beecham research

BROADCOM

Edge AI for Enterprise: overcoming the challenges of resource-constrained distributed networks

Edge AI combines the localised processing of edge computing with the intelligence of AI models and algorithms, keeping data as close to its source or site as possible. However, despite its multiple benefits, the implementation is a challenge.

As resources become more distributed, the way the network communicates changes. Intelligent edge hardware must increasingly communicate with each other as well as be able to send and receive data from on-premises and centralised cloud systems, creating the need for multiple high bandwidth traffic directions. This is leading to the evolution of a new network architecture. Managing this whilst also optimising application performance, minimising network congestion and ensuring data security poses significant challenges.

To aid enterprises in this new age of AI, Broadcom has introduced VeloRAIN: standing for VeloCloud Robust Artificial Intelligence Networking, VeloRAIN underpins VeloCloud software to optimise AI workloads across distributed enterprise wide-area networks, ensuring quality of experience.

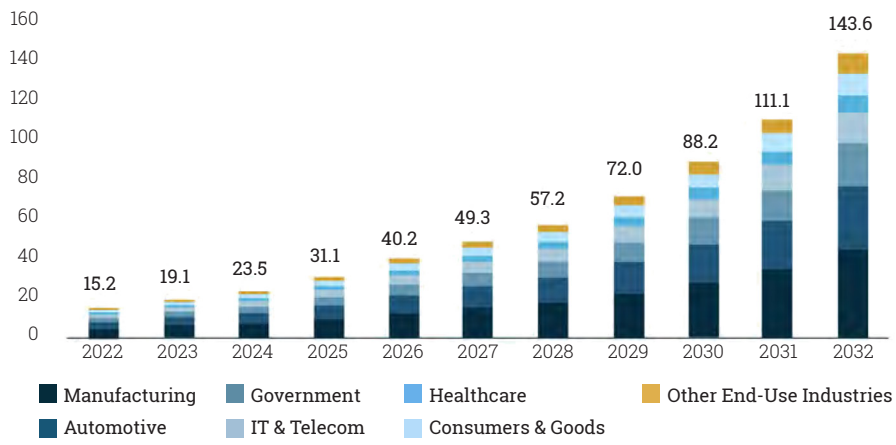
The revenue size for the global Edge AI market is large, with an estimated value of around $20 billion in 2024. The market will see robust growth across industries in the coming years. Per annum growth is expected to be around 25%.

In their 2024 State of Edge AI Report, Wevolver valued the revenue of the global edge AI market as $15.2 billion in 2022, growing to reach $143.6 billion by 2032 with a CAGR of 25.18%. The revenue split by end-user is as shown in **Figure 1**.

North America holds the greatest proportion of the Edge AI market, capturing nearly two fifths of the total revenue. This is thanks to the presence of major tech hubs like Silicon Valley and continuous private and governmental investments in the research and development of AI.

Europe and Asia-Pacific also hold significant shares of the market. Combined with North America, these three regions are responsible for almost 90% of the market.

Figure 1: Projected revenue growth in the Global Edge AI Market split by end-user (2022-2032).



Key drivers for the intelligent edge

There are multiple drivers for the growth of the market, with the key ones being:

