# Resonant Energy Recycling SRAM Architecture

Riadul Islam, *Member, IEEE*, Biprangshu Saha, and Ignatius Bezzam, *Member, IEEE*



***Abstract*—Although we may be at the end of Moore's law, lowering chip power consumption is still the primary driving force for the designers. To enable low-power operation, we propose a resonant energy recovery static random access memory (SRAM). We propose the first series resonance scheme to reduce the dynamic power consumption of the SRAM operation. Besides, we identified the requirement of supply boosting of the write buffers for proper resonant operation. We evaluated the resonant 144KB SRAM cache through SPICE and test chip using a commercial 28nm CMOS technology. The experimental results show that the resonant SRAM can save up to 30% dynamic power at 1GHz operating frequency compared to the state-of-the-art design.**

***Index Terms*—SRAM, series resonance, low-power, caches, bitline discharge.**

## I. Introduction

Connecting an unlimited amount of high-speed embedded memories such as static random access memories (SRAMs) to the microprocessor or a system-on-chip (SOC) and having them as piggyback on computing is playing a pivotal role in designing a high-performance computing system and data centers. The embedded memory consumes the significant portion of a microprocessor and enjoys more aggressive design rules compared to the rest of the logic. However, the cache memories remain in the critical path of a general-purpose computing and designing large SRAMs with a bounded performance and power budget becomes a very thorny problem that needs to be dealt with immediately and carefully.

Among all the memories in a cache architecture, the SRAMs are essential for efficient program execution. The SRAM provides the performance that is close to the processor speed, which is much faster than the main memory; however, it consumes significantly more area and power per/bit than dynamic-RAM or DRAM. Due to large size and high-speed, SRAMs consumes about 10%–20% of total dynamic power in a microprocessor power-arc [1]. To reduce microprocessor power, researchers applied many low-power techniques; among them, resonant energy recovery (ER) clocking is used widely. In this work, we introduce resonant SRAM architecture to reduce effectively SRAM power even for non-cyclical operation and consequently enable ultra-low-power computing.

R Islam is with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, MD 21250, USA e-mail: riaduli@umbc.edu.

B Saha is with the Si2Chip Technologies, Road 1B, Gayatri Tech Park, Bengaluru, Karnataka 560066, India e-mail: biprangshu.saha@si2chip.com

I Bezzam is with the Rezonent Inc., 1525 McCarthy Blvd, Milpitas, CA 95035, USA e-mail: i@rezonent.us

This work was supported in part by Rezonent Inc. and by the UMBC startup grant.



### A. Prior Work and Motivations

An SRAM consists of an array of data storage cells and peripheral circuits to control the memory and allow us to read/write with a bit-level precision. The SRAM's reliability depends on the cell's robustness and the peripheral circuitry to noise; and process, supply voltage, and chip temperature (PVT) variation. Besides, researchers identified that a significant amount of dynamic and static power consumed by SRAMs, especially at sub 10nm technology node with increased SRAM density and many SRAM cuts in a single chip. As a result, there has been a tremendous amount of work on SRAM design to improve SRAM design efficiency [2]–[4]. However, this work's primary goal is to reduce the SRAM power without affecting the cell density and performance of the memory.

The most widespread low-power techniques in IC design are dynamic voltage and frequency scaling (DVFS) [5], resonant LC clocking [6]–[8], current-mode (CM) clocking [9], and etc. Among different low-power techniques, LC resonant clocking is very interesting due to its constant phase and magnitude. However, in the proposed research, we apply LC resonance to reduce SRAM power consumption. Previously, researchers applied resonant clocking in SRAMs to save power [10]. This method used resonant ER latches in the address, wordline, and input latches to save energy.

