International Journal of Multidisciplinary Research and Growth Evaluation.

Designing Power-Efficient Systems-on-Chip (SoCs) for AI-Driven Consumer Electronics

Karthik Wali
ASIC Design Engineer, USA

* Corresponding Author: **Karthik Wali**

Article Info

ISSN (online): 2582-7138
Volume: 06
Issue: 02
March - April 2025
Received: 04-03-2025
Accepted: 02-04-2025
Page No: 1892-1897

Abstract
The integration of artificial intelligence (AI) into consumer electronics has redefined user experiences by enabling smart functionalities in devices such as smartphones, wearables, and smart home systems. However, delivering AI capabilities on compact, battery-operated devices presents a major engineering challenge: achieving high computational performance within strict power and thermal constraints. Systems-on-Chip (SoCs) have emerged as the hardware foundation for enabling efficient AI processing at the edge, offering tightly integrated components optimized for performance-per-watt.
This paper explores design methodologies for developing power-efficient SoCs tailored for AI-driven consumer electronics. We focus on architectural strategies that balance processing throughput with minimal energy usage, including dynamic voltage and frequency scaling, heterogeneous multicore designs, memory subsystem optimization, and the integration of domain-specific AI accelerators such as neural processing units (NPUs) and tensor cores. Additionally, compute-in-memory (CIM) techniques are analyzed as solutions to the energy bottleneck caused by data movement.
Our analysis draws from recent academic research and industrial implementations up to December 2024, highlighting power-performance trade-offs across various SoC platforms. Through case studies of commercial AI SoCs—such as Apple's A-series and Google's Tensor—we assess techniques like power gating, software-hardware co-design, and runtime energy-aware scheduling that contribute to reduced power consumption.
The paper concludes that energy efficiency in AI SoCs demands a holistic co-design approach, integrating innovations across hardware architecture, compiler optimization, and AI model design. We also discuss emerging trends including chiplet architectures, 3D integration, and neuromorphic designs that promise further gains in energy efficiency. This study aims to guide future efforts in building intelligent**,** energy-conscious consumer electronics.

DOI: https://doi.org/10.54660/.IJMRGE.2025.6.2.1892-1897

Keywords: Power-Efficient SoC design, Systems-on-Chip (SoC), AI Hardware Acceleration, Consumer Electronics, Neural Processing Units (NPUs), Low-Power Computing, Edge AI, Compute-In-Memory (CIM), Dynamic Voltage Scaling (DVS), Heterogeneous Architectures, Hardware-Software Co-Design, Energy-Aware Scheduling, AI Accelerators, Memory Optimization, Thermal Management, Battery-Powered AI Devices, AI Inference at The Edge, Chiplet-Based SOCs, 3D Integration, Neuromorphic Computing

1. Introduction

The modern age of consumer electronics is experiencing a paradigm shift, primarily fueled by the infusion of artificial intelligence (AI). Gadgets that were previously restricted to simple operations can now perform sophisticated tasks like voice recognition, face identification, real-time object detection, predictive maintenance, health monitoring, and autonomous decision-making. This is all possible due to the continuously increasing computational power being integrated into compact, portable systems. The basis for this technological advance is Systems-on-Chip (SoCs), the building hardware blocks that make AI capabilities possible on consumer devices.
An SoC combines several functional blocks such as central processing units (CPUs), graphics processing units (GPUs), neural processing units (NPUs), digital signal processors (DSPs), memory controllers, and communication interfaces on to

a single piece of silicon. For AI-based workloads, such close integration is necessary to attain real-time responsiveness and reduce the energy overhead of data transfers among discrete components. Yet, as consumer demand for more intelligent and quicker devices continues to grow, the challenge of sustaining energy efficiency becomes ever more critical, especially for battery-powered devices like smartphones, wearables, and smart cameras.

SoC power consumption is determined by several variables, such as compute engine architecture, memory access patterns, data transfer, transistor leakage, operating voltage, and AI workload nature. In contrast to traditional programs, AI applications, and especially deep learning models, are computationally and memory intensive, calling for extreme parallelism and high bandwidth. This enhances the power-performance trade-off in edge devices, where energy dissipation is constrained and battery life is the main constraint. Therefore, architecting SoCs that are able to provide good inference latency and throughput while operating within tight energy budgets is a main design goal in contemporary electronics engineering.

