EE577A Phase 2 Lab Report

In Memory Computing using Vector Bitwise Boolean Operations in 10T-SRAM Cell Utilizing Multi-Vt Structure For Power Optimization

Group Name: 4TSRAM

Group Members:
Ziqiao Fan
Zeyu Xie
Chunxiao Lin
Mustafa Altay Karamuftuoglu

November 27, 2020

Objective

In phase 2, we have followed multiple design knobs and added our own way of implementation to decrease the power consumption and leakage. First of all, we started with decreasing the VDD of SRAM to 0.9 Volt. Since we decreased the VDD and didn't change the wordline driven voltage which is 1 Volt, this also allowed us to achieve wordline overdrive feature while making the access transistor stronger. Since our access transistor became stronger, we decreased its size to a minimum value 120n. Additionally, in order to decrease the leakage, we changed all SRAM transistors and replaced them with high threshold transistors (hvt) from the library. Moreover, the precharge voltage of bitlines for write and read assigned 0.9 volt as well. Following this, we implemented further Vdd decrement on the write operation by making it 0.25 volt only during write and idle state (Vdd is 0.9 while read and subthreshold operation during write). Since we didn't change the bitline voltage (letting the precharge voltage as 0.9 volt), charge sharing feature allowed us to achieve faster write operation by flipping the Q/Q_bar (the voltage became approximately 0.48 volt due to charge sharing only during the write and slowly decreased due to leakage) when we activated the word-line. This voltage change is achieved by putting an inverter and activating it with write and read wordlines.implementing the multi-vt design into the SRAM. Furthermore, we drive the ground path of each row's read transistors by an upsized inverter (transistor stacking and making VDD when we don't read). As a result, this inverter enabled us to get lower leakage on read bitlines. Since we didn't change any peripheral circuit, their power consumption became dominant when compared to the previous phase.

Implementation List:

- **Multi-vt on SRAM (High Vt implementation)**
- **Voltage reduction on Vdd of SRAM and Precharge circuit**
- **Word-line overdrive to strengthen access transistor**
- **Subthreshold write operation with charge sharing (write assist by bitlines)**
- **Resized SRAM transistors to the minimum value (everything 120n with high Vt)**
- **Transistor stacking on ground path**
- **Negative Vgs due to the high voltage on ground path driven by an inverter for read**

SRAM Array Design

For lowering the VDD of the SRAM cell, we added additional two inverters between the row decoder and the SRAM cell that have an internal VDD of 0.9V. The output of the inverter would provide a constant 0.9V voltage supply. Erased some of the inverters to decrease the delay from decoder and added new gates to balance out the path due to logical effort. Before reading, we want to pull up the Q value to 0.9V, we provide a delay cell to let the give it enough time to achieve that voltage.

1. Multi-VT 10T-SRAM Cell

Description

Unlike the 6T-SRAM cell, we have additional 4 NMOS transistors which isolate the read operation. Since the read operation is isolated, pull-down NMOS transistors of 10T-SRAM can be lowered down. However, we still need to properly size the pull-up and access transistors similar to the 6T-SRAM cell. When this SRAM contains logic 1 on Q, the bitline value will be pulled down to the logic 0 due to the activation of NMOS on the read path.

