Chapter 7

Ultra-Low-Power Embedded SRAM Design for Battery-Operated and Energy-Harvested IoT Applications

Arijit Banerjee

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76765

Abstract

Internet of Things (IoT) devices such as wearable health monitors, augmented reality goggles, home automation, smart appliances, etc. are a trending topic of research. Various IoT products are thriving in the current electronics market. The IoT application needs such as portability, form factor, weight, etc. dictate the features of such devices. Small, portable, and lightweight IoT devices limit the usage of the primary energy source to a smaller rechargeable or non-rechargeable battery. As battery life and replacement time are critical issues in battery-operated or partially energy-harvested IoT devices, ultra-low-power (ULP) system on chips (SoC) are becoming a widespread solution of chip makers' choice. Such ULP SoC requires both logic and the embedded static random access memory (SRAM) in the processor to operate at very low supply voltages. With technology scaling for bulk and FinFET devices, logic has demonstrated to operate at low minimum operating voltages (V_{MIN}). However, due to process and temperature variation, SRAMs have higher V_{MIN} in scaled processes that become a huge problem in designing ULP SoC cores. This chapter discusses the latest published circuits and architecture techniques to minimize the SRAM V_{MIN} for scaled bulk and FinFET technologies and improve battery life for ULP IoT applications.

Keywords: IoT, SoC, ULP, SRAM, FinFET, assists, canary sensor SRAM

1. Introduction

The revolutionizing Internet of Things (IoT) devices connect us to a new horizon of smart wearable gadgets, home appliances, health monitors, home automation controllers, etc. According to a growth projection of IoT devices by CISCO in 2013, the number of these

connected IoT devices could reach 50 billion by the year 2020 [1]. Among these IoT devices, a significant amount of products would be of wearable or portable categories. Thus, the portability and form factor of such smaller devices restrict the use of power source to smaller batteries. Besides, these mobile IoT devices could harvest energy from ambient light, body heat, etc. energy sources. Based on the power consumption of these IoT devices, the battery life would vary for different applications [2]. However, power storage capacity of smaller batteries, both non-rechargeable and rechargeable, is limited. Therefore, all of these so-called battery-operated portable devices are limited by the battery life, and battery replacement of millions of IoT devices per year could result in millions of dollars in replacement cost.

On the other hand, energy harvesters could transform light, radio wave, and vibration energy to electrical energy that could be a potential solution for battery life and replacement issues. However, the limited harvested power [2] from various energy sources may be insufficient to power IoT devices for applications requiring milliwatt or even hundreds of microwatts of power with the constraints of a smaller form factor. Also, guaranteed availability of energy sources may not be available for long-term application usage. Therefore, batteries remain the primary power source and choice for most of the IoT applications. However, due to self-leakage and energy consumption in IoT applications, the battery life and replacement time of IoT devices are major concerns, which last much less than the shelf-life of batteries of about 10 years. As battery energy density doubles every 10 years [3], which is much slower than Moore's law of the number of transistors doubling every 2 years [3], the low-power circuit solutions show great promises to empower IoT devices for longer battery life.

Every modern-day electronic gadget that has a digital processor in its circuit board, starting from the household micro-oven to Apple's iPhone and the commercial Amazon's cloud servers, uses a fast and power-efficient on-chip memory called the static random access memory (SRAM). The SRAM has three operations: one can write some desired data into a particular memory address location or read some data from a specific memory address or hold the written data to access in the future. Hence, the usual metrics to evaluate an SRAM are (1) the ability to write (write-ability), (2) ability to read (readability), and (3) ability to retain data (data retention) without any operation. Also, there is another metric called read stability that evaluates the stability of unselected bitcell columns while writing a data in selected columns. A simple architecture of the SRAM is given in **Figure 1(a)**, which shows it has an address bus to select an address for a write or read operation. The other pins are a data input bus DIN, a data output bus DOUT, a chip enable signal EN, a synchronous clock signal Clk, and a write and read select signal WRRD. More advanced SRAMs can have additional pins, such as test pins, write and read margin control pins, power management pins, etc. SRAMs are nonvolatile: disconnecting the power supply from the SRAM would result in loss of memory data stored previously. The SRAM typically shares the power rail with the microprocessor's digital circuits (logic core). The SRAM and microprocessor logic core can also have a separate supply rail at the cost of power rail routing, silicon area of the DC-DC converter, chip design time, and overheads. **Figure 1(b)** and **(c)** shows the two usual topologies used in system on chip (SoC) integrated circuits. The advantage of SRAM power rail topology shown in **Figure 1(b)** over **Figure 1(c)** is that it saves silicon area required by the additional on-chip DC-DC converter blocks; those are usually very large compared to the other blocks in the SoC. Due to the square-law dependency of power with supply voltage, one of the best ways for

http://dx.doi.org/10.5772/intechopen.76765

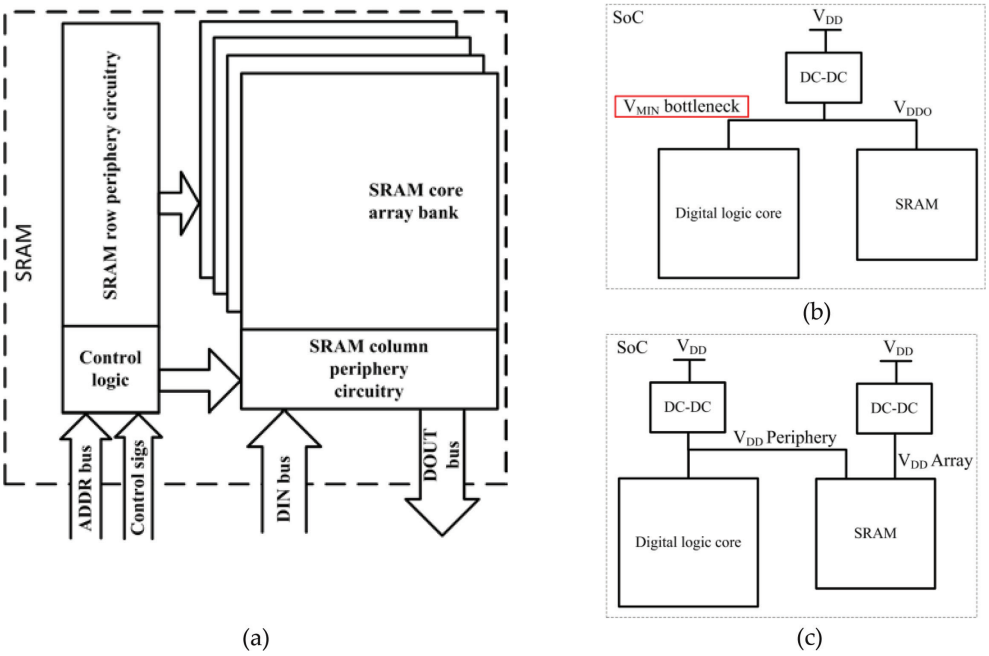

Figure 1. (a) SRAM architecture, (b) digital core and SRAM sharing the same rail, and (c) digital core and SRAM having dual-rail architecture.

low-power operation of an SoC is to lower the supply voltage (V_{DD}) and operate the entire digital microprocessor block at the scaled V_{DD}. Digital logic has been demonstrated to work at subthreshold [4] supply voltages [4–6] (100 mV and lower) that is lower than the threshold voltage (V_T) of bulk MOSFETs, as shown in **Figure 2**, in a MOSFET I_D-V_{GS} curve. However, the conventional 6T SRAM bitcell (**Figure 3(a)**) being a ratioed logic, which shares the same M5 and M6 transistors

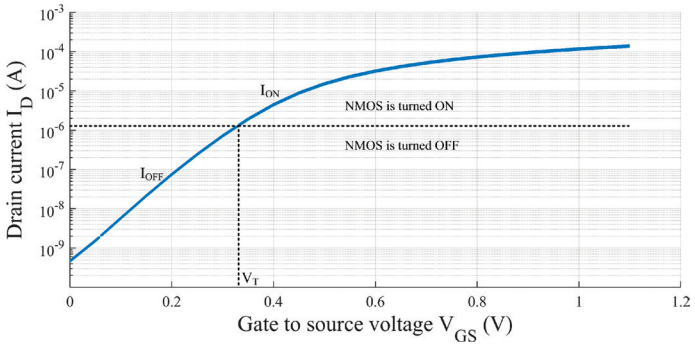

Figure 2. Drain current (I_D) vs. gate-to-source voltage (V_{GS}) plot for an NMOS transistor showing on and off states in 130 nm bulk predictive technology model from Arizona State University. Below the threshold voltage (V_T) of the MOSFET, the transistor is still operable.

