Analysis on Techniques Used to Scale Dynamic 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. There are many techniques aimed for active power reduction on SRAM arrays that are being published and already in use. This report investigates number of different approaches in reducing active power in SRAM and explores its performance in lowering power and how much area and speed of the memory it sacrifices in return. Immediately after, we will simulate each technique to compare our results to reported data for correctness. Afterwards, there will be design exploration on our side where we will try to combine multiple techniques, or add extra features to the technique presented for more efficient power reduction while trying to mitigate the effect of tradeoffs.

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. Although complexity of the applications that these devices perform have increased significantly along with boost in operating speed and better architecture in processor and its system around it, battery technology undergone very small improvement compared to other aspects listed above. Performance of a chip will continue to improve significantly but the battery efficiency will continue its slow growth as time passes as shown in Figure 1.

In this report, we investigate techniques used to reduce active power of the memory; one of the most frequently used components on the hardware of mobile devices. As software applications on small devices get more complex, processing units often need a temporary storage bigger than the flip-flops that it has on side. Therefore, caches made of SRAMs

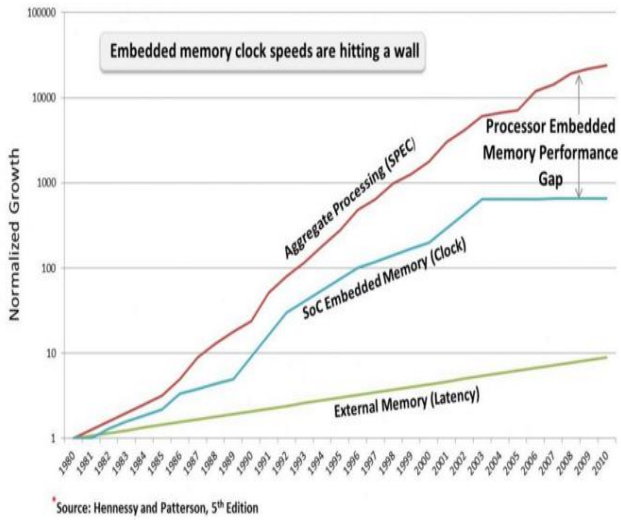

Fig.1 Graph of Processor Speed, SRAM and DRAM gap

are used to consistently write and read blocks of data for computations. Large ratio of the processor instructions are associated with the access of the memories which correlates to higher dynamic power consumed by the SRAMs. Thus, it is important to investigate how to mitigate dynamic power burned by memories since it is part of significant amount of energy that most of mobile devices use. Many techniques of reducing active power of the SRAMs have been formulated and have been in used or explained in papers. This report will explain and re-iterate ways of reducing dynamic power of the SRAM and go further into what is its limitation and explore possible errors. Later, our group will simulate the SRAM model and its extra features to see if the

results support lowering the dynamic power and comment on it. Furthermore, we will make our own design/modification for possible improvements.

II. LOW DYNAMIC POWER TECHNIQUES

In this section, we will take a look at different approaches made to reduce dynamic power consumption of the SRAM. We will analyze on what type of SRAM they tested/simulated on, the ways of reducing power, how much they were able to save power and finally, the tradeoff for making such design choice.

A. 10T SRAM Cell

One of the first considerations to reduce power in SRAM would be modifying the cells. One variation adds two NMOS transistors to an 8T SRAM cell such that the read path is controlled to be connected to one of the bit lines for write operation depending on what was written last, as shown in Figure 1. This design saves dynamic power as there are less discharge on read and write lines of the cells. However, there are clear tradeoffs in this design as we require adding two more transistors and wires for extra lines which add more area as number of cells grow large. Driving extra line RWL and RBL means additional area in the controller/decoder logic of the entire SRAM which also can lead to slower access time to data in a cell. The reported power reduction of 10T SRAM cell design relative to the 8T cell was 66% less during reads, 22.64% less writing a '0', 30.68% less writing a '1', and a 64% reduction of bit line leakage at V_{DD} = 1V. [1]

