Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Choosing the Right Infrastructure Stack for Your AI Application: A Comprehensive Framework for Modern AI Systems

Reeshav Kumar

Independent Researcher, USA

ARTICLE INFO

Received: 01 Oct 2025

Revised: 02 Nov 2025

Accepted: 12 Nov 2025

ABSTRACT

The rapid growth of AI applications has fundamentally transformed how organizations approach system architecture and infrastructure design, creating a need for a comprehensive framework to select appropriate infrastructure components. This article presents a systematic methodology for understanding and implementing AI infrastructure through a layered architecture approach that decomposes the AI stack into six critical elements: (i) data and governance, (ii) storage systems, (iii) compute resources, (iv) model toolchains, (v) orchestration platforms, (vi) serving infrastructure including retrieval and augmentation systems, and observability and safety mechanisms. The article examines all six components along core AI system performance dimensions of quality, latency, throughput, cost, and energy efficiency, and analyzes inherent trade-offs. Unique bottlenecks and optimization strategies are identified for each infrastructure component, from data quality assurance challenges in governance layers to resource utilization inefficiencies in compute environments. This article presents a decision framework that encompasses workload characterization, technical evaluation criteria, economic analysis, and risk assessment, to enable AI practitioners to make informed infrastructure choices aligned with application-specific requirements and business constraints.

Keywords: AI Infrastructure, Layered Architecture, Performance Optimization, Decision Framework, Enterprise Architecture

[I] Introduction

The rapid growth of artificial intelligence applications across all sectors has not only redefined but also revolutionized how organizations design their infrastructure and system architecture. AI-based infrastructure scaling approaches have not only revolutionized cloud cost optimization strategies but also enabled organizations to achieve dynamic resource provisioning with minimal intervention, thereby transforming operational efficiency. The infrastructure underpinning AI systems can often make or break the final success or failure of deployments, as much as model selection and algorithmic tuning.

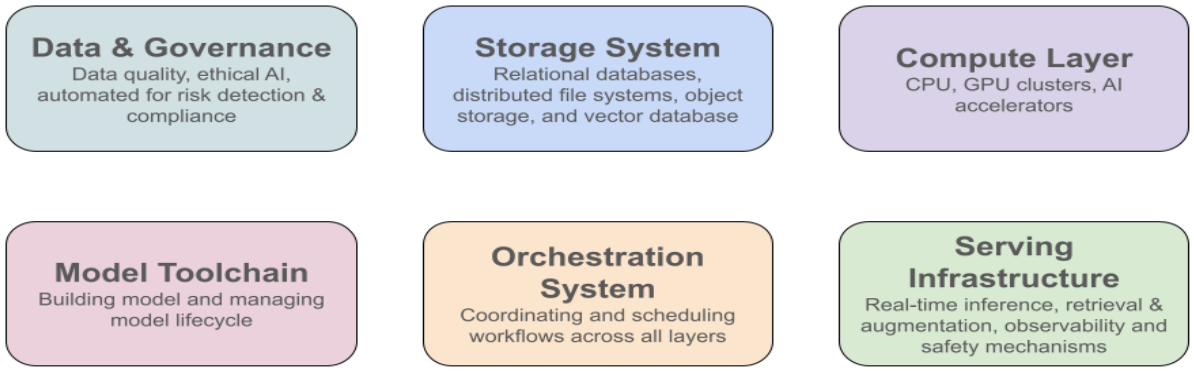

Fig. 1. Layers in enterprise AI infrastructure stack

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The complexity of contemporary AI infrastructure stems from the need to balance multiple competing objectives simultaneously. Organizations must optimize for model accuracy and quality while maintaining low latency for user-facing applications, controlling operational costs, and increasingly considering energy efficiency for sustainable computing practices. Next-generation scaling algorithms now utilize machine learning methods to forecast resource requirements, resulting in more efficient use of cloud computing resources and substantial cost savings for enterprise deployments. Enterprises implementing systematic methods for integrating AI into their IT strategies have significantly higher rates of successful implementations and operational efficiency levels [2]. These systematic methodologies also enable organizations to establish robust governance mechanisms and integrate AI seamlessly with existing enterprise workflows.

