The AI Revolution in Networking, Cybersecurity, and Emerging Technologies

OMAR SANTOS | SAMER SALAM | HAZIM DAHIR

THE AI REVOLUTION IN NETWORKING, CYBERSECURITY, AND EMERGING TECHNOLOGIES

IGPs perform route computation based on the Dijkstra shortest path algorithm, relying on statically configured link weights (or costs) that reflect certain link properties, such as bandwidth or delay. More dynamic solutions, which offer better routing optimization, leverage call admission control (CAC) and compute traffic engineered paths using constraint-shortest paths. These constraint-shortest paths can be computed based on dynamically measured bandwidth usage. Alternatively, a path computation element (PCE) may be used to perform global optimization of traffic in the network based on collected topology and resource information. Traffic engineering and PCE are typically used in Multiprotocol Label Switched (MPLS) networks.

A shared characteristic among all of these routing optimization mechanisms is that they are reactive in nature. A failure or SLA violation must first be detected and must persist for a period of time before a rerouting action is taken. Furthermore, even after rerouting occurs, there is no visibility or guarantee that the SLA will be met after traffic is steered toward the alternative path, especially if that path is computed on the fly.

AI, and ML in particular, ushers in a paradigm shift in routing optimization that allows network routing to be predictive rather than reactive. This predictive approach enables traffic to be rerouted from a path, before the occurrence of an impending (predicted) failure or SLA violation, onto alternative paths that meet the application's SLA requirements. It complements the reactive mechanisms that have governed routing technologies so far.

The first step toward predictive route optimization is building statistical and ML models trained with historical network data. These models rely on a variety of network KPIs and statistical features (e.g., Welch spectral density, spectral entropy) to forecast (or predict) the occurrence of an event of interest, such as node failure or link congestion. Different models and routing optimization approaches offer different forecasting horizons (how long in advance can the model predict an event) and varying forecasting granularity (general trend versus occurrence of a specific event).

Mid-term and long-term prediction approaches, which can predict events days and weeks in advance, model the network to determine where and when remedial actions should be taken to adapt routing policies and change configurations based on the network's observed state and performance. While beneficial, they are generally less efficient when compared to short-term prediction approaches that can forecast events within minutes or hours, thereby enabling quick closed-loop remediation of temporary failures or transient degradation in network performance. For instance, such a predictive system can accurately predict an SLA violation and find an alternative path that meets the SLA in the same network. Of course, the availability of an alternative path is highly contingent upon the network topology and its dimensions. Determining the alternative path is a nontrivial task due to the following issue: The proactive rerouting of traffic from path X to path Y can potentially eliminate the poor QoE of the original traffic on path X, but can also impact the traffic that is already in place along path Y. To alleviate this problem, the routing optimization system must predict not only poor QoE along a given path but also significantly better QoE on alternative paths, while taking into account the new traffic mix on those alternative paths. This can be guaranteed by ensuring that route optimization is performed by a central engine that has a global view of the network and the constraints associated with shared resources between traffic flows.

Working implementations of various statistical and ML-driven models have demonstrated the possibility of predicting future events and taking proactive actions for short-term and long-term predictions with high accuracy, thereby avoiding many issues that would have detrimentally impacted application performance and user experience. For example, Cisco's predictive engine is in production in more than 100 networks around the world. These technologies can lead the way toward fully autonomic self-optimizing networks.

Radio Resource Management

Wireless networks are ubiquitous. In fact, the number of wireless endpoints on the Internet has exceeded the number of wired endpoints since 2014. Most users have multiple wireless devices that are always on (e.g., smartphone, tablet, wearable device, connected thermostat). All of these devices consume bandwidth and wireless spectrum. The spectrum is the physical layer in wireless networks, and it propagates away from access points in all directions. If two adjacent access points share the same channel, then their corresponding overlapping cells will end up sharing the spectrum that is normally reserved for each. This phenomenon, referred to as co-channel interference, results in less throughput to the users of these cells.

Radio resource management (RRM) is the process of continuously analyzing the RF environment of a wireless network and automatically adjusting the access points' power and channel configuration, among other parameters, to help mitigate interference (including co-channel interference) and signal coverage problems. RRM increases the overall capacity of the wireless network and provides automated self-optimization functionality to account for dynamic environment changes (e.g., noise, interference, number of users, traffic load). As Wi-Fi technology evolves, the increase in frequency from 2.4 GHz to 5 GHz and 6 GHz requires the spacing between access points to decline. At the same time, deployments have migrated from providing simple coverage to handling dense capacities for thousands of clients. All of this makes RRM even more critical to the operation of wireless networks.

Traditionally, RRM solutions have operated based on dynamic measurements collected from every access point regarding its neighbors. RRM examines the recent historical data (several minutes' worth of information) and optimizes the network's operations based on current network conditions. This process is effective as long as RRM is configured correctly for the type of RF network coverage required. RRM does require manual fine-tuning of parameters depending on the network administrator's learnings about the idiosyncrasies of the environment in which the network is operating. With such fine-tuning, RRM can optimize a deployment of any size or density.

