TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 08/16 pp696–718
DOI: 10.26599/TST.2022.9010064
Volume 28, Number 4, August 2023

Near-Threshold Wide-Voltage Design Review

Yan Zhao, Jun Yang*, Chao Chen, Weiwei Shan, Peng Cao, Yongliang Zhou, Ziyu Li, and Tai Yang

Abstract: This paper presents a comprehensive review of near-threshold wide-voltage designs on memory, resilient logic designs, low voltage Radio Frequency (RF) circuits, and timing analysis. With the prosperous development of wearable applications, low power consumption has become one of the primary challenges for IC designs. To improve the power efficiency, the prefer scheme is to operate at an ultra low voltage of Near Threshold Voltage (NTV). For the performance variation and degradation, a self-adaptive margin assignment technique is proposed in the low voltage. The proposed technique tracks the circuit states in real time and dynamically allocates voltage margins, reducing the minimum supply voltage and achieving higher energy efficiency. The self-adaptive margin assignment technique can be used in Static Random Access Memory (SRAM), digital circuits, and analog/RF circuits. Based on the self-adaptive margin assignment technique, the minimum voltage in the 40 nm CMOS process is reduced to 0.6 V or even lower, and the energy efficiency is increased by 3–4 times.

Key words: Near Threshold Voltage (NTV); wide-voltage memory; resilient logic design; low voltage Radio Frequency (RF); timing analysis

1 Introduction

For the flourishing of handheld applications, such as 4G systems, WLAN, WSN, Internet of Things (IoT) etc., low voltage and low power consumption of Radio Frequency (RF) transceivers, the core CPU, and DSP have become design bottlenecks for IC designs. Intel pointed out that power consumption will be one of the primary challenges according to Moore's Law in 2013[1]. In order to improve the power efficiency significantly, to operate the system at a Near Threshold Voltage (NTV) is a general solution. For the most efficient energy utilization, NVT is an excellent choice in the voltage domain. Meanwhile, NTV solution can also provide promising system performance with technology innovations. Figure 1a shows the energy-efficiency of CMOS circuits versus operation region. It is indicated that the NTV range is the optimal energy efficiency range. As shown in Fig. 1b, the operating frequency reduces linearly as the supply voltage decreases to NTV range, power consumption declines exponentially to make efficiency progressively improve 5–10 times with voltage decreasing.

However, operating under NTV environment also brings some design challenges, among which are performance variation and performance degradation. Under the NTV conditions, circuit performances are more susceptible to threshold voltage, process deviation, operating voltage, and temperature variation than being operated at regular supply voltage. Compared with regular supply voltage, a 20-time variation would be observed in a typical System on a Chip (SoC) system[2]. Moreover, the threshold voltage deviations will bring about extremely different effects in circuit operating

• Yan Zhao, Jun Yang, Chao Chen, Weiwei Shan, Peng Cao, Yongliang Zhou, Ziyu Li, and Tai Yang are with National Application Specific Integrated Circuit (ASIC) Center, Southeast University, Nanjing 210096, China. E-mail: 17352916606@163.com; dragon@seu.edu.cn; chen_seu@aliyun.com; wwshan@seu.edu.cn; caopeng@seu.edu.cn; zhouyongliang@seu.edu.cn; 220191403@seu.edu.cn; 220191436@seu.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2021-09-06; revised: 2021-11-22; accepted: 2022-12-07

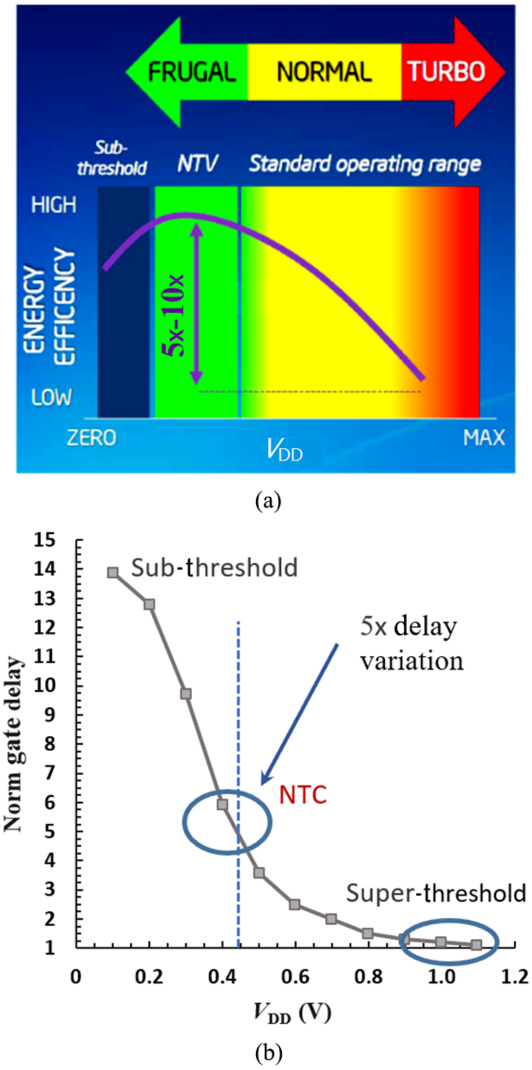


Fig. 1 Property of CMOS circuits versus operating voltage. (a) Energy efficiency of CMOS circuits versus operating region, and (b) delay variation versus operating voltage (NTC means near threshold conversion).

states of the regular voltage and sub-threshold voltage. Process deviations and threshold voltage changes only cause 3 times differences of driving current extreme value under normal working voltage. However, for the sub-threshold circuit state, the differences in the extreme value of the sub-critical current are increased by 700 times and the difference in the extreme value of the leakage current are expended 900 times caused by process variation and threshold voltage deviation[2]. For regular operating voltage, critical voltage variations and process deviations only bring about 10% of the difference of driving current variation. Unfortunately, for 400 mV supply voltage NTV scenario, differences will substantially increase to 54%. 1000 cycles of the Monte Carlo simulation analysis of a 30-stage inverter chain with the supply voltage of 265 mV show that the longest delay is 7 times longer than the shortest delay[3]. In addition, the path delay differences fixed in the chip are distributed much wider than the delay differences caused by process changes. From this point of view, performance degradation for NTV applications is another thorny issue. When the power supply voltage is lower than 0.6 V, most of the transistors in the circuit can only be biased in the NTV region, while the transistors working in the sub-threshold region face the decline of characteristic frequency, which limits the bandwidth and gain of the RF circuit[4]. Figure 2 shows the characteristic frequency of an NMOS common source amplifier versus gate source voltage[5]. The drain voltage of the NMOS transistor is fixed at 0.3 V. By scanning the gate voltage, it is indicated that the characteristic frequency of the transistor decreases rapidly with the gate source voltage decreasing. When the gate source voltage decreases from 0.5 V to 0.3 V, the transistor transits from strong inversion region to

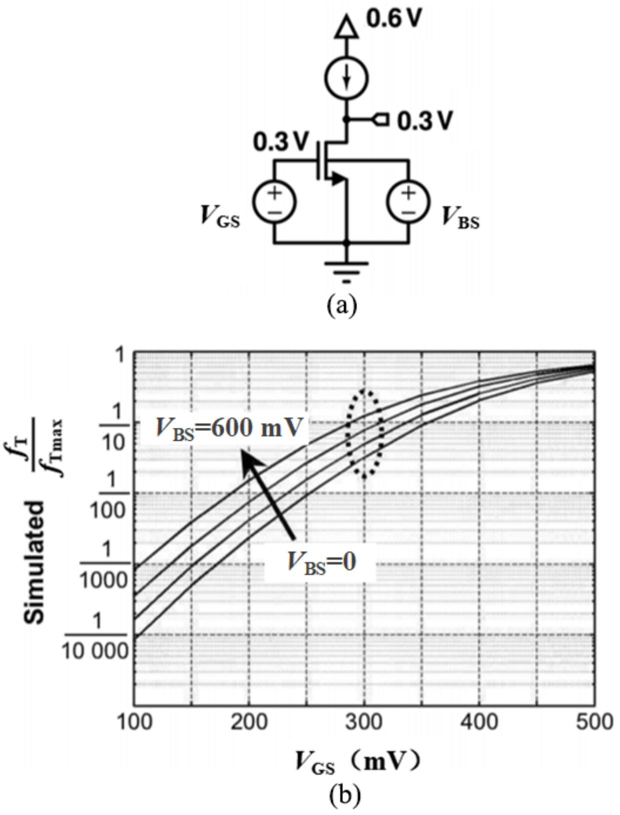


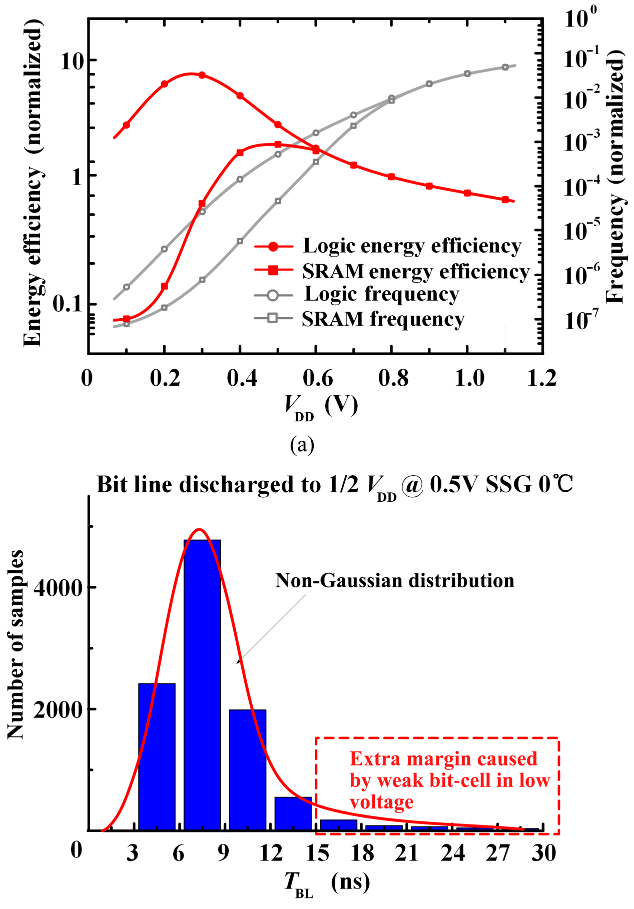
Fig. 2 Transistor characteristic frequency versus gate source voltage. (a) Circuit schematic for transistor testing, and (b) transistor characteristic frequency versus gate source voltage at 0.6 V supply voltage.

NTV region, and its characteristic frequency decreases to 1/100. The characteristic frequency of a typical 130 nm CMOS process is about 100 GHz. When the gate source voltage is equal to 0.3 V, the characteristic frequency drops below 2 GHz.

To overcome performance degradation, and Process, Voltage, Temperature (PVT) variation, conventional solutions assign adequate margin for all the possible process corners and performance variations. However, the static margin design method can hardly work properly under low voltage and NTV process, which limits the minimum operating voltage. In this paper, a self-adaptive margin assignment technique is proposed. The proposed technique tracks the circuit states in real time and assigns the margin dynamically, which can significantly reduce the minimum supply voltage and achieve high energy efficiency simultaneously. The concept and methodology of the self-adaptive margin assignment technique can be employed in Static Random Access Memory (SRAM), digital circuits, and analogue circuits, including speculative dual sampling SRAM circuit, which tracks the bit line read signal and dynamically adjusts read timing; half path dynamic AVFS, which tracks timing margin and dynamically adjusts clock frequency and supply voltage; dynamic biasing technique for analog/RF circuits, which tracks the key parameters, such as biasing current, voltage gain and transconductance etc., and dynamically adjusts the working current and load impedance to compensate any possible performance degradation in real time. Based on the above technology, the minimum voltage under 40 nm CMOS process is reduced to 0.6 V, and the energy efficiency is also improved by 3–4 times.

2 Wide-Voltage Memory

In order to meet the energy efficiency requirements of SOC, wide voltage SRAM is gradually becoming a research hot-spot in the industry. With the decrease of power supply voltage and local process fluctuation, the design requirement of the circuit is becoming more rigorous. In the near threshold region, too pessimistic design margin greatly increases the readout delay of the storage array, and the performance of SRAM is seriously degraded. In Fig. 3a, the energy efficiency and frequency trends of logic and SRAM are displayed with the normalized power supply voltage scaling. As V_{DD} reduces to the near-threshold region, the frequency decreases slightly and the energy efficiency of each operation is significantly improved. Once the power supply voltage enters the sub-threshold region, the energy efficiency and frequency drop sharply due to leakage energy dominance[6]. And at Ultra-Low Voltage (ULV) below 0.3 V, the degradation of SRAM's performance is more drastically than that of logic. Figure 3b illustrates that the SRAM read access time made up of bit-line enabling time (T_{BL}) shows a non-Gaussian distribution with supply voltage decreasing to the near-threshold region. Ideally, Minimum Word-line Enabling Time (MWET), which is defined to ensure the differential voltage between the Bit-Lines (BL), should achieve the minimum voltage recognizable by the Sense Amplifier (SA), so the MWET is designed to cover 6 standard deviations (σ) bit-cells discharge delay for correct reading in conventional voltage. However, due to the low voltage condition, the non-Gaussian distribution of bit-cell delay results in extra MWET margin to ensure



(a)

(b)

Fig. 3 SRAM characteristics. (a) Normalized supply voltage scaling trends of the frequency and energy efficiency of SRAM, and (b) SRAM read access time in the near-threshold region.

the correct reading operation. More MWET means worse performance[7].

