International Journal of Computational and Electronics Aspects in Engineering
Volume 7, Issue 5, November 2026, pp 646 - 658
https://doi.org/10.26706/ijceae.7.5.20261009

A Leakage-Aware Cross-Layer Analytical Framework for Energy-Efficient Computing: Integrating CMOS Power Modeling, Thermodynamic Constraints, and Energy–Delay Optimization

Mohammed A. Noaman Al-hayanni*

College of Electrical Engineering, University of Technology, Baghdad, Iraq.

***Correspondence:** mohammed.a.noaman@uotechnology.edu.iq



Article – Peer Reviewed
Received: 30 August 2026
Accepted: 27 October 2026
Published: 10 November 2026

Copyright: © 2026 RAME Publishers
This is an open access article under the CC BY 4.0 International License.
https://creativecommons.org/licenses/by/4.0/

Cite this article: Mohammed A. Noaman Al-hayanni, "A Leakage-Aware Cross-Layer Analytical Framework for Energy-Efficient Computing: Integrating CMOS Power Modeling, Thermodynamic Constraints, and Energy–Delay Optimization", *International Journal of Computational and Electronics Aspects in Engineering*, RAME Publishers, vol. 7, issue 5, pp. 646-658, 2026.
https://doi.org/10.26706/ijceae.20261009

Abstract: The increasing need for high-performance computing within the limits of power has made energy efficiency an important factor in computer designs in recent years. Traditional dynamic voltage and frequency scaling (DVFS) techniques aim to reduce dynamic power consumption by adjusting the voltage. However, too much voltage can lead to high execution delays as well as a higher impact of static leakage energy. Consequently, aggressive low-power operations are not effective in near-threshold computing systems. In this paper, we propose a leakage-aware model with three layers of computing system abstraction: thermodynamic limits, CMOS power and delay, and architectural performance. The proposed model combines the minimum energy boundary of Landauer's theory with the dynamic and leakage energy of CMOS systems and provides an optimization strategy based on the Energy-Delay Product (EDP). We show a mathematical formula to compare the relationship between the supply voltage, leakage current, threshold voltage, switching activity, and execution delay. The numerical calculations using CMOS technology parameters confirm the presence of an optimal voltage range in which the reduction in dynamic energy is compensated by the delay due to leakage.Sensitivity analysis further shows that variations in leakage current, threshold voltage, and workload characteristics significantly influence the optimal operating point. The proposed framework provides an analytical basis for leakage-aware voltage management and hardware–software co-design strategies in energy-constrained computing systems.

Keywords: Energy-Efficient Computing; Cross-layer optimization; CMOS Leakage Modeling; EDP; DVFS; Near-Threshold Computing

1. Introduction

The continuous increasing of computational workloads, such as Artificial Intelligence (AI), cloud computing and edge-based applications, have increased the demand for computing systems that can provide high performance under tight energy and thermal constraints [1], [2] [3] . Thus, energy efficiency became an important design factor at many levels of computing systems, from semiconductor devices to processor architectures. Historically, the scaling of transistors, guided by the principles of Moore's Law and Dennard scaling, strongly supported the improvements in computational efficiency [4][5]. The performance is improved while the power density is reduced simultaneously. However, further technology scaling reduces the effectiveness of these trends significantly according to the increase in leakage mechanisms, short-channel effects and limitations of further voltage reduction [6] .

Dynamic Voltage and Frequency Scaling (DVFS) became a common practice as a useful technique in reducing the computational energy consumption [7]. The dynamic switching power is quadratically dependent on supply voltage and thus, DVFS can provide significant

savings in active power consumption. However, there are limits to the benefits of voltage scaling. As the operating voltage approached the near-threshold region [8], the switching speed of the transistor is reduced considerably, which caused the longer execution time and the higher static leakage energy accumulation. This resulted in inefficient operating conditions where the decrease in dynamic energy is offset by an increase in leakage energy when the supply voltage is reduced without taking into account the delay effects caused by leakage part [9] [10].

The interaction between dynamic energy reduction and leakage energy accumulation represents a fundamental challenge in modern low-power computing design. Existing optimization approaches often investigated different abstraction levels independently. Device-oriented research usually examined physical constraints and transistor leakage behaviors, whereas circuit-level models investigated characteristics of power dissipation. On the other hand, architectural strategies emphasized the execution of workloads and optimization of performance [11]. This separation limited the ability to capture the nonlinear relationship between physical constraints, circuit behavior, and system-level performance.

Near-threshold computing showed the possibility of large energy reduction by operating CMOS circuits near the minimum energy point [12][13]. However, the optimal operating voltage is significantly influenced by technological parameters, leakage characteristics, temperature conditions, and workload patterns. Therefore, determining an efficient operating voltage point necessitates a comprehensive analysis of circuit-level energy dynamics alongside architectural execution demands.

Thermodynamic principles establish a basic framework for computational energy efficiency [14]. Landauer's principle established the minimum theoretical energy required for irreversible information processing, representing the physical limit toward which future computing technologies approached. However, practical CMOS systems operated several orders of magnitude above this limit according to switching losses, leakage currents, and architectural overheads. So, understanding the connection between this physical limitation and the operation of real circuits is vital to the formulation of efficient energy optimization strategies.

Recent studies have expanded the scope of energy optimization beyond voltage scaling to the use of workload characteristics, advanced transistor technology, and adaptive control at the system level [15], [16]. More recent energy-efficient computing methods incorporate circuit-level power modeling and architectural enhancements in order to overcome the limitations of conventional Dynamic Voltage and Frequency Scaling (DVFS) methods. However, most current methods have only focused on specific abstraction levels and do not explore how physical energy constraints, leakage behavior, and performance that depends on workload interact.

