4. CURRENT SENSE AMPLIFIER

4.1 INTRODUCTION

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

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

4.2 SENSING SRAM CELLS

4.2.1 Voltage Sense Amplifier System

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

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

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

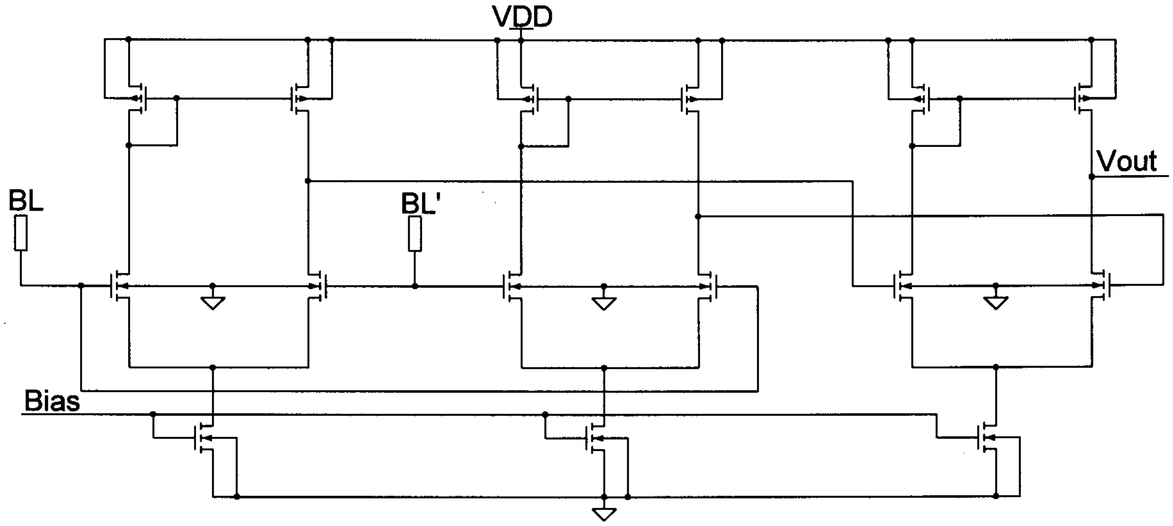


Figure 4.1 Voltage sense amplifier.

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

4.2.2 Current Sense Amplifier Systems

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

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

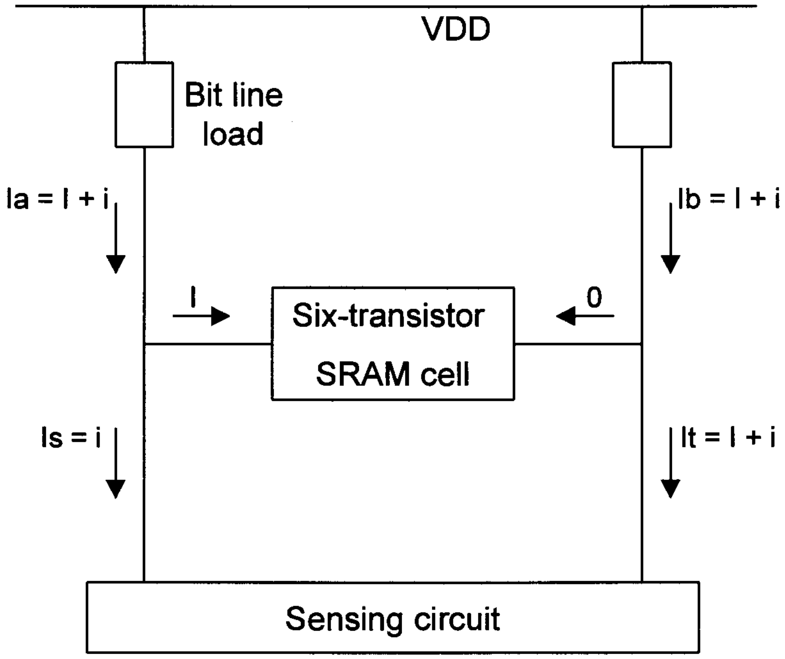


Figure 4.2 Current-mode sensing theory for an SRAM cell.

Current Sensing Circuit

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

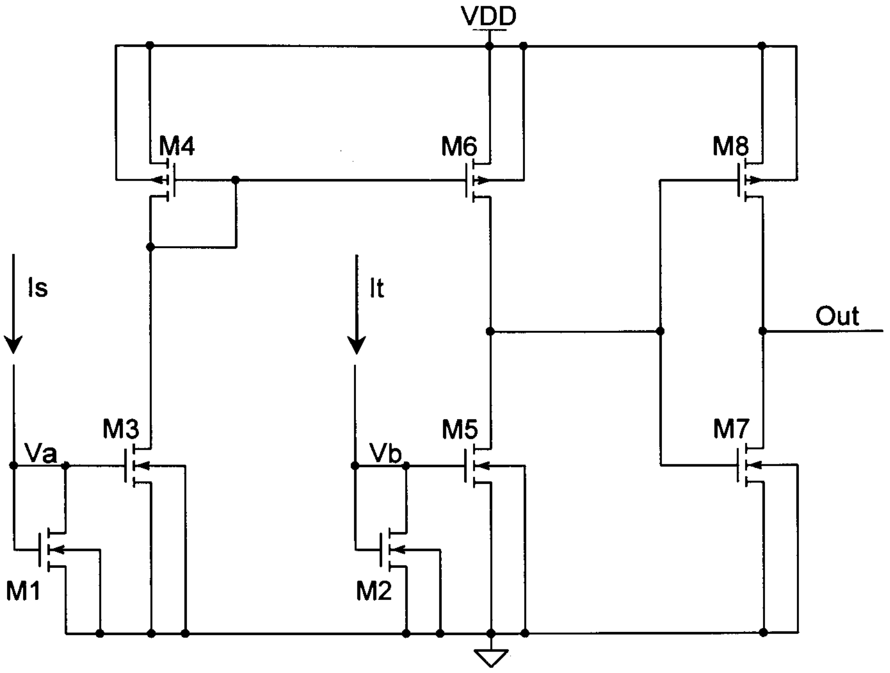


Figure 4.3 Differential current sense circuit.

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

Current Conveyor Circuit

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

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

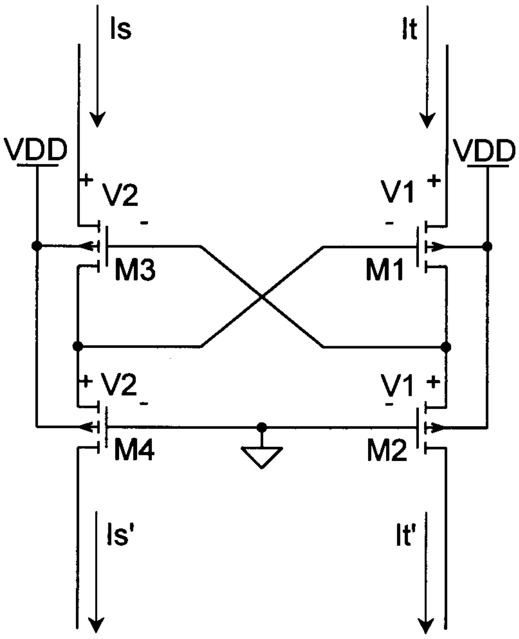


Figure 4.4 Differential input, differential output current conveyor circuit.

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

Latched Sense Amplifier

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

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

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

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

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

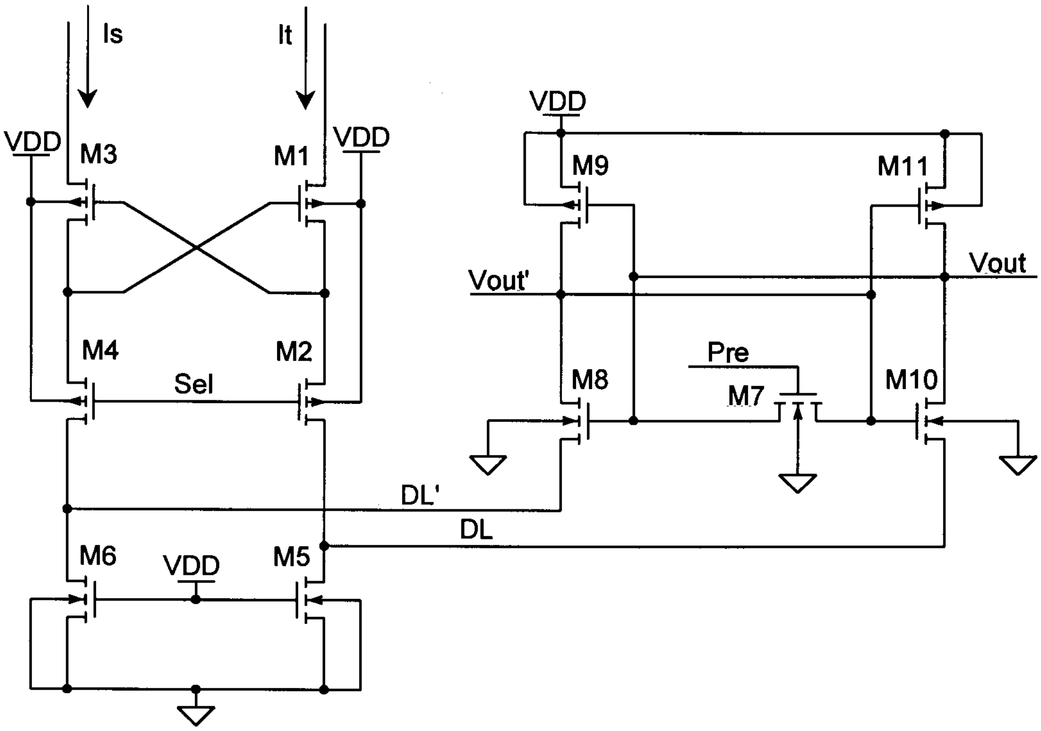


Figure 4.5 Current-mode sensing system using a cross-coupled inverter pair as a sense amplifier.

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

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

Clamped Bit Line Sense Amplifier

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

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

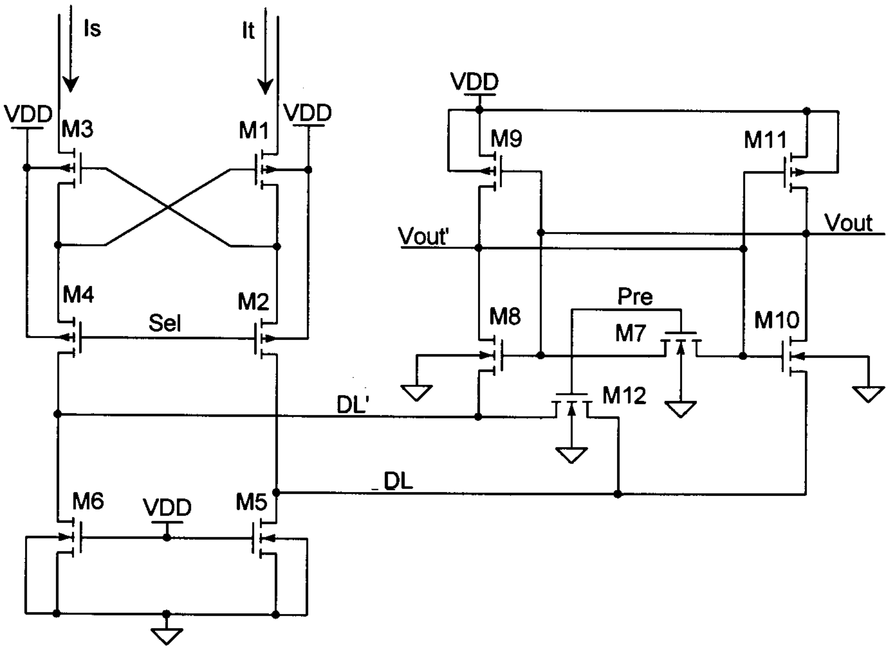


Figure 4.6 Clamped bit line current-mode sense amplifier.

