SRAM Voltage Stacking

Elnaz Ebrahimi, Rafael Trapani Possignolo, and Jose Renau
Dept. of Computer Engineering, University of California, Santa Cruz, Santa Cruz, CA, 95064
Email:{eebrahim,rpossign,renau}@ucsc.edu

***Abstract*—Current delivery is a major challenge in chip design. Reduction of the nominal voltage due to technology scaling has worsened the problem. Voltage stacking has been proposed as a way to alleviate the problem by delivering power in a serial rather than the conventional parallel way. Several studies have proposed techniques to stack logic designs. This paper applies the voltage stacking technique to SRAMs. By dividing the SRAM into two logic domains, we are able to double the supply voltage VDD while reducing the current draw significantly. Since SRAMs have a predictable activity pattern, each stack consumes the same amount of power, therefore, the stack voltage $2VDD$ will distribute evenly between the stacks and the current demand will decrease up to 44%. The combined effects of increasing VDD and decreasing current allow the design of Voltage Regulators to be 10%-15% more efficient.**

I. Introduction

Delivering power to the logic in a chip is one of the major challenges in current chip design [7]. As technology scales down, the reduction in voltage supply has led to a sharp increase in the total current that needs to be delivered to a chip. Decreasing voltage has been shown to reduce efficiency of Voltage Regulators [6], while increasing current also has several drawbacks such as: increase in voltage noise and losses due to parasitics [4], [7], increase in the number of pins dedicated to power [11], and electromigration. Hence, voltage stacking of CPU cores has been proposed by several groups to mitigate these problems [4], [7], [11].

Voltage stacking is an alternative method to deliver power to components that are placed in series rather than in parallel. The charge between the layers is recycled, *i.e.*, it passes through multiple components [9]. Thus, for the same power budget, voltage stacking allows for delivering less current at an increased voltage level. This effectively reduces the total chip current by n, where n is the number of stack levels used in the design [6].

Previous work has focused on the voltage stacking of cores and logic components of a chip, however, current chips dedicate a large portion of their area to SRAMs and cache blocks [3], and thus SRAMs account for a considerable amount of the chip power. Researchers have been able to reduce the current drawn from on-chip SRAMs by focusing on reducing their total power consumption [1], [3], [5]. Although reducing the power consumption is a goal by itself, it has limited impact when it comes to reducing the total current. Voltage stacking resolves the problem by greatly reducing the current drawn by SRAM components.

The main challenge in voltage stacking is to balance the activity between the stack levels to maintain the voltage at each level roughly constant [2], [7]. An additional Voltage Regulator (VR) is generally used in the intermediate node to guarantee that the voltage stays within specified values [6]. SRAMs have fairly predictable activity during operation, making them ideal candidates for voltage stacking. Voltage stacking SRAM banks has been proposed as a technique to reduce the standby power of those components [3]. In this approach, when the banks are not in use, power switches change the banks to a stacked configuration where only half of VDD is applies to each bank. This reduces the leakage power during the standby mode.

We propose applying voltage stacking to SRAMs. In order to guarantee that the activity is the same across the stack levels, the stacking is done by splitting each word and stacking the word parts, which also guarantees that the access to both stack levels occurs in the same clock cycle. Besides the RAM core, sense amplifiers and prechargers are also stacked. Level converters are used where needed to guarantee that each component receives the appropriate voltage level.

Our experiments show that SRAM stacking reduces the current drawn by the SRAM by 40% during the write, 36% during the read and 44% during the standby mode. Overall, it leads to an average of 36% to 44% of current reduction depending on the activity. This can yield to a reduction in IR drop. The reduced current also has a linear impact on the number of power delivery pins.

To the best of our knowledge, this is the first work that proposes voltage stacking in SRAMs at all times. The proposed approach reduces the pressure of on-chip power grid design, by reducing the current, especially when combined with core voltage stacking. This is also the first paper to propose voltage stacking without the additional VR.

II. Related Work

Voltage stacking has been proposed in the context of CPU cores to increase the efficiency of VR as well as reduce the high current demands for current chips [7], [9]. Voltage stacking does not reduce the power consumption in the chip, but rather allows for operating at a higher voltage and lower current level. This is beneficial for VR design, because it reduces both its power loss and area [7].

Voltage stacking has also been proposed to reduce the number of pins dedicated to power on a chip [11]. Since the number of power pins is roughly proportional to the current, reducing the current by a factor of n (in an n stacked configuration) would result in a reducing the number of power pins by the same factor.