Table 1. Comparison Between Existing Energy Optimization Approaches and Proposed Framework

Method	Thermodynamic Limit	Leakage Model	Voltage Optimization	Workload Interaction	Architectural Impact
DVFS [7], [8], [10]	No	Limited	Yes	Limited	Partial
Near-threshold computing[12], [17]	No	Yes	Yes	Limited	Partial
CMOS energy models [13], [18], [19], [20], [21], [22]	No	Yes	Limited	No	No
Proposed framework	Yes	Yes	Yes	Yes	Yes

To address these challenges, this work proposes a leakage-aware cross-layer analytical framework that connects three complementary abstraction levels: thermodynamic constraints, CMOS power and delay modeling, and architectural performance metrics. The framework integrates the minimum energy boundary defined by Landauer's principle with dynamic and leakage energy models and evaluates the resulting trade-off using the Energy–Delay Product (EDP) metric. The primary contributions of this research are outlined below:

1. Cross-Layer Analytical Integration: Proposes an integrated analytical model that unifies thermodynamic energy limits, CMOS power dissipation mechanisms, and architectural execution characteristics.
2. Leakage-Aware Energy-Delay Optimization: Formulate a mathematical framework to analyze the tradeoff between dynamic energy reduction and leakage-induced delay penalties during voltage scaling.
3. Sensitivity analysis of operating conditions: A thorough study of the shifts of the optimal operating voltage according to leakage current, threshold voltage variations and workload switching activity.

4. Adaptive Voltage Management Insights: Leakage-aware DVFS design guidelines and hardware-software co-design for energy-constrained computing systems.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed cross-layer analytical framework, and describe the thermodynamic, circuit and architectural models. In Section 3, we present the leakage-aware energy formulation under voltage scaling. The EDP optimization model is formulated in Section 4. Section 5 describes the methodology and results of numerical evaluation. Sensitivity analysis under different technology and workload conditions is presented in Section 6. Section 7 presents limitations and directions for future research. The paper is concluded in Section 8.

2. Cross-Layer Analytical Framework

Energy-efficient computing design requires simultaneous assessment of physical constraints, circuit-level energy requirements, and system-level performance requirements. Standard optimization approaches tend to concentrate on one level of abstraction and may not take into account the interaction between voltage scaling, leakage accumulation, and workload performance. In this study, we offer a comprehensive analysis framework that combines thermodynamic limitations, CMOS power modeling, and architectural performance evaluation and provides the framework to optimize energy and delay. Three layers are considered in the framework, as shown in Figure 1.

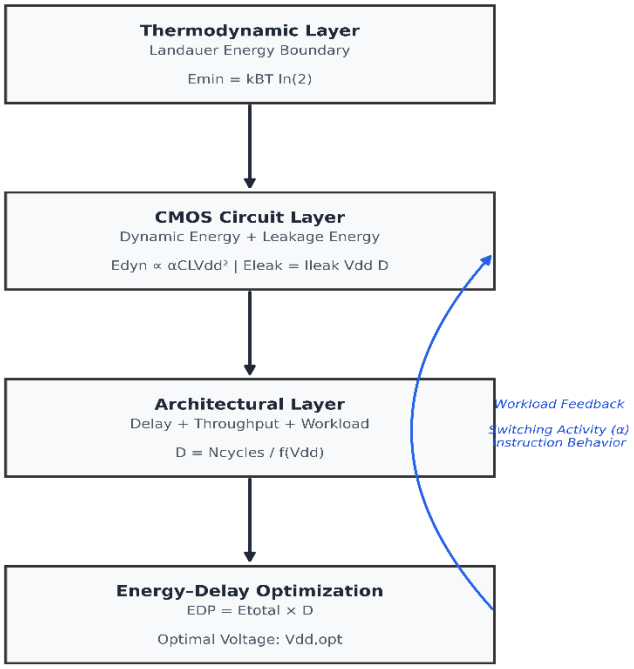


Figure 1. Proposed leakage-aware cross-layer analytical framework integrating thermodynamic limits, CMOS energy mechanisms, and architectural performance characteristics for EDP optimization.

The first layer signifies the thermodynamic limits of computation. At least according to Landauer's principle, there is a fundamental minimum energy required for irreversible information processing. The second layer describes the dynamics of CMOS circuits, namely the dynamic switching energy and the static leakage energy from the transistors. The third layer shows the architecture execution dynamics, in which the voltage of the supply affects the clock frequency, execution time, and overall computational speed. Between these layers, we can identify the trade-offs between energy consumption and performance

2.1 Thermodynamic Foundation

At the most basic physical level, computation is limited by the laws of thermodynamics. According to Landauer's principle, every logically irreversible process, like the erasure of a bit, necessitates a certain minimum amount of energy to be dissipated [19]. The least theoretical energy needed for one irreversible bit operation is defined as [23]:

$$E_{min} = k_B T ln(2) \quad . \tag{1}$$

Where E_{min} represents the least amount of energy lost during an irreversible process, k_B denotes Boltzmann's constant, and T signifies the absolute temperature. At a typical room temperature of 300 K, the theoretical minimum energy threshold is estimated to be $E_{min} = 2.87\times10^{-21}$ J. This is only a physical value and not the practical one because current CMOS systems use much more energy due to switching losses, leakage currents, and circuit overhead. To better understand the difference between real and thermodynamic performance, we introduced the thermodynamic efficiency factor [23][24]:

$$\eta_{th} = \frac{E_{min}}{E_{op}} \quad . \tag{2}$$

E_{op} represents the total energy consumed by the computational system. In later discussions on circuit and architectural design, E_{op} includes the switching energy and leakage energy during the whole computation process. This illustrates the distance from the physical energy threshold and shows that although it is optimal to reduce the supply voltage initially, reducing the supply voltage too much can lead to leakage becoming a serious problem and reaching critical levels

2.2 Circuit-Level CMOS Power and Energy Modeling

The circuit layer describes the electrical mechanisms responsible for energy consumption in CMOS processors. The overall power consumption consists of dynamic switching power and leakage power [18][19]:

$$P_{total} = P_{dyn} + P_{leak} \quad . \tag{3}$$

Where P_{dyn} is switching power, and P_{leak} is static leakage power.

