A Comparative Analysis of Low Temperature and Room Temperature Circuit Operation

Zhichao Chen, Ali H. Hassan, *Senior Member, IEEE*, Rhesa Ramadhan, *Student Member, IEEE*, Yingheng Li, Chih-Kong Ken Yang, *Fellow, IEEE*, Sudhakar Pamarti, *Senior Member, IEEE*, and Puneet Gupta, *Fellow, IEEE*

***Abstract*— Low-temperature (LT) conditions can potentially lead to lower power consumption and enhanced performance in circuit operations by reducing the transistor leakage current, increasing carrier mobility, reducing wear-out, and reducing interconnect resistance. We develop PROCEED-LT, a pathfinding framework to co-optimize devices and circuits over a wide performance range. Our results demonstrate that circuit operations at LT (−196 °C) reduce power compared to room temperature (RT, 85 °C) by 15× to over 23.8× depending on performance level. Alternatively, LT improves performance by 2.4× (high-power, high-performance) − 7.0× (low-power, low-performance) at the same power point. These gains are further improved in low-activity circuits and when using multivoltage configurations. Meanwhile, we highlight the need for improvement in V_{th} variation to leverage benefits at cryogenic temperatures.**

***Index Terms*— 77 K, circuit optimization, cryogenic computing, FinFET, Pareto optimization, process variations, transistor aging.**

I. Introduction

DRIVEN by the burgeoning needs of artificial intelligence and deep learning, the computing demand is escalating at an unprecedented pace. However, the performance boost from technology scaling is stagnant. Increased leakage currents restrict reductions in threshold voltage (V_{th}), while heightened dynamic power constrains the elevation of the supply voltage (V_{dd}). Consequently, these factors collectively limit performance improvements. Furthermore, the threshold voltage (V_{th}) cannot be scaled down aggressively, complicating efforts to reduce the power-supply voltage (V_{dd}) for power conservation. Moreover, enhancing performance by increasing the clock frequency is becoming increasingly challenging as the dynamic power consumption rises accordingly [2]. Thus, innovative approaches are needed to advance modern transistors and interconnect to fulfill the exponentially growing computing demand.

Low-temperature (LT) computing, or operating the computer system at a liquid nitrogen temperature (e.g., 77 K), has emerged as a promising avenue for boosting circuit performance and power efficiency. On the transistor side, LT conditions offer enhancements in subthreshold slope, and they aid in lowering leakage current and allow the reduction of V_{th}. A higher carrier mobility and a lower V_{th} at LT correspondingly enable the reduction of V_{dd} while maintaining the performance with less power consumption [3], [4]. As for interconnects, LT could decrease the bulk wire resistivity with the temperature drop [5], which, in turn, permits the use of thinner wires, further reducing capacitance. Nevertheless, the benefits of LT are not without challenges. LT conditions lead to an increase in the transistor's V_{th} due to the bandgap widening and shifts of the Fermi potential [15], potentially hindering the ability to meet frequency constraints and fully utilize the advantages inherent to LT operation.

At the circuit level, recent studies have successfully exploited the benefits of LT computing by developing LT-optimized cores with smaller micro architectural elements [6], substituting traditional L2 and L3 caches with innovative non-SRAM technologies that minimize chip area and power usage (STT-MRAM [7], [8], GC-eDRAM [9], [10], [11], and 1T Floating Body RAM [12], [13]). This progress also extends to probing downsized interconnect materials such as aluminum at single nanometer scale [14] and efficiently implementing in-memory computing at cryogenic temperatures [16], [17].

Most studies manually perform a simple design technology co-optimization (DTCO) process to determine an optimal combination of V_{dd} and V_{th} for a specific design and its constraints [11], [18], [19], [20], [22], [23], [24]. Initially, the researchers reduce V_{th} to augment the I_{on}/I_{off} ratio, ensuring the circuit's functionality under LT conditions. Researchers commonly refer to this method as "V_{th} engineering." The methodologies for the tuning, however, are diverse. For example, some researchers prefer to tune V_{th} by matching the LT off-leakage current to the off-leakage levels of an off-the-shelf device at 300 K, where the leakage current is deemed reasonable [11], [18], [19], [20]. Efforts also include matching V_{th} at 300 K for low-voltage designs [22], [23].

Received 3 July 2024; revised 25 October 2024; accepted 24 November 2024. Date of publication 11 December 2024; date of current version 31 December 2024. This work was supported by the Defense Advanced Research Project Agency (DARPA) through the Low Temperature Logic Technology (LTLT) Project. *(Corresponding author: Puneet Gupta.)*

Zhichao Chen, Ali H. Hassan, Rhesa Ramadhan, Chih-Kong Ken Yang, Sudhakar Pamarti, and Puneet Gupta are with the Department of Electrical and Computer Engineering, University of California at Los Angeles (UCLA), Los Angeles, CA 90095 USA (e-mail: zhichaochen@ucla.edu; aehassan@ucla.edu; rhesamr@g.ucla.edu; yangck@ucla.edu; spamarti@ee.ucla.edu; puneetg@ucla.edu).

Yingheng Li was with the Department of Electrical and Computer Engineering, UCLA, Los Angeles, CA 90095 USA. He is now with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: yil392@pitt.edu).

Digital Object Identifier 10.1109/TVLSI.2024.3508673

Alternatively, some optimize V_{th} to curtail the leakage power of the circuit based on specified targets of reducing the total power consumption to offset the associated cooling power cost [11]. Once V_{th} is decided, the search commences for the minimum V_{dd} that satisfies the power-delay (PD) constraints. Ultimately, this simple DTCO method yields a set of V_{dd} and V_{th} values (for both nMOS and pMOS) for a specific design that can deliver the optimal PD tradeoffs within the specified constraints. Determining the optimal operating point is critical for circuit performance. Choosing the correct V_{th} that can produce the optimum PD tradeoff is very crucial because the subthreshold leakage current and the delay are sensitive to V_{th} [25]. Moreover, using the lowest V_{dd} possible for the circuit is important to leverage the significant leakage power reduction at LT condition. Nonetheless, this tuning strategy in DTCO is time-consuming, involves massive manual intervention, and can only derive limited sets of V_{th} and V_{dd} configurations. Utilizing a register-transfer level (RTL) compiler to obtain synthesis results requires tens of minutes and, therefore, is impractical for the analysis of the circuit operations across multiple performance targets.

PROCEED [26] provides the ability to derive optimal V_{dd}, V_{th}, and transistor sizing based on the analysis of comprehensive chip characteristics. PROCEED facilitates a more efficient evaluation of specific voltage configurations, typically taking less than 1 min per configuration, and the output is fit into a Pareto front to converge optimal solutions through iterative refinement. Despite its contributions, the existing version of PROCEED is confined to supporting limited technologies and potentially generates suboptimal voltage configurations. Details about the PROCEED framework will be described in Section III.

To address these limitations and align with the advancements in current technology nodes, such as FinFETs, this study introduces PROCEED-LT. Using this tool, we enable the analysis across a spectrum of temperature conditions, particularly LT environments, and focus on exploring the potential benefits and challenges of LT circuit operations. Compared to [6], [11], [18], [20], [23], and [24], this work introduces a novel consideration of device aging, IR drop, and other reliability factors under LT. In addition, it exhibits the benefits of LT operations across a broad range of voltage configurations at the circuit level. This work makes main contributions as follows.

1) A more accurate interconnect model and a device wear-out model are incorporated into PROCEED-LT. The FinFET-based circuit optimization is supported, and an automated multivoltage configuration with less manual intervention is enabled.
2) PD optimization by V_{dd} and V_{th} optimization is supported. The PD benefits of LT circuit operations versus RT are demonstrated when considering the aging effect, the IR drop, and the multivoltage configuration.
3) The various power improvements ranging from 7.97× to 21.88× under LT conditions across different activity factors are explored on two digital circuit benchmarks.
4) The challenge of the higher V_{th} variation sensitivity at LT due to the process variation and the scaled-down operating voltages deteriorating the benefits of LT-computing is proposed.
5) It is demonstrated that under LT conditions, the circuit can achieve higher peak performances (>2×) than RT at equivalent power constraints.

We demonstrate the advantages of LT circuit operations by utilizing PROCEED-LT to evaluate a small microprocessor design (ARM Cortex M3) and an emerging neural networks accelerator design (ACOUSTIC [21]). PROCEED-LT supports the generation of a Pareto front spanning a significant range of power and delay tradeoffs at LT. This capability is essential for uncovering the benefits of LT circuit operations across diverse performance ranges. Meanwhile, we should consider the required cooling cost to remove the heat dissipated from the circuit. The energy benefit of circuit operations at LT should consume at least 10.65× less power than RT to overcome the cooling cost and achieve overall power efficiency [6]. We reveal that the optimization conducted at LT by PROCEED-LT, specifically at −196 °C, can yield power improvements ranging from 10× to more than 20× compared to standard room temperature (RT) of 85 °C with equivalent performances.

The structure of this article is organized in the following manner. Section II delves into the potential benefits of LT computing. Section III explains the challenges of PD evaluation and optimization at LT using PROCEED and the corresponding improvements. Section IV details the experiment results and analysis of our PROCEED-LT's study on optimizing Cortex M3 and ACOUSTIC. Finally, the article reaches its conclusion in Section V.

II. Desirable Traits of LT Circuit Operations

A. Potential of Leakage Current Reduction

As the technology node scales, researchers are chasing higher operating frequencies in large-scale circuits. To address the incurring challenge of dynamic power rise, circuit designers tried to reduce both V_{dd} and V_{th}. However, with the reduced V_{th}, the static power, which is caused by the leakage current of off devices, will grow exponentially. LT computing, which is known as cryogenic computing as well, aims to operate computers at exceedingly LTs. Thanks to the temperature decreasing, the leakage current shrinks exponentially, which significantly reduces the static power and enables continuous reduction of both V_{dd} and V_{th} with no performance loss [27].