Today's AI workloads require advanced infrastructure stacks that can accommodate a range of workloads, including data preparation and model training, as well as real-time inference and ongoing monitoring. Intelligent scaling solutions have enabled businesses to automatically scale computational resources to match workload patterns, resulting in optimal cost-performance ratios in cloud environments. Enterprise AI implementation frameworks require careful strategic goal assessment, evaluation of technical requirements, and risk reduction approaches before deploying AI solutions at scale. Contemporary research on AI infrastructure optimization highlights the critical role of automated scaling mechanisms in achieving cost-effective cloud operations. In contrast, comparative studies of enterprise AI frameworks demonstrate the value of structured implementation approaches in ensuring successful technology adoption across diverse organizational contexts [1][2].

These mechanisms, which include tools and processes for monitoring and ensuring the safety of AI systems, are crucial for the success of an AI system.

[II] The Modern AI Infrastructure Stack: A Layered Architecture Approach

Production AI infrastructure stack comprises five critical underlying systems, each with unique constraints and optimization trade-offs that are crucial to the overall system's success. The enterprise AI infrastructure stack can be broken down into six unique layers, each targeting different aspects of AI system capabilities [3]. A solid understanding of constraints for each underlying system in the multi-layered AI architecture is essential for making informed choices to ensure success.

(i) **Data and governance**: At the foundation level, data and governance systems define the policies, procedures, and technical frameworks for ensuring data quality, compliance, and ethical AI practices. Robust and well-defined AI governance frameworks that guide data management capabilities and regulatory compliance processes are not only essential but also reassuring for the success of an enterprise AI system [3]. This layer includes data lineage tracing, bias detection and mitigation, model explainability frameworks, and regulatory compliance mechanisms. The governance layer has become increasingly vital as all enterprises face increasing scrutiny for AI fairness, transparency, and accountability. Product enterprise systems also use automated monitoring for real-time risk evaluation and compliance.

(ii) **Storage systems:** Provide persistent data management features needed to support large-scale AI operations. Cloud service optimization studies reveal that careful deployment of scalable storage architectures allows organizations to support exponentially increasing amounts of data while ensuring optimal performance attributes [4]. Storage systems comprise traditional relational databases, distributed file systems, object storage solutions, and vector databases, all of which are tailored for similarity search operations. The storage layer must support various types of data, including structured tabular data, unstructured text, images, and embeddings, with throughput and latency profiles that meet the needs of downstream applications.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

(iii) **Compute layer:** Comprises processing resources to drive AI workloads, including traditional CPU clusters, GPU accelerators, and dedicated AI accelerator chips (ASICs). Smart resource allocation and dynamic scaling controls are crucial for optimizing compute usage across AI workloads and demand profiles [4]. Many enterprises require compute layers that can accommodate both training workloads, which require massive parallel processing capacity, and inference workloads, which emphasize low latency and power efficiency. The compute layer also takes into account considerations of hybrid cloud, edge computing, and distributed processing architectures.

(iv) **Model toolchains:** Include tools and processes that manage the entire lifecycle of the AI model, from training, optimization, deployment, and monitoring. The model toolchain ensures the creation of scalable and reproducible models through efficient experiment tracking, version control, performance tuning, packaging, and deployment. Popular tools include PyTorch, MLflow, and TensorRT.

(v) **Orchestration systems**: Coordinate and automate execution of workflows across all layers of the AI infrastructure stack. Orchestration systems manage scheduling tasks, resource allocation, and inter-dependencies across the AI stack, ensuring all components operate in sync. Tasks managed by the orchestration system include scheduling and coordinating data pipeline management, model training, hyperparameter tuning, and continuous integration (CI)/continuous deployment (CD) processes, thereby increasing the efficiency, accuracy, and availability of AI systems. Orchestration systems are crucial for maintaining operational reliability and scalable performance in enterprise AI systems by coordinating across all system components [3]. Orchestration systems incorporate both real-time and batch processing paradigms, managing dependencies between various stages of workflows to ensure fault tolerance and recovery mechanisms. Examples include AirFlow and Kubernetes.

(vi) **Serving infrastructure**: Transform static models into responsive services and manage real-time inference, delivering trained models to end users and applications under various workloads and demands. Serving infrastructure also manages real-time inference, retrieval, and augmentation to enrich model inputs with relevant external data. Observability tools monitor model performance, latency, and resource usage, while safety mechanisms ensure content and reliability standards are met, maintaining trust and stability for large-scale deployment.

