UNIVERSITY of HOUSTON

PROJECT REPORT

"DESIGN OF A 16 BIT SRAM"

By

GOKUL BALAGOPAL - 1590661

Under the guidance of

Dr. Wanda Wosik
Associate Professor
Department of Electrical and Computer Engineering
University of Houston

Table of Contents

1. Introduction

Static Random Access Memory (SRAM) is a type of semiconductor memory that uses bi-stable latching circuitry (flip-flop) to store each bit. SRAM has been widely used as the representation of memory for its fast operation and lesser power consumption. This trend of Static Random Access Memory (SRAM) along with CMOS technology scaling in different processors and system-on-chip (SoC) products has fuelled the need of innovation in the area of SRAM design.

Now a day, low power and high speed SRAMs have turned into a discriminating part of numerous VLSI chips. This is particularly valid for microchips, where the on-chip reserve sizes are developing with every era to extend the expanding disparity in the speeds of the processor and the Cache memory [1]. SRAM is utilized as Cache memory which is quick and used to accelerate the assignment of processor and memory interface. With the new technologies in VLSI innovation, the speed of the logic gates has expanded essence, but comparing the memory speed is not enhanced [2]. Thus, for high speed PCs SRAM memories are critical to enhance operating speed and DRAM are utilized as a part of Main Memory where Density has more significance than Speed. In this work we have focused on the outline of high performance 6T SRAM which can be utilized as high speed and low power memory for PCs.

2. SRAM & Its Peripherals

2.1 Introduction to SRAM cell

Depending upon the utilization of a clock, SRAM can be partitioned as synchronous SRAM and asynchronous SRAM. In synchronous SRAM, all the inward flags and timing will be controlled by the clock edge [1]. Information in, control motions and location identifies with the clock signal, it is for the most part utilized as a cache memory while Asynchronous SRAM is autonomous of clock recurrence. All the inner signals and timings are introduced by the location move. The extent of offbeat SRAM fluctuates from 4 KB to 64 MB. Because of the quick access time of Asynchronous SRAM, it is suitable as main memory for cache less embedded processors which are utilized as a part of modern hardware, estimation of frameworks, organizing hardware. [3]

The operation of SRAM can be partitioned into three states:

a. Standby mode - word line is not initiated, so the address and data lines are kept withdrawn from SRAM memory cells, subsequently cells keep the information as it is and no read and write operation is there. Power consumption in this mode is reduced the most.
b. Reading data from cells - Assume we are reading data 0 which has already stored in the memory cell. The Read cycle begins with pre-charging the bit line and bit line bar, after pre-charge operation word line gets actuated as by particular row address and one of the bit line starts discharging through the cell. Here logic 0 is stored in the cell initially, hence Bit line voltage starts discharging through the ground and simultaneously bit line bar voltage starts charging to VDD. Then sense amplifier senses the difference between the voltages on two bit lines and gives proper output, i.e. reads 0 or read 1. If the bit line voltage is greater than the bit line, bar voltage than the output of the sense amplifier as logic 1 which indicates read 1 operation. Similarly if bit line voltage is less than bit line bar voltage then output of sense amplifier indicate read 0 Operation [3]. Presence of sense amplifier increases the speed of operation of memory as it senses the small difference between voltages on bit lines otherwise it takes lot of time to perform any read operation.
c. Write operation begins with applying data need to be written on bit lines. Suppose we need to write logic 0, then the bit line will get discharged to 0 and bit line bar voltage is charging to 1. At that point the word line will get actuated, and proper information gets to keep in to the cell.

2.2 Peripherals of SRAM circuit

Peripherals include row decoders, pre-charge circuit, column multiplexers/Decoders, Isolator, Sense amplifiers and write drivers [4].

2.2.1 Row Decoder

At whatever point memory takes into consideration arbitrary address based access address decoders must be available. The Boolean function of the decoder is comparable to n-input AND logic gates, where the extensive fan-in AND operation is actualized in a various levelled structure. The configuration of these decoders has a noteworthy impact on the speed and power dissipation of the memory. Two classes of decoders that is row decoder, whose task is to empower one memory row out of 2^M, where M is the width of particular fields in address word. While considering these decoders, it is imperative to keep the complete memory floor plan in context. When the number of inputs is more than or equal to four then the speed of operation of the decoder is effected so pre-decoders are to be used which reduces the large fan-in such that the speed of the decoder is improved. The principal level is the pre-decoder where two groups of address inputs and their complements are first decoded to initiate one of the pre-decoder yield wires separately to obtain the

partially decoded outputs. The pre-decoder yields are consolidated at the following level to enable the word line. The decoder delay comprises of word line wire delay, interconnect delay of pre-decoder and gate delays in the critical path. As the wire RC delay develops as the square of the wire length, the wire delays inside the decoder structure, particularly of the word line, gets to be critical in extensive SRAMs. From delay analysis, it was observed that the NOR based decoder is quicker than the NAND based decoder.

2.2.2 Column Decoder

This circuit is used to select particular column in the memory array. The typical column decoder/Mux in which the outputs of 2 to 4 decoder are used to enable pass transistors. Depending on the output of decoder only of the bit line or bit line bar is selected and the above circuit acts as a 4 to 1 multiplexer. COL_EN signal is used to enable the column decoder/Mux circuitry. As we discussed earlier the frequency of operation of memory is strongly affected by number of columns as well as number of rows. Hence to maintain good frequency of operation and an aspect ratio of 4x4, we are using DWL (Divide Word line Architecture). Consider the design of 16 cell array, there is totally 16 numbers of SRAM cells present in a cell row and these 16 cells are divided into 4 portions such that the output of SRAM memory is having a size of 8 bits. So 2 to 4 multiplexer is used to select one of the portion out of 4 portions.

2.2.3 Pre Charge

This circuit is used to pre-charging the both bit lines voltages to supply voltage and pre-charging operation should perform before every write and read operation. The pre-charge circuit which consists of pull up PMOS transistors and an equalizer is used to equalize the voltage on both bit lines. The pull up PMOS transistors are controlled by PR signal i.e. The Transistor M3 shown in the schematic of Pre-charge is an equalizer which is used to equalize the voltage on both bit lines. Pre-charge circuit should provide large driving current to drive the bit lines which are having large parasitic capacitances, so the transistor sizes of pre-charge circuit need to be increased.

2.2.4 Sense Amplifier

Sense Amplifiers plays a crucial role in the design of memories to achieve performance, reliability and functionality of memory circuits. Normally sense amplifiers perform various operations like voltage amplification, reduction in delay, power reduction and restoration of original signal. Generally sense amplifiers are used in the memories to speed up the read operation. Sense amplifier takes the small signal difference bit line voltage as input and gives full swing single ended output [6]. Access time and power consumption of memory is affected by the sense amplifier hence the performance of memory is improved by reducing both sensing delay and power dissipation.

When SE is logic low then both bit line voltages are charges to supply voltage, when SE is logic high then sense amplifier is getting ON and one of the bit line voltage discharges to ground via pull down transistor,. It takes BL and BL_bar voltages as an input and generates single ended output. When BL voltage is greater than the BL_bar voltage then current through is increases and simultaneously current through decreases to maintain as a constant, then the drop across is decreases hence output voltage increases, which interprets output as the logic 1[1]. Similarly when BL voltage is less than BL_bar voltage then it indicates output as the logic 0. In this way, Sense amplifier plays a crucial role in the memory read operation.

2.2.5 Write Driver

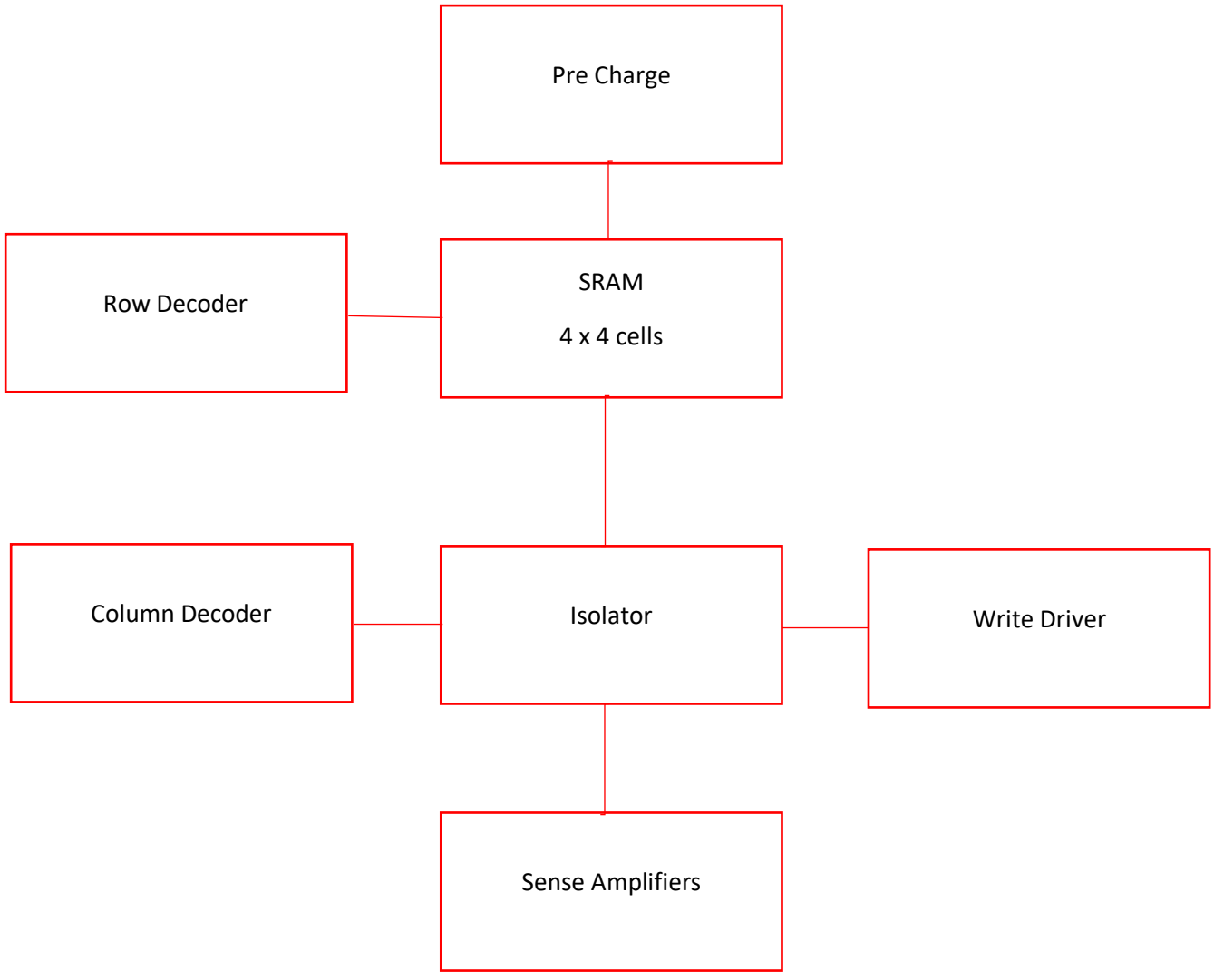
The tremendous bit line swing can bring about huge power dissipation in write operation and during read operation, the bit line voltage swing is normally limited to 180mV, and consequently the write

cycle can consume around 1/8th more power than a read operation. Initially, before write operation both bit line voltages are charging to supply voltage and the write operation is performed by enabling WR_EN signal. Suppose if we want to write logic 0 in to the memory cell, then the BB line voltage charges to supply voltage VDD and BT line voltage is discharges to lower potential i.e. ground. The data stored in bit line, BT and bit line 45 bar, BB is accessed by enabling word line. The sizing of transistors in write driver is quiet large to provide large driving current[3]

$$(W/L)write\_driver = 6(W/L)isolator \text{ which gives approx. } (9u/1.5u)$$

2.2.6 Isolator

As word lines have large parasitic capacitance the output of decoder cannot drive the last cell in a memory row. So there must be a buffer exists between decoder and monolithic memory array to drive the last cell in a row. If the word line driver is capable to drive the worst case, i.e. last cell, then we can access all cells in a row. Figure 2.4 shows the schematic of the typical word line driver. It is a circuit which is nothing but the cascading connection of the AND gate with an even number of inverters [13]. To drive the word line which is having a large parasitic capacitance we need to design a stack of inverters with increase in size such that it should capable to drive the worst case cell.

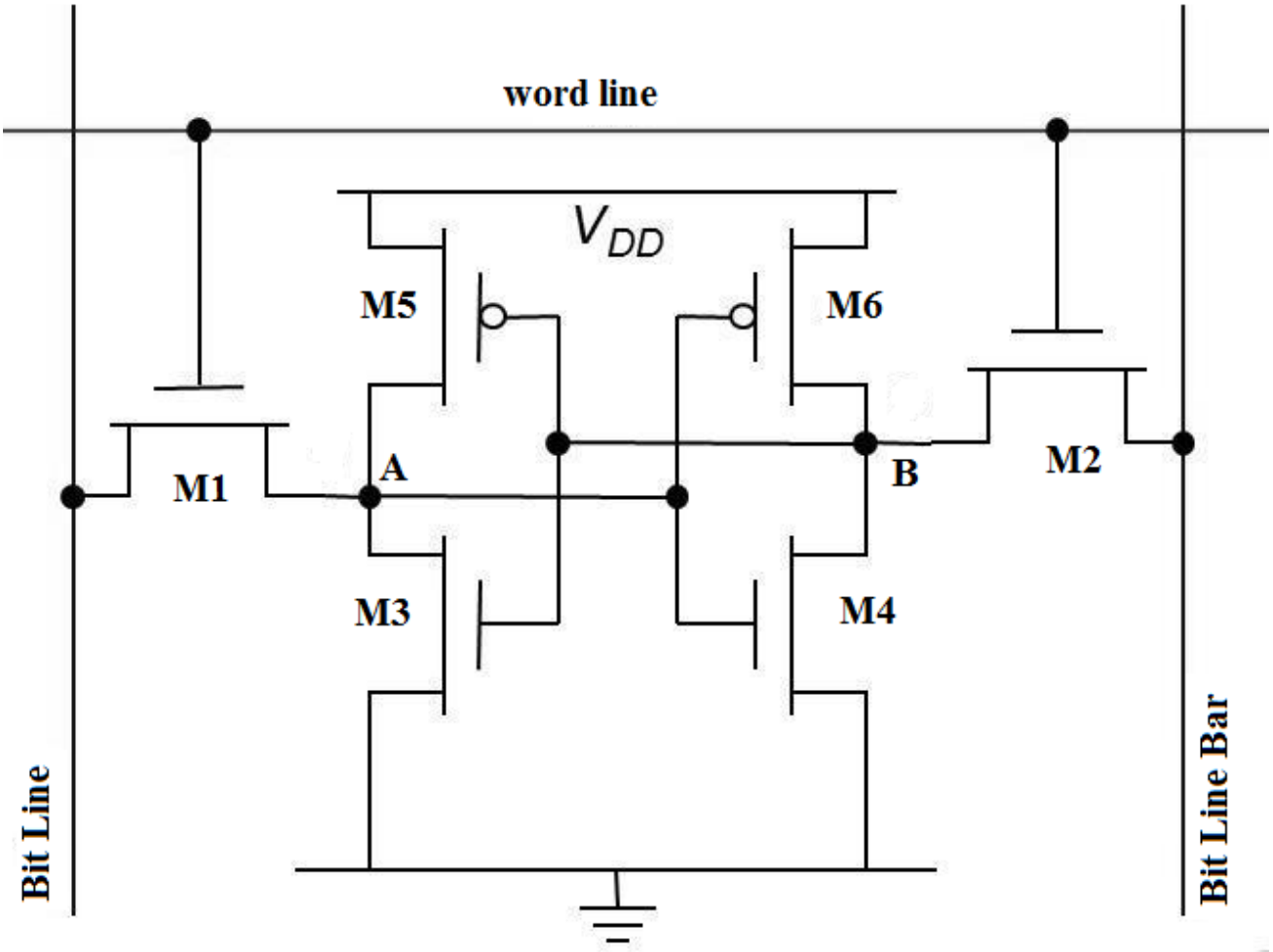


3. Design of SRAM & its Peripherals

3.1 6T SRAM Memory Cell

3.1.1 Design

Depending on the type of load used at the inverter of Flip-Flop, the SRAM Cells are classified into three categories such as 4-Transistor cells, 6-Transistor cell and Thin-Film-Transistor (TFT) cell. We chose to implement using 6-Transistor cells because of its advantages with respect to other cells. This topology have good noise immunity because of its large Noise margin and low static power dissipation because of less leakage current in cell.