The property of bit-cell has a direct impact on SRAM performance, similarly, the SA and the timing control also play a leading role in reading access time. Ideally, SA senses the smallest voltage difference over the BLs to determine the data[8], which results in near-zero bit-line power consumption and near-zero sensing time. But because of nonidealities of large bit-line capacitance and SA differential offset, the operations of the SA slow down, and the offset of the sense amplifier limits the minimum required BL swing in the read operation. Therefore, for eliminating the impact of the offset voltage, the differential bit line voltage of the accessed SRAM cell that is greater than the offset voltage is generated by bit-cells to obtain a correct readout. However, recent works have developed various SA offset remission techniques[9] to robust the read operation. But the offset can not be eliminated because the device threshold voltage affected by intra-die variations, such as random dopant fluctuations, is inescapable, which is more pronounced in ULV. To improve the throughput of ULV SRAM and decrease the extra and unnecessary delay margin, the concept of SRAM timing speculation is presented in Refs. [7, 10]. Traditional speculative schemes were based on the double sensing method, which can be interpreted as the reading operation is twice requested, the first sensing corresponds to speculative reading, and the second sensing corresponds to confirm the reading. By comparing the first and the second sensing data, the logic will confirm the correctness of the read operation. However, a separate sensing method will require an additional sampling period and extra sequential logic circuits, it will reduce the timing proceeds.

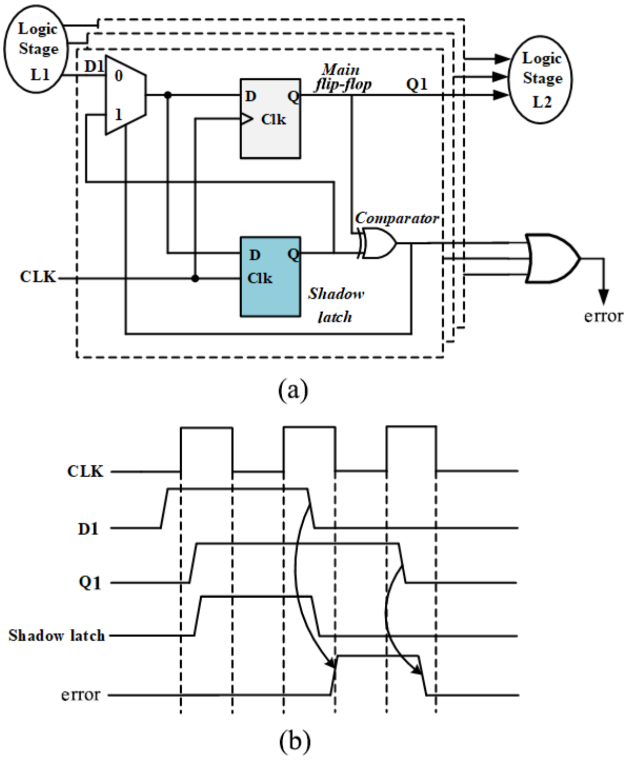
2.1 Low voltage SRAM

2.1.1 Related researches

2.1.1.1 Razor error detection

Traditional near-threshold SRAM reading assisting technique usually enhances the stability of bit-cell to reduce the supply voltages[11, 12], which is different from timing speculative technology. SRAM timing error-tolerant technique could improve the timing performance significantly by abandoning the delay drag of the weakest bit-cell in the array and detecting read fault. However, the logic speculative technique is different from that of SRAM.

As the fault-tolerant Flip-Flop (FF), Razor commonly utilized in voltage management techniques[13] is first presented in digital circuits for eliminating the margin of over design through in-situ timing error detection. The Razor, which relies on a combination of circuit-level technologies and architectural designs, is used to detect error efficiently and correct the critical path delay failures in the SoC system, the concept of Razor is demonstrated in Fig. 4a for a pipeline stage. In the work, the main flip-flop is realized with an augmented shadow latch controlled by a delayed clock. The monitoring unit is used to judge whether the timing of critical path in SoC system is wrong. The monitoring unit includes a D-Flip-Flop (DFF), a shadow latch, and an XOR gate. In addition to the function of ordinary trigger, it can also be used to detect timing error. The DFF is used to sample the input data, and the shadow latch is used to confirm whether the data sampled by the DFF are correct. The discussion is divided into the following two situations: (1) No establishment time violation occurred. The flip of data occurs before the rising edge of the clock, so the DFF and shadow latch are sampled correctly, and the XOR gate outputs low level; (2) Establishment time violation occurred. The data flip occurs after the rising edge of the clock. At this time, the path is in



(a)

(b)

Fig. 4 Diagram of Razor. (a) Pipeline stage of Razor, and (b) timing diagram of Razor.

violation of the establishment time, so the DFF collects the wrong data, and the latch can sample the data flip when the clock is high-level, so the XOR operation result of the output of the DFF and the output of the latch is high-level. Figure 4b shows the timing diagrams that explain its working principle, the main DFF is used to sense the input data, and the shadow DFF is used to confirm whether the data sensed by the main DFF are correct. The discussion is divided into the following two situations: (1) no timing violation occurs. The switching of data occurs before the rising edge of the clock, the data latched by the main DFF and shadow DFF are all correct, and the XOR gate outputs 0; (2) the timing violation occurs after the rising edge of the clock. At this time, the timing violation occurs in the path, the main DFF senses the wrong data, and the shadow DFF can latch the correct data, so the XOR gate outputs 1. Then the system will recover the wrong data through the error correction mechanism to ensure the correctness of the function.

2.1.1.2 Dynamic voltage scaling capable SRAM

Karl et al.[10] proposed a solution of Dynamic Voltage Scaling (DVS) for SRAM which applies the Razor mechanism to the reading operation. As shown in Fig. 5a, this method uses two standard differential latch-type sense amplifiers to double-sample the BL swing at the SA boundary on the read path, and checkup the sensing result by comparing the two SA outputs. Figure 5b indicates that the main SA triggered by the rising edge of the SAE1 (SAE denotes sense amplifier enable), which is generated from the falling edge of the clock. When the SRAM enters to the precharge phase of the cycle, the output is immediately stored in the unclocked SR latch1 to obtain the stable static output bus. The shadow SA is enabled by the rising edge of SAE2 which is later than the SAE1, while the differential voltage between the bit-lines is more significant than the main SA sensed. The shadow SA re-sampling operation must guarantee to overcome data-related leakage current, offset voltage caused by process changes, and internal voltage related to actions in the sense amplifier. The outputs of the two SAs are sent to the XOR gate for judging, if the result of the XOR is low, the output of the main SA is correct; if the result is high, the output is wrong. If the system determines that the output of the main SA is an error, the multiplexer will output the shadow SA result, the system will ensure the correctness of output through the error correction. By adopting a correction circuit and an embedded timing error detection, this scheme could dynamically converge to a minimum operating voltage.

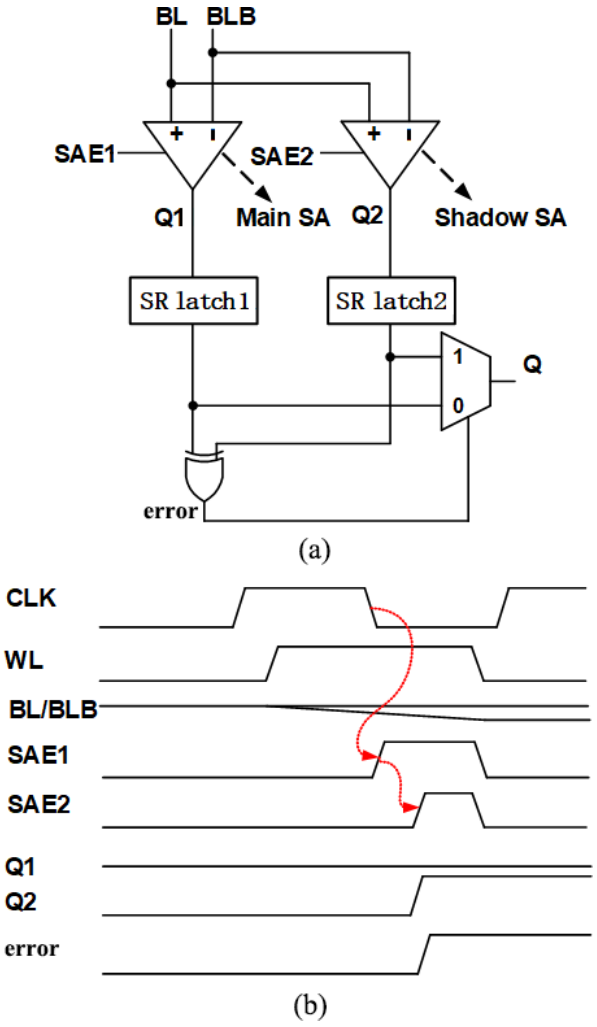


Fig. 5 DVS for SRAM. (a) Circuit diagram of DVS, and (b) timing diagrams of DVS (BLB means bit-line bar).

2.1.1.3 Dual-port SRAM with error detection

For mitigating variations in the internal sensing delay that is needed to develop the targeted BL voltage, Khayatzadeh et al.[11] introduced a Razor-style Error Detection And Correction (EDAC) technique for SRAM arrays, which utilizes the double sampling method to speculatively complete an SRAM read access in one cycle for the common case. The Razor Read (RR) circuit shown in Fig. 6a is a pipelined structure with the SR latch, XOR, flip-flop, and the multiplexer to form a speculative mechanism.

As shown in Fig. 6b, SA is first activated at the middle of the word line, the register will storage the SA output, and the data are immediately sent to the system for processing. Then the BL keeps on discharging, at the

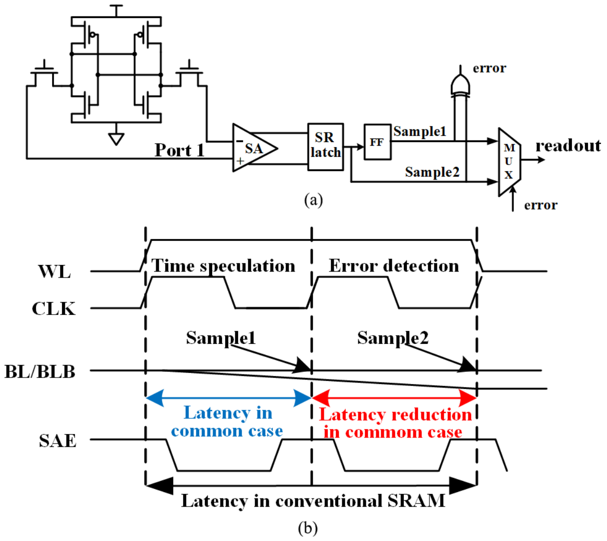


Fig. 6 RR circuit. (a) Circuit schematic of RR circuit, and (b) timing diagrams of RR circuit.

falling edge of the word line, the SA will start the second sensing operation, which will guarantee the correctness of the sampling data. The twice-sensing results of the sense amplifier are sent to the combinational logic circuit to judge the correctness of the data. If the XOR gate output is high, the error flag of the Razor SRAM will be activated by the error control block through any existing roll-back mechanism available in the processor, and the EDAC capability in Razor-class processors will be used to correct the error data; if the XOR gate output is low, the first sensing data are proved to be correct, and the system will not perform error correction mechanism. In most cases, the read output is available after the first clock cycle. The few slow-bit cells, which just in the statistical tail, will trigger errors infrequently. In summary, margin-less RR enables a considerable speed-up compared to a conventional read. This is particularly useful in the near-threshold regime, where variations cause a long tail in the read current. But the extended cycles, which are used to accommodate the slow bit cells, will cause unnecessary loss of performance, which can be further optimized.

2.1.2 Our works on low voltage SRAM

2.1.2.1 Double sensing with selective bit-line voltage regulation

DVS capable SRAM[10] and dual-port memory with error detection[11] are all based on the Double Sensing (DS) technique, the first sensing corresponds to speculative reading, and the second sensing corresponds to confirm the reading. However, the common shortcomings of the above solutions are that the interval between speculative reading and confirm reading is large and the error flag is generated too later especially at ULV, which restricts its application in SoC systems. Yang et al.[7] proposed an intelligent solution to this problem. The schematic and timing of the Double Sensing with Selective Bitline Voltage Regulation (DS-SBVR) scheme are displayed in Fig. 7a, the architecture of DS-SBVR includes a latch style SA, a dynamic latch, a configurable sharing capacitor, and an XOR logic in each column. By sharing charge between the capacitor and BL/BLB, the voltage of BL/BLB can be regulated.

Detecting the weak bits is the key idea of the bit-line voltage regulation mechanism[12], as illustrated in the timing diagrams of Fig. 7b, the risk sensing and the confirm sensing will be implied in the timing speculation cycle case. Once the risk sensing finishes executing, BL is lowed by the sharing capacitor and the risk result with decreased voltage difference ΔBL_2 will be checked by the confirm sensing. Then the data are temporarily stored in the dynamic latch, and the results of the two read operations are compared by the XOR logic, if the XOR output is low, it means that the speculative data are the same as the confirmed data, the sensing result is identified as the correct data; contrarily, if the XOR output is high, it means that the two sensing results are different, and DS-SBVR will generate the error flag, identify the bit-cell as a weak bit, and extend an extra cycle to ensure correct readout. The reason why the risk sensing operation is failed is that ΔBL_1 is not large enough to cover the SA offset, so during the error correction cycle, a conservative drop is generated, and ΔBL_3 will ensure to cover the offset of the SA.

In summary, previous SRAM speculation techniques allow designers to avoid margin adjustments for rare transient states that cause the worst-case read current. The re-sensing operation allows a speculative sensing phase with detection/correction to maximize the SRAM's cycle time in the presence of changes in read current caused by leakage current.