B. 5T SRAM Cell

Another variation of the SRAM cell uses five transistors in total in a cell as in Figure 2. Based off the 6T SRAM cell, the transistor connecting the complement of the bit line is removed. While read operations and writing '0' can be done in a similar manner to the 6T SRAM cell, writing '1' uses a global write signal since there is no longer any feedback. It uses a simplified sense amplifier due to there being only one bit line. One notable technique that was used with the 5T SRAM cell chip was to utilize the leakage from the actual array to raise the voltage of the bottom rail of the cell. In a more standard 6T SRAM cell, leakage is reduced by increasing the negative supply voltage; 5T SRAM technique reduces power through collecting the leakage rather an external power supply [2]. Although this design reduces both power and area of the SRAM, it has tradeoffs in read/write stability as well as speed of the access time of the cells that needs to be improved on. The power reduction relative to the standard 6T SRAM cell was ~30% for reads, ~80% for writing '0', and ~9% for writing '1'.

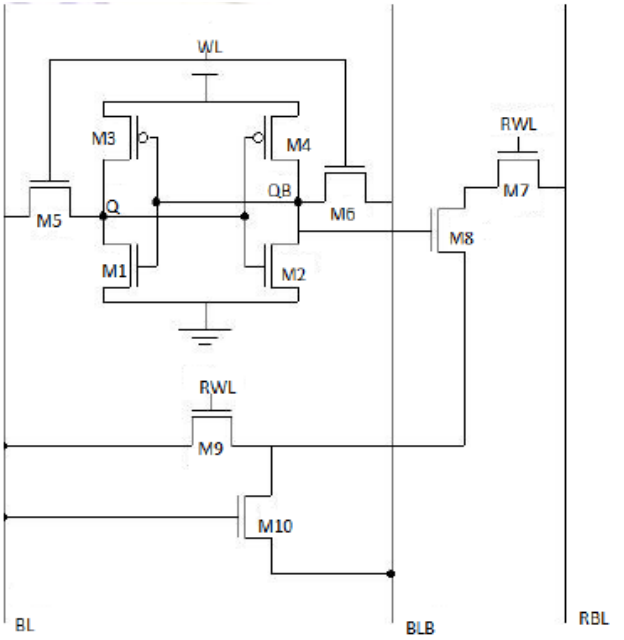

Fig. 1 10T SRAM cell which differs from the standard 8T SRAM because of M9 and M10

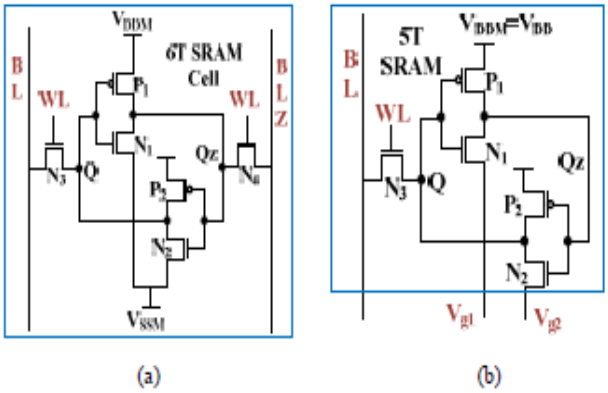

Fig. 2 A standard 6T SRAM cell (a) and the proposed 5T SRAM cell (b)

C. Clamping diode cell with new bit line/ word line architecture utilizing extra decoder

In this following design variation, a small change to the 6T SRAM cell was made by clamping diode SRAM cell, which is essentially adding a pair of

transistors to the negative rail as seen in Figure 3. The NMOS transistor would be controlled by a control signal such that when it is on, the node would act as a virtual ground but otherwise in standby mode, the NMOS is switched off, turning the PMOS off, raising the negative rail and reducing gate leakage due to the lowest V_{DS} [3].