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

4.3 CURRENT SENSE AMPLIFIER DESIGN

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

4.3.1 Reference Current

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

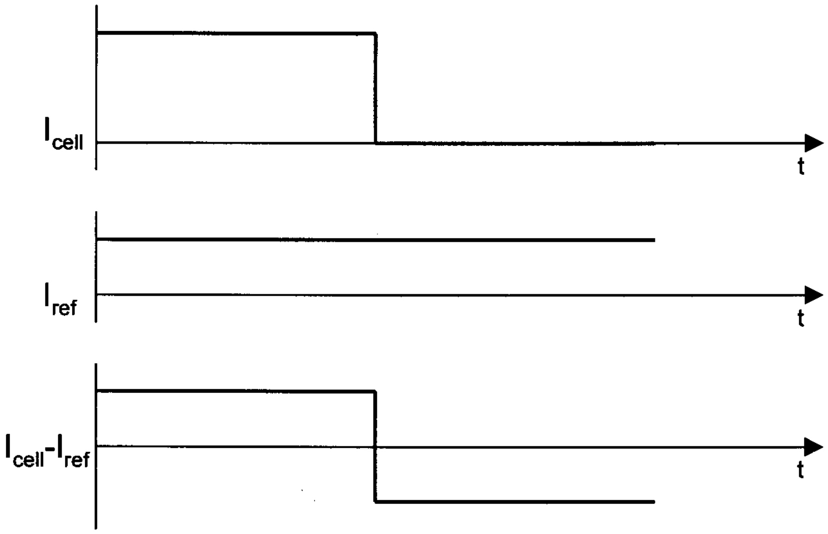
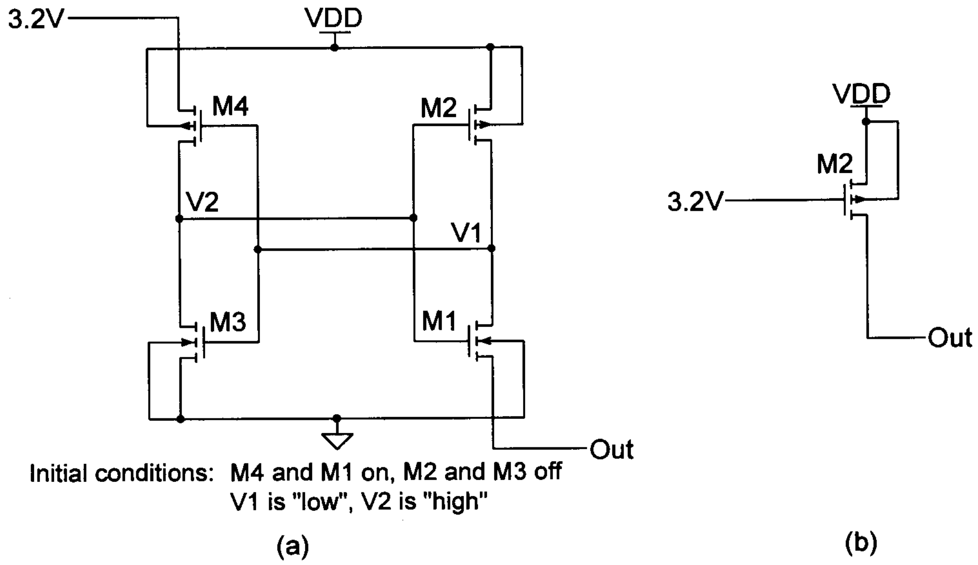


Figure 4.7 Generating a differential current by using a constant reference current of half the cell current amplitude.

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

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



Initial conditions: M4 and M1 on, M2 and M3 off
V1 is "low", V2 is "high"

(a) (b)

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

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

4.3.2 Current Switching

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

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

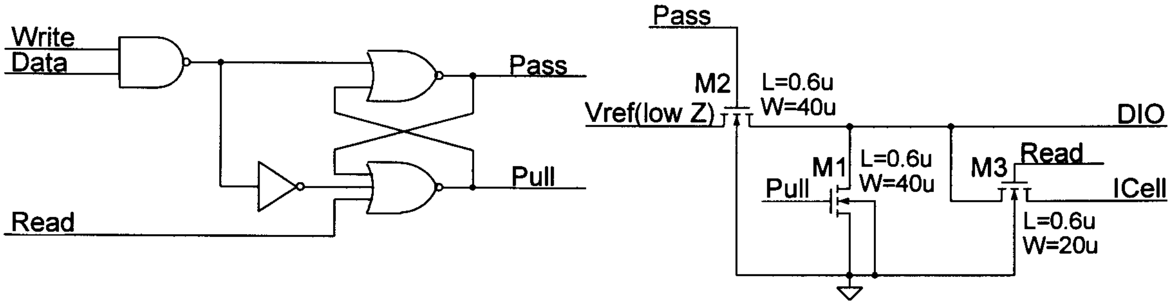


Figure 4.9 Modified *D*/*O*-line driver to allow the pull-down *M1* to be disconnected and the output current to be guided to the *ICell* node.

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

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

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

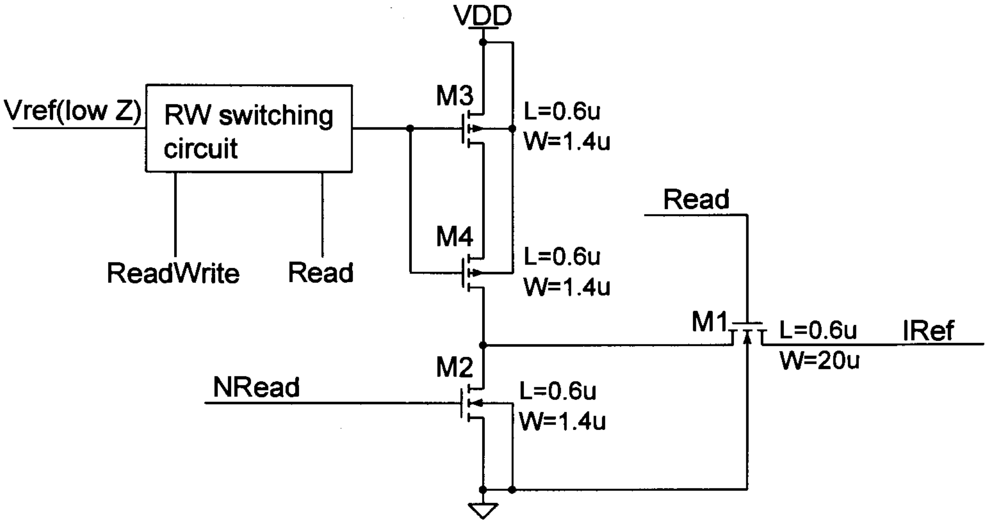


Figure 4.10 Reference current generator circuit.

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

4.3.3 Current Conveyor and Loads

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

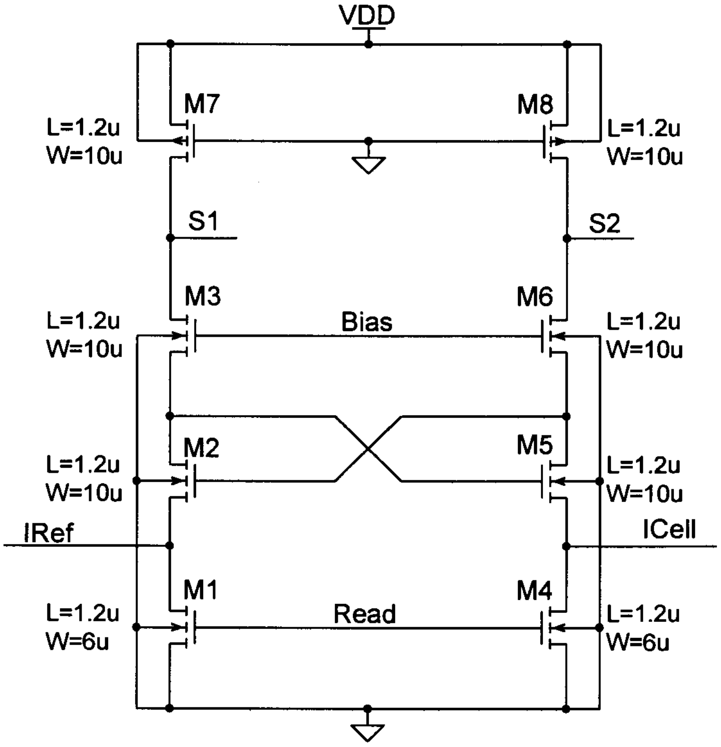


Figure 4.11 Current conveyor circuit for the sensing system.