To focus the sizes of the transistors in 6-T cell, various outline criteria must be taken into account such that it should not destroy the information present in the cell during read operation and similarly it should allow modification of data in write operation. So design of proper (W/L) of transistors is must to satisfy above two requirements. The stability and speed of SRAM cell can be increased by increasing the supply voltage but increase in supply voltage leads to increase in the power dissipation of circuit as power dissipation varies with square of supply voltage. So always there exists a trade-off between power and speed. So we have to design a SRAM cell such that it consumes less power with better speed.

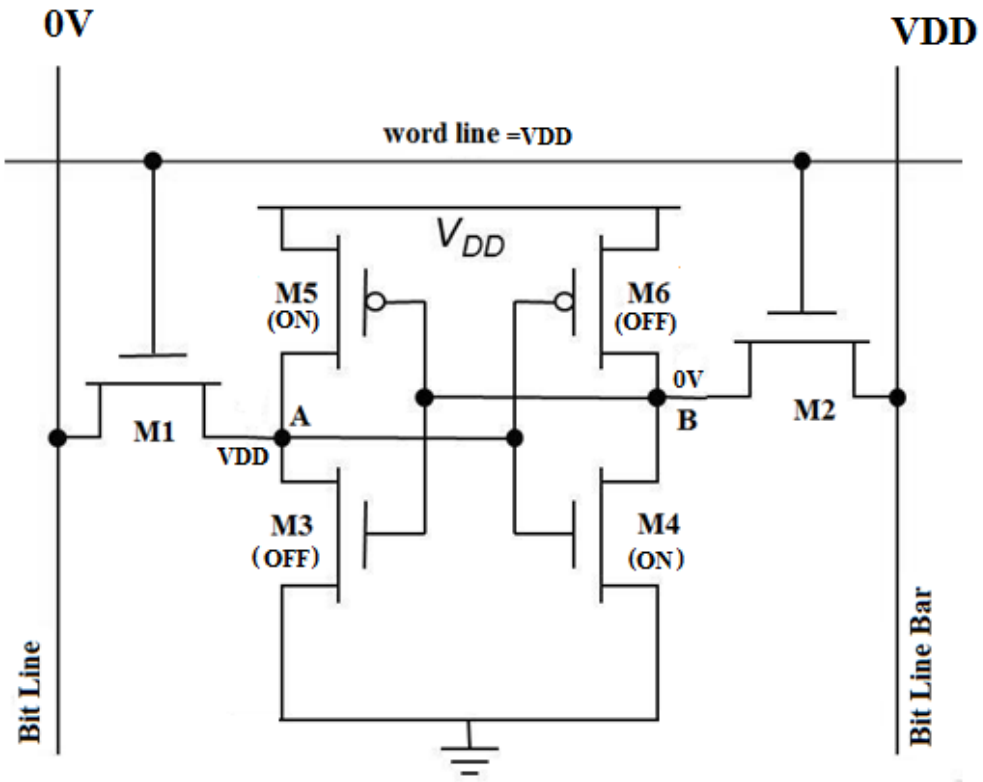
The operation of SRAM memory is classified into three modes i.e. standby mode, write mode and read mode:

Standby mode:
In this mode word line is not activated, so both the access transistors M1 and M2 are disconnected from memory cell, hence in this mode memory cell retains its previous data as long as power supply is provided. Here the column capacitances are charges to supply voltage, through M5 and M6. In this mode memory cell consumes less power[3].

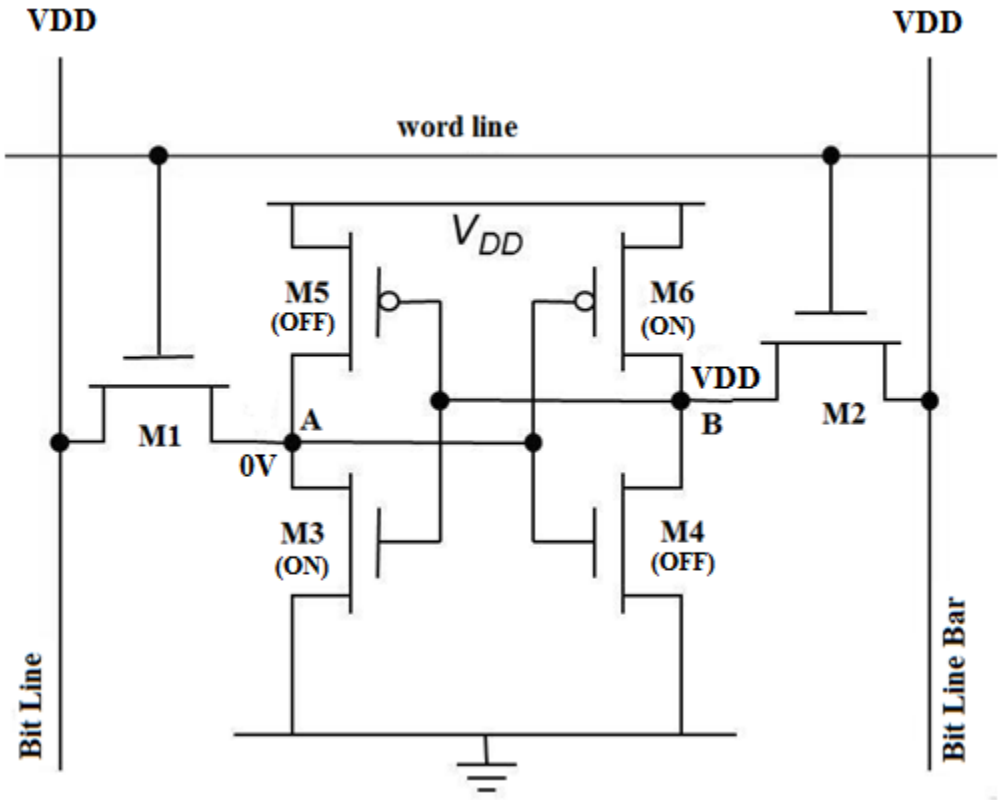
Write Mode:
Suppose if we want to perform write zero operation by assuming initial stored data in memory cell as logic one. So in the beginning of write operation the voltages present at two nodes A and B are VDD and 0V respectively hence initially the two transistors M3 and M6 are operated in cut off mode whereas the transistors M4 and M5 are operated in linear region. Now by using write driver circuit the column voltage of bit line is forced to logic zero. Now the pass transistors M1 and M2 are activated by using word line whose address is given by row decoder. In order to achieve basic requirements which we discussed earlier, we should have

$$\frac{\left(\frac{W}{L}\right)_5}{\left(\frac{W}{L}\right)_1} < \frac{\mu_n}{\mu_p} \frac{2(V_{DD}-1.5V_{T,n})V_{T,n}}{(V_{DD}+V_{T,p})^2} \text{ ---------- (A)}$$

The following Figure depicts the operation of write zero.



Read Mode:
Before the Read operation the two bit line voltages charge to supply voltage VDD. During



read mode, suppose if we want to perform read 0 operations, then assume initial data stored in the memory cell is logic 0.The figure above depicts the read zero operation of 6T SRAM Cell and the initial voltages present at the storing nodes A and B is 0V and VDD respectively. Hence transistors M4 & M5 are operated in cut-off region whereas the transistors M3 & M6 are operated in linear region. The pass transistors M1 & M2 are activated by using word line which is controlled by row decoder circuit. Now there is no current flow in transistor M2 why because the voltages at node B and bit line bar voltages are almost equal which is equal to VDD. On the other hand, the transistors M1 & M3 will exhibit a nonzero current and the bit line voltage is discharges to ground from VDD. Now both the bit line voltages are given to the inputs of sense amplifier [6] and it will produce logic 0 as output why because the bit line voltage is less than the bit line bar voltage. The necessary condition to be satisfied in the read operation is given as follows,

$$\frac{\left(\frac{W}{L}\right)_1}{\left(\frac{W}{L}\right)_3} < \frac{2(V_{DD}-1.5V_{T,n})V_{T,n}}{(V_{DD}-2V_{T,n})^2} \quad ----------(B)$$

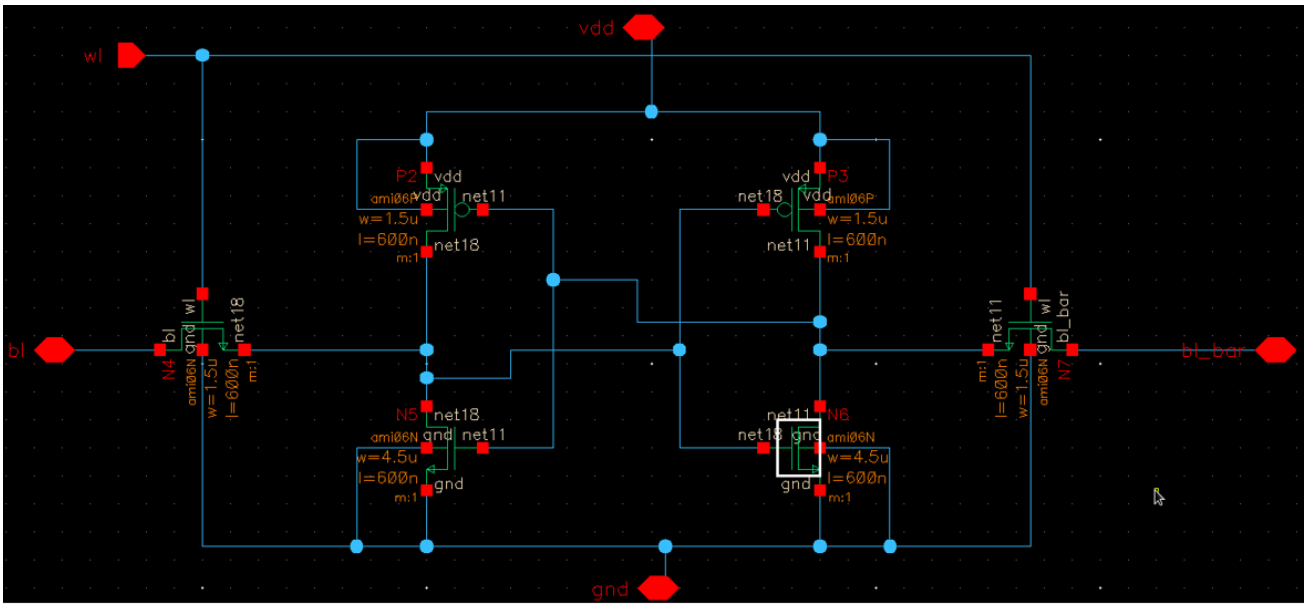
By using above two conditions (A) and (B) mathematically, we can design the 6T-SRAM cell. Otherwise the below condition needs to be satisfied which is obtained from practical observation in order to perform correct operation i.e. either read or write operation.

$$\left(\frac{W}{L}\right)_{pull-up} < \left(\frac{W}{L}\right)_{access} \ll \left(\frac{W}{L}\right)_{pull-down} \quad ----------(C)$$

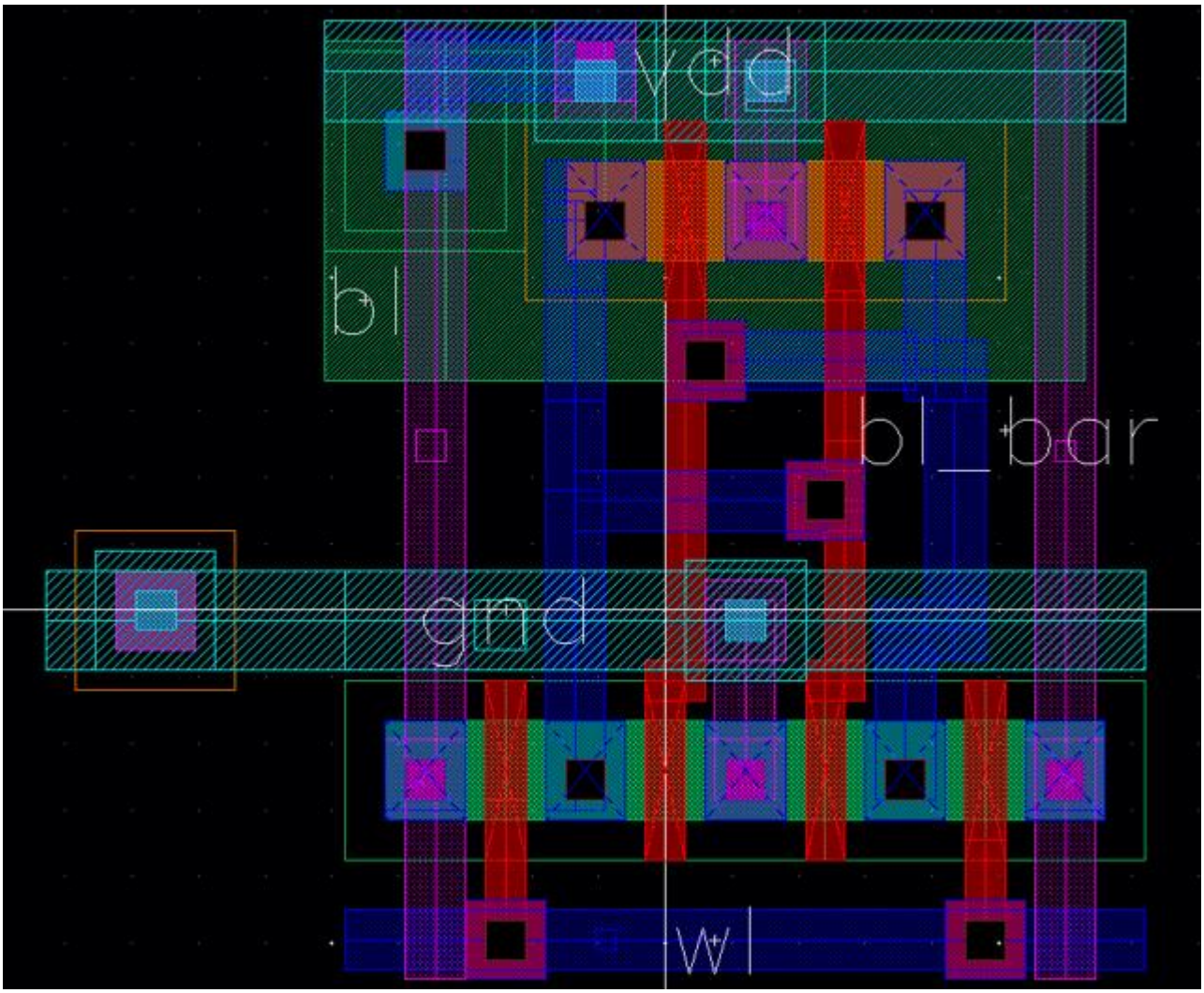
So finally the following transistor sizes will satisfy all the above conditions (A), (B) and (C) and perform correct read and write operations i.e.

