A Fast SRAM for Cache Applications Implemented Using SiGe HBT BiCMOS Technology

by

Hadrian Olayvar Aquino

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE

Major Subject: Computer and Systems Engineering

Approved:

____

John F. McDonald, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

May 2009

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENT

The work in this thesis was performed using the resources of the FRISC Research laboratory in Rensselaer Polytechnic Institute. I would like to thank Professor John F. McDonald, my adviser, for giving me the opportunity to work on this project. I would also like to express my gratitude to Philip Jacob and Aamir Zia who were always there to answer my questions. Thank you to Sherry Liu for the CACTI simulation results, summarizing the results in a pie chart and for continuing the work after this thesis.

I would also like to thank my family and friends for their love and support.

ABSTRACT

A fast SRAM was designed using Current Mode Logic to implement the peripheral circuits around a CMOS SRAM core using Silicon Germanium Heterojunction Bipolar Transistors. BiCMOS technology was used from IBM 8HP. The process features 0.12 μm as the minimum channel length for MOSFETs and 0.12μm emitter width HBTs.

The address decoder, output multiplexer and sense amplifier were implemented using CML. The write circuit and word line driver was implemented in BiCMOS. The memory array was done in CMOS. A layout of a 4KB memory was completed and tested for functionality.

1. Introduction

1.1 Objective

It is the objective of this thesis to design and to analyze a BiCMOS SRAM, the result of which can be used in the research for SRAM application as cache memory in a high performance computer system.

1.2 Research History

This thesis continues the efforts done by Sumit Suhag in [10]. This previous work provided an initial design for a BiCMOS SRAM. The memory bank was made up of 128 x 256 6T SRAM cells (4KB). The CMOS decoder in the design was a cascade of two input AND gates. The sense amplifier was an ECL buffer acting as a differential amplifier. The write circuit is made up of tri-state inverters on the bit lines. Figure 1 shows the word line driver used in [10].

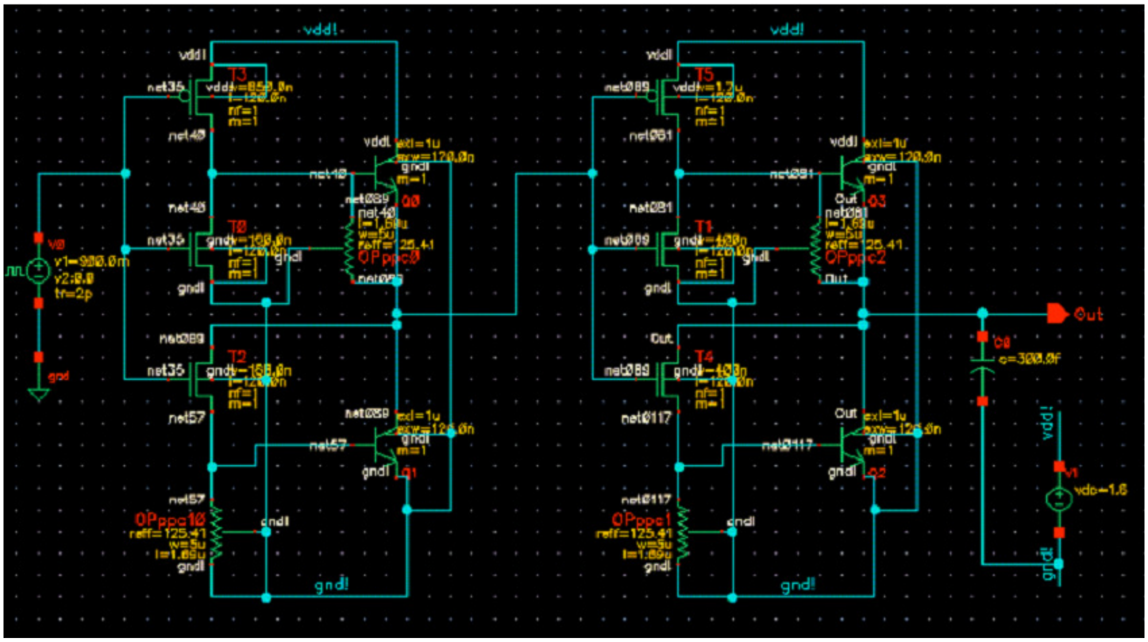

Figure 1: BiCMOS Inverter Word Line Driver used in previous work [10].

Figure 2 shows the circuit used for simulations in [10]. It included a memory cell in the middle, an AND gate to act as the decoding circuit and the sense amplifier. The 4KB memory array's word line and bit line capacitances were estimated at 278.016 fF and

105.15456 fF respectively. These capacitances were added to the word and bit line. The methodology used for calculating these capacitances are described in more detail at the conclusion of this thesis.

The transistors at the top act as the precharge circuit. The logic gates at the bottom act as the write circuit. It should be noted that the previous design assumed CMOS level inputs.

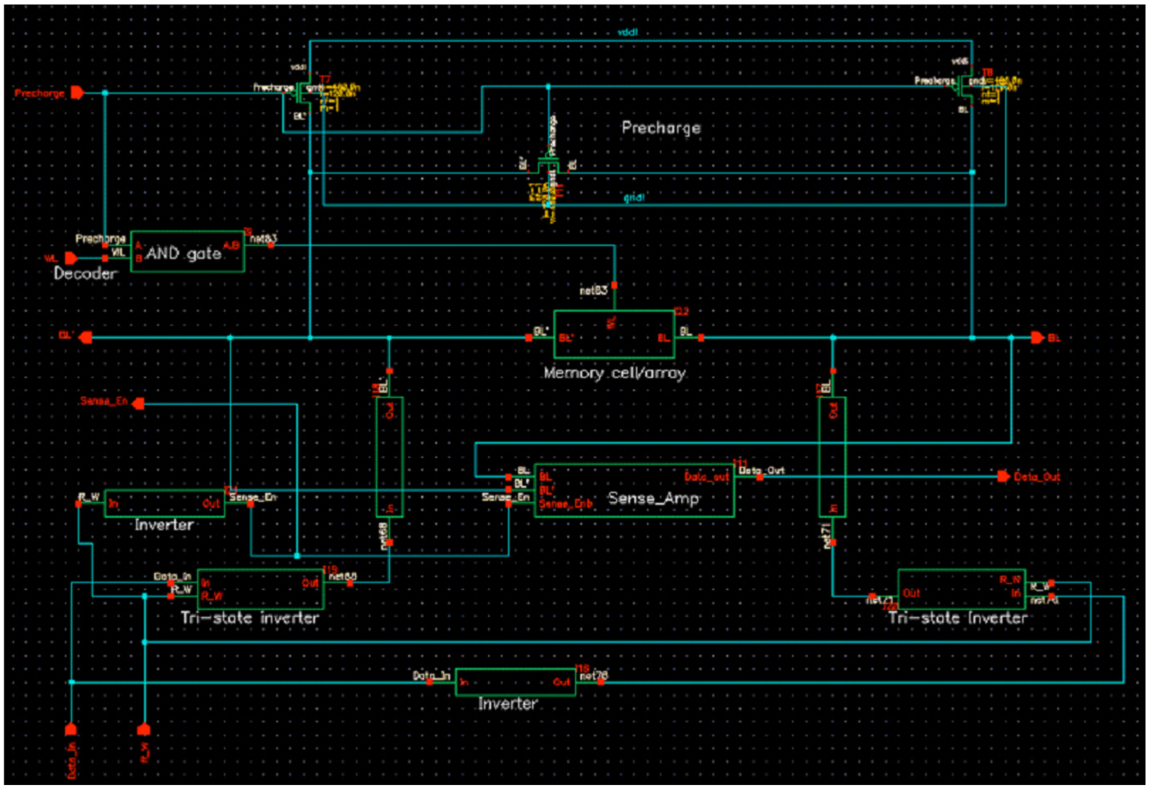

Figure 2: Simulation circuit in previous work [10].

The current thesis presents a BiCMOS SRAM for use as an L1 cache for a bipolar microprocessor. A layout of the design was assembled and a simplified but accurate layout was simulated.

The design makes use of a CML decoder and a simplified write circuit. The CML decoder necessitated the design of a CML/ECL to CMOS translator as well as modifications to the word line driver. The sense amplifier was modified to maintain its previous output during precharge. CML/ECL level inputs are assumed and CML/ECL level outputs are produced.

1.3 A Brief History of SiGe HBTs and SiGe HBT BiCMOS [1][9]

The idea of the Heterojunction Bipolar Transistor dates all the way back to the BJT patents filed by William Shockley in 1948. The operational theory of HBTs was pioneered by Herbert Kroemer and was largely in place in 1957. However, because of the difficulties in growing device quality SiGe films, the first SiGe HBT was not demonstrated until 1987 by an IBM team at the IEDM.

This first SiGe base mesa transistor was fabricated using MBE to grow the collector, base and emitter layers without breaking vacuum and then defined using dry-etching techniques. The transistor demonstrated an increase in the collector current as predicted, however it was far from an ideal device and was unusable.

In the early 1980s, IBM was using ion implanted base bipolar technology. During that time the ability to make the bipolar transistor's base region narrower was the means to improve device performance. However, scaling limitations in the technology became evident and became a motivation to look into silicon epitaxy.

At the time, silicon epitaxy was a high temperature process with temperatures well in excess of 1000°C. The high temperature made it unsuitable for the precision required because of the increased relaxation rates as well as dopant redistribution. Therefore to achieve a film of arbitrary dopant and chemical content, a low temperature epitaxy was needed.

Sometime in the 1980s Bernard Meyerson, who was working for IBM, dropped a piece of silicon. This seemingly innocuous event led to the invention of Ultra High Vacuum/ Chemical Vapor Deposition. While cleaning this wafer he observed that it reacted much differently from how conventional wisdom at the time predicted it should. That is, it was observed that the wafer was hydrophobic. It was soon observed that a passivation layer of hydrogen terminated silicon bonds persisted after an HF etch. This passivation layer allowed high temperature steps to be removed and still have high quality epitaxy.

Soon after fabricating a silicon epitaxial base bipolar transistor using UHV/CVD, attention went to the developing a SiGe epitaxial base bipolar transistor.

The first CVD grown SiGe HBT was demonstrated in 1989 by a team composed of members from Stanford and Hewlett Packard, which was followed shortly thereafter by the first SiGe HBT grown by UHV/CVD by a group from IBM.

The first SiGe BiCMOS technology was reported by a group from IBM on December 1992 at the IEDM. The technology was dubbed the High Performance Transistor Generation 6 (HPT6), and was the result of a program to build the mainframe dubbed H10C. However, that same year IBM suspended all bipolar and SiGe activity in favor of CMOS and parallel architecture and the H10C was never built.

1.4 Process Overview

1.4.1 UHV/CVD Epitaxy

Silicon and Germanium have a 4.17% lattice mismatch at 300K which increases slightly with temperature. This lattice mismatch between the SiGe film and the Si substrate causes one of two things to happen when a SiGe film is deposited on a Si substrate. Either the SiGe film is forced to adopt the lattice constant of the underlying Si or the SiGe film can relax during growth to its natural lattice constant.

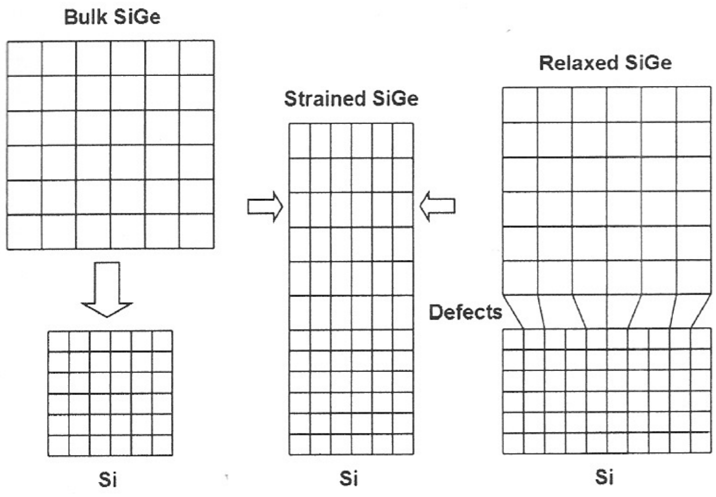

Figure 3: 2D representation of both strained and relaxed SiGe on a Si substrate. [2]

In the case that SiGe adopts the lattice constant of the underlying Si; the SiGe film is forced into biaxial compression. Because of the additional strain energy, it embodies a higher energy state than for an unstrained film.

However, when the film thickness reaches a certain “critical thickness” the strain energy becomes too large. This causes the film to relax during growth, causing defects that affect device performance. Critical thickness is defined as the thickness where the force of the dislocation segment found at the Si-SiGe interface is equal to the component of force per unit length acting on the threading component of the dislocation in growth plane.

The critical thickness of the film is dependent on the Ge fraction of the film and the temperature at which the film is grown. To make the critical thickness as thick as possible, the temperature the film is grown in must be as low as possible. This also means that all processes after the growth cannot exceed that temperature. Should the film be exposed to high temperatures, it may relax.

Using UHV/CVD allows us to perform the growth in relatively low temperatures. The wafer is first prepared by dipping it in a dilute 10:1 H_20/HF solution for 10-30 seconds. This creates a hydrogen ad layer that reduces the reactivity of the interface. This de-wets completely when extracted from the bath. The wafer is then placed in a load lock. When the pump down reaches 10^{-6} torr, the wafers are transferred under flowing hydrogen to the process chamber where the growths are started immediately. The gas sources are SiH_4, GeH_4, B_2H_6 and PH_3. The growth is performed at a temperature range of 400-500C. The growth rate can vary between 0.1-100 Å/min and is a function of temperature and film Ge content. Typical rates range in 4-40 Å/min.

1.4.2 The IBM SiGe BiCMOS 8HP Process

The 8HP process is a base-after-gate (BAG) process, where the base of the HBT is built after the bulk of the CMOS process has been completed. This is done to minimize the thermal budget of the SiGe HBT process.

The process starts with the formation of the subcollector. This is followed by isolation and formation of the collector reach-through. After that the process proceeds with the standard CMOS processing steps. The bulk of the CMOS process is completed leaving only the source-drain implant and activation. The source-drain implants are left until after the bipolar process because the dopants used (phosphorus and boron) tend to diffuse through the gate during the low-temperature cycles in the bipolar process.

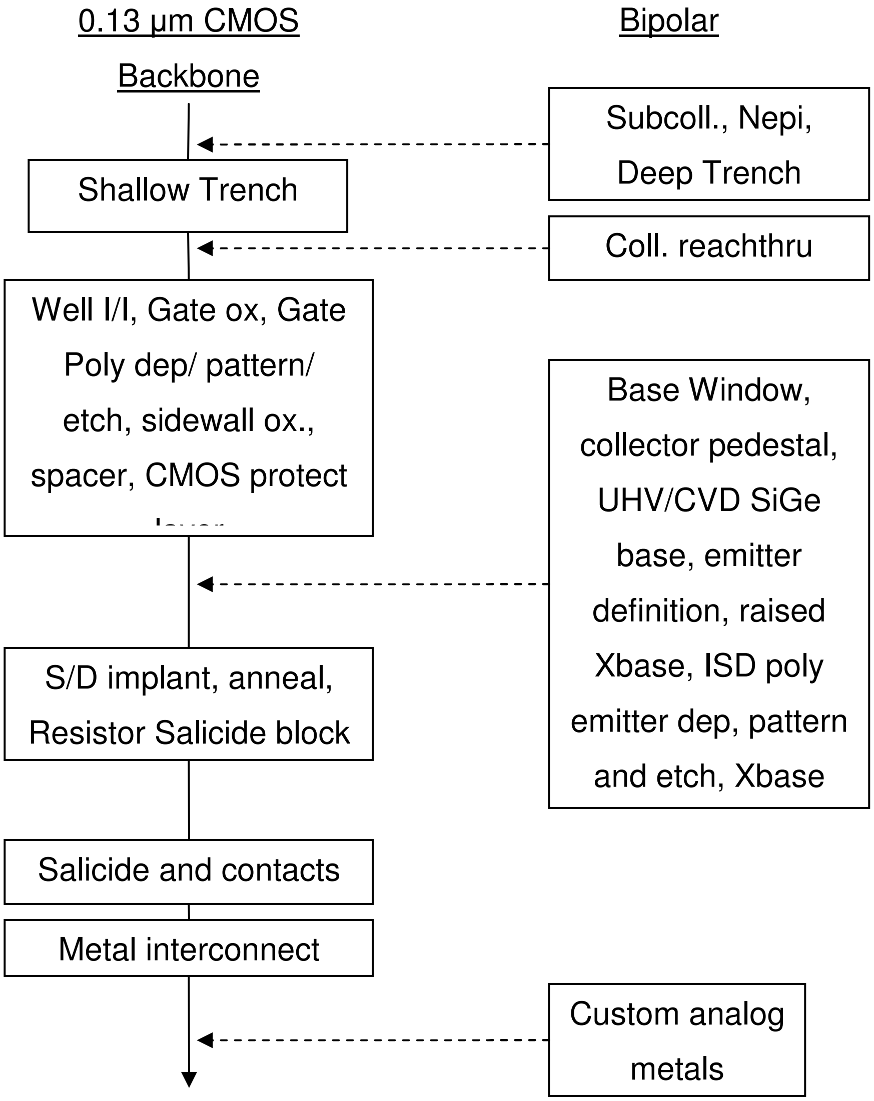
0.13 µm CMOS Backbone
Bipolar
Subcoll., Nepi, Deep Trench
Shallow Trench
Coll. reachthru
Well I/I, Gate ox, Gate Poly dep/ pattern/ etch, sidewall ox., spacer, CMOS protect
Base Window, collector pedestal, UHV/CVD SiGe base, emitter definition, raised Xbase, ISD poly emitter dep, pattern and etch, Xbase
S/D implant, anneal, Resistor Salicide block
Salicide and contacts
Metal interconnect
Custom analog metals

Figure 4: Process flow of BiCMOS 8HP in a 0.13µm CMOS backbone. [8]

The bipolar process begins with an etch to define the bipolar window opening. This is followed by the nonselective epitaxial growth of the UHV/CVD SiGe base. Then, the raised extrinsic base is introduced. This is to reduce the collector-base capacitance. Then an opening for the emitter is formed with inside spacers. In situ doped polysilicon is then deposited. The emitter polysilicon is then patterned and etched followed by the base polysilicon.

