The impact of AI on government networks

Verizon is developing a robust model with a strong focus on cybersecurity and critical workloads that demand low latency connections to meet the challenge posed by agencies who want to use Artificial Intelligence products to improve their efficiency. The proliferation of AI technologies places significant strain on network resources, resulting in challenges such as bandwidth limitations, latency issues, network congestion, and increased security vulnerabilities.

This document proposes a formula to evaluate AI workload impact while also considering cybersecurity implications and the performance of critical, low-latency workloads: Impact Score (IS) = (U * AIW * CS) / (B * A * LL), where U is the number of users, AIW is the average AI workload, CS is a cybersecurity risk factor, B is available bandwidth, A is an adjustment factor, and LL is a low-latency requirement factor. This formula helps us determine if the network is overloaded, optimally loaded, or underutilized, while also assessing security risks and ensuring performance for latency-sensitive applications.

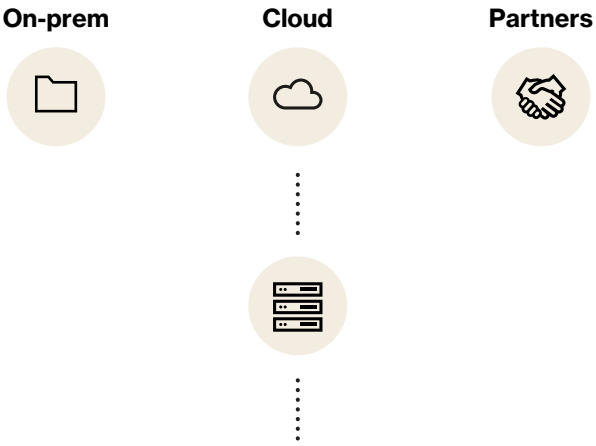
To support this evaluation, data collection is essential. This includes user statistics, AI workload data, bandwidth measurements, network traffic analysis, hardware performance metrics, security threat intelligence, and latency measurements for critical workloads. The collected data will be used to calculate Key Performance Indicators (KPIs) that incorporate cybersecurity and low-latency metrics, which will be integrated into our network monitoring tools for real-time analysis and alerts. The framework can also be used in or as an estimation tool where the formula takes projected data to create the high level output.

Our framework emphasizes visualizing results, generating performance reports, conducting root cause analysis, and performing security audits to inform decision-making and optimize network performance. Verizon's approach includes analyzing AI traffic patterns, quantifying resource demands, developing capacity planning models, implementing dynamic resource allocation, and deploying advanced cybersecurity measures. The goal is to ensure network scalability, optimize performance for AI applications and critical workloads, enhance security posture, and provide network operators with clear visibility and control.

Recommendations include adopting best practices for AI workload optimization, integrating the formula with monitoring and security systems, developing specific policies for AI workload management and security incident response, and prioritizing low-latency connections for critical applications. Ultimately, this framework will enable organizations to effectively manage the increased bandwidth required by larger AI implementations.

verizon

Background and motivation



The surge in AI technologies has led to unprecedented demand on existing network infrastructures. Networks not designed for heavy AI workloads are experiencing issues related to bandwidth limitations, latency, and overall network efficiency.

This whitepaper outlines a practical approach for assessing AI workloads' impact on traditional networks through a mathematical formula. It also presents recommendations for enhancing network performance and stability.

The "why" of Verizon's model is very simple. The Organization's information is all over, literally. From cloud services to partner services information requires three critical things (see below)

Challenges

- Bandwidth limitations due to high-volume data transfers. Bandwidth limitations focusing on Cyber limitations as well.
- Network congestion from AI-driven applications. Competing with standard network traffic generated from collaboration, live meetings and other human network activities.
- Latency issues affecting AI model training and real-time data processing.

Problem statement

The growing demand for AI-driven applications poses critical challenges for traditional network infrastructures. As AI workloads increase, network efficiency and performance can degrade, leading to:

- Poor user experiences due to congestion
- High latency impacting real-time AI operations
- Underutilized networks resulting from inefficient resource allocation.
- Data occlusion[1]

This formula provides a straightforward way to measure the potential strain that AI workloads will place on a network.

At its core, the formula works by creating a simple ratio. It calculates the total data demand generated by all AI users and divides it by the network's effective available capacity. The resulting score shows whether the network has enough resources to handle the AI traffic.

Think of it like assessing traffic on a highway. The formula checks if the number of cars trying to use the highway (AI data demand) is greater than the number of cars the highway can actually handle (network capacity). A score greater than one signals a potential traffic jam (network overload).

AI workload impact formula

The Impact Score (IS) quantifies the total AI-driven load on your network relative to its effective capacity. A score greater than 1 indicates that the demand from AI workloads exceeds the network's available bandwidth.