$$\left(\frac{W}{L}\right)_{pull-up} = (1.5u/1.5u), \quad \left(\frac{W}{L}\right)_{access} = (3u/1.5u) \quad \& \quad \left(\frac{W}{L}\right)_{pull-down} = (4.5u / 1.5u)$$

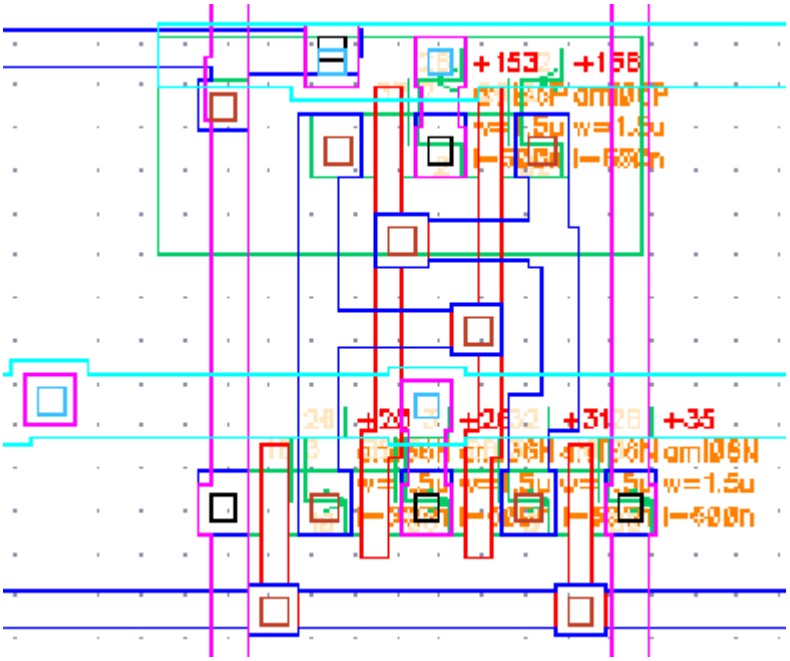
3.1.2 Schematic of 6T Cell



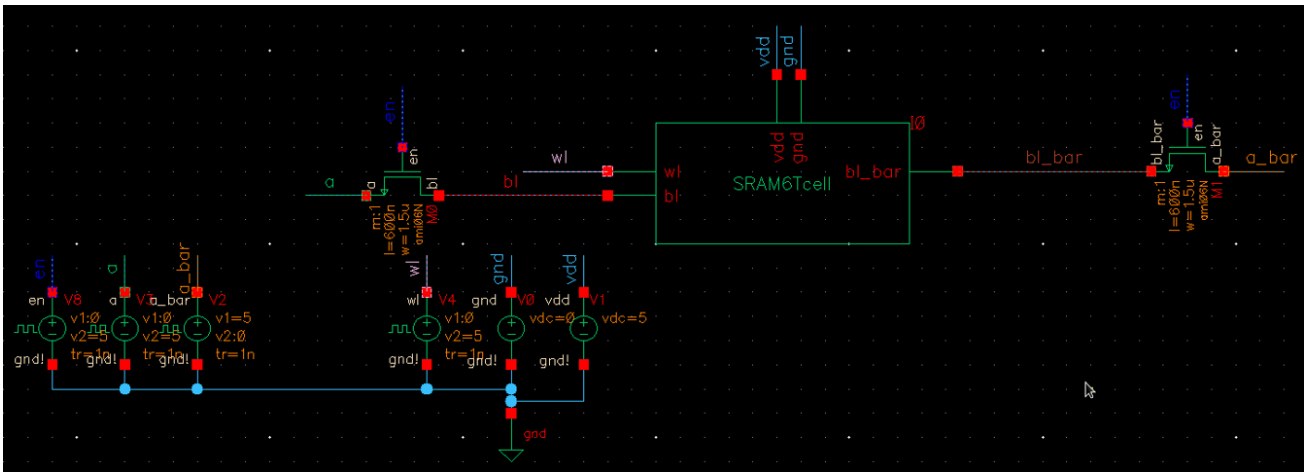
3.1.3 Layout of 6T Cell



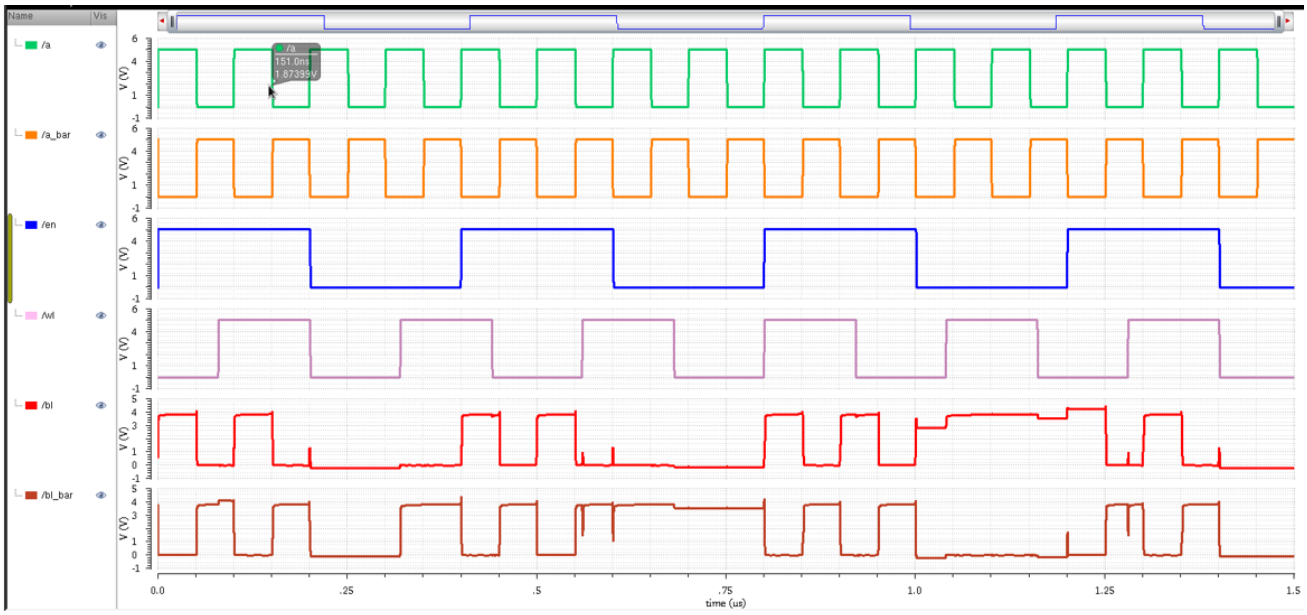
3.1.4 Extracted View



3.1.5 Test bench for 6T Cell

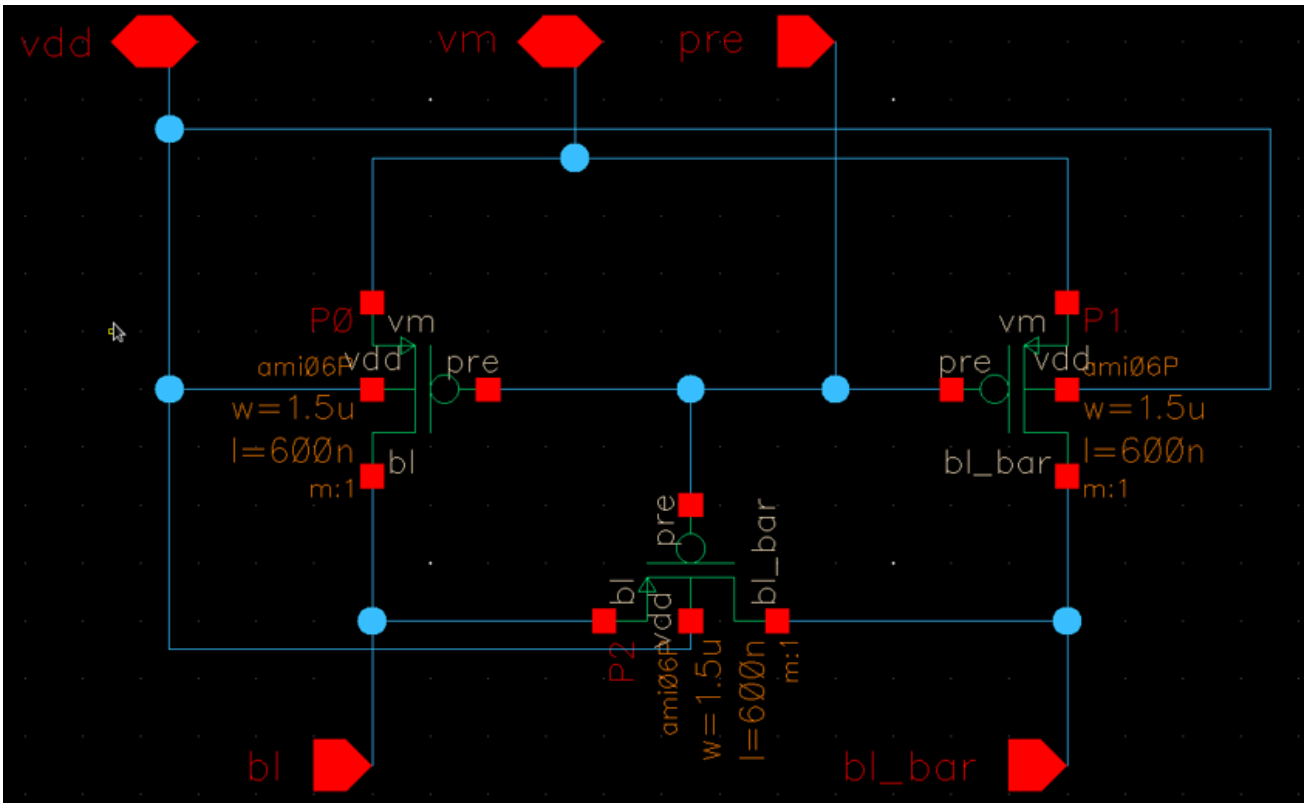


3.1.6 Simulation of 6T Cell

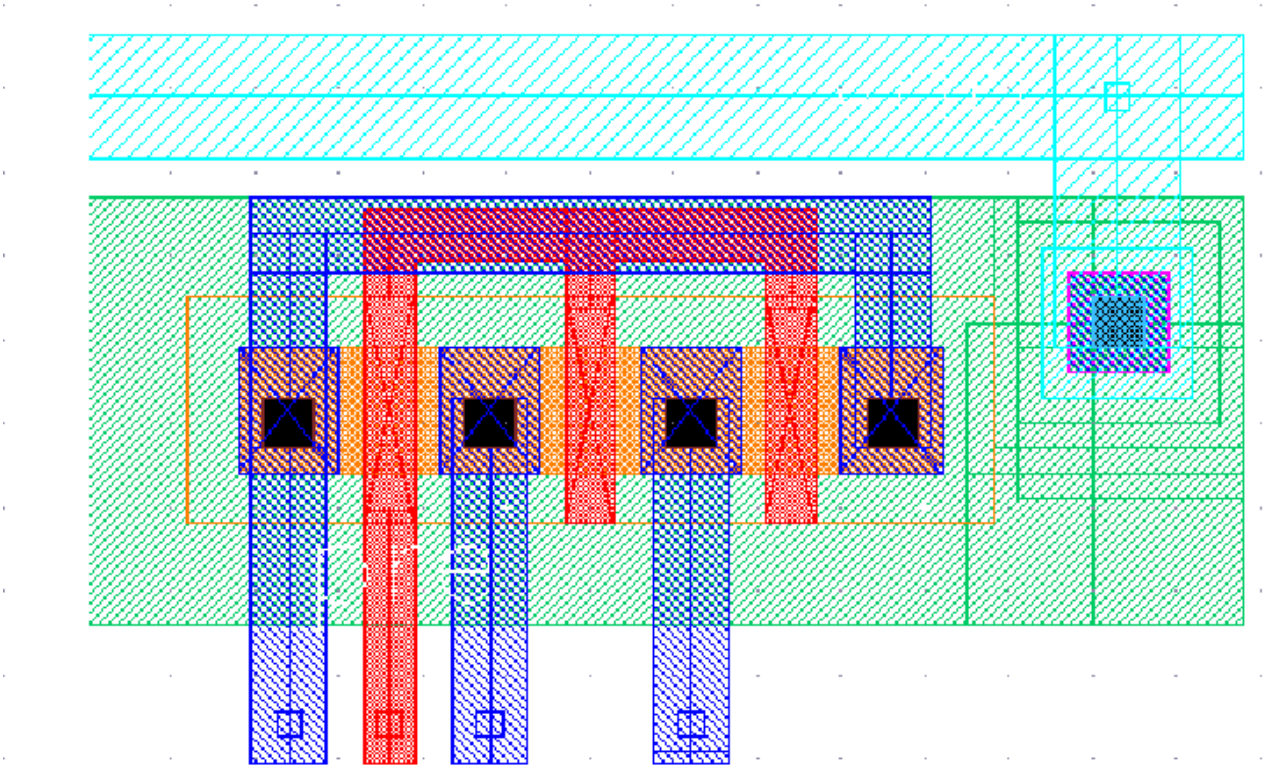


3.2 Pre Charge

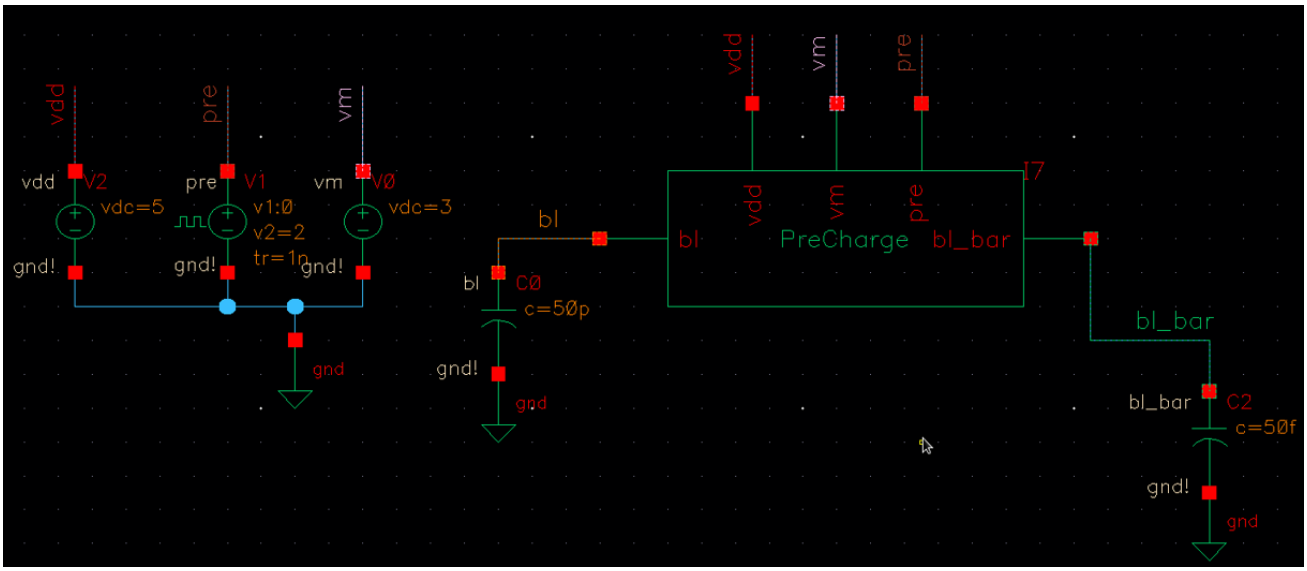
3.2.1 Schematic



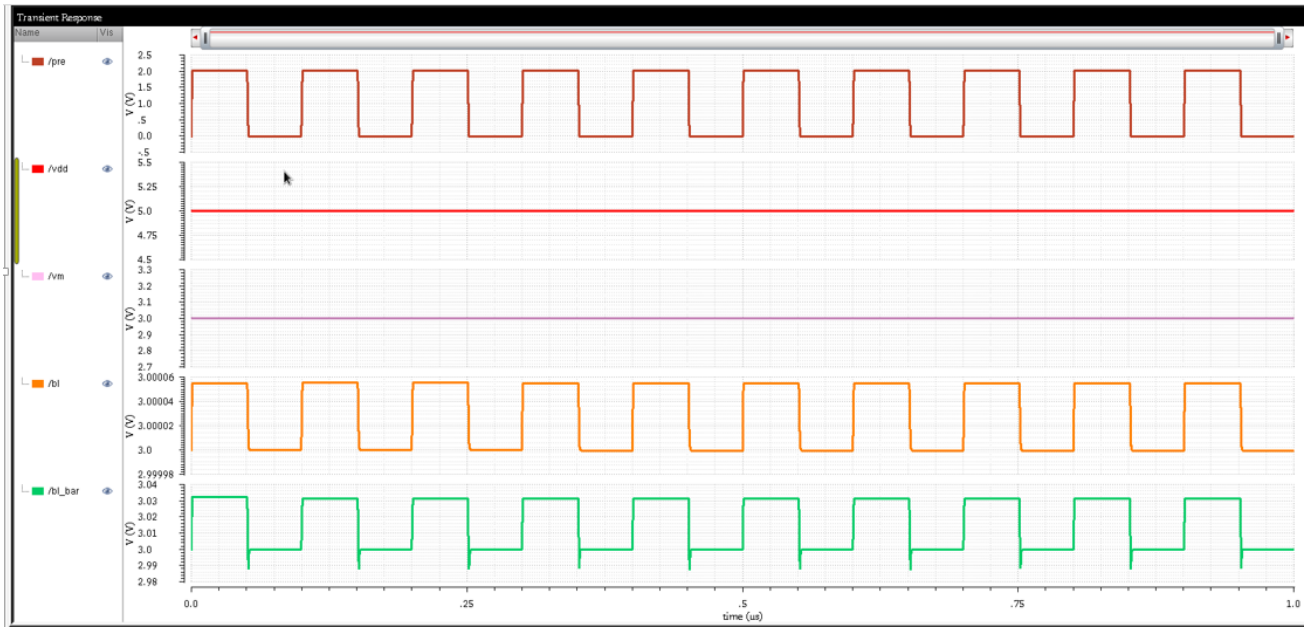
3.2.2 Layout of Pre Charge



3.2.3 Test Bench for Pre Charge

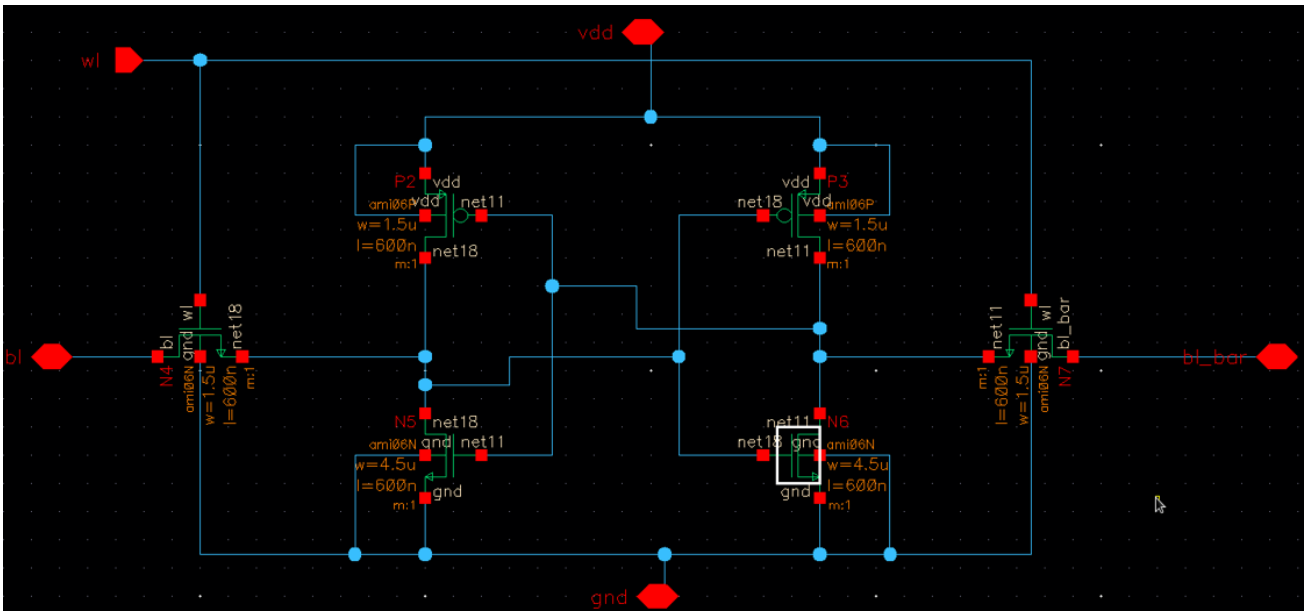


3.2.4 Simulation of Pre Charge

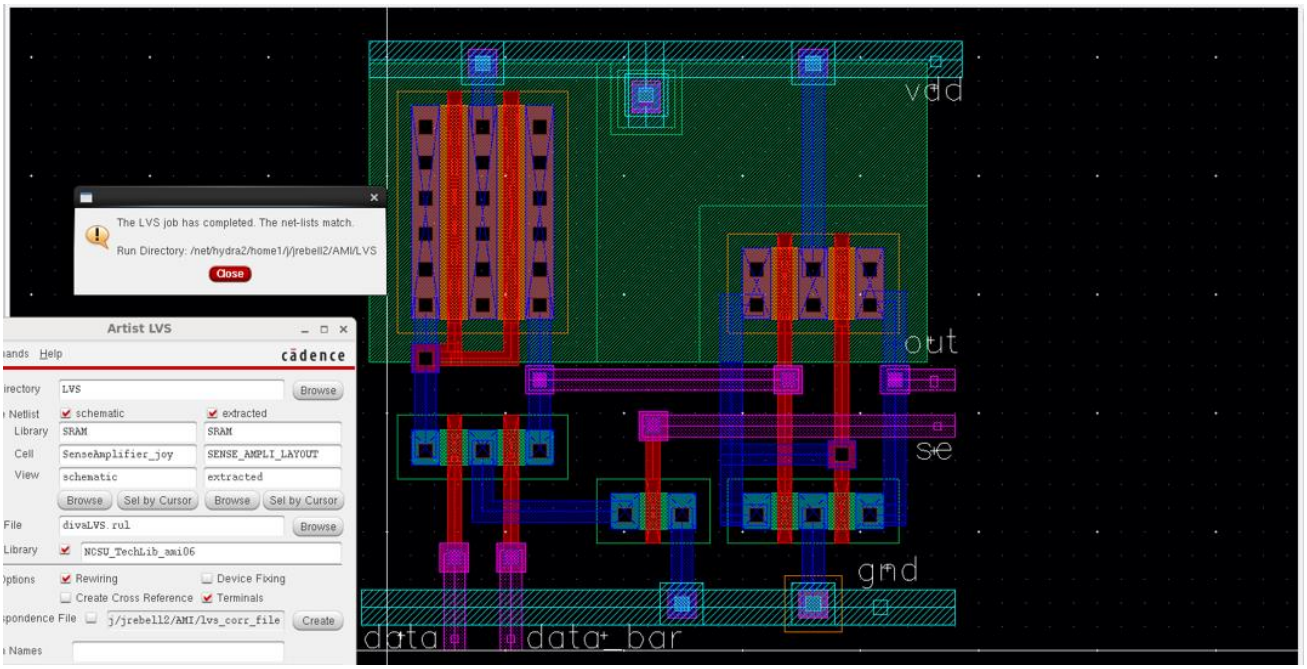


3.3 Sense Amplifier

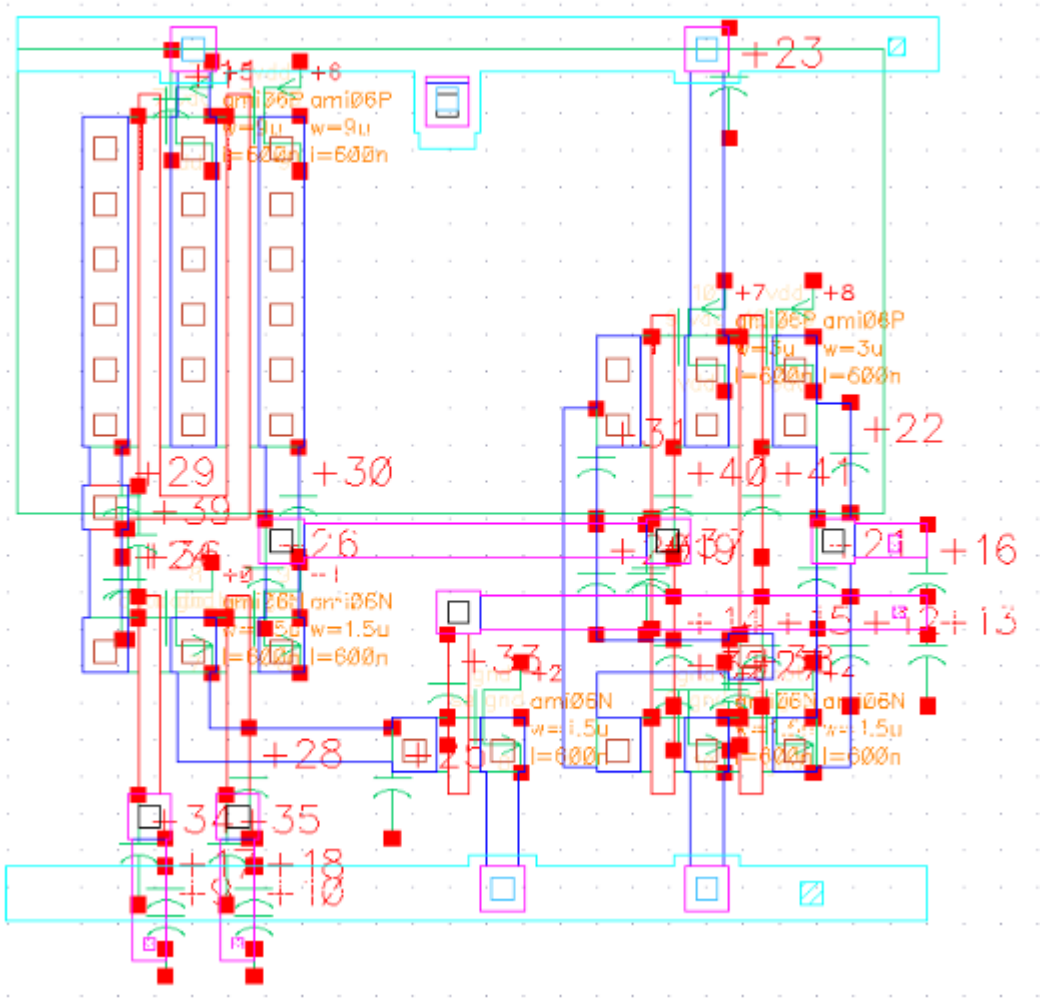
3.3.1 Schematic



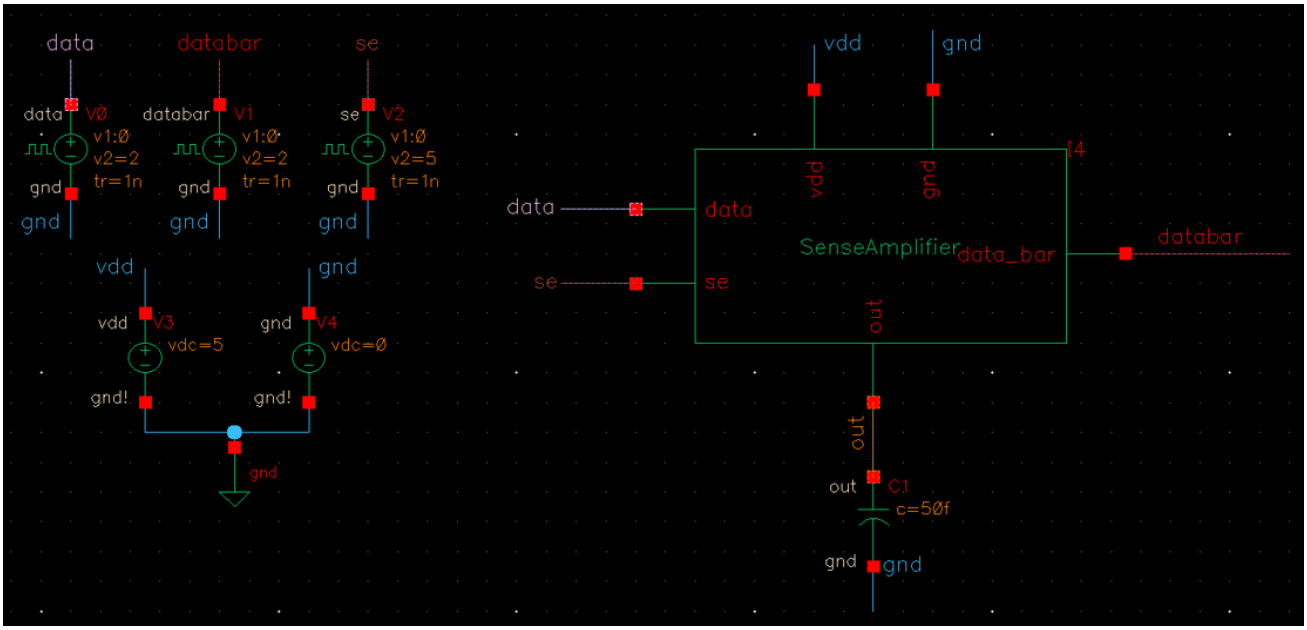
3.3.2 Layout of Sense Amplifier



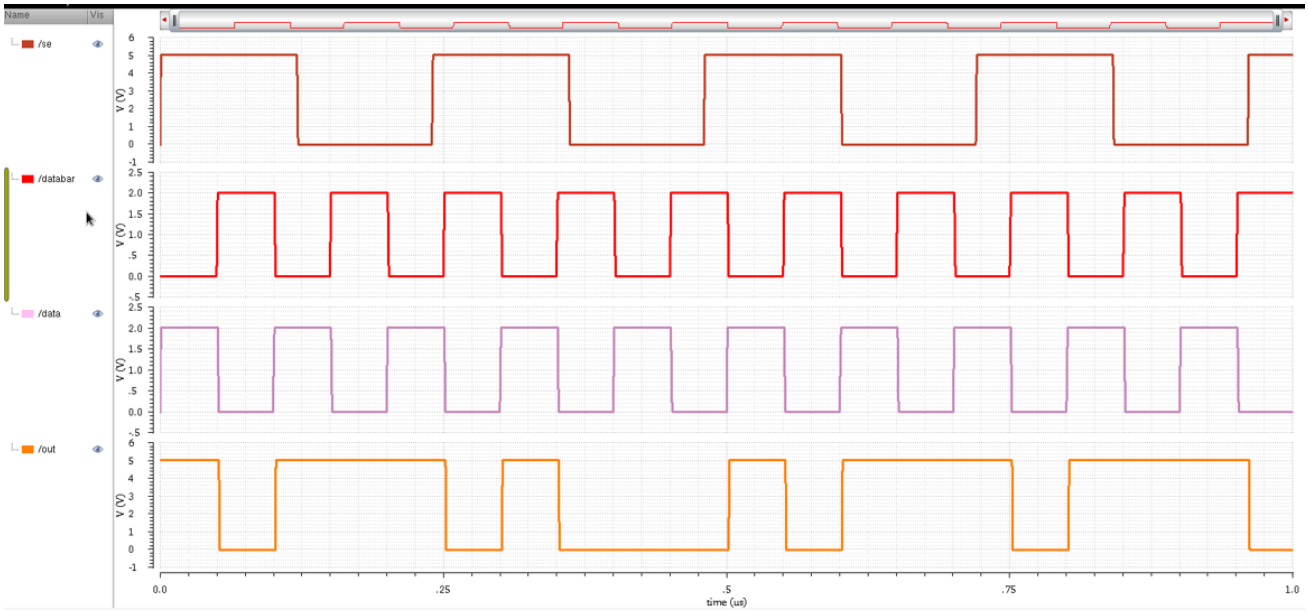