After the bipolar process, the CMOS protect layer is removed. The source-drain implant and the activation RTA are then done. After this salicidation, contact formation and metallization complete the process.

1.5 Performance Advantages

Silicon and Germanium have bandgaps of 1.12eV and 0.661eV, respectively. The band gap of SiGe lies somewhere between the bandgap of Si and Ge and depends on composition. The compressive force found in growing SiGe films produces an additional reduction in the bandgap.

In the operation of a BJT, when the emitter-base (EB) junction is forward biased, electrons are injected into the base from the emitter across the EB potential barrier. The electrons diffuse across the base until they are swept into the electric field in the collector-base (CB) junction.

The introduction of Ge into the base lowers the EB potential barrier increasing electron injection for the same bias. This leads to higher collector current, and therefore, higher current gain. In typical circuits that do not require high current gain, this higher current gain from the introduction of Ge can be offset by higher base doping which improves dynamic switching and noise characteristics.

In most of today’s SiGe HBTs, the composition is graded with the Ge content increasing from the EB junction to some maximum value near the CB junction where it is rapidly ramped back down. This creates an accelerating electric field that improves minority carrier transport reducing the base transit time and therefore improving the switching speed.

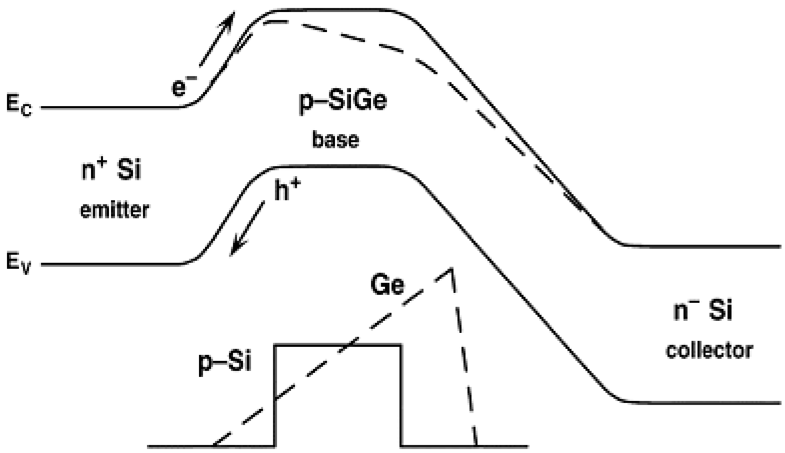

Figure 5: Energy band diagram for Si BJT and graded base SiGe [2]

2. BiCMOS Memory Design

2.1 CML/ECL Logic

One logic family that makes use of bipolar transistors is Current Mode Logic (CML). Current is steered through a tree, causing the desired voltage changes at the outputs. It is based on the differential circuit consisting of an emitter coupled pair of NPN transistors biased by a constant current source. This is known as a CML buffer or, if its outputs are reversed, the CML inverter and is shown in the Figure 6 below.

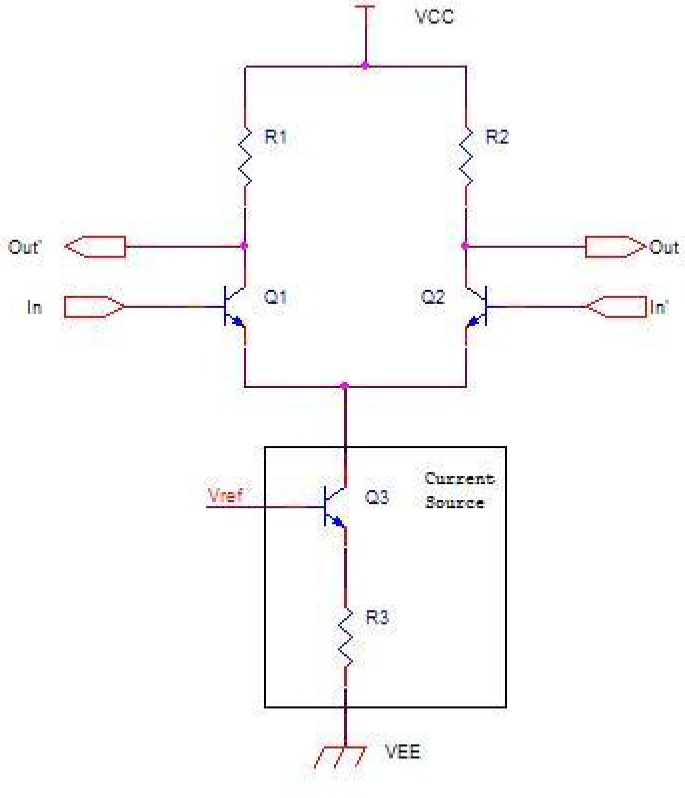

Figure 6: CML buffer

In order to achieve high switching speeds transistors Q1 and Q2 are kept within cut-off or active modes of operation. Saturation of the device must be avoided as their switching speed is greatly reduced in the saturation region. In order to saturate the device, a large amount of charge needs to accumulate in the base of the device. In order to turn off a device that is in saturation, a large amount of charge must be removed from the base. These are slow operations and severely degrade the switching speed of the device.

Given the differential input voltage v_d=*In-In'*, and that the transistors are matched, and assuming they operate exclusively within forward active and cut-off, the ratio of the currents through these transistors can be expressed as:

$$\frac{i_{c1}}{i_{c2}} \cong e^{\frac{v_d}{v_T}}$$

Neglecting the base currents it can be observed that:

$$i_{c1} + i_{c2} \cong I_s$$

It can be seen then that the current I_S is almost completely steered to one of the two transistors controlled by the differential voltage v_d with just a few orders of v_T . When v_d is 4.6 times v_T (at 300K that would be a v_d of about 119mV), 99% of the current is flowing through one transistor.

Currents i_{c1} and i_{c2} can be found from the observed relationships above to be:

$$i_{c1} = I_s \frac{e^{\frac{v_d}{v_t}}}{1 + e^{\frac{v_d}{v_t}}}$$

$$i_{c2} = I_s \frac{1}{1 + e^{\frac{v_d}{v_t}}}$$

The output voltages can than be calculated.

$$v_{o1} = -R_c\left(i_{c1}\right) = -R_c I_s \frac{e^{\frac{v_d}{v_t}}}{1 + e^{\frac{v_d}{v_t}}}$$

$$v_{o2} = -R_c\left(i_{c2}\right) = -R_c I_s \frac{1}{1 + e^{\frac{v_d}{v_t}}}$$

$$v_o = -R_c I_s \frac{e^{\frac{v_d}{v_t}} - 1}{1 + e^{\frac{v_d}{v_t}}}$$

CML gates can have single ended inputs. In this case the remaining input is pegged at a constant voltage. Here, $\underline{v_d}$ becomes a difference between the input voltage and the reference voltage. However, differential gates are preferred because they have better noise immunity and speed. They do, however, require that the input signals are skew free and may thus present a challenge in routing the wires.

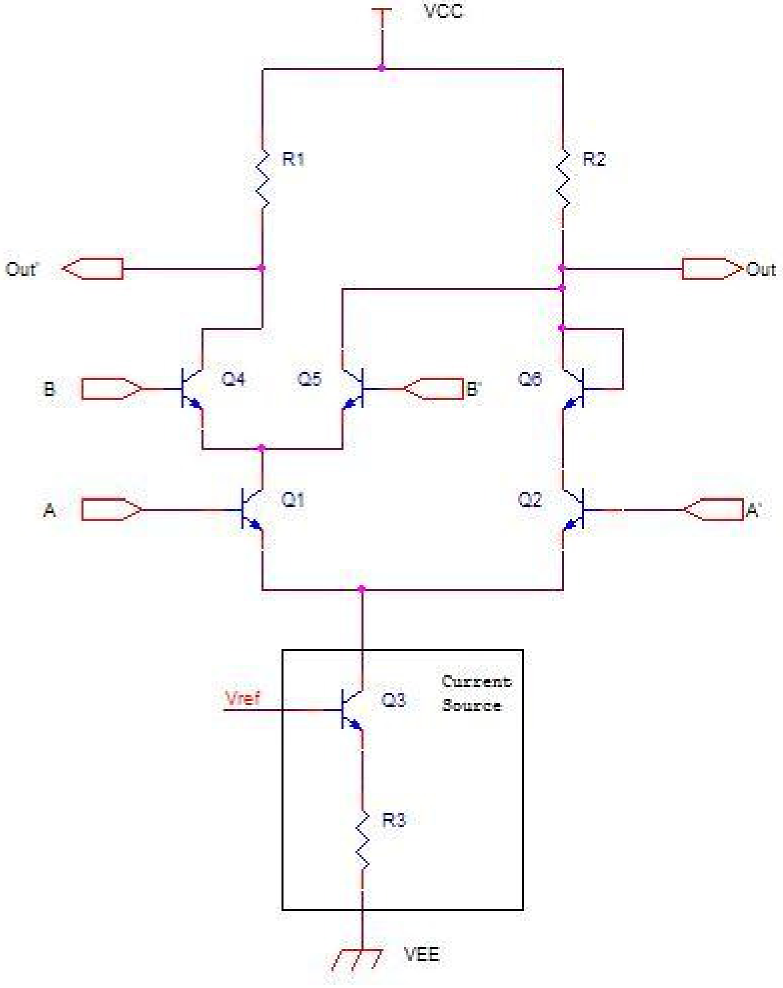

Figure 7: CML 2 input AND gate

As previously described, CML gates operate by steering current through one of two output resistors to create a desired voltage drop. To create a properly operating CML gate, there must be a unique path from the output to the current source for each of the possible input combinations.

Figure 7 shows a stacked topology implementation of a 2 input AND gate in CML. Current is passed through the path through transistors Q4 and Q1 only when both inputs are logic high. With current passing through that path, almost all the current passes through R1 causing a voltage drop, and almost no current is passing through R2 causing the voltage below R2 to be pulled up to VCC. Transistors Q2, Q5 and Q6 cover all the other input combinations. Using DeMorgan's rules an OR gate can also be built using the same circuit by switching both the complimentary inputs and outputs.

In order for proper operation of a stacked topology, different voltage levels must be used for the inputs of the different stacks. If the same input levels are used, then since the collector voltage of the lower stack is equal to the emitter voltage of the upper stack, the lower device then saturates. It is therefore necessary to create multiple logic levels to have more complex gates. Level 1 signals generally have a swing of 0V to -350mV. Level 2 signals have a swing of -850mV to -1.2V.

To shift the voltage levels, an output buffer is added to the output of the CML gate. Adding this output buffer to a CML gate creates an ECL (Emitter Coupled Logic) gate. ECL gates can also be used to improve the driving capability of the gate. Figure 8 shows an ECL buffer.

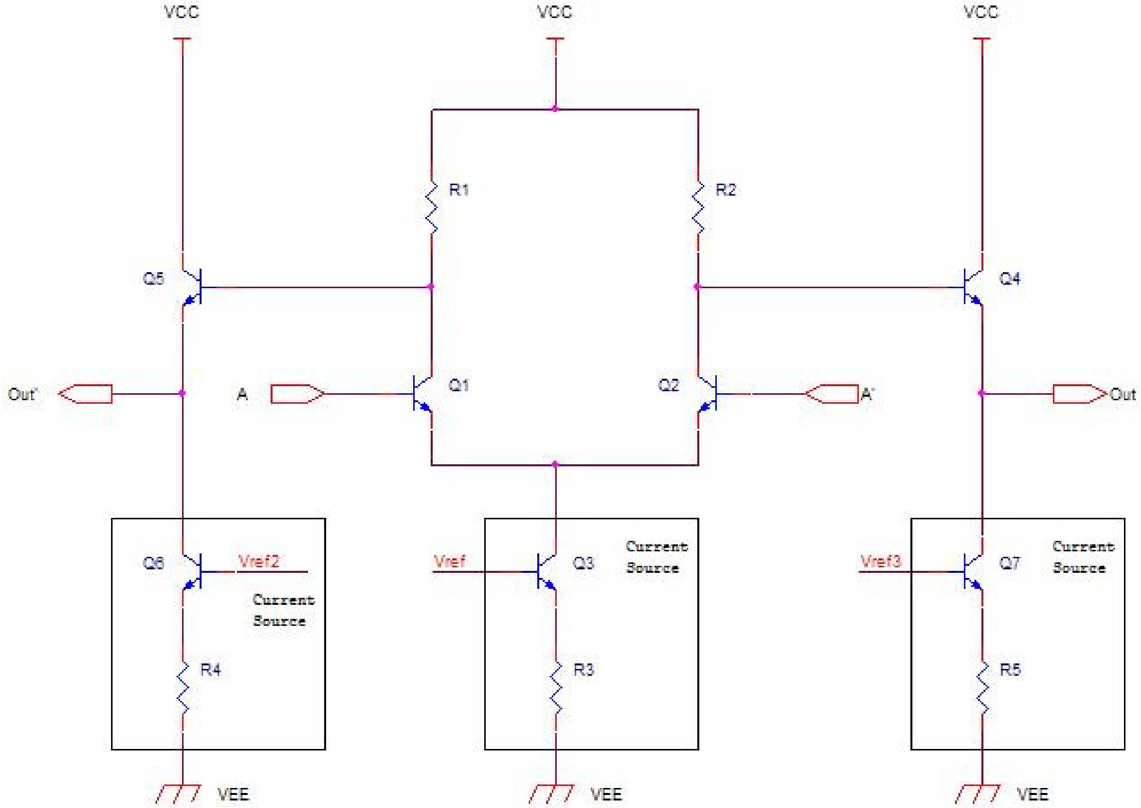

Figure 8: ECL buffer

The output buffer offsets the CML output by v_{BE}. These output buffers can be cascaded to have more logic levels. The number of logic levels, however, is limited by the supply voltage because the number of v_{BE} drops cannot exceed the supply voltage.

2.2 SRAM Overview

The 6-transistor (6T) SRAM cell is used. It has a single word line and complementary bit lines. It consists of a pair of cross-coupled inverters and an access transistor for each bit line. The cell is activated by raising the word line and accessed through the bit lines. The word lines connect all cells in a row and bit lines connect all cells in a column. Data is stored in the cross-coupled inverters at nodes A and A'.

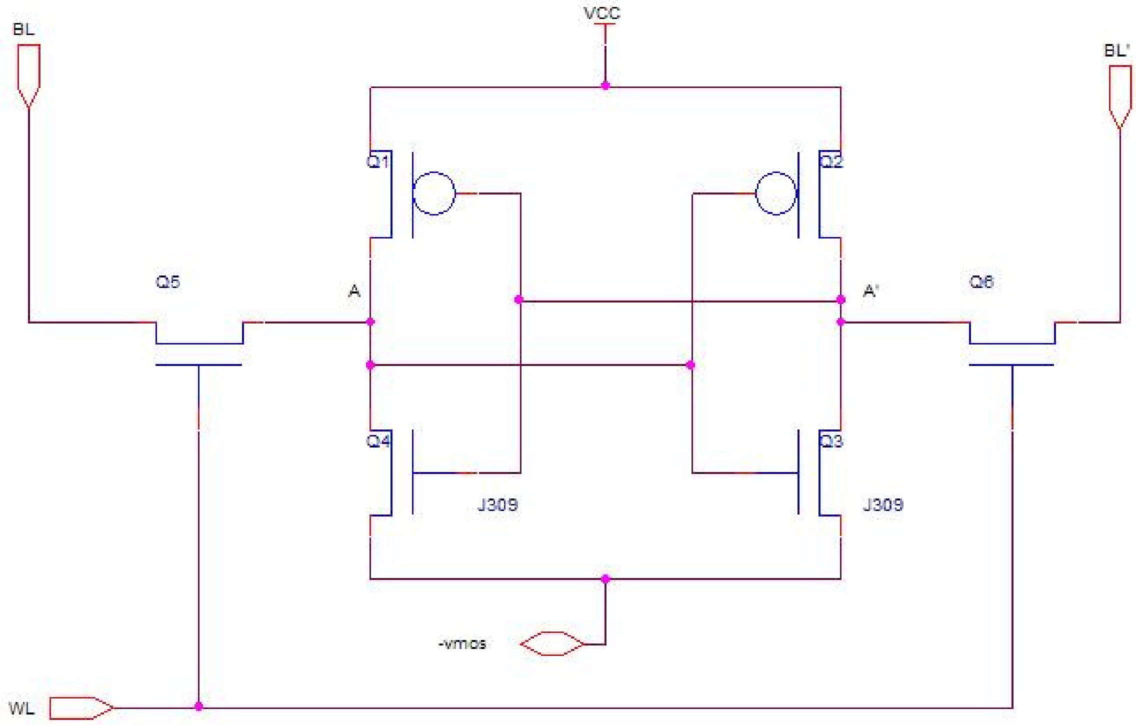

Figure 9: 6T SRAM cell.

