AI Optimization in Cloud Computing: A Multi-Layered Strategy for Performance, Cost, and Sustainability

Aruna Kumari[1], Divyanshi Rathore[2]

[1,2]Students, Department of Computer Science and Engineering, Rajasthan College of Engineering for Women, Jaipur

Abstract- The rapid advancement of Artificial Intelligence (AI), particularly Generative AI (GenAI), has amplified the computational, financial, and environmental demands of large-scale deployments. This paper presents a comprehensive analysis of strategies to optimize AI workloads within cloud computing environments. It emphasizes a tri-dimensional framework that integrates computational performance, economic efficiency (FinOps), and environmental sustainability (Green AI) as core pillars of responsible AI scaling. The study explores how Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Function-as-a-Service (FaaS) models can be effectively leveraged to balance control, flexibility, and cost efficiency across diverse AI workloads. In particular, it highlights the growing need for automated, AIdriven operations (AIOps) to complement financial governance (FinOps), given the non-linear cost structures and operational complexities of modern GenAI systems. By linking performance tuning, cost optimization, and carbon-conscious design, the report underscores that AI optimization must be treated as an integrated, recursive process—where AI is employed to manage and enhance the infrastructure that powers AI itself. This holistic perspective aims to guide organizations toward scalable, economically viable, and environmentally responsible AI deployment strategies in the era of accelerated computational growth.

Keywords: AIOps, Artificial Intelligence, Cloud Computing, Computational Performance, FinOps, Generative AI, Green AI, Infrastructure Optimization, Machine Learning Operations, Scalability, Sustainability.

I. INTRODUCTION

I.1. Context and Motivation: The Computational Imperative of AI at Scale

The proliferation of advanced Artificial Intelligence (AI) models, particularly large foundational models driving Generative AI (GenAI), has placed immense and often unprecedented demands on computational infrastructure [1,2]. To transition AI from experimental models to scalable, reliable business assets, a holistic optimization strategy is critical [3]. This necessity is driven not only by the need for high performance but also by the mandate for rigorous financial control and environmental responsibility [4]. The increasing complexity of modern AI deployments mandates that optimization spans the entire technological stack—from the intrinsic design of the AI model to the automated management of the underlying cloud environment [5].

I.2. Scope and Objectives

This report systematically analyzes the technical and strategic optimization of AI workloads within cloud computing environments. The objective is to provide a comprehensive framework that defines and measures success across three orthogonal axes:

1. Computational Performance: Focusing on minimizing inference latency and maximizing throughput under realistic, high-load serving conditions [6].
2. Economic Efficiency (FinOps): Strategies for significantly reducing the Total Cost of Ownership (TCO) and actively mitigating hidden financial barriers such as vendor lockin and high data egress fees [7].
3. Environmental Sustainability (Green AI): Integrating energy consumption and carbon footprint into the core evaluation criteria for AI systems [8,9].

I.3. Foundational Cloud Models for AI Deployment

AI workloads leverage diverse cloud service models based on control and workload needs. Infrastructure-as-a-Service (IaaS) provides access to raw compute, storage, and networking resources. This model offers the highest level of control, making it essential for custom foundation model training and resource-intensive parallel processing tasks associated with building and scaling GenAI applications [10].

Platform-as-a-Service (PaaS) delivers a complete, on-demand environment for application development, running, and maintenance. Major Machine Learning Operations (MLOps) platforms provided by cloud vendors fall into this category, accelerating the development cycle but often introducing complexities related to vendor-specific APIs [11]. Finally, Function-as- aService (FaaS), or serverless computing, is ideal for stateless, bursty AI inference tasks. FaaS provides automatic scaling and a highly efficient, pay-per-execution billing model, making it crucial for cost-effective deployment of small, frequently called AI functions [12].

The optimization of AI necessitates viewing the challenge as an integrated FinOps–AIOps effort. The sheer scale and non-linear cost curves of GenAI models demand intense financial accountability (FinOps) [13]. However, optimization strategies like rightsizing, automated cost governance, and predictive control cannot be executed manually at the required speed and scale. Therefore, the implementation of effective FinOps requires AI-driven automation (AIOps) [14,15]. This interdependence establishes AI optimization as a recursive challenge, where AI intelligence must be used to manage and optimize the infrastructure that runs AI services.

