Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Energy-Efficient Semiconductor Physical Design and Its Role in Sustainable Computing Systems

Rambabu Mandalapu
Independent Researcher, USA

ARTICLE INFO
Received: 05 March 2026
Accepted: 08 March 2026

ABSTRACT
With environmental, economic, and social considerations dominating computer system design, energy efficiency has become one of the primary concerns. This article will present the basic theory behind energy-efficient semiconductor physical design and discuss its implications for energy-sustainable computer systems. Topics include static and dynamic power dissipation in IC circuits, technology scaling, and power analysis methodology. For power optimization, the following physical design methodology and techniques are discussed: multi-voltage domain design, clock tree synthesis, and clever placement and routing techniques. The article reviews state-of-the-art power management techniques such as dynamic voltage and frequency scaling, power gating, and near-threshold operation and their impact on energy usage and sustainability, as well as environmental, economic, and policy aspects of sustainability. While carbon-aware design optimization and lifecycle assessment strategies are covered, energy-efficient physical design techniques continue to be a critical aspect of responsible technology development. Along with technological advancement and economic drivers, regulation in an increasingly power-constrained computing landscape has elevated the importance of energy-efficient physical design techniques.

Keywords: Energy-Efficient Design, Semiconductor Physical Design, Power Optimization, Sustainable Computing, Carbon-Aware Architectures

Introduction

Energy efficiency is a key design requirement of all computing systems today, due to growing environmental and social pressures. Semiconductor manufacturers are faced with a rapid demand for higher performance, with lower power consumption becoming the most important design constraint [1]. Physical design has strong implications for power consumption, thermal characteristics, and overall performance of the integrated circuit. Microprocessor design can be energy-efficient by saving 10% to 50% of energy through the optimization of circuit architecture and implementation strategies [1]. The increasing number of computing devices has caused a rapid growth of the worldwide digital system in terms of energy dissipation. This has required the development of techniques to reduce energy consumption while maintaining a given level of computation. The basics of energy conservation are based on supply voltage and power consumption. Dynamic power is quadratically proportional to the voltage levels [2]. Integrated circuits dissipate power from various sources, including dynamic switching, short-circuit power, and leakage power. Since the magnitude of leakage currents increased as technology nodes shrank, leakage now accounts for 30-40% of total power in idle mode, which has changed the design model [2]. This article presents an overview of energy-efficient semiconductor physical design implications on energy-efficient computing systems, describing their technical implementations as well as their importance to sustainability in the computing industry.

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Fundamentals of Power Consumption in Semiconductor Design

Static and Dynamic Power Components

The two primary forms of power dissipated by an IC include static power and dynamic power. Static power is consumed when the circuit is not being driven and is mainly due to leakage currents through transistors' junctions and subthreshold conduction [3]. Technology scaling has led to an important increase in leakage power due to smaller transistor dimensions and a thinner gate oxide layer, which accounts for 30-50% of total power in advanced process technology nodes (22 nm and below) [3]. Dynamic power consumption is due to the switching of the signals, charging/discharging of capacitive loads, and short-circuit currents. Dynamic power is quadratic in voltage, $P\_dynamic = \alpha \cdot C \cdot V^2 \cdot f$, where α is switching activity, C is capacitance, V is supply voltage, and f is frequency. Because of this quadratic dependence, voltage scaling is the most effective power-saving technique, achieving up to 70-80% power savings between nominal and near-threshold voltages [5].

In the idle state, static power is dominant, and in the working state, dynamic power dominates. Understanding the impact of both the components helps to identify and direct the optimization method in different modes of operation. Further complicating matters are temperature dependencies: leakage currents double for every 15 degrees Celsius rise in temperature, but maximum dynamic switching rate may also slow down at higher temperatures [3]. Solutions for these static/dynamic power tradeoffs in static physical design flow include multi-threshold voltage libraries, power gating schemes, and adaptive voltage-frequency scaling (AVS).

Technology Scaling and Power Density Challenges

These scaling effects have yielded enormous improvements, dramatically increasing the number of devices per die and their performance. Under Moore's Law, a doubling of transistor counts (and performance) every 18-24 months is the historical trend; however, increasing power density and device reliability issues have forced limits on further scaling. Below 45 nm, increased electric fields lead to more leakage due to quantum tunneling and also to more reliability issues, such as time-dependent dielectric breakdown [4]. Power density scales worse than linearly, causing thermal issues where performance is limited by the amount of heat the system can dissipate [3]. Because the classical Dennard scaling laws have been shown to break down around 2005, there is a need for alternative methods to increase transistor density per die while improving integration efficiency, as voltage scaling is no longer possible [5].