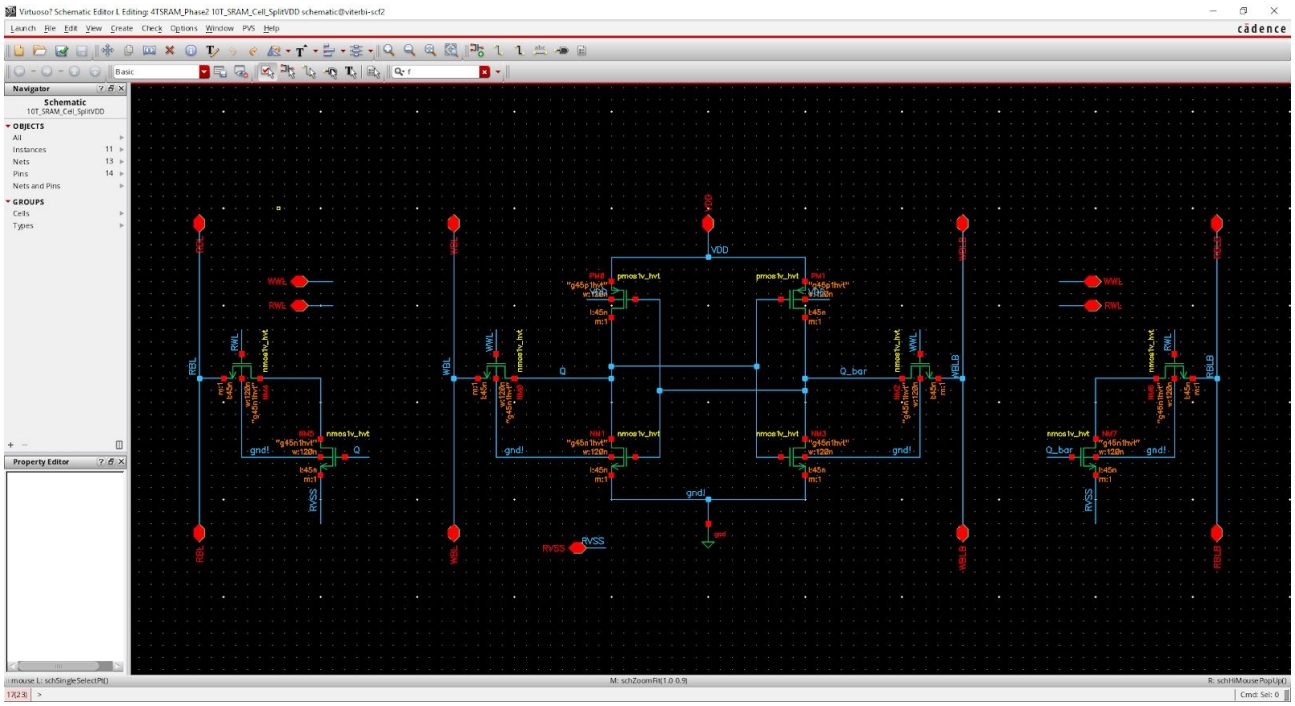

Figure 1 Multi-VT 10T-SRAM cell

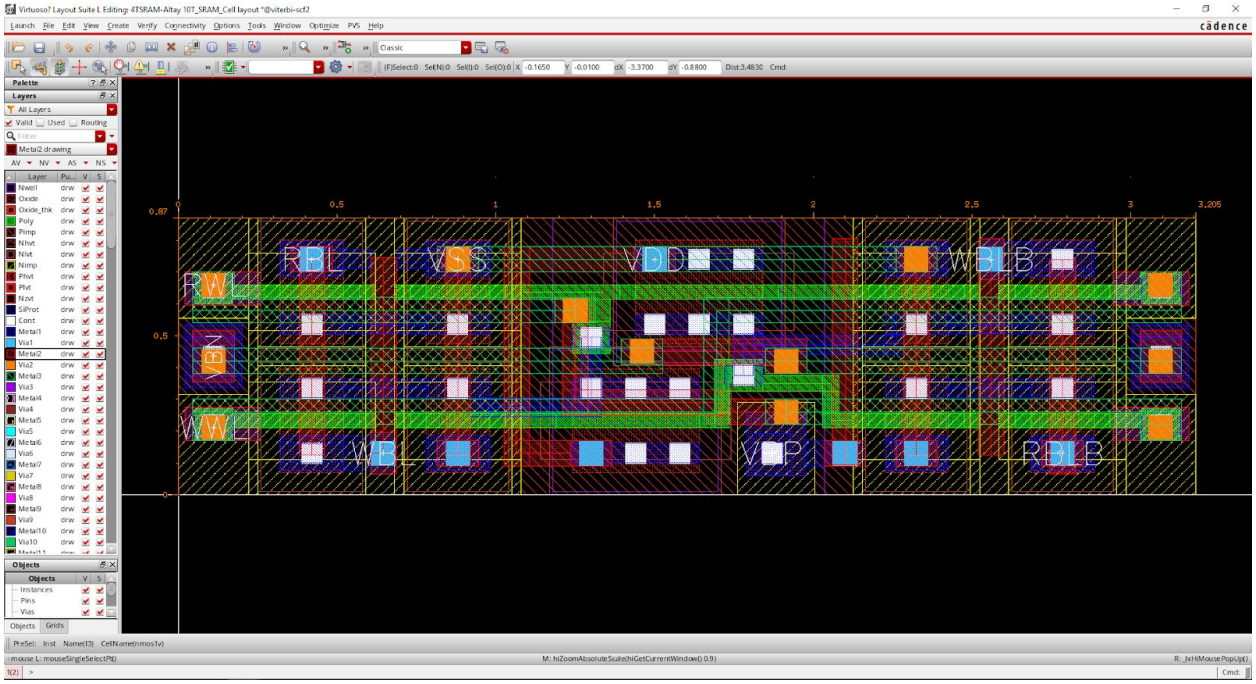

Figure 2 Old version SRAM layout

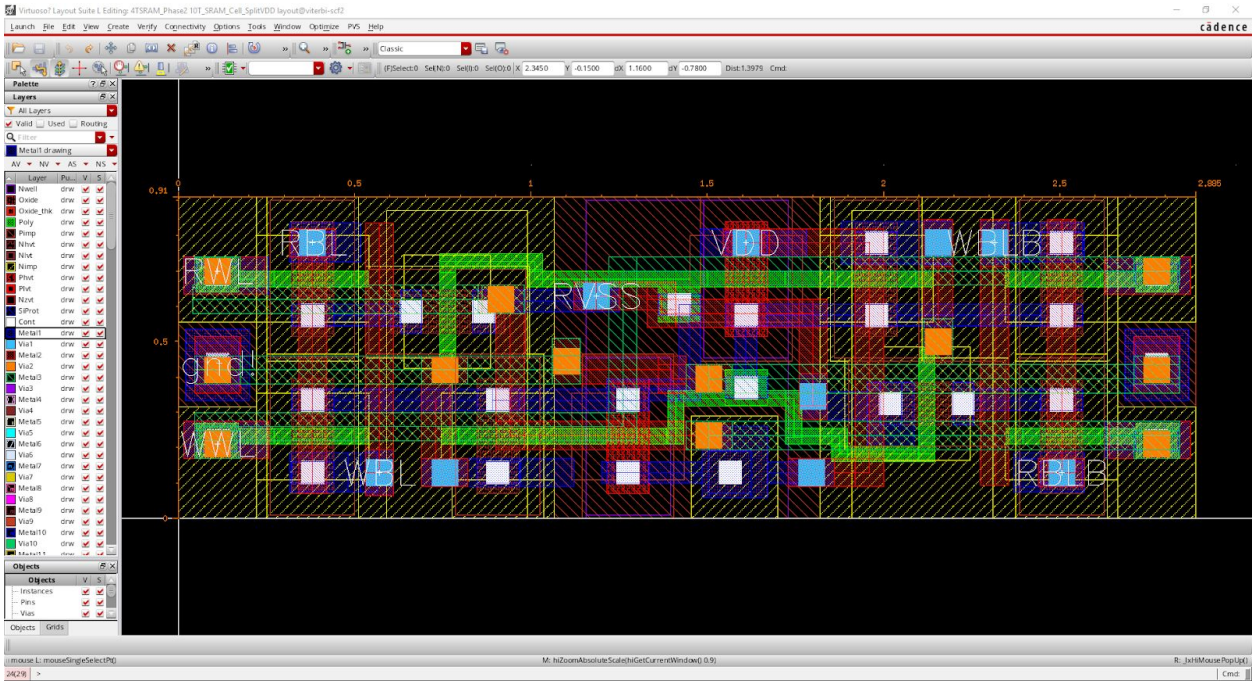

Figure 3 New version SRAM layout

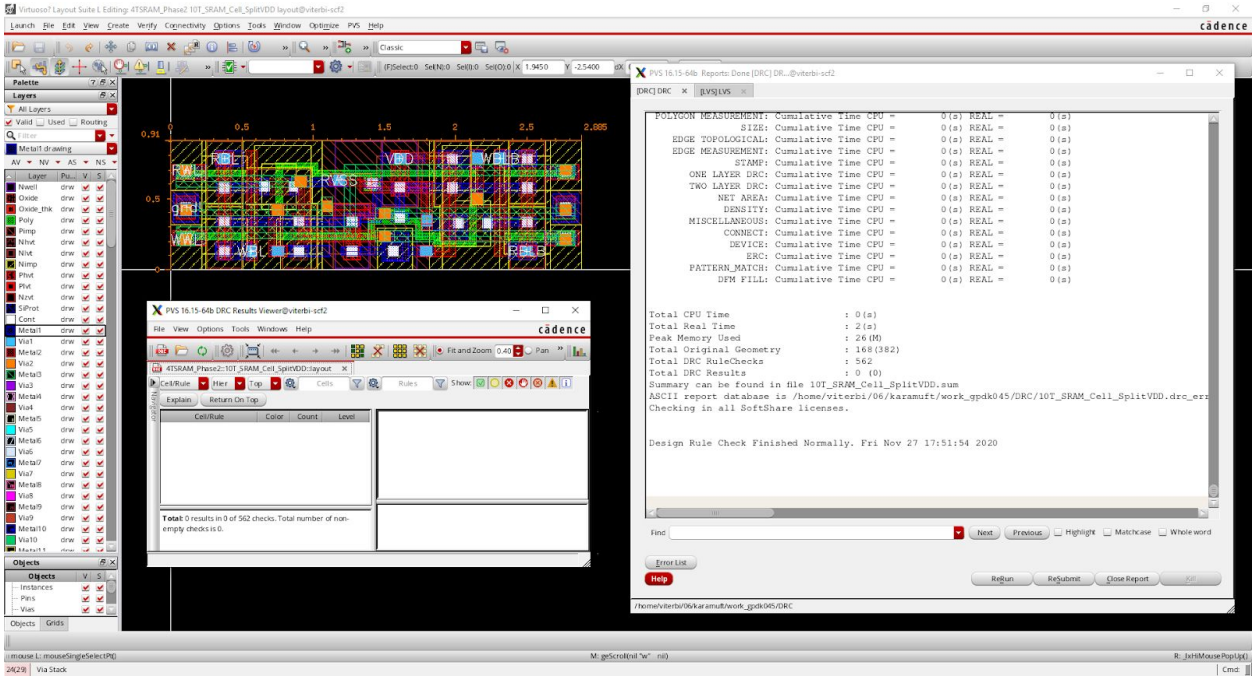

Figure 4 New version SRAM DRC

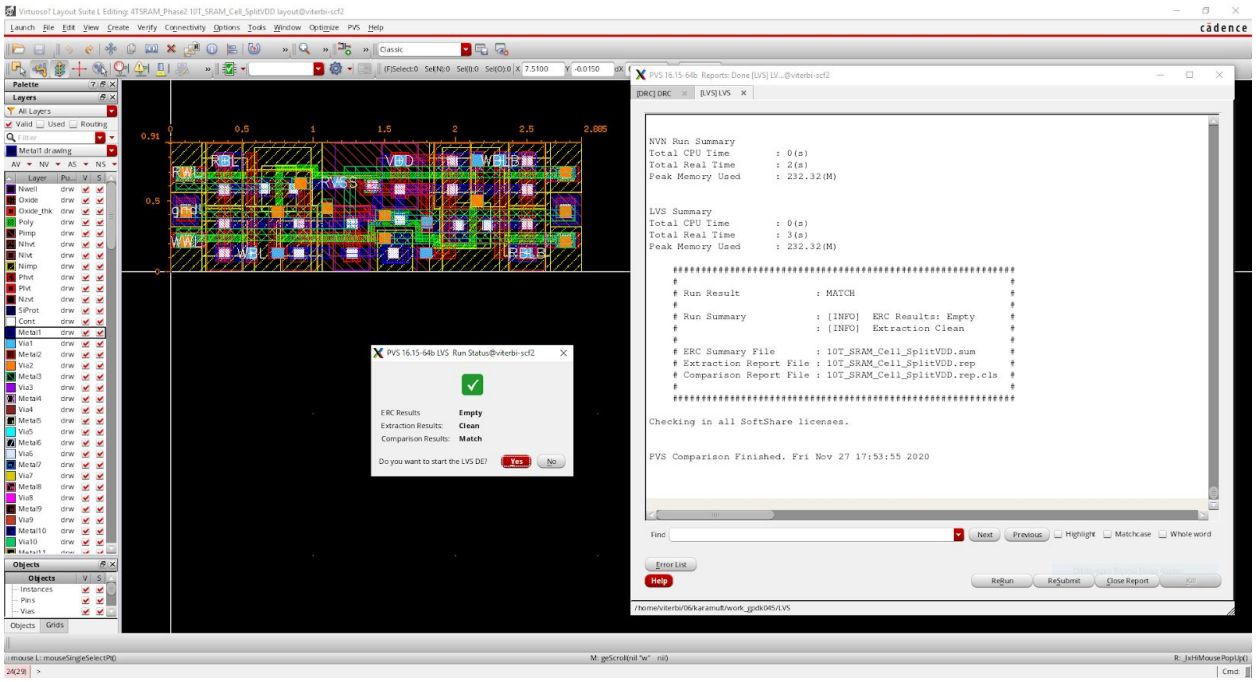

Figure 5 New version SRAM LVS

2. Sense Amplifier

Description

This sense amplifier design is current based unlike the one in the previous phase. Here, we use the BL and BLB voltage on the gate of a transistor. Due to the voltage difference on these asymmetric transistors, the current flowing to the ground will be different. Depending on asymmetric transistor strengths, the value within the cross-coupled inverters will flip according to the BL and BLB values. In phase 2, we have two asymmetric amplifiers. One of them used for small sensing and both of them used for in-memory computation (XOR operation).

