beecham research

BROADCOM

Edge AI driving Telcos to improve operations, offer differentiated services, and monetise 5G.

The telco landscape is under pressure: with customers demanding more data, faster speeds and greater coverage and reliability across all regions on Earth, the need to improve operational efficiency, deliver superior services, and unlock new revenue streams while leveraging 5G networks is huge. Furthermore, as resources become more distributed, the way networks communicate is changing.

By investing in lightweight AI models, collaborative ecosystems, and predictive network optimisation, Telcos can harness edge AI to enable real-time decision-making, improve customer experiences, and establish a sustainable, future-ready network infrastructure. However, challenges such as infrastructure scalability, low latency demands, security, energy efficiency, and compliance must be addressed.

To aid telcos in implementing successful edge AI strategies, Broadcom now offers its innovative VeloRAIN architecture. Standing for VeloCloud Robust Artificial Intelligence Networking, VeloRAIN builds on existing technology to optimise AI workloads across distributed wide-area networks, ensuring quality of experience, customisable workloads and seamless, scalable operations.

Telco operations: the current landscape

The Challenges

Demand for data is soaring. Consumers and businesses around the world are spending more time connected, indicating a significant shift in the way we live and work. According to Analysys Mason, worldwide data traffic is growing at a rate of 24% per year, leading to an expected consumption of 1.8 petabytes by 2026.

As more data is consumed, it becomes more difficult for telcos to manage operations. They need to ensure a reliable connection and deploy robust networking mechanisms to handle routing, congestion or network outages. As networks grow there are further challenges in terms of data management, adapting to new technologies (like 5G) and ensuring overall infrastructure scalability.

In addition, customer expectations are rising. The notion of 'always-on' connectivity that is fast, reliable and secure is increasingly being taken for granted. Enterprises are developing applications requiring real-time processing and low-latency transmission, most of which have the complication of dynamic high-bandwidth requirements. Furthermore, customers are increasingly considering the values of the providers they choose, making it important to achieve operational effectiveness and efficiency in an ethical, energy-efficient and sustainable way.

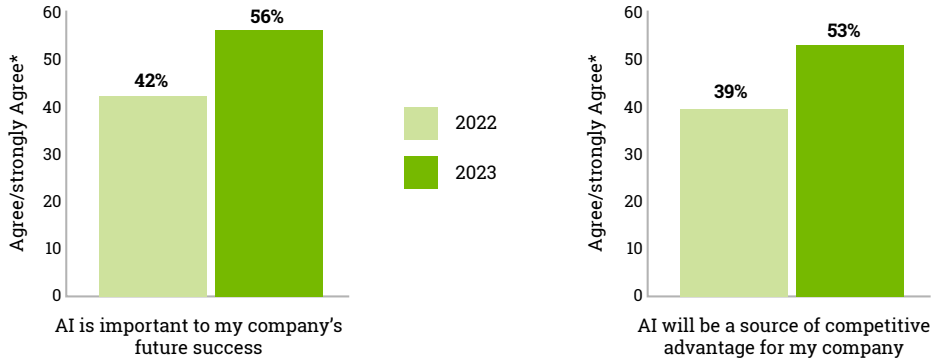
Managing these challenges while adhering to the stringent regulations of the telecommunications market is a substantial task. In addition, the multi-competitor landscape gives telcos little pricing power in the market. In fact, Analysys Mason predicts that across 2023-2028 the global telecoms service revenue will record a CAGR of only 1%.

The rise of AI in telecommunications

To manage the wealth of challenges now apparent in the telco landscape, companies are increasingly turning to AI to help solve their problems.

In a recent AI telecommunications report, NVIDIA stated that 90% of the 400+ participants surveyed were currently engaged with AI (either at the assessment/pilot stage or at the implementation/using stage). This is due to a strong belief that AI is important to a company's future success. The number of respondents believing this has notably grown from the previous year's survey – see **Figure 1**.