Variability due to the manufacturing process is another dimension of power optimization: device behavior varies statistically on a die and a wafer; for instance, the threshold voltage may vary by 2.5 V for an advanced process, which causes a variation of leakage power from 2 to 3 times [3]. These variations affect performance and power and require design methodologies, including statistical variation and corner analysis. Multi-gate transistors such as FinFETs and gate-all-around nanowires are a part of an attempt to simultaneously provide better power density and scaling and alleviate the scaling end of life given above [4]. Physical design must keep up with dynamic voltage-frequency scaling, near-threshold computing for slow and non-critical paths, and three-dimensional integration to reduce the power of interconnects.

Power Analysis and Estimation Methodologies

Correct power estimation throughout the design process is necessary for proper optimization and confirmation of efficiency goals. Broadly, the techniques depend on the design stage and the required precision. New technologies can help to reduce the effect of 16% [4]. Gate-level static timing analysis uses power models to estimate switching activity and capacitive loading as a function of circuit topology and input patterns with high accuracy when properly calibrated [3]. Dynamic simulation simulates switching activity under a particular set of workload conditions and produces detailed

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

power estimates; it is more accurate but slower to run [3]. Statistical power analysis approaches use Monte Carlo and probabilistic modeling to account for the stochasticity in process parameters and operating condition variations [5]. Probabilistic models for power distribution are essential for yield-aware optimization, which accounts for manufacturing variability and the uncertainties in process parameters and operating conditions. Early-stage power estimation at the architectural level is typically carried out using simpler models that can be computed with lower time complexity than gate-level tools, allowing for more rapid exploration of the design space. Power estimation at later stages typically proceeds with parasitic extraction, interconnect modeling, dynamic power estimation for modern designs, and thermal analysis using coupled electro-thermal simulation [3].

The accuracy-complexity trade-off determines which method is used. Accurate transistor-level SPICE simulations can take hours to days to simulate a circuit block but give the most accurate results. Power models are verified against measurements on silicon, and a typical calibration results in more accurate estimations [3]. Integrating power analysis into design flows allows for iterative power optimization and verification against power constraints at various stages of design to catch violations earlier and to guide target energy efficiency.

Methodology	Design Stage	Key Applications
Architectural	Early planning	Design space exploration
RTL Statistical	Pre-synthesis	Power budgeting
Gate-level Static	Post-synthesis	Sign-off verification
Dynamic Simulation	Post-layout	Workload profiling
SPICE Transistor	Critical paths	Accuracy validation

Table 1: Power Analysis Methodologies Comparison [3, 5]

Physical Design Techniques for Power Optimization

Multi-Voltage and Power Domain Strategies

Multi-voltage design is a technique that separates circuits into domains with different supply voltages. This allows higher voltages to be used for performance-critical paths and lower voltages to be used for non-performance-critical paths. Since the timing requirements of functional blocks can be heterogeneous, this can save up to 30% of power in datapaths without increasing area or delay [6]. As the voltages of the functional units are scaled down, and the speed of the system is reduced, but the power dissipation is reduced as well. Voltage scaling algorithms determine the voltages of the functional units such that the timing constraints are satisfied and the power is minimized [6]. Graph-theoretic methods have been proposed to minimize power for variable-voltage operation but incur high area overheads for additional functional units and supporting circuitry [6]. The increase in performance from microarchitecture enhancements diminishes with an increasing number of cores due to Pollack's Rule, where doubling the logic in a processor core will only yield 40% more performance [7]. Multi-core architectures with voltage scaling gain a performance increase of 70-80% compared to the same implementation of monolithic cores, with independent control of voltage per core [7]. Level shifters allow for signal transmission across the voltage domain but incur area and power overhead, requiring trade-offs. Domain isolation can be used to prevent a power supply from polluting other domains' voltages when it is turned off. In fine-grained power gating, individual domains can be turned off during inactivity to gain extra power savings, while other domains continue to operate.

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supply voltage assignment optimization algorithms for registers utilize timing margins and workload characteristics to model performance and power constraints. Experiments show that register allocation and binding algorithms that minimize power can yield average register file power savings of 20% when formulated as minimum cost clique covering problems and optimally solved in polynomial time by max-cost flow algorithms [6]. In contrast, adaptive voltage scaling changes supply levels dynamically based on real-time feedback on workload and environment. This requires complicated controls and power management capabilities. 3D integration of memory with processors is a promising technology, where the memory is placed between the processor and the package after thinning. Si vias are used for signal and power routing, allowing voltage to be tuned per layer [7].

