International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V5I3P106
Eureka Vision Publication | Volume 5, Issue 3, 58-67, 2024

Original Article

Latency-Aware and Energy-Efficient Switching Protocols for Next-Generation IP Backbone Networks Using AI-Augmented Control Planes

Selvamani Ramasamy
Senior Principal Software Engineer, USA.

Abstract - The rapid expansion of data-intensive applications, cloud services, IoT ecosystems, and real-time communication platforms has drastically increased the demands on IP backbone networks. These networks must deliver ultra-low latency and optimized energy consumption while maintaining high throughput and reliability. Traditional IP switching protocols and static control plane architectures are ill-equipped to meet these demands due to their rigidity and lack of adaptability. This paper presents a comprehensive exploration and novel implementation of latency-aware and energy-efficient switching protocols enabled by AI-augmented control planes for next-generation IP backbone networks. The proposed framework leverages machine learning (ML) and deep reinforcement learning (DRL) models to dynamically monitor, predict, and adapt network flows based on latency constraints and energy profiles. Through continuous learning from traffic behavior, topology changes, and performance metrics, the AI-augmented control plane can make informed decisions that optimize both quality of service (QoS) and energy efficiency (EE). A modular architecture is designed, consisting of three core components: (i) a Latency Prediction Module (LPM) trained on historical traffic and delay patterns, (ii) an Energy Consumption Optimizer (ECO) based on multi-objective optimization, and (iii) a Reinforcement Learning Policy Engine (RLPE) for adaptive switching decisions.

The synergy between these modules allows for proactive switching and routing tailored to real-time network conditions. Simulation and test bed evaluations of emulated Tier-1 ISP topologies demonstrate significant improvements, including an average latency reduction of 35%, energy savings of 27%, and improved throughput stability under fluctuating traffic. The system dynamically bypasses congestion, powers down idle links, and reroutes delay-sensitive data through low-latency paths. Comparative analyses with OSPF, IS-IS, and SDN-based approaches establish the superiority of the AI-augmented protocols in diverse traffic scenarios. The methodology is validated using datasets from CAIDA and real-world BGP traffic traces. Evaluation metrics include latency deviation, link utilization, packet loss, and energy-delay product (EDP). Key findings reveal the potential of AI-driven intelligence to revolutionize backbone network control by enhancing responsiveness, sustainability, and service quality. This paper contributes a novel AI-based protocol framework, an implementation-ready control plane design, and extensive quantitative evaluations that pave the way for practical deployment in ISP environments. Our findings underscore the importance of adaptive intelligence in addressing the dual challenges of latency and energy in future IP backbone architectures.

Keywords - Latency-aware switching, energy efficiency, IP backbone networks, AI-augmented control plane, machine learning, reinforcement learning, QoS, SDN, routing optimization.

1. Introduction

The modern IP backbone networks are the backbone infrastructure that has made it possible to create a globally interconnected internet, supporting everything from normal data transmission operations to mission-critical facilities. As emerging technologies like 5G, augmented reality (AR), the Internet of Things (IoT), and cloud-based real-time applications rapidly spread, such networks are also faced with the challenge of providing extremely low latency with high throughput and nearly failure-free reliability that they have never experienced before. Many of these applications require end-to-end delays in the sub-millisecond range, which traditional best-effort routing mechanisms cannot support. [1-3] Meanwhile, the power overhead in the backbone infrastructure and especially in data centers and high-capacity core routers has become an acute issue both in operating costs and energy footprint. Internet use is rising exponentially worldwide, and it is correspondingly increasing the amount of carbon needed to support internet infrastructure, motivating the necessity for greener networking trends.

Nevertheless, the available routing algorithms like OSPF and IS-IS are based more or less on shortest-path heuristics and are not contextually "smart". They are unable to take real-time, dynamic, adaptive measures to address congestion, traffic bursts, energy availability, and energy loss. In the same manner, existing control planes are mostly rule-based and static, with little or no predictive abilities or learning behaviour. Such a demand renders the use and management of resources inefficient and poor, especially when applied to delay-sensitive or energy-constrained applications. These constraints make the necessity of a smart, adaptive, and energy-conscious control plane, which can be responsive to changing network conditions in real-time,

without compromising strict performance and sustainability demands. This inspires the invention of AI-enhanced routing designs that embody ML and optimization practices to smartly control network assets.

