Dilshan Wijeratne, 2021, 9:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

AI-Based Traffic Engineering in SD-WAN Networks

Dilshan Wijeratne
Open University of Sri Lanka

Abstract- The rapid proliferation of cloud-native applications, hybrid work models, and bandwidth-intensive services has fundamentally challenged the static nature of traditional Wide Area Networks (WAN). Software-Defined WAN (SD-WAN) introduced a centralized control plane to decouple network software from hardware, yet the manual definition of steering policies often fails to account for the highly volatile nature of internet transport circuits. This review examines the paradigm shift toward AI-based Traffic Engineering (TE) within SD-WAN architectures. By leveraging Machine Learning (ML) and Deep Learning (DL) models, SD-WAN controllers can now transition from reactive threshold-based switching to proactive, predictive traffic steering. We categorize current methodologies, focusing on the use of Reinforcement Learning (RL) for dynamic path optimization and Long Short-Term Memory (LSTM) networks for forecasting link congestion. This article explores how AI-driven TE optimizes Quality of Experience (QoE) for mission-critical applications—such as VoIP and real-time video—by analyzing multi-dimensional telemetry including jitter, latency, and packet loss in real-time. Furthermore, the review addresses the critical challenges of model interpretability in network operations, the "cold start" problem in new deployments, and the necessity for federated learning to ensure data privacy across multi-tenant SD-WAN environments. By synthesizing recent academic breakthroughs and industrial implementations, this paper provides a strategic roadmap for building "Self-Driving WANs." The findings suggest that AI-integrated traffic engineering not only reduces operational expenditure (OPEX) by automating complex routing decisions but also provides the "Cognitive Intelligence" required to manage the unpredictable performance of commodity internet underlays in a global digital economy.

Keywords: SD-WAN, Traffic Engineering, Reinforcement Learning, Network Optimization, Quality of Experience (QoE).

I. INTRODUCTION

The evolution of corporate networking has been defined by a constant struggle to balance cost, performance, and complexity. Historically, enterprises relied on Multiprotocol Label Switching (MPLS) to provide guaranteed bandwidth and low latency for their branch offices. While MPLS offered high reliability, its exorbitant cost and lack of flexibility made it poorly suited for the cloud era, where traffic is increasingly destined for SaaS providers rather than centralized data centers. Software-Defined Wide Area Networking (SD-WAN) emerged as the solution to this problem, allowing organizations to aggregate multiple transport links—including broadband, LTE/5G, and MPLS—into a single logical overlay.

By using a centralized controller to set policies, SD-WAN promised to simplify management and reduce costs. However, as the number of applications and network nodes has grown, the "Management Gap" has widened. Human operators can no longer manually tune the thousands of steering rules required to maintain optimal performance in a dynamic environment where the underlying internet "weather" changes by the second. This operational bottleneck is the primary driver behind the integration of Artificial Intelligence into SD-WAN Traffic Engineering (TE).

AI-based Traffic Engineering represents a move from "Policy-Based" to "Intent-Based" networking. In a traditional SD-WAN, an administrator might write a rule: "If latency on Link A exceeds 150ms, move Voice traffic to Link B." This is a reactive, binary decision that often leads to "flapping"—where traffic repeatedly bounces between links, further degrading performance. AI-driven TE, conversely, utilizes predictive analytics to forecast congestion before it happens. By training on historical telemetry data, Machine Learning models

can identify the "signatures" of an impending brownout. This allows the SD-WAN controller to proactively shift traffic, ensuring a seamless experience for the end-user. The introduction of AI into this space is not just an incremental improvement; it is a fundamental reimagining of the network as an autonomous, self-optimizing organism. It moves the network from being a "passive pipe" to an "intelligent fabric" that understands the specific requirements of the applications it carries.

The necessity for AI-driven TE is further amplified by the shift toward encrypted traffic and the death of Deep Packet Inspection (DPI). As more applications adopt TLS 1.3 and other encryption standards, the SD-WAN's ability to identify traffic based on payload becomes limited. AI-based classification models solve this by focusing on "Traffic Fingerprinting"—analyzing the timing, size, and sequence of packets to identify the application type without needing to see the unencrypted data.

Once identified, the AI-TE engine can apply the appropriate "Service Level Agreement" (SLA) to that flow. This section of the review sets the stage for a granular analysis of how AI architectures—from Reinforcement Learning to Graph Neural Networks—are being used to manage the complex trade-offs between cost and performance in the modern WAN. We will explore the transition from "Static Underlays" to "Autonomous Overlays" and analyze how the fusion of big data and algorithmic intelligence is creating a more resilient, scalable, and efficient global infrastructure.

