On Low-Leakage CMOS Switches

Bo Wang[1], Shiwei Wang[2] and Man-Kay Law[3]

Division of Information and Computing Technology, College of Science and Engineering
Hamad Bin Khalifa University, Doha, Qatar[1]
Centre for Electronics Frontiers, Zepler Institute for Photonics and Nanoelectronics
University of Southampton, UK[2]
State Key Laboratory of Analog and Mixed-Signal VLSI, AMSV, University of Macao, Macao, China[3]

***Abstract*—Continuing CMOS process scaling to favor the design of high-performance digital systems has resulted in many issues for precision analog design, and one of which is the detrimental transistor leakage. This paper focuses on the analysis and design of low-leakage switches. Specifically, transistor leakage mechanisms and the evolution of low-leakage switch design techniques are revisited. Different schemes to achieve transistor channel and body leakage reduction are discussed. In addition, we propose a low-leakage switch that can operate for a wide temperature range. At 200 °C, it achieves 130× and 8× lower leakage than the transmission gate and the popular analog T-switch, respectively.**

***Index Terms*—CMOS switch, low-leakage CMOS switch, CMOS switch array, transistor leakage compensation.**

I. Introduction

CMOS process scaling has contributed to the performance improvements of digital systems such as speed-boosting, power saving, and form-factor reduction, etc. [1]. However, it raises many issues for analog design, one of which is the increased transistor leakage due to the lower threshold voltage (e.g., 0.2 V standard V_{th} in 7 nm FinFET process [2]) and thinner equivalent gate oxide (e.g., sub-nm). For charge-based circuits like switched-capacitor circuits [3], sample-and-hold circuits (S/H), highly duty-cycled circuits [4], or circuits operating at high temperatures [5], device leakages must be minimized to maintain signal integrity.

Over the years, transistor leakage mechanisms have been analyzed and modeled comprehensively [6]. Various process techniques, such as halo doping, high-κ metal gate, triple gate oxide, etc., were adopted to reduce transistor leakages. Nevertheless, many designs in deep sub-micron processes still need to use high-V_{th} devices for critical signal paths like switch or amplifier input pairs to attain an acceptable leakage level. Though using a smaller supply could result in lower transistor leakage [7], it sacrifices the signal dynamic range and is not an ideal solution.

Among all the circuits, the CMOS switch is a vital circuit block. It plays a role in almost all analog circuits for sampling, channel multiplexing, dynamic element matching, chopping, etc. A robust low-leakage switch is therefore essential to underpin high system performance. In this paper, different circuit techniques to suppress switch leakages (especially in its off-state) are revisited and discussed in terms of functionality, complexity, and performance. A new low-leakage switch optimized for high-temperature operation is presented and compared with the prior arts. Also, the difficulties of designing switch arrays for emerging applications like high-density neural interface are discussed. For clarity, techniques presented in this paper only employ standard CMOS processes.

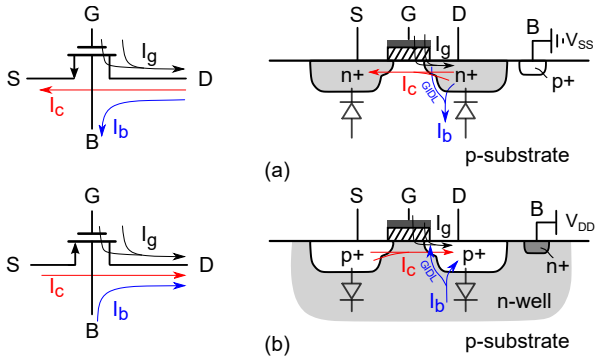

Fig. 1. Simplified leakage model for (a) NMOS, and (b) PMOS transistor (only showing drain leakages). The source and drain are interchangeable, and the leakage current directions could vary with transistor bias conditions.