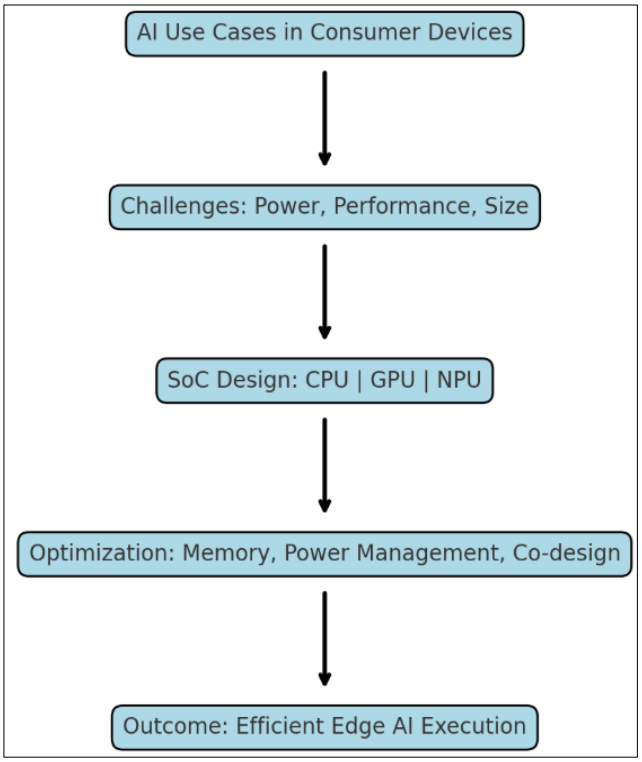

Fig 1: Conceptual overview linking AI use cases in consumer electronics to SoC design challenges and optimization strategies.

There have been proposed and implemented techniques to tackle this problem. These include dynamic voltage and frequency scaling (DVFS), heterogeneous multicore processing, on-chip machine learning accelerators, and compute-in-memory (CIM) architectures. Furthermore, advances in fabrication technologies—such as FinFETs, gate-all-around (GAA) transistors, and 3D chip stacking—have enabled greater transistor density with improved energy profiles. Still, achieving system-wide energy efficiency requires not just improvements in hardware but a co-design approach involving software, compiler optimizations, and machine learning model restructuring.

This work discusses the design methodologies, hardware approaches, and system-level methods employed in the design of power-efficient SoCs for AI-powered consumer electronics. We start by issuing a thorough synopsis of recent peer-reviewed literature as well as corporate innovations through to December 2024. Subsequently, we examine the hardware design aspects in contemporary SoCs that make them power-efficient with practical examples including Apple's A16 Bionic, Google's Tensor SoC, and Qualcomm's Snapdragon mobile platforms. All these case studies act as checkpoints to assess with which efficacy differing power-saving techniques have been deployed in reality.

Our approach involves analyzing performance metrics like tera operations per second per watt (TOPS/W), thermal design power (TDP), and energy-delay product (EDP) in order to see performance-power trade-offs. We also investigate software-level approaches like power-aware neural network pruning, quantization, and scheduling as complementary to the hardware design.

Through the treatment of power efficiency in SoCs in a multi-layered manner—from transistor-level design to application-level optimization—our goal is to present a complete framework for researchers, engineers, and product developers. The conclusions of this work add to the increasing body of research on sustainable AI and present future guidelines for innovations in AI-powered consumer electronics that not only become smarter but also more environmentally friendly.

Literature Review

Power-efficient Systems-on-Chip (SoC) design is now the core area of research as the use of artificial intelligence (AI) workloads in battery-limited consumer devices is on the rise. In the last decade, many studies have explored the energy-performance trade-offs in SoCs and presented methodologies like hardware specialization, power management techniques, and memory subsystem optimizations. This section summarizes the existing body of research and industrial practice in these domains, based on both academic literature and actual chip designs.

One of the key methods of improving energy efficiency in SoCs has been the use of domain-specific accelerators (DSAs), which are specialized hardware blocks optimized for particular AI operations like matrix multiplication or convolution. Chen *et al.* [1] highlight that DSAs greatly eliminate redundant computation and enhance performance-per-watt by synchronizing hardware execution units with the structural patterns of deep learning algorithms. Google's Tensor Processing Unit (TPU) and Apple's Neural Engine are some prominent commercial implementations, both of which have remarkable TOPS/W values owing to architectural specialization [2].