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

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

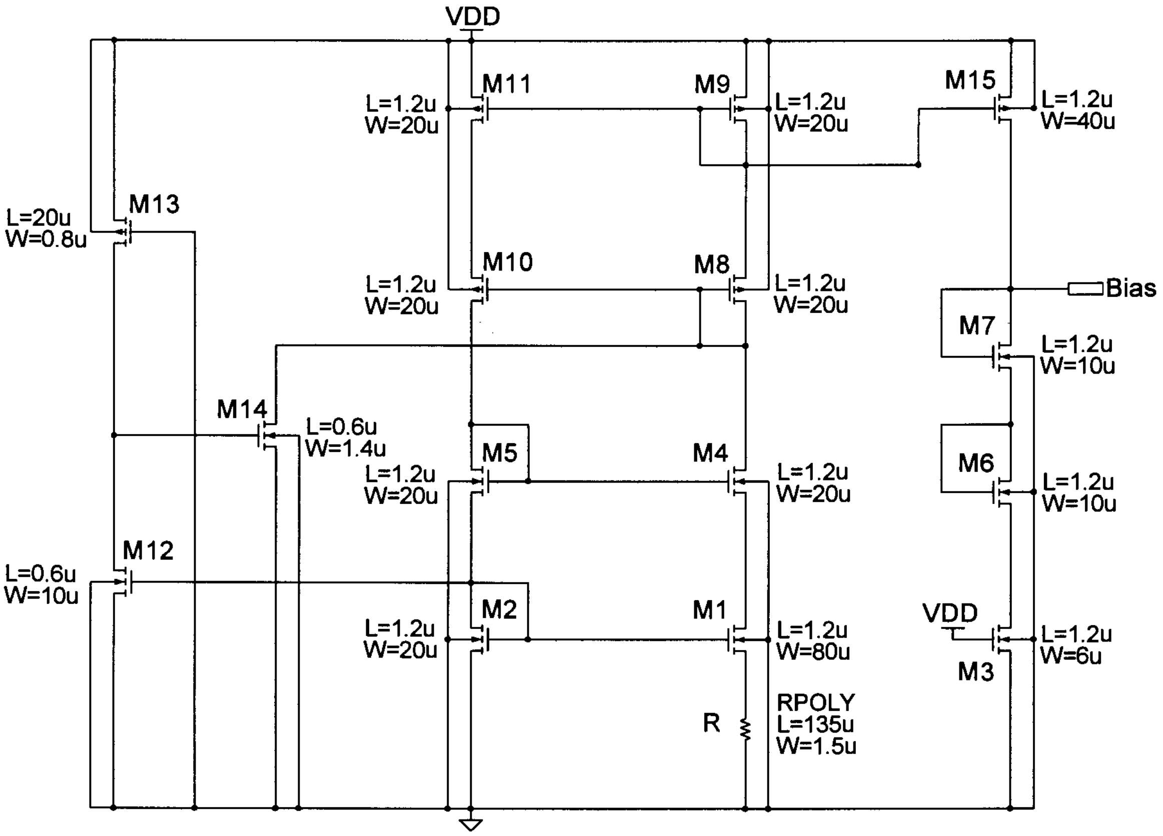


Figure 4.12 Bias network for the current conveyor.

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

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

4.3.4 Clamped Bit Line Sense Amplifier

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

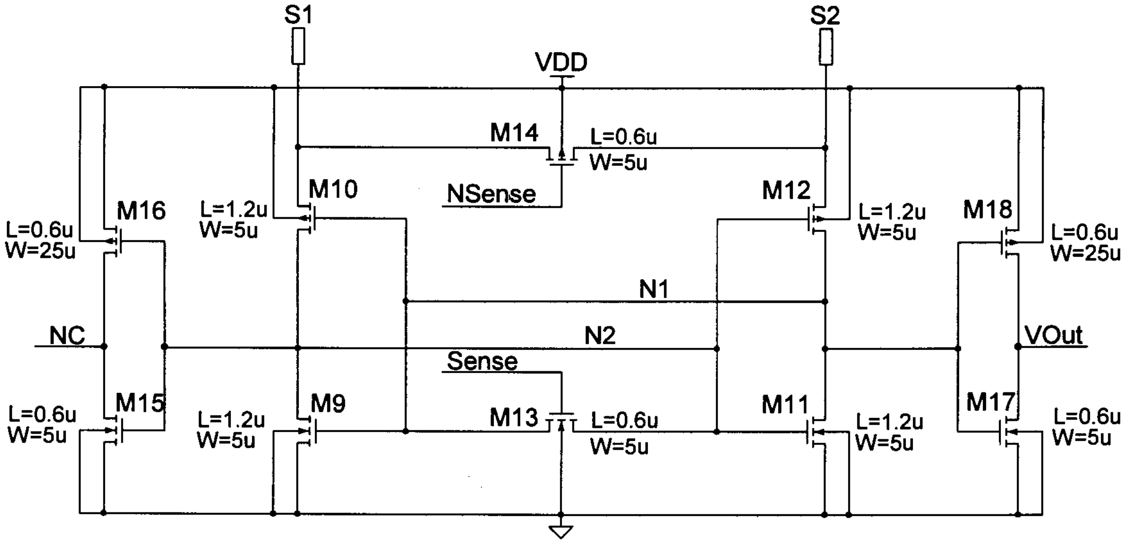


Figure 4.13 Clamped bit line current sense amplifier.

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

4.4 SIMULATION

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

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

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

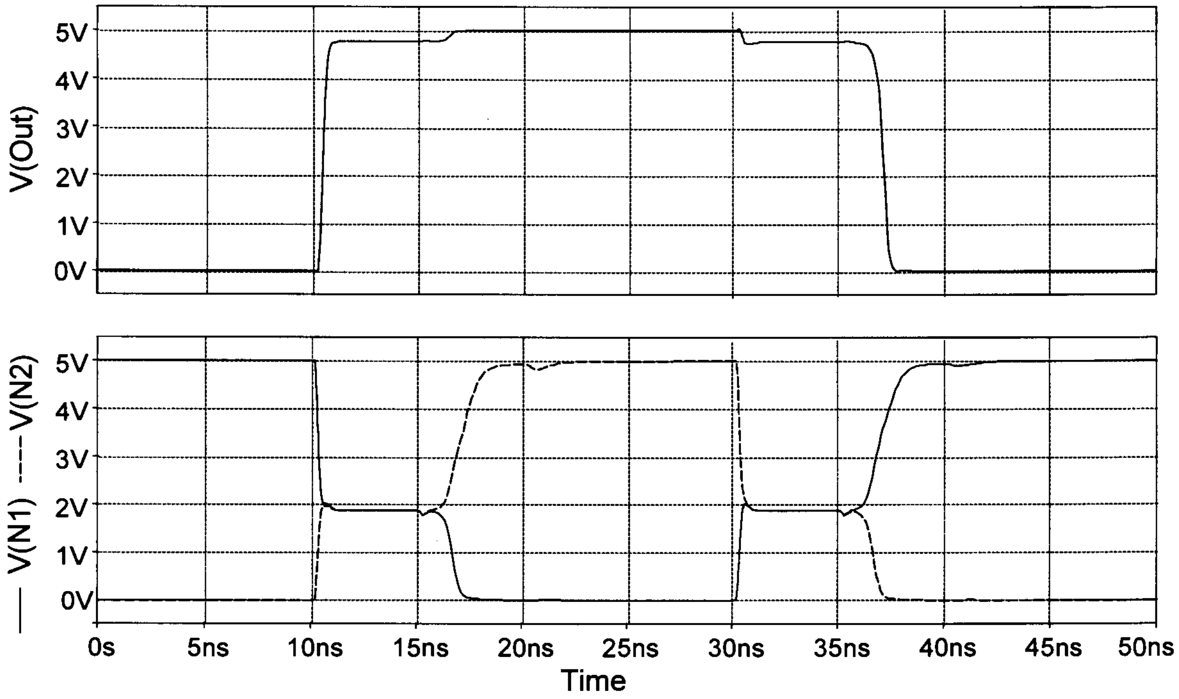


Figure 4.14 Simulated output voltage and internal voltages of the sense amplifier. The metastable condition can clearly be seen.

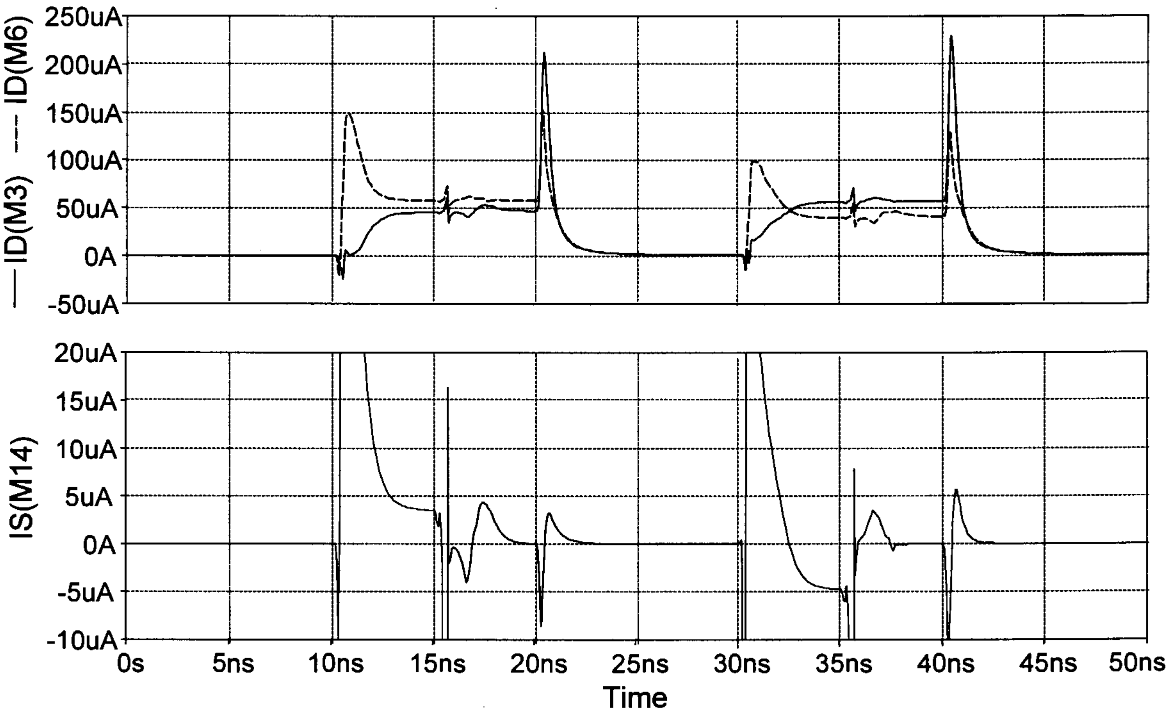


Figure 4.15 Simulated current conveyor currents and bit line clamp device current. The device indices refer to Figures 4.11 (top) and 4.13 (bottom).

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

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

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

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

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

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

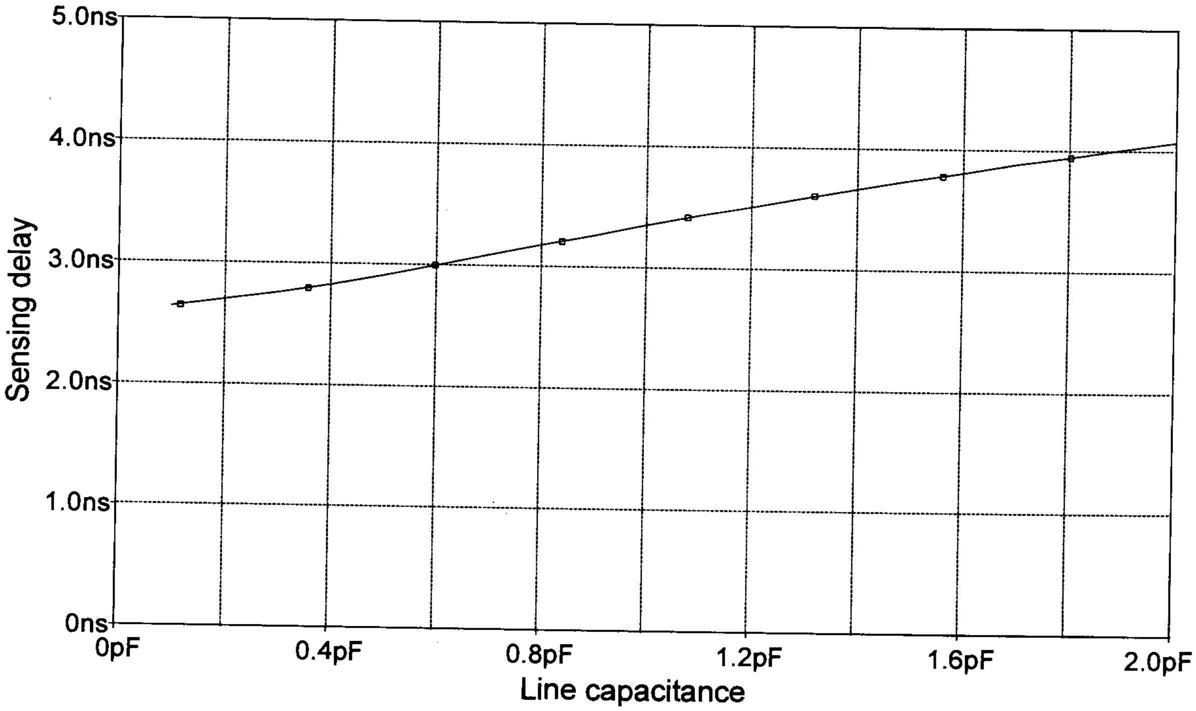


Figure 4.16 Read to valid differential current as a function of the capacitance of the input lines to the current conveyor.

