ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 7, Issue 4, April 2018

AREA- EFFICIENT, LEAKAGE SUPPRESSED, POWER OPTIMIZED POWER GATING SCHEME FOR SINGLE RAIL SRAM MEMORIES

[1]SUMITH K S ,[2] SENTHILKUMARAN V
[1]Pg scholar, Department of ECE, Mahendra Engineering College, Namakkal District, T.N, India
[2]AP, Department of ECE, Mahendra Engineering College, Namakkal District, T.N, India

Abstract-Low power supply operation with leakage power reduction is the prime concern in modern nano-scale CMOS memory devices. In the present scenario, low leakage memory architecture becomes more challenging, as it has 30% of the total chip power consumption. Since, the SRAM cell is low in density and most of memory processing data remain stable during the data holding operation, the stored memory data are more affected by the leakage phenomena in the circuit while the device parameters are scaled down. Reducing the leakage power in embedded SRAM memories is critical for low-power applications. Raising the source voltage of SRAM cells through diode transistor in standby mode reduces the leakage currents effectively. However, in order to preserve the state of the cell in standby mode, the source voltage cannot be raised beyond a certain level. To achieve that, the size of the required diode transistor becomes larger, as the supply voltage shrinks in the nano-CMOS technologies. In this thesis, an area efficient power gating technique is presented. Proposed scheme reduces the area overhead by 1-5% compared to conventional schemes, when applied to a 64Kb SRAM macro at 28nm CMOS technology at 0.85V supply voltage. This thesis explores the design and analysis of 64Kb SRAM, focusing on optimizing delay and power in 28nm planer high-K gate last process technology. Various low power techniques such as divided wordline architecture, binary tree decoder structure, self timed clocking using tracking circuits to limit bitline and I/O line swings are studied and designed. High speed techniques were inherently obtained by the process along with implementation of butterfly architecture. Floor plan and layout was done (addressing the constraints on form-factor of each sub-block) in an efficient way, meeting the Mentor Caliber DRC and LVS specifications for the TSMC foundry.

I.INTRODUCTION

To accomplish high-density chip, ultra-low power dissipation, and high performance, complementary metal oxide semiconductor (CMOS) devices have been scaled since last 30 years. As a result, the propagation delay time has been reduced by 30% per technology leading to the microprocessor performance being doubled every two years. Scaled technology has reduced the supply voltage to obtain low power consumption. Additionally, a scaled technology also has reduced device parameters such as threshold voltage, channel length and gate oxide thickness. However, the scaled technology has two drawbacks. First, a low-VTH device has an exponential increase in sub-threshold leakage. Sub-threshold leakage rises by ten times for every 0.1-volt decrease of the threshold voltage. The second problem is the reduction of worst-case performance due to threshold variation at lower supplies. As technology scales down, leakage current in a sub-micron region becomes more significant and is comparable with the dynamic power dissipation. Figure 1.1 shows the full chip leakage power dissipation based on the international technology roadmap for semiconductor (ITRS). Various components affecting the sub-threshold leakage, gate leakage, and junction leakage are depicted in Figure 1.2. However, finding and modelling of the several leakage mechanisms are essential for evaluation and minimization of leakage current for low power application.

