World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

Check for updates

AI-driven dynamic power allocation between CPU and GPU for optimal performance and battery life

Pratikkumar Dilipkumar Patel *

Arizona State University, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 2801–2815

Publication history: Received on 16 April 2025; revised on 27 May 2025; accepted on 29 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0841

Abstract

Dynamic power allocation strategies for heterogeneous computing systems have emerged as a crucial advancement in optimizing AI workload performance while managing energy consumption. This article explores the fundamental challenges posed by static power allocation in CPU-GPU systems and presents AI-driven solutions that enable intelligent redistribution of power resources based on real-time computational demands. The integration of machine learning techniques for workload characterization and power prediction allows these systems to anticipate phase-dependent behavior and proactively adjust power distribution, significantly improving both energy efficiency and computational throughput. Various implementation approaches are examined, from hardware-level composable architectures to operating system facilitation mechanisms, highlighting the tangible benefits observed across diverse computing environments from data centers to edge devices. Despite impressive advancements, several challenges persist, including prediction accuracy limitations, implementation complexity, and privacy concerns. Future directions point toward deeper hardware integration of AI capabilities, increasingly granular power control mechanisms, and standardized interfaces across heterogeneous components to further enhance the effectiveness of dynamic power allocation in next-generation computing systems.

Keywords: Dynamic Power Allocation; Heterogeneous Computing; AI Workloads; Energy Efficiency; CPU-GPU Optimization

1. Introduction

The computational landscape for artificial intelligence (AI) applications has evolved dramatically, with heterogeneous computing architectures, specifically systems integrating central processing units (CPUs) and graphics processing units (GPUs) emerging as the dominant paradigm. These hybrid systems capitalize on the complementary strengths of both processor types: CPUs excel at sequential processing with sophisticated branch prediction and deep cache hierarchies, while GPUs offer massive parallelism ideal for the matrix operations fundamental to modern AI algorithms [1]. As organizations increasingly deploy sophisticated AI models, the power demands and distribution challenges of these heterogeneous systems have become critical concerns for both performance optimization and operational sustainability.

AI workloads present unique power utilization patterns characterized by extreme dynamism and high intensity. Recent research has documented that large-scale AI training clusters now operate at megawatt scales with power density exceeding 35 kW per rack [1]. More significantly, these workloads exhibit distinctive temporal characteristics, with power draw fluctuations that can span from microseconds to hours. At the microsecond level, modern GPUs demonstrate remarkable power transients, with slew rates exceeding 1000 A/µs during phase transitions between compute-intensive and memory-intensive operations [1]. These rapid fluctuations present substantial challenges for

* Corresponding author: Pratikkumar Dilipkumar Patel.

conventional power delivery architectures that were designed for more predictable, gradual changes in computational load.

The performance implications of these power constraints are substantial. Under static power allocation schemes, where fixed power budgets are assigned to CPUs and GPUs, system efficiency suffers significantly. When multiple applications run concurrently on integrated CPU-GPU platforms constrained by a total power budget, traditional approaches fail to adapt to the changing resource needs of diverse workloads [2]. This rigid allocation frequently results in scenarios where one processing unit becomes power-starved while the other operates below its potential, creating artificial performance bottlenecks that compromise overall system throughput.

Power management challenges extend beyond individual server units to data center infrastructure. The temporal diversity of AI workloads spanning microsecond voltage regulator dynamics to daily load patterns requires multi-layered power management solutions [1]. Conventional data centers typically provide for peak power demand plus a safety margin, resulting in substantial unutilized capacity during average operation. This overprovisioning represents both capital inefficiency and operational waste, particularly problematic as AI deployments scale globally.

Research into co-run scheduling techniques with dynamic power capping has demonstrated potential efficiency gains of 20-40% compared to static allocation approaches [2]. These techniques continuously monitor application behavior and workload characteristics across both CPU and GPU subsystems, redistributing power allocations in real-time based on resource sensitivity and utilization patterns. Such adaptive approaches can identify when certain applications would benefit more from additional power allocation to either the CPU or GPU component, optimizing the performance-per-watt metric crucial for sustainable AI deployment.

The intersecting challenges of power delivery, thermal management, and performance optimization in heterogeneous computing systems call for sophisticated solutions that transcend traditional static approaches. Dynamic power management frameworks that leverage workload characterization, predictive modeling, and adaptive allocation represent a promising direction for addressing the unique demands of AI applications [1, 2]. As AI continues its rapid integration across industries, the effectiveness of these power management strategies will increasingly determine not only computational efficiency but also the economic and environmental sustainability of large-scale AI deployments.

2. Background: Heterogeneous Architectures and Power Challenges

Heterogeneous computing architectures represent a paradigm shift in computational system design, fundamentally altering how processing resources are organized and utilized. These architectures intentionally integrate disparate processing units primarily CPUs and GPUs to capitalize on their complementary strengths across varied computational tasks. This approach has gained significant traction as traditional homogeneous systems have struggled to maintain performance scaling under power constraints [3].

CPUs and GPUs exhibit fundamentally different architectural philosophies. Modern CPUs typically incorporate between 4-64 sophisticated cores designed for versatile instruction handling, branch prediction, and speculative execution, with each core capable of independent task management. These cores operate at high frequencies (typically 3-5 GHz) and incorporate substantial cache hierarchies (up to 128MB in high-end server CPUs) to minimize memory latency [3]. This design prioritizes single-threaded performance and instruction-level parallelism, making CPUs exceptionally efficient for serial processing, complex decision-making algorithms, and tasks with unpredictable memory access patterns.

In stark contrast, GPU architectures embrace massive parallelism, with contemporary devices featuring thousands of relatively simple processing cores (up to 10,752 CUDA cores in high-end models) organized into streaming multiprocessors [3]. These cores operate at more modest frequencies (typically 1-2 GHz) but achieve extraordinary aggregate throughput for suitable workloads. The GPU memory subsystem is likewise optimized for bandwidth rather than latency, with contemporary HBM2e interfaces delivering up to 3.2 TB/s of memory bandwidth [3]. This architectural approach makes GPUs supremely efficient for data-parallel operations where the same instruction sequence executes across large datasets precisely the computational pattern dominant in deep learning, scientific simulation, and graphical rendering.

