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*.

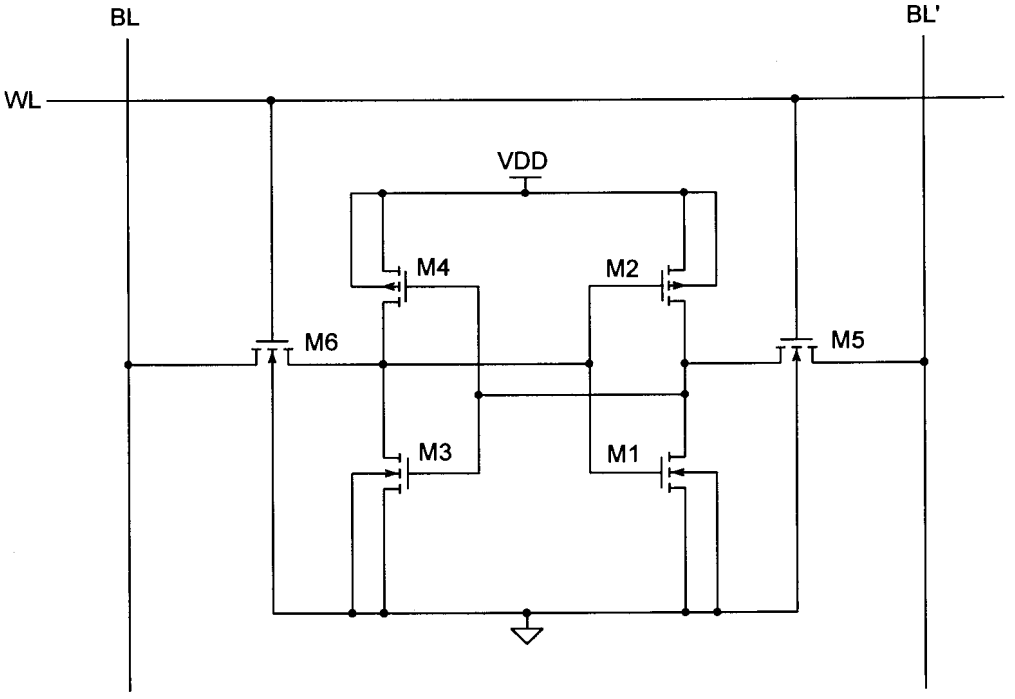


Figure 2.1 Standard six-transistor CMOS SRAM cell.

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.

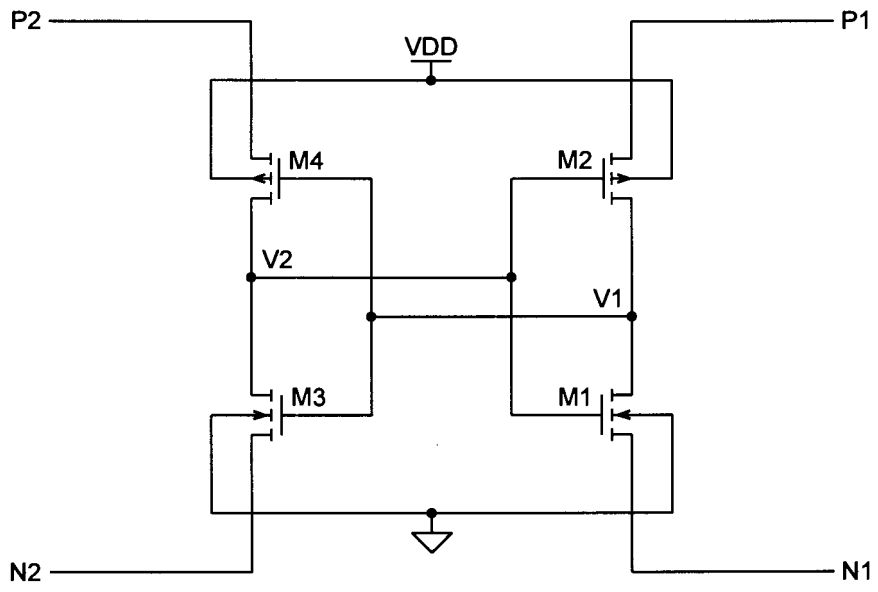


Figure 2.2 Four-transistor CMOS SRAM cell.

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].

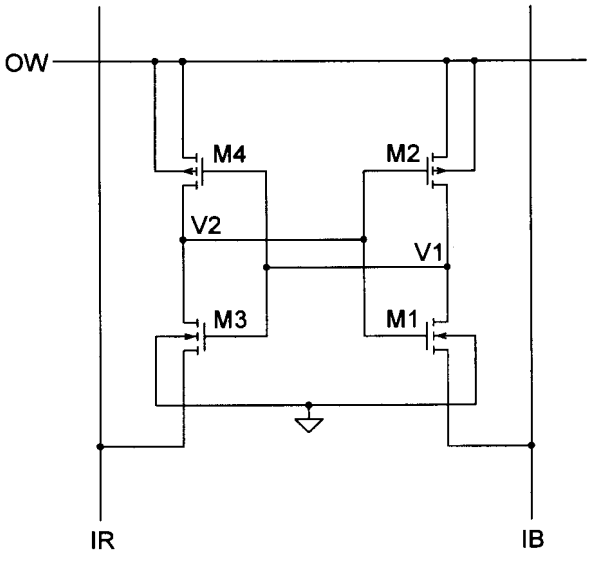


Figure 2.3 Smallest area four-transistor cell array implementation.

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.

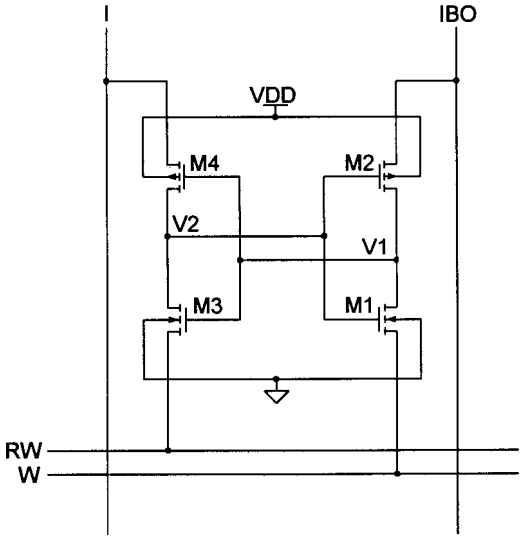


Figure 2.4 Advanced four-transistor cell array implementation.

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.

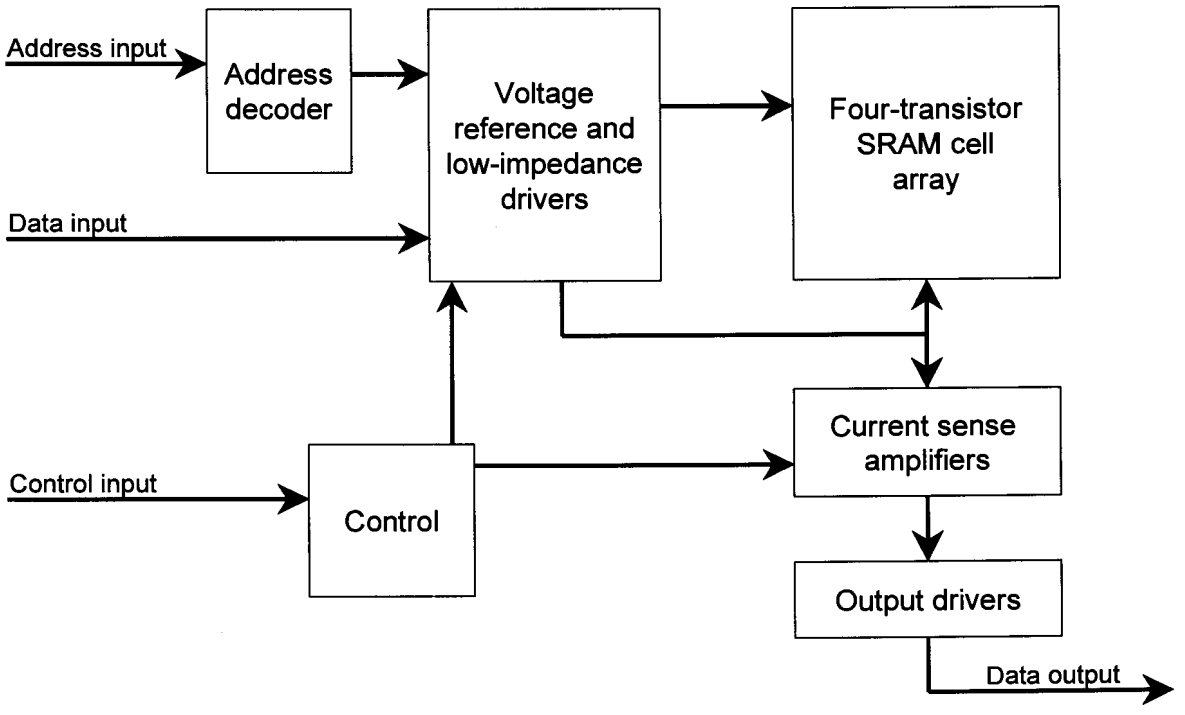


Figure 2.5 Block diagram of the SRAM system.

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.

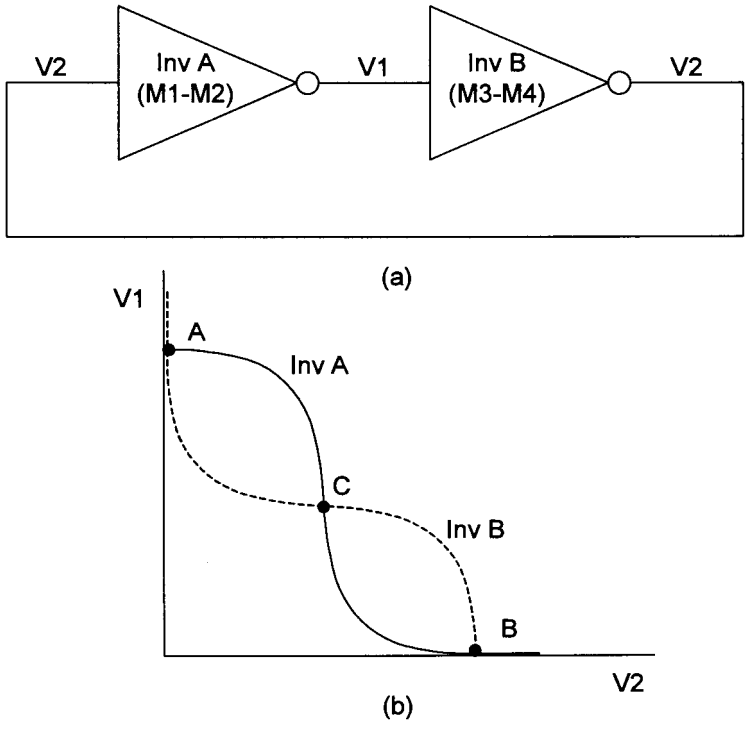


Figure 2.6 (a) A pair of cross-coupled inverters with (b) their voltage transfer characteristics showing the three operating points.

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

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

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

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

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.

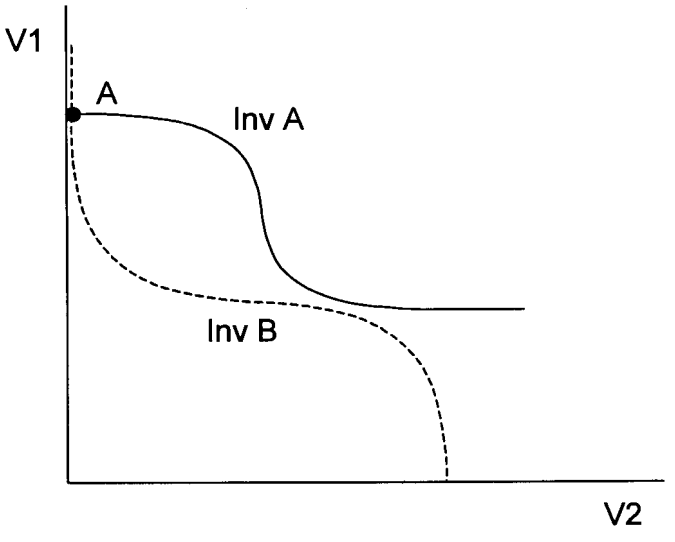


Figure 2.7 A cross-coupled inverter pair transfer curve when the ground node of inverter A is raised. Only one stable operating point exists.

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

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

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.

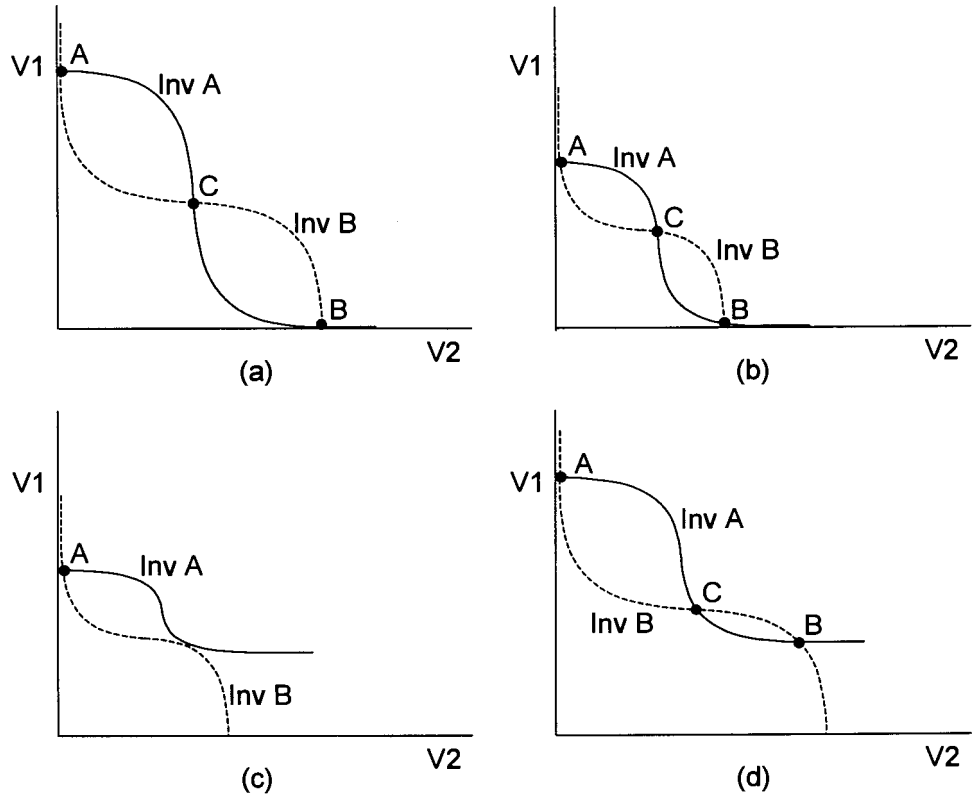


Figure 2.8 (a) The voltage transfer characteristic of a pair of cross-coupled inverters. (b) The supply voltage has been lowered and (c) one ground node has been raised to create a single operating point. (d) The ground node is raised without lowering the power supply and two stable operating points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

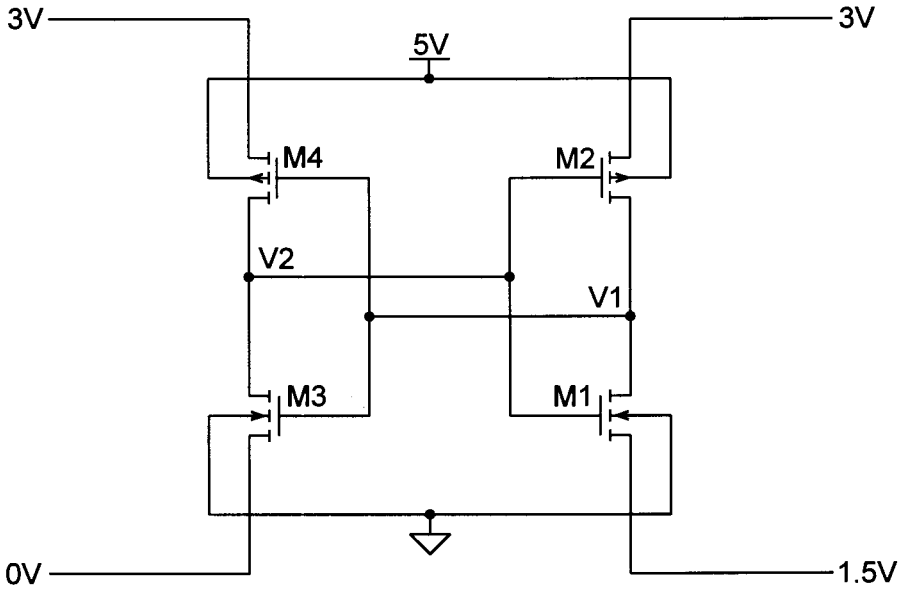


Figure 2.9 Four-transistor SRAM cell with write condition applied.

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.

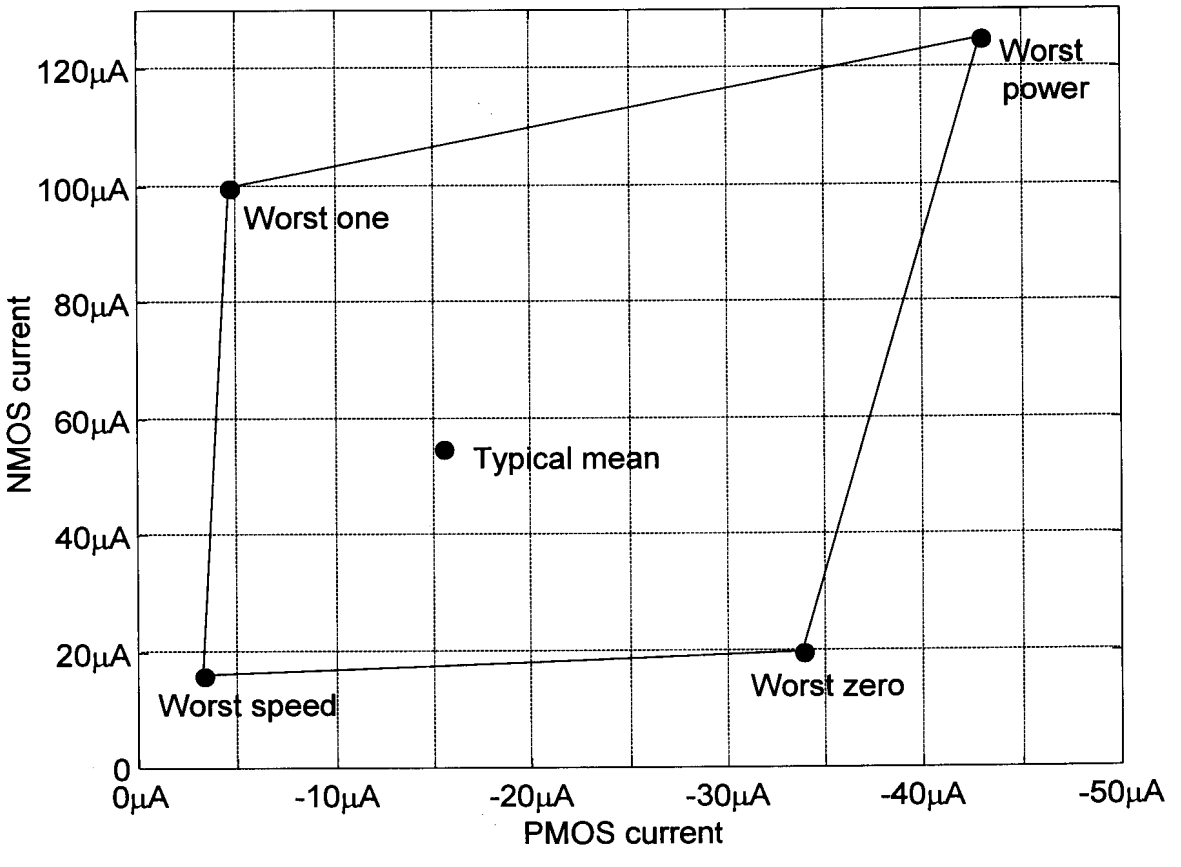


Figure 2.10 Drain current of saturated NMOS and PMOS devices at $|V_{GS}|$=1.3V and a *W/L* ratio of 1.4µm/0.6µm.