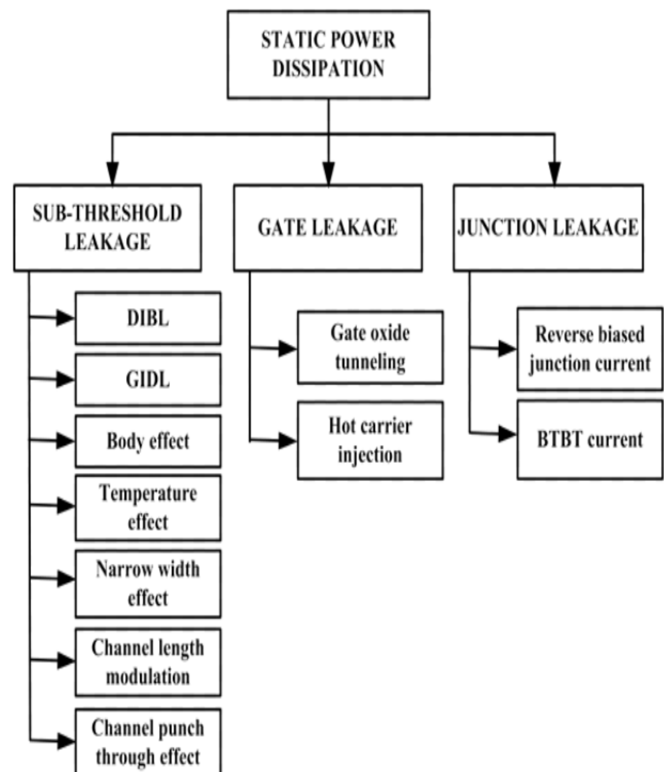

Figure 1.1 Leakage currents components

Generally, device non-conducting current (IOFF) depends on the supply voltage, threshold voltage, length of the channel, surface/channel doping profile, drain/source junction depth and gate oxide thickness. For long channel devices IOFF mainly originates from the drain-source

All Rights Reserved © 2018 IJARECE

reverse bias junctions. Short-channel device needs low power supply in order to reduce power dissipation. Hence, the reduced threshold voltage causes exponential increase in IOFF current due to the weak-inversion region.A conventional 6T SRAM cell consists of two inverters connected back to back and two access NMOS transistors as shown in Figure 1.2.

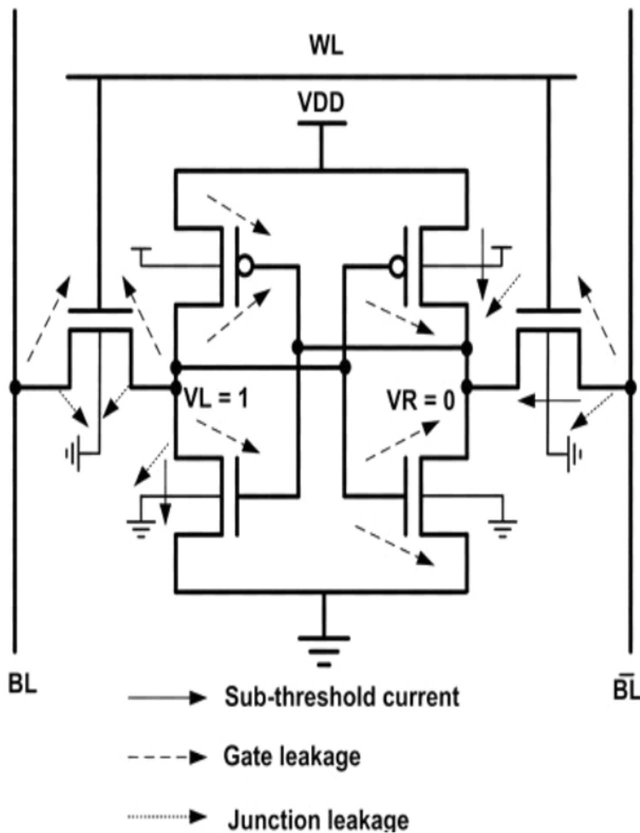

Figure 1.2 Leakage phenomena in basic SRAM cell

II.SRAM ARCHITECTURE AND SPECIFICATIONS

The following chapter provides a brief explanation of the 64kb SRAM architecture being implemented, its specifications description and basic operation.

A.SRAM Architecture

The architecture of a SRAM includes 1024X64 memory cells with control circuitry to decode addresses, and IO circuits to implement the required read and write operations. Figure 2.1 shows the block diagram of designed SRAM.

To design 64kb memory the basic requirements for no rows and no of columns should be multiple integer of 2^x if this condition is not meeting then the memory configuration invalid.