4.5 LAYOUT

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

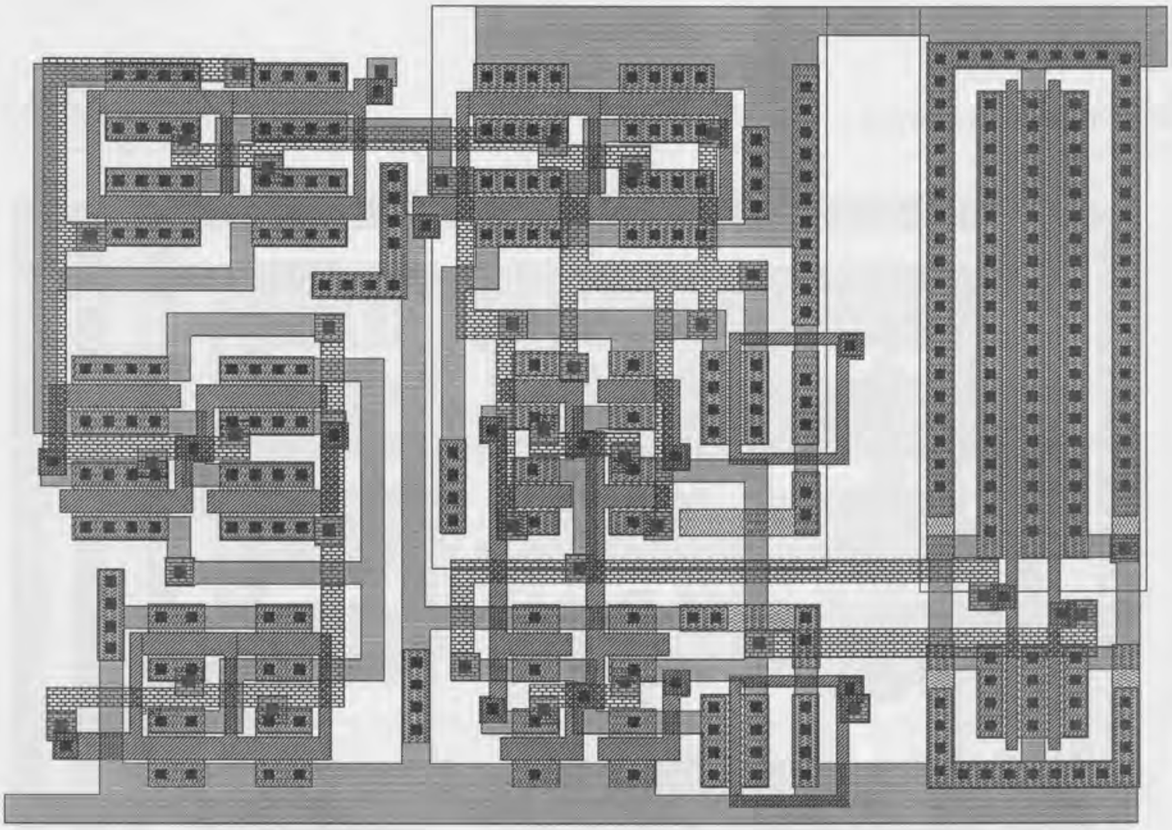


Figure 4.17 Layout of a single current conveyor and clamped bit line current sense amplifier. The devices on the right are the noisy and thus heavily guarded output inverters.

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

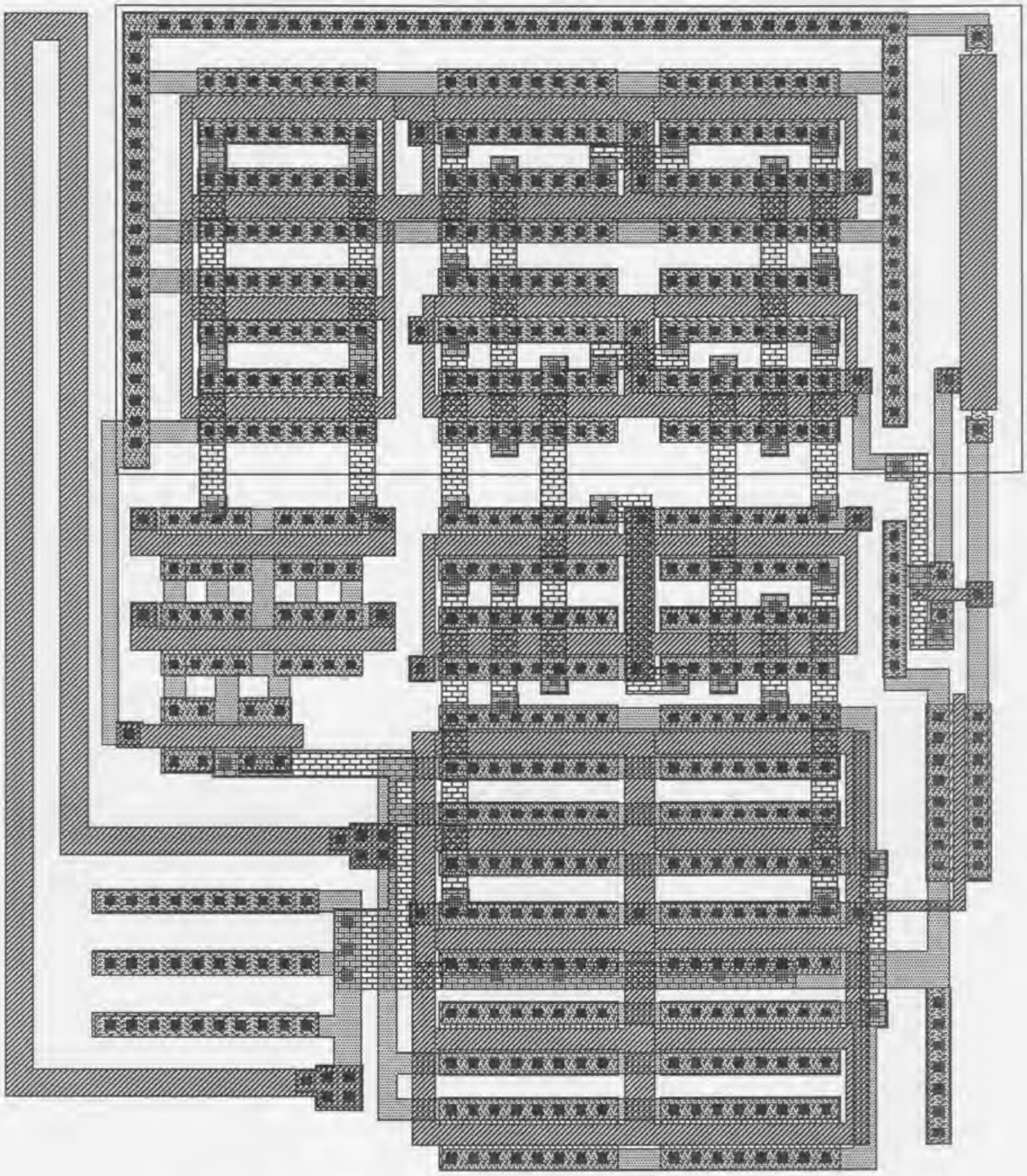


Figure 4.18 Layout of the bias network of the current conveyor.

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

4.6 CONCLUSION

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

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

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

5. SRAM SYSTEM BASED ON THE FOUR-TRANSISTOR CELL

5.1 INTRODUCTION

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

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

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

5.2 SRAM INTERFACE AND CYCLES

5.2.1 SRAM Control Signals [32]

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

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

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

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

5.2.2 SRAM Read Operation [32]

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

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

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

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

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

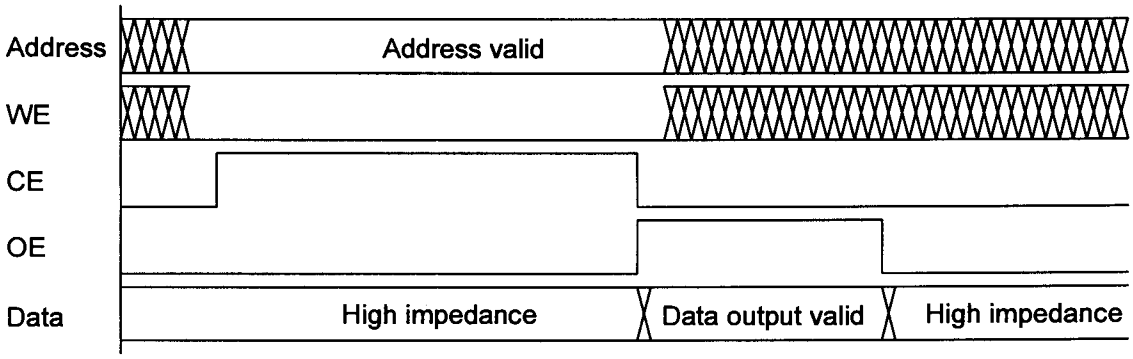


Figure 5.1 SRAM read cycle.

5.2.3 SRAM Write Operation [32]

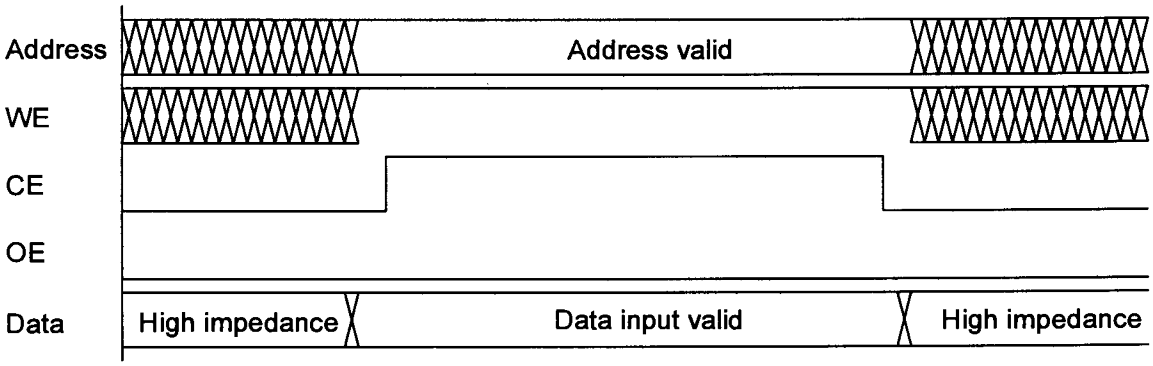


Figure 5.2 SRAM write cycle.