2.1.2.2 Extended to Magnetoresistive Random Access Memory (MRAM)/cache

To alleviate the deteriorating "power wall" problem, more and more applications need to extend their operating voltage to a wide voltage range including the near-threshold region. However, due to process fluctuations, the read delay distribution of SRAM cells and MRAM cells at near-threshold voltages shows more severe long-tail characteristics than those at nominal

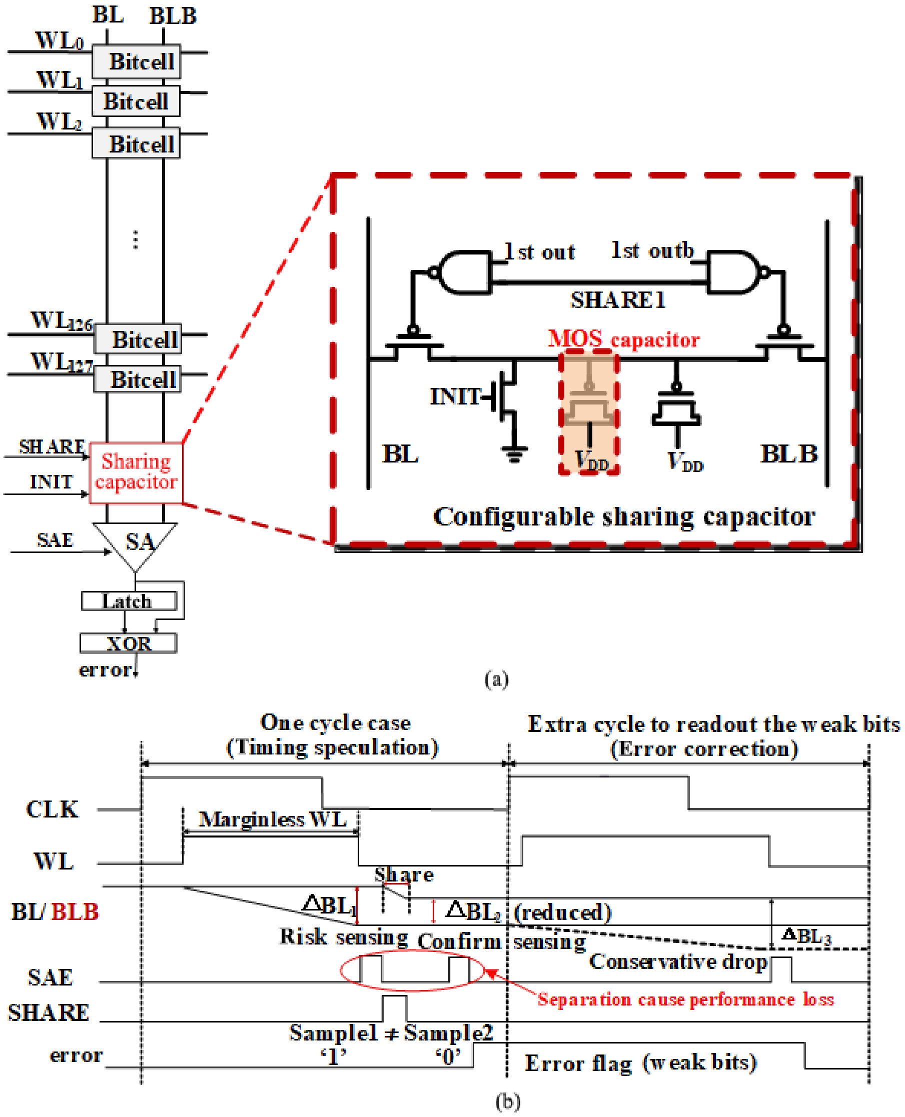


Fig. 7 Diagram of the DS-SBVR scheme. (a) Schematic of the DS-SBVR scheme, and (b) timing diagrams of DS-SBVR.

voltages. We proposed a Timing Speculative (TS) cache to increase the cache frequency and improve energy efficiency under low power supply voltage[13]. In the TS cache, the voltage difference of the BLs is continuously evaluated twice by the SA, and the access timing error can be detected in real time. On the other hand, we also applied this technology to the design of wide-voltage MRAM[14]. We proposed a self-timed voltage mode detection scheme called ST-VSS, which can achieve the best timing based on the discharge capability of the bit cell. The proposed single SA structure uses a multiplexer at the input of SA. Its continuous sensing operation is realized by input offset flip. Through the Built-In Self-Test (BIST) method, the dual SA structure is reconfigured to the opposite offset state to monitor each other's sensing results. After successful reading, the sensing operation can be terminated immediately. In order to solve the problem of cache and MRAM performance degradation at near threshold voltage, a speculation technique is proposed. By using an efficient speculation mechanism, Refs. [13, 14] break the restriction that all memory accesses must be completely correct. Low-cost error detectors can quickly identify erroneous readings and correct them in an extended period.

3 Resilient Logic Design

3.1 Related researches

As the feature size of transistors under advanced technology is getting smaller and smaller, the circuit becomes extremely sensitive to static manufacturing deviations and dynamic environmental fluctuations. In order to overcome the above-mentioned differences, designers often add a certain margin to the circuit design to ensure correct circuit operation, which is the so-called worst-case design. However, as the deviation of the path delay under low voltage becomes more and more significant, this design margin is often detrimental to the designer and will make the circuit unable to exert the best performance. Therefore, the traditional worst-case design efficiency is getting lower and lower, and an adaptive performance compensation in post-silicon is urgently needed as a countermeasure[15]. The resilient design usually refers to the ability to adaptively adjust the chip power supply voltage after tape out, and monitor the circuit timing in real time through the on-chip timing monitoring unit. Such a system is usually called Adaptive Voltage Frequency Scaling (AVFS) system. When the timing of the chip is relatively loose, the timing margin can be compressed by increasing frequency or reducing voltage; when the timing of the chip is relatively tight, it is necessary to increase the operating voltage or reduce the operating frequency. According to different monitoring methods, AVFS technology is mainly divided into two categories: direct monitoring and indirect monitoring. The following describes the research status and development of the two AVFS technologies at home and abroad.

3.1.1 Research status of direct monitoring AVFS technology

The core idea of direct timing monitor AVFS technology is to directly monitor the critical path inside the chip in real time. Based on the idea, the working voltage and frequency of the chip are trimmed in time according to the monitoring results to eliminate the PVT variations and timing when the timing is eased. The purpose is to ensure the correct function when the time is stressed. The direct monitoring AVFS design method can effectively suppress the influence of local deviations in the chip. The direct monitoring AVFS technology was first proposed by the University of Michigan[16] in 2003 and perfected the Razor I structure in 2006, creating the basis for the error-correcting direct monitoring AVFS technology. This structure uses a circuit structure of a shadow latch, a flip-flop, and an exclusive OR gate to monitor the timing of the circuit. The shadow latch always keeps the correct output data. The XOR gate compares the results of the latch and the flip-flop to output a timing error signal, and then the data recovery mechanism is executed. This method is applied to a 64-bit processor and can achieve up to 50% energy consumption gains. In the following 10 years, the University of Michigan has successively proposed Razor II[17], Bubble-Razor[18], Razor-lite[19], iRazor[20], and other different direct-monitoring online sequential monitoring solutions, making the monitoring unit gradually simplify. Especially in the iRazor design, the number of transistors in the monitoring unit has increased by 1.46 on average on the basis of the standard latch unit, which is the most streamlined structure proposed so far. However, this structure has threshold loss and cannot work normally in the near-threshold area, so it is not suitable for use in near-threshold designs. Many other companies and research institutions have also conducted in-depth discussions on the direct monitoring AVFS technology. The AVFS technology based on Razor-latch presented by ARM Company[21] in 2013 introduced the pulse latch, only added 1 transistor on the clock end, greatly reduced the power consumption of the clock network, and finally obtained 34% of the power consumption. The sparse latch replacement method proposed by Columbia University[22] can be applied to ULV in 2015, and analyzed the reason why the traditional error correction method in the ULV area cannot work normally. This solution is applied to a 16-bit microprocessor chip under 65 nm and can achieve up to 42% energy consumption gains. The online timing monitoring system based on pulse latch from Southeast University[23, 24] utilizes multiple pulse signal generators to generate pulse clocks, compresses the monitoring window, and provides greater convenience for the repair of short paths, at a cost of area only 1.4%. This solution is applied to Advanced Encryption Standard (AES) circuits, and can achieve up to 64.3% power consumption gains. Leuven University[25] put forward a soft-edge flip-flop in 2018, which delays the master latch clock in the flip-flop relative to the slave latch for a period of time, creating a window similar to the time borrowing from the latch. The data can be sampled correctly when the timing is tight. This method is designed and implemented on the Cortex-M0 microprocessor, and can achieve up to 36% of power consumption gains. Shanghai Jiao Tong University[26] proposed the sparse insertion of error detection registers. Compared with the traditional EDAC technology, the total area is reduced by 26%–33%, and the number of error detection registers is reduced by 2.9–4.3 times. The metastability condition detection and correction scheme of Shanghai Jiao Tong University[27] introduced a circuit based on double-sampling to detect whether the input data are too close to the clock edge of the receiver, which is defined as a metastable condition. The technique ameliorates energy efficiency and the data rate by 27.2% and 21.1%, respectively.

Another AVFS system is warning monitoring system. Singapore's A*STAR Institute of Microelectronics[28] first proposed the adaptive voltage scaling scheme based on half-path monitoring in 2013, placing the monitoring unit in half of the path, predicting potential timing errors in the circuit, and applying the scheme to FFT. In the processor, the problem that the adaptive voltage regulation system cannot respond within the current cycle is solved, but the maximum frequency of the chip supply voltage at 0.5 V is only 500 KHz. Most of the direct monitoring AVFS technology monitor path endpoints (i.e., destination registers), however, the Short Path (SP) issues of above techniques are non-ignorable because a timing error signal of EndPoint Latches (EDLs) is produced at the detection window of signal transition, which may be the normal signal of the SP of the current clock period or the late signal of the critical path of the previous period. How to adjust the frequency and voltage in time to ensure the correct function of the chip is also a problem when the timing is tight.

3.1.2 Research status of indirect monitoring AVFS technology

The core idea of the indirect timing monitor type AVFS technology is to simulate the delay characteristics of the real critical path in the chip by copying the critical path or using the critical path constructed by general-purpose units. Use the on-chip monitoring circuit to monitor the delay information and feed it back to the upper-level power management system to adjust the operating voltage of the chip, thereby effectively reducing the timing margin reserved in the chip design, and at the same time suppressing the impact of global deviation in the chip. The indirect monitoring AVFS technology was first proposed by the Swiss company CSEM[29] in 1990. The designer implanted a ring oscillator on the chip to monitor the PVT changes on the chip. Sony[30] put forward a method of duplicating the critical path based on a configurable delay chain in 2005. Its circuit structure is composed of a configurable number of inverters, NAND gates, and other logic gates and interconnections. The control signal is manually configured into different path structures to simulate the critical path delay. When applied to 0.18 μm technology chip to play MPEG4 video, the power consumption can be reduced by 40%. The critical path structure of the United States IBM[31] used logic units, such as NOR gates, NAND gates, adders, transmission gates, and long wires to simulate the critical path delays. The path is placed according to the hot spot distribution of the chip. The test results of this method on the POWER6 processor show that the monitoring effect is better. However, this critical path structure also has the disadvantages of complex structure and large area. The technique introduced by American Intel Corporation[32] in 2010 adopted an adjustable inverter chain to simulate the critical path of the chip. Post-silicon tests show that this solution can achieve a 41% increase in throughput and a 22% gain in power consumption. Making the replication path close to the true critical path is the goal of indirect monitoring. In Ref. [33], a universal optimal gate structure can better simulate the actual critical path, and the path structure in the chip can be deduced through theoretical deduction. This method is taped out under the 40 nm process, and the dynamic power consumption after AVFS turned on can be reduced by 27%. In 2018, In Ref. [34], an online monitoring method for Cortex-M3 microprocessors was put forward, which combined critical path replication and in-situ monitoring to deal with the effects of on-chip and inter-chip deviations. The replication critical path in this method includes three thresholds: high threshold, conventional threshold, and low threshold, which provides a basis for the selection of the ratio of different threshold units in the comprehensive optimization work. This solution can work in a wide voltage range of 0.5 V to 1 V, and the power consumption gain brought by AVFS is 14%. Seoul National University[35] found in 2018 that the mismatch of delay between the replication critical path and the actual path in a wide voltage range from near the threshold to the over-threshold may lead to a decline in the revenue of DVFS/AVFS. The method of constructing replicated critical paths with multiple threshold units reduces the deviation between the replicated path and the actual path. After being applied to the Cortex-A53 processor, the deviation can be reduced by up to 91%. An online monitoring program[36] was employed to combine direct monitoring and indirect monitoring in 2018. It uses direct monitoring for some critical paths on the film and uses this as a basis to guide indirect monitoring. This solution solves the defects of indirect monitoring path mismatch and high cost of direct monitoring, and can obtain up to 53% of power consumption gains within a wide voltage range of 0.5 V to 0.9 V.

3.2 Our works

In recent years, our research in the field of resilient design has mainly focused on half-path monitoring, short-path padding, and adaptive clocking.

3.2.1 Half-path monitoring AVFS system

Southeast University proposed a half-path monitoring AVFS system[37] in 2019. As shown in Fig. 8, a new Transition Detector (TD) circuit with only 7 transistors is proposed. It has a wide operating range from standard threshold voltage to over-threshold voltage. Compared with traditional latch, it has only 29% dynamic energy, 10% area overhead, as well as 14% leakage power.

The TD is carefully inserted at the half-path point of the critical path in order to generate an error signal after the falling edge of the clock. Therefore, when combined with error correction methods such as clock gating, timing errors can be predicted in advance. The system contains multiple TDs inserted in different paths. In order to cluster all error signals, dynamic OR gates are used to reduce propagation delay, as shown in Fig. 9. In order to cope with the change of the insertion point and reduce the insertion rate of the TD, a selection method suggests adopting the halfway insertion point combined with the intermediate points. The test chip is manufactured using a 40-nm CMOS process. Silicon measurements show that compared with the traditional

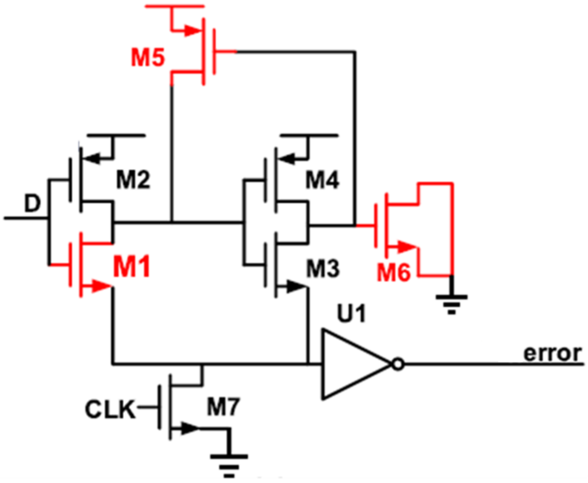


Fig. 8 Schematic of the proposed TD.