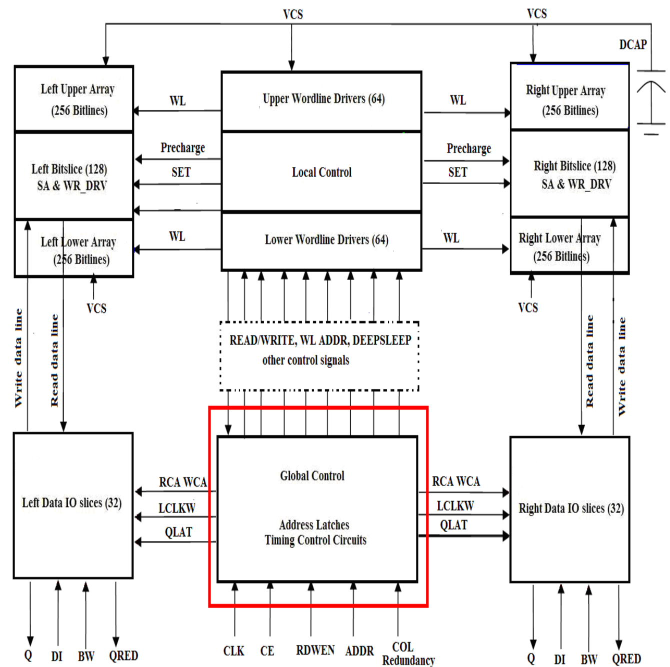

Figure 2.1. SRAM block diagram

B.SRAM Top Level Specifications

The SRAM was designed for the specification shown in the following Table 1.

Table 1 SRAM Specifications

Basic Operation	Write-Read
Storage Mode	Volatile
Access Mode	Random
Storage Cell Operation	Static
Storage Capacity	64Kb, Mux 8
Operating frequency	1GHz
Redundancy	Single column
Power Supply	Vdd - 0.7-0.95 V (peripheral circuits),
Array architecture	M2 bitlines, M4 global bitlines, M3 wordlines
Technology	28nmplaner high-K gate last process
Performance	Ultra High Density ,High Speed, Low-Power, High-Reliability
Environmental Tolerance	Commercial, Space, Military, High Temperature
Read Effect	Nondestructive
Logic System	Binary
Application	Networking

All Rights Reserved © 2018 IJARECE

C. Leakage current in 6T CMOS SRAM cells

The data retention current of SRAM cells has been dramatically increasing as the device sizes are shrinking and thus resulting in high power consumption. The dominant leakage mechanisms in a 6T SRAM cell in CMOS nano-meter technologies are sub-threshold leakage and gate leakage. Since last few years, gate leakage has been well controlled at the process level through the use of HIGH-K metal gates . Sub-threshold leakage is still a challenge for low power SRAMs and the transistors which dissipate this leakage in a 6T SRAM cell are shown in Figure.2.2. Scaling the cell bias voltage by raising the source voltage of SRAM cells reduces the sub threshold leakages. Raising the source voltage can be achieved by ground-gating the cells using a sleep transistor.

In general, the source voltage should be raised as much as possible so that maximum leakage reduction can be achieved. However, in order to preserve the state of the cells in the standby mode, this voltage must not exceed a certain level.With the technology scaling down to smaller geometries, the exacerbated variation of device parameters causes significant inter-die and intra-die variations which lead to instability in the SRAM cells.