Gu *et al.* show how voltage stacking reduces voltage noise (IR and $L\frac{dt}{dt}$) and IR drop in the power grid, which can ultimately reduce power in the parasitics of the system [4]. Having reduced the noise, we would be able to reduce the voltage margins, and increase the power savings. However, we do not evaluate it in this manuscript, because it requires a very good parasitics characterization to obtain meaningful results.

Cabe *et al.* propose to dynamicall stack SRAM blocks while they are inactive to decrease the leakage power during the standby phase of the operation. This technique uses power switches that select the stacked mode when in standby or regular mode when performing regular operations [3]. However,

it provides a constant VDD, regardless of the circuit state (stacked or not), which means when stacked, half the VDD is applied to each stack level. Thus, when VDD is doubled, the full VDD is applied to each stack level at all times. We propose maintaining the circuit stacked during all phases of the operation. It provides a decrease in leakage power as well as other benefits of voltage stacking, such as power pin reduction, VR efficiency increase, and voltage noise reduction.

III. SRAM Stacking Model

Voltage stacking helps reduce the current draw when applying a higher power supply voltage to the logic blocks in the design. To take advantage of the charge recycling, instead of running the SRAM at VDD, we divide it into two logic domains connected in series, and apply $2VDD$. If each logic domain consumes the same amount of power, the voltage will distribute evenly between them. The logic domain loads however have to be selected in such way that they have well-balanced charge utilization to achieve a high efficiency [9]. If the power consumption of the 2 stacks is the same, as the voltage supply is multiplied by n, where n is the number of stacks, the current draw will be reduced to $\frac{1}{n}$. We use 2 stacks and theoretically it should lead to 50% reduced current in the SRAM as shown in equation 1.

$$\begin{aligned} p1 &= p2 \\ V_1 \times I_1 &= V_2 \times I_2 \text{ where } V_2 = 2 \times V_1 \\ I_2 &= \frac{I_1}{2} \end{aligned} \tag{1}$$

In terms of circuit design, Figure 1 and 2 show how the conventional model differs from the stacking model. In Figure 1, the two circuits are running in parallel and the same voltage differential is applied to each. The total current drawn from the power source is the summation of the current of all the components in the circuit. In Figure 2, the stacks run in series and the configuration reduces the current draw. And the IR drop across the stacked components is reduced by a factor of n where n is the number of circuits in series. V_{mid} will fluctuate depending on the load and impedances present in each stack. If the impedances are similar, the V_{mid} balances in the middle and becomes the most efficient stacked configuration.

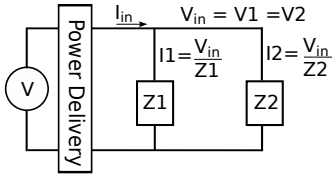

Fig. 1: In the conventional powery delivery mode, the same voltage is applied to each component in the circuit.

Figure 3 shows a high level view of the stacked SRAM. The SRAM is 1Kb, 32 32-bit words, with 2 read ports and 1 write port. To divide it into two vertical logic domains, we cut the 32-bit wordline in 2, which leaves 16 bit for each stack. Bits 0-15 go to the bottom stack and bits 16-31 go to the top stack. Consequently, each stack will have 16x32 bitcells, 16 precharge circuits, 16 write drivers and 3x32 wordline buffers.

The two stacks operate at 0-VDD and VDD-$2VDD$ where VDD is 1.1V. The voltage level of all the signals entering

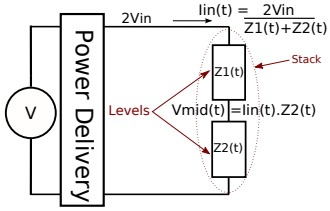

Fig. 2: In the Voltage Stacked mode, current is recycled through multiple components in series, reducing the total current drawn.

the top stack (2.2V) will need to be shifted; hence, level shifters are placed in 3 locations. First, the input data to write drivers has to be converted to the same voltage level as the top stack. The wordlines will pass both stacks in a row of bitcells, therefore, each decoder will drive two wordlines. Secondly, the top stack wordline voltage level is shifted. To avoid having the wordline be driven by the level shifters, we place them before the read and write address buffers. Lastly, the sense amplifier outputs exiting the top stack will have to be converted back to VDD; hence, level shifters are placed right after the sense amplifiers. The read and write address decoders, placed in the middle of the RAM core, are the only components which are not part of the stacked architecture.

The schematic of the level shifter used throughout the SRAM is shown in Figure 3. There are two capacitors C1 and C2, which are sized 15fF each. The design is adopted from the Lee's 16-core design [6].