The integration of these processing units introduces complex power management challenges. Research has documented distinctive power consumption profiles between CPUs and GPUs, particularly under AI workloads. Comprehensive benchmarking of GPU behavior under various computational loads has revealed that power consumption can fluctuate by 30-45% when transitioning between compute-bound and memory-bound operations, even when executing the same application [4]. These power fluctuations occur at millisecond timescales, presenting significant challenges for traditional power delivery and management systems designed for more gradual load transitions.

The efficiency implications of static power allocation in these heterogeneous systems are substantial. Experimental analysis of fixed power budgets in integrated CPU-GPU systems reveals performance inefficiencies of up to 27% compared to theoretical optimums [4]. In typical scenarios, static allocation results in either power starvation during peak computational periods or significant power wastage during phases where either the CPU or GPU experiences reduced utilization. The performance impact is particularly pronounced in applications with phase-dependent processing requirements, where computational demands shift between CPU and GPU resources throughout execution.

Dynamic Voltage and Frequency Scaling (DVFS) techniques have emerged as essential tools for managing these challenges. Comprehensive evaluations of DVFS implementations on NVIDIA K20 GPUs demonstrate energy savings of 10-25% across benchmark suites, with minimal performance degradation when correctly tuned [4]. These savings are achieved by dynamically adjusting GPU core voltages from 0.7V to 1.0V and frequencies from 324 MHz to 758 MHz in response to application characteristics. The performance-per-watt improvements are particularly significant in memory-bound applications, where reducing core frequencies has minimal impact on execution time while substantially decreasing power consumption.

The problem is further complicated in multi-application environments where diverse workloads compete for shared power resources. Analysis of concurrent GPU applications indicates that without intelligent power management, performance variations of up to 35% can occur depending on co-scheduled workloads [3, 4]. This variability stems from both resource contention and thermal interactions, where power-intensive applications can induce thermal constraints that affect all concurrently running tasks.

As AI applications continue to scale in computational intensity and deployment scope, addressing these heterogeneous power management challenges has become essential for sustainable computing. Dynamic power allocation frameworks that intelligently distribute power resources based on workload characteristics represent a promising approach to maximizing computational efficiency within fixed power constraints.

In stark contrast, GPU architectures embrace massive parallelism, with contemporary devices featuring thousands of relatively simple processing cores (up to 10,752 CUDA cores in high-end models) organized into streaming multiprocessors [3]. These cores operate at more modest frequencies (typically 1-2 GHz) but achieve extraordinary aggregate throughput for suitable workloads. The GPU memory subsystem is likewise optimized for bandwidth rather than latency, with contemporary HBM2e interfaces delivering up to 3.2 TB/s of memory bandwidth [3]. This architectural approach makes GPUs supremely efficient for data-parallel operations where the same instruction sequence executes across large datasets precisely the computational pattern dominant in deep learning, scientific simulation, and graphical rendering. The NVIDIA Blackwell architecture exemplifies these design principles, featuring 5th generation CUDA cores and advanced power management capabilities specifically optimized for AI workloads. Blackwell-based GPUs implement sophisticated power monitoring and control mechanisms that operate at multiple granularities, from individual tensor cores to streaming multiprocessors, with transition latencies as low as 20-30μs between power states. Additionally, the memory hierarchy incorporates independent power states for different partitions that can be dynamically adjusted based on access patterns, allowing precise matching of power allocation to computational demands throughout varying execution phases [5].

Table 1 Heterogeneous Architecture Characteristics [3, 4]

Component	Specification
CPU Core Count	4-64 cores
CPU Frequency Range	3-5 GHz
GPU Core Count	Up to 10,752 CUDA cores
GPU Frequency Range	1-2 GHz
HBM2e Memory Bandwidth	Up to 3.2 TB/s
GPU Power Fluctuation Range	30-45%
Performance Inefficiency with Static Allocation	Up to 27%

3. AI-Driven Workload Characterization and Power Prediction

The effective implementation of dynamic power allocation strategies in heterogeneous computing systems necessitates sophisticated methods for workload characterization and power consumption prediction. Artificial intelligence techniques, particularly machine learning algorithms, have emerged as powerful tools for analyzing the complex, non-linear relationships between application characteristics and their corresponding power requirements. These AI-driven approaches enable systems to anticipate power demands and optimize resource allocation with unprecedented precision.

Machine learning models excel at identifying subtle patterns in multivariate data that would be impractical to detect through traditional heuristic approaches. Research has demonstrated that supervised learning algorithms can predict GPU performance metrics with remarkable accuracy, achieving mean absolute percentage errors as low as 7.3% for execution time and 6.8% for energy consumption across diverse benchmark suites [6]. These predictive capabilities are especially valuable in heterogeneous systems, where determining optimal workload distribution between CPU and GPU resources requires accurate forecasting of performance and power characteristics on both processing units.

The feature selection process for these predictive models represents a critical design consideration. Comprehensive studies have identified that CPU performance counters readily accessible hardware metrics that track microarchitectural events like cache misses, branch mispredictions, and instruction throughput can serve as highly informative proxies for estimating GPU behavior [6]. This approach offers significant advantages in production environments, as it enables systems to make preliminary power allocation decisions based on CPU-only execution profiles before committing to specific resource distribution strategies. Experimental validation across 224 distinct kernels from 55 applications demonstrated that models trained on just 10 carefully selected CPU performance counter features could achieve 91.2% prediction accuracy for GPU execution characteristics [6].

The temporal dimension of power prediction introduces additional complexity, as AI workloads frequently exhibit phase-dependent behavior with distinct power profiles at different execution stages. Advanced modeling approaches address this challenge through statistical characterization of temporal power variations. Time-series analysis of GPU power consumption during deep learning training has revealed that power fluctuations follow distinctive patterns, with standard deviations ranging from 18W to 47W depending on the specific neural network architecture and batch size configuration [7]. These fluctuations occur at multiple time scales, from millisecond-level variations during kernel execution to longer periodic patterns across training epochs [7].

Table 2 AI Power Prediction Performance [6, 7]

Metric	Accuracy/Value
ML Prediction Error (Execution Time)	7.30%
ML Prediction Error (Energy)	6.80%
CPU Feature to GPU Prediction Accuracy	91.20%
GPU Power Standard Deviation Range	18-47W
Component-Specific Prediction Accuracy	85-97%
Performance Variance Reduction	Up to 37%

Microarchitecture-aware power modeling provides further insights for fine-grained power prediction. Detailed analysis of GPU pipeline stages has enabled the development of component-specific power models that account for the unique contributions of shader cores, memory controllers, and interconnect networks [7]. These disaggregated models achieve per-component prediction accuracies between 85-97%, allowing systems to identify specific power bottlenecks and optimize accordingly. For example, transformer-based models have been observed to stress memory subsystems differently than convolutional architectures, with up to 32% higher power consumption in memory controllers despite similar overall power envelopes [7].