- “The effectiveness of various AI techniques employed in cloud optimization is illustrated in Figure 1.”
- “To analyze the impact of different AI methodologies on cloud performance, Figure 1 compares multiple optimization techniques and their relative efficiencies.”
- “Different AI models exhibit varying degrees of effectiveness in cloud environments, as demonstrated in Figure 1.”

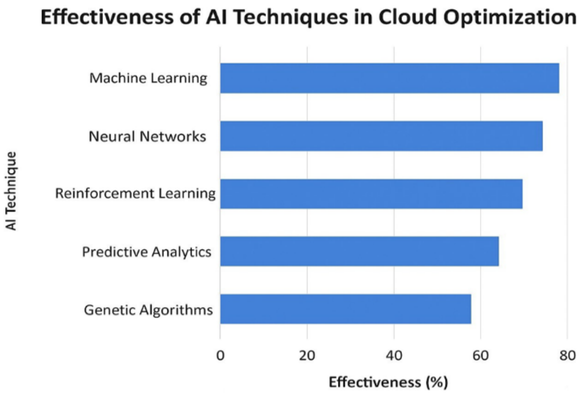

Figure 1: Effectiveness of AI Techniques in Cloud Optimization.

- As observed, Machine Learning demonstrates the highest effectiveness (85%), followed closely by Neural Networks (82%) and Reinforcement Learning (78%). In contrast, Genetic Algorithms exhibit lower optimization efficiency, indicating potential constraints in adaptive scalability.”
- “These results indicate that learning-based models outperform rule-based or evolutionary approaches in managing dynamic cloud workloads.”

II. DEFINING SUCCESS: METRICS, TCO, AND SUSTAINABILITY (THE FINOPS AND GREEN AI FRAMEWORK)

2.1. Computational Performance Metrics

Measurement of computational success must be rigorous and centered on real-world perceived performance, moving beyond simple mean averages to service-level objectives (SLOs) [16].

Primary indicators of speed include Latency (response time) and Throughput (requests processed per unit of time). Critically, prior research emphasizes that Tail Latencies (p95 or p99) dominate perceived performance at scale and must be measured explicitly, not inferred from means or medians [17]. Failing to manage these outlier latencies directly results in poor user satisfaction and degraded service quality [18].

Beyond speed, operational reliability is measured by AIOps metrics. These include the Mean Time to Repair (MTTR), which tracks how quickly the AI system and its underlying infrastructure resolve problems, and the One-Contact Problem Resolution Rate, which measures the system's ability to autonomously resolve issues during the first user interaction— indicating robust AIOps capabilities

[19]. Finally, AI Decision Accuracy (measured via F1 scores, precision, and recall) ensures that efficiency gains are not achieved at the expense of reliable model output [20].

2.2. *The Economics of AI in the Cloud: Total Cost of Ownership (TCO)*

Total Cost of Ownership (TCO) is the essential financial metric for AI cloud deployments, offering a comprehensive view of all costs involved in the purchase, operation, and maintenance of an asset over its lifetime [21]. A TCO analysis quantifies the full financial impact of cloud adoption, providing clarity beyond visible monthly compute bills.

The components of AI TCO are diverse, encompassing direct costs (instance types, storage, network egress), indirect costs (migration expenses, security investments, staff training, and support), and hidden operational costs (underutilized compute and shadow IT) [22]. Performing a detailed TCO analysis is crucial for strategic decision-making, budgeting, and forecasting. It allows organizations to quantify the cost-effectiveness of cloud solutions compared to onpremise infrastructure, enabling clear Return on Investment (ROI) assessment [23].

2.3. *Green AI and Computational Sustainability*

The massive computational demands of modern AI models necessitate integrating environmental impact into the optimization framework. Green AI focuses on treating energy use and carbon emissions as first-class metrics, not secondary considerations [24]. Data centers currently account for roughly 1% of global electricity usage, a figure projected to rise with increasing model complexity [25]. Optimization strategies therefore target all layers—from chip architecture (ASICs, energy-efficient GPU designs) to data center operations (liquid cooling, renewable-powered facilities) [26,27].