RRM can take advantage of AI to analyze multidimensional RF data and deliver actionable insights for management simplicity. Using historical data, ML models can discover client behavior and network patterns and trends. By analyzing not only access point telemetry but also client device (e.g., mobile phone) telemetry, AI algorithms can make data-driven inferences to optimize the performance of the wireless network and enhance how wireless endpoints operate over time. They can also provide insights into the effectiveness of current configurations and settings, and even recommend adjustments to the most optimal configuration for the network. With AI, RRM can perform network-wide holistic optimizations without being subject to the shortcomings of greedy localized optimizations that could lead to cascading network changes and possibly even disruptions. These optimizations result in significant reduction of co-channel interference and channel changes during peak usage periods, in addition to major improvement in wireless signal-to-noise ratio.

Energy Optimization

Environmental sustainability is top of mind for many governments, businesses, and organizations. Many of these entities are establishing green initiatives and setting sustainability goals to lower energy consumption and limit greenhouse gas emissions. IT infrastructure in general, and networks in particular, have a role to play in the sustainability journey, especially given the fact that the majority of network devices are powered on all the time, and the energy use of network infrastructure has been increasing over the years. To illustrate, data transmission networks around the world consumed 260 to 360 TWh in 2022, or 1% to 1.5% of the global electricity use. This constituted an increase of 18% to 64% from 2015.[1]

The good news is that there is room to drive energy savings from networks by building higher-efficiency hardware and implementing software mechanisms that reduce the power consumption of devices in a manner that is proportional to their traffic load. As a matter of fact, there are many deployment scenarios where network devices can be completely powered down during periods of time when they are not in use. For example, consider wireless access points in a stadium when games are not in season, or access points in university classrooms outside of lecture hours. The conventional approach to turning off networking devices when they are not needed is to either perform the task manually or rely on automation systems that enable the administrator to configure time-based scheduling templates. In those templates, the administrator specifies the time when to power on or off the equipment based on the day of the week.

While these simple solutions work for smaller networks with highly predictable usage patterns, in general the approach does not apply to larger networks with more complex usage patterns. As an example, consider an office building where employees sometimes come in outside of normal business hours to work on urgent deliverables or to handle project escalations. The last thing that they would want to face, in these situations, is the network being down because of static power-saving schedules. Another problem with static scheduling is that it introduces operational overhead for the network administrators who must keep up with configuring and maintaining these schedules in a dynamic and continuously changing business environment.

By monitoring client densities, connection times, traffic volumes, and network usage patterns, AI can identify both the zones of the network where energy optimization can be applied and the points of time when those optimizations should be activated or deactivated. This guarantees that the network infrastructure will draw energy in a manner that is proportional to its utilization. ML algorithms can dynamically learn the time schedules of wireless users in a given network deployment, determine the daily and weekly seasonality patterns, and then drive the automatic powering down and powering up of access points in a specific zone of a building or floor depending on predictions derived from those patterns. Such an AI system can dynamically react to out-of-profile usage by continuously

[1] www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks#overview

monitoring clients in a wireless cell to quickly adapt to a sudden unanticipated surge in demand. This ensures network availability and user quality of experience at all times.

AI can also drive more granular power optimization within network devices. An ML trend analysis algorithm can monitor the volume of traffic flowing over the individual member links of an Ethernet link aggregation group, and then automatically power down/up the transceivers of one or more members of the group depending on traffic load. For instance, if the group consists of five member links, and the AI agent predicts that the traffic load will not exceed the capacity of two member links for the coming minute, then it can safely power down three member links for that duration. Similarly, an AI solution can monitor the energy draw from multiple supplies within a switch stack, and can decide to turn off a subset of the supplies to increase the efficiency yield of the remaining power supplies. This is predicated upon the fact that power supplies achieve better efficiency at higher load. The AI algorithm would be adapted to the specifics of the power supply characteristics so that it could determine the right thresholds for taking a supply offline or online.

AI algorithms can help optimize the energy consumption of computer networks by allowing them to dynamically adapt to changes in usage and demand, thereby reducing energy waste and greenhouse gas emissions.

AI for Network Security

Network security entails the protection of the networking infrastructure from unauthorized access, misuse, or data theft. It involves mechanisms, systems, strategies, and procedures to create a secure infrastructure for users, devices, and applications to operate in. Network security combines multiple layers of defenses that augment one another. These defenses are positioned at the edge as well as within the network. Each layer enforces policies and implements controls that allow authorized users to gain access to network resources, but block malicious actors from carrying out threats or exploiting security vulnerabilities. Numerous mechanisms are employed to collectively achieve network security. Among these are the following key components:

- Access control
- Anti-malware systems
- Firewalls
- Behavioral analytics
- Software and application security

AI is playing a transformative role in several of these components, as discussed in the following subsections.