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

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

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

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

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

5.2.4 SRAM Timing [33]

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

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

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

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

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

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

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

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

5.3 THE SRAM SYSTEM

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

5.3.1 Global System

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

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

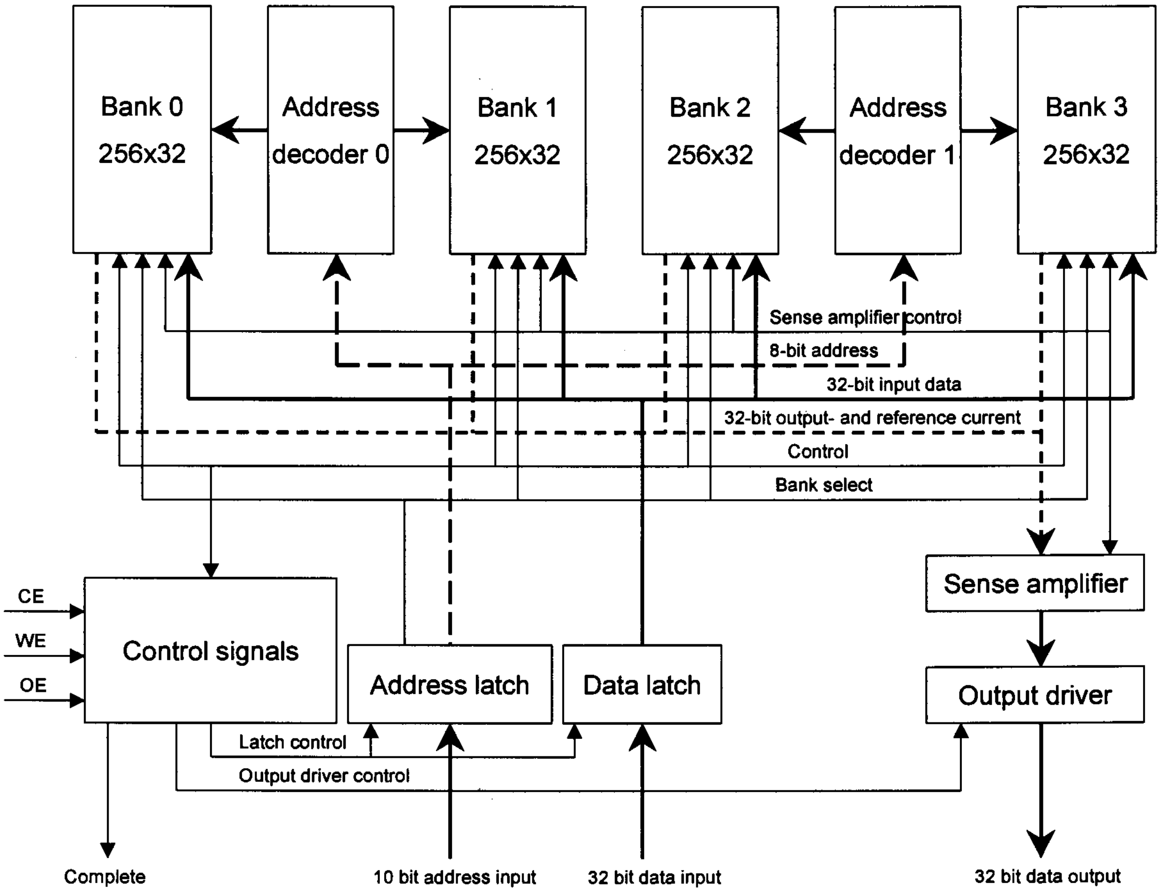


Figure 5.3 Top level block diagram of the SRAM system.

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

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

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

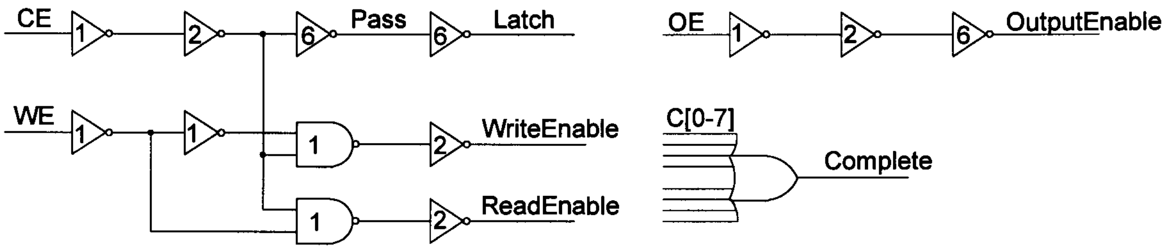


Figure 5.4 Circuit diagram of the control signals block in Figure 5.3.

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

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

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

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

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

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

5.3.2 Sense Amplifier

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

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

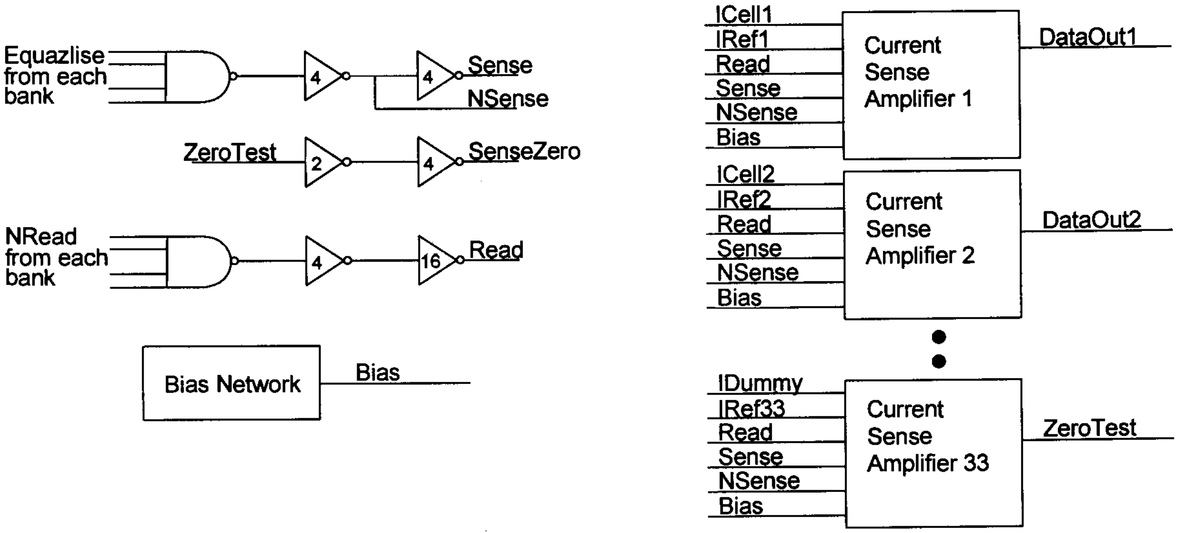


Figure 5.5 Complete current sense amplifier system.

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

5.3.3 Memory Bank

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

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

- the array of 256 words with 32 bits each,

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

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

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

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

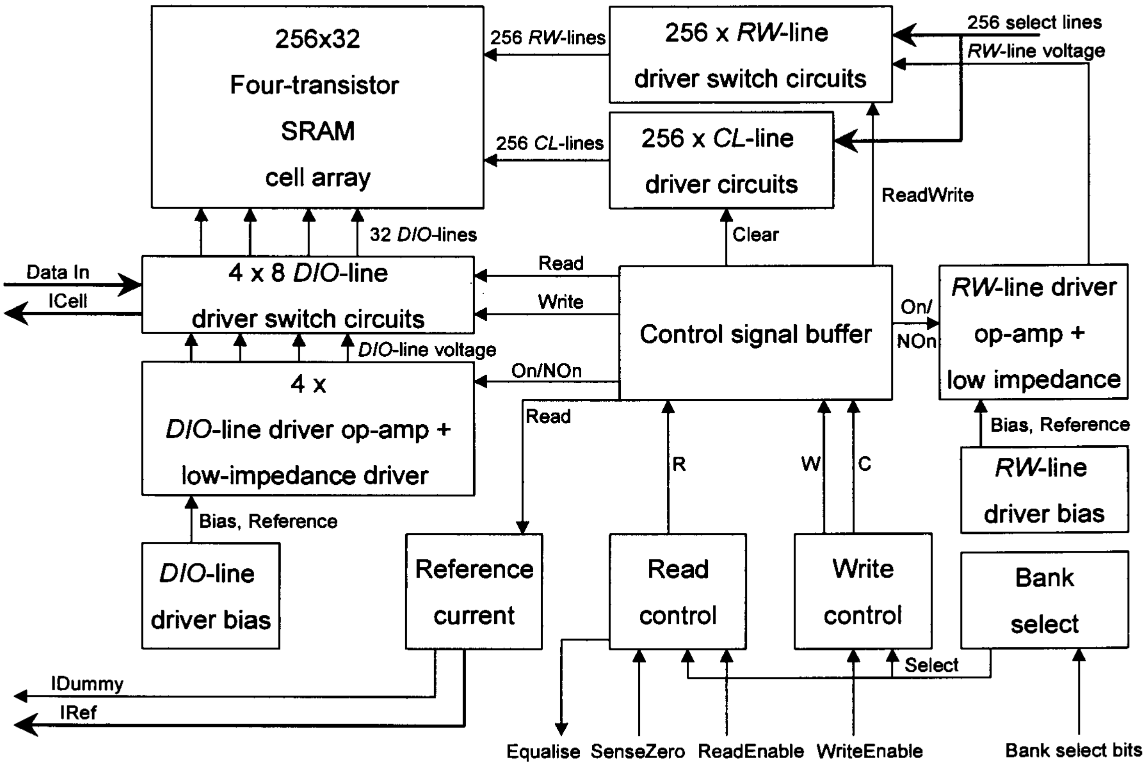


Figure 5.6 Block diagram of a single memory bank.