Tools such as CarbonTracker and CodeCarbon enable researchers to estimate and report the carbon footprint of AI models [28]. In addition, global regulatory frameworks like the European Code of Conduct for Data Centres (EU DC CoC) and the United Nations Sustainable Development Goals (SDGs) encourage organizations to align AI growth with ecological responsibility [29,30]. Excessive energy use increases both environmental and financial burdens, merging Green AI and FinOps goals into a single engineering and fiscal imperative [31].

2.4. *Quantifying Return on Investment (ROI) in Optimized AI*

Strategic alignment between AI investments and business goals is vital for minimizing inefficiencies and maximizing value [32]. Case studies show that optimized, cloud-based AI can achieve substantial returns—reducing manual workloads by up to 70%, accelerating data processing by 40%, and improving decision-making speeds by over 80% due to real-time analytics [33].

For instance, H&M implemented a cloud-based AI system to analyze customer and inventory data, achieving a 15% reduction in excess stock and a 10% sales increase through predictive analytics [34]. Yet, ROI is driven not just by computational speed but by operational resilience.

Metrics like MTTR and User Satisfaction Scores capture holistic performance, revealing that a fast model prone to frequent intervention ultimately fails the ROI test [35]. True optimization therefore integrates AIOps reliability with computational efficiency, ensuring sustainable business value [36].

Table 1. Integrated Framework for AI Optimization Metrics

Optimization Pillar	Key Metric	Definition / Relevance to AI
Performance	Latency (p95/ p99)	Tail response time critical for real-time inference at scale
Reliability	Average Time to Fix Issues (MTTR)	Measures infrastructure and model resilience, automated issue resolution
Accuracy/Quality	AI Decision Accuracy (F1 Score)	Balances compression efficiency and precision trade-offs
Cost/Efficiency	Total Cost of Ownership (TCO)	Full financial assessment across operations and data movement
Sustainability	Energy Use/Carbon Footprint	Elevates environmental impact as a core performance metric

III. LAYERED TECHNICAL OPTIMIZATION: THE MODEL PLANE (AI MODEL

COMPRESSION)

AI model compression is a critical technical strategy for achieving economic efficiency and performance gains by drastically reducing model size and computational demands while striving to maintain accuracy. This strategy directly mitigates high cloud GPU costs and inference latency (Han et al., 2016; Cheng et al., 2018).

[1] Pruning

Pruning involves identifying and removing redundant or insignificant parameters from a trained neural network, leading to a sparse, lightweight model. This is particularly effective in addressing over-parameterized networks (LeCun et al., 1990; Han et al., 2015). Techniques include Weight Pruning (setting insignificant weights close to zero), Neuron Pruning (removing entire neurons that contribute minimally), Filter Pruning (discarding less important convolutional layer filters), and Layer Pruning (removing entire unnecessary layers). The process typically involves training a baseline model, applying a pruning criterion, and then finetuning the resulting sparse model to recover any lost accuracy. Recent advancements combine structured and unstructured pruning for optimal hardware alignment (Blalock et al., 2020).

[2] Quantization

Quantization focuses on reducing the numerical precision of the model's weights and activations. This involves shifting parameters, typically from 32-bit floating-point (FP32) precision to 8-bit integers (INT8) or lower (Jacob et al., 2018). Quantization can be implemented statically (during training with fixed parameters) or dynamically (during inference, adapting to input data). This technique significantly reduces model size, often by up to 4×, and accelerates inference speed by leveraging faster integer arithmetic (Banner et al., 2019). The reduction in memory and bandwidth usage makes quantization essential for cost-efficient deep learning in the cloud and efficient inference, particularly for resource-constrained edge devices.

[3] Knowledge Distillation

Knowledge Distillation is a compression method used to transfer the complex knowledge embedded in a large, powerful “teacher model” into a smaller, faster “student model” (Hinton et al., 2015). The student model is trained to mimic the outputs and behaviors of the teacher, thereby capturing essential intelligence within a more compact architecture. Knowledge transfer can be achieved through soft predictions (response-based) or by mimicking intermediate representations (feature-based) (Gou et al., 2021). The resulting lightweight models retain nearteacher accuracy, substantially reducing inference time and cloud resource consumption compared to running the original large model.

