Analysis on Techniques Used to Scale Power on Modern SRAM

Jimin Yoon (jimin2016@berkeley.edu), Gary Choi (gchoi@berkeley.edu)

Abstract — As the scaling of CMOS is coming to its limit, modern digital design problems have shifted focus from improving performance to reducing energy consumed by the chip. Reducing memory power consumption is one of the main focus in energy-efficient digital design. We present two schemes that scales power on ordinary 6-Transistor SRAMS: Using different supply voltage for read/write operations and using sleep transistor to float the rail of the inactive cells. In addition to adding two extra transistors on each SRAM cells, we built SRAM peripherals that controls multiple supply voltage in different operations and extra decoder that allows inactive SRAM sub-arrays to reduce unnecessary leakage when not used. Given lenient initial supply voltage of 1V and secondary supply voltage of 0.7V, proposed techniques achieved roughly 41 to 46% decrease in power during a read operation, 26 to 86% decrease in power during a write operation, and 19 to 63% decrease in power while inactive.

I. INTRODUCTION

Designing low power SRAMs in many different VLSI chips became important as CMOS scaling almost reached its limit and consumer market favors miniature devices such as smart phones and watches. For these mobile devices, keeping the power consumption of the device as low as possible is extremely important because they are supplied by batteries that provide power for temporary amounts of time. Performance of a chip will continue to improve significantly but the battery efficiency will continue its slow growth as shown in Figure 1 which demands low-power design.

In this report, we investigate techniques used to reduce overall power of the ordinary 6-T SRAM design. We will first create 6-T cell model and calculate sizing of the transistors for sufficient read/write stability. In addition to the 6-T cell, we build peripheral circuits such as decoder for the WL access, pre-charge, write driver and sense amps to detect and manipulate BL of SRAM cells. Next, we present our two techniques to reduce power consumed in different memory operations as well as lowering leakage of the inactive SRAM sub-arrays in SRAM arrays. At the end, power reduction compared to the original 6-T model for each technique and two techniques combined will be measured. Tradeoffs for the technique in area and stability will also be analyzed qualitatively in later part of the report.

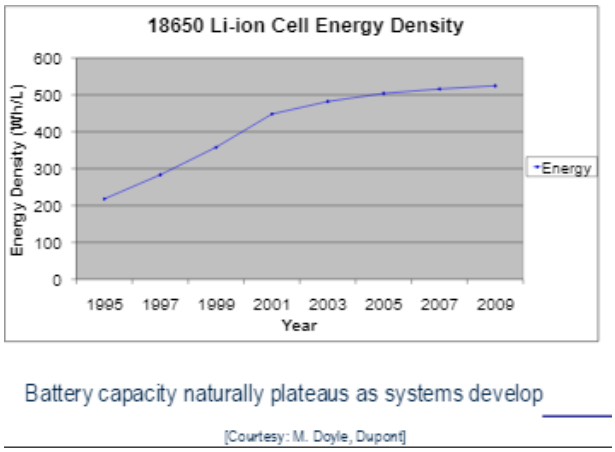

Fig.1 Technology growth in Battery Capacity

II. 6-T SRAM SETUP

We now analyze the periphery circuit of the 6-T SRAM and the setup of transistor sizing in 6-T cell. SRAM read/write stability as well as functionality with full operation will be shown afterwards to show correct setup on our end. As for the peripherals, pre-charge, sense amplifier and write-driver were used to simulate the operations of the 6-T cell. Setup of decoder and word-line logic will be discussed later with the array setup and explanation of the techniques.

A. Pre-charge Circuit

The circuit shown in Figure 2 is responsible for charging the bit line to half of V_{DD} when the cell is inactive. When the PRE signal is low, BL and BLN are

pulled to the VDD input (which is connected to a voltage source that is half of the actual V_{DD} if we want to pre-charge the bit lines to that value) by the two side PMOS's. The middle PMOS is responsible for equalizing the bit lines to $V_{DD}/2$. We want to pre-charge the bit lines such that tiny differences in voltage can be detected by our sense amp during a read and the word driver doesn't have to pull as much current during a write.