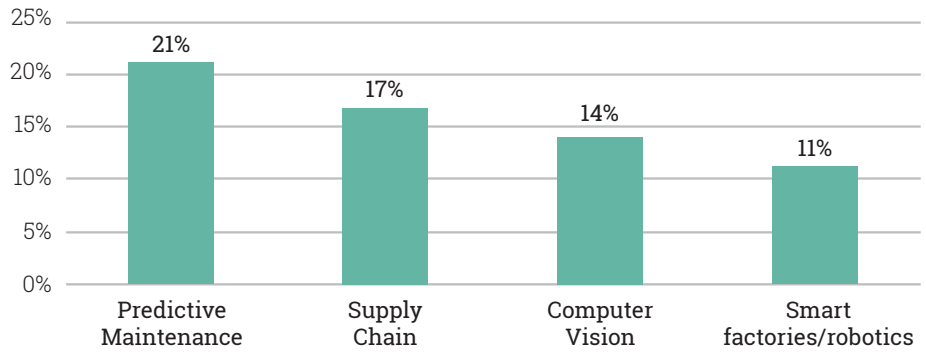
- **Emerging Data Privacy and Security Needs** -processing data locally can reduce the risk of data breaches and mitigate the risk of unauthorised access.
- **Regulatory Compliance** – local processing and analysis of data can help meet data storage and privacy regulations.
- **Scalability and Cost Efficiency** - distributing computing resources across the network avoids the need for massive infrastructure upgrades in centralised data centres and can reduce costs associated with data storage.
- **Emergence of Latency-Sensitive Applications** - including machine monitoring, predictive maintenance, and autonomous control systems.
- **Increased Adoption of Industry 4.0 Technologies** - to create smart factories and revolutionise manufacturing processes.

According to survey data from Broadcom's own State of the Enterprise Edge report, the top benefit organisations hoped to achieve by implementing edge AI solutions was faster response times for latency-sensitive applications, with 68% of respondents citing this. This was followed by 58% of respondents citing improved bandwidth/reduced network congestion as a major benefit to edge AI solutions.

In a similar survey by PSA Certified, 51% of decision makers cited efficiency as a key benefit of deploying AI at the edge, while 41% cited faster data processing/reduced latency.

In terms of applications, the same survey found predictive maintenance to be the top use case for edge AI – see **Figure 2**.

Figure 2: Applications cited by survey respondents as drivers for employing edge AI. Source: Broadcom, 2024, State of the Enterprise Edge Report.



Key challenges in edge AI deployment and management

Despite the significant benefits edge AI promises, challenges remain that negatively impact the implementation, management and scalability of such solutions.

Network Resources

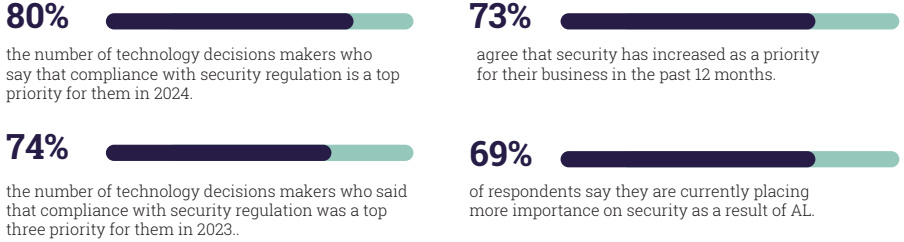
In Flexential's State of AI Infrastructure Report 2024, more than four in five of those surveyed reported "some kind of performance issue with their AI workloads in the past 12 months." Respondents cited bandwidth issues as the most common issue. Unreliable connections and difficulty scaling data centre space and power ranked second and third respectively.

This matches the findings of Broadcom's State of the Enterprise Edge report, which claimed "The driving challenge for scaling edge solutions and AI workloads at the distributed edge is network connectivity issues across locations (57%)."

Security

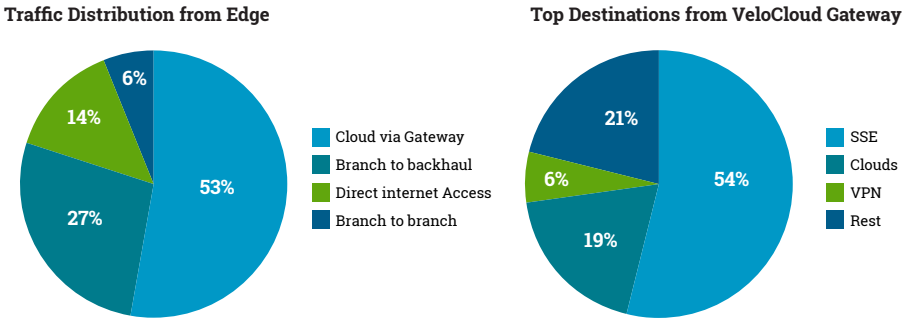
Security is also high-in-mind for edge AI solutions. Of those surveyed in Flexential's State of AI Infrastructure Report 2024, 95% of respondents believe their vulnerability to cyberthreats has increased due to their investments in AI. In fact, data privacy or security is a significant driver for companies moving AI applications and workloads away from the public cloud.