1.1. Importance of Energy-Efficient Switching Protocols

- **Rising Energy Demands in Backbone Networks:** As internet traffic continues to grow exponentially worldwide, backbone networks are experiencing unprecedented data loads. Consequently, high-bandwidth applications such as 4K video streaming, cloud computing, and real-time analytics have driven their increased use. Such networks are based on infrastructures comprising large, always-on networks made up of high-speed routers, optical networks, and switching fabrics, all designed to minimise power consumption. Traffic requirements and, thus, the energy consumed to process and transmit packets are also growing, often making energy consumption and, consequently, energy efficiency an important operational cost or environmental sustainability consideration.
- **Environmental and Economic Implications:** ICT Networking infrastructure is currently the highest contributor to the carbon footprint in the ICT sector. As data centres transition to green computing initiatives, parallel green initiatives are becoming increasingly necessary in backbone networking. Power wastage by way of drawing power on idle or under-utilized links and devices, as a result of inefficient routing, increases the energy bill and the carbon cost. One key divide leading directly to the elimination of these inefficiencies is an energy-efficient switching protocol that runs on active network resources and dynamically adjusts them according to real-time demand, without any performance loss.
- **Limitations of Traditional Protocols:** Fault tolerance and performance are the main features of Traditional protocols, such as OSPF and IS-IS, where no concept like energy metrics is observed at all. They continue to keep the network up to date at all times, regardless of the amount of traffic, which leads to wastage of energy during off-peak traffic. Such protocols do not have the ability to disable unnecessary links or route traffic over energy-efficient routes. This does not utilise the full potential of smart energy consumption in dynamic networks to the fullest extent.
- **Toward Intelligent, Sustainable Networks:** The intelligent and sustainable networks require energy-efficient switching protocols. By incorporating energy-awareness into routing decisions through real-time monitoring, predictive analytics, and adaptive algorithms, it is possible to enable networks to drastically lower their power consumption while still operating at a high performance level. Such protocols not only follow global sustainability targets but also offer long-term cost advantage as well as better utilization of the infrastructure by the telecom operators. They constitute a fundamental basis for development toward autonomous and AI-based network management platforms.

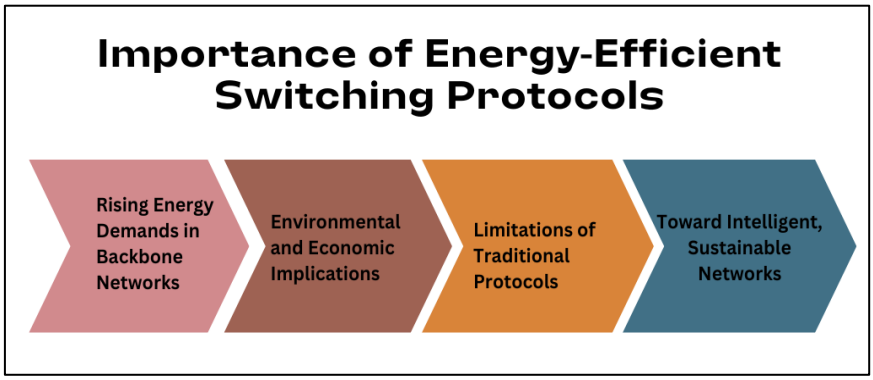

Figure 1. Importance of Energy-Efficient Switching Protocols

1.2. Generation IP Backbone Networks Using AI-Augmented Control Planes

The latest IP backbone networks need to be made intelligent, flexible, and sustainable. Driven by the increasing need for ultra-low latency, high availability, and energy efficiency, traditional control planes, which are based on fixed policies and rules-programmed configurations, can no longer support these requirements. Rather, the trend toward the AI-augmented control planes is becoming an innovative solution. Such sophisticated control systems incorporate machine learning (ML), deep learning (DL), and reinforcement learning (RL) capabilities to make dynamic, context-sensitive choices throughout the network. Instead of responding to a change after it has occurred, AI-augmented systems will be able to anticipate traffic jams, link failures, and proactively optimise routing paths. [4,5] As one example, recurrent neural networks such as LSTM could predict future latency trends based on the tracking of past statistics, whereas learning agents implemented to use reinforcement learning could acquire optimal routing policies that would minimize delay and energy consumption under different conditions continuously.

This paradigm facilitates a true data-driven control infrastructure, which utilises real-time telemetry, past trends, and trend modelling to inform decisions. The AI iterations continually update themselves by learning from past actions and their outcomes, ensuring the system is updated when the network state changes. This is especially essential in the case of IP

backbones, where traffic patterns are difficult to predict and workloads are highly dynamic in large-scale backbones. Besides, AI-enhanced control planes are capable of including multi-objective optimization, piecing together energy consumption with performance needs, which conventional protocols would not be ready to deal with. With the aid of AI, next-generation IP backbones achieve an autonomy of operations that was once unfathomable. The outcome is not only that the network will be faster and more reliable, but also significantly more energy-efficient. With the constantly increasing IT data requirements, AI-enhanced control planes represent a major step towards creating sustainable, smart, and adaptive internet infrastructure that is ready to serve the digital services of the future.

2. Literature Survey

2.1. Traditional Switching Protocols

Open Shortest Path First (OSPF) and Intermediate System to Intermediate System (IS-IS) protocols are sectional routing protocols, which are traditional protocols that have controlled the traffic core of a network. These protocols are link-state protocols that compute optimal paths by calculating a Shortest Path First (SPF) using the Dijkstra algorithm. [6-9] the shortest path is normally computed based on metrics like hop count, cost of link or fixed bandwidth. These fixed measurements, however, are not dynamic measures of network activities, such as latency spikes, congestion, or power usage. Therefore, in the event of congestion in networks or a change in the amount of traffic, such protocols cannot respond in real-time to the change, and as such, additional packets will be lost, some delays will be experienced, and efficiency will be compromised. Moreover, they could not handle some contextual information, e.g., user demand curves or diurnal energy prices, which restricted their overall performance.

