Global Multidisciplinary Journal
eISSN: 2791-173X pISSN: 2791-2760

VOLUME04 ISSUE10
Published 31-10-2025 *Page No. 84-92*

Heterogeneous GPU Architectures, Energy-Aware Thermal Management, and Validation Strategies for Next-Generation High-Performance Computing

Dr. Rafael M. Cortez

Global Institute of Computational Engineering, University of Lisbon

ABSTRACT

This article synthesizes theoretical perspectives and applied insights drawn from a curated set of contemporary and foundational works to present an integrative, publication-ready examination of graphics processing units (GPUs) as central engines of modern high-performance computing (HPC), machine learning, and real-time multimedia systems. We develop a coherent narrative that traces GPU evolution and architectural principles, explores GPU programming models and their implications for large-scale data mining and accelerated computing, interrogates energy, power, and thermal management across device-to-application layers, and details validation and manufacturing strategies for acoustic and thermal integrity. Methodologically, the work adopts a cross-disciplinary descriptive synthesis grounded in primary references, combining architectural analysis, systems-level power and thermal modeling concepts, and process- and design-oriented validation approaches. Results are presented as a rich descriptive analysis that elucidates (1) how architectural choices have shaped parallel programming paradigms and application performance, (2) the complex trade-offs between performance, energy consumption, and thermal constraints in GPU-centric systems, (3) mechanisms for integrated CPU–GPU power management in constrained environments such as mobile gaming, and (4) scalable acoustic and thermal validation strategies needed in modern GPU manufacturing. We interpret these findings to argue for a layered, co-designed approach that couples architectural innovations (including 3-D integration and GPU-in-memory concepts) with machine-learning-aided power/thermal management and scalable manufacturing validation. The discussion highlights limitations of current approaches—particularly the challenges in generalizing thermal models across heterogeneous stacks and the nascent state of AI-driven thermal control for GPUs—and proposes a future research agenda that emphasizes co-design, domain-specific cooling techniques, hardware/software power coordination, and standardized validation pipelines. This integrative treatment aims to inform researchers, system designers, and manufacturing engineers seeking to align GPU architecture, system-level energy efficiency, and robust validation practices in the era of AI-scale computing.

KEYWORDS

GPU architecture; energy efficiency; thermal management; GPU validation; heterogeneous computing; GPU-in-memory

INTRODUCTION

Over the past two decades the graphics processing unit (GPU) has migrated from a narrowly defined role as a fixed-function graphics accelerator to a versatile, massively parallel compute engine underpinning diverse

computational workloads including scientific simulation, data mining, machine learning, cloud gaming, and cryptocurrency mining (Dally, Keckler & Kirk, 2021; Peddie, 2023). This trajectory reflects an iterative co-evolution of hardware design, programming models, and application demand: hardware vendors generalized SIMD and SIMT execution models and exposed programmable shader and compute pipelines (Moya et al., 2005; Kirk, 2007), while software frameworks like CUDA popularized general-purpose GPU programming and catalyzed the porting of dense linear algebra, image processing, and machine learning kernels to GPUs (Kirk, 2007). The performance dividends of such specialization are now being reframed by a second-order set of constraints: energy consumption, thermal dissipation, acoustic emissions, and manufacturing-validation complexity (Shankar, 2023; Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025). These constraints arise because GPUs, by virtue of their high compute density and memory bandwidth, concentrate power and heat into localized die regions and package volumes that are increasingly heterogeneous (e.g., 3-D stacked dies, GPU-in-memory) (Kurshan & Franzon, 2025; Wen, Yang & Zhang, 2017). Consequently, contemporary GPU system design requires an integrated view that couples architecture, energy-aware runtime management, cooling technologies, and validation methodologies.

The problem statement this article addresses is multifold. First, while the literature has exhaustively characterized GPU architectural evolution and application acceleration (Dally et al., 2021; Peddie, 2023; Cano, 2018), there remains a need for comprehensive synthesis that explicitly connects architecture to energy/thermal behavior, and from there to validation strategies at manufacturing scale. Second, emerging packaging and integration approaches—such as 3-D stacking and heterogeneous die assembly—change heat-flow paradigms and complicate established validation and acoustic/thermal testing regimes (Kurshan & Franzon, 2025; Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025). Third, system-level power management techniques that must coordinate CPU and GPU operations—particularly in constrained devices like mobile systems and cloud gaming servers—are in active development but require clearer guidance on trade-offs, algorithms, and evaluation approaches (Pathania et al., 2014; Hou et al., 2014; Mills & Mills, 2016). Finally, machine-learning-driven power and thermal management methods are promising but need to be positioned relative to traditional model-based controls and practical constraints in deployment (Pagani et al., 2020; Shankar, 2023). The literature gap, therefore, is an integrated, theory-rich, and practically oriented treatment that spans architecture, runtime power control, cooling design, and manufacturing validation.