Clock Tree Synthesis and Distribution Optimization

In terms of dynamic power dissipation, clock networks can contribute up to 45% of the power in high-performance designs and are often the largest power consumer in many other architectures [6]. As a consequence, the objective of clock tree synthesis is to control capacitive load and switching activity while achieving appropriate timing. By carefully setting the size and placement of buffers, one can minimize power dissipation and clock skew. Further reduction in switching power can be accomplished through clock gating, which turns off the clock signal to unused portions of the circuit. Clock gating is among the best techniques for reducing power consumption, especially when the activity is predictable or the granularity of the control is high, though that adds complexity to the design. Automatic clock gating insertion during synthesis identifies registers and logic blocks that are candidates for clock gating through structural analysis that identifies safe locations at which to gate the clock while maintaining the functional correctness of the design. Glitches and hazards (spurious switching activity) account for about 15-20% of the total dissipated energy in a circuit on average (varies from 9-38% depending on individual circuit characteristics) [6]. For example, 32-bit pipelined multipliers have been found to dissipate up to three times the standby power due to hazard activity than functional activity; hence, minimizing spurious transitions through logic synthesis and path balancing is essential [6].

Another major subsystem that benefits from low-power techniques is cache, which employs clock gating. Intel's 65 nm 16 MB on-die L3 cache for the dual-core Xeon 7100 uses low-power leakage and dynamic energy-saving techniques such as clock gating, where only 0.8% of the array blocks are powered up on each cache access. Intel's 45 nm Enterprise Xeon 8-core processor employs state-preserving techniques to reduce cache leakage by more than 2× in SRAM arrays and peripherals over unoptimized caches [8]. H-tree, mesh and hybrid clock tree topologies are typically used to optimize the clock power and timing characteristics of an integrated circuit. Although mesh trees reduce skew, they incur a power penalty due to their higher capacitance, typically 20-40% higher than optimized trees. Resonant clock distribution recovers switching energy through inductive coupling and can reduce power by 30-50% in some cases. The additional complexity, however, limits the common use of this approach [6].

Placement and Routing for Reduced Power

Layout choices affect power consumption due to interconnect length, capacitance, and activity. In particular, interconnect capacitance plays a greater role than ever in determining the total capacitance of a gate. In advanced designs, interconnect capacitance has been estimated as 40% to 70% of total capacitance, and it is expected to increase further with process scaling [6]. Power-aware placement algorithms take timing and power into account. For example, they group logic elements that communicate frequently to reduce the capacitance and dynamic power of the interconnect network. Activity-driven placement techniques place high-activity cells in places where the most frequently switching nets have low interconnect power. Minimizing capacitance has great potential in reducing power dissipation. Power is proportional to the capacitances seen at each gate [6]. For adders, multipliers, memory arrays, and address decoders, pre-characterization provides accurate capacitance

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

estimates. For random logic, analytic models such as those based on the number of input and output signals, the size of the circuit such as the number of states for finite-state machines or the number of cubes in sum-of-products expressions, and the technology/library [6] provide good estimates. Thermal-aware placement also prevents hot spot formation due to performance and reliability concerns. This is critical for thousand-core chips, where resilient microarchitectures use error detection, fault isolation, and hardware reconfiguration due to local hotspots on chip surfaces. More advanced caches can favor smart placement and data management, as shown in the next section. 3D stacked DRAM caches can be very energy efficient, as data can be fetched at the granularity of page sizes, and only data needed when the page is resident is fetched to avoid bandwidth/energy waste [8]. Predicting useful blocks with the help of spatial correlation reduces cache/energy bandwidth demand. These caches can outperform conventional/block-based 3D stacked caches concerning energy efficiency [8]. Sub-block-level cache policies exploit the predictability of cache accesses to specific sub-blocks in a cache line and turn on the necessary sub-blocks while turning off the unneeded ones after some predicted number of accesses, achieving better savings than block-level cache policies like decay caches or drowsy caches [8].