If the word line is not asserted, the access transistors isolate the cell from the bit lines and the cross-coupled inverters continue to reinforce each other. During a read operation, the bit lines are pre-charged to ‘1’. Assume that node *A* is ‘0’ and thus *A’* is ‘1’. When the word line is raised BL, will be pulled down through transistors Q5 and Q4. BL’ is kept at ‘1’ by transistors Q2 and Q6. However, *A* is also being pulled up by BL, thus proper sizing of the transistors must be done to prevent a ‘1’ from being written into *A*. Also, NFETs are used for the access transistors as they are better at passing ‘0’s than ‘1’s.

During a write operation, the bit lines are pre-charged to high. Assume that node *A* is ‘0’ and a ‘1’ is to be written into it. BL’ is then pulled down by the write circuit. When the word line is asserted, *A’* is pulled down to ‘0’ and the cross-coupled inverter as well as BL pull up *A* to ‘1’.

The memory peripherals were tested with an assumed load of a 4KB SRAM core. It consists of a memory array of 128 rows by 256 columns of cells.

A layout of the SRAM cell was done. Figure 10 shows the variations of inverter placement in the SRAM cell.

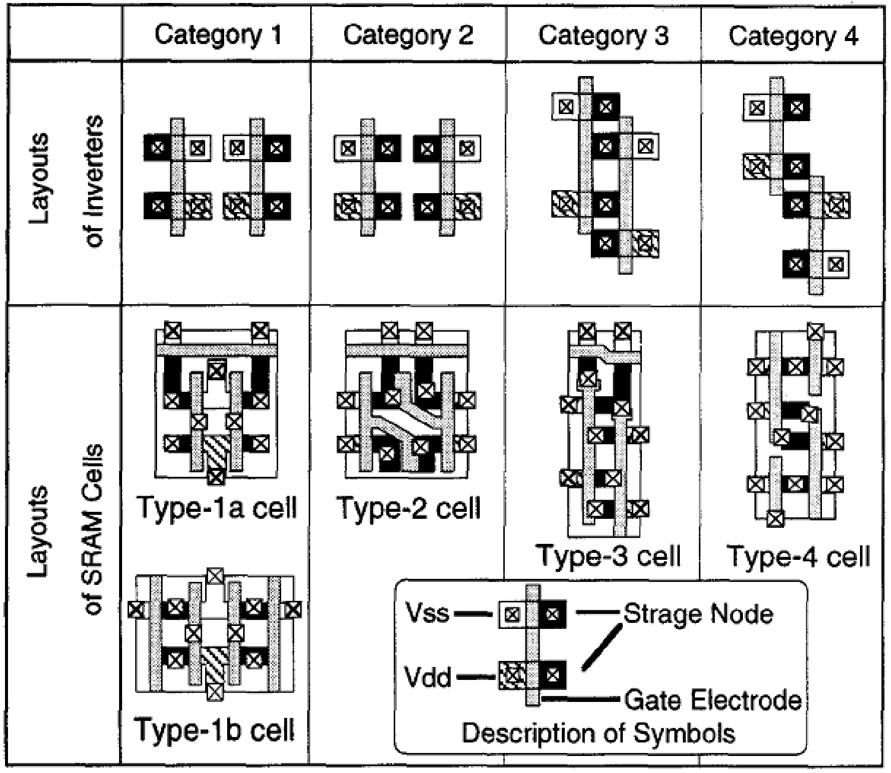

Figure 10: Variations of the inverter layouts and SRAM cell layouts. [18]

The SRAM layout done follows the inverter orientations in category 2 found in Figure 10. Below is an image of the SRAM cell layout. It has an area of 6.8906 μm^2.

The 4KB SRAM bank has the measured dimensions 590.4 μm x 448 μm for an area of 264.499 x 10^3 μm^2. Source/Drain sharing was done between cells and NWELL contacts were placed after every 16th cell.

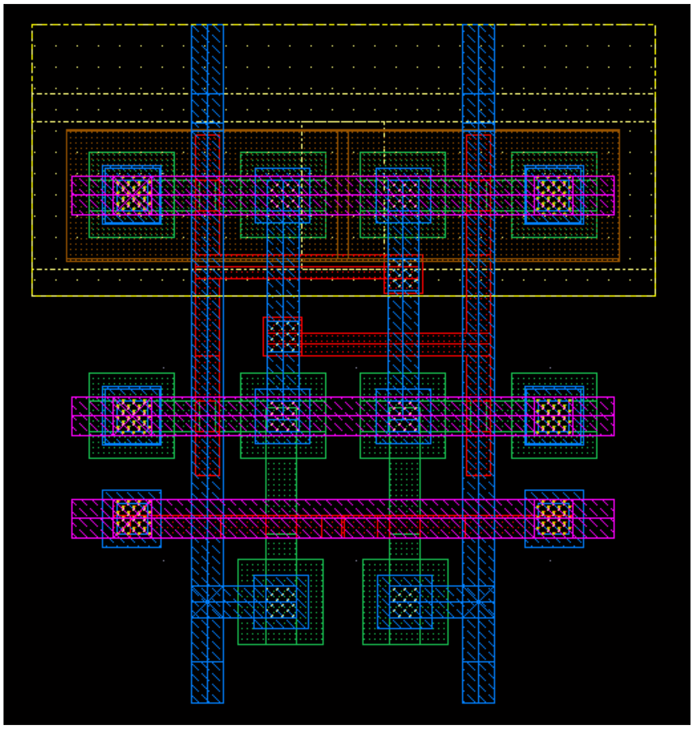

Figure 11: SRAM Cell layout

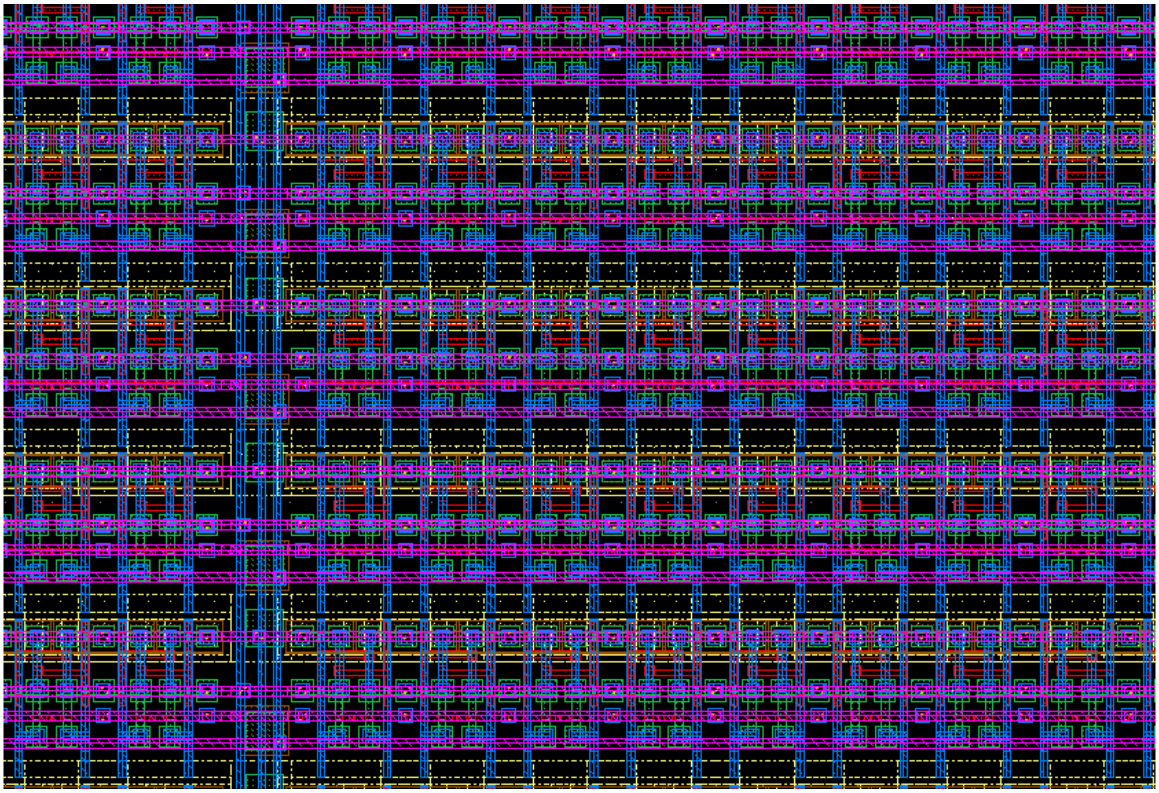

Figure 12: Zoomed- in image of the layout of the SRAM bank

2.3 Address Decoder

2.3.1 2 to 4 Decoder

Because it is in the nature of ECL logic to have both true and complimentary versions of data, N to 2^N decoding can be performed by N-input AND gates for each output.

A 2 to 4 decoder is made up of four AND gates. A stacked CML topology is used. Figure 13 shows the schematic of the decoder. The layout for the AND gate is shown in Figure 14.

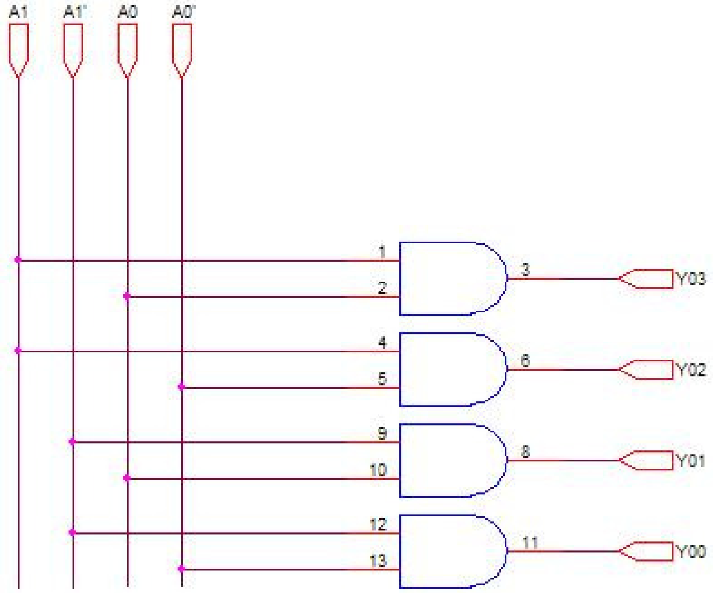

Figure 13: 2 to 4 decoder.

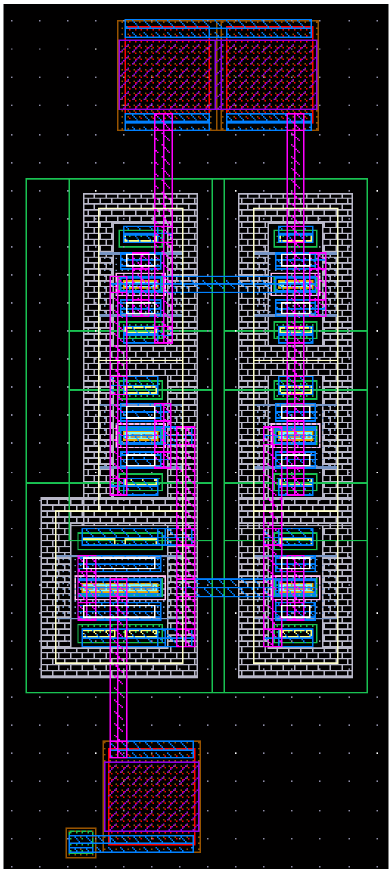

Figure 14: Layout of AND gate.

2.3.2 4 to 16 Decoder

The 4 to 16 decoder was done in two stages to reduce the fan in of the gates used. In addition, if the decoder was done in a single stacked CML AND gate, the supply voltage would have to be increased (made more negative) to accommodate the added logic levels. Two versions of the 4 to 16 decoder were done. One version uses a cascade of CML AND gates. The second version makes use of wired-or logic and single-ended NOR gates.

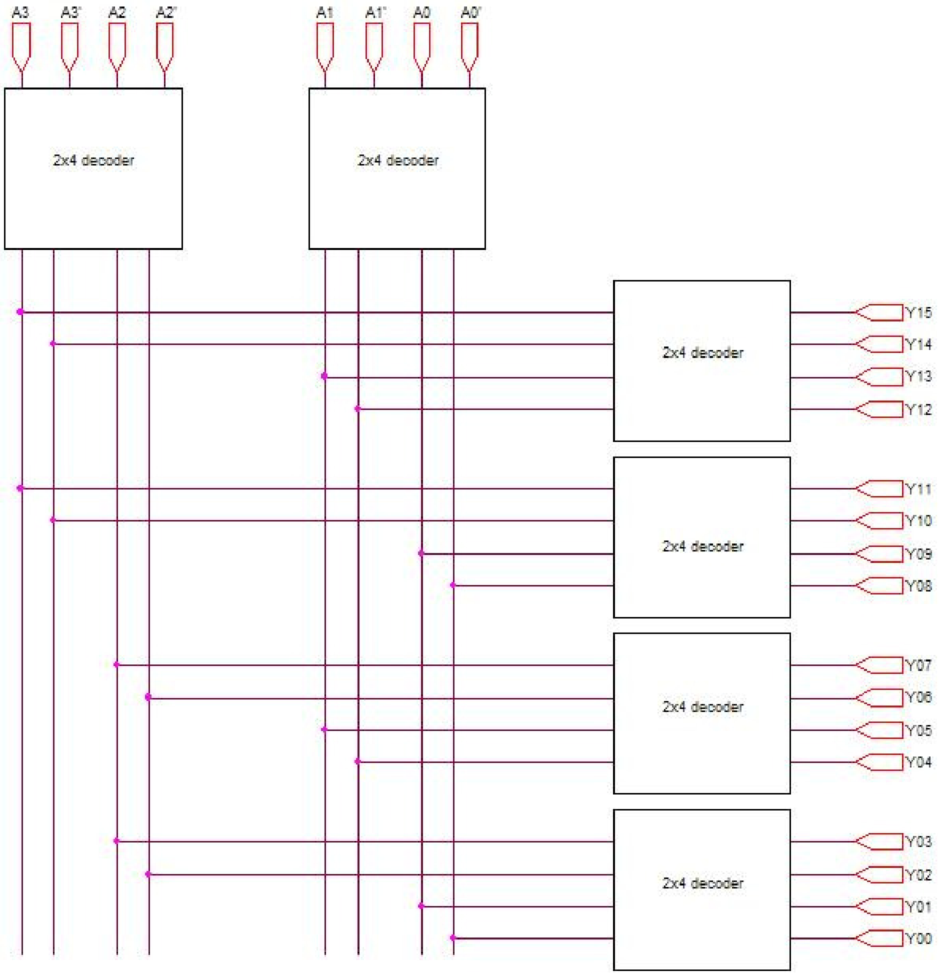

Figure 15: Block diagram of the 4 to 16 decoder.

The first 4 to 16 decoder, as mentioned, is made up of two stages of CML AND gates such as shown in Figure 7. Each CML AND gate acts as a 2 to 4 decoder. Figure 15 shows a block diagram of this decoder.

Because both true and complimentary versions of the input are available DeMorgan's rules can be used liberally. The outputs then take the following form.

$$Y_{00} = \overline{A_3}\cdot\overline{A_2}\cdot\overline{A_1}\cdot\overline{A_0} = \overline{(A_3+A_2)+(A_1+A_0)}$$

$$\overline{Y_{00}} = (A_3+A_2)+(A_1+A_0)$$

$$Y_{01} = \overline{A_3}\cdot\overline{A_2}\cdot\overline{A_1}\cdot A_0 = \overline{(A_3+A_2)+(A_1+\overline{A_0})}$$

$$\overline{Y_{00}} = (A_3+A_2)+(A_1+\overline{A_0})$$

$$\vdots$$

$$Y_{07} = \overline{A_3}\cdot A_2\cdot A_1\cdot A_0 = \overline{(A_3+\overline{A_2})+(\overline{A_1}+\overline{A_0})}$$

$$\overline{Y_{00}} = (\overline{A_3}+\overline{A_2})+(\overline{A_1}+\overline{A_0})$$

$$\vdots$$

$$Y_{15} = A_3\cdot A_2\cdot A_1\cdot A_0 = \overline{(\overline{A_3}+\overline{A_2})+(\overline{A_1}+\overline{A_0})}$$

$$\overline{Y_{00}} = (\overline{A_3}+\overline{A_2})+(\overline{A_1}+\overline{A_0})$$

The decoder then becomes two stages of OR gates. The inputs are first predecoded using wired-or logic. It is then decoded using a single ended input CML OR/NOR gate. Figure 16 shows a partial schematic of this decoder. Figure 17 shows the layout of the decoder.