The integration of multiple prediction granularities from application-level characterization to component-specific modeling enables comprehensive power forecasting across diverse operating conditions. Ensemble methods that combine predictions from multiple specialized models have demonstrated robust performance across varied workloads, achieving average prediction errors below 8.5% even for previously unseen applications [6]. This generalization capability is crucial for practical deployment, as systems must adapt to continuously evolving AI software frameworks and model architectures.

Beyond steady-state prediction, capturing transient power behavior represents a frontier challenge for AI-driven workload characterization. Research has identified distinct power signatures during phase transitions in GPU workloads, with initial power ramps exhibiting slopes of 40-120W/ms depending on the specific transition type [6]. These transients can trigger protective throttling mechanisms if not properly anticipated, highlighting the importance of predictive models that capture not just average power consumption but also temporal dynamics.

The practical implementation of these AI-driven prediction techniques enables proactive power management strategies that significantly outperform reactive approaches. By forecasting workload power requirements before execution, systems can pre-allocate appropriate power budgets, reducing performance variance by up to 37% compared to reactive allocation methods [6, 7]. This predictive capability is particularly valuable in multi-tenant GPU environments, where workload interference can amplify power fluctuations and exacerbate allocation challenges.

As AI workloads continue to grow in complexity and scale, the sophistication of power prediction methodologies must evolve accordingly. Emerging research directions include transfer learning approaches that adapt power models across different hardware generations, reinforcement learning techniques that optimize prediction accuracy for specific operational contexts, and federated learning methods that leverage distributed power telemetry to improve global prediction performance while preserving privacy and security [6, 7].

3.1. Dynamic Power Allocation Techniques

The ever-increasing computational demands of modern AI workloads have catalyzed the development of sophisticated power management techniques that dynamically redistribute power resources between CPU and GPU components. These approaches leverage both hardware and software innovations to maximize system efficiency while maintaining performance targets within overall power constraints. Dynamic allocation techniques operate across multiple levels of abstraction, from hardware-level resource composition to software-driven workload optimization.

Composable infrastructure represents a paradigm shift in hardware resource management, particularly for GPU-accelerated AI workloads. Traditional static hardware configurations often result in significant resource underutilization, with studies indicating that GPU utilization in enterprise environments averages only 15-30% across deployments [8]. Composable GPU architectures address this inefficiency by disaggregating physical GPUs from server chassis and placing them in resource pools that can be dynamically allocated based on application requirements. This approach enables unprecedented flexibility in resource scaling, allowing systems to provision precisely tailored GPU configurations for specific AI workloads.

The implementation of composable GPU architectures yields substantial efficiency improvements across diverse AI applications. Benchmark testing across large language model deployments has demonstrated that dynamic GPU allocation can improve throughput by 47-83% compared to static configurations while simultaneously reducing power consumption by 21-35% [8]. These gains stem from the ability to precisely match GPU resources to specific model requirements, avoiding both over-provisioning (which wastes power) and under-provisioning (which degrades performance). For instance, transformer-based models with high parameter counts benefit from configurations emphasizing GPU memory capacity, while CNN-based applications typically extract greater benefit from raw computational throughput.

PCIe fabric-based composable systems enable this flexibility by creating a dynamic interconnect between CPUs and GPU resources with near-native performance characteristics. Advanced implementations achieve latency penalties below 3.8μs compared to direct-attached configurations while maintaining 94-97% of native bandwidth [8]. This performance preservation is critical for latency-sensitive AI inference applications, where responsiveness requirements may dictate strict service-level agreements.

The efficiency benefits of composable GPU architectures extend beyond individual workloads to multi-tenant environments. In production deployments supporting multiple concurrent AI applications, dynamic GPU allocation has demonstrated the ability to support 2.3-3.1× higher workload density compared to static provisioning approaches [8]. This consolidation directly translates to power efficiency gains at the data center level, with typical implementations reducing overall power consumption by 22-38% while maintaining equivalent computational output.

At a more granular level, dynamic power sharing technologies enable fine-grained power redistribution between CPU and GPU components within a single system. Intel's Dynamic Power Share technology exemplifies this approach, implementing sophisticated power monitoring and allocation mechanisms that continuously balance resources between CPU and integrated or discrete GPU components [9]. This technology operates through a coordinated power management framework that monitors real-time workload characteristics and dynamically adjusts the power budget allocation based on application needs.

Table 3 Dynamic Allocation Techniques Performance [8, 9]

Technique	**Impact**
Dynamic GPU Allocation Throughput Improvement	47-83%
Dynamic GPU Allocation Power Reduction	21-35%
Composable GPU System Latency Penalty	<3.8μs
Composable GPU Bandwidth Preservation	94-97%
Workload Density Improvement	2.3-3.1×
Dynamic Power Sharing Performance Improvement (Integrated)	14-23%
Dynamic Power Sharing Performance Improvement (Discrete)	7-18%
Mobile Sustained Performance Improvement	15-28%

The underlying power sharing architecture incorporates thermal and power sensing capabilities that monitor conditions across more than 1,000 on-die sensors with sampling rates of 1ms or better [9]. These measurements feed into predictive models that forecast workload power requirements across both CPU and GPU domains. Based on these predictions, the power management controller dynamically adjusts the power distribution between components, with transition times as low as 50-100μs to accommodate rapidly changing workload characteristics [9].

The efficiency gains from dynamic power sharing are particularly pronounced in mixed workloads that alternate between CPU-intensive and GPU-intensive phases. Benchmark testing across representative AI workflows has demonstrated performance improvements of 14-23% for integrated graphics configurations and 7-18% for discrete GPU setups compared to static power allocation approaches [9]. These improvements stem from the system's ability to shift power resources to the component currently experiencing the highest computational demand, rather than maintaining fixed power allocations that may be suboptimal for the current workload phase.