Routing for low power consists of minimizing wire capacitance through layer assignment, wire sizing, and path selection. Lower metal positions have lower capacitance (30% to 50% less than upper layers) but have higher density. These competing goals for power and routability must be carefully tuned [6]. The shield insertion and wire spacing control crosstalk and capacitive loading. Shielding reduces crosstalk-induced power by 15-25% but adds capacitance overhead. Buffer insertion for long interconnect involves several trade-offs, including the costs of drive strength versus the power penalty of additional buffer stages. Buffers should be sized optimally to reduce short-circuit power to below 15% of the dynamic power when the rise/fall times are matched [6]. Power-supply aware routing (a.k.a. power delivery network (PDN) routing) controls current and reduces resistive losses on the power distribution network. This is important because the interconnect power usage within a poorly routed design can use 10-20% of the total design power. Interconnect estimates can be extracted from companion placement results or stochastic/procedural interconnect models, which are used for early power estimation. [6] Via minimization also reduces resistance and improves reliability through decreased series resistance and improved electromigration resistance, which indirectly improves power efficiency through improved electrical performance metrics such as voltage droops and signal integrity across the die.

Technique	Power Component Targeted	Typical Savings (%)	Implementation Overhead	Key Applications
Multi-voltage domains	Dynamic, Static	20-40	Level shifters, Regulators	Multi-core processors
Clock gating (Fine-grain)	Dynamic (Clock)	30-50	Control logic, Gating cells	Cache arrays, Datapaths
Register binding optimization	Dynamic (Registers)	15-25	Allocation algorithms	Pipelined datapaths
Layer-aware routing	Dynamic (Wire)	10-20	Routing complexity	Long interconnects
3D DRAM cache management	Dynamic, Static	NA	3D integration cost	Memory hierarchies
Cache sub-block management	Dynamic, Leakage	30-50	Prediction logic	L2/L3 caches

Table 2: Physical Design Power Optimization Techniques and Achievable Savings [6, 7, 8]

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Advanced Power Management Architectures

Dynamic Voltage and Frequency Scaling

Dynamic voltage and frequency scaling (DVFS) keeps power consumption in line with demand by reducing voltage and frequency during low-demand periods. The usage model for DVFS is based on the dynamic nature of application workloads, which may vary considerably over time, in order to achieve meaningful savings. Customarily, scaling a technology results in a 30% decrease in gate delay, which translates to a 43% increase in operating frequency, which yields a 65% reduction in energy per transition. This translates to a 50% reduction in power at the new frequency [9]. Microprocessors double their clock frequency every generation, more than the 43% increase from technology scaling, due to aggressive architecture scaling and high levels of parallelism [9]. Power scales as the square of voltage, so there is potential for large power savings if performance is adequate to use a lower voltage, as a half-voltage change results in a quarter of the power dissipation [10]. Power reductions via frequency scaling alone are limited by the linear dependence between execution time and frequency. Using voltage-frequency scaling conserves power while at the same time reducing execution time. Modern voltage regulators can switch in tolerances of tens of millivolts at currents of hundreds of amperes at nearly 1V [10]. Newer microprocessors will pull over 100A at full utilization, requiring distributed power architectures (DPA): minimizing the distance between the power sources and chips will help meet the transient response requirements of such an arrangement [10].

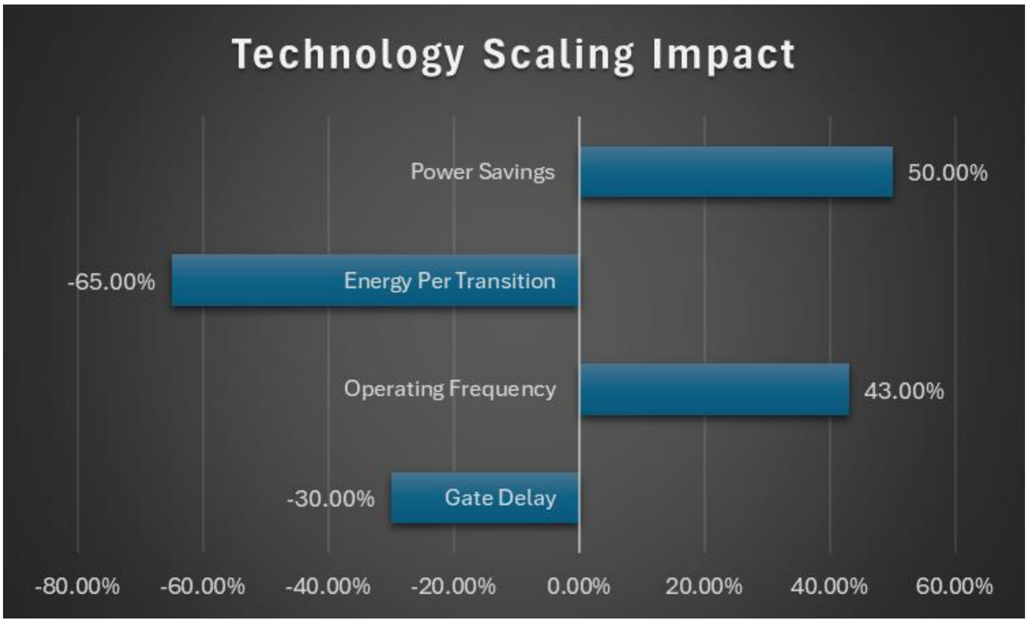

Figure 1: Technology Scaling Impact [9]