IV. Setup

We implmented the voltage stacking technique on SRAMs generated by FabMem. FabMem is a multiported RAM and CAM compiler for design space exploration and given the configuration, it generates netlists and layouts and estimates read/write delay and energy consumption [10]. FabMem uses the NCSU FreePDK, the Open-Access-based PDK for the 45nm technology node.

Using FabMem, we generated an SRAM with a configuration similar to the size of a typical Register File: A 2-read 1-write 1Kb consisting of 32 32-bit words. We use one RF for the base case SRAM and another for the stacked version.

A few alterations had to be made to some of the SRAM components such as the sense amplifier and the precharge circuits. In our experiments, we use current latch mode sense amplifier [8] as opposed to the FabMem default voltage controlled sense amplifier circuitry, because the current latch mode works with the stacked SRAM design. The precharge circuitry used is shown in figure 3.

All the energy related measurements were taken using Synopsys HSpice version I-2013.12-1.

V. Evaluation

In this section, we discuss the overall results and share insights as how effective SRAM voltage stacking is.

The two SRAM netlists are simulated using HSpice at $VDD = 1.1V$ with frequency of 500MHz. The simulation results compare the stacked SRAM against the non-stacked SRAM, which we refer to as the base case in this section. Figure 4 shows the energy consumption breakdown. Each colored pattern shows the total energy consumption of a particular component. The components included in the graph are all part of the stacked architecture. The address decoders

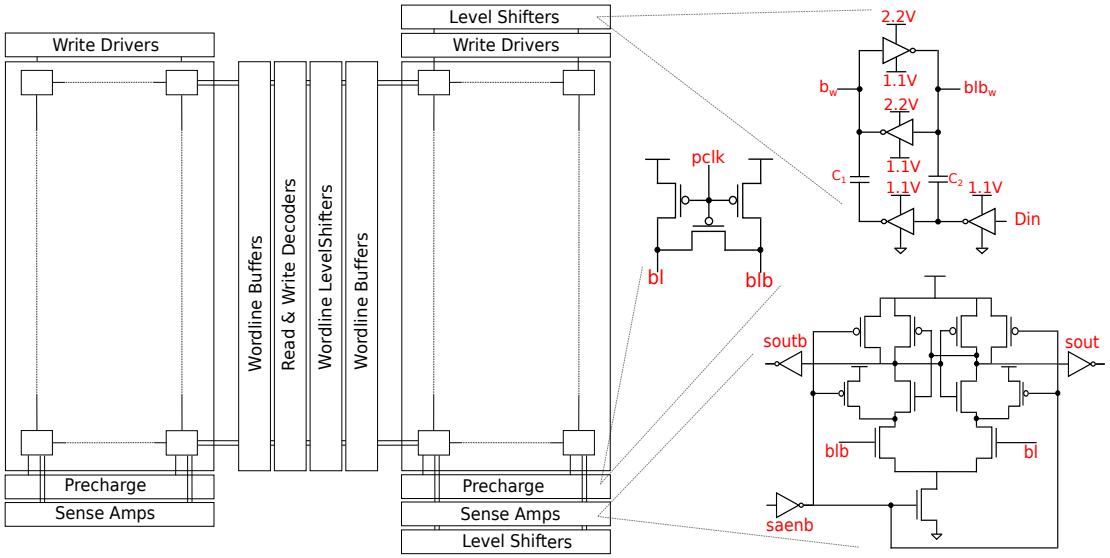

Fig. 3: Voltage Stacked SRAM

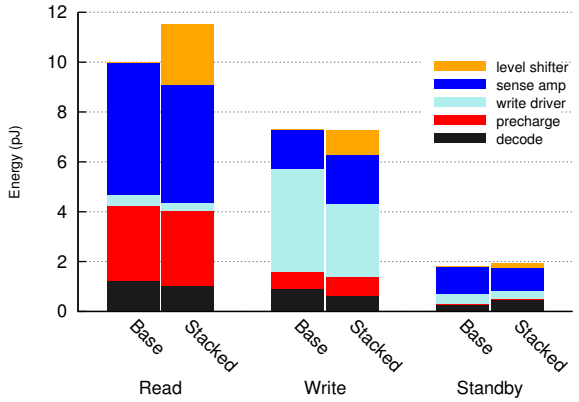

Fig. 4: Energy breakdown in voltage stacked SRAM vs. non-stacked.

are not included, but the buffers associated with them are. For convenience, we call this component "decoder".