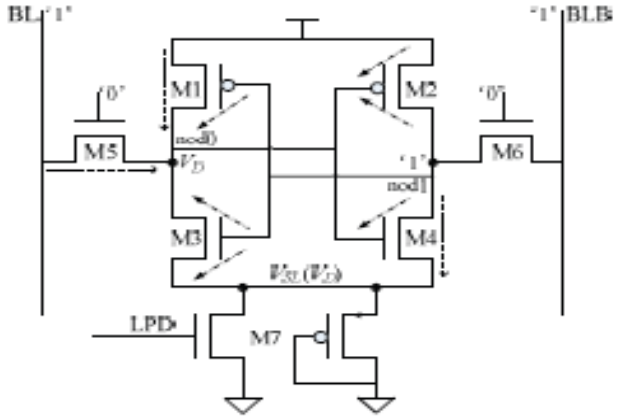

Fig. 3 Clamping Diode SRAM Cell

Another technique in addition to the clamping diode SRAM cell is adding extra decoder to implement a divided word line or bit line. Normally a decoder and pre-decoder are responsible for controlling the vertical and horizontal cells of our memory array separating the array into individual cells. What was proposed was to add another decoder and partitioning the array into sub-arrays as shown in Figure 4. This method reduces number of active cells in SRAM array down to ¼ of total cells when accessing the array to reduce the amount of power consumed in SRAM array [3]. Non-active cells in the SRAM array controlled by the extra decoder would only consume leakage power during the read/write as un-accessed cells have lessened the affect of charge/discharge of word and bitlines to reduce dynamic power used.

Clear tradeoff in this design is the increase in area of SRAM controller due to the extra decoder. However, the report showed data that read/write stability and performance of the memory wasn't significantly impacted. For instance, the performance of the memory was worsened by around 4% at maximum while stability of the data was acceptable even in wide range of temperatures with the Clamping Diode SRAM cell design [3]. For the results on power reduction, their simulation indicated a saving of 66.6% leakage power and 27.9% dynamic power.

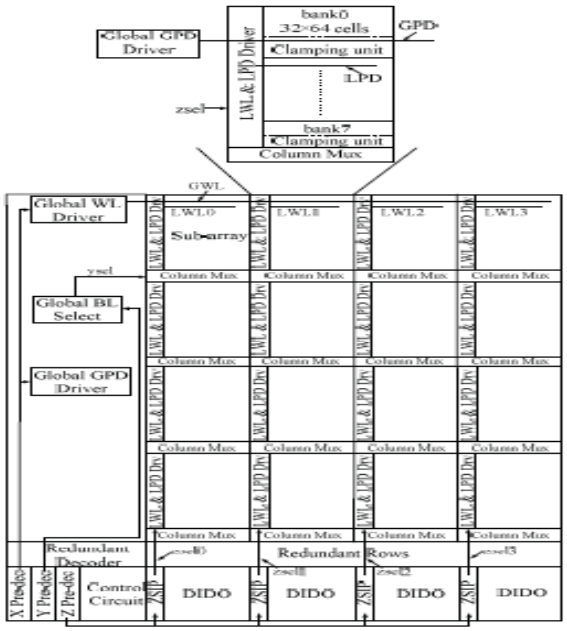

Fig. 4 SRAM divided into sub-arrays with addition of extra decoder

D. Drain Induced Barrier Lowering (DIBL) and bit line loaded with NMOS

Our last technique we investigated reduces SRAM power using circuitry outside of SRAM arrays. First, cell leakage would be reduced utilizing the DIBL effect. It assumes that the chip will have a DC-DC converter which will lower the top rail of the chip when it is not being accessed to lower threshold voltages of the transistors which lowers leakage power. To lower bit line power, a negative voltage is applied to inactive lines. By itself, this change would lower the reliability of the device because the oxide of the pass gate becomes overstressed. To fix this, the bit line is loaded with an NMOS transistor as shown in Figure 5. The bias conditions imposed by introducing the NMOS ensures that the pass gates are never overstressed [4].

Even though this technique sounds attractive, there might be complexity in designing such a memory as there has to be a controller that feeds two different voltage levels to both supply and bitlines of the cells in SRAM arrays depending on whether the cell is being accessed or not. This can have dramatic effect in increasing the chance of less stability and more error in data retention in SRAM as the rails and bitlines are fluctuating between two voltage levels. DC-DC converter circuit as well as the controller mentioned above may require additional unwanted

area. According to this technique, the proposed bit line scheme noticeably lowered the power; however the reduction wasn't as drastic as the DIBL effect when switching from 5V to 1V in rail which lowered leakage power by 99.2%.