Figure 3: Security remains a major concern for organisations investing in AI, according to those surveyed as part of the PSA Certified Report 2024.



This is corroborated by data collected by Broadcom's VeloCloud SD-WAN Gateways, showing that there is a clear trend towards a distributed architecture and away from a data centric approach – see Figure 4.

Figure 4: VeloCloud deployment data shows movement towards distributed architecture



Nevertheless, of those surveyed by PSA Certified, more than two thirds consider the benefits of AI to outweigh the risk, demonstrating considerable appetite for this technology.

Respondents cited bandwidth issues as the most common issue. Unreliable connections and difficulty scaling data centre space and power ranked second and third respectively.

Edge AI workflows

"The edge" is not a singular mechanism, but its own complex ecosystem of hardware, software, data centres and network connection. These elements can be further categorised into the User Edge and the Service Provider Edge – see **Figure 5**.

Historically, edge components have had limited processing power, storage and battery life, and therefore lacked the capacity to run AI software. Instead, data had to be sent to the cloud for AI processing. This meant that real-time insights and decision-making could not be achieved. It also meant that a successful response within the IoT solution relied on the gateway, network, and cloud connection – a system with a lot of potential break points.

Now, thanks to significant developments in chipsets, sophisticated edge hardware capable of running AI algorithms now exist. The use of 7-nanometer and 5-nanometer manufacturing processes creates smaller and more powerful chips, facilitating higher data loads and low-latency computing.

Figure 5: The Edge Continuum. Source: Linux Foundation.

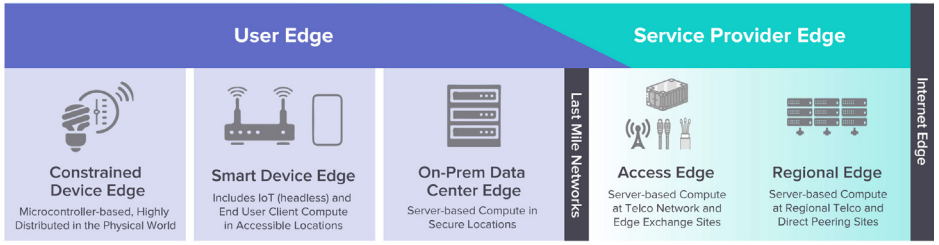
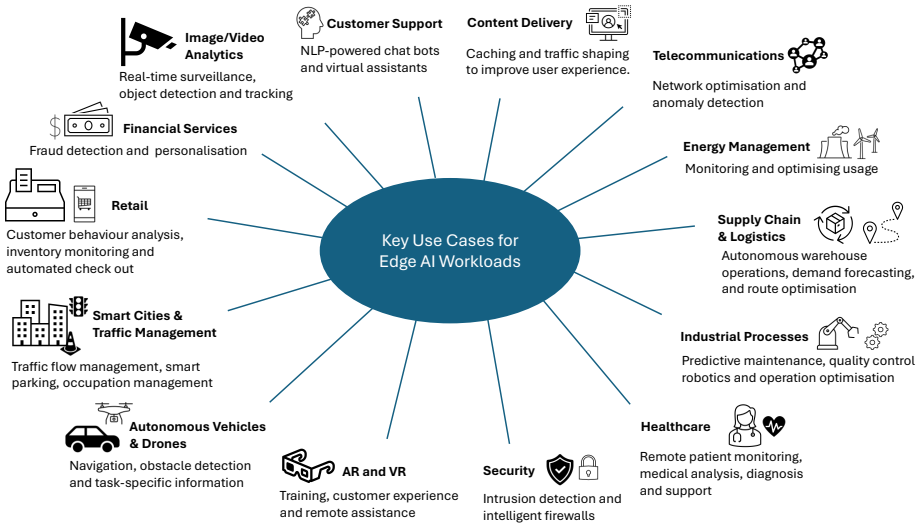