This article responds by offering a descriptive and analytic synthesis based strictly on the provided references, mapping connections between GPU architecture and energy/thermal phenomena, and articulating validation strategies that manufacturers and system designers can use to ensure acoustic and thermal compliance at scale. Rather than presenting new empirical data, the contribution lies in the comprehensive conceptual integration, critical analysis of trade-offs, and explicit direction for future research and engineering practice. The approach taken deliberately avoids summarization in favor of in-depth theoretical elaboration: each architectural and systems-level point is expanded with implications, counter-arguments, and nuanced interpretation in order to provide a publication-ready, extended treatment for academic and engineering audiences.

METHODOLOGY

This study follows a structured, text-based integrative methodology designed for theoretical synthesis and critical interpretation. The method prioritizes rigorous, citation-anchored reasoning and layered analysis rather than empirical experimentation. The main steps are described below.

Literature selection and scope. The analysis is strictly confined to the references provided by the requestor.

These sources span foundational GPU histories and architectural surveys (Dally et al., 2021; Peddie, 2023; Peddie, 2023—chapter), architectural performance studies (Moya et al., 2005; Kirk, 2007), applications in data mining and cloud gaming (Cano, 2018; Hou et al., 2014), energy-focused analyses across layers (Shankar, 2023; Pagani et al., 2020; Mills & Mills, 2016), integrated power management approaches (Pathania et al., 2014), and emerging packaging/thermal research including 3-D stacking and GPU-in-memory (Kurshan & Franzon, 2025; Wen et al., 2017). Manufacturing validation perspectives are provided by a 2025 synthesis of scalable acoustic and thermal validation strategies (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025) and domain-specific cooling structure advances (Wang et al., 2022). Additional cross-domain reference points include memristor-based neuromorphic devices (Wu et al., 2024) which inform discussion of future heterogeneous co-design, and thermal/cooling advances (H. Wang et al., 2022). The provided corpus thus supports a layered exploration across architecture, software, energy/thermal management, and manufacturing.

Analytic synthesis. Each major thematic axis—GPU architecture and programming models; energy/power/thermal interactions; run-time power and scheduling policies; cooling and structural design; manufacturing validation strategies—was analyzed by extracting the claims, design principles, and empirical insights contained in the references, then elaborating the implications and interactions among them. Major claims from the literature are cited inline and then expanded with layered argumentation: theoretical ramifications, potential counterpoints, and the practical consequences for designers and manufacturers.

Trade-off analysis. For each system-level decision point recognized in the literature (e.g., specialization vs. generality, performance vs. energy efficiency, 2-D vs. 3-D integration), the analysis identifies the design variables, articulates how they interact, and reasons about optimization spaces. Citations are provided for the empirical or conceptual bases of these interactions (Dally et al., 2021; Pagan i et al., 2020; Wen et al., 2017).

Validation and manufacturing pipeline mapping. The article synthesizes manufacturing-scale validation strategies by mapping processes—from silicon-level thermal characterization, package-level acoustic testing, system-level thermal profiling, to end-of-line manufacturing checks—drawing on the 2025 validation strategies reference and on energy/workload insights from Shankar (2023) and Pagani et al. (2020). This section is descriptive, emphasizing process design decisions and scalable methodologies.

Limitations of method. The methodology explicitly avoids primary experimental data collection or simulation not present in the referenced corpus. Instead, it focuses on rigorous conceptual integration. This choice enables a deep theoretical elaboration but limits the work's ability to validate claims empirically beyond what the references provide.

Ethical and reproducibility considerations. All claims are anchored to the reference set and are presented transparently with explicit citations. Because the work is synthetic rather than empirical, reproducibility is a matter of repeated literature analysis rather than experimental replication.