B. Potential of Improved Interconnect

To achieve a higher clock frequency, smaller and faster devices are deployed as technology scales. However, the clock frequency's enhancement is concurrently hindered by the speed of data transfer, which will be lower with the increasing wire latency. Unlike transistors, the latency of interconnect cannot be significantly improved as technology nodes shrink. Global interconnects that communicate signals across the chip are difficult to scale in length, and their latencies are mostly maintained [30]. Fortunately, the resistivity of some metal materials (e.g., copper) can linearly decrease with temperature. As a result, the latency of the interconnect is substantially

reduced, thereby allowing for a safe increase in clock frequencies under LT conditions [5]. In addition, as the resistance in the interconnect lowers, the IR drop, an issue that significantly affects voltage delivery and thermal behavior, shows promising improvement.

C. Potential of Circuit Wear-Out Slowdown

The phenomenon of circuit wear-out due to device aging can affect end-of-life reliability and performance of the electronic design [28]. Device aging primarily stems from bias temperature instability (BTI) and hot carrier injection (HCI), which cause irreversible shifts in V_{th} and degrade mobilities over long periods. As shown in the following equations, BTI- and HCI-induced ΔV_{th}'s exhibit an exponential dependence on the channel temperature T_c:

$$\Delta V_{\text{th}}(\text{BTI}) \propto V_{\text{gs}}^{m_1} \cdot e^{-\frac{E_a}{kT_c}} \cdot t^{n_1} \tag{1}$$

$$\Delta V_{\text{th}}(\text{HCI}) \propto V_{\text{ds}}^{m_2} \cdot e^{-\frac{E_a}{kT_c}} \cdot t^{n_2} \tag{2}$$

where E_a denotes the activation energy, t denotes the stress time, and $m_{1,2}$, $n_{1,2}$ are factors obtained from characterization experiments [34], [35], [36]. A reduction in operation temperature can significantly impact ΔV_{th} during long-term device aging. Research indicates that at LT, ΔV_{th} resulting from BTI can be mitigated under the same voltage stress time in both planar or FinFET devices [36], [37], [38]. In the context of HCI, [39], [40] suggest that cryogenic temperatures might exacerbate hot carrier degradation and reduce the device lifetime, especially for NFET. Nonetheless, it is also mentioned that a slight reduction in V_{ds} by reducing V_{dd} at LT can counterbalance these adverse effects and offer additional advantages by lowering power consumption. Cryogenic conditions can dramatically reduce leakage currents in scaled semiconductors, which leads to a decrease in static power consumption, thereby weakening the thermal stress. Moreover, LT computing allows for a reduction in V_{dd} while preserving the performance. Then, the voltage stress on the device can be effectively decreased, and the wear-out process can be decelerated, resulting in a lower V_{th} shift.

III. Optimizing Circuits for LT Operations With PROCEED-LT

A. PROCEED Framework

Fig. 1 presents an overview of the PROCEED framework. The input of PROCEED is the interconnect information (i.e., wire resistance R and capacitance C), the processor benchmark design (e.g., design logic depth histogram (LDH) and average fan-out), the operating activity factor, the variation information, and constraints (highest or lowest V_{dd}, V_{th}, transistor size, and chip area). After the Pareto-based optimization process, PROCEED then outputs the Pareto curve of PD over a wide range of delay and power. Each point on the Pareto curve represents a specific combination of single or multiple V_{dd} and V_{th} values.

Instead of complete and exact optimization using RTL simulation, PROCEED predicts the best performance and power tradeoffs by using essential design information and determining V_{dd} and V_{th}. Since the path delay is roughly proportional to the logic depth, PROCEED uses LDH extracted from a synthesized benchmark processor to predict the path delays. PROCEED divides the logic paths of a processor into n bins based on the logic depth distribution. Logic paths grouped within a bin then share identical delay (stages). Employing a larger number of bins enhances the accuracy at the expense of computation time.

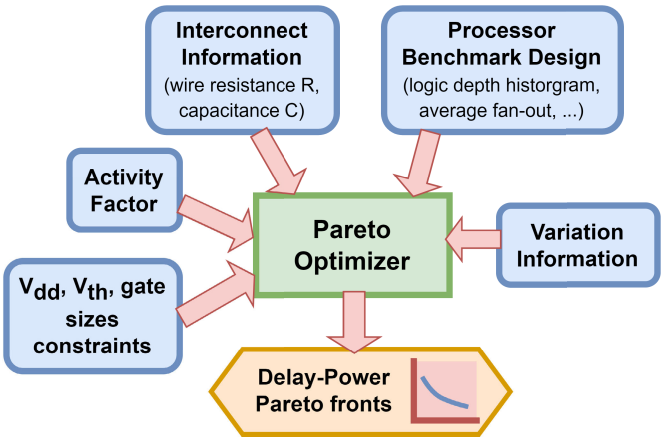


Fig. 1. PROCEED framework.

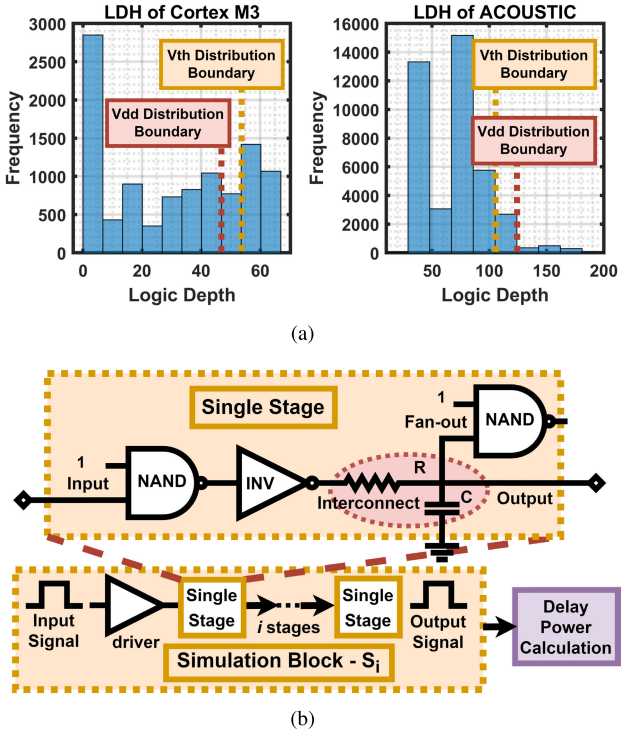


Fig. 2. (a) Logic path depth histogram and the example of V_{dd}, V_{th} distribution. (b) Circuit schematic of the single stage used for simulation and optimization.

Fig. 2(a) shows the LDH of Cortex M3 and ACOUSTIC. For example, we divide the Cortex M3 circuit with the deepest path of 67 stages into $n = 10$ bins. Each bin is modeled by the corresponding simulation block S_i, which is made of i gate stages shown in Fig. 2(b). The delay of the bin i is then estimated by the delay of i copies of the single stage.

Each gate stage consists of the interconnect load including a resistor and a capacitor generated from synthesized results and logic gates of specific sizes based on the design. The delay and power of S_i can be represented by a function of a set of tuning parameters variables. Let the vector $\mathbf{X}$ represent the tuning parameters for optimization

$$\mathbf{X} = (\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}) \tag{3}$$

$$\mathbf{y_i} = \left(V_{\mathrm{dd},i}, V_{\mathrm{th},i}, W_{i,1}, \ldots, W_{i,2i}\right), \quad i = 1, 2, \ldots, n \tag{4}$$

where $\mathbf{y_i}$ is the tuning parameter vector for simulation block S_i, and $V_{\mathrm{dd},i}$, $V_{\mathrm{th},i}$, and $W_{i,j}$ are the supply voltages, threshold voltages, and the sizes of gates and inverters in S_i, respectively. After picking a starting point $\mathbf{X_0}$ for each iteration of the optimization process, the delay D and power P of the circuit configured by $\mathbf{X}$ can be given by the following equations:

$$D(\mathbf{X}) = D(\mathbf{X_0}) + \mathbf{G}_D(\mathbf{X_0})^T(\mathbf{X} - \mathbf{X_0}) + \frac{1}{2}(\mathbf{X} - \mathbf{X_0})^T\mathbf{H}_D(\mathbf{X} - \mathbf{X_0}) \tag{5}$$

$$P(\mathbf{X}) = P(\mathbf{X_0}) + \mathbf{G}_P(\mathbf{X_0})^T(\mathbf{X} - \mathbf{X_0}) + \frac{1}{2}(\mathbf{X} - \mathbf{X_0})^T\mathbf{H}_P(\mathbf{X} - \mathbf{X_0}) \tag{6}$$

where $\mathbf{G}$ and $\mathbf{H}$ are the gradient vector and the Hessian matrix, respectively. By analyzing the output signal of S_i, PROCEED extracts the delay and power simulation results of blocks from SPICE without using a synthesis tool. The critical path delay $D(\mathbf{X})$ is represented by the highest delay of all bins by (7), which is approximated by (8) with the order of the norm $K = 200$ in our experiments. The approximation enables the calculation for $\mathbf{G}$ and $\mathbf{H}$

$$D(\mathbf{X}) = \max\left(D_{S_1}(\mathbf{y_1}), D_{S_1}(\mathbf{y_2}), \cdots, D_{S_n}(\mathbf{y_n})\right) \tag{7}$$

$$D(\mathbf{X}) \approx \|\mathbf{D}\|_K, \quad \mathbf{D} = \left(D_{S_1}(\mathbf{y_1}), D_{S_1}(\mathbf{y_2}), \cdots, D_{S_n}(\mathbf{y_n})\right). \tag{8}$$

The power $P(\mathbf{X})$ in (9) is a scaled sum of $P_{S_i}(\mathbf{X})$ by using the design information (i.e., LDH and the number of gates) and activity factors, where W_i is the copies of S_i used in the canonical circuit construction. Both dynamic power and leakage power obtained from the SPICE simulation will be calculated

$$P(\mathbf{X}) = \sum_{i=1}^{n} W_i \cdot P_{S_i}(\mathbf{X}). \tag{9}$$

The optimization of PROCEED also includes the constraint of the area metric. The area of the gates is simulated using a quick layout estimator UCLADRE [42] or from an actual cell library. The detailed optimization objective function, derivation of $\mathbf{G}$ and $\mathbf{H}$, and definition of trust region around $\mathbf{X_0}$ are discussed in [26]. The tuning parameter vector $\mathbf{X}$ is optimized using gradient descent by $\mathbf{G}$ and $\mathbf{H}$ without manual intervention. After each iteration, only tuning parameters that can generate a Pareto point can be preserved, otherwise are abandoned.