One of the recent works, researchers applied supply boosting for SRAMs as a combination of capacitive and inductive boosting, as shown in Figure 1 [11]. This

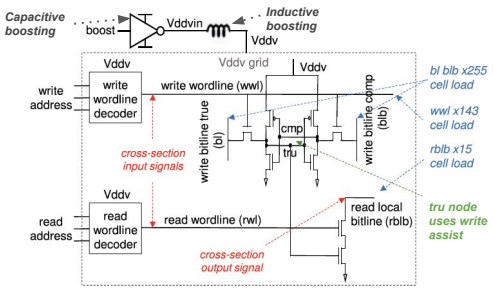


Fig. 1: To enable low supply voltage operation researchers applied both capacitive and inductive boosting of the input voltage; figure modified from [11].

approach uses a transistor-based capacitive coupling for initial supply boosting on 14-nm SOI FinFET technology.The initial enhanced supply voltage further amplified using a resonant inductor to achieve a meager 0.3V supply voltage operation. This method uses two additional transistors per cell as reading and write-assist, compared to the standard six transistors (6T) SRAM cell. However, this method requires a sizeable 4nH inductor for $144 \times 256$b SRAM for 0.3V operation. Another similar approach uses novel cascaded inductive and capacitive booster to reduce 8T SRAM supply voltage down to 0.24V considering 14nm SOI FinFET technology [12]. However, there is no real guideline on the appropriate inductor size for the 25.5Kb SRAM.

To enable low-power memory operation, we introduce resonant bit lines and inductive supply boosting for the write drivers using conventional 6T SRAM cells on a 28nm CMOS technology. We applied the proposed technique on a generic K2 cryptoprocessor (GCrP) with 144KB of SRAM, which enables up to 20% lower SRAM dynamic power compared to the conventional CMOS SRAM implementation with no leakage penalty.

### *B. Main Contributions*

In this work, we reduced the embedded SRAM power by introducing series resonance on the cache memory. In particular, the critical contributions of this work are:

- The first series resonant SRAM architecture.
- The first inductor sizing technique considering discharge time and maximum resonant swing.
- The significant reduction of the SRAM dynamic power without changing the conventional 6T cell architecture.

### *C. Paper Organization*

The rest of this paper is organized as follows. In Section II, we first introduce the resonant techniques. Section III presents the proposed SRAM architecture.

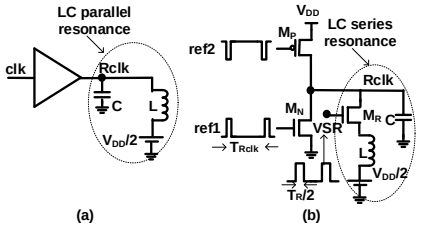


Fig. 2: (a) Parallel resonance exhibits power saving at a limited frequency range [8], (b) A series resonance uses pulsed signals to maintain rail-to-rail voltage swing, and the switched controlled inductor helps it to operate efficiently in a wide frequency range [6].

In Section IV, the power efficiency of the proposed resonant design with existing industry-standard schemes is investigated with simulation and experimental results. Finally, Section V concludes the paper.

## II. Resonant Background

The ER resonant clocking can be classified as standing wave [13], rotary [14], and LC resonant [6], [15]. Among various resonant schemes, rotary clocks have fixed amplitude but a variable phase. In contrast, a standing wave clock has a constant phase but a varying amplitude. LC resonants mimic the conventional CMOS clocking accurately with a higher slew rate; however, it exhibits tremendous potential to save dynamic power.

In a conventional CMOS design, half of the switching energy is wasted in charging a capacitive node (i.e., 0-to-1 transition); and the other half is wasted in the discharging phase (i.e., 1-to-0 transition). The LC resonance stores some of the discharge energy in the magnetic field on an inductor (L) and recycles during the charging phase to charge the capacitor (C). To maintain the resonance, we need an external source to compensate for the resistive loss. LC resonant clocking can be categorized as parallel and series resonance. At resonance, conventional LC parallel resonance cancels out inductive and capacitive reactance, as shown in Figure 2(a).

On the other hand, the series LC resonance requires additional transistor switching ($M_R$), as shown in Figure 2(b). The pull-up ($M_P$) and pull-down ($M_N$) switches help maintain rail-to-rail voltage operation. The ($VSR$) signal is generated from the rising and falling edges of the input clock (clk) signal. The primary advantage of series resonance compared to the parallel resonance is the wideband frequency operation. The required timing of $T_R$ is only a fraction of the overall resonance clock ($T_{Rclk}$), and inductor sizes are an order of magnitude less.