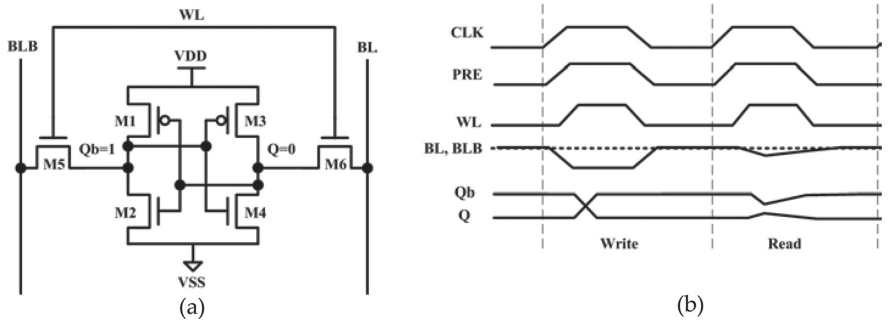

Figure 3. (a) Conventional 6T SRAM and (b) its write and read waveforms.

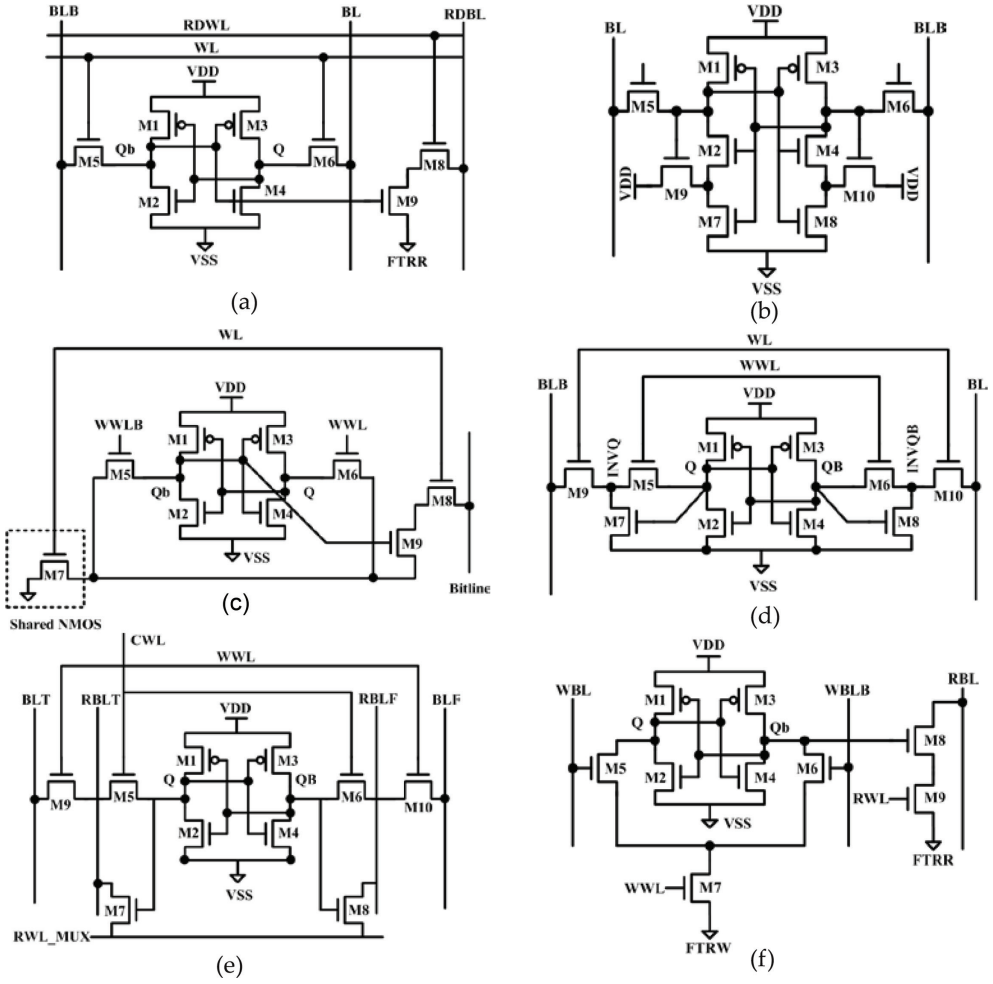

Figure 4. (a) Conventional 8 T SRAM bitcell, (b) Kulkarni's 10 T bitcell, (c) Chiu's 8 T bitcell, (d) Chang's 10 T bitcell, (e) Feki's bitcell, and (f) Arijit's 9 T bitcell.

for write as well as read operation, faces write-ability and read stability challenges across process variation, and the minimum operating voltage (V_{MIN}) of SRAM increases. Thus, sharing the same power rail of the logic core with SRAMs limits the voltage scaling of the SRAM with logic core for low-power operations. Additionally, with technology scaling in nanometer domain, the 14-nm FinFETs experience huge process variation [7], and the conventional high-density (HD) 6T FinFET bitcell (**Figure 3(a)**) with 1:1:1 M1:M2:M5 beta ratios has insufficient write-ability and read stability across process variation. With further technology scaling in 7 nm and smaller processes, it will be very challenging to make the conventional 6T SRAM memory to work, which has been there for decades.

There are mainly two available solutions to address these challenges of 6T SRAM by trading off SRAM area such as alternative bitcells and a write-read peripheral assist (PA) to improve V_{MIN} of 6T SRAM bitcell. The alternative bitcells are a class of bitcells that has lower V_{MIN} or lower energy consumption than the conventional 6T SRAM bitcell. A very popular alternative bitcell is 8 T bitcell, as shown in **Figure 4(a)**. Here, the write and read path are decoupled to improve the write-ability, readability, and read stability of the 8 T SRAM compared to the 6Ts. However, after a certain V_{DD}, even alternative bitcells are inoperable, and one of the popular SRAM schemes comes into the play for further V_{MIN} lowering: peripheral write and read assist techniques. Although the PAs reduce the worst-case SRAM V_{MIN}, it does not remove the SRAM V_{MIN} guardbanding across process variation, which costs additional area and energy penalty in the typical and best case dies. A couple of recently published works address this V_{MIN} guardbanding issue by tracking it using canary sensor SRAM. The canary SRAM extends the SRAM operating range by reducing V_{MIN} guardbanding across process variation, which promises to enable a multitude of IoT applications. This chapter will discuss aforementioned three major techniques that could enable ULP low-V_{MIN} SRAMs for IoT applications as follows. Before delving details into these topics, we need to understand the SRAM design metrics as follows.