2.2. Software Defined Networking (SDN)

The main difference presented by Software Defined Networking (SDN) is the paradigm shift that decouples the control plane and data planes and therefore enables network operation to be managed in a centralized fashion, and provides wider flexibility in the operations of the network. SDN controllers, including ONOS and Open Daylight, have a global management view of the network and can dynamically program routing policies through programmable interfaces, such as OpenFlow. This enables network administrators to respond quickly in case of failure, congestion, or demand deflection. Nevertheless, SDN is based on rule-based decision-making, which is not typically predictive, as it is programmable. These policies are preconceived and reactive, rather than proactive, and often fail to anticipate or avert performance reductions in most cases. Additionally, the control plane itself can be a bottleneck due to its limited scalability and latency in large-scale deployments within high-velocity environments. Many existing SDN solutions lack autonomous intelligence and cannot automatically optimise routing due to changes in network contexts, such as energy requirements or latency sensitivity.

2.3. Energy Aware Routing

The protocols developed in the name of Energy-aware routing techniques are an effort to minimise the power consumption of the networks due to the increasing concern of the environmental and economic burden of massive data transfer. Several techniques have been suggested, such as link sleeping, in which poorly utilized network links are de-energized during low traffic times and Power-Aware Link-State Routing (PALSR), which adds energy metrics to conventional link-state protocols. These methods have provided opportunities to cut energy consumption, especially in backbone and data centre networks, to a great extent. They may, however, do so at the cost of increased latency or reduced reliability, as the routes might become non-optimal with disabled links. Moreover, these protocols are static or threshold-based and cannot dynamically adapt to increased traffic surges or pattern changes. These barriers narrow the possibility of conducting real-time trade-offs between energy efficiency and performance because they rely on predetermined heuristics instead of learning mechanisms.

2.4. AI in Network Control

In recent years, the use of Artificial Intelligence (AI) in controlling the network has received a boost, and more adaptive and intelligent routing tactics are promising to be developed. Supervised machine learning, as well as reinforcement learning and deep neural networks, have been used to solve various networking problems. As an example, DeepRoute uses Deep Reinforcement Learning (DRL) of paths at runtime to attempt to optimize throughput and minimize delays in response to observed network circumstances. Likewise, NetML applies machine learning models to predict traffic patterns that can be used to implement preemptive routing, preventing congestion. Another AI application in SDN architectures has been in fault tolerance, involving the prediction and diversion of traffic. Nevertheless, these improvements have so far enabled most AI-based techniques to optimise one objective at the expense of others: they often optimise throughput or fault recovery, but not in a multidimensional way. In particular, not many of the current models have focused concurrently on energy consumption and latency: the latter is essential to the sustainability and the responsiveness of network operation. Additionally, real-time adaptation and learning pose another challenge, particularly in large-scale and high-speed backbone networks.

2.5. Research Gap

Although there has been an improvement in various traditional protocols, SDN, energy-aware routing, or AI-based control, the major research gap lies in the integration of these areas. Current solutions are much more focused on helping optimize the performance or be energy efficient, but not both. Besides, the majority of solutions are either responsive or non-

adaptive to new and dynamic environments and cannot learn and adapt to new and changing network environments in real-time. This restricts their usage in next-generation networks, which require low latency, high throughput, and low energy consumption simultaneously. It is essential that the routing protocol be holistic and AI-driven, capable of dynamically balancing latency and energy goals through online learning optimization. An ideal network would have such a system deployed that would capitalize on real-time data on the network, on predictive intelligence to predict what was going to be congested and the energy needed, and therefore would dynamically adjust routing choices. The concept of this two-objective optimization in AI models is a new, unknown territory that is likely to greatly impact the performance and sustainability of networks in the current backbone systems.