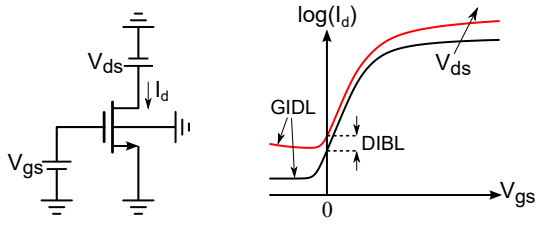

Fig. 2. Illustration of DIBL and GIDL of an NMOS transistor.

II. Transistor Leakage Mechanism Revisit

A simplified transistor leakage model is shown in Fig. 1, which does not include the punch-through current between the source and drain as it will be avoided in actual designs either by using long channel devices or by limiting the signal swing. To explain different leakage mechanisms, in the following, we use the leakage currents from/to the drain terminal of an NMOS transistor as an example, including I_{c} through the channel, I_{b} to the body, and I_g from the gate [8].

The transistor's channel leakage I_{c} is the result of subthreshold conduction enhanced by the drain-induced barrier lowering

(DIBL) effect (see Fig. 2), and can be expressed as (valid for $V_{gs} \ll V_{th}$) [9]

$$I_c = \underbrace{I_{const}\frac{W}{L}\cdot e^{(V_{gs}-V_{th})/n\phi_T}\cdot\left(1-e^{-V_{ds}/\phi_T}\right)}_{\text{subthreshold conduction}}\cdot\underbrace{\left(e^{\eta V_{ds}/n\phi_T}\right)}_{\text{DIBL}} \quad (1)$$

where I_{const} of 100 nA is commonly used to define the transistor's threshold voltage (channel current equals to 100·W/L nA when $V_{gs} = V_{th}$ [10]), ϕ_T is the thermal voltage, n is the subthreshold swing coefficient, and η is the DIBL coefficient (a small value, e.g., 0.08 [6]). As in (1), I_c can be reduced exponentially by reducing V_{gs}, $|V_{ds}|$ or increasing V_{th}. Meanwhile, I_c increases with temperature via the complex temperature dependency of V_{th} and ϕ_T.

The transistor's drain to body leakage I_b mainly consists of the reverse-biased parasitic diode leakage (the small band-to-band tunneling occurs in this diode is not discussed here) and the gate-induced drain leakage (GIDL). It can be expressed as

$$I_b \approx \underbrace{I_s(e^{-V_{db}/n\phi_T}-1)}_{\text{parasitic diode leakage}} + \underbrace{I_{GIDL}(V_{dg},V_{ds},V_{db})}_{\text{GIDL}} \quad (2)$$

where I_s is a function of drain diffusion area, transistor doping profile, and temperature ($\propto T^3$). I_{GIDL} is temperature-independent and has a complex dependency on the transistor biasing [6], and it increases drastically when V_{dg} exceeds the silicon bandgap (e.g., becomes a few orders of magnitude larger than the parasitic diode leakage at room temperature [11]). As shown in Fig. 2, GIDL is the minimum achievable drain leakage of an off-state transistor. According to (2), one can reduce I_b by minimizing the transistor size and voltage drops between different terminals of the transistor.

The transistor's gate to drain leakage I_g is due to tunneling and hot carrier injection between the gate and channel, and the edge direct tunneling via the overlap region of the gate and drain [12]. This current increases exponentially with gate-oxide thickness scaling. It is hard to handle the gate tunneling leakage at the circuit level unless using a lower supply.

III. Low-Leakage CMOS Switch Design

After figuring out the transistor's leakage mechanisms, different circuit techniques can be used to design low-leakage switches, mainly to suppress the transistor's channel and source/drain-to-body leakages.

A. Channel Leakage Reduction Techniques

1) Stacked MOS Switch: as in (1), the most straightforward way to reduce a transistor's off-state channel leakage is to minimize its $|V_{ds}|$. It is feasible via transistor stacking as in Fig. 3. In the worst case, with $V_{in} = 0$ and $V_{out} = V_{dd}$, or vice versa, the intermediate node voltages are $v_{mn} \approx \eta V_{dd}$ and $v_{mp} \approx (1-\eta)V_{dd}$ [13]. With small η, the DIBL effects of $M_{1a,\,2b}$ can be reduced by a few to a few tens of times (depending on the adopted process) compared to that of a simple transmission gate [14]. It can be improved further if more transistors are stacked, at the cost of an increased transistor size to maintain its on-resistance. However, large transistors are not preferred as they will increase body leakages and introduce larger charge injection and clock feedthrough during switching.