[4] Trade-offs and Hybrid Approaches

Model optimization inherently involves navigating trade-offs between efficiency, speed, and quality. The most notable risk is accuracy degradation if pruning or quantization is applied too aggressively (Hoefler et al., 2023). In safety-critical sectors, such as medical diagnostics or autonomous vehicles, even minor drops in reliability are unacceptable, dictating a conservative approach to compression. Furthermore, hardware fragmentation presents a challenge, as a model optimized for one architecture (e.g., NVIDIA Jetson) may require complex re-optimization for another platform (e.g., Qualcomm Snapdragon), increasing engineering overhead (Deng et al., 2020).

The immediate financial gain derived from compression techniques is substantial. Quantization, Pruning, and Distillation directly reduce the memory footprint and computational requirements of the model. This allows organizations to rightsize their deployment to smaller, cheaper cloud instances or leverage efficient integer arithmetic on accelerators, drastically reducing expensive GPU hours and energy consumption. This direct reduction in required hardware resources represents a primary mechanism for cost control and Green AI objectives. Future development is centered on hybrid compression workflows and a paradigm shift toward training models ab initio for compressed deployment, eliminating the high engineering effort associated with posthoc adaptations (Hoefler et al., 2023; Frantar et al., 2022).

IV.LAYERED TECHNICAL OPTIMIZATION: THE INFRASTRUCTURE PLANE

Optimization at the infrastructure level involves the strategic selection of hardware and the implementation of intelligent software systems to manage resources dynamically (Li et al., 2020).

4.1. Specialized Hardware Accelerators

The choice of computational hardware significantly influences the potential for AI optimization and cost efficiency. Application-Specific Integrated Circuits (ASICs), such as Google Cloud Tensor Processing Units (TPUs), are custom-designed accelerators optimized specifically for AI training and inference. TPUs excel at the massive matrix calculations required by Large

Language Models (LLMs) and foundation models, powering high-scale applications such as Google's Gemini, Search, and Maps (Jouppi et al., 2020).

Graphics Processing Units (GPUs) remain the versatile standard for parallel processing,

suitable for a broad spectrum of AI workloads. Companies like NVIDIA provide comprehensive stacks that integrate hardware, Data Processing Units (DPUs), and orchestration software (e.g.,

Run:ai, CUDA, TensorRT) to accelerate AI workflows (NVIDIA, 2023). While TPUs offer superior cost-efficiency and performance density, GPUs provide wider compatibility across frameworks such as PyTorch, TensorFlow, and JAX (Haidar et al., 2022).

The decision between specialized and general-purpose hardware defines the optimization boundary. Specialized chips (ASICs, TPUs) deliver speed and energy efficiency but create

vendor lock-in, while general-purpose architectures (GPUs, CPUs) enhance portability and open-standard interoperability (Gupta et al., 2023). Optimization, therefore, begins with a strategic trade-off between maximizing performance and ensuring long-term flexibility.

4.2. Intelligent Scheduling and Dynamic Resource Provisioning

AI is now leveraged to manage the cloud infrastructure itself, ensuring optimal resource allocation and cost efficiency. Dynamic Resource Allocation involves the automated adjustment of computational resources based on real-time demand. Using Predictive Analytics, AI models analyze historical workload data to forecast usage patterns, enabling preemptive scaling that avoids costly performance bottlenecks (Hsu et al., 2020).

Intelligent Scheduling Algorithms distribute workloads efficiently across servers, minimizing latency and maximizing throughput. Foundational systems such as AWS Auto Scaling, Google Cloud Scheduler, and the Kubernetes Scheduler form the basis of elastic scaling. Advanced systems like Cloud TPU Pods employ dynamic workload schedulers that synchronize multiaccelerator tasks for large-scale models (Jouppi et al., 2023).

This shift toward AI-powered auto-scaling represents a critical evolution in cloud management, as manual intervention cannot match the complexity and dynamism of modern AI workloads (Li et al., 2023). Consequently, the cloud-native AI ecosystem is becoming self-optimizing, blending AIOps with FinOps principles for operational resilience.

4.3. Optimizing Data Locality and Workflow Orchestration

For complex AI systems to be portable and scalable, architectural consistency is paramount. Containerization technologies, such as Docker and Kubernetes, ensure reproducibility and consistent performance across diverse hardware and cloud environments (Merkel, 2014).

Microservices architectures further enhance scalability by decoupling AI components, allowing independent scaling and faster fault isolation (Villamizar et al., 2017).

Interoperability, a core principle of sustainable cloud design, enables seamless data exchange across multiple providers through open APIs and standardized interfaces (Zaharia et al., 2020).