Subsequent implementations (e.g., of IBM POWER7) use per-core frequency scaling via digital phase-locked loops (DPLLs) with frequency steps as small as 28 MHz and slew rates of 50-110% of the nominal frequency [11]. Whereas previous SMPs and multicore systems typically change the frequency for all cores in a chip simultaneously, thus making DVFS slow and not appropriate for adaptive power management, the POWER7 uses an asynchronous core chiplet design that allows each of the eight cores to be frequency scaled independently. The EnergyScale firmware controls power and performance based on activity and workload profile per core [11]. Algorithms model future processing requirements based on previous behavior and application profiling. The hardware low-activity detect

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

(LAD) mechanisms can change the frequency of each core on an order of microsecond granularity, based on qualified IPC (instructions per cycle) data. Based on configurable IPC values, they reduce frequency when IPC is below programmable thresholds, thus taking advantage of even very brief low-activity periods that firmware cannot detect.

Power Gating and Retention Strategies

Power gating is a technique to greatly reduce static power consumption by cutting off supply from non-active blocks. Leakage currents are a proportional part of power consumption, and leakage increases exponentially with voltage scaling. To meet performance, the threshold voltage of a transistor can be reduced. However, this exponentially increases leakage [9]. Header and footer switches or header/footer gated level shifters can control the connection between domains. Switch size can trade off the area overhead and wake-up latency against the amount of power savings. Retention registers can preserve state during power-off and wake-up without the overhead of saving and restoring all state information. The retention registers are placed in separate power domains so that their latches stay powered when all other combinational logic is powered down. This maximizes leakage savings. For example, the POWER7 architecture implements retention voltage strategies that support all eight core chiplets going to retention when they go to sleep (the retention voltage being the voltage level needed to hold latch and array states for an extended period, while minimizing leakage) [11]. Heavy sleep mode latency on such architectures is typically 1 ms to 20 ms depending on the configuration of the system, with worst-case latencies exceeding 50 ms in certain situations that require ramping the power plane down to retention voltage and then back up again on wakeup [11]. This is also an advantage regarding the retention voltage when the frequency and voltage cannot be dropped to the lowest levels due to CICR (coherence-induced cycle restrictions) [11].

Balloon latches and retention cells are alternatives, but they provide different tradeoffs in terms of energy savings, area, and complexity. The granularity and duration of these events affect system-level savings. More advanced control policies are required to balance the overhead involved in state transitions with the power savings achieved in the powered-down states. During power-up transitions, ground bounce and voltage droops can lead to functional failures or timing violations, as the supply voltages are close to 1V, with a tolerance of a few hundred millivolts. Isolation cells prevent unknown states from spreading between powered and unpowered domains during transitions of power states while still maintaining correct functionality. This is useful with software visible power states, as the decision about what power state the domain should be in can be made by the OS based on what is needed from the application and hardware. The off-chip microcontroller of the POWER7 processor autonomously determines modes of operation and power-performance trade-offs at runtime per customer policy, POWER Hypervisor feedback, and operating system feedback [11]. This hierarchical control approach builds upon classic dynamic techniques such as clock gating and includes adaptive power management features that sense environmental conditions and workload characteristics.

Near-Threshold and Subthreshold Operation

Near-threshold voltage operation is a low-power design methodology in which the supply voltages are set to within a small margin of the threshold voltages of the transistors. This can be useful if speed can be traded off for a large reduction (by factors of 5 to 10) in power consumption. Due to the exponential relationship of subthreshold current to threshold voltage, near-threshold operation can lead to meaningful power savings through careful selection of the voltage. However, designs operating near the threshold are very sensitive to process variation and device mismatches, affecting timing and functionality. Each process generation introduces wider variations in VT, which range between 30 mV and 50 mV in recent process nodes, also having a rapidly increasing impact on delay and power when the supply approaches VT [9]. Constant electric field scaling obeys a supply voltage and threshold voltage scaling, which reduces the energy-delay product and increases the subthreshold leakage current exponentially, resulting in an active-leakage power trade-off. Adaptive body biasing and

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

statistical design alleviate the effects of processing variation and provide reliable operation across process corners and environmental conditions by dynamically compensating the threshold voltage variations [9].