The dynamic power is mainly according to the charging and draining of capacitive loads during the switching operations of the transistors [18][19]:

$$P_{dyn} = \alpha\, C_L V_{dd}^2 f \quad . \tag{4}$$

Where α is the switching activity factor, C_L is the effective load capacitance, V_{dd} is the supply voltage, and f is the operating frequency. This relationship describes reduction is a useful technique to reduce dynamic power since dynamic power decreases with the square of supply voltage. Leakage power is present even in the idle state of the circuit, unlike dynamic power [20]. The smaller the transistors become, the more and more important it is. The leakage power is given by:

$$P_{leak} = I_{leak} V_{dd} \quad . \tag{5}$$

Where I_{leak} is the leakage current. The leakage current is highly dependent on the threshold voltage, the temperature of the transistor and variations in the manufacturing process. Aggressive voltage scaling decreases the switching energy but increases the execution time such that the leakage energy accumulates.

2.3 Architectural Performance Model

The architectural layer connects circuit behavior with workload execution. For a computational workload requiring N_{cycles} clock cycles, execution delay is [7]:

$$D = \frac{N_{cycles}}{F(V_{dd})} \quad . \tag{6}$$

Where D signifies the execution delay, N_{cycles} denotes the complexity of the workload, and $F(V_{dd})$ refers to the operating frequency. This frequency is influenced by the supply voltage [17]. Based on the characteristics of CMOS delay, a decrease in V_{dd} results in a slower switching speed for transistors, thereby extending the execution time.

The throughput can be represented as:

$$Throughput = \frac{1}{CPI\times T_{cycle}} \quad . \tag{7}$$

Where CPI denotes cycles per instruction, and T_{cycle} represents the clock period. Thus, decreasing V_{dd} reduces dynamic energy; however, it also lowers frequency. This frequency reduction extends execution time, which in turn increases leakage energy; this forms the basis of the proposed Energy-Delay Product optimization.

3. Leakage Aware Energy Modeling Under Voltage Scaling

Supply voltage reduction is a practical way of reducing energy consumption in CMOS-based computing systems [5]. However, reducing dynamic switching energy leads to an increasing execution delay as the operating voltage approaches the threshold voltage of the transistor. As the computation time increases, the leakage power will not stop, and the increase in execution time will lead to more leakage energy, so it is important to take into account voltage-sensitive dynamic energy and

delay-related leakage energy for the total energy consumption of a computing task. The total amount of energy required to complete a task can be expressed as:

$$E_{total} = E_{dyn} + E_{leak} \quad . \tag{8}$$

Where E_{dyn} is the dynamic switching energy, and E_{leak} is the accumulated leakage energy.

3.1 Dynamic Energy Component

Dynamic energy is associated with charging and discharging the effective capacitance during transistor switching activities [13]. For a workload that necessitates clock cycles, dynamic energy can be represented as:

$$E_{dyn} = \alpha\, C_L V_{dd}^2 N_{cycles} \quad . \tag{9}$$

Where α is the switching activity factor, C_L is the load capacitance, V_{dd} is the supply voltage, and N_{cycles} are the number of clock cycles to run. The dynamic energy decreases quadratically with the supply voltage, and as a consequence, lowering V_{dd} leads to significant savings when the dynamic power is the dominant one. The problem is that, by reducing the voltage, the transistor switching speed is also reduced.

3.2 Leakage Energy Accumulation Mechanism

In contrast to dynamic energy, leakage energy depends on the duration of computation [5], [9], [14]. The leakage component is calculated as:

$$E_{leak} = P_{leak} D \quad . \tag{10}$$

$$\text{Since: } P_{leak} = I_{leak} V_{dd} \quad . \tag{11}$$

the leakage energy becomes:

$$E_{leak} = I_{leak}\, V_{dd} D \quad . \tag{12}$$

The execution delay is determined by the operating frequency as explained in equation (6). Thus:

$$E_{leak} = I_{leak} V_{dd} \frac{N_{cycles}}{F(V_{dd})} \quad . \tag{13}$$

This equation highlights the fundamental limitation of aggressive voltage scaling. Although reducing V_{dd} decreases the instantaneous leakage power, it also results in a decrease in the operating frequency. This increase in execution time allows for leakage current to remain active for an extended period, thereby increasing the overall leakage energy.

3.3 Near-Threshold Voltage Region

When the supply voltage approaches the transistor threshold voltage (V_{th}), the delay behavior becomes highly nonlinear [17] , and the operating frequency can be approximated using the alpha-power law model [22]:

$$f(V_{dd}) \propto I_{leak} \frac{(V_{dd} - V_{th})^{\delta}}{V_{dd}} \quad . \tag{14}$$

Where δ is the velocity saturation parameter.

As V_{dd} approaches V_{th}, the $(V_{dd} - V_{th})^{\delta}$ reaches 0 which causes, a significant reduction in frequency, a rapid increase in execution delay, and higher leakage energy accumulation, which in turn causes the non-linear behavior of the total energy. At high voltage, dynamic energy dominates, and reducing voltage improves efficiency. At near-threshold voltage, leakage energy becomes dominant, and further voltage reduction becomes inefficient. This produces an energy minimum region where dynamic energy reduction and leakage energy increase are balanced.