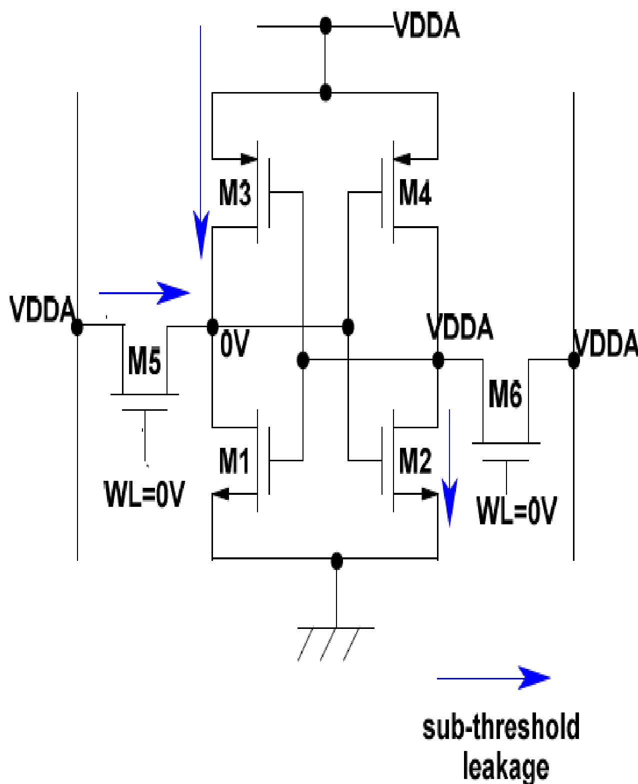

Figure.2.2 Leakage current in 6t CMOS SRAM cells

D.Obstacles in SRAM Scaling

There are, however, obstacles on the way of continuous scaling of SRAM. One of them is that SRAM cell power delay-area product is not scaling as efficiently as that of logic circuits. This phenomenon is known as the non-scaling problem of SRAM. Multiple effects of device dimensions scaling over technology generations yielded faster transistors. However, the saturation current of the scaled transistors often remained on the previous level. The read current of an SRAM cellIreaddepends on the saturation current of the access transistor and the speed often depends on the small-signal slew rate as :

$$t\Delta VBL = (CBL\Delta VBL)/ Iread \quad (1.1)$$

Equation 1.1 shows that if Iread remains at the previous level, then SRAM speed can be improved by reducing the bit line differential developed during a read operation ΔVBL and/or the capacitance of the bit lines CBL. The bit line capacitance CBL, which consists of the diffusion and wire capacitance components, only partially benefits from scaling of the transistor dimensions. The wire capacitance component that is responsible for at least a half of the total bit line capacitance CBLis not scaling down at the same rate as the speed of the logic circuits. The voltage differential required on the bit lines for reliable sensing ΔVBLis not scaling at the same rate as the logic speed. Moreover, the ΔVBL may not scale down and can even increase with the process technology scaling to allow for larger tolerances and margins due to the increasing process variability and difficulty of parameter matching in nano-scaled technologies, which also tend not to scale down.

While the non-scaling problem can be addressed by the modifications to the SRAM architecture, such modifications often involve certain tradeoffs. The extra logic and bypass paths that support these architectural workarounds come at a cost of extra chip area and power dissipation. Another approach reduces the bit line capacitance CBL seen by any given cell by reducing the number of cells per a bit line or introducing a hierarchical bit line structure with global and local bit lines. However, the area and power penalty paid for applying this strategy has reached its practical limits.

III. GENERAL DESIGN TECHNIQUES OF HIGH SPEED, HIGH DENSITH, LOW POWER SRAM

In subsequent sections the salient design of SRAM cell, sense amplifier, control circuit, wordline driver and I/O circuit is discussed. The key aspect of this chapter is to give intricate details of the design techniques involved.

A. Low Power Design Techniques

To reduce the power consumption the techniques used are:

All Rights Reserved © 2018 IJARECE

- SRAM partitioning: For large SRAMs, significant improvements in delay and power can be achieved by partitioning the cell array into smaller sub arrays, rather than having a single monolithic array. Typically, a large array is partitioned into a number of identically sized sub arrays (commonly referred to as macros), each of which stores a part of the accessed word, called the sub word, and all of which are activated simultaneously to access the complete word. The macros can be thought of as independent RAMs, except that they might share parts of the decoder.

Each macro conceptually looks like the basic structure shown in Figure 2.1. During an access to some row, the word line activates all the cells in that row and the desired sub word is accessed via the column multiplexers. This arrangement has two drawbacks for macros that have a very large number of columns: the word line RC delay grows as the square of the number of cells in the row, and bitline power grows linearly with the number of columns. Both these drawbacks can be overcome by further sub dividing the macros into smaller blocks of cells using the Divided Word Line (DWL) technique. In the DWL technique the long word line of a conventional array is broken up into k sections, with each section activated independently thus reducing the word line length by k and hence reducing its RC delay by k^2.

The design has 128 rows. To reduce the bitline capacitance the array is divided into two banks. each bank consisting of 64 bitcells per bitline. The design was done taking into consideration of the required differential development on the bitline and bitline bar during read operation which is around 100mV. To reduce the RC delay in the wordline path, Butterfly architecture is implemented, placing 256 bitlines on either side of the row decoder.

- Active power reduction: The decoder is implemented in a tree structure by which only specific paths along the decoder will be active. By using a three-stage decode architecture, the number of transistors, fan-in and the loading on the address input buffers is reduced. As a result, both speed and power are optimized .

