RESEARCH ARTICLE

SAMRIDDHI Volume 18, Issue 3, 2026 Print ISSN: 2229-7111 Online ISSN: 2454-5767

AI inference routing and cost management

Prasad Prabhu

USA

Abstract

Artificial intelligence (AI) inference routing has become a fundamental capability for delivering scalable, responsive, and cost-efficient AI services across cloud, edge, and hybrid computing environments. As organizations increasingly deploy large language models and other computationally intensive AI applications, selecting optimal execution paths for inference requests is essential to maintaining service quality while controlling operational expenses. This study examines the principles, architectures, and optimization techniques that underpin AI inference routing and cost management. It explores routing strategies based on workload distribution, network awareness, adaptive scheduling, reinforcement learning, and collaborative cloud-edge inference. The discussion further evaluates cost drivers, including computational resource consumption, network bandwidth, storage utilization, energy demand, and model-serving overhead, alongside techniques such as dynamic scaling, request batching, model optimization, and intelligent resource allocation that improve efficiency. Performance is assessed using metrics such as latency, throughput, scalability, resource utilization, and cost per inference request. The study also identifies emerging trends, including agentic AI orchestration, heterogeneous infrastructure management, and 6G-enabled edge intelligence, which support more autonomous and resilient inference ecosystems. The findings demonstrate that intelligent routing combined with proactive cost management enhances system performance, improves infrastructure utilization, and promotes economically sustainable AI deployment while maintaining reliability, responsiveness, and high-quality inference outcomes across distributed computing environments.

Keywords: AI inference routing, cost management, cloud-edge computing, large language models, intelligent scheduling, resource optimization, adaptive routing, edge AI, scalable AI services, operational efficiency.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2026); DOI: 10.18090/samriddhi.v18i03.04

Introduction

Artificial intelligence (AI) inference has become the operational backbone of modern intelligent applications, enabling real-time predictions, decision support, content generation, and automation across industries. The rapid adoption of large language models (LLMs), computer vision systems, and multimodal AI has significantly increased inference workloads, creating substantial pressure on computing infrastructure, networking resources, and operational budgets. Consequently, efficient inference routing has emerged as a critical mechanism for directing requests to the most appropriate cloud, edge, or hybrid computing resources while maintaining low latency, high availability, and economic efficiency (Yu, Goudarzi, & Toosi, 2025; Parashar, 2026).

Traditional routing mechanisms primarily emphasize request distribution and server utilization. However, contemporary AI services require intelligent routing strategies capable of considering computational capacity, network conditions, model availability, energy consumption, and service-level objectives simultaneously. Recent advances integrate adaptive decision-making, reinforcement learning, and active inference techniques to continuously optimize request allocation under changing workloads and heterogeneous environments. These approaches improve responsiveness while reducing unnecessary computational overhead and infrastructure expenditure (Wang, Sedlak, & Dustdar, 2026; Zhang, Dong, & Ota, 2021).

The growing convergence of cloud computing, edge intelligence, and next-generation communication networks has further transformed AI deployment architectures. Rather than relying exclusively on centralized cloud servers, inference tasks are increasingly distributed across multiple execution layers to balance processing speed, bandwidth utilization, and resource efficiency. Collaborative cloud-edge inference enables latency-sensitive applications—including autonomous systems, industrial automation, healthcare monitoring, and smart cities—to deliver faster responses

Corresponding Author: Prasad Prabhu, USA, e-mail: prasadmp.net@gmail.com

How to cite this article: Prabhu, P. (2026). AI inference routing and cost management. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 18(3), 64-72.

Source of support: Nil

Conflict of interest: None

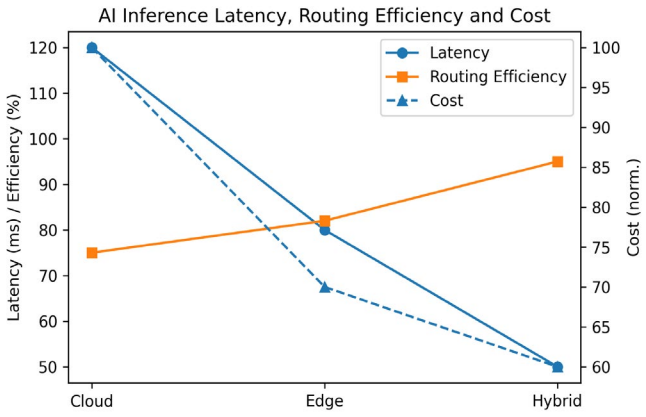

Figure 1: Hybrid cloud-edge deployment achieves the lowest inference latency and operational cost while maintaining the highest routing efficiency compared with cloud-only and edge-only architectures.

without sacrificing model accuracy. Emerging frameworks such as cost-aware collaborative inference and network-aware routing illustrate how intelligent orchestration can dynamically select optimal execution paths based on workload characteristics and infrastructure conditions (Zheng et al.; Fu, Qin, & Jiang, 2026; Farooq et al., 2026).

