affecting another word in the array. This allows large portions of the array to be left out, as long as the absence thereof is compensated by adding the loading effect. This gives a reliable estimate of circuit operation as well as performance characteristics.

As a second step, reducing the word length is considered as being an optimal method of reducing the size of the simulation. A large amount of redundancy in the form of parallel circuits that hardly interact and do not increase the information that can be gathered from a simulation, is removed. It was therefore decided to simulate a system of 8 by 8 cells. Eight bits remain in a word because this is the group connected to one *D*/*O*-line driver. The number of words is reduced to eight as well. All peripheral circuits remain as they are because these play an important role in the delay specification. The delays achieved for the small system are basically identical to what can be predicted for the complete system because all loading effects are modelled. As far as power dissipation is concerned, the results of the simulation of the reduced system will be extrapolated to the complete system.

The cells of the array that have been omitted, need to be added as capacitance. Each cell can contribute a high or a low capacitance depending on its state. The loading is data dependent so for any given process there is a worst case delay and worst case power dissipation, depending on data conditions. The delays are longest if the capacitance is high, because this slows down the source driver circuits. The slower circuits also cause the deviations to be applied for longer, and this results in higher power dissipation.

A simulation is run on the 8x8 array where words 0, 1 and 2 are initialised with the hexadecimal value H"FF" and words 3 to 7 with H"00". The following simulation is performed to test that words can be written and read without modifying others in the system:

- cycle 1: 0 - 20ns - word 3 is written with H"FF",

- cycle 2: 20 - 40ns - word 2 is written with H"00",
- cycle 3: 40 - 60ns - word 4 is written with H"AA",
- cycle 4: 60 - 80ns - word 1 is read, the required answer is H"FF",
- cycle 5: 80 - 100ns - word 2 is read, the required answer is H"00",
- cycle 6: 100 - 120ns - word 3 is read, the required answer is H"FF",
- cycle 7: 120 - 140ns - word 4 is read, the required answer is H"AA",
- cycle 8: 140 - 160ns - word 5 is read, the required answer is H"00".

To be able to view the simulation results in terms of digital signals, the input signals are generated digitally and the output signals are directed through a digital buffer. The analogue to digital interfaces of these circuits are set to have no loading effect on the circuit and no delay. They therefore do not influence the simulation conditions, but do make the data easier to analyse.

The output of the simulation in the digital domain is given in Figure 5.14. Identical results are obtained regardless of process conditions, but the delays differ. This proves that the circuit functions correctly, independent of process conditions. The control of the SRAM via the two signals *CE* and *WE* can be seen. The completion of a cycle is indicated by the rising edge on the *Complete* signal. For a read cycle the *Complete* signal arrives shortly after the data is latched into current sense amplifiers. During sensing the output value of each sense amplifier is "one", as was discussed in conjunction with the sense amplifier design. The global system as designed, especially the control circuits, have herewith been proven to be operational for all process and data conditions.

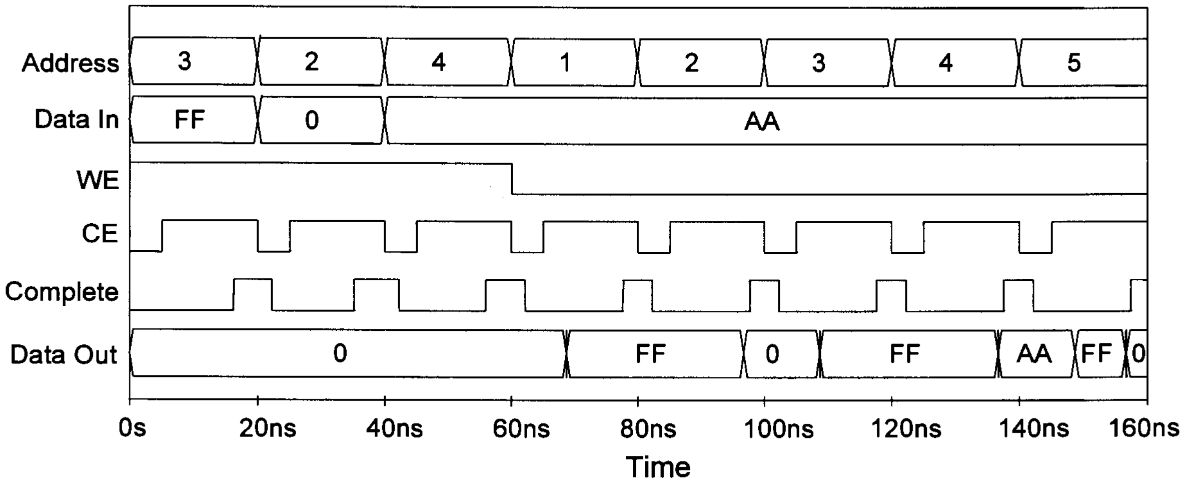


Figure 5.14 Simulation results of the 8x8 system showing correct functional operation.

5.4.1 Timing Specifications

From the simulations several specification can be derived, notably the four timing specifications mentioned in Section 5.2.4. The timing specifications shown in Table 5.1 for the zero output load condition were measured from the simulation results in the following manner:

- read access time: the time difference between the *CE* activation and the correct data value appearing on the output nodes,

- write access time: the time difference between the activation of the *CE* signal and the rising edge on the *Complete* signal,

- read cycle time: the read access time is added to the time difference between the valid data and the *Complete* signal activation and the address decoder system delay,

- write cycle time: the write access time added to the address decoder delay.

From the data in Table 5.1 it can be seen that the performance of the system is actually quite independent of process conditions with the exception of the worst case speed transistor model. The most important timing specifications are the best, typical and worst read access times which are 8.1ns, 11.7ns and 19.8ns respectively. Comparing the figures with those combined from Table 3.4 and 4.1

shows that about 60% of the access time is delay in the cell and the immediate peripheral circuits. The rest is accumulated in the various control circuits and distribution of signals. It is evident that the read and write access times are about equal. This means that having the two cycle write method does not influence performance, because the worst of the read and write cycle time is taken to be the cycle time of the memory system.

Table 5.1 Simulated timing specifications for all process conditions.

Process corner		Read access time (ns)	Read cycle time (ns)	Write access time (ns)	Write cycle time (ns)
Transistor	Resistor				
TM	TM	11.7	16.4	11.2	15.0
TM	WP	12.1	16.8	11.0	14.8
TM	WS	11.7	16.4	11.5	15.3
WO	TM	12.5	17.6	11.3	15.4
WO	WP	12.6	17.7	11.2	15.3
WO	WS	12.5	17.6	11.5	15.6
WP	TM	8.1	11.4	7.3	10.1
WP	WP	8.1	11.4	7.1	9.9
WP	WS	8.1	11.4	7.3	10.1
WS	TM	19.8	26.5	18.3	23.6
WS	WP	18.6	25.3	18.3	23.6
WS	WS	18.4	25.1	18.3	23.6
WZ	TM	11.3	15.6	10.9	14.5
WZ	WP	11.3	15.6	10.8	14.4
WZ	WS	11.3	15.6	11.0	14.6

A benchmark system implemented in a 0.6μm process [10] has a read access time of 20ns at 30pF load capacitance. If the time taken by the output driver to charge such a load is taken into account, the equivalent specification for the four-transistor SRAM cell system is 12.8ns. Considering that the system in [10] uses a voltage-mode sense amplifier and is substantially larger, the timing specifications may be considered to be in the same order of performance.

5.4.2 Power Dissipation

Several power specifications can be measured from the simulations performed and these can be extrapolated to the complete system. The static power dissipation, as well as power dissipation during write cycles and read cycles, are considered.

Static Power Dissipation

The bias currents of the bias circuits of the sense amplifier, *RW*-line driver and the *DIO*-line driver, as well as the op-amps that are part of the of the last two circuits, were measured. The total system consists of one sense amplifier, four *RW*-line drivers and 16 *DIO*-line drivers, The total bias currents are given in Table 5.2. The best, typical and worst static power dissipation specifications are therefore 7.66mW, 11.89mW and 17.98mW respectively. This is high compared to typical SRAM systems [10], but is a result of the analogue nature of a large percentage of the circuits, combined with the fact that the fast turn-on time requires that the low current bias networks are not turned off. It was observed that turning on the bias networks only when the analogue circuits are required, results in very long turn-on time, which was considered unacceptable.

Most SRAM systems only require biasing for the sense amplifier, because all other circuits are usually digital in nature. The percentage of the total static power required for biasing the sense amplifier is only 4.5%. If this were the only static power dissipation present in the system, this performance figure would be more competitive. An interesting point to mention about the static power dissipation is that as process conditions worsen, or when the overall quality of the devices decreases, the bias currents increase. This correlates well with the requirement of achieving a constant transconductance bias. As the devices become weaker the current has to increase to counter the effects, and hence the larger bias currents.

Table 5.2 Static bias current in mA for the complete SRAM system approximated using data from the 8x8 system simulation runs.

Transistor model Resistor model	TM	WP	WS	WO	WZ
TM	2.38	2.75	2.69	2.14	2.05
WP	3.19	3.60	3.45	3.08	2.89
WS	1.85	2.19	2.19	1.57	1.53

Write Cycle Power Dissipation

The average dynamic power dissipation of the circuit over time is found by integrating the instantaneous power delivered by the power supply over time and dividing by the total time. This is done for both the write cycle and the read cycle by performing the integration over the area of interest. During the first 60ns, the three write cycles are performed and in the last 100ns only read cycles are carried out. The average power for the write cycles can be found by considering only the first 60ns and that for the read cycle by integrating only the last 100ns. The two scenarios are shown in Figures 5.15 and 5.16. The figures show the average power dissipation for the various process corners.

In Figure 5.15 the initial constant region between 0ns and 5ns is the static power dissipation of the simulated circuit. The maximum, typical and minimum average power dissipation values during the write cycle are 44mW, 41mW, 35mW respectively. These values hold for the 8x8 array, but with all load capacitances added as if the system were complete. They are also frequency dependent and have been specified for each cycle lasting 20ns, leading to a cycle frequency of 50MHz. As the cycle frequency decreases the power dissipation will also decrease.

To extrapolate the power figures to the complete system, the values may be multiplied by four. This is however an overestimate of the power dissipation of an 8-words x 32-bits array, because the power dissipation of all peripheral circuits as well as the *RW*-line driver is now budgeted four times instead of once. The values should rather be multiplied by 2.5. It is believed that this is a more accurate estimate, because it contributes five high power dissipation driver circuits. There are 2 high power driver circuits (*RW* and *D/O*) in the 8x8 array simulation. This multiplied by 2.5 equals 5 driver circuits (1 *RW* and 4 *D/O*) as is required for the complete system.

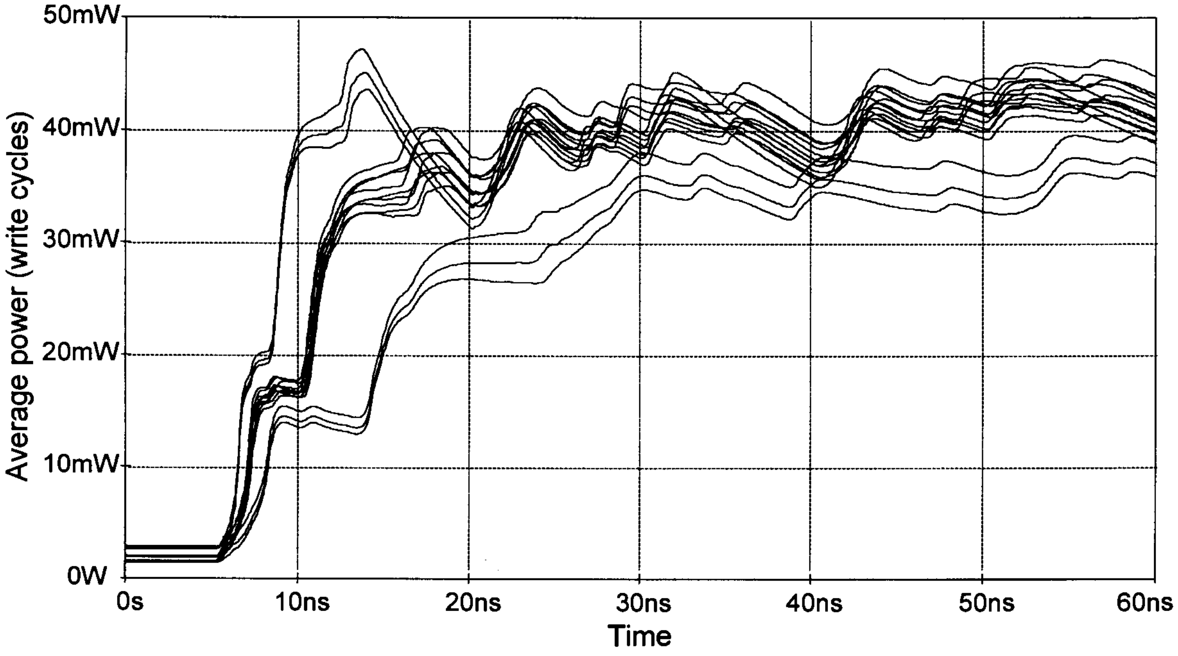


Figure 5.15 Simulated average power for all process corners for the 8x8 SRAM system during the write cycle at a cycle frequency of 50MHz.

The other three banks do not contribute power dissipation, except for the negligible static power. The final step is to add the wasted write currents of the omitted cells of the array. To do this the wasted current pulse flowing in a cell is averaged over 20ns, the cycle time used in the simulation. This yields wasted power dissipation per cell of 6.1µW, 1.8µW and 0.2µW for the worst, typical and best cases respectively. Considering typical data conditions, where one quarter of all cells in the array are unintentionally read, and worst case data conditions, where all cells are read, the total estimated power dissipation of the SRAM system is as given in Table 5.3.

Here the results of designing to reduce the wasted write currents can be seen. In the typical mean case the wasted write current contributes only about 10% of the total power dissipation. This is due to reduced and process dependent *D*/*O*-line voltages that keep the peak current level below 45µA, and on-chip control circuits that keep the pulse width as narrow as possible by activating the *D*/*O*-line drivers for the shortest possible time, to successfully write the cells.

Table 5.3 Estimated total power dissipation in mW at 50MHz cycle frequency during the write cycle of the complete SRAM system, based on results from simulating the reduced 8x8 system.

Condition	Best	Typical	Worst
Typical data	87.9	106.1	122.1
Worst data	89.1	116.8	158.4

Read Cycle Power Dissipation

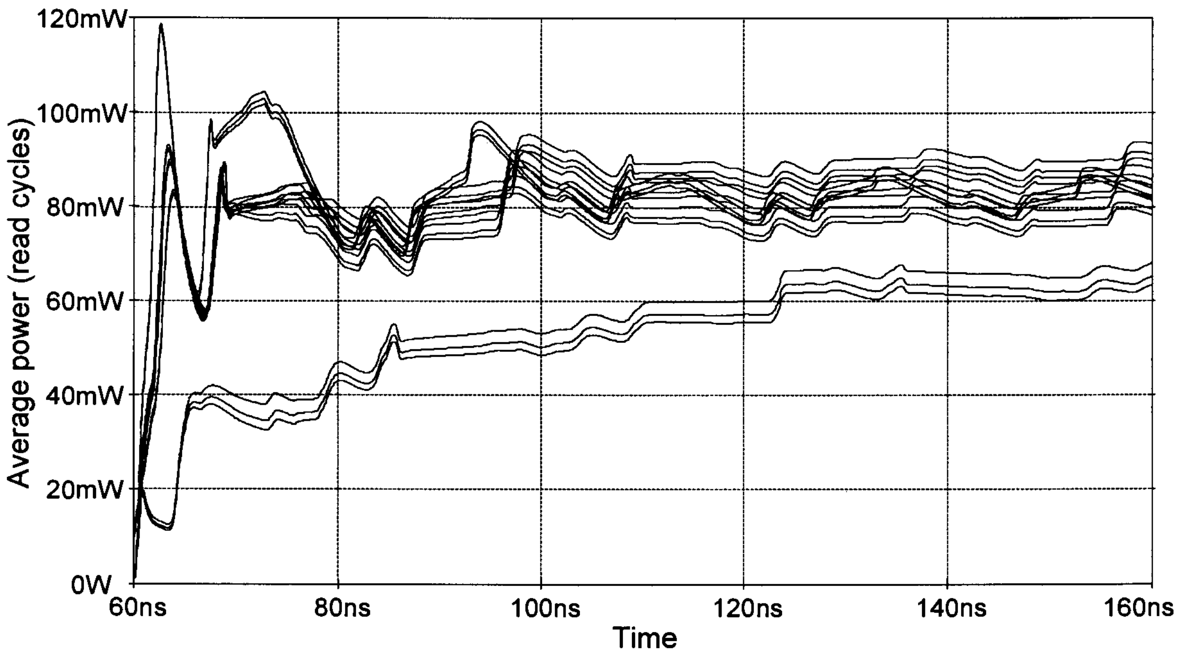


Figure 5.16 Simulated average power for all process corners for the 8x8 SRAM system during the read cycle at a cycle frequency of 50MHz.

From Figure 5.16 the best, typical and worst read cycle power dissipation is estimated to be 60mW, 80mW and 87mW, respectively. To extend this figure to a complete system no extra power dissipation is present on the memory bank except for the additional cell and reference read currents, but the extra 24 current sense amplifiers and current conveyors need to be added. These consume high currents when active, estimated at twice as much as a single *RW*-line driver circuit. It was therefore decided to multiply the power consumption figures for the 8x8 system by three to obtain the estimated power figures for the complete system. This leads to a power consumption during the read cycle of 180mW, 240mW and 261mW (best, typical, worst) at 50MHz.

These estimated power dissipation figures are competitive with benchmark systems [10] which have an active power dissipation of 231mW at 40MHz cycle frequency, considering the fact that the total estimated power dissipation lies somewhere between that of the read cycle and that of the write cycle, depending on the relative frequency of each.

5.5 LAYOUT

In order to analyse if using the four-transistor SRAM cell yields a smaller layout on a system level, the layouts of all circuit building blocks were combined. The floor-plan of the system layout is shown in Figure 5.17, and the layout in Figure 5.18.

From the layout it is evident that the peripheral circuits, as well as the routing of two signals per bit to the sense amplifier, and one data signal per bit to the memory banks take up a significant amount of area. The high power output drivers also require very wide power supply tracks so that the current density of the metal interconnect is not exceeded. These wide tracks can be seen on either side of the sense amplifier and output driver. Table 5.4 lists some important dimensions and characteristics of the layout.

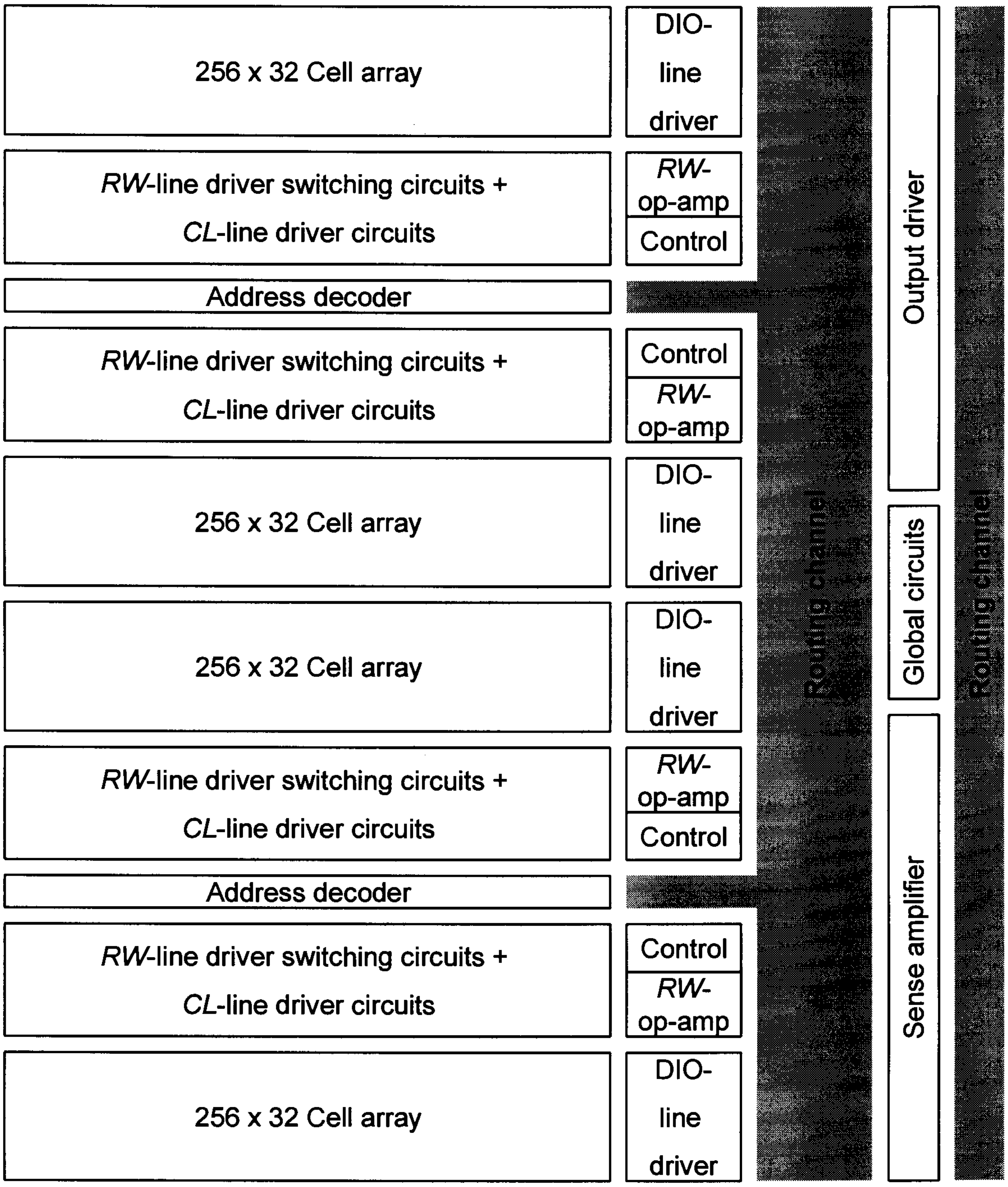


Figure 5.17 Floor-plan of the SRAM system.

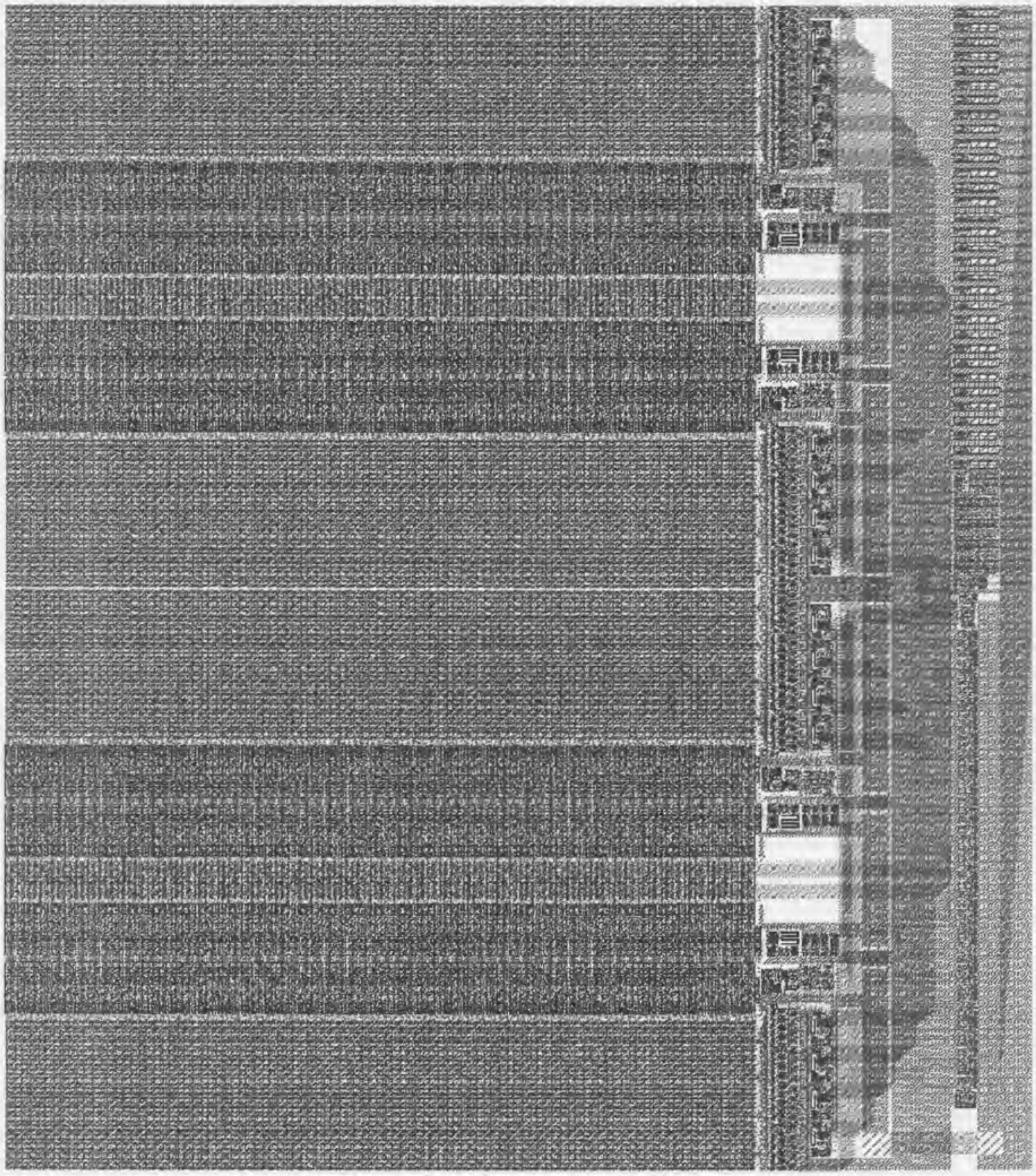


Figure 5.18 Layout of the complete four-transistor cell SRAM system in the 0.6μm CMOS process.

Table 5.4 Characteristics of the four-transistor SRAM system layout.

Characteristic	Value
Layout area - 256 x 32 cell array	0.74mm^2
Layout area - *RW*-line driver + *CL*-line driver	0.87mm^2
Layout area - *D*/*O*-line driver and control	0.22mm^2
Layout area - one memory bank	1.83mm^2
Layout area - address decoder	0.27mm^2
Layout area - all other peripheral circuits and routing	1.74mm^2
Layout area - complete system	9.59mm^2
Area ratio - memory cells to peripheral circuits for one bank	0.68
Area ratio - memory cell to peripheral circuits for the system	0.44

From the table, one of the issues associated with using the four-transistor SRAM cell can be identified. The cell may be small, but the peripheral circuits are quite large. This leads to a situation where less than a third of the chip area is used for memory cells. The layout of the system in [10] has a memory cells to peripheral circuits area ratio of 1.07, which means that about one half the chip area is used for memory cells.

5.6 COMPARISON TO A SIX-TRANSISTOR SYSTEM

The results of the previous section need to be placed in context. The system presented in [10] uses a process with similar characteristics, but the system is larger and is therefore not accurately comparable. The power consumption comparison is based on estimates for the four-transistor SRAM system, and is therefore not accurate.

For this reason it was decided to design an identical six-transistor SRAM cell system that can be simulated to obtain a more accurate comparison. During the design of this system it was decided to reuse most of the circuits designed for the four-transistor SRAM cell system. All global circuits, including the sense amplifier were reused. The read access mechanism between the sense amplifier and the bit lines is accomplished by a second current conveyor. This creates a cascaded current sensing data path with two current conveyors, as proposed in [26].