RESULTS

The results are presented as a rich, descriptive analysis organized around four principal findings that emerge from integrating the references: (1) the architectural lineage and its implications for parallelism and programmability; (2) energy and thermal dynamics across system layers; (3) runtime power/thermal management approaches and their trade-offs; and (4) scalable manufacturing validation strategies for acoustic and thermal integrity.

1. Architectural lineage, programmability, and implication for compute density

GPU architecture has evolved from fixed-function graphics to general-purpose compute through successive

design innovations: programmable shaders, SIMT execution, wide vector pipelines, and specialized memory hierarchies (Moya et al., 2005; Kirk, 2007; Dally et al., 2021; Peddie, 2023). The net effect of these innovations is a hardware substrate optimized for throughput-oriented, data-parallel tasks. GPUs provide large numbers of relatively simple cores, rapid context switching within warp/wavefront units, and high-bandwidth memory interfaces that together maximize arithmetic intensity for amenable kernels (Dally et al., 2021; Cano, 2018). This compute density creates both opportunity and constraint: it enables exceptional throughput for tasks like matrix multiplication and convolutional kernels common in machine learning, but it also localizes power dissipation and raises cooling requirements (Dally et al., 2021; Shankar, 2023).

A critical implication is that the programming model—embodied in CUDA, OpenCL, and shader languages—must expose sufficient parallelism and locality while giving runtime systems control over resource allocation (Kirk, 2007; Cano, 2018). When programmers and compilers can map computation to GPU memory hierarchies effectively, the architecture's bandwidth and parallelism produce substantial accelerations (Moya et al., 2005). Conversely, poor mapping increases off-chip memory traffic and energy costs, diluting throughput gains (Cano, 2018).

2. Energy and thermal dynamics across layers: device to application

Energy consumption in GPU systems is a cross-layer phenomenon requiring coordinated understanding from transistor-level switching to workload scheduling (Shankar, 2023; Pagani et al., 2020). At the device level, switching activity, leakage currents, and the physical characteristics of interconnects dictate baseline power envelopes (Shankar, 2023). At the microarchitectural level, many-core parallelism intensifies instantaneous power draw when large groups of cores execute simultaneously; memory controllers and memory channels similarly concentrate energy in interfaces. At the system and application layers, algorithmic choices (e.g., precision reduction, operator fusion) influence arithmetic intensity and memory access patterns, thereby affecting energy consumption (Cano, 2018; Shankar, 2023).

Thermally, the die's power density determines local temperature gradients that influence reliability (electromigration, threshold shifts) and performance (frequency throttling, voltage droop) (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025; Wang et al., 2022). Packaging decisions—2-D vs. 3-D stacking, presence of thermal vias, integrated heat-spreaders—substantially change heat-flow pathways (Kurshan & Franzon, 2025; Wen et al., 2017). Notably, 3-D integration, while promising for reduced interconnect latency and higher integration density, complicates cooling because stacked dies introduce internal heat sources that do not have direct exposures to ambient heat sinks (Kurshan & Franzon, 2025). This reality motivates structural cooling innovations and co-design with software power scheduling.

3. Runtime power/thermal management: policies and trade-offs

Runtime strategies to manage power and thermal constraints occupy a spectrum from conservative, rule-based throttling to predictive, learning-based controllers (Pathania et al., 2014; Pagani et al., 2020; Shankar, 2023). Integrated CPU–GPU power management is particularly important for applications requiring tight coordination between the host and accelerator, such as 3-D mobile gaming where battery and thermal envelopes are strict (Pathania et al., 2014; Hou et al., 2014). Techniques include dynamic voltage and frequency scaling (DVFS) for both CPU and GPU domains, workload redistribution, kernel fusion to reduce memory traffic, and frame-rate regulation in gaming contexts to balance user-perceived performance against energy consumption (Pathania et

al., 2014; Hou et al., 2014; Mills & Mills, 2016).

Learning-based approaches introduce the potential for more granular, anticipatory control by predicting thermal trajectories and adjusting operating points preemptively (Pagani et al., 2020; Shankar, 2023). The advantage is improved responsiveness and potential energy savings; the counterpoint is the complexity and potential brittleness of models that must generalize across workloads and hardware variations. Furthermore, while machine-learning-based controllers can reduce energy and avoid thermal violations under sampled workloads, they introduce new validation challenges: controllers themselves must be validated to behave safely under adversarial or unobserved workloads (Pagani et al., 2020).

