In partnership with:

Red Hat

STL PARTNERS

EDGE AI: HOW WILL AI IMPACT THE ADOPTION OF EDGE INFRASTRUCTURE?

Webinar: Questions and Answers

Edge AI: How will AI impact the adoption of edge infrastructure?

Questions and Answers

This document outlines the questions and answers received from the STL Partners and STL webinar, ***Edge AI: How will AI impact the adoption of edge infrastructure?****, which was hosted on Thursday 28th September 2023.*

You can watch the recording of the session, and also access the slides, using the link here. *In this document, we seek to address the questions raised in the webinar that we were unable to address in the time available.*

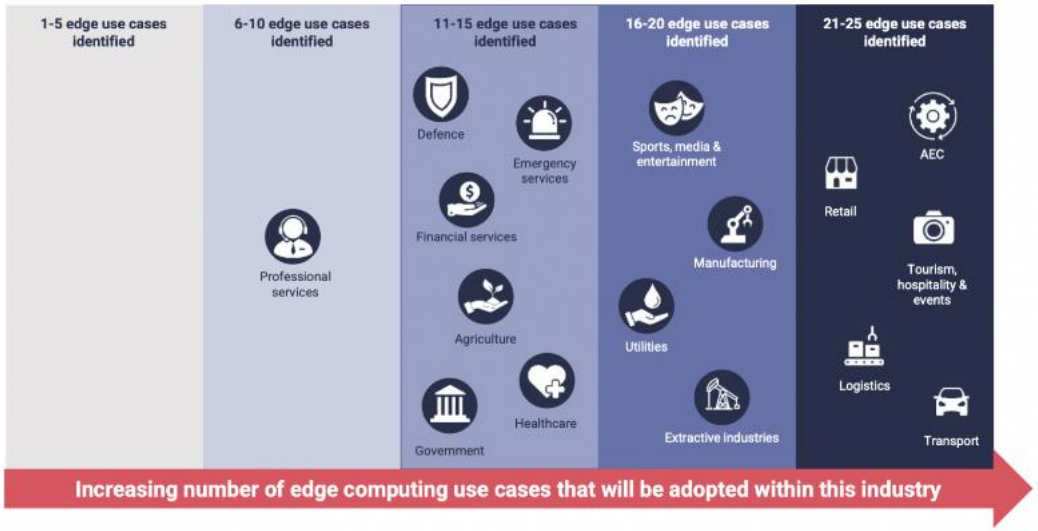
If you have any questions not addressed in the webinar or this Q&A document, or want to hear more about our latest research or from our panellists, please contact:

• *Yesmean Luk, Principal Consultant – yesmean.luk@stlpartners.com*

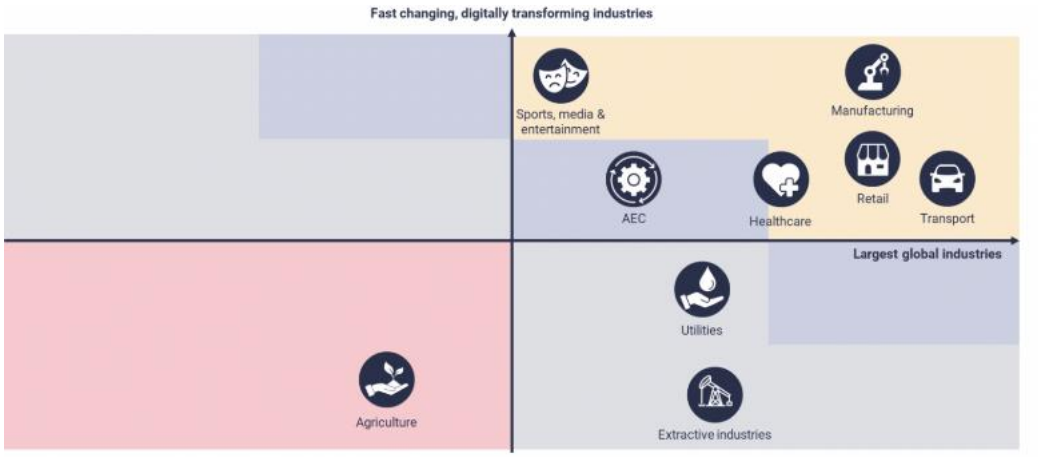
• *Jack Hurley, Consultant – jack.hurley@stlpartners.com*

1. **Are there any industry verticals that are more mature than others in terms of edge/AI adoption?**

Edge computing is poised to make a significant impact across industries. The adoption of edge on some verticals is already being seen, whilst other verticals are lagging. Edge computing is more likely to be adopted by enterprises due to the scope of benefits it offers – increased latency, reliability, and security of edge permitting a proliferation of new use cases.