The implementation of dynamic power sharing requires sophisticated coordination between hardware and software components. At the hardware level, voltage regulators must support rapid adjustment capabilities with slew rates sufficient to accommodate millisecond-scale power transitions without introducing voltage instability. Software control mechanisms leverage operating system power management frameworks to implement policies that balance performance requirements against power consumption targets. Advanced implementations incorporate workload-aware optimization that recognizes specific application signatures and applies pre-optimized power distribution templates tailored to their characteristic requirements [9].

Real-world validation of these dynamic allocation approaches demonstrates their practical benefits across diverse computing scenarios. In mobile environments with strict thermal constraints, dynamic power sharing enables systems to maintain 15-28% higher sustained performance under extended workloads by intelligently redistributing thermal headroom between CPU and GPU components as application requirements evolve [9]. In data center environments, similar techniques applied at rack scale have demonstrated the ability to support 42% higher computational density within the same power envelope, significantly improving the performance-per-watt metric critical for sustainable AI deployments [8].

As AI workloads continue to evolve in complexity and scale, these dynamic power allocation techniques will play an increasingly critical role in managing the tension between computational performance and energy efficiency. The integration of these approaches with AI-driven predictive models presents particularly promising opportunities, enabling systems to anticipate workload shifts before they occur and proactively adjust power allocations to optimize both performance and efficiency.

3.2. Operating System and System Software Facilitation

The successful implementation of AI-driven dynamic power allocation strategies fundamentally depends on sophisticated operating system and system software support. These software layers serve as the critical intermediaries between hardware capabilities and application requirements, creating the infrastructure necessary for intelligent power management across heterogeneous computing resources. As computing architectures evolve toward increasingly heterogeneous designs, operating systems must adapt to effectively orchestrate these diverse computational resources while optimizing for both performance and energy efficiency.

Modern operating systems face unique challenges in supporting dynamic processors and power allocation mechanisms. Traditional OS designs were developed with relatively static hardware configurations in mind, where processor capabilities remained largely consistent throughout execution. Contemporary heterogeneous systems, however, can dynamically reconfigure their computational resources in response to workload characteristics, requiring fundamentally different OS approaches [10]. These challenges are magnified in AI workloads, where computational demands can shift rapidly between sequential and parallel execution phases, each with distinct power profiles and resource requirements.

The Chameleon project represents a significant advancement in operating system support for dynamic processors. This Linux-based framework enables rapid hardware reconfiguration to optimize both performance and energy efficiency across diverse workload types. Experimental evaluations have demonstrated that Chameleon can improve energy efficiency by 35-55% compared to static configurations while maintaining comparable performance levels across benchmark suites [10]. These efficiency gains stem from the system's ability to adapt processor resources to match application characteristics dynamically, rather than maintaining fixed configurations optimized for specific workload types.

The core architecture of Chameleon introduces several innovations critical for dynamic resource management. Its modular design incorporates specialized components for hardware monitoring, workload characterization, and reconfiguration decision-making. The monitoring subsystem collects over 32 distinct performance metrics at microsecond intervals, providing high-resolution visibility into application behavior [10]. This telemetry feeds into sophisticated classification algorithms that identify execution phases and predict upcoming resource requirements with 91-96% accuracy across tested benchmarks.

Particularly relevant for AI workloads is Chameleon's support for heterogeneous core management, which enables intelligent task distribution between CPU and GPU resources. The system implements a combination of static analysis and runtime monitoring to identify code regions suitable for GPU execution, achieving speedups of 2.1-4.7× for applicable portions while maintaining energy efficiency [10]. This capability is essential for AI applications that exhibit phase-dependent behavior, where certain algorithmic components may benefit from GPU acceleration while others run more efficiently on CPUs.

The reconfiguration mechanisms in Chameleon support both coarse-grained (core allocation) and fine-grained (voltage/frequency scaling) adjustments, with transition latencies as low as 2-5μs for frequency changes and 10-15μs for power gating operations [10]. These rapid transition capabilities are crucial for adapting to the dynamic nature of AI workloads, which can exhibit significant variations in computational intensity over millisecond timescales. The system's decision engine incorporates both rule-based heuristics and machine learning models to determine optimal configurations, with the latter demonstrating 15-23% better energy efficiency compared to static allocation approaches.

At a more accessible level, commercial operating systems are increasingly incorporating dynamic power management capabilities specifically designed for heterogeneous computing systems. Ubuntu Linux 25.04 exemplifies this trend with its integration of NVIDIA Dynamic Boost technology, a sophisticated power-sharing mechanism that dynamically redistributes power between CPU and GPU components based on workload demands [11]. This implementation showcases how modern operating systems are evolving to accommodate hardware-level power management techniques within standardized, user-friendly environments.

The Ubuntu implementation of Dynamic Boost operates through a specialized system daemon (nvidia-powerd) that continually monitors the power consumption and utilization patterns of both CPU and GPU components. This daemon samples hardware performance counters at intervals of 10-50ms, providing near-real-time visibility into system behavior [11]. Based on these measurements, the daemon dynamically adjusts power allocation between components, with adjustments as large as 15-35W shifting between CPU and GPU resources depending on the specific workload characteristics.

Benchmark testing on Ubuntu systems with Dynamic Boost enabled has demonstrated performance improvements of 8-17% in graphics-intensive applications and 4-7% in general computational workloads compared to systems without dynamic power allocation [11]. These gains are particularly pronounced in applications that alternate between CPU-intensive and GPU-intensive phases, as the system can redistribute power resources to the component currently experiencing the highest computational demand.

The integration of these capabilities into mainstream operating systems represents a significant step toward democratizing advanced power management techniques. While early implementations of dynamic power allocation required specialized hardware and proprietary software stacks, their incorporation into widely used operating systems like Ubuntu makes these efficiency benefits accessible to a broader range of users and applications [11]. This accessibility is crucial for the widespread adoption of energy-efficient computing practices, particularly as AI applications continue to proliferate across diverse computing environments.

Looking forward, operating system support for dynamic processors and power allocation is likely to evolve in several key directions. First, the integration of more sophisticated AI-driven prediction models within the OS kernel will enable increasingly proactive resource management, anticipating workload shifts before they occur rather than reacting to already-changed conditions. Second, improved coordination between application-level hints and system-level decisions will allow for more targeted optimization, with applications providing explicit guidance about their upcoming resource requirements. Finally, enhanced telemetry and visualization tools will give both developers and system administrators deeper insights into power consumption patterns, enabling more informed decision-making about application design and infrastructure provisioning [10, 11].