A 8x8 array, together with all peripheral circuits, was simulated using an identical setup to that shown in Figure 5.14, and the characteristics extracted. An approximate layout area was found by placing circuit blocks together and estimating the routing channel dimensions based on those of the four-transistor SRAM system. Table 5.5 compares some characteristics of the two systems.

Table 5.5 Comparison of the four-transistor and six transistor SRAM cell systems.

Characteristic	Four-transistor system	Six-transistor system	Percentage change
Layout area - 256x32 cell array	0.74mm^2	1.194mm^2	38.2% less
Layout area - complete system	9.59mm^2	7.83mm^2	22.4% more
Area ratio - memory cells to total area	0.31	0.61	49.7% less
Typical read access time	11.7ns	8.6ns	36.0% more
Typical write access time	11.2ns	5.7ns	96.5% more
Typical read cycle power dissipation	80mW	75mW	6.7% more
Typical write cycle power dissipation	41mW	35mW	17.1% more
Typical static power - compete system	11.9mW	533μW	2131% more

The table gives a clear indication that the four-transistor SRAM cell system performs worse in all areas, when compared to a similar six-transistor SRAM cell system. The cells are smaller, but it was not possible to transform this into a system area gain. This is due to the significant overhead associated with the line driver circuits. The read access time is longer although it must be mentioned that a smaller differential current is being sensed. The typical differential currents present at the current sense amplifier are in the order of 120μA if the six-transistor cell is being sensed. This is very large compared to the 5μA in the case of the four-transistor cell system. The higher differential current causes a quicker response. The write access time is almost double as long. This is the result of having a write method composed of two sub-cycles. Power dissipation while reading is almost equivalent and that for writing is about 17% more. This is the contribution made by the wasted write currents. The worst specification, as far as comparisons are concerned is the static power dissipation. The complete four-transistor SRAM cell system dissipates about 22 times more static power than the six-transistor cell system. This is a direct result of the analogue circuits used, which all require biasing. The total bias current could be reduced by sharing bias networks, although this would reduce the static power dissipation by only 10%.

Some other reasons for the weak performance of the system when compared to an equivalent implementation based on the six-transistor cell will be explored in the next chapter.

5.7 CONCLUSION

The four transistor SRAM cell array, together with the line driver circuits and the current sense amplifier, were used as building blocks to design a complete SRAM system. Even though the system is based on an analogue cell, it functions just like any other SRAM system at the external ports. To ensure that a standard interface can be used to control the system, self timing control circuits that can adapt to the speed of the memory array were designed. These timing circuits allow the system to always operate at maximum speed. This means that the time the line driver circuits are on has been reduced to the absolute minimum under all conditions. This has positive consequences as far as the power dissipation is concerned. Even though the prospects did not look good given the fact that wasted write currents exist, the power dissipation has been kept low. It is estimated to be in the same order as typical SRAM systems [10]. The timing is also in the same order as comparative systems. As far as the layout of the system is concerned, a low efficiency was achieved, if it is defined as being the percentage area of the layout dedicated to memory cells. This is only about 30.7%.

A more reliable performance comparison was made by designing and simulating a similar system based on the six-transistor SRAM cell. Here the four-transistor cell system compares poorly. The system based on the six-transistor cell outperforms the one based on the four-transistor cell in all three aspects, which are layout area, power dissipation as well as speed. This comparison was based on simulations of identically sized systems.

6. CONCLUSION

6.1 WHAT WAS GIVEN?

In 1995 a four-transistor SRAM cell, where the access transistors are omitted, was proposed [1]. The access of the cell was achieved via the source nodes of the four transistors of the cross-coupled inverter pair. A paper presented at the International Symposium on Circuits and Systems in May 2000 [2] described the operation of the cell in detail and stated that the functionality has been proven. A method of creating an array of cells was also proposed. A promising 14.7% reduction in area created the need for further investigation. Some issues such as high power dissipation and degraded noise margins due to reduced power supply voltages would also need consideration.

6.2 WHAT WAS THE AIM?

The proposed four-transistor SRAM cell was the starting point of the research described in this document. It was decided to design a complete SRAM system based on the four-transistor cell. This would allow the concept of the cell as a building block for an SRAM system to be investigated and would indicate if the cell area advantage could be transformed into a system area advantage.

6.3 WHAT HAS BEEN ACCOMPLISHED?

This document described the design of the SRAM system based on the four-transistor SRAM cell. The first step was a closer investigation of the cell itself. The proposed write access mechanism was found to lack reliability as the process conditions change and to suffer from high power dissipation during the write cycle. A different write method has therefore been proposed that not only is reliable as the process conditions vary, but also wastes 87.5% less power and results in further reduction of the cell size with one line fewer to route.

The noise margins of the cell under read and write access were analysed and found to be a useful tool in designing the magnitude of the voltage deviations that need to be applied to the source nodes of the cell during access. A 38% reduction in area was achieved in comparison to a six-transistor cell with identical noise margin of 0.6V.

The source driver circuits for realisation of the access to the cell, were designed. Here the power and speed characteristics of the cell were slightly improved by designing the voltage deviations to fit with the given process conditions. Constant transconductance biasing systems were used to ensure that the performance of the analogue circuits does not vary drastically as the device quality varies. The driver circuits use a combination of two feedback loops to obtain invariance to all conditions, as well as fast charging of the array capacitance to the required voltage level.

The current output of the cell is sensed using a clamped bit line latching current sense amplifier, based on a cross-coupled inverter pair. This is supplemented with a current conveyor, so that the sense amplifier can be isolated from the sensitive source nodes of the SRAM array. This aids in maintaining the noise margins.

The final step was to design the complete system. Self timed control circuits generate the required signals for the two-cycle write method and current sensing. A standard SRAM system interface has been created. This system was compared to an identical system based on the six-transistor cell. Here it was found that the latter outperforms the four-transistor SRAM cell system in all the measured specifications.

6.4 WHAT CAN BE LEARNED FROM THIS?

Although it was not possible to achieve an area advantage on system level it has been shown that the four-transistor SRAM cell can be used to create a system that performs well.

The fundamental restriction on performance is the high capacitance contributed by each cell to the common access lines. This capacitance can potentially be three times greater than that of the six-transistor cell. To drive this load at an identical speed, the driver circuits need to be three times larger. Added to this is the fact that analogue circuits are required which consume larger areas to achieve the same performance of their digital counterparts. The high capacitance limits the speed and the maximum array size that can be implemented. The more the design has to be split up into banks, the higher becomes the area overhead required for the peripheral circuits.

The first iteration of the system design has provided a circuit and a set of specifications. A design route where none of the specifications was neglected was chosen, resulting in a system where each specification is average rather than one being exceptionally good and the others therefore weak. The average path was important to investigate the relationship between the performance parameters. Higher speeds require more power and larger circuits, but reduced area circuits typically also mean longer delays and higher power dissipation. Implementing the complete system has therefore been a vehicle to achieve greater understanding of its most important building blocks.

6.5 WHAT IS THE NEXT STEP?

The first step towards improving characteristics is having a set of specifications to build upon. This is what has been delivered and from the results it is evident that SRAM systems based on the six-transistor cell will probably always be faster. The reduced noise margin of the four-transistor cell is the cause of this. A six-transistor cell with the same noise margin has quite strong access devices. This allows fast reading due to high differential currents or fast bit line discharge, and it also allows strong static and dynamic write conditions to be created with ease. As far as power dissipation is concerned, the high static current and the wasted write currents, although significantly reduced, will always pose a restriction.

The reduced area of the cell should rather be put to good use in systems where small area is the only important factor. The peripheral circuits may then be designed to be as small as possible, disregarding speed and power dissipation specifications. Given this framework, different methods of writing the cell could be devised that allow certain peripheral circuits to be reduced in complexity or size.

For example, during the design of the driver circuits, the importance of not having a high voltage drop present on the *D/O*-line and the *RW*-line at the same time was discussed. This could lead to the undesirable situation that the cell may be unintentionally written. This idea can be expanded to creating a new method of writing the cell. Assume that a resistance *R* is placed in the *D/O*-line and that the potential of the *RW*-line is reduced by *Y*, as shown in Figure 6.1. A current *I* flowing in the opposite inverter indicates that the cell is in the "clear" state. It is now desired to flip the state of the cell. The current flowing can be used to create a voltage drop over the resistance in the *D/O*-line, thereby applying the required *D/O*-line deviation *X* to create the static write conditions. This scheme has the advantage that the data to be written is applied to the array by means of a high or low impedance in the *D/O*-line. This may reduce the area significantly and be the first step towards a system based on the four-transistor SRAM cell that is smaller than one based on the six-transistor cell.

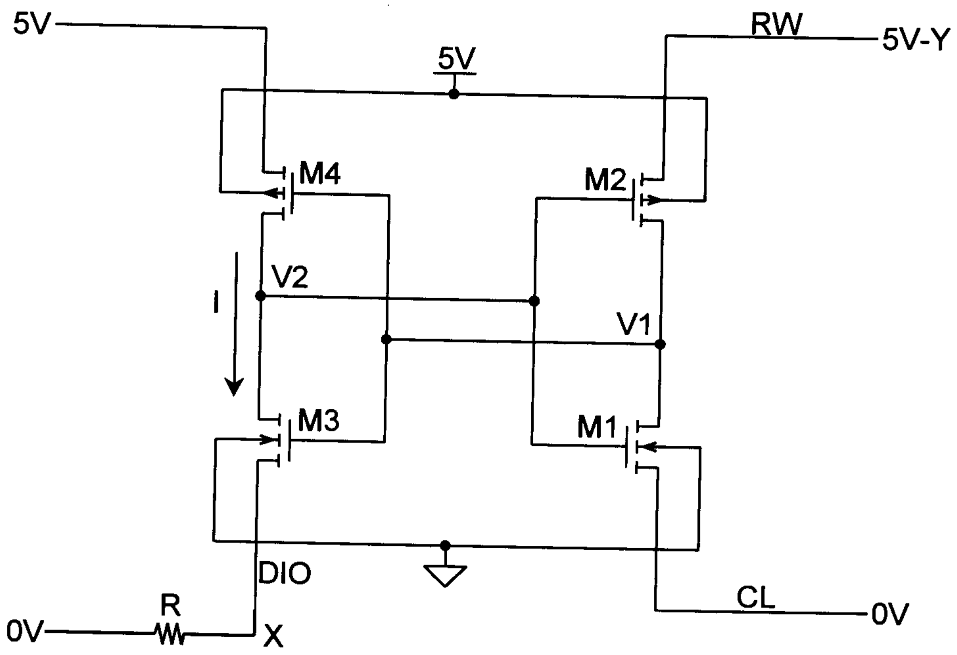


Figure 6.1 Alternative cell write method.

REFERENCES

[1] E Seevinck, Project Proposal Report, University of Pretoria, January 1995.

[2] T-H Joubert, E Seevinck, M du Plessis, "A CMOS Reduced-Area SRAM Cell", *Proceedings of the IEEE 2000 International Symposium on Circuits and Systems*, pp III-335-8, 28-31 May 2000, Geneva, Switzerland.

[3] S Yamamoto, N Tanimura, K Nagasawa *et al*, "A 256K CMOS SRAM with Variable Impedance Data-Line Loads", *IEEE Journal of Solid-State Circuits*, Vol. 20, No. 5, pp 924-8, October 1985.

[4] N Okazaki, T Komatsu, N Hoshi *et al*, "A 16ns 2K X 8 Bit Full CMOS SRAM", *IEEE Journal of Solid-State Circuits*, Vol. 19, No. 5, pp 552-6, October 1984.

[5] R Hollingsworth, A Ipri, C Kim, "A CMOS/SOS 4K Static RAM", *IEEE Journal of Solid-State Circuits*, Vol. 13, No. 5, pp 664-9, October 1978.

[6] H Levy, E Daniel, T McGill, "A Transistorless-Current-Mode Static RAM Architecture", *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 4, pp 669-72, April 1998.

[7] K Takeda, Y Aimoto, N Nakamura, *et al*, "A 16-Mb 400-MHz Loadless CMOS Four-Transistor SRAM Macro", *IEEE Journal of Solid-State Circuits*, Vol. 35, No. 11, pp 1631-39, November 2000.

[8] K Noda, K Matsui, K Imai, *et al*, "A 1.9-μm^2 Loadless CMOS Four-Transistor SRAM Cell in a 0.18μm Logic Technology", *IEDM Dig. Tech. Papers*, pp 22.8.1-4, 1998.

[9] N Weste, K Eshraghian, *Principles of CMOS VLSI Design, A Systems Perspective*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1985.

[10] E Seevinck, F List, J Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells", *IEEE Journal of Solid-State Circuits*, Vol. 22, No. 5, pp 748-54, October 1987.

[11] T Hirose, H Kuriyama, S Murakami, *et al*, "A 20-ns 4-Mb CMOS SRAM with Hierarchical Word Decoding Architecture", *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 5, pp 1068-73, October 1990.

[12] J Rabaey, *Digital Integrated Circuits, A Design Perspective*, Prentice-Hall Inc., New Jersey, 1996.

[13] P Gray, R Meyer, *Analysis and Design of Analog Integrated Circuits, Third Edition*, John Wiley and Sons, Inc., New York, 1993.

[14] K Anami, M Yoshimoto, H Shinohara, *et al*, "Design Considerations of a Static Memory Cell", *IEEE Journal of Solid-State Circuits*, Vol. 18, No. 4, pp 414-7, August 1983.

[15] J Uyemura, *Fundamentals of MOS Digital Integrated Circuits*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1988.

[16] Austria Mikro Systeme International AG, *0.6μm CMOS Design Rules*, Document No. 9931025, Revision 2.0, October 1998.

[17] J Lohstroh, E Seevinck, J De Groot, "Worst-Case Static Noise Margin Criteria for Logic Circuits and Their Mathematical Equivalence", *IEEE Journal of Solid-State Circuits*, Vol. 18, No. 6, pp 803-7, December 1983.

[18] J Lohstroh, "Static and Dynamic Noise Margins of Logic Circuits", *IEEE Journal of Solid-State Circuits*, Vol. 14, No. 3, pp 591-8, June 1979.

[19] Austria Mikro Systeme International AG, *0.6μm CMOS CUP Process Parameters*, Document No. 9933011, Revision B, October 1998.

[20] M Yoshimoto, K Anami, H Hirofumi, *et al*, "A Divide Word-Line Structure in the Static RAM and Its Application to a 64K Full CMOS RAM", *IEEE Journal of Solid-State Circuits*, Vol. 18, No. 5, pp 479-85, October 1983.

[21] D Johns, K Martin, *Analog Integrated Circuit Design*, John Wiley and Sons, Inc., New York, 1997.

[22] Gregorian, G Temes, *Analog MOS Integrated Circuits for Signal Processing*, John Wiley and Sons, Inc., New York, 1986.

[23] P Gray, R Meyer, "MOS Operational Amplifier Design - A Tutorial Overview", *IEEE Journal of Solid-State Circuits*, Vol. 17, No. 6, pp 969-82, December 1982.

[24] K Ishibashi, K Komiyaji, S Morita, *et al*, "A 12.5-ns 16-Mb CMOS SRAM with Common-Centroid-Geometry-Layout Sense Amplifiers", *IEEE Journal of Solid-State Circuits*, Vol. 29, No. 4, pp 411-7, April 1994.

[25] S Kayano, K Ichinose, Y Kohno, *et al*, "25-ns 256Kx1/64Kx4 CMOS SRAM's", *IEEE Journal of Solid-State Circuits*, Vol. 21, No. 5, pp 686-91, October 1986.

[26] E Seevinck, P van Beers, H Ontrop, "Current-Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's", *IEEE Journal of Solid-State Circuits*, Vol. 26, No. 4, pp 525-36, April 1991.

[27] Y Seng, S Rofail, "1.5V High Speed Low Power CMOS Current Sense Amplifier" *Electronics Letters*, Vol. 31, No. 23, pp 1991-3, November 1995.

[28] G Lahiji, A Sodagar, "High-Speed Current-Mode Sense Amplifier" *Electronics Letters*, Vol. 30, No. 17, pp 1371-2, August 1994.

[29] P Chee, P Liu, L Siek, "High-Speed Hybrid Current-Mode Sense Amplifier for CMOS SRAMs" *Electronics Letters*, Vol. 28, No. 9, pp 871-3, April 1992.

[30] L Kim, R Dutton, "Metastability of CMOS Latch/Flip-Flop" ", *IEEE Journal of Solid-State Circuits*, Vol. 25, No. 4, pp 942-951, April 1990.

[31] T Blalock, R Jaeger, "A High-Speed Clamped Bit-Line Current-Mode Sense Amplifier" ", *IEEE Journal of Solid-State Circuits*, Vol. 26, No. 4, pp 542-8, April 1991.

[32] H Veendrick, *Deep-Submicron CMOS ICs*, Kluwer BedrijfsInformatie b.v. - Deventer, The Netherlands, 1998.

[33] R Howe, C Sodini, *Microelectronics, An Integrated Approach*, Prentice-Hall Inc., New Jersey, 1997.

ADDENDUM A: C-CODE FOR NOISE MARGIN ANALYSIS

A.1 FOUR-TRANSISTOR CELL NOISE MARGIN ANALYSIS

```
#include <stdio.h>
#include <time.h>
#define MAXV 501
#define MAXQ 70
#define MAXR 70
#define One_Over_Root2 0.707106781
int main(void){
  // Declarations
  char filenameinn[12];
  char filenameinp[12];
  char filenameoutr[12];
  char filenameoutq[12];
  char filenameswitch[12];
  char indata[200];
  float rStart = 0.5;
  float rEnd = 5;
  float rSpacing = 0.5;
  float qStart = 0.5;
  float qEnd = 5;
  float qSpacing = 0.5;
  float VStart = 0;
  float VEnd = 5.0;
  float VSpacing = 0.01;
  register int i,j,k,l,r,q,v;
  register float temp, tempa;
  register float err, erri, errj;
  register float m,c;
  float NMMax, NMMin;
  int rlength, qlength, vlength;
  float Vin[MAXV], R[MAXR], Q[MAXQ];
  float Vout1[MAXR][MAXQ][MAXV], Vout2[MAXR][MAXQ][MAXV];
  float SNM[MAXR][MAXQ][2];
  char Switch[MAXR][MAXQ];
  int bar = 0;
  int barcount = 0;
  float w[MAXV], u[MAXV], s[MAXV], t[MAXV];
  FILE *in;
  FILE *out;
  time_t dt;
  // Initialisation
  printf("\e[J");
  printf("Welcome to SNM - Static Noise Margin Analysis!!!!!\n\n");
  //Creating Vectors
  printf("\n\nCreating Data Vectors\n");
  rlength = 0;
  i = 1;
  while (i == 1){
    temp = rStart+(rSpacing*rlength);
    if (temp <= rEnd){
      R[rlength] = temp;
      rlength++;}
    else{
      i = 0;}}
  qlength = 0;
```

```
i = 1;
while (i == 1){
  temp = qStart+(qSpacing*qlength);
  if (temp <= qEnd){
    Q[qlength] = temp;
    qlength++;}
  else{
    i = 0;}}
vlength = 0;
i = 1;
while (i == 1){
  temp = VStart+(VSpacing*vlength);
  if (temp <= VEnd){
    Vin[vlength] = temp;
    vlength++;}
  else{
    i = 0;}}
// Read Data from input files
printf("Reading the LO data input file\n");
i = qlength*rlength;
if ((in = fopen("Invlo.csd", "rt")) == NULL){
  fprintf(stderr, "Cannot open LO data input file\n");
  return 1;}
indata[0] = '#';
indata[1] = 'H';
r=0;
q=0;
while (i > 0){
  while ((indata[0] != '#') || (indata[1] != 'C')){
    fgets(indata, 200, in);}
  v=0;
  while ((indata[0] == '#') && (indata[1] == 'C')){
    fgets(indata, 200, in);
    sscanf(indata, "%f", &Vout1[r][q][v]);
    v++;
    fgets(indata, 200, in);}
  r++;
  if (r==rlength){
    r=0;
    q++;}
  i--;}
fclose(in);
printf("Reading the HI data input file\n");
i = qlength*rlength;
if ((in = fopen("Invhi.csd", "rt")) == NULL){
  fprintf(stderr, "Cannot open HI data input file\n");
  return 1;}
indata[0] = '#';
indata[1] = 'H';
r=0;
q=0;
while (i > 0){
  while ((indata[0] != '#') || (indata[1] != 'C')){
    fgets(indata, 200, in);}
  v=0;
  while ((indata[0] == '#') && (indata[1] == 'C')){
    fgets(indata, 200, in);
    sscanf(indata, "%f", &Vout2[r][q][v]);
    v++;
```

```
      fgets(indata, 200, in);}
    r++;
    if(r==rlength){
      r=0;
      q++;}
    i--;}
  fclose(in);
  //Analysing the Data
  printf("Analysing Noise Margins\n");
  printf("[           ]\n\e[A[");
  for(r=0;r<rlength;++r){
    for(q=0;q<qlength;++q){
      //Translate Coordinate systems
      for (v=0;v<vlength;++v){
        u[v] = One_Over_Root2*(Vout1[r][q][v] + Vin[v]);
        w[v] = One_Over_Root2*(Vout1[r][q][v] - Vin[v]);
        s[v] = One_Over_Root2*(Vin[v] + Vout2[r][q][v]);
        t[v] = One_Over_Root2*(Vin[v] - Vout2[r][q][v]);}
      // Noise Margin Algorithm
      NMMin = 0;
      NMMax = 0;
      for (v=0;v<vlength;++v){
        i=0;
        j=0;
        erri=1000;
        errj=1000;
        temp = w[v];
        // Scan for closest values
        for (k=0;k<vlength;++k){
          if((temp <= t[vlength-1]) && (temp >= t[0])){
            tempa = t[k];
            err = temp-tempa;
            if ((err >=0) && (err <= erri)){
              erri = err;
              i = k;}
            err = tempa-temp;
            if ((err >=0) && (err <= errj)){
              errj = err;
              j = k;}}
          else{
            k = vlength;
            i = -1;
            j = -1;}}
        // Calculate the noise margin
        if (i != j){
          m = (t[i]-t[j])/(s[i]-s[j]);
          c = t[i]-m*s[i];
          temp = u[v]-((w[v]-c)/m);}
        else{
          if (i != -1){
            temp = u[v] - s[i];}
          else{
            temp = 0;}}
        if (temp > NMMax){
          NMMax = temp;}
        if (temp < NMMin){
          NMMin = temp;}}
      SNM[r][q][0] = -NMMin*One_Over_Root2;
      SNM[r][q][1] = NMMax*One_Over_Root2;
    if ((NMMin >= -0.01) || (NMMax <= 0.01)){
      Switch[r][q] = '0';}
    else{
      Switch[r][q] = '1';}
    bar++;
    temp = (float)barcount/77.0;
    tempa = (float)bar/(rlength*qlength);
    if (temp < tempa){
      barcount = 0;
      while(temp < tempa){
        printf("*");
        barcount++;
        temp = (float)barcount/77.0;}
      printf("\n\e[A[");}}}
printf("\n");
bar = 0;
barcount = 0;
// Write the Switch data output file
time(&dt);
strcpy(indata,ctime(&dt));
printf("\nWriting the Switch data output file\n");
if ((out = fopen("Swit.txt", "wt")) == NULL){
  fprintf(stderr, "Cannot open switch data output file\n");
  return 1;}
fprintf(out,"Switch Data output File\n");
fprintf(out,"Written by 4TCell SNM Analysis %s\n", indata);
fprintf(out,"Vertical: q Value\n");
fprintf(out,"Horizontal: r Value\n\n");
fprintf(out,"       ");
for (i=0;i<rlength;++i){
  fprintf(out,"%1.2f  ",R[i]);}
fprintf(out,"\n\n");
for (j=(qlength-1);j >= 0; --j){
  fprintf(out,"%1.2f  ", Q[j]);
  for (i=0;i<rlength;++i){
    fprintf(out,"    %c  ",Switch[i][j]);}
  fprintf(out,"  %1.2f\n", Q[j]);}
fprintf(out,"\n       ");
for (i=0;i<rlength;++i){
  fprintf(out,"%1.2f  ",R[i]);}
fprintf(out,"\n");
fclose(out);
//Write the X Data output file
printf("Writing the R data output file\n");
if ((out = fopen("Outwr.csd", "wt")) == NULL){
  fprintf(stderr, "Cannot open R data output file\n");
  return 1;}
for(q=0;q<qlength;++q){
  fprintf(out,"#H\n");
  fprintf(out,"SOURCE='SNM Analysis' VERSION='1.0 (June 2001)'\n");
  fprintf(out,"TITLE='** Static Noise Margin of 4TSRAM Cell '\n");
  fprintf(out,"SUBTITLE='Step parameter Q = %1.3E'\n", Q[q]);
  fprintf(out,"TIME='%c%c:%c%c:%c%c' DATE='%c%c%c/%c%c/%c%c'
       TEMPERATURE='27'\n",indata[11], indata[12], indata[14],
       indata[15], indata[17], indata[18], indata[4], indata[5],
       indata[6], indata[8],indata[9], indata[22], indata[23]);
  fprintf(out,"ANALYSIS='DC Sweep' SERIALNO='00001'\n");
  if(q==0){
    i = 3;}
    else{
      i = 2;}
    fprintf(out,"ALLVALUES='YES' COMPLEXVALUES='NO' NODES='%i'\n",i);
    fprintf(out,"SWEEPVAR='r' SWEEPMODE='LINEAR'\n");
    fprintf(out,"XBEGIN='%1.3E' XEND='%1.3E'\n",rStart,rEnd);
    fprintf(out,"FORMAT='0 VOLTSorAMPS;EFLOAT : NODEorBRANCH;NODE  '\n");
    fprintf(out,"DGTLDATA='NO'\n");
    fprintf(out,"#N\n");
    if (q==0){
      fprintf(out,"'V(Min_Noise_Margin)' 'V(Max_Noise_Margin)'
          'Q(Cell_Trigger)'\n");}
    else{
      fprintf(out,"'V(Min_Noise_Margin)' 'V(Max_Noise_Margin)'\n");}
    for (r=0;r<rlength;++r){
      if (q==0){
        fprintf(out,"#C %1.3E 3\n",R[r]);
        i = 0;
        for (j=(qlength-1);j>=0;--j){
          if (Switch[r][j] == '0'){
            i = j;}}
        fprintf(out,"%1.3E:1 %1.3E:2
          %1.3E:3\n",SNM[r][q][0],SNM[r][q][1], Q[i]);}
      else{
        fprintf(out,"#C %1.3E 2\n",R[r]);
        fprintf(out,"%1.3E:1 %1.3E:2\n",SNM[r][q][0], SNM[r][q][1]);}}
    fprintf(out,"#;\n");}
  fclose(out);
  //Write the Y Data output file
  printf("Writing the Q data output file\n");
  if ((out = fopen("Outwq.csd", "wt")) == NULL){
    fprintf(stderr, "Cannot open Q data output file\n");
    return 1;}
  for(r=0;r<rlength;++r){
    fprintf(out,"#H\n");
    fprintf(out,"SOURCE='SNM Analysis' VERSION='1.0 (June 2001)'\n");
    fprintf(out,"TITLE='** Static Noise Margin of 4TSRAM Cell '\n");
    fprintf(out,"SUBTITLE='Step parameter R = %1.3E'\n", R[r]);
    fprintf(out,"TIME='%c%c:%c%c:%c%c' DATE='%c%c%c/%c%c/%c%c'
          TEMPERATURE='27'\n",indata[11], indata[12], indata[14],
          indata[15], indata[17], indata[18], indata[4], indata[5],
          indata[6], indata[8],indata[9], indata[22], indata[23]);
    fprintf(out,"ANALYSIS='DC Sweep' SERIALNO='00001'\n");
    if(r==0){
      i = 3;}
    else{
      i = 2;}
    fprintf(out,"ALLVALUES='YES' COMPLEXVALUES='NO' NODES='%i'\n",i);
    fprintf(out,"SWEEPVAR='q' SWEEPMODE='LINEAR'\n");
    fprintf(out,"XBEGIN='%1.3E' XEND='%1.3E'\n",qStart,qEnd);
    fprintf(out,"FORMAT='0 VOLTSorAMPS;EFLOAT : NODEorBRANCH;NODE  '\n");
    fprintf(out,"DGTLDATA='NO'\n");
    fprintf(out,"#N\n");
    if (r==0){
    fprintf(out,"'V(Min_Noise_Margin)' 'V(Max_Noise_Margin)'
          'R(Cell_Trigger)'\n");}
    else{
      fprintf(out,"'V(Min_Noise_Margin)' 'V(Max_Noise_Margin)'\n");}
    for (q=0;q<qlength;++q){
      if (r==0){
        fprintf(out,"#C %1.3E 6\n",Q[q]);
        i = 0;
        for (j=(rlength-1);j>=0;--j){
          if (Switch[j][q] == '0'){
            i = j;}}
        fprintf(out,"%1.3E:1 %1.3E:2
          %1.3E:3\n",SNM[r][q][0],SNM[r][q][1], R[i]);}
      else{
        fprintf(out,"#C %1.3E 2\n",Q[q]);
        fprintf(out,"%1.3E:1 %1.3E:2\n",SNM[r][q][0], SNM[r][q][1]);}}
    fprintf(out,"#;\n");}
  fclose(out);
  return 0;}
```