Figure 6 demonstrates a variety of the use cases that edge AI workloads now make possible.



Historically, edge components have had limited processing power, storage and battery life, and therefore lacked the capacity to run AI software

Like most technology, there is a sliding scale of what is possible. In general, the more complex an AI workload, the greater the requirements on processing power and network traffic. Video surveillance is one of the most common applications of edge AI and the levels of complexity it can achieve are shown in **Figure 7**.

To achieve these higher levels of complexity, legacy network architecture is not sufficient. There will be inadequate resources, creating performance and latency issues. As such, to implement successful edge AI solutions, it is important to not only consider the abilities of the hardware and software to be implemented, but also to examine the network architecture. Properly designed and with the support of robust optimisation strategies to handle dynamic workloads, this can minimise network congestion while maximising resource utilisation.

Figure 7: An example of how AI can graduate in complexity in a video use case.

COMPLEXITY

Video AI Workload Type	Capabilities	Processing and network	Data transmissions
1. Basic Detection	Detects movement or objects (e.g., vehicles, people).	Minimal processing and network traffic.	Only metadata is shared.
2. Classification	Identifies object types (e.g., car vs. person).	Moderate complexity and increased traffic.	Metadata and occasional snapshots sent.
3. Tracking	Follows objects across frames, analysing trajectories.	Higher processing and network load.	Metadata and video segments transmitted (e.g. following an individual in security footage).
4. Behaviour Analysis	Recognises actions or events (e.g. distinguishing walking from running or detecting suspicious activities)	High complexity and traffic.	Spatiotemporal data often needed.
5. Context Awareness	Begins to analyse the broader context of the scene and the relationships between objects and their environment.	Very high processing and substantial traffic.	Rich video data transmitted.
6. Predictive Analysis	Forecasts future events (e.g., collision prediction, anomaly detection).	Intensive processing and significant, continuous data sharing.	Detailed frame-by-frame data and predictions may need cloud resources for validation.
7. Real-Time Decision-Making	Immediate responses based on insights (e.g., alerts, controls)	Extreme complexity and very high network load. Constant two-way communication.	Integrates all previous layers while maintaining low latency for real-time responses.

Integrating LLM and SLM: a new network architecture

Based on neural networks, Large Language Models (LLMs) and Small Language Models (SLMs) are a type of AI designed to process, understand and reproduce human language. They are trained by consuming data sets and their accuracy hugely depends upon the quality of the data they learn from – without quality data, they can be subject to bias, factual errors and copyright infringement.

Given that SLMs require less data and are typically programmed to perform a single specific task, they are less expensive to train and (provided the training data is good) likelier to achieve an accurate performance. They are also less energy-intensive than LLMs and therefore more suited to be deployed at the edge.

LLMs are predominantly more suited to data centre or cloud environments where they have the power to handle larger amounts of data, including aggregated and historic information and larger data sets. To achieve lower latency however, these models must be deployed closer to the data source, in data centres at the network edge - between edge hardware and the cloud.

The interaction between SLMs and LLMs in a data centre can significantly influence the performance of an edge AI solution, impacting network congestion, scalability and efficiency.

A network comprised of multiple interconnected SLMs and LLMs and the cloud forms a highly flexible network architecture that operates in the following way:

Distributed Processing

- SLMs at the edge perform lightweight inference or pre-processing of data close to the source (e.g., IoT devices, smartphones).
- LLMs in data centres handle more complex, resource-intensive tasks like fine-grain reasoning or context-heavy inferences.