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

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

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

5.3.4 Read Control

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

Equalisation Cycle Timing

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

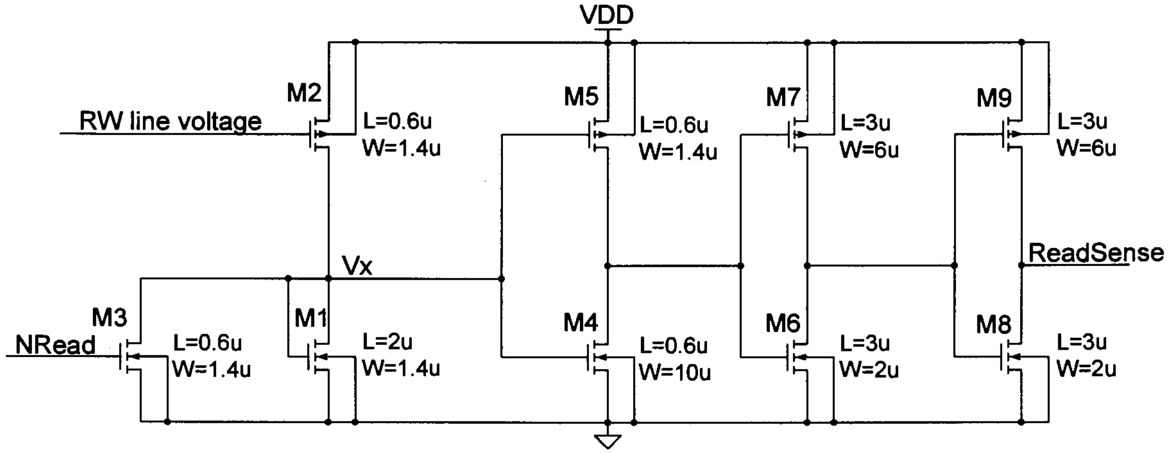


Figure 5.7 Circuit to sense completion of the equalisation cycle

Timing Circuit Structure

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

Edge to Pulse Converter

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

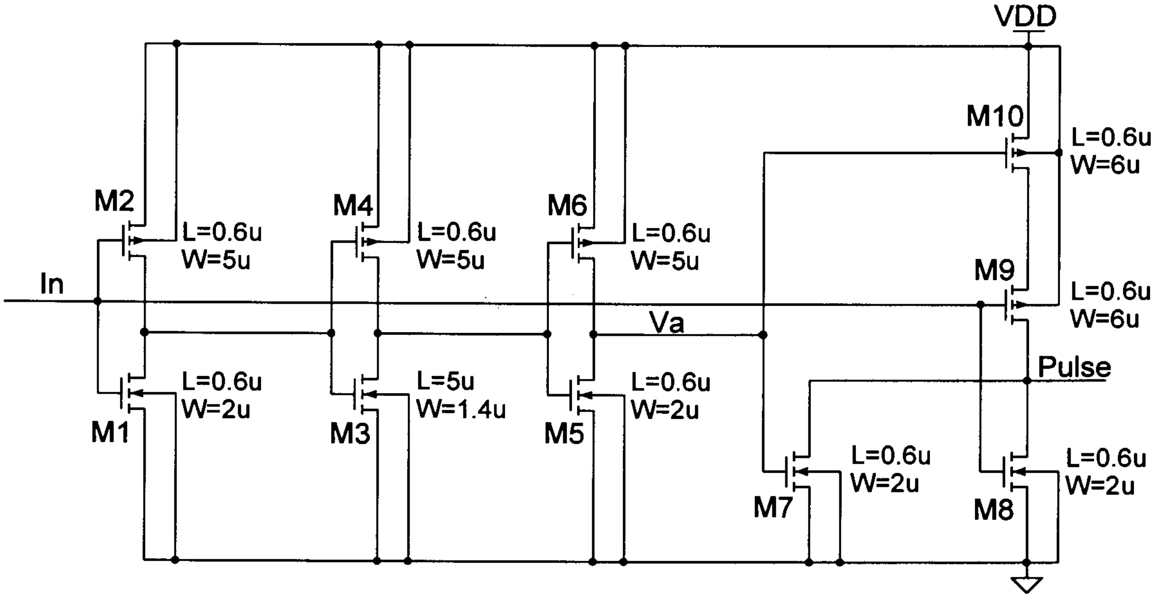


Figure 5.8 Falling edge to positive pulse converter.

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

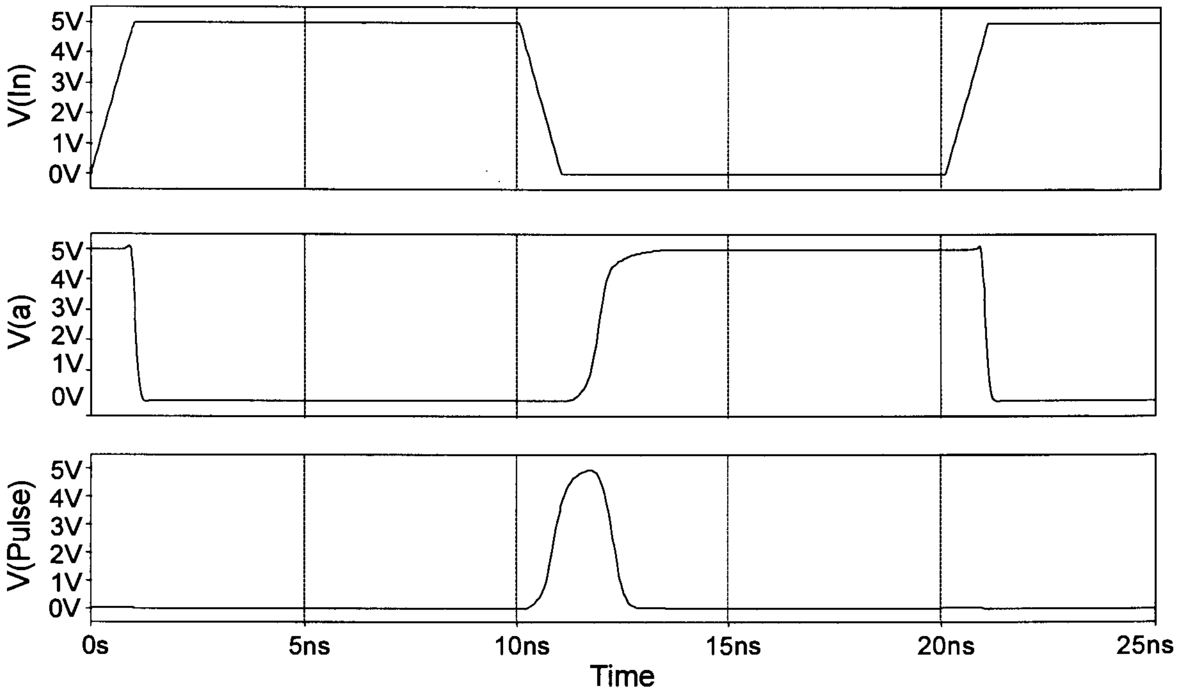


Figure 5.9 Simulation of the falling edge to pulse converter.

Read Control Circuit

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

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

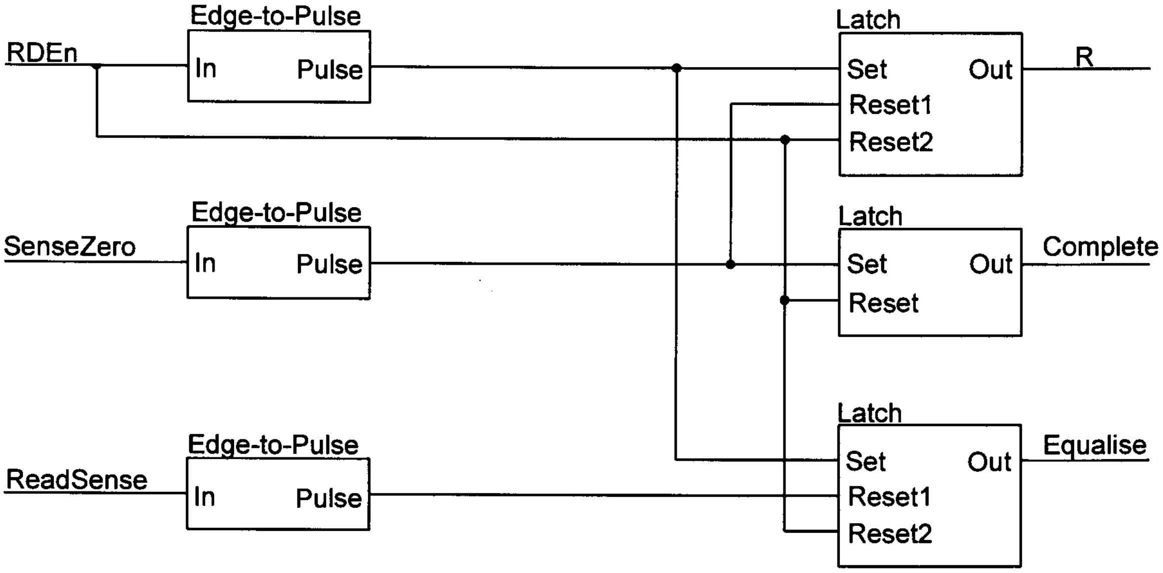


Figure 5.10 Read control and timing generator circuit.

5.3.5 Write Control

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

Timing Cells

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

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

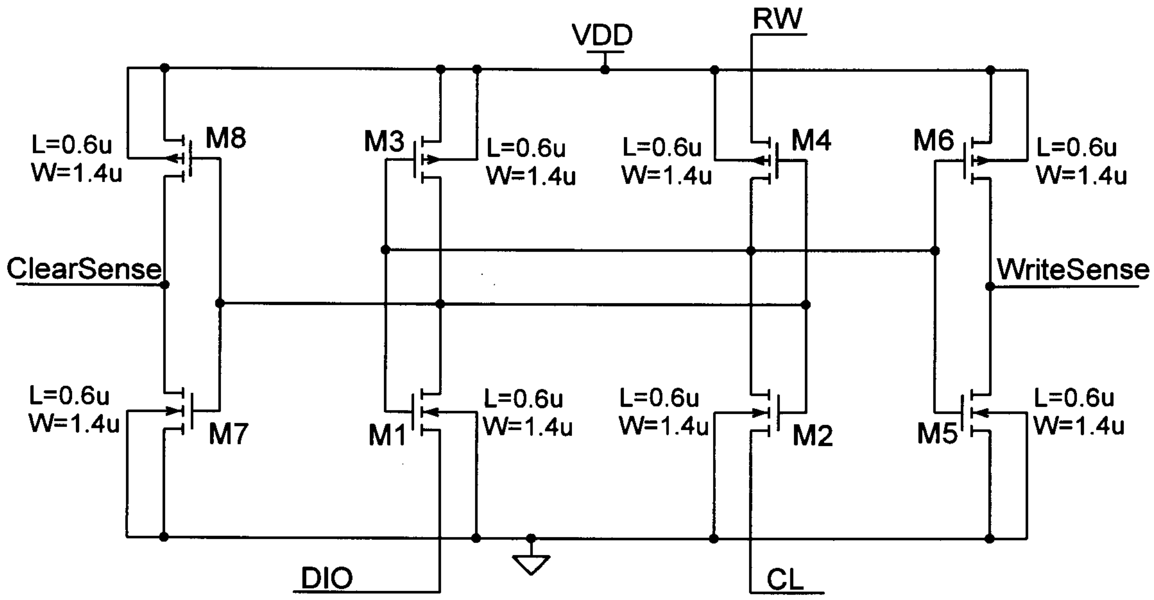


Figure 5.11 Dummy four-transistor SRAM cell with internal nodes sensed to time the write cycle.