A.2 SIX-TRANSISTOR CELL NOISE MARGIN ANALYSIS

```
#include <stdio.h>
#include <time.h>
#define MAXV 501
#define MAXQ 70
#define MAXR 70
#define One_Over_Root2 0.707106781

int main(void){
  // Declarations
  char filenamein[12];
  char filenameoutr[12];
  char filenameoutq[12];
  char indata[200];
  float rStart = 1;
  float rEnd = 3;
  float rSpacing = 0.1;
  float qStart = 1;
  float qEnd = 3;
  float qSpacing = 0.1;
  float VStart = 0;
  float VEnd = 5.0;
  float VSpacing = 0.01;
  register int i,j,k,l,r,q,v;
  register float temp, tempa;
  register float err, erri, errj;
  register float m,c;
  float NMMax, NMMin;
  int rlength, qlength, vlength;
  float Vin[MAXV], R[MAXR], Q[MAXQ];
  float Vout[MAXR][MAXQ][MAXV];
  float SNM[MAXR][MAXQ][2];
  int bar = 0;
  int barcount = 0;
  float w[MAXV], u[MAXV], s[MAXV], t[MAXV];
  FILE *in;
  FILE *out;
  time_t dt;
  // Initialisation
  printf("\e[J");
  printf("Welcome to SNM - Static Noise Margin Analysis!!!!!\n\n");
  //Creating Vectors
  printf("\n\nCreating Data Vectors\n");
```

```
rlength = 0;
i = 1;
while (i == 1) {
  temp = rStart+(rSpacing*rlength);
  if (temp <= rEnd){
    R[rlength] = temp;
    rlength++;}
  else{
    i = 0;}}
qlength = 0;
i = 1;
while (i == 1){
  temp = qStart+(qSpacing*qlength);
  if (temp <= qEnd){
    Q[qlength] = temp;
    qlength++;}
  else{
    i = 0;}}
vlength = 0;
i = 1;
while (i == 1){
  temp = VStart+(VSpacing*vlength);
  if (temp <= VEnd){
    Vin[vlength] = temp;
    vlength++;}
  else{
    i = 0;}}
// Read Data from input files
printf("Reading the HI data input file\n");
i = qlength*rlength;
if ((in = fopen("Invhi.csd", "rt")) == NULL){
  fprintf(stderr, "Cannot open HI data input file\n");
  return 1;}
indata[0] = '#';
indata[1] = 'H';
r=0;
q=0;
while (i > 0){
  while ((indata[0] != '#') || (indata[1] != 'C')){
    fgets(indata, 200, in);}
  v=0;
  while ((indata[0] == '#') && (indata[1] == 'C')){
    fgets(indata, 200, in);
    sscanf(indata, "%f", &Vout[r][q][v]);
    v++;
    fgets(indata, 200, in);}
  r++;
  if(r==rlength){
    r=0;
    q++;}
  i--;}
fclose(in);
//Analysing the Data
printf("Analysing Noise Margins\n");
printf("[                                              ]\n\e[A[");
for(r=0;r<rlength;++r){
  for(q=0;q<qlength;++q){
    //Translate Coordinate systems
    for (v=0;v<vlength;++v){
      u[v] = One_Over_Root2*(Vout[r][q][v] + Vin[v]);
      w[v] = One_Over_Root2*(Vout[r][q][v] - Vin[v]);
      s[v] = One_Over_Root2*(Vin[v] + Vout[r][q][v]);
      t[v] = One_Over_Root2*(Vin[v] - Vout[r][q][v]);}
    // Noise Margin Algorithm
    NMMin = 0;
    NMMax = 0;
    for (v=0;v<vlength;++v){
      i=0;
      j=0;
      erri=1000;
      errj=1000;
      temp = w[v];
      // Scan for closest values
      for (k=0;k<vlength;++k){
        if ((temp <= t[vlength-1]) && (temp >= t[0])){
          tempa = t[k];
          err = temp-tempa;
          if ((err >=0) && (err <= erri)){
            erri = err;
            i = k;}
          err = tempa-temp;
          if ((err >=0) && (err <= errj)){
            errj = err;
            j = k;}}
        else{
          k = vlength;
          i = -1;
          j = -1;}}
      // Calculate the noise margin
      if (i != j){
        m = (t[i]-t[j])/(s[i]-s[j]);
        c = t[i]-m*s[i];
        temp = u[v]-((w[v]-c)/m);}
      else{
        if (i != -1){
          temp = u[v] - s[i];}
        else{
          temp = 0;}}
      if (temp > NMMax){
        NMMax = temp;}
      if (temp < NMMin){
        NMMin = temp;}}
    SNM[r][q][0] = -NMMin*One_Over_Root2;
    SNM[r][q][1] = NMMax*One_Over_Root2;
    bar++;
    temp = (float)barcount/77.0;
    tempa = (float)bar/(rlength*qlength);
    if (temp < tempa){
      barcount = 0;
      while(temp < tempa){
        printf("*");
        barcount++;
        temp = (float)barcount/77.0;}
      printf("\n\e[A[");}}
  printf("\n");
  bar = 0;
  barcount = 0;}
//Write the R Data output file
```

```
time(&dt);
strcpy(indata,ctime(&dt));
printf("Writing the R data output file\n");
if ((out = fopen("Outrr.csd", "wt")) == NULL){
  fprintf(stderr, "Cannot open R data output file\n");
  return 1;}
for(q=0;q<qlength;++q){
  fprintf(out,"#H\n");
  fprintf(out,"SOURCE='SNM Analysis' VERSION='1.0 (June 2001)'\n");
  fprintf(out,"TITLE='** Static Noise Margin of 6TSRAM Cell '\n");
  fprintf(out,"SUBTITLE='Step parameter Q = %1.3E'\n", Q[q]);
  fprintf(out,"TIME='%c%c:%c%c:%c%c' DATE='%c%c%c/%c%c/%c%c'
        TEMPERATURE='27'\n",indata[11], indata[12], indata[14],
        indata[15], indata[17], indata[18], indata[4], indata[5],
        indata[6], indata[8],indata[9], indata[22], indata[23]);
  fprintf(out,"ANALYSIS='DC Sweep' SERIALNO='00001'\n");
  fprintf(out,"ALLVALUES='YES' COMPLEXVALUES='NO' NODES='1'\n",i);
  fprintf(out,"SWEEPVAR='r' SWEEPMODE='LINEAR'\n");
  fprintf(out,"XBEGIN='%1.3E' XEND='%1.3E'\n",rStart,rEnd);
  fprintf(out,"FORMAT='0 VOLTSorAMPS;EFLOAT : NODEorBRANCH;NODE  '\n");
  fprintf(out,"DGTLDATA='NO'\n");
  fprintf(out,"#N\n");
  fprintf(out,"'V(Noise_Margin)'\n");
  for (r=0;r<rlength;++r){
    fprintf(out,"#C %1.3E 1\n",R[r]);
    fprintf(out,"%1.3E:1\n",SNM[r][q][0]);}
  fprintf(out,"#;\n");}
fclose(out);
//Write the Y Data output file
printf("Writing the Q data output file\n");
if ((out = fopen("Outrq.csd", "wt")) == NULL){
  fprintf(stderr, "Cannot open Q data output file\n");
  return 1;}
for(r=0;r<rlength;++r){
  fprintf(out,"#H\n");
  fprintf(out,"SOURCE='SNM Analysis' VERSION='1.0 (June 2001)'\n");
  fprintf(out,"TITLE='** Static Noise Margin of 6TSRAM Cell '\n");
  fprintf(out,"SUBTITLE='Step parameter R = %1.3E'\n", R[r]);
  fprintf(out,"TIME='%c%c:%c%c:%c%c' DATE='%c%c%c/%c%c/%c%c'
        TEMPERATURE='27'\n",indata[11], indata[12], indata[14],
        indata[15], indata[17], indata[18], indata[4], indata[5],
        indata[6], indata[8],indata[9], indata[22], indata[23]);
  fprintf(out,"ANALYSIS='DC Sweep' SERIALNO='00001'\n");
  fprintf(out,"ALLVALUES='YES' COMPLEXVALUES='NO' NODES='1'\n",i);
  fprintf(out,"SWEEPVAR='q' SWEEPMODE='LINEAR'\n");
  fprintf(out,"XBEGIN='%1.3E' XEND='%1.3E'\n",qStart,qEnd);
  fprintf(out,"FORMAT='0 VOLTSorAMPS;EFLOAT : NODEorBRANCH;NODE  '\n");
  fprintf(out,"DGTLDATA='NO'\n");
  fprintf(out,"#N\n");
  fprintf(out,"'V(Noise_Margin)'\n");
  for (q=0;q<qlength;++q){
    fprintf(out,"#C %1.3E 1\n",Q[q]);
    fprintf(out,"%1.3E:1\n",SNM[r][q][0]);}
  fprintf(out,"#;\n");}
fclose(out);
return 0;}
```

ADDENDUM B: LAYOUT LEGEND

□ N-Well

▒ N-Active

░ P-Active

▨ Poly 1

■ Contact

▒ Metal 1

▓ Via

▦ Metal 2

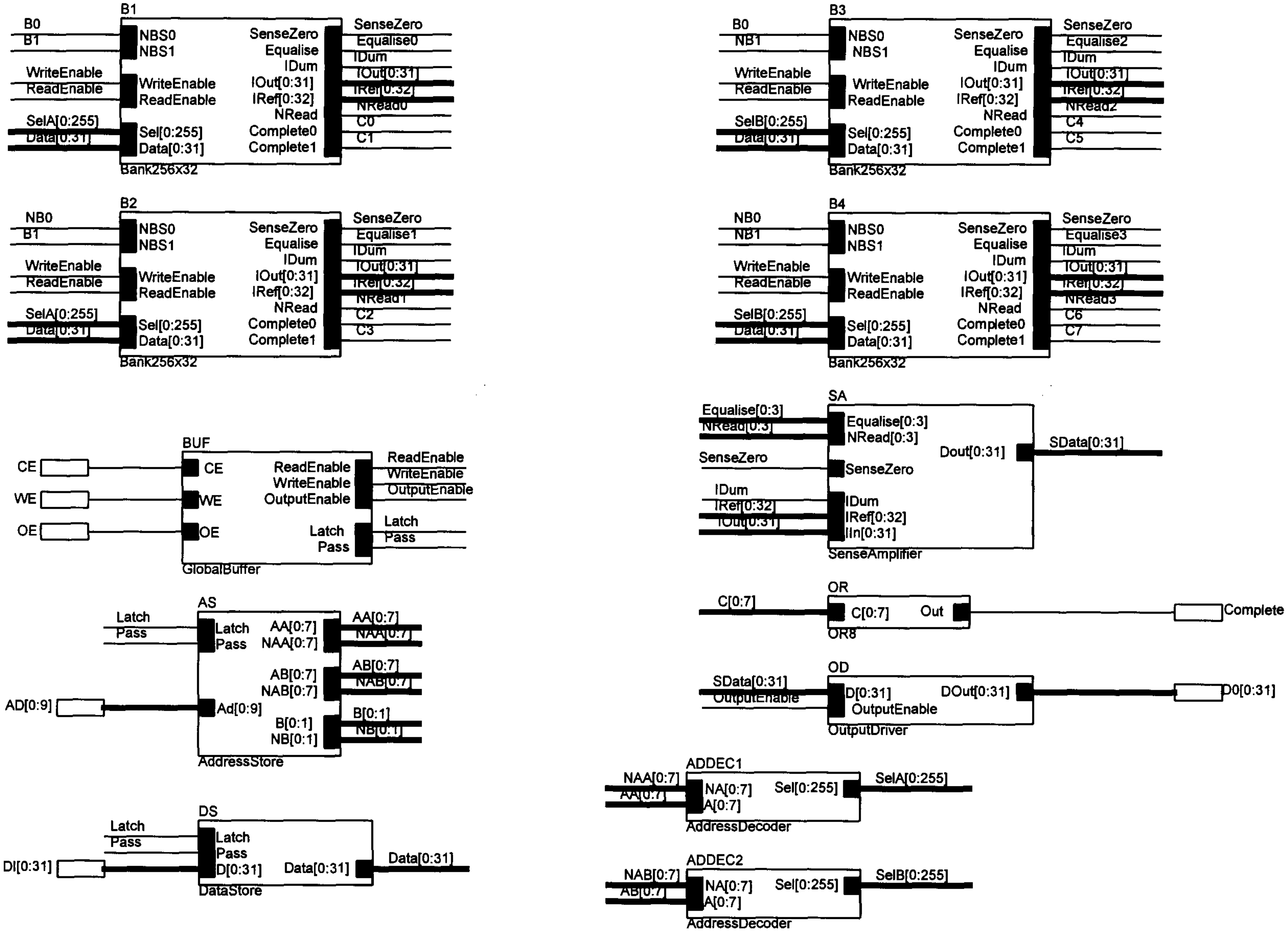


Figure C.1 SRAM System: Top level of the four-transistor SRAM cell system.

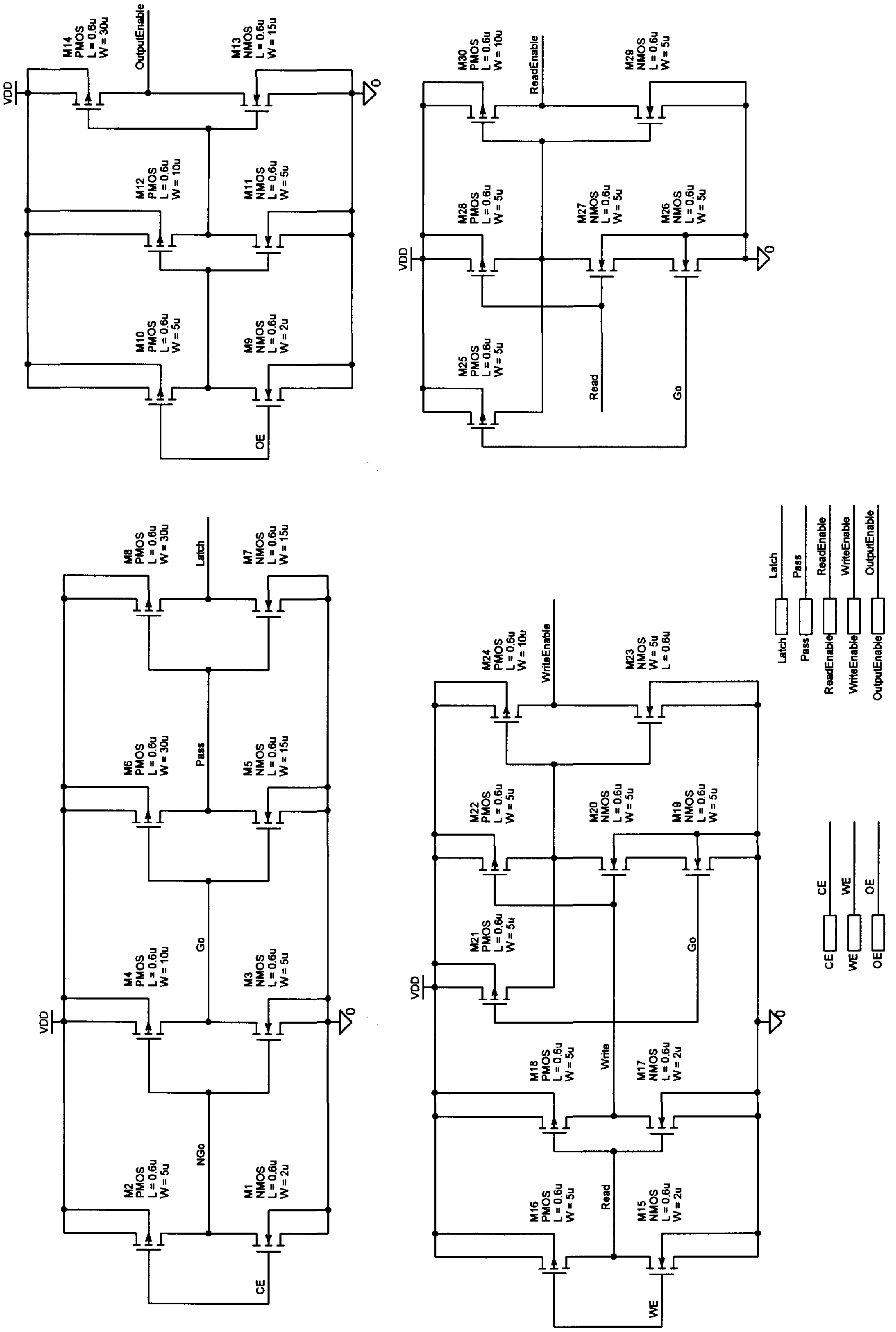


Figure C.2 Global Buffer: Buffering system for the input control signals.

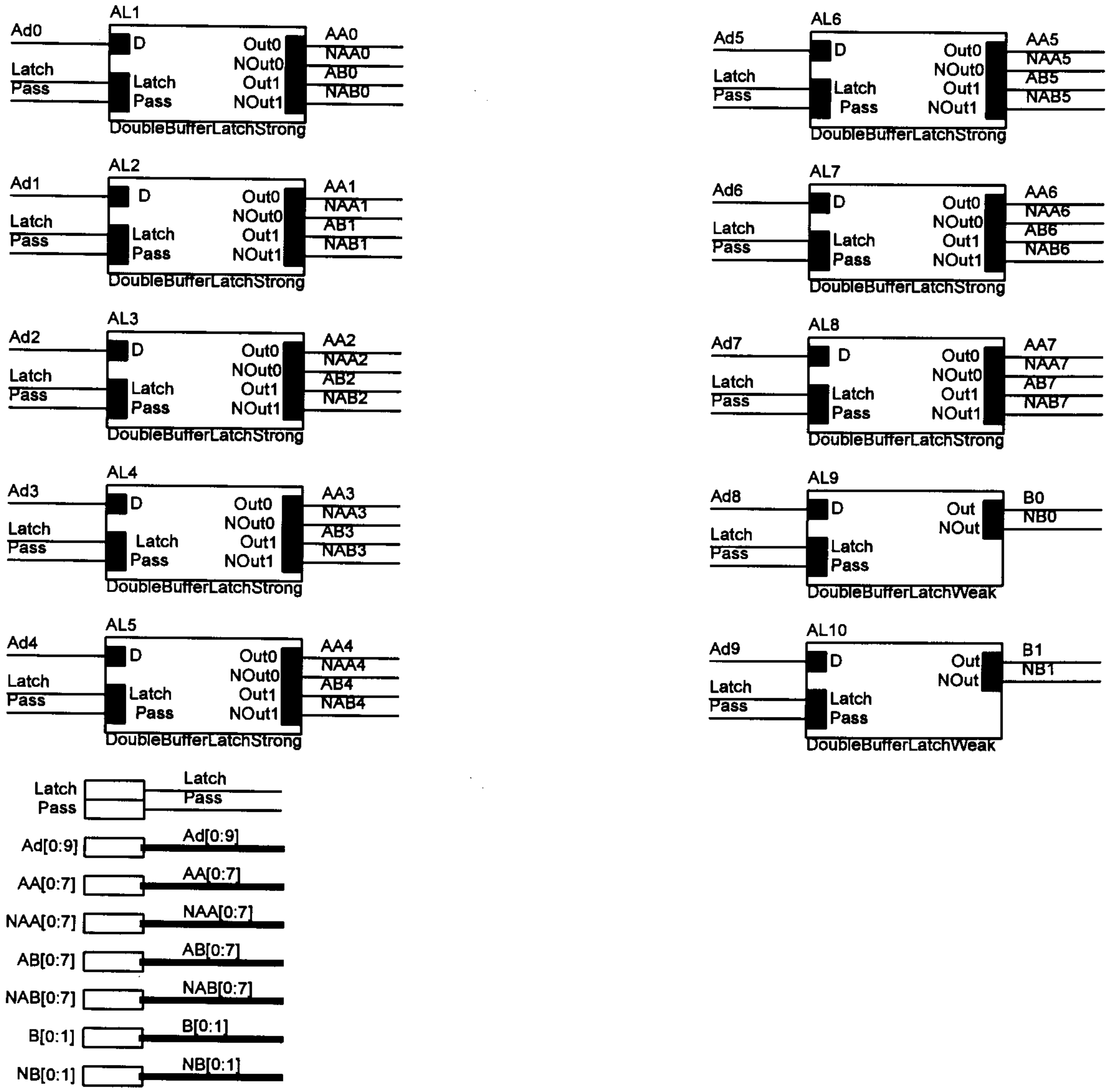


Figure C.3 Address Store: Latch system for the address inputs.

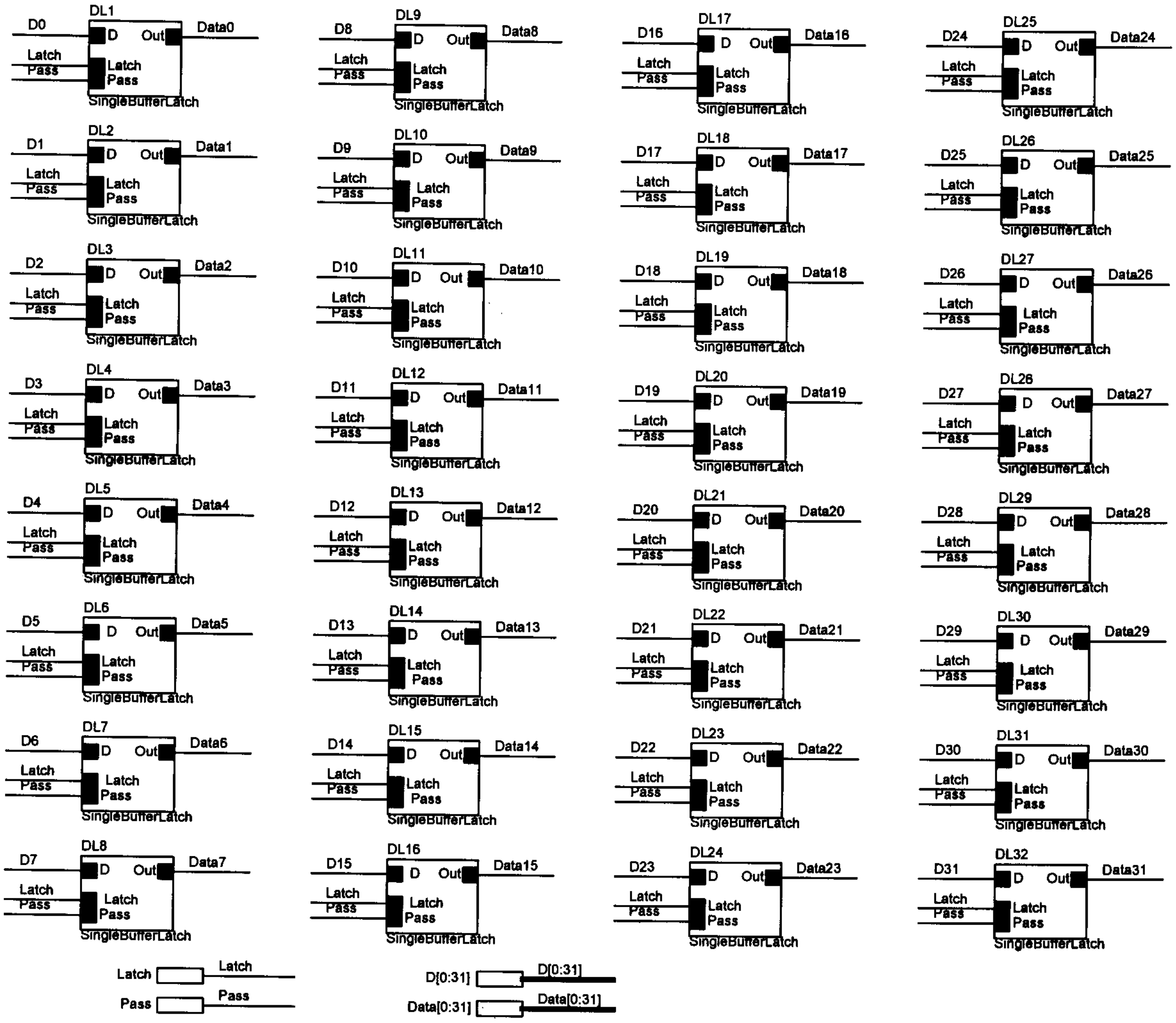


Figure C.4 Data Store: Latch system for the data inputs.

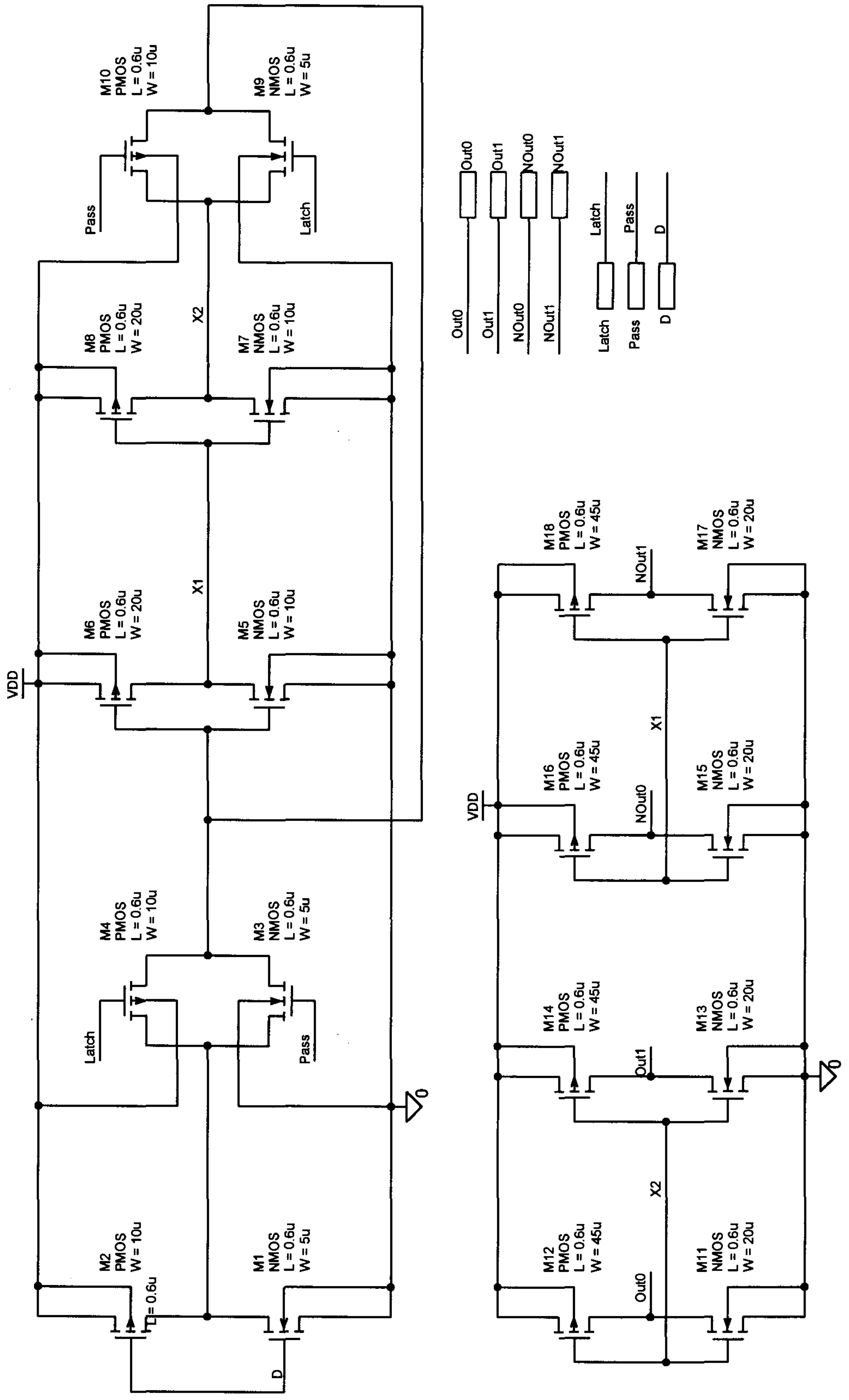


Figure C.5 Double Buffer Latch Strong: Transparent latch with double strong complementary output buffers.