Figure 3(a) shows a series resonance equivalent circuit which helps us accurately define the resonant frequency and the corresponding inductor size. The equivalent total

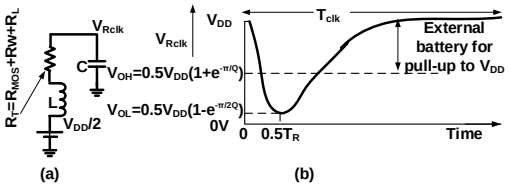


Fig. 3: (a) The series resonance equivalent circuit model help us to identify the proper $R_TLC$, (b) The output capacitive voltage is important to identify maximum resonant $V_{Rsw}$ and timing specification of a design.

resistance ($R_T$) is the combination of "ON" NMOS resistance ($R_{MOS}$), wire resistance ($R_W$), and inductor parasitic resistance ($R_L$). According to Kirchhoff's voltage law (KVL), we can write,

$$R_T i_L(t) + \int \frac{i_L(t)}{C} dt + L\frac{di_L(t)}{dt} = \frac{V_{DD}}{2} \qquad (1)$$

where $i_L$ is the inductor current [6]. The minimum inductance required considering underdamped condition is $L > \frac{R_T^2 C_L}{4}$. This is a critical condition that helps us to pick the right inductor for our design. From the Equation 1, we can express the $i_L$ as,

$$i_L(t) = \frac{V_{DD}}{2\sqrt{\frac{L}{C}}\sqrt{1-\frac{1}{4Q_f^2}}} e^{-\frac{tR_T}{2L}} sin(2\pi f_R t) \qquad (2)$$

where $f_R = \frac{1}{T_R} = \frac{1}{2\pi}\sqrt{\frac{1}{LC} - \frac{R_T^2}{4L^2}}$ represent the damping oscillation frequency, $Q_f = \frac{1}{R}\sqrt{\frac{L}{C}}$ is the quality factor. The $f_R$ value help us to identify the proper inductor, $R_T$, and capacitive load for corresponding damping frequency in Section III. The $T_R$ time corresponds to the bitline discharge time and will be discussed in detail in Section IV. Now, we can compute the voltage across the capacitor ($V_{Rclk}$) as,

$$\begin{aligned} V_{Rclk}(t) = \frac{V_{DD}}{2} + \frac{V_{DD}}{2} e^{-\frac{tR_T}{2L}} cos(2\pi f_R t) \\ - \frac{1}{2Q_f}\frac{V_{DD}}{2} e^{-\frac{tR_T}{2L}} cos(2\pi f_R t) \end{aligned} \qquad (3)$$

Figure 3(b) shows the $V_{Rclk}$ curve, where the difference between resonant high output ($V_{OH}$) and low output ($V_{OL}$) represent the voltage-swing ($V_{Rsw}$) due to the autonomous ER. Hence, we performed extensive simulations to identify the proper inductor and capacitive load to maximize this $V_{Rsw}$ in Section IV-A.

## III. Proposed Resonant SRAM Architecture

To improve the power-performance of embedded cache memory, we propose the resonant SRAM architecture. Empirically, in an SRAM write operation, many

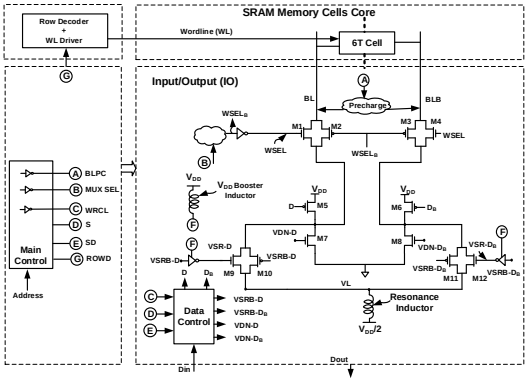


Fig. 4: Unlike existing low-power SRAMs [11], [12], we proposed the first series resonant SRAM architecture that uses ER bitlines and inductive boosting for the write driver to sustain resonance at a wide frequency range.