To realize circuits at near-threshold voltages, a special standard cell library is used. Standard cell libraries do not allow for low noise margins and high variabilities in delay. Since domino circuits are often used for high performance, they suffer from decreased noise margins and increased contention by larger keeper transistors when threshold voltage is decreased [9]. Domino gates have a latency 30% less than static CMOS gates but also a 50% increase in power consumption (which decreases with low supply voltages) [9]. Scaling the supply voltage of most special circuits, such as sense amplifiers and programmable logic arrays (PLAs), requires a redesign of the critical parts of the circuit. In contrast, design for size and topology is done under noise margin, variability, and leakage dissipation considerations, rather than speed considerations. These principles are carried over into subthreshold mode, where the voltages are below the threshold voltage and the switching is driven entirely by the subthreshold current (rather than the gate-to-source voltage). Ultra-low-power sensors and energy harvesting systems that prioritize efficiency over performance and allow currents on the order of nanowatts also benefit from low supply voltages. Memory is difficult to implement at low voltages and typically requires increased wordlines, negative bitline schemes, and asymmetric cell sizes to enable reading and holding of data at low supply voltages [10].

Sustainability Implications and Broader Impact

Environmental Benefits and Carbon Footprint Reduction

Energy-efficient semiconductor design can have an important impact on environmental sustainability due to energy and carbon emission reductions. The information and communications technology (ICT) sector has a carbon footprint of almost 3% of the global total, with data centers alone representing an estimated 1-2% of the total electricity consumption [8, 13]. Total datacenter electricity usage was estimated at 460 TWh in 2022 (more than double from 2018) and projected to reach 620 TWh in 2026, equivalent to 140% growth across these recent years. A 1% efficiency improvement occurs across billions of machines globally. Such savings can have distinct and positive impacts on the environment, as they apply to operational emissions in computing machines, cooling, and power distribution. However, the environmental burden is the sum of operational and represented carbon emissions, with the represented carbon emissions from manufacturing and production accounting for 50% of cloud computing carbon emissions and over 70% of the consumer electronics carbon emissions. Life cycle assessment approaches are required [14].

The efficiency and environmental impacts of designs exist throughout the product life cycle, from manufacturing to in-use to end-of-life. When thermal management is relaxed (e.g., relaxed temperature and transfer rate), it reduces the need for cooling and packaging and minimizes carbon footprint and material use. An extended lifespan lowers the need to manufacture new devices, creating less electronic waste and spreading the represented carbon cost across a longer lifespan. Total lifecycle carbon (C_total) can only be measured with architectural carbon modeling tools, which take into account all operating carbon (CO_2e produced in energy consumed to run the devices) and represented carbon when they are produced [12]. The Architecture Carbon Modeling Tool (ACT) defines four new sustainability metrics: carbon-delay product, carbon-energy product, carbon^2-energy product, and carbon-energy^2 product, building on the EDP and EDAP metrics, which have been formularized by computer architects for decades. ACT allows designers to optimize their hardware design space search with respect to trade-offs between operational and manufacturing overheads in terms of sustainability. Power-aware design can also improve material efficiency, as the reduced power

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

requirement allows for smaller and lighter designs, which require less raw material, and reduce the energy used in shipping the product [12].

Economic Implications and Accessibility

Power efficiency has economic value beyond just energy cost savings and is foundational to total cost of ownership and technology availability. Lower operational costs for computing technologies would make these technologies more available to resource-constrained users and organizations and contribute to digital inclusion. The cost of infrastructure can be reduced when power and cooling requirements are lower, such as in emerging markets where electrical infrastructure is not yet ubiquitous [12]. The economics of efficiency are wide-ranging. For battery-powered equipment, a longer run time makes the device more valuable to its user, as they need to change batteries less frequently and make fewer capital expenditures. Since datacenter power consumption, cooling infrastructure, and configurations need not be supported, the cost of ownership is lower. This economic advantage reduces the cost of technology adoption and increases computer resource egalitarianism across socioeconomic lines [13].