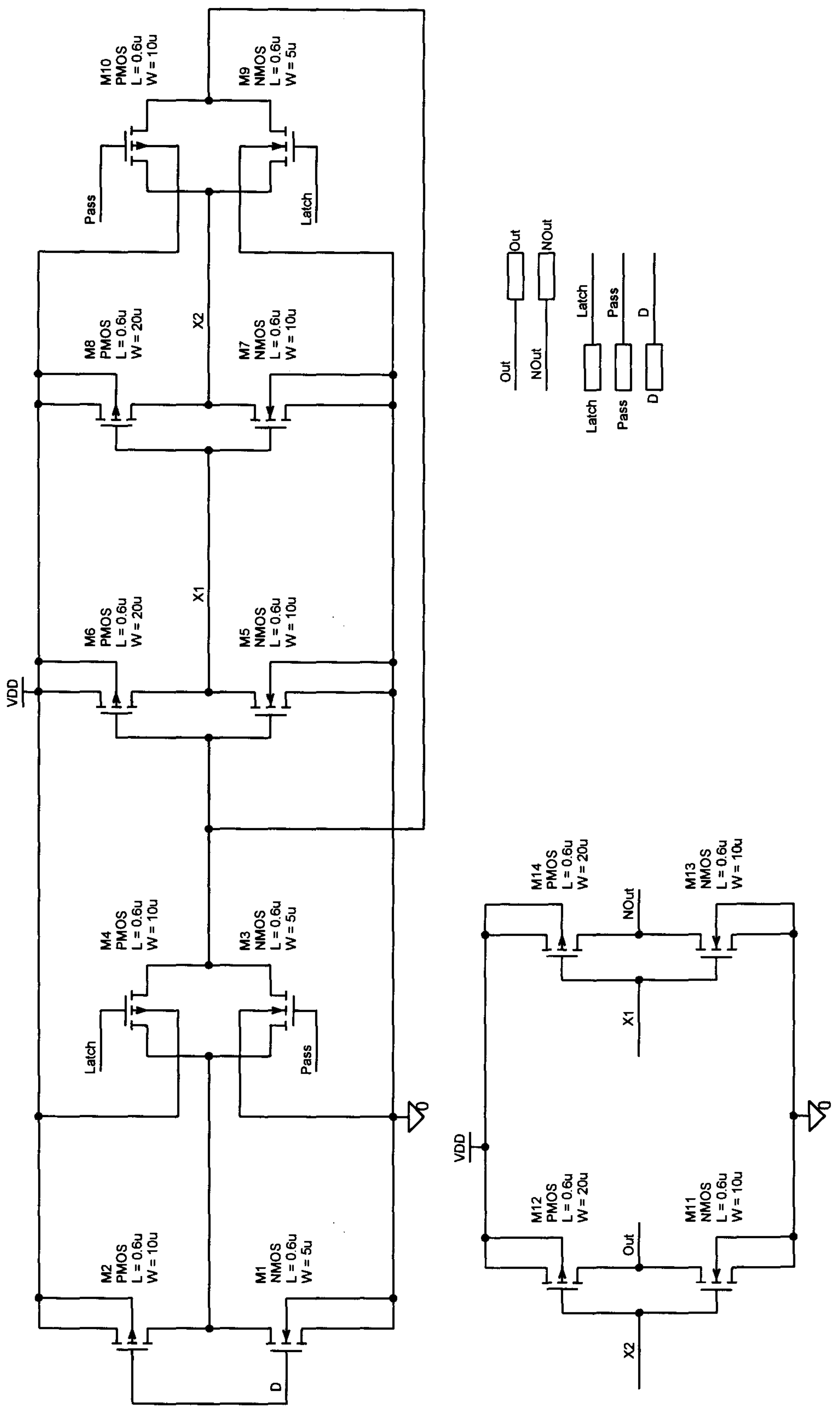


Figure C.6 Double Buffer Latch Weak: Transparent latch with weak complementary output buffers.

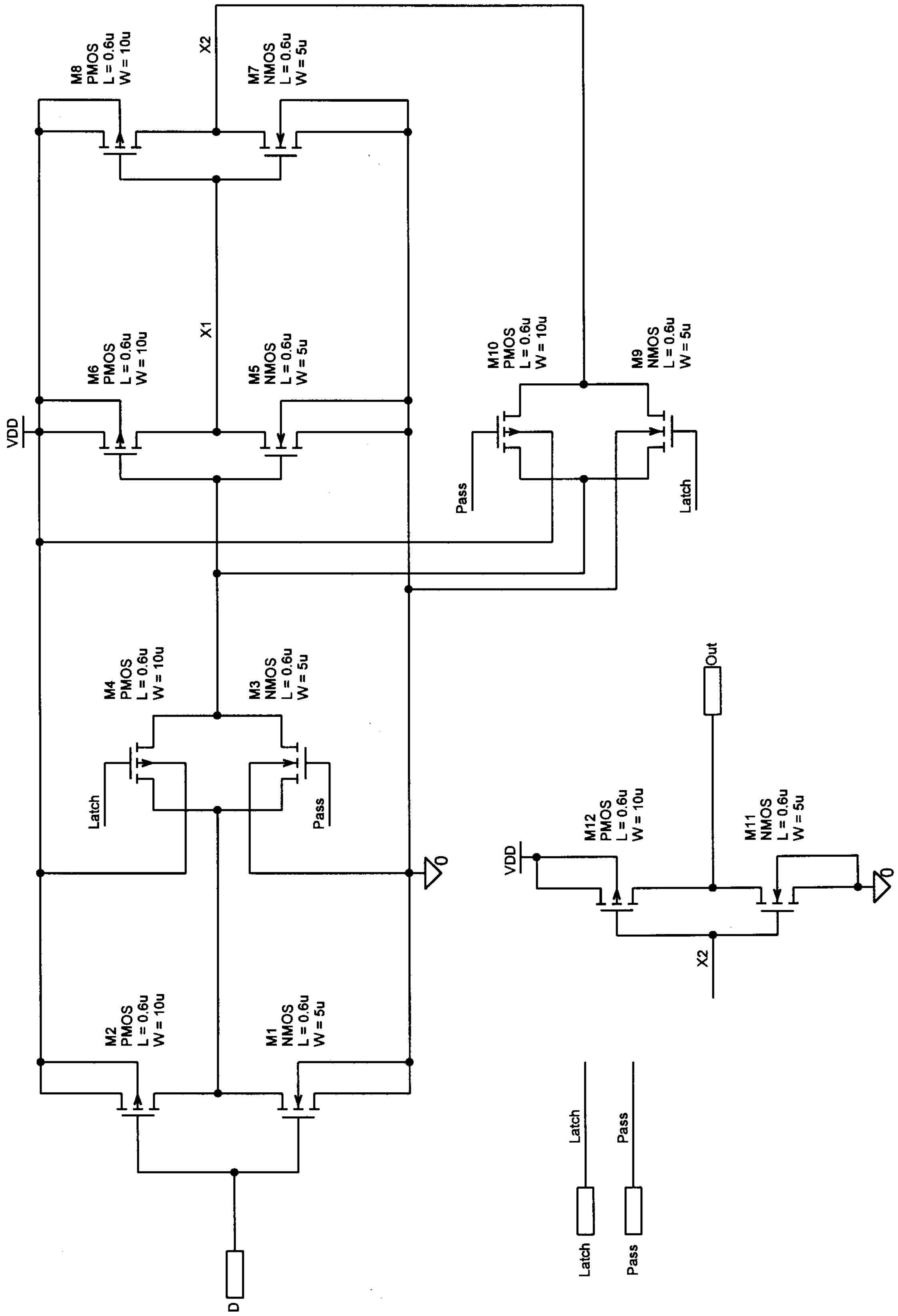


Figure C.7 Single Buffer Latch: Transparent latch with output buffer.

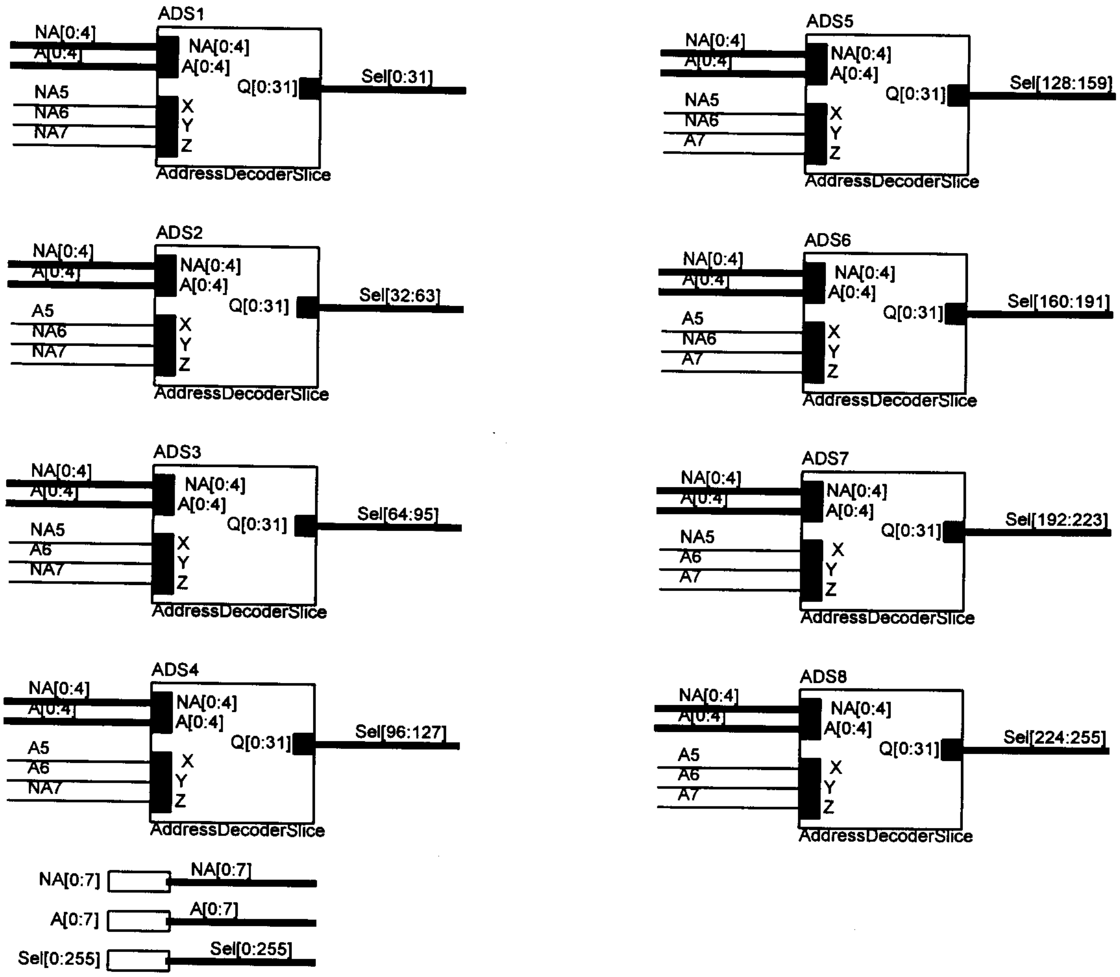


Figure C.8 Address Decoder: 8-256 line address decoder.

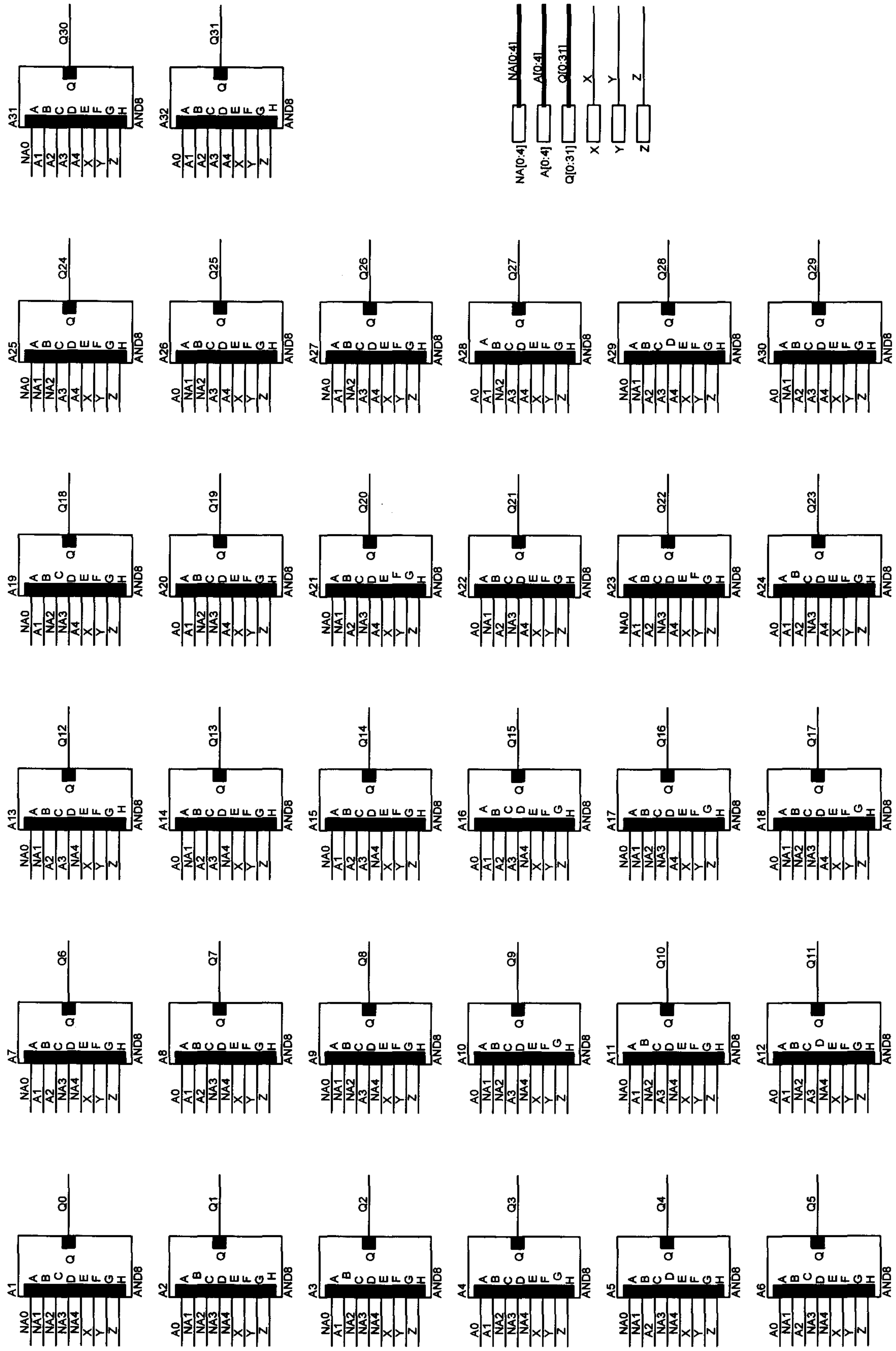


Figure C.9 Address Decoder Slice: 1/8 of the 8 - 256 line address decoder.

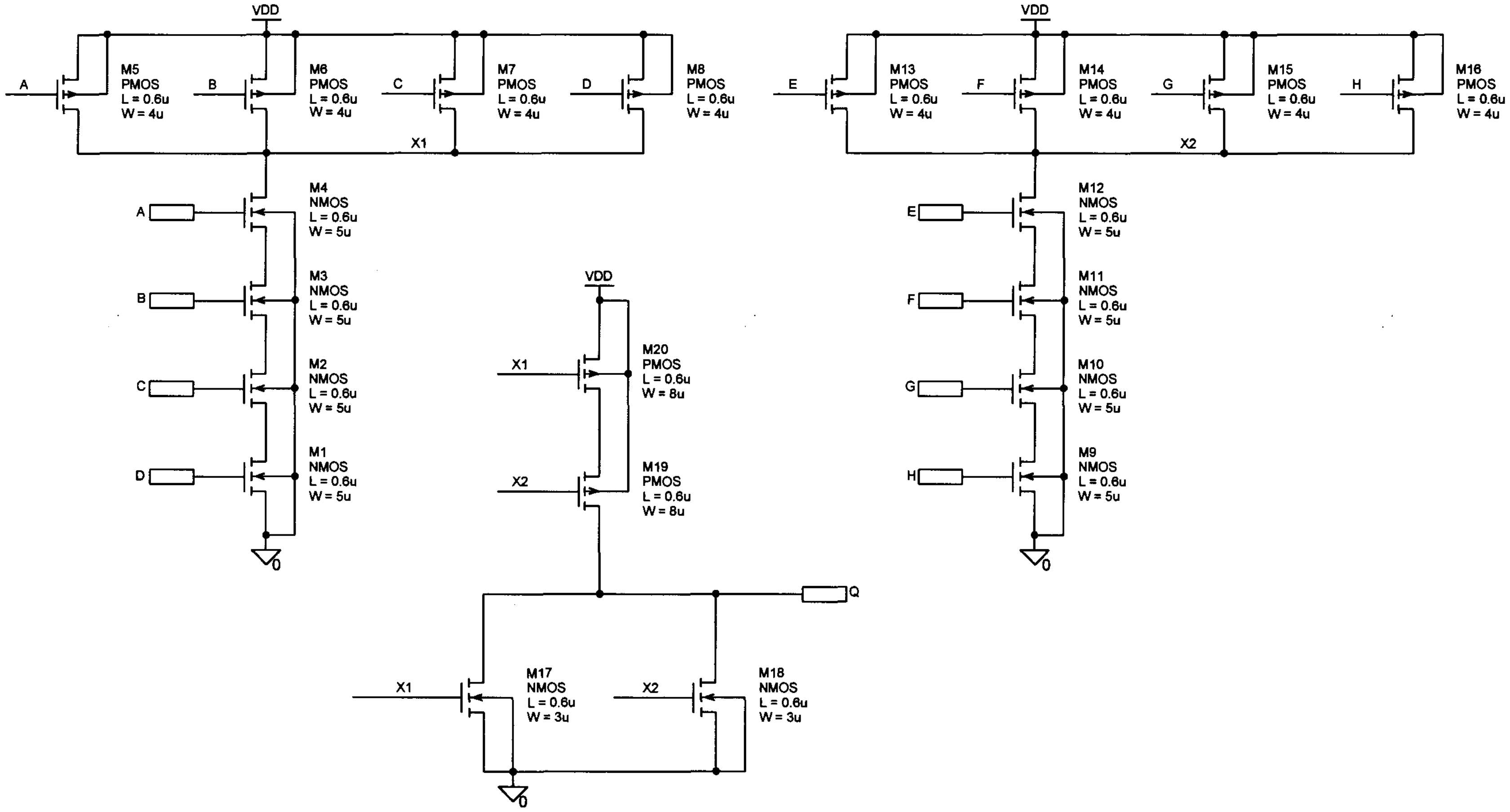


Figure C.10 And8: Eight input And-gate.

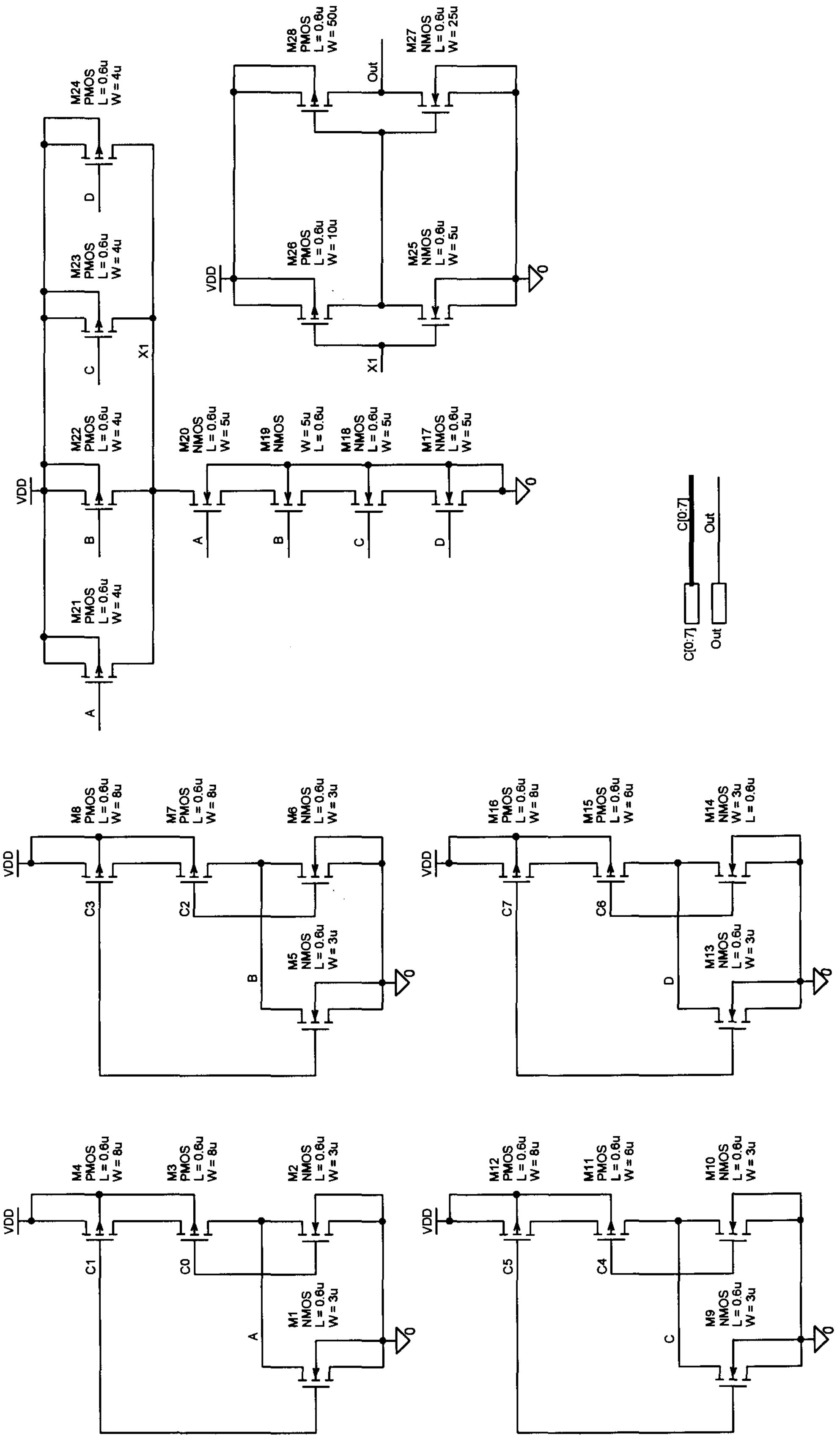


Figure C.11 Or8: Eight input Or-gate with buffered output.

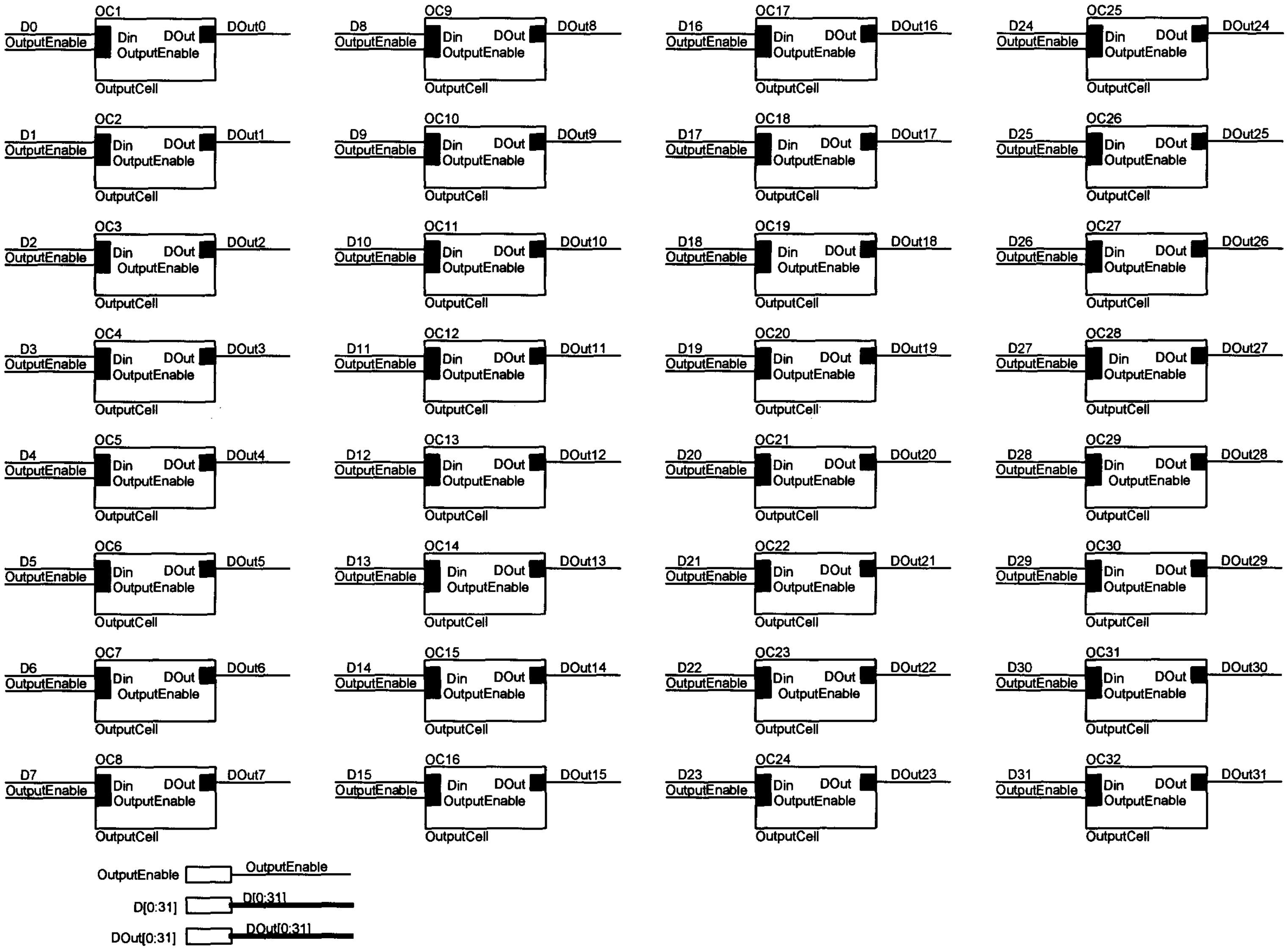


Figure C.12 Output Driver: Tri-state data output driver circuit array.