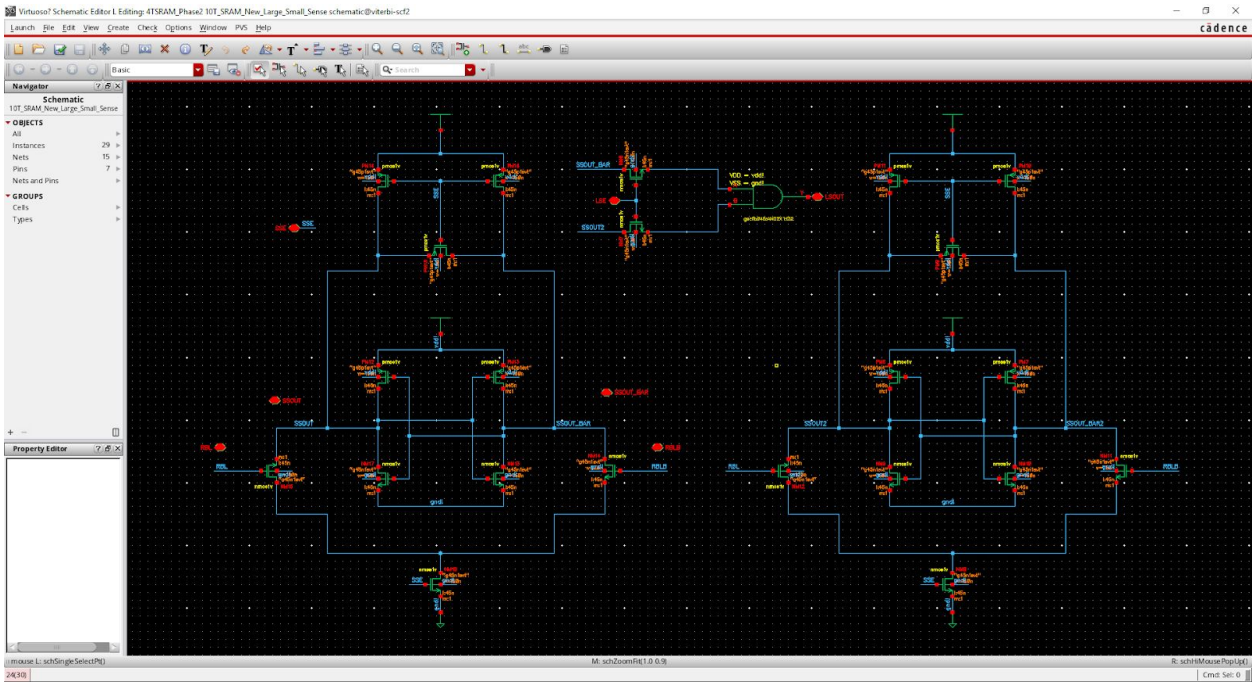

Figure 6 Asynchronous sense amplifier

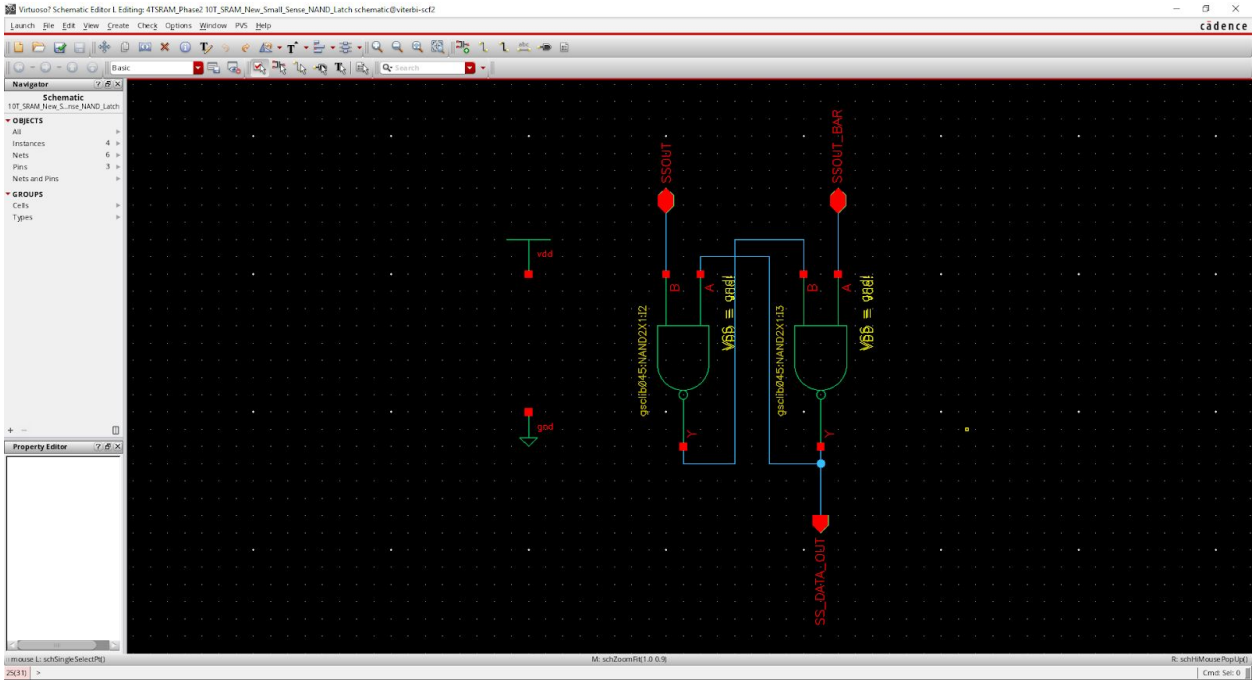

Figure 7 latch

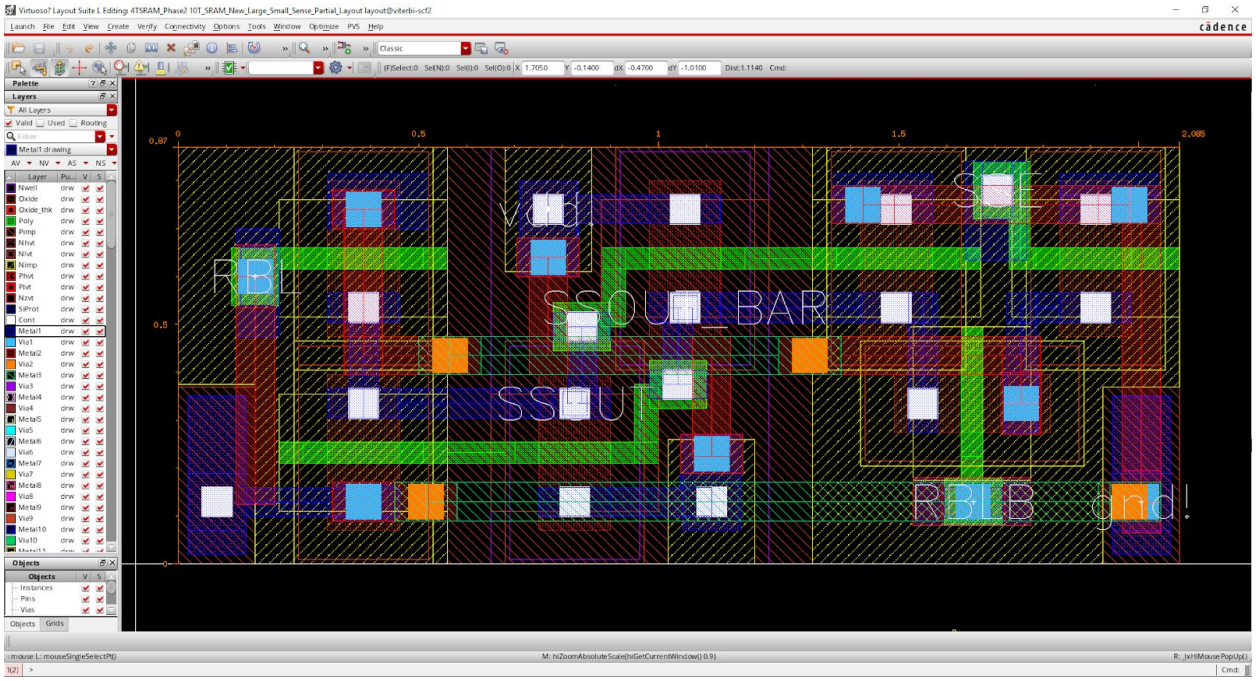

Figure 8 Sense amplifier layout

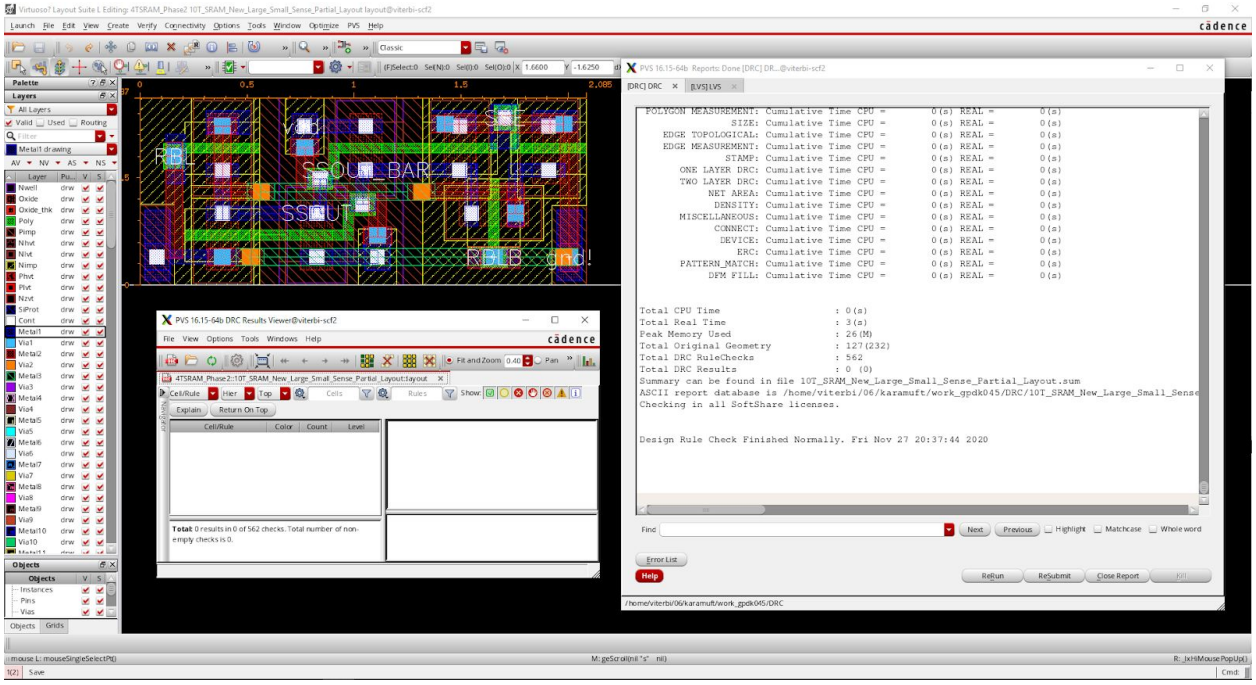

Figure 9 Sense amplifier DRC

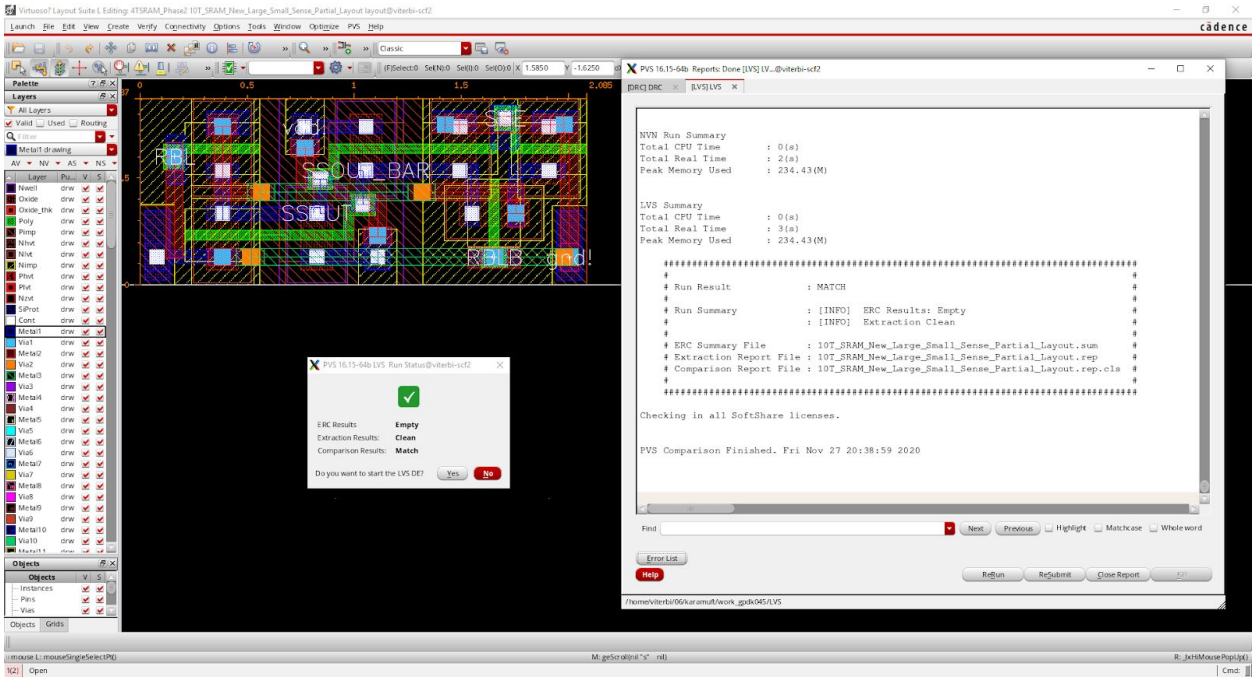

Figure 10 Sense amplifier LVS

3. Precharge circuit

Description

There are two precharge circuits and they are used to precharge the bitline and bitline_bar of write and read operations. The sizes of these transistors are assigned as 180nm.

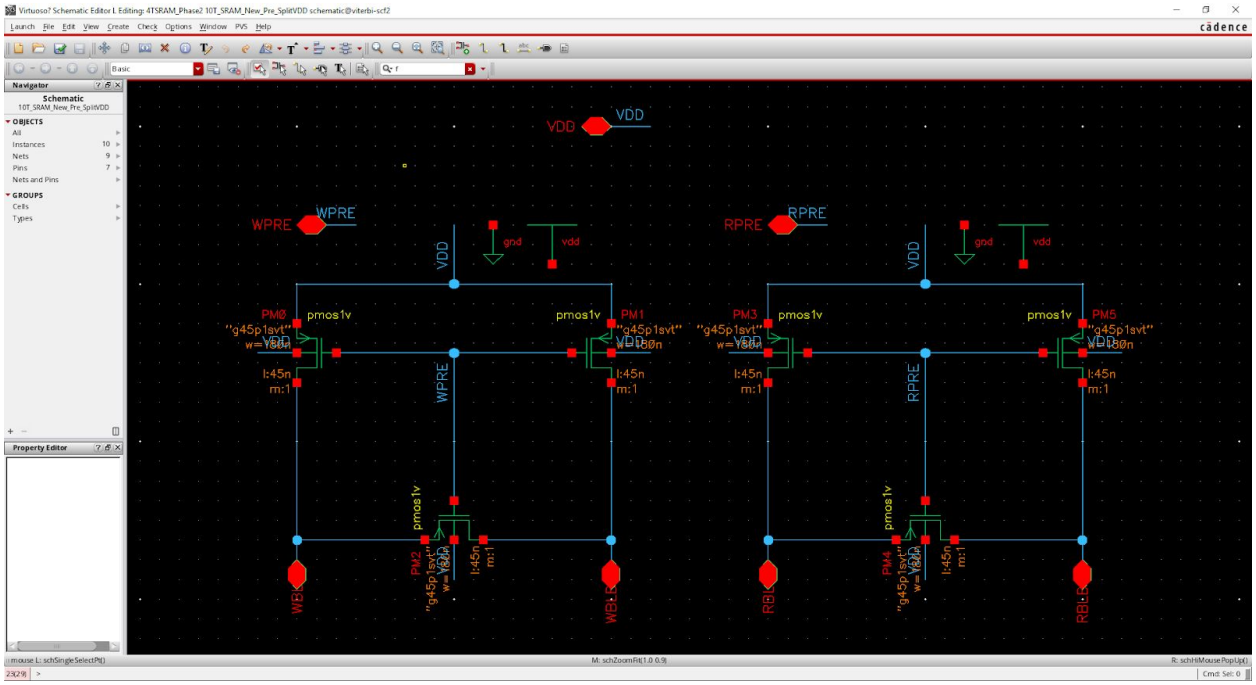

Figure 11 Precharge schematic

4. Row decoder

Description

For the row decoder, in order to reduce the delay, we removed the two inverters in the row decoder cell and rearranged them to different locations.Memory mode circuit has two NAND gates