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

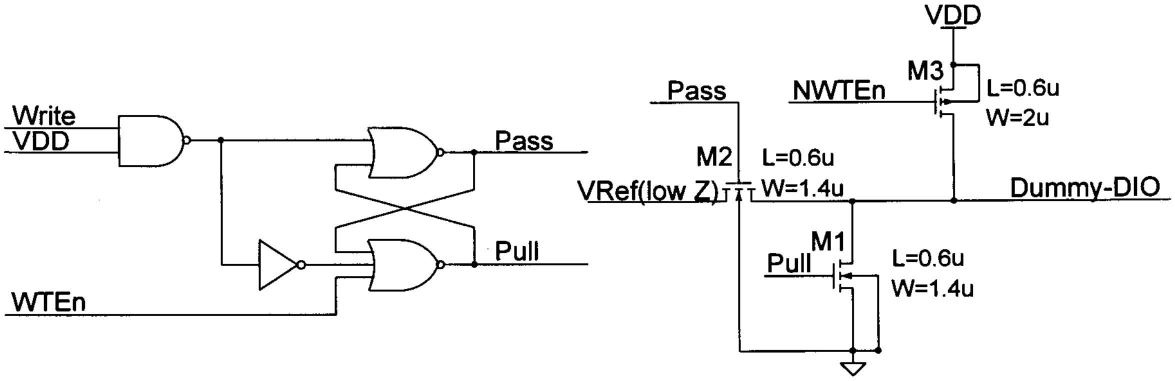


Figure 5.12 Modified *DIO*-line driver to ensure the dummy cells are in the correct initial state.

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

Write Control Circuit

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

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

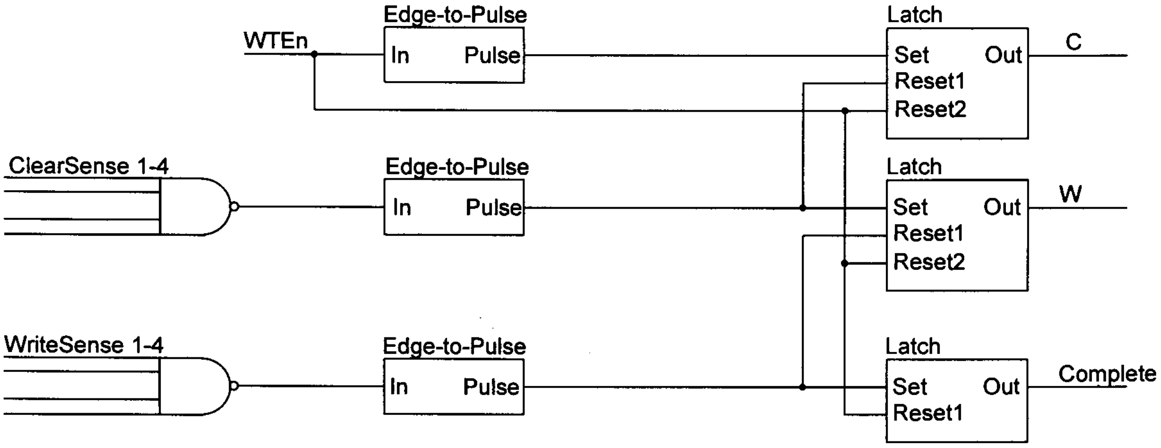


Figure 5.13 Write control signal generator circuit.

5.4 SIMULATION

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

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

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

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

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

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

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

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

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

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

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

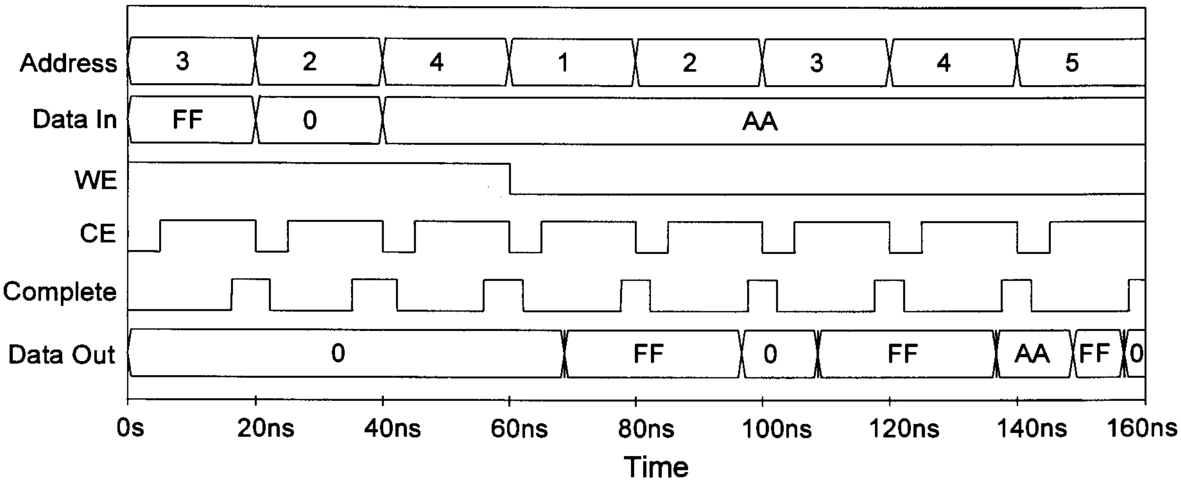


Figure 5.14 Simulation results of the 8x8 system showing correct functional operation.

5.4.1 Timing Specifications

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

- read access time: the time difference between the *CE* activation and the correct data value appearing on the output nodes,
- write access time: the time difference between the activation of the *CE* signal and the rising edge on the *Complete* signal,
- read cycle time: the read access time is added to the time difference between the valid data and the *Complete* signal activation and the address decoder system delay,
- write cycle time: the write access time added to the address decoder delay.

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

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

Table 5.1 Simulated timing specifications for all process conditions.

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

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

5.4.2 Power Dissipation

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

Static Power Dissipation

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

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

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

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

Write Cycle Power Dissipation

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

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

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

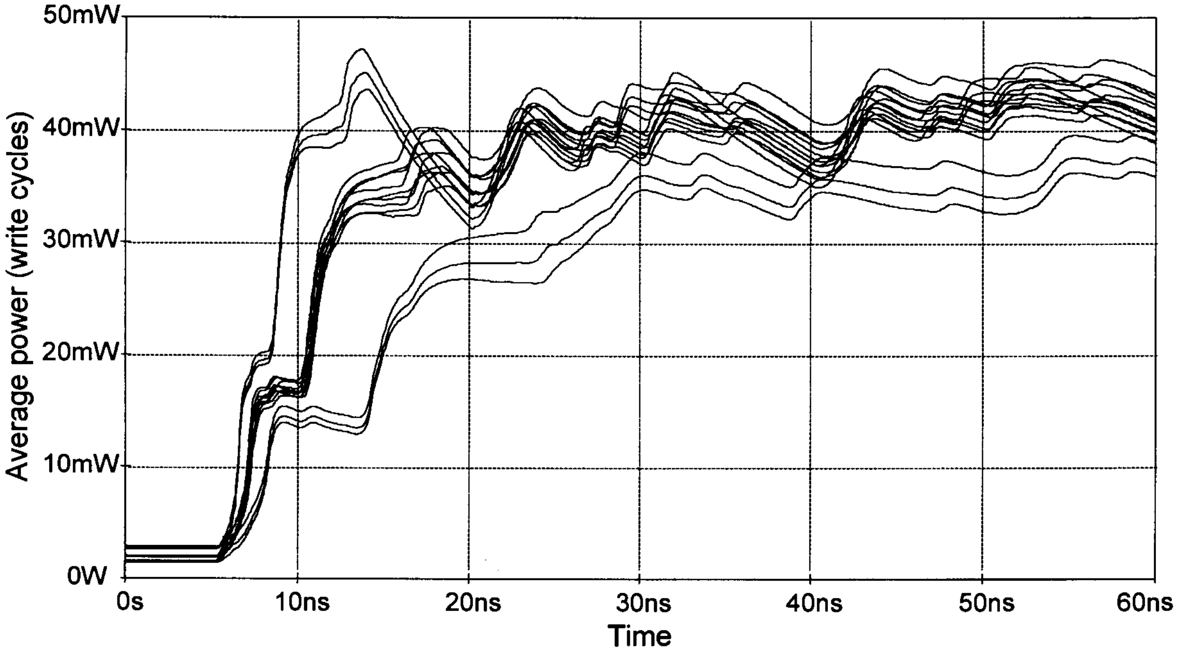


Figure 5.15 Simulated average power for all process corners for the 8x8 SRAM system during the write cycle at a cycle frequency of 50MHz.

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

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

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

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

Read Cycle Power Dissipation

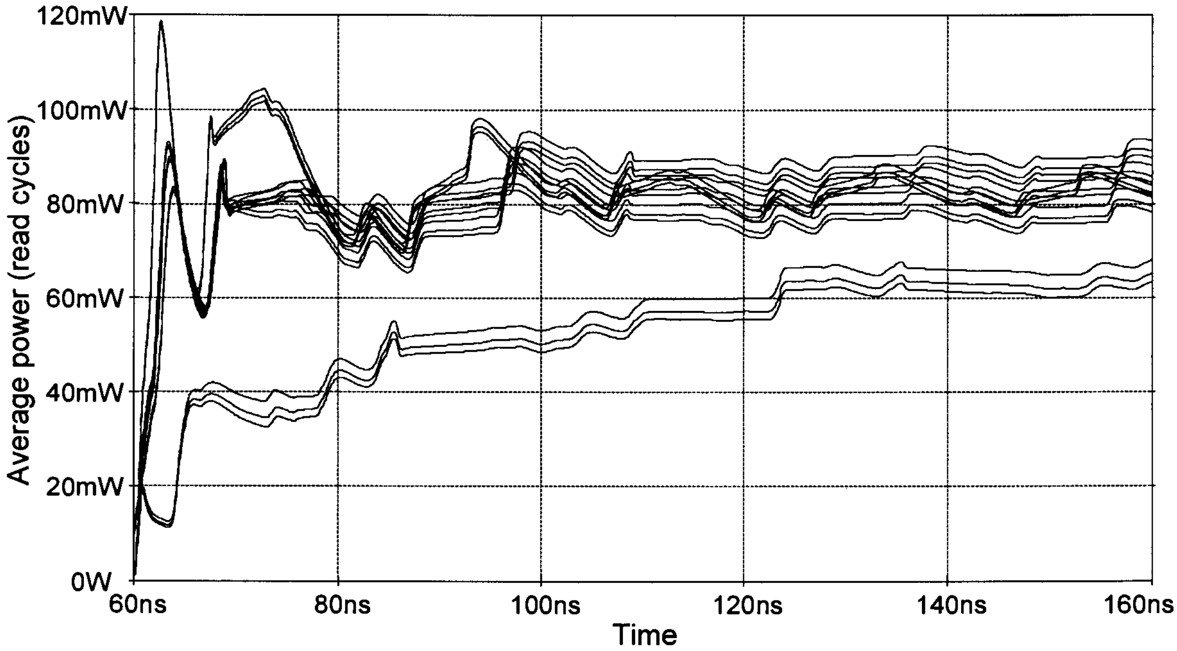


Figure 5.16 Simulated average power for all process corners for the 8x8 SRAM system during the read cycle at a cycle frequency of 50MHz.

