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

**Research Article**

# AI-Powered Dynamic Flow Steering in 5G Networks: RAN Analytics Integration with UPF for Enhanced QoS Management

Binu Kiliamkavunkal Govindan
Independent Researcher, USA

**ARTICLE INFO**

Received: 05 Dec 2025

Revised: 08 Dec 2025

**ABSTRACT**

This article presents an integrated framework for AI-powered dynamic flow steering in 5G networks that leverages real-time Radio Access Network (RAN) analytics to optimize traffic management through the User Plane Function (UPF). The technical architecture establishes seamless integration between Network Data Analytics Function (NWDAF) components and distributed UPF instances, enabling intelligent routing decisions based on current radio conditions. Machine learning algorithms, particularly reinforcement learning models, provide adaptive optimization capabilities that continuously refine steering policies through closed-loop feedback mechanisms. Performance validation across diverse deployment scenarios demonstrates that this approach maintains superior throughput and latency metrics compared to traditional management techniques, especially in challenging high-density environments and industrial IoT applications. The integration of predictive congestion management with computational offloading strategies creates a comprehensive optimization framework that addresses both network efficiency and application performance, establishing a foundation for future-ready, user-centric network experiences in evolving 5G architectures.

**Keywords:** Dynamic Flow Steering, Radio Access Network Analytics, Machine Learning Optimization, Quality of Service Management, Network Data Analytics Function

## 1. Introduction and Background

Alongside unmatched technological difficulties, fifth-generation wireless networks offer transformational potential. While trying for quick deployment schedules, telecommunication companies all over struggle with difficult resource allocation. Current 5G architectures feature intricate multi-layered designs that create operational environments vastly different from previous generations, demanding sophisticated traffic management approaches [1]. Conventional rule-based systems performed passably well in less complex network designs, but they show insufficient management of dynamic, high-density 5G environments where resource rivalry grows during peak demand periods.

More than just a small upgrade, the change from 4G ushers in basic architectural changes including network slicing, virtualized capabilities, and dense small cell installations. These innovations enable breakthrough applications while simultaneously creating complex management scenarios. Network slicing requires strict performance isolation between virtual networks operating on shared physical infrastructure, substantially increasing configuration complexity compared to conventional architectures [2]. Operators must now coordinate multiple network slices while supporting diverse service requirements, creating significant increases in administrative overhead and resource coordination challenges.

Service-based 5G core architectures require more sophisticated management than legacy systems. These frameworks introduce specialized network functions with distinct operational

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

roles, creating numerous interaction points and potential bottlenecks. The 3GPP Release 16 specifications define extensive standardized interfaces—far more than previous generations included. Each interface represents a potential failure point requiring continuous monitoring to preserve end-to-end service quality. Operational experience reveals troubleshooting complications stemming from subtle interactions between multiple network elements rather than isolated component failures [1].

Radio Access Network optimization methods demonstrate critical limitations in dynamic 5G environments despite recent technological progress. Current approaches depend heavily on reactive strategies, analyzing historical performance data to implement manual or semiautomated parameter modifications. These methods work reasonably well with predictable traffic patterns but fail completely when addressing microsecond-level fluctuations characteristic of latency-critical 5G applications. Urban deployment analysis shows reactive methods produce substantial quality variations during rapid condition changes, particularly impacting applications with stringent performance requirements [1].

Existing systems suffer from fundamental architectural separation between RAN analytics and core network components handling traffic routing and policy enforcement. This disconnection creates scenarios where routing decisions occur without real-time radio condition awareness, directly affecting user experiences. When cell congestion develops during peak periods, the User Plane Function continues directing traffic through overloaded paths while remaining unaware of deteriorating radio conditions. Metropolitan network studies document how this separation produces suboptimal performance despite adequate overall capacity, highlighting the need for integrated analytics bridging radio conditions with core network functions [1].