Cost management has become equally important as performance optimization. AI inference expenses extend beyond computational resources to include GPU utilization, storage requirements, network traffic, energy consumption, and model serving overhead. Organizations therefore seek routing mechanisms capable of minimizing operational costs while sustaining quality of service. Techniques such as adaptive load balancing, automated scaling, workload scheduling, and communication-efficient inference have demonstrated considerable potential for improving infrastructure utilization and reducing unnecessary expenditures (Jin & Yang, 2025; Chen et al., 2026).

Moreover, intelligent optimization methods inspired by reinforcement learning have demonstrated effectiveness in solving complex routing and allocation problems under dynamic operating conditions. Similar optimization principles have also been successfully applied in broader resource allocation and route optimization domains, highlighting their capability to improve efficiency while minimizing operational costs (Kodali, 2020; Zhang, Dong, & Ota, 2021).

This study examines AI inference routing and cost management by exploring routing architectures, optimization strategies, collaborative cloud-edge execution, and intelligent resource allocation techniques. It also evaluates the principal factors influencing inference performance and operational expenditure while highlighting emerging developments that support scalable, adaptive, and economically sustainable AI service delivery.

Fundamentals of AI Inference Routing

AI inference routing refers to the intelligent process of directing inference requests to the most appropriate computing resource based on workload characteristics, latency requirements, resource availability, network conditions, and operational objectives. As AI applications increasingly rely on distributed cloud-edge infrastructures and large language models (LLMs), efficient routing has become essential for achieving low response times, balanced resource utilization, and economical service delivery. Modern routing frameworks continuously evaluate infrastructure status, model availability, and communication overhead before assigning inference tasks, enabling organizations to improve both performance and cost efficiency (Yu, Goudarzi, & Toosi, 2025; Wang, Sedlak, & Dustdar, 2026).

AI Inference Architecture

AI inference architectures determine where prediction tasks are executed and how computational resources are coordinated across distributed environments. Traditional cloud-based inference offers virtually unlimited computational capacity and centralized management but may experience higher communication delays for latency-sensitive applications. Conversely, edge inference executes models closer to data sources, reducing transmission latency while supporting real-time analytics for Internet of Things (IoT), industrial automation, healthcare, and autonomous systems (Parashar, 2026; Fu, Qin, & Jiang, 2026).

Hybrid cloud-edge architectures combine the strengths of centralized and decentralized computing by dynamically selecting execution locations according to workload complexity, available resources, and network quality. This collaborative model allows lightweight requests to remain at the edge while computationally intensive inference tasks are redirected to cloud resources. Multi-model deployment further enhances flexibility by hosting different AI models across heterogeneous infrastructures, allowing routing mechanisms to select the most suitable model according to accuracy requirements, computational demand, and operational policies (Yu, Goudarzi, & Toosi, 2025; Zheng et al.).

Routing Strategies

Routing strategies govern how inference requests are distributed among available computing resources. Static routing relies on predefined allocation rules and performs effectively under predictable workloads but lacks adaptability when resource demand changes. Dynamic routing continuously monitors infrastructure conditions, enabling real-time adjustments that improve throughput and responsiveness during fluctuating traffic (Jin & Yang, 2025).

Load-aware routing considers processor utilization, memory availability, and queue length before assigning inference requests, preventing resource saturation and improving workload distribution. Network-aware routing

Table 1: Comparison of AI Inference Routing Approaches

Routing Method	*Core Principle*	*Advantages*	*Limitations*	*Typical Applications*
Static Routing	Uses predefined routing policies with fixed execution paths	Simple implementation, predictable behavior, minimal management overhead	Poor adaptability under changing workloads and network conditions	Enterprise AI services with stable traffic patterns
Dynamic Routing	Continuously adjusts request allocation according to real-time infrastructure status	Improves responsiveness, scalability, and resource utilization	Requires continuous monitoring and decision-making mechanisms	Cloud-native AI platforms and elastic inference services
Load-Aware Routing	Selects inference nodes based on processor, memory, and queue utilization	Prevents resource congestion and balances computational demand	Performance depends on accurate resource monitoring	High-volume LLM inference and distributed GPU clusters
Network-Aware Routing	Incorporates latency, bandwidth, congestion, and communication quality into routing decisions	Reduces transmission delay and improves user experience	Sensitive to dynamic network fluctuations	Edge AI, mobile networks, and AI-enabled communication systems
Reinforcement Learning-Based Routing	Learns optimal routing policies through continuous interaction with the environment	Self-optimizing, adaptive, and effective for complex infrastructures	Requires training time and computational resources	Autonomous cloud-edge orchestration and intelligent workload scheduling
Active Inference Routing	Predicts environmental changes and proactively adapts routing behavior	Enhances resilience, efficiency, and decision accuracy under uncertainty	Greater implementation complexity and higher computational requirements	Heterogeneous edge AI services, agentic AI ecosystems, and next-generation distributed inference platforms

incorporates communication latency, bandwidth, congestion levels, and transmission reliability to identify optimal execution paths, particularly within cloud-edge ecosystems and emerging AI-enabled communication networks (Farooq et al., 2026; Chen et al., 2026).

Recent research increasingly emphasizes reinforcement learning and active inference techniques for autonomous routing decisions. Reinforcement learning enables routing agents to learn optimal allocation policies from operational feedback, improving decision quality over time. Active inference extends this capability by continuously adapting routing behavior according to environmental uncertainty and changing infrastructure conditions, thereby supporting resilient and self-optimizing AI services (Zhang, Dong, & Ota, 2021; Wang, Sedlak, & Dustdar, 2026).