Layer Category	Technical Mechanisms	Operational Scope	Scalability Features	Compliance Requirements	Automation Level
Foundation (Data/Governance)	Automated Monitoring, Real-time Risk Assessment	Enterprise-wide Policy Implementation	AI-driven Governance Frameworks	Regulatory Compliance, Ethical AI	High Automation
Persistence (Storage)	Multi-format Data Handling, Distributed Architecture	Large-scale AI Operations	Exponential Growth Support	Data Protection Standards	Medium Automation
Processing (Compute)	Intelligent Resource Allocation, Dynamic Scaling	Distributed AI Workloads	Hybrid Cloud, Edge Computing	Energy Efficiency Standards	High Automation

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Model Toolchain (Training and Deployment)	Experiment Tracking, Version Control, Performance Tuning	End-to-End Model Lifecycle Management	Reproducible Model Pipelines	Model Governance Standards	High Automation
Coordination (Orchestration)	Fault Tolerance, Recovery Mechanisms	Complex Workflow Management	Batch and Real-time Processing	Operational Standards	Very High Automation
Serving Infrastructure (Responsive application)	Real-time Inference, Retrieval Augmentation, Load Balancing	Production Model Deployment	Auto-scaling, Multi-workload Support	Content Safety, Reliability Standards	Very High Automation

Table 1: Modern AI Stack Architecture: Component Analysis and Performance Characteristics [3, 4]

[III] Core Performance Dimensions and Trade-off Analysis

The effectiveness of an enterprise AI infrastructure stack needs to be assessed simultaneously across multiple dimensions, including quality, latency, cost optimization, and energy efficiency. The underlying systems in the AI infrastructure stack often present conflicting optimization objectives that must be balanced to ensure optimal system performance. Organizations implementing systematic resource efficiency frameworks achieve significant improvements in operational effectiveness and cost management across diverse technological environments [5]. Understanding these trade-offs is crucial for making informed architectural decisions that align with specific application requirements and business constraints, particularly as modern AI systems require increasingly sophisticated approaches to balancing competing performance objectives.

(i) Quality represents the fundamental measure of an AI system's effectiveness, encompassing model accuracy, precision, recall, and domain-specific performance metrics. Quality considerations extend beyond raw model performance to include data quality, feature engineering effectiveness, and the ability to maintain consistent performance across different user segments and use cases. Infrastructure decisions have a significant impact on quality through factors such as data pipeline reliability, model versioning and rollback capabilities, and the ability to perform comprehensive testing and validation [5].

(ii) **Latency** characteristics determine the responsiveness of AI applications, with requirements varying dramatically across different use cases. Studies on energy efficiency optimization in edge computing environments demonstrate that dynamic resource allocation strategies can significantly improve system responsiveness while maintaining optimal energy consumption patterns [6]. Real-time applications, such as autonomous vehicles or high-frequency trading systems, require sub-millisecond response times, whereas batch processing applications may tolerate latency measured in hours or days. Infrastructure components impact latency through network topology, data locality, caching strategies, model optimization techniques, and the choice between synchronous and asynchronous processing patterns. Metrics used to measure latency include average latency, percentile latency (such as p95, p99), and specific metrics for LLMs, including Time to First Token (TTFT) and Time Per Output Token (TPOT).

(iii) **Throughput** measures an AI system’s processing capacity using metrics such as tokens per second, requests per minute, or queries per second. Throughput affects scalability, cost, and performance under heavy load. While latency measures the speed of a single request, throughput

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

measures the total processing across multiple concurrent requests. A model can have low latency yet poor throughput when handling various users, which is crucial for scalable applications. Metrics used to measure throughput of AI systems include requests per second (RPS), tokens per second, or images per second.

(iv) **Cost** optimization encompasses both direct infrastructure expenses and indirect operational costs. Direct costs include compute resources, storage capacity, network bandwidth, and specialized hardware accelerators. Indirect costs involve personnel time for system maintenance, debugging, and optimization activities. The cost dimension presents complex trade-offs, as investments in higher-performance infrastructure may reduce operational overhead while increasing capital expenditure.

(v) **Energy efficiency** has emerged as a critical performance dimension driven by both environmental concerns and operational cost considerations. Research on dynamic resource allocation in edge computing environments demonstrates that intelligent energy management strategies can substantially reduce power consumption while maintaining system performance and reliability standards [6]. AI workloads such as LLM training and inference consume substantial energy resources. Infrastructure decisions, including hardware selection, data center location, cooling systems, and workload scheduling, have a significant impact on overall energy consumption. Organizations are increasingly considering energy efficiency as a primary optimization objective, alongside traditional performance metrics. Energy metrics for AI systems include Performance per Watt (PPW), which measures computations such as FLOPS or inferences per watt. Other key metrics include Carbon Intensity per AI task, measured by the CO_2 emitted per training run or inference batch, and energy consumption, measured in kilowatt-hours (kWh).