The expanding ecosystem of connected devices and application requirements exposes additional weaknesses in static optimization approaches. Industrial IoT deployments, augmented reality platforms, and autonomous vehicle communications present dramatically different performance needs that conventional Quality of Service classifications cannot adequately address without granular, context-aware management. Network slicing implementation studies reveal that existing management frameworks struggle to maintain consistent performance boundaries between slices during resource constraints, especially when multiple slices experience simultaneous peak demand [2]. These restrictions become clear in heterogeneous settings where mission-critical applications use the same infrastructure as best-effort traffic.

Effective answers for these optimization problems are found in machine learning approaches able to handle complicated real-time data. AI-driven dynamic flow control represents an emerging methodology that leverages continuous RAN information to make intelligent, adaptive routing decisions for individual sessions or packets. This approach enables networks to respond immediately to changing radio conditions, proactively redirecting traffic from congested areas or prioritizing critical data flows during resource limitations. Laboratory evaluations of AI-powered optimization show significant improvements in aggregate throughput and individual session stability compared to traditional management approaches, particularly under highly variable traffic conditions [1].

The Network Data Analytics Function established in 3GPP specifications provides the architectural foundation for implementing AI-driven optimization strategies. NWDAF collects data from various network functions, applying machine learning algorithms to generate actionable insights for policy and routing decisions. By feeding these analytics directly to the UPF, operators can establish closed-loop optimization that continuously adapts to network

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

conditions without manual intervention [2]. NWDAF deployments in network slicing situations show how analytics-driven resource allocation significantly enhances slice isolation and performance predictability over static allocation techniques. In settings with erratic traffic patterns—such as metropolitan areas, transportation terminals, or amusement venues where user counts vary greatly—this feature proves particularly useful. Advanced dynamic flow steering solutions use reinforcement learning algorithms that constantly improve route decisions via ongoing network environment contact. These algorithms evaluate outcomes from previous routing choices against established performance objectives, updating their models to optimize specific metrics including throughput, latency, or energy efficiency. Reinforcement learning applications in network optimization demonstrate the capability to identify unexpected optimization strategies that outperform manually designed policies, particularly in complex multi-variable environments typical of contemporary 5G deployments [1]. The self-adaptive nature of these systems enables response to evolving traffic patterns and network configurations without extensive recalibration or operator intervention.

This research examines implementation considerations and performance implications of AIpowered dynamic flow steering through integrating real-time RAN analytics with UPF functionality. Primary objectives include quantifying performance improvements achievable through dynamic flow steering across diverse deployment scenarios, evaluating architectural requirements for effective NWDAF analytics and UPF routing function integration, comparing different machine learning algorithms for network optimization, and establishing frameworks for predictive congestion management. Preliminary testing of analytics-driven optimization in network slicing environments indicates considerable potential for performance enhancement, particularly for applications with strict quality requirements [2]. The implications extend beyond theoretical advancement to address practical 5G network operation and evolution challenges, directly supporting transformative application enablement while maximizing infrastructure utilization efficiency.

| Optimization Method | Response Time | Adaptation Capability |
|---|---|---|
| Rule-based Systems | 10-30 seconds | Static configuration |
| Semi-automated Tuning | 5-15 seconds | Limited parameter adjustment |
| AI-driven Dynamic Control | 50-200 ms | Real-time adaptive routing |

Table I: Comparison of Network Optimization Approaches. [1, 2]

### 2. Technical Framework for RAN-Integrated Flow Steering

The implementation of dynamic flow steering requires a comprehensive technical framework integrating Radio Access Network (RAN) analytics with packet routing functions in the 5G core. This framework combines the Network Data Analytics Function (NWDAF) for data processing with enhanced User Plane Function (UPF) capabilities that enable intelligent traffic steering based on current radio conditions. Research demonstrates that performance improvements correlate directly with the responsiveness of this integration, highlighting the necessity for well-designed frameworks that minimize latency between measurement and action [3]. The following subsections examine the key components of this technical framework and their implementation considerations.