In addition to the Pareto-based optimization, PROCEED has the following characteristics: it supports different activity factors, process variation settings for devices, power management modeling such as DVFS, and power gating. However, PROCEED has a very simple interconnect model, and it does not take device wear-out into account, which makes it not suitable for LT computing optimization. What is more, FinFETs are now widely employed in integrated circuit design, but PROCEED does not support FinFET-based circuit optimization. PROCEED only allows users to manually set the distribution of multi-V_{dd} and multi-V_{th} before the Pareto point searching. This necessitates that users determine which circuit bins should be applied by $V_{\mathrm{dd1}}(V_{\mathrm{th1}})$ and which by $V_{\mathrm{dd2}}(V_{\mathrm{th2}})$ to maximize the improvement of the performance. PROCEED forced the distribution boundary of V_{dd} and V_{th} to be the same and fixed it after beginning optimization.

B. Improvements of PROCEED-LT

PROCEED-LT is a calibrated LT SPICE model to evaluate and optimize circuit operations at 77 K. PROCEED-LT leveraged the advancements of PROCEED in circuit optimization with added support for FinFETs, an accurate interconnect model, and a device wear-out model. The device parameters used in the SPICE simulation and optimization are derived from a commercial-grade 14-nm FinFET low-voltage-threshold (LVT) cell library. Besides, the circuit netlist generation and simulation flow are optimized, and automated optimization for multi-V_{dd}/V_{th} is adopted, increasing the efficiency of the Pareto point searching. V_{th} of NFET and PFET can be separately optimized, and the distribution of multiple V_{dd} and V_{th} can be different. The core optimization algorithm and the upgrades of PROCEED-LT are written in MATLAB. Our code is available at `https://github.com/nanocad-lab/PROCEED-LT`. Upgrades in detail are described in the following.

1) Accurate Interconnect Model: PROCEED only comes with a very simple interconnect model, which includes the resistance per unit length and capacitance per unit length. However, this simple interconnect model is not sufficient for LT computing. The reduction of the temperature will significantly influence the conductivity of the interconnect. Therefore, it is critical to use an accurate interconnect model to evaluate the characteristics of LT computing. In PROCEED-LT, the detailed interconnect information is gathered from the physical synthesis result of the processor benchmark. For example, the interconnect length of each metal layer, the resistance per unit length, and the corresponding temperature coefficient of resistance (TCR) for each metal layer and vias are gathered. Then, the total resistance of the design can be calculated based on the information of wires and vias in each metal layer. Since the interconnect capacitance is hardly affected by temperature, the capacitance per unit length of each metal layer is gathered from the physical synthesis result as well. The resistance and capacitance values of each stage in PROCEED-LT are determined by the interconnect lengths, unit capacitance values, and unit resistance values of each metal layer mentioned above.

2) Device Wear-Out Model: PROCEED-LT considers HCI and BTI mentioned above as the aging effect, and we simulate them by using Cadence RelXpert [32]. Incorporated with temperature, time of use, and voltage conditions, RelXpert will estimate the V_{th} shift caused by the BTI effect. To simulate the

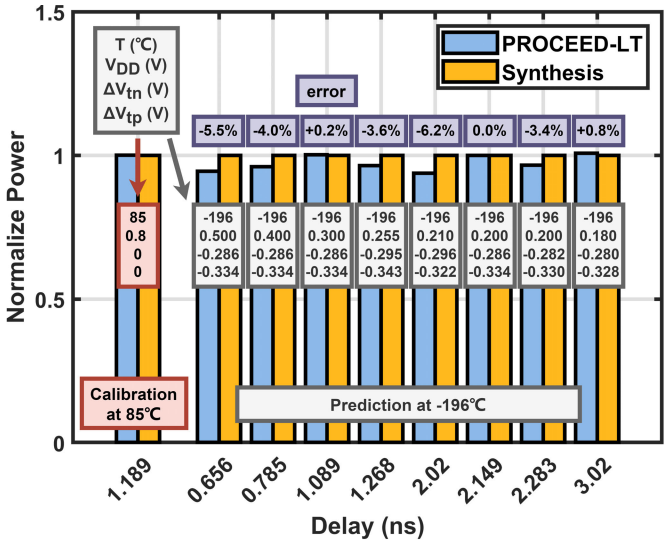


Fig. 3. Comparison of PROCEED-LT prediction and synthesis results at LT after the calibration at RT. The power is normalized by the synthesis power results.

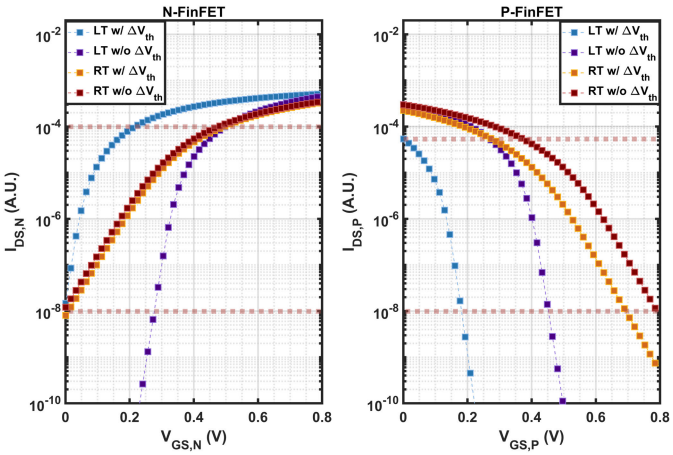


Fig. 4. I_{DS}–V_{GS} characteristics for N-FinFET and P-FinFET devices of the technology.

HCI effect, the RelXpert will simplify it as a current controlled current source to output a drain current (I_d) degradation, and PROCEED-LT then converts the I_d degradation into V_{th} shift. The V_{th} shifts of PFET and NFET from both HCI and BTI are combined. In aging evaluation, the V_{th} shifts will be applied to the worst case power and delay estimation based on the temperature and voltage configurations.

Besides, PROCEED-LT converts the effective width of the multigate FinFET device to the number of fins and, therefore, can evaluate the discrete width and optimize modern FinFET devices. In addition, PROCEED-LT adopts an automated multivoltage optimization strategy. Compared to PROCEED, PROCEED-LT allows a separate optimization for V_{th} of PFET and NFET. The distribution of multi-V_{dd} and multi-V_{th} is independent, and it is associated with the number of bins and the logic path depth distribution of the design, thereby reducing the manual intervention and increasing the efficiency of the optimal solution searching for various designs.

C. Calibration of PROCEED-LT

To validate our framework at LTs, we employed a commercial synthesis tool to assess the PD predictions generated by PROCEED-LT. Initially, we calibrate the PROCEED-LT model using a reference data point at 85 °C. Subsequently, we applied this calibrated model to predict the power and delay across various V_{dd} and V_{th} configurations at −196 °C. As depicted in Fig. 3, the PROCEED-LT model predicts power consumption with an error margin of less than 7%. This demonstrates that the predictions closely align with the synthesis results obtained from the RTL compiler under LT conditions. Generating libraries for diverse voltage configurations is extremely time-consuming, often extending over several days, and the synthesis of each configuration costs tens of minutes. In contrast, PROCEED-LT can evaluate each voltage configuration in less than 1 min and optimize a PD curve within a few hours. The feasible time required to generate a PD curve allows us to comprehensively assess the characteristics of LT circuit operations over a wide range of performance.

IV. Experiment Results

To evaluate the advantage of LT computing and how different temperature-dependent factors affect the circuit-level optimization for LT computing, we present the experiment results generated by PROCEED-LT. We evaluate two benchmarks, a microprocessor design ARM Cortex-M3, and a neural network accelerator ACOUSTIC [21]. The device model is from a hardware-calibrated 14-nm FinFET technology. A typical activity factor of 0.1 is set as the default in our benchmark given that real digital systems often have idle components [43]. We use 10 bins in the optimization. We compare the PD curves of the processor at 85 °C and at 196 °C under different circumstances. We show the importance of using accurate models to comprehensively evaluate the benefits of LT computing.

A. Device Model

This work utilizes a commercial 14-nm FinFET technology, with models specifically designed and calibrated for operations at 77 K. Fig. 4 presents the I_{DS}–V_{GS} characteristics for both N-FinFET and P-FinFET devices. These devices have been optimized for LT operation through a V_{th} shift, facilitating low-V_{dd} operation while maintaining essential device characteristics within the targeted low-V_{dd} range. This ensures consistent and reliable device performance under LT conditions.