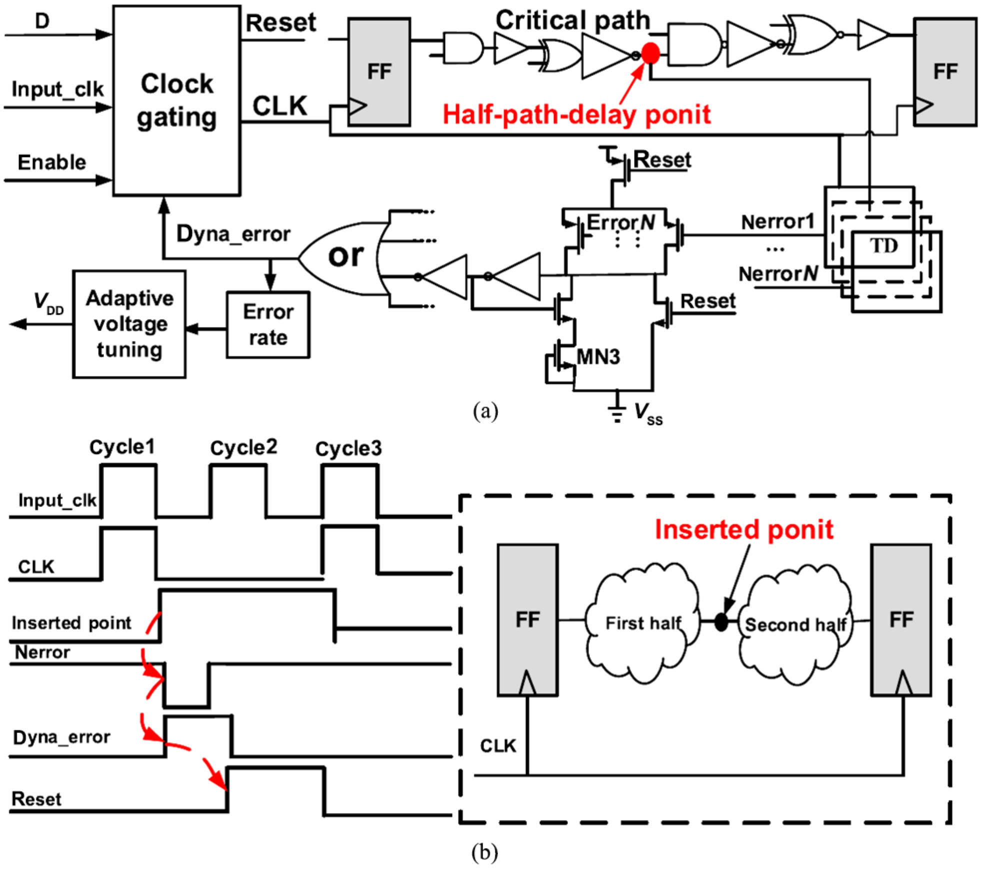


Fig. 9 Diagram of the half-path error-detection circuit. (a) Circuit structure, and (b) timing and error-detection waveforms and the insertion constraint.

0.56 V margin design, the system achieves 41.2% to 50.5% energy savings at close to the threshold voltage and has a small area overhead.

3.2.2 A one-transmission-gate short-path padding

Southeast University proposed a Transmission Gate-based SPP (TG-SPP) method[38], the short-path is extended to the negative clock phase by one transmission gate while keeping the critical paths unaffected. As shown in Fig. 10a, the path from FF1 to EDL is critical and the path from FF2 to EDL is a short-path, they share a common part of Tc2. CTG is a fundamental transmission gate which consists of parallel-connected PMOS and NMOS transistors, and it is inserted in the common part of the two paths. The CTG is turned off during the positive clock phase, the short-path data cannot propagate to the EDL until the falling clock edge arrives, as shown in Fig. 10b. On the other hand, although a CTG is inserted in the critical path, its setup timing constraint is only slightly affected. The delay overhead of the critical path is only a CTG's delay time, which is nearly as small as the delay of an inverter.

TG-SPP is implemented in SHA-256 chip using 28-nm CMOS process. The results demonstrate that the TG-SPP method reduces sequence area overhead and the glitch power by 6×, while achieving the same filling effect as the method based on the two-phase latch. In contrast to the typical margin baseline close to threshold voltage, the manufactured chip can be measured to achieve 38.6%–69.4% energy saving and 55%–405% frequency improvement.

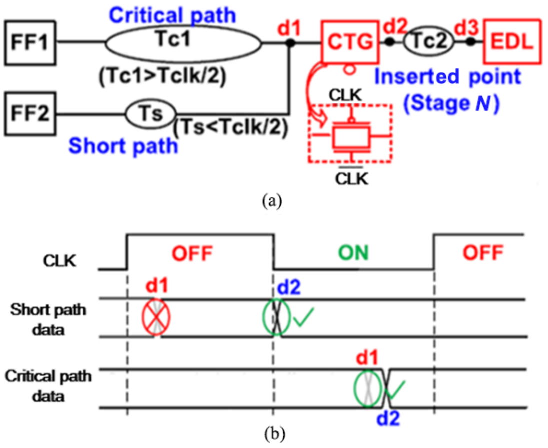


Fig. 10 TG-SPP with a transmission gate inserted in near the critical path end. (a) Illustration of the TG-SPP with a transmission gate inserted in near the critical path end, and (b) its impact on critical-path and SP data (Abbreviations: Tc1 refers to the most majority part of critical path delay; Tc2 refers to the small part of critical path delay; and Ts refers to SP delay).

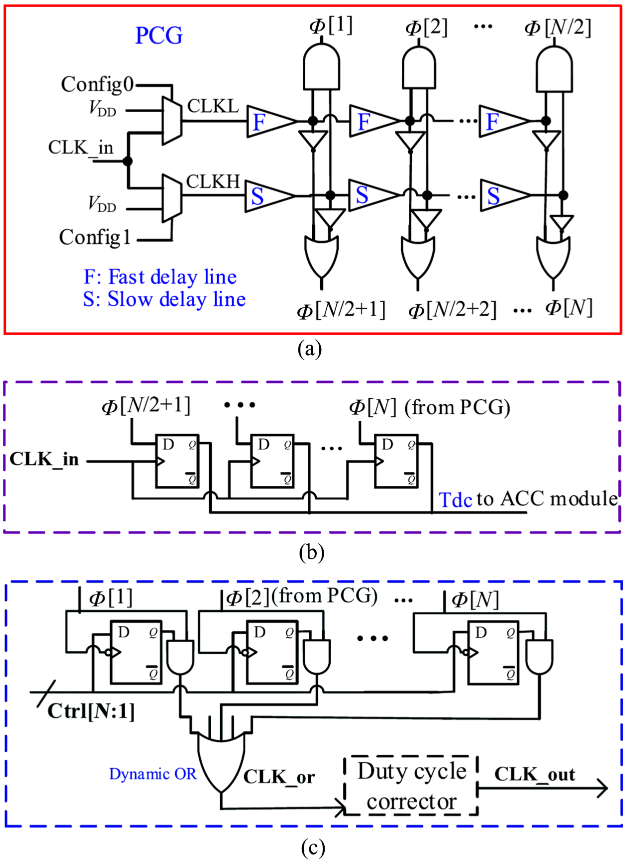
3.2.3 Adaptive clocking AVFS system

Southeast University proposed a bidirectional adaptive clocking circuit to provide zero-delay fine frequency tuning for AVFS system[39, 40]. It can compress the cycle when the timing margin is too much, or extend the clock cycle when the timing is wrong to ensure correct function. For supporting a wider frequency range, multiple phase clocks are generated based on two delay lines, an appropriate phase clock is chosen to get an extended output clock, and a balanced clock path is obtained through dynamic OR gates and time-to-digital converters. The function of the fast adaptive clock circuit is achieved by continuously selecting only a specific phase clock from the generated poly-phase clocks to tune the clock cycle. Therefore, it needs to fulfill the following requirements: (1) to generate multiphase clocks ($\varPhi[1], \varPhi[2], \dots,$ and $\varPhi[N]$) by buffers in a wide frequency range; (2) to choose one of the appropriate clocks according to AVFS requirements; and (3) when the last available clock is reached, it continuously loops between the limited phase clocks until clock stretching/compression is disabled.

As illustrated in Fig. 11, it is mainly made up of (1) a Phase Clock Generator (PCG) to produce multiple phase clocks; (2) a Phase Clock Detector (PCD) to detect the last available phase clock; and (3) a Phase Clock Selector (PCS) to pick up an output from one of the phase clocks, with control signals coming from an Adaptive Clock Controller (ACC). Implemented to a wide-operating-range AVFS system of an SHA-256 accelerator with TD latches, the AVFS system is able to respond in one clock cycle. Fabricated in 28-nm CMOS, chip measurements demonstrate that 38.6%–69.4% power gains can be obtained at near-threshold while decreasing throughput loss during error recovery.

4 Low Voltage RF

In recent years, low voltage has become one of the primary targets of digital and analog integrated circuit applications[41]. Under low supply voltage, the voltage margins of traditional radio frequency circuits shrink and the performance of transistors are severely limited. The insufficient overdrive voltages of transconductance transistors attenuate the bandwidth and intrinsic gain. Tail current transistors are in the near linear region, which leads to deterioration of Power Supply Rejection Ratio (PSRR) and Common Mode Rejection Ratio (CMRR). And load transistors located in the subthreshold region decrease output



(a)

(b)

(c)

Fig. 11 Circuit architecture of the adaptive clocking, mainly composed of (a) PCG with dual delay lines, (b) PCD, and (c) PCS.

impedance and voltage gain, the output swing is also restricted simultaneously to decrease the Signal-to-Noise Ratio (SNR). Therefore, the traditional techniques of high performance, high matching, and anti-PVT by biasing transistors in saturation region and providing sufficient voltage margin for circuit components are not suitable for low voltage designs. Key parameters, such as transconductance, intrinsic gain, and oscillation frequency at low supply voltage, are more susceptible to PVT fluctuations and signal swings. Traditional static margin solutions require the device parameters to remain stable across PVT conditions, which results the minimum supply voltage to be difficult to reduce to below 0.9 V[42]. Recently, low-voltage RF circuit schemes have been put forward one after another.

4.1 Related researches

For low voltage RF designs, front-ends mostly utilize current-reused structures[43, 44], Forward Body Bias (FBB) technique[44, 45], and other ULV front-end schemes[46, 47]. The core idea of current-reuse technique is to share the bias current of different parts of the circuit. Based on the relatively simple structure and fewer transistors stacked from the power supply voltage to the ground of the RF circuits, multiple RF modules can be stacked on top of each other to realize the multiplexing of the bias current. The current reused structure can achieve better performance with less current consumption. In addition, it can construct independent functions of multiple modules while only consuming the power consumption of one circuit module. So far, current-reuse scheme is still one of the most popular low-voltage low-power RF circuit designs. FBB is one of the solutions to lower the threshold voltage of the devices, which alleviates the problem of insufficient voltage margin and effectively reduces the supply voltage. However, junction leakage current is a critical issue, which degrades the performance of RF circuits. To minimize power consumption, ULV front-ends are preferred in RF circuit designs. In ULV Low Noise Amplifiers (LNAs), the transistors work in the sub-threshold region, leading to the deterioration of the linearity and the sensitivity to PVT fluctuations.

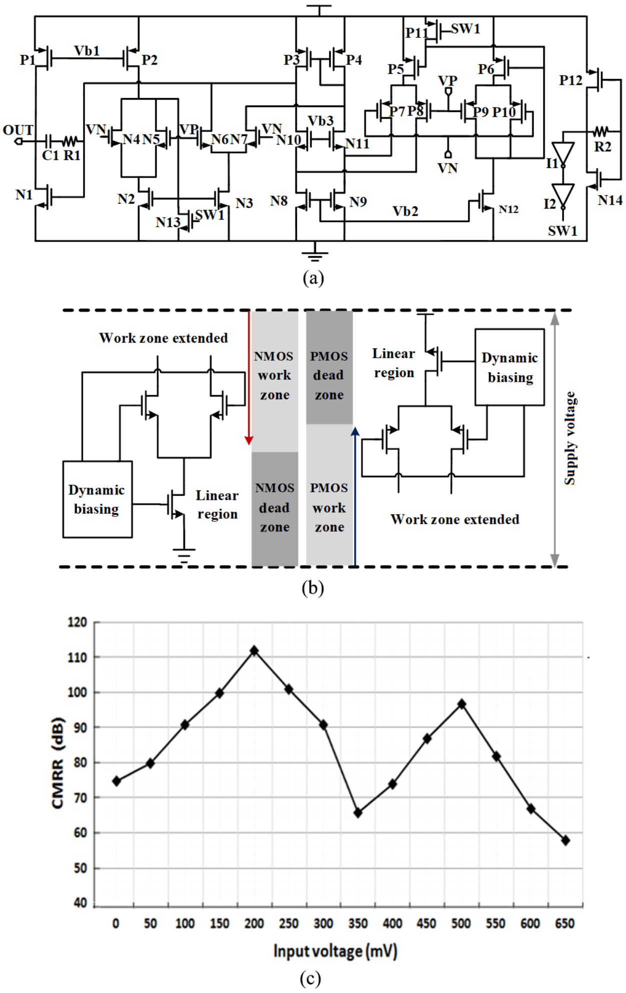
Low voltage is the most effective solution to improve the energy efficiency of the SoC. However, this is a huge challenge for mixed-signal circuits such as Phase-Locked Loops (PLLs)[48]. As critical modules that affect PLL performance, Voltage-Controlled Oscillators (VCOs) and Charge Pumps (CPs) have become difficult issues for PLL designs under low voltage[49–52]. Reference [49] proposed a Feed-forward Ring VCO (FRVCO) for low voltage PLL to effectively compensate for frequency fluctuations caused by power supply noise by adjusting the drive strength ratio between the FRVCO direct path and the feed-forward path. The promising solution based ring VCO with two voltage control points in Ref. [50] achieved low power PLL for biomedical applications. The first control point is used to coarse tune and correct frequency, and the second control point performs fine-tuning to sufficiently cover the MICS band frequency. A novel CP structure was introduced in Ref. [51] to realize the NTV PLL, achieving excellent matching characteristics of up/down currents and higher output resistance, effectively alleviating the impact of CP current changes with VCO control voltages. In Ref. [52], FBB technique is utilized for the lower threshold voltage so that ULV PLL is achieved. Recently, ADPLLs[53, 54] have been proposed for low voltage to overcome the stringent requirements of traditional analog circuits for voltage margin. However, the impact of low voltage on the voltage margins of the analog circuits leading to performance degradation is still a key issue for integrated circuits.