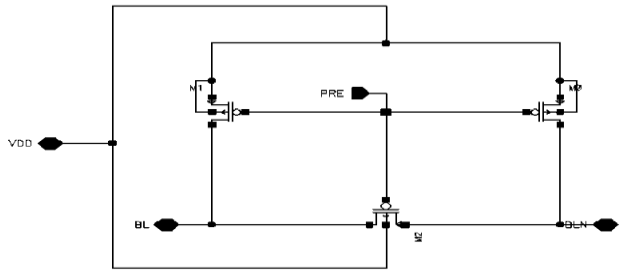

Fig.2 Pre-charge circuit [1]

B. Sense Amplifier

The sense amplifier circuit as shown in Figure 3 is responsible for detecting differences in the bit lines during a read. During a read, the SE signal is pulled high and the SEN is pulled low such that the crossed inverter circuit becomes functional. The crossed inverters are responsible for stabilizing the differences across the bit lines and consequently one of the bit lines will hold the cell contents and the other can be pulled to the logical opposite by an inverter.

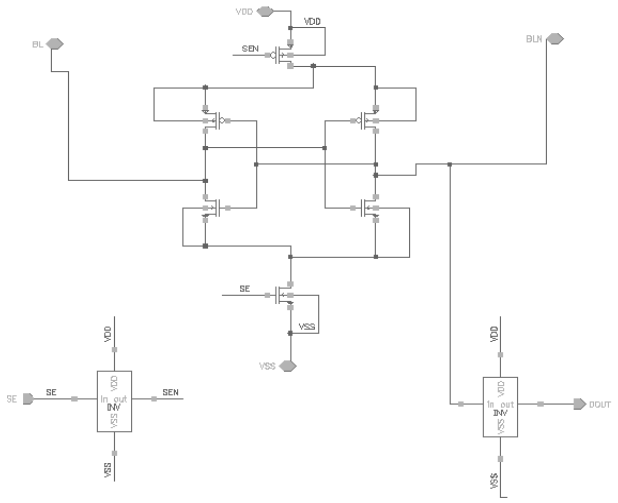

Fig.3 Sense Amplifier

C. Write Driver Circuit

The word driver as shown in Figure 4 is responsible for pulling the bit lines to the respective values that we want to write into the cell. One of the NMOS's will be switched on by the input DATA signal, pulling one bit line to low. This same voltage is applied to the gate of the other bit line's PMOS; the PMOS consequently then begins pulling VDD into the other bit line. The inverter is responsible for the symmetry that allows either bit lines to be pulled low. The sizing of the word driver is crucial in giving it enough drive strength to toggle the cell state.

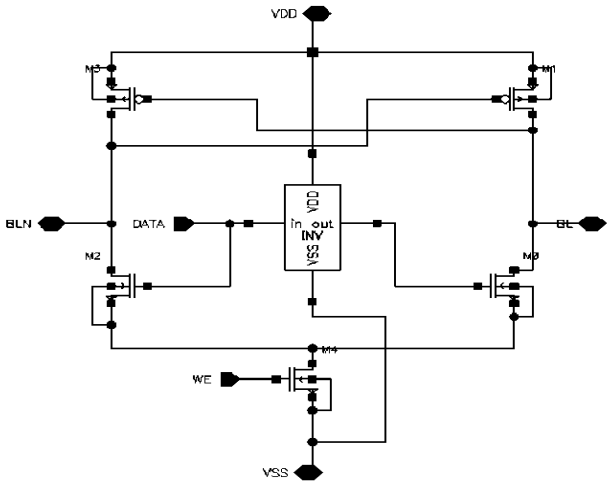

Fig.4 Write Driver circuit

D. 6-T Memory Cell

When choosing the sizing of our 6T SRAM cell, we make a few assumptions and decisions. Firstly, we assume the PMOS minimum width is 100 nm. The PMOS sizing is the base which we use to determine the sizing of the other transistors in the cell. We approximate the threshold for toggling an inverter to be 500 mV so if the state of the cell passes this threshold, its state is flipped. We also decide to aim for a 300 mV read margin when initially deciding the widths.