B. Impact of Accurate Interconnect Model

An accurate interconnect model can better show the advantage of LT computing. In the accurate interconnect model, we calculate the metal wire resistance per unit length or via resistance at LT as follows:

$$R_{w/v} = R_{0,w/v} \times \left[1 + \mathrm{TCR}_{w/v} \times (T_l - T_r)\right] \quad (10)$$

TABLE I

REDUCTION IN THE METAL RESISTANCE OF PER UNIT LENGTH AND VIA RESISTANCE OF CORTEX M3 AT −196 °C COMPARED TO 85 °C

Metal Level	ΔR_w	Via Level	ΔR_v
M1	43.6%	V1	14.6%
M2	43.6%	V2	16.3%
M3	43.6%	V3	15.5%
M4	51.8%	V4	16.3%
M5	50.0%	V5	16.3%
M6	57.9%	V6	17.3%

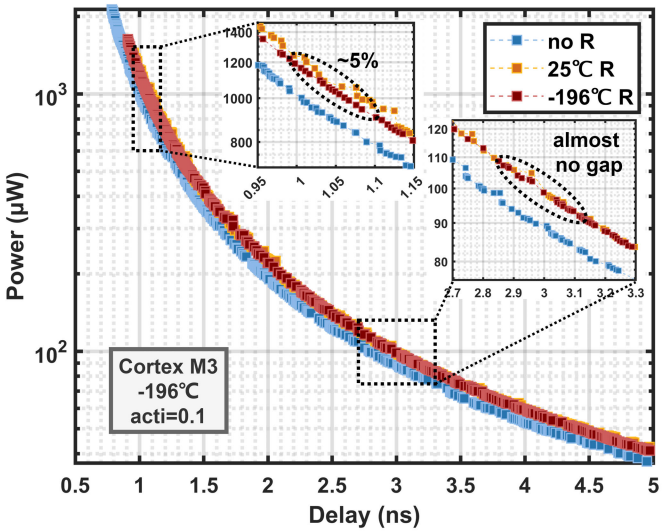


Fig. 5. PD curves for Cortex M3 at −196 °C, using 1vt,1vd configuration, along with aging effects, and various interconnect models: (a) no R: excluding interconnect resistance, (b) 25 °C R: original PROCEED interconnect model, and (c) −196 °C R: accurate interconnect model.

where $R_{0,w/v}$ denotes the target metal resistance per unit length or via resistance at RT, $\text{TCR}_{w/v}$ denotes the TCR of the corresponding metal level or via level, and T_l and T_r denote the LT and the reference temperature (25 °C), respectively. Then, the average resistance per unit length of the interconnect can be given by the weighted sum of wire and via resistances as follows:

$$R_{\text{avg}} = \frac{\sum_{i=0}^{M} WL_i \cdot R_{w,i} + \sum_{j=0}^{N} V_j \cdot R_{v,j}}{WL_{\text{tot}}} \tag{11}$$

where WL_i denotes the wire length of the ith metal level, V_j denotes the number of vias of the jth via level, and WL_{tot} denotes the total wire length of the design. For example, the reduction in the metal resistance per unit length and via resistance of different levels of Cortex M3 at LT and RT is shown in Table I. Due to the reduction in metal and via resistances as temperature decreases, the interconnect resistance at −196 °C decreases by 32% compared to its value at 85 °C.

Fig. 5 shows the PD curves of the benchmark under −196 °C. The three curves are given as follows: without interconnect resistance, with a simple interconnect model in PROCEED, and with an accurate interconnect model in PROCEED-LT. The optimized PD curves of the Cortex M3 using single V_{th} (1vt) and V_{dd} (1vd) with the aging effect and an activity factor of 0.1 is demonstrated. At −196 °C, utilizing

TABLE II

POWER COMPARISON OF CORTEX M3 BETWEEN −196 °C AND 85 °C USING DIFFERENT INTERCONNECT MODELS WITH THE ACTIVITY FACTOR OF 0.1 AT 1.5-NS CIRCUIT LATENCY

Interconnect Model	Power(mW) and improvement		
	85°C	-196°C	*Improvement*
no R	4.748	0.364	13.04X
PROCEED	5.140	0.422	12.18X
PROCEED-LT	5.242	0.418	12.54X

the PROCEED-LT interconnect model registers approximately 5% less power consumption at the high-performance range than the original PROCEED interconnect model, which bases its calculations only on RT parameters without accounting for the precise resistance of different metal layers and vias. At lower performance, it is observed that the accurate interconnect model and the simple one almost overlap, indicating that the influence of interconnect resistance discrepancies on performance is diminished. In addition, as illustrated in Table II, power consumption at 1.5 ns at 85 °C using the PROCEED-LT interconnect model is about 2% higher than that of the original PROCEED model. The power improvement between −196 °C and 85 °C at 1.5 ns can vary from 12.18× to 12.54× depending on the interconnect model used. Most wires in our evaluated benchmark are in metal layers with lower temperature coefficients of resistance (TCR) in this technology. In large-scale designs, higher metal layers, which generally have higher TCR, will be used. Hence, the influence of temperature on interconnect resistance will be more pronounced, which can be demonstrated by employing an accurate interconnect model. The minor discrepancies introduced by different interconnect models suggest that the benefits of LT in signal delay are less impacted by interconnect resistance. However, the reduced interconnect resistance at LT plays a more significant role in the benefits of improved IR drop, as will be demonstrated in Section IV.

C. Impact of the Aging Model

As we discussed in Section II-C, the processor can potentially benefit from LT conditions when taking into account the impact of aging. In this section, we show how LT conditions can mitigate the aging effect of a processor. To the best of our knowledge, the tool is modeling the aging-induced ΔV_{th} from BTI and HCI in a similar way to that described in [34], [35], and [36], such as (1) and (2). We suppose that the devices are running for ten years and use Cadence RelXpert [32] to interpret the influence from BTI and HCI as the V_{th} shifts and the I_d degradation for a wide range of V_{dd} and V_{th} configurations. The relationship between the transistor's drain current and the threshold voltage in the saturation region allows us to correlate the degradation in I_d to shifts in V_{th}. Finally, we combine the V_{th} shifts of NFET and PFET from both BTI and HCI as the aging influence of ten years under 85 °C and −196 °C. Fig. 6 illustrates the aging-induced ΔV_{th} under various stress voltages at RT and LT. It is clear from the figure that the ΔV_{th} of both nMOS and pMOS devices at LT remains minimal across a wide range of stress voltages. In contrast, at RT, ΔV_{th} increases significantly,

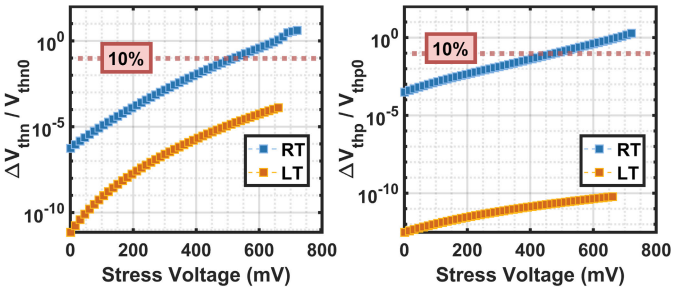


Fig. 6. Ten-year aging-induced ΔV_{th} of nMOS and pMOS devices under various stress voltages at 85 °C and −196 °C.

exceeding 10% at higher stress voltages. When evaluating the characteristics of the design, we obtain the V_{th} shifts of NFET and PFET based on the V_{dd} and nominal V_{th} conditions and consider the worst case, i.e., applying the V_{th} shifts in the delay calculation while ignoring them in the power calculation. In this way, we can assess the end-of-life reliability of the circuit at different temperatures.

Fig. 7 shows the PD curves of Cortex M3 operating at 85 °C and −196 °C with and without the aging effect. The results show that the power increases by 5%∼10% under the aging model for equivalent performance at 85 °C when the operating latency is less than 1.3 ns. After considering the aging effect, the benefit of LT is more pronounced, improving from 10.23× to 11.94× at a latency of 1.5 ns. Notably, in the range of high performance in Fig. 7(b), it is observed that the power at RT will rapidly increase compared to the curve without aging. In contrast, the power discrepancy attributable to aging effects at LT is merely 1% at 1ns. The difference in aging effects at RT and LT is attributed to the fact that the aging effect is not only influenced by temperature but also $(V_{dd} - V_{th})$. To achieve high performance, a larger $(V_{dd} - V_{th})$ is required to increase the transistor switching speeds, and it consequently exacerbates the influence of HCI. Hence, the aging effect is more severe with a higher V_{dd} value, and the performance of the processor running at higher frequencies is affected. Conversely, at −196 °C, it is possible to substantially reduce the voltage stress without compromising performance, thereby alleviating the impact of HCI. Furthermore, as shown in Fig. 7(b), the aging effect limits the circuit at 85 °C from achieving higher performance (<1.1 ns), while the limitation at −196 °C will become obvious only when the operating frequency is extremely high (<0.7 ns). We observe that the PD curves with and without the aging effect at −196 °C show great convergence until the latency is less than 0.65 ns, which suggests that performance degradation over time is minimized at LT, reinforcing the notion that LT computing could be more advantageous for applications that demand high performance coupled with long-term stability.