2. SRAM write and read design metrics

As discussed earlier SRAM has four different categories of design metrics such as write-ability, readability, read stability, and data retention. The first three categories of design metrics can have static and dynamic measures. Here the static measures are obtained using DC SPICE simulations, and dynamic measures are obtained using transient simulations. Static measures for write and read metrics are easy to evaluate and are widely being used to quantify the SRAM static V_{MIN} across process and temperature corners. On the other hand, dynamic measures for write and read metrics are more accurate to represent an actual SRAM write or read operation; however, they are harder to evaluate and time-consuming. The static measures for write-ability are called write margin (WM) and write static noise margin (WSNM). Both WM and WSNM assume an infinitely long wordline pulse. The WM during a write is defined in two ways: the margin between V_{DD} and WL while BL and BLB are fixed at V_{SS} and V_{DD} and the margin between V_{DD} and BL while WL is fixed at V_{DD}. On the other hand, WSNM has a single definition for measuring the SRAM static noise margin (SNM) when the wordline is turned on. The static measure of readability is the DC read current (I_{read}) drawn from the bitline while reading a bitcell. The static measure of read stability is read static noise margin (RSNM), which assumes an infinitely long wordline pulse too. The measurement technique of RSNM and WSNM using

the SNM measurement technique shown in [8] is widely used across industry and academia. The wordline is turned off during a standby operation, and the corresponding hold metric is called hold static noise margin (HSNM). During the read operation, the internal nodes of the 6T SRAM bitcell are read stressed, and thus, RSNM is the worst-case SNM among the RSNM and HSNM. On the other hand, the quantifiable metric to retain data at the lowest supply voltage is called the data retention voltage (DRV) or at the supply voltage at which the HSNM is almost zero. Due to the reason that the static metrics assume an infinitely long wordline pulse, the measurement of WM is optimistic, and the measurement of RSNM is pessimistic. Moreover, the static metrics does not represent the true nature of the SRAM write and read operations, which has a finite wordline pulse-width. Thus, dynamic metrics play an important role to accurately determine the write-ability, readability, and read-stability metrics and their corresponding V_{MIN} of SRAM. There can be many measures of dynamic metrics, such as dynamic write-ability and readability margins; the critical wordline pulse-width [9] for write-ability, readability, etc.; the failure rate of write-ability, readability, or read stability for a given wordline pulse-width; etc. Among these measures, the measurements of failure rates are the more popular choice to determine the V_{MIN} of SRAM. This section concludes the discussion of SRAM write and read design metrics, which paves the path for discussion to alternative bitcell in the next section.

3. Alternative bitcells for low-power IoT applications

As broad categories of ULP and mid-high performance IoT applications demand to run on modern IoT SoCs, the SoC must be operable throughout a wide range of supply voltages. The SRAM in the SoC for such IoT application is no exception. However, at a lower supply voltage, the conventional 6T high-density (smallest area) bitcell has poor write-ability, readability, and read-stability metrics, such as WM, I_{read}, and RSNM. Across process and temperature variation, these metrics degrade even more, and the conventional 6T SRAM becomes inoperable at lower supplies. Device sizing for write improvement hampers the read stability and vice versa due to shared write and read path and thus is not an option for ULP IoT applications. Moreover, near and below the subthreshold supplies, sizing does not work well to improve WM and RSNM metrics. On the other hand, at lower supply voltage, the soft error rate (SER) [10] increases. The SER can cause soft error disturb (SED) caused by high-energy particle strike that can flip the internal content of the bitcells in an SRAM. Error-correcting codes (ECC) [10] are essential to fix the SED errors; however, it requires additional ECC hardware and memory row or column to fix single-bit single-word errors. Detecting and correcting a multi-bit single-word (MBSW) error is expensive regarding ECC hardware and layout area. An MBSW error is usually lowered using bitline interleaving scheme, which is also known as column muxing. However, in a column mux scenario, selecting a 6T bitcell row for a write using the so-called wordline boosting-type peripheral write assist for V_{MIN} lowering degrades the read stability of the half-selected bitcells, which is known as the half-select issue [16]. The root of the problem in the conventional 6T is the shared write and read path that degrades both the write and read operations in a column mux scenario. Thus, separating the shared path for write and read operation is the desired solution for low-V_{DD} operation of SRAMs.

State-of-the-art alternative bitcells' [11–16] innovations in the last decade, having separate write and read path, show promises for low-V_{DD} operations. Among these bitcells the 8 T

(**Figure 4(a)**) bitcell is very popular and widely used in register files. Here, the write operation is performed using the 6T part of the 8 T bitcell, which is exactly the same as the conventional 6T operation. According to the data, one of the WBIT or WBITB lines goes high while the write wordline (WWL) is turned on. The read operation uses the two transistor read buffers M8 and M9. During a read, the read bitline (RBL) is initially precharged to V_{DD}, and after the read wordline (RWL) turns on the RBL discharges if the internal node Qb is holding a logic "1," else not. This RBL discharge directly drives an inverter or logic gate or a single-ended sense amplifier to generate the read-out signal. Although the 8 T bitcell separates the write and read paths, it suffers from read-stability issues in column mux scenario due to the half-select problem. Thus, an ultra-low voltage (ULV) operation using 8 T may not be viable in scaled technology across process and temperature variation. On the other hand, some of the other alternative bitcells that are shown in **Figure 4**, which includes Kulkarni's [12], Chiu's [13], Chang's [14], Feki's [15], and Arijit's [16] bitcells, show promise for ULV operation. Kulkarni's bitcell (**Figure 4(b)**) uses Schmidt-trigger type topology to have higher read stability and shown to operate down to 160 mV. However, due to feedback in the Schmidt-trigger-type topology, the write and read energy, as well as leakage current of the bitcell, is higher than the other state-of-the-art alternative bitcells. It also suffers from the half-select issue. Chiu's and Wang's bitcell has a unique data-aware cross-point selection in the topology itself, which not only avoid the half-select issue but also serve as a lower energy bitcell. On the other hand, Feki's bitcell has two wordlines (**Figure 4(e)**) that separated the write from read operations and has lower leakage numbers. All of these ULV alternative bitcells show improvement in V_{MIN} or dynamic energy or leakage numbers. However, it does not necessarily mean that any capacity ULV SRAM using any of these alternative bitcell would be suitable for all the ULV application. Where the battery life is extremely important, such as invasive or noninvasive ECG, EEG, or EMG monitoring for patients for a long time, a careful selection of bitcells is required based on total energy per cycle consumption and the duty cycle of the active IoT device.

With the voltage scaling in subthreshold supplies although the dynamic energy per cycle decreases, the cycle time increases due to the exponential relationship of MOSFET drain current with gate supply voltage. Thus, with voltage scaling the leakage energy per operation increases in SRAMs, and there can be a minimum energy point (MEP) [16]. Hence, arbitrary scaling down supply voltage for alternative bitcell arrays using different methods may not be fruitful from the standpoint of energy consumption or battery life. Authors in [16] compare Kulkarni's, Feki's, Chiu's, and Chang's bitcells with Arijit's 9 T that shows the MEP contours are best for Arijit's 9 T bitcell for low-energy biomedical applications due to its lower read, write, and leakage energy per operation. **Figure 5(a)** and **(b)** shows across design knobs (word width and size) the MEP comparison of the abovementioned bitcells as described in prior work [16], which is useful for selecting greener bitcells for low-energy consumption for extending battery life of biomedical devices. Note that all of the alternative bitcells have area penalty and energy tradeoffs compared to the high-density 6T SRAMs. Although alternative bitcells allow us to somewhat lower the V_{MIN} of SRAMs for low-energy operation, there is another widely used design knob, called peripheral assists (PAs), for achieving a low-V_{MIN} in SRAMs. Without the V_{MIN} lowering PAs, even for alternative bitcells, below some V_{DD} doing write and read operation is challenging, such as subthreshold V_{DD}s. The V_{MIN} lowering PAs are discussed next.