We initially run a parametric sweep on half of SRAM cells to get a sense of the cell state voltages during a read and write operation as seen in Figure 5. Let Q represent the voltage corresponding to the 0 V state of the cell. Again, assuming the threshold for toggling an inverter is about 500 mV and wanting 300 mV margins we see that Q is 200 mV during a read when the pull down to axis ratio is about 1.6. For writes, to reach 300

mV giving us a 200 mV static write margin the ratio of axis to pull up is about 0.9

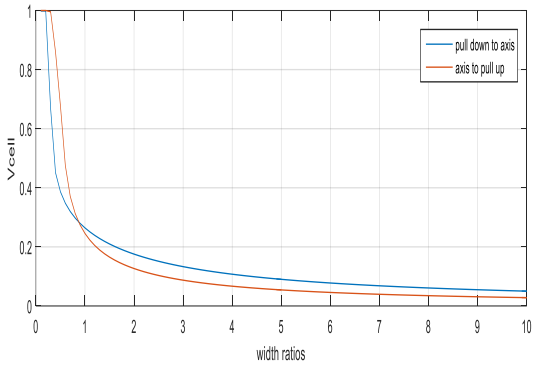

Fig.5 Plot of cell state voltages during read and write operations as widths of pull down and axis are swept in different plots

With our PMOS minimum width of 100 nm and using the proposed sizing, we get that the axis gates are 90 nm wide and the pull down transistors are 144 nm. However, the noise margin we got using these widths was unsatisfactory and the cell consequently didn't look very stable. Since the read margin is lower than expected, we changed the sizing ratio from 1.6 to 2 for pull down to axis and 0.9 to 1 for axis to pull up to give a better read margin. This results in a width 100 nm for the axis transistor and 200 nm for the pull down transistor. This gave us a better read noise margin as shown in Figure 6.

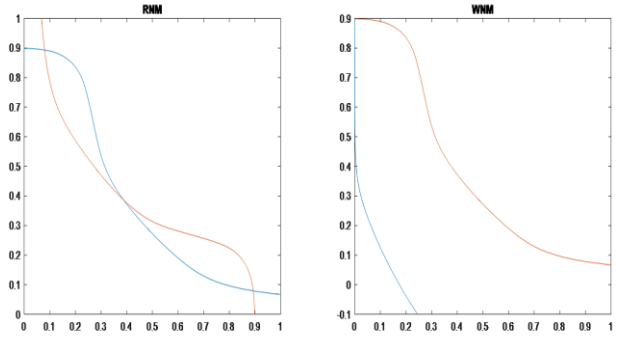

Fig. 6 Read noise margin and write noise margin after changing sizing. The read noise margin is noticeably better than with our previous sizing

We run a simple simulation where we first write a logical high into our cell, read the value, write a logical low into our cell and read the new value. In between each operation we pre-charge the bit lines to half of V_{DD}. Pulse width of the word line is 1ns for both read and write operations. We made the pulse width the same for two different operations because in real designs, it requires complex analog work for main controller to send different width of pulses separate for read and write. According to our simulation, our cell can correctly operate while retaining the correct state during and after the operations.