After incorporating the accurate interconnect resistance model and an aging model, the power of processor logic operating at −196 °C exhibits an improvement ranging from around 11× to more than 17× for equivalent performance compared to its 85 °C counterpart. The 1vt,1vd configuration and the accurate interconnect model are used in this experiment. Subsequent sections will employ the same configurations unless specified otherwise. The activity factor that

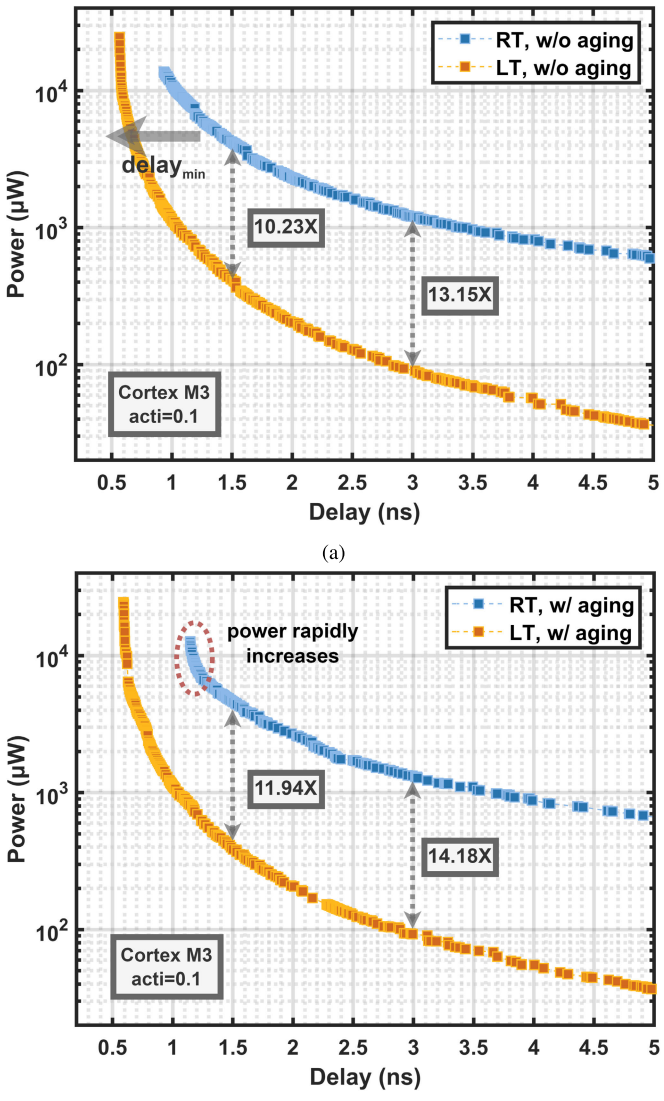


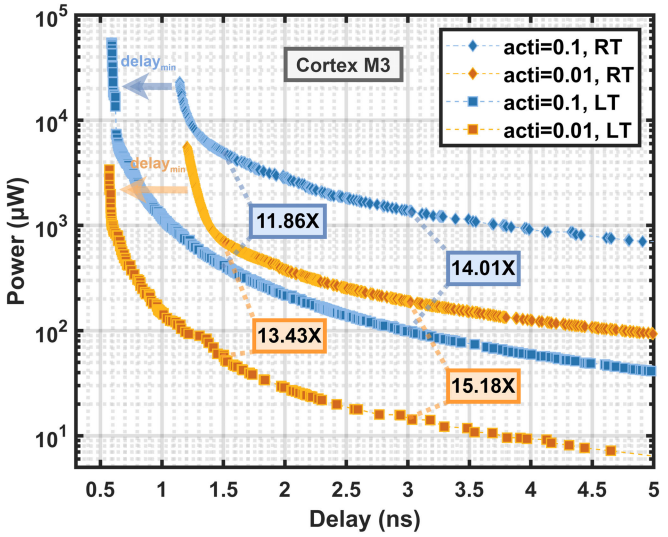
Fig. 7. PD curves of Cortex M3 at 85 °C and −196 °C with and without the aging effect using 1vt,1vd configuration along with an accurate interconnect model. (a) Without aging effects. (b) With aging effects.

we use is 0.1 so far, and forthcoming discussions will explore the influence of the activity factor on LT optimization further.

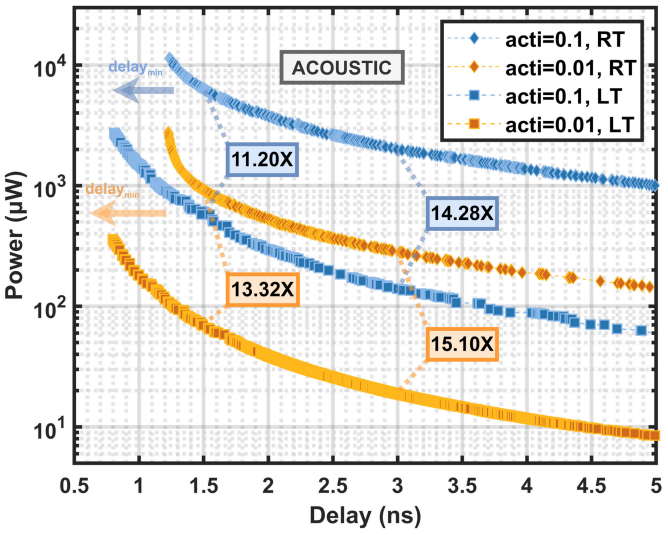
D. Impact of Activity Factors in LT Operations

This section delves into the impact that varying activity factors have on the optimization of LT computing. At LT, the leakage power of the processor is vastly reduced due to a considerable decrease in leakage current, resulting in the processor's total power being predominantly dynamic. While at RT, the processor consumes more percentage of leakage power with a lower activity factor since the activity factor only affects dynamic power and the leakage power stays the same. It is revealed that the processor optimization at LT computing benefits more from a low activity factor.

Fig. 8 depicts the PD curves of the Cortex M3 and ACOUSTIC, showcasing performance with activity factors 0.1 and 0.01 at 85 °C , as well as −196 °C. LT conditions are observed to substantially elevate peak performance across multiple activity factors, more than doubling it with comparable power

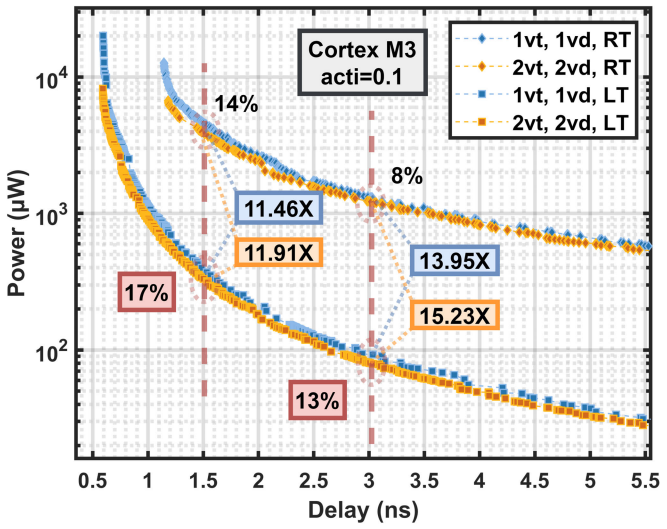


(a)

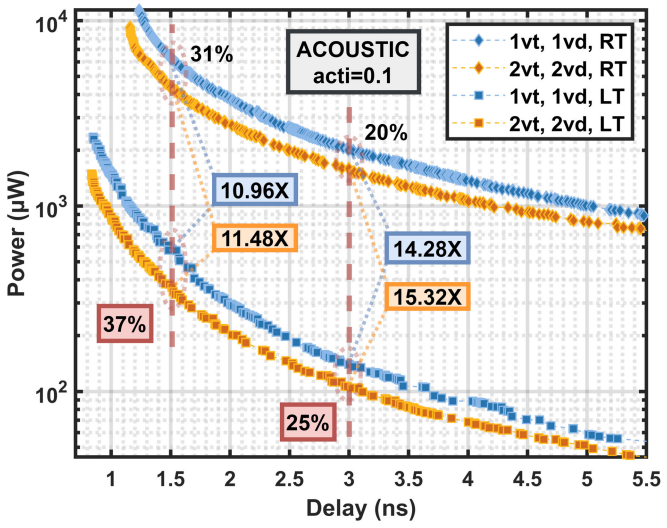


(b)

Fig. 8. PD curves of benchmarks using 1vt,1vd at 85 °C and −196 °C with different activity factors. (a) Cortex M3. (b) ACOUSTIC.



(a)



(b)

Fig. 9. PD curves of 1vt,1vd and 2vt,2vd configurations under RT and LT, along with aging effects and the accurate interconnect model. (a) Cortex M3. (b) ACOUSTIC.

usage, facilitating low delays under 1 ns. As the activity factor decreases to 0.01, Cortex M3 achieves a 13.43× improvement of power efficiency at LT rather than 11.86× with an activity factor of 0.1 at 1.5 ns. Similarly, the phenomenon can be observed in the PD curves of ACOUSTIC. Table III shows the power of various activity factors at 85 °C and −196 °C with 1.5-ns latency for Cortex M3 and ACOUSTIC. The figures and the table show that the power at −196 °C is almost proportional to the activity factor. Table III(b) shows the improvement of power efficiency of ACOUSTIC from 85 °C to −196 °C at 1.5-ns latency, noting an improvement ranging from 7.97× to a striking 21.88× as the activity factor is lowered from 0.5 to 0.001. These results show that LT computing processors benefit more from lower activity factors compared to standard RT operations.

E. Benefit of Multi-V_{dd} and Multi-V_{th} at LT

The deployment of multi-$V_{\rm dd}$ (2vd) and multi-$V_{\rm th}$ (2vt) strategies has been revealed to offer substantial benefits [26]. The 2vt2vd approach can adaptively allocate appropriate $V_{\rm dd}$ and $V_{\rm th}$ values to specific segments of a circuit, particularly those that predominantly influence power consumption or latency. This targeted allocation not only optimizes power savings but also preserves the overall performance of the chip. The actual benefit of using 2vt2vd is closely associated with the distribution of the logic path depth of the design.