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.

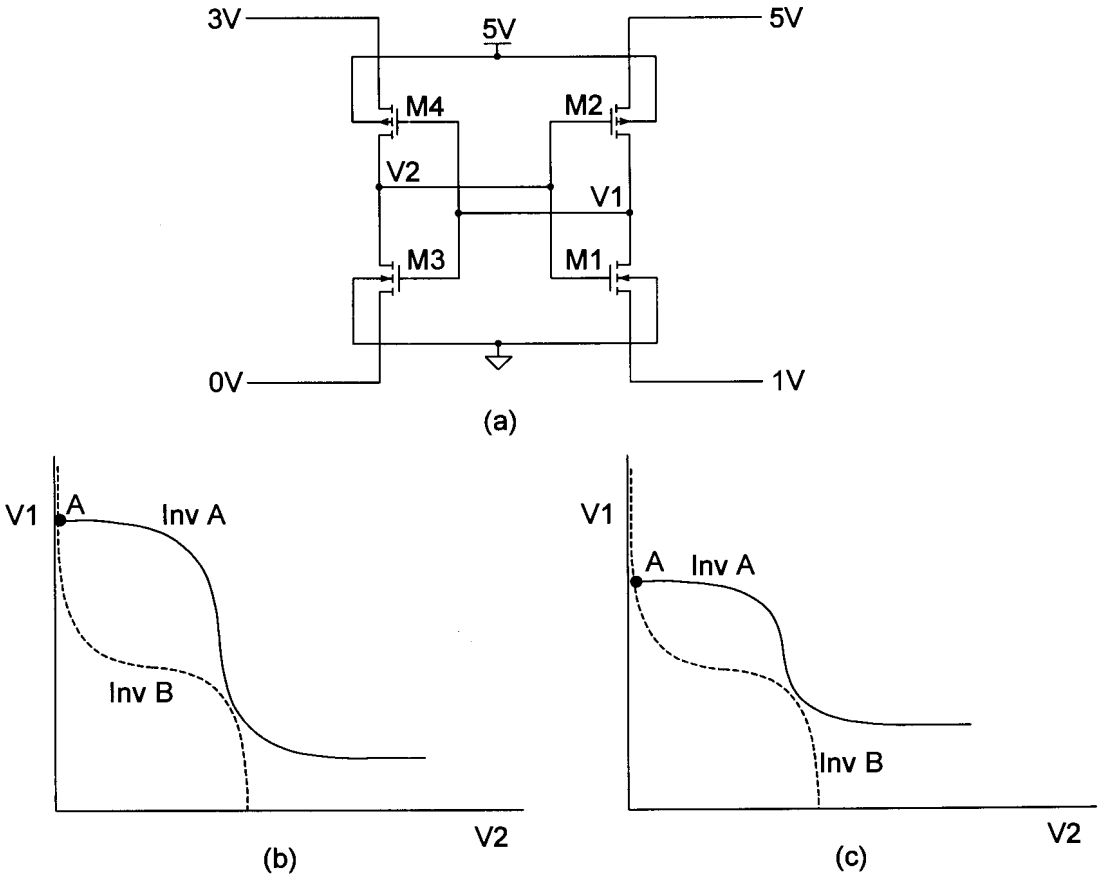


Figure 2.11 (a) Newly proposed scheme of writing the four-transistor SRAM cell. (b) The static write condition for this scheme compared to (c) the static write condition for the previously proposed scheme [2].

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.

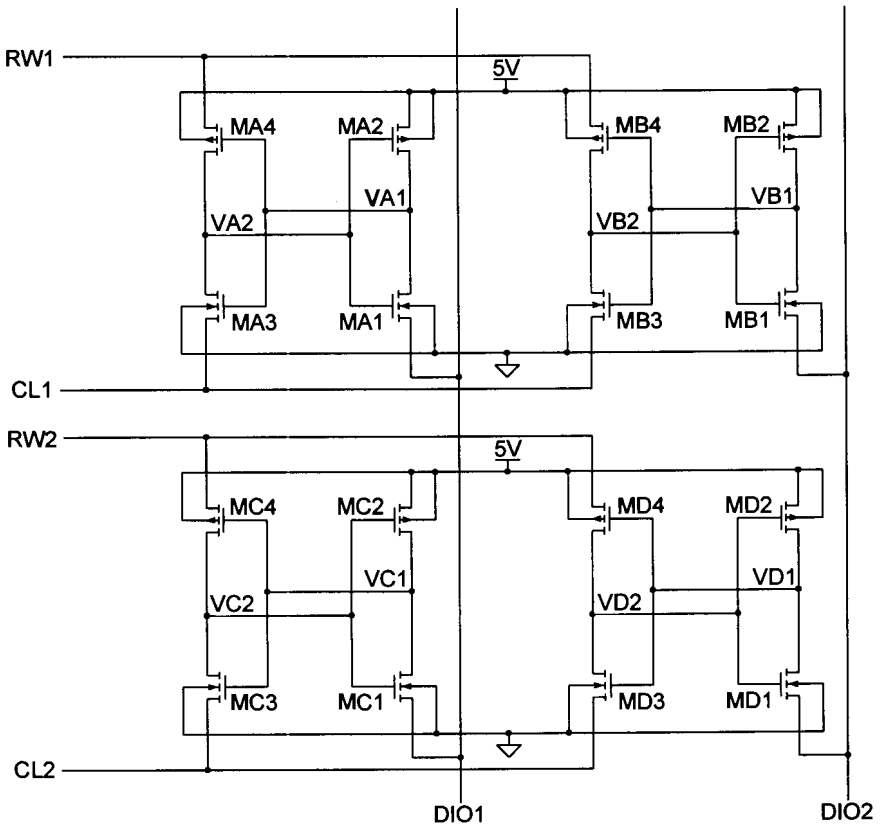


Figure 2.12 Proposed array structure incorporating the alternative write cycle.

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 *DIO*-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 *DIO*-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.

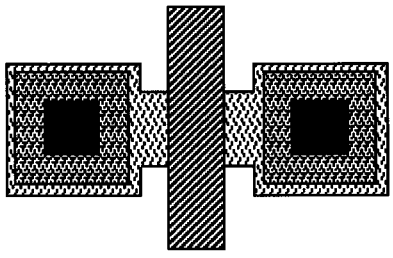
The price paid for these advantages is the two cycle write procedure which requires more time and more complex control.

2.7 CELL DESIGN

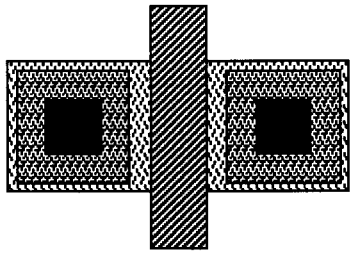
In this section a design procedure that can be applied to design the four-transistor SRAM cell for any CMOS process is discussed. Two aspects require designing, namely the device ratio between the NMOS and PMOS device and the magnitude of the voltage deviations. The design of the latter is based on a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



(a)



(b)

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).

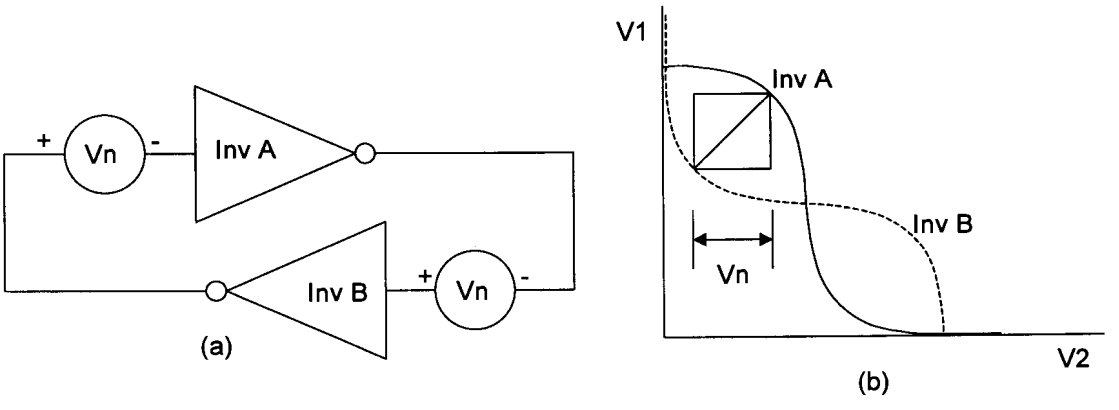


Figure 2.14 (a) Cross-coupled inverter with worst case series-voltage noise sources inserted and (b) the graphical representation of the worst case series-voltage noise margin.

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

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

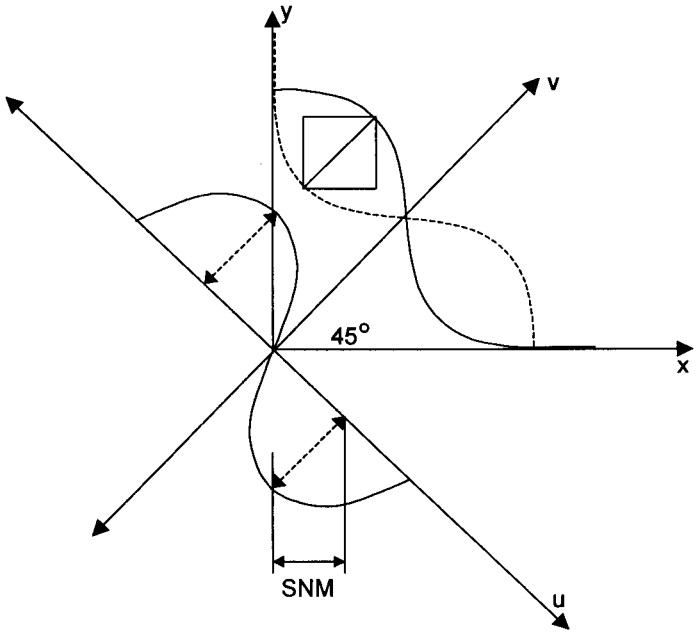


Figure 2.15 Static noise margin (SNM) estimation based on "maximum squares" in a 45° rotated coordinate system.

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.

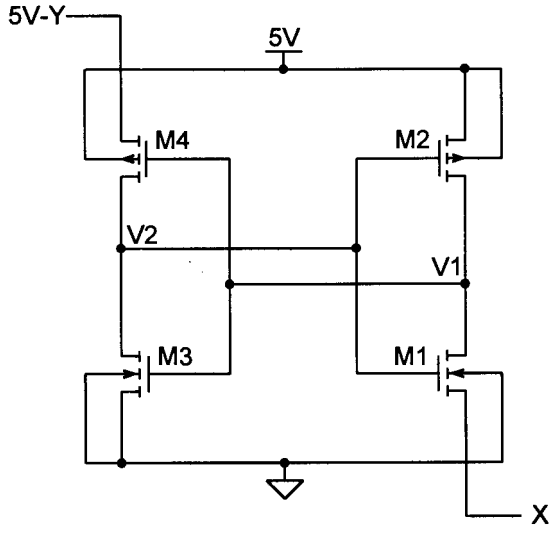


Figure 2.16 Static noise margin analysis system.

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.

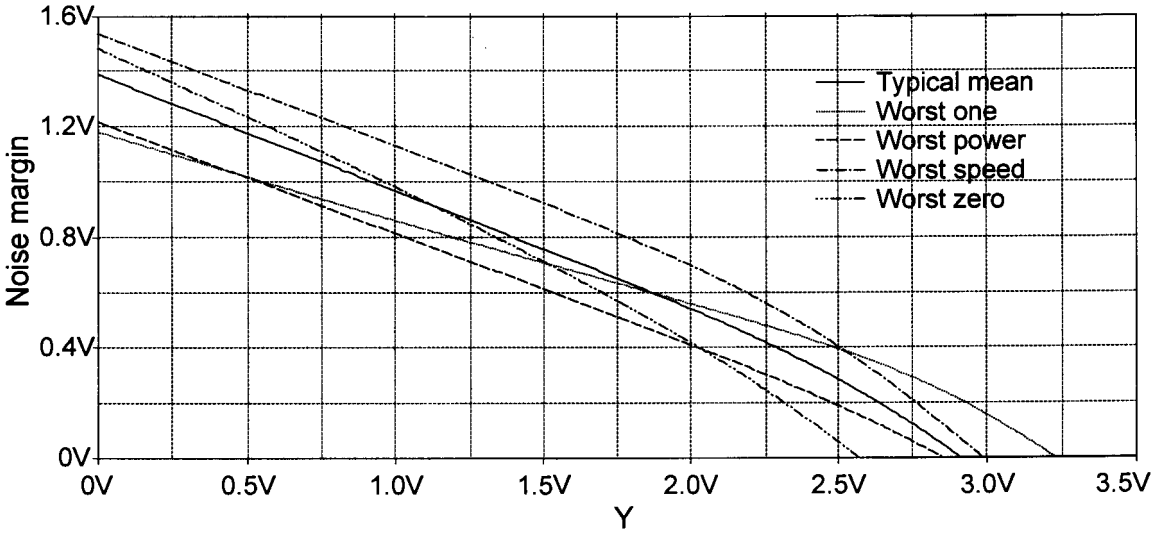


Figure 2.17 Noise margin plotted against *Y*-deviation for *X*=0 for the different simulation models.

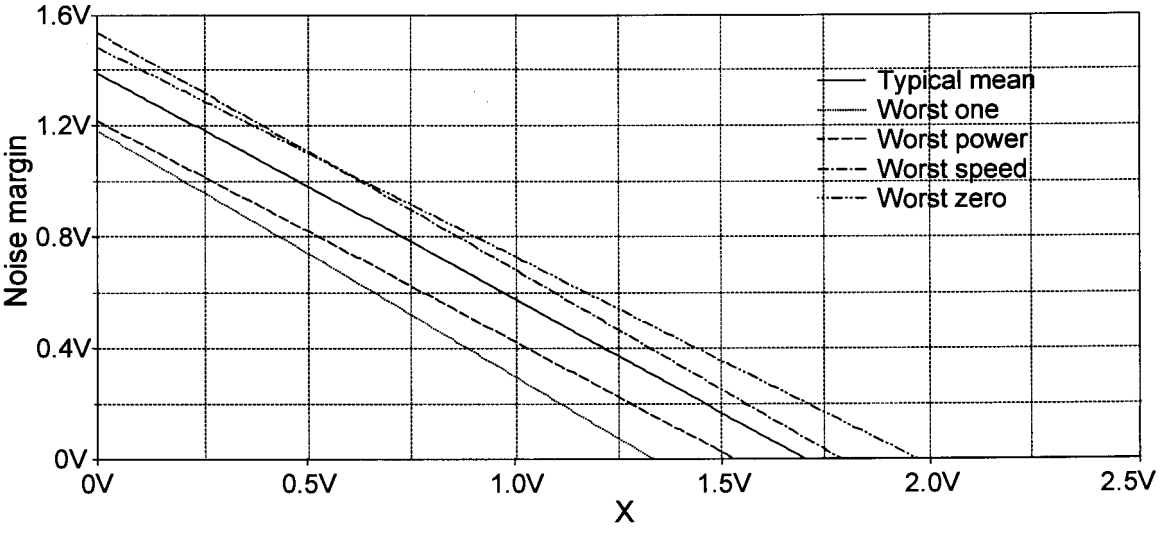


Figure 2.18 Noise margin plotted against *X*-deviation for *Y*=0 for the different simulation models.

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.

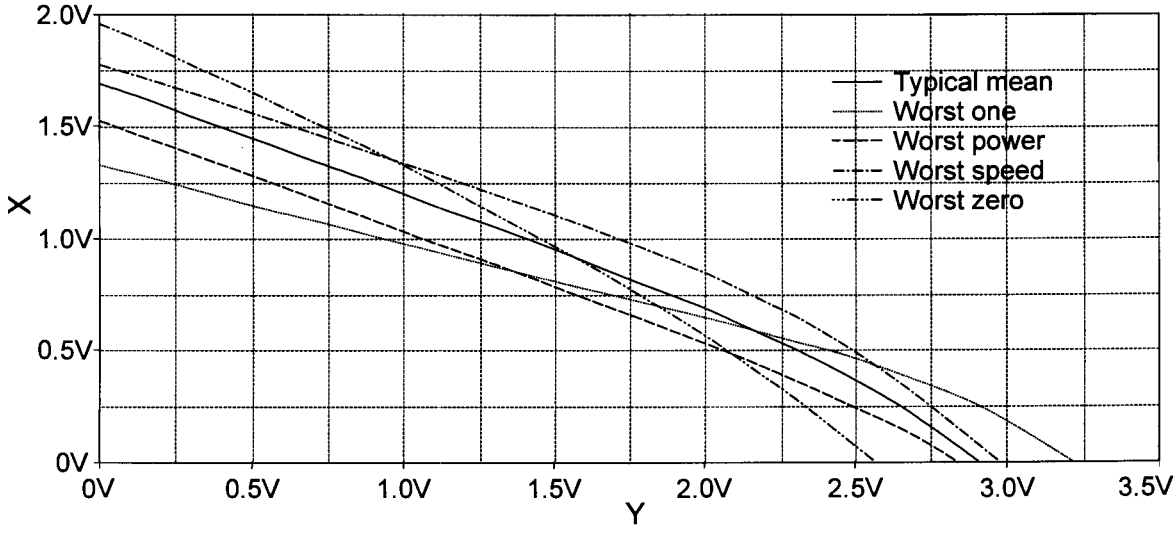


Figure 2.19 Zero noise margin trajectories for all simulation models. A point above the graphs implies that static write conditions are satisfied.

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.

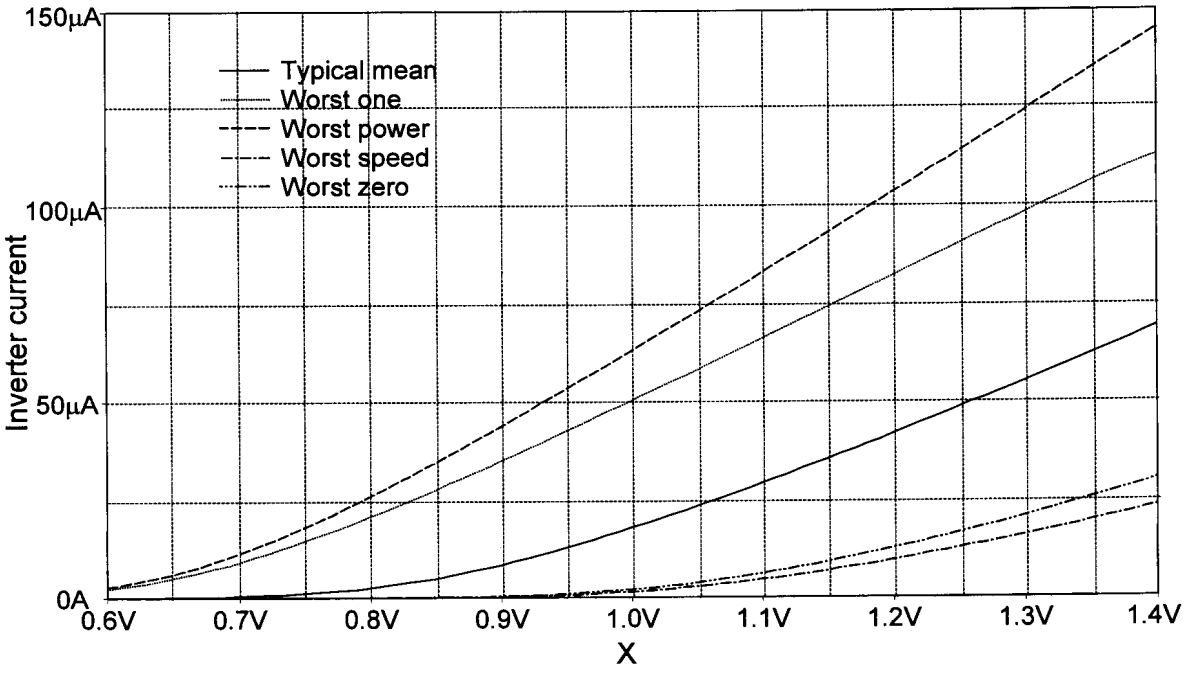


Figure 2.20 Simulated current flowing in the opposite inverter of the four-transistor SRAM cell across the five process models when a certain *X*-deviation is applied.