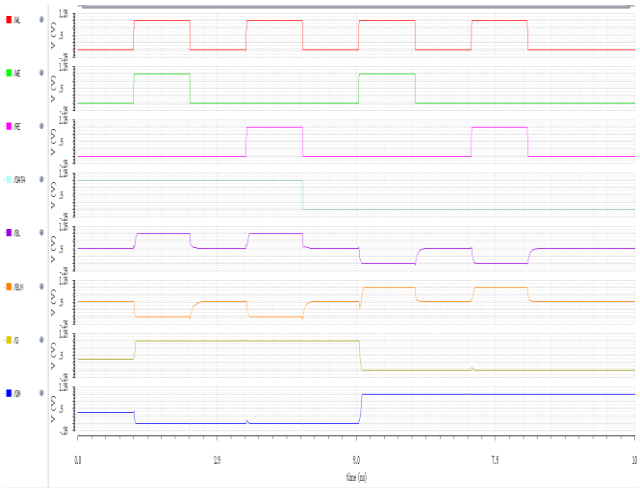

Fig.7 Simple simulation with 2 reads and 2 writes. Bitline is pre-charged to VDD/2. Q is written to VDD, read, written to 0V, then read. From top to bottom: WL, WE, RE, Data, BL, BLN, Q, QN.

III. PROPOSED TECHNIQUES

Original 6-T cell design with constant supply of rails are sufficient for read/write operation functionality. However, there are plenty of design spaces to optimize it to consume less power. We now present two techniques that are applied to the 6-T cell which requires small changes in both cell and peripheries around the SRAM array to reduce power significantly.

A. Dynamic Supply of SRAM cell

One of the obvious modification we can make to reduce power consumption of SRAM cell is to vary the supply voltage. Supply voltage is significant contributor to power in digital design as power is proportional to square of supply voltage. Although reducing supply voltage will lower the speed of the design, when designing memory cells and the overall array, there are already enough overhead in the memory interface and stack of arrays that lowering VDD doesn't heavily affect the timing. Furthermore, delay is growing linearly while power consumed scales down in square factor.

For number of variable supply voltages, we will choose two values VDDs since effect of power reduction saturates if there are too many different supplies. With the initial supply of 1V we used, second VDD will be set to 0.7V as optimum ratio between different supplies are around 0.7 for greater power

reduction. [2] Second voltage supply of 0.7V can be built using simple digital voltage regulator. However, we will not discuss in depth on creating an efficient voltage regulator but take its extra tradeoffs into account such as area and extra active current for the regulator. Furthermore, normal digital design uses second supply for other parts of the chip outside of memory which we can use to supply the dynamic voltage control.

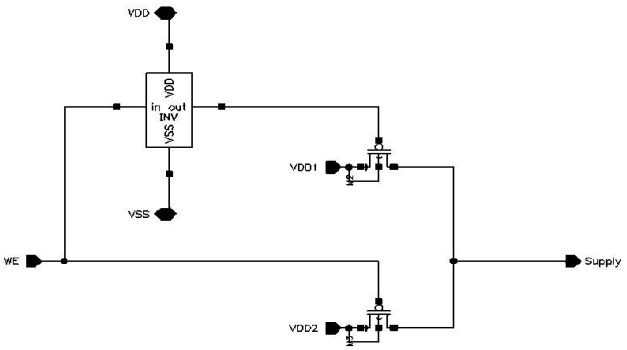

Fig. 8 Voltage Control for SRAM Cell Supply

SRAM cell can be designed in a way that it constantly uses second VDD of 0.7V. However, given our sizing setup we had in previous section, there is a possibility that read operation can fail. Read stability is improved as drive strength of the NMOS of inverters is increased relative to the access NMOS transistor. Lowering the supply voltage would definitely hurt the read operation since there are less current flowing to the NMOS of the inverter in a cell. In contrast, lowering the supply of the cell will improve write stability as drive strength ratio of NMOS transistor to PMOS of the inverter is increased. When the cell is inactive, we can still keep the supply of the cell to be lower than standard 1V as it will store the voltage above 0.5V (0.7V stored for 0.7V supply) when writing a '1'. Therefore, we have a voltage control for 6-T cell supply that provides 1V for read operation and 0.7V rest of the time. [3] Figure 8 is a circuit for providing dynamic voltage supply using a simple PMOS switches controlled by Read-Enable signal. When RE is high, PMOS1 is turned on to feed VDD1 of 1V. Otherwise, VDD2 of 0.7V will be provided to the supply of the cells.