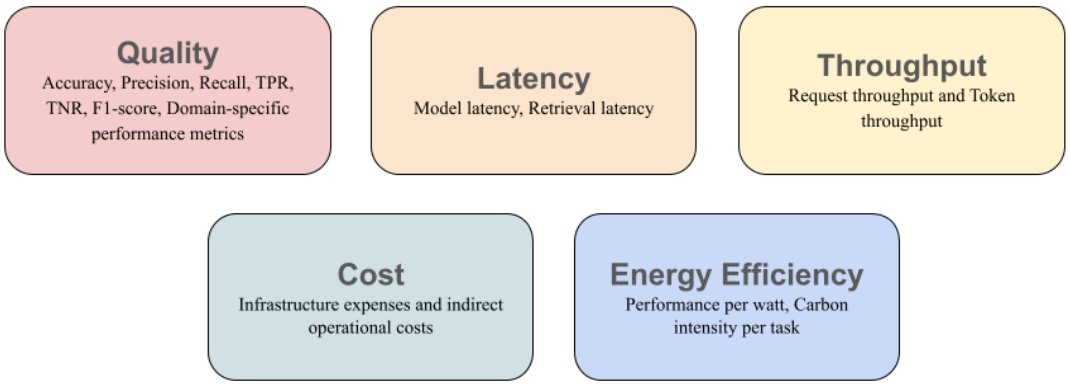

Fig 2: Core trade-off dimensions for AI infrastructure performance

Primary Dimension	Secondary Impact	Optimization Focus	Constraint Factor	Business Priority	Technical Complexity
Quality	Resource Utilization	High-Performance Standards	Infrastructure Reliability	Critical	High
Latency	Energy Consumption	Response Time	Network Architecture	Variable	Medium-High
Throughput	Scalability Requirements	Concurrent Processing	System Capacity	High	Medium

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Cost	Performance Level	Resource Efficiency	Budget Limitations	High	Medium
Energy	System Performance	Power Optimization	Environmental Standards	Growing	Medium-High

Table 2: AI Infrastructure Performance Dimensions: Trade-offs and Optimization Strategies [5, 6]

[IV] Layer-Specific Analysis: Bottlenecks and Optimization Strategies

Each layer of the AI infrastructure stack has specific bottlenecks and optimization opportunities that must be addressed with specialized methods and technologies. A study on optimizing AI performance in industry reveals that hybrid computing architecture methods, based on big data technologies, enable organizations to tackle layer-specific issues more effectively than conventional monolithic solutions [7]. Recognition of these layer-specific behaviors enables more precise decision-making and resource planning, particularly as firms strive to meet performance demands with efficiency in operating diverse AI workloads.

The data and governance layer experiences a bottleneck involving data quality validation, compliance checking, and bias discovery at scale. Legacy data quality tools often struggle to handle the volume and speed of data required in contemporary AI applications. In contrast, hybrid computing designs offer greater capabilities in addressing the needs for large-scale data processing and governance [7]. The most effective strategies in optimization are automated profiling of data, statistical anomaly detection, and machine learning-based monitoring of data quality. Governance bottlenecks typically arise due to human review processes and the absence of automated compliance checking. Solutions include integrating automated bias detection software algorithms, explainable AI frameworks, and real-time monitoring systems that can alert for potential problems.

Storage layer bottlenecks usually include throughput, consistency in distributed systems, and query performance for advanced analytical workloads. Conventional relational databases often struggle to scale effectively in managing the large data volumes and complex query patterns typical of AI applications, necessitating the development of novel data management and storage optimization techniques. Optimization techniques involve introducing distributed storage architectures, using special vector databases to support similarity search operations, and applying data partitioning and indexing optimization strategies tailored for AI workloads. Purpose-built AI databases evolving to meet the needs of AI applications resolve many historical storage bottlenecks but create new consistency models and query optimization considerations.