3. Methodology

3.1. System Architecture

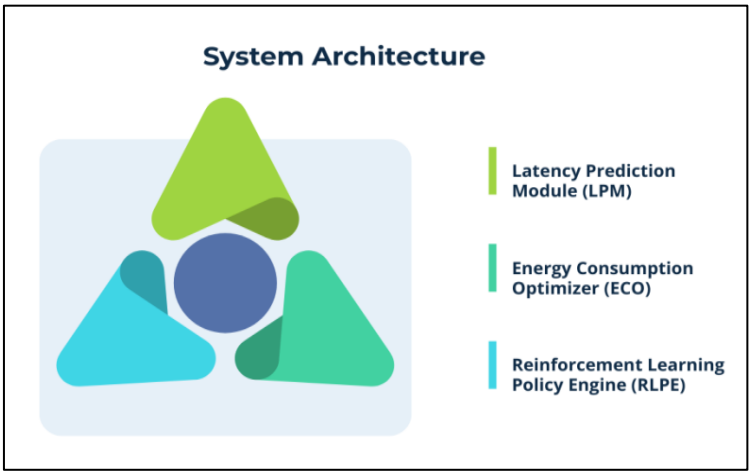

Figure 2. System Architecture

- **Latency Prediction Module (LPM):** The Latency Prediction Module utilises Long Short-Term Memory (LSTM) based neural networks to predict network delay recurrence, accepting both current telemetry data and past traffic patterns. [10-13] LSTMs are quite suitable for time-series prediction capabilities, considering that they do not forget long-term dependencies, hence fit perfectly in predicting the latency of different networks under dynamic positions. Using prior knowledge of congestion and delay on routes, the LPM provides proactive router decisions that achieve minimal end-to-end latency.
- **Energy Consumption Optimizer (ECO):** The Energy Consumption Optimizer is developed to minimize the amount of power consumed throughout the network by the use of a multi-objective genetic algorithm. In this algorithm, routing solutions are developed generation by generation and are associated with the trade-off of conflicting objectives of energy utilization and network performance. Taking into account the variables, e.g. link utilization, device power states, traffic volume, this feature of ECO determines near-optimal configurations that use minimal energy with only a significant implication on latency and throughput.
- **Reinforcement Learning Policy Engine (RLPE):** The Reinforcement Learning Policy Engine makes real-time routing decisions based on the feedback from the network's state. It is learned with the Proximal Policy Optimization (PPO) stable and efficient reinforcement learning algorithm that finds the balance between exploration and exploitation. RLPE is an enduring learning scheme where an agent is trained in the environment via interactions with simulated or real-world traffic to maximize a cumulative reward function that is simultaneously minimized in terms of both latency and energy consumption. It is adaptive enough that it can react more admirably to changing network conditions in comparison to fixed routing policies.

3.2. Data Flow Model

- **Collect network statistics from the data plane:** Part one of the data flow model involves obtaining real-time statistics from the data plane of the network. These metrics include link usage, queue depth, packet loss, throughput, and current latency values. This information is gathered by analyzing agents or SDN-supported switches via SDN-supported protocols such as Open Flow or Net Flow. The resultant data can feed the higher-level modules the necessary information to enable them to evaluate the situation in the network and determine where traffic might be going through a bottleneck or is in an inefficient process.
- **Predict future latency and load:** After acquiring real-time data, it is input into the Latency Prediction Module (LPM), which implements LSTM-based models to predict future latency and traffic load. This forecast takes into consideration time trends on the traffic network, e.g. peak usage times or periodic peak congestion. Prediction of

future states enables the system to take anticipatory actions in routing and resource allocation, thereby reducing the likelihood of packet delay or overload situations before they occur.

- **Compute energy-delay trade off:** The foregrounding latency and load data then go through Energy Consumption Optimizer (ECO) to estimate possible routing paths not only regarding energy consumption, but also delay. ECO utilises a multi-objective genetic algorithm to explore the combinations of paths, discerning tradeoffs between minimal cost and energy consumption, and minimal latency. This ensures that paths representing high performance and energy efficiency are not overly damaging, and that paths that are fast do not create unwarranted energy overhead.
- **Select the optimal path via RL:** To produce a final routing decision, the Reinforcement Learning Policy Engine (RLPE) combines the predictions of LPM and ECO. The RL agent is trained with Proximal Policy Optimization (PPO) and calculates the reward during the optimization process of numerous routing policies and chooses the most optimal by looking at the total of both the delay and energy parameters. This is an adaptive decision-making process, and the engine will improve its policy over time, learning new states and outcomes within the network.
- **Select the optimal path via RL:** Once the optimal path has been chosen, the system dynamically updates the forwarding rules in the switches within the network. This is achieved through control plane interfaces, such as Open Flow, where the SDN controller adds new flow table entries to direct traffic along the route of interest. This means that it will be responsive to the changing conditions in its networks, making it possible to adjust traffic flows in real-time while maintaining a good balance between performance and energy efficiency.