We do the same simulation as we did for the normal 6-T cell. Value of '1' is written and read and then value of '0' is written and read. When '1' is written to cell, 0.7V is written to Q while Q is pulled up to 1V when read. Cell keeps the voltage value of 0.7V to store logic equivalence of '1' when inactive. Output to the decoder which is inverted signal of BLN, is correctly outputting voltage of 1V when reading '1' in Figure 9.

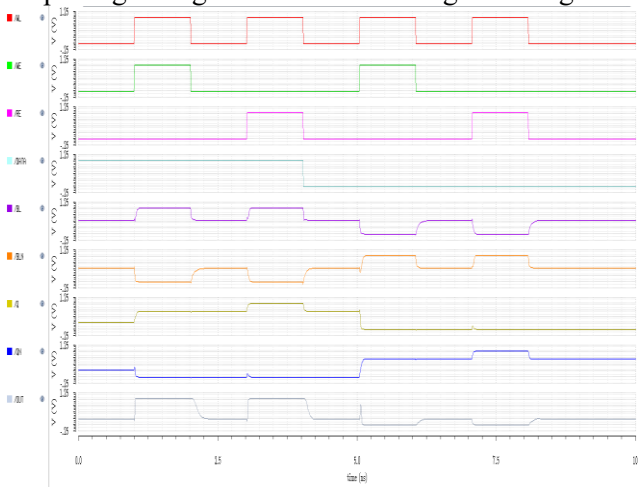

Fig.9 Simple simulation with 2 reads and 2 writes. Bitline is pre-charged to VDD/2. Q is written to VDD, read, written to 0V, then read. From top to bottom: WL, WE, RE, Data, BL, BLN, Q, QN, Output to Decoder.

B. Sleep Transistor with Clamp Diode

Another method of reducing a power consumption of SRAM is to reduce leakage current when cell is inactive which is part of significant portion of total power burned in SRAM. To reduce leakage current, we implement a clamping diode and sleep transistor on each cell as shown in Figure 10. When M7 is on, it serves as a small resistance and V_{SL} is pulled to ground and the SRAM cells functions as normal. However, when it is off, V_{SL} is raised to higher floating voltage, reducing sub-threshold and gate leakage. To prevent the floating value from fluctuating from noise, a PMOS is added in parallel with M7. If some noise source manages to increase V_{SL}, the diode turns on and pulls it back down to a lower voltage V_D so that the cell may remain stable [4]. One design choice is whether to turn on M7 when active or inactive. If we choose to turn it off while inactive, we'll have a floating voltage while the cell is inactive resulting is lower leakage when the cell isn't being accessed. On the other hand, having the higher V_{SS} during an active operation lowers the power of the operation. Our simulation reflected this result although we observed that with our test set up and transistor sizes, the tradeoffs were not very even so we chose to focus on turning M7 off while inactive. In addition, the floating voltage is dependent on the characteristics of the sleeping diode and is another design space that can be explored.

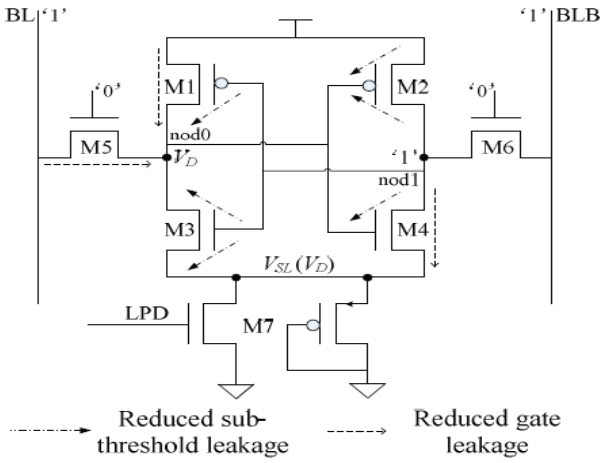

Fig.10 SRAM cell with Clamping Diode

We made this change to our SRAM cell by sizing both M7 and M8 to 100 nm and confirmed that the V_{SS} source of the cell is indeed raised to a higher voltage with the clamp diode as shown in Figure 11. We also confirmed that it is still functional even with the higher V_{SS}. We used the same test as when choosing our normal SRAM size and the results are very similar to Figure 7 of normal 6-T Cell.