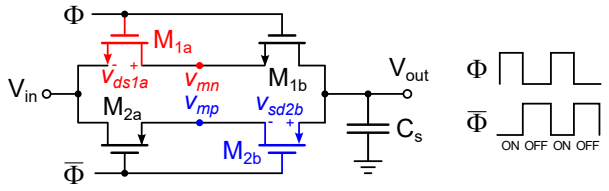

Fig. 3. Transistor stacking to reduce $|V_{ds}|$ thus minimizing subthreshold conduction and DIBL.

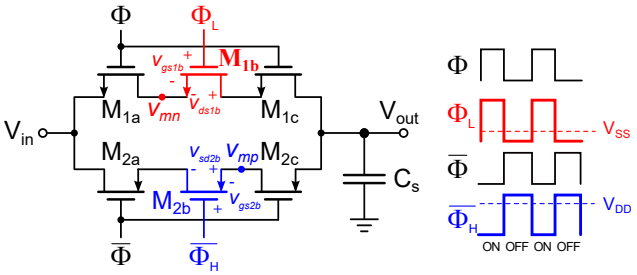

Fig. 4. Super cut-off CMOS switch design using gate control voltages beyond the ground and supply rail.

2) Super Cut-off Switch: besides using transistor stacking to reduce $|V_{ds}|$, the gate-source voltage of a transistor can also be manipulated. Fig. 4 shows such a super cut-off switch design [15]. When it is off, the gate voltage of the NMOS M_{1b} is set below ground potential while that of the PMOS M_{2b} is set beyond supply rail. $M_{1a,\,2a}$ and $M_{1c,\,2c}$ are used to relieve the gate-oxide stress of the negatively biased $M_{1b,\,2b}$. This switch enjoys the advantages of both transistor stacking and exponential leakage reduction due to the lower (higher) V_{gs} of M_{1b} (M_{2b}). Therefore, depending on the degree of negative and over-supply biasing, channel leakage of this switch is usually lower than that of using transistor stacking alone (e.g., a few times). Particularly, this scheme shows superior performance at high temperatures when V_{th} becomes small. However, it requires a charge pump to generate a negative bias and is not design-friendly.

3) Analog T-switch: to suppress the switch's channel leakage without severely increasing its on-resistance or using negative bias, the analog T-switch shown in Fig. 5 is an option [16]. In its off-state, $v_{mn,\,mp}$ are set to a fixed voltage V_{cm} (typically $V_{dd}/2$) by $M_{3a,\,3b}$, which is controlled by non-overlapping clocks to avoid accidentally shorting the input/output to V_{cm}. As a result, $M_{1b,\,2a}$ enjoys a large reverse gate-source bias. For certain range of V_{out}, its channel leakage can be two to three orders of magnitude lower than that of transistor stacking scheme. This is because, due to the small η factor in (1), unless making $|V_{ds}| \approx 0$, reducing $|V_{ds}|$ is less efficient in suppressing I_c as compared to that of reducing V_{gs} by the same magnitude. However, because $v_{mn,\,mp}$ are fixed, DIBL effects of $M_{1b,\,2a}$ are signal dependent and are maximized when V_{out} is close to the ground or supply rail. For example, when $V_{out} \approx 0$, the switch leakage is even a few times larger than that of the transistor

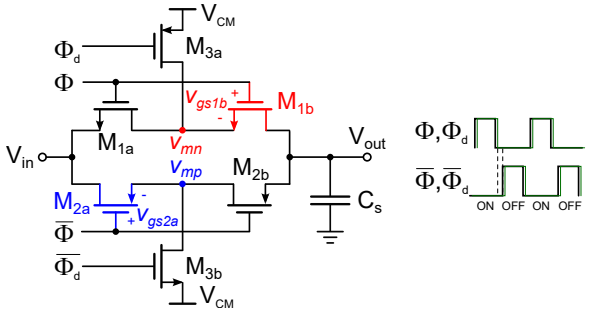

Fig. 5. Analog T-switch by applying a constant voltage for the internal nodes $v_{\mathrm{mn,\,mp}}$ in Fig. 3.