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

## NWDAF Architecture and Analytics Interfaces

The NWDAF serves as the centralized analytics engine processing telemetry from diverse network functions to derive actionable insights. As defined in 3GPP TS 23.288, it implements a service-based architecture with standardized interfaces for both data collection from producer functions and analytics distribution to consumer functions. Research reveals that interface performance significantly impacts steering effectiveness, with optimized implementations achieving reduced latency through specialized data serialization and connection management [3]. Analysis of signaling patterns demonstrates that these interfaces must accommodate variable data volumes, particularly during mobility events, necessitating adaptive flow control mechanisms to prevent congestion during peak periods
[4].

The data collection interfaces connect NWDAF to functions including the Access and Mobility Management Function (AMF) and Session Management Function (SMF), establishing pathways for continuous streaming of RAN metrics such as Physical Resource Block utilization and channel quality reports. Research identifies that these interfaces must implement sophisticated filtering to extract relevant measurements from the overwhelming volume of raw telemetry, as unfiltered data would introduce unnecessary latency for critical decision pathways [4]. The analytics distribution interfaces connect NWDAF to consumer functions, critically, the SMF that controls UPF behavior. Studies show that subscriptionbased models with event-triggered notifications achieve optimal performance for dynamic steering applications, minimizing both latency and signaling overhead compared to polling approaches [3].

## UPF Integration with Real-Time RAN Measurements

The integration of UPF with real-time RAN measurements enables dynamic flow steering through analytics-driven decision logic, influencing packet routing on a per-flow basis. This creates a closed-loop architecture where NWDAF-processed insights inform UPF routing decisions, which generate performance outcomes feeding back into the analytics system. Research demonstrates that implementations achieving the shortest possible control loops deliver the most significant performance improvements, particularly for latency-sensitive applications [3].

The integration pathway flows from RAN elements through the NWDAF to the SMF, which translates analytics into policy rules governing UPF behavior via the N4 interface using enhanced Packet Forwarding Control Protocol (PFCP). Studies identify that standard PFCP mechanisms require significant enhancement to support the granularity and update frequency necessary for effective steering, particularly regarding complex condition evaluations and rule prioritization [3]. Analysis of core network signaling reveals that these rule updates constitute a significant portion of control plane traffic during network volatility, necessitating optimization techniques such as differential updates to reduce overhead [4].

Enhanced UPF implementations incorporate specialized packet inspection capabilities correlating traffic flows with QoS requirements and current radio conditions. Research shows that the computational complexity of this correlation increases non-linearly with active flows and policy sophistication, potentially creating bottlenecks during peak traffic [3]. High-performance implementations address this through specialized hardware acceleration for packet classification and rule evaluation while carefully balancing inspection depth against throughput requirements [4].

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

## Key Performance Indicators (KPIs): Accessibility, Retainability, Integrity

Effective flow steering decisions require comprehensive visibility through well-defined Key Performance Indicators spanning multiple dimensions. Research demonstrates that frameworks incorporating a balanced combination of KPIs across accessibility, retainability, and integrity dimensions achieve superior performance compared to single-dimension approaches, particularly in heterogeneous environments [3]. Analysis reveals that KPI measurement itself constitutes a significant portion of control plane traffic, highlighting the importance of efficient collection mechanisms that minimize overhead while maintaining accuracy [4].

Accessibility KPIs measure the network's ability to establish requested connections, providing early indicators of emerging congestion. Studies show these metrics offer crucial warning of deteriorating conditions before impacting active sessions, enabling preemptive routing adjustments that maintain service continuity [3]. Analysis identifies distinctive signaling signatures preceding congestion events that sophisticated analytics can detect before widespread degradation occurs [4]. Retainability KPIs quantify the network's ability to maintain established connections, identifying areas where resource constraints lead to premature session termination. Research demonstrates these metrics provide essential visibility into cell-level stability issues that may not appear in aggregate measurements [3]. Flow steering mechanisms utilize these indicators to implement protective routing for vulnerable sessions, significantly reducing service interruptions during periods of instability [4].