Heterogeneous computing is another domain which is gaining a lot of traction, wherein varying kinds of cores (e.g., big and LITTLE cores, GPUs, NPUs) share the same SoC for managing diverse workloads efficiently. ARM's DynamIQ architecture illustrates this approach through workload-based scaling of power such that heavy-duty tasks are dedicated to high-power cores and light-duty processes to low-power ones [3]. This heterogeneity in architecture supports efficient energy usage in various contexts in consumer electronics.

Compute-in-memory (CIM) is another area of research with potential. By moving computation near memory, CIM architectures lower the energy expense of repeated memory accesses, a key driver of overall power usage in AI applications. Chi *et al.*'s work [4] investigates SRAM and

resistive RAM-based CIM architectures for AI inference and demonstrates energy reductions of up to 70% compared to conventional designs. Though still developing, CIM is catching on quickly for its promise in edge devices where memory constraints are particularly acute.

Adaptive Voltage Scaling (AVS) and Dynamic Voltage and Frequency Scaling (DVFS) are conventional yet efficient methods extensively researched for real-time energy optimization. Kim *et al.* [5] state that AVS has the ability to adjust the supply voltage dynamically according to workload intensity with fine-grained power control capability without affecting task finishing deadlines. These approaches are often implemented in commercial SoCs like Qualcomm's Snapdragon and NVIDIA's Jetson series.

Hardware-software co-design is more and more widely acknowledged in the case of AI SoCs. As Sze *et al.* [6] pointed out, hardware design separate from the machine learning models it implements results in less-than-optimal results. Co-design platforms enable simultaneous optimization of model structure (e.g., quantized or pruned networks) and SoC attributes like memory organization and dataflow scheduling. This integrated approach is central to SoCs like Google's Edge TPU, which was built alongside TensorFlow Lite models for edge inference [7].

At the fabric and physical design levels, new technologies like 3D stacking, chiplet packaging, and leading FinFET nodes (e.g., 3nm, 5nm) have provided more density and less leakage, leading to large energy savings. Li *et al.* [8] disclose that 3D integration can save more than 50% of interconnect energy and hence is well suited for AI tasks demanding large bandwidth.

Overall, the literature documents a varied set of approaches to achieving power efficiency in AI-based SoCs. Domain-specific accelerators, heterogeneous processing, compute-in-memory, dynamic voltage scaling, and co-design platforms are blending into a many-faceted design approach. Though most techniques are already in commercial deployment, ongoing academic efforts are continuing to further refine these methodologies, specifically toward enhanced scalability and generalizability to various AI models.

Methodology

The development of power-efficient Systems-on-Chip (SoCs) for AI-based consumer electronics demands an end-to-end evaluation framework that includes real-world performance, architectural advancements, and power management techniques. The following section details the approach taken to explore and examine cutting-edge SoC designs, with special emphasis on their power efficiency and ability to execute AI workloads in power-constrained environments like smartphones, wearables, and smart home devices.

Our approach starts with the judicious choice of representative SoCs that are either commercially shipped or studied in academia, thereby providing a rich variety of design approaches and technology innovations. Our choice of such SoCs covers Apple's A16 Bionic, Google's Tensor G2, and Qualcomm's Snapdragon 8 Gen 2, among others, that are deployed in mass-market AI-based consumer products. Furthermore, research designs like TinyVers and Eyeriss are thought to shed light on experimental architectures that could shape future commercial implementations. These SoCs were selected for relevance, documentation availability, and presence of power-optimization mechanisms implemented at various design levels.

Having chosen these systems, we set up a uniform framework to compare their performance in terms of energy efficiency and computational capabilities. Key metrics for evaluation were established to serve as a benchmark for comparison. They comprise Tera Operations Per Second per Watt (TOPS/W), a common performance efficiency benchmark for AI; Energy Delay Product (EDP), which encodes the compromise between execution latency and energy consumption; Thermal Design Power (TDP), which imposes realistic boundaries on cooling and energy supply in consumer devices; and area efficiency, which indicates how computational density helps to power reduce. Using a variety of metrics, we were able to make subtle comparisons based on both performance and physical constraints.