3.3.3 Extracted View



3.3.4 Test Bench for Sense Amplifier

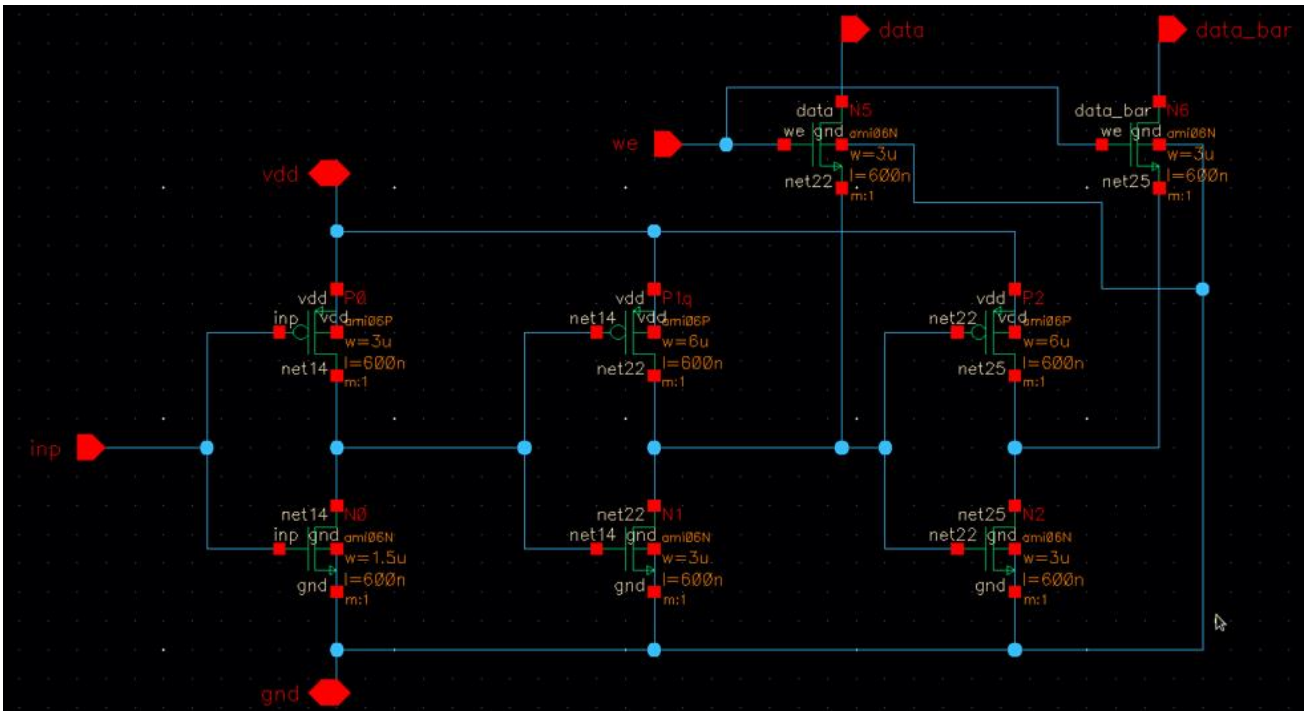


3.3.5 Simulation of Sense Amplifier

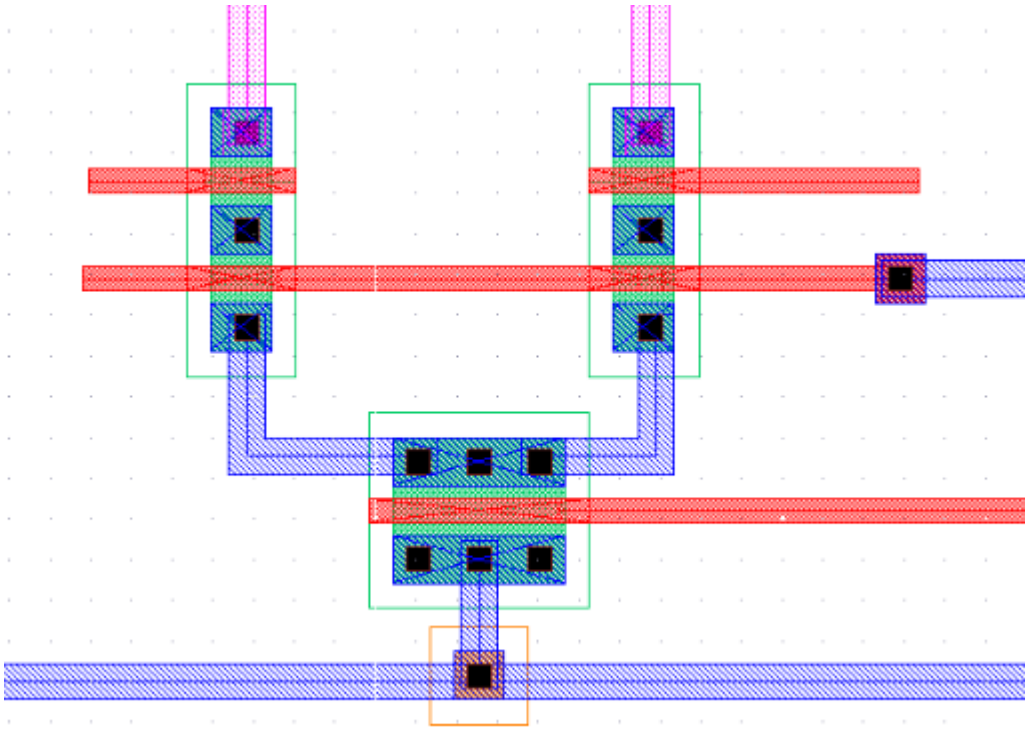


3.4 Write Driver

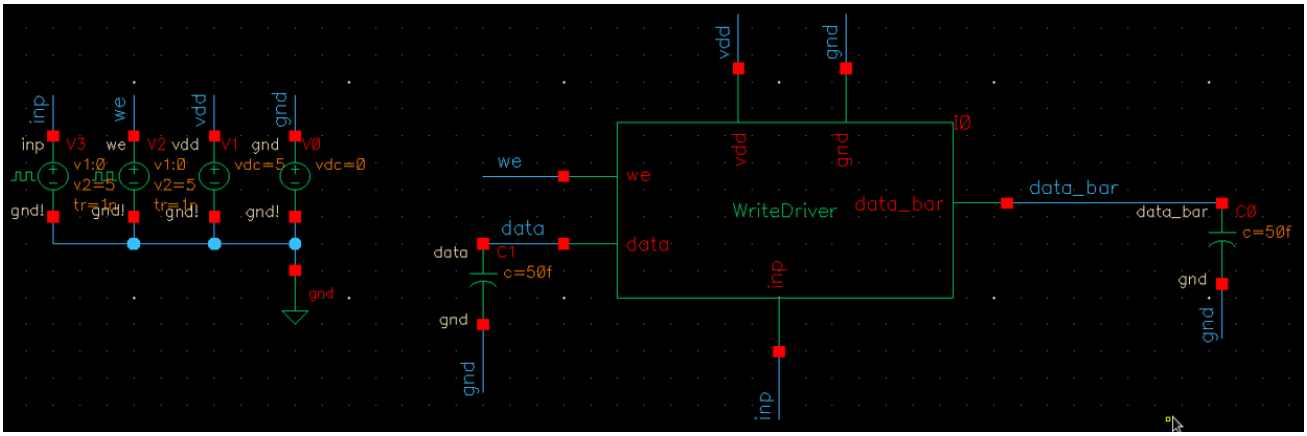
3.4.1 Schematic



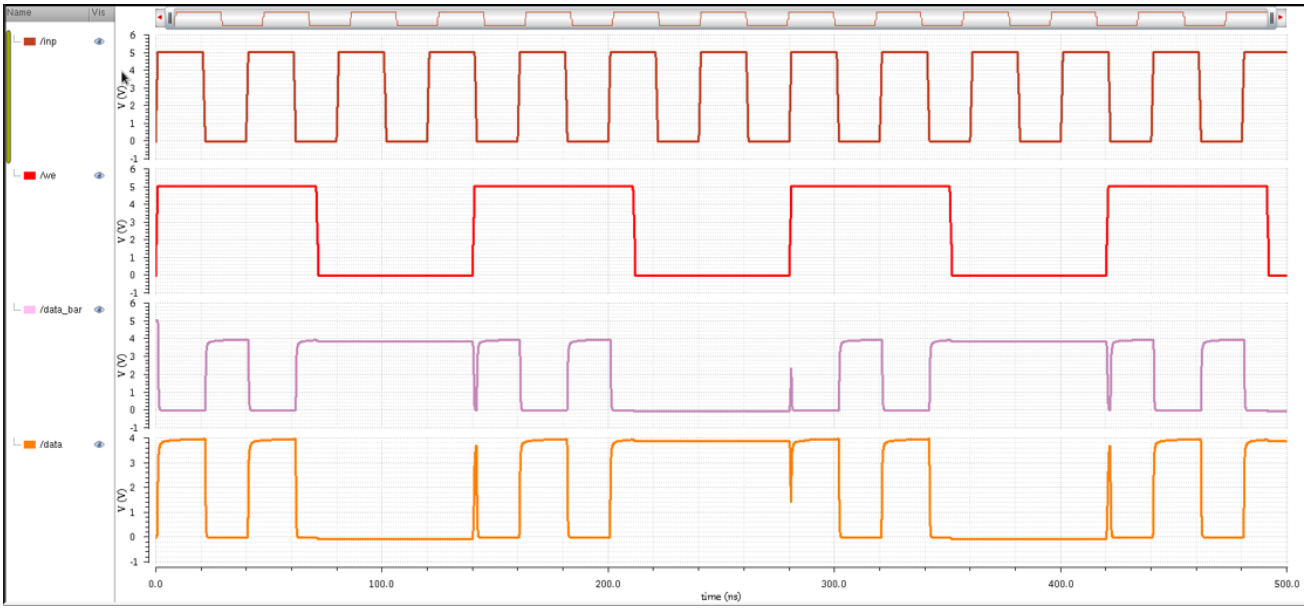
3.4.2 Layout of Write Driver



3.4.3 Test Bench for Write Driver

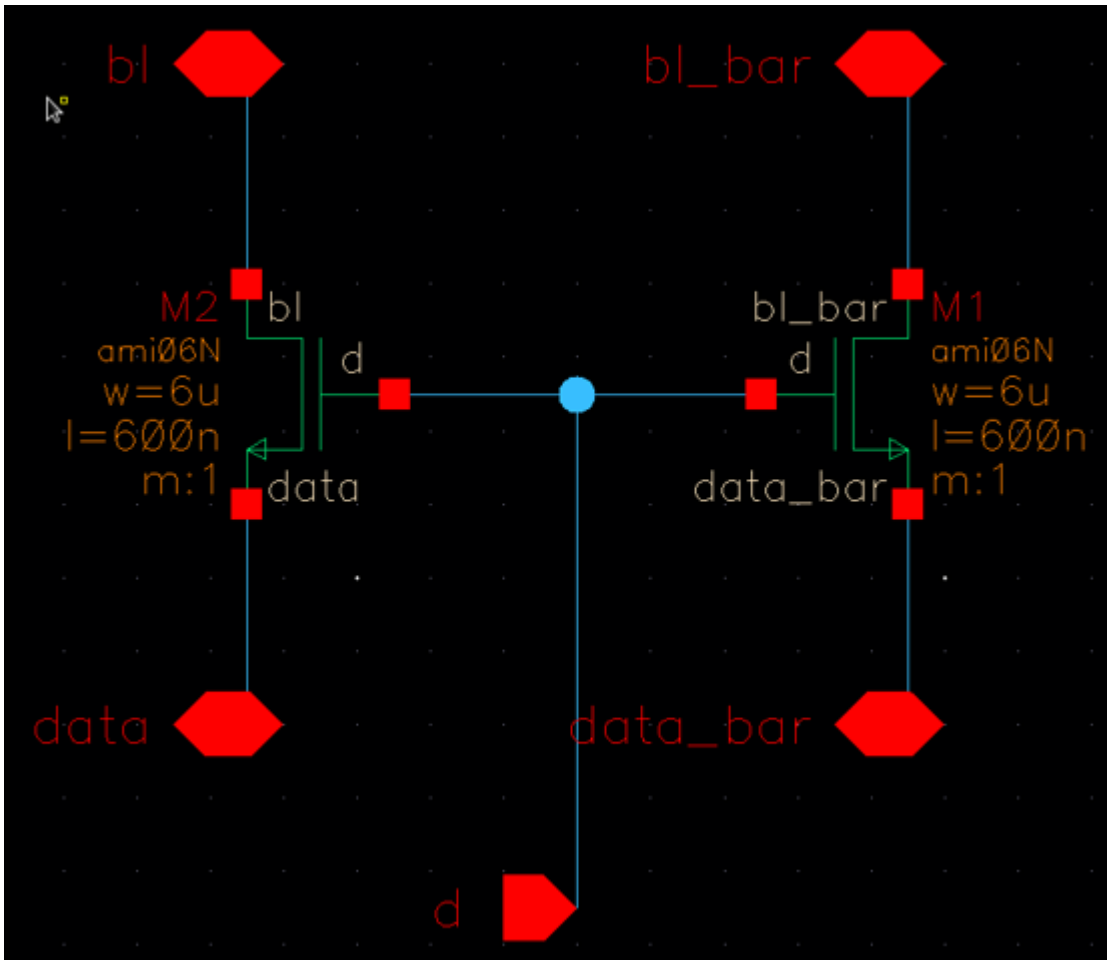


3.4.4 Simulation of Write Driver

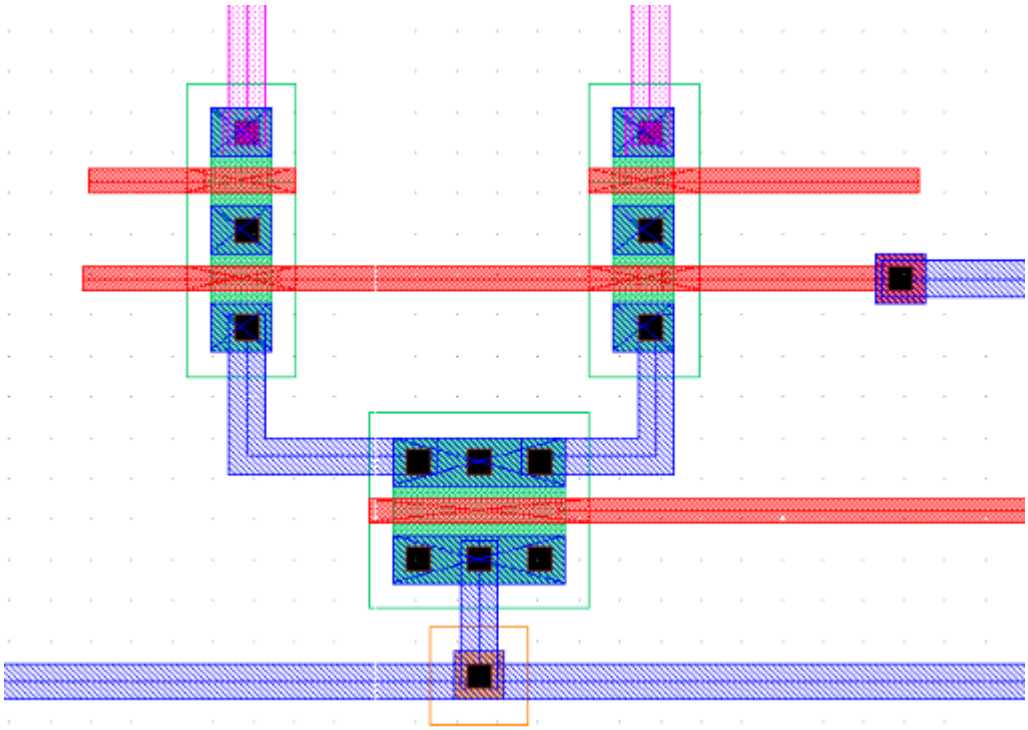


3.5 Isolator

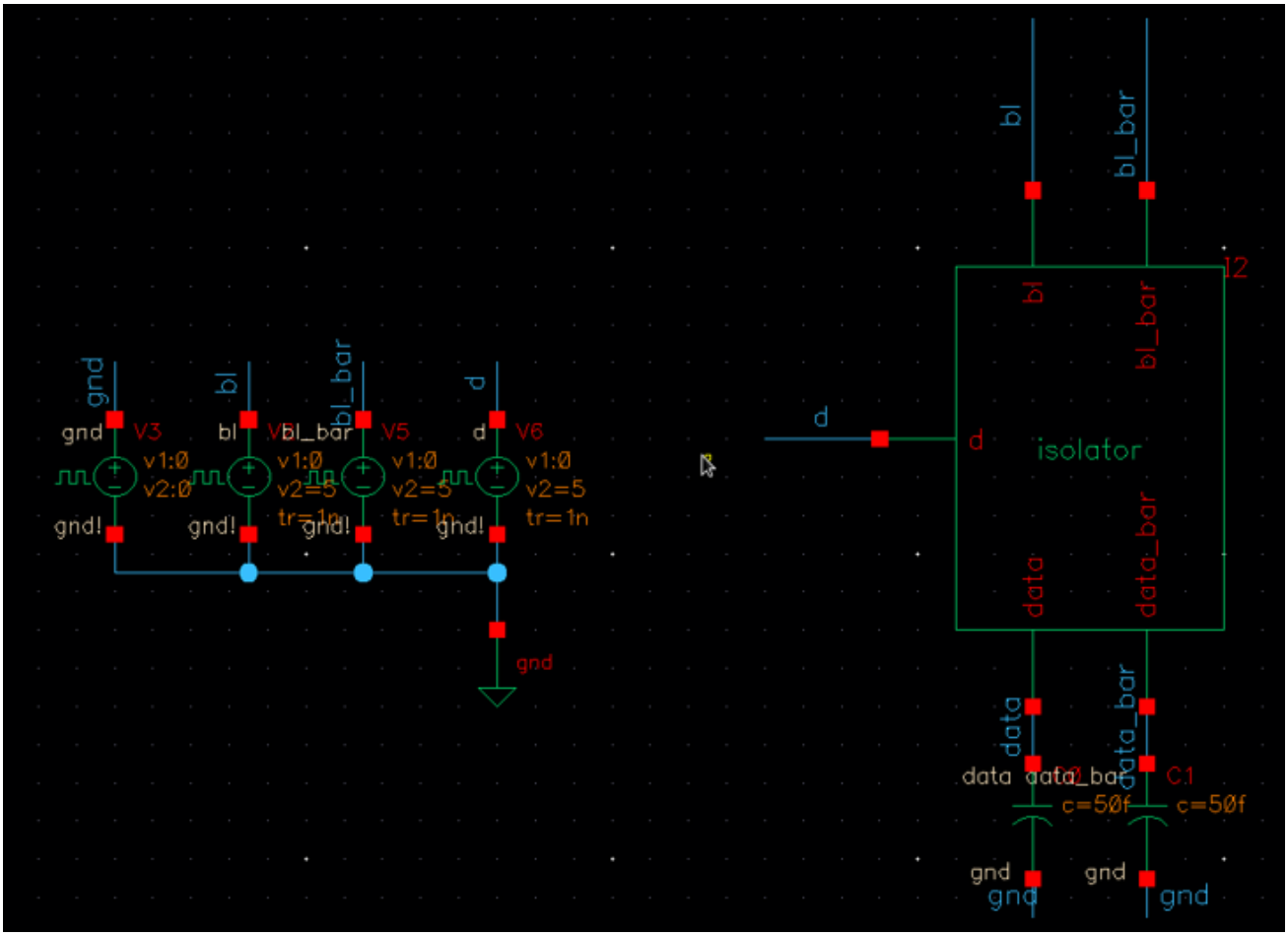
3.5.1 Schematic



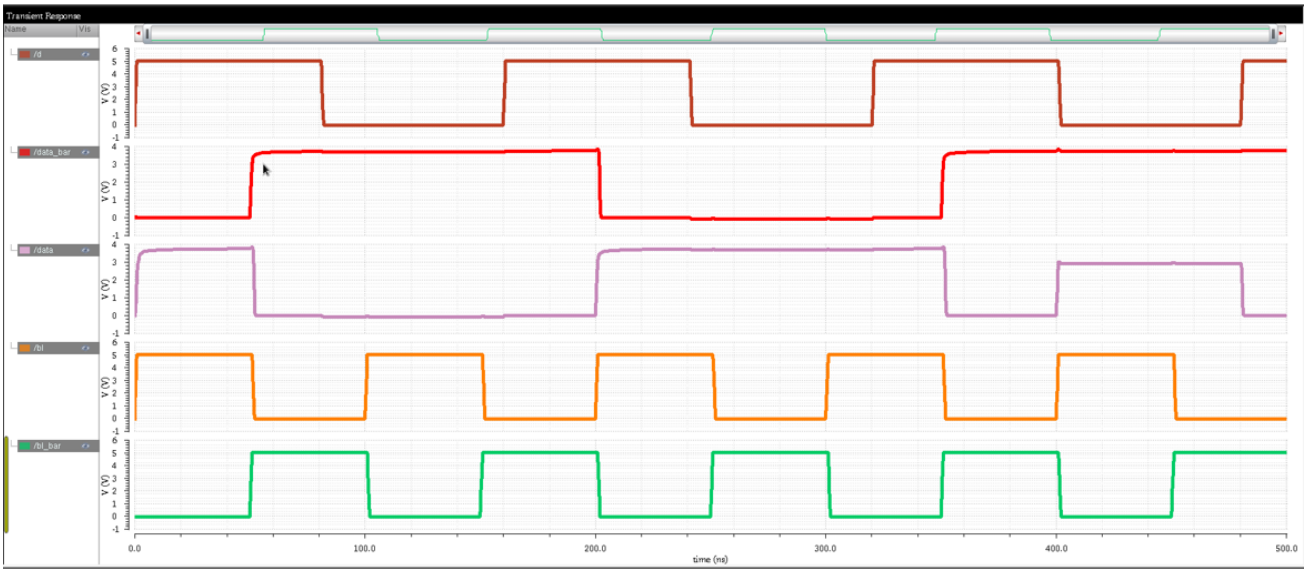
3.5.2 Layout of Isolator



3.5.3 Test Bench for Isolator

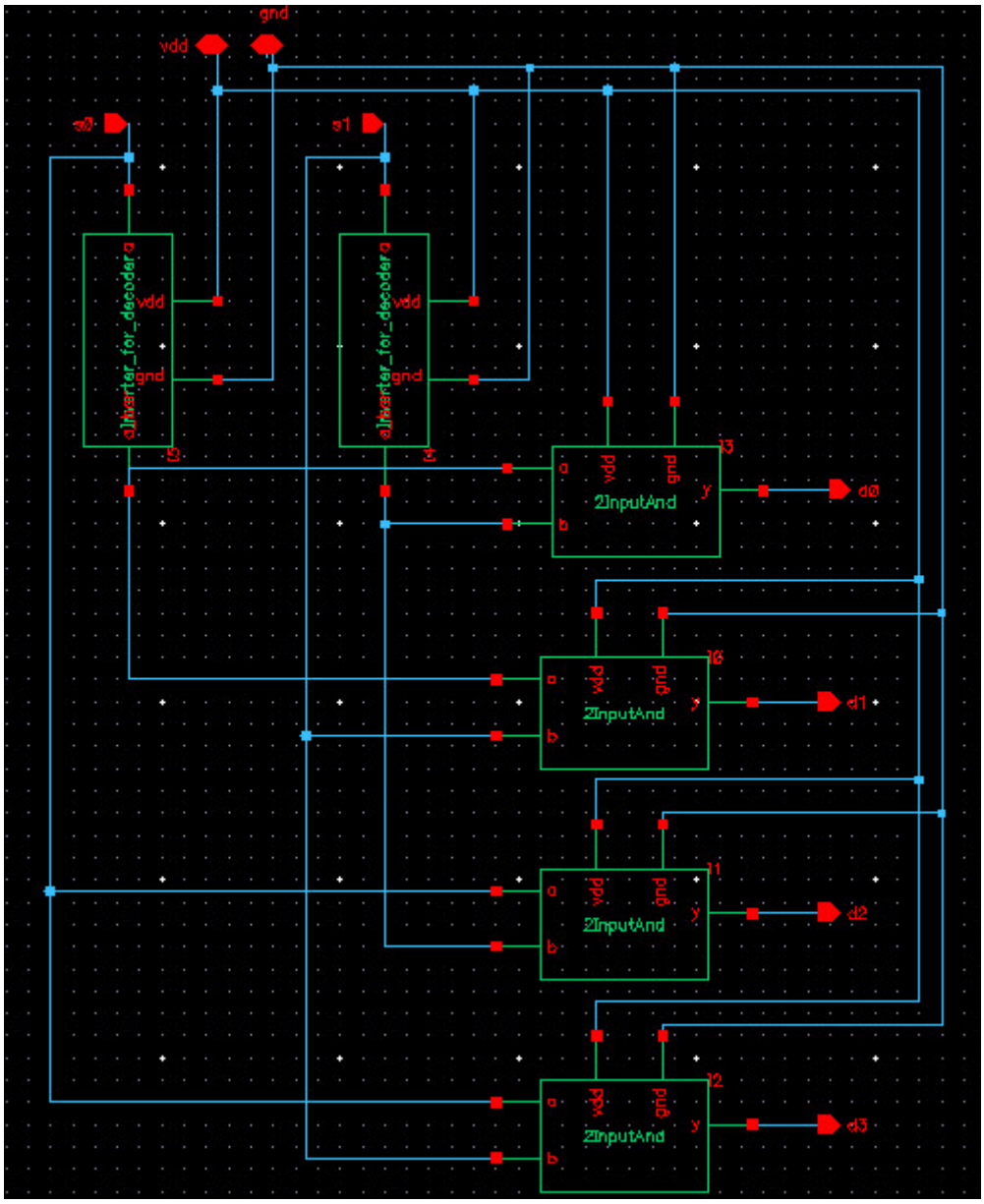


3.5.4 Simulation of Isolator

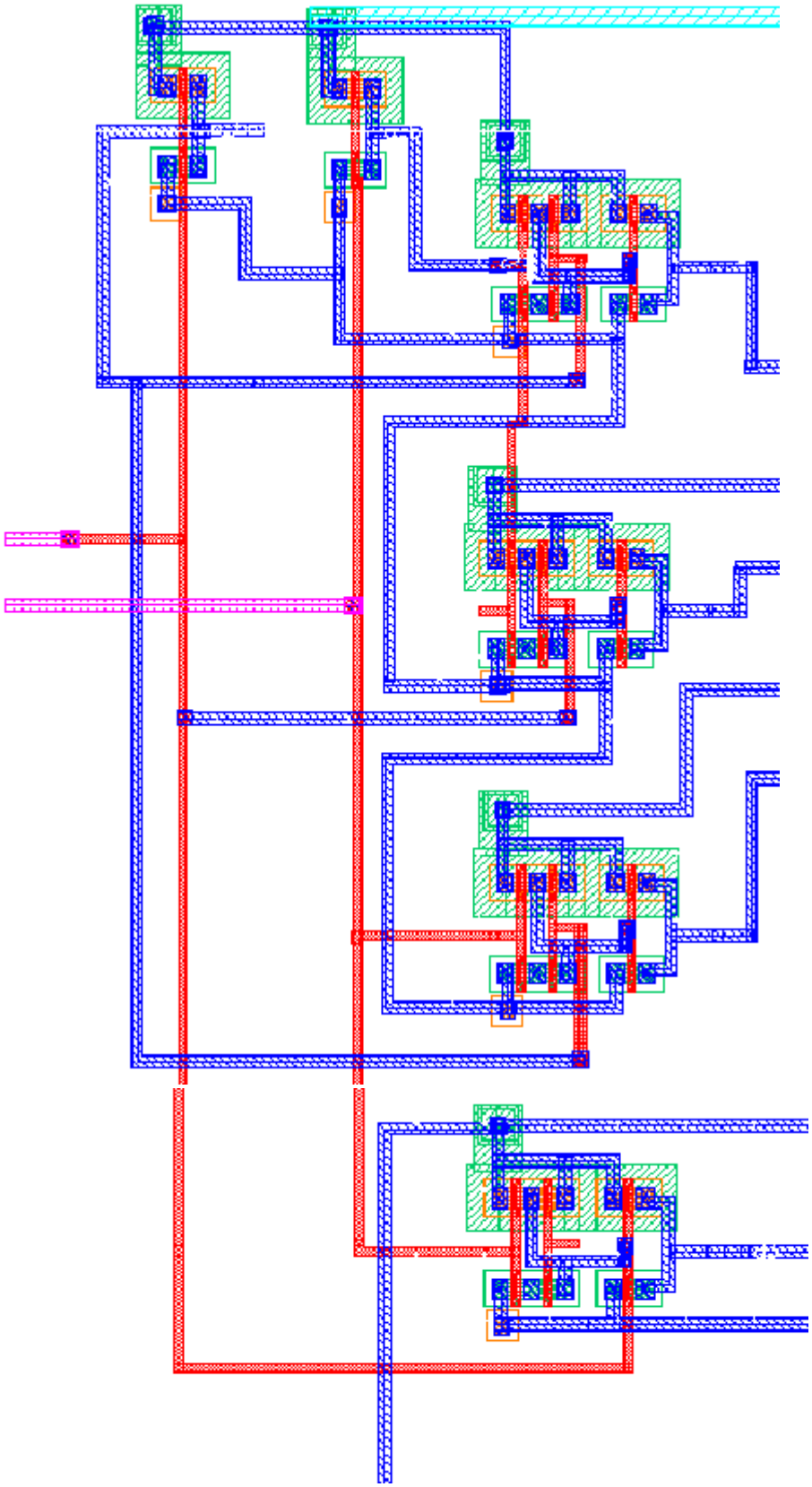


3.6 Row Decoder/ Column Decoder

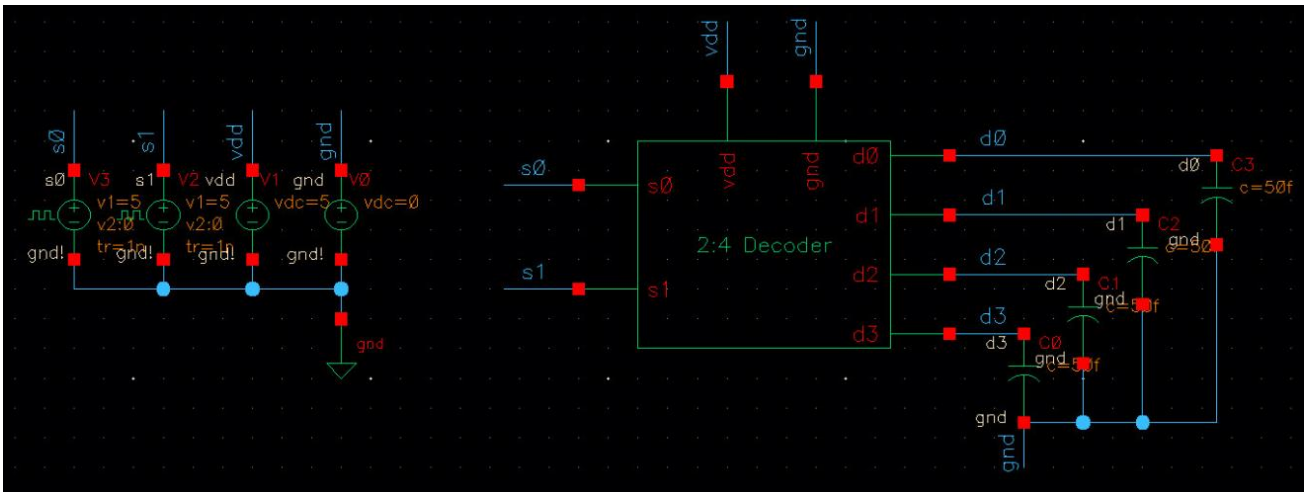
3.6.1 Schematic



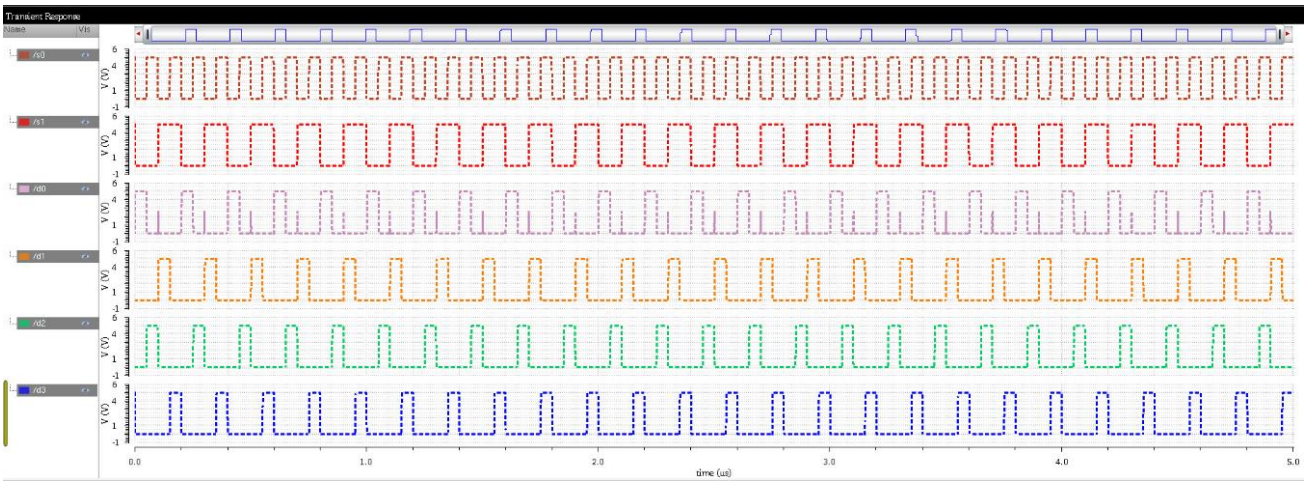
3.6.2 Layout of Row/Column Decoder



3.6.3 Test Bench for Row/Column Decoder

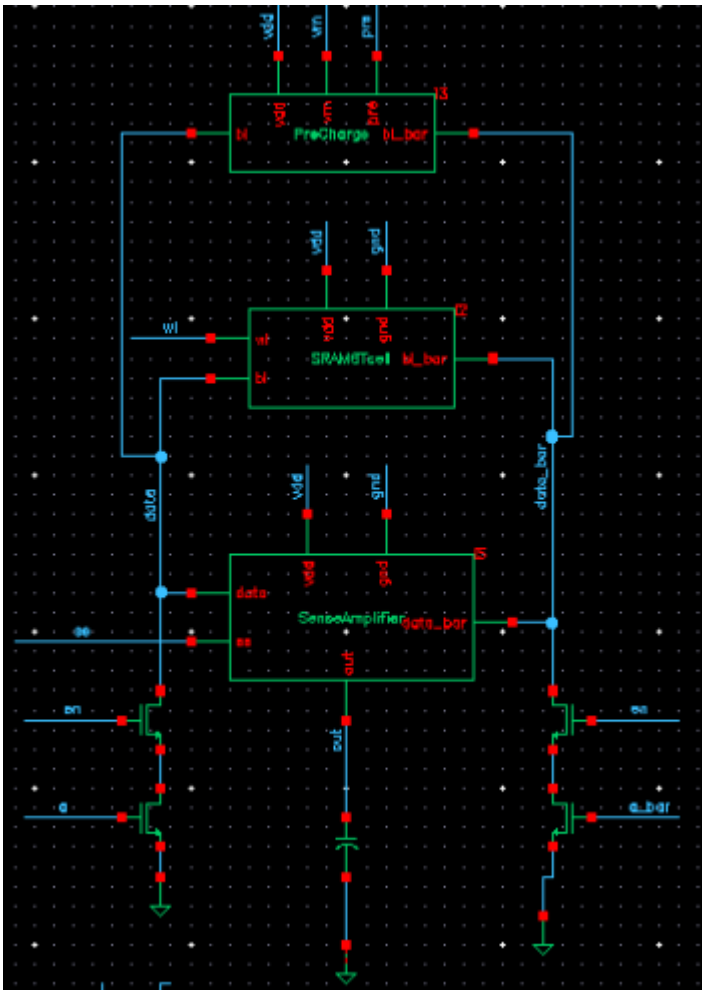