- Deep-sleep technique: This technique is used to power down the array, thus saving the dc power consumption.

B.LAYOUT OF 64Kb SRAM

Once the schematic design is done and functionality and margins are verified for different process corners, layout is the next important task in the chip design cycle. In this project, 64Kb SRAM floor-plan and layout is done in 28nm planer high-K gate last process technology, considering different area optimization techniques along with critical signal routing with lesser parasitics. The tool used for the layout design is Cadence Virtuoso Layout Editor. Mentor Calibre LVS and DRC tools are used to verify the correctness of the layout.

C.Array

The bitcell design is used from TSMC foundry. M2 is used for bitlines, M4 for global bitlines and M3 for wordlines. Figure 5.1 shows one segment of a bank, which contains 64X256 bitcells.

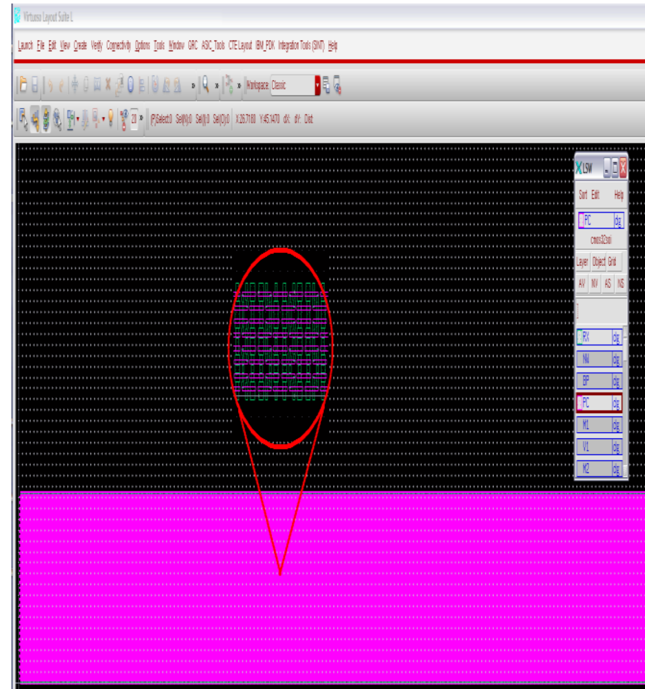

Figure 3.1 One segment layout- 64X256 bit cells

D.Power Gating Implementation

The proposed 'area efficient, leakage supressed, power optimized power gating for single rail SRAM memories' is shown in Figure 3.2. The scheme is applicable to SRAM memories in both the modes whether data retention is important or not. These two modes are explained in the next sections.

1.Light Sleep mode

Light sleep mode is defined as the leakage saving mode where SRAM data needs to be retained, so voltage across SRAM cell should be more than the minimum data retention voltage.

As shown in Figure.3.2.Same transistor (M13) is used as 'on transistor' as well as diode. For illustration, with light sleep mode entry, LS will be logic '1' which makes M9 'off' and M12 'on'. When M12 is 'on', drain and gate of M13 are connected through M12 and form a diode structure. When SRAM comes out of light sleep, LS goes to logic '0', which makes M9 'on' and M12 'off'. LSB node goes to logic '1', M13 is 'on' and VSSC node goes to 0 and thus SRAM goes to active mode.

All Rights Reserved © 2018 IJARECE

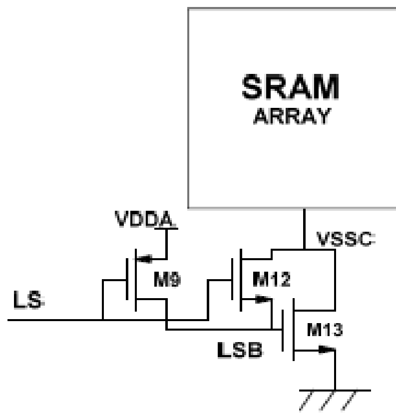

Figure3.2. Diode and ON transistor interchangeable power gating scheme

2.Shut down mode

Shut down mode is defined as the mode where SRAM data need not be retained. So, in this mode, the VSSC level can be raised further to achieve better leakage savings. The logic which controls gate of 'on' transistor (M13), is modified to incorporate shut down mode. The circuit is as shown in Figure.3.3. This circuit supports both light sleep mode as well as shut down mode. The truth table for both the modes is shown in table1. When SD='1', irrespective of LS signal, memory goes to shut down mode and hence leakage savings are more compared to light sleep case. LSD signal is generated from LS and SD in the control block.