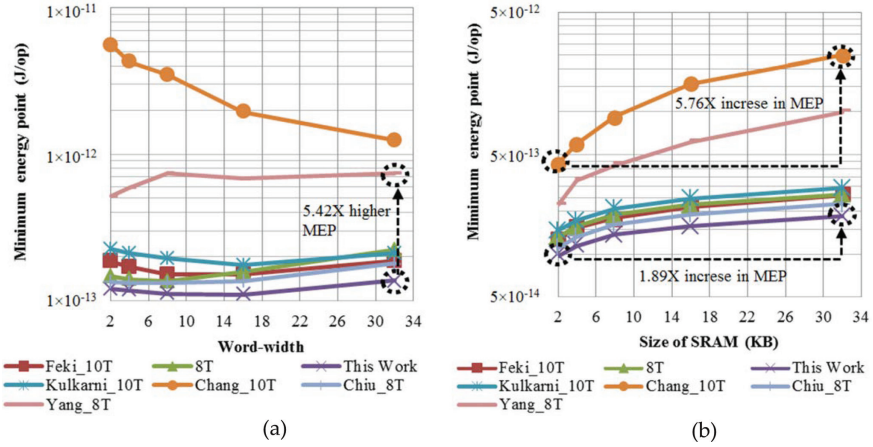

Figure 5. Minimum energy point (MEP) vs. (a) word width and (b) size of SRAM [16].

4. Write and read peripheral assist techniques for low-V_{MIN} applications

In moderate- to high-speed IoT applications, such as 100 MHz–1GHz, an alternative bitcell may not be the choice of SRAM designer due to high timing as well as area penalty. Thus, the lowest area 6T bitcell is still a popular choice for mid- to high-speed IoT applications. However, the 6T V_{MIN} is heavily guardbanded due to process and temperature variation. Thus, lowering V_{MIN} requires write and read peripheral assist [17] (PA) techniques. Moreover, alternative bitcells in ULV application involves the help of PA to have correct write and read functionality across process variation. We define the PAs as a class of circuit techniques used in SRAM periphery that improves the write-ability, readability, and read stability of SRAM bitcells. Mainly, a PA technique would either bump up or lower the wordline or bitline control voltages of the SRAM to make the write or read operation successful, as shown in **Figure 6(a)**. A PA can also decrease the SRAM cycle time by shortening the write or read operations. The PAs are transient in nature and can be classified into write-ability, readability, and read-stability PAs. For the conventional 6T SRAM bitcell, the control signals are mainly wordline and bitline. Thus, an example of write-ability PA would be wordline boosting (WLB) [17] or negative bitline (NBL) [17]. Although V_{DD} and ground (V_{SS}) signals are usually static, they can serve as control signals for SRAM write operation. Thus, V_{DD} lowering and V_{SS} rising [17] are also write assist techniques. On the other hand, to improve the readability or differential development or shorten the differential development time, one can apply a small percentage of WLB (as bigger percentages could induce read-stability issues in half-selected bitcells in column mux scenario) or negative V_{SS} (NVSS). Applying a suppressed wordline in write improves the read stability in half-selected bitcells, which have better RSNM numbers; however, it degrades the write-ability in selected bitcells. Additionally, column-wise boosting the V_{DD} or making the V_{SS} negative during the write operation in half-selected

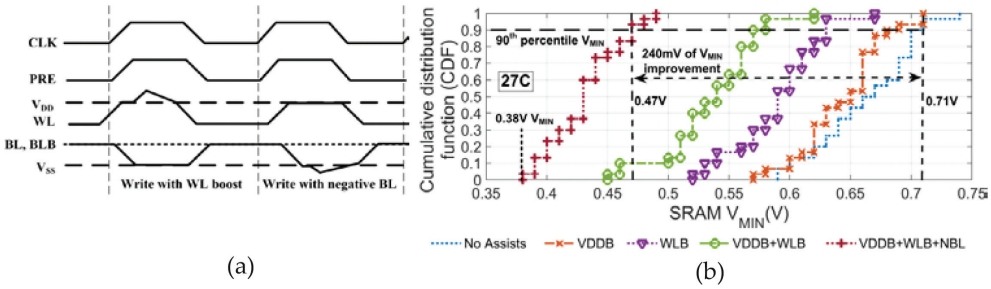

Figure 6. (a) Example of write assists using wordline boost and negative bitline techniques and (b) measured CDF of 256 kb SRAM V_{MIN} showing 90th percentile V_{MIN} improvement of 240 mV using combined assists [V_{DD} boosting (VDDB), WL boosting (WLB), negative bitline (NBL)] [20].

bitcells also improves the read stability. Note that usually increasing the percentage of assist further enhances the write-ability, readability, and read stability but has a limit called assist line contour [17], which is controlled by the V_{MAX} of the process technology. The list of possible PAs for write-ability, readability, and read stability can be found in [17]. PAs can affect the V_{MIN} and yield of SRAMs differently in different technology. Thus, evaluations of PAs are necessary for new scaled-technologies, as past technology trends may not hold true in newer ones.

More than a decade ago, when bulk CMOS technology scaling at 65 nm and lower was facing challenges of higher process variation, the single write or read PAs showed enormous promises to improve the V_{MIN} and yield of 6T SRAMs. However, with the introduction of scaled 28 nm technology, the process variation was so high that the HD 6T bitcell was not writeable in all process corners, especially for the worst case. Post 28 nm bulk the FinFETs become a device fabrication option, and the trend of write-ability issues in the HD 6T bitcell persisted due to huge process variation. Thus, from 28 nm onward applying a particular single write or a read assist may not lower the SRAM V_{MIN} across process variation anymore. Authors in [18] show the use of dual write and read PAs that reduces the V_{MIN} and improves the yield. Moreover, authors in [19] discussed some appropriate combination of PAs (CPAs) that could lower the V_{MIN} further for FinFETs at near-subthreshold supplies, such as a combination of negative bitline with boosting the V_{DD}, etc. One could employ different CPAs based on V_{MIN} lowering application needs. Because write-ability and read stability are more important metrics in FinFET SRAM design, and they often contradict the use of certain assists, such as wordline boosting for write improvement, the SRAM designer must make a careful selection of CPA. Usually, a widely used CPA combination for FinFETs nowadays is V_{DD} underdrive with wordline underdrive [18] schemes.

Moreover with technology scaling the metal width and pitch scale. Thus, there exist challenges of electro-migration, IR drop, and cross talk issues, which could restrict the use of a specific assist or limit the size of an SRAM bank. With the explosion of IoT application needs, ULP SoCs are targeted to run ultra-low energy as well as high-speed applications from time to time. Thus, voltage scaling down to near-subthreshold or deep-subthreshold supplies for SoC is a need nowadays. As logic V_{MIN} easily scales down to lower V_{DD}s, but SRAM V_{MIN} is

comparatively higher due to process and temperature variation, the overall V_{MIN} of IoT SoCs increase that has a logic core and SRAM sharing the same power rail. Note that splitting the logic and SRAM power rail, called dual-rail technology, requires additional silicon area and power routing costs due to additional DC-DC converters. Thus, lowering SRAM V_{MIN} is essential for wide-range ULP IoT applications depending on the speed requirements of the applications. Authors [20] show for the first time the use of three combined PAs (NBL + VDDB + WLB) that lowered the conventional 6T V_{MIN} from 0.71 V (90%) to 0.47 V as shown in **Figure 6(b)** using a measured cumulative distribution function. The work reports the total V_{MIN} improvement as 240 mV in a commercial bulk 130 nm technology. This work shows more than 300X active power lowering using the triple CPA technique.

However, it is not imperative that always lowering the V_{MIN} would help to reduce the SRAM energy consumption. Lowering V_{MIN} requires energy penalty due to the use of assists, which could lead to the cause of overall SRAM energy could increase in some case. The total SRAM energy could increase with certain higher assist percentages and lower V_{MIN} is not always an intended requirement for low-energy applications. However, if the SRAM shares the same power rail with the logic core, as shown in **Figure 1(b)**, the energy savings from voltage scaling in the logic core could be much higher than the energy increase in SRAM, and thus, in this scenario only it might help.