Source: STL Partners

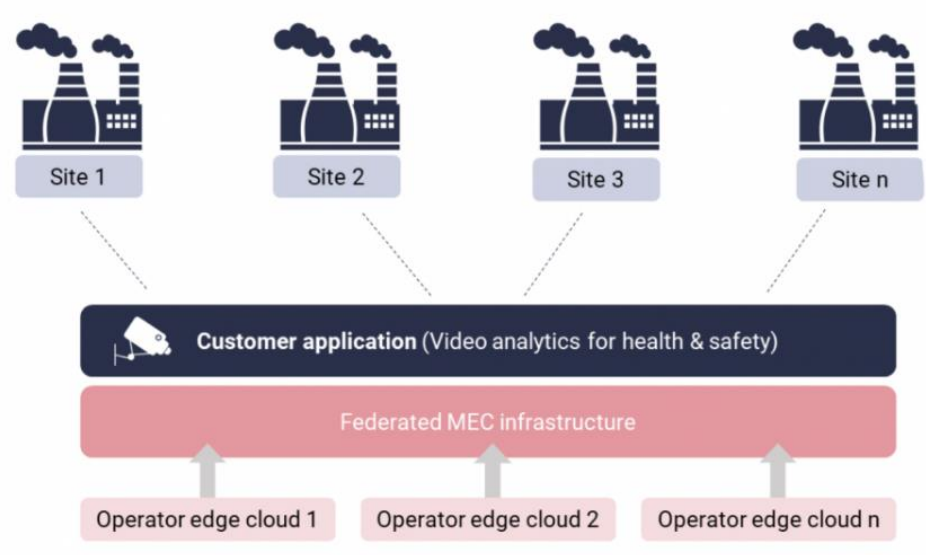


Source: STL Partners

Generally, industries that are more digitally mature are better positioned to adopt edge solutions sooner, such as manufacturing and logistics verticals (see graphics above). Each industry is approaching edge differently, with the extent of advancement dependent on specific use case demand and requirements within the vertical. The involvement of telcos also fluctuates greatly between verticals, as they require the ability to build strong anchor customers who they can test and develop new solutions.

2. **What have you seen in federated learning leveraging edge?**

Federating the telco edge is the process of interconnecting operator edge platforms enabling consistent delivery of edge computing services across networks and national boundaries.



Ideally, an edge application developer will interface with a single platform gaining access to global/regional edge cloud services, easing the access to work with telecoms operators and improve the proposition they can take to customers. The developer should be able to interact with the platform without having to worry about navigating multiple relationships with operators in different countries. Edge federation also provides roaming support and low latency interconnection between the edge clouds of operators in the same area geographically.

It has application in various domains, such as healthcare where medical AI models are trained using patient data from edge devices whilst maintaining data privacy. It also benefits smartphones, IoT, autonomous vehicles, and manufacturing.

For further details, see our article which explains what the telco edge is, as well as its benefits and challenges: https://stlpartners.com/articles/edge-computing/what-is-telco-federated-edge/

3. **What kind of model monitoring is applicable to LLMs?**

When deploying LLMs in a production environment, it is critical to monitor behaviour and performance to ensure that they consistently provide accurate and reliable results.

Applicable model monitoring techniques to LLMs:

Model Metrics Tracking: Track relevant performance metrics such as accuracy, precision, recall, and perplexity, as deviations can signal issue with the model's performance.

Data Drift Monitoring: Data drift can impact model performance requiring re-training or fine-tuning to eliminate the issue. Monitoring ensures consistency over time and reduces the chance of drift occurring.

Bias and Fairness Monitoring: LLMs can produce biased or unfair outputs. Monitoring can mitigate these biases which is pertinent with user-generated content or sensitive topics.

Response Time and Latency Monitoring: These are important in guaranteeing the model provides prompt outputs, monitoring this ensures that user experience remains high which is imperative in real-time applications.

Usage Patterns and Traffic Analysis: This is critical in resource utilisation and capacity planning when handling fluctuations in demand. The volume of requests, usage patterns, and peak times can be identified through monitoring.