3.3. Benefits of Dynamic Power Allocation

The implementation of AI-driven dynamic power allocation between CPU and GPU components yields transformative benefits across computing ecosystems, from data centers to edge devices. These benefits manifest in several critical dimensions: energy efficiency, performance optimization, thermal management, and system longevity. Understanding these advantages quantitatively demonstrates why dynamic power allocation represents a crucial advancement in sustainable computing for AI applications.

Energy efficiency improvements constitute perhaps the most immediate and measurable benefit of dynamic power allocation. In AI-focused computing systems, power consumption represents both a significant operational expense and an environmental concern. Quantitative analysis of dynamic power allocation in edge computing environments has demonstrated energy savings of 32-47% compared to static allocation approaches across diverse AI workloads [13]. These savings are particularly pronounced in environments with varying computational demands, where traditional static allocation methods frequently result in substantial power wastage during low-utilization periods. For large-scale deployments, these efficiency improvements translate to meaningful operational cost reductions, with studies indicating potential annual savings of $378-$512 per computing node in typical data center environments [13].

The emergence of AI-focused personal computing devices has further emphasized the importance of efficient power allocation. In AI PCs, specialized memory subsystems designed for heterogeneous computing demonstrate the tangible benefits of dynamic resource management. LPDDR5X memory configured with dynamic power allocation capabilities has shown power consumption reductions of up to 83% compared to standard DDR5 configurations while maintaining comparable performance for AI workloads [12]. Similarly, LPCAMM2 memory modules with intelligent power distribution between CPU and neural processing units (NPUs) achieve power efficiency improvements of 5.8-7.1× compared to traditional architectures [12]. These advancements are critical for extending battery life in mobile AI systems, with test results showing runtime extensions of 2.1-3.4 hours for typical mixed-workload scenarios [12].

Performance optimization represents another significant benefit of dynamic power allocation. By intelligently redistributing power resources based on real-time workload characteristics, these systems can ensure that critical computational bottlenecks receive priority access to available power budgets. Benchmark testing across representative AI applications has demonstrated performance improvements of 17-29% for inference tasks and 11-23% for training operations compared to static allocation approaches [13]. These gains stem from the system's ability to allocate additional power to either CPU or GPU components based on the specific requirements of different execution phases within AI workloads.

The performance benefits of dynamic allocation are particularly evident in multi-tenant environments where diverse applications compete for shared resources. In edge computing scenarios supporting multiple concurrent AI services, dynamic allocation has demonstrated the ability to improve aggregate throughput by 28-42% compared to static approaches, while simultaneously reducing energy consumption by 23-35% [13]. This dual optimization of both performance and efficiency underscores the fundamental advantage of intelligent power management in heterogeneous systems.

Memory subsystem innovations play a crucial role in realizing these performance benefits. Modern AI workloads exhibit distinctive memory access patterns that differ significantly from traditional computing tasks, with characteristic data movement requirements that can dominate overall system power consumption. Memory technologies specifically designed for heterogeneous AI systems incorporate dynamic power allocation capabilities that adapt to these unique requirements. For instance, LPDDR5X configurations with dynamic bandwidth allocation can shift power resources between GPU and CPU memory controllers based on workload demands, achieving throughput improvements of 22-31% for memory-intensive AI operations compared to static configurations [12].

Thermal management advantages constitute another significant benefit of dynamic power allocation. By intelligently distributing power resources, these systems can avoid localized thermal hotspots that might otherwise trigger performance throttling. Experimental measurements in edge computing environments have demonstrated that dynamic allocation approaches maintain peak temperatures 7-12°C lower than static allocation methods under equivalent workloads [13]. This improved thermal distribution enables sustained performance for extended durations, particularly beneficial for long-running AI training operations that might otherwise experience thermally-induced performance degradation over time.

The reliability and longevity implications of these thermal improvements are substantial. Component aging mechanisms in semiconductor devices are strongly temperature-dependent, with typical acceleration factors of 1.3-2.5× for every 10°C increase in operating temperature [12]. By maintaining lower and more consistent thermal profiles, dynamic power allocation can extend the operational lifespan of computing systems. Reliability modeling based on field data suggests potential lifetime extensions of 15-30% for systems employing adaptive power management compared to traditional fixed allocation approaches [12].

Beyond these primary benefits, dynamic power allocation enables several secondary advantages. For instance, these techniques facilitate more efficient provisioning of computing infrastructure by reducing the need for worst-case power capacity planning. Data centers employing dynamic allocation can support 18-27% higher computational density within the same power envelope, significantly improving capital efficiency [13]. Similarly, the ability to adapt to varying environmental conditions enables more resilient operation in challenging deployment scenarios, with systems demonstrating performance stability improvements of 8-14% under fluctuating ambient temperature conditions [13].

4. Challenges and Limitations of AI-Based Dynamic Power Management

While AI-driven dynamic power allocation offers compelling benefits for heterogeneous computing systems, its implementation faces numerous technical, operational, and ethical challenges that must be addressed to achieve widespread adoption. These challenges span multiple domains, from algorithmic complexity to hardware integration, and vary significantly across deployment environments from data centers to edge devices.

The fundamental challenge of developing accurate power prediction models stems from the inherent complexity of modern computing workloads, particularly in AI applications. Research on edge AI deployments has documented prediction error rates of 12-27% when standard machine learning approaches are applied to power forecasting without domain-specific optimizations [14]. These errors primarily result from the dynamic, phase-dependent nature of AI workloads, which can exhibit power fluctuations of 35-78W within milliseconds as execution transitions between different algorithmic components. Conventional prediction models struggle to capture these rapid transitions, particularly when operating with limited historical data or facing previously unseen workload patterns.

Resource constraints represent another significant challenge, particularly in edge computing environments. The implementation of sophisticated AI-based power management introduces computational overhead that must be carefully balanced against its benefits. Measurements across representative edge devices indicate that naive implementations can consume 4-7% of system resources for monitoring and prediction, potentially negating a substantial portion of the energy efficiency gains [14]. This overhead becomes particularly problematic in resource-constrained environments such as IoT devices or mobile platforms, where available compute capacity and memory are severely limited.