bit cells switches and make it the most dynamic power-consuming phase as all the bitlines with high capacitive load discharges during the write operation and charges back, irrespective of mux factor. Unlike conventional low supply voltage SRAMs, we recycle the discharged energy from the bitline load capacitances. The proposed architecture uses an on-chip inductor is conditionally attached to bitlines biased by another supply voltage ($\frac{V_{DD}}{2}$) to store the discharge energy considering in series resonance topology. While charging back in the recovery phase, the stored energy is used to charge the load capacitance towards $V_{DD}$. The inductor in the series resonance circuit stores the discharge energy in the form of a magnetic field, which in turn empties the load capacitance charge to a greater extent, hence stores the electric charge in $\frac{V_{DD}}{2}$ node. In the recovery phase, the same series resonance circuit pulls out the same amount of charge from the $\frac{V_{DD}}{2}$ node, leaving zero net currents from that node in the whole cycle, ensuring no additional power drawn from the inductor bias supply.

The proposed SRAM architecture is based on the existing 6T bit cell and uses conventional read-write methodology with peripheral circuitry, as shown in Figure 4. The write driver connects to bitlines load through the transmission gates (M9-M12) and controlled by VSR-D and VSRB-D and their complements. To enable series resonance inductor is placed between node VL and write driver transmission gates. We ensure the full rail-to-rail swing using a pair of NMOS transistors (M7-M8) parallel to the write driver's series resonance transmission gates. To achieve a reasonably high $Q_f$, we use low threshold voltage devices in the series resonance path (M1, M9, M4, and M12).

It is vital to use a shared inductor to reduce the size of the inductor. We need N number of write drivers

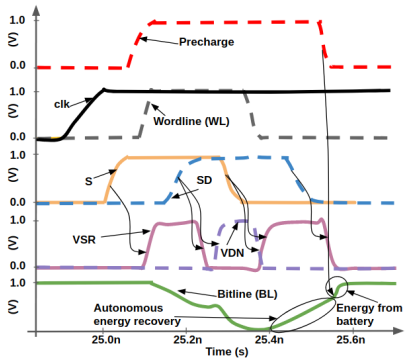


Fig. 5: The internal control signals S and SD produce the VSR and VDN signals; the former helps the bitline to discharge on the resonant inductor path, while the latter confirms the full rail-to-rail swing of the bitline in the resonant write cycle.

for N number of bits. A shared inductor connects all the write drivers to the VL node. Hence total load capacitance increases N folds, and series path effective resistance (M1-M2 and M9-M10) decreases N folds. The proposed architecture makes it possible to achieve the target frequency of operation with a low value of the inductor and high $Q_f$.

As the transistor source is connected to the $\frac{V_{DD}}{2}$ node through the inductor, driving the gate of these NMOS transistors by $V_{DD}$ turns these "ON" loosely. To overcome this problem, we generate a bump voltage with another resonant path to get the bump without wasting the power. The M9 and M12 transistor gate capacitances are the load capacitance for this path. For all the write drivers, we used a shared booster inductor.

## IV. Simulations and Test Chip Results

### A. Simulations Results

We performed extensive simulations of peripheral and SRAM core arrays using TSMC 28nm CMOS technology. For series resonant operation, we generate two voltage pulses (VSR and VSD) using the signal S and a delayed version of S, named SD. Both the signals are generated from the SRAM input clk signal. We generate the VSR signal using a 2-input XOR gate where the input signals are S and SD. Similarly, we extract the VSD signal using a 2-input AND gate where the input signals are S and SD. During the write operation, VSR and its complementary signals discharge the bitlines using transmission gate M9-M12 transistors through the inductive path, as shown in Figure 4. However, to fully discharge the bitline, we need the VDN signal. Figure 5 shows the simulation results of $512 \times 128$ bits resonant SRAM control signals during the write operation.