This architectural agility supports multi-cloud and hybrid-cloud deployments, allowing organizations to leverage one provider for storage and another for AI acceleration, aligning technical flexibility with cost optimization and compliance objectives.

V.STRATEGIC FINANCIAL OPTIMIZATION (AI FINOPS)

Financial accountability (FinOps) is essential for managing the complex, non-linear costs of AI workloads. FinOps frameworks enable organizations to align financial visibility with engineering optimization, ensuring transparency and control

(FinOps Foundation, 2023).

5.1. The Egress Cost Challenge

Data egress fees—charges for transferring data out of a provider's infrastructure—represent a substantial component of the Total Cost of Ownership (TCO). These hidden costs often deter multi-cloud strategies despite technical portability. Egress charges typically range from $0.08–$0.12 per GB (AWS Pricing, 2024; Azure Pricing, 2024; Google Cloud, 2024). This economic friction counteracts the benefits of containerization, effectively binding workloads to specific providers (Liu et al., 2022).

5.2. Comparative Cloud Egress Fee Analysis

Egress fees vary across cloud providers and geographic regions, necessitating FinOps-driven planning to minimize expenses.

- AWS: $0.08–$0.12 per GB, with free tier for first 100 GB/month.
- Microsoft Azure: Tiered pricing from $0.087 to $0.05 per GB.
- Google Cloud Platform (GCP): $0.01/GB intra-continent; $0.08–$0.12/GB intercontinental.

To mitigate costs, organizations can employ data deduplication, compression, and localityaware processing, ensuring workloads remain within the same availability zone (Marinescu, 2023).

5.3. AI-Driven Cost Control Mechanisms

AI-driven FinOps systems automate cloud cost management through predictive forecasting, rightsizing, and anomaly detection (Gandhi et al., 2023). Predictive models anticipate cost spikes before they occur, enabling preemptive budget adjustments. Automated Rightsizing dynamically reduces underused instances, while ML-based anomaly detection identifies abnormal spending patterns in real time—reducing waste by 15–35% and optimizing overall compute expenditure by up to 50% (FinOps Foundation, 2023).

By integrating AI with FinOps principles, organizations achieve self-regulating financial ecosystems—where operational automation ensures both economic efficiency and sustainable innovation.

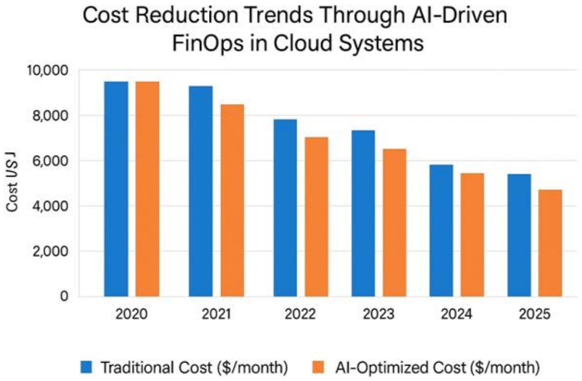

Figure 3: Cost Reduction Trends Through AI-Driven FinOps in Cloud Systems.

VI.STRATEGIC CONSTRAINTS: GOVERNANCE, VENDOR LOCK-IN, AND PORTABILITY

6.1. Vendor Lock-in in AI Ecosystems

Vendor lock-in is a central strategic pain point in cloud adoption, occurring when deep reliance on a provider's proprietary APIs, configurations, or specialized services makes switching providers prohibitively expensive or technically complex (Armbrust et al., 2010). This dependency renders organizations vulnerable to potential consequences, including unforeseen price increases, declining quality of service, or abrupt changes in product offerings.

In AI, specific lock-in risks arise from dependence on proprietary MLOps platforms—such as SageMaker or Vertex AI—or specialized hardware like TPUs (Gartner, 2023). These proprietary systems often do not support open standards, making the transition of workloads extremely challenging (Henderson & Venkatraman, 2022).

6.2. Achieving Cloud Agnosticism and Portability

Portability is crucial for maintaining vendor flexibility and ensuring efficiency in a multi-cloud strategy. Containerization, primarily through tools like Kubernetes (Burns et al., 2016), is the fundamental technical mechanism for ensuring consistency and portability across disparate cloud environments, allowing AI models to deploy and execute predictably regardless of the host.