4.2 Our works

For the serious limitation of voltage margin of traditional RF/analog circuits, the self-adaptive margin assignment technique is proposed. The deviations caused by insufficient voltage margin are canceled by dynamically changing the component biasing states and automatically tracking parameter variation to release voltage margin. Based on the idea of the self-adaptive margin assignment technique, this paper proposes a master-slave Operational Transconductance Amplifier (OTA) structure and master-slave global gain automatic calibration technology for low voltage RF and analog circuits. In the master-slave OTA structure, the master amplifier keeps track of the current variation of the slave amplifier and dynamically adjusts the gate-source voltage V_{GS} of tail current transistors to stabilize the drain current I_{D}. The master-slave global gain auto-calibration technology enables the master stage to automatically adjust the current for achieving precise transconductance. The master stage monitors transconductance variations in the real time. The slave stage tracks the states of the master stage, so the slave stage achieves the same accurate transconductance as the master stage, and eliminates the overall PVT fluctuations.

4.2.1 Master-slave OTA

Figure 12 indicates the circuit schematic of the proposed master-slave OTA structure[55]. The OTA employs complementary transconductance (gm) stages, and a common gate stage is applied to superimpose the currents of the gm stages. Here, the principle of the master-slave structure is explained with the PMOS input stage. In Fig. 12a, the master stage is composed of the tail current transistor P6, the differential transconductance transistors P9 and P10, and the current source N12. The slave stage copies the differential transconductance transistors and tail current transistor of the slave stage in equal proportions, which are P7, P8, and P5, respectively. In addition, the slave stage also contains the current mirror loads P3 and P4. The gate of P5 is connected to the gate of P6, and the gates of P7 and P8 are connected to the gates of P10 and P9, respectively. Both P5 and P6 work in the linear region to save the drain-source voltages, so more voltage margins are allocated to transconductance transistors and load transistors to obtain greater voltage gain. The drains of P9 and P10 are connected to the drain of the current from source of N12, and are also connected to the gate of P6, constructing a negative feedback loop. The negative feedback loop dynamically revises the gate voltage of P6 as the input voltage varies so that the current of P6 is always equal to the current from source of N12. Since the transconductance transistors and tail current transistor of the slave stage are in a copy relationship with the master circuit, the V_{DS} of P5 and P6 are always the same under any input voltages. Therefore, the dynamic adjustment of the gate voltage of P5 keeps the current flowing through P5 unchanged, that is multiple of the current from source of N12. As shown in Fig. 12c, although the tail current sources work in the linear region, the proposed master-slave OTA structure achieves high CMRR. The proposed master-slave OTA structure is suitable for circuits used in PLLs and RF transceivers.



(a)

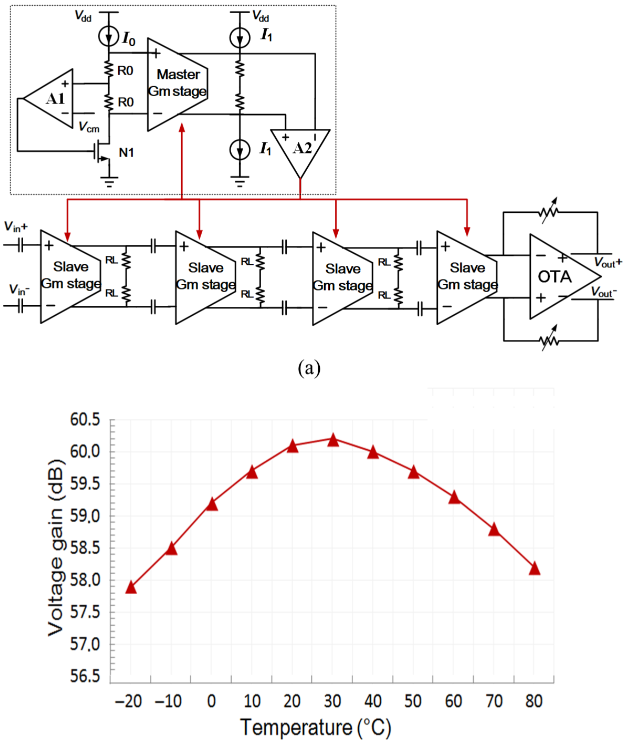
(b)

(c)

Fig. 12 The proposed master-slave OTA. (a) Circuit schematic of the master-slave OTA, (b) circuit schematic of the OTA's dynamic biasing with low V_{DS}, and (c) CMRR of the master-slave OTA.

4.2.2 Master-slave global gain auto-calibration technology

Figure 13 illustrates the circuit diagram of the presented master-slave global gain automatic calibration technology. The master transconductance stage gets a fixed input voltage from the negative feedback loop constructed by the amplifier A1 and a constant output current by the amplifier A2 negative feedback loop, so the master transconductance stage achieves accurate transconductance. The product of transconductance and output resistance is the voltage gain of the transconductance amplifier, so the voltage gain is approximately a ratio of two resistors and the circuit voltage gain is independent of PVT deviation. Slave transconductance stages replicate the master transconductance stage, so PVT deviations are eliminated synchronously. This master-slave global gain automatic calibration technology is applied to the Programmable Gain Amplifier (PGA), which achieves the highest gain of 60 dB, and the step gain fluctuation is less than 1 dB.



(a)

(b)

Fig. 13 The proposed master-slave global gain automatic calibration technology. (a) Circuit schematic of the master-slave global gain automatic calibration technology, and (b) voltage gain of the PGA adopted the master-slave global gain automatic calibration technology.

5 Timing Analysis

Low-voltage technology has widespread applications in the IoTs, intelligent devices, and other fields for power consumption reduction and higher energy efficiency. However, the delay fluctuation becomes larger and larger with the decrease of voltage, which can no longer be ignored. As a result, the traditional timing analysis method is no longer applicable, which brings new challenges to timing analysis. Seok et al.[56] and Borkar et al.[57] found that the working frequency under low voltage (400 mV) should be reduced by 90% to ensure the normal operation of the circuit compared with 1.1 V working voltage. Therefore, the circuit with low voltage may not work properly if the margin is too small; if the margin is too large, the benefit of energy consumption brought by voltage reduction may not be worth the loss. The new challenges for timing analysis under low voltage are as follows: First, the relationship between delay and the process parameters under low voltage is nonlinear, and the delay obeys non-Gaussian distribution instead of Gaussian distribution[58]. Second, the correlation between different cells in a circuit path is different and difficult to characterize, which should not be neglected and brings a great challenge to path delay statistical modeling. Third, it is hard to find the best/worst PVT combination in the nominal voltage region due to the sensitivity to PVT corners and the nonlinear relationship between PVT and delay, which means that it is necessary to do timing analysis at hundreds of thousands of PVT corners at the same time and take months to build a library with process fluctuation.

5.1 Conventional approaches for timing analysis

5.1.1 Related researches

Research methods for timing analysis based on local fluctuations under low voltage and advanced technology are mainly divided into the following categories: Monte Carlo (MC) simulation method, discrete numerical method, look-up table-based method, and analytic method. The traditional Monte Carlo method uses pseudo-random sequence for sampling and its convergence rate is very slow. In order to improve the convergence rate, different sampling methods have been proposed, such as Latin Hypercube Sampling (LHS)[59], Markov Chain Monte Carlo (MCMC) method[60], Linear Sampling method (LS)[61], mixture importance sampling method[62], and Quasi-Monte Carlo (QMC) method[63].

The Monte Carlo simulation method is generally used as the golden reference, but the obvious disadvantage is the time-consuming simulation, which is not acceptable for IC design. In the discrete numerical methods, the random variable cell delay is expressed as the polynomial function of the various process parameters, such as the first-order polynomial under high voltage, the second-order polynomial under low voltage due to the nonlinear relationship between delay and process parameters. In order to further reduce computation effort and consider correlation coefficient between cells, various statistical distributions were adopted, including the inverse Gaussian distribution[64], Weibull distribution[65], skew-normal distribution[66], and log-skew-normal distribution[67]. Then with the method of discrete numerical calculation, the maximum and sum operation can be implemented. These methods have the advantage of high precision at the cost of non-negligible computational complexity. The look-up table-based timing analysis methods are the core of Static Timing Analysis (STA) tools, mainly containing On-Chip Variation (OCV)[68], Advanced OCV (AOCV)[69], Parametric OCV (POCV)[70], and Liberty Variation Format (LVF). These methods are based on libraries with look-up tables composed of input transition time and output load capacitance for PVT corners. The advantage is that it can be characterized in advance, but the disadvantage is that the characteristic time is long. What's more, the interpolation calculation on the non-lattice points will bring errors due to nonlinear characteristics for cell timing, especially for low voltages. The analytical methods are to establish a statistical delay model based on the current formula considering process fluctuation. In order to analyze digital circuits around the threshold voltage, Keller et al.[71] presented a simple, continuous, and accurate trans-regional compact model with a broad working range: from several times of the thermal voltage to near twice of the threshold voltage in advanced processes. To evaluate the additional cycle time margin imposed by local variations quickly, Aliolo et al.[72] introduced a new architecture to evaluate the maximum delay variability in logic paths. They also proposed a concept of X_{cell}. The ratio X_{cell} denotes the delay variability of a standard cell, and $\sigma_{\text{FO4}}/\mu_{\text{FO4}}$ of the proposed Fan-Out of 4 (FO4) cell is essentially constant across different processes and PVT corners. Shiomi et al.[73] proposed architecture-level Statistical Static Timing Analysis (SSTA) models which use log-normal distribution to fit path delay distribution at near-threshold voltage. But their model cannot be applied to arbitrary circuit structures because the fitting coefficients in this method are dependent on path, process, and load. A fast tool for obtaining the variation (σ/μ) of delay of every logic path in a comprehensive design for every process corner was introduced in Ref. [74] without prior knowledge of the design, extensive Monte Carlo simulation, or deep understanding of device physics.

5.1.2 Our works

Local variation effects have posed huge challenges to timing estimation. To solve this problem, a semi-analytical statistical delay model is proposed, which combines analytical and simulation-based method for the subthreshold region[75]. The innovation points are as follows: (1) When the input is fast or slow, the sources of gate delay variation are divided into process parameter variation of the current stage and input slew variation. (2) When the input is fast or slow, the output slew variance sources are divided into threshold voltage variation of the current stage and the input slew variation of the previous stage. (3) Based on the delay and output slew deviation, the correlation problem is transformed into an independent problem related to gate delay variation with step input. The model can be extended to the multi-PVT scenarios since the relative variability of step input delay varies independently with process, temperature, voltage, and load. In this work, first, the variance, variability, and relative variability of the gate delay variations are discussed. Some details about gate delay variation properties, such as the output slew variations, delay variations of stages, and path delay variations, are studied. To prove the effectiveness of our model, comparisons with SPICE results, as well as other related works, are implemented. Compared with other related analytical works, the average accuracy improvements are 8.3× , 9.6×, and 2.7× in variance, variability, and max delay, respectively. When the path delay analysis accuracy and running time are basically the same, the characteristic time cost and storage data (from TB to GB) are 3 orders of magnitude lower than the current LVF model in the industry, which provides the possibility for practical low-voltage timing analysis.

5.2 Machine learning for timing analysis

5.2.1 Related researches

In addition to traditional methods for timing analysis, learning-based schemes have also been actively studied recently. At present, machine learning has been widely used in timing prediction of different backgrounds, especially to solve the trade-off between the convergence speed and accuracy. Based on analysis of the logical structural parameters and electrical parameters that affect transition time and incremental delay in Signal-Integrity (SI) mode, Ref. [76] established a machine learning model to predict SI mode path timing with non-SI mode. In Ref. [77], a machine learning model with path-stage bigrams was presented to predict expensive Path Based Analysis (PBA) results using comparatively cheap Graph Based Analysis (GBA) results, it has significantly reduced pessimism, and a low GBA analysis turnaround time is maintained simultaneously. In order to reduce the pessimism of timing prediction in the placement stage, Ref. [78] used the placement information and circuit structure before routing to predict the delay and transition time of each stage after routing, and then obtained the arrival time, required time, and margin of the path through addition and subtraction operation. The loss function proposed in Ref. [78] can significantly reduce the optimism of the prediction, and their model has obvious advantages over commercial tools in accuracy. Low voltage design needs a lot of time on multi-corner analysis, so Ref. [79] pursued a learning-based method to fit wire slew and delay to estimate through a sign-off STA tool. By adopting these models, the accuracy of delay and slew estimations is increased and the gap between the internal timer and sign-off timer is narrowed finally. Machine learning also provides a promising way to solve the time-consuming problem of simulation at low voltage with multiple PVT corners. Reference [80] proposed a machine learning method based on multivariate linear regression to estimate timing results at unobserved corners using timing delay at observed corners, and put forward a kind of backward selection method based on a greedy algorithm to reduce the number of observed corners. In Ref. [81], in order to get a subset of the worst PVT corner at the minimum cost, Ganapathy et al. gradually iterated and expanded the training set using Gaussian process regression modeling.