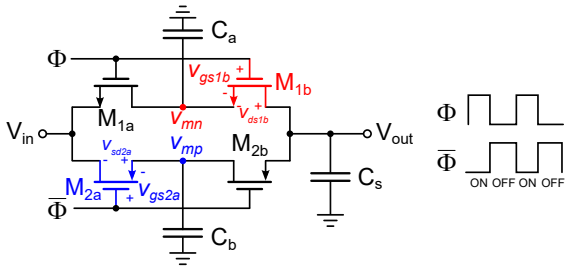

Fig. 6. Cap-hold T-switch to address the DIBL issue of Analog-T switch, allowing rail-to-rail output.

stacking scheme. Therefore, the analog-T switch is not suitable for applications with rail-to-rail signals.

4) Cap-hold T-switch: to utilize the advantage of analog T-switch while addressing its detrimental DIBL issue when V_{out} approaches the ground or supply rail, a cap-hold T-switch shown in Fig. 6 can be utilized. During sampling, V_{in} is also sampled on the hold capacitors $C_{\mathrm{a,b}}$, making $|V_{\mathrm{ds}}|$ of all transistors being initially zero, thus resulting in a near-zero channel leakage. Compared to the analog T-switch, its control is also simpler without using non-overlapping clocks. However, DIBL effect could still emerge if V_{out} varied without refreshing $v_{\mathrm{mn,\,mp}}$, making it more suitable for S/H circuits with stable V_{out}. This switch occupies a larger area due to the hold capacitors, whose size can be determined by the switching speed, error tolerance, operating temperature, etc. The hold capacitors also load the input during sampling.

5) V_{ds}-Regulated Switch: to avoid using large hold capacitors, especially when the switching rate is low (e.g., sub-Hz), and to minimize DIBL when V_{out} varies, an active buffer can be used [17]. As shown in Fig. 7, when the switch is off, M_{p} is negatively biased ($V_{\mathrm{sg}} < 0$) and its $|V_{\mathrm{ds}}| \approx 0$ as v_{mp} follows V_{out}. The leakage from V_{out} is thus negligible. The buffer can be of ultra-low bandwidth for low-speed operation, for example, using a leakage-based amplifier with thick-gate input transistors [18]. Though a single switch's power consumption is small (e.g., nA), the total power of a large switch array still hurts, making it only applicable for a few critical nodes. Meanwhile, to maintain a small on-resistance for a wide output range (limited by the input and output range of the buffer), M_{p} is controlled by a bootstrap circuit, whose interconnection path with V_{out} must be low-leakage as well, thus bringing in new

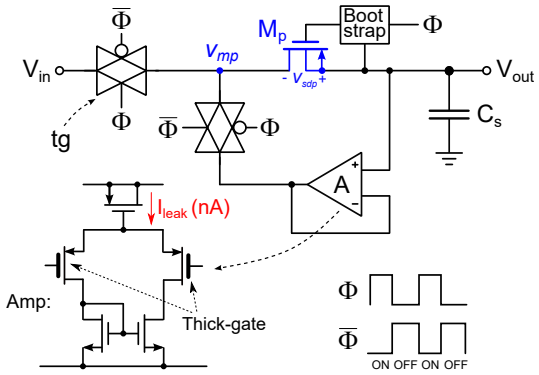

Fig. 7. Subthreshold leakage reduction with regulated source-drain voltage using an active buffer.