Integrity KPIs assess connection quality, measuring the network's ability to deliver consistent performance meeting application requirements. Studies confirm these provide the most direct visibility into actual user experience, enabling steering decisions that optimize for perceivable quality rather than technical metrics [3]. Steering mechanisms leverage these indicators to implement application-aware routing that aligns path selection with specific requirements—directing latency-sensitive traffic through minimal-delay routes while optimizing bandwidth-intensive applications for throughput, enabling networks to simultaneously support diverse application portfolios with conflicting requirements [4].

## Distributed UPF Architectures and Deployment Considerations

The distribution of UPF instances throughout the network fundamentally influences steering effectiveness. Research demonstrates that edge-positioned instances achieve dramatically improved responsiveness compared to centralized architectures, enabling steering decisions that react to changing conditions before significantly impacting user experience [3]. Analysis shows distributed architectures also reduce core network congestion by localizing control plane interactions, preventing signaling storms from propagating during localized disturbances [4]. However, this distribution introduces complexity in synchronization, resource allocation, and management that must be addressed through appropriate deployment strategies.

Geographic distribution considerations balance competing priorities, including latency minimization, resource efficiency, and operational complexity. Research indicates that optimal distribution patterns depend heavily on specific network characteristics including coverage topology and traffic distribution [3]. Hybrid approaches have emerged as predominant, with core UPF instances handling stable traffic while edge instances manage applications with stringent requirements or sensitivity to radio conditions [4]. Computational resource allocation must account for the additional processing requirements of steering

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

functionality. Studies show that steering-enabled instances require substantially greater computational capacity compared to traditional implementations, particularly for control plane processing handling rule management [3].

Resilience architectures must adapt to accommodate distributed deployments, implementing appropriate failover mechanisms maintaining service continuity. Research demonstrates that traditional redundancy approaches often prove inadequate due to the complex stateful nature of steering operations [3]. Advanced implementations employ state synchronization mechanisms continuously replicating session information and decision context across redundant instances, enabling seamless failover without disrupting active flows [4]. Scalability must address both vertical scaling (increasing individual instance capacity) and horizontal scaling (adding instances). Studies show steering-enabled implementations exhibit different scaling characteristics than traditional gateways, with control plane processing often becoming the limiting factor rather than forwarding capacity [3].

| Network Slice Type | Latency Requirement | Throughput Demand |
|---|---|---|
| Enhanced Mobile Broadband (eMBB) | 10-20 ms | 1-10 Gbps |
| Ultra-Reliable Low Latency (URLLC) | 0.5-1 ms | 100 Mbps - 1 Gbps |
| Massive IoT (mIoT) | 100-1000 ms | 1-100 kbps |

Table II: Performance Metrics Across 5G Network Slices. [3, 4]

### 3. Machine Learning Algorithms for Network Optimization

The intelligent decision-making capabilities that power dynamic flow steering rely on sophisticated machine learning algorithms designed for network optimization. These algorithms process high-dimensional data from RAN measurements to derive optimal routing decisions balancing throughput maximization, latency minimization, and resource efficiency. Recent research classifies these approaches into supervised, unsupervised, and reinforcement learning categories, with reinforcement learning demonstrating particular efficacy for dynamic steering due to its ability to adapt without extensive labeled training data [5]. Implementation in production environments requires careful consideration of realtime decision-making requirements, non-stationary network conditions, and alignment with 5G's disaggregated deployment model.

## Reinforcement Learning Models for Optimal Routing Path Selection

Reinforcement learning (RL) offers a mathematical framework that naturally aligns with sequential decision-making requirements of dynamic flow steering. RL models frame routing as a Markov Decision Process where states represent network conditions, actions represent routing decisions, and rewards reflect performance improvements. This formulation enables learning optimal policies through continuous interaction with the network environment without requiring explicit rule programming. Research demonstrates that properly implemented models discover non-intuitive optimization strategies outperforming conventional approaches, particularly with complex interdependencies between radio conditions, traffic patterns, and application requirements [5].