Furthermore, the implementation of AI-TE addresses the "Complexity Tax" of multi-cloud networking. In an environment where an application might be hosted in AWS, Azure, and a local data center simultaneously, the path selection process involves a massive number of variables.

AI models can process these variables at machine speed, selecting the path that offers the best "Path Quality" based on a multi-objective optimization function. This review will explore the role of "Explainable AI" (XAI) in ensuring that network engineers can trust these autonomous decisions. By providing the "Reasoning" behind a routing change, XAI bridges the gap between machine logic and human oversight. Ultimately, AI-based Traffic Engineering is the foundational technology that will enable the "Self-Healing WAN," where the network can detect, diagnose, and remediate its own performance issues without a single human keystroke.

II. DEEP LEARNING ARCHITECTURES FOR TRAFFIC FORECASTING AND CLASSIFICATION

The efficacy of any traffic engineering system depends on its ability to accurately identify and predict the nature of the traffic it manages. In the SD-WAN context, this is a two-fold challenge: classifying the application in real-time and forecasting the future bandwidth requirements of the network. Deep Learning (DL) has emerged as the superior methodology for these tasks. Convolutional Neural Networks (CNNs) are increasingly used for "Spatial Traffic Analysis."

By converting the first few packets of a flow into a 2D or 1D image-like representation of byte frequencies and packet sizes, CNNs can identify applications with over 95% accuracy, even within encrypted tunnels. This allows the SD-WAN controller to immediately distinguish between a low-priority software update and a high-priority video call, applying the appropriate TE policy from the very first packet.

For the temporal aspect of traffic engineering, Recurrent Neural Networks (RNNs) and specifically Long Short-Term Memory (LSTM) units are the industry standard. Network traffic is not a series of independent events; it is a time-series with deep seasonal and cyclical patterns. LSTMs are designed to remember long-range dependencies, allowing them to predict "Micro-Bursts" of traffic based on historical patterns.

This section explores how these models are used to perform "Predictive Congestion Management." If

the AI forecasts that the broadband link will become congested in the next 30 seconds due to a scheduled data backup, the TE engine can begin pre-emptively moving real-time traffic to the MPLS or 5G circuit. We also analyze the rise of "Transformer" architectures in traffic forecasting, which use self-attention mechanisms to weigh the importance of different historical events, providing even higher accuracy in volatile, multi-cloud environments. By combining spatial classification and temporal forecasting, DL-based SD-WANs achieve a level of "situational awareness" that is impossible for human-defined rules.

III. REINFORCEMENT LEARNING FOR AUTONOMOUS PATH OPTIMIZATION

The most complex part of traffic engineering is "Decision Making"—choosing the best path among many variables. This is increasingly handled by Reinforcement Learning (RL). In an RL-based SD-WAN, the TE engine acts as an "Agent" that interacts with the network "Environment." The agent receives "Rewards" for maximizing the QoE of applications and "Penalties" for violating SLAs or increasing costs.

Over millions of iterations—often performed in a simulated "digital twin" of the network—the AI learns the optimal routing policy for every conceivable network state. This section deep-dives into the use of "Deep Q-Networks" (DQN) and "Proximal Policy Optimization" (PPO) in SD-WAN controllers. Unlike static routing tables, an RL agent is "Adaptive"; if a specific link starts exhibiting intermittent jitter, the agent learns to avoid it for sensitive traffic without needing a human to rewrite the policy.

The expansion of this section focuses on "Multi-Objective RL," where the agent must balance competing goals, such as "Minimize Latency," "Maximize Throughput," and "Minimize Data Transit Costs." In a hybrid WAN where some links are metered (LTE/5G) and others are flat-rate (Broadband), the RL agent performs a "Cost-Benefit Analysis" in real-time for every flow. We also explore the challenge of "Exploration vs. Exploitation"—how the agent balances trying a new, potentially better path with staying on a known good one. The beauty of RL-based TE is its ability to handle "Unforeseen Scenarios." If a fiber-cut occurs on a major backbone, the RL agent can automatically discover the best alternate path based on its learned understanding of the network's topology and performance characteristics. This section highlights how RL transforms SD-WAN from a "Scripted" system into a "Self-Learning" ecosystem that constantly evolves to protect the user's experience.