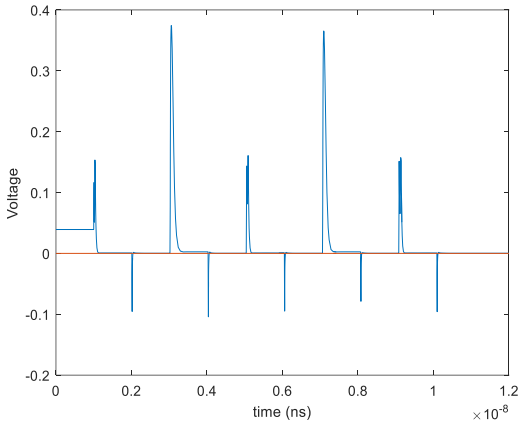

Fig.11 Voltage of the V_{SS} source of our cell. The red plot is the normal cell and stagnant at 0 V throughout the simulation. The blue plot is our clamp diode cell's voltage and even at points where it seems to be 0 V, it is actually at several hundred microvolts as opposed to just 0 V.

IV. PROPOSED TECHNIQUES ON SRAM ARRAY

We have built two versions of 32x32 SRAM Array where one has 6-T cell while other has two extra transistors for floating bottom rail during inactive mode. SRAM Array of extra transistor cell has two versions of its own: SRAM Array made of 8x8 sub-arrays, and another made of 16x16 subarrays. For the basic peripheral circuits, we have 5-to-32 row and column decoders for address and data access. Pre-charge, write driver and sense-amp will be multiplied to 32 copies to accommodate 32 different BL and BLN wires. Extra peripherals required for our proposed techniques will be discussed below.

A. Dynamic Voltage Supply

Dynamic voltage control will be done by the same circuit described in previous section. PMOS switches with same input to each gate where one is inverted will determine the all the SRAM Array cells' supply voltage level. When it is doing a read operation, supply will be raised to original 1V, otherwise supply will be lowered to 0.7V. As mentioned before, stability is a significant issue with this technique. The supply circuit needs to reliably switch between the two voltages for the whole array such that the correct rail voltages can be provided to active cells.

B. Extra Decoder for Sleep Transistor Technique

The architecture we're using for our circuit takes advantage of the clamp diodes we implemented in our cell. When the cells are inactive, or not being accessed, the clamp diode raises the inactive cell's V_{SS} voltage. When we want to access the cell, the clamp diode is turned off and the cell's V_{SS} is effectively pulled to ground to ensure stable reads and writes. We use an extra decoder to divide our array similar to Figure 12. The decoder is responsible for determining which partitions' clamp diodes will be turned on and off. Only one block will have its clamp diode turned off and that block is the one we want to read or write from. The other blocks will have the clamp diode turn on to reduce leakage.

The overall goal of this architecture is to toggle between lowering leakage when a cell is inactive and then switching to a higher stability once the cell is being accessed. Our current simulation uses a 32x32 array. Since each row is not too long, our partitions are strictly row-wise. However, as Figure 12 indicates, this technique can be scaled column-wise as well.