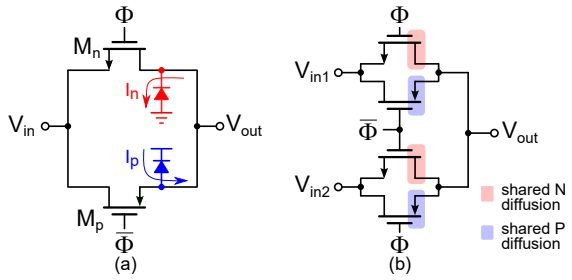

Fig. 8. (a) Counterbalance of the PMOS and NMOS drain to body leakages; (b) diffusion area sharing to obtain 2× drain to body leakage reduction.

design challenges.

Techniques discussed above have different tradeoffs in leakage performance, switch on-resistance, design complexity, silicon area, etc. One can select a suitable topology based on the target application. For example, for an S/H without silicon area limitation, the cap-hold switch is an optimal choice. If a rail-to-rail operation is required without considering design complexity, the super cut-off switch outweighs.

B. Drain to Body Leakage Reduction Techniques

For transistor operating at high temperatures, its drain to body leakage becomes significant (e.g., hundreds of pA to a few nA) and must be suppressed. For a PMOS-only switch, this can be achieved by regulating its body, drain, and source to the same potential, as shown in Fig. 7. This is not the best solution as it shifts the design challenge to a low-leakage bootstrap circuit when the input has a wide voltage range. For a transmission gate designed in standard processes, the body of NMOS transistor is not accessible, thus mandating other techniques to mitigate its body leakage. Note that without using gate bias beyond the ground or supply rail, GIDL of a transistor is much smaller than its parasitic diode leakage at high temperatures and is thus not discussed here [11].

1) Diode Leakage Counterbalance: as shown in Fig. 8(a), the parasitic diode leakages of NMOS and PMOS are in the opposite directions. Therefore, one can tune the source/drain diffusion areas in their layouts to achieve leakage counterbalance with $I_{\mathrm{n}} \approx I_{\mathrm{p}}$. Note that, for a reverse-biased diode, the dependency of I_{n} (I_{p}) on V_{out} is negligible if $V_{\mathrm{out}} > 3\phi_{\mathrm{T}}$

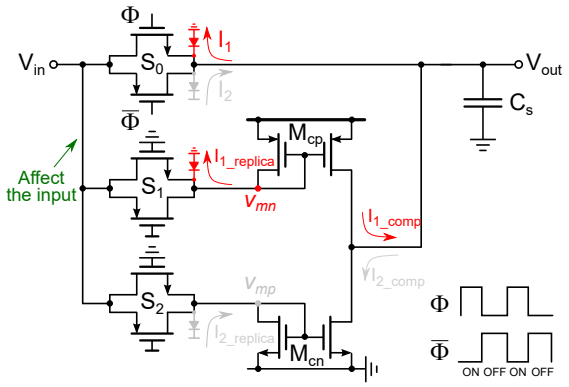

Fig. 9. Drain to body leakage compensation using switch dummies [19].

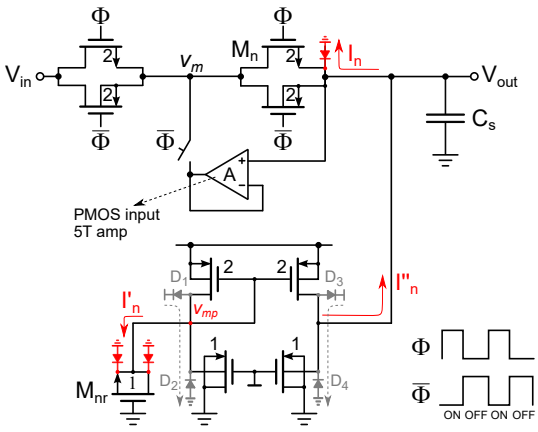

Fig. 10. Proposed unidirectional (only output node optimized) low-leakage switch for high-temperature operations (diodes shown in this schematic are all parasitic, aspect ratios of critical transistor are also listed in the figure).