5.2.2 Our works

Aiming to improve prediction precision and avoid unacceptable simulation costs, two kinds of timing prediction frameworks based on machine learning for wide supply voltage design are proposed, which use feature engineering and data argumentation, respectively.

In the first work[82], in order to enhance the correlation across PVT corners, feature engineering based on dilated Convolutional Neural Network (CNN) and an ensemble model are adopted to increase prediction accuracy. The main contributions of this work can be summarized as below: (1) In order to reduce error and accelerate the convergence, dilated CNN is used to extract features, and it can be proved that ensemble learning model is very beneficial to the timing prediction problem for wide supply voltage; (2) The path delay at high voltage is utilized to predict the path delay at low voltage with the same and different process parameters, whose values of average relative root mean squared error are 2.3% and 4.7%, respectively. As Fig. 14 shows, the first step is to extract features using dilated CNN with the aim of capturing the path delay correlations and other higher-level characteristics at different temperatures. When it comes to predicting models, an ensemble model is adopted. As Fig. 15 indicates, the ensemble model combines a variety of individual models Linear Regression (LR) and LightGBM (LGBM) together to improve the predictive power and stability. Here we use stacking to construct our ensemble model. Experimental results show that the proposed feature engineering and modeling method have good robustness and high accuracy, no matter within or across PVTs. In addition, the presented method can provide new solutions for such timing prediction problems.

Since simulation or/and library characterization at low voltage are extremely time-consuming, the second timing prediction framework[83] solves the problem of data insufficiency issue in timing prediction when training machine-learning based models. The main contributions are summarized as follows: (1) Conditional Generative Adversarial Networks (named CTGAN) is used to capture the path delay distribution for each specific design at different PVTs, analyze its statistical properties and synthesize new data according to the captured distribution. (2) Synthetic Minority Oversampling Technique for Regression (named SMOTER) is employed to synthesize new data to solve the problem of data imbalance in training. (3) Due to high prediction accuracy and fast training speed, LightGBM is applied to estimate path delay. As shown in Fig. 16, the original path delays of known circuits (ckt) at wide supply voltage range are acquired by simulation. The features are at high voltages and the labels are at low voltages, and the inputs of CTGAN are those original path delays, as well as random noise, to produce synthetic data. Then the synthetic path delays

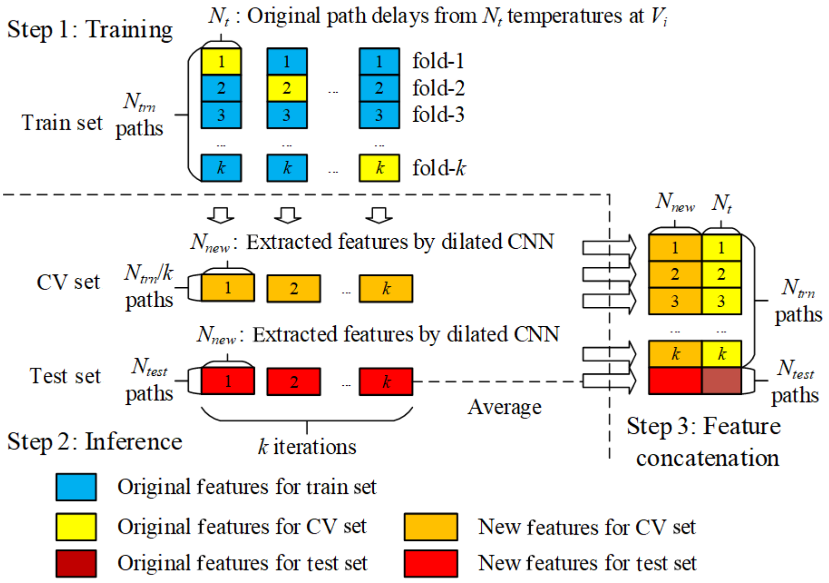

Step 1: Training
N_t: Original path delays from N_t temperatures at V_i
Train set
N_{trn} paths
fold-1
fold-2
fold-3
fold-k
N_{new}: Extracted features by dilated CNN
CV set
N_{trn}/k paths
Test set
N_{test} paths
N_{new}: Extracted features by dilated CNN
Average
Step 2: Inference
k iterations
N_{new} N_t
N_{trn} paths
N_{test} paths
Step 3: Feature concatenation
Original features for train set
Original features for CV set
Original features for test set
New features for CV set
New features for test set

Fig. 14 Feature extraction.

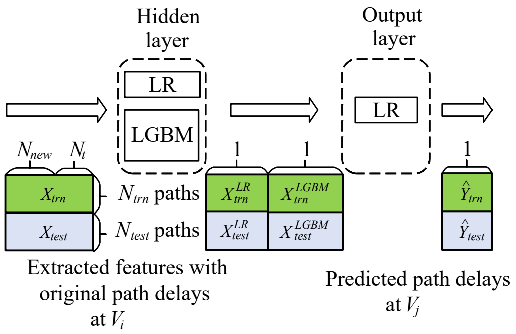


Fig. 15 Ensemble model.

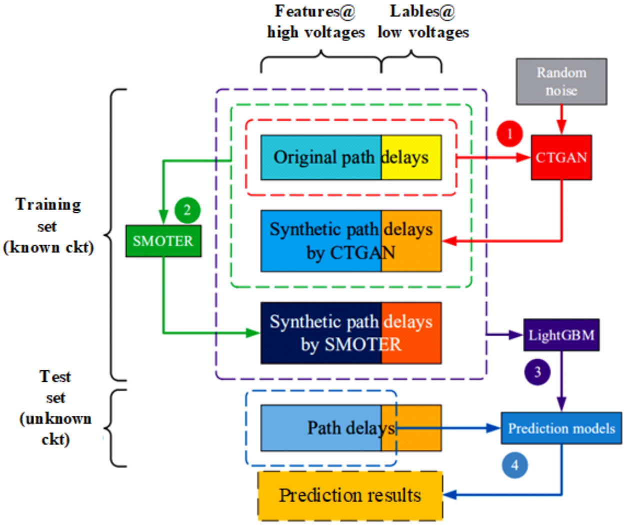
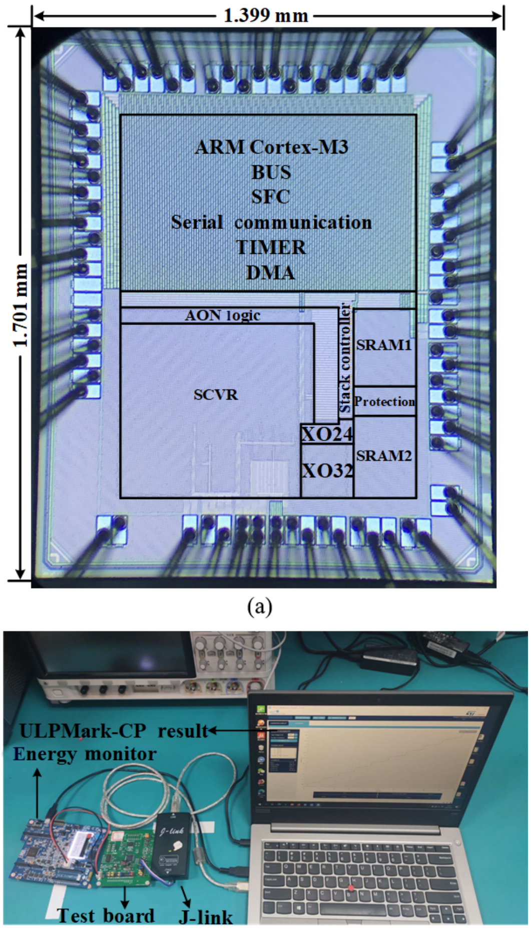


Fig. 16 Structure of the proposed timing prediction framework.

and the original path delays are concatenated and fed into SMOTER to balance data. Finally, the original data and the augmentation data are utilized to train the LightGBM model, and it can predict unknown circuits' path delays at low voltage with less sampling effort and high prediction accuracy. The proposed framework is validated on the TAU2019 benchmark and significant prediction improvement is obtained on both accuracy and efficiency.

6 Dynamic Voltage-Stacking MCU

With the proposed self-adaptive margin assignment techniques for SRAM, digital logic circuits, and analogue circuits, a dynamic voltage-stacking MCU is designed and introduced in Ref. [84]. Previous voltage stacking architecture cannot achieve dynamic switching between flat mode and stack mode, which results in high dynamic power at normal state, in addition, Switched-Capacitor Voltage Regulator (SCVR) still consumes part of the energy in sleep state. In the proposed MCU, two operating modes are supported: a flat mode at the normal state and a stack mode at the sleep state. The instances of the XO32, the retention memory, and the Real-Time Clock (RTC) in parallel are powered by the SCVR at the flat mode. And the SRAM1 (level1), the XO32, the SRAM2 (level2), and the RTC (level3) are connected in series at the stack mode. To save power consumption, the on-chip SCVR is shut down. The measurements show that this MCU improves the ULPMark-CP score by 10.6% than the top1 in the ULPMark score list. Sleep current is 31.2% of the top2 in ULPMark score list (the data of top1 is not available). Figure 17 shows the die photo and the test board, the active area is 2.38 mm^2. For a fair comparison,



(a)

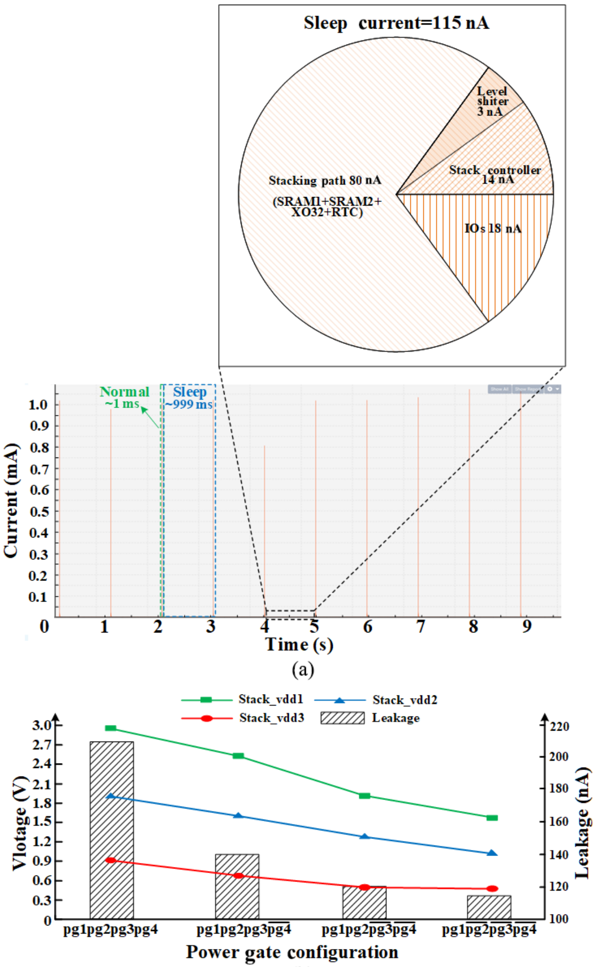
(b)

Fig. 17 Diagram of MCU. (a) Die photograph of the MCU, and (b) testboard.

the test chip can be configured into two modes, one mode is the flat mode at normal operation state and sleep state with SCVR, the other mode is the dynamic switching mode of flat/stack architecture in which flat architecture is at normal operation state and stack architecture is at sleep state without SCVR.

Figure 18 illustrates the measured current consumption with ULPMark-CP. The MCU performs the benchmark program within 1 ms, then sleeps beyond 999 ms, and over and over. With different power gate (pg) configuration in the stacking path, the Stack_vdd3, Stack_vdd2, and Stack_vdd1 are reduced by 44%, 42%, and 45.8%, respectively. When pg2–pg4 are enabled, the average sleep current is reduced to only 115 nA, concluding stacking path current (SRAM1+SRAM2+XO32+RTC) 85 nA, leakage current in level shifters 10 nA, stack controller 5 nA, and IOs leakage current 15 nA.

The screenshot of ULPMark-CP is indicated in Fig. 19, as well as the comparison table. When the flat architecture is configured with SCVR, the sleep current of the MCU is 170 nA and ULPMark score is 920. When the dynamic switching mode of flat/stack architecture is configured, the sleep current is decreased into 115 nA, and the score is increased to 1205. Even compared with ON Semiconductor RSL10 (top1 in ULPMark score list) and Ambiq Apollo512-KBR (top2), the dynamic voltage stacking solution increases the scores by 10.6% and 219%, respectively.



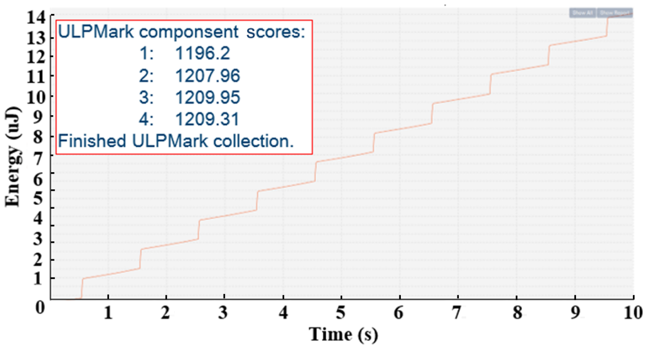
(a)

(b)

Fig. 18 Measured ULPMark-CP. (a) Measured ULPMark-CP current waveform, and (b) stacking voltage and leakage with different power gate configuration during sleep state.

7 Conclusion

This paper reviews near-threshold wide-voltage designs on memory, resilient logic designs, low voltage RF circuits, and timing analysis. To improve the power efficiency significantly and overcome the performance degradation across PVT, the system is operated at a near threshold voltage to save supply voltage, and a self-adaptive margin assignment technique is adopted



ULPMark componsent scores:
1: 1196.2
2: 1207.96
3: 1209.95
4: 1209.31
Finished ULPMark collection.