Schools, small businesses, and home users in developing markets have additional motivation to adopt energy-efficient designs for lower TCO. In markets sensitive to operating costs and environmental impact, energy efficiency is becoming an important differentiator such that systems are evaluated and purchased on the basis of their five-year TCO and capital cost. Legal compliance and incentives, such as tax credits, rebate programs, and favoritism in public procurement, acknowledge the economic reward of efficiency. The long-term persistence of a given installed technology will depend on a variety of factors, including the management of the operating costs of equipment over long service lives. Power-efficient designs enable business models like edge computing and distributed processing architectures that minimize infrastructure investment and enable new service delivery. In addition, workloads for machine learning, such as large language models that require hundreds or thousands of GPU units running for weeks or months to train, as well as large amounts of high-power specialized hardware, create new economic incentives for power management techniques [13]. For inference services, horizontal and vertical autoscaling of the number of replica pods and GPUs is performed to optimize the costs while keeping the number of pods and GPUs aligned with the requested volume and service quality.

Policy and Standards Development

Such technical capabilities from power-efficient design methodologies inform policies and regulations related to sustainability. Climate change represents arguably the biggest threat facing humanity and is addressed through regulatory measures to reduce carbon emissions from all sectors of the economy [13]. Standards provide the measurement protocols, reporting formats, performance levels, and definitions that allow meaningful comparison of products or suppliers. They also drive procurement practices, regulatory compliance programs, and consumer information initiatives that shape market behavior. Globally, policies are increasingly driving market pull for electronics innovation through energy efficiency standards. Examples include the E.U. Energy-Related Products Directive, which specifies minimum efficiency requirements, phase-out timelines for inefficient technologies, and rules for testing and enforcement. Product labeling and energy efficiency ratings provide information to enable procurement choices that take account of environmental and business considerations. International climate agreements on how the information technology sector uses energy, including the Paris Agreement, set targets and time frames for energy consumption reduction [14].

Agile SoC design techniques can reduce carbon footprint by improving design reuse and designer productivity [13]. For example, the ESP (Embedded Scalable Platform) project shows that heterogeneous SoCs with 14 different accelerators, in addition to RISC-V cores, can be designed and

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

brought into production in 12 nm technology in three months or fewer than thirty person-months by small design teams [13]. This flexibility makes it more feasible to adopt low-carbon design decisions iteratively, allowing sustainable computing to be adopted without the lengthy multi-year development cycles of customary computing. Yet many important difficulties remain for fully realizing carbon-aware design optimization. Challenges include lack of cross-stack carbon-aware design tools that fit into design flows; difficulties in estimating carbon footprint due to lack of manufacturing transparency and location-specific/time-specific energy sources; lack of guidance on how to optimize approaches for carbon-awareness, as communities learn what design parameters have the most impact on carbon footprint [14]; and the enormous design space of combinations across process technologies, logic/memory devices, computer architectures, 3D integration techniques, and application use cases that are optimized across power, performance, area, and (now) carbon.

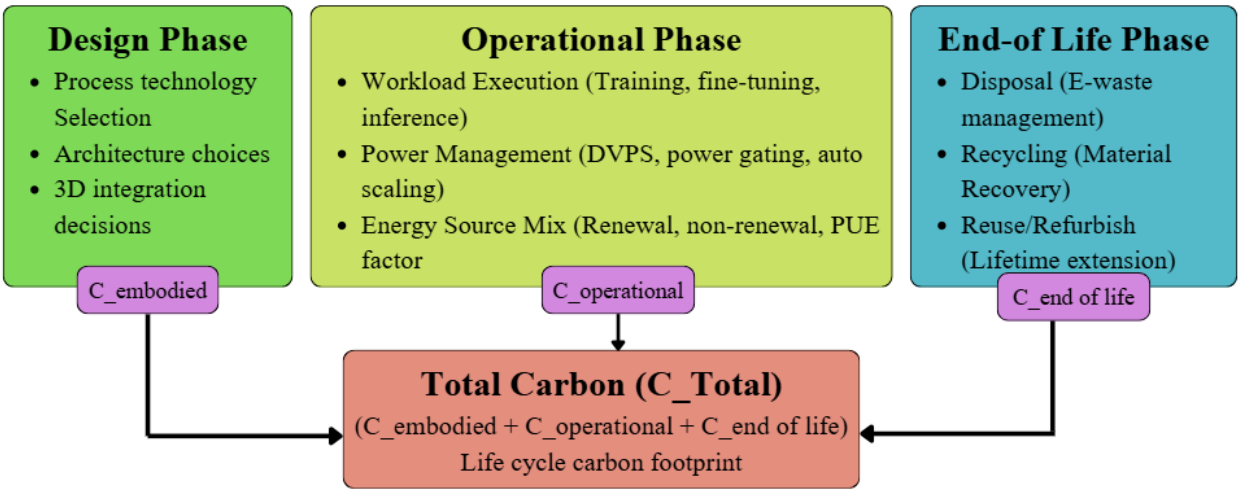

Figure 2: Life-Cycle Carbon Footprint Analysis Framework [12, 13, 14]