The simulation includes periods of initialization, write, read, followed by standby. Each write or read period consists of 10 write or read operations. Overall, the power consumption increases by 20%, 23%, and 13% during write, read, and standby modes respectively. During the read operations, the stacked SRAM consumes a bit more energy than the base case, however, looking at the breakdown, the excess energy usage is due to having the level shifters in the top stack. For instance they draw 8% more current during the read operation than the precharge circuits. The other components such as sense amplifier, precharger, decoder, and write drivers almost consume the same amount of energy. The small differences are due to the partitioning and changes in load capacitance needed for stacking. As the SRAM size increases, the overhead of level shifters will become less of an issue. The standby energy consumption is the same: At $2VDD$ the SRAM has the same leakage as it would when operating at VDD. Figure 5 shows the current consumption in the top stack of the stacked SRAM versus the base SRAM. Overall, stacked components draw 40%, 36% and 44% less current than the base components during the write, read, and standby modes. Comparing the current and energy breakdown plots for the write operation, we notice that the write drivers do not have a balanced load, leading to less than expected current savings. For larger SRAMs however, we expect this gap to decrease, since the relative power consumption of write drivers is smaller.

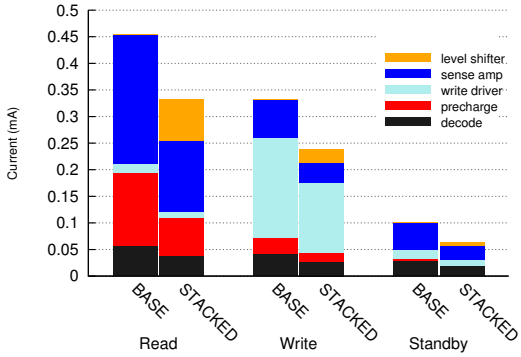

Fig. 5: Current breakdown in voltage stacked vs. non-stacked SRAM. Voltage stacking has reduced the total current drawn from the source.

We also evaluate the impact of stacking on the SRAM performance. For bitline transitions shown in Figure 6, we simulated both SRAMs with 1GHz frequency. *out2<2>* and *out2<29>* are selected from bottom and top stack respectively to show how level conversion affects the SRAM speed. As Figure 6 shows, *out2<29>* has some delay when transitioning from 0 to 1. It is 60*ps* delay which translates to 6% frequency hit. In reality, the performance hit should be even smaller, because the stacking can be applied to the slower part of the SRAM. We did not perform this optimization because it is layout dependent. Thus, SRAM stacking has limited impact on performance.

Finally, we evaluate the voltage noise level in the V_{mid} due to the fluctuations of the load. As expected in a stacked architecture V_{mid} fluctuates and has noise and this is shown

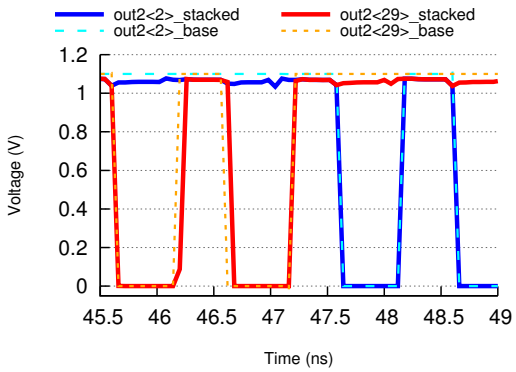

Fig. 6: The stacked SRAM output delay is comparable to the non-stacked. *out2<2>* is one of the fastest bit compared against *out2<29>*, one of the slowest bit of the top stack.

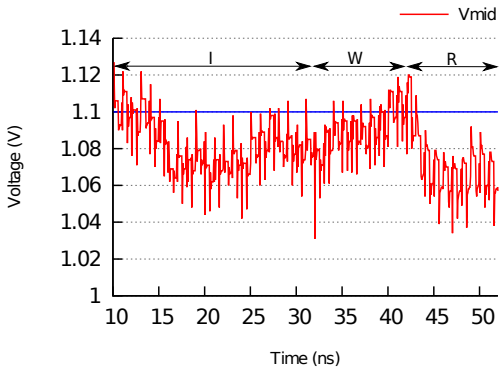

Fig. 7: Voltage noise for Stacked SRAM is within acceptable levels, even without an extra VR.

in Figure 7 during initialization, write, and read operations. However, it is within the accepted range without placing a VR in the design. In case the voltage noise is not within the accepted range, a VR can be placed at V_{mid} to guarantee that it stays at the *VDD* level.