Although write and read PAs usually improve the V_{MIN} guardbanding, it does not remove it entirely. Moreover, due to design and application of PAs, we trade off additional silicon area and energy for SRAM and overall SoC sharing the same power rail with SRAM. Additionally, with the design for the worst-case approach, the nominal and the best-case corner dies suffer from additional area and energy overhead. Thus, an important research question emerges: how to minimize this additional V_{MIN} guardbanding of SRAMs across process and temperature variation? The answer lies in tracking the SRAM V_{MIN} using in situ canary sensor SRAMs that helps to apply CPA for individual dies differently across process and temperature variation. The next section describes the canary SRAM techniques.

5. Canary sensor SRAMs for V_{MIN} tracking and guardband lowering

The story of canary SRAMs ties to the story of the canary in a coalmine. Eighteenth-century coal miners used to carry this beautiful yellow canary bird for poisonous gas, such as methane detection. A moderate presence of such gases could be fatal to human beings. If there is a significant level of methane being present in the mines, the canaries used to feel sick. By observing the canaries, thus, the miners get enough time to evacuate the coalmines. The Canaries from the standpoint of a circuit could mean a weak circuit that fails earlier than the main circuit. The canary circuits are first being introduced as canary flip-flops [21]. Later in the year 2007, the authors in [22] show canary techniques could reduce the data retention voltage (DRV) of an SRAM, thus saving a huge amount of leakage power for ULP applications. This work indicates that canary SRAM bitcells are a modified version of the SRAM cells; those use an additional bias control knob to weaken the DRV of sets of canaries to tune to fail them

earlier than the population of the core SRAM bitcells [22]. A bias generator circuit is used to generate the bias voltage for the canaries in a row. A failure detector senses the canary retention failures in a closed loop. Thus, canaries could achieve a failure point before the SRAM DRV in each dies fabricated that lower the SRAM DRV and leakage current. This technique avoids the design for the worst case using canary-based DRV tracking.

In the year 2014, the authors in [23] demonstrated a theory for dynamic write V_{MIN} tracking for the conventional 6T SRAM. This work introduces the term reverse assists (RA) as one of the canary design knobs. As discussed earlier, the peripheral assists (PA) improve the write-ability or readability of the SRAMs. On the contrary, the RA **Figure 7(a)** degrades the write-ability or readability of the canaries to fail earlier than the population of core SRAM, as shown in **Figure 7(b)**. Thus, with the increase of the RA percentage, the canary distribution of write V_{MIN} would shift to the right-hand side from distribution A to B to C. A user can tune the failure point of the canaries by selecting the proper reverse assist percentages or settings [23]. Another input design knob that helps to tune the canaries at a desired V_{MIN} failure point is the failure threshold condition (F_{th}) [22], which defines the no. of canary failures that correspond to a threshold failure point. The work also derives a mathematical formulation for dynamic write V_{MIN} tracking as shown in Eqs. (1) and (2) [23]. Here the meanings of the variables of the equations are described in [23]. Here, the two equations relate the input and output SRAM design knobs and metrics to the canary design knobs and metrics. The work explains how to calculate the output metric named canary chip failure probability. The intended SRAM bit failure rate vs. V_{MIN} data is calculated first. Then from the canary failure rate vs. V_{MIN} data, the corresponding canary bit failure rate p_f is calculated. This serves as the input data to Eq. (2) [23] for the calculation of the output metric of canary chip failure probability P_{fc}. The authors show that one can achieve the desired canary-chip-failure-probability either by selecting a smaller no. of canaries with larger reverse assist voltage strength or the vice versa, as shown in **Figure 8** [23]. The work further shows that for a fixed reverse assist voltage to track the V_{MIN} of a bigger SRAM, more number of canaries are required. Moreover, with the same reverse assist voltage, increasing the SRAM yield requires more number of canary bitcells to track the corresponding SRAM's V_{MIN} and so on.

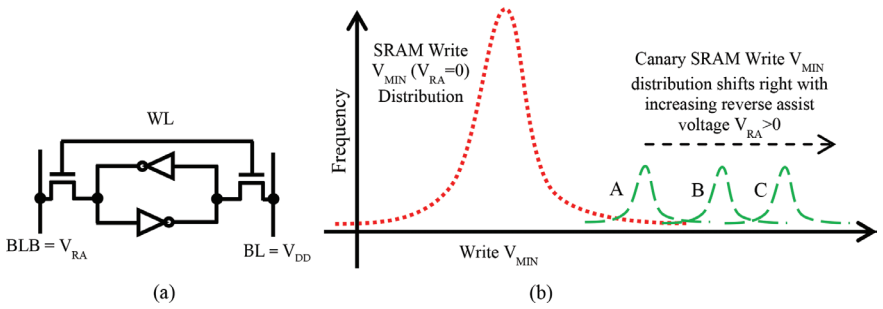

Figure 7. (a) SRAM write operation using bitline-type reverse assist and (b) write V_{MIN} distributions with a reverse assist (A, B, Cs are canary V_{MIN} distributions) [23].

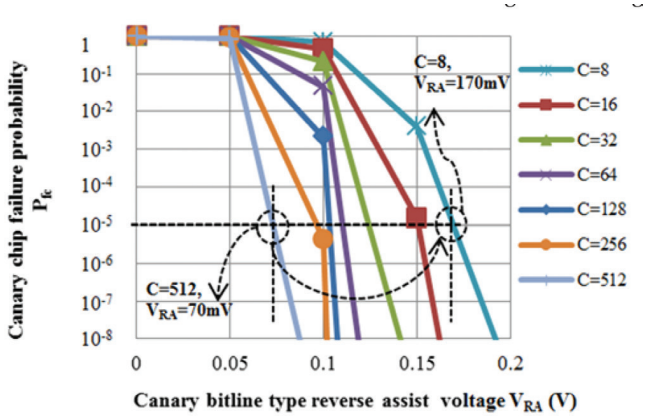

Figure 8. Canary chip failure probability vs. reverse assist voltage for 1 million SRAM bitcells with 95% yield @ TT_85C [23].

The work [23] shows the proposed bitline-type peripheral reverse assist circuit, as shown in **Figure 9**. The peripheral RA circuit uses a configurable NMOS-NMOS voltage divider to pass the generated voltage using an analog demultiplexer to the BL or BLB lines controlled by data D and data-bar Dbar. The reverse assist voltage generation can be disabled for normal write mode by asserting the AON signal to logic "0." The proposed block diagram of the integrated canary SRAM architecture is shown in [23], which is physically adjacent to the SRAM itself that shares the power rails. However, for independent write and read operations, at the canary and SRAM boundary, the bitlines are disconnected. The advantage of canary being an independent memory permits simultaneous operation to track voltage droops occurring at the SRAM-canary power rails to take actions if the canary SRAM fails. Such actions include either stopping the SRAM operation or lowering the SRAM clock frequency to prevent voltage scaling further or selecting an apt PA to lower the V_{MIN} further. The proposed algorithm in this work starts with an initial V_{RA} voltage and writes and reads canary rows to compare if the data written is correct. If the canary write operation is successful, the V_{RA} is increased else lowered gradually to reach the minimum V_{RA} settings. Unless the minimum V_{RA} setting is reached, the dynamic voltage and frequency scaling (DVFS) is allowed else the DVFS has to be stopped, as reaching the minimum V_{RA} would indicate the SRAM V_{MIN} is reached. The minimum V_{RA} setting would vary with the SRAM and canary input design knobs. The work also shows the area and power tradeoffs for SRAM and canary design knobs. It shows that for an increase in the number of canary bits, the normalized canary area and power overhead are amortized in bigger SRAM and increase with smaller capacity and so on. The work [23] showcases interesting results revealing that due to write V_{MIN} tracking, the canaries can save a minimum of 31% in SS corner dies and a maximum of 51.5% in FS corner dies compared to the worst-case SF corner dies.