We computed the required $T_R$ pulse width depending on the SRAM MUX factor or number of connected

TABLE I: For a fixed resonant inductor, the $\frac{T_R}{2}$ time reduces with the reduction of the total number of associated columns.

| MUX factor | # of columns | Total cap (pF) | Inductor (nH) | $\frac{T_R}{2}$ (ps) |
|---|---|---|---|---|
| 1 | 256 | 10.10 | 0.621 | 248.0 |
| 2 | 126 | 5.07 | 0.621 | 176.0 |
| 4 | 64 | 2.53 | 0.621 | 125.0 |

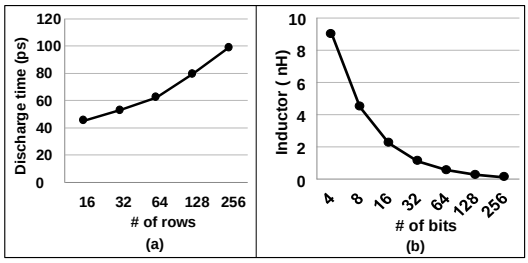


Fig. 6: (a) The discharge time increases with the increase in the number of rows, and this analysis helps us to precisely define the $\frac{T_R}{2}$ time for operating a specific frequency, (b) we identified the resonant inductor sizing by the varying number of bits for a target voltage swing and $\frac{T_R}{2}$ time.

columns. According to our analysis, the $\frac{T_R}{2}$ time reduces with the reduction of the number of associated columns for a fixed inductor. The Table I shows the results of this analysis. We can adjust the pulse width by controlling the delay between S and SD signals. However, we may need to change the pulse width for proper resonant operation due to process variation. To tackle this issue, we have a 4-bit register controlled input signal to the tuned circuit that generates the S and SD signals.

The performance of resonant SRAM depends on the bitline discharge time. We tuned the SRAM instance for a particular number of bits and varied the number of rows to compute the discharge time. Figure 6(a) shows the results of this analysis. As expected, the discharge time increases with the increase of the number of rows. This analysis helps us to define the $\frac{T_R}{2}$ time and optimize our design for a target frequency range.

To properly size the resonant inductor, we consider a target frequency range of 200MHz-1GHz and a bitline discharge time down to 100ps. Also, we use the target resonant voltage swing is approximately two-third of the $V_{DD}$. We varied the number of bits connected to the VL line of Figure 4 and identified the inductor size that achieves the target design parameters. Figure 6(b) shows the results of this analysis. Clearly, increasing the number of bits reduces the inductance requirement. This analysis also helps us to achieve the target power gain in the form of $V_{Rsw}$ even with the varying inductor size. The primary reason for this constant voltage swing is the parallel write drivers' resistance reduces with the

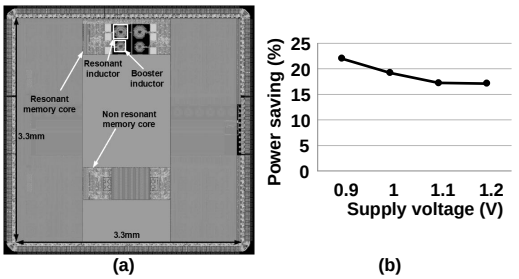


Fig. 7: (a) Die photograph of the test chip, (b) The proposed memory power saving decrease with the increase of supply voltage.

reduction of inductance, resulting in a fixed $Q_f$.

We performed extensive simulations considering corner cases, supply voltage, and temperature variations. Using fast-fast (FF) devices, 1.1V supply voltage, the maximum bitcell write time is 37ps at $125°C$., using slow-slow (SS) devices, 0.81V supply voltage, and the maximum bitcell write time is 79.6ps at $-40°C$.

### *B. Test Chip Experimental Results*

We verified the proposed resonant SRAM architecture by designing a GCrP using a commercial 28nm CMOS technology. We integrate 144KB of SRAM for the GCrP, as shown in Figure 7(a). For comparison, we created a similar GCrP with the same amount of conventional SRAMs. We embed a 2nH resonant inductor and 0.5nH booster inductor for each 8KB of memory instance.

The total resonant memory area is $0.36mm^2$, which consumes only 2% extra silicon area for the additional switching transistor that recycles the energy than the conventional 6T-based SRAM design. We used the top two metals for inductors. The primary goal of this test chip is to verify the power efficiency of the resonant SRAM. The resonant SRAM operating supply voltage ranges from 0.9V to 1.2V, which results in 22% to 17% overall memory power saving compared to the conventional industry standard SRAM architecture with no leakage penalty, as shown in Figure 7(b). The primary reason is the use of the same 6T cells. We set the chip operating frequency 200MHz to 1GHz, which results in 20% to 30% overall memory power saving compared to the non-resonant memory.

