Prepared for

velocloud™
by Broadcom

The AI-Ready WAN: How VeloCloud Closes the AI Networking Gap

January 2025 EMA White Paper
By **Shamus McGillicuddy**, VP of Research
Network Infrastructure and Operations

EMA™
IT & DATA MANAGEMENT RESEARCH,
INDUSTRY ANALYSIS & CONSULTING

Table of Contents

Executive Summary

More and more enterprises are developing artificial intelligence (AI) solutions. This white paper identifies the unique requirements that AI applications have for wide-area networks (WANs). It explores these requirements in depth and explains how AI-ready software-defined WAN (SD-WAN) solutions, like VeloCloud SD-WAN and the underlying VeloRAIN architecture, can ensure that enterprises can optimize their networks for AI.

Enterprise AI Investment is Accelerating

When OpenAI initially released ChatGPT in November 2022, it set off a tidal wave of new investment in generative artificial intelligence. Adoption of ChatGPT was so rapid and ubiquitous that companies immediately recognized an opportunity to drive business value through AI innovation. In fact, enterprise investment in generative AI technology spiked from $2.3 billion in 2023 to $13.8 billion in 2024.[1] There is no sign of a slowdown.

Companies recognize that they must modernize their networks to facilitate their AI ambitions. Enterprise Management Associates (EMA) research found that AI projects are already a major influence on how 28% of companies build and manage their networks.[2] Much of their attention is focused on modernization of data center Ethernet fabrics for AI workload connectivity. However, companies must also modernize their WANs to facilitate secure and reliable connectivity because data at the edge of an enterprise is critical to AI success.

More enterprise data now lives outside the data center and the cloud, such as in factories, oil fields, research labs, and hospitals. This provides critical insights into business operations, and AI applications must access those insights for training and inference. Thus, WAN modernization will be a critical AI focus over the next few years. Enterprises will discover that AI workloads have unique characteristics and behaviors that will strain existing WAN architecture.

Companies must modernize their networks to facilitate their AI ambitions.

[1] https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/#data-sources
[2] EMA, "Network Management Megatrends 2024," May 2024.

How AI Workloads Impact Wide-Area Networks

Unprecedented Traffic Symmetry Scrambles WAN Bandwidth Paradigm

Enterprises have built out their existing WAN architecture to serve web application traffic, which is highly asymmetric. Web applications send large volumes of data downstream to users and receive very little data in return. For instance, a user clicks on a URL and all the components of a website are sent downstream to the user's device. The user clicks on a button to purchase a product, and a new webpage is served to facilitate the transaction. At each step, only a few packets move upstream, and thousands move downstream. The typical web application has a 9-to-1 ratio of data that users download versus upload.

Consequently, many WAN and internet services are built with this web application ratio in mind, delivering much more download bandwidth to the last mile than upload bandwidth. Cable internet services and DSL typically offer download speeds that are five or ten times faster than their upload speeds. Cellular services like 5G typically offer upload speeds that are half the rate of download speeds. Generally, only fiber connections offer symmetric download and upload bandwidth, but fiber to the edge is relatively rare in today's WAN underlays, which leverage software-defined technology to maximize the value of low-cost broadband.

AI training and inference traffic break this asymmetric paradigm. AI workloads need high-bandwidth connectivity in both directions. For example, companies will upload large files to the AI cloud for both training and inference. AI applications might process security camera footage, medical imaging, or transportation data for traffic management or airport operations. AI workloads have a low tolerance for delayed uploads, so enterprises must ensure significant upstream bandwidth for edge data.

AI workloads need high-bandwidth connectivity in both directions.

AI Workloads Have Low Tolerance for Latency

Most network managers may think they are familiar with latency challenges already. For instance, the covid pandemic led to a permanent spike in real-time communications traffic, and these applications are sensitive to latency, jitter, and loss. However, latency is only an annoyance on a Zoom call. With AI applications, latency can derail critical and complex data processing.