3.6.4 Simulation of Row/Column Decoder

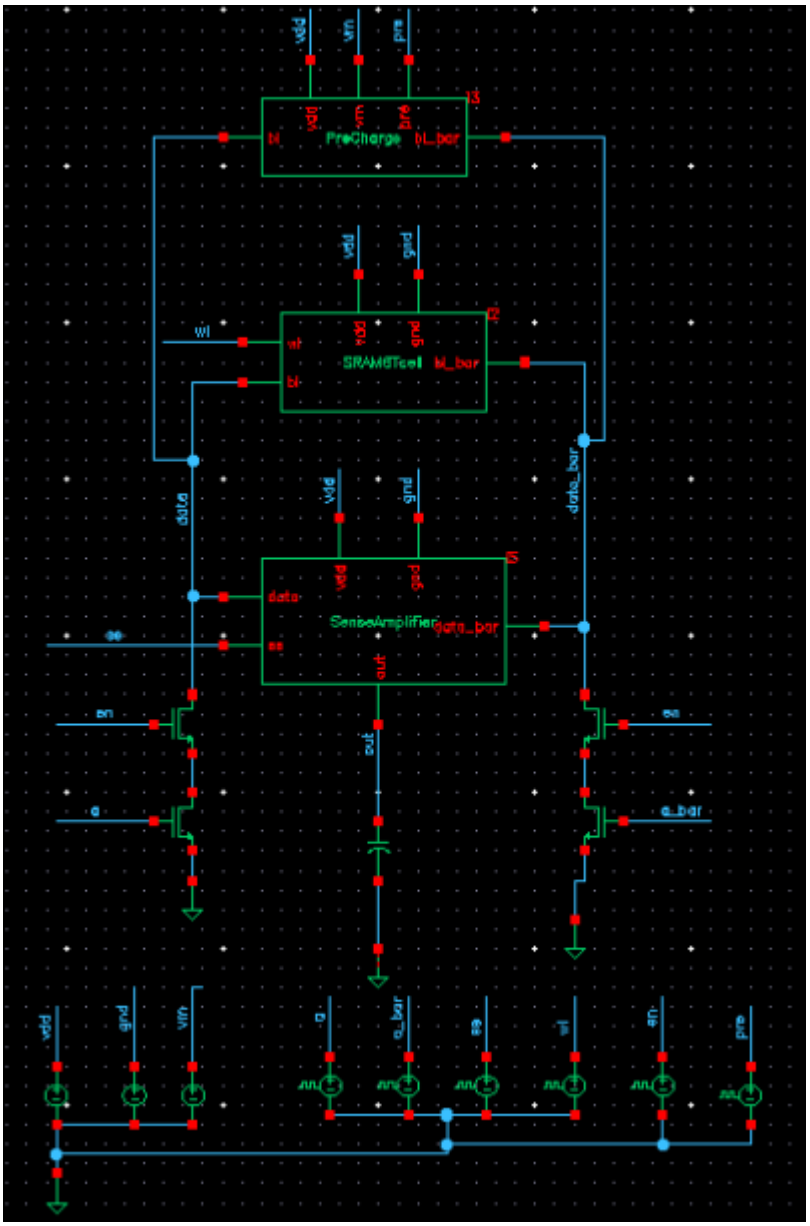


3.7 Design of 1 x 1 SRAM Memory

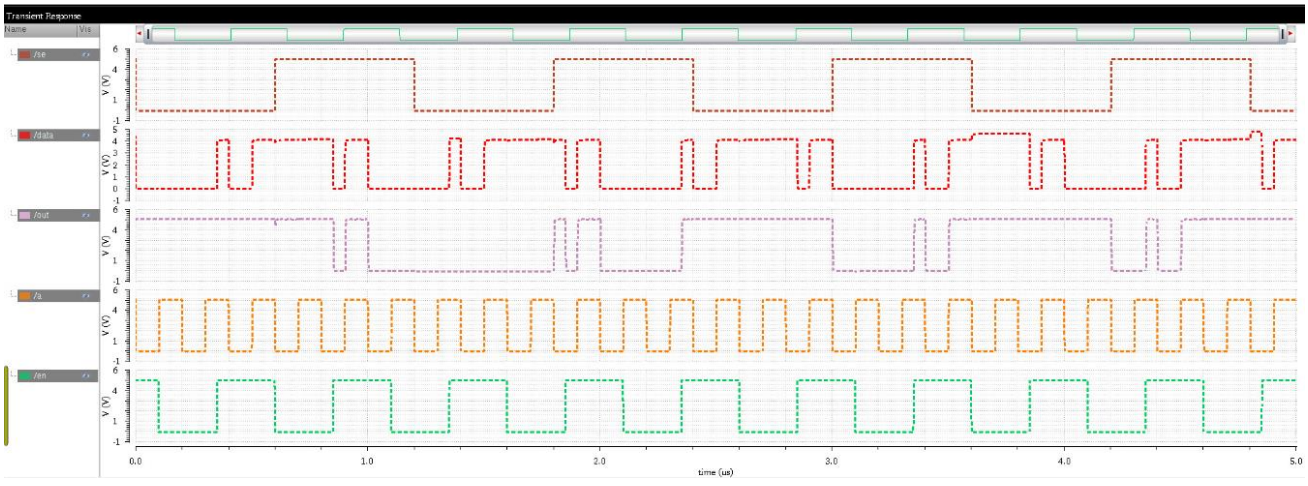
3.7.1 Cell View



3.7.2 Test Bench

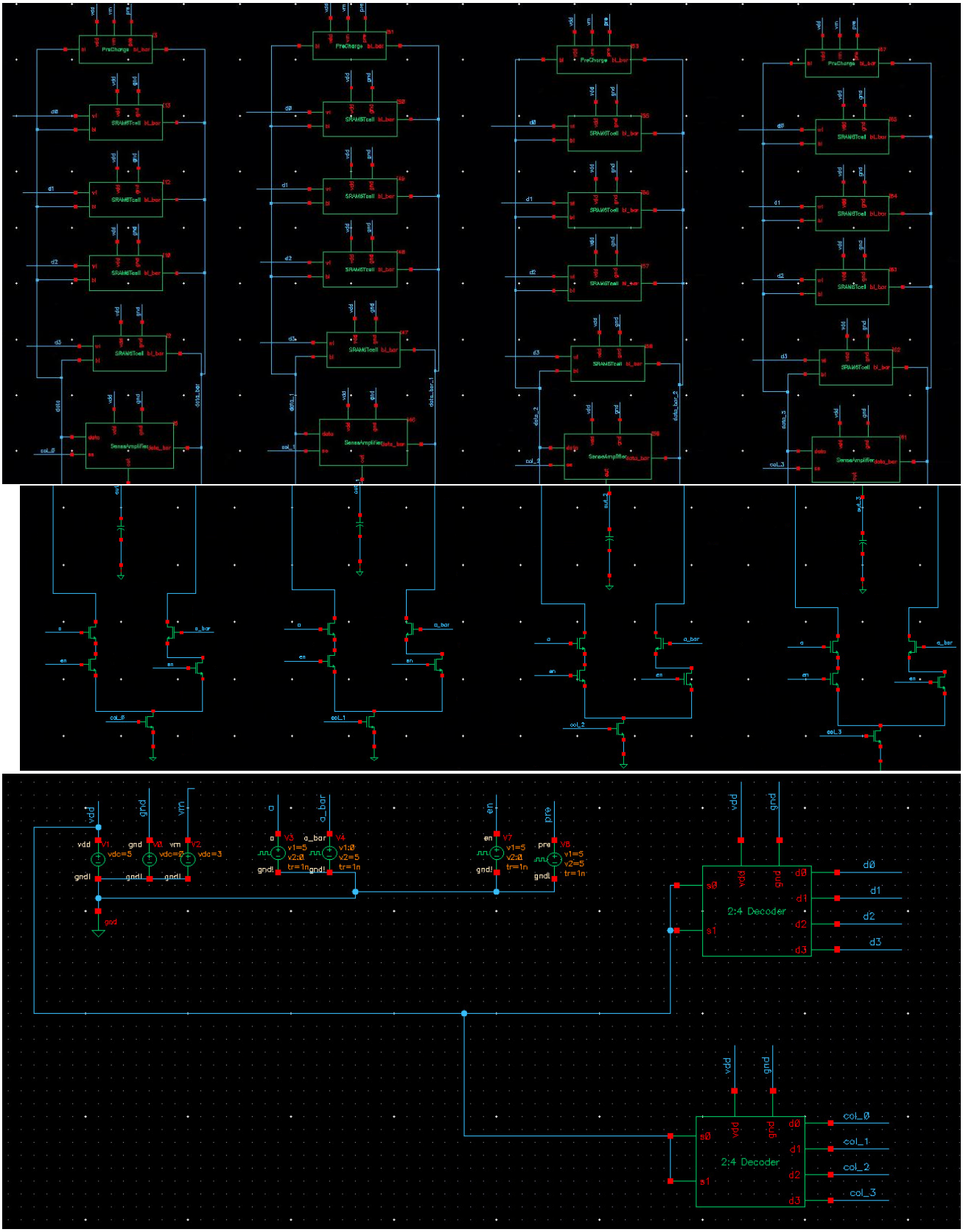


3.7.3 Simulation

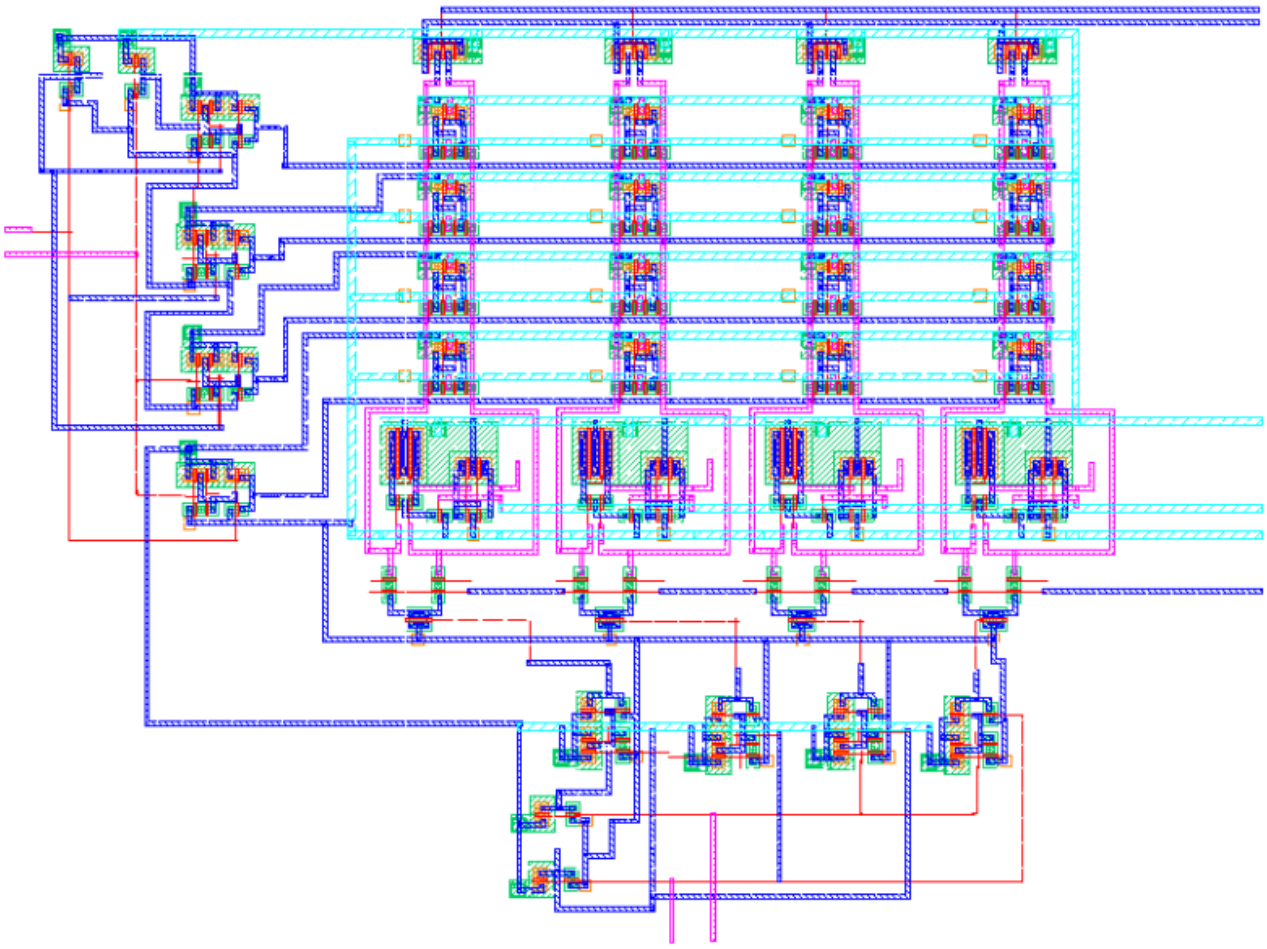


3.8 Design of 4 x 4 SRAM Memory

3.8.1 Cell View

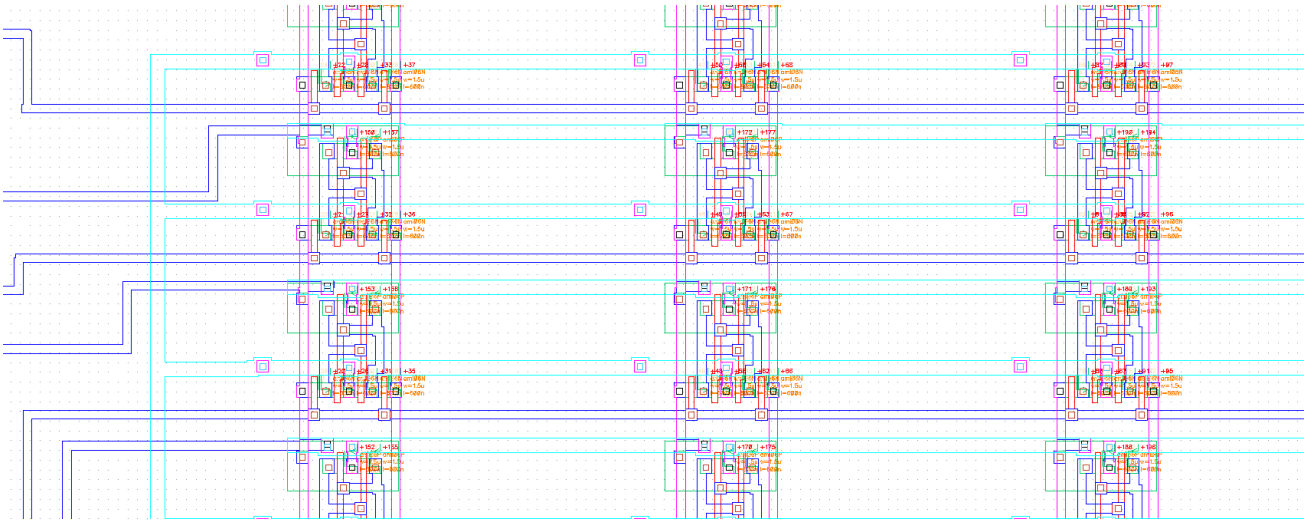


3.8.2 Layout Cell View

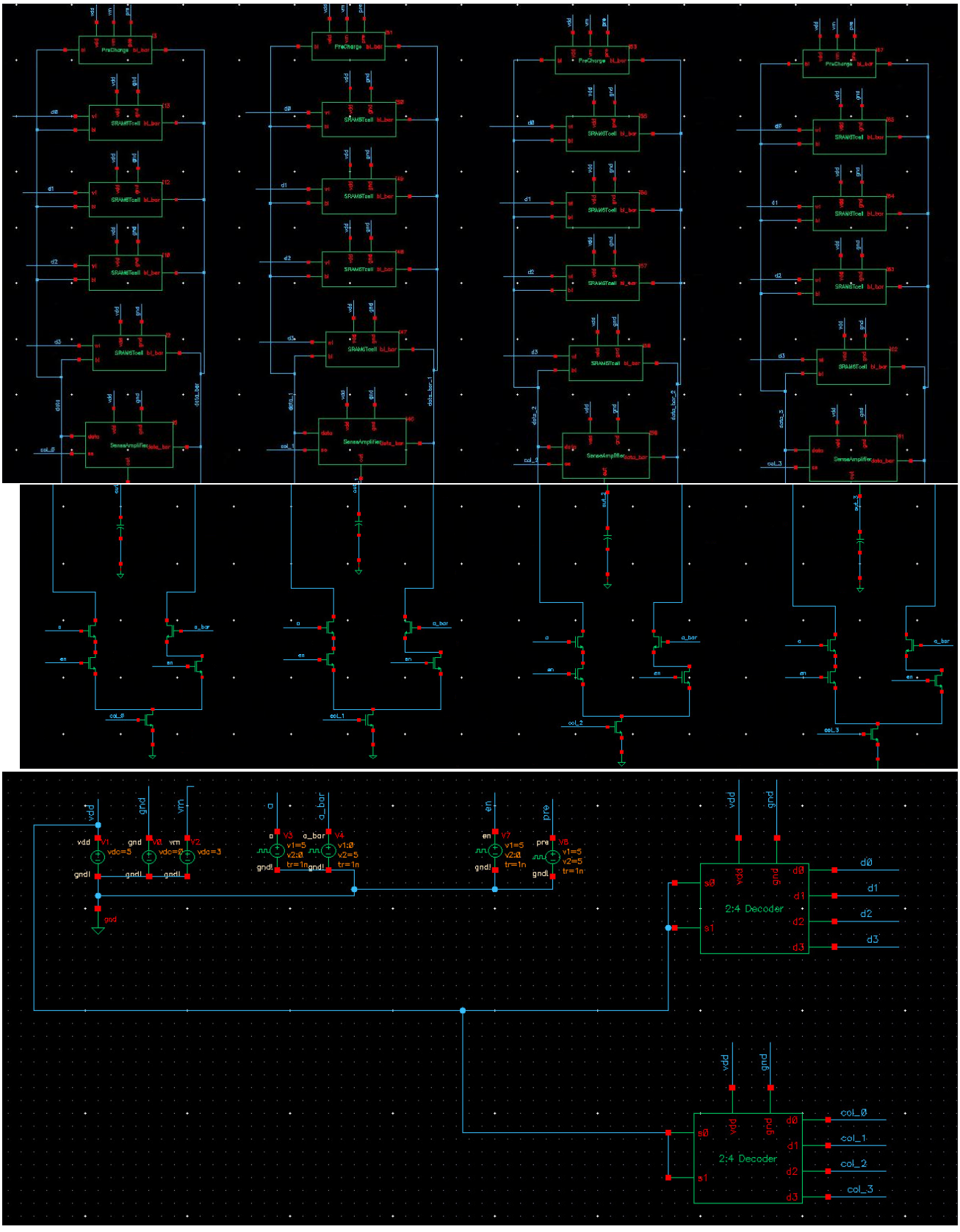


3.8.3 Extracted View

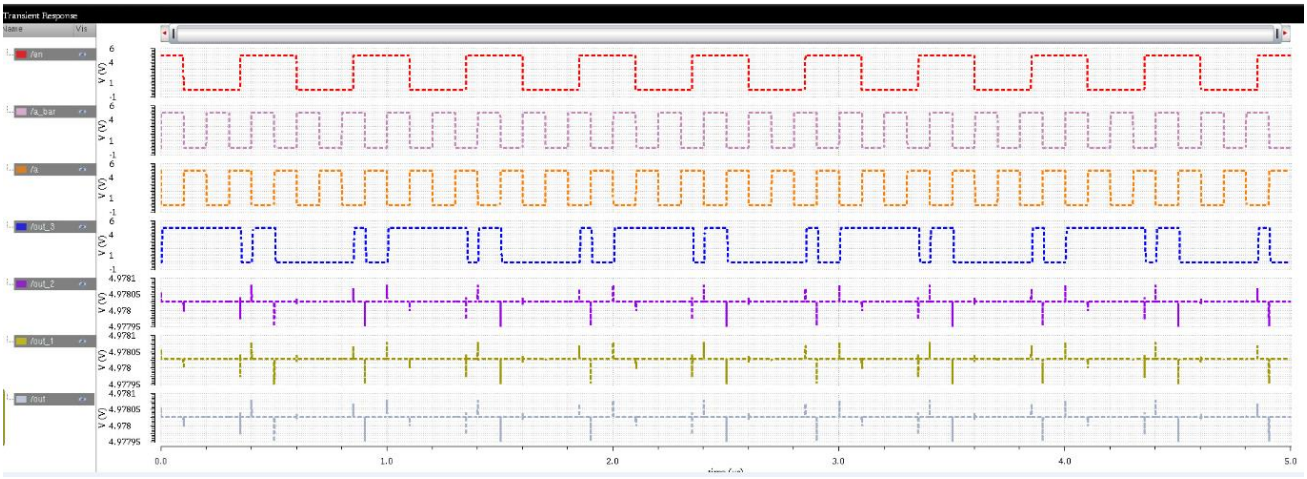
For the ease of reading, we are showing the extracted view for the few columns of SRAM.



3.8.4 Test Bench



3.8.5 Simulation



4. Conclusion

In this we designed 16Kb memory using memory banking method which gives better performance compared to monolithic architecture. The SRAM is divided in to parts such as read data path and another one is row decoder path. The layout of complete design is drawn in an optimised manner, such that we can achieve minimum delay in above two portions. Post