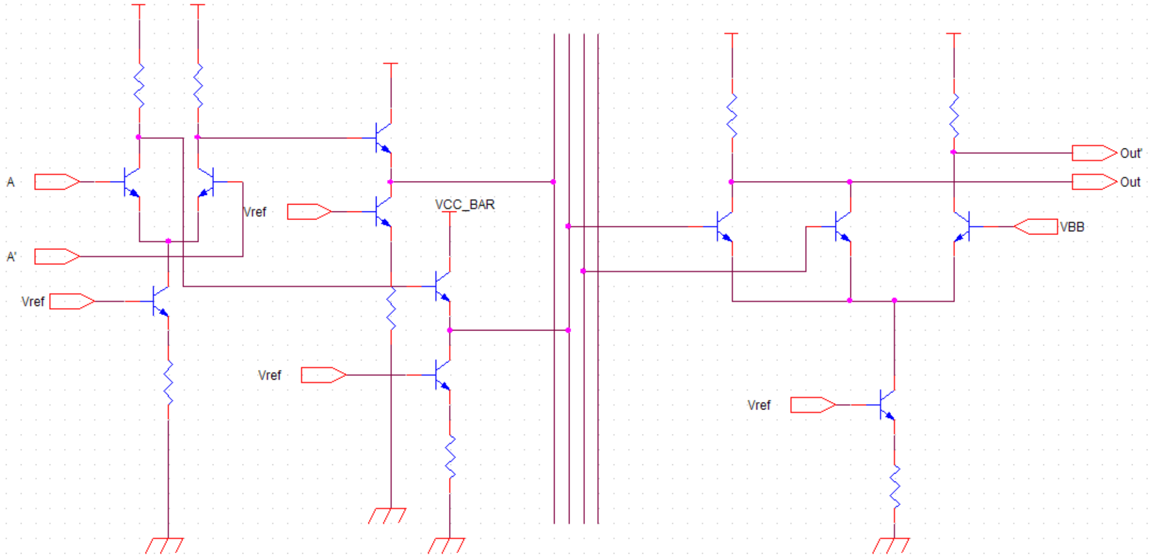

Figure 16: Partial schematic of the 4 to 16 decoder.

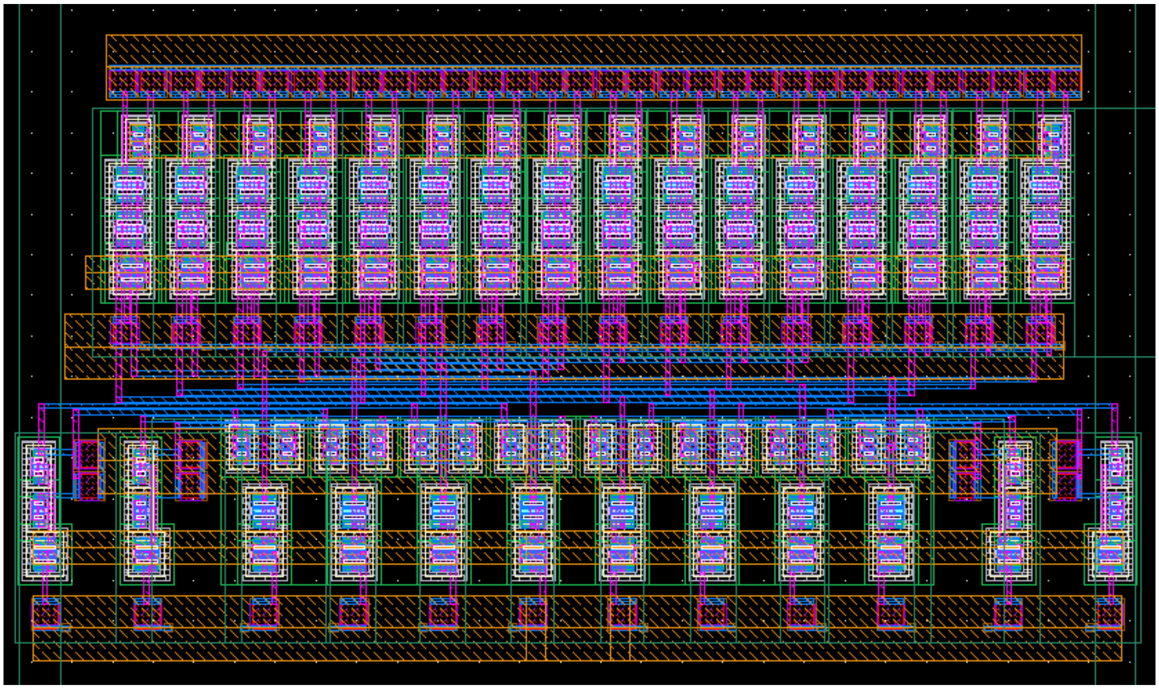

Figure 17: Layout of the 4 to 16 decoder.

It is found that this setup is slightly slower than using a cascade of CML AND gates. However, if a reference voltage is readily available then this setup provides a simpler circuit with fewer transistors and without the need of skew free complimentary inputs, at a very slight cost in performance. When used in the 7x128 decoder, near identical outputs where achieved.

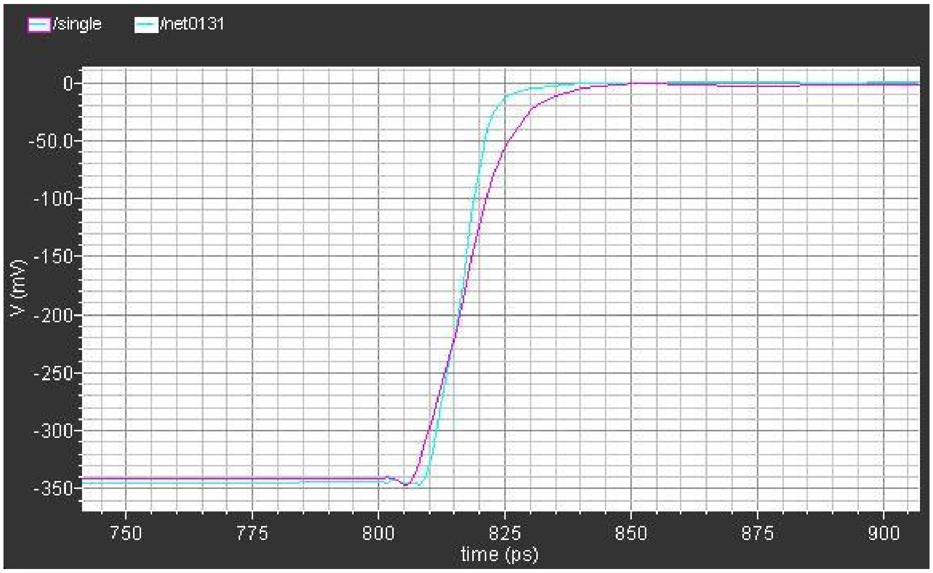

Figure 18: Comparison of the two implementations of a 4 to 16 decoder.

2.3.3 3 to 8 Decoder

The 3 to 8 decoder is done in the same way as the 4 to 16 decoder. A_1 and A_0 are predecoded through a wired-or while A_3 is passed through a level shifter so as to match the output level of the wired-or. Then, once again CML OR/NOR is employed to do the final decoding. The schematic and layout are shown below.

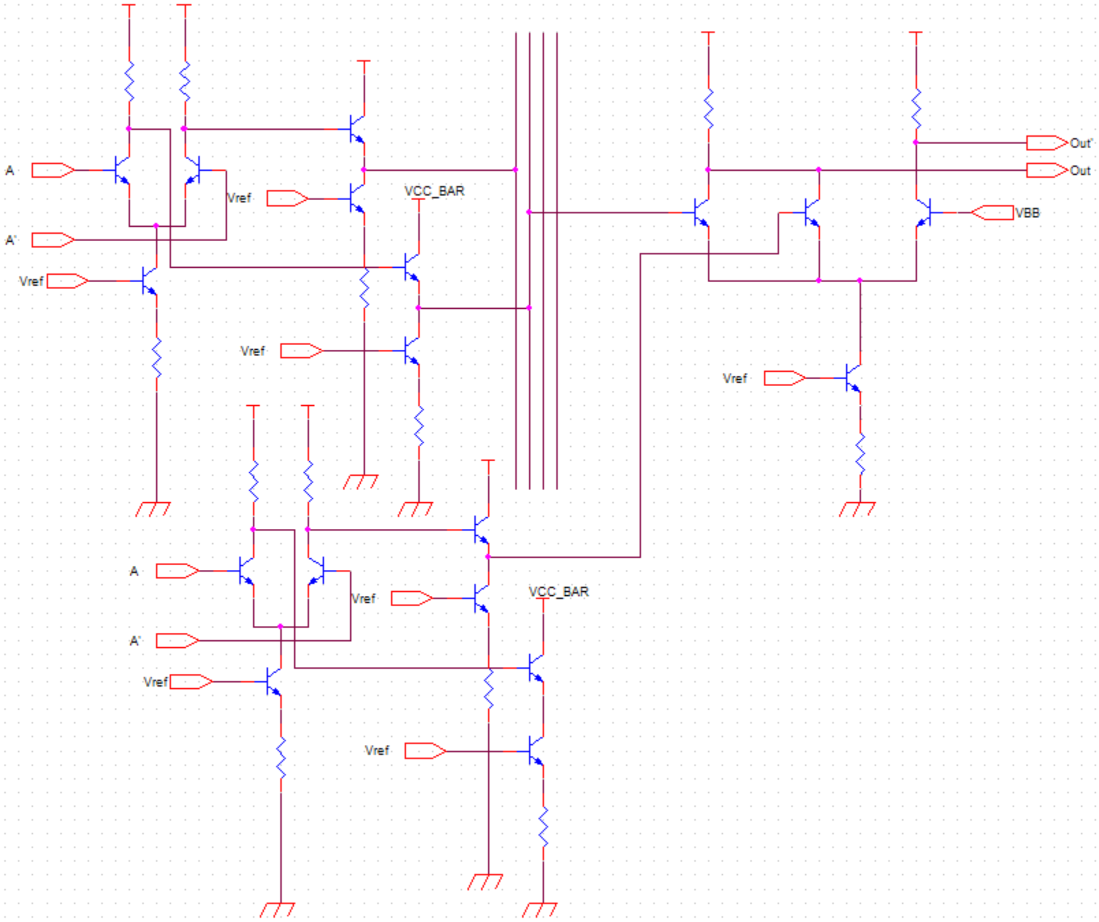

Figure 19: Partial schematic of the 3 to 8 decoder.

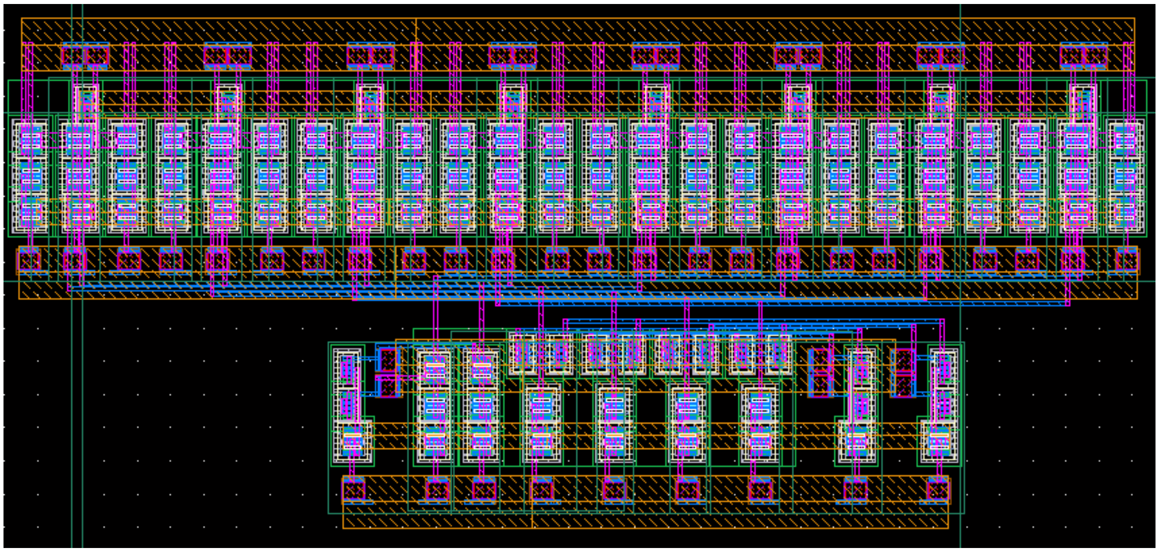

Figure 20: Layout of the 3 to 8 decoder.

2.3.4 7 to 128 Decoder

The 7 to 128 decoder is completed by passing the outputs of the 3 to 8 and 4 to 16 decoder through AND gates.

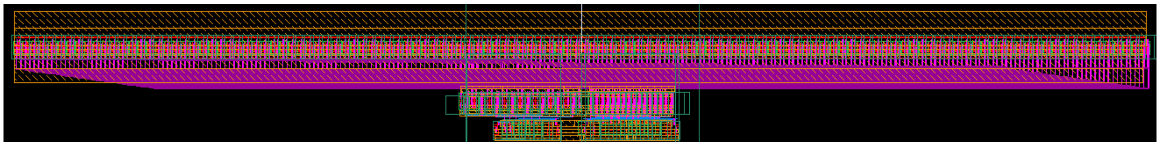

Figure 21: Layout of 7 to 128 decoder.

2.4 CML/ECL to CMOS Translator

The CML/ECL to CMOS translator takes an input with a voltage swing in CML/ECL logic (-0.85 to -1.2V) and translates it to full swing for use with the CMOS circuits. It consists of a CMOS differential amplifier whose outputs connect to CMOS an inverter which amplifies the voltage swing to CMOS levels. Below are the schematic and layout of the circuit. The layout for this circuit and the driver were designed so that they match the height of the memory cell.

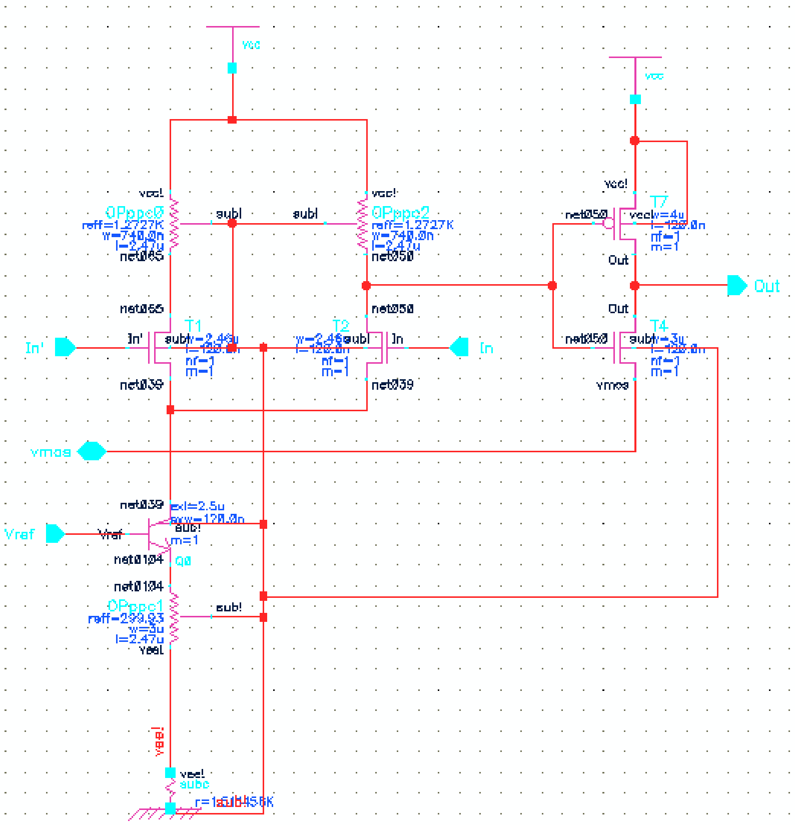

Figure 22: CML/ECL to CMOS translator Schematic.

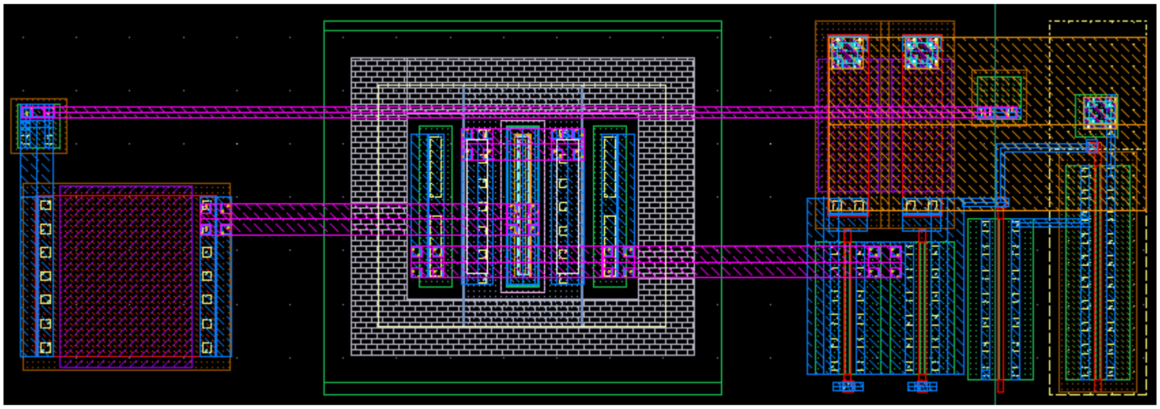

Figure 23: CML/ECL to CMOS translator Layout.