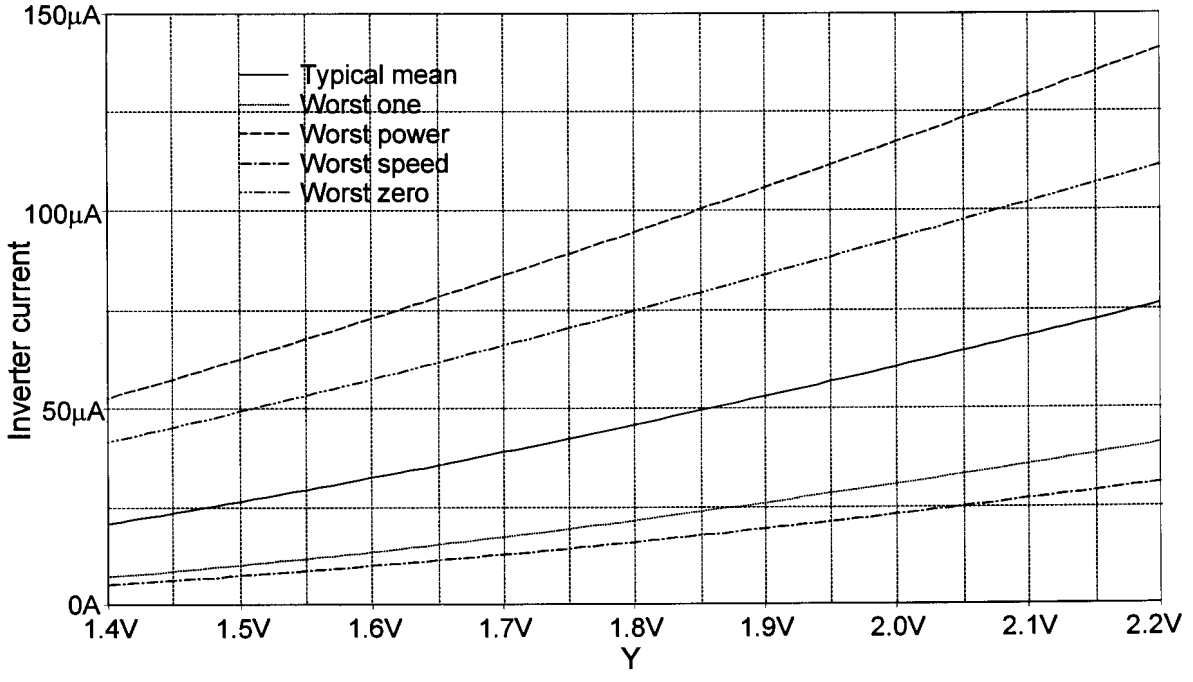


Figure 2.21 Simulated current flowing in the opposite inverter of the four-transistor SRAM cell across the five process models when a certain *Y*-deviation is applied.

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 *DIO*-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.

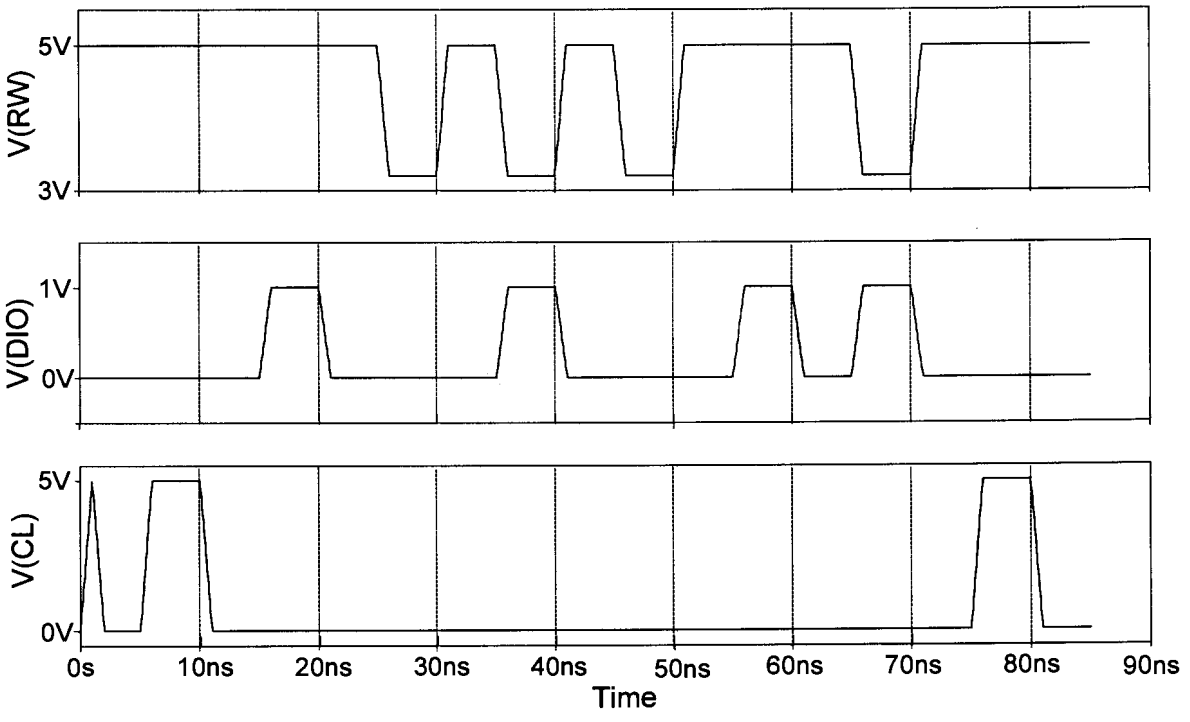


Figure 2.22 Control signals applied to the four-transistor SRAM cell for the typical mean case.

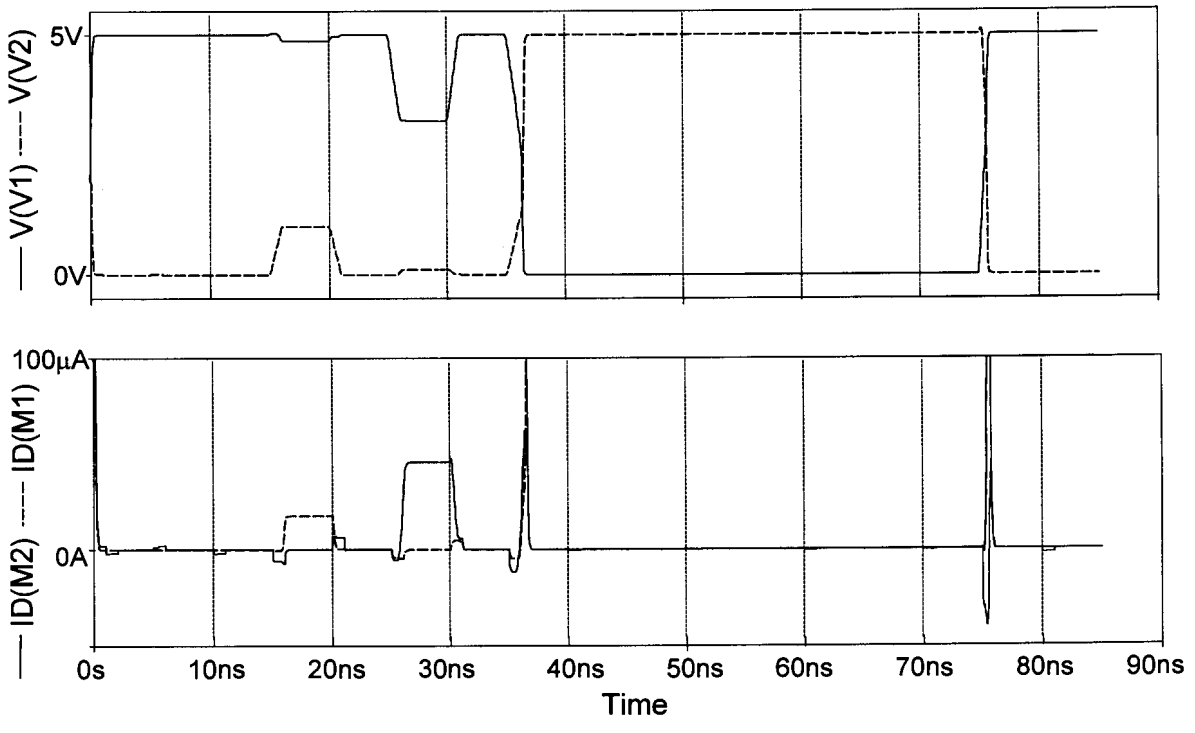


Figure 2.23 Response of the four-transistor SRAM cell using the typical mean model.

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.

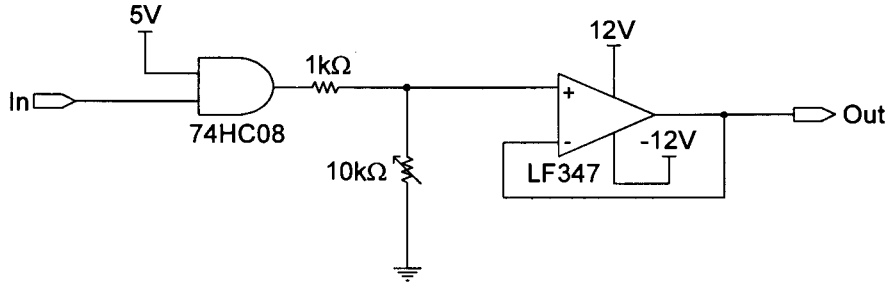


Figure 2.24 Discrete NMOS source driver circuit.

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.

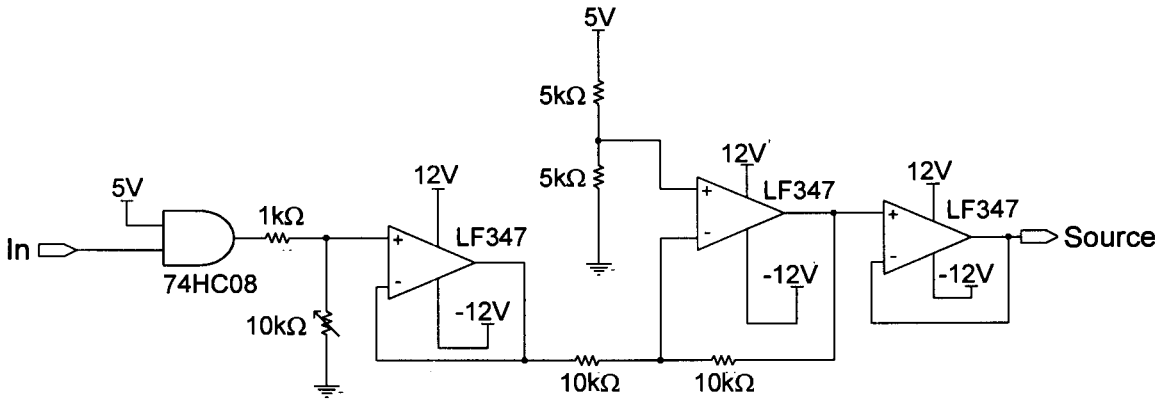


Figure 2.25 Discrete PMOS source driver circuit.

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.

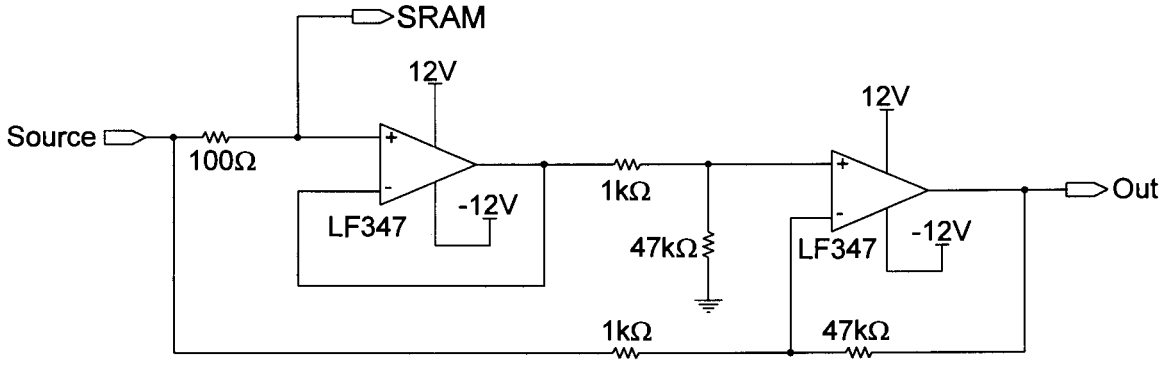


Figure 2.26 Discrete current-to-voltage converter.

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 *DIO*). 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 *DIO*-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 *DIO*-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.

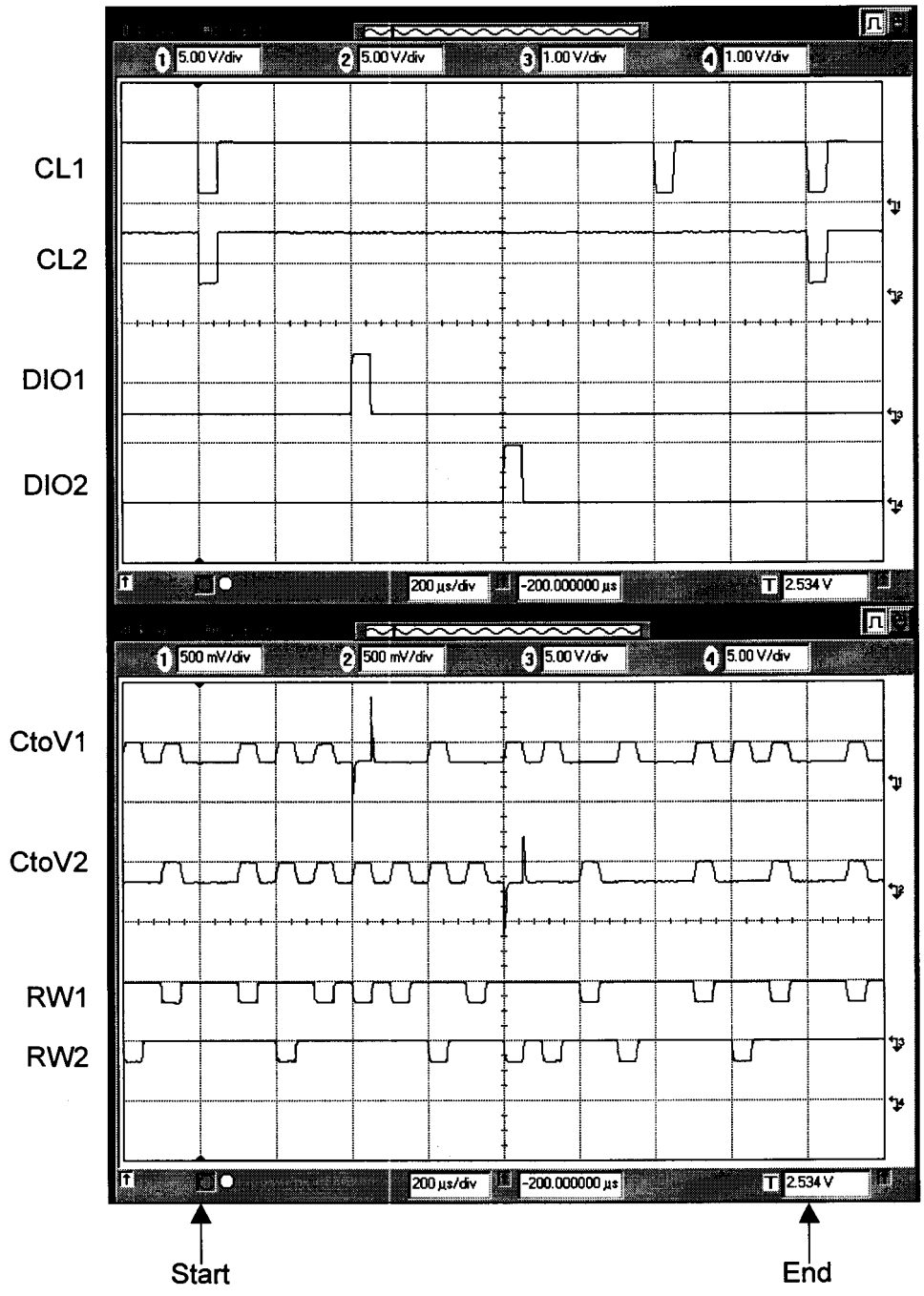


Figure 2.27 Experimental results for the 2x2 SRAM array showing the four described procedures between the "Start" and "End" indicators.

The plots also show the deviations used. The *DIO*-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.

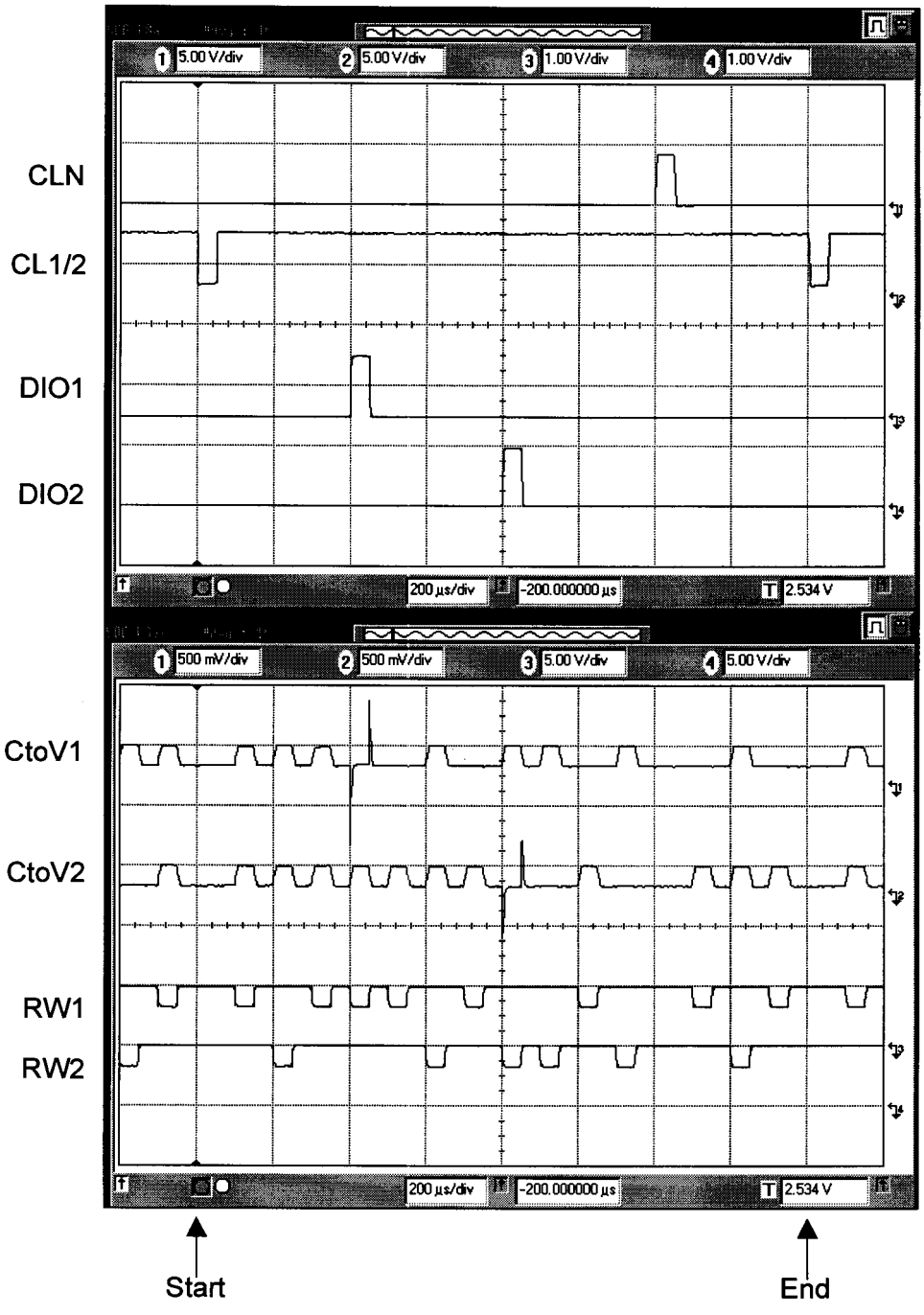


Figure 2.28 Experimental results for the sequence that tests clearing the cell via the NMOS source.