Figure 1: Interests and Expectations from AI. Source: NVIDIA Telco State of AI in Telecommunications 2024 Trends



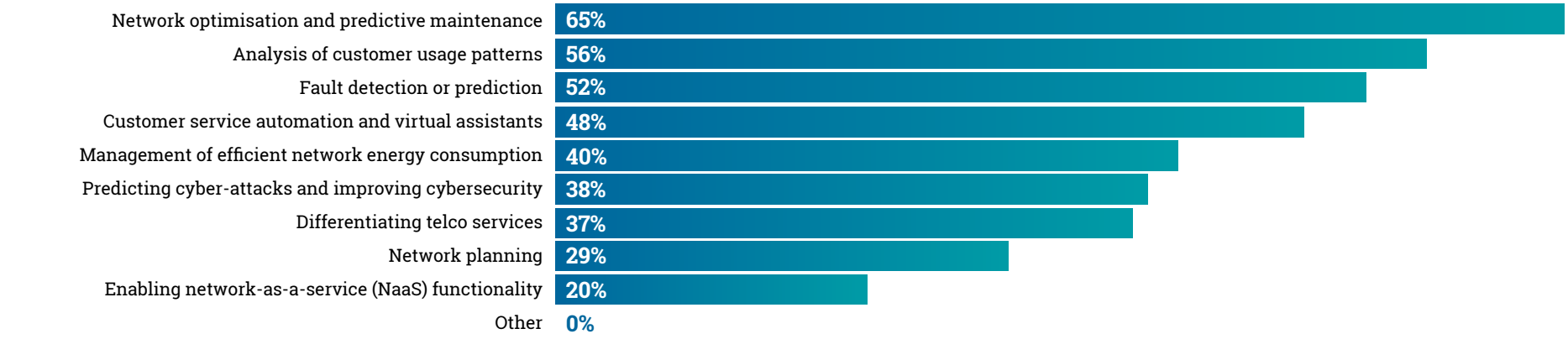
*Agree/Strongly Agree are the top two points on a 7-point scale

There are many reasons for use of AI in telecommunications. Companies are investing in a variety of AI technologies (including machine learning, deep learning, generative AI, high performance computing and digital twins) to transform telecoms from a manual and reactive management of networks system to a more proactive and dynamic operational system.

According to Telecom's Annual Industry Survey 2023 Report, industry leaders consider network optimisation and predictive maintenance to be the most important application of AI, followed by customer usage patterns and fault detection or prediction – see **Figure 2**.

However, telcos must consider where to implement AI to get the most from their solutions and gain the greatest benefits. Cloud is no longer the only feasible option; now, edge technology has become sophisticated enough to enable the integration of more advanced computing and processing solutions, including AI.

Figure 2: Telecom's Annual Industry 2023 Survey - the best applications of AI in telecoms.

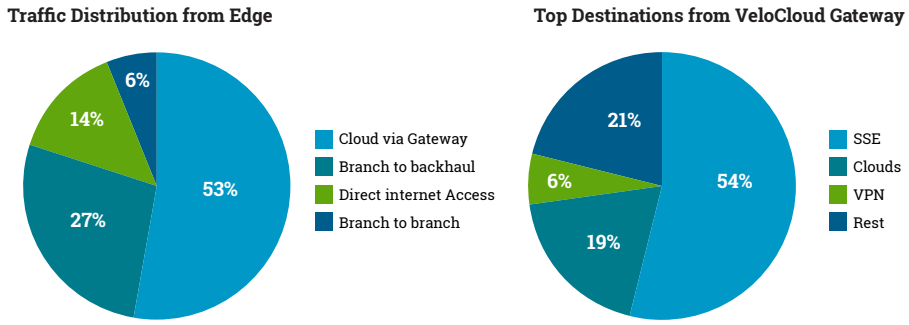


Beyond the Cloud: telco AI adoption at the edge

Traditionally, AI workloads were confined to centralised data centres. However, with Gartner predicting that 75% of enterprise traffic will originate outside the data centre by next year, the shift toward edge AI is pivotal. While large AI models may remain in the cloud for managing historical and large-scale data, edge computing empowers telcos and enterprises to process data locally.

This is corroborated by data collected by Broadcom's VeloCloud SD-WAN Gateways – see **Figure 3**.

Figure 3: VelCloud deployment data shows movement towards dustributed architecture
Source: Broadcom



There are many benefits to this distributed approach, including:

Enhanced Operational Efficiency
Reducing the need to transmit massive datasets back to a central location alleviates network congestion, improves performance and speed, and lowers operational costs. AI models can also be used to identify network trends and deficiencies in real-time, enabling telcos to fine-tune the network dynamically.

Unique Customer Experiences
By leveraging edge AI, telcos can offer intelligent, application-specific services (such as differentiating between banking and video streaming traffic) and ensure optimal network performance. Ultra-low latency improves the speed with which it can query data or generate reports, enhancing customer satisfaction, leading to additional applications. Furthermore, AI can be used to improve engagement through personalised virtual experiences and assistance.