The reliability challenges of AI-based power management are magnified in environments with unstable power sources. Edge AI systems powered by renewable energy face unique difficulties, as they must adapt not only to workload variations but also to fluctuating energy availability. Field studies of solar-powered edge AI deployments have documented energy availability variations of 65-92% depending on weather conditions and time of day [14]. In these scenarios, power management algorithms must make complex trade-offs between immediate performance, long-term sustainability, and application quality of service. Experimental implementations have demonstrated the difficulty of these trade-offs, with even sophisticated adaptive algorithms experiencing service degradation periods of 8-14% during extended low-power conditions [14].

Spatial multitasking GPUs present particularly complex challenges for dynamic power management due to their highly parallel architecture and shared resources. When multiple applications execute concurrently on partitioned GPU resources, complex interference patterns emerge that conventional power management approaches fail to address effectively. Research has documented performance variations of 18-42% for identical workloads depending on co-location patterns and resource allocation decisions [15]. These variations stem from shared resource contention across multiple dimensions, including computational cores, memory bandwidth, cache capacity, and power delivery infrastructure.

The time-sensitive nature of many AI applications further complicates dynamic power allocation. In real-time systems such as autonomous vehicles or industrial control applications, power management decisions must respect strict latency constraints while maximizing energy efficiency. Experimental evaluations have shown that naive power allocation strategies can introduce latency variations of 15-38ms in critical processing paths, potentially exceeding acceptable thresholds for safety-critical applications [15]. Addressing these constraints requires sophisticated QoS-aware power management frameworks that can provide statistical guarantees on worst-case execution times while still capturing efficiency opportunities.

From an implementation perspective, the diversity of hardware architectures across heterogeneous computing systems presents substantial integration challenges. Modern GPUs implement proprietary power management interfaces with varying capabilities, from simplified DVFS controls to sophisticated hardware-level power capping mechanisms. This heterogeneity complicates the development of generalized power management solutions, with compatibility testing across representative GPU platforms revealing implementation differences that necessitate architecture-specific optimizations for 53-78% of power management operations [15]. These differences extend beyond interface variations to fundamental behavioral characteristics, including power state transition latencies that range from microseconds to milliseconds depending on the specific hardware implementation.

The scalability of dynamic power allocation presents another significant concern, particularly in large-scale systems with many processing elements. Current implementations have demonstrated effective resource management for systems with up to 8-16 concurrent applications, but experimental evaluations reveal scaling limitations as workload complexity increases [15]. Beyond certain thresholds typically around 24-32 concurrent processes the computational overhead of coordination and optimization can grow superlinearly, limiting practical applicability for highly multiplexed environments. These limitations stem from both algorithmic complexity and hardware constraints, as power delivery infrastructures in current systems are not designed for fine-grained, rapidly changing allocation patterns across many components.

Data privacy and security considerations introduce additional complexities for AI-driven power management. Effective power prediction often requires detailed telemetry about application behavior, which may inadvertently expose sensitive information about workload characteristics or user activities. Analysis of power traces from representative workloads has demonstrated that sophisticated side-channel attacks can recover up to 37-62% of key information from cryptographic operations based solely on power consumption patterns [14]. These vulnerabilities necessitate careful consideration of privacy-preserving monitoring approaches that can provide sufficient information for power optimization while protecting sensitive workload characteristics.

Despite these challenges, promising research directions are emerging. Self-adaptive approaches that combine rule-based heuristics with learning components have demonstrated robust performance across diverse operating conditions, achieving 82-91% of theoretical optimal efficiency while maintaining adaptability to changing environments [14]. Similarly, hardware-software co-design approaches that integrate power management awareness across the entire system stack have shown the potential to address many current limitations, reducing coordination overhead by 45-67% compared to purely software-based implementations [15].

Table 4 Challenges in Dynamic Power Management [14, 15]

Challenge	Impact/Metric
Edge AI Prediction Error Rates	12-27%
Power Fluctuation Range	35-78W
Monitoring System Overhead	4-7%
Renewable Energy Availability Variation	65-92%
Service Degradation Periods	8-14%
GPU Multitasking Performance Variation	18-42%
Latency Variation in Real-time Systems	15-38ms
Architecture-Specific Optimization Requirement	53-78%
Side-channel Attack Information Recovery	37-62%

5. Case Studies and Existing Systems

The practical implementation of AI-driven dynamic power allocation has progressed from theoretical research to deployed systems across diverse computing environments. These real-world implementations provide valuable insights into both the achievements and challenges associated with intelligent power management in heterogeneous CPU-GPU systems. Examining these case studies reveals the tangible benefits already being realized and illuminates promising directions for future development.

Large-scale AI data centers represent perhaps the most compelling demonstration of dynamic power management's impact. The power demands of modern AI infrastructure have reached unprecedented levels, with hyperscale facilities now routinely operating at power densities of 35-50 kW per rack for AI-optimized servers [16]. These extreme power requirements have catalyzed the development of sophisticated power management solutions that dynamically optimize resource allocation across heterogeneous computing elements. Micron's implementation in their AI validation clusters demonstrates how memory-aware power management can yield significant efficiency improvements. By implementing dynamic power allocation between CPU, GPU, and memory subsystems, their engineering team achieved energy efficiency improvements of 23-31% across representative large language model workloads [16].

The architecture of these optimized systems incorporates multiple power management domains with independent control capabilities. Power distribution units with per-outlet monitoring and control functions enable dynamic reallocation at rack scale, with measurement resolution of ±0.5% and adjustment capabilities at 50W increments [16]. This granular control extends to the server level, where platform management controllers implement model-specific power policies that continuously rebalance resources between processing elements. In production environments, these systems have demonstrated the ability to operate consistently at 92-97% of theoretical peak efficiency across diverse AI workloads, substantially outperforming traditional static allocation approaches [16].

The memory subsystem plays a particularly critical role in these optimized architectures. Modern AI models, especially large language models and diffusion models, exhibit extreme memory bandwidth sensitivity that significantly impacts power consumption patterns. Measurements across representative transformer-based architectures indicate that memory operations can account for 28-42% of total system power consumption during inference and 18-27% during training [16]. Dynamic power management systems address this challenge through intelligent memory configuration that adapts to workload characteristics. For instance, HBM3E memory operating in dynamically adjusted power states has demonstrated energy efficiency improvements of 4.7× for bandwidth-intensive operations and 2.8× for capacity-bound operations compared to fixed-configuration DDR5 implementations [16].