Conclusion

In summary, energy-efficient semiconductor physical design is the key to sustainable computing systems through systematic power reduction needed to tackle environmental, economic, and social issues. This article has surveyed essential approaches to multi-voltage physical design, clock distribution, smart placement, and power management that are leading the way to make power efficiency the first-order design criterion. Quantifiable energy savings and maintained performance are possible with power-aware design flows. Besides energy savings while the product is in use, the power-aware design also lowers the represented and operational carbon footprint on environmental grounds during life-cycle assessment. These economic benefits democratize the technology and enable economically sustainable business models for delivering the technology to a wide variety of customers. There has been some recognition of energy efficiency within policies or standards across sectors. Technical feasibility, economic incentives, and regulatory pressures frame energy-efficient physical design as a priority for computing infrastructure that can address global sustainability priorities and keep pace with increasing demand for computational resources.

References

[1] Thomas D. Burd and Robert W. Brodersen, "Energy efficient CMOS microprocessor design," Proceedings of the Twenty-Eighth Annual Hawaii International Conference on System Sciences, IEEE,

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1995. Available: https://www.researchgate.net/profile/Tom-Burd/publication/3596033_Energy_efficient_CMOS_microprocessor_design/links/5493dbe60cf25de74db4fe7a/Energy-efficient-CMOS-microprocessor-design.pdf

[2] Jan Rabaey, "Low Power Design Essentials," Springer Science & Business Media, 2009. Available: https://books.google.com/books?hl=en&lr=&id=A-sBy_nmQ8wC&oi=fnd&pg=PR2

[3] Sparsh Mittal, "A survey of techniques for improving energy efficiency in embedded computing systems," International Journal of Computer Aided Engineering and Technology, 2014. Available: https://arxiv.org/pdf/1401.0765

[4] Onur Mutlu, "The RowHammer problem and other issues we may face as memory becomes denser," Design, Automation & Test in Europe Conference & Exhibition, IEEE, 2017. Available: https://arxiv.org/pdf/1703.00626

[5] Anantha P. Chandrakasan, et al., "Low-power CMOS digital design," IEICE Transactions on Electronics, 1992. Available: https://www.academia.edu/download/4219478/10.1.1.136.1616.pdf

[6] Massoud Pedram, "Power minimization in IC design: Principles and applications," ACM Transactions on Design Automation of Electronic Systems (TODAES), 1996. Available: https://dl.acm.org/doi/pdf/10.1145/225871.225877

[7] Shekhar Borkar, "Thousand core chips: a technology perspective," Proceedings of the 44th Annual Design Automation Conference, 2007. Available: https://dl.acm.org/doi/pdf/10.1145/1278480.1278667

[8] Sparsh Mittal, "A survey of architectural techniques for improving cache power efficiency," Sustainable Computing: Informatics and Systems, 2014. Available: https://hal.science/hal-01103026/file/Mittal_SUSCOM_CacheSurvey.pdf

[9] Shekhar Borkar, "Design challenges of technology scaling," IEEE Micro, 2002. Available: https://bioee.ee.columbia.edu/courses/ee6321/papers/00782564.pdf

[10] Alexander Lidow et al., "The semiconductor roadmap for power management in the new millennium," Proceedings of the IEEE, 2001. Available: https://ieeexplore.ieee.org/abstract/document/931468/

[11] Michael Floyd et al., "Adaptive energy-management features of the IBM POWER7 chip," IBM Journal of Research and Development, 2011. Available: https://ieeexplore.ieee.org/abstract/document/5742026/

[12] Udit Gupta et al., "ACT: designing sustainable computer systems with an architectural carbon modeling tool," Proceedings of the 49th Annual International Symposium on Computer Architecture, 2022. Available: https://dl.acm.org/doi/pdf/10.1145/3470496.3527408

[13] Tamar Eilam et al., "Reducing datacenter compute carbon footprint by harnessing the power of specialization: Principles, metrics, challenges and opportunities," IEEE Transactions on Semiconductor Manufacturing, 2024. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10618912

[14] Mariam Elgamal et al., "Carbon-efficient design optimization for computing systems," Proceedings of the 2nd Workshop on Sustainable Computer Systems, 2023. Available: https://dl.acm.org/doi/pdf/10.1145/3604930.3605712

Copyright © 2026 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.