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).

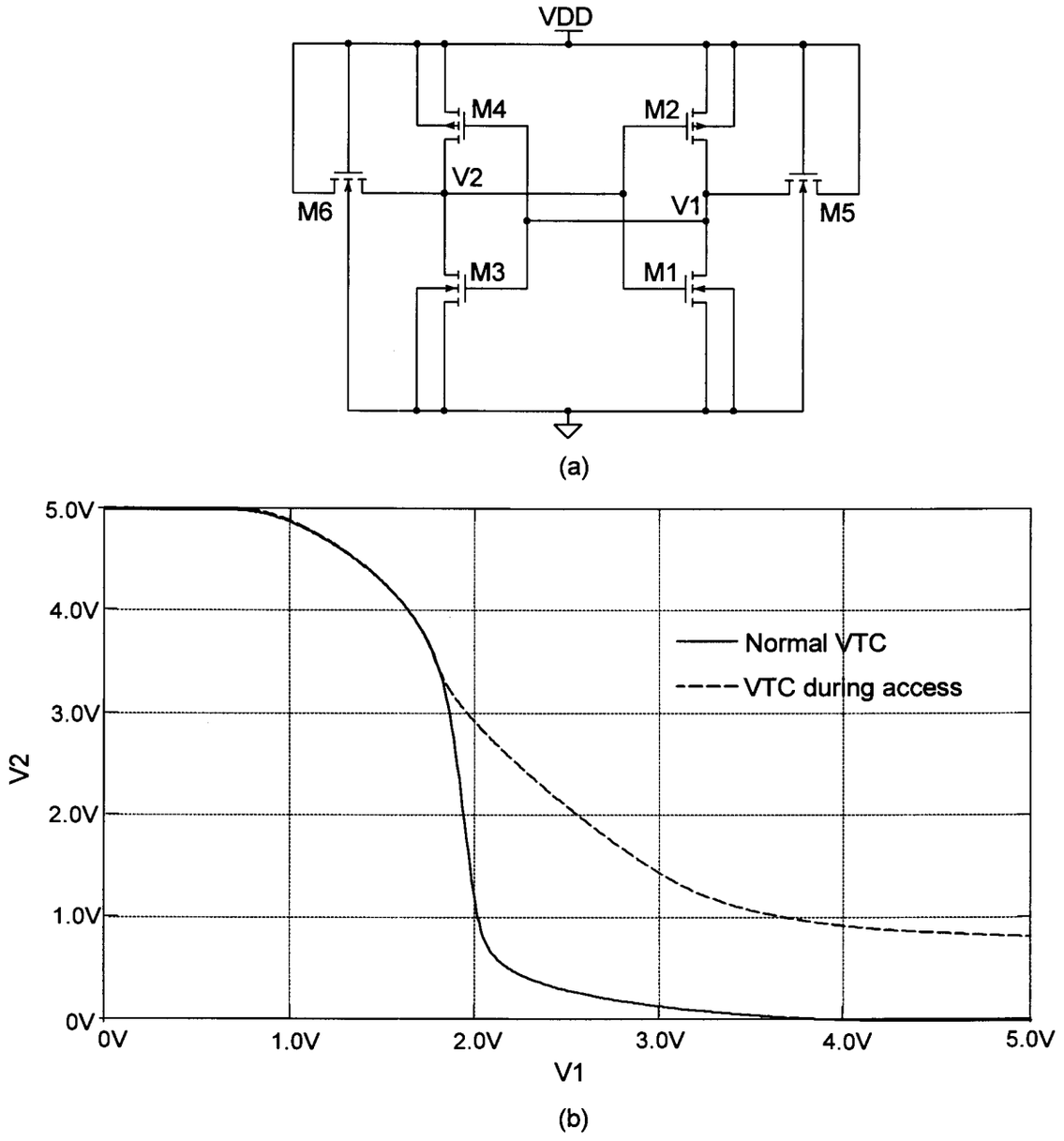


Figure 2.29 (a) Six-transistor SRAM cell during initial read access and (b) the effect this has on the voltage transfer characteristic.

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.

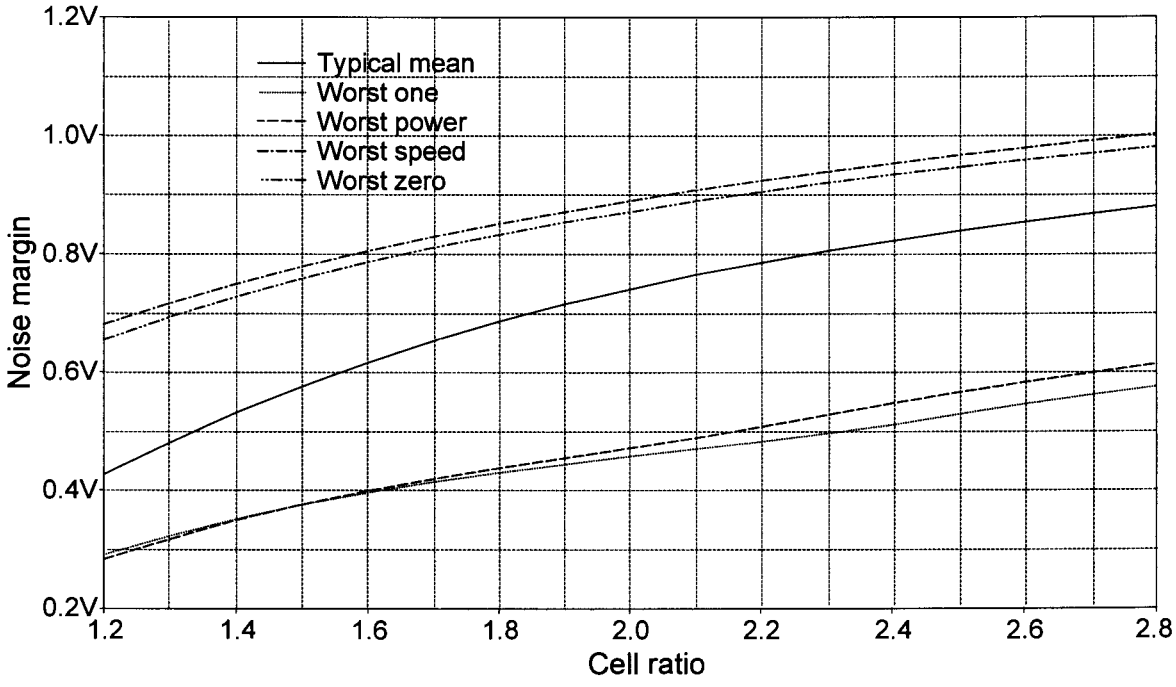


Figure 2.31 Static noise margin of a six-transistor SRAM cell during read access as a function of cell ratio for different simulation models.

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.

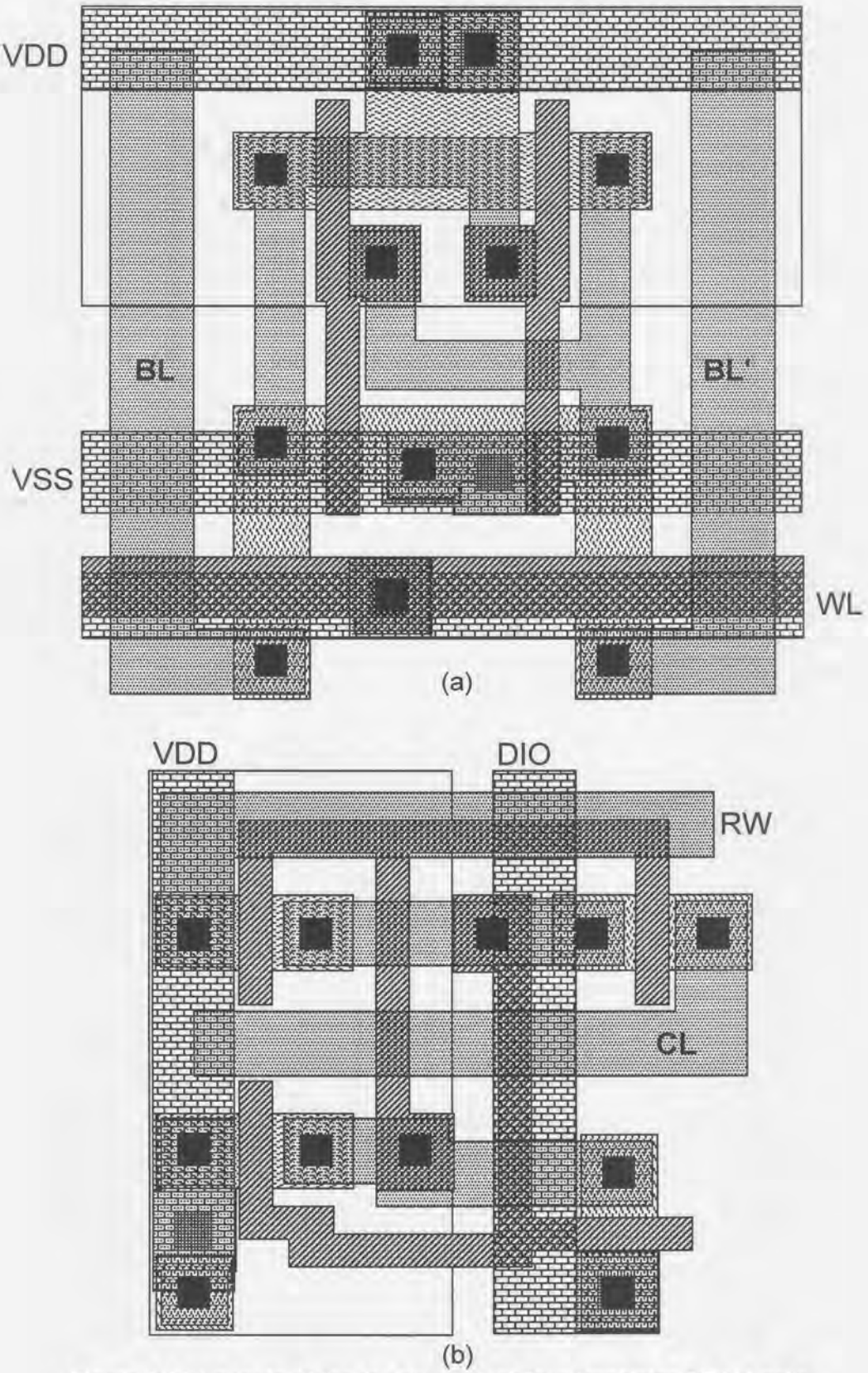


Figure 2.32 Layouts of the (a) six- and (b) four-transistor SRAM cells (Legend in addendum B).

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 *DIO*-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 *DIO*-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 *DIO*-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 *DIO*-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.

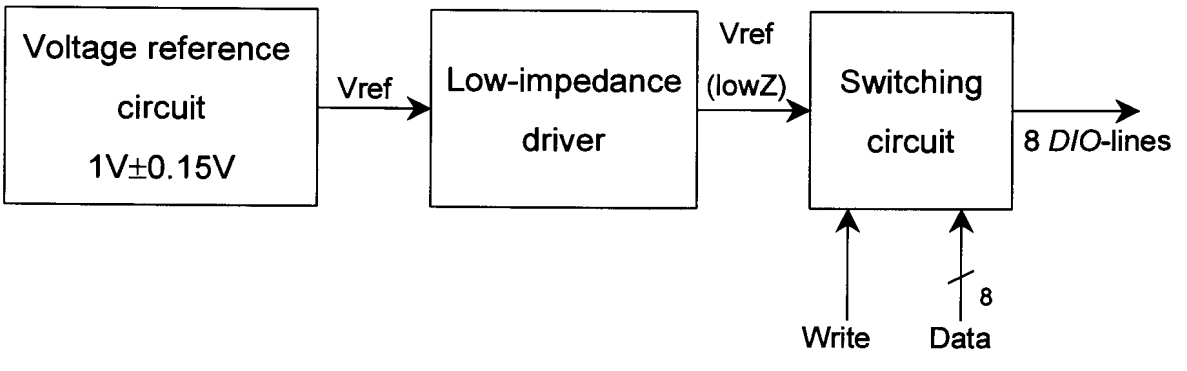


Figure 3.1 Functional block diagram of the *DIO*-line driver circuit.

The low-impedance driver circuit assures that the capacitance associated with the *DIO*-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 *DIO*-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 *DIO*-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 *DIO*-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 *DIO*-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 *DIO*-line to be switched between the two voltages.

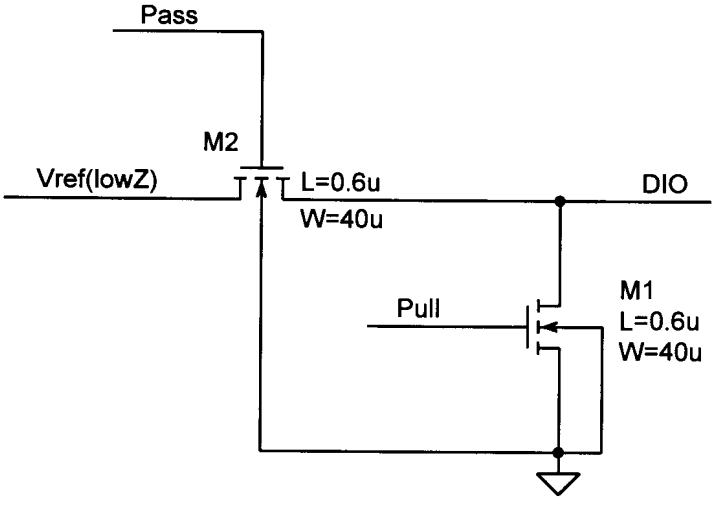


Figure 3.2 Circuit to switch the *DIO*-line between ground and the reference voltage.

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 *DIO*-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 *DIO*-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.

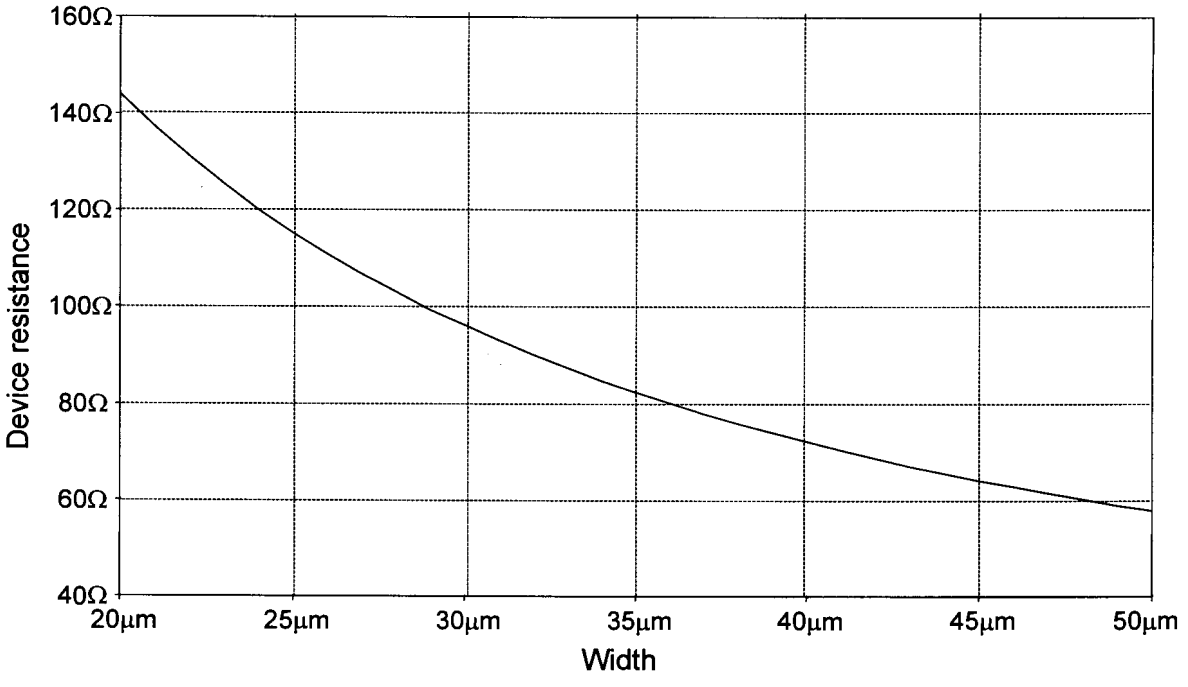


Figure 3.3 Resistance of the pull-down device as a function of the device width. The gate length is 0.6μm.

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.

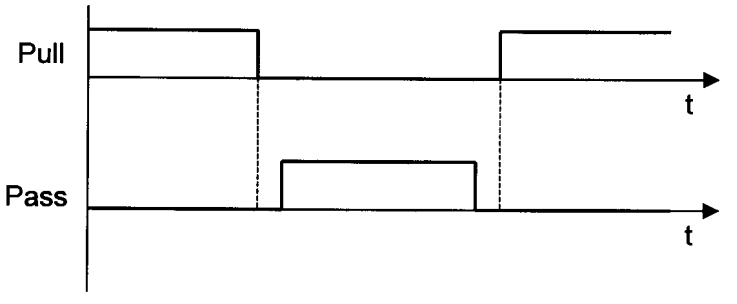


Figure 3.4 Timing pattern for the circuit in Figure 3.2 to limit short-circuit current and load on the driver circuit.

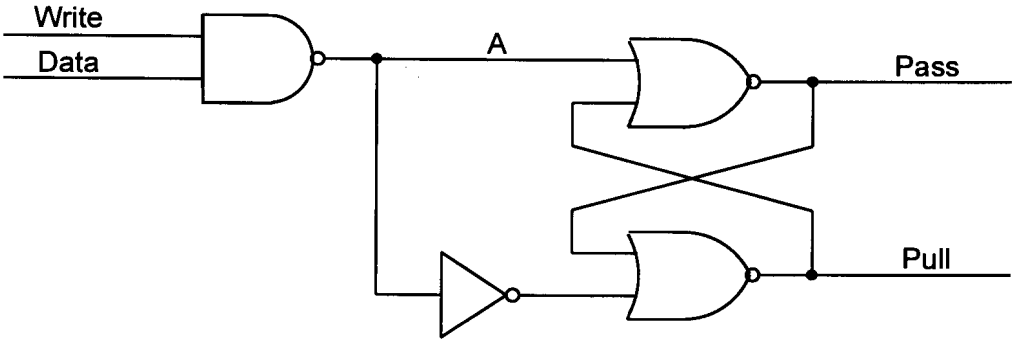


Figure 3.5 Control circuit to activate the switches of Figure 3.2 ensuring that short-circuit current is low.