Enhanced Security and Privacy
Enterprises prefer on-premises processing to prevent sensitive data from leaving their sites. Processing data locally and sending only inferences to the cloud addresses privacy concerns.

Limited edge resources: the need to create new workflows

Despite many potential benefits, the increasing move of AI to the edge comes with its own set of challenges, negatively impacting implementation, management and scalability.

Typically, as the use of AI proliferates through the edge, customers are noticing three things:

- **AI applications are bursty and latency sensitive** - GenAI interactions typically involve high traffic when the client uploads a large request. This results in a pause as the server processes and prepares a response, then high traffic again when a large response is sent back to the client.
- **AI traffic is encrypted, challenging network optimisation** – if AI workloads cannot be distinguished from other workloads, the network cannot prioritise them. This means they may not receive the bandwidth necessary for optimal performance.
- **AI traffic changes traffic patterns** - AI applications can radically change the upstream/downstream balance and make WAN traffic unpredictable. In a typical network, video streaming traffic is heavily skewed towards the downstream. AI workloads often flip that pattern on its head, as many use cases send video data upstream in massive volumes causing bandwidth and network congestion problems.

As such, legacy architecture cannot handle the higher demands and complexity of AI workflows. Therefore, a new network topology is required.

Enhancing operational efficiency: the power of agentic AI networks

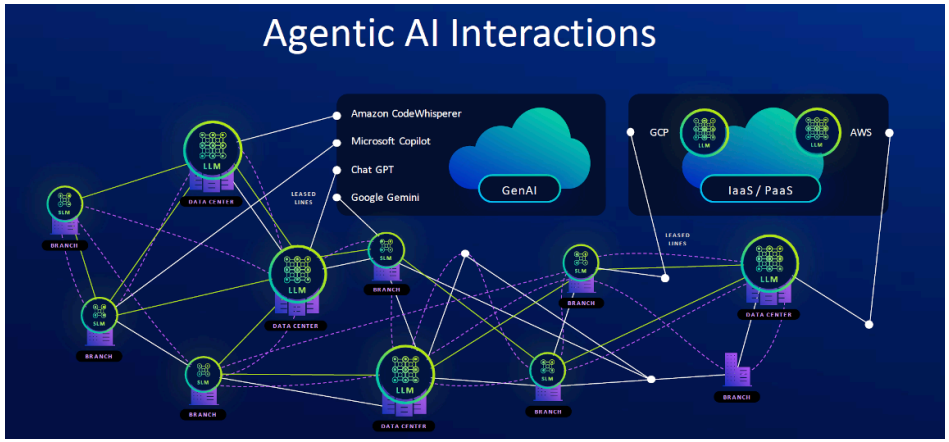
Based on neural networks, Large Language Models (LLMs) and Small Language Models (SLMs) are a type of AI designed to process, understand and reproduce human language.

Deploying LLMs and SLMs at the edge allows for immediate data analysis and decision-making. SLMs, being more resource-efficient, are particularly suitable for edge devices with limited computational power. This setup enables applications such as natural language processing and real-time analytics to operate efficiently at the edge, improving performance and user experience. LLMs, however, are more suited to data centre environments as they have the power to handle larger amounts of data, including aggregated and historic data. 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.

In this paradigm, edge devices equipped with intelligent gateways perform local data processing and inference tasks. These gateways handle data at the point of collection, reducing upstream traffic by filtering, summarising or encoding it, minimising unnecessary bandwidth usage.

SLMs can also offload partial computation, where only unresolved queries or more intricate tasks are forwarded to LLMs for detailed processing. These LLMs can then refine, validate or augment the outputs, with results transmitted back to the edge devices, ensuring consistent and context-aware responses.

Figure 4: An interconnected network of SLMs and LLMs facilitate agentic AI interactions.



Deploying LLMs and SLMs at the edge allows for immediate data analysis and decision-making

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.