We adopt 2vt2vd configurations, as described in Section III-A. As presented in Fig. 9(a) and (b), we compare PD curves of 2vt2vd and 1vt1vd at 85 °C and −196 °C for two benchmarks, with the aging effect considered. The adoption of 2vt2vd configurations under LT conditions notably diminishes power consumption, showing an additional 5% in comparison to that at RT for both two benchmarks. Take the PD curves of ACOUSTIC as an example. 2vt2vd can achieve 11.48× power efficiency improvement, while 1vt1vd can only gain 10.96× as the temperature decreases from RT to LT at 1.5 ns. As the temperature decreases, the leakage power is substantially reduced. PROCEED-LT adjusts the lower bound

TABLE III
TOTAL POWER COMPARISON OF VARIOUS ACTIVITY FACTORS BETWEEN −196 °C AND 85 °C AT 1.5-NS CIRCUIT LATENCY. (A) CORTEX M3. (B) ACOUSTIC

(a)

Activity factor	Power (mW) and improvement		
	85°C	-196°C	*Improvement*
0.2	10.164	0.803	12.66X
0.1	4.957	0.418	11.86X
0.05	3.032	0.231	13.12X
0.01	0.685	0.051	13.43X
0.001	0.132	0.007	18.86X

(b)

Activity factor	Power (mW) and improvement		
	85°C	-196°C	*Improvement*
0.5	24.532	3.076	7.97X
0.3	16.615	1.531	10.85X
0.1	6.367	0.578	11.02X
0.01	0.920	0.069	13.25X
0.001	0.189	0.009	21.88X

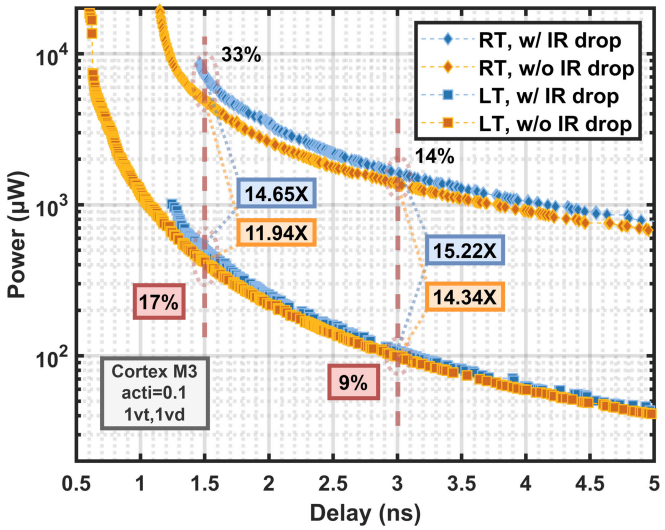


Fig. 10. PD curves of Cortex M3 using the accurate interconnect model, 1vt,1vd configuration, and an activity factor of 0.1, considering aging effects, with and without IR drop at 85 °C and −196 °C.

of the V_{th} set lower during the optimization adaptively, which expands the search space for solutions because a broader array of V_{th} and V_{dd} combinations becomes feasible with 2vt2vd. Such an expanded search space potentially renders the Pareto optimization more effective, optimizing the employment of the 2vt2vd strategy.

Furthermore, it is noted that the benefits realized from the 2vt2vd configuration vary between benchmarks. In the case of ACOUSTIC, the 2vt2vd setup can achieve 37% power improvement compared to 1vt1vd at the equivalent performance (1.5 ns), whereas the Cortex M3 demonstrates a more modest improvement of 17%. The reason for this discrepancy is that in the ACOUSTIC accelerator, the majority circuit that dominates the power of the whole design does not have long path depth, hence not dominating the delay. Conversely, the circuit paths influencing delay, characterized by greater depths, constitute a smaller fraction of the total circuits. This discrepancy allows PROCEED-LT to negotiate the balance between total power and performance more effectively. On the contrary, as depicted in Fig. 2(a), the Cortex M3 features a relatively even distribution of logic paths across varying depths. Therefore, the advantage of using 2vt2vd configuration is shrunk. In summary, the 2vt2vd strategy is effective for power reduction without sacrificing performance, and at cryogenic temperatures, the benefits of using 2vt2vd can be more evident.

F. Potential Improvement of IR Drop at LT

The static IR drop, which is usually ignored in the analysis of cryogenic computing, refers to the voltage drop that occurs when current flows through a resistive path in the circuit [44]. In digital circuits, an excessive IR drop can lead to an insufficient voltage supply to certain parts of the chip, causing timing issues and suboptimal thermal behavior. As the interconnect resistance decreases at LT, the voltage drop on the power network can potentially be mitigated [45].

We compare the PD curves of Cortex M3 using the 1vt1vd configuration with and without IR drop at 85 °C and −196 °C in Fig. 10. Initially, we assume a 10% of supply voltage drop at 85 °C and compute the effective resistance that causes the corresponding IR drop. Then, considering the TCR, the effective resistance value at −196 °C is calculated, and the IR drop at LT of each Pareto point can be evaluated during the optimization. The current used to calculate the IR drop is based on the average current when the circuit operates with an activity factor of 0.1. We apply the IR drop in the delay calculation and exclude it from the power calculation, representing a worst case scenario for chip operation. It is shown that at lower frequencies of 3∼4 ns, the impact of IR drop at LT is 5% less on average than that of RT. The benefit of LT computing is more apparent at higher performance. To achieve high performance at RT, a high $(V_{dd} - V_{th})$ is required, resulting in a rapid increase in the operating current. In contrast, at LT, the required $(V_{dd} - V_{th})$ to achieve a comparable performance is lower, which can potentially reduce the current. The effect of IR drop is more pronounced at 85 °C, with a 33% increase in power at lower delays, almost 2× of the 17% increase observed at 196 °C for equivalent performance (1.5 ns). The power efficiency improvement due to the temperature reduction increases from 11.94× to 14.65× at 1.5 ns after considering the IR drop. Overall, with the same circuit architecture, LT computing demonstrates greater resilience to the impact of IR drop compared to RT.

G. Impact of V_{th} Variation Under LT

Attention should be paid to the V_{th} variation of the devices, which can result from a variety of process factors in manufacturing. We utilize Monte Carlo simulations in Cadence to get the 3σ of V_{th} shift of the device. Similarly, we apply $+3\sigma$ ΔV_{th} in delay calculation and -3σ ΔV_{th} in power calculation to assess the worst case scenario.

The impact of V_{th} variation on PD curves of Cortex M3 at RT and LT can be seen in Fig. 11. When applying the same 3σ V_{th} variation at LT (variation), the power increases to almost 2× at the same performance compared to the no variation

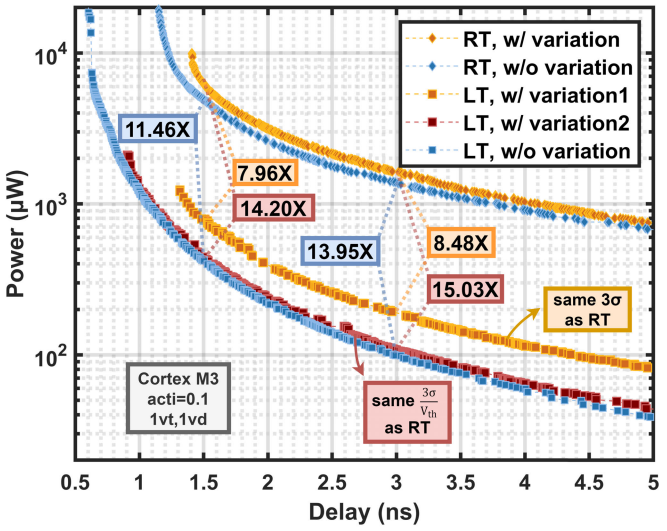


Fig. 11. PD curves of Cortex M3 using the accurate interconnect model, 1vt,1vd configuration, and an activity factor of 0.1, considering aging effects, with and without V_{th} variation under RT and LT. Variation1: the same 3σ as RT in V_{th} variation; variation2: the same $\frac{3\sigma}{V_{th}}$ as RT in V_{th} variation; and w/o variation: excluding V_{th} variation influence.

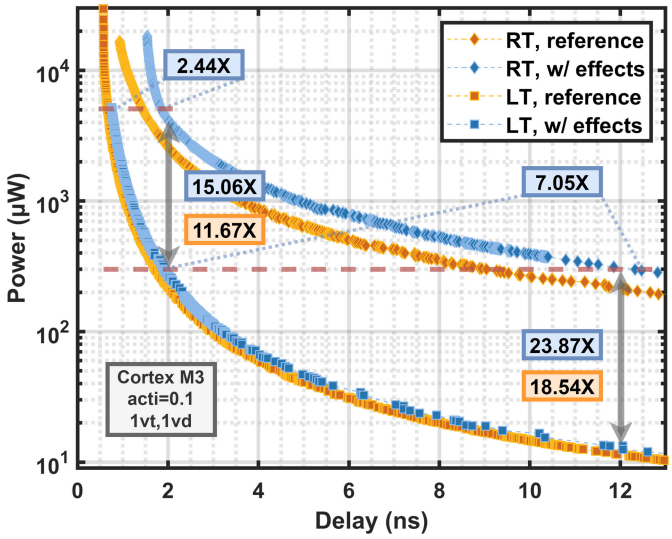


Fig. 12. PD curves of Cortex M3 using the accurate interconnect model and 1vt,1vd configuration at 85 °C and −196 °C. Reference: without considering any other effect; W/ effects: considering effects of aging, IR drop, and V_{th} variation.

scenario, whereas, at RT, the power gap is only around 17%. The comparable 3σ ΔV_{th} values have a tangible impact on the PD characteristics, with more significant effects at cryogenic temperatures. The reason is that at LT, V_{dd} and V_{th} are both scaled down while maintaining the same performance at RT, which amplifies the sensitivity of circuit characteristics to V_{th} changes. Hence, a comparable V_{th} variation imposes a greater impact at LT. In Fig. 11, if the V_{th} variation at LT maintains an identical $\frac{3\sigma}{V_{th}}$ value as RT's (variation2), the PD curves can have a power penalty (around 10%) with no variation case. This observation highlights a significant challenge for LT computing circuit designers: the magnified impact of process variation at cryogenic temperatures. To mitigate the effect of V_{th} variation at LT that is equivalent to the impact observed at RT with the current process, it is necessary to improve the current process to reduce the 3σ ΔV_{th} variation as the V_{dd} and nominal V_{th} are scaled down. This adjustment is crucial for ensuring that LT circuits can leverage the benefits of lower temperatures without disproportionately suffering from the adverse effects of V_{th} variations.