Authors in [24] first show a working prototype of the canary SRAM in a commercial 130 nm technology that reveals the necessary properties of canary SRAM to track SRAM V_{MIN}. The work further shows a proof of concept V_{MIN} tracking canaries that fail earlier than the SRAM starts to fail, which is controllable using the canary design knobs (F_{th} and RAS) post-fabrication. The architecture of the SRAM is shown in **Figure 10**, which is similar to the [23].

http://dx.doi.org/10.5772/intechopen.76765

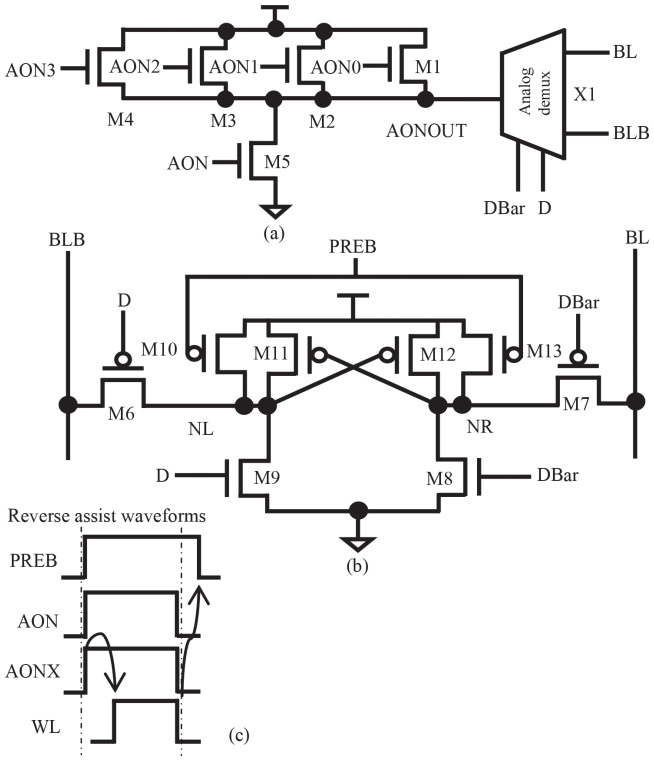

Figure 9. (a) Canary SRAM reverse assist circuit. (b) Canary write driver. (c) Reverse assist waveforms [23].

The testchip includes an 8kb core SRAM, 512 kb canary SRAM, a memory BIST (MBIST), a canary BIST (CBIST), and boundary scan chain blocks. The 6T bitcells used in both canary and core SRAM are same; it uses an external BLVRA voltage to apply as reverse assists to the canary SRAM. Both the MBIST and CBIST architecture are similar to a traditional MBIST [25]; however, they are specialized in measuring the number of bit failures in the core and canary SRAM. This work characterizes some important properties of canary SRAM that helps to track the core SRAM write V_{MIN} (WV_{MIN}). The authors show that using BLVRA and WLVRA reverse assists across different voltage, frequency, and temperatures (VFT), the canary failure curve shifts distinctly compared to each other. Without this distinction in shifting of failure curves, canary SRAM would not work, as there will be no way to tell if the input design knobs are changed, such as VFT. As discussed earlier, this work shows the first silicon proof that canaries can be tuned to fail earlier than the core SRAMs.

With the intuition presented in [23, 24] the authors in [20] show a closed-loop 256 kb self-tuning SRAM that can automatically track the SRAM V_{MIN} using canaries and apply apt write-read PAs to improve the V_{MIN} based on frequency needs for ULP IoT application. This work shows a 67% extension of operating voltage from 1.2 to 0.38 V deep into subthreshold supplies. Reverse assists are used to track the core SRAM V_{MIN} using canaries to allow

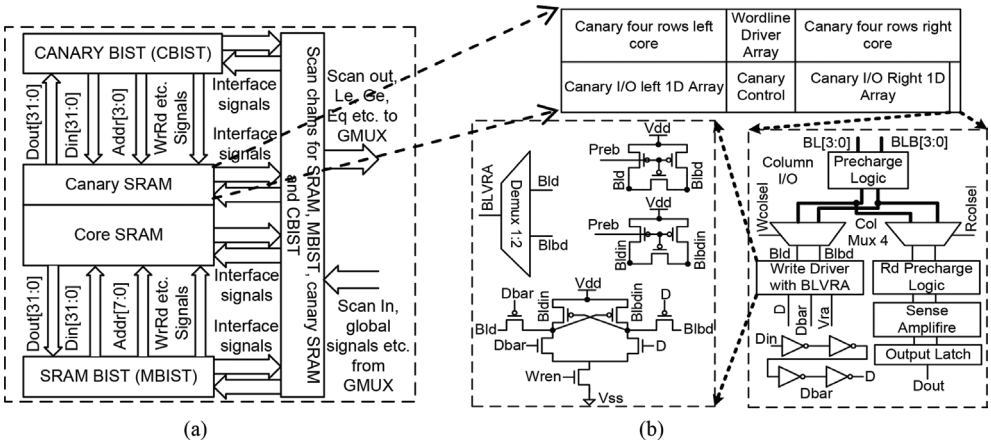

Figure 10. (a) Block diagram (not in scale) of the memory block and (b) block diagram (not in scale) of the canary SRAM column periphery (I/O) and BL-type reverse assist [24].

a closed-loop control of the system supply voltage at an intended operating frequency. The system uses write and read combined PA (CPA) along with in situ canary sensor SRAM-based V_{MIN} tracking to maximize the operating range of the SRAM into subthreshold supplies. This work meets the design needs for SRAMs of highly variable IoT applications while retaining the density of the conventional 6T bitcells. As the battery-operated or harvested energy IoT devices have an operating range of 10kHz to 10 MHz [26, 27], it is needed as a highly versatile feature to expand the 6T SRAM operating range to ULV supply voltages for low power operation. PAs can lower SRAM V_{MIN}; however, selecting the best CPA depends on the supply voltage that could influence the power-performance tradeoff.

This work uses write assists NBL and WLB along with read stability assist VDB to achieve a 90% V_{MIN} of 0.47 V compared to the other assist combinations as well as a no-assist case (shown in the CDF plot in **Figure 6** [20]). However, CPA alone requires a V_{MIN} guardbanding that ensures all chips functioning across PVT variation, hampering potential power savings. Canaries play a vital role to extend the power saving achieved using CPA using runtime determination of V_{MIN} that allow us to reduce the guardbanding. The block diagram of the proposed system is shown in **Figure 11**. The SRAM testchip comprises a 256 kb SRAM with 2 kb integrated canaries, a PA controller (ASC), a frequency to digital converter (FDC), an MBIST, and a CBIST. This architecture shares the SRAM periphery with canary sensors, such as write drivers, sense amplifiers, pre-charge circuits, etc. The RA circuit uses a wordline slope degrading programmable control for canaries.