Generative AI workloads participate in distributed application architectures where data is processed in parallel across clusters of servers. Large language models (the foundation of generative AI) run centrally and receive inputs from multiple small language models (SLM) that are running their own processing at the edge. Latency across these architectures delays or breaks distributed AI application processing, which leads to inaccurate training and unreliable inferences.

An AI-ready WAN architecture should have per-application latency tracking and quality of service guarantees around latency. Ideally, the WAN should also be flexible, intelligent, and programmable enough that application developers can deploy their AI applications without having to worry about network design.

AI Applications Challenge WAN Infrastructure Scalability

IT organizations must recognize that the scale of their existing WAN underlays and overlays is not prepared for the demands of AI applications. For example, a user prompt can cause an AI application to request a massive amount of data instantly. The application will connect to dozens or hundreds of edge sites and trigger large file uploads. These uploads will saturate the WAN circuits at data centers and cloud onramps. If there isn't enough bandwidth available in these circuits, the AI application will fail to deliver a satisfactory result to the user.

Thus, WAN circuits must have enough bandwidth to accommodate massive spikes in traffic. With upgrades to WAN connections, enterprises will also need SD-WAN vendors that offer highly scalable appliances with high throughout.

Security must also scale for AI. Because AI applications demand large-scale peer-to-peer connectivity, network security must scale with them. For example, a central LLM will connect with multiple SLM agents in edge locations. These peer-to-peer connections can expand rapidly as a company's overall AI solution grows and encompasses more of the business. SD-WAN solutions must scale out their capacity for establishing an overlay of encrypted tunnels across the WAN. Centralized management of edge security components, such as firewalls, will also be essential. Enterprises will also need to leverage the power of the cloud to scale out core cybersecurity solutions, like data loss prevention and cloud application security brokers, to ensure that they can handle a higher volume of AI traffic.

Because AI applications demand large-scale peer-to-peer connectivity, network security must scale with them.

Bursty Traffic Patterns Demand a Dynamic WAN

Network engineers usually design and build a network based on carefully observed business requirements, such as number of end users and the applications they consume. Network teams identify patterns in digital communications and establish a network that accommodates them. Unfortunately, AI applications are unpredictable.

AI traffic bursts can happen anywhere at any time, based on ongoing training and unexpected insights derived from inferences. For instance, AI applications request massive file uploads from the network edge in response to business events, new insights, new prompts, and outside events that might impact an industry. Networks must adapt to these bursty patterns of AI traffic.

An SD-WAN solution must steer and shape traffic in response to changes in application traffic. By pinning this traffic steering to individual application flows rather than the WAN underlay, an SD-WAN solution can ensure these dynamic network adjustments are aligned with the unique requirements of AI traffic bursts.

How to Build an AI-Ready WAN

Most enterprise WANs are highly abstracted today. The WAN underlay is typically diverse, static, and relatively inflexible, with a variety of public and private connections provided across corporate sites. SD-WAN overlays abstract this connectivity and bring flexibility to the network, and they are increasingly intelligent in how they leverage the underlay to deliver network services.

An SD-WAN solution that can adjust its overlay dynamically in response to the unique requirements of AI can ensure the success of distributed AI application architectures and edge AI deployments. In other words, SD-WAN must leverage the power of AI itself to accommodate the AI revolution.

Enterprises should look for SD-WAN solutions that can identify the presence of AI applications on the network, optimize the network for the unique requirements of AI traffic, and provide intelligent insights and workflows to network managers so they can proactively manage the WAN for AI traffic.

These AI-ready networks must also be highly secure. Data is the currency and the lifeblood of any AI application. These applications consume and process massive volumes of sensitive corporate data that threat actors will target, whether it's at rest or in motion.

SD-WAN must leverage the power of AI itself to accommodate the AI revolution.

SD-WAN solutions must leverage their foundational components to protect this data. Firewalls and other network security functions at the network edge can prevent intrusion into data repositories while allowing approved applications to upload data on demand. Network segmentation can prevent unauthorized access and lateral movement by threat actors against AI applications and data. Finally, end-to-end encryption of SD-WAN tunnels will protect data in motion as massive files upload to the data center and the cloud.