and one inverter to achieve only wordline selection during the memory computing. The normal mode circuit has only two buffers and one inverter which makes it relatively simple. This decoder controller is always active since the decoder on the left hand side will be activated in both normal and in-memory computation.

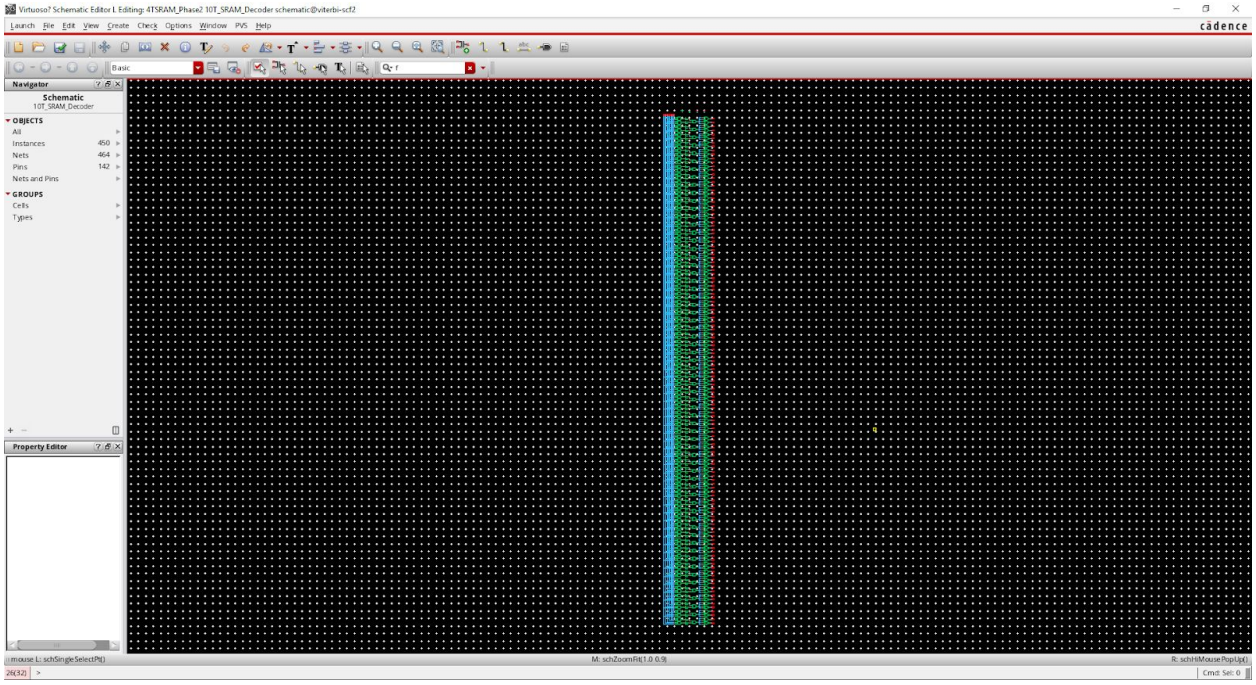

Figure 12 Decoder overview

Figure 13 Decoder zoomed schematic

5. Write Driver

Description

In order to achieve correct write operation on the SRAM columns, each column has its own write driver circuitry to pull down the write BL and write BLB. The sizes of these NMOS transistors determine the pull-down speed of the BL and BLB voltages. The related pictures are shown in below. For phase 2, we reduced the size of NMOS accordingly for the delay. The sizing of NMOS is listed in page below.