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

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

5.5 LAYOUT

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

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

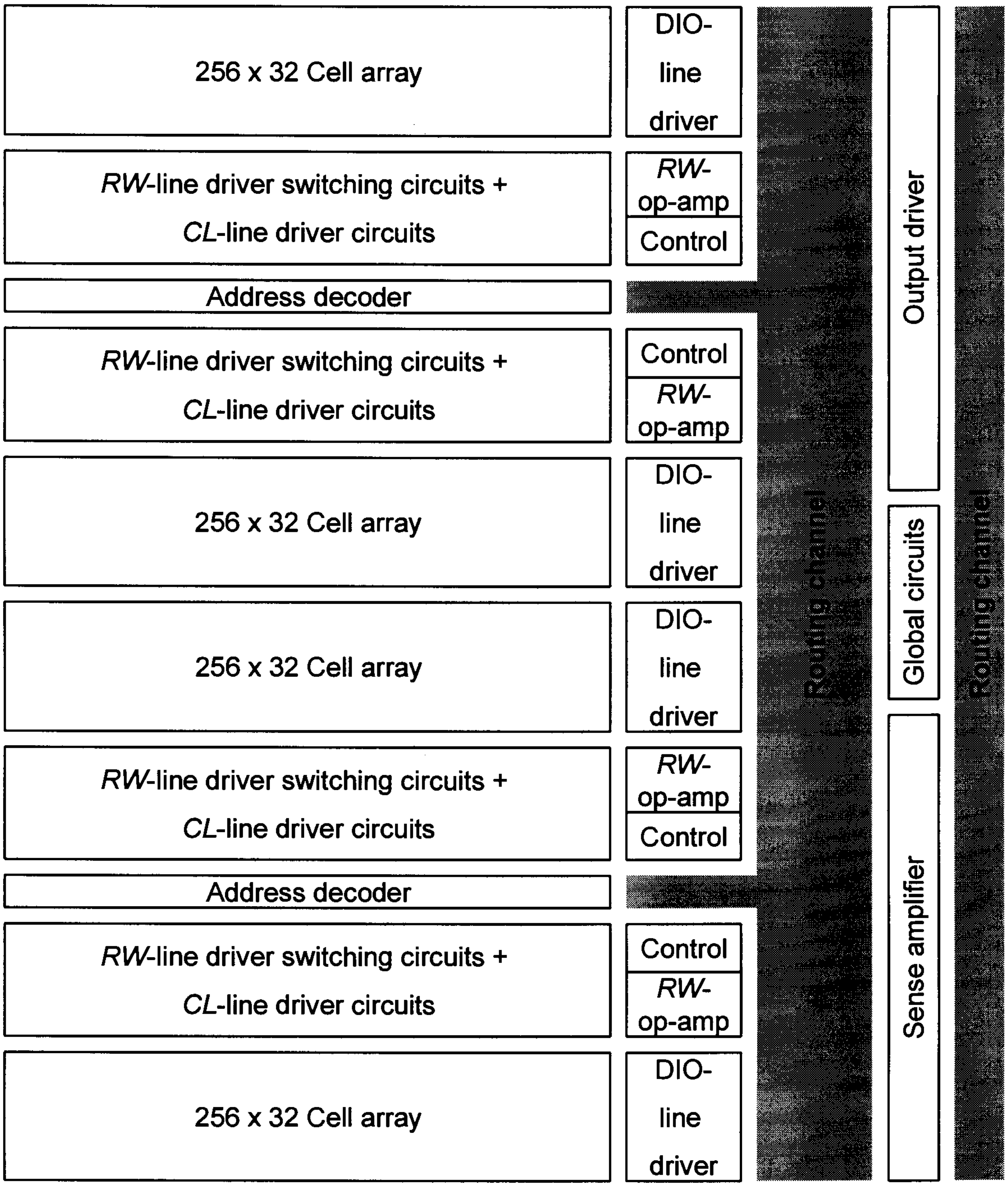


Figure 5.17 Floor-plan of the SRAM system.

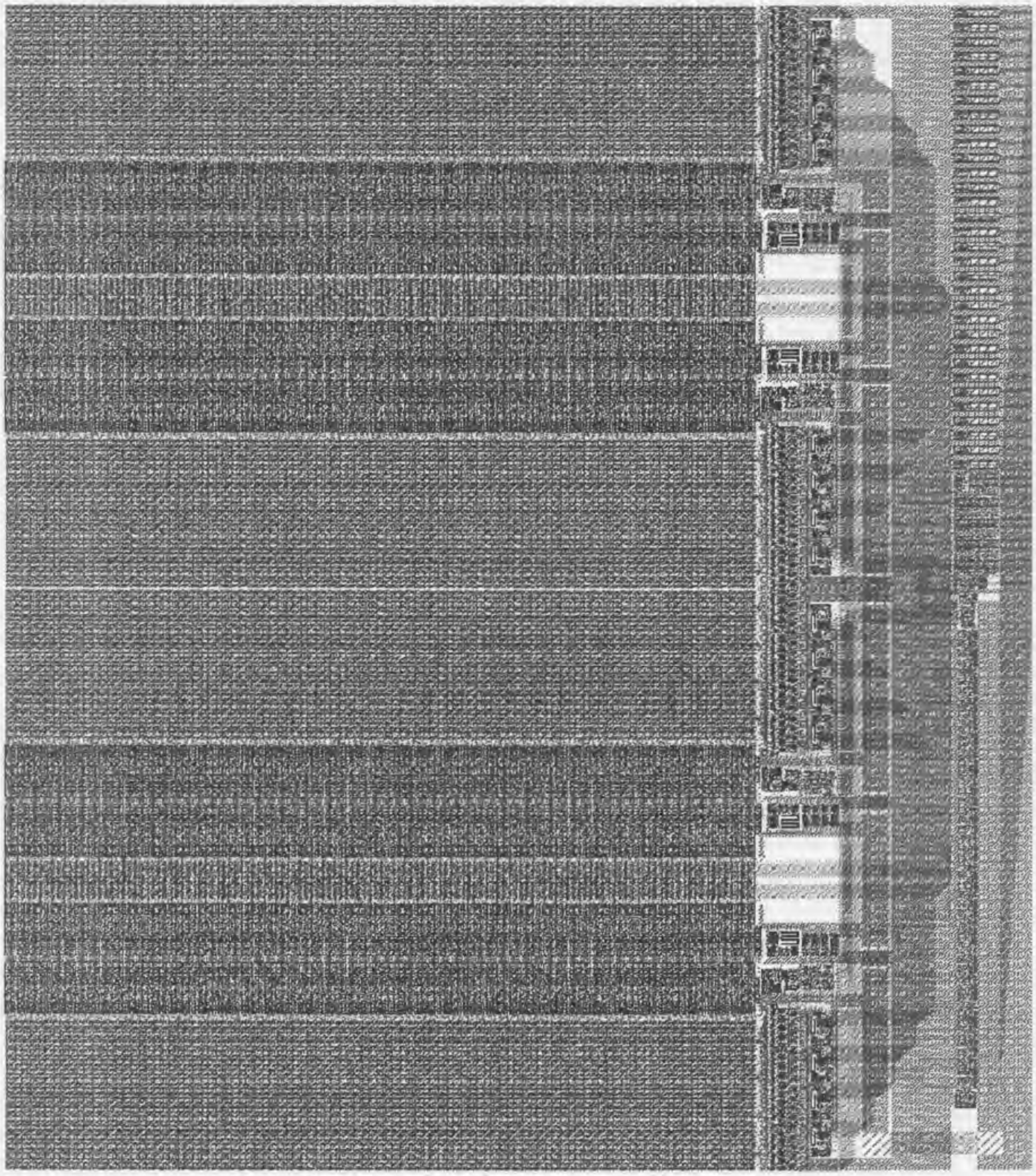


Figure 5.18 Layout of the complete four-transistor cell SRAM system in the 0.6μm CMOS process.

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

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

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

5.6 COMPARISON TO A SIX-TRANSISTOR SYSTEM

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

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

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

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

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

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

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

5.7 CONCLUSION

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

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

6. CONCLUSION

6.1 WHAT WAS GIVEN?

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

6.2 WHAT WAS THE AIM?

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

6.3 WHAT HAS BEEN ACCOMPLISHED?

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

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

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

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

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

6.4 WHAT CAN BE LEARNED FROM THIS?

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

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

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

6.5 WHAT IS THE NEXT STEP?

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

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

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

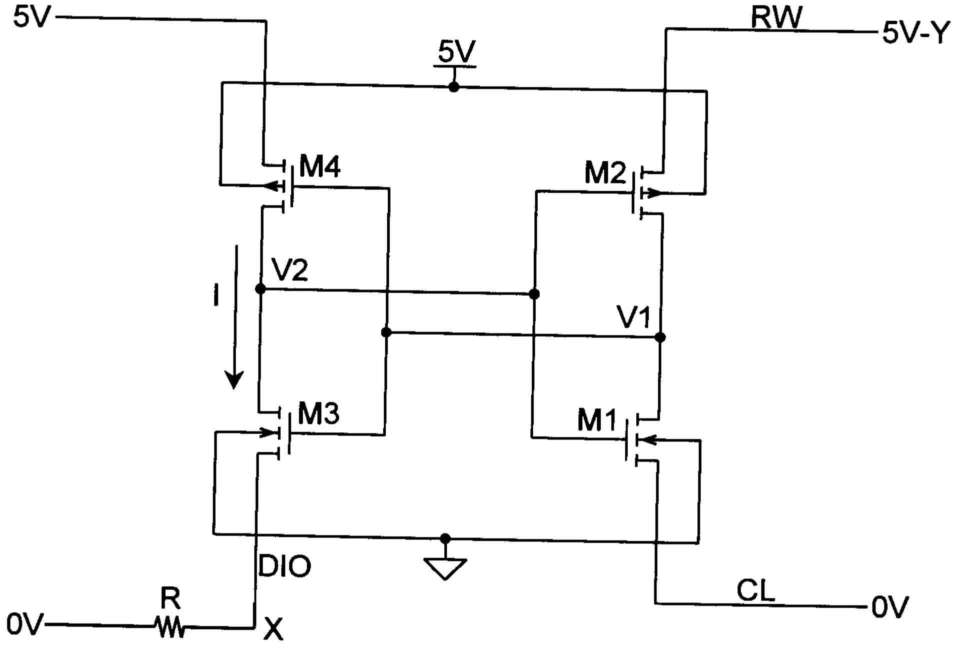


Figure 6.1 Alternative cell write method.