At a more granular level, research implementations have explored sophisticated algorithmic approaches to GPU power management. The GAMESS (General Atomic and Molecular Electronic Structure System) project represents a particularly illustrative case study in scientific computing applications. This quantum chemistry application leverages multiple GPUs for computationally intensive tensor operations, creating a complex power management challenge due to its phase-dependent execution patterns [17]. Researchers implemented a dynamic power allocation system that continuously monitors GPU utilization metrics and redistributes power budgets accordingly, achieving energy efficiency improvements of 17-24% compared to static allocation approaches [17].

The GAMESS implementation is particularly noteworthy for its integration of application-specific knowledge into the power management framework. By instrumenting key computational kernels with power management hooks, the system can anticipate upcoming phase transitions and proactively adjust power allocations before execution patterns change. This predictive capability enables power state transitions to complete before new computational phases begin, eliminating performance penalties that would otherwise occur from reactive management approaches. Benchmark results demonstrate that this proactive approach reduces execution time variability by 62-78% compared to reactive methods, while maintaining equivalent energy efficiency improvements [17].

The implementation details of the GAMESS power management system reveal the practical challenges of dynamic allocation in production environments. The system operates within a global power cap that constrains total consumption across all GPUs, typically set 15-20% below the theoretical maximum to accommodate power supply inefficiencies and system-level overhead [17]. Within this constraint, the allocation algorithm continuously monitors 12 distinct GPU performance metrics sampled at 50ms intervals, using these measurements to compute utilization scores that guide power distribution decisions. The allocation algorithm implements a proportional-integral-derivative (PID) control approach that has demonstrated stability across diverse workloads, with convergence times of 150-350ms following major phase transitions [17].

The performance impact of this dynamic allocation is substantial. For a representative dataset involving molecular dynamics simulations, the system achieved execution time improvements of 12-18% compared to static allocation approaches with equivalent total power consumption [17]. These improvements were particularly pronounced for complex simulations involving transition metals and large orbital basis sets, where computational phases exhibit highly variable power efficiency characteristics. The system's ability to identify these efficiency differences and redistribute power accordingly represents a significant advancement over conventional approaches that maintain fixed allocations regardless of workload characteristics.

Edge computing environments present different but equally compelling case studies of dynamic power allocation. Micron's AI PC solutions demonstrate how these techniques can be adapted to resource-constrained environments where battery life and thermal management are primary concerns. Their implementation combines hardware-level power management capabilities with software-driven workload characterization, continuously redistributing power between CPU, GPU, and NPU components based on application requirements [16]. Field testing across representative usage scenarios has demonstrated battery life extensions of 2.1-3.8 hours compared to systems without dynamic allocation, while maintaining equivalent application performance [16].

6. Future Directions and Research Opportunities

The landscape of AI-driven dynamic power allocation in heterogeneous computing systems is poised for transformative advancement as emerging technologies and research initiatives converge to address current limitations. These future directions span hardware architectures, software frameworks, and algorithmic approaches, collectively promising substantial improvements in both performance and energy efficiency for next-generation computing systems.

The integration of AI capabilities directly into processor hardware represents perhaps the most significant architectural trend. Rather than treating power management as an external function, next-generation CPUs and GPUs are incorporating dedicated neural network accelerators optimized specifically for power-related decision making. These specialized circuits, occupying just 2-3% of total die area, can process telemetry data and implement sophisticated allocation policies with minimal latency and energy overhead [18]. Preliminary implementations have demonstrated response times of 1-2μs for power state transitions, compared to 25-40μs in conventional software-driven approaches, enabling much finer-grained adaptation to workload variations [18].

The sophistication of these integrated AI systems extends beyond simple reactive policies to incorporate predictive capabilities. On-chip neural networks trained on extensive workload traces can anticipate computational demands 10-50ms before they occur, enabling proactive power allocation that eliminates performance penalties associated with reactive approaches [18]. This predictive capability is particularly valuable for applications with phase-dependent behavior, where performance can improve by 12-18% compared to reactive management techniques while simultaneously reducing energy consumption by 7-14% [18].

Beyond integration into existing architectures, AI is fundamentally reshaping processor design methodologies. Traditional design approaches relied heavily on human expertise and heuristic optimization, limiting exploration of the vast design space for power-efficient architectures. AI-driven design automation tools are dramatically expanding this exploration capability, evaluating thousands of potential microarchitectural configurations to identify optimal power-performance trade-offs [18]. These tools have already influenced commercial processor designs, with AI-optimized floor plans demonstrating power delivery efficiency improvements of 8-15% compared to conventional human-designed layouts [18].

The memory subsystem represents another critical frontier for AI-driven optimization. Future architectures will likely implement fine-grained, content-aware memory power management that adapts not just to access patterns but to the specific data being processed. Research prototypes have demonstrated techniques that can identify computational patterns associated with specific AI operations (convolution, attention mechanisms, etc.) and dynamically reconfigure memory power states to match these requirements, achieving energy savings of 22-37% compared to current approaches [19].

The emergence of specialized AI accelerators within heterogeneous systems introduces new coordination challenges and opportunities. Future research directions include the development of hierarchical power management frameworks that can effectively coordinate allocation decisions across increasingly diverse computational elements. These frameworks must balance local optimization (within each accelerator) against global efficiency objectives, a complex challenge that conventional approaches struggle to address [19]. AI-driven techniques show particular promise for this coordination role, with reinforcement learning approaches demonstrating the ability to navigate complex trade-offs across multiple objective functions while adapting to changing environmental conditions.

GPU architectures are evolving to incorporate increasingly granular power management capabilities. Future GPUs will likely implement tensor-level power gating, allowing individual tensor processing units to be selectively powered down when not needed for specific computational phases [19]. This approach represents a significant advancement over current core-level or SM-level power management, potentially reducing idle power consumption by 65-78% during sparse computation phases according to simulation studies [19]. Combined with dynamic voltage and frequency scaling at similar granularity, these techniques could enable unprecedented matching of power allocation to computational requirements.

Compiler and runtime system innovations represent another promising research direction. Future systems will likely implement sophisticated power-aware compilation techniques that generate multiple code variants optimized for different power-performance trade-offs [19]. At runtime, dynamic selection mechanisms can choose the appropriate variant based on current system conditions and power availability. Preliminary implementations have demonstrated energy efficiency improvements of 14-26% across diverse benchmark suites compared to conventional approaches that optimize code for a single operating point [19].