Cost Components

Effective AI inference routing requires a comprehensive understanding of the factors contributing to operational expenditure. Compute costs represent one of the largest expenses, particularly when deploying GPU-intensive large language models that process substantial volumes of inference requests. Network transmission costs increase as data moves between geographically distributed cloud and edge resources, while storage expenses arise from maintaining trained models, intermediate datasets, and cached inference outputs (Parashar, 2026).

Energy consumption has also become a critical consideration because AI inference workloads require considerable processing power across distributed infrastructures. Networking-aware optimization seeks to minimize unnecessary communication while improving overall energy efficiency, supporting environmentally sustainable AI deployment (Chen et al., 2026). Additional overhead originates from model serving, orchestration platforms, monitoring systems, and resource management services that maintain continuous availability. Consequently, cost-aware routing frameworks evaluate these components collectively to balance service quality with infrastructure expenditure while maximizing resource utilization (Zheng et al.; Fu, Qin, & Jiang, 2026).

AI Inference Cost Management Frameworks

Cost Drivers in AI Inference

AI inference cost management focuses on minimizing operational expenditure while preserving inference quality, response speed, and service availability. As AI workloads expand across cloud and edge infrastructures, several factors significantly influence deployment expenses. Compute resources, particularly GPU and accelerator utilization, remain the largest contributor because complex models require substantial processing capacity. Network transmission costs increase when inference requests and model outputs

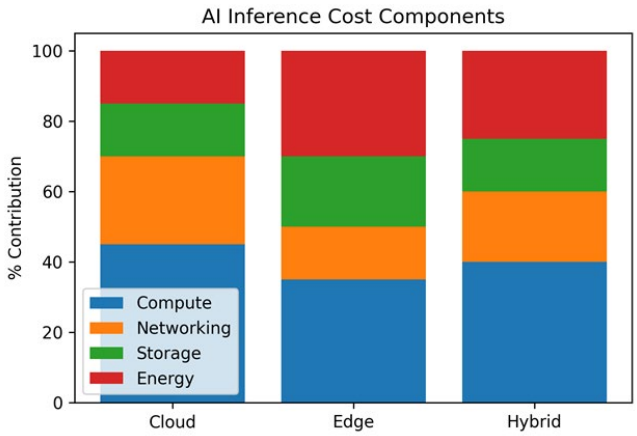

Figure 2: Hybrid cloud-edge deployment provides a more balanced distribution of compute, networking, storage, and energy costs, resulting in improved overall cost efficiency.

traverse distributed environments, especially in edge-cloud collaborations. Additional expenditures arise from storage requirements, memory allocation, model replication, and energy consumption associated with continuous inference operations (Parashar, 2026; Chen, X., Yu, H., Ni, W., Niyato, D., Zhang, R., Wang, X., ... & Xu, S., 2026).

The increasing adoption of large language models further amplifies operational costs through higher token processing demands, concurrent user requests, and heterogeneous hardware utilization. Consequently, intelligent routing mechanisms are essential for assigning inference tasks to the most economical computing resources while satisfying latency and quality-of-service requirements. Efficient request distribution reduces unnecessary resource consumption and improves infrastructure utilization across geographically dispersed computing nodes (Yu, S., Goudarzi, M., & Toosi, 2025; Zheng, W., She, J., Zhang, W., Chen, Y., Chen, L., Kundu, S., ... & Yao, H.).

Intelligent Cost Optimization Techniques

Modern AI inference platforms employ multiple optimization strategies to improve resource efficiency and reduce operational expenses. Adaptive workload allocation dynamically directs requests toward servers with available computational capacity, preventing resource bottlenecks and unnecessary overprovisioning. Elastic resource scaling automatically increases or decreases computing instances according to workload intensity, enabling organizations to pay only for required resources while maintaining consistent performance (Jin, Y., & Yang, Z., 2025).

Additional optimization methods include request batching, model quantization, model pruning, and

Table 2: AI Techniques for Routing Optimization

AI Technique	*Routing Objective*	*Performance Benefit*	*Cost Impact*	*Representative Reference*
Q-Learning	Learn optimal routing policies through continuous interaction	Lower latency, balanced traffic distribution, improved throughput	Reduces resource waste and infrastructure congestion	Zhang, Dong, & Ota (2021)
Reinforcement Learning	Dynamically optimize routing decisions under changing workloads	Adaptive load balancing and higher system scalability	Minimizes unnecessary compute allocation	Kodali (2020)
Predictive Traffic Management	Forecast workload demand before congestion develops	Stable response times and proactive resource allocation	Prevents over-provisioning and improves utilization	Jin & Yang (2025)
Active Inference Routing	Continuously adjust routing using environmental observations	Higher resilience and adaptive decision-making	Improves operational efficiency across heterogeneous systems	Wang, Sedlak, & Dustdar (2026)
Network-Aware Routing	Select routes based on communication quality and bandwidth	Lower transmission delay and improved Quality of Service	Reduces networking overhead	Farooq, Forgeat, Cyras, Bothe, & Chowdhury (2026)
Cost-Aware Edge-Cloud Routing	Allocate inference between edge and cloud resources	Balanced latency and inference accuracy	Minimizes communication and computing expenses	Zheng et al.
Agentic AI Routing	Coordinate autonomous routing decisions across distributed AI services	Intelligent orchestration, energy efficiency, and scalable management	Optimizes overall infrastructure expenditure	Chen et al. (2026)
Cloud-Edge Collaborative Routing	Partition workloads according to computational requirements	Improved scalability and resource utilization	Balances cloud and edge operating costs	Yu, Goudarzi, & Toosi (2025); Parashar (2026); Fu, Qin, & Jiang (2026)