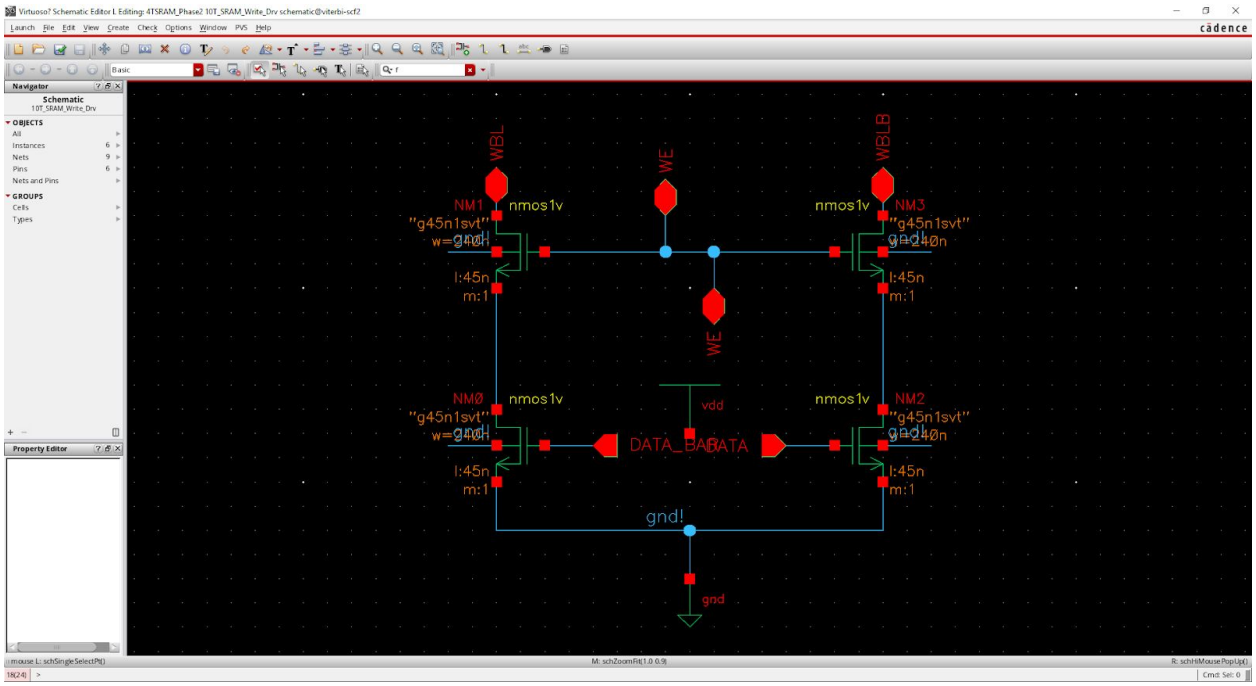

Figure 14 Write driver schematic

6. DFF

Description

Since we don't have any multiplexer structure, we have placed DFF to large sensing and small sensing parts separately. With every rising edge of the clock pulse, we sample the data coming from the sense amplifiers.

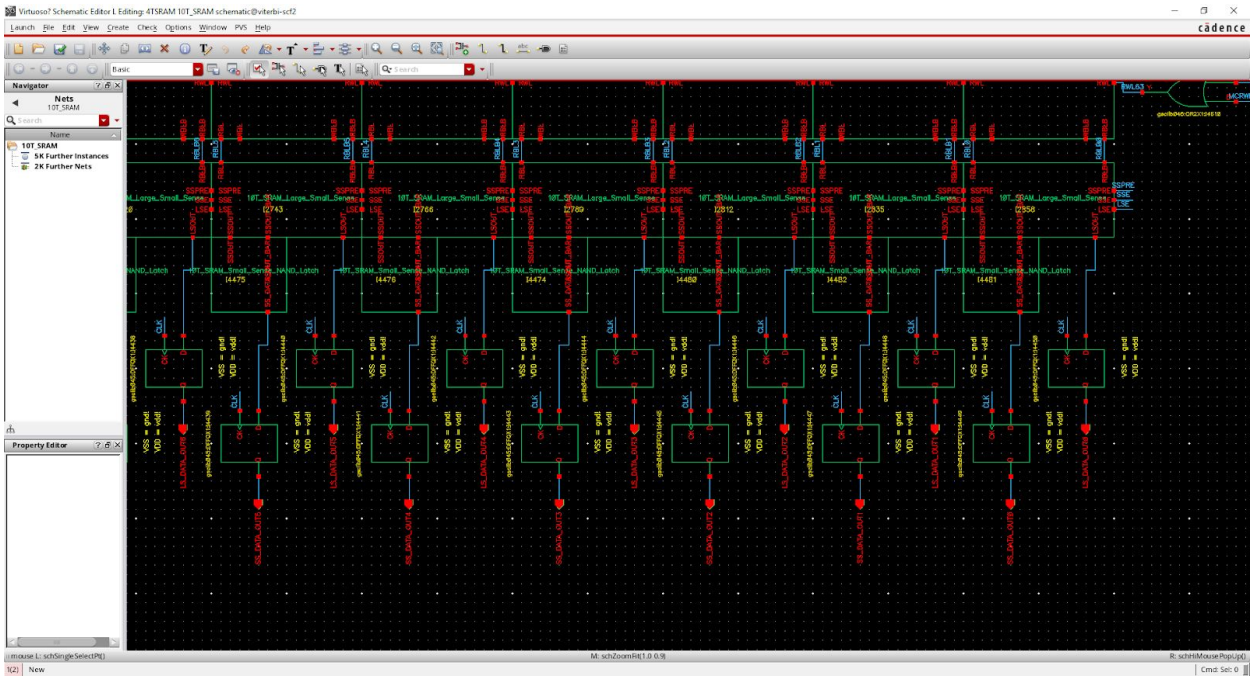

Figure 15 DFF placements

7. Overall 64x64 SRAM Schematic

Description

The SRAM structure consists of cross-coupled inverters which are used to hold the state in CMOS. This static storage in a powered cell does not require a refresh operation. To write a new state, it is required to force the nodes into the opposite state. The writing operation is achieved by enabling the write wordline and it is connected to the gate of access transistors. The read operation is achieved by peripheral transistors that are placed in a way that reading will not affect the data. Overall, we have 10 transistors (4 for the two inverters, 2 for the access, and 4 for BL and BLB read lines). In the project, we are asked to design a 1K bit (64x64) SRAM circuit.

In normal mode, we activate the decoder on the left hand side given in the figure below. In in-memory computing, we activate both of the decoders and by using logic OR gate, we achieve multiple wordline access on SRAM structure. Since small sense and large sense circuits have their own enable signals, we can connect the read BL and BLB paths to the corresponding sense circuits.

During the in-memory computing mode, we enable both of the sense amplifiers. The asymmetry within these circuits are mirrored to achieve the correct truth table. By giving the SSOUT_BAR from sense amplifier 1 and SSOUT2 to the AND gate, we will achieve XOR operation.