The stacked SRAM in our experiments does not include any voltage regulators to maintain V_{mid} at a constant level. Note that the overall efficiency of a VR depends on *VDD* and the output current. VR efficiency depends on a large number of factors at design time, but with 2*VDD* and roughly 40% less current, one can estimate the increase in 10%-15%, thus saving 10%-15% system-wide power [2]. *IR* drop is also reduced, since it is proportional to the current, thus in our design, we expect around 40% reduction in *IR* drop (which could be converted into voltage margin reduction, but this is not evaluated).

The reduced current draw also impacts the number of pins used for power delivery. Ardestani et. al in their core unfolding work also mention that the number of pins and pads is mainly determined by the total current flowing through them. To maintain the current per pin constant, it is possible to reduce the number of pins used in the design [2]. Therefore, our proposal could allow a reduction of up to 40% in the number of power pins.

Voltage Stacked SRAMs have a small area increase; the level shifters are the main source of overhead. Even in a small SRAM such as the one we have evaluated, the level shifters represent less than 6% of the total number of transistors.

VI. Conclusion

As the technology scales, and power supply is reduced, delivering power to the logic on a chip becomes a major challenge as the current demand increases drastically. Increased current has many drawbacks [4], [6], [7], [11]. Voltage stacking is an alternative for delivering power to components on a chip that are stacked or placed in series. A large portion of the chip area belongs to SRAMs and caches. In this paper, we focus on SRAMs and stack an SRAM the size of a typical Register File. Our design keeps the SRAM in stacked mode at all times. Other research work have been proposed where they keep the SRAM banks in stacked mode during the standby mode [3]. We divide the SRAM words into 2 stacks and are able to reduce current by 36%-44% while not letting energy consumption be more than 23%. In addition, doubling *VDD* and reducing the current could save as much as 10%-15% power from the increased VR efficiency [7].

Acknowledgements

We thank Doctor Matthew Guthaus for his assistance. This work is supported by the NSF grants CNS-1059442-003, CNS-1318943-001, CCF-1337278, and CCF-1514284. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] B. S. Amrutur, "Design and Analysis of Fast Low Power SRAMs," Stanford, CA, USA, Tech. Rep., 2000.

[2] E. K. Ardestani, R. T. Possignolo, J. L. Briz, and J. Renau, "Managing Mismatches in Voltage Stacking with CoreUnfolding," *ACM Trans. Archit. Code Optim.*, vol. 12, no. 4, pp. 43:1–43:26, Nov. 2015.

[3] A. C. Cabe, Z. Qi, and M. R. Stan, "Stacking SRAM Banks for Ultra Low Power Standby Mode Operation," in *Proceedings of the 47th Design Automation Conference*, ser. DAC '10, pp. 699–704.

[4] J. Gu and C. Kim, "Multi-story power delivery for supply noise reduction and low voltage operation," in *Proceedings of the 2005 international symposium on Low power electronics and design*. ACM, 2005, pp. 192–197.

[5] M. Huang, J. Renau, S.-M. Yoo, and J. Torrellas, "Cache Decomposition for Energy-Efficient Processors," in *International Symposium on Low Power Electronics and Design*, Aug 2001.

[6] S. K. Lee, T. Tong, X. Zhang, D. Brooks, and G.-Y. Wei, "A 16-Core Voltage-Stacked System with an Integrated Switched-Capacitor DC-DC Converter," in *VLSI Circuits (VLSI Circuits), 2015 Symposium on*, June 2015, pp. C318–C319.

[7] S. Lee, D. Brooks, and G. Wei, "Evaluation of Voltage Stacking for Near-threshold Multicore Computing," in *Low Power Electronics and Design (ISLPED), 2012 IEEE International Symposium on*. ACM, 2012, pp. 373–378.

[8] B. Mohammad, P. Dadabhoy, K. Lin, and P. Bassett, "Comparative Study of current mode and Voltage mode Sense Amplifier Used for 28nm SRAM," in *Microelectronics (ICM), 2012 24th International Conference on*, Dec 2012, pp. 1–6.

[9] S. Rajapandian, K. Shepard, P. Hazucha, and T. Karnik, "High-voltage Power Delivery Through Charge Recycling," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 6, pp. 1400–1410, 2006.

[10] T. Shah, "FabMem: A Multiported RAM and CAM Compiler for Superscalar Design Space Exploration," Master's thesis, North Carolina State University, 2010.

[11] Y. Zhan and S. S. Sapatnekar, "Automated module assignment in stacked-vdd designs for high-efficiency power delivery," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 4, no. 4, p. 18, 2008.