H. Peak Performance Improvement at LT

High peak performance is essential for applications demanding real-time processing and high-speed data analysis. LT computing can achieve higher peak performances even by considering multiple effects, such as the aging, the IR drop, and the V_{th} variation.

In Fig. 12, we compare the performance at RT and LT under equivalent power limitations. We apply an identical $\frac{3\sigma}{V_{th}}$ value as RT to assess the V_{th} variation at LT. A high operational voltage stress is necessary to satisfy the high-performance requirement at RT, exacerbating the aging and the IR drop effects and leading to a sharp increase in power consumption. In contrast, at LT, these factors, which often degrade performance at higher temperatures, are more effectively mitigated. Under equivalent power constraints in the reference case, for example, 5 mW, the improvement in performance at LT compared to RT can reach 2.15×. This advantage becomes more pronounced when multiple effects are considered, achieving a 2.44× improvement. At the same performance, LT operations exhibit power reduction ranging from 15× to over 23× with these effects. Furthermore, as the power limitation decreases, the performance gap between LT and RT enlarges. Notably, the achievable peak performance at LT increases to 1.65× that of RT, while this improvement rises to 2.04× with a 3.51× power improvement when considering these effects similarly. In addition, Table IV presents a comparison of the Cortex M3 energy breakdown results between LT and RT, taking into account the effects of aging, IR drop, and V_{th} variation. PROCEED-LT optimizes the total power of the circuit, rather than focusing solely on either dynamic energy or leakage energy. Both the dynamic energy and leakage energy at 2.0-ns latency under LT conditions are significantly reduced by over 13× compared to their levels under RT conditions. In conclusion, the experiment results demonstrate that the circuit operations under LT conditions are more resilient against various performance-degrading factors.

TABLE IV

POWER BREAKDOWN COMPARISON OF CORTEX M3 AT 2.0-NS LATENCY WITH AN ACTIVITY FACTOR OF 0.1 BETWEEN −196 °C AND 85 °C, CONSIDERING EFFECTS OF AGING, IR DROP, AND V_{TH} VARIATION

Category	Power (mW) and improvement		
	85°C	-196°C	*Improvement*
Dynamic	3.772 (92.50%)	0.245 (91.76%)	15.39X
Leakage	0.306 (7.50%)	0.022 (8.24%)	13.90X
Total Power	4.078	0.267	15.05X

V. CONCLUSION

In summary, in order to reveal the circuit-level advantages and challenges of LT computing a wide power-performance range, this work develops PROCEED-LT, a Pareto-based device-circuit co-optimization tool for LT computing, adding support for temperature-dependent interconnect

model, an aging model, and FinFET devices and improves the optimization of multi-V_{dd} or multi-V_{th} configurations. LT computing promises lower interconnect resistance, slowed down device wear-out, low leakage, and high drive current at lower voltages. For high-performance regimes, LT delivers 10× power improvement compared to RT, which improves further to nearly 12× when reduced aging is accounted for. When improved IR drop due to improved interconnect resistance and reduced currents are taken into account, the LT-RT energy difference increases to 14×. Low leakage in LT operation for low-activity factor scenarios further improves its energy benefit over RT to as much as 21×. Furthermore, achievable peak performance in LT can improve by over 2× compared to RT. Our results further indicate that control of V_{th} variation to the same percentage levels as RT is critical to preserve LT benefits.

This work has primarily focused on device-circuit co-optimization for cryogenic operation. A similar effort is needed to optimize the interconnect stack for LT. For example, failures due to electromigration in interconnects will improve significantly at LTs [14]. Improved reliability and resistance warrant reengineering the interconnect stack for LT operation along with power and clock distribution strategies.

ACKNOWLEDGMENT

Any opinions, findings, or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

REFERENCES

[1] A. Mehonic and A. J. Kenyon, "Brain-inspired computing needs a master plan," *Nature*, vol. 604, no. 7905, pp. 255–260, Apr. 2022, doi: 10.1038/s41586-021-04362-w.

[2] N. Romli, K. Minhad, M. Reaz, and M. Amin, "An overview of power dissipation and control techniques in CMOs technology," *J. Eng. Sci. Technol.*, vol. 10, no. 3, pp. 364–382, Mar. 2015.

[3] H. L. Chiang et al., "Cold CMOS as a power-performance-reliability booster for advanced FinFETs," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265065.

[4] S. Datta, W. Chakraborty, and M. Radosavljevic, "Toward attojoule switching energy in logic transistors," *Science*, vol. 378, no. 6621, pp. 733–740, Nov. 2022, doi: 10.1126/science.ade7656.

[5] R. A. Matula, "Electrical resistivity of copper, gold, palladium, and silver," *J. Phys. Chem. Reference Data*, vol. 8, no. 4, pp. 1147–1298, Oct. 1979, doi: 10.1063/1.555614.

[6] I. Byun, D. Min, G.-H. Lee, S. Na, and J. Kim, "CryoCore: A fast and dense processor architecture for cryogenic computing," in *Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA)*, May 2020, pp. 335–348, doi: 10.1109/ISCA45697.2020.00037.

[7] E. Garzon, R. De Rose, F. Crupi, A. Teman, and M. Lanuzza, "Exploiting STT-MRAMs for cryogenic non-volatile cache applications," *IEEE Trans. Nanotechnol.*, vol. 20, pp. 123–128, 2021, doi: 10.1109/TNANO.2021.3049694.

[8] E. Garzón, L. Yavits, A. Teman, and M. Lanuzza, "STT-MRAM technology for energy-efficient cryogenic memory applications," in *Proc. IEEE 14th Latin Amer. Symp. Circuits Syst. (LASCAS)*, Feb. 2023, pp. 1–4, doi: 10.1109/lascas56464.2023.10108316.

[9] D. Min, I. Byun, G.-H. Lee, S. Na, and J. Kim, "CryoCache: A fast, large, and cost-effective cache architecture for cryogenic computing," in *Proc. Twenty-Fifth Int. Conf. Architectural Support for Program. Lang. Operating Syst.*. Switzerland: ACM, Mar. 2020, pp. 449–464, doi: 10.1145/3373376.3378513.

[10] Y. Shu, H. Zhang, H. Sun, Q. Deng, and Y. Ha, "CSDB-eDRAM: A 16Kb energy-efficient 4T CSDB gain cell eDRAM with over 16.6s retention time and 49.23uW/Kb at 4.2K for cryogenic computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2023, pp. 1–5, doi: 10.1109/iscas46773.2023.10181628.

[11] D. S. Kang and S. Yu, "Design-technology co-optimization for cryogenic tensor processing unit," in *Proc. IEEE Asia–Pacific Conf. Circuits Syst. (APCCAS)*, Nov. 2022, pp. 1–4, doi: 10.1109/APCCAS55924.2022.10090326.

[12] W. Chakraborty et al., "Pseudo-static 1T capacitorless DRAM using 22 nm FDSOI for cryogenic cache memory," in *IEDM Tech. Dig.*, Dec. 2021, pp. 40.1.1–40.1.4, doi: 10.1109/IEDM19574.2021.9720578.

[13] W. Chakraborty et al., "Multi-bit per-cell 1T SiGe floating body RAM for cache memory in cryogenic computing," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Jun. 2022, pp. 302–303, doi: 10.1109/VLSITechnologyandCir46769.2022.9830483.

[14] R. Saligram, S. Datta, and A. Raychowdhury, "Design space exploration of interconnect materials for cryogenic operation: Electrical and thermal analyses," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 11, pp. 4610–4618, Nov. 2022, doi: 10.1109/TCSI.2022.3195636.

[15] A. Beckers, F. Jazaeri, and C. Enz, "Theoretical limit of low temperature subthreshold swing in field-effect transistors," *IEEE Electron Device Lett.*, vol. 41, no. 2, pp. 276–279, Feb. 2020, doi: 10.1109/LED.2019.2963379.

[16] W.-C. Wang et al., "Cool-CIM: Cryogenic operation of analog compute-in-memory for improved power-efficiency," in *IEDM Tech. Dig.*, Dec. 2023, pp. 1–4, doi: 10.1109/iedm45741.2023.10413880.

[17] Y. Shu, H. Zhang, Q. Deng, H. Sun, and Y. Ha, "CIMC: A 603TOPS/W in-memory-computing C3T macro with boolean/convolutional operation for cryogenic computing," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Mar. 2023, pp. 1–2, doi: 10.1109/cicc57935.2023.10121295.

[18] R. Saligram, D. Prasad, D. Pietromonaco, A. Raychowdhury, and B. Cline, "A 64-bit arm CPU at cryogenic temperatures: Design technology co-optimization for power and performance," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–2, doi: 10.1109/CICC51472.2021.9431559.

[19] R. Saligram, W. Chakraborty, N. Cao, Y. Cao, S. Datta, and A. Raychowdhury, "Power performance analysis of digital standard cells for 28 nm bulk CMOS at cryogenic temperature using BSIM models," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 7, pp. 193–200, 2021, doi: 10.1109/JXCDC.2021.3131100.

[20] P. Wang, X. Peng, W. Chakraborty, A. Khan, S. Datta, and S. Yu, "Cryogenic performance for compute-in-memory based deep neural network accelerator," in *Proc. 2021 IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2021, pp. 1–4, doi: 10.1109/ISCAS51556.2021.9401756.