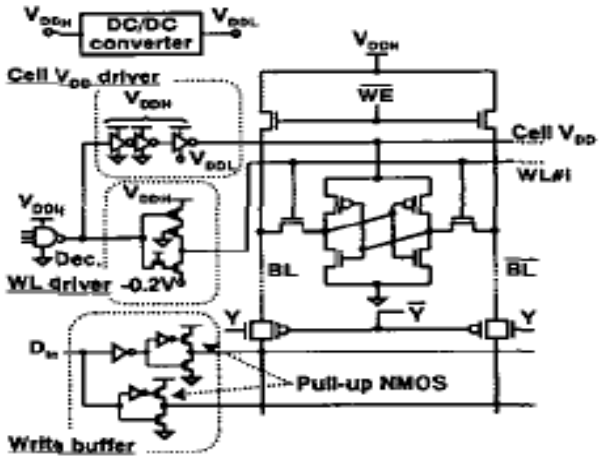

Fig. 5 Proposed bitline structure, replacing PMOS transistors with NMOS transistors

III. POSSIBLE DESIGN METHOD AND PLANS

We plan to simulate and gain data of our own for few of the techniques we investigated above. Reconstructing the circuit and simulating on our end will give us exact guidelines on how to make our design approach to improve reducing dynamic power of SRAM and mitigate the tradeoffs for each technique. Simulations and testing will be mostly done in Cadence and parasitic will be applied for accurate measurement of power, performance and stability of accessing the SRAM. Furthermore, we may be able to do some layout of the SRAM which will allow us to obtain even more accurate data as well as correct number for the area.

For the current design plan, we are looking at following list of techniques to reduce the active power of the SRAM:

1. Clamping diode cell with utilization of extra decoder
2. Lowering supply of the SRAM arrays using simple DC-DC converter
3. Extra features to help improve performance and stability of read/write

We believe that combining these three design points will meet the goal for reducing the dynamic power of the SRAM. Point 3 will require features to mitigate the effects of the tradeoffs in design. Example design we have in mind is to use simpler version of TVC-WA and WLUD (SRAM assist circuits) around the SRAM for read and word lines [5]. Key design goal for the project is to reduce the power of the SRAM. After the main goal is reached, we will measure how much the tradeoff has affected performance and stability of the memory and mitigate them as much as possible. Currently, we are looking to mitigate the increase in area of the SRAM and its surrounding circuitry to be less than 50%.

IV. CONCLUSION

In this report we have investigated at few techniques for reducing the dynamic power of the SRAM and explored tradeoffs for each of them. Some of the selected techniques will be base of our design which will help reduce the active power if the data provided from the references holds to be true. However, there will be tradeoffs associated with reducing power such as degrade in performance and stability which will be mitigated by our own extra features such as the assist circuit for the read/write lines. Further investigation will be done on the power-reduction techniques through simulation on our own end and to find appropriate methods for constructing simple DC-DC circuitry and assist techniques for lessening the tradeoffs.

REFERENCES

[1] G. Prakash, U. Dutta, M. T. Khan, "Dynamic Power Reduction in SRAM", IJERA, Vol. 2 Is. 5, Oct. 2012

[2] Hooman Jarollahi, et al, Dynamic Power Reduction in a Novel CMOS 5T-SRAM for Low-Power SoC, Proc. Intl. Conf. on Computer Design, CDES 2010, Jul. 12-15, pp. 169-175. (This work was first presented by HJ on Jun. 3, 2010 as part of his M.A.Sc. Thesis defense).

[3] 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.

[4] Kanda, Kouichi, et al. "Two orders of magnitude leakage power reduction of low voltage SRAMs by row-by-row dynamic V dd control (RRDV) scheme." ASIC/SOC Conference, 2002. 15th Annual IEEE International. IEEE, 2002.

[5] Karl, Eric, et al. "A 4.6 GHz 162Mb SRAM design in 22nm tri-gate CMOS technology with integrated active V MIN-enhancing assist circuitry." Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International. IEEE, 2012.