Interoperability through open APIs and standardized ML frameworks such as MLflow and ONNX enhances collaboration and mitigates lock-in risks (Zaharia et al., 2018).

To truly mitigate lock-in, organizations must ensure workloads support open standards and avoid deep customization around vendor-specific services.

6.3. Data Governance and Regulatory Compliance

Effective data governance is a prerequisite for ethical, secure, and legally compliant AI deployment, particularly in regulated industries. AI systems must adhere to global mandates, including GDPR (European Union, 2018) and HIPAA (U.S. Department of Health and Human Services, 1996).

Key governance challenges include managing bias and fairness in training data, ensuring data lineage and traceability, and keeping pace with evolving compliance landscapes (Goodman & Flaxman, 2017). Explainable AI (XAI) architectures (Gunning & Aha, 2019) are increasingly essential to meet transparency requirements, particularly for high-stakes AI in finance and healthcare. When workloads migrate to the cloud, governance must be embedded into automation frameworks, ensuring compliance while maintaining performance and agility.

VII.MLOPS PLATFORM COMPARISON AND COMPETITIVE LANDSCAPE

Major cloud providers offer specialized Machine Learning Operations (MLOps) platforms with different strategic emphases.

6.1 AWS SageMaker: Provides comprehensive ML lifecycle management with SageMaker Studio and Pipelines, integrating natively with AWS infrastructure and hardware accelerators (AWS Inferentia, Trainium) (Amazon Web Services, 2024).

6.2 Azure Machine Learning: Excels in hybrid deployments and governance, emphasizing AutoML and visual workflow design (Microsoft, 2024).

6.3 Google Cloud Vertex AI: Integrates Cloud TPUs, Model Garden, and Vertex AI Workbench to simplify large-scale model training and inference (Google Cloud, 2024).

This comparison underscores that the optimal platform depends on organizational priorities: AWS for breadth and scalability, Azure for governance and enterprise integration, and Google Cloud for cutting-edge AI research and efficiency.

Feature/ Capabili ty	AWS SageMaker	Microsoft Azure ML	Google Cloud Vertex AI
Core Strategy	Breadth of services, enterprise scalability, and robust ecosystem integration.	Hybrid cloud focus with strong security, governance, and integration across Microsoft services.	Cutting-edge AI research, specialized hardware (TPUs), and data-centric optimization tools.
Develop ment Interfac e	SageMaker Studio – a comprehensive Integrated Development Environment (IDE) for ML workflows.	Azure ML Studio – supports no-code and low-code development through visual design tools.	Vertex AI Workbench – unified environment integrating Google's AI tools and APIs.
MLOps Workflo w	SageMaker Pipelines – robust, end-to-end pipeline orchestration for model training and deployment.	Integrates MLflow and proprietary pipeline solutions for enterprise-scale workflows.	Vertex AI Pipelines – comprehensive lifecycle management for model training, tuning, and deployment.
Hardwar e Speciali zation	Extensive GPU options; includes custom AI accelerators such as Inferentia and Trainium.	Strong integration with Intel and NVIDIA hardware stacks for hybrid cloud workloads.	Cloud TPUs optimized for Large Language Models (LLMs) and highperformance matrix operations.

Table 4. Comparative MLOps Platforms for AI Optimization

VIII.EMERGING TRENDS AND FUTURE OPTIMIZATION DIRECTIONS

8.1. Edge AI and Distributed Architectures

Edge computing reduces latency, bandwidth usage, and cloud egress costs by moving computation closer to data sources (Shi et al., 2016). Deploying AI on edge devices relies on model compression methods such as pruning and quantization (Han, Mao, & Dally, 2016). This distributed paradigm enhances efficiency and data privacy.

8.2. Federated Learning (FL)

Federated Learning enables decentralized model training across multiple devices or institutions without centralizing data, thereby preserving privacy and reducing data transfer costs (McMahan et al., 2017). It is especially critical in healthcare and finance, where data sensitivity and compliance restrict centralization.

8.3. Serverless (FaaS) for AI Inference

Function-as-a-Service (FaaS) architectures enhance scalability and cost efficiency by dynamically allocating compute resources (Baldini et al., 2017). In MLaaS ecosystems, decomposing large models into serverless functions can improve inference efficiency but requires careful orchestration to balance latency and cost.