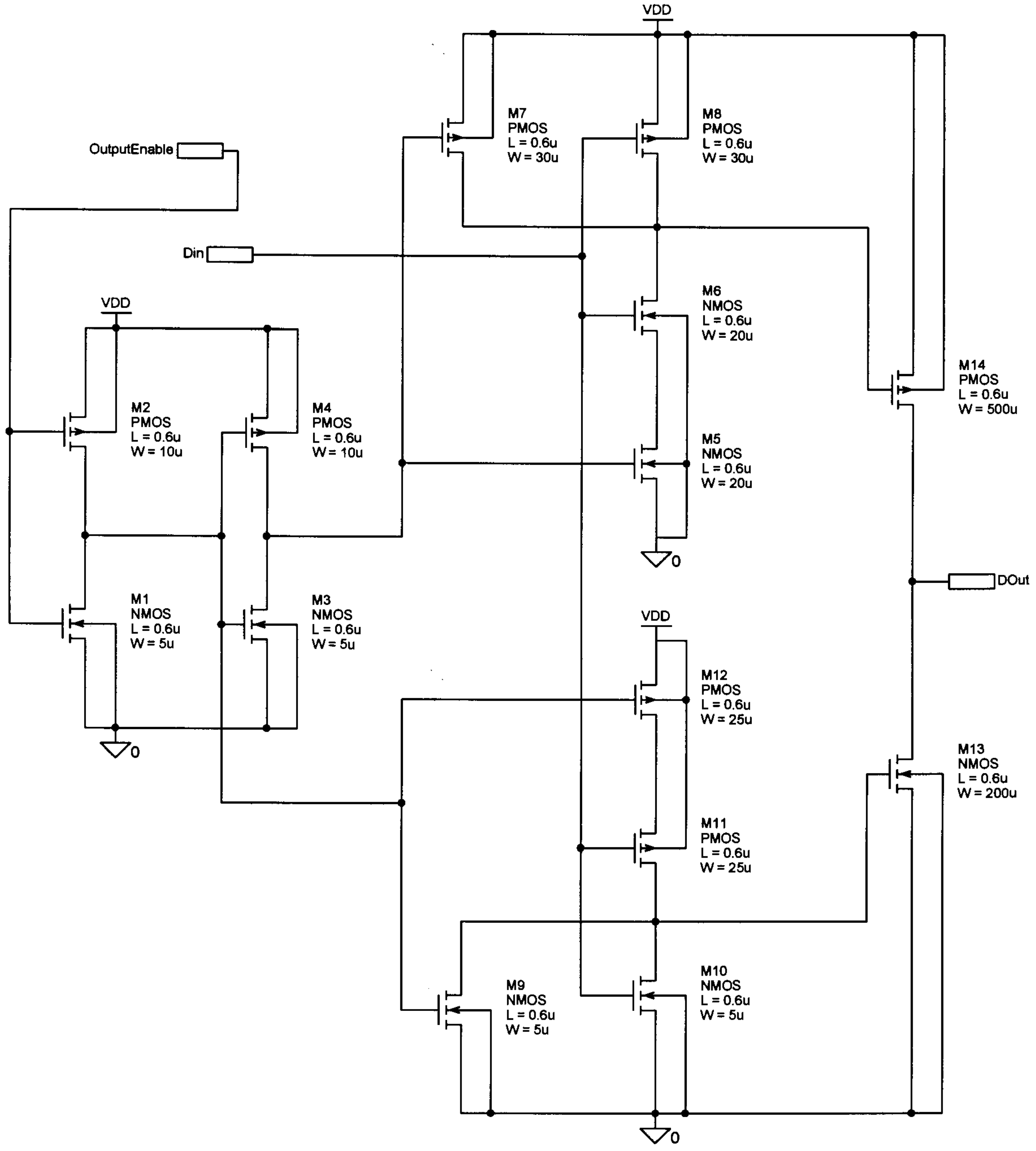


Figure C.13 Output Cell: Tri-state data output driver circuit.

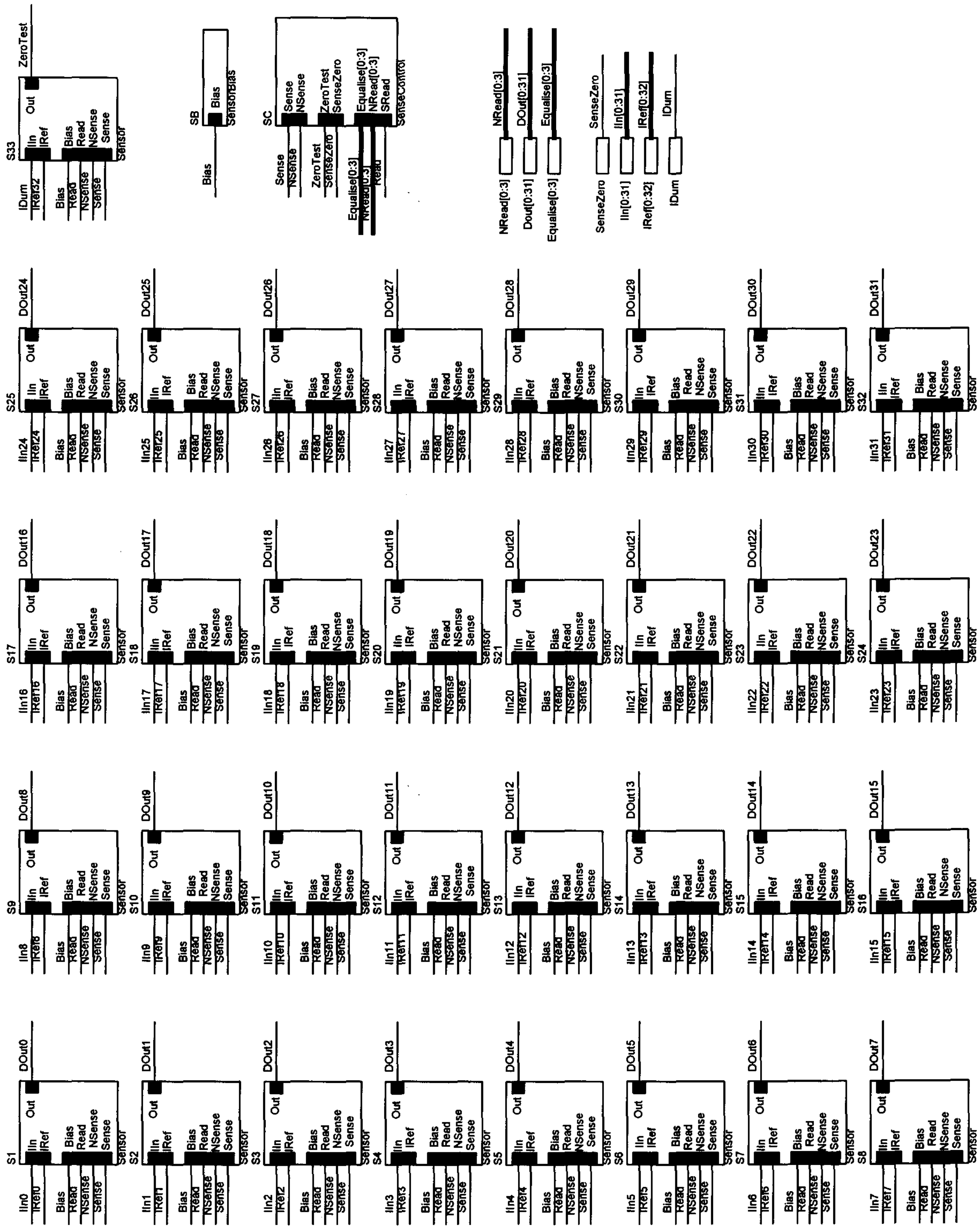


Figure C.14 Sense Amplifier: Complete sense amplifier system.

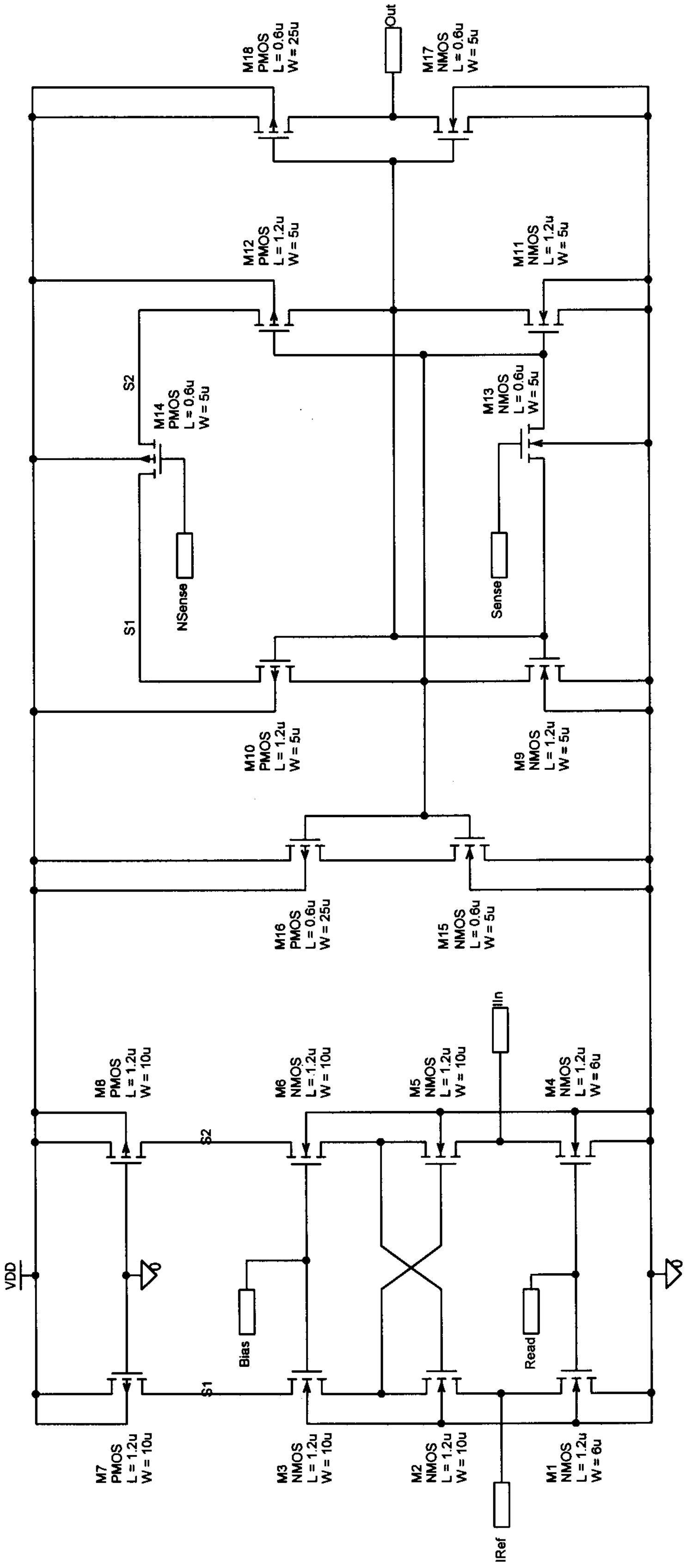


Figure C.15 Sensor: Current conveyor and clamped bit line sense amplifier.

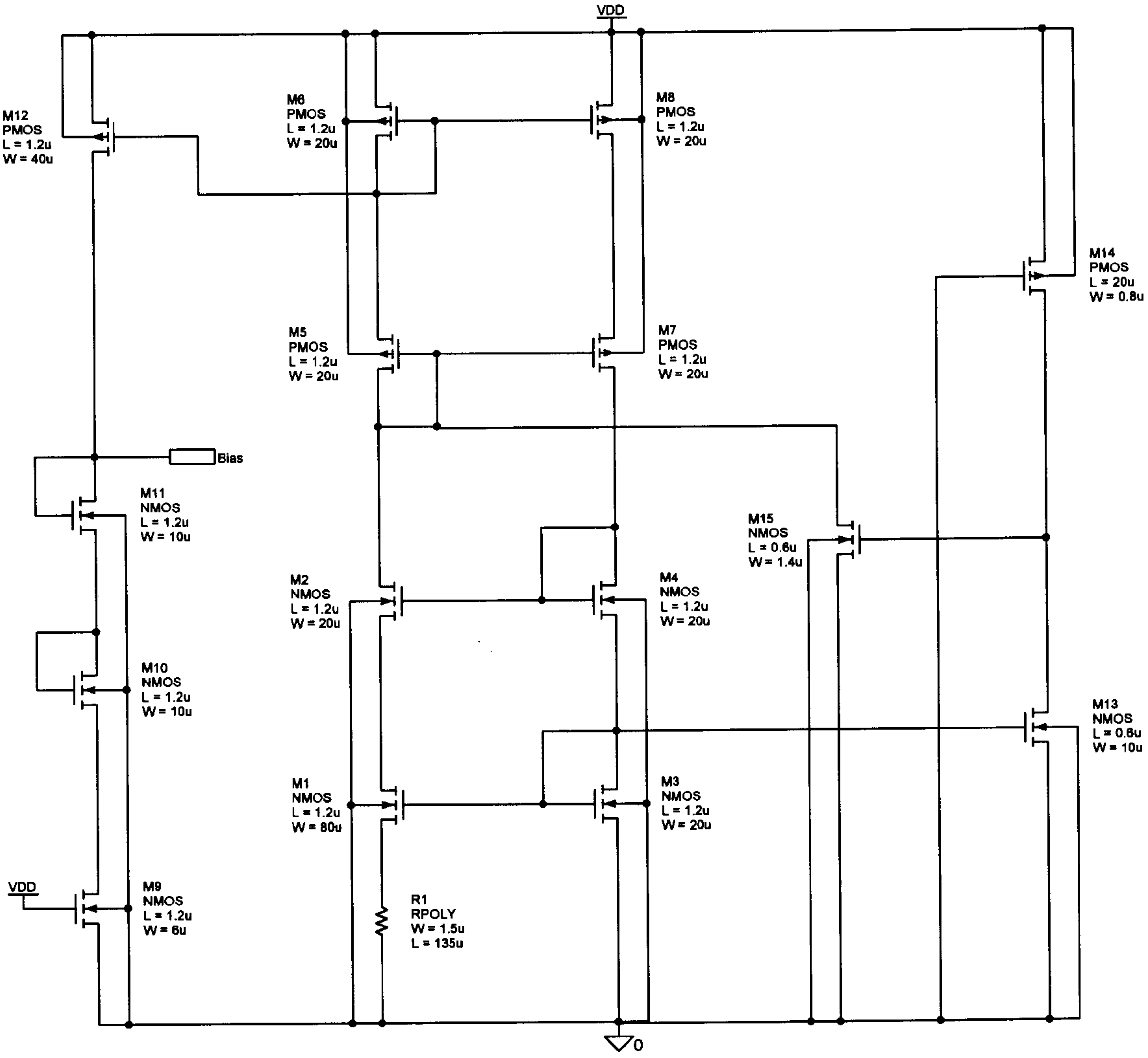


Figure C.16 Sensor Bias: Constant transconductance bias network for the current sense amplifier.

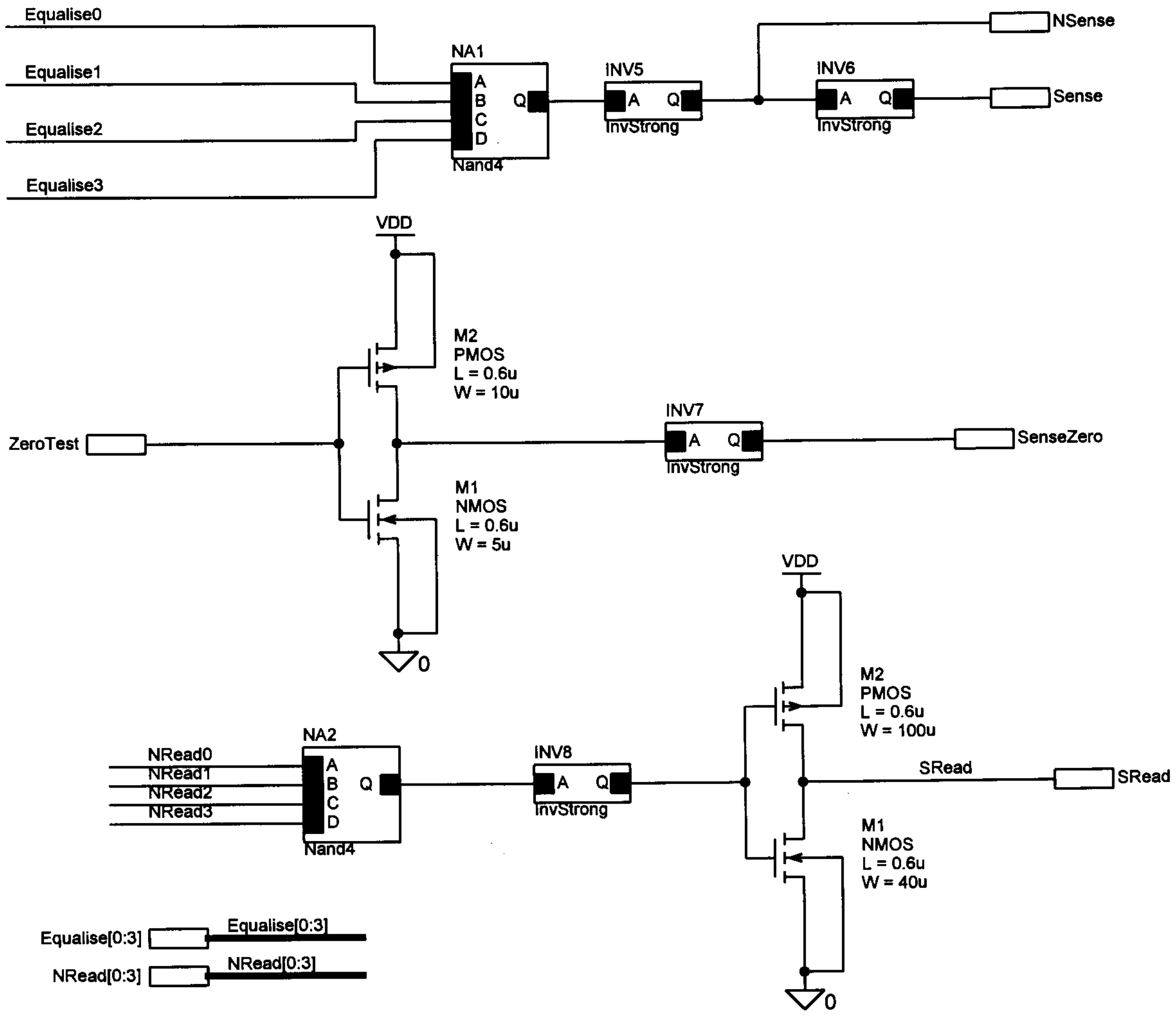


Figure C.17 Sense Control: Sense amplifier peripherals and control circuits.

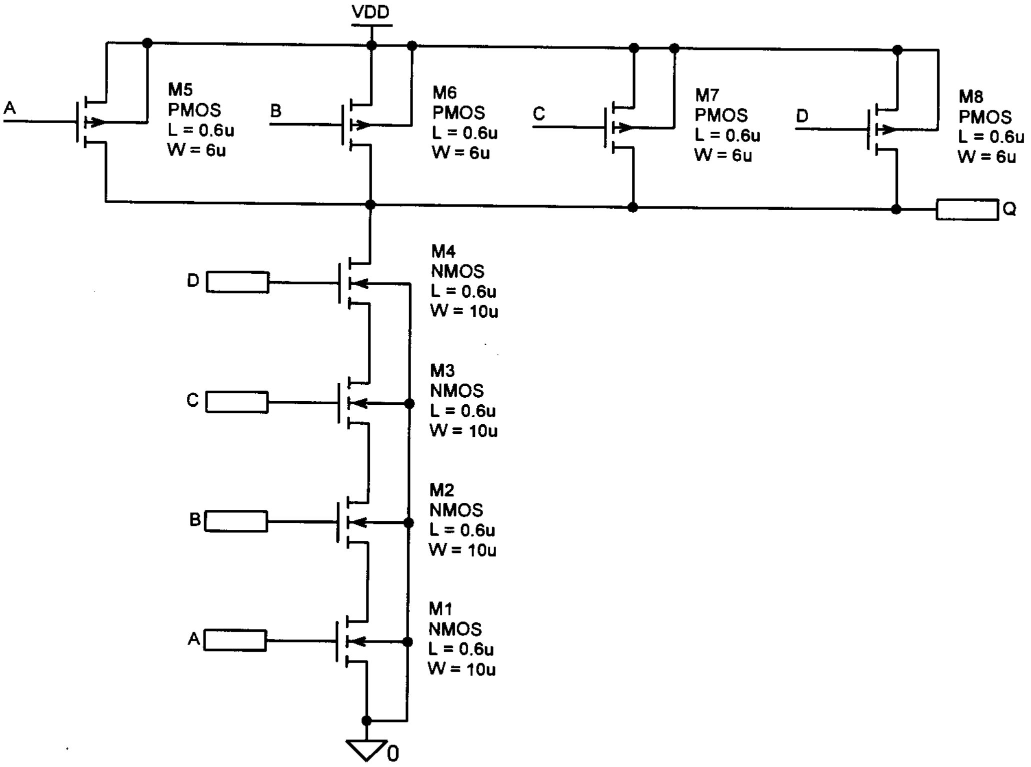


Figure C.18 Nand4: Four-input Nand-gate.

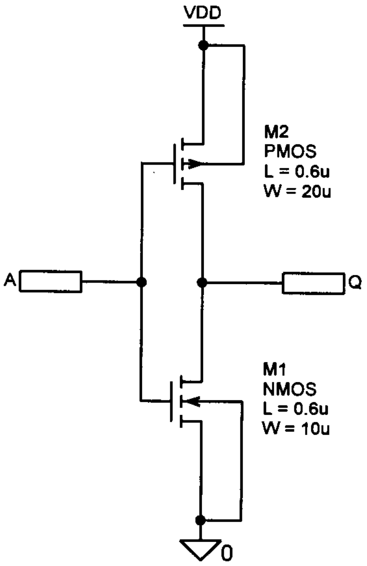


Figure C.19 Inv Strong: High driving strength inverter.

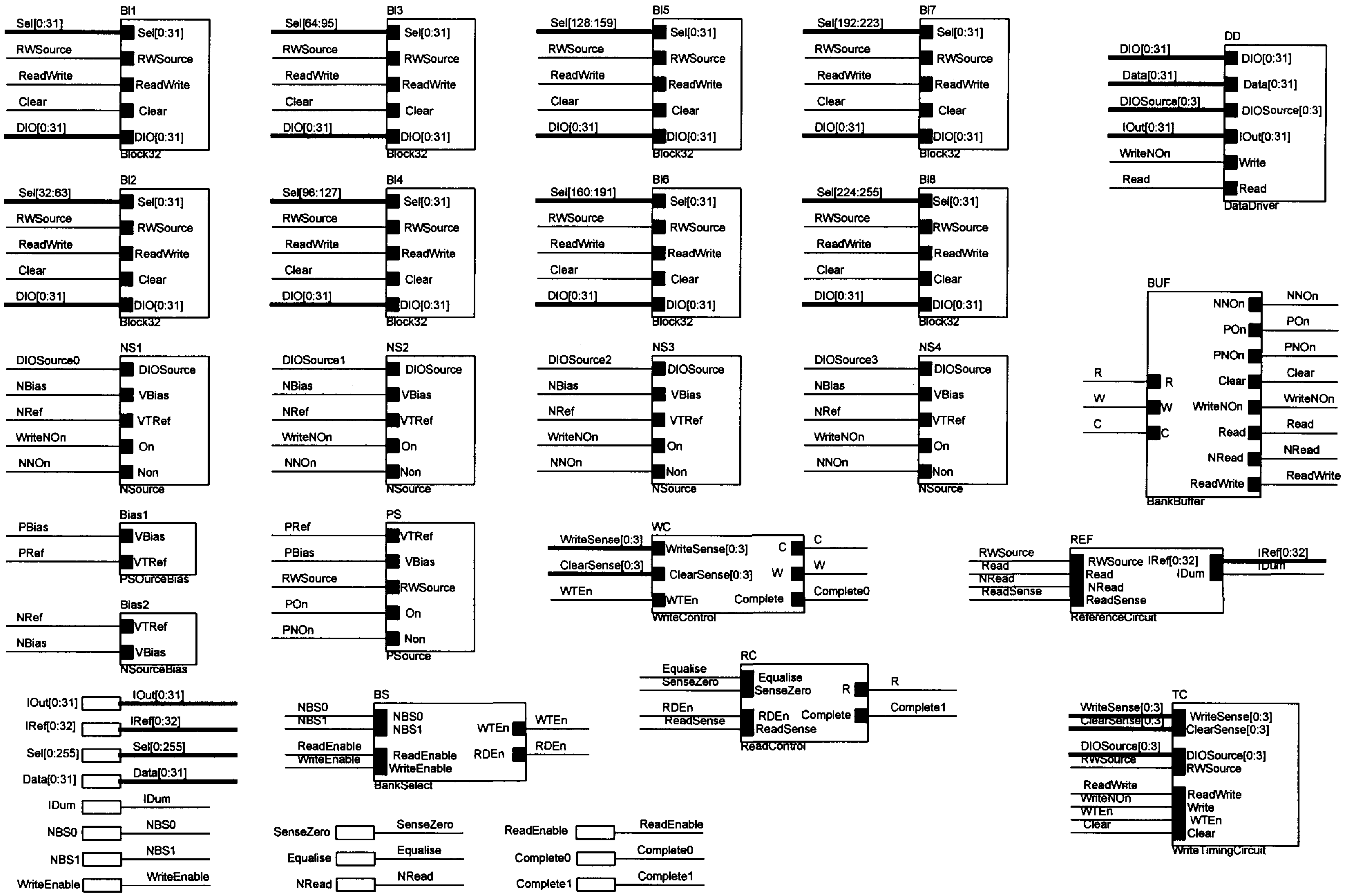


Figure C.20 Bank 256x32: One memory bank with all the associated peripheral circuits.