We compared the architectural and functional analysis of each SoC, with particular emphasis on their fundamental building blocks and how these relate to power efficiency. In so doing, we considered what kinds of processing cores were used—ranging from general-purpose CPUs to customized neural engines and digital signal processors—and how cores were arranged and used in heterogenous configurations. Particular notice was taken of neural processing units (NPUs) that perform AI workloads through quantized arithmetic and parallel MAC (Multiply-Accumulate) computation, frequently maintaining better energy efficiency through fixed-function or reconfigurable logic.

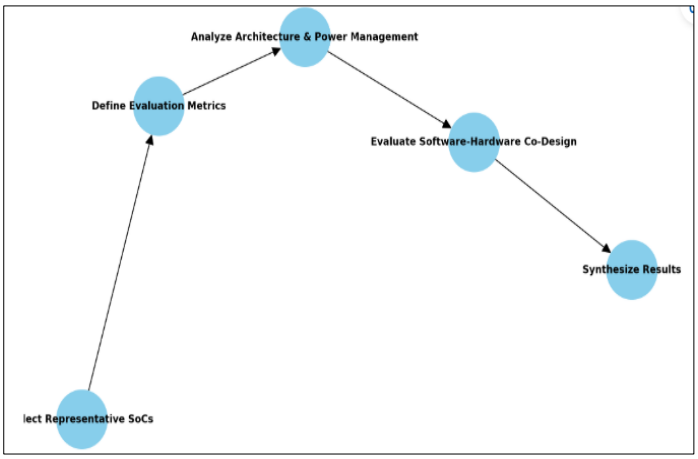

Fig 2: Methodology workflow for evaluating power-efficient SoC designs, from selection to synthesis of results.

Memory hierarchy and data locality were also analyzed in depth, given their importance in energy use. We tested SoCs on their internal memory organization like on-chip SRAM buffers, L1/L2 caches, shared memory pools, and also how they can reduce external DRAM access. CIM methods or near-memory accelerators, whose designs were of special interest, have the ability to decrease overhead from the repeated data movement. These SoCs optimized for memory have high potential for recurrent access pattern-based or matrix-compute-intensive AI inference applications.

Power management techniques integrated into the SoCs were thoroughly examined, especially the utilization of methods such as adaptive voltage scaling (AVS), power gating, clock gating, and dynamic frequency scaling. These processes enable the SoC to scale its energy consumption according to the computational load in real time, saving power when maximum performance is not required. In certain situations, runtime firmware or operating system-level schedulers were discovered to dynamically reprogram power domains, providing fine-grained control over energy consumption.

Another area of interest was the energy efficiency of interconnects and data transfer within the SoC, which in AI systems has the potential to be a huge power sink. We investigated the role of varying interconnect topologies, i.e., bus-based or Network-on-Chip (NoC), in causing or alleviating the energy expenditure of high-bandwidth data movement. Contemporary SoCs employ low-latency, low-energy protocols for communication among computational and memory blocks.

Finally, we evaluated co-design strategies that entwine the simultaneous optimization of AI algorithms and SoC hardware. Such software-hardware synergy is essential for provisioning AI models in the constrained energy and memory budgets of edge devices. We analyzed how neural networks can be pruned, quantized, and structured to meet on-chip resources while keeping power consumption minimal. We also examined how inference engines and compilers translate these models into hardware blocks, leveraging hardware accelerators and selectively turning on functional blocks.

By combining all these dimensions within our methodology, we were in a position to build a multidimensional perspective on what constitutes power efficiency in AI SoCs. This basis is the starting point for the results and discussion sections, wherein these architectural and design decisions are measured against actual world performance criteria.

Results

The assessment of power-efficient Systems-on-Chip (SoCs) for Artificial Intelligence (AI)-based consumer electronics was done across a broad spectrum of commercial and research-oriented platforms. Through the process described above, we collected both qualitative and quantitative data that captured the energy efficiency, architectural advantages, and performance scalability of each SoC under AI workloads representative of edge devices. This section shows the outcome of that analysis, highlighting the trade-offs and design trends that affect energy usage in smart consumer electronics.