3.4 Total Energy Behavior During Voltage Scaling

According to the sum of the two energy components:

$$E_{total} = \alpha C_L V_{dd}^2 N_{cycles} + I_{leak} V_{dd} \frac{N_{cycles}}{f(V_{dd})} \quad . \tag{15}$$

The first term decreases as the voltage decreases (dynamic energy). The second term (leakage energy) increases when the delay expansion due to voltage scaling is important [7][9] . Thus, the minimum total energy condition is not satisfied at the

minimum supply voltage. Instead, it occurs at an intermediate operating point, where the reduction in dynamic energy is exactly compensated for by the increase in leakage energy. This behavior gives us the physical basis to find the optimal voltage point using the EDP formulation developed in the following section.

4. Energy–Delay Product Optimization

Energy efficiency in computing systems cannot be evaluated only through energy minimization because reducing energy consumption often increases execution time [13]. Thus, this study adopts the EDP metric, which provides a combined evaluation of energy consumption and computational performance [21], the EDP is defined as:

$$EDP = E_{total} \times D \quad . \tag{16}$$

Where *EDP* value decreases this will lead to a more efficient operating condition because it represents a favorable balance between energy consumption and computational speed.

4.1 Delay-Aware EDP Formulation

The execution delay is expressed in equation (6), substituting the total energy and delay equations into the EDP expression gives [7]:

$$EDP = (\alpha C_L V_{dd}^2 N_{cycles} + I_{leak} V_{dd} \frac{N_{cycles}}{f(V_{dd})}) \times \frac{N_{cycles}}{F(V_{dd})} = \alpha C_L V_{dd}^2 \frac{N_{cycles}^2}{f(V_{dd})} + I_{leak} V_{dd} \frac{N_{cycles}^2}{f(V_{dd})^2}. \tag{17}$$

This equation illustrates the two major contributors to EDP. The dynamic energy penalty and leakage delay penalty. The second term becomes dominant when the operating voltage approaches the threshold region because the frequency decreases rapidly.

4.2 Optimal Supply Voltage Determination

The optimal supply voltage is obtained by minimizing the EDP function, $\frac{\partial \mathrm{EDP}}{\partial V_{dd}} = 0$, and the solution represents the point where the energy reduction obtained from voltage scaling is exactly balanced by the delay-related leakage increase [7].

$$V_{dd.opt} = argmin(EDP(V_{dd})) \quad . \tag{18}$$

Where $V_{dd,opt}$ represents the energy-efficient operating voltage.

Because $f(V_{dd})$ has a nonlinear relationship with voltage, the optimization is solved numerically over the allowable voltage range. The optimal voltage achieved is not at the point of minimum dynamic energy. It is a balance between various competing factors. When supply voltage is high, the switching speed of transistors is rapid, and leakage energy is accumulated for a very short time, so dynamic energy is the most important factor. At low supply voltages, dynamic energy is lower, transistor delays are long, and leakage energy is accumulated during a long execution time. So, the optimum operating point is in the middle between these two extremes. This is why the most energy-efficient operating condition is typically found close to, but not necessarily exactly in, the near-threshold region.

5. Numerical Evaluation and Simulation Setup

We studied the proposed cross-layer analytical framework with numerical simulations to understand how voltage scaling, leakage energy accumulation, execution delay, and overall EDP relate to each other. This is primarily the phase where we seek to find the optimal operating voltage range for energy consumption and computational efficiency. The evaluation goes through three steps:

1. Calculating dynamic switching energy for different supply voltages.
2. Estimation of leakage energy accumulation according to voltage-dependent execution delay.
3. Determination of the minimum EDP operating point.

Voltage scaling analysis varies the supply voltage from standard operating conditions to the near-threshold region. To guarantee reproducibility, for each specified voltage level, the suggested analytical model computes total energy consumption, execution delay, and EDP are clearly documented. Table 2 summarizes the CMOS and workload parameters used in the evaluation [25].

Table 2. CMOS Technology and Workload Parameters Used for Numerical Evaluation

Parameter	Value	Parameter	Value
Technology node	45 nm	Leakage current I_{leak}	10 µA
Supply voltage range V_{dd}	0.3–1.2 V	Switching activity factor α	0.15
Threshold voltage V_{th}	0.25 V	Workload cycles N_{cycles}	1000000000 cycles
Velocity saturation parameter δ	1.3	Temperature T	300 K
Load capacitance C_L	1e-9 F		

5.1 Voltage Scaling and Energy Components

The first evaluation analyzes the variation of dynamic and leakage energy as a function of supply voltage. At higher supply voltages, the transistor switching speed is high, the execution delay is minimized, and leakage accumulation time is reduced. However, dynamic switching energy becomes dominant according to the V_{dd}^2 relationship. When voltage scaling approaches the threshold voltage, the dynamic energy continues decreasing, transistor switching speed decreases, and execution delay increases rapidly. Consequently, leakage energy becomes the dominant energy component, and additional voltage reduction increases total energy consumption. This behavior produces the characteristic minimum-energy region predicted by the proposed framework.

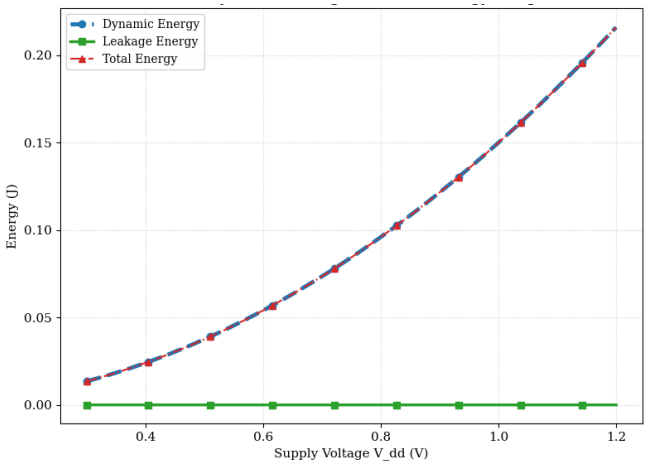


Figure 2. Variation of dynamic, leakage, and total energy components under supply voltage scaling.