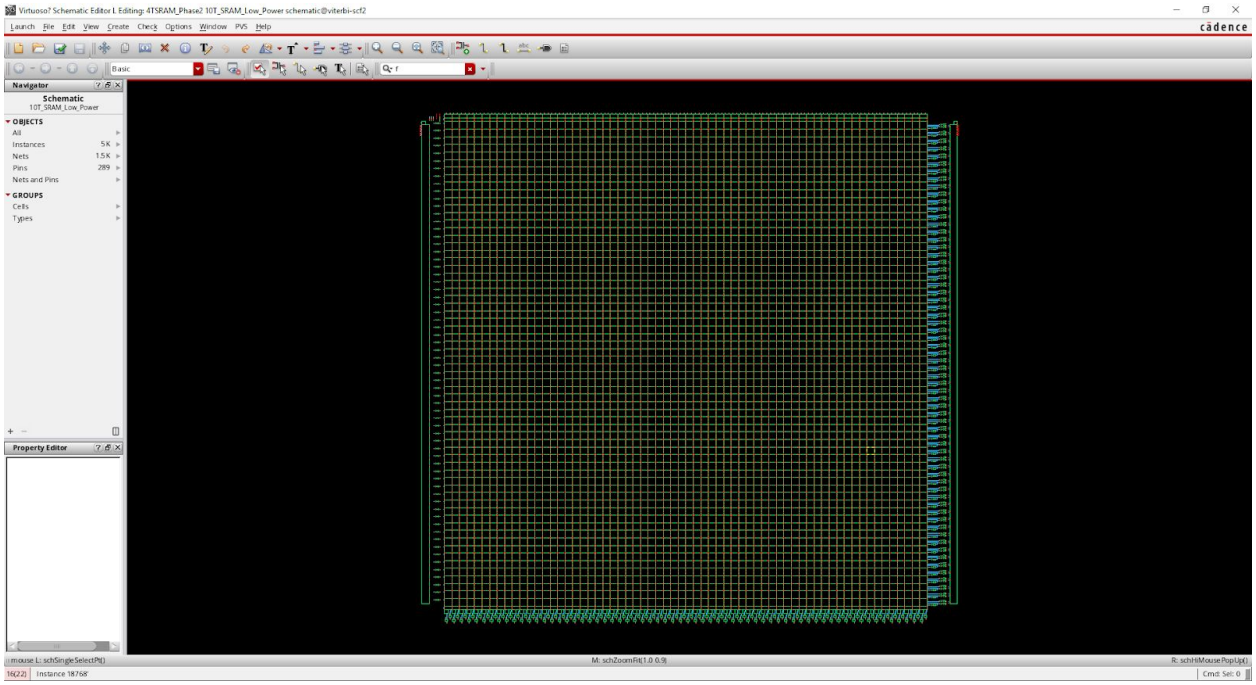

Figure 16 64x64 SRAM schematic overview

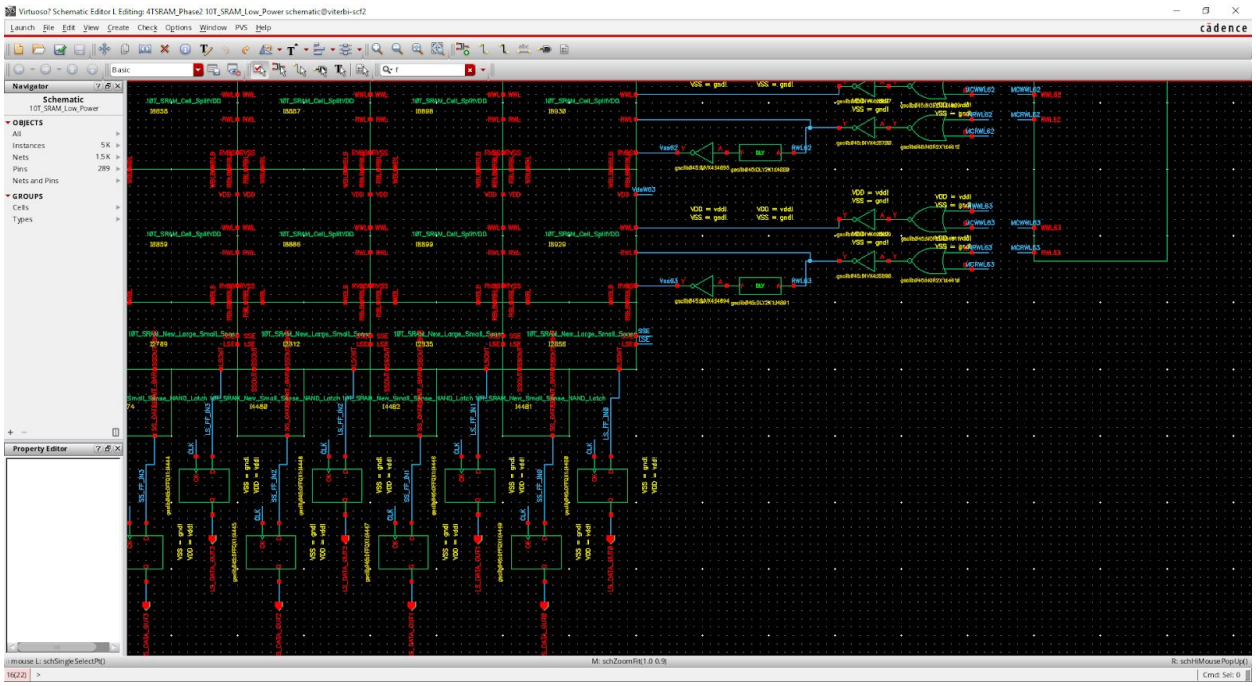

Figure 17 10T SRAM 64x64 Schematic Zoomed-1

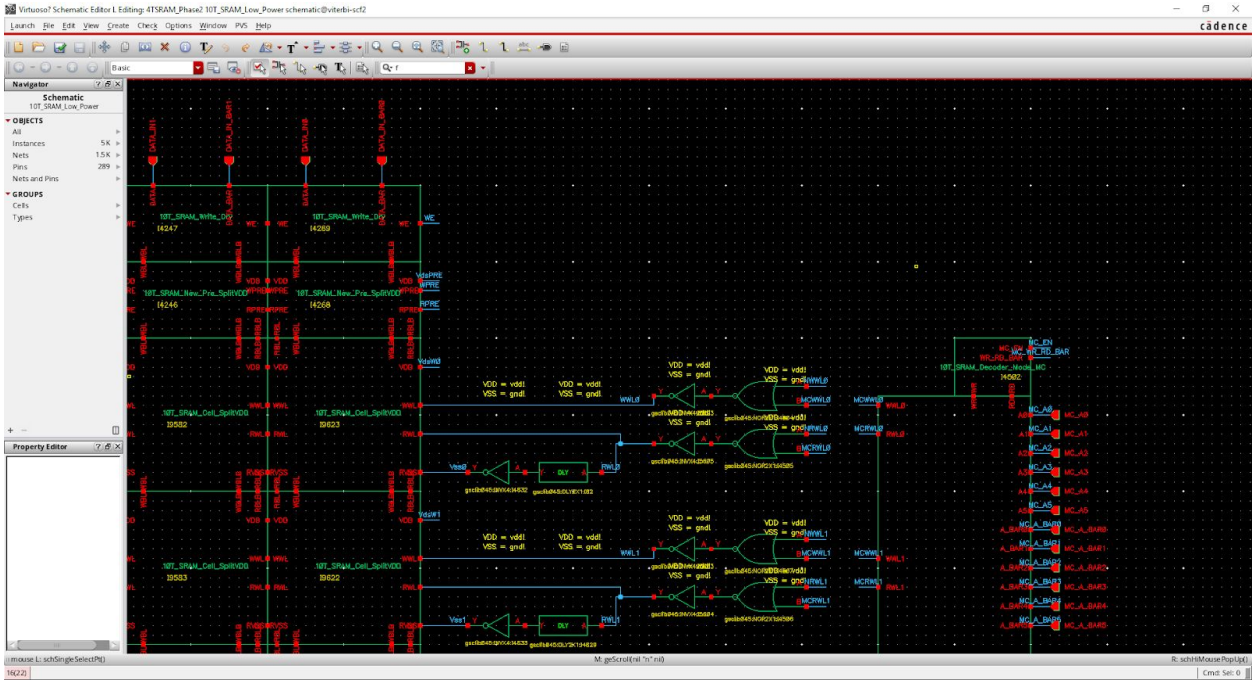

Figure 18 10T SRAM 64x64 Schematic Zoomed-2

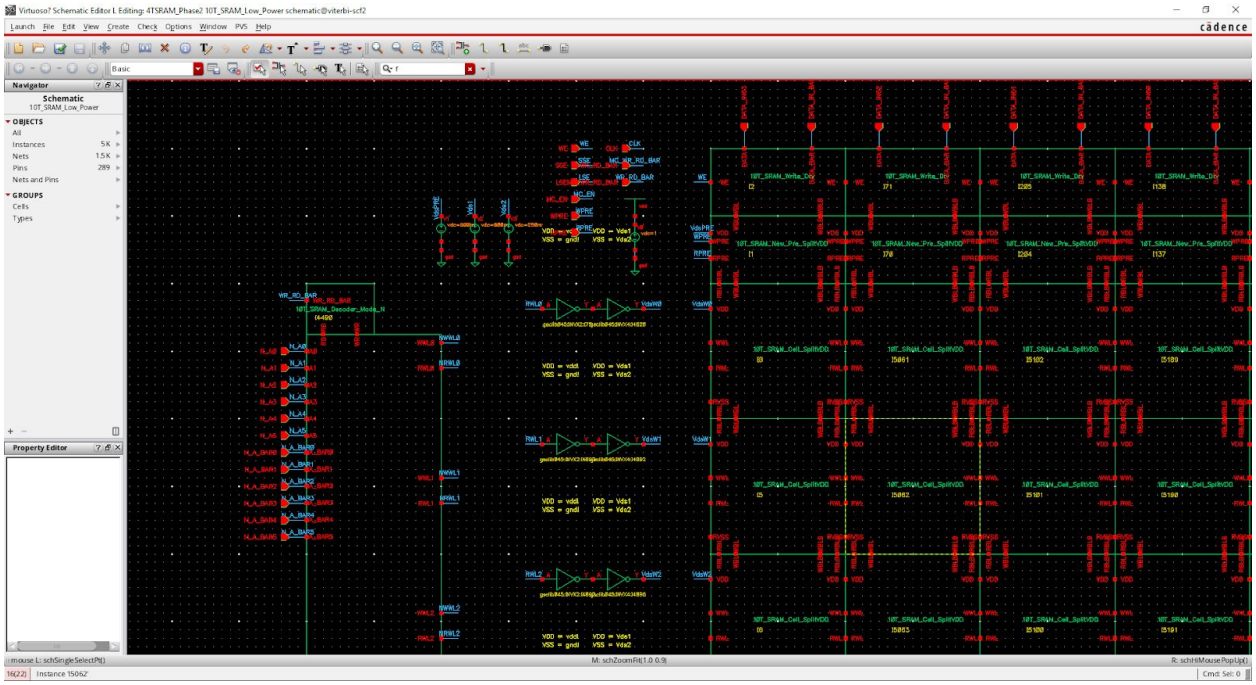

Figure 19 10T SRAM 64x64 Schematic Zoomed-3

Simulation

Due to the power saving techniques, the SRAM operations became slow. The decoder delay is observed as 164ps. Write delay, read delay and memory computation delay are observed as 586ps, 615ps and 664ps, respectively. This result is also affected by the vector file since we couldn't manage to time the signals appropriately due to 1 hour each simulation time.