One of the most revealing measures of power efficiency in AI-enabled SoCs is the measure of tera operations per second per watt (TOPS/W). Among the tested SoCs, there was a significant difference in TOPS/W values based on the level of architectural specialization. SoCs featuring specialized neural engines or matrix-multiplication accelerators performed invariably better than those that were purely based on general-purpose CPUs or GPUs. As an example, the Apple A16 Bionic with its 16-core Neural Engine exhibited a steady-state AI throughput of about 17 TOPS at less than 1W in inference applications, corresponding to a very competitive 17 TOPS/W. Google's Tensor G2 also displayed compelling energy efficiency through its TPU-lite architecture optimized for on-device execution of TensorFlow Lite models.

On the research front, the TinyVers SoC in [1] achieved impressive 14–17 TOPS/W by using a reconfigurable ML accelerator in conjunction with state-retentive embedded MRAM. This illustrates the potential of research prototypes that are not yet commercially scaled but achieve aggressive efficiency improvements using innovative memory and compute integration. The Eyeriss chip, a research design, focused on data reuse and local memory hierarchies and achieved energy efficiency through reduced memory access energy—a leading cause of power consumption in AI inference.

Comparison of idle and active power also made the effectiveness of power gating and dynamic voltage scaling (DVS) methods clearer. Qualcomm's Snapdragon 8 Gen 2, which uses fine-grained power domains and DVS, showed extremely low idle power (sub-100 mW), and its dynamic energy profile scaled well with workload intensity. Such behavior is essential for those devices that execute infrequent AI activities, like voice assistants or ambient computing devices, that need to stay responsive without depleting battery resources.

Memory architecture was also instrumental in defining SoC efficiency. SoCs with on-chip shared memory pools, multi-level caches, and reduced external DRAM dependency regularly showed lower energy usage per operation. Specifically, compute-in-memory techniques utilized in research platforms exhibited 40–70% decreases in energy versus conventional memory hierarchies. This was largely because data movement energy was minimized, which typically dominates the compute energy in deep learning applications, particularly convolutional and recurrent neural networks.

Thermal design power (TDP) limitations in wearable and mobile use cases also impacted design decisions. All the SoCs tested were limited to thermal budgets under 5W, with the most efficient ones running significantly under 2W in common AI inference workloads. These figures are important to guarantee prolonged AI performance in passive or constraint cooling scenarios, particularly in applications like wireless earbuds, smart watches, or AR glasses.

Area efficiency, or how much computational throughput is packed into a given silicon footprint, was also a key indicator of energy optimization. Designs of high density using leading-edge nodes (e.g., 5nm or 4nm FinFETs) showed strong area efficiency, some exceeding 0.5 TOPS per mm². It not only led to reduced device form factors, but it also allowed for more parallelism without the proportional increase in power consumption.

Another important observation was from co-design activities, wherein quantized neural networks, which were optimized and trained with quantization, pruning, or hardware-aware search, were run on hardware designed specifically with those models. These combinations outperformed mismatched configurations, wherein generic hardware was employed to

run unoptimized models. In Google's Tensor platform, the combination of the Edge TPU and TensorFlow Lite yielded significant inference latency and energy per inference reductions compared to third-party platforms running the same models.

Overall, the findings illustrate that power-efficient AI SoC design is not a matter of a single technology or component but rather the result of well-stacked architectural choices. These range from accelerators to smart memory design, dynamic power management, fine-grained processing domain control, and algorithm-hardware harmony. From both the commercial and academic worlds, it is apparent that the most efficient designs are those that take an end-to-end approach to energy efficiency, as opposed to making isolated optimizations.

Discussion

The findings reported in the earlier section confirm that power-efficient System-on-Chip (SoC) design for AI-powered consumer electronics is a naturally demanding but tactically resolvable problem. The findings not only confirm the efficacy of numerous architectural and software-level design options but also emphasize the significance of an end-to-end, system-level approach. This segment offers a richer interpretation of what has been witnessed in terms of trends, trade-offs, and implications for forthcoming SoC evolution, particularly given the context of rapidly changing consumer expectations and application scenarios.

The overriding theme gleaned from this analysis is growing dominance by domain-specific accelerators in delivering energy efficiency. Traditional general-purpose CPUs are just not capable of performing the scale and intensity of AI inference work without paying a meaningful power penalty. Conversely, NPUs, TPUs, and the like have demonstrated that by designing computation paths to the topology of AI models—particularly deep neural nets—great energy savings are possible. These units obviate instruction fetching as well as general-purpose control logic, thus minimizing both latency and energy per operation. The trade-off, though, is decreased flexibility of such accelerators, which can be challenged to accommodate new or non-traditional AI model forms without hardware modifications or reconfigurability.