($<V_{dd}$-$3\phi_T$) [10]. It means this scheme works for a wide range of output, though not rail-to-rail. For circuits like an analog multiplexer, diffusion sharing can further reduce the source/drain-to-body leakage by 2× as in Fig. 8(b). The leakage counterbalance scheme does not consume extra resources, but the large doping variations could degrade its performance.

2) Diode Leakage Compensation: to mitigate this issue, a design shown in Fig. 9 can be used [19]. When the main switch S_0 is cut-off, if the NMOS parasitic diode leakage I_1 dominates, a matched switch S_1 replicates this current and compensates I_1 via a current mirror M_{cp}, and vice versa. It works well at high temperatures in a 1 μm SOI process [19]. However, because the terminal voltages $v_{mn,\,mp}$ of the dummy switches cannot track V_{out}, and the current mirrors do not function when V_{out} approaches V_{ss} or V_{dd}, this scheme is not accurate nor can operate for rail-to-rail output. Even worse, in deep-submicron processes, channel leakages of $S_{1,2}$ could introduce large errors to both V_{in} and V_{out}.

C. Proposed Switch for High Temperature Operation

To achieve low-leakage operation for a wide voltage and temperature range without shifting the design difficulties to other circuits, we designed a unidirectional switch in a standard process, as shown in Fig. 10. Its channel leakage is suppressed using V_{ds}-regulation, while a scheme improved from [19] is used for reverse-biased diode leakage compensation. When cut-off, a matched transistor M_{nr} replicates the drain to body leakage of M_n, which compensates I_n via a body leakage compensated current mirror (the leakages in $D_{1,3}$ are compensated by $D_{2,4}$ and will not flow to the main switch). The buffer is a PMOS-input current-mirror-loaded amplifier consuming 20 nA of current.

For comparison, different switches are implemented to have the same on-resistance as a minimum-sized transmission gate in a 0.18 μm process. Fig. 11(a) presents the switch leakages at different temperatures and Fig. 11(b) shows the switch leakages at different output voltages. Limited by the current mirror, the allowed output voltage range is from 0.1 V to 1.6 V to ensure low-leakage operation, which is comparable to that of the super cut-off and the analog-T switch.

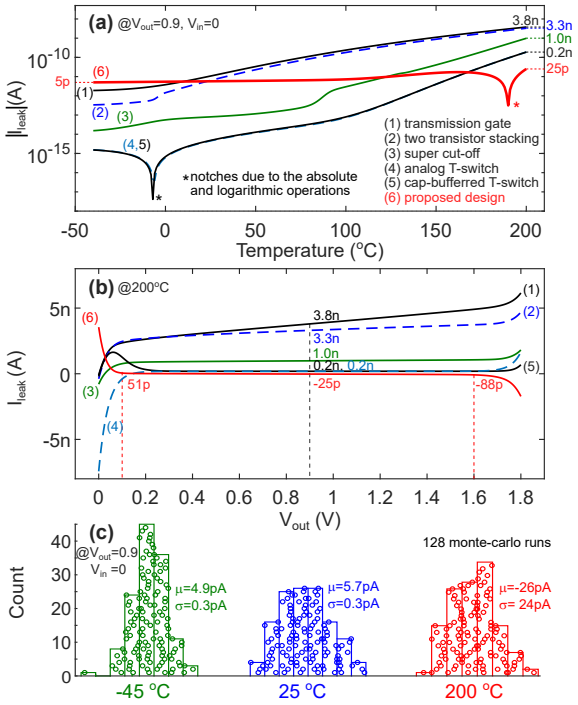

Fig. 11. Simulated performance of different switches (note: 1. V_{ds}-regulated scheme of Fig. 7 not included as it relies on the bootstrap design; 2. the under/over-drive voltage of the super cut-off switch is 0.1 V; 3. $C_{a,b}$=50 pF, 1.5 kHz switching rate for the cap-hold T-switch). (a) absolute switch leakage vs. temperature with a mid-rail output; (b) switch leakage vs. V_{in} at 200 °C; (c) switch leakage at different temperatures and process conditions.