## V. Conclusion

In this paper, we presented the first series resonant SRAM architecture to reduce memory power. The proposed architecture uses a booster inductor for the write drivers and a resonant inductor to recycle energy from the SRAM bitlines. We fabricated a test chip using TSMC 28nm CMOS technology. The proposed resonant SRAM can save up to 30% dynamic power with only a 2% area penalty than the conventional CMOS SRAM.s

## References

[1] J. Rabaey, "Low Power Design Essentials. second edition," *Springer Science and Business Media*, January 2009.

[2] E. Karl, Z. Guo, J. Conary, J. Miller, Y. Ng, S. Nalam, D. Kim, J. Keane, X. Wang, U. Bhattacharya, and K. Zhang, "A 0.6 v, 1.5 GHz 84 Mb SRAM in 14 nm FinFET CMOS technology with capacitive charge-sharing write assist circuitry," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 222–229, 2016.

[3] N. Maroof and B. Kong, "10T SRAM using half- $v_{dd}$ precharge and row-wise dynamically powered read port for low switching power and ultralow RBL leakage," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1193–1203, 2017.

[4] R. V. Joshi and M. M. Ziegler, "Programmable supply boosting techniques for near threshold and wide operating voltage sram," in *IEEE Custom Integrated Circuits Conference*, 2017, pp. 1–4.

[5] K. J. Nowka, G. D. Carpenter, E. W. MacDonald, H. C. Ngo, B. C. Brock, K. I. Ishii, T. Y. Nguyen, and J. L. Burns, "A 32-bit PowerPC system-on-a-chip with support for dynamic voltage scaling and dynamic frequency scaling," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1441–1447, Nov 2002.

[6] I. Bezzam, C. Mathiazhagan, T. Raja, and S. Krishnan, "An energy-recovering reconfigurable series resonant clocking scheme for wide frequency operation," *IEEE Transactions on Circuits and Systems I*, vol. 62, no. 7, pp. 1766–1775, July 2015.

[7] H. Fuketa, M. Nomura, M. Takamiya, and T. Sakurai, "Intermittent resonant clocking enabling power reduction at any clock frequency for near/sub-threshold logic circuits," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 2, pp. 536–544, February 2014.

[8] V. Sathe, S. Arekapudi, A. Ishii, C. Ouyang, M. Papaefthymiou, and S. Naffziger, "Resonant-clock design for a power-efficient, high-volume x86-64 microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 140–149, January 2013.

[9] R. Islam and M. Guthaus, "Low-power clock distribution using a current-pulsed clocked flip-flop," *IEEE Transactions on Circuits and Systems I*, vol. 62, no. 4, pp. 1156–1164, April 2015.

[10] N. Tzartzanis and W. C. Athas, "Energy recovery for the design of high-speed, low-power static RAMs," in *International Symposium on Low Power Electronics and Design*, 1996, pp. 55–60.

[11] R. V. Joshi, M. M. Ziegler, and H. Wetter, "A low voltage sram using resonant supply boosting," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 3, pp. 634–644, 2017.

[12] R. V. Joshi, M. M. Ziegler, K. Swaminathan, and N. Chandramoorthy, "Cascaded and resonant SRAM supply boosting for ultra-low voltage cognitive IoT applications," in *IEEE Custom Integrated Circuits Conference*, 2018, pp. 1–4.

[13] F. O'Mahony, C. P. Yue, M. A. Horowitz, and S. S. Wong, "Design of a 10GHz clock distribution network using coupled standing-wave oscillators," in *IEEE/ACM Design Automation Conference*, June 2003, pp. 682–687.

[14] B. Taskin, J. Wood, and I. S. Kourtev, "Timing-driven physical design for VLSI circuits using resonant rotary clocking," in *IEEE International Midwest Symposium on Circuits and Systems*, vol. 1, August 2006, pp. 261–265.

[15] R. Islam, "Low-power resonant clocking using soft error robust energy recovery flip-flops," *Journal of Electronic Testing*, June 2018.