Any networking solution must be inherently secure to protect itself from attacks. SD-WAN solutions should integrate with security service edge solutions, which can layer cloud-based cybersecurity onto the SD-WAN overlay and provider deeper protection against threats and unauthorized access to data.

VeloRAIN: An AI-Native SD-WAN Architecture

To meet the demands of AI applications, Broadcom is building its VeloCloud SD-WAN platform on the VeloRAIN (VeloCloud Robust Artificial Intelligence Networking) architecture. VeloRAIN boosts the power of VeloCloud's existing SD-WAN traffic steering and WAN optimization technologies with new AI and machine learning technology to dynamically optimize the network for edge AI application demands.

VeloRAIN boosts the power of VeloCloud's existing SD-WAN traffic steering and WAN optimization technologies with new AI and machine learning technology.

With VeloRAIN, VeloCloud SD-WAN offers AI-based application identification. Its algorithms can recognize leading AI applications based on their network behavior, then profile AI applications and AI users to ensure the network recognizes and accommodates their requirements.

The solution also offers AI-based network optimization. Machine learning technology can adjust VeloCloud's Dynamic Multipath Optimization (DMPO) algorithm to accommodate rapid bursts in traffic. AI-based dynamic application-based slicing will ensure quality of experience by quickly allocating bandwidth to AI applications as they hit the network.

VeloRAIN architecture enables AIOps for SD-WAN management. This includes a dynamic policy framework that can automatically adjust network policies for the ever-changing needs of AI applications, which improves performance and security. This AIOps capability also features enhanced telemetry based on a combination of retrieval augmented generation and SLMs. It can automatically capture and analyze packets and provide intelligent insights for automated troubleshooting and incident response.

Broadcom offers a series of new edge appliances with up to 100 Gbps of throughput and support for up to 20,000 concurrent overlay tunnels, supporting the augmented VeloCloud capabilities that VeloRAIN enables. These appliances can cluster to scale beyond these limits, too. This ensures that an end-to-end WAN can scale to the requirements of AI from the data center to the branch. These appliances also offer high-availability pairing with fast failover to ensure AI applications never lose connectivity.

A security mindset is foundational to the VeloRAIN architecture. It supports encrypted communication between solution components, and the hardware includes a trusted platform module that provides a secure storage component for sensitive information, like certificates and encryption keys.

VeloRAIN architecture helps VeloCloud software leverage the cloud-based Symantec security service edge capabilities to protect AI traffic with a full suite of Symantec security solutions and integrations with leading third-party vendors. For example, Symantec Data Loss Prevention prevents sensitive AI data from leaving the enterprise perimeter. Cloud Access Security Broker gives visibility and control into AI application activity and prevents use of unsanctioned or "shadow" AI applications.

EMA Perspective

AI investment is growing exponentially as vast segments of the global economy embrace the technology. This investment has massive implications for today's WAN. The unique characteristics and requirements of AI applications challenge existing WAN architecture. To accommodate distributed AI application architecture, IT organizations must adopt SD-WAN solutions that leverage the power of AI to optimize and secure the network. These networks must also scale capacity to meet the requirements of AI applications. Broadcom enhanced its VeloCloud SD-WAN solution with VeloRAIN architecture, an AI-powered evolution that is built to meet the WAN requirements of AI transformation.

IT organizations must adopt SD-WAN solutions that leverage the power of AI to optimize and secure the network.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading IT analyst research firm that specializes in going "beyond the surface" to provide deep insight across the full spectrum of IT management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help its clients achieve their goals. Learn more about EMA research, analysis, and consulting services at **www.enterprisemanagement.com**. You can also follow EMA on **X** or **LinkedIn**.

2770 Arapahoe Road, Ste 132 - 544, Lafayette, CO 80026 +1 303.543.9500 **www.enterprisemanagement.com**

4473.121124