Compute layer bottlenecks focus on maximizing resource utilization efficiency, optimizing scheduling, and increasing the imbalance between computational demand and available resources. Optimization studies of AI workloads utilizing edge computing and cloud orchestration demonstrate that distributed computing techniques can significantly enhance resource utilization while meeting the computational demands of current AI applications [8]. GPU usage tends to remain suboptimal due to data loading bottlenecks, memory limitations, and inefficient parallelization techniques. Optimization techniques involve the use of dynamic resource allocation, utilizing specialized AI accelerators for specific types of workloads, and designing more effective model architectures that are less computationally intensive yet maintain similar quality.

Orchestration layers have bottlenecks associated with workflow complexity, dependency management, and fault tolerance. As the AI pipeline becomes more advanced, traditional workflow management tools are less effective in handling the dynamic nature of AI workloads and the requirement for real-time adaptability. Edge computing and cloud orchestration research proves that intelligent orchestration techniques can dynamically control sophisticated AI processes while adapting resource allocation across distributed environments [8]. Optimization techniques include adopting event-

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

driven architectures, leveraging containerization and microservices patterns, and creating intelligent scheduling algorithms that can adapt resource allocation across multiple competing workloads.

Strategy Type	Data/Governance Layer	Storage Layer	Compute Layer	Orchestration Layer	Technology Maturity
Automation	Automated Profiling, Bias Detection	Distributed Management	Dynamic Allocation	Event-driven Architecture	Mature
Architecture Innovation	Hybrid Computing	Vector Databases	Edge Computing	Microservices Patterns	Emerging
Performance Optimization	ML-based Monitoring	AI-optimized Indexing	Specialized Accelerators	Intelligent Scheduling	Developing
Scalability Solutions	Real-time Compliance	Purpose-built Databases	Distributed Computing	Container Orchestration	Mature

Table 3: AI Infrastructure Layer Bottlenecks: Analysis and Targeted Optimization Solutions [7, 8]

[V] Decision Framework for Infrastructure Component Selection

The choice of suitable infrastructure components should be guided by an organized assessment framework that considers both technological needs and organizational limitations. Studies on the formulation of decision-making structures for selecting infrastructure projects have shown that systematic methodologies during the front-end planning stages enable organizations to make better investment choices and deliver improved project performance [9]. The decision framework outlined here provides a systematic method for component choice that is flexible enough to be applied in various organizational settings and requirements, particularly in situations where limited resources and technical constraints necessitate careful consideration of trade-offs among conflicting goals.

The structure begins with the characterization of the workload through an exhaustive analysis of the volume of data, the processing pattern, latency requirements, and scalability expectations. Research on methodologies for selecting infrastructure projects emphasizes the importance of end-to-end front-end planning, which involves an extensive technical and economic evaluation of solution alternatives [9]. This characterization step must estimate crucial indicators, such as rates of data ingestion, model training frequency, inference request volumes, and expected growth patterns. Comprehending these baseline needs enables a more accurate estimation of various infrastructure alternatives, avoiding both over-engineering and under-provisioning of resources, particularly in situations where budget limitations and technical skill shortages necessitate prudent resource allocation decisions.

Technical evaluation factors include performance properties, integration potential, operational complexity, and vendor ecosystem considerations. Performance testing should involve benchmarking under a realistic workload, evaluating scalability limits, and analyzing failure modes and recovery. Integration evaluation would relate to compatibility with existing systems, the quality of APIs and documentation, and the availability of qualified personnel for implementation and maintenance. Periodic technical assessment procedures are essential for ensuring optimal system performance and resource efficiency [10].

Economic analysis is an essential part of the decision model, involving total cost of ownership calculations that go beyond up-front licensing or subscription charges. A cost analysis must account

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for infrastructure resources, staff time, training and certification expenses, and potential migration costs. The model needs to consider the cost implications of various scaling approaches, as well as the financial impact of performance bottlenecks on business processes. Infrastructure optimisation research stresses the need for thorough economic modelling that includes both direct and indirect costs of various technologies and implementation options [10].

Risk assessment encompasses vendor stability, technology maturity, security concerns, and compliance issues. Key risks to consider include a vendor's financial standing, aligning the vendor's roadmap with organizational requirements, and identifying alternative solutions in case migration becomes unavoidable. Security assessment emphasizes data protection features, access control measures, and compliance with applicable regulatory standards. The decision-making framework employs a systematic scoring approach that assigns weights to various criteria based on their organizational importance, enabling the objective ranking of different infrastructure alternatives while also accounting for subjective choices and strategic factors.