Layout simulations are completed and investigated the power analysis of complete design. Initially single 6T SRAM cell is designed which operates at a frequency of 8GHz and PVT analysis for the 6T cell is performed because the design of SRAM cell is the heart of whole design. Stability analysis for single 6T SRAM is performed and designed 6T cell with static noise margin, Read Noise Margin and Write Noise Margin of 240mV,115mV and 425mV respectively for a supply voltage of 1V. The Layout of single cell is drawn in a symmetric manner, such that contacts of one cell are shared with contacts of another cell while making a cell row and cell array. Hence layout area of cell array is minimised which results reduction in the power dissipation of circuit.

All peripherals like pre-charge, row decoder, word line driver, column multiplexer, write driver and sense amplifier are designed .The layouts of all above peripherals are drawn in an optimised manner and individual simulations of all designs are performed. The Frequency of memory is affected with number of rows and columns i.e. frequency of operation is divided by a factor of two as number of rows doubles and similarly frequency is reduced by a factor of four when number of columns doubles, hence memory banking method is used in the design of 16Kb to achieve better frequency of operation.

5. References

[1] Koichi Takeda, Yasuhiko Hagihara, et.al, ―A Read-Static-Noise-Margin-Free SRAM Cell for Low-VDD and High-Speed Applications‖, IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 41, NO. 1, JANUARY 2006.

[2] Byung-Do Yang ―A Low-Power SRAM Using Bit-Line Charge-Recycling for Read and Write Operations‖, IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 45, NO. 10, OCTOBER 2010

[3] Sung-Mo Kang and Yusuf Leblebici ‖CMOS Digital Integrated Circuits‖, TATA McGRAW-HILL EDITION 2003.

[4]. Kunihiko Yamaguchi, Hiroaki Nambu, et. al. ―A 1.5-ns Access Time, 78-pm' Memory-Cell Size, 64-kb ECL-CMOS SRAM‖, IEEE Journal of Solid-State Circuits, Vol. 27, No.2. February 1992