IV. GRAPH NEURAL NETWORKS FOR RELATIONAL NETWORK INTELLIGENCE

Traditional traffic engineering treats network nodes as isolated entities or simple lists. However, a Wide Area Network is a "Graph"—a complex web of interconnected branch offices, data centers, and cloud regions. Graph Neural Networks (GNNs) are a new class of AI models designed specifically to process relational data. In an AI-based SD-WAN, GNNs are used to model the "Topological Intelligence" of the network.

By representing the WAN as a graph where nodes are sites and edges are transport links, GNNs can predict how a change in one part of the network (e.g., a congestion event in the London hub) will ripple through the rest of the global infrastructure. This allows for "Global Traffic Engineering" rather than just local link-switching.

This section explores the use of GNNs for "Intent-Based Path Computation." Instead of calculating a path based on simple hop-counts, the GNN analyzes the "Relational Risk" of a path. For example, if two different ISPs share the same physical fiber-conduit, the GNN recognizes that they represent a "Common Point of Failure" and ensures that primary and backup traffic are routed over truly diverse paths.

We also analyze the role of GNNs in "Cloud-Edge Correlation." As organizations move workloads to the "Edge," the GNN helps the SD-WAN controller decide which edge node should handle a specific

request based on the current graph state of the entire network. This "Graph-Awareness" is essential for large-scale enterprise networks with hundreds of sites, where the interdependencies between links are too complex for traditional matrix-based optimization algorithms. By turning the network topology into a "Latent Space" that the AI can reason about, GNNs provide the SD-WAN with a "God's-Eye View" of the global digital infrastructure.

V. QUALITY OF EXPERIENCE (QOE) MODELING AND PREDICTIVE STEERING

The ultimate metric for any SD-WAN is not "Throughput" or "Uptime," but "Quality of Experience" (QoE)—how the end-user perceives the application's performance. Traditional TE focuses on "Quality of Service" (QoS) metrics like latency and jitter, but these do not always correlate with user satisfaction. For instance, a 1% packet loss might be unnoticeable for a web download but catastrophic for a Zoom call. AI-based SD-WANs utilize "QoE Modeling" to bridge this gap. By training on subjective user feedback and objective network telemetry, AI models create a "QoE Score" for every application flow. The TE engine then uses this score as its primary steering metric, moving traffic before the user even notices a degradation.

This section examines the use of "Regression Models" and "Random Forests" to map network conditions to application-specific QoE. We discuss the "App-Specific Sensitivity" of these models—how the AI learns that voice traffic is sensitive to jitter, while file transfers are sensitive to throughput. The expansion of this section also covers "Predictive Steering" for SaaS applications. Since the SD-WAN controller doesn't own the "Last Mile" to a SaaS provider like Microsoft 365, it must use AI to probe and predict the performance of different "Cloud Gateways."

By analyzing the "Historic Performance" of various internet on-ramps, the AI can steer the user to the gateway that is currently offering the best QoE. This proactive approach turns the SD-WAN into a "SaaS Accelerator," ensuring that critical business tools remain responsive regardless of the local internet conditions. We conclude by looking at "Closed-Loop Remediation," where the AI continuously verifies that its steering decisions actually resulted in an improved QoE score, creating a self-correcting feedback loop.

VI. HANDLING NON-IID TELEMETRY AND DATA DRIFT IN SD-WAN

A significant challenge in AI-based TE is that network data is "Non-IID" (not independent and identically distributed). Network traffic is highly non-stationary; what is "normal" on a Tuesday morning is not normal on a Saturday night. Furthermore, network conditions exhibit "Concept Drift"—the underlying performance of an ISP link can change permanently due to infrastructure upgrades or routing changes in the provider's core.

If an AI model is not designed to handle this drift, its TE decisions will become inaccurate over time. This section explores the use of "Adaptive Learning" and "Online Training" to keep SD-WAN models fresh. Instead of a "Static" model, the AI uses "Streaming Analytics" to update its weights as new data arrives.

We also analyze the "Data Silo" problem in SD-WAN. Telemetry data is often distributed across thousands of branch routers, making it expensive to backhaul all raw data to a central controller for training. This section explores "Federated Learning" (FL) as a solution. In an FL-based SD-WAN, each branch router trains its own "Local Model" on its own data and only sends the "Model Updates" to the central controller.

The controller aggregates these updates to create a "Global Model" which is then pushed back to the branches. This significantly reduces bandwidth usage and ensures that the AI can learn from a global dataset without compromising the data privacy of individual sites. We also discuss "Transfer Learning," where a model pre-trained on a generic network dataset is "Fine-Tuned" for a specific enterprise's environment, allowing for rapid deployment and high accuracy even in the early stages of a new SD-WAN rollout.