For a mid-rail output and at 200 °C, the designed switch has 130× and 8× lower leakage than the transmission gate and the analog-T switch, respectively. The robustness of this switch relies on the matching between M_{nr} and M_n, as well as the accuracy of the current mirror to duplicate the compensation current I'_n. As in Fig. 11(c), based on the monte-carlo simulation results, this design is robust against device spreads as it does not rely on absolute process parameters. This scheme is also verified in a 65 nm process, and the improvements are in the same orders of magnitude. Note that the leakage

TABLE I
QUALITATIVE COMPARISON OF DIFFERENT CMOS SWITCHES

	Switch size[†]	Complexity	Leakage[#]
Transmission gate	Small	Low	High
Transistor stacking	Medium	Low	Medium
Super cut-off	Large	Medium	Low
Analog T-switch	Medium	Medium	Low
Cap-hold T-switch	Medium	Low	Low
V_{ds}-regulated	Medium	High	Low
Proposed	Medium	High	Ultra-Low

[†]for the core transistor size required to achieve the same R_{on} as a minimum sized transmission gate.
[#]for leakage at high temperatures (>150 °C).

reduction of our design at high temperatures is at the expense of more transistor counts (chip area) and extra static power consumption. Table I shows a qualitative comparison and the switch design tradeoffs.

D. Discussion

The above discussion focuses on a single switch design. For systems that require a switch array for channel multiplexing, leakage is also one of the most critical concerns. For example, in high-density bio-sensing and neural interfacing applications such as active microelectrodes arrays (MEAs) and neural probes, a large-scale switch bank/matrix is needed for micro-electrode selection [20] [21]. During operation, most of the switches in the bank/matrix are in their off-states, which load the signal line and could degrade the signal integrity if a large leakage current exists. Even worse, for *in-vivo* applications, leakage of the switch draws safety concerns since it induces lasting charge transfers to or from the brain tissue if connected to the microelectrodes directly [22]. In bidirectional neural interfacing applications, another parallel switch is required to route, for example, a stimulator to the same electrode as recording. This parallel switch usually has a large size to accommodate the stimulation current, and its leakage makes it more challenging to maintain the recording signal's integrity. Practices used to address this issue include using an active electrode with in-situ switch bank/matrix driving circuits or using a bootstrapped body-driven switch, both of which require additional circuit overhead and decrease the achievable density of the recording circuitry. These applications are still awaiting new area-efficient low-leakage switch designs.

IV. CONCLUSION

In this paper, transistor leakage mechanisms and the evolution of various low-leakage CMOS switch designs are briefed. Different techniques to reduce and/or compensate a transistor's channel and parasitic diode leakages are revisited. A new low-leakage switch design that outperforms the prior arts is also presented. In addition, this paper discussed the existing challenges to design a high-density CMOS switch array for signal multiplexing in emerging applications such as MEAs and high-density neural probes.

ACKNOWLEDGMENT

This publication was made possible by NPRP grant NPRP11S-0104-180192 from the Qatar National Research Fund (a member of Qatar Foundation).

REFERENCES

[1] Altera Corporation. White Paper: Reducing Power Consumption and Increasing Bandwidth on 28-nm FPGAs, March 2012.

[2] S. Wu et al., "A 7nm CMOS platform technology featuring 4th generation FinFET transistors with a 0.027um^2 high density 6-T SRAM cell for mobile SoC applications," *IEEE Int'l. Electron Devices Meeting,* pp. 2.6.1–2.6.4, Dec. 2016.

[3] T. T. Zhang et al., "A 310 nW 14.2-bit iterative-incremental ADC for wearable sensing systems," *IEEE Int'l. Symp. Circ. Syst.,* pp. 1-4, May 2017.

[4] P. M. Nadeau et al., "Ultra Low-Energy Relaxation Oscillator With 230 fJ/cycle Efficiency," *IEEE J. Solid-State Circuits,* vol. 51, no. 4, pp. 789–799, April 2016.