The standardization of power management interfaces across heterogeneous components remains a significant challenge and opportunity. Current systems implement vendor-specific interfaces with varying capabilities and control granularity, complicating the development of unified management approaches [18]. Industry initiatives to establish common power management frameworks could substantially accelerate progress by enabling more consistent optimization approaches across diverse hardware platforms.

Looking further ahead, quantum-classical hybrid systems present unique power management challenges that will require novel approaches. The extreme cooling requirements of quantum processing units create energy consumption profiles fundamentally different from conventional electronics, necessitating innovative allocation strategies that consider both computational efficiency and cooling overhead [18]. Research in this direction remains nascent but represents a critical frontier as quantum computing capabilities continue to advance.

7. Conclusion

The integration of AI-driven dynamic power allocation in heterogeneous CPU-GPU computing environments represents a significant advancement in addressing the escalating energy demands of modern computational workloads. Through sophisticated workload characterization and predictive modeling, these systems achieve remarkable improvements in both performance and energy efficiency across diverse deployment scenarios. The demonstrated benefits include substantial energy savings ranging from 20-47%, performance enhancements of 12-83%, and significant improvements in thermal management leading to extended system longevity. The implementation of these technologies' spans multiple abstraction levels, from hardware-level resource composition to operating system facilitation mechanisms, each contributing unique capabilities to the overall power management framework. Despite impressive progress, challenges persist in prediction accuracy, implementation complexity, and cross-vendor standardization. Future directions point toward deeper hardware integration of AI capabilities, increasingly granular control mechanisms, and comprehensive frameworks that coordinate power allocation across increasingly diverse computational elements. As AI applications continue to proliferate and scale, effective dynamic power allocation will become increasingly critical for sustainable computing, driving continued innovation at the intersection of machine learning and power management technologies. The evolution of these approaches will play a decisive role in determining not only the computational efficiency but also the economic and environmental sustainability of AI deployments worldwide.

References

[1] Yuzhuo Li, and Yunwei Li, "AI Load Dynamics–A Power Electronics Perspective," arXiv, 2025. https://arxiv.org/html/2502.01647v2

[2] Qi Zhu, et al., "Co-Run Scheduling with Power Cap on Integrated CPU-GPU Systems," Research, https://research.csc.ncsu.edu/picture/publications/papers/ipdps17.pdf

[3] Hyperstack, "Rent NVIDIA DGX B200 GPU – Boost Your AI Workloads," Hyperstack, https://www.hyperstack.cloud/nvidia-blackwell-b200#:~:text=Interconnect-,5th%20Generation%20NVLink%3A%201.8TB%2Fs%2C%20PCIe%20Gen6%3A,FP8%2FFP6%20Tensor%20Core

[4] Rong Ge, et al., "Effects of Dynamic Voltage and Frequency Scaling on a K20 GPU," Science - Texas State University, accessed March 23, 2025, https://userweb.cs.txstate.edu/~mb92/papers/pasa13.pdf

[5] PNY, NVIDIA RTX PRO 6000 Blackwell Server Edition, PNY. https://www.pny.com/nvidia-rtx-pro-6000-blackwell#:~:text=NVIDIA%20Blackwell%20Architecture-,CUDA%20Parallel%20Processing%20Cores,752%20(5th%20Gen)

[6] Ioana Baldini, et al., "Predicting GPU Performance from CPU Runs Using Machine Learning," ResearchGate, 2014. https://www.researchgate.net/publication/292845940_Predicting_GPU_Performance_from_CPU_Runs_Using_Machine_Learning

[7] Gene Wu, et al., "GPGPU Performance and Power Estimation Using Machine Learning" https://users.ece.utexas.edu/~derek/Papers/HPCA2015_GPUPowerModel.pdf

[8] Sumit Puri, "Optimizing AI Model Performance Through Dynamic GPU Allocation," Liqid, 2024. https://www.liqid.com/blog/optimizing-ai-model-performance-through-dynamic-gpu-allocation

[9] Intel, "What is Intel® Dynamic Power Share for Intel® Graphics?," 2024. Intel, https://www.intel.com/content/www/us/en/support/articles/000090047/graphics/intel-arc-dedicated-graphics-family.html

[10] Sankaralingam Panneerselvam and Michael M. Swift, "Chameleon: Operating System Support for Dynamic Processors," https://pages.cs.wisc.edu/~swift/papers/asplos12_chameleon.pdf

[11] Arol Wright, "Ubuntu Linux 25.04 Will Be a Great Upgrade for Games," How to Geek, 2025. https://www.howtogeek.com/ubuntu-25-04-nvidia-dynamic-boost/

[12] Micron, "The Role of the Memory Subsystem in Achieving AI PC Efficiency," Micron Technology, https://www.micron.com/content/dam/micron/global/public/products/white-paper/ai-pc-white-paper.pdf

[13] Lavanya Shanmugam, et al., "Dynamic Resource Allocation in Edge Computing for AI/ML Applications: Architectural Framework and Optimization Techniques," Journal of Knowledge Learning and Science Technology, 2023. https://jklst.org/index.php/home/article/download/173/146/494

[14] Julia Oberauner, "Dynamic Power Management for Edge AI: A Sustainable Self-Adaptive Approach," netidee, 2024, https://www.netidee.at/dynamic-power-management-edge-ai-sustainable-self-adaptive-approach

[15] Hoda Sedighi, et al., "Efficient Dynamic Resource Management for Spatial Multitasking GPUs," IEEE Transactions on Cloud Computing, 2025. https://www.computer.org/csdl/journal/cc/5555/01/10778657/22qSX1nspqw

[16] Larry Hart, "AI drives power consumption, Micron drives power efficiency," Micron, 2024, https://www.micron.com/about/blog/storage/ai/ai-drives-power-consumption-micron-drives-power-efficiency

[17] Masha Sosonkina, et al., "Runtime Power Allocation Based on Multi-GPU Utilization in GAMESS," Scientific Research Publishing, 2022. https://www.scirp.org/journal/paperinformation?paperid=119960

[18] The Social Media Monthly, "The Role of AI in CPU and GPU Design: How AI is Shaping Future Processors," The Social Media Monthly, 2025, https://thesocialmediamonthly.com/the-role-of-ai-in-cpu-and-gpu-design-how-ai-is-shaping-future-processors/

[19] Adrien Payong, "Future Trends in GPU Technology," DigitalOcean, 2024, https://www.digitalocean.com/community/conceptual-articles/future-trends-in-gpu-technology