VII. EXPLAINABLE AI (XAI) AND NETWORK OPERATOR TRUST

The move toward "Self-Driving Networks" creates a "Trust Gap" between the AI and the network engineer. If an AI-TE engine moves all of the company's financial traffic to a 5G link, the engineer needs to know "Why." In mission-critical environments, "Black Box" decisions are a major risk. "Explainable AI" (XAI) is the technological layer that provides transparency into the AI's decision-making process. This section explores XAI techniques like "SHAP" (SHapley Additive exPlanations) and "LIME" (Local Interpretable Model-agnostic Explanations) applied to traffic engineering. These tools can highlight the specific telemetry features—such as a 10ms spike in jitter—that triggered a routing change.

This section also addresses the "Human-in-the-Loop" (HITL) model. We discuss how AI-TE engines provide "Recommendations" with a "Confidence Score." For low-stakes traffic, the system acts autonomously; for high-stakes traffic, it presents its "Reasoning" to the engineer for a final "Click-to-Approve." This synergy allows the organization to benefit from machine speed while maintaining human accountability.

We analyze the role of "Visualization Dashboards" that turn complex neural network weights into intuitive "Risk Maps" and "Topology Overlays." By making the AI's logic "Human-Readable," XAI transforms the SD-WAN controller from a mysterious oracle into a transparent, trusted advisor. This transparency is also vital for "Regulatory Compliance" and "Post-Mortem Analysis," ensuring that every routing decision can be audited and justified after the fact.

VIII. SCALABILITY AND REAL-TIME INFERENCE AT THE EDGE

For AI-based TE to be effective, it must operate at the "Edge" of the network—on the branch routers themselves. However, these routers often have limited CPU and memory compared to cloud servers. "Real-Time Inference" requires the AI models to be "Right-Sized" for the hardware they run on. This section explores "Model Compression" techniques such as "Pruning," "Quantization," and "Knowledge Distillation." These methods allow a massive deep learning model to be shrunk down to a size that can run on a low-power ARM or MIPS processor without significant loss in accuracy.

We also examine the role of "Hardware Acceleration" in SD-WAN appliances. Many modern routers now include "Neural Processing Units" (NPUs) or "Tensor Cores" designed specifically to accelerate AI math. This section discusses the "Inference Latency" requirements for TE. If the AI takes 5 seconds to calculate a new path, the congestion might have already caused the voice call to drop.

We analyze "Fast-Path vs. Slow-Path" AI architectures: a fast, simple model handles immediate link-switching, while a more complex model handles long-term path optimization in the background. Furthermore, we discuss the "Scale-Out" challenge—how to synchronize AI models across a network with 5,000 branch sites to ensure a consistent TE policy. By optimizing the "Data Pipeline" and the "Inference Engine," AI-based SD-WANs ensure that the "Intelligence" doesn't become a "Bottleneck" for the very traffic it is trying to optimize.

IX. SECURITY-AWARE TRAFFIC ENGINEERING AND AI FUSION

In a modern SD-WAN, Traffic Engineering and Security are increasingly fused into a single architecture known as SASE (Secure Access Service Edge). AI-based TE must now be "Security-Aware." This means the TE engine doesn't just choose the "Fastest" path, but also the "Safest" path. For example, if the AI detects that a specific internet gateway is currently under a DDoS attack or exhibiting "BGP Hijacking" signatures, it will steer sensitive traffic away from that path, regardless of its latency score. This section explores the fusion of "Anomaly Detection" and "Path Selection."

We discuss the use of "Multi-Agent Systems" where a "Security Agent" and a "TE Agent" negotiate the best path for a flow. If the Security Agent identifies a flow as "Suspicious," the TE Agent might route it through a "Scrubbing Center" or a "Cloud Sandbox" for deeper inspection, even if that path is slower. We also analyze the threat of "Adversarial AI"—where an attacker tries to "Fool" the TE engine into routing traffic over a compromised link by spoofing "Good Telemetry."

This section highlights the necessity for "Robust AI" models that can distinguish between "True Performance" and "Manipulated Metrics." By integrating security into the TE decision-making process, the AI-driven SD-WAN becomes a "Self-Defending Fabric" that protects the data as much as it optimizes the delivery, creating a unified, intelligent gateway to the multi-cloud world.