[5] B. Wang et al., "A 10.6 pJ·K^2 Resolution FoM Temperature Sensor Using Astable Multivibrator," *IEEE Trans. Cirt. and Syst. II,* vol. 65, no. 7, pp. 869–873, July 2018.

[6] Harshit Agarwal et al., Technical Manual: BSIM-BULK106.2.0 MOSFET Compact Model, June 2017. [Online]. Available: http://bsim.berkeley.edu/models/bsimbulk/, Accessed on: Oct. 2020.

[7] K. Roy et al., "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE,* vol. 91, no. 2, pp. 305–327, Feb. 2003.

[8] J. Kang et al., "Computational study of gate-induced drain leakage in 2D-semiconductor field-effect transistors," *IEEE Int'l. Electron Devices Meeting,* pp. 31.2.1–31.2.4, Dec. 2017.

[9] F. Fallah and M. Pedram, "Standby and Active Leakage Current Control and Minimization in CMOS VLSI Circuits," *IEICE Trans. Electron.,* vol. E88-C, no. 4, pp. 509–519, 2005.

[10] Chenming Hu, *Modern Semiconductor Devices for Integrated Circuits*, 1st ed. Pearson, 2009.

[11] Domenik Helms, "Leakage Models for High-Level Power Estimation," Ph. D dissertation, OFFIS Institute for Computer Science, Oldenburg, Germany, 2009.

[12] K. N. Yang et al., "Characterization and modeling of edge direct tunneling (EDT) leakage in ultrathin gate oxide MOSFETs," *IEICE Trans. Electron.,* vol. 48, no. 6, pp. 1159–1164, June 2001.

[13] Neil H. E. Weste and David M. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. Addiuson Wesley, 2009.

[14] S. Narendra et al., "Scaling of stack effect and its application for leakage reduction," *Int'l. Symp. on Low Power Electronics and Design,* pp. 195–200, Aug. 2001.

[15] K. Ishida et al., "Subthreshold-leakage suppressed switched capacitor circuit based on super cut-off CMOS (SCCMOS)," *IEEE Int'l. Symp. on Cirt. and Syst.,* pp. 3119–3122, 2005.

[16] K. Ishida et al., "Managing subthreshold leakage in charge-based analog circuits with low-V/sub TH/ transistors by analog T- switch (AT-switch) and super cut-off CMOS (SCCMOS)," *IEEE J. Solid-State Circuits,* vol. 41, no. 4, pp. 859–867, April 2006.

[17] J. T. Xu et al., "Low-leakage analog switches for low-speed sample-and-hold circuits," *Microelectronics Journal,* vol. 76, pp. 22–27, June 2018.

[18] H. Wang et al., "A Reference-Free Capacitive-Discharging Oscillator Architecture Consuming 44.4 pW/75.6 nW at 2.8 Hz/6.4 kHz," *IEEE J. Solid-State Circuits,* vol. 51, no. 6, pp. 1423–1435, June 2016.

[19] L. Zou et al., "Sample-and-hold circuit with dynamic switch leakage compensation," *Electronics Letters,* vol. 49, no. 21, pp. 1323–1325, Oct. 2013.

[20] S. Wang et al., "A Compact Quad-Shank CMOS Neural Probe With 5,120 Addressable Recording Sites and 384 Fully Differential Parallel Channels," *IEEE Trans. on Biomedical Circ. and Syst.*, vol. 13, no. 6, pp. 1625–1634, Dec. 2019.

[21] X. Yuan et al., "Extracellular Recording of Entire Neural Networks Using a Dual-Mode Microelectrode Array With 19,584 Electrodes and High SNR," *IEEE J. Solid-State Circuits,* Early Access, March 2021.

[22] T. Jochum et al., "Integrated Circuit Amplifiers for Multi-electrode Intracortical Recording," *J. Neural Eng.*, vol. 6, no. 1, pp. 1–26, Jan. 2009.