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.

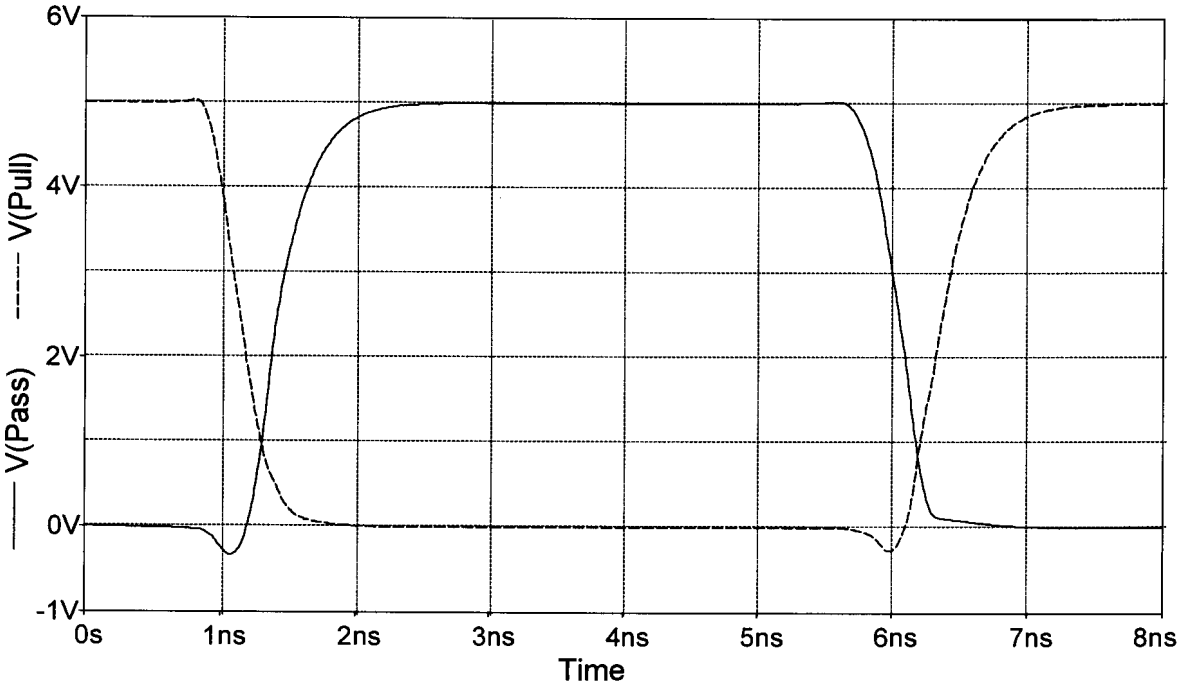


Figure 3.6 Simulation showing the correct delays between the *Pull* and the *Pass* signal.

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.

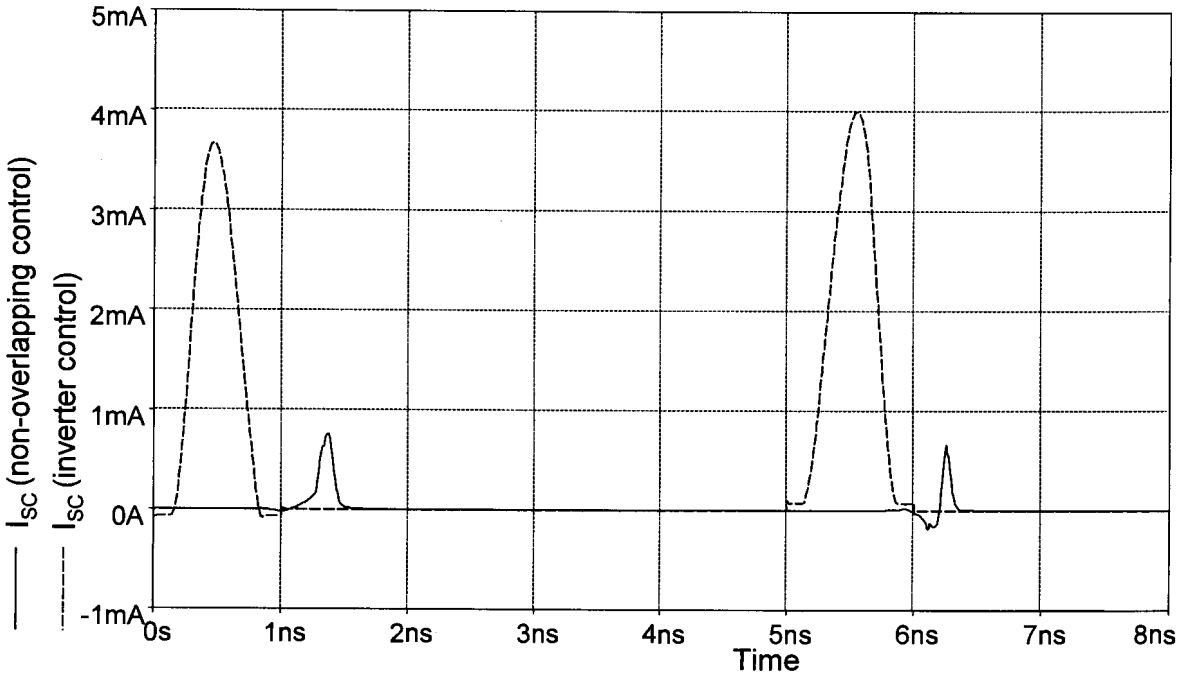


Figure 3.7 Simulation showing the difference in peak short-circuit current between the non-overlapping *Pull* and *Pass* signal control and inverter control.

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).

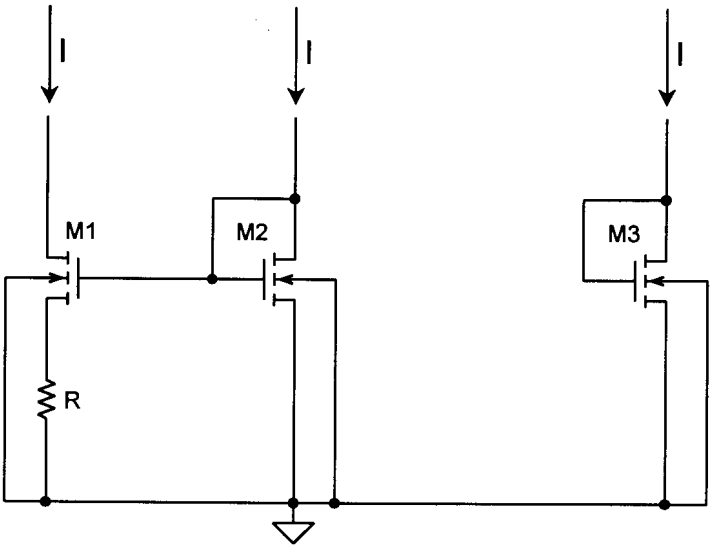


Figure 3.8 Basic topology of the voltage reference circuit.

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-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.

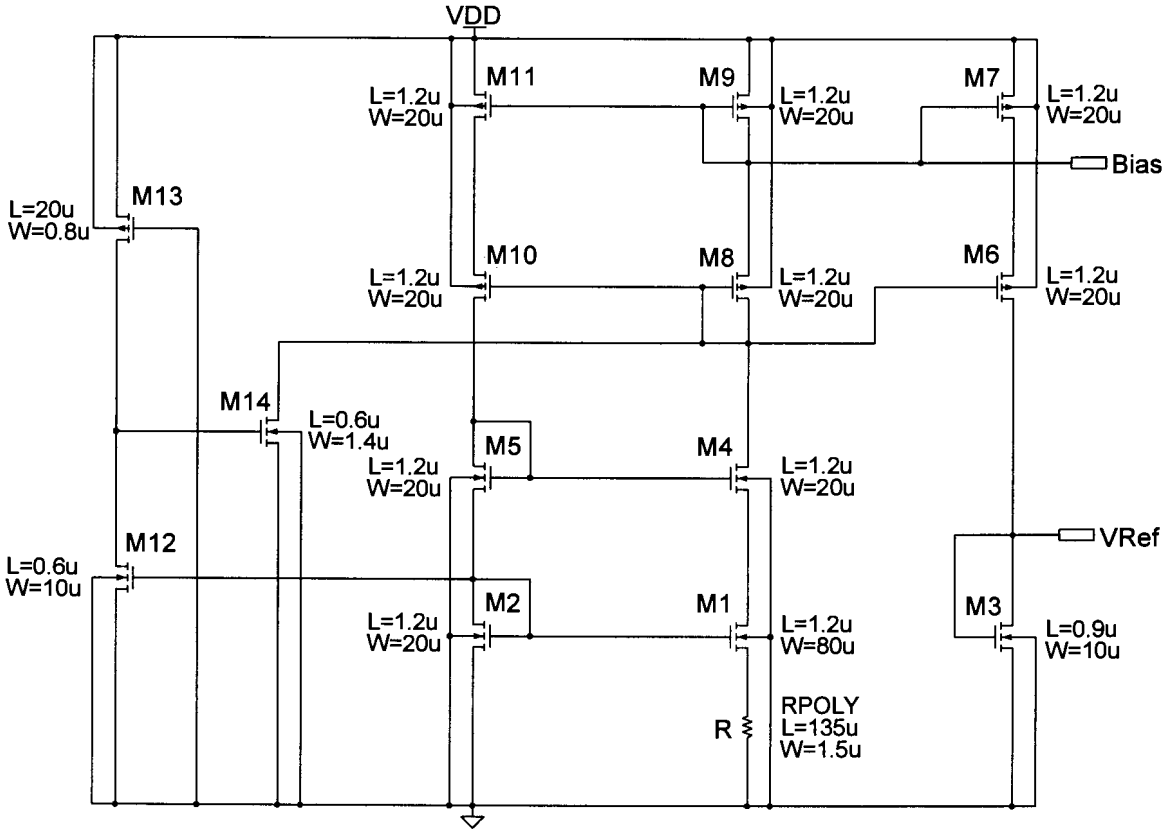


Figure 3.9 Constant transconductance bias circuit and reference voltage generator.

- 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 *DIO*-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 *DIO*-line is 4.05pF, and that the total load for one *DIO*-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.

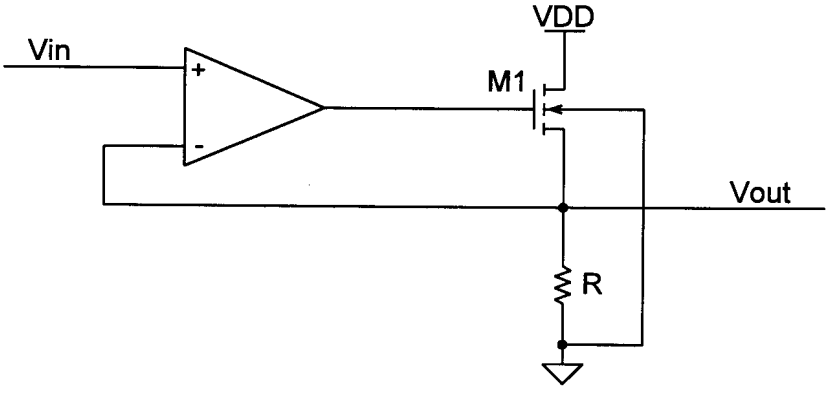


Figure 3.10 Proposed circuit topology for the low-impedance driver.

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 *DIO*-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 *DIO*-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.

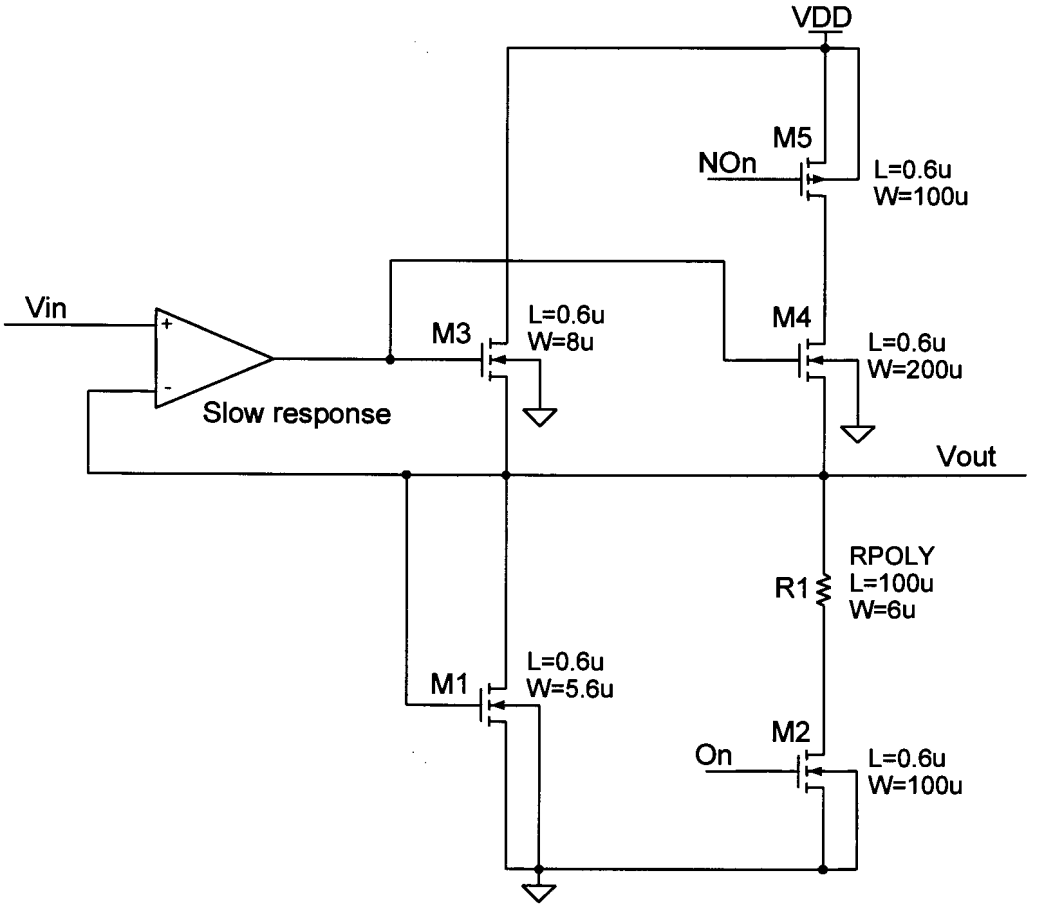


Figure 3.11 Low power, low-impedance driver circuit.

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 *DIO*-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.

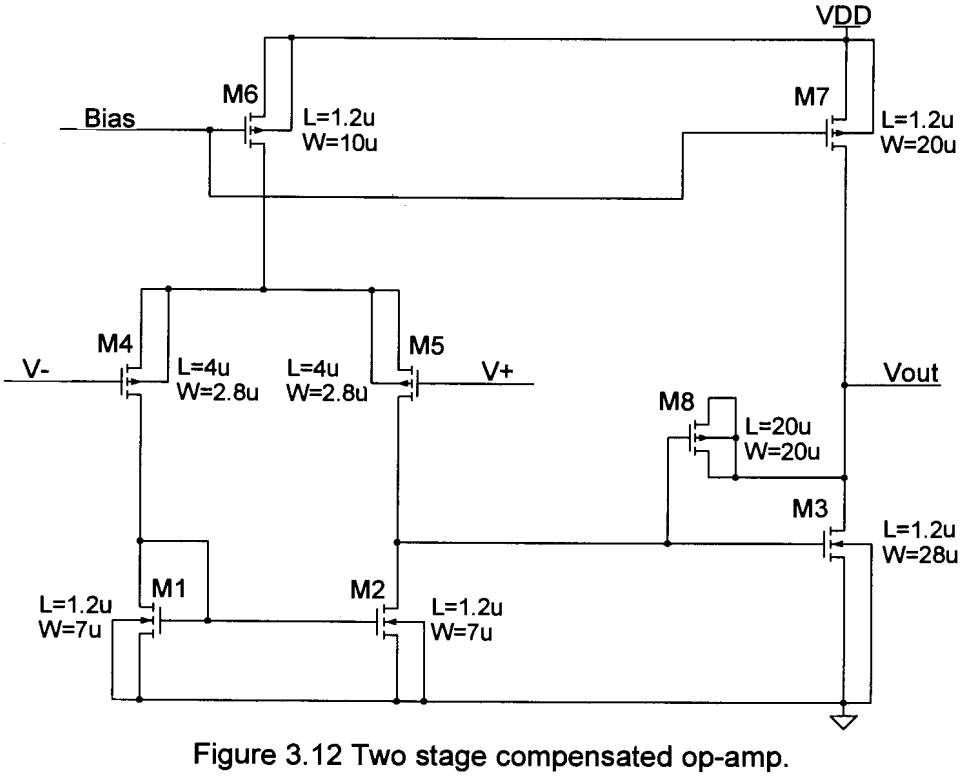


Figure 3.12 Two stage compensated op-amp.

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

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

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

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

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.

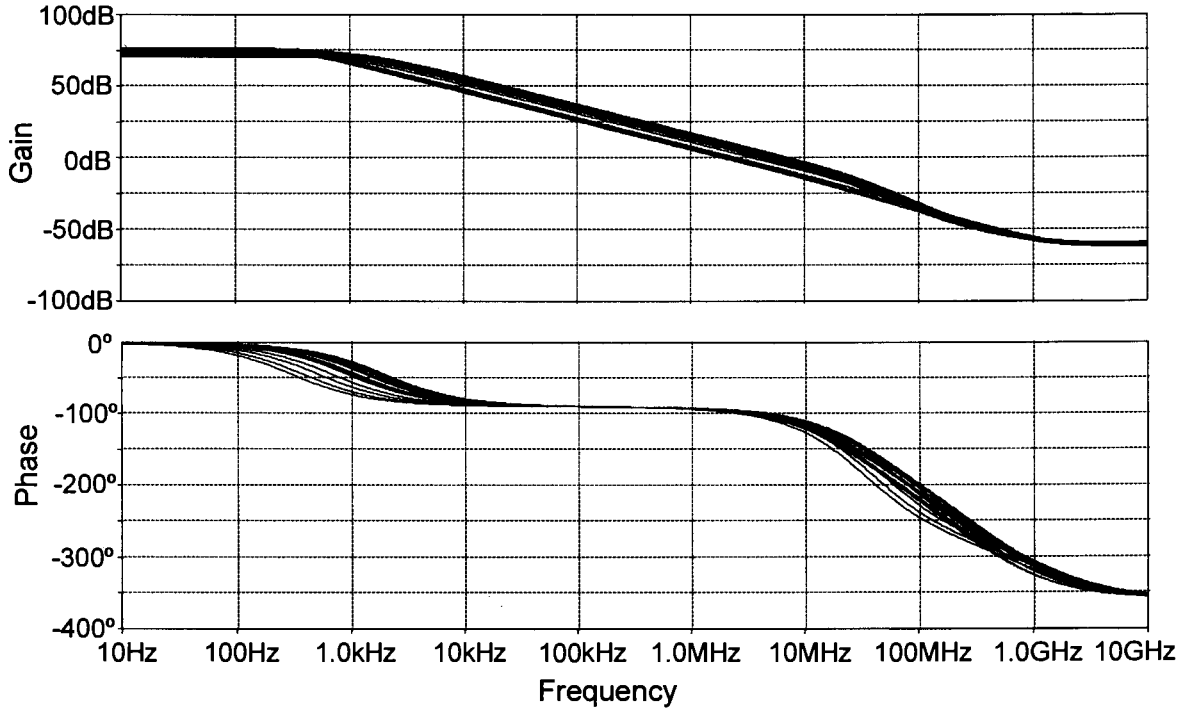


Figure 3.13 Op-amp gain and phase response for different process conditions.