4. Scalable manufacturing validation: acoustic and thermal strategies

Manufacturing-scale validation of acoustic and thermal properties requires pipelines that can handle high throughput while maintaining sensitivity to subtle deviations. The 2025 study on scalable acoustic and thermal validation identifies key components of a robust validation pipeline: reproducible thermal benchmarking (standardized workloads and sensors), package-level acoustic characterization, statistical process control to detect drift, and end-of-line testing procedures that are non-destructive and time-efficient (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025). Thermal validation must account for variability induced by packaging tolerances, coolant attachment quality, and assembly-induced mechanical stress; acoustic validation must capture fan-induced tonalities and flow noise that can impact product perception in consumer devices (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025; Mills & Mills, 2016).

A central result is that scalable validation is most effective when thermal/acoustic testing is combined with targeted instrumentation during design-for-test (DFT) and early silicon debugging phases; coupling these early measurements with manufacturing process control provides leading indicators that prevent large-scale failures (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025). This insight underscores the need for cross-disciplinary teams that bridge design, test engineering, and manufacturing operations.

DISCUSSION

The results above yield multiple interpretive threads that inform both the theoretical understanding and practical engineering of GPU-based systems. Below we discuss key insights, explore competing perspectives, identify limitations, and propose a prioritized future research agenda.

Architectural specialization vs. energy efficiency: a nuanced trade-off

GPU specialization—i.e., creating an architecture finely tuned for data-parallel, throughput-oriented tasks—yields large performance gains (Dally et al., 2021; Peddie, 2023). However, specialization concentrates power and heat in workloads where many functional units are simultaneously active. The counter-argument is that specialization can be harnessed to improve energy efficiency if the hardware and software co-design match the workloads' resource profiles; examples include precision reduction (using FP16 or mixed-precision), operator fusion, and locality-enhancing data layouts (Cano, 2018; Shankar, 2023). Thus, the research question shifts from "Is specialization good?" to "How can specialization be guided by energy-aware runtime and compiler techniques to maximize energy-proportional performance?"

Integrated power/thermal control: model-based versus learning-based approaches

Model-based control (physical and control-theory-inspired) offers interpretability and predictable worst-case behaviors—attributes valuable in safety-critical or manufacturing contexts (Pagani et al., 2020). Machine-

learning-driven controllers promise adaptability and better average-case performance (Pagani et al., 2020; Shankar, 2023). Yet, they present challenges: the risk of overfitting to observed workloads, increased validation complexity, and potential failure modes when presented with adversarial or rare workloads. A hybrid approach, where interpretable model-based safety bounds constrain learning-based optimizers, appears promising: controllers can use learning methods for optimization within bounded, verifiable envelopes enforced by robust model-derived constraints (Pagani et al., 2020). Future work should formalize these bounds and define certification procedures for ML-driven thermal controllers.

Packaging, 3-D integration, and thermal path innovation

3-D stacking and GPU-in-memory architectures offer latency and bandwidth advantages that are crucial for future AI workloads (Kurshan & Franzon, 2025; Wen et al., 2017). Yet, these integrations complicate thermal design: internal dies trap heat and reduce effective thermal conductivity pathways to ambient heat sinks (Kurshan & Franzon, 2025). Cooling innovations—such as embedded microfluidic channels, thermal through-silicon vias, and high-efficiency heat spreaders—are necessary complements (Wang et al., 2022; H. Wang et al., 2022). The counter-argument is that adding cooling complexity increases manufacturing cost and may introduce reliability concerns (e.g., sealing microfluidics). Thus, economic and reliability trade-offs must be quantified. Design-space exploration tools that combine thermal simulation with cost and reliability models will be essential for making informed packaging decisions.

Manufacturing validation: balancing throughput, sensitivity, and interpretability

Scalable validation requires tests that are both fast and diagnostically informative. Burn-in or long-duration thermal cycling tests are sensitive but time-consuming; conversely, short synthetic benchmarks may miss intermittent defects. The 2025 validation strategies recommend a staged approach: rapid screening tests at end-of-line, followed by targeted longer-term assays for samples exhibiting anomalies, and continuous manufacturing process monitoring to detect drift (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025). A challenge remains in defining standards: different vendors and product segments (data center vs. consumer) accept differing acoustic and thermal profiles. Therefore, industry-standard benchmark suites and reporting formats would improve comparability and reduce time-to-market ambiguities.