Assessment Category	Technical Weight	Economic Weight	Risk Weight	Strategic Weight	Organizational Priority
Workload Characterization	High	Medium	Low	High	Foundation
Performance Evaluation	Very High	Medium	Medium	Medium	Core Capability
Integration Assessment	High	Low	Medium	High	Operational Excellence
Cost Analysis	Medium	Very High	Low	High	Financial Sustainability
Security & Compliance	Medium	Low	Very High	Very High	Business Continuity
Vendor Ecosystem	Medium	Medium	High	Medium	Strategic Partnership

Table 4: Systematic Infrastructure Component Selection: Decision Framework Analysis and Evaluation Criteria [9, 10]

[VI] Conclusion

We present a systematic model to guide enterprises in selecting appropriate AI infrastructure components and making informed trade-off decisions that balance demands, performance requirements, and organizational limitations to achieve optimal system performance. The layered architecture approach also enables identification of bottlenecks within each infrastructure element, and can assist organizations in navigating trade-offs between quality, latency, cost, and energy efficiency. This decision paradigm encompasses workload characterization, technical analysis, economic analysis, and risk assessment, providing a formalized approach that can be tailored to meet the unique requirements of each organization. As AI systems become increasingly complex, the framework presented in this article can guide optimal infrastructure investments, enabling enterprises to achieve successful AI deployments while maintaining operational effectiveness and strategic coherence with their business goals.

Journal of Information Systems Engineering and Management
2025, 10(62s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

References

[1] Prasen Reddy Yakkanti, "AI-Driven Infrastructure Scaling for Cost Optimization in Cloud Environments: A Systematic Review," ResearchGate, March 2025. https://www.researchgate.net/publication/390325939_AI-Driven_Infrastructure_Scaling_for_Cost_Optimization_in_Cloud_Environments_A_Systematic_Review

[2] Nasir Kahan, "Frameworks for Implementing AI in Enterprise IT Strategy: A Comparative Study," ResearchGate, November 2024. https://www.researchgate.net/publication/393795192_Frameworks_for_Implementing_AI_in_Enterprise_IT_Strategy_A_Comparative_Study

[3] George Christopher et al., "Integrating Artificial Intelligence into Enterprise Architecture for Enhanced Scalability and Efficiency," ResearchGate, February 2025. https://www.researchgate.net/publication/388856249_Integrating_Artificial_Intelligence_into_Enterprise_Architecture_for_Enhanced_Scalability_and_Efficiency

[4] Saloni Sharma & Ritesh Chaturvedi, "Optimizing Scalability and Performance in Cloud Services: Strategies and Solutions," ResearchGate, December 2021. https://www.researchgate.net/publication/388799085_Optimizing_Scalability_and_Performance_in_Cloud_Services_Strategies_and_Solutions

[5] Ramachandran K K, "Optimizing IT Performance: A Comprehensive Analysis of Resource Efficiency," ResearchGate, December 2023. https://www.researchgate.net/publication/377598053_Optimizing_it_performance_a_comprehensive_analysis_of_resource_efficiency

[6] Mohan Harish Maturi et al., "Optimizing Energy Efficiency in Edge-Computing Environments with Dynamic Resource Allocation," ResearchGate, July 2024. https://www.researchgate.net/publication/381577573_Optimizing_Energy_Efficiency_in_Edge-Computing_Environments_with_Dynamic_Resource_Allocation

[7] Maya Utami Devi et al., "Optimizing AI Performance in Industry: A Hybrid Computing Architecture Approach Based on Big Data," ResearchGate, December 2024. https://www.researchgate.net/publication/387377833_Optimizing_AI_Performance_in_Industry_A_Hybrid_Computing_Architecture_Approach_Based_on_Big_Data

[8] Warren Liang, "Optimizing AI Workloads with Edge Computing and Cloud Orchestration," ResearchGate, February 2024. https://www.researchgate.net/publication/389097974_Optimizing_AI_Workloads_with_Edge_Computing_and_Cloud_Orchestration

[9] Seng Hansen et al., "Methods in Developing a Decision-Making Framework for Infrastructure Project Selection during Front-End Planning Phase in a Developing Country," ResearchGate, May 2018. https://www.researchgate.net/publication/337211816_Methods_in_Developing_a_Decision-Making_Framework_for_Infrastructure_Project_Selection_during_Front-End_Planning_Phase_in_a_Developing_Country

[10] Bamidele Matthew et al., "Infrastructure Optimization," ResearchGate, September 2025. https://www.researchgate.net/publication/395652588_Infrastructure_Optimization