(a)

	Architecture	Process	Voltage	Frequency	CPU	SRAM size@sleep	I/O	Sleeping current	ULPMark score
ON Semiconductor RSL10	Flat	55 nm	3 V	24 MHz@run 32 KHz@sleep	32-bits ARM Cortex-M3	8 KB	I2C/UART/SPI	–	1090
Ambiq Apollo512-KBR	Flat	40 nm	3 V	24 MHz/ 1 MHz@run 32 KHz@sleep	32-bits ARM Cortex-M4	8 KB	I2C/UART/SPI	369 nA@3V	378
This paper	Flat	40 nm	3 V	24 MHz@run 32 KHz@sleep	32-bits ARM Cortex-M3	8 KB	I2C/UART/SPI	170 nA@3V	920
	Flat/stack dynamic Switching	40 nm	3 V	24 MHz@run 32 KHz@sleep	32-bits ARM Cortex-M3	8 KB	I2C/UART/SPI	115 nA@3V	1205

(b)

Fig. 19 Summary of ULPMark-CP. (a) Measured ULPMark-CP scores and (b) performance summary of voltage-stacking MCU.

to track circuit states and assign voltage margins dynamically. The self-adaptive margin assignment technique is used in SRAM, digital circuits, and analog/RF circuits. Employing the self-adaptive margin assignment technique, the minimum supply voltage in the 40 nm CMOS process is reduced to 0.6 V and the energy efficiency is increased by 3–4 times.

References

[1] S. K. Hsu, A. Agarwal, M. A. Anders, S. K. Mathew, H. Kaul, F. Sheikh, and R. K. Krishnamurthy, A 280 mV-to-1.1 V 256b reconfigurable SIMD vector permutation engine with 2-Dimensional shuffle in 22 nm Tri-gate CMOS, *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 118–127, 2013.

[2] F. Moradi, D. T. Wisland, H. Mahmoodi, A. Peiravi, S. Aunet, and T. V. Cao, New subthreshold concepts in 65 nm CMOS technology, in *Proc. 10^{th} Int. Symp. on Quality Electronic Design*, San Jose, CA, USA, 2009, pp. 162–166.

[3] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, et al., Energy-efficient subthreshold processor design, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 17, no. 8, pp. 1127–1137, 2009.

[4] S. Chatterjee, Y. Tsividis, and P. Kinget, 0.5-V analog circuit techniques and their application in OTA and filter design, *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2373–2387, 2005.

[5] A. Balankutty, S. A. Yu, Y. Feng, and P. R. Kinget, A 0.6-V zero-IF/low-IF receiver with integrated fractional-N synthesizer for 2.4-GHz ISM-band applications, *IEEE J. Solid-State Circuits*, vol. 45, no. 3, pp. 538–553, 2010.

[6] P. Yang, X. Ye, Y. Zhao, W. Zhang, S. Huang, Y. Huang, and Y. Wang, An error detecting scheme with input offset regulation for enhancing reliability of ultralow-voltage SRAM, *Microelectron. Reliab.*, vol. 114, p. 113788, 2020.

[7] J. Yang, H. Ji, Y. Guo, J. Zhu, Y. Zhuang, Z. Li, X. Liu, and L. Shi, A double sensing scheme with selective bitline voltage regulation for ultralow-voltage timing speculative SRAM, *IEEE J. Solid-State Circuits*, vol. 53, no. 8, pp. 2415–2426, 2018.

[8] H. Attarzadeh and M. Sharifkhani, An auto-calibrated, dual-mode SRAM macro using a hybrid offset-cancelled sense amplifier, *Microelectron. J.*, vol. 45, no. 6, pp. 781–792, 2014.

[9] N. Verma and A. P. Chandrakasan, A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy, *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, 2008.

[10] E. Karl, D. Sylvester, and D. Blaauw, Timing error correction techniques for voltage-scalable on-chip memories, in *Proc. IEEE Int. Symp. on Circuits and Systems*, Kobe, Japan, 2005, pp. 3563–3566.

[11] M. Khayatzadeh, M. Saligane, J. Wang, M. Alioto, D. Sylvester, and D. Sylvester, 17.3 a reconfigurable dual-port memory with error detection and correction in 28 nm

FDSOI, in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2016, pp. 310–312.

[12] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS, *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, 2009.

[13] S. Shen, T. Shao, X. Shang, Y. Guo, M. Ling, J. Yang, and L. Shi, TS cache: A fast cache with timing-speculation mechanism under low supply voltages, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 28, no. 1, pp. 252–262, 2020.

[14] Y. Zhou, H. Cai, L. Xie, M. Han, M. Liu, S. Xu, B. Liu, W. Zhao, and J. Yang, A self-timed voltage-mode sensing scheme with successive sensing and checking for STT-MRAM, *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 67, no. 5, pp. 1602–1614, 2020.

[15] Y. Masuda, J. Nagayama, H. Takeno, Y. Ogawa, Y. Momiyama, and M. Hashimoto, Comparing voltage adaptation performance between replica and in-situ timing monitors, in *Proc. 2018 IEEE/ACM Int. Conf. on Computer-Aided Design*, San Diego, CA, USA, 2018, pp. 1–8.

[16] S. Das, D. Roberts, S. Lee, S. Pant, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, A self-tuning DVS processor using delay-error detection and correction, *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 792–804, 2006.

[17] S. Das, C. Tokunaga, S. Pant, W. H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, Razorii: In situ error detection and correction for PVT and SER tolerance, *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, 2009.

[18] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. M. Harris, D. Blaauw, and D. Sylvester, Bubble razor: Eliminating timing margins in an ARM cortex-M3 processor in 45 nm CMOS using architecturally independent error detection and correction, *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 66–81, 2013.

[19] I. Kwon, S. Kim, D. Fick, M. Kim, Y. P. Chen, and D. Sylvester, Razor-lite: A light-weight register for error detection by observing virtual supply rails, *IEEE J. Solid-State Circuits*, vol. 49, no. 9, pp. 2054–2066, 2014.

[20] Y. Zhang, M. Khayatzadeh, K. Yang, M. Saligane, N. Pinckney, M. Alioto, D. Blaauw, and D. Sylvester, iRazor: Current-based error detection and correction scheme for PVT variation in 40-nm ARM Cortex-R4 processor, *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 619–631, 2018.

[21] P. N. Whatmough, S. Das, and D. M. Bull, A low-power 1-GHz Razor FIR accelerator with time-borrow tracking pipeline and approximate error correction in 65-nm CMOS, *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 84–94, 2014.

[22] S. Kim and M. Seok, Variation-tolerant, ultra-low-voltage microprocessor with a low-overhead, within-a-cycle in-situ timing-error detection and correction technique, *IEEE J. Solid-State Circuits*, vol. 50, no. 6, pp. 1478–1490, 2015.

[23] X. Shang, W. Shan, J. Xu, M. Lu, Y. Xiang, L. Shi, and J. Yang, A 0.46 V–1.1 V transition-detector with in-situ timing-error detection and correction based on pulsed-latch design in AES accelerator, in *Proc. 2018 IEEE Asian Solid-State Circuits Conf.* (*A-SSCC*), Tainan, China, 2018, pp. 1–4.

[24] W. Shan, X. Shang, L. Shi, W. Dai, and J. Yang, Timing error prediction AVFS with detection window tuning for wide-operating-range ICs, *IEEE Trans. Circuits Syst. II*: *Express Briefs*, vol. 65, no. 7, pp. 933–937, 2018.

[25] H. Reyserhove and W. Dehaene, Margin elimination through timing error detection in a near-threshold enabled 32-bit microcontroller in 40-nm CMOS, *IEEE J. Solid-State Circuits*, vol. 53, no. 7, pp. 2101–2113, 2018.

[26] W. Jin, S. Kim, W. He, Z. Mao, and M. Seok, In situ error detection techniques in ultralow voltage pipelines: Analysis and optimizations, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 3, pp. 1032–1043, 2017.

[27] C. Lin, W. He, Y. Sun, B. Pei, Z. Mao, and M. Seok, 25.8 a near-threshold-voltage network-on-chip with a metastability error detection and correction technique for supporting a quad-voltage/frequency-domain ultra-low-power system-on-a-chip, in *Proc. 2020 IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2020, pp. 394–396.

[28] J. Zhou, L. Xin, Y. H. Lam, C. Wang, K. H. Chang, J. Lan, and M. Je, HEPP: A new in-situ timing-error prediction and prevention technique for variation-tolerant ultra-low-voltage designs, in *Proc. 2013 IEEE Asian Solid-State Circuits Conf.*, Singapore, 2013, pp. 129–132.

[29] M. A. Ealey and J. F. Mark, Continuous facesheet low voltage deformable mirrors, *Opt. Eng.*, vol. 29, no. 10, pp. 1191–1198, 1990.

[30] S. Akui, K. Seno, M. Nakai, T. Meguro, T. Seki, T. Kondo, A. Hashiguchi, H. Kawahara, K. Kumano, and M. Shimura, Dynamic voltage and frequency management for a low-power embedded microprocessor, in *Proc. 2004 IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2005, pp. 28–35.

[31] K. Agarwal and K. J. Nowka, Dynamic power management by combination of dual static supply voltages, in *Proc. 8^{th} Int. Symp. on Quality Electronic Design*, San Jose, CA, USA, 2007, pp. 85–92.

[32] K. A. Bowman, J. W. Tschanz, S. L. L. Lu, P. A. Aseron, M. M. Khellah, A. Raychowdhury, B. M. Geuskens, C. Tokunaga, C. B. Wilkerson, T. Karnik, et al., A 45 nm resilient microprocessor core for dynamic variation tolerance, *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 194–208, 2011.

[33] I. Ikenaga, M. Nomura, S. Suenaga, H. Sonohara, Y. Horikoshi, T. Saito, Y. Ohdaira, Y. Nishio, T. Iwashita, M. Satou, et al., A 27% active-power-reduced 40-nm CMOS multimedia SoC with adaptive voltage scaling using distributed universal delay lines, *IEEE J. Solid-State Circuits*, vol. 47, no. 4, pp. 832–840, 2012.

[34] R. Salvador, A. Sanchez, X. Fan, and T. Gemmeke, A Cortex-M3 based MCU featuring AVS with 34nW static power, 15.3pJ/inst. active energy, and 16% power variation across process and temperature, in *Proc. ESSCIRC 2018 - IEEE 44^{th} European Solid State Circuits Conf.* (*ESSCIRC*), Dresden, Germany, 2018, pp. 278–281.

[35] J. Kim, K. Choi, Y. Kim, W. Kim, K. Do, and J. Choi,

Delay monitoring system with multiple generic monitors for wide voltage range operation, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 26, no. 1, pp. 37–49, 2018.

[36] M. Saligane, J. Lee, Q. Dong, M. Yasuda, K. Kumeno, F. Ohno, S. Miyoshi, M. Kawaminami, D. Blaauw, and D. Sylvester, An adaptive body-Biaslna SoC using in situ slack monitoring for runtime replica calibration, in *Proc. 2018 IEEE Symp. on VLSI Circuits*, Honolulu, HI, USA, 2018, pp. 63–64.

[37] W. Shan, X. Shang, X. Wan, H. Cai, C. Zhang, and J. Yang, A wide-voltage-range half-path timing error-detection system with a 9-transistor transition-detector in 40-nm CMOS, *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 66, no. 6, pp. 2288–2297, 2019.

[38] W. Shan, W. Dai, C. Zhang, H. Cai, P. Liu, J. Yang, and L. Shi, TG-SPP: A one-transmission-gate short-path padding for wide-voltage-range resilient circuits in 28-nm CMOS, *IEEE J. Solid-State Circuits*, vol. 55, no. 5, pp. 1422–1436, 2020.

[39] W. Shan, W. Dai, L. Wan, M. Lu, L. Shi, M. Seok, and J. Yang, A bi-directional, zero-latency adaptive clocking circuit in a 28-nm wide AVFS system, *IEEE J. Solid-State Circuits*, vol. 55, no. 3, pp. 826–836, 2020.

[40] W. Shan, L. Wan, X. Liu, X. Shang, W. Dai, S. Shao, J. Yang, and L. Shi, A low overhead, within-a-cycle adaptive clock stretching circuit with wide operating range in 40-nm CMOS, *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 65, no. 11, pp. 1718–1722, 2018.

[41] L. Liu and Z. Wang, Analysis and design of a low-voltage RF CMOS mixer, *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 53, no. 3, pp. 212–216, 2006.

[42] H. Lakdawala, M. Schaecher, C. T. Fu, R. Limaye, J. Duster, Y. Tan, A Balankutty, E. Alpman, C. Lee, S. Suzuki, et al., 32nm x86 OS-compliant PC on-chip with dual-core atom® processor and RF WiFi transceiver, in *Proc. 2012 IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2012, pp. 62–64.

[43] J. Zhang, D. Zhao, and X. You, A 20-GHz 1.9-mW LNA using g_m-boost and current-reuse techniques in 65-nm CMOS for satellite communications, *IEEE J. Solid-State Circuits*, vol. 55, no. 10, pp. 2714–2723, 2020.

[44] M. Parvizi, K. Allidina, and M. N. El-Gamal, Short channel output conductance enhancement through forward body biasing to realize a 0.5 V 250 μW 0.6-4.2 GHz current-reuse CMOS LNA, *IEEE J. Solid-State Circuits*, vol. 51, no. 3, pp. 574–586, 2016.

[45] C. H. Chang, A forward-body-bias CMOS LNA with ultra-low device junction leakage using intrinsic self-balanced pseudo resistor, *IEEE Trans. Circuits Syst. II*: *Express Briefs*, vol. 66, no. 4, pp. 697–701, 2019.

[46] D. Markovic, C. C. Wang, L. P. Alarcon, T. T. Liu, and J. M. Rabaey, Ultralow-power design in near-threshold region, *Proc. IEEE*, vol. 98, no. 2, pp. 237–252, 2010.