The typical mean performance figures are:

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

3.3.8 *DIO*-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 *DIO*-lines are switched and their capacitance is maximum,
- the minimum capacitance is switched, that is a single *DIO*-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 *DIO*-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.

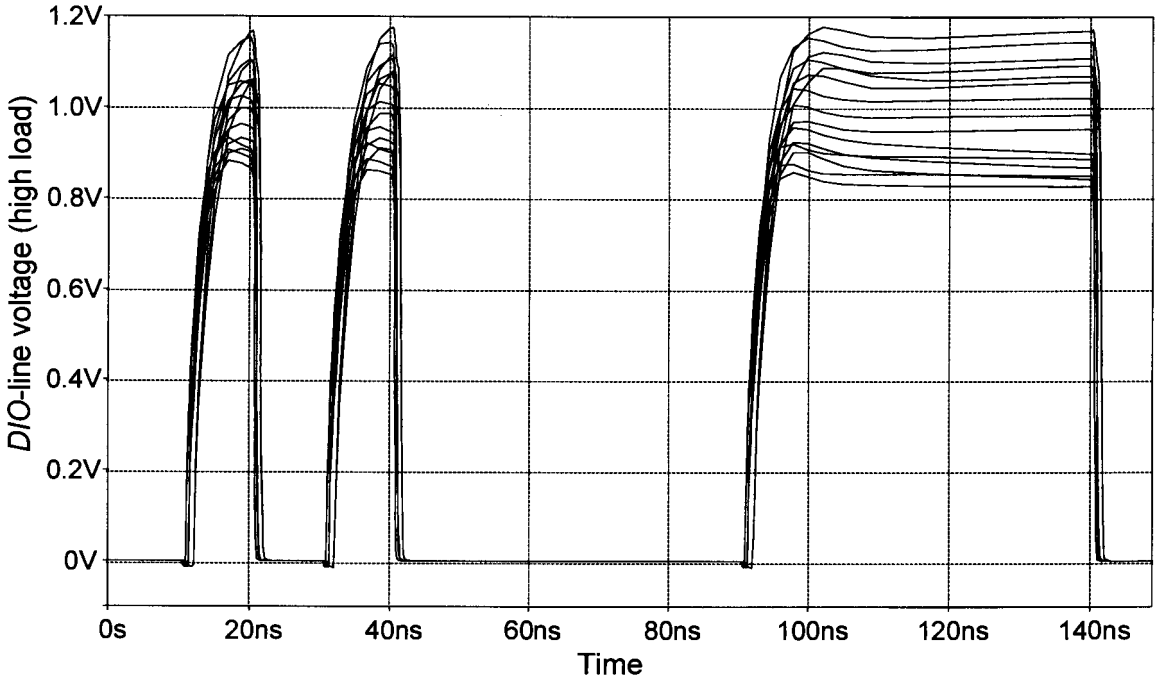


Figure 3.14 *DIO*-line voltage for the maximum load condition across all process models.

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.

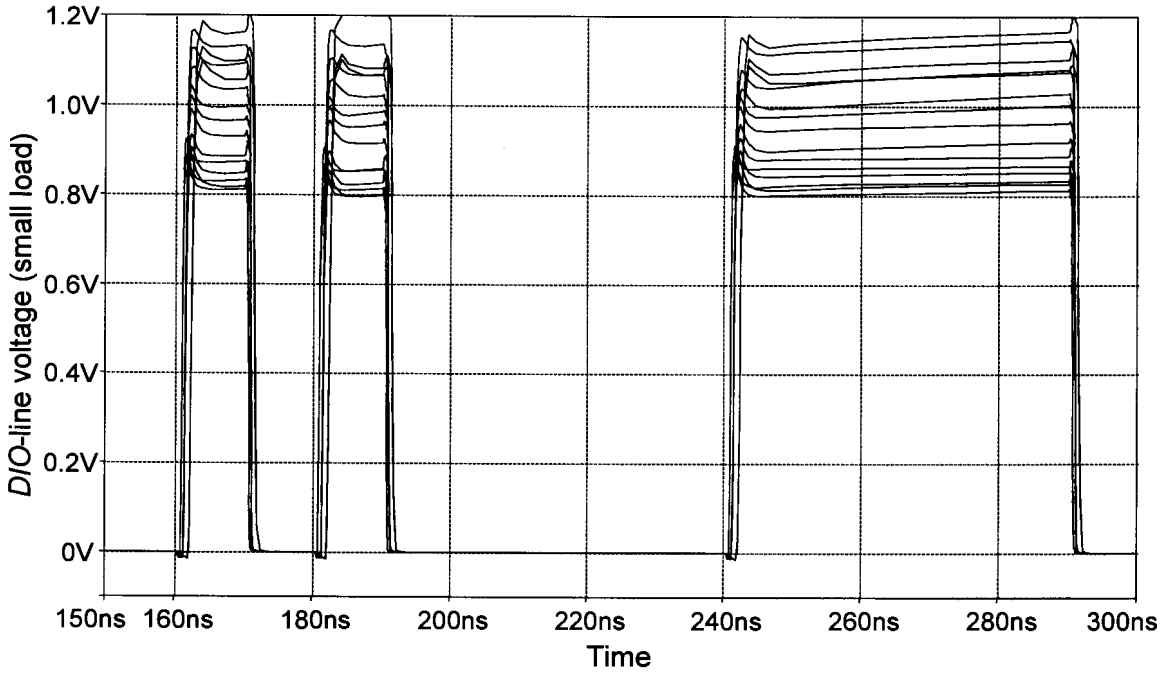


Figure 3.15 *DIO*-line voltage for the minimum load condition across all process models.

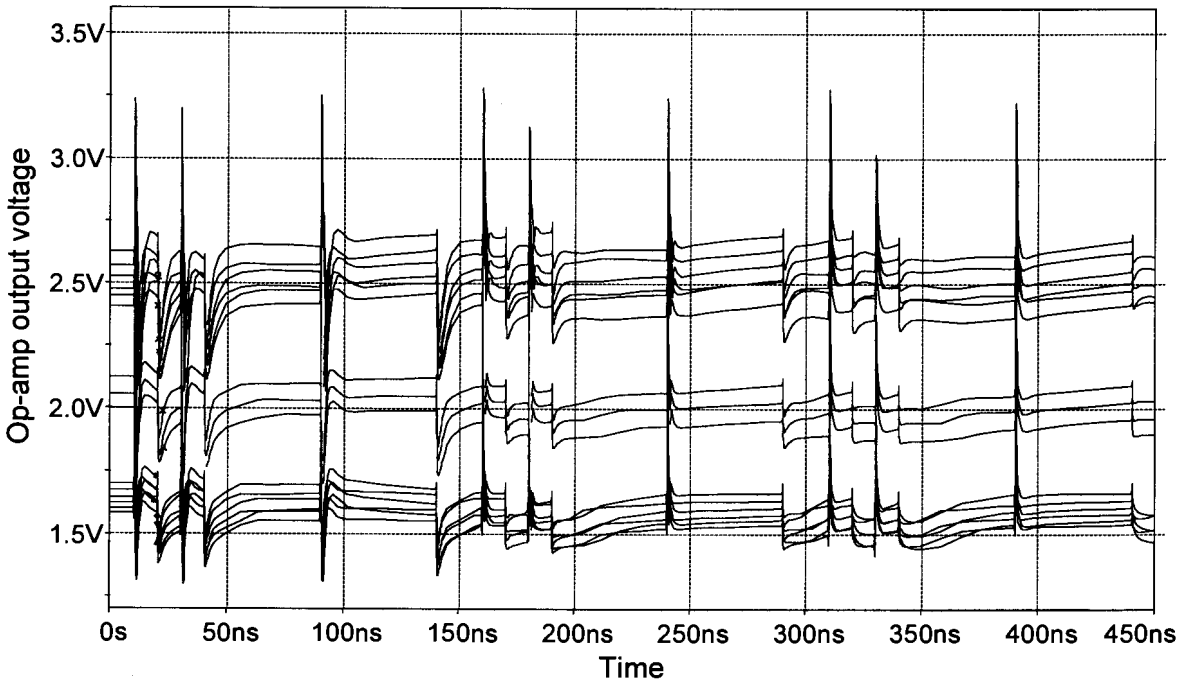


Figure 3.16 Simulation of the operational amplifier output voltage for different processes and load conditions.

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 *DIO*-line driver, as can be seen by comparing its functional block diagram (Figure 3.17) to that of the *DIO*-line driver (Figure 3.1).

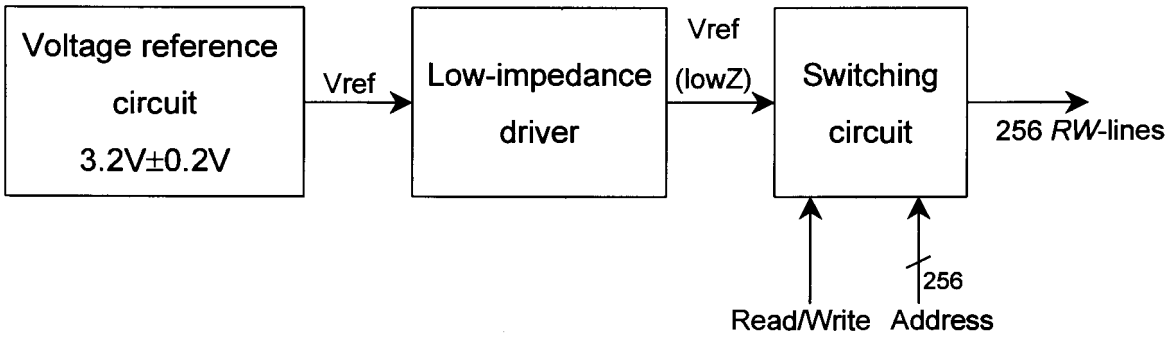


Figure 3.17 Functional block diagram of the *RW*-line driver circuit.

Basically the circuit has to perform an identical function to the *DIO*-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 *DIO*-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.

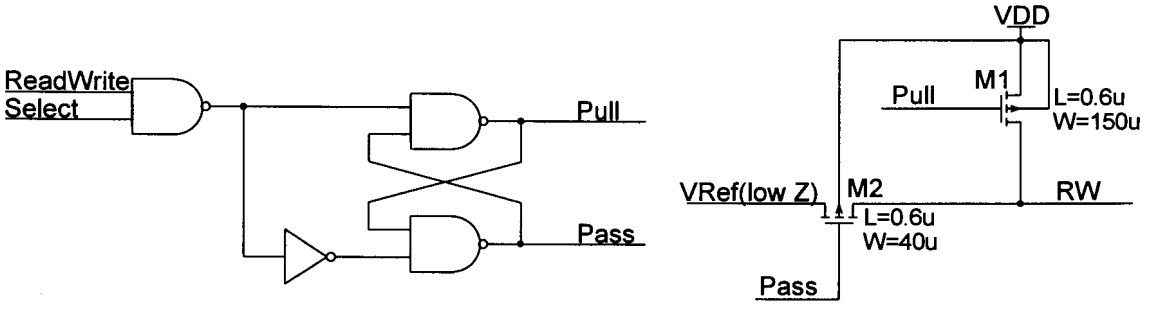


Figure 3.18 Switching circuit for the *RW*-line driver including the control latch.

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 *DIO*-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.

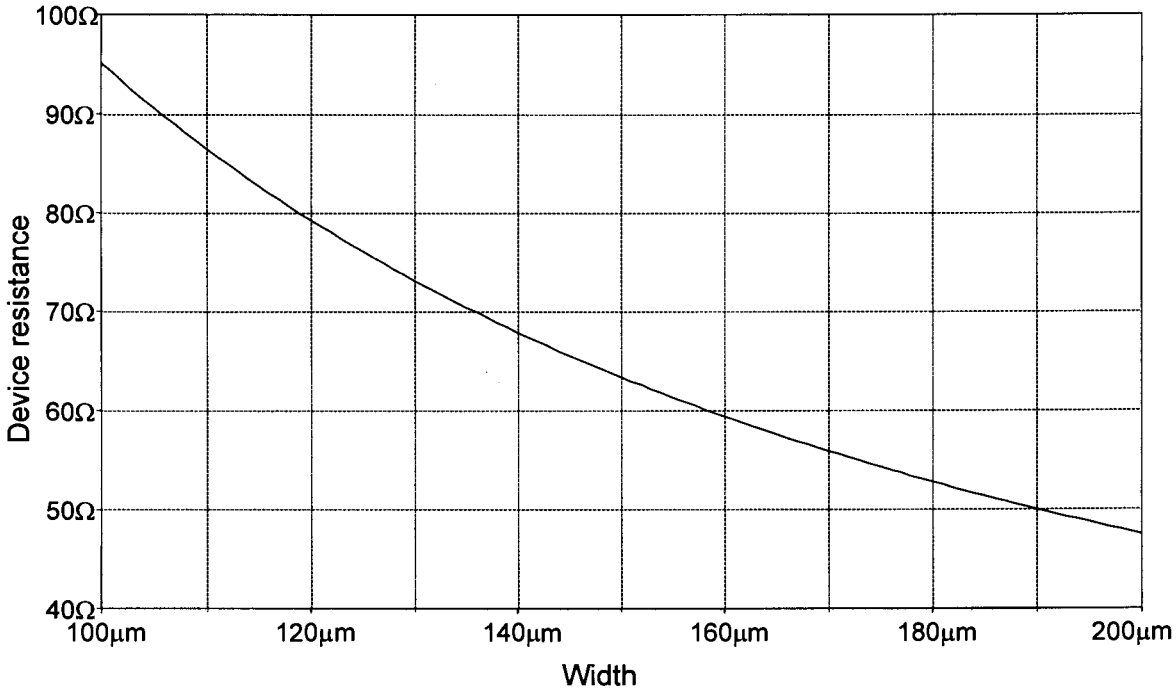


Figure 3.19 PMOS device resistance at V_{DS}=0.05V as a function of the gate width. The gate length is 0.6μm.

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.

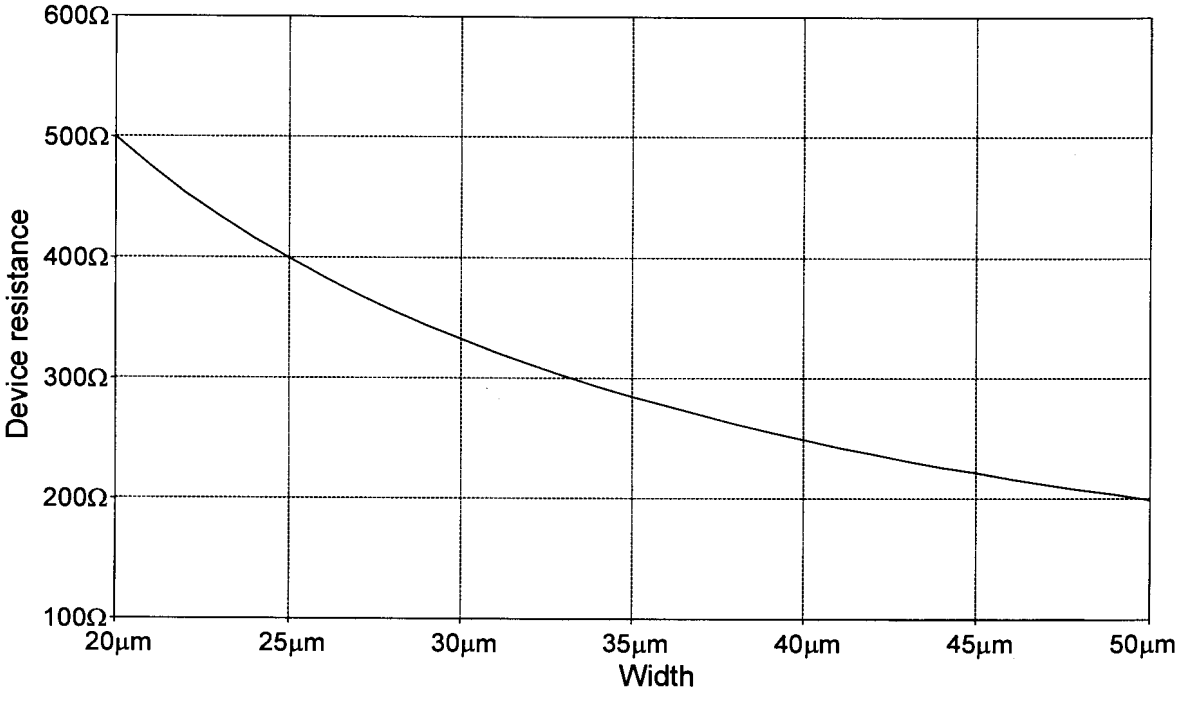


Figure 3.20 PMOS device resistance at V_{DS}=1.8V as a function of the gate width. The gate length is 0.6μm.

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 *DIO*-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.

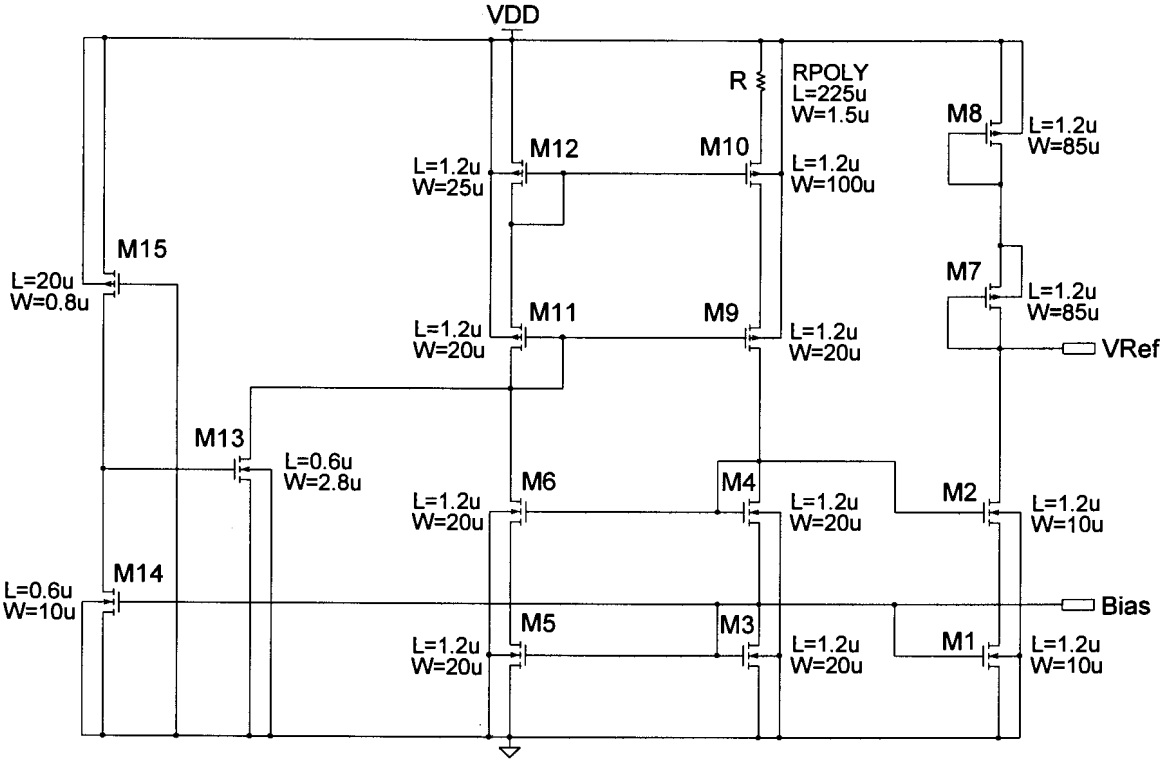


Figure 3.21 Reference voltage generator for the *RW*-line driver.