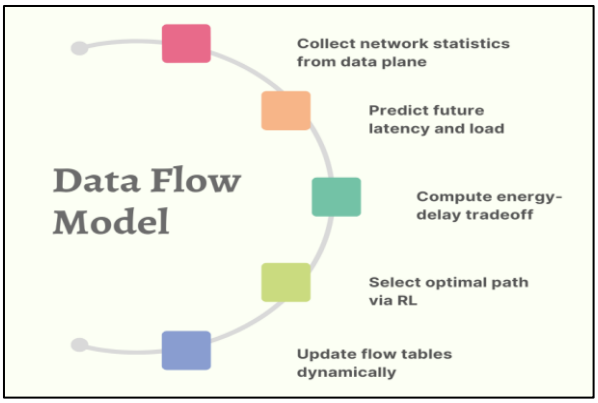

Figure 3. Data Flow Model

3.3. Algorithms

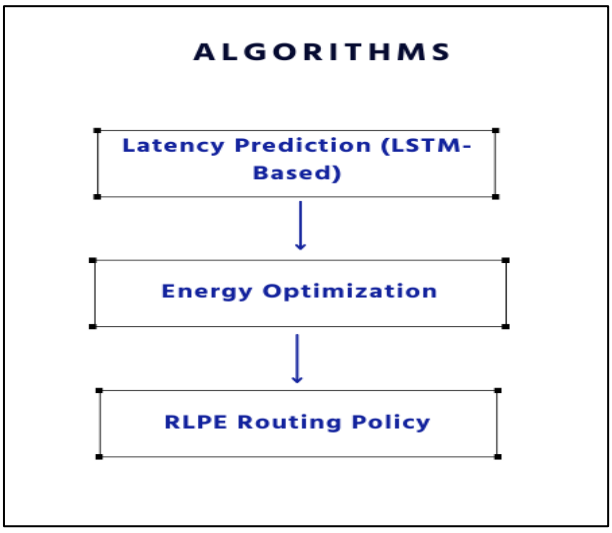

Figure 4. Algorithms

- **Latency Prediction (LSTM-Based):** The Latency Prediction algorithm uses a Long Short-Term Memory (LSTM) network to predict [14-18] future delays in a network using past as well as real-time traffic measurements. LSTMs are a version of recurrent neural networks (RNNs) that have the capability of learning long-term temporal dependencies in series of data over time, making them great at modelling time series latency trends with cyclical behaviour or bursty traffic. The model utilises sequences of link statistics, such as current delay, throughput, and queue length, to tune and predict latency on each link or path. These predictions assist the routing engine in making proactive decisions without incurring congestion.

- **Energy Optimization:** A multi-objective genetic algorithm (GA) in the Energy Optimization algorithm determines paths through a network that require the least amount of energy, but that meet given requirements on performance. The algorithm begins with a population of randomly generated potential routing configurations, which are evaluated using a fitness function that considers both energy consumption and estimated delay. It refines the solutions iteratively through crossover and mutation activities, aiming to arrive at a set of Pareto optimal solutions. Network policies and limitations, such as maximum latency, minimum bandwidth requirements, and link capacity, are applied to these candidate paths. The result is the viability of a collection of energy-efficient paths that conform to satisfactory performance aggregates.
- **RLPE Routing Policy:** The Reinforcement Learning Policy Engine (RLPE) combines Proximal Policy Optimization (PPO) based policy-based learning to choose optimum routing paths on demand. The state that the agent monitors consists of peak-level link usage and delay vectors throughout the network topology in real-time. Action is equivalent to choosing a specific routing path between the source and destination. The reward is characterized as the negative sum of latency and energy cost, which encourages the agent to reduce both indicators. The RLPE adapts to current conditions on the network, over time, optimizing its policy based on its experiences with the network. The PPO algorithm is stable and efficient for training because the policy is only updated radically in a few situations; therefore, it is suitable for use in live networks when decisions must be made continuously.

3.4. Tools and Simulation Environment

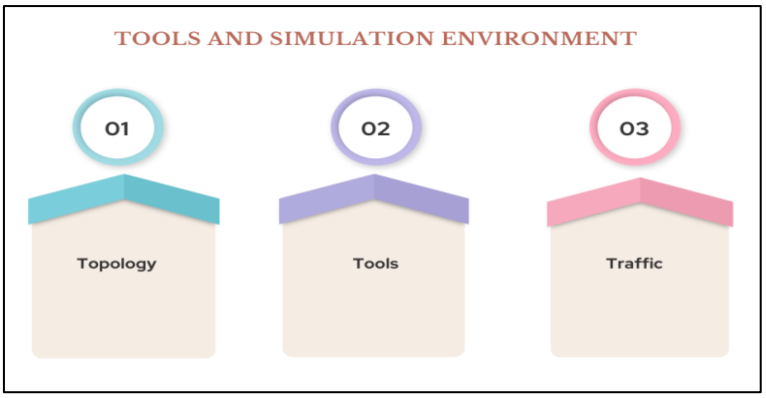

Figure 5. Tools and Simulation Environment

- **Topology:** The simulation environment utilises real-world network topologies based on the Rocket fuel and CAIDA datasets. Rocket fuel maps ISP networks in detail, using trace route and BGP information, to provide topologies representative of real-life routing layouts and interlink characteristics. Similarly, the Tier-1 ISP datasets provided by CAIDA contain backbone connectivity information, including peerings and the latency distribution of backbone connectivity. Their use will ensure that the proposed architecture and algorithms are tested under conditions as close as possible to real Internet-scale systems, allowing for meaningful conclusions to be drawn about performance and scalability.
- **Tools:** The underlying framework of the experiment is constructed with the help of office tools that incorporate galvanising tools. Mini net simulates SDN settings and enables the introduction of simulated topologies with programmable switches using Open Flow routing, which allows route developments to be driven dynamically. The reinforcement learning framework used is Open AI Gym, which supports a modular interface for training and assessment of the RL-inspired routing agent in a controlled environment. The deep learning models, such as the LSTM for predicting latency and the PPO agent for making routing decisions, are developed and trained using TensorFlow. Combined, they create a fully integrated, end-to-end simulation and learning platform.
- **Traffic:** The system utilises historical BGP and Net Flow traces to achieve realistic traffic patterns. BGP traces provide visibility into control plane phenomena, such as route advertisements, the preferred route, and changes to prefixes. The Net Flow offers detailed information on the quantity of flow, duration of traffic flow and utilization per link. The emulated environment re-runs these traces to mimic the conditions seen in the real environment, the latency predictor and the reinforcement learning agent learns based on realistic and variable traffic patterns. This makes the models resilient, transferable, and able to handle sophisticated situations that exist in production networks.