Access Control

Network access control is the process by which individual users and devices are recognized for the purpose of enforcing security policies that prevent potential attackers from gaining access to the network. This can be achieved by either completely blocking noncompliant endpoint devices or giving them only limited access. The first step in network access control is determining which endpoints are connecting to the network. Put simply, you cannot protect the network from what you cannot see. Once the endpoint devices are identified, the proper access control policies can be applied.

AI plays an instrumental role in endpoint visibility by gathering deep context from the network and supporting IT systems. By combining deep packet inspection (DPI) with ML, it can help make all network endpoints visible and searchable. DPI helps collect deeper context for the endpoint communication protocols and traffic patterns, while ML aids in clustering or grouping the endpoints that share similar behavior for the purpose of labeling and identifying them. In other words, AI-powered analytics help to profile the endpoints by aggregating and analyzing data from various sources, including DPI data collected from switches or routers, identity services managers, configuration management databases, and onboarding tools. The collected data is compared to known profile fingerprints. If a match is found, the endpoint is successfully profiled. Otherwise, ML kicks in to cluster the unknown endpoints based on statistical similarity. Groups of similar endpoints can then be labeled automatically using crowdsourced nonsensitive data (e.g., manufacturer, model) or the groups can be presented to the network administrator for manual labeling.

Moreover, AI can help with endpoint spoofing detection. It is possible to use ML to build behavior models for known endpoint types that are functioning under normal operating conditions. Anomaly detection algorithms can then be applied to the DPI data to analyze it against the behavior models and determine if an endpoint is being spoofed.

Anti-malware Systems

Anti-malware systems help detect and remove malware from the network. Malware, short for "malicious software," is an umbrella term that includes computer viruses, trojans, worms, ransomware, and spyware. Robust anti-malware systems not only scan for malware upon entry to the network, but also continuously monitor network traffic to detect anomalous behavior. Such monitoring is required because sometimes malware may linger in a dormant state in an infected network for days, weeks, or even longer.

The malware threat landscape is changing with the rapid rise of encrypted traffic in the enterprise. Encryption provides greater privacy and security for enterprises that communicate and transact business online. These same benefits, however, can enable threat actors to evade detection and to secure their malicious actions. The traditional mechanisms for malware detection can no longer assume that traffic flows are "in the clear" for inspection, as visibility across the network becomes increasingly difficult. At the same time, traditional threat inspection using decryption, analysis, and re-encryption is often not practical or even feasible for performance and resource consumption reasons—not to mention that it compromises data privacy and integrity. As a result, more sophisticated mechanisms are needed to assess which traffic is malicious and which is benign.

This is where AI comes in: It supports encrypted traffic analysis. In the AI-enabled approach, the anti-malware system collects metadata about network flows. The metadata includes the size, temporal characteristics (e.g., interarrival times), and byte distribution (the probability that a specific byte value appears in the payload of a packet within a flow) of the sequence of packets in a flow. The system also monitors for suspicious characteristics such as self-signed security certificates. All of this information can be collected on traffic flows, even if they are encrypted. The system then applies multilayer ML algorithms to inspect for any observable differences that set apart malware traffic from the usual flows. If indicators of malicious traffic are identified in any packets, they are flagged for further analysis and potential blocking by a security appliance such as a firewall. In addition, the flow is reported to the network controller to ensure that the traffic is blocked throughout the entire network.

Firewalls

Firewalls are network security appliances that monitor incoming and outgoing network traffic and determine whether to permit or block specific flows based on a configured set of security policies. The creation and management of security policies is often an extremely complex endeavor, albeit a critical function of network security hygiene. The process of making simple modifications to policies that won't interfere with or override previous rules is both time-consuming and technically challenging, as there is almost no room for error.

The dynamic nature of networks requires a large volume of frequent policy changes across all firewalls that are deployed, and the complexity of maintaining all these policies across the network creates a significant risk that exposes an attack surface into the network. Innovations in conversational AI and ML can simplify policy management, increase efficiency, and improve threat response. Intelligent policy assistants that leverage generative AI enable security and network administrators to describe granular security policies using natural language; the system can then automatically evaluate how to best implement them across different systems of the security infrastructure. These policy assistants can reason over the existing firewall policies to implement and simplify rules.

Behavioral Analytics

Behavioral analytics is a security mechanism for threat detection that focuses on understanding the behaviors of users and systems (e.g., servers, databases) within the IT environment. With this understanding, behavioral analytics can detect subtle changes in known behavior that might signal malicious activity. This approach differs from other security mechanisms, such as anti-malware systems, that solely focus on signature detection. Behavioral analytics employs big data analytics, AI, and ML algorithms. It can be performed on every element of the IT infrastructure: users, end-devices, applications, networks, and the cloud environment. For the purposes of our discussion here, we will focus on network behavioral analytics.

Network behavioral analytics is concerned with monitoring network traffic to detect unusual activity, including unexpected traffic patterns or traffic to known suspicious sites. The system continuously analyzes traffic and events to track unusual usage of inherently insecure protocols such