2.5 Word Line Driver

In order to achieve proper signals on the large capacitance on the word line, a driver must be used. If a CMOS driver was to be used it would be made up of a cascade of inverters progressively increasing in size. The number of stages (*N*) required would be equal to

$$N = \log_{3.59}\left({C_L}\big/{C_{in}} \right) \quad \textbf{[17]}$$

where C_L is the load capacitance and C_{in} is the input capacitance of the first inverter. The inverters would then increase in size by a factor of

$$f = \sqrt[N]{\frac{C_L}{C_{in}}}$$

times the size of the previous stage, with the first inverter being unit size. Because of the large capacitance to be driven, the CMOS driver would have to be very large.

Because of this a BiCMOS driver is used instead of a CMOS driver. The basis of the BiCMOS driver is the BiCMOS inverter shown in Figure 24. When V_{in} is high then M_1 is on, causing Q_1 to conduct resulting in a low output. When V_{in} is low then M_2 is on, causing Q_2 to conduct resulting in a high output. In steady state, Q_1 and Q_2 are never on simultaneously.

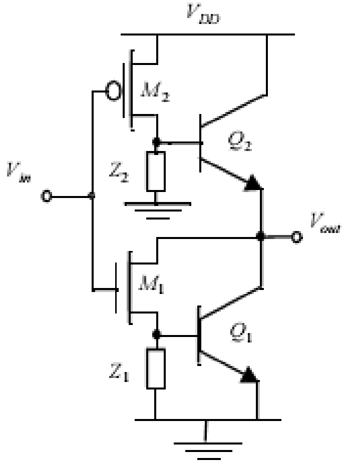

Figure 24: Generic BiCMOS Inverter

The impedances Z_1 and Z_2 are necessary to remove the base charge in transistors Q_1 and Q_2 when they are being turned off. Variations of this circuit include replacing these impedances with NMOS.

Figure 25 and Figure 26 shows the BiCMOS driver implemented. The impedance Z_2 is replaced by NMOS and additional circuitry is added to achieve rail to rail swing as the BJT prevents it.

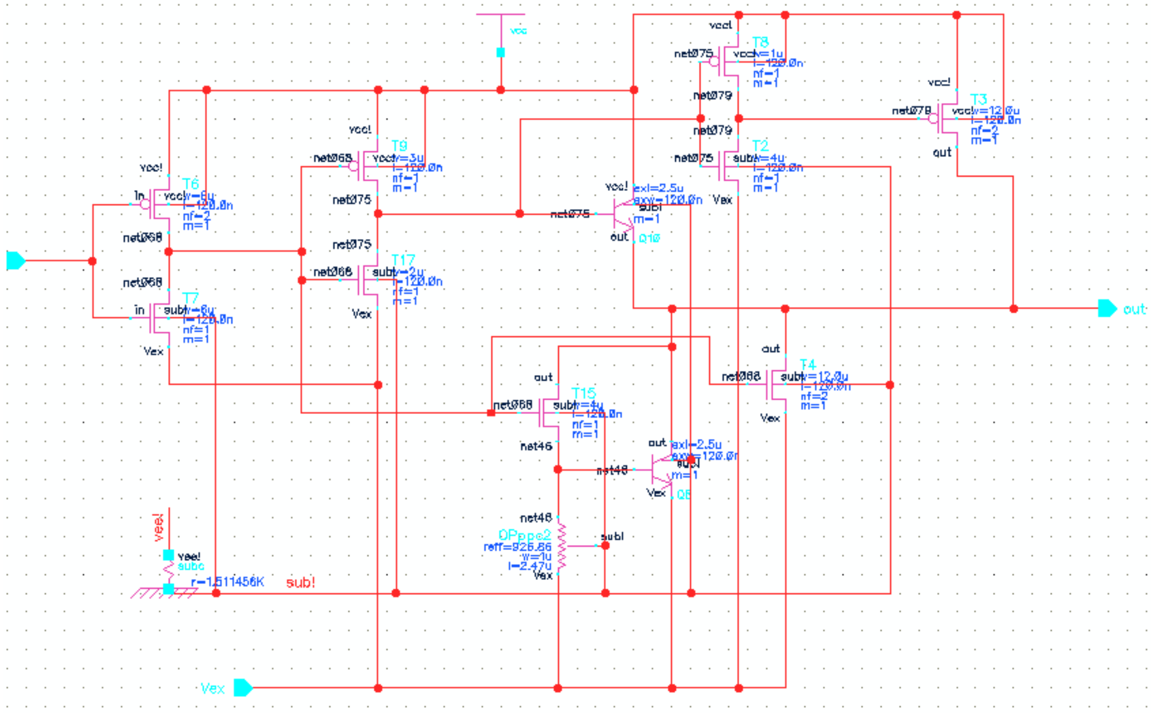

Figure 25: BiCMOS Driver Schematic.

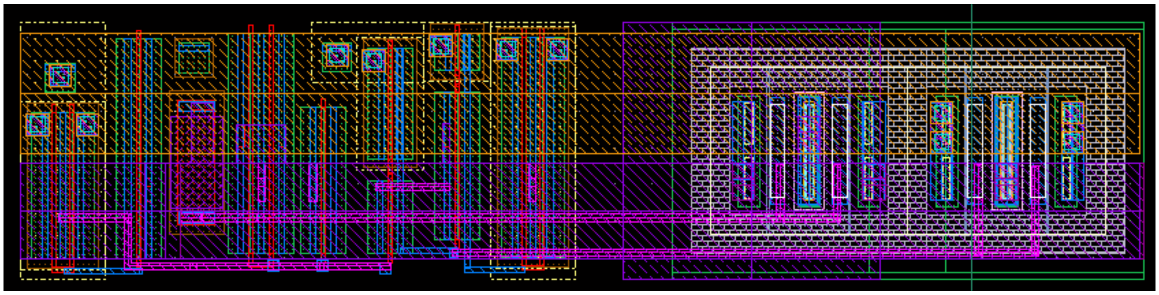

Figure 26: BiCMOS Driver Layout.

2.6 Precharge Circuit

For proper operation of the SRAM the bit lines must be prepared for read and write operations. The SRAM Cell is designed to be weak at passing ‘1’s in order to inadvertently write data during a read operation. Because of this reasons it is necessary to precharge the bit lines before any operation. Figure 27 shows the precharge circuit. It is composed of three PMOS transistors. Transistors Q1 and Q2 are used to pull up the bit line. The third PMOS is used to equalize the potential across the bit lines as it is necessary for there potential to be equal for fast sensing during a read operation.

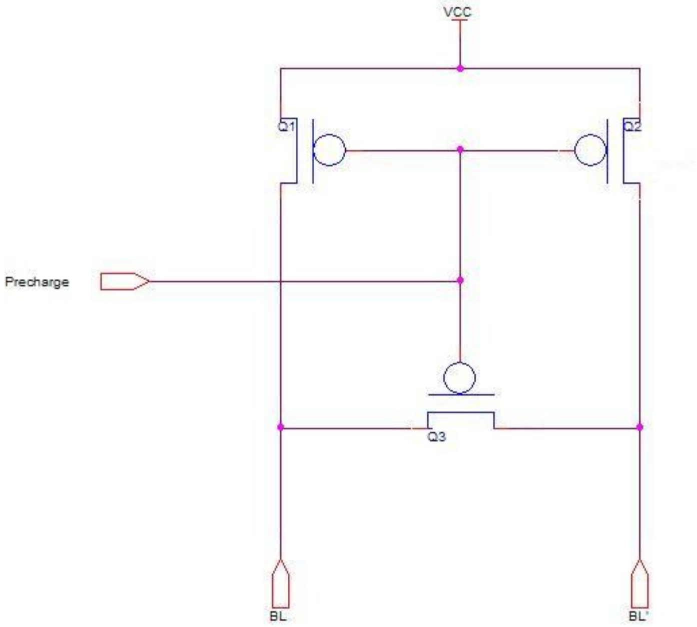

Figure 27: Precharge Circuit

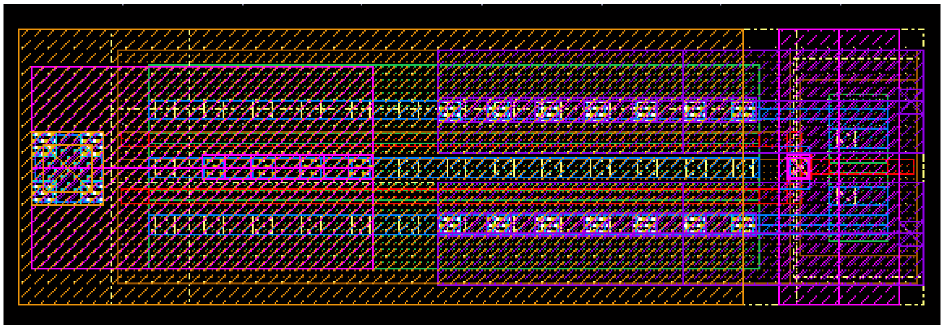

Figure 28: Layout of the precharge circuit.

2.7 Read Circuit

As the number of cells on each bit line increases, the capacitance on the bit line increases. This capacitance has a large effect during a read operation because the current from a memory cell is typically low. Because of this, a sense amplifier is necessary.

Before a read operation the bit line conditioning circuit precharges the bit lines to the same voltage. Then, when the word line is raised, a differential amplifier amplifies the resulting potential difference between the bit lines. A feedback circuit turns the differential amplifier into a latch.

Below are the schematic and layout of this circuit. The transistors Q1 and Q2 act as a differential amplifier. They translate slight changes on the bit lines into ECL level. Q4 and Q5 store this value when the bit line is not being read.

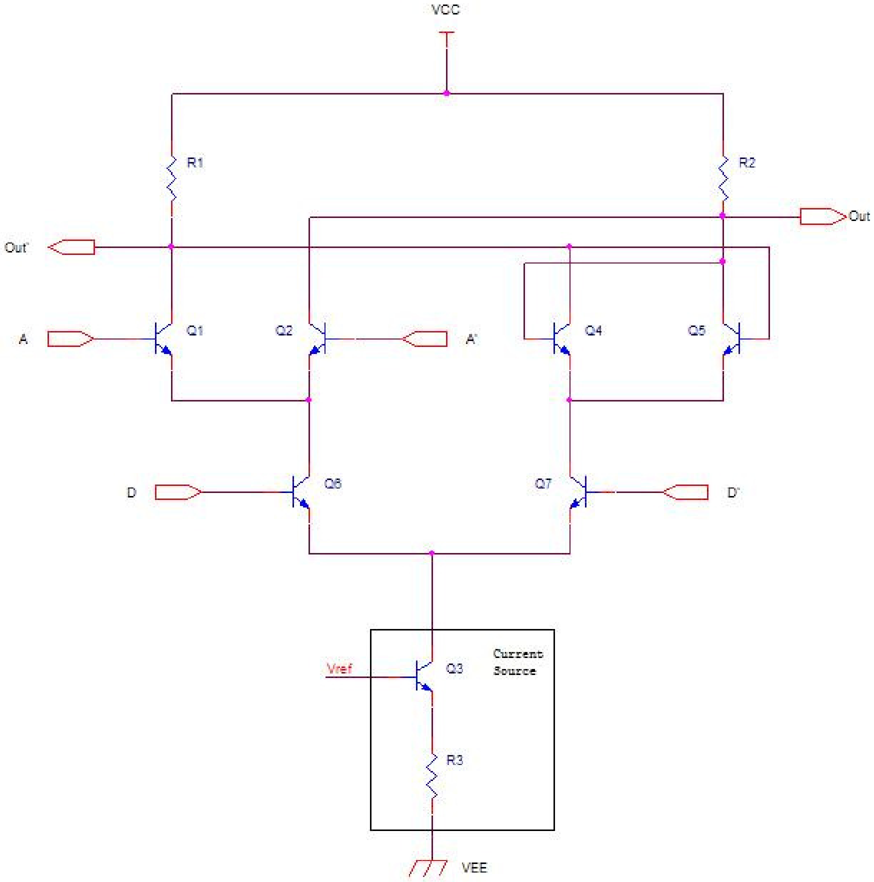

Figure 29: Schematic of the Read Circuit (D-Latch)

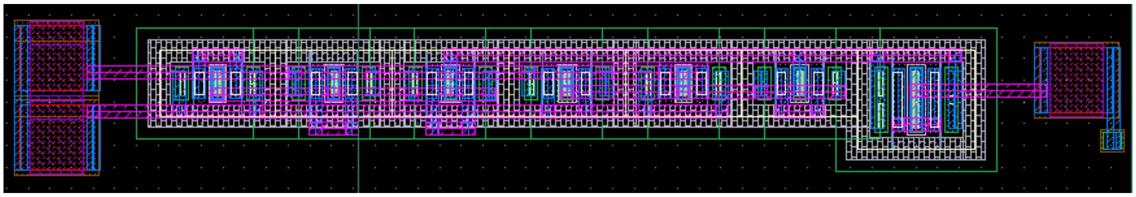

Figure 30: Layout of the read circuit.

2.8 Write Circuit

During a write operation, as previously mentioned, the bit lines are initially precharged. While the write pulse is low the transistors on the bit line are off allowing the bit line to be precharged. After the bit lines are precharged and the write pulse is high, DI and its complement are allowed to pass. The write circuit reverses the data so DI is connected to the write circuit that connects to BL’ and DI’ is connected to the write circuit that connects to BL. If DI is ‘1’ then BL’ is pulled down, otherwise BL’ is kept high. If DI’ is ‘1’ then BL is pulled down. Figure 31 shows the schematic of the circuit after the CML/ECL to CMOS translator. Figure 32 shows the layout including the translators.

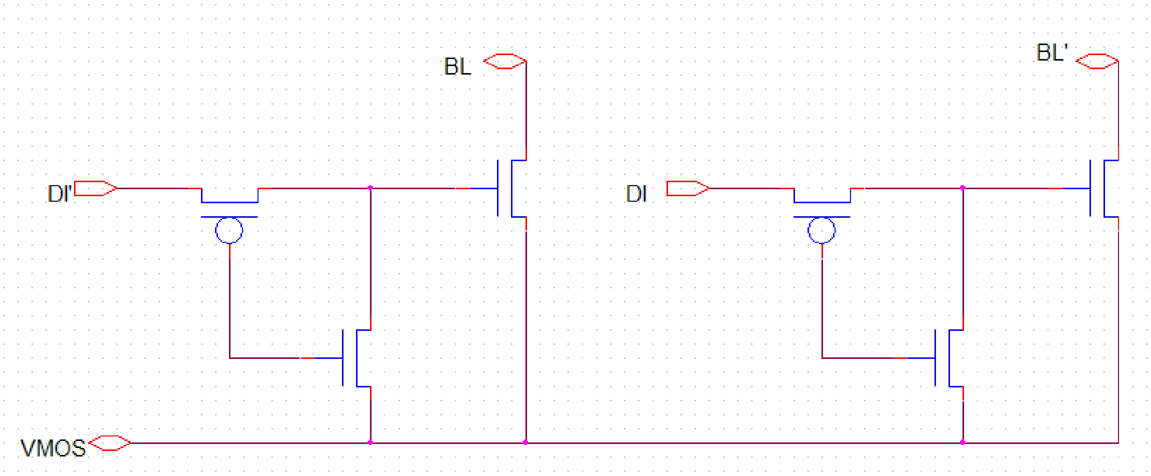

Figure 31 Write Circuit

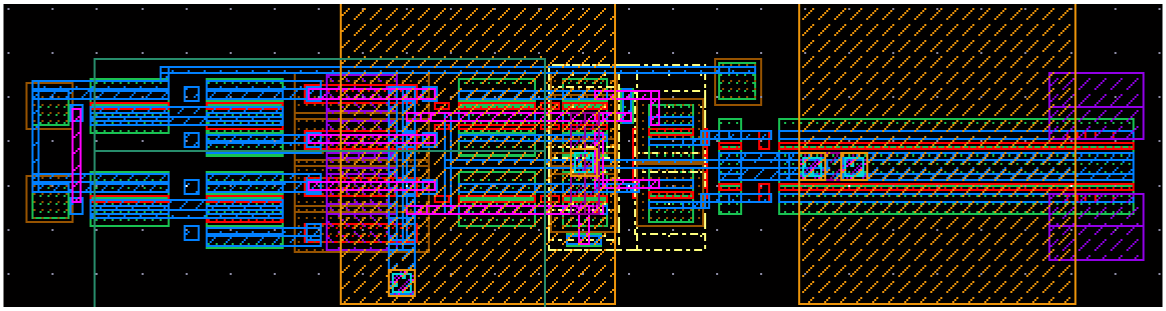

Figure 32: Layout of the write circuit.

3. Results

A 4KB SRAM core is first assumed. The load capacitances on the word line and bit line were taken from [10]. In this work the estimated word line capacitance is 278.016fF. The bit line capacitance is estimated at 105.15456fF.

As previously described, the memory core is made up of 128 rows by 256 columns of cells. The word line capacitance is from two NMOS for 256 cells in one row and wire capacitance. The bit line capacitance is from 128 access transistors and wire capacitance. These capacitances were estimated in [10] where the data was taken from the BiCMOS 8HP kit. This was done as follows.

Gate capacitance = 1.3 +/- 0.25 fF/μm

Worst case gate capacitance = 1.55 fF/ μm

Source/drain capacitance = 0.33 +/- 0.045 fF/ μm

Worst case S/D capacitance = 0.3345 fF/ μm

The capacitance on the word line is from the gate capacitance of 512 access transistors and the wire capacitance on the word line.