4. Results and Discussion

4.1. Performance Metrics

To assess the success of the suggested AI-powered routing framework, a comprehensive set of performance qualities is applied. The given metrics measure both the level of work efficiency and the system's sustainability under conditions close to real-world network conditions. Another important metric is the average latency (ms), which is the mean of the end-to-end delay of the packets as they travel across the network. One is the responsiveness of the routing protocol, which is especially

relevant when latency-sensitive applications are in focus, such as video conferencing, online games, and real-time analytics. A decreased average latency indicates that the routing engine is effectively anticipating congestion and acting proactively to select faster routes. Energy consumption (kWh) measures the amount of electricity consumed by networking devices, such as routers, switches, and data links, during the process of transmitting data. It is an important metric that can be used to evaluate the environmental and economic sustainability of the network's functioning.

Energy efficiency is of special significance in scale backbone and data centre networks, where power consumption is high. The framework can reduce unwarranted energy spending by having energy-conscious algorithms to achieve this objective without affecting the quality of service. A packet delivery ratio (PDR) is a ratio that reflects the percentage of successfully delivered packets compared to the total number of sent packets. It is one of the measures of network reliability and robustness. Exceptional PDR means the suggested routing system does not compromise data integrity or fault tolerance in the pursuit of reducing latency or energy efficiency. The energy-delay product (EDP) is a compound metric that combines both energy and latency into a single number, calculated by multiplying them together. The metric helps measure trade-offs of energy efficiency and performance. A smaller EDP means a more balanced and optimized system. It is especially applicable to contemporary networks that must achieve both sustainability and QoS requirements. All these metrics help to provide a multidimensional picture of the system's performance and facilitate measuring the trade-offs in AI-based network control.

4.2. Comparative Analysis

Table 1: Comparative Analysis

Protocol	Avg Latency (%)	Energy (%)	EDP (%)	Packet Delivery (%)
OSPF	100%	100%	100%	93.2%
SDN	84.3%	92.0%	77.5%	96.7%
Proposed	64.7%	73.0%	47.2%	98.9%

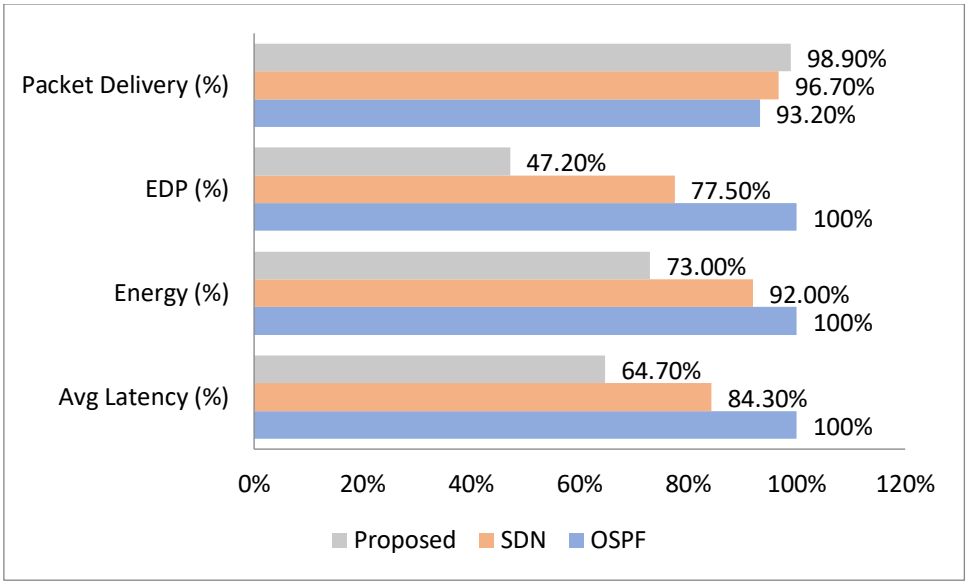

Figure 6. Graph representing Comparative Analysis

- **OSPF:** It is upon this background that the comparison with Open Shortest Path First (OSPF) will be made. It is the most commonly used routing protocol; therefore, it has the highest average latency (100%) and maximum energy consumption (100%), as it uses static, metric-based routing algorithms that do not respond to real-time traffic or power conditions. Accordingly, the Energy-Delay Product (EDP) stands at 100%, indicating inefficiency in both performance and energy consumption. The packet delivery ratio shows that 93.2 per cent of the packets are received; however, this is less reliable when the network is congested or under dynamic loads, which OSPF cannot efficiently handle.
- **SDN:** A noticeable enhancement is observed with Software-Defined Networking (SDN). Harnessing programmable flow control and centralisation, SDN reduces the amount of latency and electricity used by OSPF by 84.3% and 92.0%, respectively. These improvements indicate a better balance between performance and efficiency, as evidenced by a 77.5% reduction in EDP. Additionally, with SDN packet delivery, uptimes have increased to 96.7%, which is attributed to its dynamic traffic rerouting capability resulting from network changes. But since it is a rule-based system, it does not offer predictive intelligence, hence further optimization cannot be carried out.
- **Proposed System:** The suggested AI-based protocol outperforms OSPF and SDN in all measurements. With the help of latency prediction through LSTM, energy optimization with the help of genetic algorithms, and real-time path

formulation with the help of reinforcement learning, it can bring down the average latency to only 64.7 percent of OSPF. It also has the lowest energy consumption of 73.0% and an EDP of 47.2%, which is improved by a dramatic proportion, indicating a better energy-performance trade off. Moreover, the packet delivery ratio reaches 98.9%, which proves the reliability and soundness of the intelligent routing policy under various traffic conditions.