Implications for cloud gaming and mobile systems

Cloud gaming systems (Hou et al., 2014) and mobile gaming (Pathania et al., 2014) illustrate the practical constraints of thermal and energy management. Cloud gaming benefits from centralized hardware but must balance energy cost at scale and ensure acoustic considerations in datacenters; mobile gaming faces battery and thermal skin temperature limits. Strategies such as frame-rate capping, adaptive resolution, and edge-based rendering distribute computational loads to manage system-level energy (Pathania et al., 2014; Hou et al., 2014). These techniques highlight the importance of application-aware controllers and user-perceived quality metrics that form the optimization objectives.

Limitations of current understanding and the role of heterogeneous devices

A critical limitation of current approaches is the incomplete characterization of heterogeneity at multiple scales: processing elements (CPU, GPU, TPU, neuromorphic accelerators), memory technologies (HBM, embedded DRAM), and packaging (2-D vs. 3-D). Emerging devices like transition-metal-dichalcogenide-based memristors (Wu et al., 2024) hint at new co-design opportunities but also raise new thermal and reliability questions when co-packaged with CMOS. The literature suggests that a holistic co-design approach—pairing device-level innovation with system-level power management and manufacturing-aware validation—will be the path

forward. However, achieving such integrated design cycles requires new modeling tools that bridge device physics, architecture, runtime behavior, and manufacturing variability.

Future research directions and a prioritized agenda

Based on the synthesis, the following research priorities emerge:

1. Co-design frameworks that integrate thermal-aware architectural exploration with compiler and runtime optimizations. These frameworks should quantify both performance and energy/thermal outcomes for design points including precision choices, memory hierarchies, and parallelism degrees (Dally et al., 2021; Cano, 2018; Shankar, 2023).
2. Hybrid controller designs that combine model-based safety envelopes with learning-based optimization. Research should demonstrate provable safety bounds and validation methodologies for ML-enabled thermal controllers (Pagani et al., 2020; Shankar, 2023).
3. Packaging studies that quantify cost–performance–reliability trade-offs for 3-D stacking, including experimental and simulated analyses of microfluidic cooling, thermal vias, and heat-spreader materials (Kurshan & Franzon, 2025; Wang et al., 2022; Wen et al., 2017).
4. Standardized, scalable acoustic and thermal validation benchmarks for manufacturing that balance throughput and diagnostic power. Work should define benchmark workloads, sensor placements, statistical process control metrics, and reporting standards (Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025).
5. Cross-domain studies of heterogeneous-integration effects where neuromorphic/memristive elements are co-located with CMOS GPUs, including thermal interactions and co-managed power policies (Wu et al., 2024).
6. Economic and life-cycle assessment models that analyze the cost and environmental impact of advanced cooling and validation practices (Mills & Mills, 2016; Shankar, 2023).

Each of these directions addresses a specific gap identified by the synthesis and collectively supports the long-term sustainability and performance trajectory of GPU-centric systems.

LIMITATIONS

This article's central limitation is methodological: it is strictly a literature-based synthesis and does not present original empirical measurements, simulations, or experiments beyond the interpretive integration of the cited works. Consequently, conclusions about absolute performance or thermal behavior under particular hardware configurations should be treated as conceptual and hypothesis-generating rather than definitive. Another limitation arises from the reference set itself: while it includes seminal and recent works, the constrained corpus may omit other contemporaneous studies that could nuance particular claims (for instance, vendor-specific whitepapers or confidential manufacturing data). Finally, the heterogeneity of product segments (consumer, professional, datacenter) means that the generalizations herein need contextual adaptation; solutions appropriate for datacenter GPUs may not scale down to mobile contexts without significant redesign.

CONCLUSION

GPUs occupy a pivotal role in contemporary and future computing landscapes, powering scientific simulation, large-scale data mining, machine learning, cloud gaming, and more (Dally et al., 2021; Cano, 2018; Hou et al., 2014). Their success is rooted in architectural choices that prioritize parallelism, memory bandwidth, and

throughput. However, these same choices create concentrated power and thermal challenges that cascade across design, runtime management, and manufacturing validation. This article has synthesized the literature to present an integrated view: to achieve sustainable and high-performance GPU systems, designers must embrace co-design principles that couple architectural innovations—including 3-D stacking and GPU-in-memory—with energy-aware runtime controls and scalable acoustic/thermal validation pipelines (Kurshan & Franzon, 2025; Wen et al., 2017; Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing, 2025).