5.3 Energy–Delay Product Optimization Results

We obtain the EDP curve by combining total energy and execution delay. At voltages above this point, energy consumption is increased according to excessive dynamic switching. While at voltages below this point, delay expansion increases leakage energy, and performance degradation becomes dominant. Thus, the optimum voltage represents a balance between energy reduction and performance preservation.

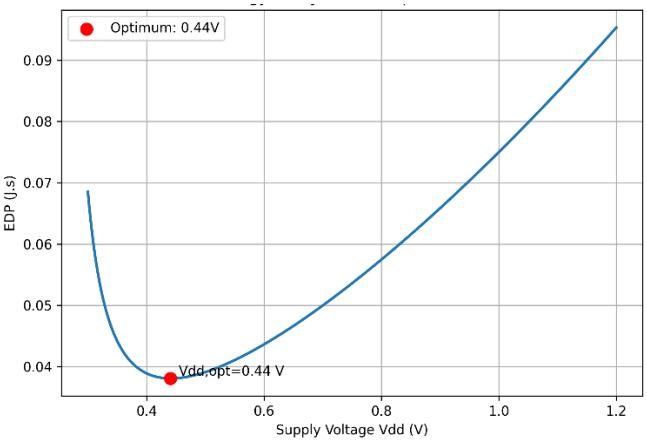


Figure 3. Energy–Delay Product variation with supply voltage and identification of the optimal operating point.

We summarize the quantitative evaluation of operating regions in Table 3.

Table 3. Energy and Performance Trade-Off at Different Voltage Regions

Operating Region	V_{dd}	Dynamic Energy	Leakage Energy	Delay	EDP
Super-threshold	High	High	Low	Low	—
Energy-optimal region	$V_{dd,opt}$	Moderate	Balanced	Moderate	Minimum
Near-threshold	Low	Low	High	High	Increased

5.4 Discussion of Numerical Findings

The numerical evaluation confirms that supply voltage scaling alone is insufficient for minimizing computational energy. Additionally, voltage reduction provides significant dynamic energy savings, the associated frequency reduction increases execution time and extends the period during which leakage current is active. The obtained results support the proposed cross-layer approach because the optimal operating voltage cannot be determined from circuit-level energy alone. Instead, it requires simultaneous consideration of transistor-level leakage behavior, voltage-dependent delay, and workload characteristics. These results provide the foundation for the sensitivity analysis presented in the following section.

6. Sensitivity Analysis

The optimal operating voltage of a CMOS-based computing system is not a fixed value because it depends on technology parameters, environmental conditions, and workload characteristics. Process variation, temperature-dependent leakage behavior, and differences in application activity can shift the balance between dynamic energy reduction and leakage energy accumulation. To assess the reliability of the suggested framework, a sensitivity analysis is conducted by individually altering the Leakage current, Threshold voltage, and Switching activity factor. The effect of each parameter is analyzed in relation to total energy consumption, execution delay, EDP, and optimal supply voltage.

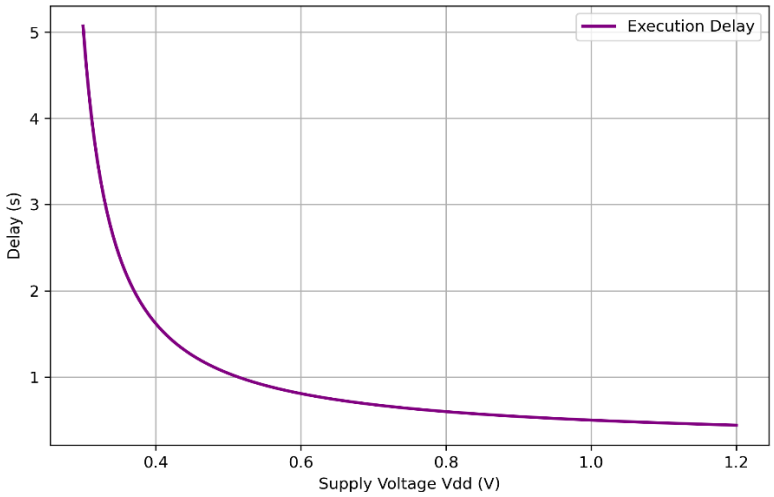


Figure 4. Delay and Throughput Behavior Under Voltage Scaling

6.1 Sensitivity to Leakage Current Variation

In scaled CMOS, leakage current is influenced by subthreshold conduction factors as well as gate leakage, and it plays a significant role in energy efficiency. With shrinking transistors, leakage components are more prominent, and static energy consumption is increased. Sensitivity analysis shows that the leakage current increases during operation, and leakage energy increases throughout the whole operating voltage range. As a consequence, the EDP shifts to higher supply voltages (e.g., ultra-low voltage), so no ultra-low voltage operation is possible. In fact, the leakage energy increases during long running times, leading to leakage losses higher than the savings from dynamic switching.