Table 3. Performance Metrics for AI Inference Routing

Performance Metric	*Description*	*Optimization Objective*	*Business Impact*
Inference Latency	Time required to generate inference results	Minimize response delay	Improved user experience
Throughput	Number of requests processed per unit time	Maximize processing capacity	Higher service availability
Resource Utilization	Efficient use of CPUs, GPUs, TPUs, and memory	Increase infrastructure efficiency	Lower operational expenditure
Energy Efficiency	Energy consumed during inference execution	Reduce power consumption	Sustainable AI deployment
Cost per Inference	Financial cost associated with each inference request	Minimize execution expenses	Better return on investment
Routing Accuracy	Correct selection of execution destination	Improve routing decisions	Enhanced workload distribution
Scalability	Ability to accommodate increasing workloads	Maintain performance under demand growth	Enterprise readiness
Network Utilization	Efficient bandwidth consumption	Reduce communication overhead	Improved distributed performance

automated scheduling. Request batching processes multiple inference queries simultaneously, improving hardware utilization and lowering per-request costs. Quantized models reduce memory consumption and computational complexity without substantially affecting prediction accuracy. Intelligent schedulers continuously evaluate infrastructure status, network conditions, and workload characteristics before selecting optimal execution locations. Reinforcement learning further strengthens routing decisions by learning efficient allocation policies from historical system behavior, thereby reducing congestion and improving overall operational efficiency (Zhang, C., Dong, M., & Ota, 2021; Kodali, 2020).

Cloud-Edge Collaborative Inference

Cloud-edge collaborative inference combines centralized cloud resources with distributed edge devices to achieve balanced performance, scalability, and economic efficiency. Latency-sensitive inference tasks are executed closer to end users at the network edge, whereas computationally intensive operations remain within high-performance cloud infrastructures. This collaborative architecture minimizes communication delays, lowers bandwidth utilization, and optimizes infrastructure costs through intelligent workload partitioning (Fu, Y., Qin, P., & Jiang, 2026).

Emerging routing frameworks further enhance collaboration by considering network conditions, resource availability, service demand, and energy efficiency during routing decisions. Networking-aware optimization enables inference requests to follow execution paths that minimize communication overhead while maximizing throughput and reliability. Active inference routing and AI-assisted network orchestration continuously adapt routing decisions according to changing environmental conditions, supporting resilient and cost-effective AI services across heterogeneous infrastructures (Wang, Z., Sedlak, B., & Dustdar, 2026; Farooq, H., Forgeat, J., Cyras, K., Bothe, S., & Chowdhury, M. M. U., 2026). Consequently, cloud-edge collaboration establishes a practical foundation for sustainable AI deployment by balancing computational demand, operational expenditure, and service responsiveness across distributed computing ecosystems.

Intelligent Routing Optimization for Scalable AI Services

Machine Learning-Based Routing

Intelligent routing optimization enables AI inference platforms to efficiently distribute requests across cloud, edge, and hybrid infrastructures while maintaining low latency and minimizing operational expenses. Unlike static routing methods, machine learning-driven approaches continuously analyze workload characteristics, infrastructure utilization, network conditions, and application requirements to determine the most appropriate inference destination. This adaptive capability enhances service reliability, improves scalability, and maximizes computing resource efficiency, particularly in environments supporting large language models (LLMs) and distributed AI applications (Yu, Goudarzi, & Toosi, 2025).

Reinforcement learning has emerged as one of the most effective techniques for dynamic routing optimization. By learning from previous routing outcomes, reinforcement learning agents continuously refine decision policies to balance response time, computational efficiency, and infrastructure utilization. Q-learning further strengthens routing performance by selecting actions that maximize long-term rewards based on changing traffic patterns and resource availability, leading to improved throughput and reduced congestion across distributed inference systems

(Zhang, Dong, & Ota, 2021). Similar optimization concepts have also demonstrated significant benefits in large-scale logistics and network optimization problems, highlighting the effectiveness of reinforcement learning for complex routing decisions (Kodali, 2020).

Predictive traffic management enhances routing intelligence by forecasting workload fluctuations before congestion occurs. Using historical inference requests and real-time telemetry, predictive models proactively redistribute workloads, preventing resource bottlenecks while maintaining consistent service quality. Automated scaling mechanisms complement predictive routing by provisioning additional computing resources during traffic surges and releasing idle capacity when demand decreases, thereby improving cost efficiency without compromising performance (Jin & Yang, 2025).