[47] M. Tamura, H. Takano, S. Shinke, H. Fujita, H. Nakahara, N. Suzuki, Y. Nakada, Y. Shinohe, S. Etou, T. Fujiwara, and Y. Katayama, A 0.5 V BLE transceiver with a 1.9 mW RX achieving –96.4 dBm sensitivity and 4.1 dB adjacent channel rejection at 1 MHz Offset in 22 nm FDSOI, in *Proc. of IEEE ISSCC Dig. Tech.*, San Francisco, CA, USA, 2020, pp. 468–470.

[48] T. S. Chao, Y. L. Lo, W. B. Yang, and K. H. Cheng, Designing ultra-low voltage PLL using a bulk-driven technique, in 2009 *Proc. ESSCIRC*, Athens, Greece, 2009, pp. 388–391.

[49] S. G. Kim, J. Rhim, D. H. Kwon, M. H. Kim, and W. Y. Choi, A low-voltage PLL with a supply-noise compensated feedforward ring VCO, *IEEE Trans. Circuits Syst. II*: *Express Briefs*, vol. 63, no. 6, pp. 548–552, 2016.

[50] B. Ghafari, L. Koushaeian, and F. Goodarzy, New architecture for an ultra low power and low noise PLL for biomedical applications, in *Proc. 2013 IEEE Global High Tech Congress on Electronics*, Shenzhen, China, 2013, pp. 61–62.

[51] J. W. Moon, S. G. Kim, D. H. Kwon, and W. Y. Choi, A 0.4-V, 500-MHz, ultra-low-power phase-locked loop for near-threshold voltage operation, in *Proc. 2014 Custom Integrated Circuits Conf.*, San Jose, CA, USA, 2014, pp. 1–4.

[52] H. H. Hsieh, C. T. Lu, and L. H. Lu, A 0.5-V 1.9-GHz low-power phase-locked loop in 0.18-μm CMOS, in *Proc. 2007 IEEE Symp. On VLSI Circuits*, Kyoto, Japan, 2007, pp. 164–165.

[53] H. Liu, D. Tang, Z. Sun, W. Deng, H. C. Ngo, K. Okada, and A. Matsuzawa, A 0.98 mW fractional-N ADPLL using 10b isolated constant-slope DTC with FOM of –246 dB for IoT applications in 65nm CMOS, in *Proc. 2018 IEEE Int. Solid - State Circuits Conf.*, 2018, San Francisco, CA, USA, pp. 246–248.

[54] M. Tamura, H. Takano, S. Shinke, H. Fujita, H. Nakahara, N. Suzuki, Y. Nakada, Y. Shinohe, S. Etou, T. Fujiwara, et al., 30.5 A 0.5 V BLE transceiver with a 1.9 mW RX achieving –96.4 dBm sensitivity and 4.1 dB adjacent channel rejection at 1 MHz offset in 22 nm FDSOI, in *Proc. 2020 IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, USA, 2020, pp. 468–470.

[55] C. Chen and J. Wu, 0.6-V 2.1-mW RF receiver based on passive mixing and master–slave common-mode rejection technique in 65ńm CMOS, *Electron. Lett.*, vol. 52, no. 5, pp. 335–336, 2016.

[56] M. Seok, S. Hanson, Y. S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, The phoenix processor: A 30 pW platform for sensor applications, in *Proc. 2008 IEEE Symp. on Vlsi Circuits*, Honolulu, HI, USA, 2008, pp. 188–189.

[57] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, Parameter variations and impact on circuits and microarchitecture, in *Proc. 2003 Design Automation Conf.*, Anaheim, CA, USA, 2003, pp. 338–342.

[58] F. Frustaci, P. Corsonello, and S. Perri, Analytical delay model considering variability effects in subthreshold domain, *IEEE Trans. Circuits Syst. II*: *Express Briefs*, vol. 59, no. 3, pp. 168–172, 2012.

[59] W. L. Loh, On Latin hypercube sampling, *Ann. Statist.*, vol. 24, no. 5, pp. 2058–2080, 1996.

[60] S. Reh, P. Lethbridge, and D. Ostergaard, Quality based design and design for reliability of micro electro mechanical systems (MEMS) using probabilistic methods, in *Proc. 2000 In. Conf. on Modeling and Simulation of Microsystems*, San Diego, CA, USA, 2000, pp. 708–711.

[61] H. Awano and T. Sato, Efficient transistor-level timing yield estimation via line sampling, in *Proc. 53rd ACM/EDAC/IEEE Design Automation Conf.*, Austin, TX, USA, 2016, pp. 1–6.

[62] R. Kanj, R. Joshi, and S. Nassif, Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events, in *Proc. 2006 43rd ACM/IEEE Design Automation Conf.*, San Francisco, CA, USA, 2006, pp. 69–72.

[63] R. E. Caflisch, Monte Carlo and Quasi-Monte Carlo methods, *Acta Numer.*, vol. 7, pp. 1–49, 1998.

[64] J. Chen, S. Cotofana, S. Grandhi, C. Spagnol, and E. Popovici, Inverse Gaussian distribution based timing analysis of sub-threshold CMOS circuits, *Microelectron. Reliab.*, vol. 55, no. 12, pp. 2754–2761, 2015.

[65] L. Zhang, J. Shao, and C. C. P. Chen, Non-gaussian statistical parameter modeling for SSTA with confidence interval analysis, in *Proc. 2006 Int. Symp. on Physical Design*, San Jose, CA, USA, 2006, pp. 33–38.

[66] S. Ramprasath, M. Vijaykumar, and V. Vasudevan, A skew-normal canonical model for statistical static timing analysis, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 24, no. 6, pp. 2359–2368, 2016.

[67] H. A. Balef, M. Kamal, A. Afzali-Kusha, and M. Pedram, All-region statistical model for delay variation based on log-skew-normal distribution, *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 35, no. 9, pp. 1503–1508, 2016.

[68] K. J. Chang, Accurate on-chip variation modeling to achieve design for manufacturability, in *Proc. 4th IEEE Int. Workshop on System-on-Chip for Real-Time Applications*, Banff, Canada, 2004, pp. 219–222.

[69] J. Hong, K. Huang, P. Pong, J. D. Pan, J. Kang, and K. C. Wu, An LLC-OCV methodology for statistic timing analysis, in *Proc. 2007 Int. Symp. on Vlsi Design, Automation and Test*, Hsinchu, China, 2007, pp. 1–4.

[70] A. Mutlu, J. Le, R. Molina, and M. Celik, A parametric approach for handling local variation effects in timing analysis, in *Proc. 2009 46th ACM/IEEE Design Automation Conf.*, Francisco, CA, USA, 2009, pp. 126–129.

[71] S. Keller, D. M. Harris, and A. J. Martin, A compact transregional model for digital CMOS circuits operating near threshold, *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 22, no. 10, pp. 2041–2053, 2014.

[72] M. Alioto, G. Scotti, and A. Trifiletti, A novel framework to estimate the path delay variability on the back of an envelope via the fan-out-of-4 metric, *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 64, no. 8, pp. 2073–2085, 2017.

[73] J. Shiomi, T. Ishihara, and H. Onodera, Microarchitectural-level statistical timing models for near-threshold circuit design, in *Proc. 20th Asia and South Pacific Design Automation Conf.*, Chiba, Japan, 2015, pp. 87–93.

[74] Y. Zhang and B. H. Calhoun, Fast, accurate variation-aware path timing computation for sub-threshold circuits, in *Proc. 15th Int. Symp. on Quality Electronic Design*, Santa Clara, CA, USA, 2014, pp. 243–248.

[75] J. Guo, P. Cao, M. Li, Y. Gong, Z. Liu, G. Bai, and J. Yang, Semi-analytical path delay variation model with adjacent gates decorrelation for subthreshold circuits, *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 40, no. 5, pp. 931–944, 2020.

[76] A. B. Kahng, M. Luo, and S. Nath, SI for free: Machine learning of interconnect coupling delay and transition effects, in *Proc. 2015 ACM/IEEE Int. Workshop on System Level Interconnect Prediction* (*SLIP*), San Francisco, CA, USA, 2015, pp. 1–8.

[77] A. Kahng, U. Mallappa, and L. Saul, Using machine learning to predict path-based slack from graph-based timing analysis, in *Proc. 2018 IEEE 36th Int. Conf. on Computer Design* (*ICCD*), Orlando, FL, USA, 2018, pp. 603–612.

[78] E. C. Barboza, N. Shukla, Y. Chen, and J. Hu, Machine learning-based pre-routing timing prediction with reduced pessimism, in *Proc. 2019 56th ACM/IEEE Design Automation Conf.* (*DAC*), Las Vegas, NV, USA, 2019, pp. 1–6.

[79] A. B. Kahng, S. Kang, H. Lee, S. Nath, and J. Wadhwani, Learning-based approximation of interconnect delay and slew in signoff timing tools, in *Proc. 2013 ACM/IEEE Int. Workshop on System Level Interconnect Prediction*, Austin, TX, USA, 2013, pp. 1–8.

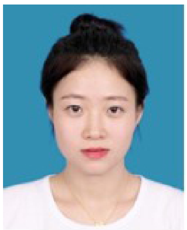
[80] A. B. Kahng, U. Mallappa, L. Saul, and S. Tong, "Unobserved corner" prediction: Reducing timing analysis effort for faster design convergence in advanced-node design, in *Proc. 2019 Design, Automation & Test in Europe Conf. & Exhibition* (*DATE*), Florence, Italy, 2019, pp. 168–173.

[81] S. Ganapathy, R. Canal, A. Gonzalez, and A. Rubio, Circuit propagation delay estimation through multivariate regression-based modeling under spatio-temporal variability, in *Proc. 2010 Design, Automation & Test in Europe Conf. & Exhibition,* Dresden, Germany, 2010, pp. 417–422.

[82] W. Bao, P. Cao, H. Cai, and A. G. Bu, A learning-based timing prediction framework for wide supply voltage design, in *Proc. 2020 on Great Lakes Symp. on VLSI*, New York, NY, USA, 2020, pp. 309–314.

[83] P. Cao, W. Bao, K. Wang, and T. Yang, A timing prediction framework for wide voltage design with data augmentation strategy, in *Proc. 2021 26th Asia and South Pacific Design Automation Conf.*, Tokyo, Japan, 2021, pp. 291–296.

[84] X. Li, Y. Xu, L. Ren, W. Ge, J. Cai, X. Liu, and J. Yang, 29.8 115nA@3V ULPMark-CP score 1205 SCVR-less dynamic voltage-stacking scheme for IoT MCU, in *Proc. 2021 IEEE Int. Solid- State Circuits Conf.*, San Francisco, CA, USA, 2021, pp. 100–102.



Yan Zhao received the BEng degree in electronic engineering from Hefei University of Technology, Hefei, China in 2018. She is currently a PhD candidate in electronic engineering at Southeast University, Nanjing, China. Her research mainly focuses on low-power radio frequency integrated circuit design.

Jun Yang received the BEng and PhD degrees in electronic engineering from Southeast University, Nanjing, China in 1999 and 2004, respectively. He is currently a professor at the National ASIC Center, Southeast University. He has authored or coauthored over 50 technical articles in conferences and journals, including IEEE International Solid-State Circuits Conference (ISSCC), Design Automation Conference (DAC), the *Journal of Solid-State Circuits*, *IEEE Transactions on Circuits and Systems I*: *Regular Papers* (TCASI), and *IEEE Transactions on Very Large Scale Integration* (*TVLSI*) *Systems*. He has authorized over 100 Chinese and U.S. invention patents. His current researches focus on SRAM design, in-memory computing, and near-threshold design.

Chao Chen received the BEng, MEng, and PhD degrees in electrical engineering from Southeast University, Nanjing, China in 2007, 2010, and 2013, respectively. In January 2014, he joined Southeast University. He is now an associated professor at the National ASIC Centre. His research interests are on analog and Radio Frequency (RF) circuits and systems, including RF front-end, frequency synthesizer, and analog-digital converters. His recent research focuses on low voltage and ultra-low-power transceivers.

Yongliang Zhou received the MEng degree in circuit and system from Anhui University, Hefei, China in 2017. He is currently a PhD candidate in electronic engineering at Southeast University, Nanjing, China. His current research interests include emerging device-circuit interaction design and computing in memory.

Ziyu Li received the BEng degree in electronic science and technology from Southeast University, Nanjing, China in 2019, where he is a master student in electronic engineering. His research mainly focuses on variation resilient adaptive VLSI circuits and ultralow-power System on Chip (SoC) design.

Tai Yang received the BEng degree in electronic science and technology from Southeast University, Nanjing, China in 2019, where she is a master student in electronic engineering. Her research mainly focuses on the statistical timing analysis and optimization.

Weiwei Shan received the BEng degree in microelectronics from Tianjin University, China in 2003, and the PhD degree in microelectronics from Tsinghua University, China in 2009. She was a visiting professor at Columbia University, New York, NY, USA from 2017 to 2019. She is currently a professor at the National ASIC Center, Southeast University, Nanjing, China. Her research mainly focuses on variation resilient, adaptive VLSI circuits, ultra-low-power SoC design, and countermeasure techniques of security circuits. She has authored or coauthored over 50 technical articles in conferences and journals, including IEEE International Solid-State Circuits Conference (ISSCC), *Journal of Solid-State Circuits*, *IEEE Transactions on Circuits and Systems I*: *Regular Papers* (TCASI), *IEEE Transactions on Circuits and Systems* II: *Express Briefs* (TCASII), *IEEE Transactions on Computer Aided Design of Integrated Circuits* & *Systems* (TCAD), and authorized over 25 invention patents. She was a recipient of the 2014 State Scientific and Technological Progress Award, the 2017 A-SSCC Distinguished Design Award, and the 2018 GLSVLSI Best Paper Candidate.

Peng Cao received the BEng and PhD degrees in microelectronics and solid-state electronics from Southeast University, Nanjing, China in 2002 and 2010, respectively. He joined research at the University of Waterloo, Waterloo, ON, Canada from 2016 to 2017, as a visiting scholar. He is currently an associate professor at the National ASIC Center, Southeast University. His research interests are focused on statistical timing analysis and low-voltage VLSI designs.