Formula

Impact Score (IS) = (B×A) (U×AIW)

Variable definitions

- **U:** The total number of simultaneous users generating AI traffic.
- **AIW:** The Average AI Workload per user, measured in megabits per second (Mbps). This represents the typical bandwidth an AI application requires during use.
- **B:** The total provisioned bandwidth of the network link, measured in megabits per second (Mbps).
- **A:** An Adjustment Factor (from 0 to 1) that reflects the real-world bandwidth availability. For example:
  - **A = 1:** The full network link is dedicated to AI traffic.
  - **A = 0.5:** Half of the link's capacity is consumed by other services or unavailable due to contention.

1 Data Occlusion is a term coined in this framework to discuss the concept of data not being fully visible to the AI system of the organization. Hidden or Occluded data results in a reduced data set available to the AI system,.

verizon

How to interpret the results

- **IS > 1 (Network Overload):** The total demand from AI workloads is greater than the network's available capacity. Users will likely experience slowdowns, latency, and poor application performance.
- **IS = 1 (Optimal Load):** The demand perfectly matches the available network capacity. The network is being used efficiently, but there is no room for additional load.
- **IS < 1 (Underutilized Network):** The network has more than enough capacity to handle the current AI workload. There is spare bandwidth available for other applications or future growth.

Once you have the Impact Score, you can derive each user's share of that load via the Per User Output (PUO):

PUO = ISU\text{PUO} \;=\; \frac{\text{IS}}{U}PUO=UIS

In effect, PUO measures whether an individual user's AI demand exceeds (PUO > 1), matches (PUO = 1), or falls below (PUO < 1) the network's per-user capacity.

Detailed explanation of each variable

Number of Users (U):
This represents the count of active users simultaneously consuming AI services—chatbots, inference engines, data-analysis pipelines—across your network. For instance, in a call center scenario you might have 200 agents each interacting with an AI assistant. You would therefore set U=200U = 200U=200.

Average AI Workload per User (AIW):
AIW captures the average throughput required by a single user's AI interactions, measured in Mbps. It is typically estimated by observing the size of inference requests and responses over time. For example, if each user invokes a vision-analysis API that on average sends 3 Mb of data and receives 2 Mb back every second of active use, you would approximate AIW=5\text{AIW} = 5AIW=5 Mbps.

Available Bandwidth (B):
This is the nominal capacity of the network link dedicated to AI traffic—say, a 1 Gbps fiber trunk or a 200 Mbps Internet connection. If your enterprise has purchased a 10 Gbps circuit exclusively for AI workloads, you would use B=10,000B = 10{,}000B=10,000 Mbps.

Adjustment Factor (A):
Real-world conditions—background traffic, link oversubscription, or scheduled maintenance—often reduce the bandwidth you can reliably devote to AI. You model this via AAA, which ranges from 0.5 (only half the link is effectively available) to 1 (the full link is available). If your 10 Gbps circuit routinely sees 30 percent non-AI traffic, you might conservatively set A=0.7A = 0.7A=0.7.

Conclusion

In conclusion, Verizon's strategic approach to managing the impact of using AI products on your network infrastructure, as outlined in this whitepaper, is crucial for ensuring optimal performance, security, and scalability. By implementing the proposed Impact Score formula, collecting comprehensive data, and focusing on key performance indicators that include cybersecurity and low-latency metrics, we can effectively manage the increasing demands of AI workloads.

Our detailed framework, encompassing real-time analysis, visualization, and proactive capacity planning, empowers network operators to make informed decisions. Moreover, by embracing best practices, integrating the formula with monitoring systems, and developing specific AI workload management policies, Verizon will be well-positioned to adapt to the evolving landscape of AI technologies. Ultimately, this framework is not just about addressing current challenges but also about proactively preparing for future advancements, ensuring our network remains robust, secure, and capable of supporting the critical applications that drive our business and serve our customers.

Learn more

For more information, visit verizon.com/federal.

verizon

Formula references

Cisco Systems. (2023). Global Networking Trends Report.

- Covers network bandwidth availability, user demand growth, and AI workload trends.
- https://www.cisco.com/c/en/us/solutions/enterprise-networks/2023-networking-trends.html

Google Cloud. (2022). Optimizing AI Workloads on Distributed Infrastructure.

- Discusses performance implications of average AI workloads (AIW) and low-latency architecture.
- https://cloud.google.com/architecture

NIST. (2022). Framework for Improving Critical Infrastructure Cybersecurity (Version 1.1).

- Offers guidance on quantifying cybersecurity risk (CS) factors in enterprise environments.
- https://www.nist.gov/cyberframework

Gartner. (2024). AI Infrastructure Readiness: Metrics for Network Load and Resilience.

- Discusses balancing AI traffic, user concurrency (U), and adjustment factors (A) under variable loads.

IEEE Communications Surveys & Tutorials. (2021). A Survey on Low-Latency Networks for AI-Driven Applications.

- Provides technical insights into latency-sensitive application requirements (LL) and network stress indicators.

verizon