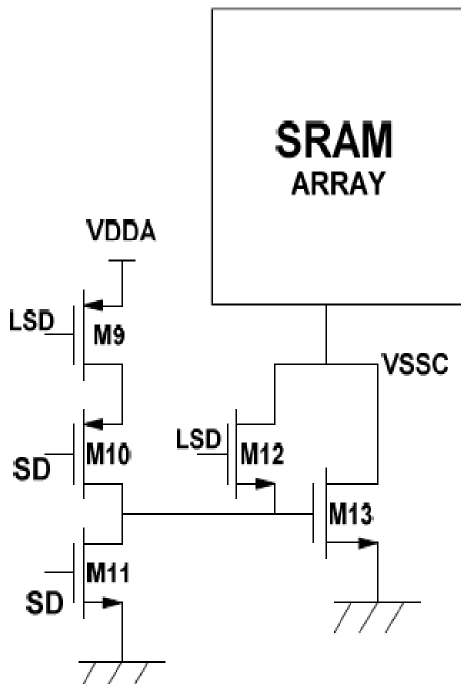

Figure 3.3 SRAM Light sleep and shut down mode

IV.OBSERVATION AND RESULTS

A. Area Comparison

Table 2 Area comparison

Sl. No	Power Gating	X	Y	Area
1	Traditional	65.85	324.645	21377 square micron
2	Proposed	63.39	328.85	20845 square micron

Achieved **2.5 %** area improvement over tradition power gating approach.

B. Power comparisons

Table 3 Power comparison

Sl. No	Corner (Process_Voltage_Temp) fast_125c_0p935v	Power
		Dynamic power (Power Dissipation (uW/MHz) CK Read)
1	Traditional power gating	1.01E+01
2	Proposed approach	1.00E+01

Achieved **1 %** power improvement over tradition power gating approach.

C. Leakage comparisons

Table 4 Leakage comparison

Sl. No	Corner (Process_Voltage_Temp) fast_125c_0p935v	Leakage	
		Static Pwr (uW) (Normal Mode)	Static Pwr (uW) (Light Sleep Mode)
1	Traditional power gating	7.10E+03	5.63E+03
2	Proposed approach	5.82E+03	5.55E+03

Achieved **22 %** leakage improvement over tradition power gating approach in normal mode and **1.2 %** in Light sleep mode

D. 64Kb SRAM Layout

Butterfly architecture is used for layout of 64Kb SRAM. The array is divided into 2 banks. Each bank is in-turn divided into left and right half segments. Each bank size is 64Kb. Figure 4.1 shows the entire 64Kb SRAM layout. Total area of the macro is 20845 square micron. Highest metal used in the design is M4. The layout is MentorCalibre LVS and DRC clean. TSMC foundry guidelines are used for layout.