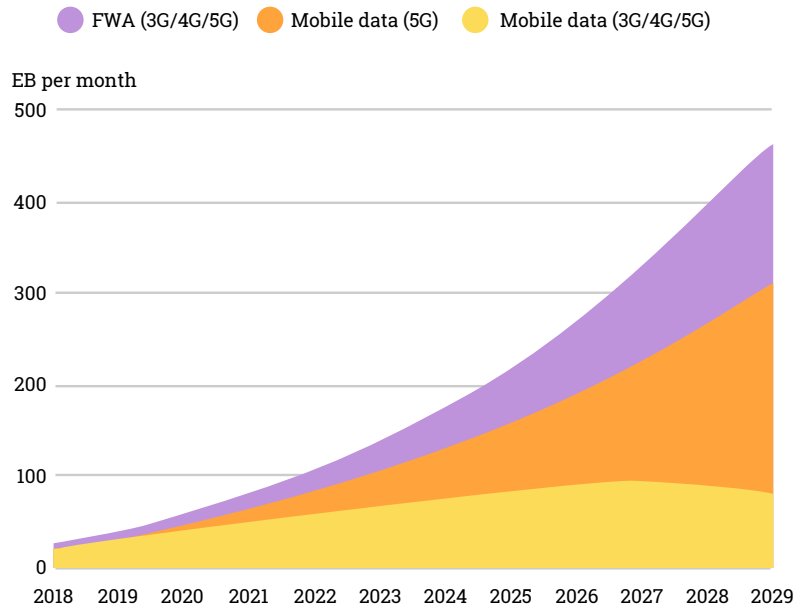
As the number of edge devices grows, the infrastructure faces challenges in efficiently handling large volumes of irregular traffic

Edge AI and 5G: facilitating new revenue opportunities?

Another factor in the development of edge AI solutions for telcos is the rise of 5G. The deployment of this communication technology continues to expand globally, facilitating higher bandwidth and low latency in a greater range of locations.

According to Ericsson's June 2024 Mobility Report, 5G is anticipated to become the dominant mobile access technology (by subscription) in 2028, with its share of mobile data traffic predicted to grow to 75% in 2029 – see **figure 6**.

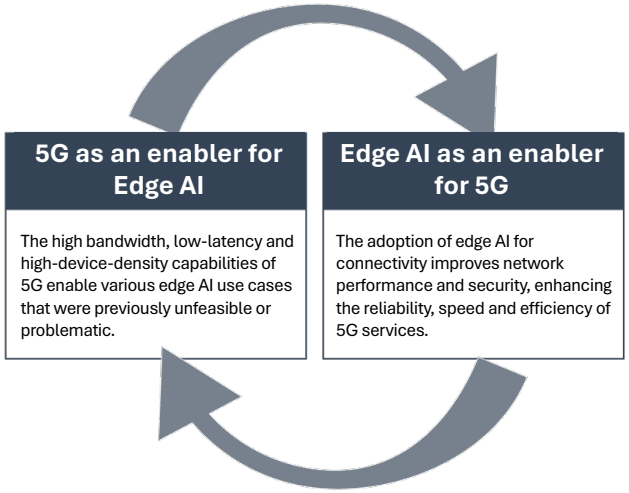
Figure 5: Ericsson's projections for global mobile network data traffic.



As 5G connectivity expands, so do the opportunities for edge computing and AI. Both these technologies are more power-hungry and time-sensitive than legacy technologies.

Therefore, 5G acts as an enabler for edge AI and edge AI for 5G.

Figure 6: 5G and edge AI enable each other, boosting the potential of each technology.



The services and applications possible with edge AI can add huge value to enterprise operations. Therefore, telcos must start looking at ways to enhance their connectivity offerings and instead of embracing a 'race-to-the-bottom' price mentality, start competing by providing unique, feature-rich services. Such services can be offered at greater expense, improving the potential revenue and profit margins for telcos.

While AI is already being used in 4G and 5G networks to streamline operational systems, the expectation is that the technology will be more heavily woven into 6G. Perhaps in the same manner that 5G is considered cloud-native and has been designed with cloud and automation in mind, 6G is likely to be AI-native and designed with intelligent and self-adaptive capabilities, deeply integrated into its architecture.

Advancing telco Edge AI adoption with Broadcom

Broadcom is a global infrastructure technology leader whose solutions power the most complex IT environments in the world. Partnering 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.

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. It introduces intelligent capabilities to identify encrypted application traffic – a significant challenge for typical network optimisation solutions. This advancement allows for better application-based Quality of Service (QoS) and security across all enterprise endpoints.

VeloRAIN also 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 VeloCloud SD-WAN Edge 4100 and 5100 modules are advanced SD-WAN appliances, with throughput up to 100 Gbps. When combined with services on VeloRAIN architecture, these appliances are equipped to process the growing networking demands of large enterprises and AI traffic.