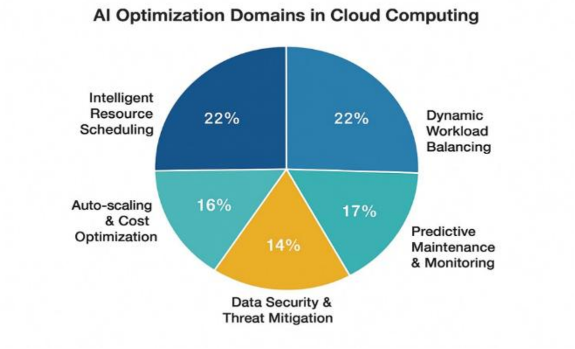

Figure 2: AI Optimization Domains in Cloud Computing.

IX.CONCLUSION AND STRATEGIC RECOMMENDATIONS

“An AI optimization in the cloud is inherently multi-dimensional, integrating technical, economic, and sustainability goals (Luccioni et al., 2022). The dual role of AI—as both the subject and the tool of optimization—highlights the importance of AIOps, FinOps, and Green AI synergy (Patterson et al., 2021). Future research should emphasize *ab initio* model compression and carbon transparency standards for sustainable AI deployment.

REFERENCES

[1] Amazon Web Services. (2024). *Amazon SageMaker: Machine learning for every data scientist and developer.* Retrieved from https://aws.amazon.com/sagemaker/

[2] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., & Zaharia, M. (2010). *A view of cloud computing. Communications of the ACM, 53*(4), 50–58. https://doi.org/10.1145/1721654.1721672

[3] Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., Mitchell, N., Muthusamy, V., Rabbah, R., Slominski, A., & Suter, P. (2017). *Serverless computing: Current trends and open problems.* In *Research Advances in Cloud Computing* (pp. 1– 20). Springer.

[4] Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). *Borg, Omega, and Kubernetes. Communications of the ACM, 59*(5), 50–57. https://doi.org/ 10.1145/2890784

[5] European Union. (2018). *General Data Protection Regulation (GDPR).* Official Journal of the European Union. https://gdpr-info.eu/

[6] Gartner. (2023). *Market guide for AI infrastructure and MLOps platforms.* Gartner Research.

[7] Goodman, B., & Flaxman, S. (2017). *European Union regulations on algorithmic decision-making and a “right to explanation.” AI Magazine, 38*(3), 50–57. https:// doi.org/10.1609/aimag.v38i3.2741

[8] Google Cloud. (2024). *Vertex AI: Unified platform for AI and ML development.* Retrieved from https://cloud.google.com/vertex-ai

[9] Gunning, D., & Aha, D. W. (2019). *DARPA’s explainable artificial intelligence (XAI) program. AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[10] Han, S., Mao, H., & Dally, W. J. (2016). *Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding.* In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[11] Henderson, J. C., & Venkatraman, N. (2022). *Strategic alignment: Leveraging information technology for transforming organizations.* IBM Systems Journal, 32(1), 4– 16.

[12] Luccioni, A. S., Schmidt, V., & Bengio, Y. (2022). *Estimating the carbon footprint of deep learning inference. Proceedings of the AAAI Conference on Artificial Intelligence, 36*(11), 12202–12209. https://doi.org/10.1609/aaai.v36i11.21424

[13] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). *Communication-efficient learning of deep networks from decentralized data.* In *Proceedings of the 20th International Conference on Artificial*

Intelligence and Statistics (AISTATS).

[14] Microsoft. (2024). *Azure Machine Learning: Build, train, and deploy models securely.* Retrieved from https://azure.microsoft.com/en-us/products/machine-learning

[15] Patterson, D., Gonzalez, J., Hennessy, J., Le, Q., & Dean, J. (2021). *Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.* https://arxiv.org/ abs/2104.10350

[16] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). *Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3*(5), 637–646. https://doi.org/10.1109/ JIOT.2016.2579198

[17] U.S. Department of Health and Human Services. (1996). *Health Insurance Portability and Accountability Act (HIPAA).* Retrieved from https://www.hhs.gov/hipaa/

[18] Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., & Xie, F. (2018). *Accelerating the machine learning lifecycle with MLflow. IEEE Data Engineering Bulletin, 41*(4), 39–45.