Figure 20 64x64 SRAM write read operations waveform

Figure 21 64x64 SRAM write read operations waveform annotated

8. Low Power Calculation

Description

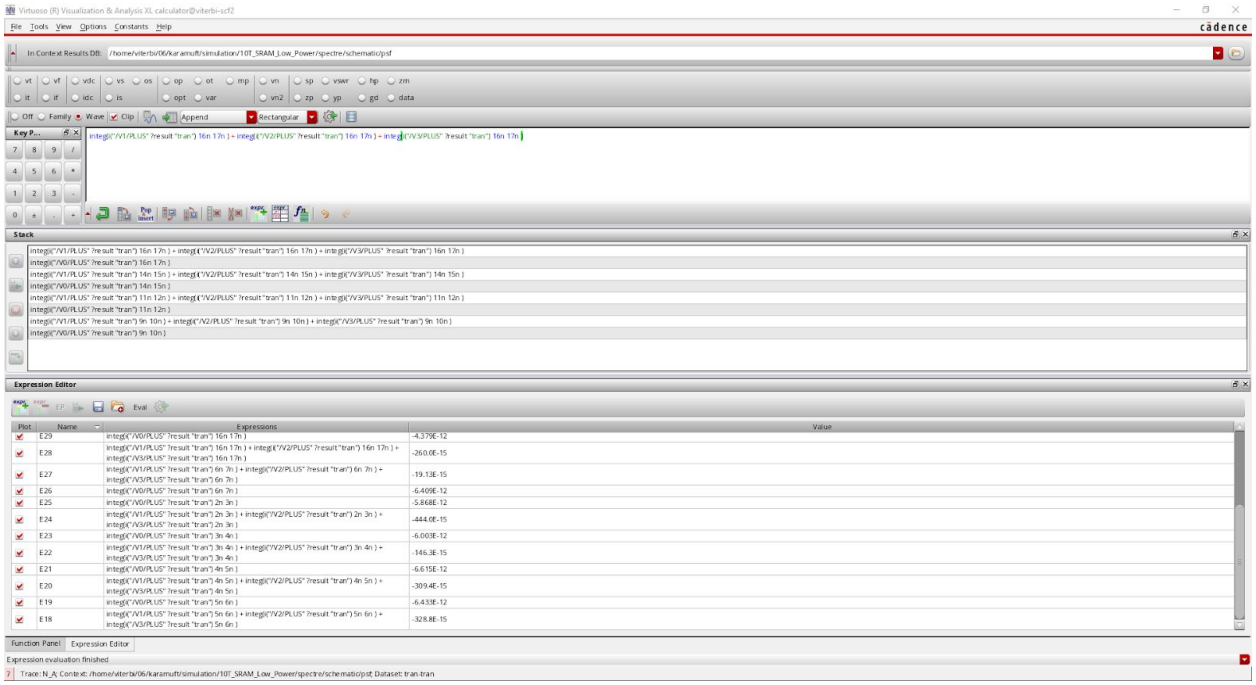

Figure 22 low power

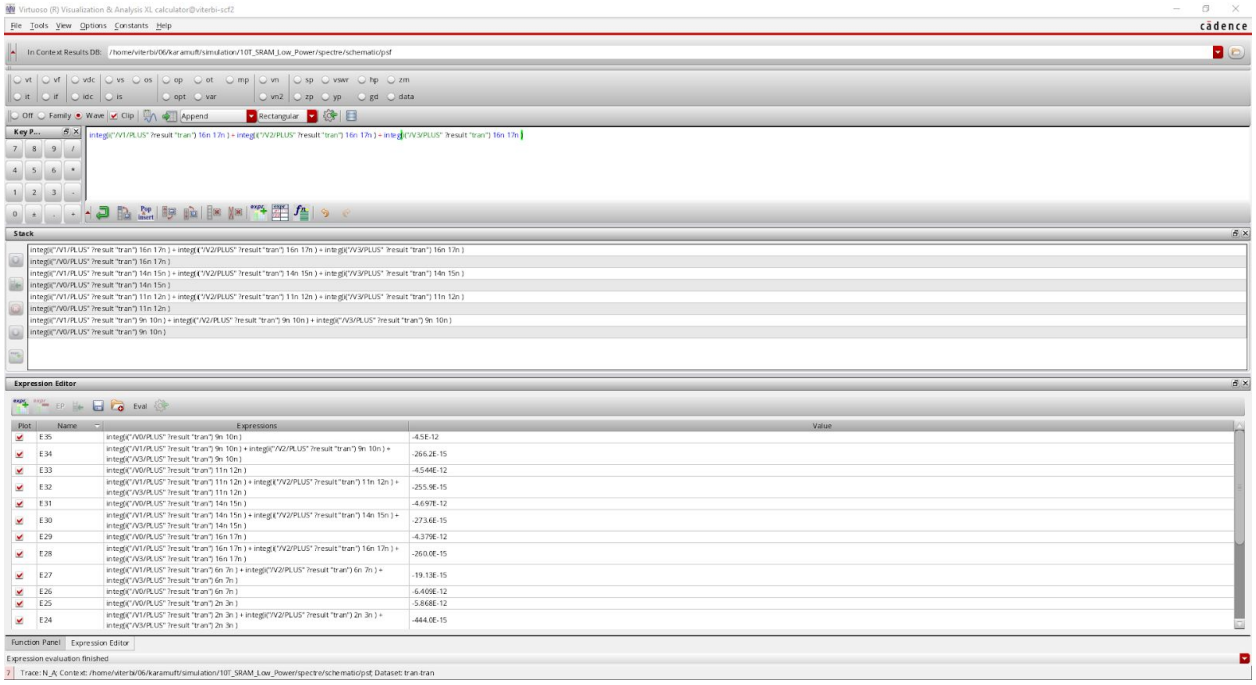

Figure 23 low power

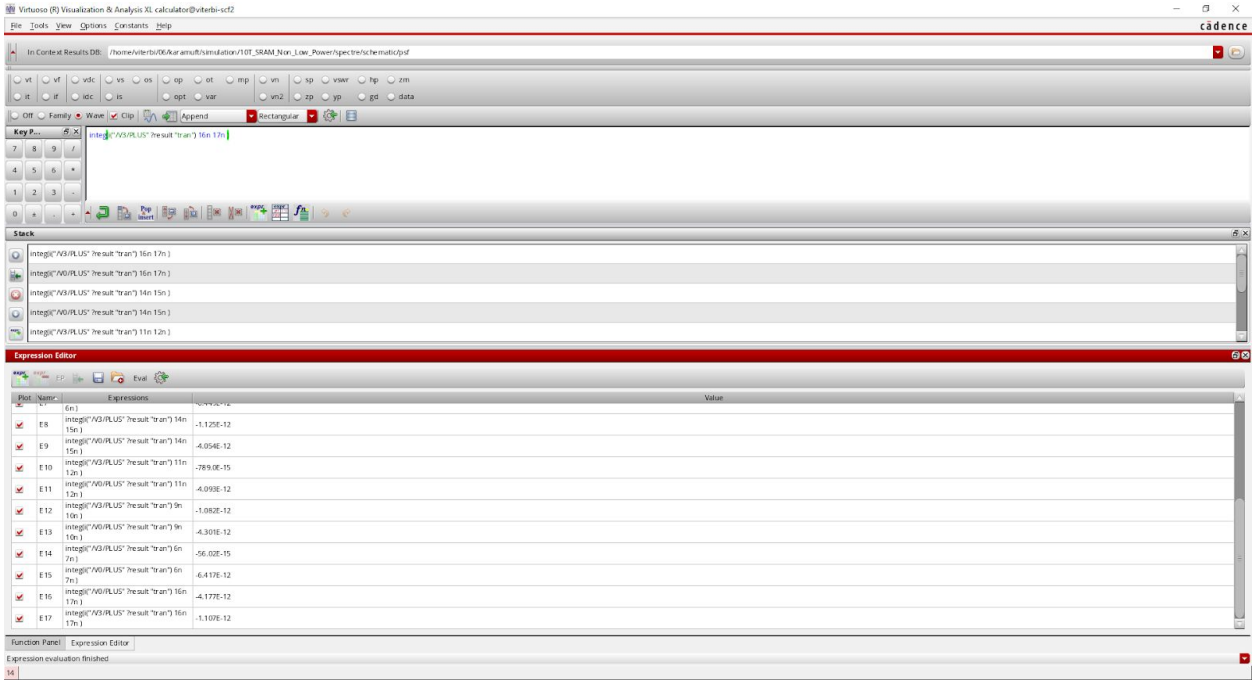

Figure 24 No low power

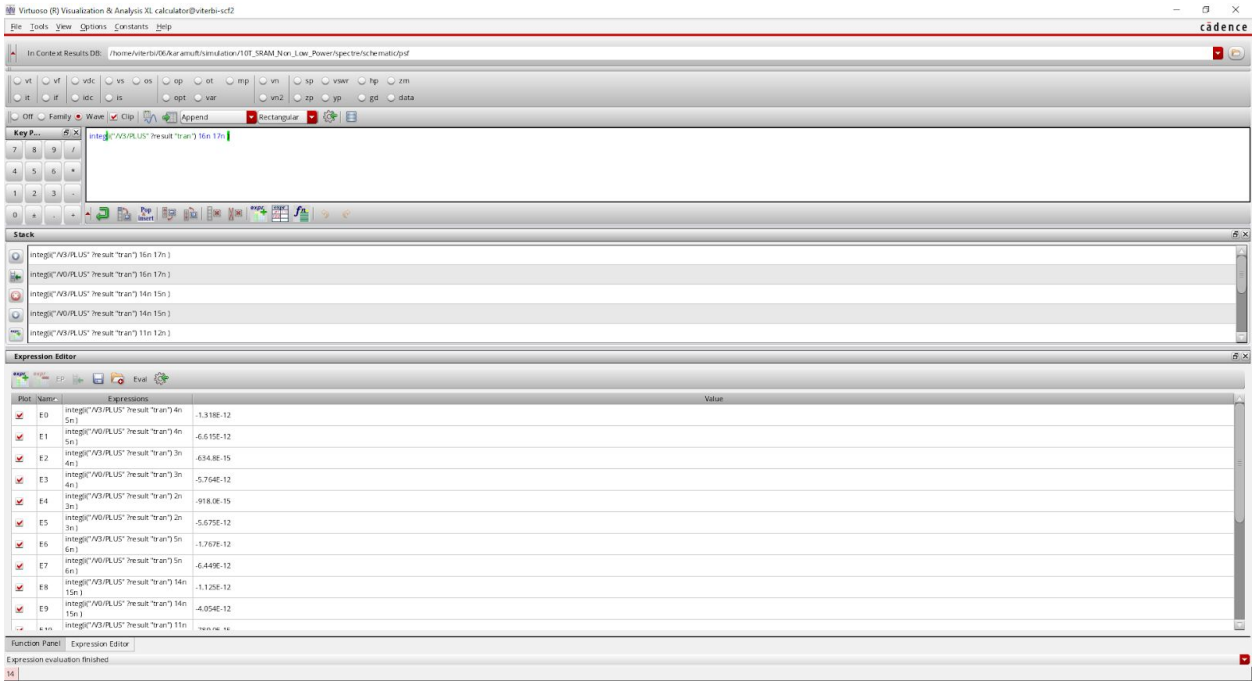

Figure 25 No low power

Operation	Previous Peripheral Energy	Previous SRAM + Precharge Circuit Energy	Previous Total	New Peripheral Energy	New SRAM + Precharge Circuit Energy	New Total
Read 5	-5.675E-12	-918.0E-15	-6.593E-12	-5.868E-12	-444.0E-15	-6.312E-12
Read A	-5.764E-12	-634.8E-15	-6.399E-12	-6.003E-12	-146.3E-15	-6.149E-12
Average Read	-5.720E-12	-776.4E-15	-6.496E-12	-5.936E-12	-295.2E-15	-6.231E-12
Write 5	-6.615E-12	-1.318E-12	-7.933E-12	-6.615E-12	-309.4E-15	-6.924E-12
Write A	-5.675E-12	-1.767E-12	-7.442E-12	-6.433E-12	-328.8E-15	-6.762E-12
Average Write	-6.145E-12	-1.543E-12	-7.688E-12	-6.524E-12	-319.1E-15	-6.843E-12
Idle	-6.417E-12	-56.02E-15	-6.473E-12	-6.409E-12	-19.13E-15	-6.428E-12
MC1 (0 XOR 0)	-4.301E-12	-1.082E-12	-5.383E-12	-4.500E-12	-266.2E-15	-4.766E-12
MC2 (0 XOR F)	-4.093E-12	-789.0E-15	-4.882E-12	-4.544E-12	-255.9E-15	-4.800E-12
MC3 (F XOR 0)	-4.054E-12	-1.125E-12	-5.179E-12	-4.697E-12	-273.6E-15	-4.971E-12
MC4 (F XOR F)	-4.177E-12	-1.107E-12	-5.284E-12	-4.379E-12	-260.0E-15	-4.639E-12
Average MC	-4.156E-12	-1.026E-12	-5.182E-12	-4.530E-12	-263.9E-15	-4.794E-12

As shown in the waveform, the clock frequency is assigned as 1 GHz. Therefore, if we want to find the average power, we need to divide the energy given above by 1xE-9. As a result, we will obtain the power and calculate the average power consumption. By following this, the average power consumption can be found at mW levels. As it can be seen from the table above, we have decreased the overall power consumption but the values are dominated by the peripheral circuits' power consumption.