Data Flow

- SLMs reduce upstream data traffic by filtering, summarising, or encoding data before sending it to LLMs, minimising unnecessary bandwidth usage.
- Raw data is replaced with intermediate representations or prompts optimised for LLM consumption.

Hybrid Model Inference

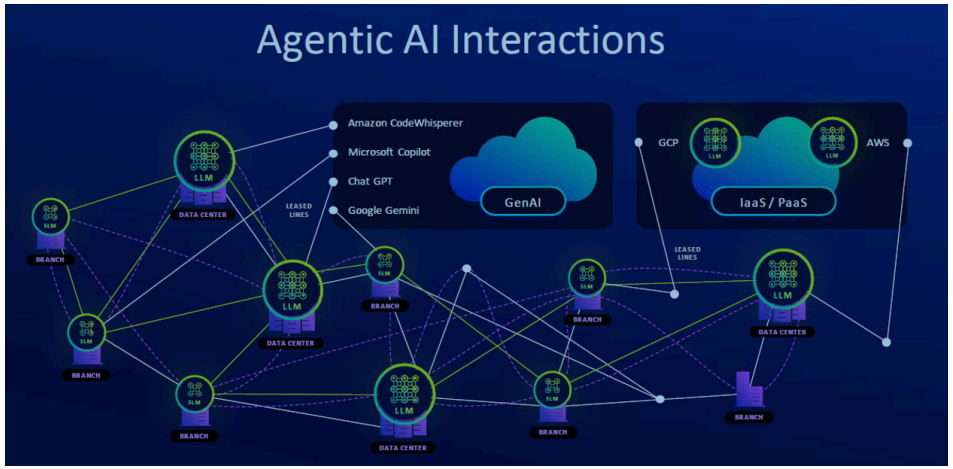
- SLMs can offload partial computation, where only unresolved queries or intricate tasks are forwarded to LLMs for detailed processing.

Feedback Loop

- LLMs can refine, validate, or augment the outputs from SLMs, with results sent back to the edge devices, ensuring consistent and context-aware responses.

Figure 8: An interconnected network of SLMs and LLM facilitate agentic AI interactions

Source: Broadcom



Network challenges and considerations of the agentic AI model

To achieve the benefits of the agentic AI model, the solution must attain an adequate balance between performance and complexity. The edge is naturally a resource-constrained environment where restrictions such as processing power, memory, storage capacity and network bandwidth can affect the AI workload. This can lead to congestion when transmitting large or frequent data payloads to data centres, which in turn could impact latency to the detriment of time-sensitive tasks.

In addition, workloads can be dynamic and unpredictable, particularly at the SLM nodes. This may result in erratic surges in communication with LLMs, straining network resources, energy consumption and the central models. Furthermore, as the number of edge devices grows, the infrastructure faces challenges in efficiently handling large volumes of irregular traffic.

There are nevertheless several design considerations developers can implement to help mitigate network congestion associated with this model. Potential solutions include:

- **Efficient Data Encoding:** developing lightweight, compressed data representations from SLMs to reduce payload sizes sent to the LLMs.
- **Prioritisation Mechanisms:** designing communication pipelines that prioritise critical data while deferring or batching non-urgent requests.
- **Model Cooperation:** optimising SLM and LLM architectures for symbiotic interaction, ensuring that SLMs handle most tasks independently.
- **Edge Storage:** cache intermediate results and frequently accessed data at the edge to limit repetitive queries to the data centre.
- **Dynamic Workload Distribution:** use AI-driven network orchestration to balance processing between edge and core based on real-time traffic and model availability.

Ultimately, the SLM-LLM interaction is a key enabler for scalable and efficient AI systems, particularly in environments with limited network capacity or high latency requirements. Properly designed, this interaction minimises network congestion while maximising resource utilisation, but it requires robust optimisation strategies to handle dynamic workloads and ensure seamless communication between edge and data centre infrastructure.