All Rights Reserved © 2018 IJARECE

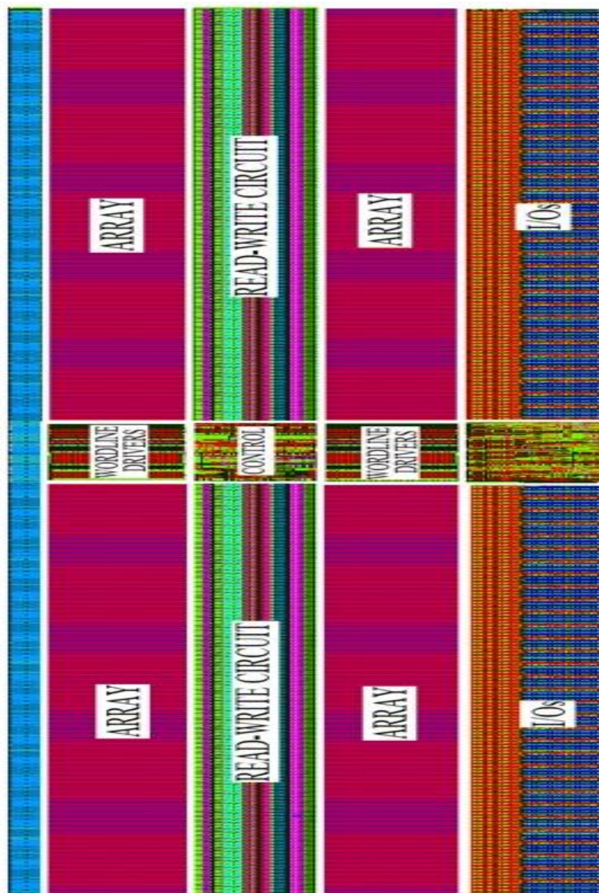

Figure.4.164Kb SRAM Layout

V.CONCLUSION AND FUTURE SCOPE

In this thesis, design techniques for an area efficient, leakage suppressed and power optimized power gating scheme for single rail SRAM were looked at. The key leakage/power optimization is implemented in the SRAM array and area optimization in periphery.Finally the SRAM was designed with above mentioned guidelines and also taking in to account of the possible parasitics from the layout. The simulations were carried out on the schematic netlist, the functionality and different margins were verified. The floor plan and layout was done for the design and was checked for LVS and DRC rules following the guidelines from the DRC deck.The area efficient, leakage suppressed and power optimized power gating scheme for dual rail SRAM technique, FINFET technology, and other novel cell designs will be worth investigating for designing future high performance high capacity RAMs. Also beyond CMOS, a breakthrough is needed into the future. Nano microelectronics also needs to be explored.

REFERENCES

[1]An Area Efficient Diode and On Transistor Interchangeable Power Gating Scheme with Trim Options for Low Power SRAMs ,Ankur Goel1, Donald Evans2, Richard Stephani2, Venkateswara Reddy1, Dharmendra Rai1, Veerabadra Chary1, Sathisha N1.

[2] Area-efficient, high-speed, dynamic-circuit-based sensing scheme for dual-rail SRAM memories,AnkurGoel,Dharmendra Kumar Rai,SumithKaippalathingal Soman.

[3] A. Bhavnagarwala, S. V. Kosonocky, M. Immediato, D. Knebel, andA.-M. Haen, “pico-joule class, 1 GHz, 32 kB × 64 b DSP SRAM with self reverse bias,” in Proc. Symp. VLSI Circuits Dig. Tech. Papers, Jun. 2003, pp. 251–252.

[4] T.Enomoto, Y.Oka and H.Shikano, “Self-controllable voltage level (SVL) circuit and its low power high speed CMOS circuit applications,” IEEE Journal of Solid State Circuits, 38(7):1220- 1226, July 2003.

[5] AnkurGoel and BaquerMazhari. Gate leakage and its reduction indeep submicron SRAM. In proceedings of International Conference onVLSI Design, Jan. 2005.

[6]K. Zhang et al., “SRAM design on 65-nm CMOS technology withdynamic sleep transistor for leakage reduction,” IEEE Journal Solid-State Circuits, vol. 40, no. 4, pp. 895–901, Apr. 2005.

[7] M. Khellah et al., “A 256-Kb dual-VCC SRAM building block in 65-nm CMOS process with actively clamped sleep transistor,” IEEE J.Solid- State Circuits, vol. 42, no. 1, pp. 233–242, Jan. 2007.

[8] Hao-I Yang, Wei Hwang, “Impacts of NBTI/PBTI and ContactResistance on Power-Gated SRAM With High-Metal-Gate Devices,”IEEE Transactions on VLSI systems, VOL. 19, NO. 7, July, 2011.

[9] Pramod Kolar, Eric Karl, Uddalak Bhattacharya, FatihHamzaoglu,HenryNho, Yong-Gee Ng, Yih Wang, Kevin Zhang, “A 32 nm High-kMetal Gate SRAM With Adaptive Dynamic Stability Enhancement forLow-Voltage Operation,” IEEE Journal of Solid State Circuits, VOL.46, NO. 1, Jan. 2011.

All Rights Reserved © 2018 IJARECE