Word line capacitance = 512 * (1.55 fF/ μm * 0.16 μm) + 0.2 fF/ μm * 590.4 μm

Word line capacitance = 126.976 fF + 118.08 = 245.056 fF

The capacitance on the bit line is due to 128 access transistors and the wire capacitance on the bit line.

Bit line capacitance = 128 * 0.3345 fF/ μm * 0.16 μm + 0.2 fF/ μm * 448 μm

Bit line capacitance = 6.85056 fF + 89.6 = 96.451 fF

The operation of the SRAM is as follows. First the bit lines are precharged to prepare them for operation. During this time the decoder address should be set to all zero's making that memory row unusable. The precharge signal is low to precharge and high for off. Once the precharge signal is off an address is accepted and decoded through a 7 to 128 decoder. The output of the 7 to 128 decoder is then passed through a CML to CMOS translator and then to the BiCMOS driver that connects to the word line.

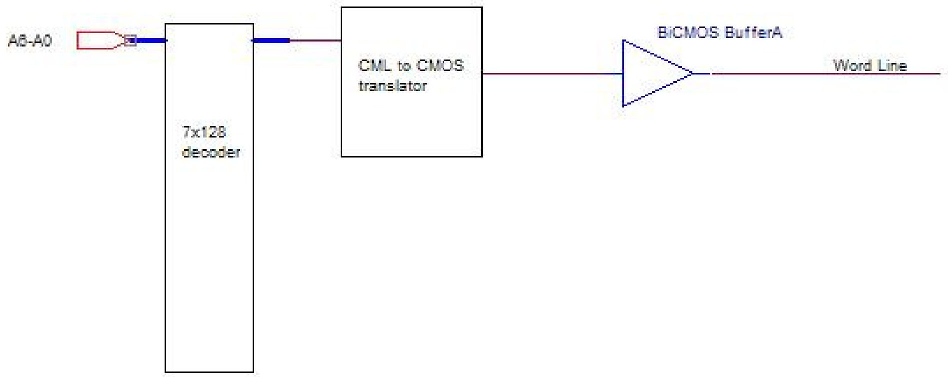

Figure 33: Block Diagram of the path from decoder to the word line for the SRAM

During a write operation the appropriate bit lines are pulled down according to the DI (data input). The activated word line allows access to a row of SRAM cells and the data is written.

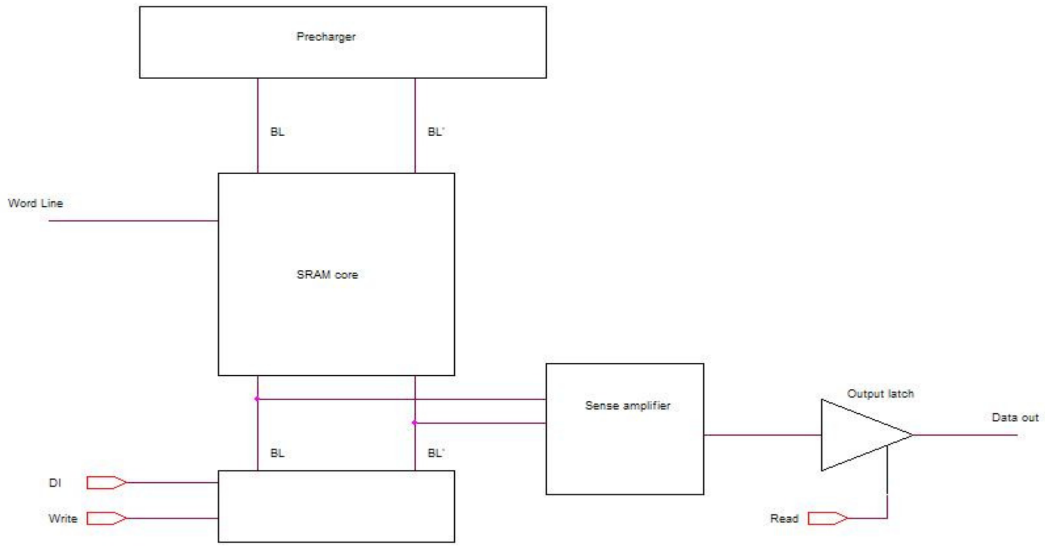

Figure 34: Block diagram including the word line, circuits around the bit line and the output latch for the SRAM.

During a read operation the write circuit does not pull down any bit line. The activated row of cells then pulls down the appropriate bit lines. This is slow as the cells transistors are small and only a small amount of current passes through. The sense amplifier picks up on small changes on the bit line and amplifies it. Finally, the output is then passed through the output latch.

Below are the results of a simulation for the 4KB SRAM schematic.

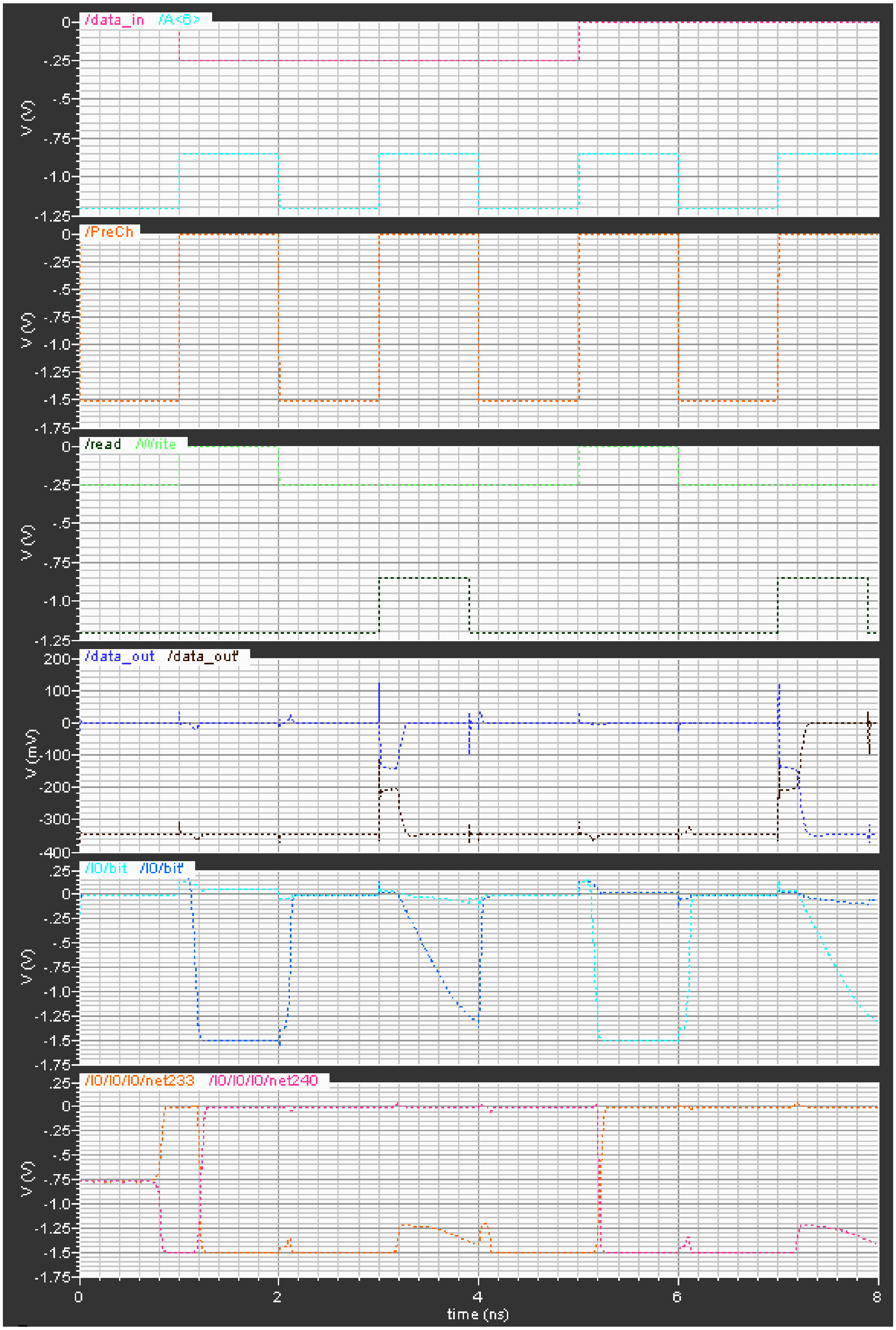

Figure 35: Simulation results for the Schematic

From 0-1ns the precharge signal is asserted. At 1ns a write operation begins. DI is set to low (-250mV) therefore bit (bit line) is pulled down to -1.6V. The inner nodes of the cell reach rail to rail data at approximately 1.4ns.

At 2-3ns the precharge signal is asserted again and the address points to zero. At 3 ns a read operation begins. BL can be seen being pulled down slowly by the cell. The sense amplifier detects and amplifies this change. Its output is initially meaningless because it is a differential amplifier and initially both bit lines are 0V. At 3.25ps data out reaches rail (-250mV).

The entire operation is repeated from 4 to 8 ns with DI set to high (0V). The inner nodes of the cell can be seen inverting between 5.35ns and 5.4ns. Data out is seen to change reaching 0V at 7.25ps.

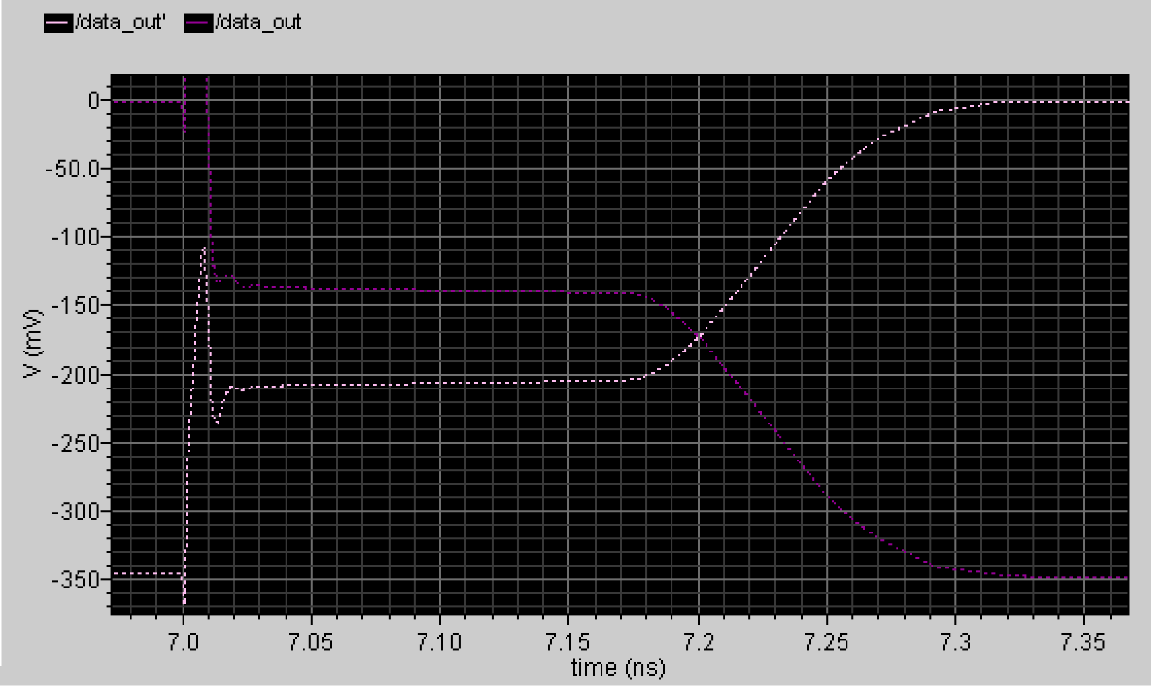

Figure 36: Output waveform of schematic.

Above is a zoomed in waveform from the same simulation showing data out and its compliment between about 6.9ns and 7.35 ns. It shows that the output reaches its mid-value at 7.2ns, and reaches rail (0V) at approximately 7.3ns, 300ps after the read operation starts.

Shown in Figure 37 is the layout of the SRAM. Its area is found to be 1863.68µm x 1552.78 µm or 2.89 cm^2. The sections in Figure 37 are the following:

A. The 7x128 decoder.

B. The memory array with the precharge circuits above.

C. The CML/CMOS converters and the BICMOS word line drivers.

D. The write circuits and the sense amplifers.

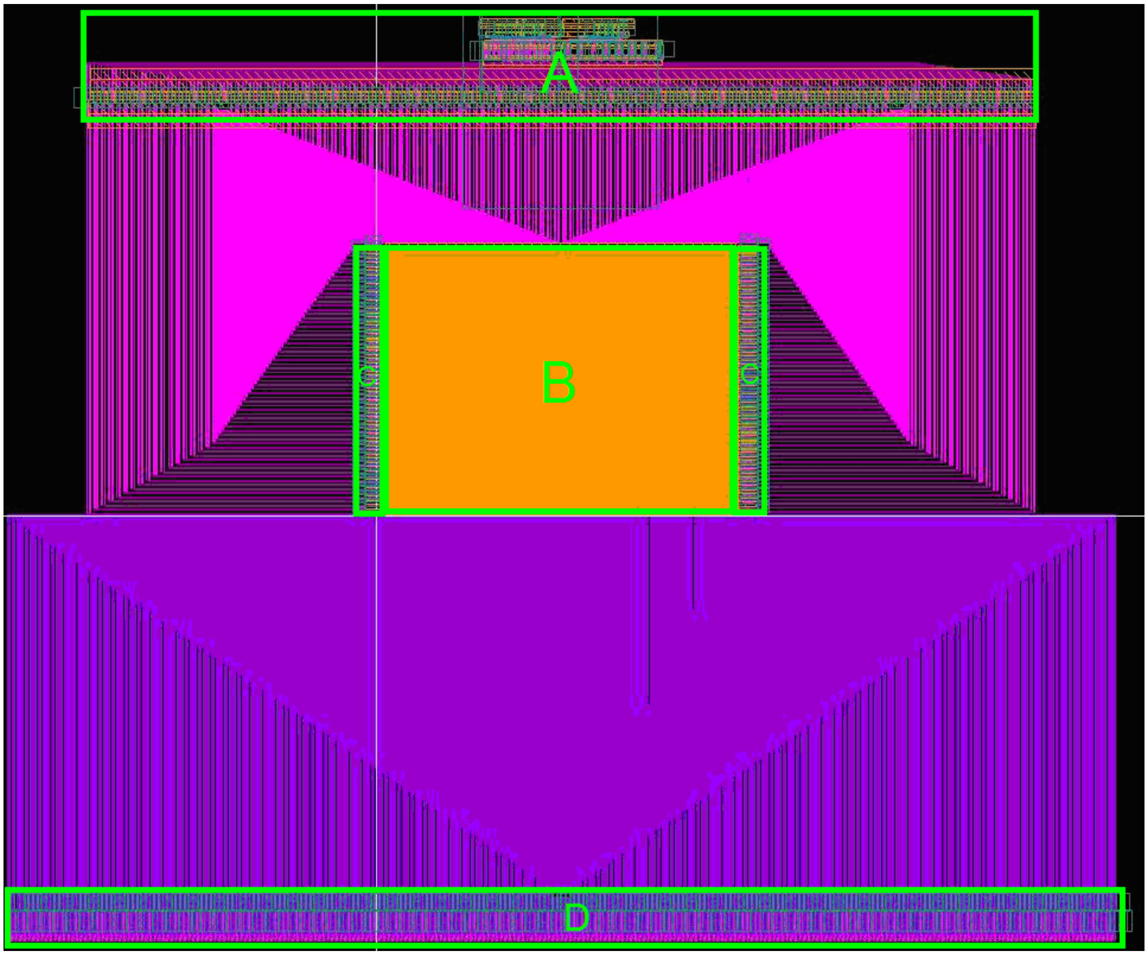

Figure 37: Layout of the entire design.

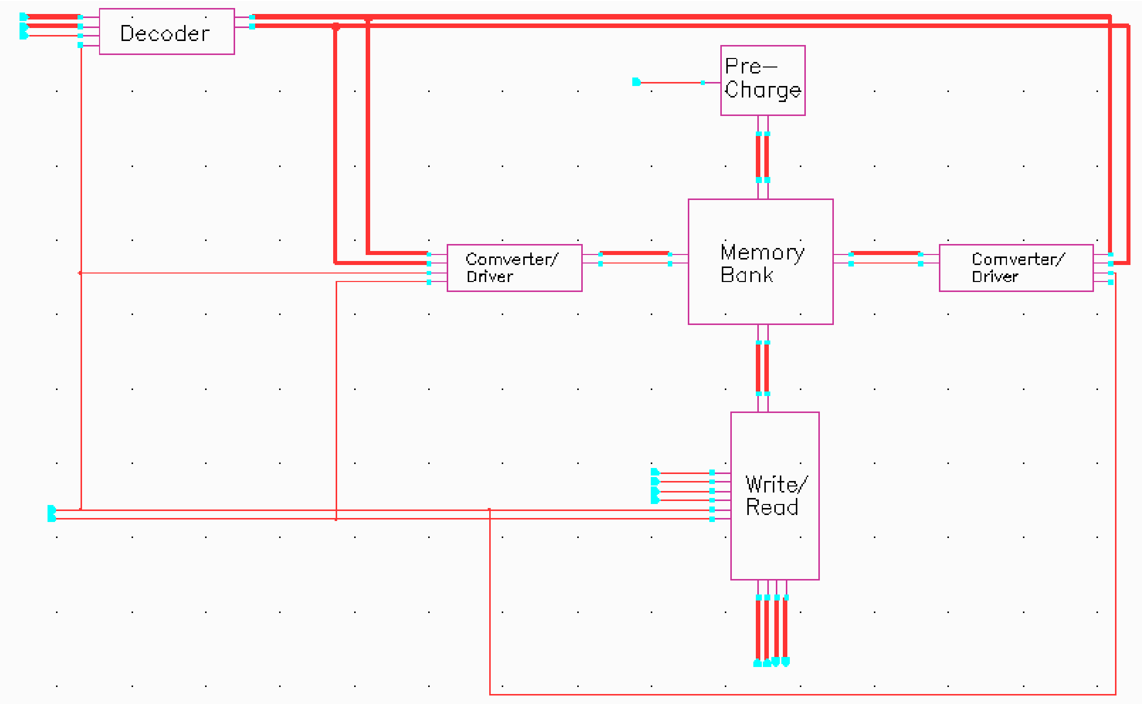

Figure 38: Schematic of the entire design.