Deep Q-Networks (DQNs) typically implement these models, approximating value functions through neural architectures processing multidimensional input vectors including cell load metrics and interference measurements. Research identifies critical design considerations

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

including the trade-off between model complexity and inference latency, with operational implementations requiring careful balancing of predictive accuracy against computational efficiency [6]. For production environments, constrained policy optimization enforces performance boundaries during learning phases, preventing exploration from generating routing decisions that could degrade network performance [7]. Multi-agent reinforcement learning extends this paradigm to distributed environments where multiple UPF instances coordinate despite limited information sharing. Hierarchical structures combining local optimization with periodic global parameter synchronization achieve a favorable balance between optimization quality and communication efficiency [6].

## Predictive Congestion Management and Preemptive Steering

Predictive congestion management employs forecasting models to anticipate network bottlenecks before they materialize, enabling preemptive flow steering actions that redistribute traffic away from cells approaching capacity. Research demonstrates that advanced forecasting detects congestion precursors significantly earlier than thresholdbased monitoring, providing critical time for preventive interventions that maintain consistent performance [5]. These capabilities prove particularly valuable in rapidly changing environments where reactive approaches fail to respond quickly enough to prevent performance degradation.

Recurrent Neural Networks, particularly LSTM variants, form the foundation for time-series forecasting, capturing complex temporal dependencies indicating imminent resource exhaustion. Research identifies several enhancements improving prediction accuracy, including attention mechanisms dynamically focusing on relevant historical patterns and multivariate models simultaneously forecasting multiple interdependent metrics [5]. Contextual analytics complement time-series forecasting by incorporating external information influencing network demand but invisible in raw measurement data. Fusion algorithms integrate these signals with time-series predictions, significantly improving forecast accuracy for locations with variable usage patterns [6]. The predictive models drive preemptive steering actions implemented through graduated response strategies that initially affect only new sessions, then progressively extend to existing sessions based on priority levels as predicted congestion increases. Research demonstrates that properly timed interventions prevent performance deterioration typically associated with reactive approaches [7].

## Task Segmentation and Computational Offloading Strategies

Task segmentation and computational offloading extend flow steering beyond traffic management to address computational aspects of modern applications. These approaches decompose processing-intensive workloads into components distributed across available computing resources. Research demonstrates that integrated approaches combining network-aware task placement with flow steering achieve substantially higher application performance compared to separate optimization of networking and computation [5]. The task segmentation process begins with application profiling, characterizing workloads according to resource requirements, dependencies, and constraints. Advanced techniques combine static code examination with dynamic execution tracing to create comprehensive application models without requiring developer modifications [6].

Offloading decision algorithms evaluate factors including resource availability, network conditions, and application requirements to determine optimal task placement. Research demonstrates particular advantages for reinforcement learning models that develop sophisticated offloading policies through continuous interaction with the execution

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

environment [7]. Integration of offloading decisions with flow steering creates a unified optimization approach jointly managing data movement and computational placement. This enables coordinated decisions considering the entire application delivery chain rather than optimizing network and computation independently. Research demonstrates that unified approaches achieve significantly higher application performance during resource contention, particularly for applications with tight coupling between computational and network requirements [5]. Energy efficiency optimization represents another crucial aspect, with models analyzing energy implications of execution locations to identify opportunities extending battery life through strategic workload migration [6].

## Feedback Loop Mechanisms for Adaptive Bearer Treatment

Feedback loop mechanisms provide the foundation for continuous adaptation, establishing pathways through which performance outcomes influence future decisions. Research identifies critical design principles, including appropriate temporal scaling, stability guarantees, and convergence properties ensuring consistent improvement without oscillation [6]. The monitoring component establishes visibility through continuous measurement of indicators across multiple protocol layers and network segments. Specialized measurement architectures provide high-resolution visibility while minimizing overhead, employing adaptive sampling techniques focusing resources on areas experiencing performance volatility [5].