The work [20] employs a self-tuning algorithm described in [20] that tracks SRAM V_{MIN} dynamically, which also adjust the supply voltage and the selection of PAs. The algorithm uses the FDC to measure and convert the input clock frequency to a digitized output and to initialize the off-chip low dropout (LDO) regulator to an initial V_{DD} programmed by a given look-up table (LUT). Using the ASC the algorithm chooses required PAs based on the LUT,

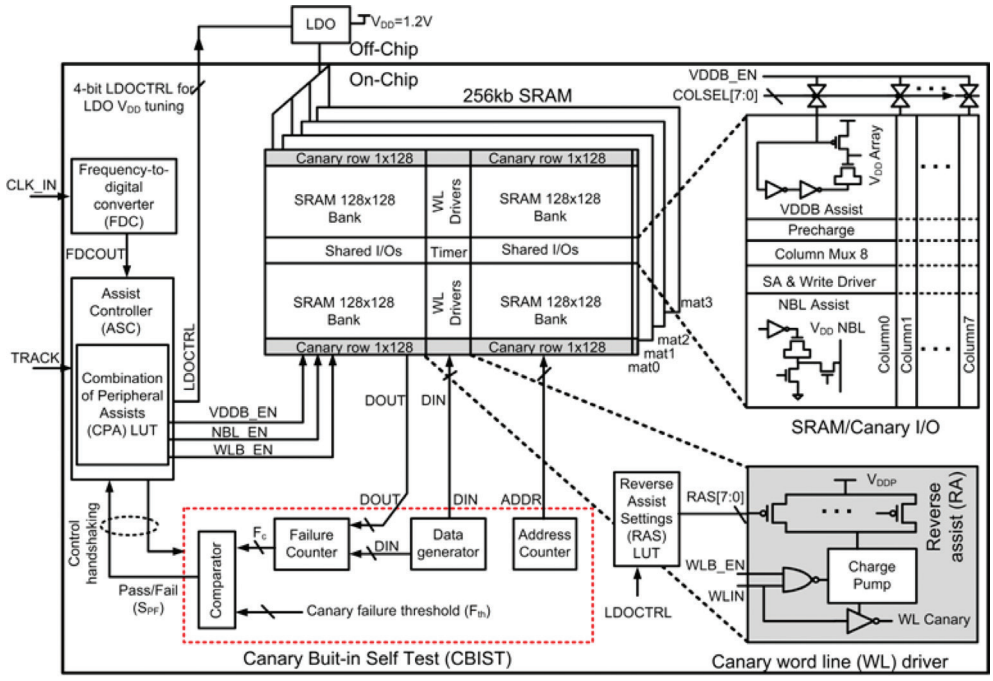

Figure 11. System-level block diagram for the 256 kb 6T self-tuning SRAM subsystem showing subcomponents [20].

and the CBIST iterates the canary write and read operations across all canary addresses to compare and determine if the canary failure (Fc) crosses the canary failure threshold condition (F_{th}). Based on the comparison, the CBIST generates a control signal for the ASC to increase or decrease the LDO supply voltage accordingly. Therefore, the closed loop tracking using self-tuning completes once the canary failure point is reached, which indicates the approaching SRAM V_{MIN}. Once the canary V_{MIN} is tuned to the SRAM V_{MIN} using F_{th} and RA settings, the worst-case SRAM bitcells are mapped into canaries, and the canary sensors track properties of the worst-case SRAM bitcells across a range of voltage, frequency, and temperature (VFT) variations. The authors show measured tracking of SRAM V_{MIN} across VFT variation as shown in **Figure 12** [20]. The canary sensors, system components without the BISTs, and CPA have reported overheads of 0.77, 1.8, and 2.8%, respectively. The work allows V_{DD} scaling using CPA at the 90th percentile worst-case V_{MIN} of 0.47 V with guardbands that reduces 337X active power. Moreover, enabling canary-based V_{MIN} tracking provides a 4.3X power savings by removing the V_{MIN} guardbanding to achieve up to 1444X active power savings at 0.38 V [20]. The authors show using CPA and in situ canary-based tracking down to 0.38 V gives a 12.4X leakage savings, too. The canary-based V_{MIN} tracking is scalable to lower technologies such as 45 and 32 nm, which shows promise to reduce the effect of process variation in FinFET SRAM in the highly-variant 7 nm and beyond technology nodes for a wide range of IoT applications.

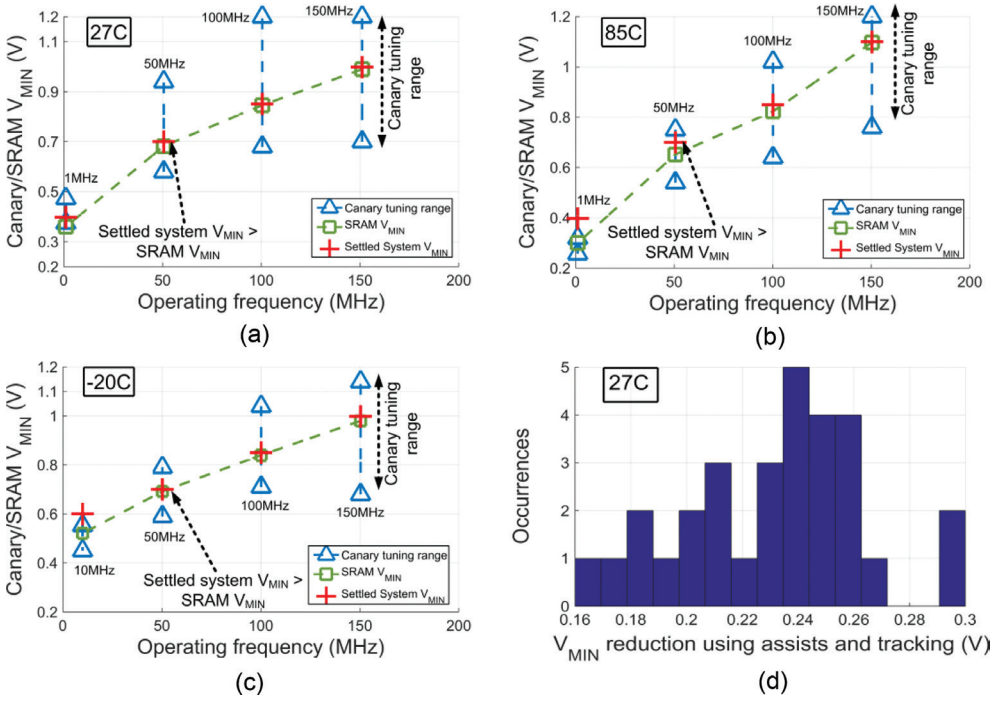

Figure 12. Measured canary V_{MIN} tracking across clock frequencies [1 or 10, 50, 100, and 150] MHz and temperatures (a) 27°C, (b) 85°C, and (c) −20°C, showing V_{MIN} tuning range, and (d) the distribution of overall V_{MIN} reduction using assist and tracking [20].

6. Discussion

Energy consumption in billion node IoT networks is expected to increase, as the total no. of IoT devices may reach 50 billion by the year 2020. A portion of these massive numbers of IoT devices will be plugged into the outlets in homes, factories, and outdoor settings. On the other hand, a huge number of IoT devices will be battery-operated or energy-harvested portable systems. The billions of IoT devices plugged into the outlets will draw power from the energy grid resulting in millions of dollars in energy bills and will increase the carbon footprint of this planet. Moreover, with a shorter battery life and replacement time, supporting billions of battery-operated IoT devices will require a massive production of portable batteries increasing the carbon footprint of Earth, too. Reducing the carbon footprint of these IoT devices requires reduction of power consumption, usage of low voltage operation for quadratic energy savings, and harvesting energy from the environment, which will require ULP IoT SoCs to reduce the energy cost and improve the battery life for a greener IoT electronics. However, technology scaling in the latest 7 nm FinFET and beyond will become a hindrance to lower operating voltage of the widely used embedded SRAMs, which shares the same power line with the digital core in ULP IoT SoCs. This chapter reviews some of the state-of-the-art SRAM design techniques, which are promising candidates for reducing power

consumption in greener IoT applications, such as alternative bitcell topologies, a combination of peripheral assists, and in situ canary-based V_{MIN} tracking for guardband lowering.