Figure 39 shows the simplified layout used for simulations. It includes the complete 4x16 decoder and 3x8 decoder. After that, the parts that do not branch out from the path are not included.

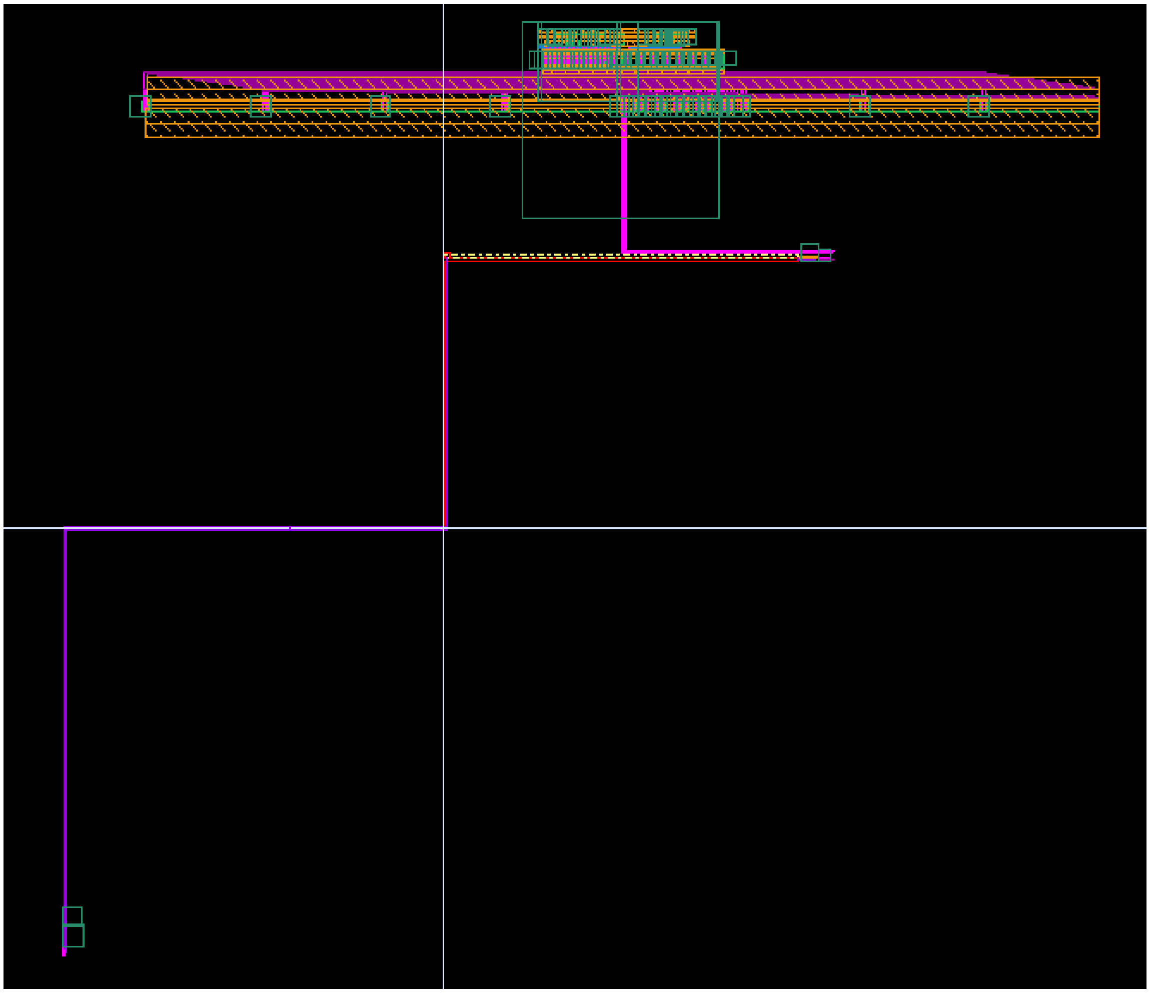

Figure 39: Layout used in simulations.

Figure 40 shows the simulation results for the layout. It confirms the functionality of the layout, the same as the simulation of the schematic. It should be noted that the time axis was doubled.

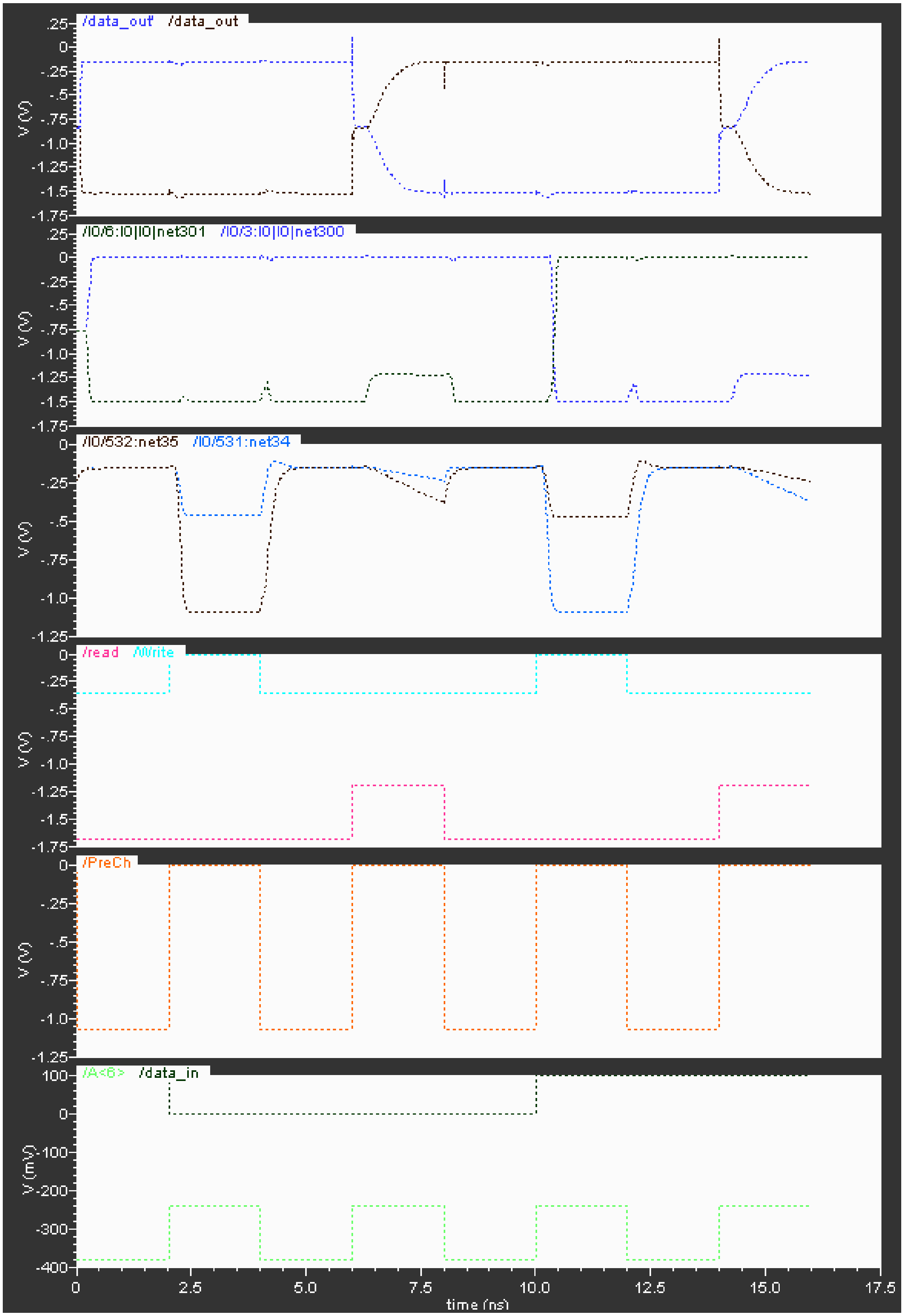

Figure 40: Simulation results for the layout.

Shown below is zoomed in image of the output waveform between about 13ns and 16ns.

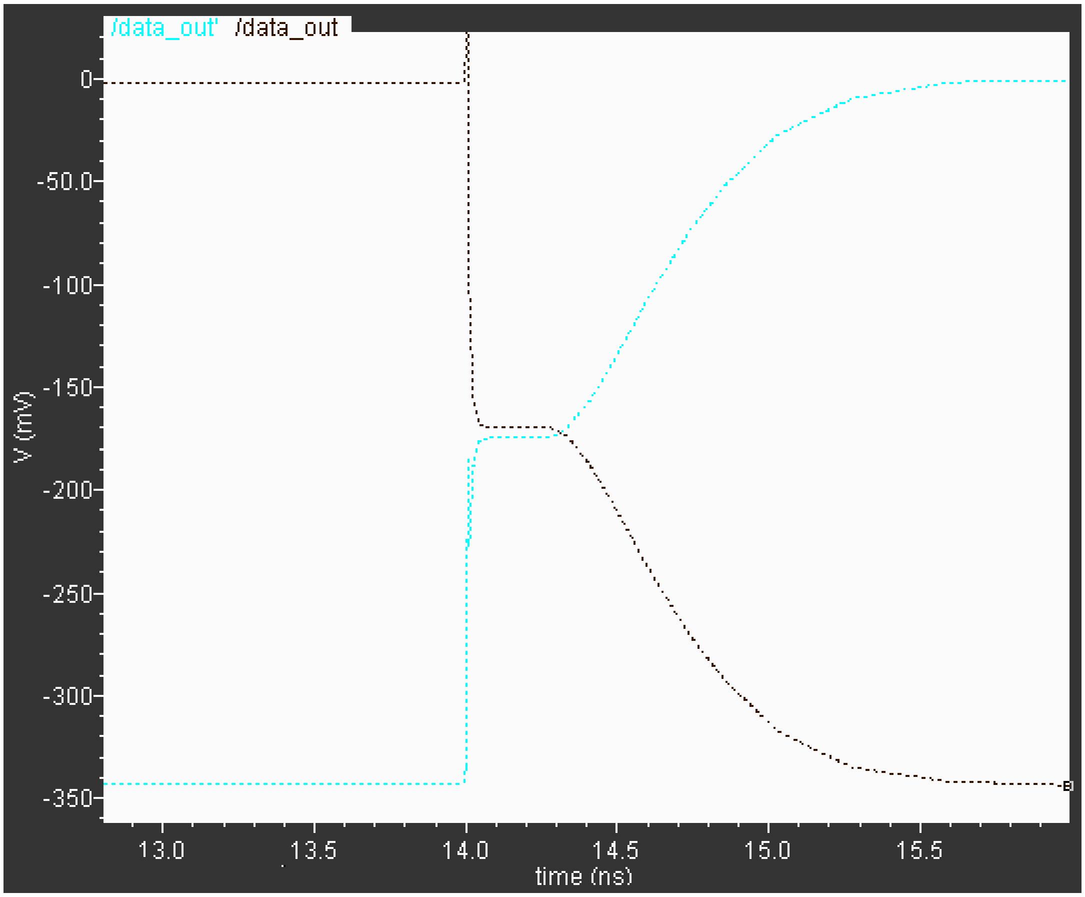

Figure 41: Output waveform of Layout.

The internal signals were probed and are shown below.

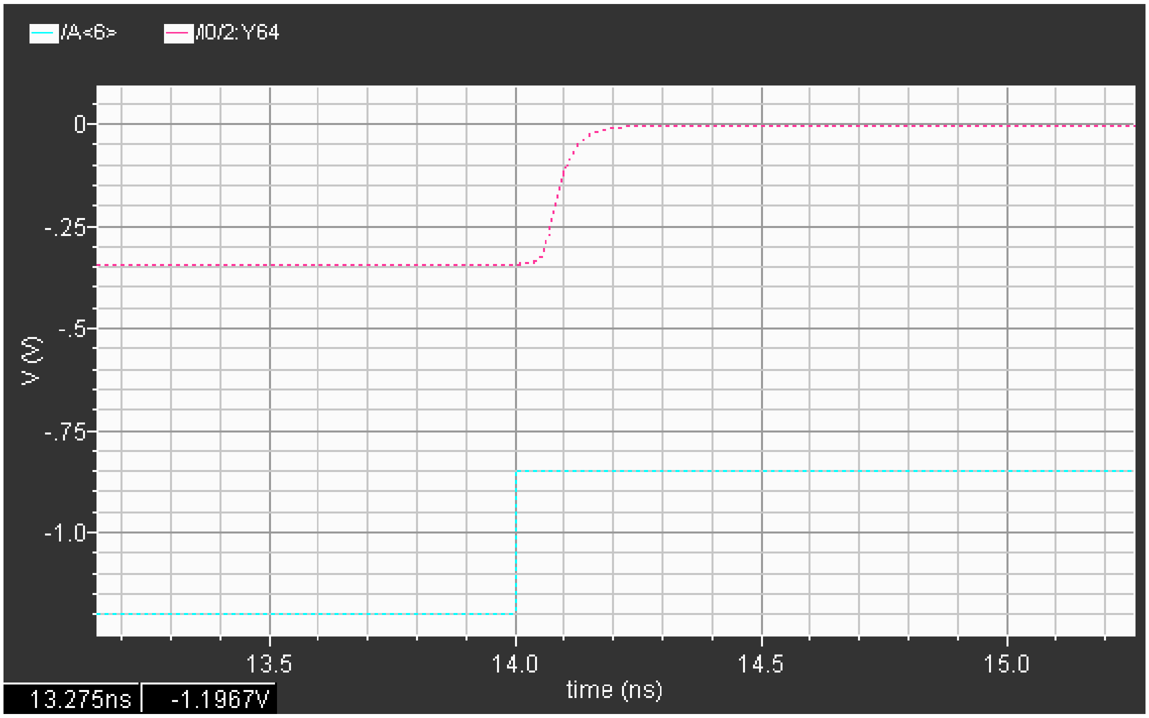

Figure 42: Decoder input and output signals.

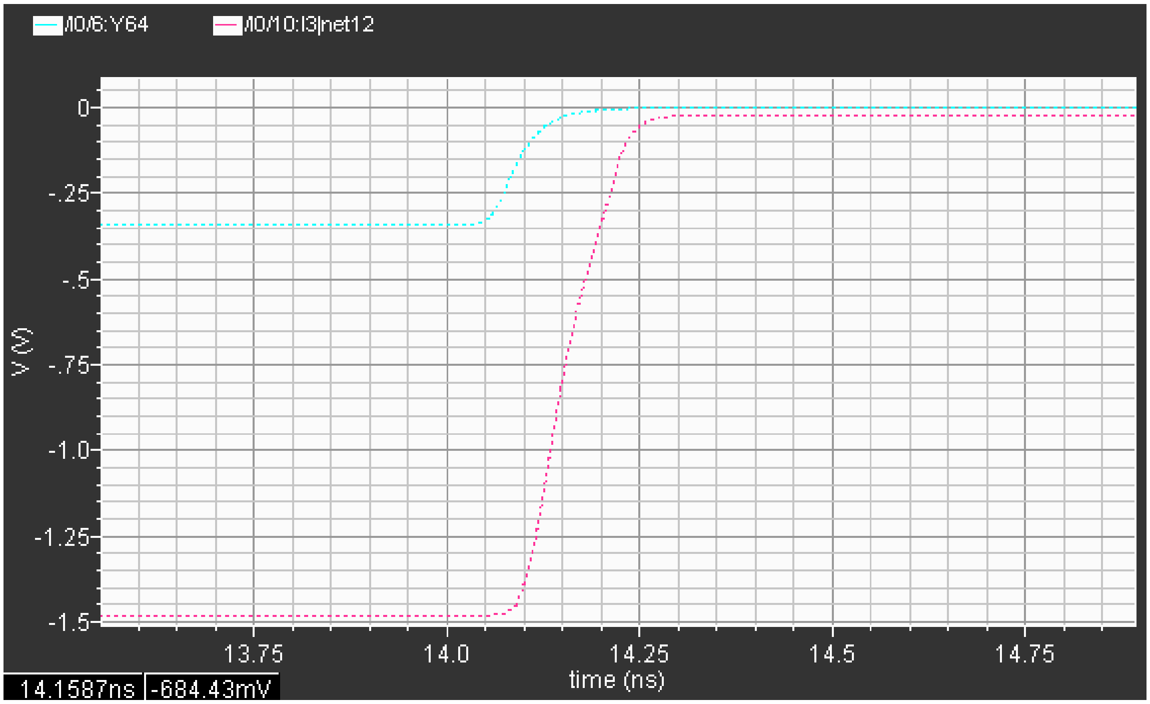

Figure 43:CML to CMOS Converter input and output signals.