Advancing Edge AI adoption with Broadcom

Broadcom is a global infrastructure technology leader whose solutions power the most complex IT environments in the world. Partnering with global companies across industries (including healthcare, utilities, automative, government, telecommunications and financial services), Broadcom's offerings range from semiconductor and infrastructure software products to network and security solutions.

Broadcom is also deeply involved in advancing Edge AI technology, including through its VeloCloud portfolio. This effort focuses on providing enterprises with the infrastructure and tools to process AI workloads at the edge, closer to where data is generated.

VeloCloud SD-WAN

VeloCloud SD-WAN is a key component of Broadcom's offerings. It enables enterprises to blend multiple connectivity options like Fixed Wireless Access (FWA) and satellite networks along with more traditional connections such as fibre, broadband and MPLS. This provides seamless and redundant connectivity, which is vital for running advanced applications at the edge. Additionally, the integration of Symantec's security capabilities ensures secure, automated cloud access for these workloads.

The VeloCloud SD-WAN Edge 4100 and 5100 modules are advanced SD-WAN appliances with throughput up to 100 Gbps, designed to meet the demands of large enterprises, regional hubs, and data centres. They offer a range of features to enhance network performance, scalability, and security.

Figure 9: Introducing VeloCloud Edge 4100/5100 models for large enterprises.

VeloRAIN: a new network architecture for AI

VeloRAIN, standing for VeloCloud Robust AI Networking, is a new network architecture designed specifically to manage AI workloads across distributed systems. Capable of understanding an application's needs and dynamically adjusting network resources to prioritise critical applications, VeloRAIN helps minimise latency, manage bandwidth and improve model interaction efficiency. This enables enterprises to use AI capabilities without compromising on network performance or user experience.

The three pillars of VeloRAIN

The VeloRAIN architecture is driven by three main aspects which enable it to optimise AI applications.

1. AI-driven application profiling
VeloRAIN identifies and prioritises applications with new intelligent capabilities. By identifying applications accurately through an enhanced machine learning (ML) system, VeloRAIN ensures that each AI and non-AI app receives appropriate network resources.

VeloRAIN will also be able to identify encrypted application traffic that was previously unreadable. This is particularly useful for sectors like healthcare and banking, where sensitive data must be securely processed while ensuring high application performance.

2. AI-based network optimisation
VeloRAIN introduces Dynamic Application-Based Slicing (DABS), an innovative technique that focuses on the application layer instead of the traditional network-based slicing. Through incorporating user profiles, VeloRAIN can understand identity and business needs enabling it to prioritise traffic, provide the necessary bandwidth and perform critical applications automatically.

For instance, businesses with multiple AI-driven applications (such as e-commerce platforms and real-time customer support bots) can implement VeloRAIN's dynamic policy framework to prioritise customer-critical processes over less time-sensitive ones. This ensures Quality of Experience per application.

3. AIOps with real-time data
With data from the vast VeloCloud SD-WAN deployment base, VeloRAIN employs AI to automate network operations (AIOps), using anonymised real-time data to dynamically adjust policies. For example, in retail stores, vision applications analysing in-store behaviour could benefit from real-time policy adjustments, ensuring seamless experiences and high-value customer retention. This dynamic approach allows enterprises to adjust network configurations on-the-fly, ensuring the best possible experience for users and applications.

Figure 10: The three pillars of VeloRAIN and compatible edge apps.

Why is VeloRAIN the right choice for AI networking?

In recent years, there has been a perception that the advanced capabilities of SD-WAN might not be necessary as enterprises utilise cloud-based models for processing- and power-heavy tasks. However, cloud-based models cannot facilitate latency-critical applications and will become increasingly hard to maintain as storage, bandwidth and power requirements grow in line with the vast amounts of data that are being generated.