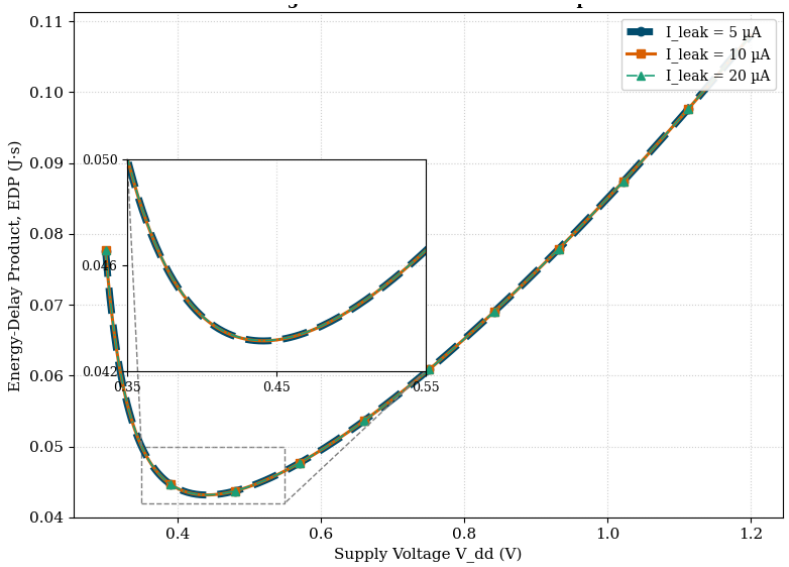


Figure 5. Influence of leakage current variation on the EDP optimization curve.

6.2 Sensitivity to Threshold Voltage Variation

The threshold voltage directly affects the switching speed and leakage characteristics of a transistor. The changes in the threshold voltage affect the voltage margin available for the transistor to operate. Higher threshold voltage results in lower transistor switching speed, higher execution delay and higher leakage related energy accumulation. Thus, $V_{dd,opt}$ shifts upward to compensate the reduction in operating speed. A low threshold voltage is providing higher switching speed and better performance. However, the leakage current increases and the static energy become significant. Thus, threshold voltage reduction is a trade-off between performance improvement and leakage penalty.

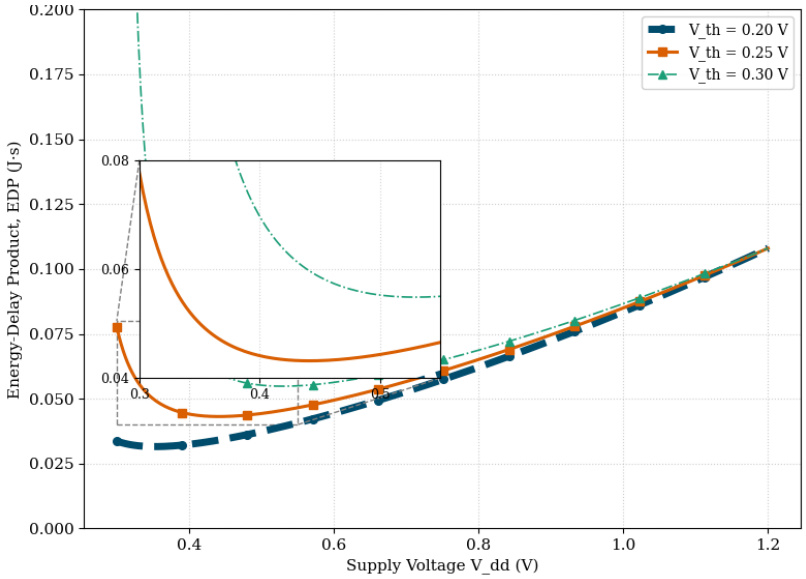


Figure 6. Effect of threshold voltage variation on energy-delay optimization.

6.3 Sensitivity to Switching Activity Factor

The switching activity factor represents the amount of circuit activity during computation. It determines the relative contribution of dynamic energy compared with leakage energy. Increasing switching activity increases the dynamic component. In high-activity workloads, such as, intensive numerical computation, signal processing, and AI acceleration. Dynamic energy dominates, voltage reduction provides greater energy benefit, and the optimal voltage may increase slightly according to the valuable of performance benefits. In workloads characterized by low activity, including memory-intensive applications and control-driven tasks, leakage energy tends to play a more significant role. Prolonged execution times lead to an accumulation of static energy, necessitating careful restrictions on voltage scaling.

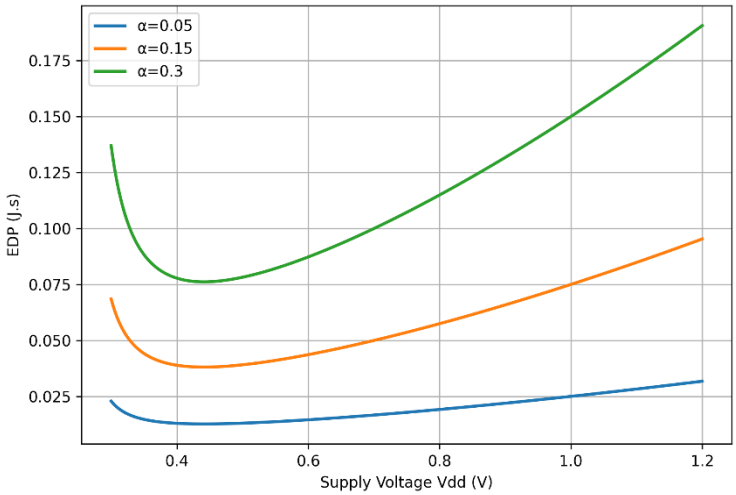


Figure 7. The impact of workload switching activity on optimal voltage selection.

6.4 Combined Process and Workload Interaction

In practical computing systems with CMOS technology, both technological and workload characteristics change simultaneously rather than separately. As such, we further examine our model by integrating the effects of the leakage current, the change in threshold voltage, and the switching activity of workloads on the energy-delay trade-off.

We have shown that the ideal operating voltage is influenced by the interaction between device properties and computation activities. Technologies with higher leakage rates increase static energy contributions, leading the optimal voltage to higher levels to reduce execution delays and to control leakage. Similarly, workloads with high activity increase dynamic energy contributions as switching energy and activity terms are linked. Thus, workloads that require intensive computation may require a different operating point from lower activity levels where leakage energy plays a much larger role.