Pins

We have divided the related control inputs and the DFF outputs of normal and in-memory read operations. The related information is given in Table 1.

Table 2: Summary of all the I/O signals

Input Pins		
Decoder Inputs	N_A<0:5>	N_A_BAR<0:5>
Memory Computation Decoder Inputs	MC_A<0:5>	MC_A_BAR<0:5>
Input Data	DATA_IN<0:63>	DATA_IN_BAR<0:63>
In-Memory Enable	MC_EN	1 to activate
In Memory-Write Read Bar	MC_WR_RD_BAR	1 to Write, 0 to Read
Write Read Bar	WR_RD_BAR	1 to Write, 0 to Read
Read Precharge	RPRE	0 to activate
Write Precharge	WPRE	0 to activate
Sense Amplifier Precharge	SSPRE	0 to activate
Write Enable	WE	1 to activate
Read Enable	SSE	1 to perform
VDD	vdd!	1
VSS	gnd!	0
VBN	gnd!	0
VBP	vdd!	1
Clock	CLK	0-1
Output Pins		
Output Data-Small Sense	SS_DATA_OUT<0:63>	
Output Data-Large Sense	LS_DATA_OUT<0:63>	

9. Vector Files

```
radix 1
io i
vname CLK
slope 0.005
period 0.5
vih 1
vil 0

0
1
0
1

0
1
0
1

0
1
0
1

0
1

0
1
0
1

0
1

0
1

0
1
0
1

0
1

0
1

0
1
0
1

0
1

0
1
```

```
radix 1 1 1 1 1
io i i i i i
vname WE WR_RD_BAR MC_WR_RD_BAR MC_EN LSE
slope 0.005
period 0.5
vih 1
vil 0

0 1 1 0 0
1 1 1 0 0
0 1 1 0 0
1 1 1 0 0

0 0 0 0 0
0 0 0 0 0
0 0 0 0 0
0 0 0 0 0

0 1 1 0 0
1 1 1 0 0
0 1 1 0 0
1 1 1 0 0

0 0 0 0 0
0 0 0 0 0

0 1 1 0 0
1 1 1 0 0
0 1 1 0 0
1 1 1 0 0

0 0 0 1 1
0 0 0 1 1

0 1 1 0 0
1 1 1 0 0

0 0 0 1 1
0 0 0 1 1

0 1 1 0 0
1 1 1 0 0
0 1 1 0 0
1 1 1 0 0

0 0 0 1 1
0 0 0 1 1

0 1 1 0 0
1 1 1 0 0

0 0 0 1 1
0 0 0 1 1

0 0 0 0 0
0 0 0 0 0
```

```
radix 1 1
io i i
vname WPRE RPRE
slope 0.005
tunit ns
vih 1
vil 0

0 0 1
0.15 1 1
1.2 0 1
1.35 1 1

2.2 1 0
2.35 1 1
3.2 1 0
3.35 1 1

4 0 1
4.15 1 1
5.2 0 1
5.35 1 1

6 1 1
6.15 1 1

7 0 1
7.15 1 1
8.2 0 1
8.35 1 1

9.2 1 0
9.35 1 1

10 0 1
10.15 1 1

11.2 1 0
11.35 1 1

12 0 1
12.15 1 1
13.2 0 1
13.35 1 1

14.2 1 0
14.35 1 1

15 0 1
15.15 1 1

16.2 1 0
16.35 1 1

17 1 1
17.15 1 1
```

```
radix 1
io i
vname SSE
slope 0.005
vih 1
vil 0

0 0
0.5 0
1 0
1.5 0

2 0
2.5 1
2.65 0
3.5 1
3.65 0

4 0
4.5 0
5 0
5.5 0

6 0
6.5 0

7 0
7.5 0
8 0
8.5 0

9.5 1
9.9 0

10 0
10.5 0

11.5 1
11.9 0

12 0
12.5 0

13 0
13.5 0

14.5 1
14.9 0

15 0
15.5 0

16.5 1
16.9 0

17 0
17.5 0
```

Figure 26 CLK, Enable, Precharge, SSE

```
radix 11 1111 11 1111 11 1111 11 1111 4444 4444 4444 4444 4444 4444 4444 4444
io i i i i i i i i i i i i i i i i
vname N_A[5:4] N_A[3:0] N_A_BAR[5:4] N_A_BAR[3:0] MC_A[5:4] MC_A[3:0] MC_A_BAR[5:4] MC_A_BAR[3:0] DATA_IN[63:48] DATA_IN[47:32] DATA_IN[31:16] DATA_IN[15:0] DATA_IN_BAR[63:48] DATA_IN_BAR[47:32] DATA_IN_BAR[31:16] DATA_IN_BAR[15:0]
slope 0.005
vih 1.0
vil 0.0

0 00 0000 11 1111 00 0000 11 1111 5555 5555 5555 5555 AAAA AAAA AAAA AAAA
1 00 0001 11 1110 00 0001 11 1110 AAAA AAAA AAAA AAAA 5555 5555 5555 5555

2 00 0000 11 1111 00 0000 11 1111 AAAA AAAA AAAA AAAA 5555 5555 5555 5555
3 00 0001 11 1110 00 0001 11 1110 AAAA AAAA AAAA AAAA 5555 5555 5555 5555

4 00 0000 11 1111 00 0000 11 1111 AAAA AAAA AAAA AAAA 5555 5555 5555 5555
5 00 0001 11 1110 00 0001 11 1110 5555 5555 5555 5555 AAAA AAAA AAAA AAAA

6 00 0001 11 1110 00 0001 11 1110 5555 5555 5555 5555 AAAA AAAA AAAA AAAA

7 00 0000 11 1111 00 0000 11 1111 0000 0000 0000 0000 FFFF FFFF FFFF FFFF
8 00 0001 11 1110 00 0001 11 1110 0000 0000 0000 0000 FFFF FFFF FFFF FFFF

9 00 0000 11 1111 00 0001 11 1110 0000 0000 0000 0000 FFFF FFFF FFFF FFFF

10 00 0001 11 1110 00 0001 11 1110 FFFF FFFF FFFF FFFF 0000 0000 0000 0000

11 00 0000 11 1111 00 0001 11 1110 FFFF FFFF FFFF FFFF 0000 0000 0000 0000


12 00 0000 11 1111 00 0000 11 1111 FFFF FFFF FFFF FFFF 0000 0000 0000 0000
13 00 0001 11 1110 00 0001 11 1110 0000 0000 0000 0000 FFFF FFFF FFFF FFFF

14 00 0000 11 1111 00 0001 11 1110 0000 0000 0000 0000 FFFF FFFF FFFF FFFF

15 00 0001 11 1110 00 0001 11 1110 FFFF FFFF FFFF FFFF 0000 0000 0000 0000

16 00 0000 11 1111 00 0001 11 1110 FFFF FFFF FFFF FFFF 0000 0000 0000 0000

17 00 0000 11 1111 00 0001 11 1110 FFFF FFFF FFFF FFFF 0000 0000 0000 0000
```

Figure 27 DATA Vector file

Specs

10T-SRAM

Sizing

BL read:	**120nm (hvt)**
BLB read:	**120nm (hvt)**
BL access:	**120nm (hvt)**
BLB access:	**120nm (hvt)**
Q pull-down:	**120nm (hvt)**
Q_bar pull-down:	**120nm (hvt)**
Q pull-up:	**120nm (hvt)**
Q_bar pull-up:	**120nm (hvt)**

Layout

Layout Area(New):	**2.510um^2 (0.91um x 2.885um)**
Layout Area(OLD):	**2.788 um^2 (0.87um x 3.205um)**
Layout Area reduction:	**5.85%**

Write Driver

Sizing

All NMOS:	**240nm**

Precharge Circuits

Sizing

All PMOS:	**180nm**

Small Sense Amplifier

Sizing

Precharge PMOS:	**150nm**
Small Sense PMOS:	**150nm**
Small Sense NMOS:	**150nm**
Asymmetric NMOS:	**120nm**
Latch NAND gates:	**x1**

Sense Amplifier Layout

Layout Area: 1.814um^2 (0.87um x 2.085um)

Pull-Down Inverter for each row

Sizing

Inverters:	**x4**

Pull-Up Inverter for each row

Sizing

Inverters:	**x2, x4**

Decoder Modes

Sizing

Inverters:	**x4**
NAND gates:	**x4**
Buffer:	**x2**

Row Decoder

Sizing

NAND gates:	**x1**
Inverters:	**x1**
NOR gates:	**x1**

DFF

Sizing

DFFQ :	**x1**

Delay Circuit

Sizing

DLY :	**x2**