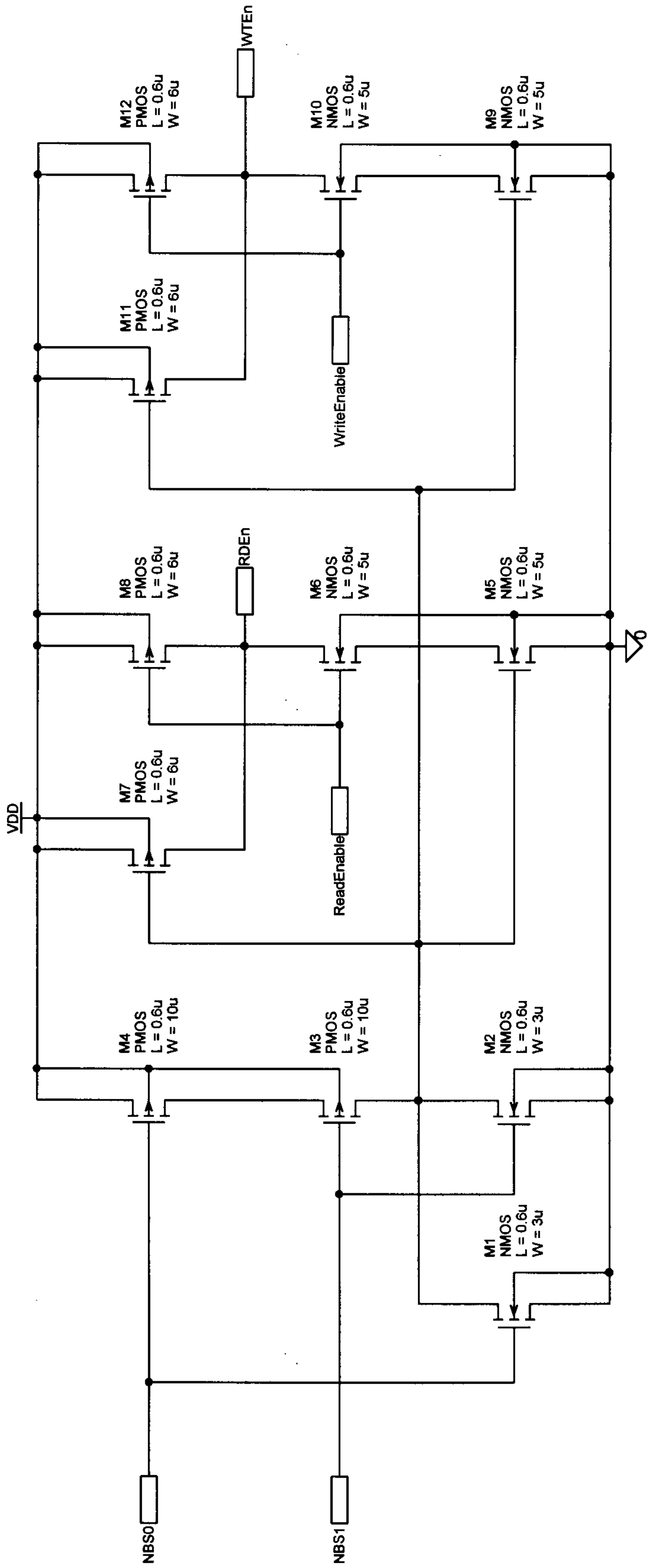


Figure C.21 Bank Select: Bank selection via control signal masking.

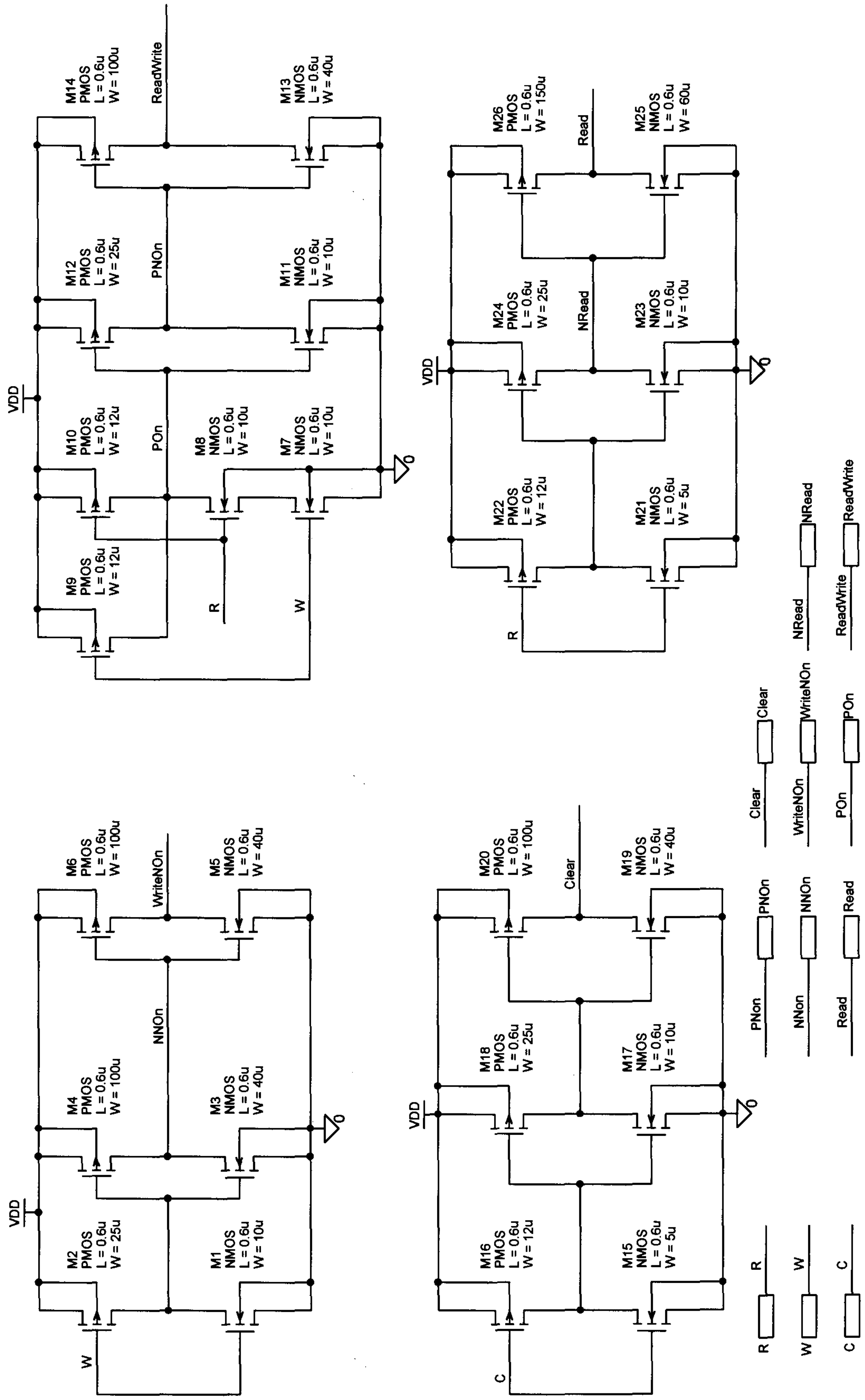


Figure C.22 Bank Buffer: Buffering system for the outputs of the control circuits.

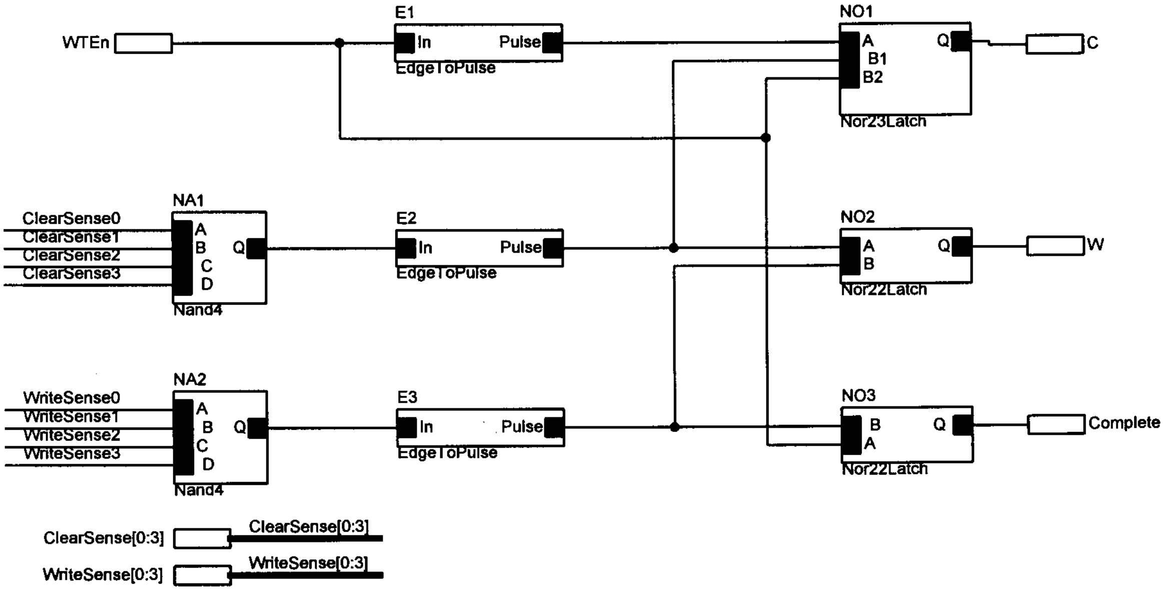


Figure C.23 Write Control: Write cycle control signal sequencer.

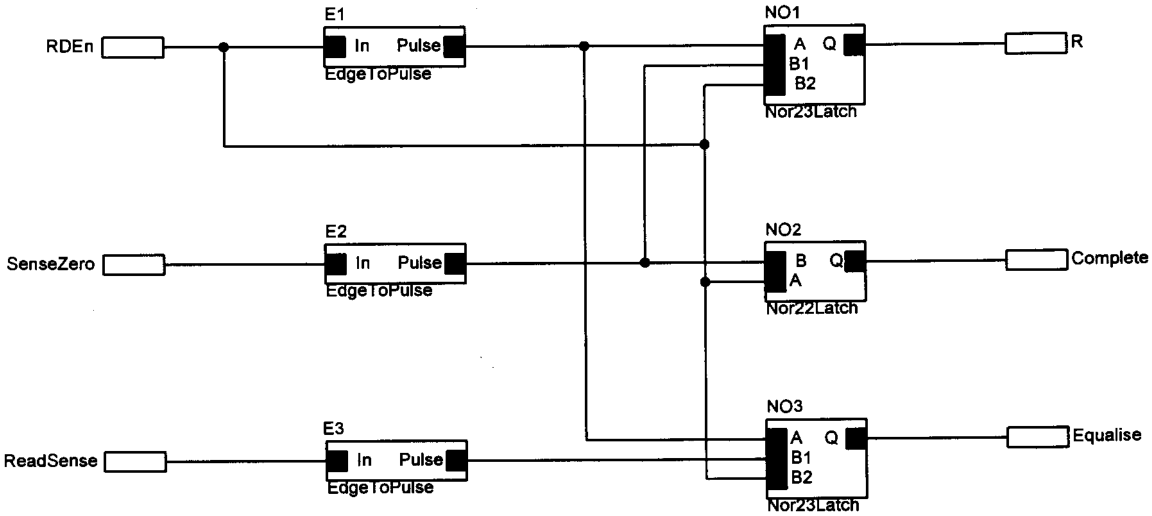


Figure C.24 Read Control: Read cycle control signal sequencer.

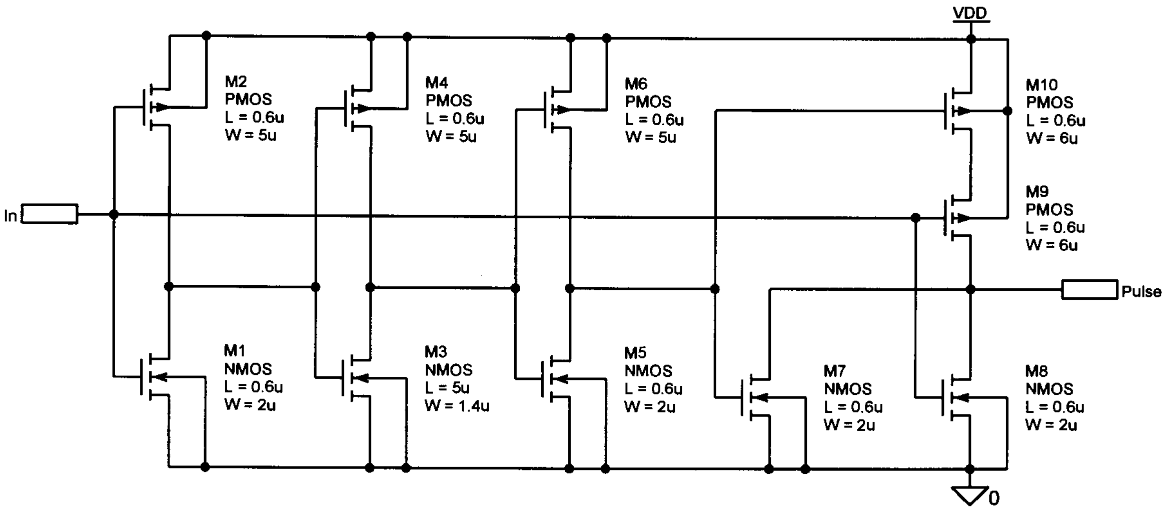


Figure C.25 Edge To Pulse: Falling edge to positive pulse converter.

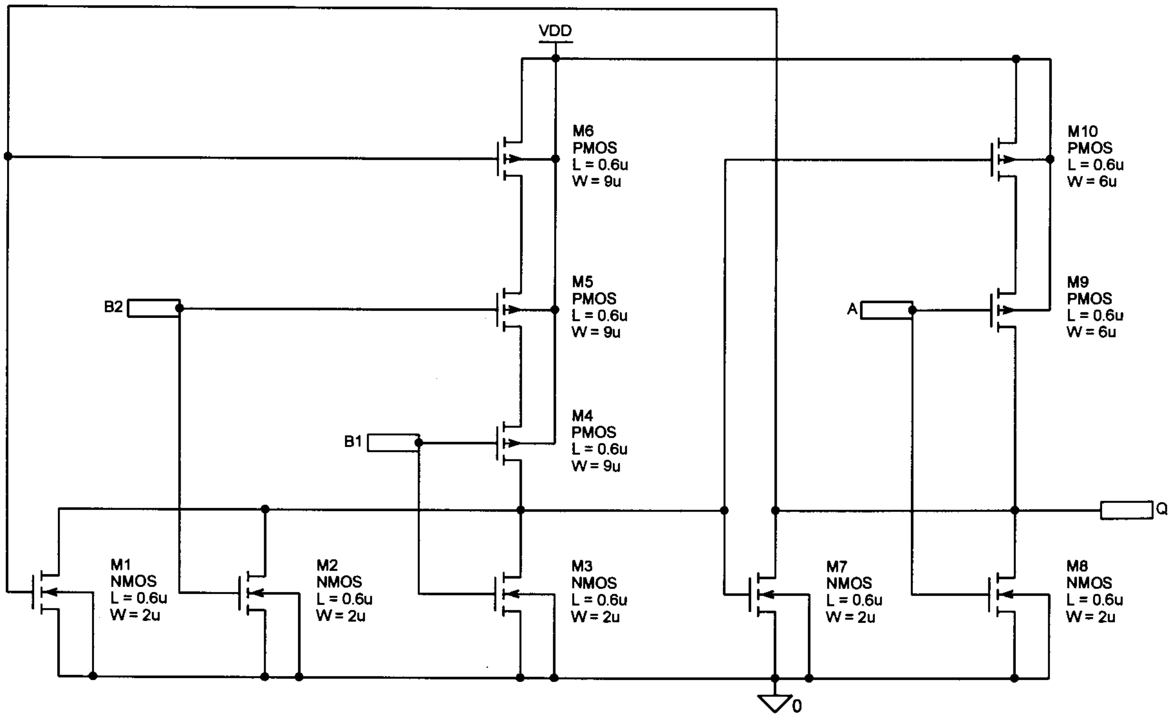


Figure C.26 Nor23 Latch: Set-reset latch with one set and two reset inputs.

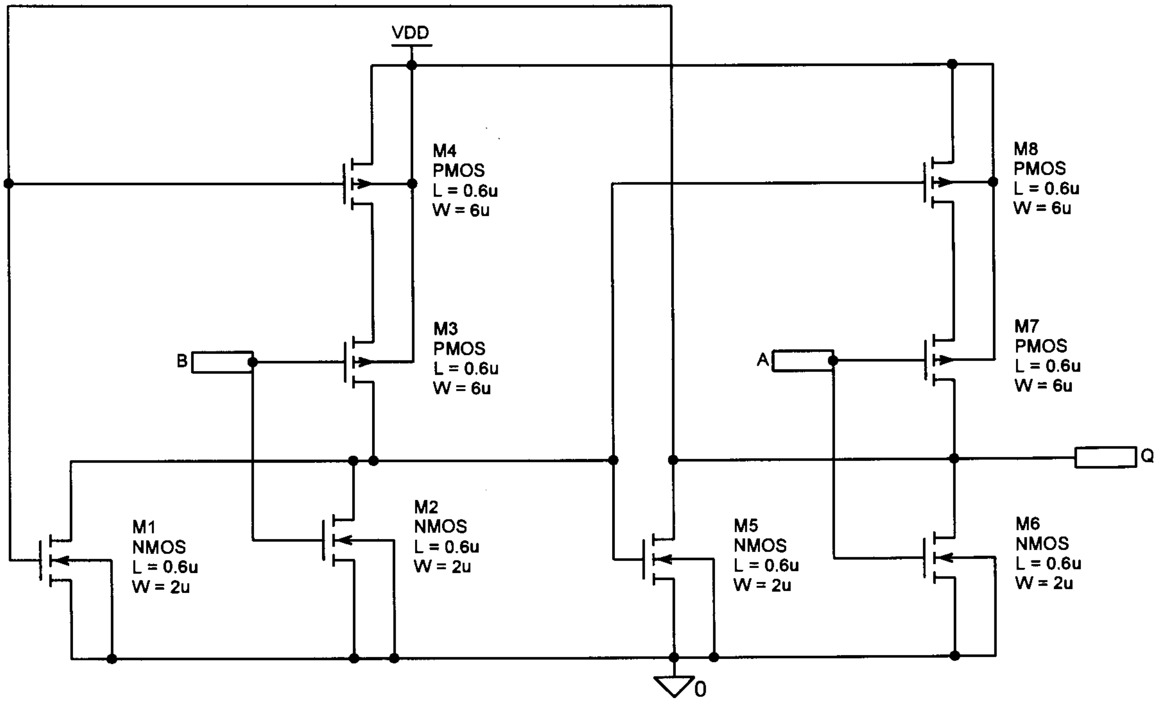


Figure C.27 Nor22 Latch: Set-reset latch with one set and one reset input

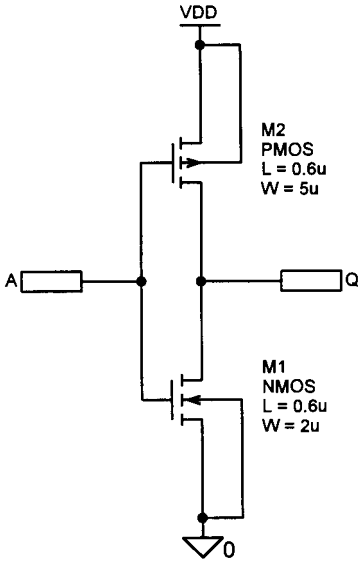


Figure C.28 Inv Weak: Low driving strength inverter.

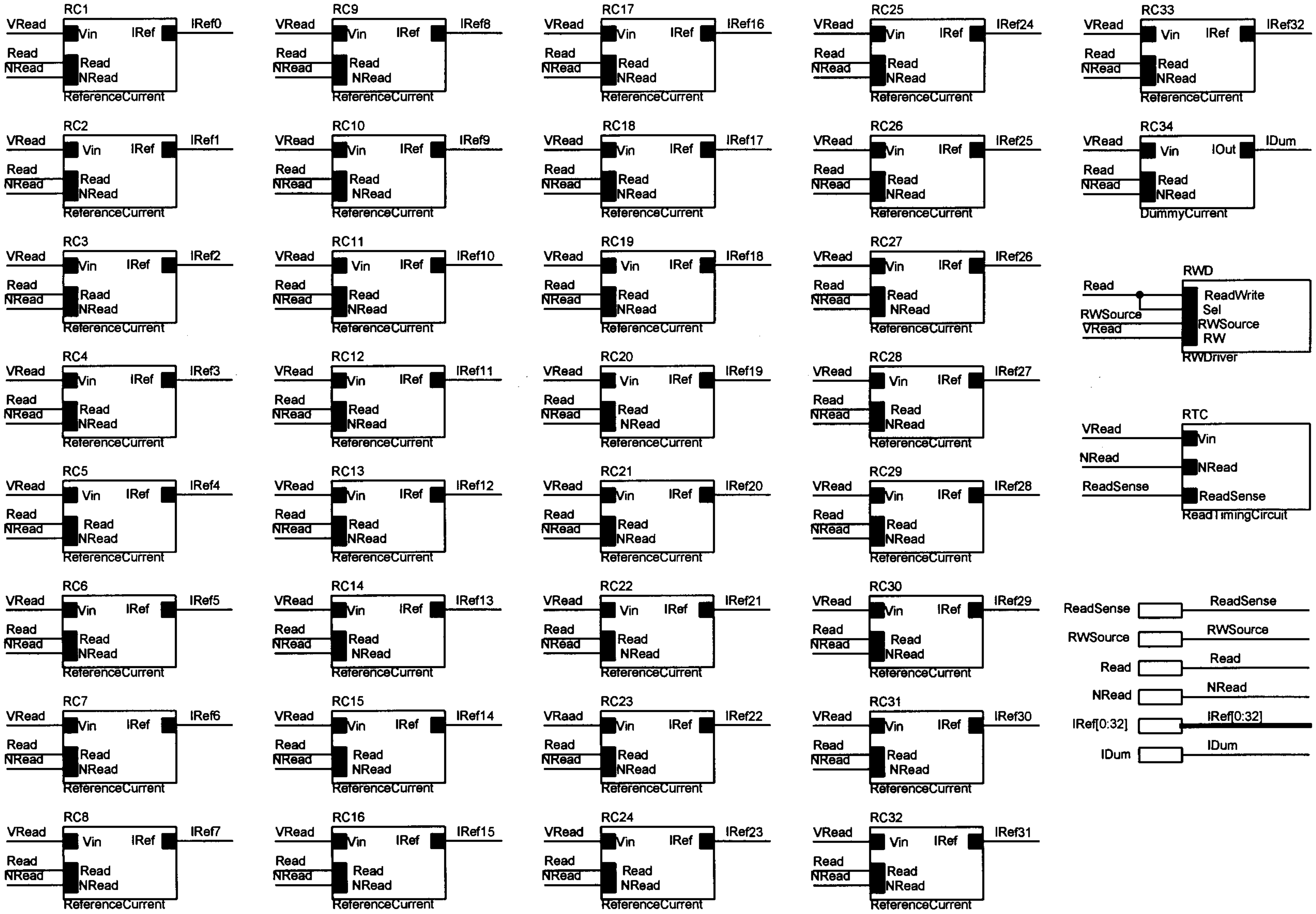


Figure C.29 Reference Circuit: Array of reference current generators with driving circuit.

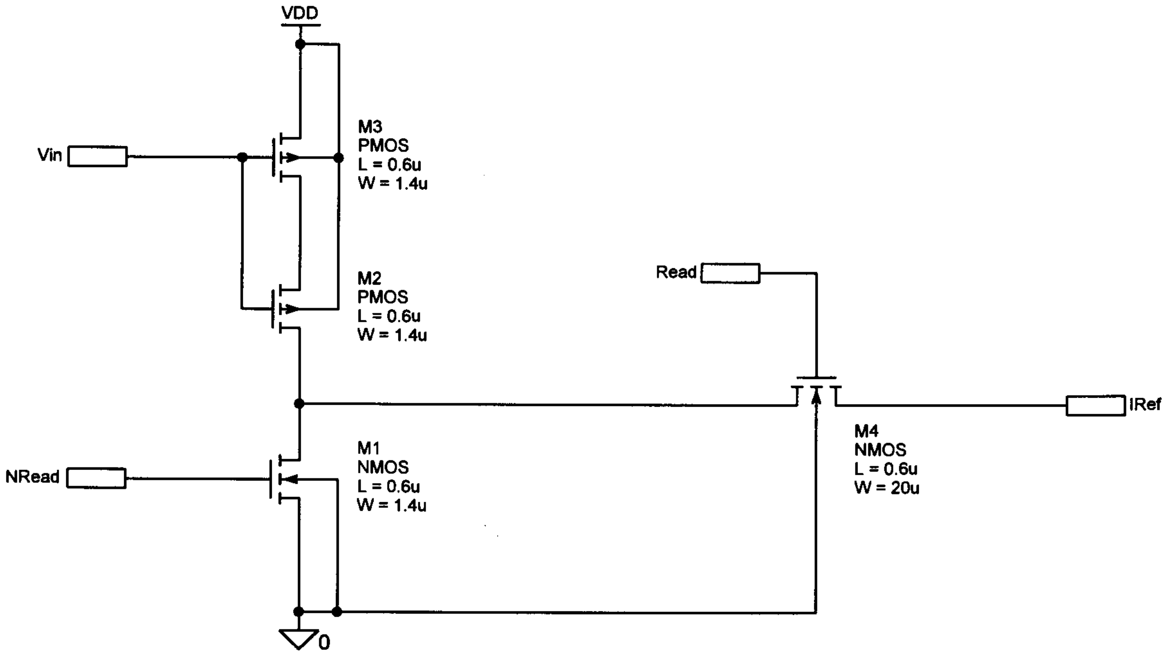


Figure C.30 Reference Current: Reference current generator with current mode multiplexing access device added.

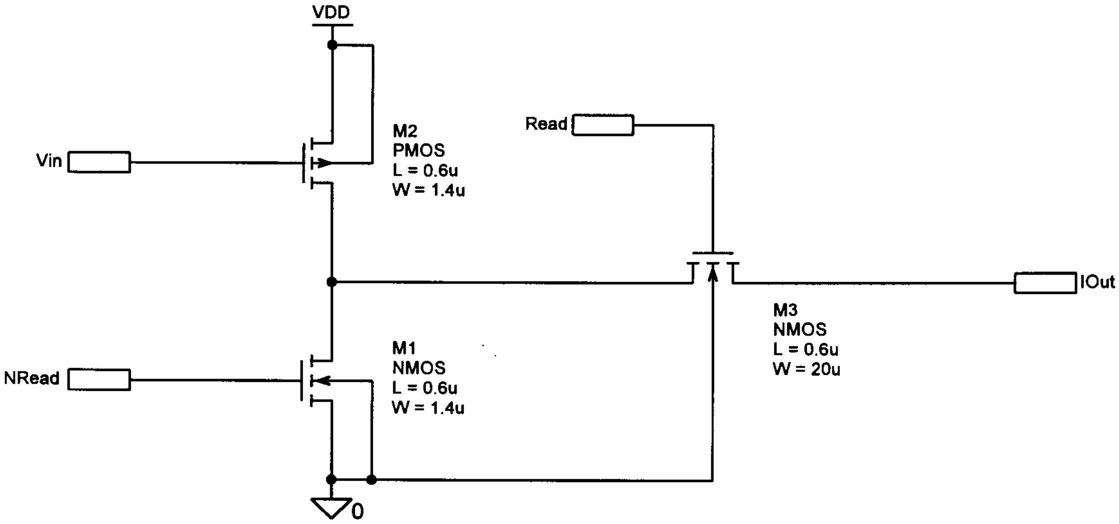


Figure C.31 Dummy Current: Dummy read current generator with current mode multiplexing access device added.