Using this approach, our study shows that the optimal voltage setting depends on both workload and technology and thus is not always the same for all scenarios. As a result, voltage management in a device-specific environment and under thermal conditions and workload dynamics needs to be taken into account to ensure good voltage management. The interaction of these factors can be seen in Table 4, for statistical process variation analysis.

Table 4. Sensitivity Analysis Summary

Parameter	Variation range	Physical Effect	Change in $V_{dd.opt}$	Change in EDP
Leakage Current I_{leak}	5-20 µA	Higher leakage shifts optimum voltage upward Increased static energy accumulation	Increases	Increases
Threshold Voltage V_{th}	0.20-0.30 V	Reduced transistor switching speed	Increases	Increases
Switching Activity α	0.05-0.30	Increased dynamic energy contribution	Slight increase	Workload dependent

6.5 Statistical Process Variation Analysis (Monte Carlo)

The previous sensitivity analyses investigated the impact of particular technology and workload variables on the optimal operational voltage. However, practical CMOS circuits are affected simultaneously by multiple sources of variability, including transistor threshold voltage fluctuations and leakage current variations caused by manufacturing process uncertainty [26], [27]. Thus, we perform a Monte Carlo analysis to evaluate the robustness of the proposed optimization framework under combined parameter variations. In this examination, the threshold voltage and leakage current are treated as independent random variables that follow normal distribution patterns [28]:

$$V_{th} \sim N\left(\mu V_{th}, \sigma V_{th}^2\right) \quad (19)$$

$$I_{leak} \sim N\left(\mu I_{leak}, \sigma I_{leak}^2\right) \quad (20)$$

The stated nominal values correspond to the baseline CMOS setup used in previous simulations. In each instance of the Monte Carlo sampling, a new set of parameters is generated, and the EDP is recalculated across the complete voltage range. The voltage that results in the minimum EDP is then identified as the ideal operating voltage.

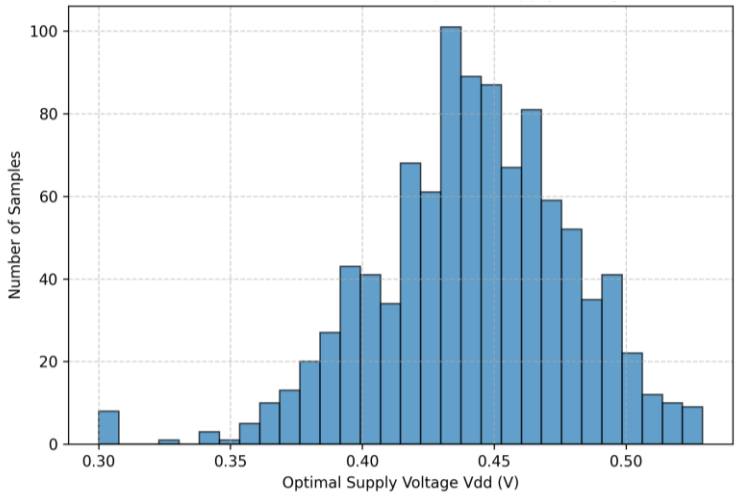


Figure 8. Statistical distribution of the ideal supply voltage derived from Monte Carlo simulations, taking into account concurrent fluctuations in threshold voltage and leakage current.

We have used 1,000 random samples to establish the statistical distribution. Our results show that the deal voltage is not fixed; it varies with the device's level. This is because the transistors' switching ability is low as the threshold voltage increases, and their energy consumption is increased due to longer running times. We show in Monte Carlo that process variation results in a range in the optimal voltage region and not in the loss of any optimization. The average value is the standard operating condition, but the standard deviation indicates how sensitive the optimal point is to manufacturing changes. Figure 8 is the statistical distribution of the optimal supply voltage obtained from the Monte Carlo simulation. A narrow distribution means the performance of the device is more stable, whereas a wider distribution means voltage regulation is more important to control device variation.

7. Discussion Limitations, and Future Research Directions.

The above insights have a significant impact on DVFS and hardware and software design and development. Traditional DVFS approaches are not suited for areas where leakage penalties are highest. Therefore, we need to set minimum voltage thresholds based on the EDP (EDP(V_{dd})) and not at the cost of the least dynamic power. Moreover, the optimal efficiency may depend on the workload and environment. In high-activity workloads, energy consumption is dominated by dynamic power, and voltage scaling can be beneficial for saving a lot of energy. Low-activity workloads, on the other hand, are more likely to leak. Consequently, current energy management strategies need to adjust to fluctuating runtime characteristics such as the switching activity factors (α), manufacturing process, and thermal situation. Future architectures should adopt these adaptive models for voltage management, energy-conscious task scheduling, and the design of accelerators instead of static ones [29]. Although such a framework relies on simplified assumptions that suggest directions for future work, first of all, CMOS is based on analytical models founded on the alpha power law and lumped leakage model. While these models capture general phenomena related to voltage scaling and transitional regions, they do not account for more advanced physical phenomena such as Drain-Induced Barrier Lowering (DIBL), short-channel effects, gate leakage mechanisms, or other characteristics of modern FinFETs or nanosheets. These models omit advanced physical effects such as DIBL, short-channel effects, gate leakage mechanisms, and modern FinFET or nanosheet characteristics. Second, the workload is modeled with constant execution cycles, omitting instruction-level diversity, memory stalls, and branch variations. Third, although leakage is heavily influenced by temperature. The current formulation lacks a fully coupled thermal model to capture the critical feedback loop where temperature increases leakage, which raises power and further elevates temperature. Future development addresses these constraints through SPICE-level device characterization, coupled thermal-electrical simulation, and hardware validation on FPGA and SoC platforms.