Network-Aware Routing

Network-aware routing incorporates communication characteristics into inference decision-making rather than relying solely on computing capacity. Parameters such as bandwidth availability, transmission delay, packet loss, and network congestion are evaluated alongside processing capability to determine the most suitable inference location. This integrated approach reduces communication overhead while improving end-to-end service responsiveness for geographically distributed AI workloads (Farooq, Forgeat, Cyras, Bothe, & Chowdhury, 2026).

Modern routing frameworks increasingly utilize active inference principles that continuously observe environmental conditions and adapt routing decisions in response to evolving infrastructure states. Instead of applying predetermined routing rules, active inference-based systems estimate future network behavior and dynamically adjust request allocation to improve stability, resilience, and overall service performance across heterogeneous edge environments (Wang, Sedlak, & Dustdar, 2026).

Collaborative edge-cloud inference further strengthens routing effectiveness by partitioning inference workloads according to computational complexity and communication requirements. Lightweight inference tasks are executed at edge nodes to reduce latency, whereas computationally intensive operations are directed toward cloud resources with greater processing capability. Cost-aware collaborative routing systems intelligently balance these execution locations, minimizing communication expenses while preserving inference accuracy and responsiveness (Zheng et al.).

Emerging Technologies

Recent developments demonstrate a transition from conventional routing mechanisms toward autonomous AI-driven orchestration capable of coordinating multiple heterogeneous inference services simultaneously. Agentic AI introduces intelligent decision agents that evaluate workload priorities, network conditions, energy availability, and resource constraints before selecting optimal execution pathways. These autonomous agents continuously refine routing strategies through ongoing environmental feedback, enabling highly adaptive and self-optimizing inference ecosystems (Chen et al., 2026).

The emergence of 6G communication infrastructures further expands routing opportunities by providing ultra-low latency, high bandwidth, and intelligent network slicing specifically designed for AI applications. Task-oriented optimization allows routing engines to allocate inference requests according to application objectives rather than simple proximity, improving both computational efficiency and service reliability across mobile edge environments (Fu, Qin, & Jiang, 2026).

Edge-cloud integration also supports large-scale big data analytics by distributing inference tasks across multiple processing layers according to workload characteristics and infrastructure availability. Such architectures improve scalability, reduce resource contention, and provide flexible deployment options for enterprise AI applications requiring continuous real-time inference (Parashar, 2026).

Performance Evaluation, Challenges, and Future Directions

5.1 Performance Evaluation

Performance evaluation is essential for determining whether AI inference routing mechanisms can satisfy application requirements while minimizing operational expenditure. Modern cloud-edge AI platforms assess routing effectiveness using multiple indicators, including inference latency, throughput, resource utilization, energy efficiency, scalability, routing accuracy, and cost per inference request. These metrics collectively measure how efficiently requests are distributed across heterogeneous computing resources while maintaining consistent service quality. Efficient routing frameworks dynamically balance workloads among cloud servers, edge devices, and specialized accelerators to reduce congestion and improve response times (Yu, Goudarzi, & Toosi, 2025).

Recent routing architectures demonstrate that adaptive decision-making significantly enhances infrastructure utilization by continuously selecting execution locations based on network conditions, computational availability, and workload characteristics. Active inference models further improve routing precision by predicting environmental changes before allocating requests, thereby reducing unnecessary migrations and service interruptions (Wang, Sedlak, & Dustdar, 2026). Likewise, networking-aware optimization minimizes communication overhead while improving overall energy efficiency in distributed AI environments (Chen et al., 2026).

Scalability remains another important evaluation criterion because enterprise AI systems must process millions of

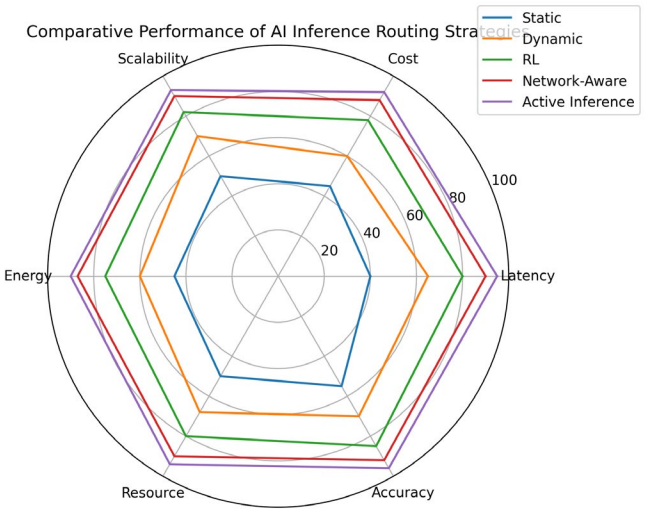

Figure 3: Active Inference Routing and Network-Aware Routing demonstrate the highest overall performance across routing metrics, whereas Static Routing exhibits the lowest adaptability and efficiency.

inference requests without degrading service quality. Automated load balancing, elastic resource allocation, and collaborative cloud-edge execution enable systems to maintain stable throughput despite fluctuating workloads (Jin & Yang, 2025; Parashar, 2026). Reinforcement learning-based routing further enhances long-term operational efficiency by continuously refining routing policies from observed traffic behavior (Zhang, Dong, & Ota, 2021).