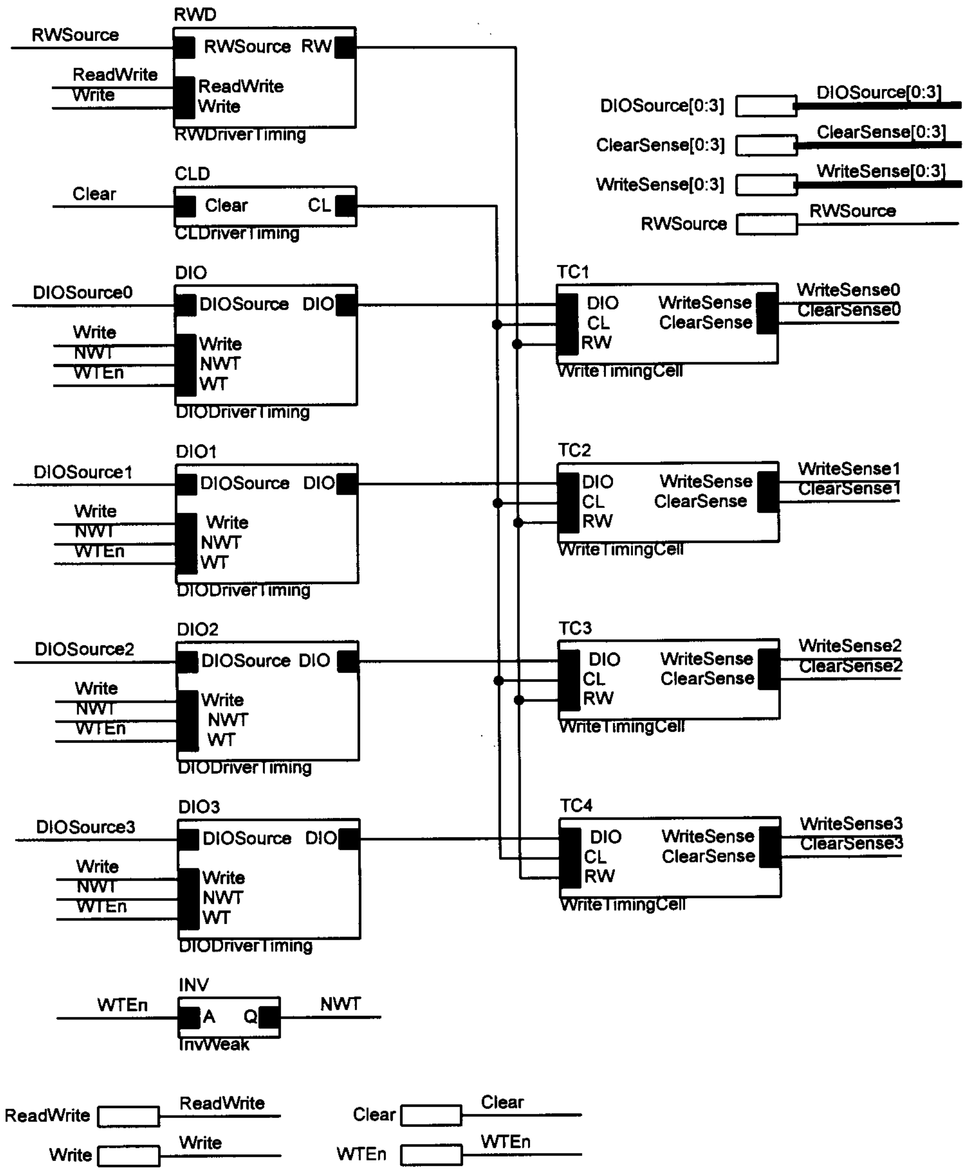


Figure C.32 Write Timing Circuit: Circuit used for sensing the timing of the write cycle.

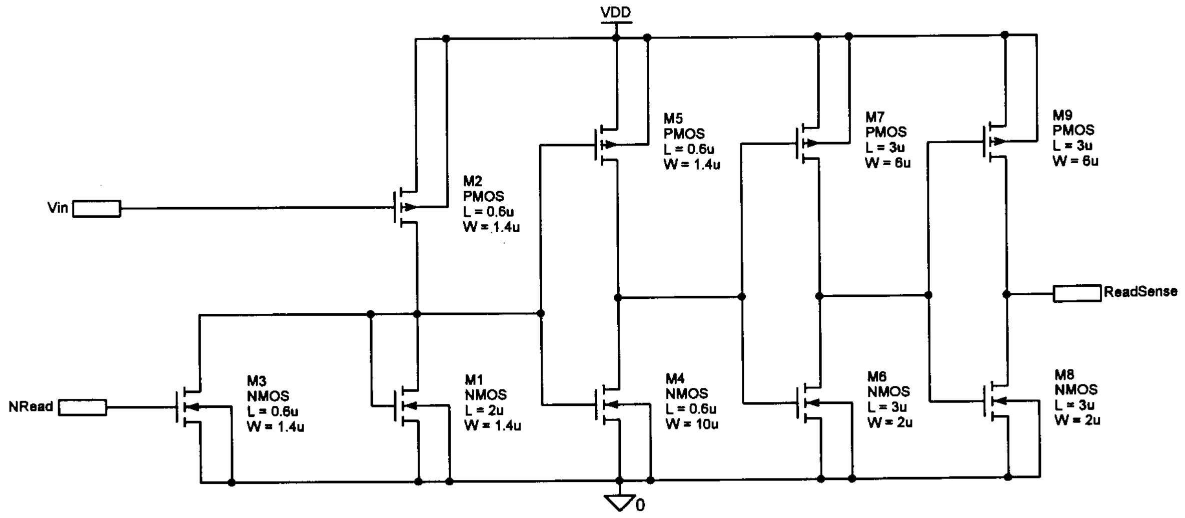


Figure C.33 Read Timing Circuit: Circuit to sense the completion of the initial read cycle phase.

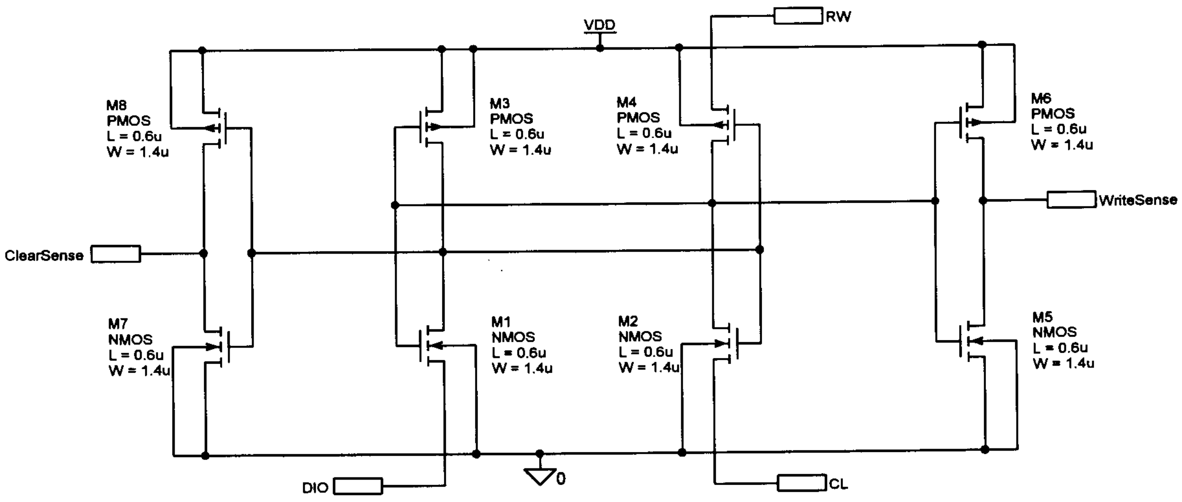


Figure C.34 Write Timing Cell: Dummy cell with access inverters to sense the write phases.

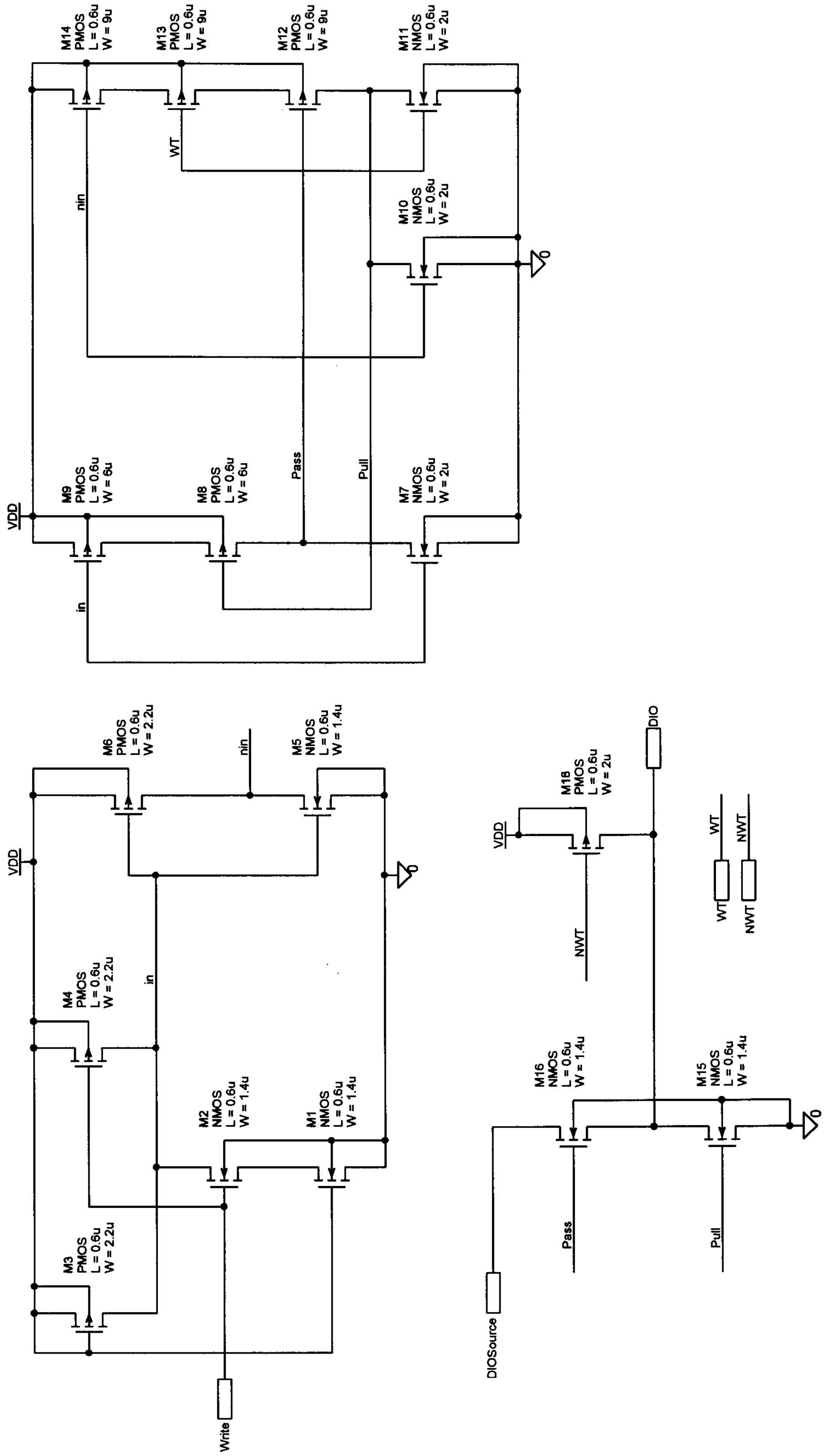


Figure C.35 DIO Driver Timing: *D*/*O*-line driver for the write cycle timing cells.

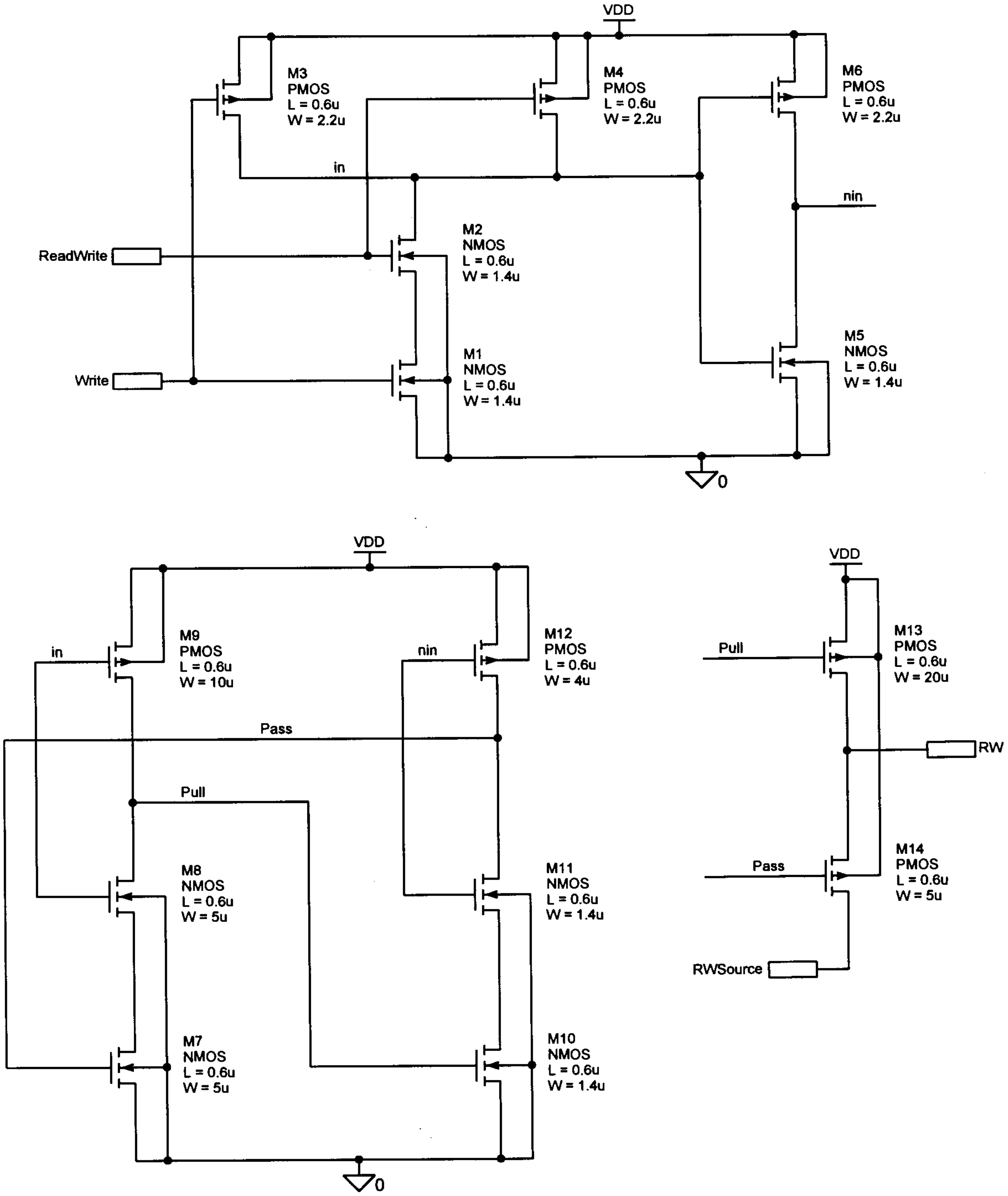


Figure C.36 RW Driver Timing: *RW*-line driver for the write cycle timing cells.

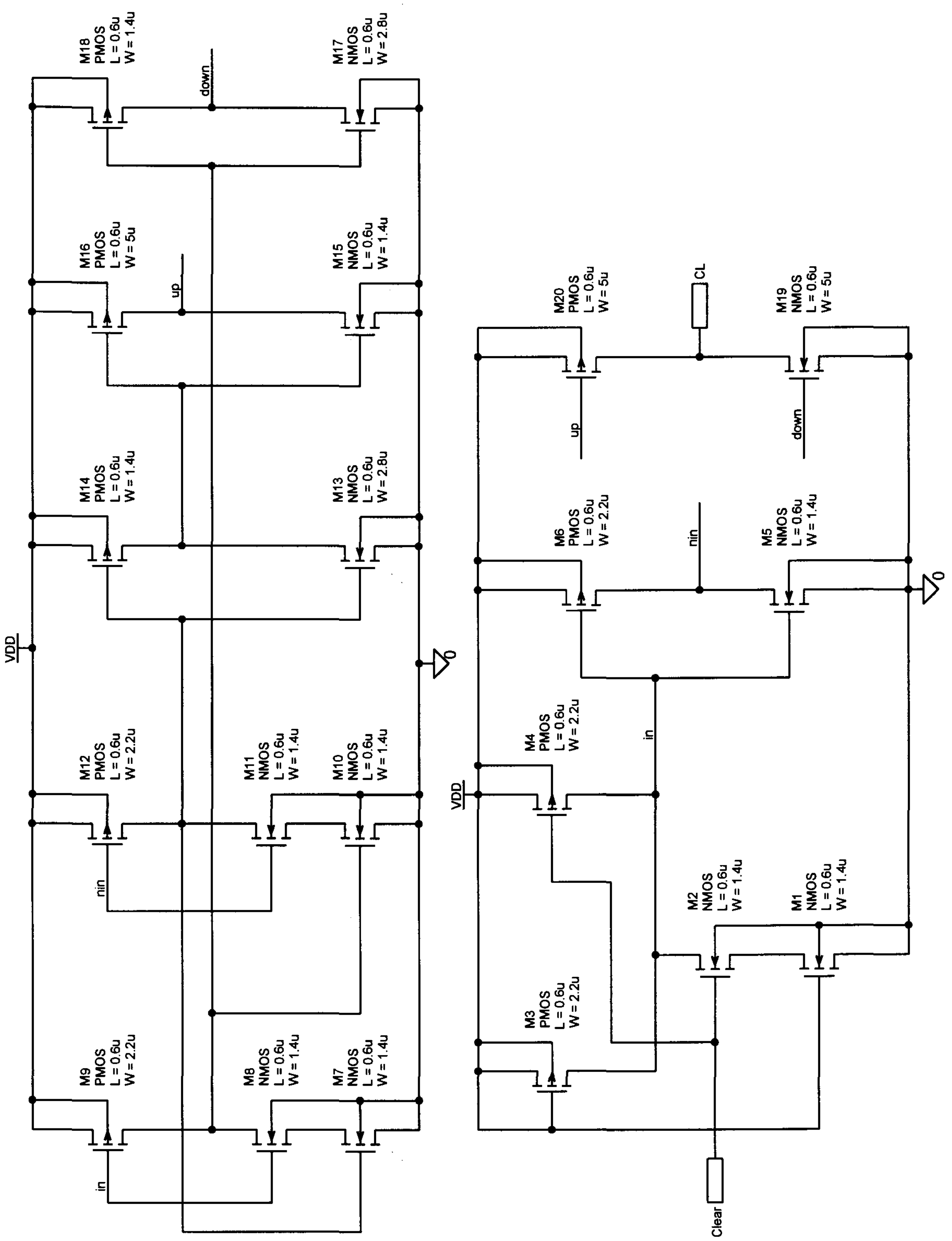


Figure C.37 CL Driver Timing: *CL*-line driver for the write cycle timing cells.

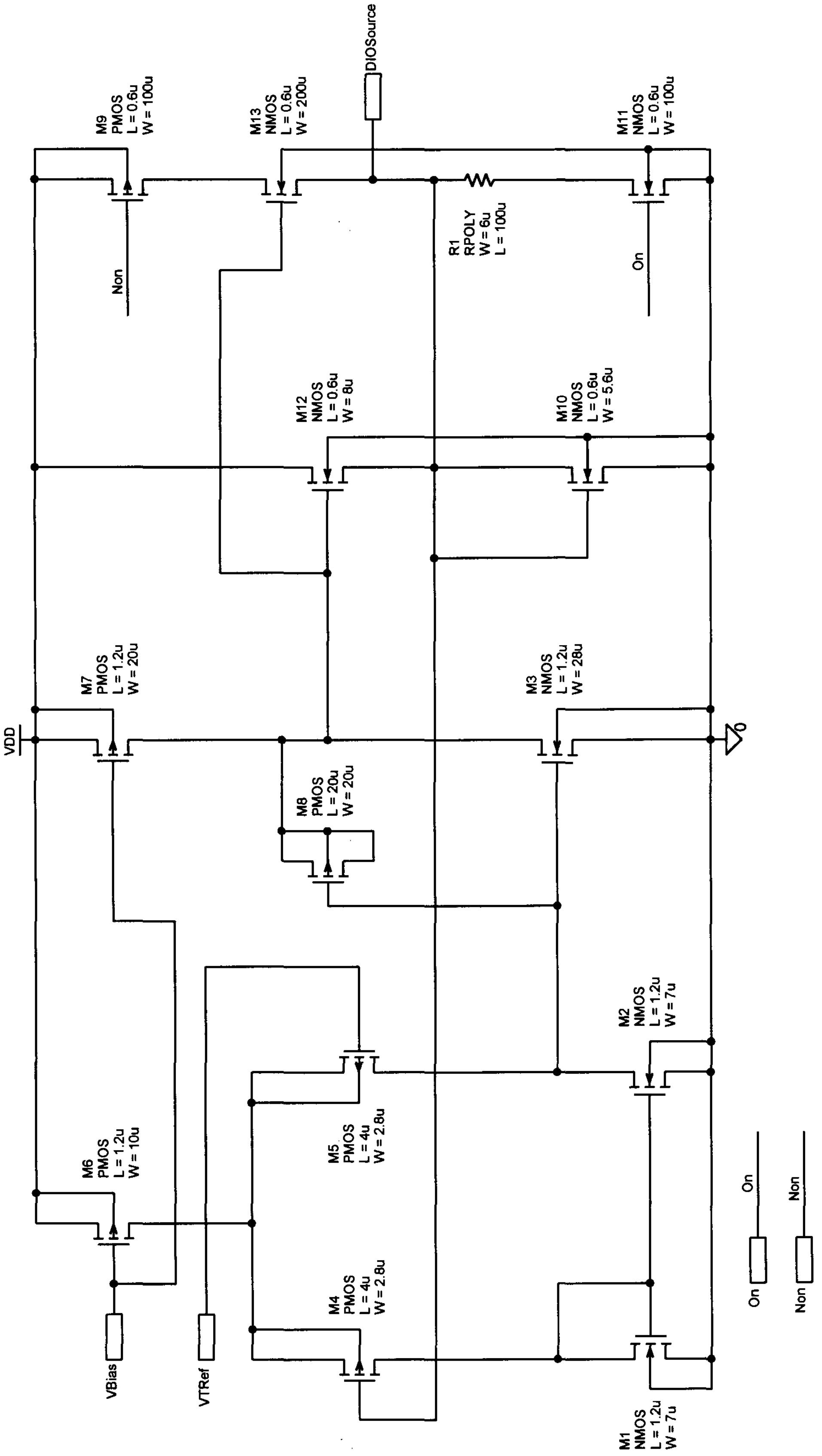


Figure C.38 NSource: *DIO*-line driver op-amp and low-impedance driver circuit.

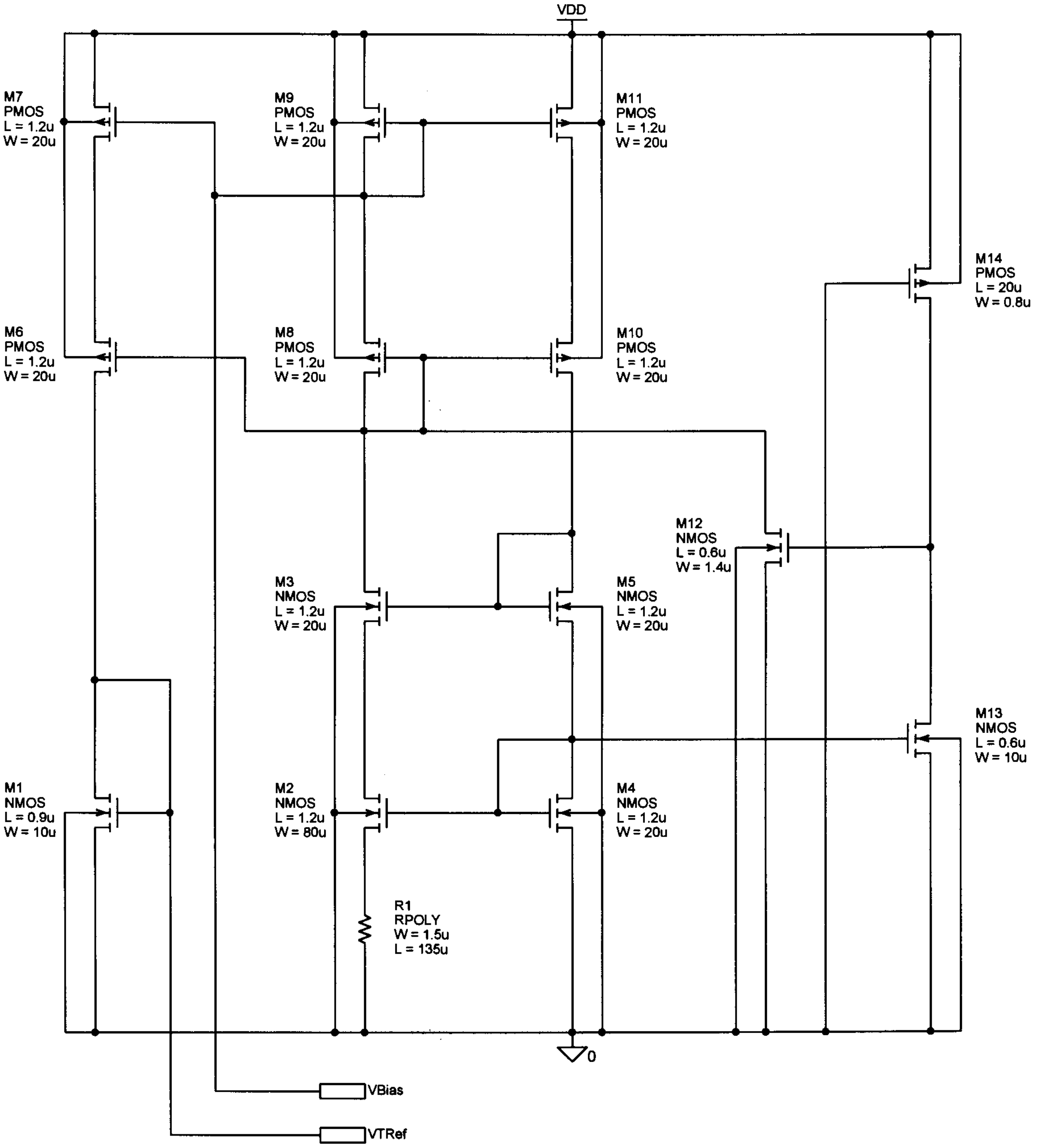


Figure C.39 NSource Bias: Reference voltage generator and bias network for the *D*/*O*-line driver circuit.