The policy adaptation engine translates performance insights into modified bearer treatment rules, influencing traffic flow handling. Hybrid approaches combine reinforcement learning with explicit domain knowledge encoded as constraints or initialization policies, ensuring adaptations remain within safe parameters while accelerating convergence [7]. Multitimescale architectural approaches implement concurrent feedback structures operating at different time horizons—millisecond-level loops managing immediate routing decisions, minute-level adaptations addressing localized congestion, and hour-level optimizations targeting broader resource allocation. Research demonstrates these layered approaches achieve more stable optimization with faster convergence than uniform feedback mechanisms [6].

Attribution mechanisms correlate specific steering decisions with resultant performance impacts, addressing the challenge of determining which decisions influenced which performance aspects. Specialized approaches combine counterfactual analysis with statistical techniques to isolate intervention effects in dynamic environments with numerous confounding factors [5]. The feedback mechanisms incorporate explicit handling for special cases including network failures or emergencies requiring immediate response. Hierarchical control structures maintain continuous monitoring for exceptional conditions while allowing learning-based optimization during normal operation, effectively addressing the fundamental challenge of combining learning-based adaptation with deterministic reliability required for critical infrastructure [7].

| Algorithm Type | Training Convergence | Optimization Accuracy |
|---|---|---|
| Deep Q-Network (DQN) | 2000-5000 episodes | 85-92% |
| Policy Gradient Methods | 1500-3000 episodes | 88-94% |
| Actor-Critic Networks | 1000-2500 episodes | 90-96% |

Table III: Machine Learning Algorithm Performance Analysis. [7]

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

## 4. Performance Analysis and Use Case Validation

The theoretical frameworks and algorithms described previously must demonstrate measurable improvements in real-world networks to justify implementation. This section presents performance analysis across diverse deployment scenarios, combining empirical measurements from testbeds, field trials, and simulation results. Standardized metrics including throughput, latency, reliability, and resource utilization enable objective comparison between AI-enhanced flow steering and traditional approaches. Research examining performance outcomes across multiple deployment categories has established that dynamic steering consistently outperforms static policies, with particular advantages in heterogeneous environments where conventional rule-based systems struggle with the complexity of interacting variables influencing optimal routing decisions [8].

### Empirical Results from Advanced Test Deployments

Advanced test environments incorporate radio equipment, edge computing resources, and core network components in configurations that model realistic deployment scenarios while supporting detailed performance analysis. These testbeds implement standards-compliant interfaces while enabling insertion of experimental components for direct comparison between conventional and AI-enhanced approaches under identical conditions. Studies reveal that AI-driven steering responds to network changes substantially faster than traditional systems, providing improved user experience during transition events such as sudden user concentration or cell failures [9]. The granularity of per-flow steering enables more efficient resource utilization compared to bearer-level management, supporting more concurrent users while maintaining consistent service quality.

Research comparing algorithm implementations shows that sophisticated deep learning approaches often achieve only marginal improvements over simpler reinforcement learning models in common scenarios, suggesting that architectural integration and measurement quality influence outcomes more significantly than algorithmic complexity [10]. Scalability evaluations demonstrate that distributed UPF architectures with embedded steering intelligence exhibit effective performance scaling up to thresholds determined primarily by control plane capacity. Energy efficiency measurements show that AI-enhanced steering can reduce network power consumption by consolidating traffic onto fewer active resources during periods of low demand, particularly valuable for deployments with sustainability objectives [8].

### Metrics of Latency and Throughput Across Deployment Scenarios

Direct indications of user experience influence come from performance indicators for throughput and latency. Standardized evaluation frameworks incorporate multiple deployment archetypes including dense urban, suburban residential, transportation corridors, venue-based, and industrial environments. In dense urban scenarios characterized by high user density, dynamic flow steering demonstrates throughput improvements consistently outperforming traditional approaches during both typical operation and peak demand periods. The advantage becomes particularly pronounced when cell congestion would otherwise create bottlenecks, with AI-driven approaches maintaining more consistent throughput by redistributing traffic to less congested resources before performance deterioration occurs [9].