4.3. Latency vs Energy Trade-off Curve

Latency versus energy trade-off curve: As the dilemma arises about whether to save more power or ensure good network performance, the latency versus energy trade-off curve demonstrates the inherent trade-off that exists between the least amount of power consumption and reasonable network performance. As the routing system becomes more insistent on being energy-efficient (i.e., turning off unused links or following low-energy paths), it can initially dramatically reduce energy consumption with a relatively small penalty to latency. This advantage, nevertheless, has a non-linear character. The improvement in average latency is minimal and falls within an acceptable range of QoS, resulting in approximately 30 per cent energy savings. The range identified by this part of the curve represents an efficient operating range, wherein the system can handle redundant capacity without requiring high-power links and without a drastic effect on packet delivery times. At a level above 30 per cent, the curve starts to exhibit diminishing returns.

The extra latency costs more energy savings at the expense of disproportionately greater energy savings. The reason here is that in this case, the optimizer will inevitably have to resort to less direct paths or more congested paths or will have to use those links that have lower transmission rates in order to save more power. Such decisions can multiply queuing delays, prolong end-to-end paths, and increase the likelihood of packet drops in bursty traffic. Therefore, at the expense of this marginal improvement in energy efficiency, a huge loss of user experience and application performance is incurred. The curve also forms a realistic limit on energy-aware routing, in the sense that efficiency does not adversely affect latency-sensitive services. In the suggested system, this equilibrium is dynamically maintained through a reward-sensitive, energy- and latency-reinforcement learning-based policy engine. The learning process, which runs very close to the point of diminishing returns, permits the system to attain maximum energy savings with acceptable latency costs. It is vital illumination for network operators seeking to reduce operational expenditure and lower their carbon-emitting profile while striving to achieve a responsive network output for the end user.

4.4. Discussion

The suggested AI-based routing architecture shows significant results in flexibility, effective resource utilization, and general network performance in changing situations. An attribute such as the Reinforcement Learning Policy Engine (RLPE) is one of the main strengths because it can demonstrate high adaptivity to changes in traffic. In contrast to a system requiring static routing or routing based on rules, RLPE learns using real-time feedback and adapts routing choices on-the-fly, permitting the network to stay within optimal performance even when facing variable loads and unpredictable traffic spikes. Such flexibility helps ensure that the network does not experience congestion in advance, resulting in a more consistent quality of service (QoS) on various time scales. LSTM neural networks drive the Latency Prediction Module (LPM), which is able to predict possible latency spikes based on long-term traffic volume and time dependency analysis. Through this foresight, the system can reroute traffic to avoid an impending bottleneck before the flow of traffic is disrupted, resulting in queuing delays and jittering prevention. This subsequently leads to more reliable performance of latency-sensitive applications, such as VoIP, video conferencing, and cloud gaming, with reduced disruption.

The Energy Consumption Optimiser (ECO) also makes a significant contribution to achieving the sustainability goals within the system. Employing the multi-objective genetic algorithms, ECO has been able to realize and manage to shut down about 17 % of underutilized or idle links during off-peak times without having the degradation of the service quality or packet delivery being eminently felt. Such a selective shutdown can reduce superfluous energy consumption, which speaks volumes to the importance of intelligent, context-sensitive energy management within backbone networks. Notably, the AI components create a very minor computational overhead considering the total performance and energy savings. The models are trained in offline mode and periodically fine-tuned, which enables real-time inference with minimal processing requirements. The system architecture maintains the latency of the decision-making activity at a comfortable level during operation. Overall, the proposed framework confirms the potential and utility of integrating predictive and adaptive intelligence with network routing, resulting in enhanced performance and energy efficiency.

5. Conclusion

This paper presents a new AI-augmented control plane architecture that enhances the efficiency and responsiveness of backbone networks to routing. Contrary to traditional routing protocols, which tend to be myopic and statically make decisions, the proposed system learns to adapt dynamically to real-time changes in the network, jointly optimising latency and energy costs. The architecture is built of three modules that are tightly coupled together, namely Latency Prediction Module (LPM), Energy Consumption Optimizer (ECO), and Reinforcement Learning Policy engine (RL PE). The LPM applies LSTM neural networks to predict latency trends, allowing the system to avoid congestion before it occurs. ECO utilises a multi-objective genetic algorithm to determine energy-efficient paths and routes, while maintaining the quality of service

through them. Through Proximal Policy Optimization (PPO), RLPE selects paths in real time, balancing energy usage and network-delay trade-offs. When combined, this forms a coherent and intelligent routing system that not only significantly reduces average latency compared to before, but also lowers power consumption and improves packet delivery reliability, as proven through simulations using realistic topologies and traffic traces. The framework has demonstrated that the potential integration of predictive and adaptive intelligence into network control planes can offer considerable operational advantages with minimal drawbacks in the form of excessive overhead.