VeloCloud SD-WAN is designed to blend multiple connectivity options like Fixed Wireless Access (FWA) and satellite networks with more traditional connections such as fibre, broadband and MPLS, creating true converged infrastructure that will enable telcos to offer their enterprise customers simplified deployment, unified management and scalability. The ability to prioritise network flows over any underlay 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.

Figure 7: Introducing VeloCloud SD-WAN Edge 4100/5100 models for large enterprises.

Why is VeloRAIN the right choice for telcos?

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. However, as the use of AI proliferates through the edge, customers are struggling with 'bursty' applications, network optimisation and changing traffic pattens.

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, Broadcom's Edge AI strategy empowers telcos and enterprises to overcome operational challenges, improve user experiences, and capitalise on 5G and emerging technologies. By decentralising AI workloads, prioritising privacy, and introducing advanced connectivity solutions, Broadcom positions itself at the forefront of the edge revolution, offering tools that drive efficiency, reliability, and innovation.

Case Study

Enhanced Fixed Wireless Access is Solving Challenges for Distributing AI in Remote Locations: How a Leading Service Provider Leveraged VeloCloud SD-WAN to Incorporate Satellite for AI at the Edge

A telco of integrated voice, data, network, cloud, and mobility IT solutions throughout the United States needed an intelligent software layer to help manage the varying needs of the unique workloads proliferating at the edge. The layer needed not only to understand the applications' needs, but also dynamically adjust network resources to prioritise critical applications, ensuring optimal quality of service while deprioritising less critical traffic.

The provider turned to VeloCloud SD-WAN technology to enable its fixed wireless access (FWA) offering. VeloCloud SD-WAN not only supports satellite connectivity but includes advanced intelligence to meet the unique needs of AI-driven applications.

A look at their service provider offerings

When delivering an SD-WAN service, the telco offers a fully managed network fulfilling all services from broadband and ethernet to FWA and satellite connectivity. This facilitates one-stop accountability and control. To serve more customers in more places, the company wanted to expand its satellite connectivity offerings. After performance issues with other satellite providers, they engaged with a fast-growing leader in Low Earth Orbit (LEO) satellite service. These satellites orbit much closer to Earth, dramatically reducing latency and enabling high-speed internet access anywhere in the world.

However, although LEO satellite networks provide immense potential, ensuring high-quality dependable connections can be challenging. Factors like handovers between satellites, atmospheric conditions and mountainous regions can impact the service's performance. With its proven ability to address broadband performance issues, VeloCloud SD-WAN was chosen to provide a seamless extension to LEO and mobile connections for FWA services.

Circling Back to Superior Experiences for AI

Satellite and cellular transport are supported by all VeloCloud Edge platforms. The intelligence of the VeloCloud SD-WAN provides dynamic connectivity that significantly improves the quality of experience, addressing issues like jitter that can impact satellite communications. With Dynamic Multipath Optimisation (DMPO) remediation, the service provider can increase usable bandwidth by up to six times on satellite links.

VeloCloud Robust Artificial Intelligence Networking (VeloRAIN) is an AI-enhanced architecture that allows VeloCloud to further optimise AI applications. AI-driven application profiling and network optimisation identifies and prioritises applications ensuring that each AI and non-AI app receives appropriate network resources.

VeloCloud also applies machine learning to perform channel estimation for networks, including satellite, 4G, and 5G. By accurately predicting channel conditions, VeloCloud can adaptively manage bandwidth, latency, and jitter, crucial for supporting asymmetric traffic patterns of AI-driven applications.

The combined service offering featuring VeloCloud SD-WAN and satellite connectivity has enabled the communications provider to support a variety of customers with challenging connectivity requirements and prepare them for the unique demands of AI workloads at the edge.

For example, this enabled them to assist one customer - a Florida construction company located in an area that lacked reliable, high-performance wired broadband services. Using VeloCloud SD-WAN technology, coupled with third-party satellite service to provide high-performance connectivity, the telco's FWA service was able to deliver approximately 225 Mbps download speeds and 30 Mbps uploads, with 19 millisecond latency. As the construction company evolves further toward edge AI applications (such as smart sensors for predictive equipment maintenance), the scalability of VeloCloud's SD-WAN service will facilitate delivery of the dynamic connectivity they require.

Unlocking growth and new market opportunities

With its robust satellite and VeloCloud SD-WAN solution, the communications provider is positioning itself to tap into an FWA for enterprise market opportunity that Gartner predicts will grow to $6 billion by 2026. More importantly, it is revolutionising digital landscapes, enhancing operational capabilities, and empowering its customers with a path toward edge AI applications.