[21] W. Romaszkan, T. Li, T. Melton, S. Pamarti, and P. Gupta, "ACOUSTIC: Accelerating convolutional neural networks through or-unipolar skipped stochastic computing," in *Proc. Des. Automat. Test Europe Conf. Exhib. (DATE)*, 2020, pp. 768–773, doi: 10.23919/DATE48585.2020.9116289.

[22] J. P. G. van Dijk et al., "Cryo-CMOS for analog/mixed-signal circuits and systems," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Mar. 2020, pp. 1–8, doi: 10.1109/CICC48029.2020.9075882.

[23] H. Bohuslavskyi et al., "Cryogenic characterization of 28-nm FD-SOI ring oscillators with energy efficiency optimization," *IEEE Trans. Electron Devices*, vol. 65, no. 9, pp. 3682–3688, Sep. 2018, doi: 10.1109/TED.2018.2859636.

[24] D. Kang and S. Yu, "Cryo-CMOS design-technology co-optimization of low noise amplifier for silicon qubit readout," *Microelectronic Eng.*, vol. 262, Jun. 2022, Art. no. 111837, doi: 10.1016/j.mee.2022.111837.

[25] D. Fitrio, A. Stojcevski, and J. Singh, "Subthreshold leakage current reduction techniques for static random access memory," in *Proc. SPIE*, Feb. 2005, pp. 673–683, doi: 10.1117/12.582332.

[26] S. Wang, A. Pan, C. O. Chui, and P. Gupta, "PROCEED: A Pareto optimization-based circuit-level evaluator for emerging devices," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 1, pp. 192–205, Jan. 2016, doi: 10.1109/TVLSI.2015.2393852.

[27] Z. Yan, "HotLeakage: A temperature-aware model of subthreshold and gate leakage for architects," Dept. Comput. Sci., Univ. Virginia, Charlottesville, VA, USA, Tech. Rep. CS-2003-05, 2002.

[28] S. S. Sapatnekar, "What happens when circuits grow old: Aging issues in CMOS design," in *Proc. Int. Symp. onVLSI Design, Autom., Test (VLSI-DAT)*, Apr. 2013, pp. 1–2, doi: 10.1109/VLDI-DAT.2013.6533827.

[29] M. B. Yelten, "Holistic device modeling: Toward a unified MOSFET model including variability, aging, and extreme operating conditions," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 9, pp. 2635–2640, May 2022, doi: 10.1109/TCSII.2022.3171136.

[30] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001, doi: 10.1109/5.920580.

[31] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "FinCACTI: Architectural analysis and modeling of caches with deeply-scaled FinFET devices," in *Proc. Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, 2014, pp. 290–295, doi: 10.1109/ISVLSI.2014.94.

[32] *Virtuoso® RelXpert Model Reference Guide*, Release V1.0_3.0, Cadence Des. Syst., San Jose, CA, USA, Nov. 2019.
[33] S. Kasap, *Principles of Electronic Materials and Devices*, 4th ed., New York, NY, USA: McGraw-Hill, 2017.
[34] H. Amrouch et al., "Reliability challenges with self-heating and aging in FinFET technology," in *Proc. IEEE 25th Int. Symp. On-Line Test. Robust Syst. Design (IOLTS)*, 2019, pp. 68–71.
[35] H. Jiang, "The impact of self-heating on HCI reliability in high-performance digital circuits," *IEEE Electron Device Lett.*, vol. 38, no. 4, pp. 430–433, Apr. 2017.
[36] P. Srinivasan and T. Nigam, "Critical discussion on temperature dependence of BTI in planar and FinFET devices," in *Proc. IEEE Electron Devices Technol. Manuf. Conf. (EDTM)*, Jan. 2017, pp. 33–35, doi: 10.1109/edtm.2017.7947497.
[37] C.-Y. Chen et al., "Negative bias temperature instability in low-temperature polycrystalline silicon thin-film transistors," *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 2993–3000, Dec. 2006, doi: 10.1109/TED.2006.885543.
[38] A. Grill et al., "Reliability and variability of advanced CMOS devices at cryogenic temperatures," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2020, pp. 1–6, doi: 10.1109/irps45951.2020.9128316.
[39] Y. Zhang, J. Xu, T.-T. Lu, Y. Zhang, C. Luo, and G. Guo, "Hot carrier degradation in MOSFETs at cryogenic temperatures down to 4.2 K," *IEEE Trans. Device. Mat. Reliab*, vol. 21, no. 4, pp. 620–626, Dec. 2021, doi: 10.1109/TDMR.2021.3124417.
[40] W. Chakraborty, U. Sharma, S. Datta, and S. Mahapatra, "Hot carrier degradation in cryo-CMOS," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–5, doi: 10.1109/IRPS45951.2020.9129312.
[41] V. Sriramkumar, J. Duarte, C. Hu, and et. al., "BSIM-CMG 107.0.0 multi-gate MOSFET compact model technical manual," BSIM Group Berkeley Univ., California, BA, USA, Tech. Rep., 2013. [Online]. Available: http://www.wrcad.com/xictools/docs/model_docs/bsimcmg-1.0.7/BSIMCMG107.0.0_TechnicalManual_20130712.pdf
[42] R. S. Ghaida and P. Gupta, "DRE: A framework for early co-evaluation of design rules, technology choices, and layout methodologies," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 9, pp. 1379–1392, Sep. 2012, doi: 10.1109/TCAD.2012.2192477.
[43] S. Harris and D. Harris, "1 - from zero to one," in *Digital Design And Computer Architecture*. Cambridge, MA, USA: Morgan Kaufmann, 2022, pp. 1–50, doi: 10.1016/B978-0-12-820064-3.00001-5.
[44] C. Halford and B. Geden, "IR-drop analysis," Adv. Layout Solutions, Aldermaston, U.K., Tech. Rep., 2009. [Online]. Available: http://pdf2.solecsy.com/558/5faac10b-e1d3-4218-bf90-8658d42f62a3.pdf
[45] S. Bhattacharya, D. Das, and H. Rahaman, "Temperature dependent IR-drop and delay analysis in side-contact multilayer graphene nanoribbon based power interconnects," in *Proc. 20th Int. Symp. VLSI Design Test (VDAT)*, May 2016, pp. 1–2.

Zhichao Chen received the B.Eng. degree in VLSI design and system integration from the Department of Electronic Science and Engineering, Nanjing University, Nanjing, China, in 2023. He is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, University of California at Los Angeles (UCLA), Los Angeles, CA, USA, under the supervision of Prof. P. Gupta.

His current research interests include advanced packaging, computer-aided design, and design-technology co-optimization.

Ali H. Hassan (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Department of Electronics and Communications Engineering, Cairo University, Giza, Egypt, in 2014 and 2018, respectively. He is currently working toward the Ph.D. degree in electrical engineering at the University of California at Los Angeles (UCLA), Los Angeles, CA, USA.

His current research interests include low-power high-speed PHY research in the electrical, optical, and short-distance millimeter-wave domains and cryo-CMOS circuits for high-performance low-power computing.

Mr. Hassan was a recipient of the 2021 UCLA ECE Department Fellowship and the 2022 UCLA Summer Mentored Research Fellowship.

Rhesa Ramadhan (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the Institut Teknologi Bandung, Bandung, Indonesia, in 2022.

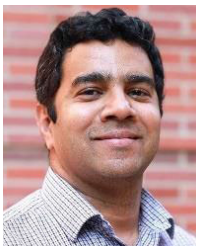
He was a member of the Microelectronics Center at the Institut Teknologi Bandung and the NanoCAD Laboratory at the University of California at Los Angeles, Los Angeles, CA, USA. During his research, his interests included VLSI design, low-power hardware accelerator systems, computer architecture, and artificial intelligence.

Yingheng Li received the B.Sc. and M.Sc. degrees in electrical engineering from Colorado State University, Fort Collins, CO, USA, in 2019 and 2021, respectively. He is currently working toward the Ph.D. degree in computer science at the University of Pittsburgh, Pittsburgh, PA, USA.

He was a member of the NanoCAD Laboratory at the University of California at Los Angeles, Los Angeles, CA, USA. His current research interests include uncertainty quantification, optimization algorithms, and machine learning.

Chih-Kong Ken Yang (Fellow, IEEE) was born in Taipei, Taiwan. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1992 and 1998, respectively.

He joined the University of California at Los Angeles, Los Angeles, CA, USA, in 1999, where he has been a Professor since 2009 and the Department Chair since 2020. His current research area is high-performance mixed-mode circuit design for VLSI systems. His research interests include clock generation, high-performance signaling, low-power digital functional blocks, analog-to-digital conversion, voltage converters, and building blocks for computer networks.



Sudhakar Pamarti (Senior Member, IEEE) received the B.Tech. degree in electronics and electrical communication engineering from Indian Institute of Technology at Kharagpur, Kharagpur, India, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at San Diego, San Diego, CA, USA, in 1999 and 2003, respectively.

He is currently a Professor of Electrical and Computer Engineering with the University of California at Los Angeles, Los Angeles, CA, USA. His current research interests include analog, mixed-signal, and RF integrated circuit design, specifically in developing signal processing techniques to overcome circuit impairments.

Dr. Pamarti was an IEEE Solid-State Circuits Society Distinguished Lecturer.

Puneet Gupta (Fellow, IEEE) received the B.Tech. degree in electrical engineering from Indian Institute of Technology Delhi, New Delhi, India, in 2000, and the Ph.D. degree from the University of California at San Diego, San Diego, CA, USA, in 2007.

He is currently a Faculty Member with the Department of Electrical and Computer Engineering, University of California at Los Angeles, Los Angeles, CA, USA. He has authored more than 200 articles, 18 U.S. patents, a book, and a book chapter in the areas of system-technology co-optimization and variability/reliability-aware architectures.