Practical engineering solutions will likely be hybrid: combining interpretable, model-based safety constraints with adaptive, learning-based optimizers; pairing advanced cooling (e.g., microfluidics, thermal vias) with manufacturing process controls; and adopting standard, scalable benchmark suites for thermal and acoustic validation. The prioritized research agenda proposed here aims to catalyze advances along these dimensions, paving the way for GPU systems that deliver both raw computational throughput and sustainable, reliable operation across product segments.

REFERENCES

1. Dally, W.J.; Keckler, S.W.; Kirk, D.B. Evolution of the graphics processing unit (GPU). IEEE Micro 2021, 41, 42–51.
2. Peddie, J. The History of the GPU-Steps to Invention; Springer: Berlin/Heidelberg, Germany, 2023.
3. Peddie, J. What is a GPU? In The History of the GPU-Steps to Invention; Springer: Berlin/Heidelberg, Germany, 2023; pp. 333–345.
4. Cano, A. A survey on graphic processing unit computing for large-scale data mining. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018, 8, e1232.
5. Shankar, S. Energy Estimates Across Layers of Computing: From Devices to Large-Scale Applications in Machine Learning for Natural Language Processing, Scientific Computing, and Cryptocurrency Mining. In Proceedings of the 2023 IEEE High Performance Extreme Computing Conference (HPEC), Boston, MA, USA, 25–29 September 2023; pp. 1–6.
6. Hou, Q.; Qiu, C.; Mu, K.; Qi, Q.; Lu, Y. A cloud gaming system based on NVIDIA GRID GPU. In Proceedings of the 2014 13th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Xianning, China, 24–27 November 2014; pp. 73–77.
7. Pathania, A.; Jiao, Q.; Prakash, A.; Mitra, T. Integrated CPU-GPU power management for 3D mobile games. In Proceedings of the 51st Annual Design Automation Conference, San Francisco, CA, USA, 1–5 June 2014; pp. 1–6.
8. Mills, N.; Mills, E. Taming the energy use of gaming computers. Energy Effic. 2016, 9, 321–338.
9. Teske, D. NVIDIA Corporation: A Strategic Audit; University of Nebraska-Lincoln: Lincoln, NE, USA, 2018.
10. Scalable Acoustic and Thermal Validation Strategies in GPU Manufacturing. International Journal of Data Science and Machine Learning, 2025, 5(01), 193-214. https://doi.org/10.55640/ijdsml-05-01-19
11. Moya, V.; Gonzalez, C.; Roca, J.; Fernandez, A.; Espasa, R. Shader performance analysis on a modern GPU architecture. In Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'05), Barcelona, Spain, 12–16 November 2005; pp. 10–364.
12. Kirk, D. NVIDIA CUDA software and GPU parallel computing architecture. In Proceedings of the International

Symposium on Memory Management (ISMM), Montreal, QC, Canada, 21–22 October 2007; Volume 7, pp. 103–104.

13. Pagani, S.; Manoj, P. D. S.; Jantsch, A.; Henkel, J. Machine Learning for Power, Energy, and Thermal Management on Multicore Processors: A Survey. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 2020, 39, 101–116.

14. Kurshan, E.; Franzon, P. 3-D Stacking for AI Systems: Emulating Heterogeneous 3-D Architectures for AI. IEEE Transactions on Components Packaging and Manufacturing Technology 2025, 15, 1161–1169.

15. Wu, X.; Kim, Jae Gwang; Hou, Aolin; Wang, Shiren. Transition Metal Dichalcogenides-Based Memristors for Neuromorphic Electronics. Journal of Neuromorphic Intelligence 2024, 1.

16. Wen, W.; Yang, J.; Zhang, Y. T. Optimizing power efficiency for 3D stacked GPU-in-memory architecture. Microprocessors and Microsystems 2017, 49, 44–53.

17. Wang, H.; Wu, Q.; Wang, C.; Wang, R. Z. A universal high-efficiency cooling structure for high-power integrated circuits. Applied Thermal Engineering 2022, 215.

18. Teske, D. NVIDIA Corporation: A Strategic Audit; University of Nebraska-Lincoln: Lincoln, NE, USA, 2018.