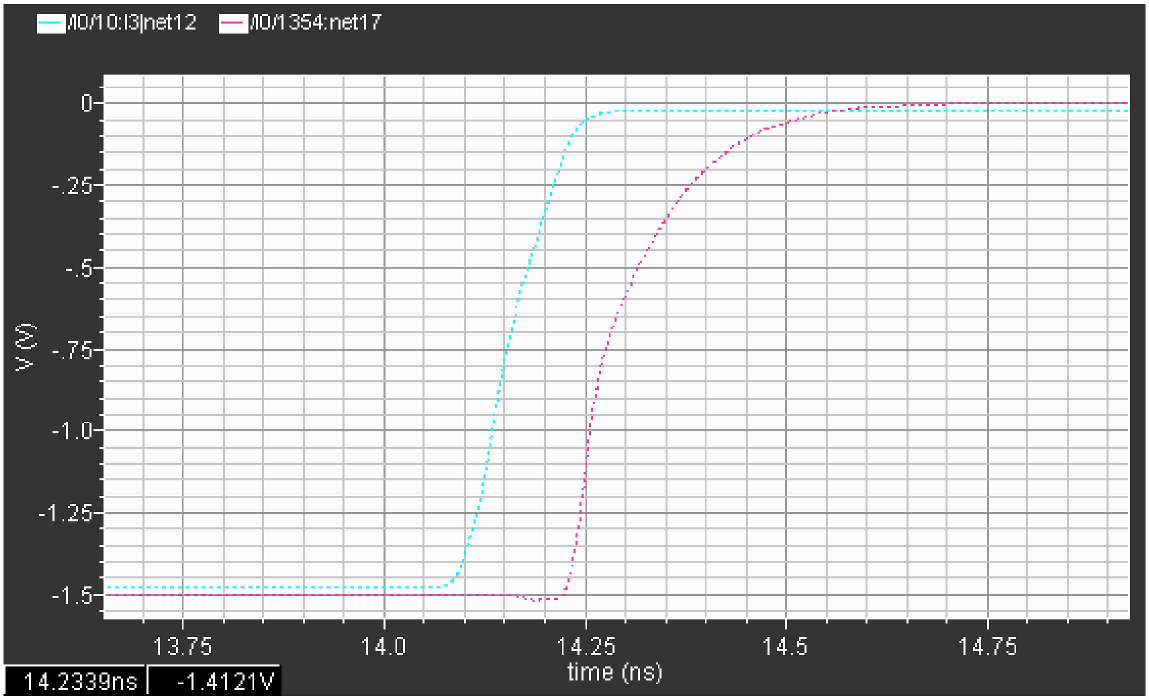

Figure 44: Word line driver input and ouptut signals.

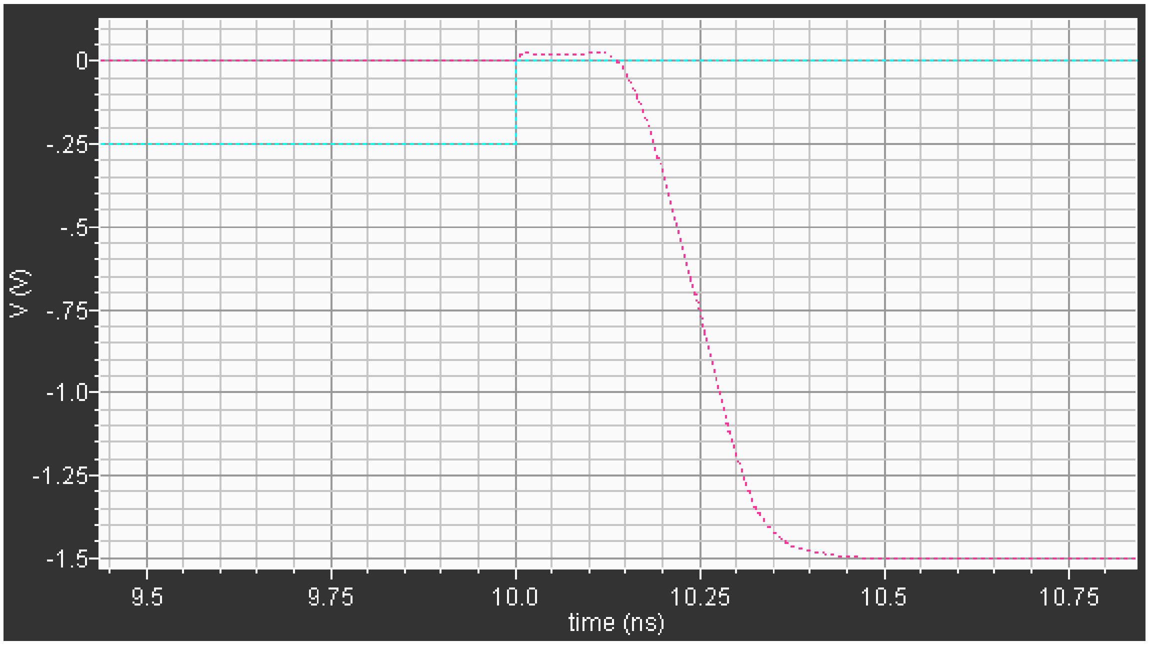

Figure 45: Write request and bit line.

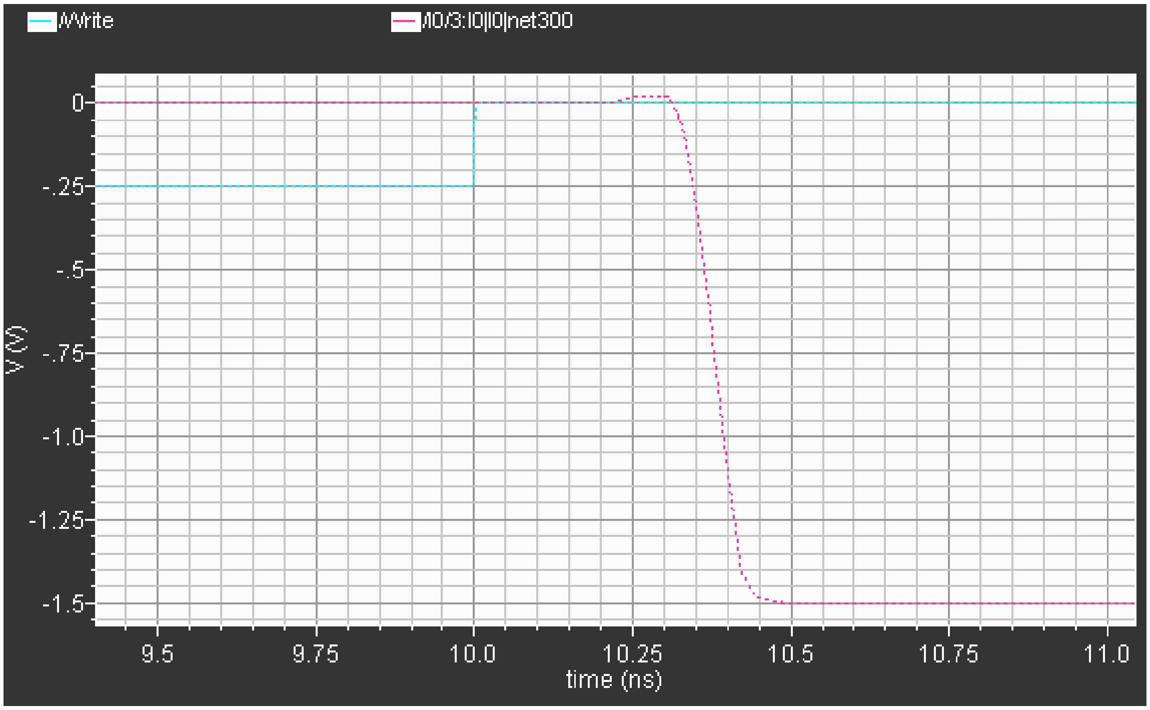

Figure 46: Write request and SRAM cell inner node.

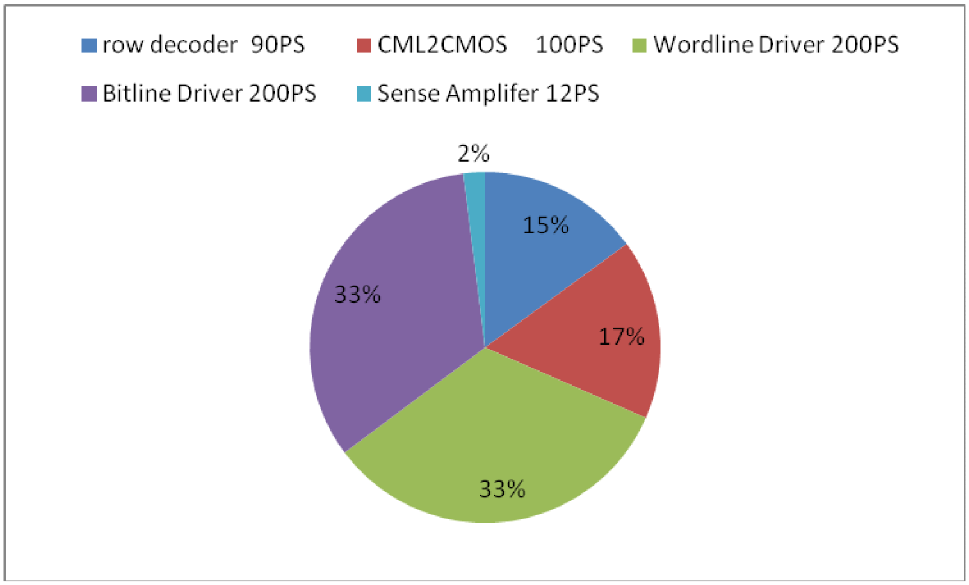

Figure 47: Breakdown of delay during a read operation from the post-layout simulation.

Figure 47 shows a pie chart illustrating the breakdown of the delay during a read process. In comparison Figure 48 shows the breakdown of the delay from a CACTI simulation of a similarly sized memory bank. It shows a significant speed up in decoding by using a CML decoder. However, much of this time gained is lost during the translation from CML to CMOS. Overall, however, the BiCMOS memory shows a speed up.

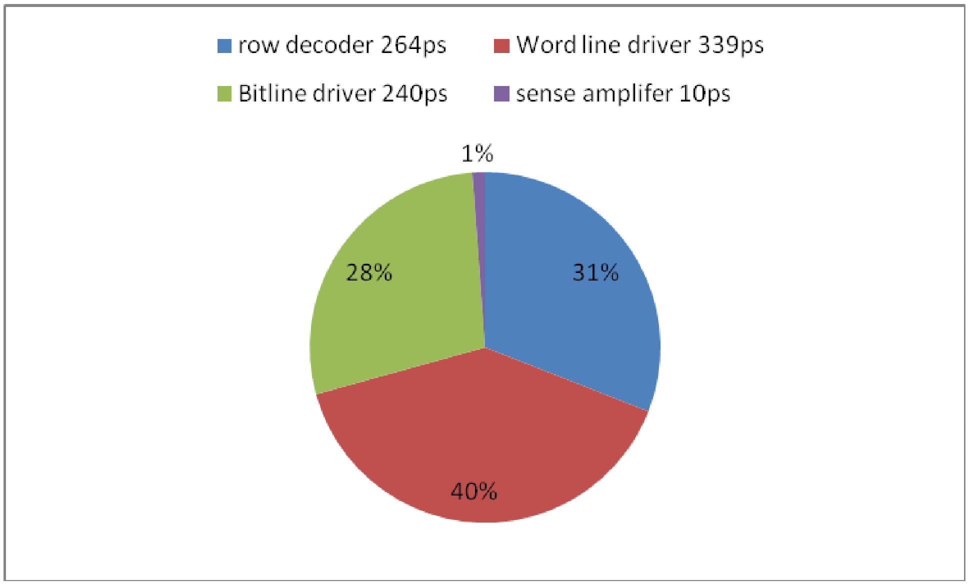

Figure 48: Breakdown of delay during a read operation from CACTI simulation.

4. Conclusion

4.1 Conclusion

An SRAM circuit is designed and tested. A CMOS memory array is used for better power dissipation and noise immunity. The peripherals were done in CML because of its speed.

The three major sources of delay are the driving of the word line and bit line due to the large capacitance that is associated with them and the translation from CML to CMOS. The slow bit line was compensated for by using a sense amplifier. The word line capacitance was compensated for by using a BiCMOS driver however much of the delay is still remains.

4.2 Future Work

Further improvements to the BiCMOS driver as well as the CML to CMOS translator can be done as this is still a large source of delay. It is found that much of the speed up gained by using a CML decoder is lost from the translation of the signal. Research to whether the speed gained from using CML is better as the decoder increases in size can be done.

Improvements to the bit line circuitry warrant further study. Additional research on sense amplifiers or the addition of the bit line driver can be done.

Some peripherals are assumed in this simulation. The voltage sources -1.5V as well as -1V were assumed and are necessary for proper operation. The input address is also assumed to be zero during precharge.

Finally, thermal analysis of the circuit can be performed. As previously stated, the percent of current passing through one path in CML is proportional to the ratio of the differential input voltage and the thermal voltage. Increasing the thermal voltage significantly can affect performance.

5. References:

[1] John D. Crestler, Silicon Heterostructure Handbook, Taylor and Francis Group, FL, 2006

[2] J.D. Cressler, G. Niu, *Silicon* Germanium Heterojunction Bipolar Transistors, Artech House, 2003.

[3] P. Narozny, M. Hamacher, H. Dambkes, H. Kibbel, E. Kasper, "*Si/SiGe Heterojunction Bipolar Transistor with Graded Gap SiGe Base made by Molecular Beam Epitaxy*", Electron Devices Meeting, pp. 562-565, 1988.

[4] C. Peiyi, "*Development of SiGe Materials and Devices*", Solid State and Integrated Circuit Technology, Vol 1 pp. 570-574, 2001.

[5] D. Houghton, "*SiGe CVD, Fundamentals and Device Applications*", Aixtron Inc, July 2004

[6] P.S. Chen, Y.T. Tseng, M.-J. Tsai, C.W. Liu, "*High Throughput UHV/CVD SiGe and SiGe:C Process for SiGe HBT and Strained Si FET*", Semiconductor Manufacturing Technology Workshop, pp. 145-148, Dec 2002

[7] S.-M. Lee, B.R. Ryum, T.-H. Han, D.-H. Cho, B. Kim, K.E. Pyun, "*Atmospheric Pressure CVD Grown SiGe Epitaxial Base Heterojunction Bipolar Transistor Using a TiSi$_2$ Base Electrode*", Journal of the Korean Physical Society, Vol 30, No 2, pp. 315-319, April 1997

[8] A.J. Joseph, J.S. Dunn, *"Industry Examples at the State of the Art: IBM",* Silicon Heterostructure Handbook, Taylor and Francis Group, FL, 2006

[9] D.L. Harame, B.S. Meyerson, *"The Early History of IBM's SiGe Mixed Signal Technology"*, IEEE Transactions on Electron Devices, Vol 48, No 11, pp 2555-2567, November 2001

[10] S. Suhag, *"High Speed BiCMOS Memory for Cache Applications"*, MS Thesis, Rensselaer Polytechnic Institute, Troy, New York, August 2007

[11] N. LiCausi, *"A Survey of SRAM Performance Enhancements"*, MS Thesis, Rensselaer Polytechnic Institute, Troy, New York, July 2007

[12] O. Erdogan, "*A Three-Port Pipelined Register File Implemented Using SiGe HBT BiCMOS Technology"*, Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, New York, December 2006

[13] J.M. Rabaey, Digital Integrated Circuits: A Design Perspective, Prentice Hall, New Jersey, 1996

[14] M. Alioto, G. Palumbo, Model and Design of Bipolar and MOS Current-Mode Logic, Springer, 2005

[15] H. Nambu, K. Kanetani, K. Yamasaki, K. Higeta, M. Usami, M. Nishiyama, K. Ohhata, F. Arakawa, T. Kusunoki, K. Yamaguchi,A. Hotta, N. Homma. *"A 550-ps Access 900MHz 1-Mb ECL-CMOS SRAM"*, IEEE Journal of Solid-State Circuits, Vol 35, No 8, pp 1159-1168, August 2000

[16] T. Soon-Hwei, L. Poh-Yee, M.S. Sulaiman*, "A Low-Power High Speed 1Mb CMOS SRAM",* Proceedings of the Third IEEE International Workshop on Electronic Design, Test and Applications, Delta 2006, January 2006

[17] N.H.E. Weste, D. Harris, CMOS VLSI Design: A Cicuits and Systems Perspective, Pearson Education, MA, 2005

[18] M. Ishida, T. Kawakami, A. Tsuji, N. Kawamoto, M. Motoyoshi, N.Ouchi, *"A Novel 6T-SRAM Cell Technology Designed with Rectangular Patterns Scalable beyond 0.18μm Generation and Desirable for Ultra High Speed Operation"*, Electron Devices Meeting, 1998, IEDM '98 Technical Digest., International, pp 201-204, December 1998