X. THE FUTURE OF AI-TE: 6G, IOT, AND BEYOND

As we look toward the future, the scope of AI-based Traffic Engineering will expand from branch offices to the "Extreme Edge"—including IoT devices and mobile users on 6G networks. This section explores the concept of "Network Slicing" driven by AI. In a 6G environment, the AI-TE engine will be responsible for dynamically creating "Virtual Slices" for different use cases—such as a high-reliability slice for a remote-surgery robot and a high-bandwidth slice for an 8K video stream. This requires a level of "Granular Orchestration" that is far beyond the capabilities of today's SD-WAN.

We also examine the role of "Sustainable AI" in traffic engineering. As data centers and networks consume a growing share of the world's energy, AI models are being used to optimize "Energy-Aware Routing"—steering traffic to links and nodes that are powered by renewable energy or that are currently in a "Low-Power State."

This section concludes by looking at the "Full Autonomy" vision, where the SD-WAN controller uses "Generative AI" to "Write its own code" and "Create its own protocols" in response to novel network challenges. This "Evolutionary Networking" represents the final frontier of traffic engineering, where the human role shifts from "Administrator" to "Governor," overseeing a vast, self-creating digital ecosystem that powers the global economy.

XI. CONCLUSION

AI-based traffic engineering represents the definitive future of SD-WAN, transforming the network from a static, policy-driven infrastructure into an autonomous, intent-aware ecosystem. By leveraging the predictive power of LSTMs, the strategic decision-making of Reinforcement Learning, and the relational intelligence of Graph Neural Networks, AI-TE resolves the inherent volatility of the modern cloud-centric WAN.

This review has demonstrated that the transition to "Cognitive Networking" not only improves the Quality of Experience for the end-user but also significantly reduces the operational burden on IT departments through automated, closed-loop remediation. However, the path toward full autonomy requires a rigorous focus on "Explainability" to maintain human trust and "Model Compression" to ensure real-time performance at the edge.

As we move into an era of 5G/6G and massive IoT, the ability to manage the global "Traffic Flow" with machine-speed intelligence will be the deciding factor in an organization's digital resilience. Ultimately, AI-driven traffic engineering ensures that the network is no longer a bottleneck for innovation, but a dynamic, self-optimizing catalyst for the next era of global digital transformation.

REFERENCES

1. Jangala, V. K. (2015). Observability and monitoring of microservices using Splunk and New Relic. International Journal of Engineering Development and Research, 3(3), 1–15.
2. Vangoor, V. K. R. (2016). AI-driven monitoring and alerting systems for enterprise-scale Linux deployments. International Journal of Science, Engineering and Technology, 4(1), 11.
3. Parimi, S. S. (2016). Analyzing the effectiveness of SAP systems in streamlining healthcare

supply chains, reducing costs, and improving service delivery.

4. Koukuntla, S. (2018). Event-driven architectures in cloud computing: Tools, patterns, and tradeoffs. International Journal of Trend in Scientific Research and Development, 2(3), 2909–2913.
5. Jangala, V. K. (2016). API gateway security implementation using JWT and Apigee in cloud-native applications. International Journal of Current Science, 6(2), 34–43.
6. Vangoor, V. K. R. (2017). Self-optimizing DevOps pipelines for enterprise infrastructure using machine learning models. International Journal of Trend in Scientific Research and Development, 1(6), 8.
7. Parimi, S. S. R. (2016). Predictive analytics for financial forecasting in SAP ERP systems using machine learning. International Journal of Creative Research Thoughts.
8. Jangala, V. K. (2018). Database performance tuning strategies for high-volume transaction systems. International Journal of Scientific Development and Research, 3(8), 274–282.
9. Vangoor, V. K. R. (2018). AI-based optimization of automated server deployment using Kickstart and Satellite systems. International Journal of Trend in Research and Development, 5(6), 5.
10. Parimi, S. S. (2018). Exploring the role of SAP in supporting telemedicine services, including scheduling, patient data management, and billing. SSRN Electronic Journal.
11. Parimi, S. S. (2018). Optimizing financial reporting and compliance in SAP with machine learning techniques. SSRN Electronic Journal.
12. Mandati, S. R. (2019). The basic and fundamental concept of cloud balancing architecture. South Asian Journal of Engineering and Technology, 9(1), 4.
13. Mandati, S. R. (2020). System thinking in the age of ubiquitous connectivity: An analytical study of cloud, IoT and wireless networks. International Journal of Trend in Research and Development, 7(5), 6.
14. Mandati, S. R., Rupani, A., & Kumar, D. S. (2020). Temperature effect on behaviour of photo catalytic sensor (PCS) used for water quality monitoring.