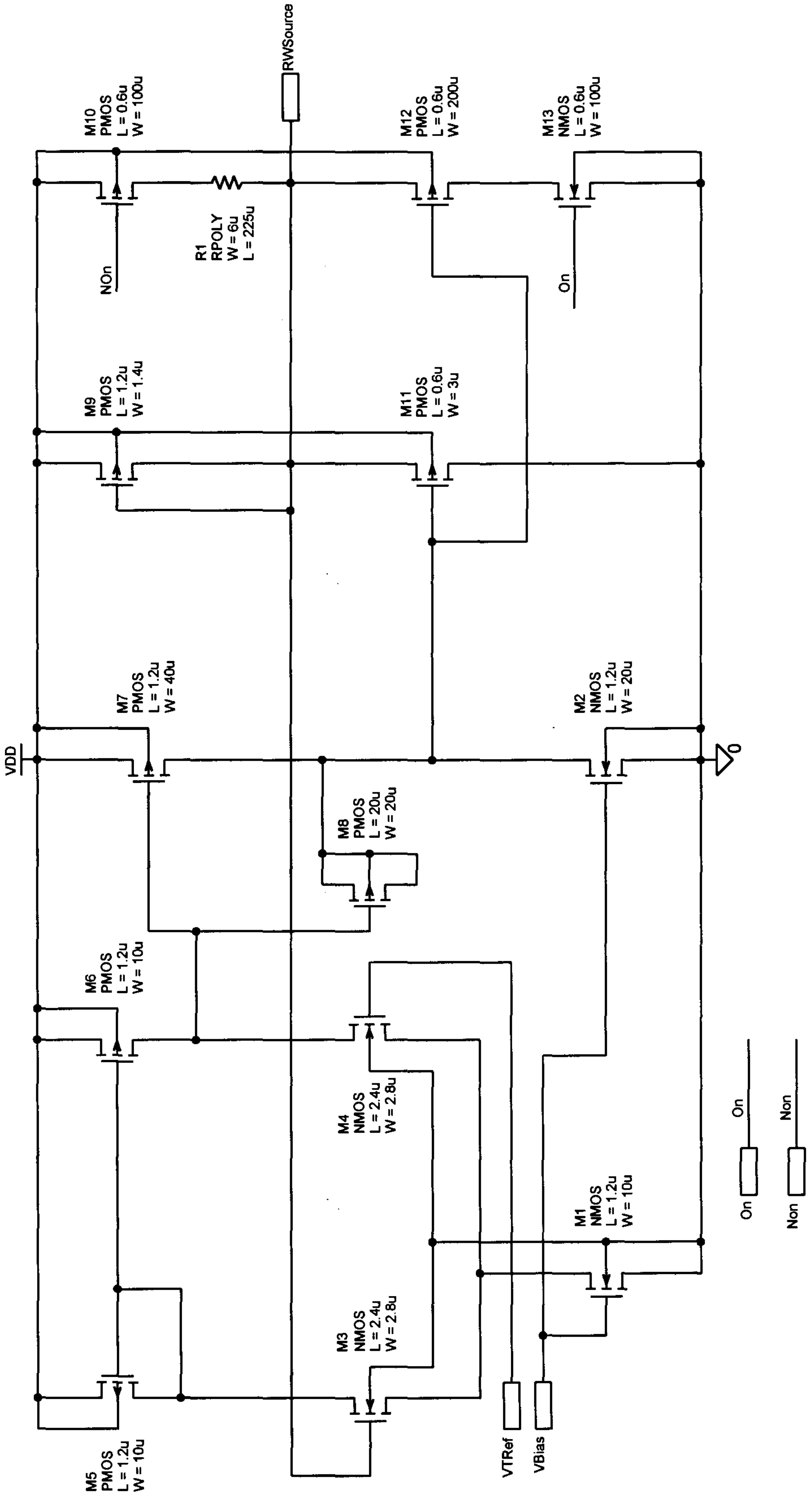


Figure C.40 PSource: *RW*-line driver op-amp and low-impedance driver circuit.

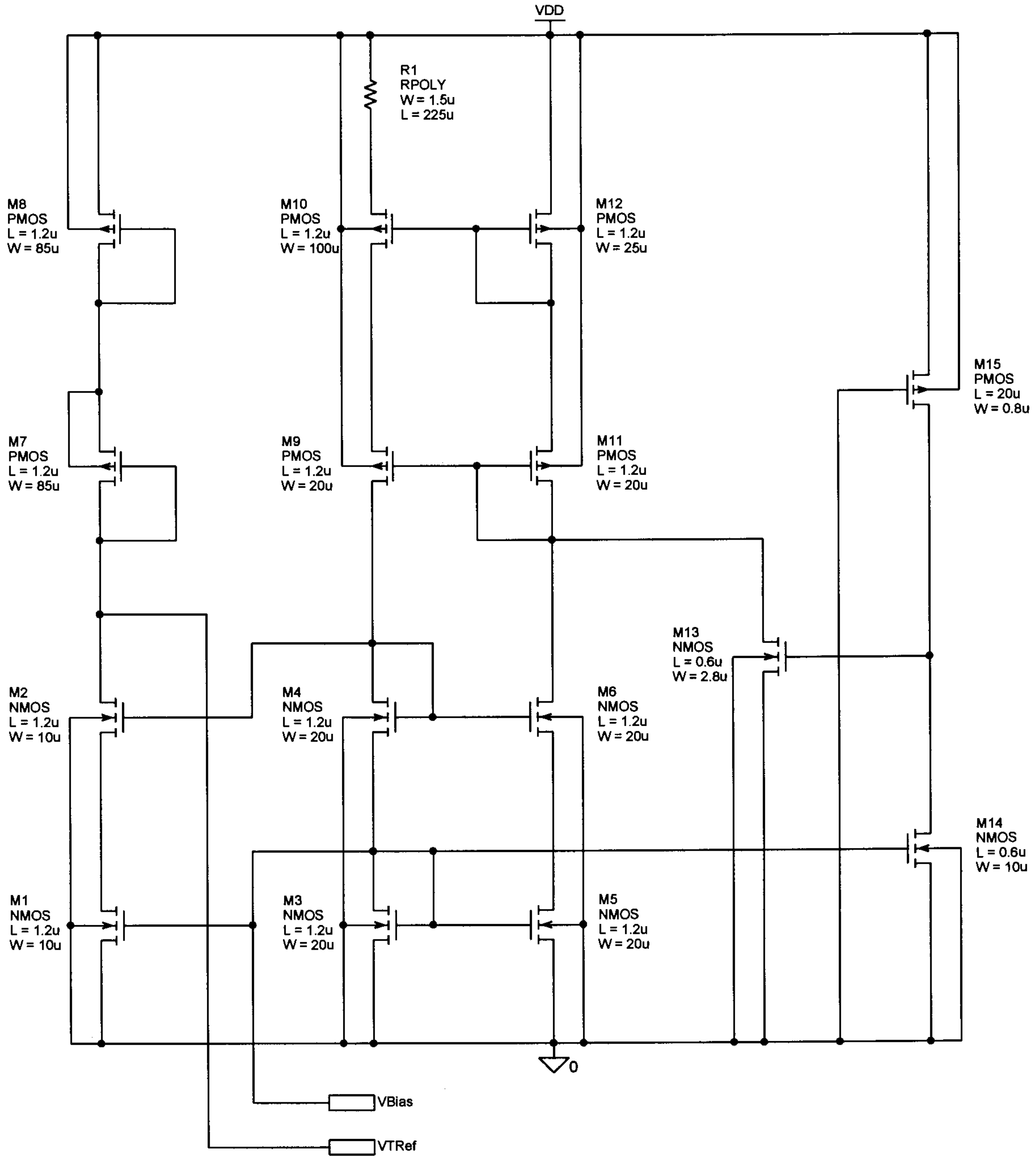


Figure C.41 PSource Bias: Reference voltage generator and bias network for the *RW*-line driver circuit.

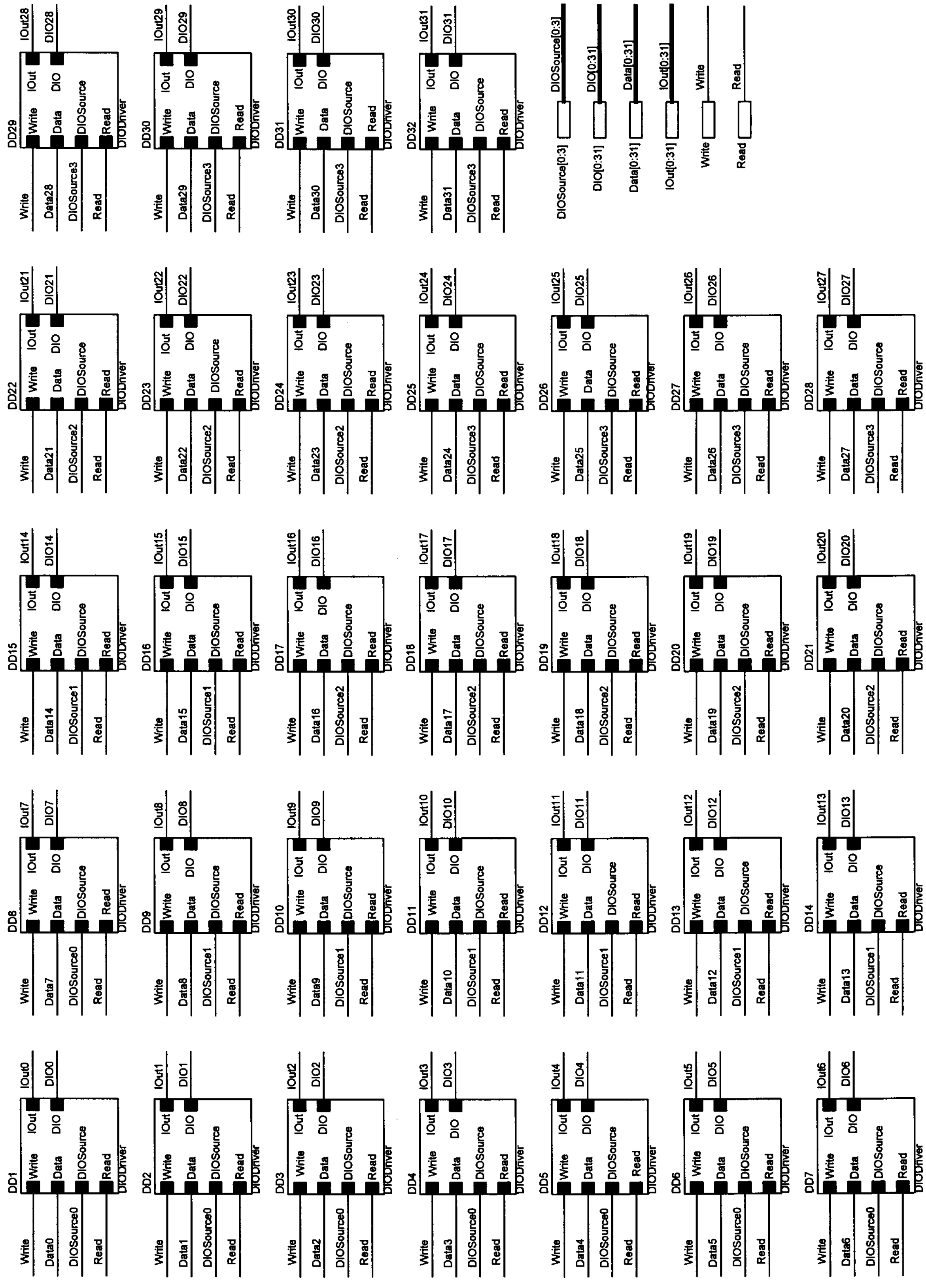


Figure C.42 Data Driver: Array of *DIO*-line driver switching circuits.

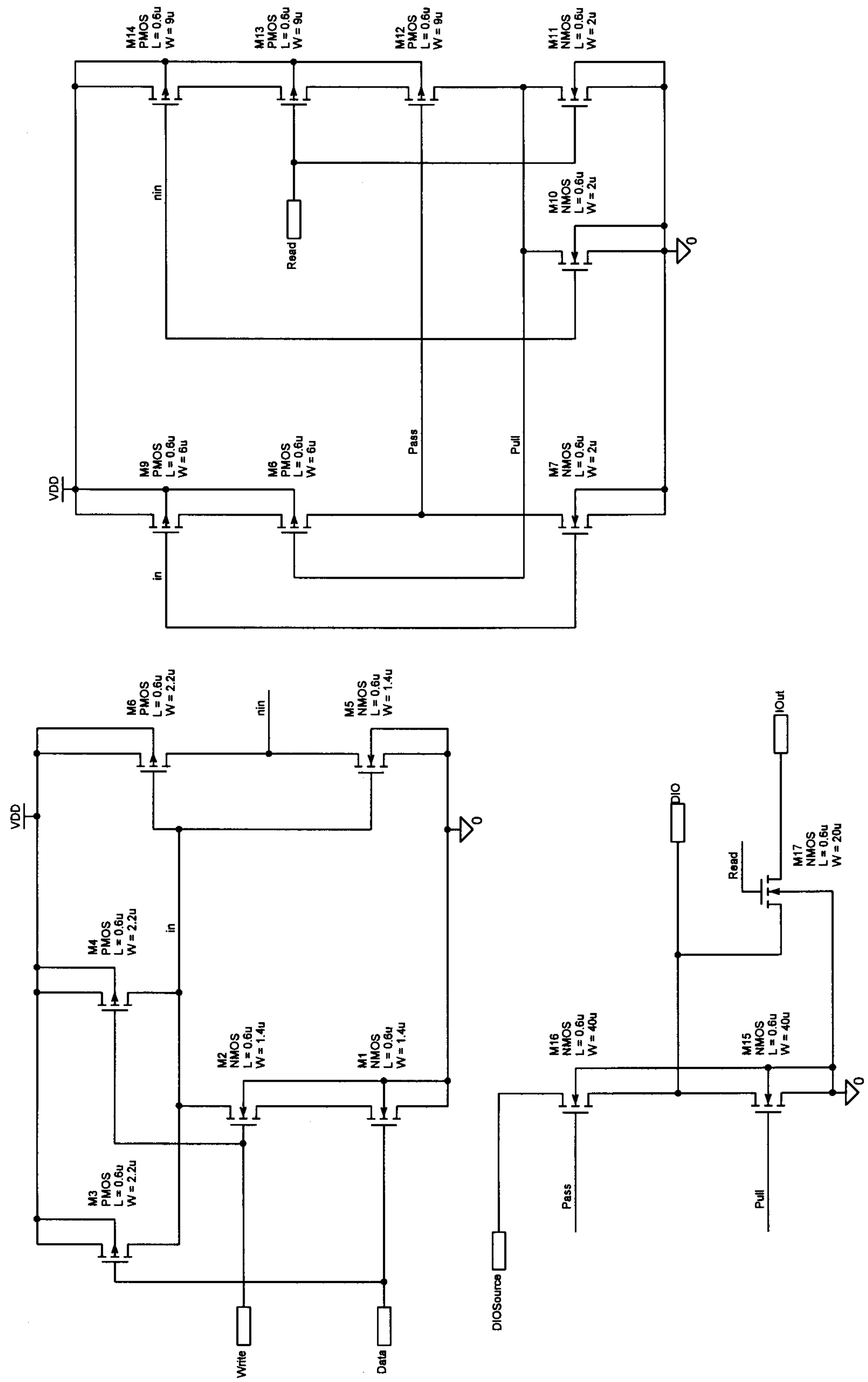


Figure C.43 DIO Driver: *DIO*-line driver switching circuit with build in current-mode multiplexing device.

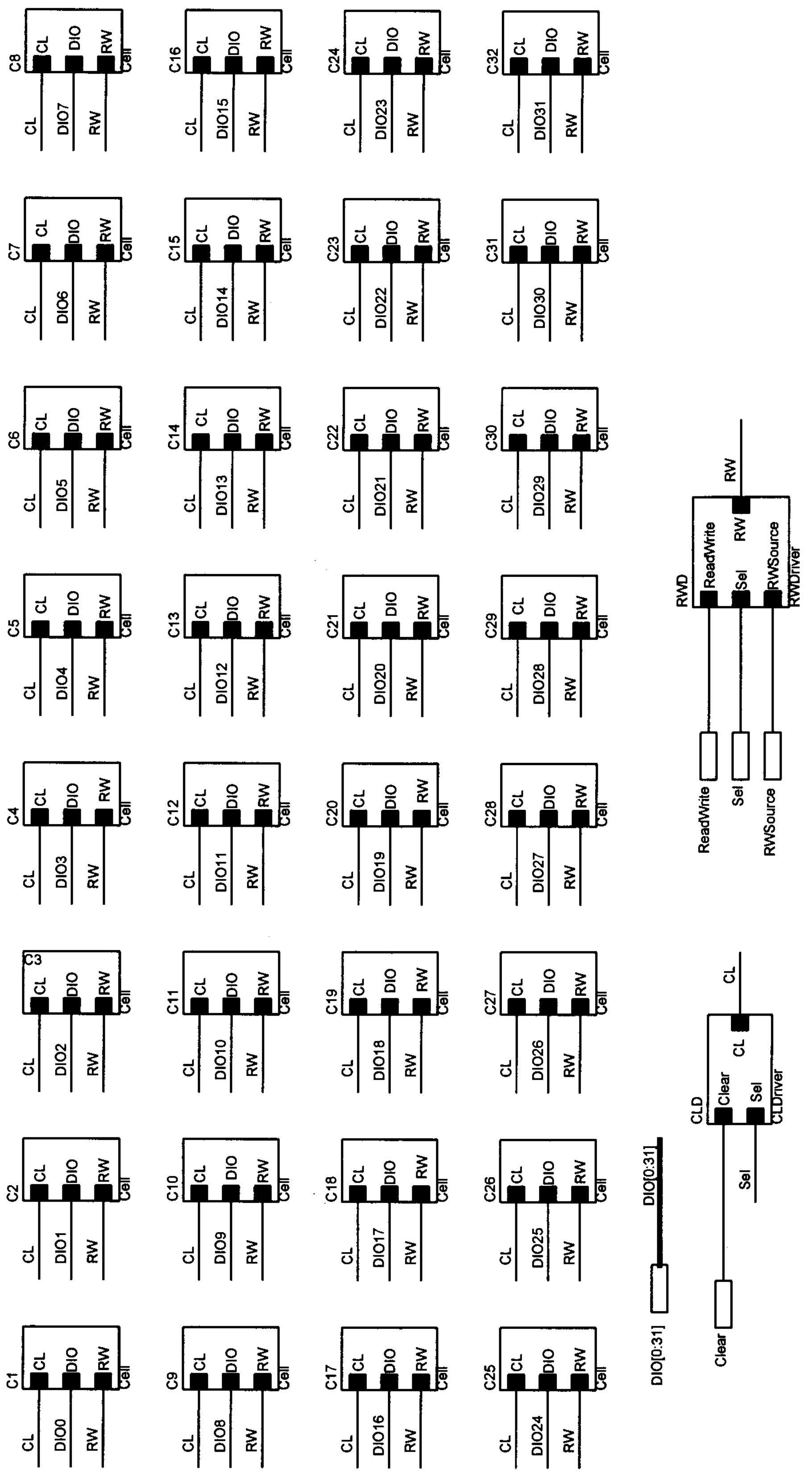


Figure C.45 Row: One row of 32 cells with the *RW*-line driver switching circuit and the *CL*-line driver circuit.

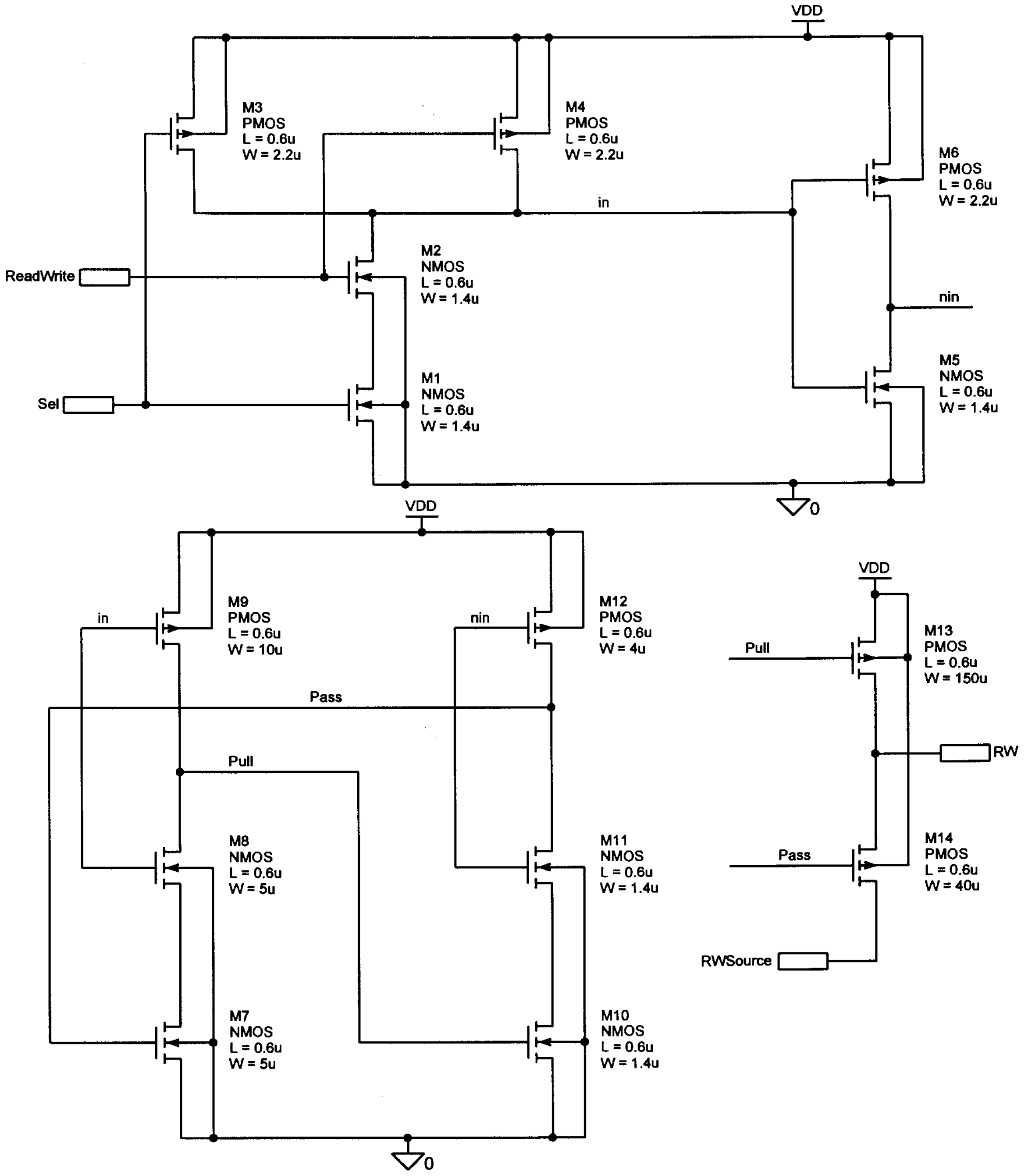


Figure C.46 RW Driver: *RW*-line driver switching circuit.

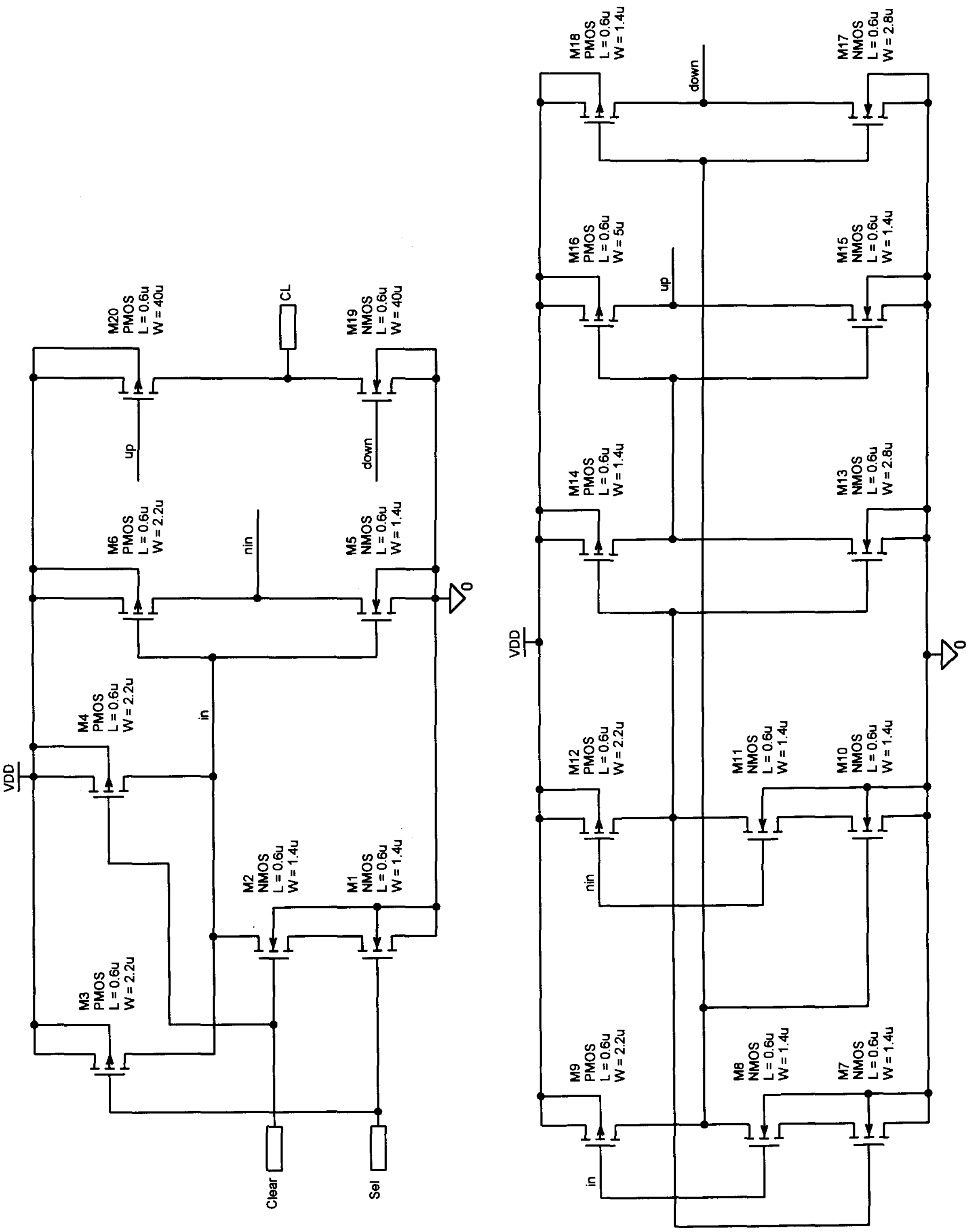


Figure C.47 CL Driver: *CL*-line driver circuit.

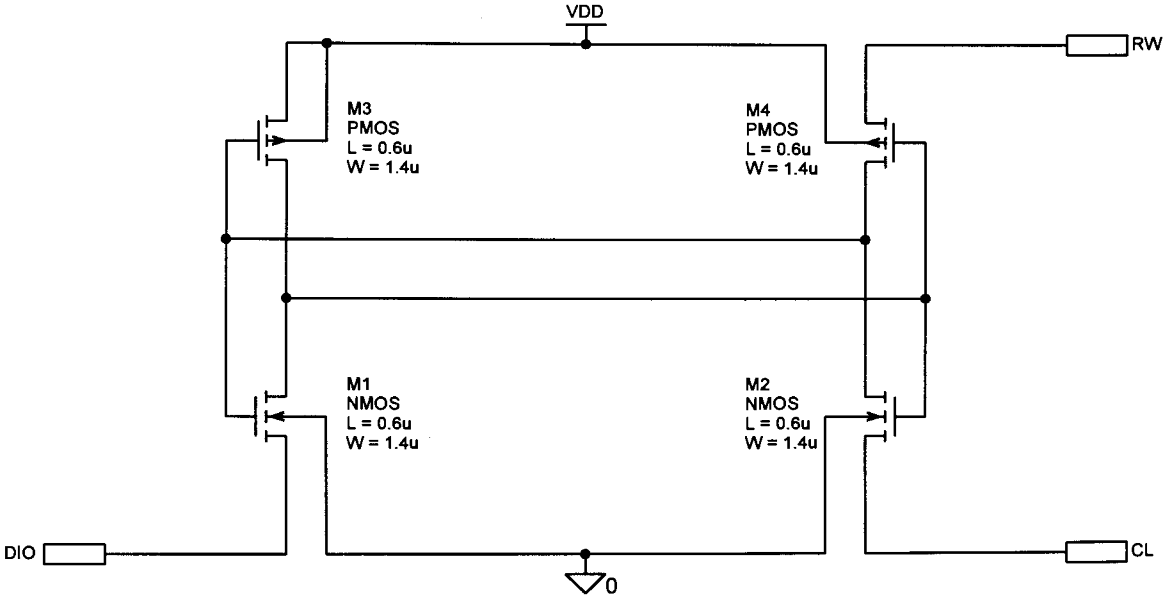


Figure C.48 Cell: Four-transistor SRAM cell.