A distributed model facilitates quick, localised processing of data, supporting low-latency tasks and freeing up space in the centralised cloud to perform more computationally intensive tasks.

VeloRAIN provides a robust framework for optimising the SLM-LLM interactions of an agentic AI model, addressing critical challenges in latency, bandwidth, and scalability. By enabling smarter workload distribution and resource utilisation, it empowers AI systems to deliver faster, more reliable, and cost-effective services across edge and cloud infrastructures. This makes VeloRAIN particularly valuable in scenarios where real-time processing, low network availability, or large-scale deployments are key considerations.

Ultimately, it addresses the unique demands of AI workloads (characterised by bursty, latency-sensitive traffic patterns) and offers secure, optimised connectivity across enterprise networks, from data centres to the edge.

Case Study

Advancing Greenhouse Innovation and Farm-to-Fork Time with AI a the Edge

Innovation at the edge is happening everywhere, including massive greenhouse growing environments. One of the largest greenhouse-grown produce leaders in North America uses VeloCloud SD-WAN to enable distributing AI networking outside the data centre. The solution supports more consistent, standardised growing methods, faster produce delivery to consumers, and better-tasting fruits and vegetables.

Unlocking crop insights at the edge

The company has long relied on technology to help grow crops like tomatoes, strawberries, and peppers on more than 250 greenhouse acres in Canada and the United States. Each of its1.8 million plants generates approximately 25GB of data per month. The produce company has nearly 200,000 sensors throughout its greenhouses—and is continually adding more. Its growers use this data to understand plant growth and optimise plant care.

The company also uses AI cameras and systems to grade each vegetable based on criteria like size and colour, which helps determine their ripeness. As IT and OT merge, and AI plays an increasingly critical role in its business, the company needed a solution to support its growing edge-based workloads. VeloCloud SD-WAN enables the produce company to move its vast flow of data from the edge to the core.

Flexible greenhouse connectivity

The produce company connects its huge network of greenhouses and growers with VeloCloud SD-WAN. Approximately half of the data produced by each plant is sent to the core for processing, and the rest is processed on databases at the edge. The solution's built-in LTE/5G option gives the company the flexibility needed for deployment in rural areas that would be challenging for a traditional terrestrial network. With VeloCloud SD-WAN, the produce firm can choose between using a cloud solution to upload the data or allowing data to go directly back to the core, depending on its origination point. Each location can connect seamlessly, regardless of its location.

The speed and ease of use of VeloCloud SD-WAN has enabled the company to support its diverse connectivity requirements. The solution has helped the organisation understand and look for latency issues in its IoT devices and discover potential issues before the growers even notice—and prevent network problems from affecting plant production.

Security and simplicity

The produce company uses VeloCloud SD-Access to provide easy, secure remote access for remote users and IoT devices anywhere, optimising connections for speed and reliability and bridging the needs of IT and OT. The company can get connections up and running within a few hours, depending on the use case, and make changes on the fly if needed.

VeloCloud SD-Access supports a variety of key growing processes. The solution is installed on a 5G-enabled scouting robot that examines and tracks the fruit and colour of each pepper plant, letting growers know the ideal time to harvest.

The solution is also key for highlighting its innovative growing methods on a mobile greenhouse that travels to schools and events, demonstrating the AI-enabled systems that adjust factors such as lights and irrigation. In the past, the mobile centre could not flow data quickly enough to demonstrate these capabilities. Now instead of one 5G connection, the company can individualise three 5G connections to get data back to its main AI core for fast processing.

Visibility across a continent

One of the organisation's biggest challenges is keeping a close eye over hundreds of acres of operations throughout North America. Using VeloCloud SD-WAN, the produce company can manage them all as one. The organisation can tie all its facilities together, scale its growing and business operations more effectively, and keep all its processes consistent.

For consumers, the end result of these edge-centric innovations is the best possible vegetables—and rapid delivery at the peak of freshness.