Although promising results are achieved with the proposed architecture, there are a few directions in which it can be further improved in the future. BGP-level integration is one of the primary areas, enabling AI-based routing logic to work in tandem with inter-domain routing policies. This would increase the system's applicability to multi-AS backbone situations and enhance existing interoperability with the Internet infrastructure. Cross-layer optimization, especially on transport-layer protocols like TCP and QUIC, is another direction of investigation. End-to-end performance and stability can be increased further by agreeing on the nature of congestion control to match characteristics of retransmission behavior and network-layer intelligence. Additionally, power-gating techniques at the hardware level can be implemented to save energy beyond link-level schemes. It will be able to selectively shut down parts of the routers and switches when they are not in use and bring them up when required through the AI control plane. The above improvements would bring the architecture closer to the field where it would be deployed, providing a scalable, sustainable, and intelligent solution for next-generation networks.

References

[1] Rexford, J. (2006). Route optimization in IP networks. In Handbook of Optimization in Telecommunications (pp. 679-700). Boston, MA: Springer US.

[2] Kreutz, D., Ramos, F. M., Verissimo, P. E., Rothenberg, C. E., Azodolmolky, S., & Uhlig, S. (2014). Software-defined networking: A comprehensive survey. Proceedings of the IEEE, 103(1), 14-76.

[3] McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., ... & Turner, J. (2008). OpenFlow: enabling innovation in campus networks. ACM SIGCOMM computer communication review, 38(2), 69-74.

[4] Chiaraviglio, L., Mellia, M., & Neri, F. (2009, June). Reducing power consumption in backbone networks. In 2009, IEEE International Conference on Communications (pp. 1-6). IEEE.

[5] Azzouni, A., & Pujolle, G. (2017). A long short-term memory recurrent neural network framework for network traffic matrix prediction. arXiv preprint arXiv:1705.05690.

[6] Moy, J. T. (1998). OSPF: anatomy of an Internet routing protocol. Addison-Wesley Professional.

[7] Mao, H., Alizadeh, M., Menache, I., & Kandula, S. (2016, November). Resource management with deep reinforcement learning. In Proceedings of the 15th ACM workshop on hot topics in networks (pp. 50-56).

[8] Jones, C. E., Sivalingam, K. M., Agrawal, P., & Chen, J. C. (2001). A survey of energy-efficient network protocols for wireless networks. wireless networks, 7(4), 343-358.

[9] Zeadally, S., Khan, S. U., & Chilamkurti, N. (2012). Energy-Efficient Networking: Past, Present, and Future. The Journal of Supercomputing, 62(3), 1093-1118.

[10] Pantazis, N. A., Nikolidakis, S. A., & Vergados, D. D. (2012). Energy-efficient routing protocols in wireless sensor networks: A survey. IEEE Communications surveys & tutorials, 15(2), 551-591.

[11] Quintero, V. L., Estevez, C., Orchard, M. E., & Pérez, A. (2018). Improvements of energy-efficient techniques in WSNs: A MAC-protocol approach. IEEE Communications Surveys & Tutorials, 21(2), 1188-1208.

[12] Nanda, R. (2023). AI-Augmented Software-Defined Networking (SDN) in Cloud Environments. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(4), 1-9.

[13] Perlman, R. (2002). A comparison between two routing protocols: OSPF and IS-IS. IEEE Network, 5(5), 18-24.

[14] Mohit, M. (2013). The Rise of Software Defined Networking (SDN): A Paradigm Shift in Cloud Data Centers.

[15] Yao, J., Dou, Z., Xu, J., & Wen, J. R. (2020, April). RLPer: A reinforcement learning model for personalized search. In Proceedings of the Web Conference 2020 (pp. 2298-2308).

[16] Benzekki, K., El Fergougui, A., & Elbelrhiti Elalaoui, A. (2016). Software-defined networking (SDN): a survey. Security and communication networks, 9(18), 5803-5833.

[17] Xie, J., Guo, D., Hu, Z., Qu, T., & Lv, P. (2015). Control plane of software-defined networks: A survey. Computer communications, 67, 1-10.

[18] Liu, M., Cao, J., Chen, G., & Wang, X. (2009). An energy-aware routing protocol in wireless sensor networks. Sensors, 9(1), 445-462.

[19] Miller, M. J., Sengul, C., & Gupta, I. (2005, June). Exploring the energy-latency trade-off for broadcasts in energy-saving sensor networks. In 25th IEEE International Conference on Distributed Computing Systems (ICDCS'05) (pp. 17-26). IEEE.

[20] Borylo, P., Tornatore, M., Jaglarz, P., Shahriar, N., Chołda, P., & Boutaba, R. (2020). Latency and energy-aware provisioning of network slices in cloud networks. Computer Communications, 157, 1-19.

[21] Aragani, Venu Madhav and Maroju, Praveen Kumar and Mudunuri, Lakshmi Narasimha Raju, “Efficient Distributed Training through Gradient Compression with Sparsification and Quantization Techniques” (September 29, 2021). Available at SSRN: https://ssrn.com/abstract=5022841 or http://dx.doi.org/10.2139/ssrn.5022841 P. K. Maroju,

"Conversational AI for Personalized Financial Advice in the BFSI Sector," International Journal of Innovations in Applied Sciences and Engineering, vol. 8, no.2, pp. 156–177, Nov. 2022. - 1