In suburban residential scenarios with pronounced temporal utilization patterns, predictive congestion management provides significant advantages by anticipating usage surges and

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

proactively implementing traffic distribution before congestion materializes [8]. Transportation corridor scenarios involving high-speed mobility present unique challenges due to frequent handovers and changing cell loads. Evaluations demonstrate that MLenhanced approaches achieve more stable throughput during handover events by incorporating mobility prediction into steering decisions, reducing performance impacts typically experienced during cell transitions [10]. Venue-based scenarios such as stadiums create extreme challenges due to extraordinarily high user density. Measurements show that dynamic steering maintains higher aggregate throughput during peak utilization while implementing application-aware prioritization that preserves essential services when capacity limits are approached.

Industrial environments present distinct requirements focused on reliability, deterministic latency, and support for massive sensor deployments. Evaluations demonstrate that AIdriven flow steering excels at maintaining strict service level agreements for critical traffic while efficiently managing lower-priority monitoring data [9]. Granular categorization and path selection guarantee that control applications stay running even at times of high overall usage by routing traffic with diverse needs via suitable network resources.

## Case studies: Industrial IoT applications, high-density settings

Thorough case studies studying particular deployments offer observations on actual implementation difficulties and performance results in real-world scenarios. A major sporting venue deployment implemented AI-driven flow steering to address extreme demands of event-day traffic, incorporating predictive analytics based on historical patterns, real-time crowd distribution monitoring, and application-aware classification [8]. Performance measurements demonstrated that the dynamic steering approach maintained viable connectivity throughout the venue even during peak periods when conventional networks typically experience congestion collapse. The implementation aggregated data from both network elements and venue systems tracking crowd distribution, employed specialized classification techniques identifying different traffic types, and implemented preemptive traffic steering during anticipated congestion periods [10].

An industrial IoT case study focused on a manufacturing facility transitioning from wired connectivity to 5G for both operational technology and information technology applications. The implementation addressed diverse requirements through sophisticated classification and prioritization mechanisms that maintained strict isolation between traffic categories while maximizing overall resource utilization. Performance measurements demonstrated successful maintenance of minimal latency and high reliability for critical control applications even during periods of high overall network utilization [8]. Technical approaches included integration with the facility's security architecture, explicit safety constraints preventing critical control traffic from experiencing experimental routing, and specialized analytics designed for industrial traffic patterns [9].

A smart city deployment presented different scaling and diversity challenges, supporting applications including connected vehicle infrastructure, public safety systems, environmental monitoring, and conventional consumer services. The implementation employed edge-based analytics processing distributed throughout the coverage area, sophisticated mobility management for connected vehicles, and explicit handling for emergency scenarios including dynamic resource reassignment during crisis events [10]. Challenges related to multi-vendor integration highlighted the importance of standardized interfaces and robust fallback mechanisms.

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

## Comparative Analysis with Traditional Network Management Approaches

Objective evaluation requires systematic comparison with traditional approaches under identical conditions. Comparative frameworks contrast dynamic steering implementations against conventional approaches, including static policy enforcement, threshold-based congestion management, and reactive optimization systems. Analysis of resource utilization efficiency demonstrates that AI-driven approaches consistently achieve more balanced distribution across available network elements, resulting in higher overall capacity utilization while reducing instances of localized congestion [8]. This improved balance stems from the fine-grained nature of flow-level steering combined with predictive capabilities that redistribute traffic before congestion materializes.

Responsiveness to changing conditions reveals significant differences, with ML-enhanced systems recognizing and responding to emerging patterns substantially faster than rulebased alternatives, particularly for complex situations not explicitly modeled in conventional policy frameworks [9]. This responsiveness becomes especially pronounced during exceptional events requiring dynamic reconfiguration of routing patterns. Traditional approaches typically require human intervention to address these scenarios effectively, introducing substantial delays compared to the rapid adaptation achieved by learning-based systems.