7. Conclusions

Technology scaling in FinFET devices 7 nm node and beyond is going to experience a higher degree of process variation, which could affect the design and production of so-called lowest area 6T SRAM memory cells used in modern IoT system on chips. Based on the latest published works, there are three key directions to solve this issue. One of the directions is to use appropriate alternative bitcells for SRAMs trading off core array area that will enable ultra-low energy and lower leakage memory operation to sustain a longer battery life for portable home automation, wearable, and biomedical IoT applications. For low-cost system on chips using 6T SRAMs supporting low-power and mid- to high-speed applications, the use of appropriate combined peripheral assists is essential for a low-V_{MIN} application. Although the combined assist lowers the V_{MIN} and improves the SRAM yield, it does not eliminate the costly V_{MIN} guardbanding due to process and temperature variation. To remove or minimize this V_{MIN} guardbanding, the in situ canary sensor SRAM shows great promises for V_{MIN} tracking across voltage, frequency, and temperature variation. Combined peripheral assists along with canary sensor SRAM show promise for improvement in the power consumption of IoT systems by more than 1000X supporting a wide range of IoT application in a single SoC. Hence, to support a wide range of greener IoT applications, SRAM designers need to choose appropriate design techniques, such as alternative bitcells, combined peripheral assist, and in situ canary sensor SRAMs to enable technology scaling for SRAMs in 7 nm node and beyond.

Conflict of interest

The author has no conflict of interest.

Author details

Arijit Banerjee

Address all correspondence to: ab9ca@virginia.edu

University of Virginia, Charlottesville, Virginia, USA

References

[1] Evans D. The Internet of Things, How the Next Evolution of the Internet Is Changing Everything. CISCO, IBSG. April 2011. http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[2] Vullers RJM, van Schaijk R, Doms I, Van Hoof C, Mertens R. Micropower energy harvesting, Solid-State Electronics, Volume 53, Issue 7, 2009, Pages 684-693. DOI: 10.1016/j.sse.2008.12.011 ISSN 0038-1101

[3] Rabey J. Low Power Design Essentials. USA: Springer; 2009. 6 p. DOI: 10.1007/978-0-387-71713-5

[4] Bryant A et al. Low-power CMOS at Vdd=4kT/q. Device Research Conference. 2001. pp. 22-23

[5] Hwang ME. Supply-voltage scaling close to the fundamental limit under process variations in nanometer technologies. IEEE Transactions on Electron Devices. Aug 2011;**58**(8):2808-2813

[6] Roy K, Kulkarni J, Hwang M. Process-tolerant ultralow voltage digital subthreshold design. In: IEEE Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems. Jan 2008. pp. 42-45

[7] Wang X, Cheng B, Brown AR, Millar C, Asenov A. Statistical variability in 14-nm node SOI FinFETs and its impact on corresponding 6T-SRAM cell design. In: 2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC), Bordeaux. 2012. pp. 113-116

[8] Seevinck E, List FJ, Lohstroh J. Static-noise margin analysis of MOS SRAM cells. IEEE Journal of Solid-State Circuits. Oct 1987;**22**(5):748-754. DOI: 10.1109/JSSC.1987.1052809

[9] Nalam S, Chandra V, Aitken RC, Calhoun BH. Dynamic write limited minimum operating voltage for nanoscale SRAMs. In: 2011 Design, Automation & Test in Europe, Grenoble. 2011. pp. 1-6. DOI: 10.1109/DATE.2011.5763081

[10] Tsiligiannis G et al. SRAM soft error rate evaluation under atmospheric neutron radiation and PVT variations. In: 2013 IEEE 19th International On-Line Testing Symposium (IOLTS), Chania. 2013. pp. 145-150

[11] Verma N, Chandrakasan AP. A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy. IEEE Journal of Solid-State Circuits. 2008;**43**:141-149

[12] Kulkarni JP, Kim K, Roy K. A 160 mV robust schmitt trigger based subthreshold SRAM. IEEE Journal of Solid-State Circuits. 2007;**42**:2303-2313

[13] Chiu Y-W, Lin J-Y, Tu M-H, Jou S-J, Chuang C-Y. 8T single-ended sub-threshold SRAM with cross-point data-aware write operation. In: 2011 Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED); 1-3 August; Fukuoka, Japan. 2011. pp. 169-174

[14] Chang I-J, Kim JJ, Park SP, Roy K. A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS. IEEE Journal of Solid-State Circuits. 2009;**44**:650-658

[15] Feki A, Allard B, Turgis D, Lafont J, Ciampolini L. Proposal of a new ultra low leakage 10T sub threshold SRAM bitcell. In Proceedings of the International SoC Design Conference (ISOCC); 4-7 November 2012; Jeju Island, Korea. 2012. pp. 470-474

[16] Banerjee A, Calhoun BH. An ultra-low energy subthreshold SRAM bitcell for energy constrained biomedical applications. Journal of Low Power Electronics and Applications. 2014;**4**(2):119-137

[17] Mann RW, Nalam S, Wang J, Calhoun BH. Limits of bias based assist methods in nanoscale 6T SRAM. In: 2010 11th International Symposium on Quality Electronic Design (ISQED); San Jose, CA. 2010. pp. 1-8

[18] Karl E et al. 17.1 A 0.6V 1.5GHz 84Mb SRAM design in 14nm FinFET CMOS technology. In: 2015 IEEE International Solid-State Circuits Conference–(ISSCC) Digest of Technical Papers; San Francisco, CA. 2015. pp. 1-3

[19] Banerjee A, Kamineni S, Calhoun BH. Multiple Combined Write-Read Peripheral Assists in 6T FinFET SRAMs for Low-VMIN IoT and Cognitive Applications. In: 2018 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED); Bellevue, WA; 23-25 July 2018:1-6

[20] Banerjee A, Liu N, Patel HN, Calhoun BH, Poulton J, Gray CT. A 256kb 6T self-tuning SRAM with extended 0.38V–1.2V operating range using multiple read/write assists and VMIN tracking canary sensors. In: 2017 IEEE Custom Integrated Circuits Conference (CICC); Austin, TX. 2017. pp. 1-4

[21] Calhoun BH, Chandrakasan AP. Standby power reduction using dynamic voltage scaling and canary flip-flop structures. IEEE Journal of Solid-State Circuits. Sept. 2004; **39**(9):1504-1511

[22] Wang J, Calhoun BH. Techniques to extend canary-based standby V_{DD} scaling for SRAMs to 45 nm and beyond. IEEE Journal of Solid-State Circuits. Nov. 2008; **43**(11):2514-2523

[23] Banerjee A, Sinangil ME, Poulton J, Gray CT, Calhoun BH. A reverse write assist circuit for SRAM dynamic write VMIN tracking using canary SRAMs. In: Fifteenth International Symposium on Quality Electronic Design; Santa Clara, CA. 2014. pp. 1-8

[24] Banerjee A, Breiholz J, Calhoun BH. A 130nm canary SRAM for SRAM dynamic write VMIN tracking across voltage, frequency, and temperature variations. In: 2015 IEEE Custom Integrated Circuits Conference (CICC); San Jose, CA. 2015. pp. 1-4

[25] Kincel A, Balaz M. MBIST for LEON3 processor core cache. In: 2013 IEEE 16th International Symposium on Design and Diagnostics of Electronic Circuits Systems (DDECS). 2013. pp. 287-288

[26] Kwong J, Chandrakasan AP. An energy-efficient biomedical signal processing platform. IEEE Journal of Solid-State Circuits. July 2011;**46**(7):1742-1753. DOI: 10.1109/JSSC.2011.2144450

[27] Roy A, Grossmann PJ, Vitale SA, Calhoun BH. A 1.3μW, 5pJ/cycle sub-threshold MSP430 processor in 90nm xLP FDSOI for energy-efficient IoT applications. In: 2016 17th International Symposium on Quality Electronic Design (ISQED); Santa Clara, CA. 2016. pp. 158-162