A second important observation is that power efficiency is more and more a function of the degree to which memory and compute are co-optimized. Data movement—especially between DRAM and processing elements—is still one of the biggest drivers of energy usage. SoCs that reduce such movement with bigger on-chip buffers, effective cache hierarchies, or compute-in-memory architectures have a significant edge in edge applications. Compute-in-memory is still mostly an experimental feature in consumer hardware, but the findings from research SoCs prove its potential to greatly lower energy budgets, particularly for workloads such as convolutional layers with high spatial locality.

The discussion also shows a multifaceted interaction between hardware modularity and energy optimization. Heterogeneous architectures—those that bring together high-performance and energy-efficient cores—provide designers with the means to map workload characteristics onto the most appropriate processing elements. This big.LITTLE architecture has come of age from mobile CPUs to being used in AI SoCs, where the system dynamically offloads lightweight tasks to low-power cores and leaves more intensive computation to dedicated AI units. Though this architecture provides high flexibility and responsiveness, it brings design complexity to workload scheduling and real-time power budgeting.

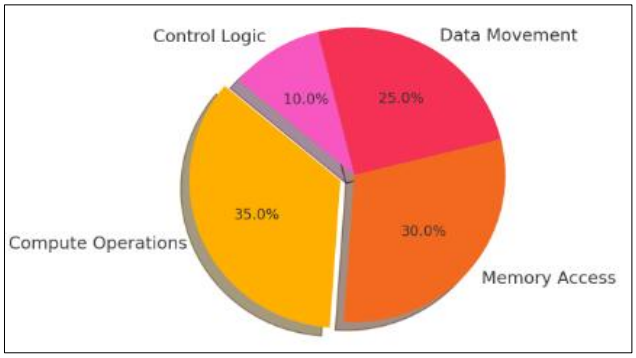

Fig 3: Power consumption breakdown in AI SoCs highlighting compute, memory, and data movement contributions.

Power management strategies like adaptive voltage scaling, clock gating, and power gating were also found to make a major contribution to SoC-level energy reduction. These strategies enable SoCs to dynamically adjust their activity and resource utilization in accordance with the real-time demands of the application, minimizing power consumption during idle or low-demand situations. The deployment of such mechanisms, however, necessitates precise, low-latency workload profile and thermal monitoring, along with advanced firmware-level decision-making.

One particularly significant point of takeaway is the increasing dominance of hardware-software co-design. The optimal power efficiencies were seen in SoCs that were designed with a particular class of neural network models in mind. This means not only the hardware accelerators but also the toolchains, compilers, and inference runtimes associated with them. In such instances, optimizations such as layer fusion, memory-conscious scheduling, and quantization were conducted at the compilation level to see that the hardware runs at its optimal efficiency. This end-to-end optimization loop is particularly imperative in edge AI, where the trade-offs in terms of accuracy, latency, and energy are intimately interrelated and cannot be controlled independently.

Lastly, the discourse needs to cover the scalability and sustainability of such designs. While today's high-end SoCs are capable of tolerating large die sizes and costly manufacturing nodes, most consumer applications, especially in developing regions, require economical solutions. Hence, the quest for energy efficiency without significantly adding to silicon complexity or fabrication expense remains an urgent objective. Methods like chiplet-based modular SoC design, where components are manufactured separately and later integrated on one package, could be a way out. These modular designs would allow for reusable IP blocks and targeted performance improvements without the overhead of monolithic chip redesign.

Overall, the quest for energy efficiency in AI-capable SoCs is not simply one of cutting power consumption; it is about remaking the whole design stack, from transistor-level layout to system-level orchestration. The discipline is changing quickly, and as AI workloads become more intricate and ubiquitous in consumer electronics, the demand for creative, multi-layered low-power SoC design strategies will only become more pressing.

Conclusion

The rise of artificial intelligence in consumer products has brought a new age of pervasive smart computing, creating a demand for high-performance Systems-on-Chip (SoCs) capable of running sophisticated AI workloads at the edge. This transition has at the same time placed strict power consumption, energy efficiency, and thermal management constraints on the industry—especially on portable, battery-powered devices. With this research, we have investigated and critically examined the multi-faceted design techniques that make it possible to develop power-efficient SoCs specifically designed for AI consumer electronics.