[5]. Jan M.Rabaey Anantha Chandrakasan Borivoje Nikolic ‖Digital Integrated Circuits‖, Pearson Education Electronics ,2003.

[6]. Ya-Chun Lai and Shi-Yu Huang, ―A Resilient and Power-Efficient Automatic Power down Sense Amplifier for SRAM Design,‖ *IEEE Trans. On Circuits & Systems*, vol. 55, no. 10, pp. 1031-1035, Oct. 2008.

[7]. Debasis Mukherjee,Hemanta Kr.Mondal,‖Static Noise Margin Analysis of SRAM Cell For High Speed Application‖, IJCSI International Journal of Computer Science Issues, Vol. 7,Issue 5,September 2010.

[8]. N. C. Li, et. al., ―CMOS tapered buffer‖, *IEEE Journal of Solid State Circuits*, vol.25, no. 4, pp. 1005-1008, August 1990.

[9]. J. Choi, et. al., ―Design of CMOS tapered buffer for minimum power-delay product‖, *IEEE Journal of Solid State Circuits*, vol. 29, no. 9, pp. 1142-1145, September 1994.

[10]. B. S. Cherkauer and E. G. Friedman, ―A unified design methodology for CMOS tapered buffers‖, *IEEE Journal of Solid State Circuits*, vol. 3, no. 1, pp. 99-110, March 1995.