Challenges

Despite substantial progress, AI inference routing continues to face several technical and operational obstacles. One major challenge involves infrastructure heterogeneity, where cloud platforms, edge servers, mobile devices, and specialized AI accelerators possess varying computational capabilities and networking characteristics. Designing routing algorithms capable of adapting to these differences without increasing latency remains difficult (Fu, Qin, & Jiang, 2026).

Another concern is fluctuating network conditions. Bandwidth availability, packet loss, and communication delays continuously change within distributed environments, making static routing strategies ineffective. Network-aware routing frameworks attempt to overcome these limitations by dynamically adjusting inference paths according to real-time communication quality (Farooq, Forgeat, Cyras, Bothe, & Chowdhury, 2026).

Cost optimization also presents considerable complexity because inference expenses extend beyond computational resources to include data transmission, storage, model synchronization, and energy consumption. Collaborative edge-cloud routing systems seek to balance these variables while maintaining acceptable inference quality, although identifying globally optimal routing decisions remains computationally demanding (Zheng et al.).

Security and privacy introduce additional concerns, particularly when inference requests involve confidential enterprise or healthcare information. Secure routing mechanisms must preserve data integrity while ensuring compliance with governance policies across distributed infrastructures (Parashar, 2026). Furthermore, highly dynamic AI workloads complicate resource forecasting, increasing the likelihood of over-provisioning or resource shortages during peak demand.

Future Directions

Future research is expected to emphasize autonomous routing systems capable of making intelligent decisions with minimal human intervention. Agentic AI is anticipated to transform inference management by enabling autonomous agents to coordinate routing, workload scheduling, and infrastructure optimization across distributed computing environments. Such systems will continuously monitor performance indicators and proactively adapt routing policies according to changing operational conditions (Chen et al., 2026).

Emerging 6G communication networks will further strengthen AI inference by providing ultra-low latency, intelligent network slicing, and improved edge connectivity. These capabilities will enable seamless execution of latency-sensitive AI applications such as autonomous transportation, industrial automation, and real-time healthcare analytics (Fu, Qin, & Jiang, 2026).

Future investigations should also prioritize sustainable AI by integrating carbon-aware routing, renewable energy utilization, and energy-efficient scheduling algorithms into inference platforms. Reinforcement learning, predictive analytics, and adaptive optimization techniques are expected to improve routing efficiency while reducing infrastructure costs and environmental impact (Kodali, 2020; Zhang, Dong, & Ota, 2021).

Another promising direction involves collaborative inference across multiple large language models operating simultaneously within cloud-edge ecosystems. Intelligent orchestration frameworks capable of selecting the most suitable model based on workload characteristics, quality requirements, and operational cost will significantly improve resource efficiency and service reliability (Yu, Goudarzi, & Toosi, 2025; Zheng et al.).

Conclusion

AI inference routing and cost management have become essential for supporting scalable, reliable, and economically sustainable AI services across cloud, edge, and hybrid computing environments. Intelligent routing mechanisms that consider workload distribution, network conditions, and

resource availability enable organizations to achieve lower latency, higher throughput, and improved infrastructure utilization while reducing operational expenditure (Yu, Goudarzi, & Toosi, 2025; Wang, Sedlak, & Dustdar, 2026). Cost-aware frameworks further strengthen deployment efficiency by balancing computational demand with communication overhead, energy consumption, and service quality, making collaborative edge-cloud inference increasingly practical (Zheng et al.; Chen et al., 2026).

The integration of reinforcement learning, adaptive scheduling, and automated scaling has demonstrated considerable potential for optimizing inference workflows under dynamic traffic conditions (Zhang, Dong, & Ota, 2021; Jin & Yang, 2025). Likewise, network-aware routing, AI-for-RAN solutions, and communication-efficient inference support more responsive decision-making in next-generation distributed infrastructures, particularly within emerging 6G ecosystems (Farooq et al., 2026; Fu, Qin, & Jiang, 2026). Complementary advances in cloud-edge processing and intelligent optimization also highlight the importance of flexible resource orchestration for real-time analytics and large-scale AI applications (Parashar, 2026; Kodali, 2020).

Overall, effective AI inference routing extends beyond request allocation to encompass strategic resource coordination, adaptive optimization, and financial efficiency. As AI workloads continue to increase in scale and complexity, integrating intelligent routing with cost-aware management will remain fundamental for delivering high-performance, resilient, and sustainable AI services across heterogeneous computing environments.

References

[1] Yu, S., Goudarzi, M., & Toosi, A. N. (2025). Efficient Routing of Inference Requests across LLM Instances in Cloud-Edge Computing. *arXiv preprint arXiv:2507.15553*.

[2] Wang, Z., Sedlak, B., & Dustdar, S. (2026). Active Inference-Based Adaptive Routing for Heterogeneous Edge AI Services. *arXiv preprint arXiv:2604.17373*.

[3] Chen, X., Yu, H., Ni, W., Niyato, D., Zhang, R., Wang, X., ... & Xu, S. (2026). Networking-Aware Energy Efficiency in Agentic AI Inference: A Survey. *arXiv preprint arXiv:2604.07857*.