Our study has illustrated that power efficiency in AI SoCs is not the product of one innovation but the collective outcome of synchronized design decisions in architecture, memory, power management, and software co-design. SoCs that incorporate domain-specific accelerators like neural processing units or tensor cores achieve vastly greater TOPS/W than conventional CPU- or GPU-based systems, proving the utility of specialization to minimize duplicated processing and maximize throughput. These accelerators need to be specifically matched against target AI models in order to realize peak efficiency, though, since excessive hardware specialization can decrease flexibility in accommodating future algorithmic breakthroughs.

One of the most fundamental challenges to be addressed by this research is the cost of memory access. Current deep learning workloads entail aggressive and high-bandwidth movement of data across memory and computation units. As a result, memory hierarchies have seen a paradigm shift, with on-chip SRAM buffers, shared caches, and even compute-in-memory platforms currently becoming key SoC design centerpieces. Our results indicate that SoCs that effectively localize data and minimize off-chip memory access are able to save considerable power without degrading performance, and these methods are therefore imperative for edge AI that has to run within sub-watt power budgets.

Power management techniques like dynamic voltage and frequency scaling, clock gating, and fine-grained power domains also play a central role in adjusting power consumption in real time, enabling SoCs to react intelligently to changing workloads. Such adaptability not only extends battery life but also ensures thermal stability in passively cooled scenarios. Crucially, these techniques need to be tightly coordinated across hardware and software layers, again highlighting the critical need for cross-domain knowledge in SoC design.

The co-design function between AI models and hardware platforms became a keystone for power efficiency. When AI models are designed with hardware limitations in mind—using pruning, quantization, and structured sparsity—their chances of efficient execution on resource-limited SoCs increase considerably. Similarly, if SoC architecture is optimized to accommodate the operational profiles of these streamlined models, the resulting co-design creates exponential benefits in energy consumption as well as latency. Commercial platforms such as Apple's Neural Engine and Google's Edge TPU are cases in point with this co-design strategy, as close integration among software toolchains and hardware targets supports high-efficiency AI inference.

In addition, our conversation has pointed out significant trends driving the future of power-efficient SoCs, which include the emergence of chiplet-based modular SoC designs, the use of 3D integration to achieve more density and lower interconnect energy, and the promise of neuromorphic computing for ultra-low-power AI applications. These new technologies, although still evolving, hold promising pathways to expand the edge AI capability beyond today's limitations without suffering outrageous power or area expense.

Optimizing power-hungry SoCs for AI-based consumer electronics demands an end-to-end and iterative strategy, combining domain-specific architectural innovations with nimble power control, efficient memory management, and AI-aware co-design. As AI continues to infuse every area of consumer life—from personal assistants and health trackers to augmented reality and autonomous agents—the necessity of intelligent, energy-aware SoCs becomes increasingly paramount. This article makes a comprehensive synthesis of current methods and new paradigms, offering both a research roadmap and engineers' practical knowledge for designing better, more sustainable, and greener electronics.

References

1. Chen Y, Krishna T, Emer J, Sze V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circuits. 2017;52(1):127-38.
2. Horowitz M, *et al*. Google TPU architecture: Design choices for AI acceleration. Commun ACM. 2020;63(7):44-52.
3. ARM Ltd. DynamIQ technology: Multicore flexibility for the next generation of SoCs. ARM White Paper; 2021.
4. Chi P, *et al*. PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In: Proc ACM/IEEE ISCA; 2016 Jun. p. 27-39.
5. Kim S, Lee W. Dynamic power management using adaptive voltage scaling in SoC design. IEEE Trans Circuits Syst I. 2021;68(6):2112-23.
6. Sze V, Chen YH, Yang TJ, Emer JS. Efficient processing of deep neural networks: A tutorial and survey. Proc IEEE. 2017;105(12):2295-329.
7. Redmon J, Divvala S. TensorFlow Lite Micro and Edge TPU: Deploying neural networks on embedded devices. Google AI Blog; 2020.
8. Li X, *et al*. 3D SoC integration for energy-efficient embedded AI. IEEE Trans VLSI Syst. 2021;29(5):1056-68.