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.

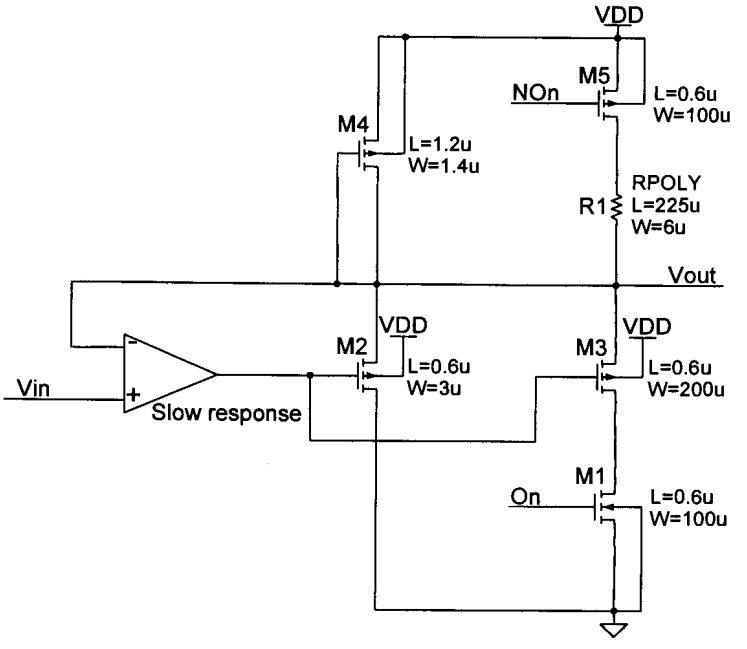


Figure 3.22 Low-impedance driver for the *RW*-line driver circuit.

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.

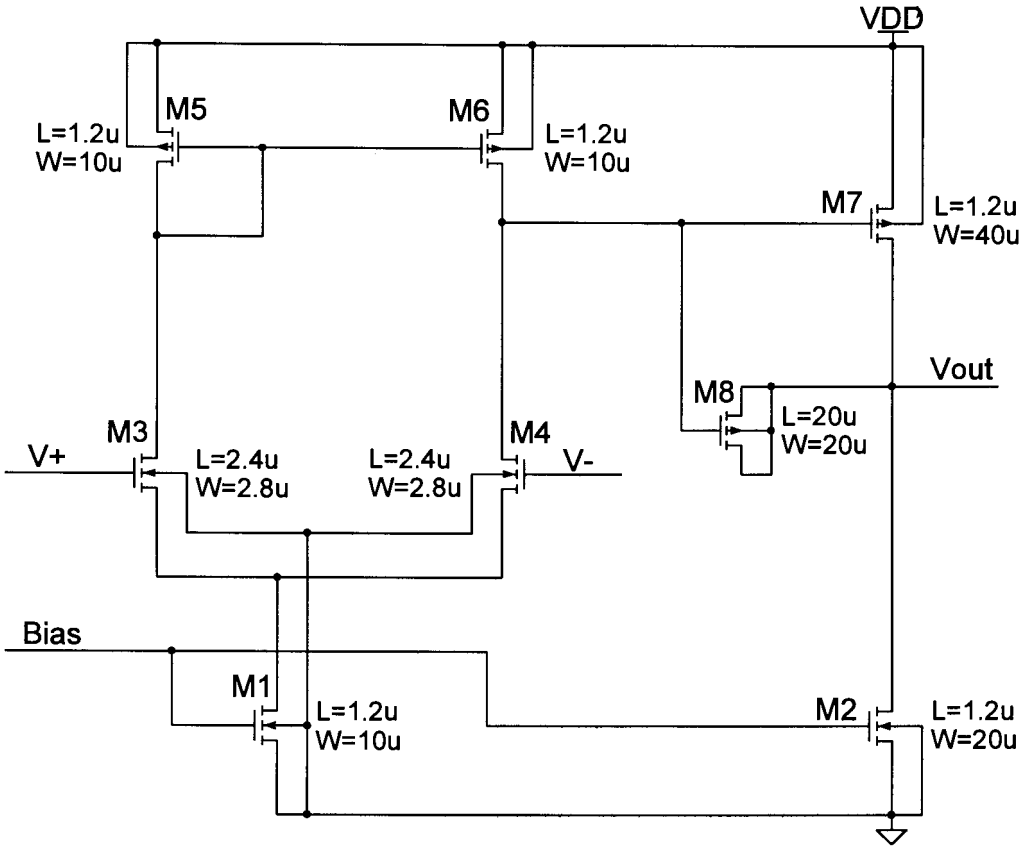


Figure 3.23 Operational amplifier circuit diagram for the *RW*-line driver circuit.

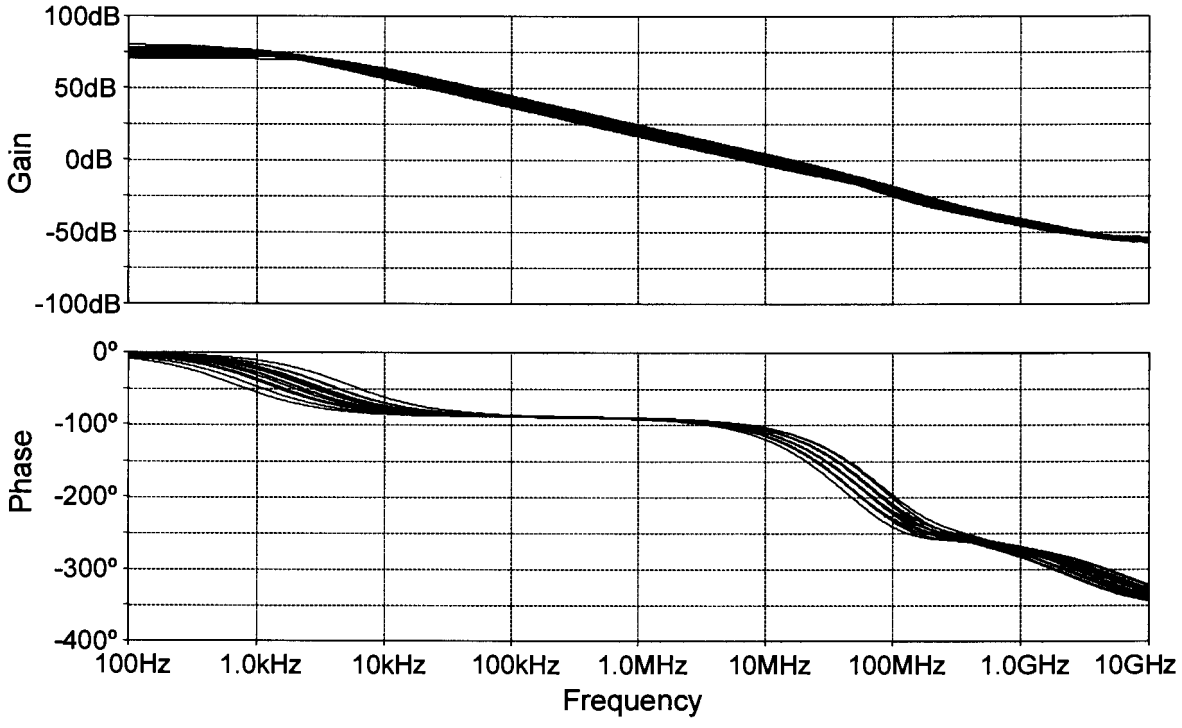


Figure 3.24 Frequency response of the *RW*-line driver operational amplifier.

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 *DIO*-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 *DIO*-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 *DIO*-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.

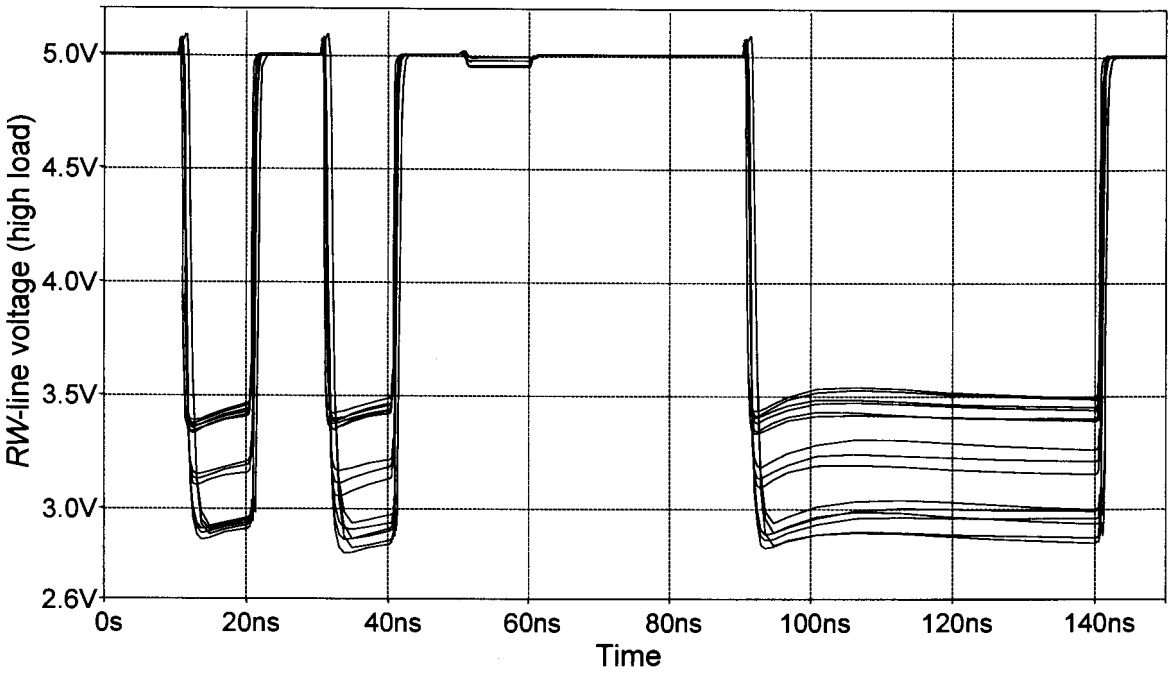


Figure 3.25 *RW*-line voltage for the maximum load condition across all process models.

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 *DIO*-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.

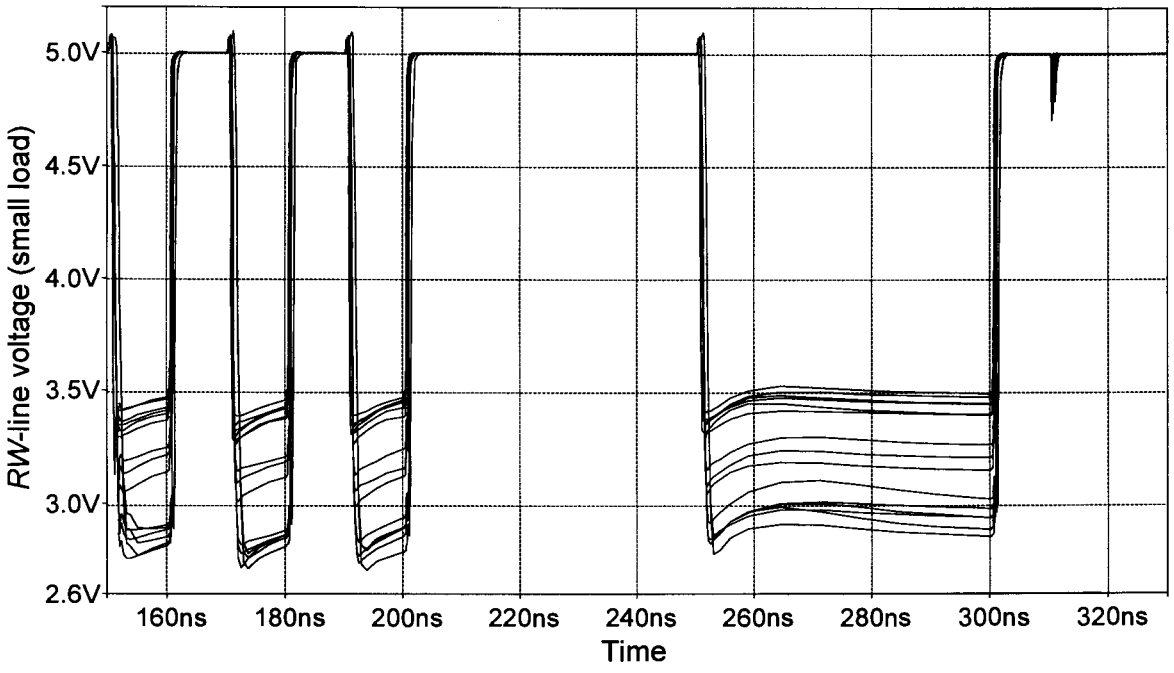


Figure 3.26 *RW*-line voltage for the minimum load condition across all process models.

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.

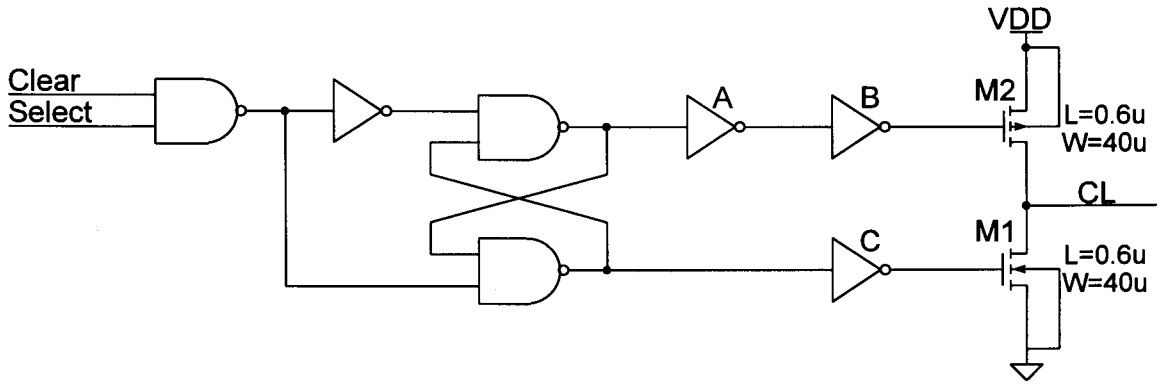


Figure 3.27. Circuit diagram of the *CL*-line driver.

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.

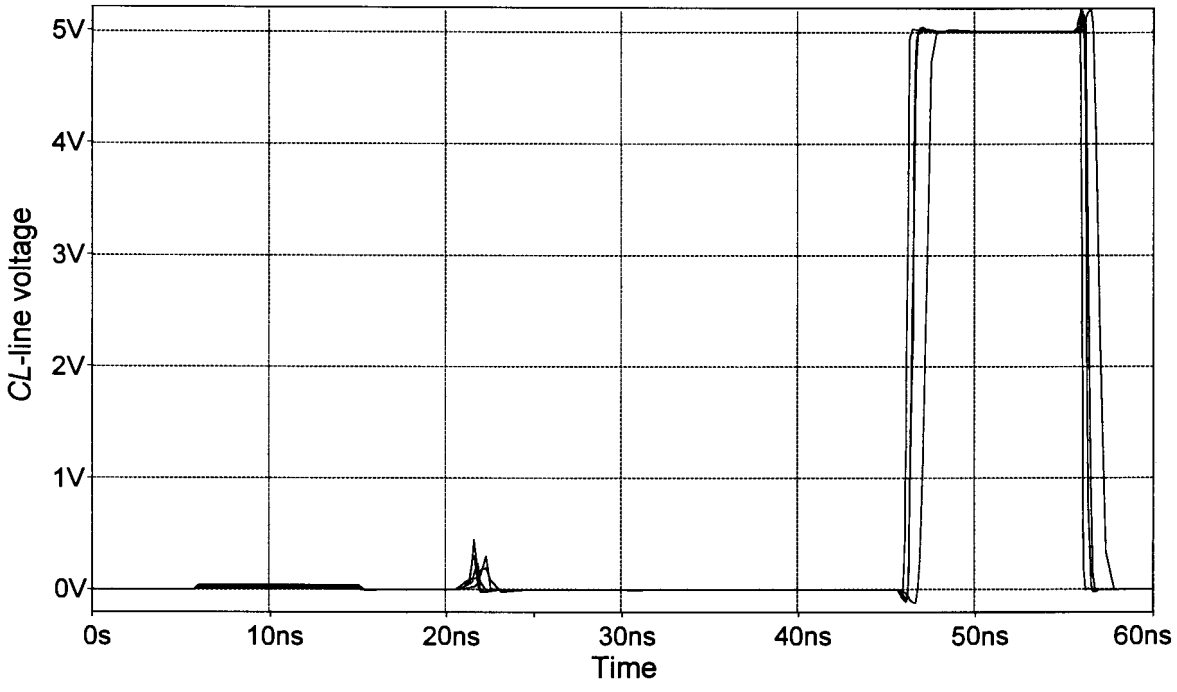


Figure 3.28 Simulation of the *CL*-line voltage for the different process models.

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.

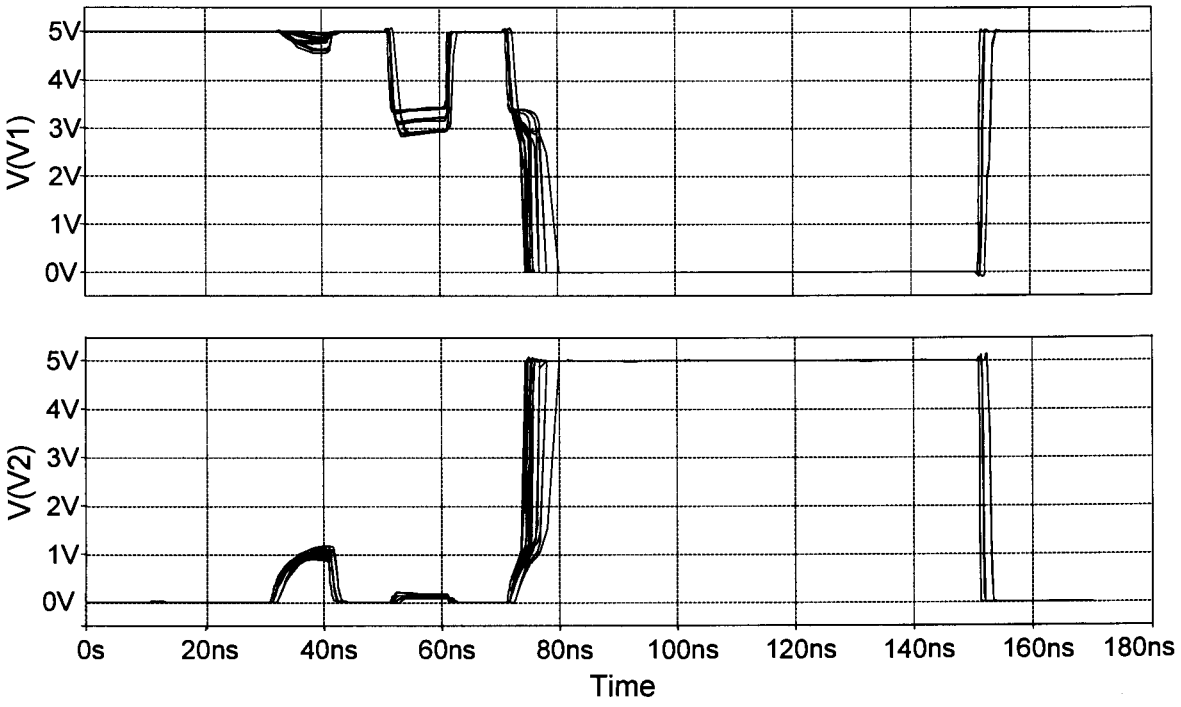


Figure 3.29 Voltages of SRAM internal nodes for different process conditions when the loading of the drivers is maximum.