[4] Farooq, H., Forgeat, J., Cyras, K., Bothe, S., & Chowdhury, M. M. U. (2026, May). STAR: Smart Network-Aware AI-for-RAN Routing. In *IEEE INFOCOM 2026-IEEE Conference on Computer Communications* (pp. 1-6). IEEE.

[5] Zheng, W., She, J., Zhang, W., Chen, Y., Chen, L., Kundu, S., ... & Yao, H. CLEAR: A Cost-Aware Routing System for Edge-Cloud Language Model Collaborative Inference.

[6] Fu, Y., Qin, P., & Jiang, C. (2026). Edge AI Inference in 6G Mobile Networks: From Communication-Efficient Techniques to Task-Oriented Optimization. *IEEE Communications Surveys & Tutorials*.

[7] Zhang, C., Dong, M., & Ota, K. (2021). Employ AI to improve AI services: Q-learning based holistic traffic control for distributed co-inference in deep learning. *IEEE Transactions on Services Computing*, *15*(2), 627-639.

[8] Parashar, P. (2026). Edge and Cloud-Based AI Inference for Big Data Processing. In *Harnessing AI Inference for Intelligent Decision-Making in Real-Time Dataflows* (pp. 1-28). IGI Global Scientific Publishing.

[9] Jin, Y., & Yang, Z. (2025, January). Scalability optimization in cloud-based AI inference services: Strategies for real-time load balancing and automated scaling. In *Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network* (pp. 266-270).

[10] Kodali, S. (2020). Optimizing Supply Chain Network Design with AI: Utilizing Reinforcement Learning and Advanced Analytics for Demand Forecasting, Route Optimization, and Cost Minimization. *American Journal of Cognitive Computing and AI Systems*, *4*, 1-36.

[11] Ezeagwuna, D. (2026). AI-Driven Intrusion Detection Systems for Cybersecurity. *Journal of Science Technology and Social Transformation*, *2*(02), 37-51.

[12] Mukherjee, C. (2026). AI-Based Detection of Deepfakes and Misinformation on Social Media. *Euro Vantage journals of Artificial intelligence*, *3*(2), 9-30.

[13] Mehta, S. (2026). Architecting AI Governance & 'Agent-ready'data layer: Semantic Governance for Autonomous AI workflows. *Journal of Science Technology and Social Transformation*, *2*(02), 29-36.

[14] Ezeagwuna, D. (2026). Digital Transformation and Business Resilience: How Service Now-Powered Automation Improves Organizational Performance, Risk Management, and Customer Satisfaction. *Journal of Data Analysis and Critical Management*, *2*(01), 48-62.

[15] Rehan, H., Akram, W., & Khan, M. B. S. (2026, May). Leveraging Gated Recurrent Units for Real-Time Tracking of Student Engagement in Healthcare LMS using LLM Analytics. In *2026 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-6). IEEE.

[16] Mehta, S. (2026). From Reactive cleanup to Real-time observability: Modernizing Enterprise Data Quality for RAG pipelines. *Journal of Science Technology and Social Transformation*, *2*(01), 35-44.

[17] Ezeagwuna, D. (2026). Enhancing Organizational Cyber Resilience Through the NIST Cybersecurity Framework. *International Journal of Technology, Management and Humanities*, *12*(01), 124-140.

[18] Kola, J. N. (2026, June). Secure Financial Analytics Platforms for Regulatory Compliance and Risk Management. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 794-801). IEEE.

[19] Rehan, H., Zimmerman, Z., & Janjua, J. I. (2026, May). Optimizer-Driven Techniques for Retuning Prompts on LLM Backend with Migration across Models for Healthcare and Cyber Security Operations. In *2026 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-6). IEEE.

[20] Kola, J. N. (2026). Enterprise Business Intelligence and Financial Analytics with a focus on Enterprise Data Architecture and Predictive Risk Integration. *Adhyayan: A Journal of Management Sciences*, *16*(01), 20-28.

[21] Ravikumar, V. (2025). Therapeutic Bot: Ethical Concerns in AI therapy for Neurodivergence. *J Int Scient Re Rep*.

[22] Mehta, S. (2025). Enhancing Enterprise Data Governance Maturity Through Cloud Based Data Integration Frameworks in Modern Organizations. *International Journal of Technology, Management and Humanities*, *11*(01), 59-67.

[23] Rehan, H., Janjua, J. I., Ali, R. A., & Khan, T. A. (2025, December). AI-Driven DNA Insights and Public Health Data Integration for

Enhancing Personal Healthcare in Critical Disease Prevention. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-6). IEEE.

[24] Ezeagwuna, D. (2025). Artificial Intelligence for IT Service Management: Developing a Framework for ROI Optimization Using Service Now and Machine Learning Technologies. *Well Testing Journal*, *34*(S4), 394-419.

[25] Rehan, H., & Janjua, J. I. (2025, October). An Intelligent Data-Driven Framework for Measuring Consumer Experience in the Digital Platforms. In *2025 IEEE International Conference on Computing (ICOCO)* (pp. 500-505). IEEE.

[26] Mehta, S. (2025). Role of Metadata Management and Data Governance in Achieving Regulatory Compliance in Enterprise Data Ecosystems. *International Journal of Technology, Management and Humanities*, *11*(02), 144-152.