8. Conclusion

This study presented a leakage-aware cross-layer analytical framework for evaluating energy-efficient computing under supply voltage scaling. The proposed framework integrates three complementary abstraction levels: the thermodynamic energy boundary defined by Landauer's principle, CMOS-level dynamic and leakage energy mechanisms, and architectural execution characteristics represented through delay and throughput metrics.

The developed analytical model shows that the energy optimization of CMOS-based computing systems must take into account the relationship between dynamic energy reduction and the accumulation of leakage energy. The lower supply voltage reduces the dynamic switching energy. However, the corresponding decrease in the operating frequency increases the execution delay and extends the period during which leakage power is consumed in the total energy consumption. Thus, the minimum energy operating condition is determined by the trade-off between these competing effects rather than by voltage reduction alone.

The EDP-based optimization approach provides a quantitative way to decide the operating voltage region that can have a good trade-off between the energy consumption and the computational performance. The sensitivity analysis shows that the optimal operating point is affected by leakage current, threshold voltage variation and workload switching activity. The energy efficient operation depends on technology features and application behavior.

The proposed framework provides an analytical foundation for leakage-aware DVFS strategies and hardware–software co-design approaches. Instead of substituting current energy management strategies, this approach offers a way to comprehend their operational boundaries and choose suitable voltage conditions across various workload and technology contexts. Subsequent efforts will aim to enhance the model by incorporating advanced transistor technologies, thermal-electrical interactions, dynamic workload simulations, and validation through hardware on actual computing platforms.

References

[1] H. A. Mutar and I. R. N. ALRubeei, "AI-Driven Power Optimization in 6G Networks: A Comprehensive Review," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 1, pp. 45–57, Mar. 2026, doi: 10.26706/ijceae.7.1.20260104.

[2] H. A. Mutar, I. J. Mohammed, and I. R. N. ALRubeei, "AI for Solar PV Forecasting, MPPT, and Energy Management: A Comprehensive Review," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 1, pp. 100–114, Mar. 2026, doi: 10.26706/ijceae.7.1.20260111.

[3] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, 2014, pp. 10–14.

[4] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[5] G. E. Moore and others, "Cramming more components onto integrated circuits," 1965, *McGraw-Hill New York*.

[6] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proceedings of the 38th annual international symposium on Computer architecture*, 2011, pp. 365–376.

[7] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A Dynamic Voltage Scaled Microprocessor System," 2000.

[8] T. D. Burd and R. W. Brodersen, "Design issues for dynamic voltage scaling," in *Proceedings of the 2000 international symposium on Low power electronics and design*, 2000, pp. 9–14.

[9] T. D. Burd and R. W. Brodersen, "Processor Design for Portable Systems."

[10] J. Zidar, T. Matić, I. Aleksi, and Ž. Hocenski, "Dynamic voltage and frequency scaling as a method for reducing energy consumption in ultra-low-power embedded systems," *Electronics (Basel).*, vol. 13, no. 5, p. 826, 2024.

[11] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003, doi: 10.1109/JPROC.2002.808156.

[12] S. Salamin, H. Amrouch, and J. Henkel, "Selecting the optimal energy point in near-threshold computing," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019, pp. 1691–1696.

[13] R. Gonzalez and M. Horowitz, "Energy Dissipation In General Purpose Microprocessors," 1996.

[14] R. Muralidhar, R. Borovica-Gajic, and R. Buyya, "Energy Efficient Computing Systems: Architectures, Abstractions and Modeling to Techniques and Standards," *ACM Comput. Surv.*, vol. 54, no. 11s, Jul. 2022, doi: 10.1145/3511094.

[15] A. A. Hussain, M. N. Hussein, and Z. F. Alnaseri, "Optimized Energy-Efficient Cluster Routing in IoT-Enabled Wireless Sensor Networks via Mapdiminution-Based Training and Discovery Algorithm," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 6, no. 2, pp. 70–80, 2025, doi: 10.26706/ijceae.6.2.20240403.

[16] I. A. Ameen and M. A. Al-Sheikh, "Wireless Sensor Networks for Smart Gardening: ESP-NOW and Blynk IoT Integration for Water and Energy Optimization," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 5, no. 3, pp. 101–115, Sep. 2024, doi: 10.26706/ijceae.5.3.20240804.

[17] J. Nunez-Yanez and N. Howard, "Energy-efficient neural networks with near-threshold processors and hardware accelerators." [Online]. Available: https://github.com/eejlny/subthreshold_hetero_mcu_ncs

[18] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEICE Transactions on Electronics*, vol. 75, no. 4, pp. 371–382, 1992.

[19] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, *Digital integrated circuits*, vol. 2. Prentice hall Englewood Cliffs, 2002.

[20] S. Borkar, "Design Challenges of Technology Scaling," 1999.

[21] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, 1997.

[22] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-threshold design for ultra low-power systems*. Springer, 2006.

[23] C. H. Bennett, "The Thermodynamics of Computation-a Review," 1982.

[24] R. Landauer, "Irreversibility and Heat Generation in the Computing Process."

[25] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45nm Design Exploration." [Online]. Available: http://www.eas.asu.edu/~ptm

[26] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proceedings of the 40th annual Design Automation Conference*, 2003, pp. 338–342.

[27] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, 2002.

[28] M. J. M. Pelgrom and A. C. J. Duinmaijer, "Matching properties of MOS transistors," in *ESSCIRC'88: Fourteenth European Solid-State Circuits Conference*, 1988, pp. 327–330.

[29] M. Kadhim, "Improving Power Quality with Active Filtering and Energy Storage Technologies in Grid-Connected Hybrid Solar-Wind Renewable Energy Systems," *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 2, pp. 291–304, Jun. 2026, doi: 10.26706/ijceae.7.2.20260317.