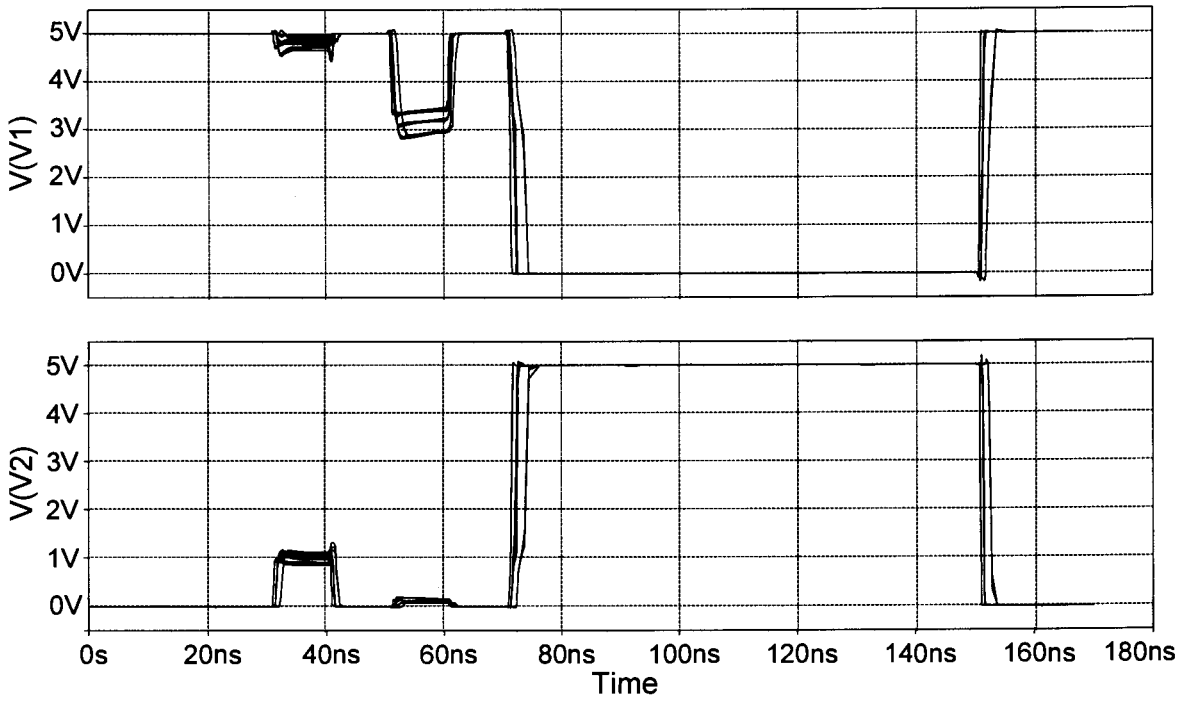


Figure 3.30 Voltages of SRAM internal nodes for different process conditions when the loading of the drivers is minimum.

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 *DIO*-line driver circuit. Comparing Figures 3.14 and 3.15 to Figures 3.25 and 3.26, shows that the *DIO*-line driver response deteriorates more than the *RW*-line driver response as the load conditions worsen, mostly because the *DIO*-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.

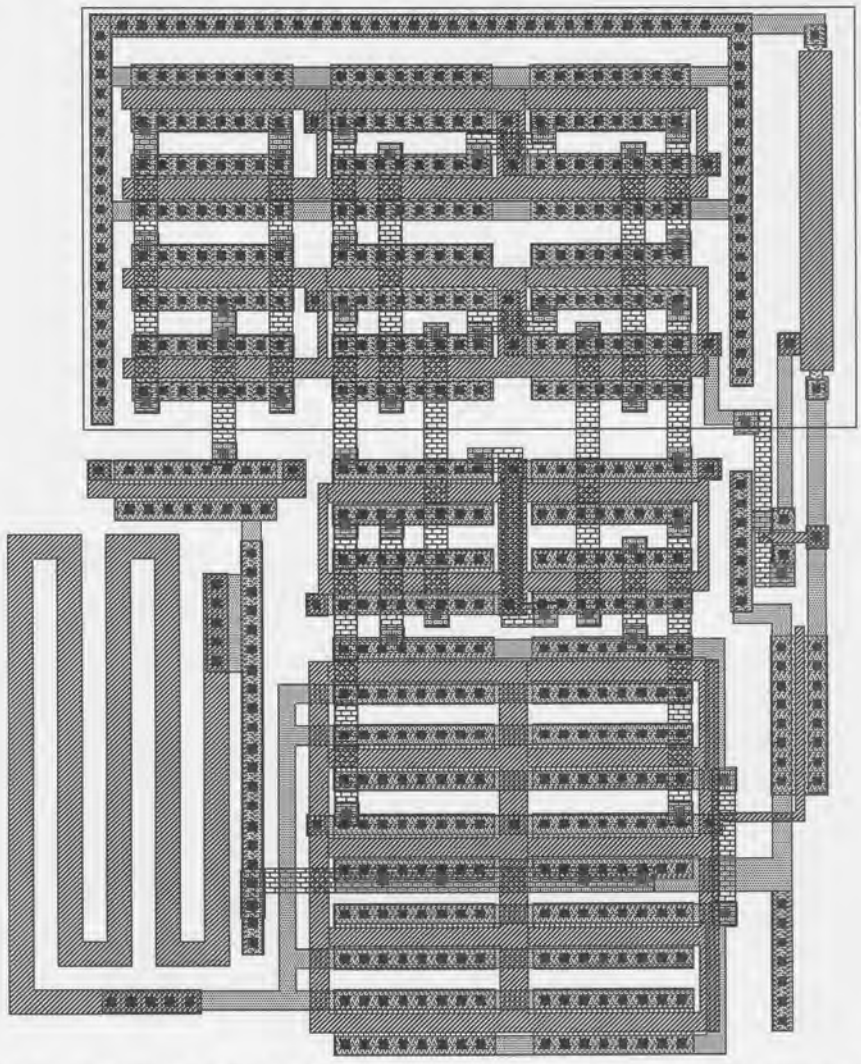


Figure 3.31 Layout of the *DIO*-line driver reference voltage and bias network.

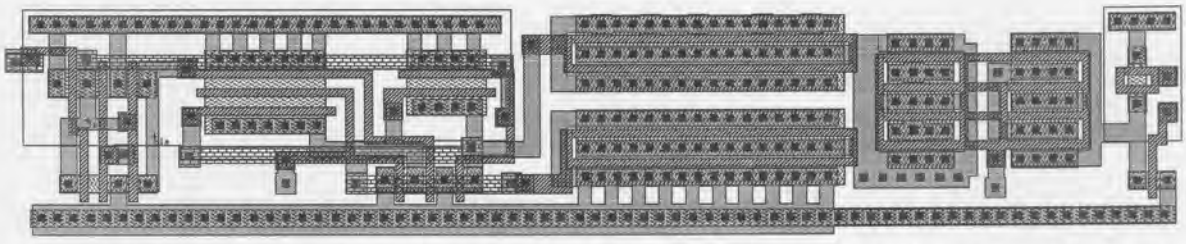


Figure 3.32 Layout of the *DIO*-line driver switch circuit with some peripherals added that are discussed in Chapter 4.

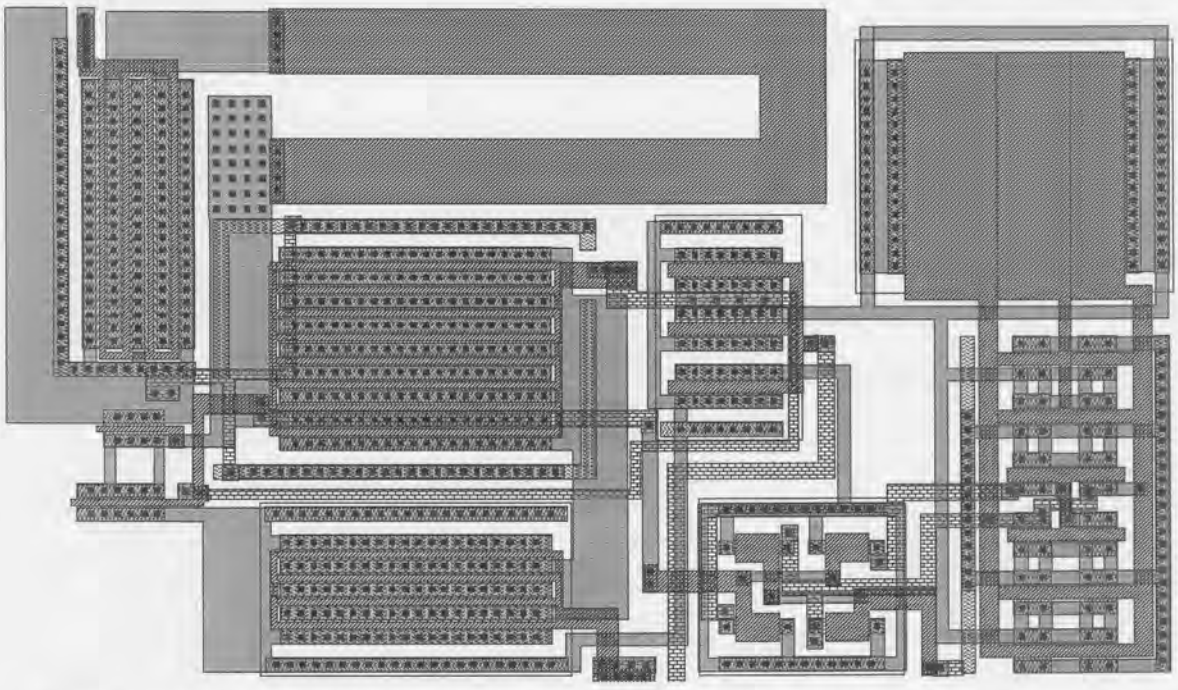


Figure 3.33 Layout of the *DIO*-line driver op-amp and low-impedance driver circuit.

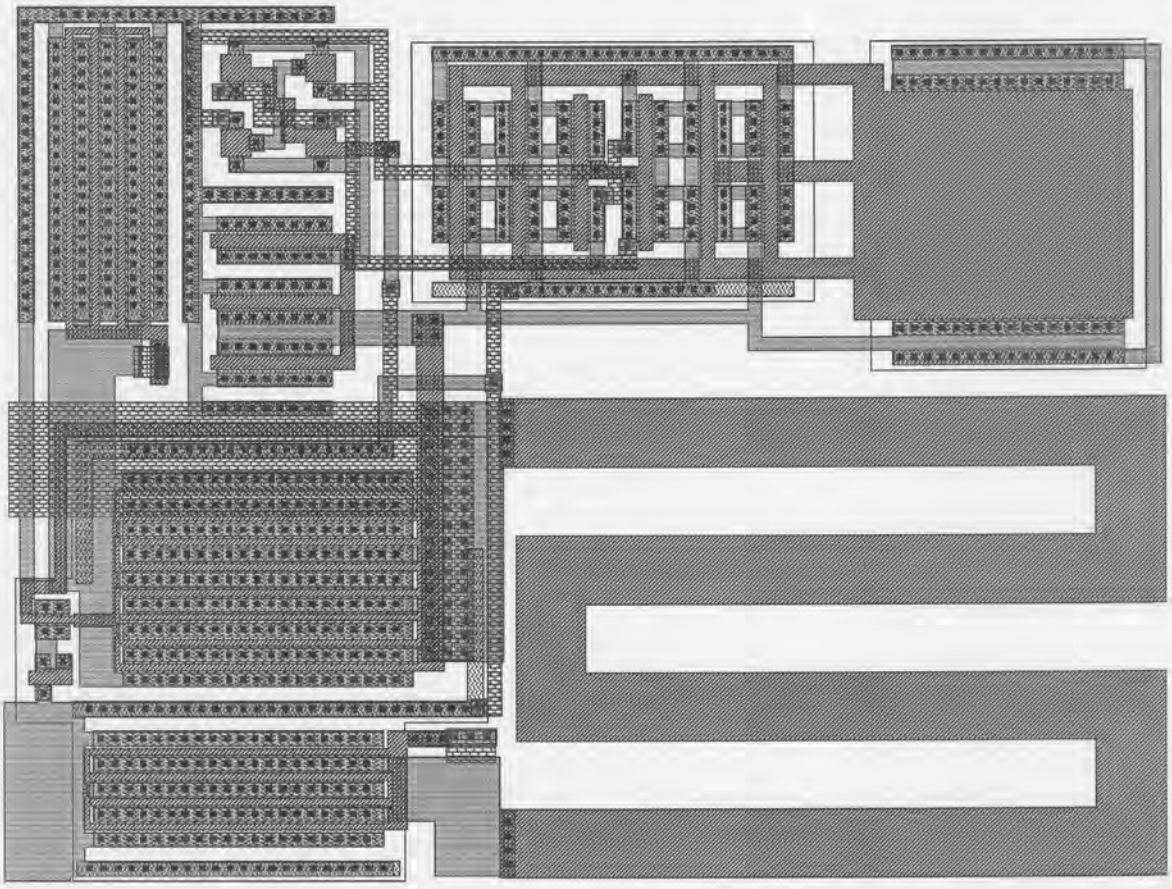


Figure 3.34 Layout of the *RW*-line driver op-amp and low-impedance driver circuit.

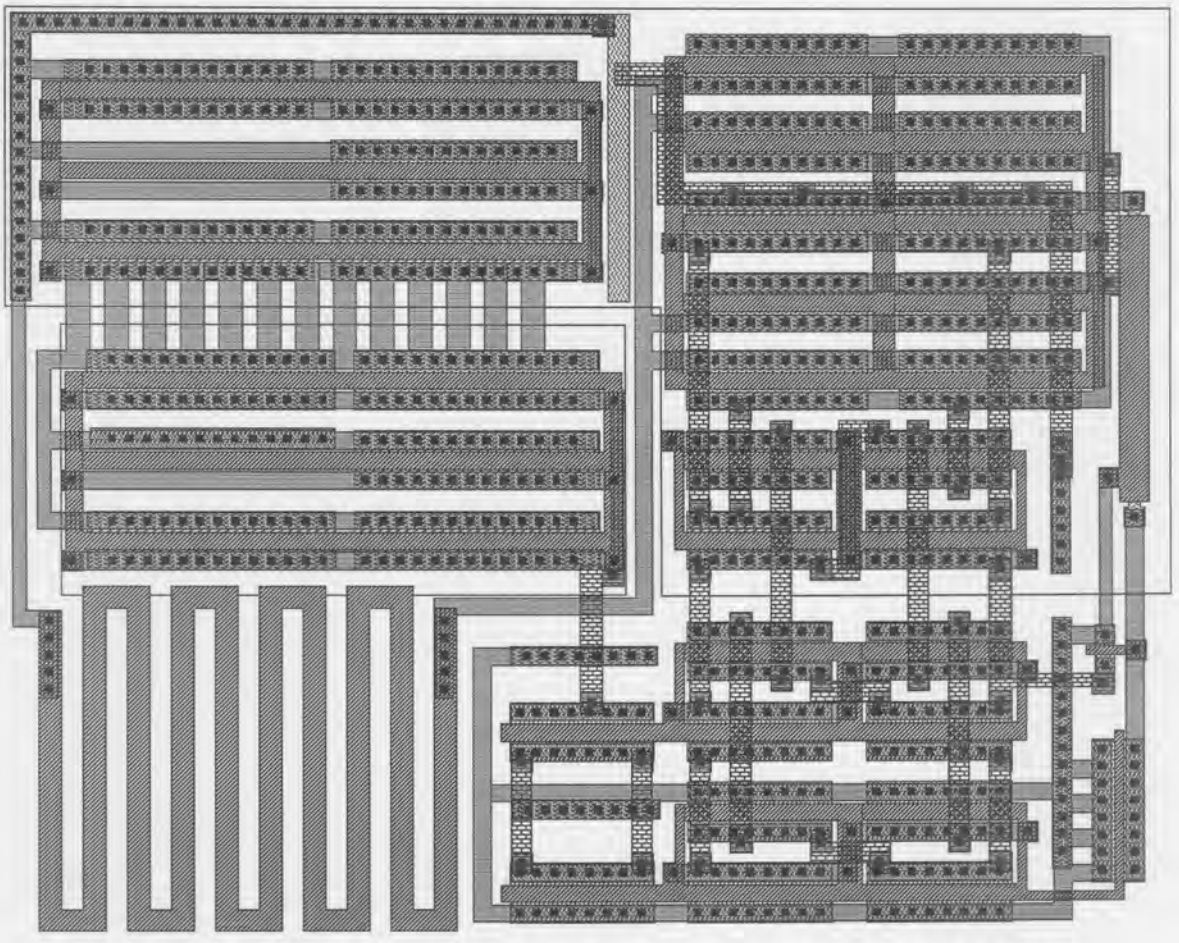


Figure 3.35 Layout of the *RW*-line driver reference voltage and bias network.

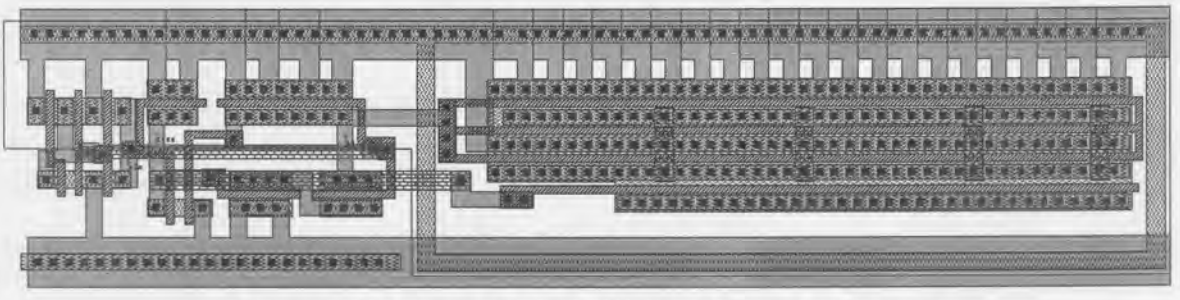


Figure 3.36 Layout of the *RW*-line driver switch circuit.

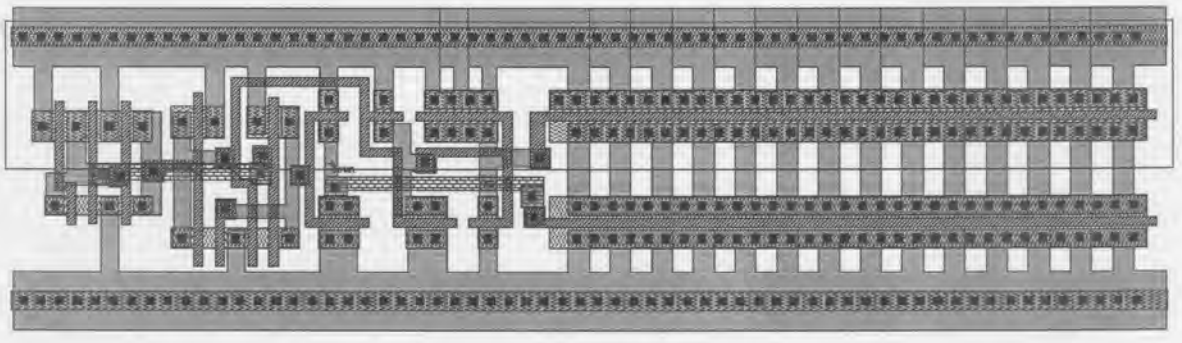


Figure 3.37 Layout of the *CL*-line driver circuit.

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.