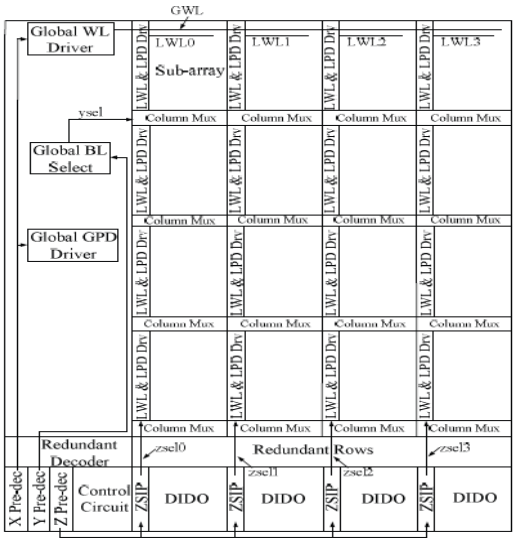

Fig.12 Example of SRAM Array with Extra Decoder in its peripherals

V. OBSERVED RESULTS AND ANALYSIS

Below are power estimates for the power consumed during read and write operations along with inactive power of the cells and the array's peripherals.

	Read	Write	Inactive
6-T	63.2uW	620.5uW	19.5uW
Multi-VDD	36.4uW	457.7uW	7.2uW
Clamp	37.6uW	153.0uW	10.2uW
Multi-VDD and Clamp	34.3uW	86.0uW	15.8uW

Table 1 Power consumption of 32x32 SRAM Array for different techniques

	Read	Write	Inactive
Pre-Charge	5.11uW	6.02uW	11.5uW
Write Driver	28.6uW	62.3uW	22.6uW
Sense-Amp	57.8uW	9.20uW	91.3uW
Decoder	12.uW	11.8uW	1.39uW
Extra Decoder	1.01uW	9.88uW	0.129uW
Dynamic Supply	0.002uW	0.007uW	0.003uW

Table 2 Power consumption of peripheral circuit in different SRAM operations.

Qualitatively, we can see that the techniques decrease the power. Multi-VDD observes lower power overall but includes the extra overhead of having two voltage sources and the proper circuitry involved in correctly switching between the two. It also suffers from the issue of higher instability due to lower cell rails during certain periods of operation which we do not test.

With the clamping diodes on the cells, we observe that the array has overall lower power relative to the control. We actually observed slightly higher read power for an individual cell but due to the architecture switching a fraction of the array to lower consumption, we get the overall power reduction. However, we have an extra decoder with its own overhead of power and area. Depending on how the array is divided, the overhead of the extra decoder may significantly affect the power of whole circuit. There is also the overhead of the new cell clamping diode which we expect to take up a considerable fraction of the new cell.

Finally we combine the techniques and observe lower power during the read and write operations. The inactive power is higher compared to the techniques individually. We suspect the higher instability of the cell while inactive because of the dynamic voltage source causes the clamping diode's voltage to fluctuate and become active more often than with the single voltage source configuration. Depending on the, traffic of the cell, there are tradeoffs to combining the techniques or using them individually along with the area and power overhead of extra peripherals.

VI. CONCLUSION

Using 32nm simulated technology and assuming 100nm minimum PMOS width, we simulated a 32x32 6T SRAM array along with relevant peripherals and proposed techniques to lower power consumption. The techniques proposed are a dynamic voltage supply and a new cell architecture and cell that uses the new architecture. The new cell architecture involves an extra decoder to partition the array to control which cells are active. The new cell uses a sleeping transistor and clamp diode to lower leakage and stabilize the cell. We tried each technique individually and combined them. Each configuration comes with tradeoffs in stability and area but overall observe lower power consumption.

REFERENCES

[1] Preeti S Bellerimatha and R. M Banakarb, "Implementation of 16X16 SRAM Memory Array using 180nm Technology", in International Journal of Current Engineering and Technology, Special Issue1 (Sept 2013)

[2] M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing Surplus Timing for Power Reduction," in Proc. of CICC'2001, pp. 89-92, May 2001.

[3] K. Zhang *et al.*, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," in *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 895-901, April 2005.

[4] Wu, Chen, et al. "SRAM power optimization with a novel circuit and architectural level technique." Solid-State and Integrated Circuit Technology (ICSICT), 2010 10th IEEE International Conference on. IEEE, 2010.