Fairness evaluation demonstrates that properly designed AI-driven approaches improve fairness compared to traditional systems by identifying and mitigating subtle biases in conventional policies. These improvements stem from the ability of learning-based systems to discover complex interaction patterns that might create unintended discrimination in resource allocation, particularly for users at coverage boundaries [10]. Operational complexity analysis shows that while AI-enhanced systems require more sophisticated initial implementation, they significantly reduce ongoing overhead through autonomous adaptation that minimizes the need for manual optimization. Cost-benefit analysis demonstrates that AI-enhanced flow steering typically requires higher initial investment but delivers substantial operational benefits creating favorable total cost of ownership over multi-year horizons [8].

| Performance Metric | Traditional Systems | AI-Enhanced Systems |
|---|---|---|
| Network Utilization | 65-75% | 85-92% |
| Energy Efficiency | 3.2 Mbps/Watt | 5.8 Mbps/Watt |
| Service Quality Score | 7.2/10 | 9.1/10 |

Table IV: Resource Utilization Before and After AI Implementation. [9]

### Conclusion

AI-powered dynamic flow steering based on real-time RAN analytics represents a transformative advancement in 5G network intelligence. By integrating NWDAF-derived insights with UPF routing decisions, operators gain unprecedented control over traffic management with surgical precision at the flow level. The implementation of reinforcement learning algorithms within distributed UPF architectures enables continuous adaptation to changing network conditions while predictive capabilities anticipate congestion before user experience degradation occurs. Performance evaluations across diverse deployment scenarios

confirm substantial improvements in throughput, latency, and resource utilization compared to conventional approaches, with particularly significant advantages in challenging environments such as high-density venues and industrial settings. The technical framework established in this article addresses fundamental limitations in static policy enforcement while providing the flexibility required for emerging applications with diverse quality requirements. As 5G continues evolving toward more complex heterogeneous deployments, the integration of AI-powered steering mechanisms with edge computing resources will define the standard for networks that dynamically balance user experience, operational efficiency, and resource utilization. The architecture presented here provides a robust foundation for future mobile networks capable of delivering consistently outstanding performance across increasingly diverse and demanding application landscapes.

## References

[1] Huijun Gao et al., "Optimizing 5G network management," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/385827239_Optimizing_5G_network_management

[2] Mateusz Kacper et al., “Performance Optimization 5G Network Slicing Architecture Implementation," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/390329399_Performance_Optimization_5G_Network_Slicing_Architecture_Implementation

[3] MINH-NGOC TRAN et al., "Design of Computing-Aware Traffic Steering Architecture for 5G Mobile User Plane," *IEEE Access*, 2024. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10570426

[4] Dimitrios Michael Manias et al., "An NWDAF Approach to 5G Core Network Signaling Traffic: Analysis and Characterization," *ResearchGate*, 2022. [Online]. Available: https://www.researchgate.net/publication/363736453_An_NWDAF_Approach_to_5G_Core_Network_Signaling_Traffic_Analysis_and_Characterization

[5] Jasneet Kaur et al., "Machine Learning Techniques for 5G and Beyond," ResearchGate, 2021. [Online]. Available: https://www.researchgate.net/publication/349147550_Machine_Learning_Techniques_for_5G_and_Beyond

[6] Rongpeng Li et al., "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," *IEEE Access*, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7886994

[7] Iñigo Amonarriz, Jose Alvaro Fernandez-Carrasco, "A Reinforcement Learning Approach for Network Slicing in 5G Networks," *ResearchGate,* 2023. [Online]. Available: https://www.researchgate.net/publication/372977611_A_Reinforcement_Learning_Approa ch_for_Network_Slicing_in_5G_Networks

[8] Shashank Singh et al., "The Role of Artificial Intelligence in 5G Network Management: A Comprehensive Study," International Journal of Research Publication and Reviews, 2023.
[Online]. Available: https://ijrpr.com/uploads/V4ISSUE11/IJRPR19445.pdf

[9] Nasir Abbas et al., "Mobile Edge Computing: A Survey," *IEEE Access*, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8030322

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

**Research Article**

[10] Khaled B. Letaief et al., "The Roadmap to 6G: AI Empowered Wireless Networks," *IEEE Access,* 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8808168