[27] Warren, Brandon. (2025). BUSINESS VALUE DRIVEN BY ENTERPRISE ARCHITECTURE TRANSFORMATION BRANDON WARREN. Journal of Tianjin University Science and Technology. 58. 10.5281/zenodo.20280991.

[28] Adepoju, S. A., & Adepoju, M. A. (2024). From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration.

[29] Khan, H. A. (2024). DATA-DRIVEN EPIDEMIOLOGICAL MODELING USING MACHINE LEARNING FOR DISEASE SPREAD FORECASTING AND PUBLIC HEALTH DECISION SUPPORT IN THE UNITED STATES. *International Journal of Applied Mathematics*, *37*(6s), 178-192.

[30] Kola, J. N. (2024). Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention. *Journal of Information Systems Engineering & Management*, *10*.

[31] Mehta, S. (2024). Architecting the Hub-and-Spoke Governance Model: Balancing Centralized Guardrails with Federated Domain Agility. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *16*(04), 206-215.

[32] Ezeagwuna, D. (2024). Enterprise Workflow Automation and Workforce Productivity: Evaluating the Economic Benefits of Service Now Adoption Across Industries. *International Journal of Technology, Management and Humanities*, *10*(04), 299-313.

[33] Kola, J. N. (2024). Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention. *Journal of Information Systems Engineering & Management*, *10*.

[34] Das, P. K., Kashem, M. A., Ferdus, Z., & Islam, S. (2019, October). Development and application of a new computerized smell generating system. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-5). IEEE.

[35] Ferdus, M. Z., Khan, M. N. I., Islam, S., & Kashem, M. A. (2019, October). VFLT: SQA Model for Cyber Physical System. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-4). IEEE.

[36] Thakkar, V. B. (2026). From linear logistics to neural supply chains: Predictive machine learning and the rise of autonomous supply chain intelligence. *International Journal of AI, BigData, Computational and Management Studies*, *7*(1), 153-161.

[37] Suseelan, S. (2026). LLM-Based Human-machine Teaming for Air Traffic Controllers. *International Journal of Humanities and Information Technology*, *8*(1), 122-129.

[38] Verma, A. (2022). Twin Poisoning: Analyzing the Impact of False Data Injection Attacks on Digital Twin-Based Decision Support Systems. *International Journal of Technology, Management and Humanities*, *8*(03), 39-61.

[39] Thakkar, V. B. (2026). A Unified Framework for Integrating Generative AI Across the Agile Software Development Lifecycle (SDLC). *International Journal of Humanities and Information Technology*, *8*(2), 11-18.

[40] Suseelan, S. (2026). Multimodal AI for Pilot Skill Assessment Using Physiological Signals. *International Journal of Technology, Management and Humanities*, *12*(02), 74-83.

[41] Thakkar, V. B. (2026). Asynchronous Generative AI: Automating Qualitative Data Workflows in Enterprise Systems. *International Journal of Science, Technology & Society*, *10*(01), 1-14.

[42] Ferdus, M. Z., Islam, S., & Kashem, M. A. (2019, October). An innovative load balancing cluster composition of wireless sensor networks. In *2019 global conference for advancement in technology (GCAT)* (pp. 1-4). IEEE.

[43] Islam, S. J., Islam, S., Ferdus, M. Z., Khan, M. N. I., Kashem, M. A., & Islam, M. S. (2020, September). Load compactness and recognizing area aware cluster head selection of wireless sensor networks. In *2020 International conference on computing and information technology (ICCIT-1441)* (pp. 1-4). IEEE.

[44] Thakkar, V. B. (2025). Algorithmic Trust, Governance, and Integration Bottlenecks of AI Tools in Legacy Project Environments. *International Journal of Humanities and Information Technology*, *7*(01), 80-89. Mazumder, P. T. (2025). Harnessing fintech innovations for anti-money laundering: a data-driven approach. *Available at SSRN 5259084*.

[45] Suseelan, S. (2025). Big Data Analytics in Air Traffic Management: Enhancing Aviation Safety, Operational Efficiency, and Predictive Decision-Making. *Journal of Science Technology and Social Transformation*, *1*(02), 61-73.

[46] Islam, S., Khan, M. N. I., Ferdus, M. Z., Islam, S. J., & Kashem, M. A. (2020, September). Improving throughput using cooperating TDMA scheduling of wireless sensor networks. In *2020 International Conference on Computing and Information Technology (ICCIT-1441)* (pp. 1-4). IEEE.

[47] Naidu, K. J. (2013). Performance Optimization Of ETL Pipelines In Distributed Data Warehouse Environments: A Network-Aware Scheduling Approach. *International Journal of Advance Industrial Engineering*, *1*(03), 63-67.

[48] Ravikumar, V. (2014). Fair and optimal resource allocation in wireless sensor networks.

[49] Naidu, K. J. (2014). Secure OLAP Reporting Architectures: Integrating Role-based Access Control and Query Execution Plan Optimization for Enterprise Analytical Environments. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *5*(02), 155-159.

[50] Kola, J. N. (2011). An Integrated Framework for Data Mining and Distributed Database Optimization in Resource-Constrained Network Environments. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *2*(02), 82-86.