Other monitoring methods include: *User feedback and reviews, abuse and toxicity detection, security auditing, model explainability, alerting and notifications, and compliance monitoring*

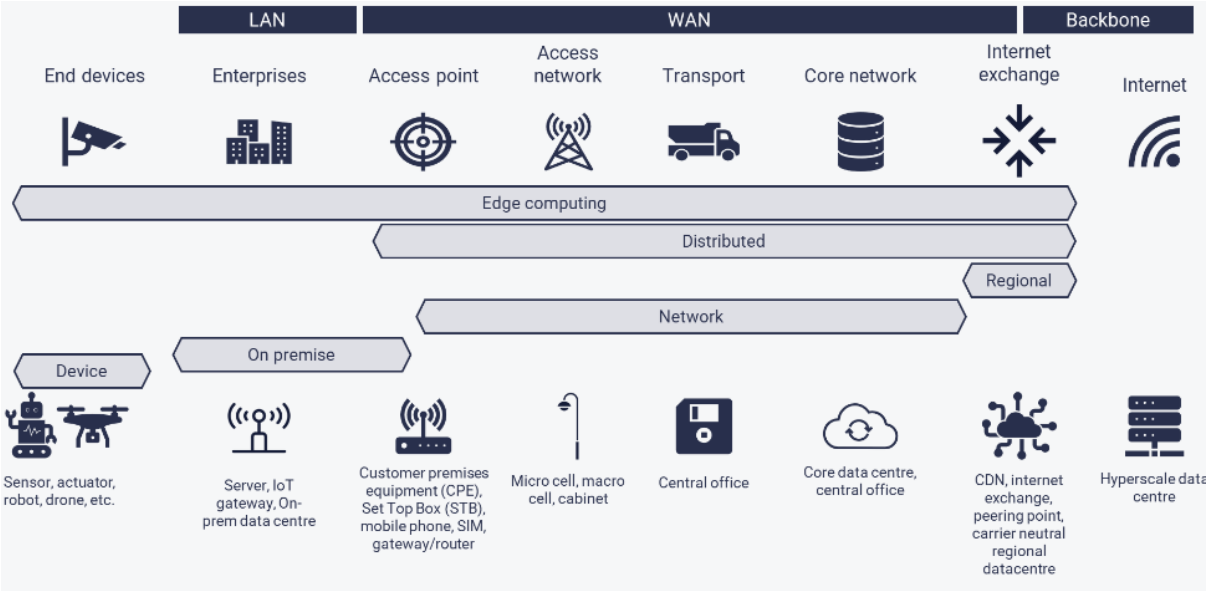
4. **Do you see HCPs moving to the edge to complement their AI offerings in the cloud?**

The hyperscale cloud providers (HCPs), namely Amazon Web Services (AWS), Microsoft Azure, and Google, are well positioned to leverage their AI capabilities at the edge and already do so. For example:

- Microsoft's Azure Stack Edge reference architecture already integrates the ability to extend machine learning inference from the cloud to the on-premise edge.
- Google also has its own purpose-built ASICs (the Cloud Edge TPU and GPU) designed to run AI workloads at the edge using the Google Distributed Cloud (GDC) platform. It also supports use of NVIDIA T4 and A100 GPUs to run edge AI workloads.
- AWS also has capabilities in the edge AI space, such as its various SageMaker capabilities to enable ML models to be run and trained on the edge. This could be used in tandem with its AWS IoT Greengrass solutions for example.

5. **Are telcos continuing to push MEC or is there a model by which they offer a solution in the far edge?**

STL breaks down edge computing into four remits based on unique physical location within a network, as demonstrated in the below diagram:



- **Device edge computing** refers to workloads running directly on physical hardware (e.g., IoT sensors, smart cameras etc.) connected to an edge computing platform, affording minimal latency and reduced backhaul. This type of edge is generally applicable for workloads that are not very compute-intensive or at a site where network connectivity is limited. This is what we would refer to as the *far edge*.
- **On-premises edge computing** refers to computing resources that reside on a customer site. This is vital for customers working with highly sensitive or proprietary data who want all of their data to remain on-premise. Edge here also offers the flexibility and elasticity benefits brought by cloud computing. This could also be referred to as the *far edge*.
- **Network edge (Multi-access edge)** eludes to points of presence (PoPs) owned by telecom operators with edge compute. It is necessary for use cases where there is not a single fixed premise or for consumers who do not want to invest in their own dedicated on-premise infrastructure. This is what STL would define as the *multi-access edge*.
- **Regional edge (far edge)** refers to carrier-neutral, small data centres or internet exchanges that are often located close to tier two or tier three cities. Servers here are rented out to many different customers wanting to run their workloads, which is known as co-location.

Telcos are actively pursuing ‘distributed edge computing’ which STL applies as an umbrella term to define the combination of on-prem and network edge (far edge and MEC) through combinations of partnerships and their own developed (in-house) solutions.

6. **Why is there no one-size fits all solution?**

   There is no one-size-fits-all approach for edge AI as both edge computing and AI separately are highly diverse technologies that rely on context. Edge AI can be applied in numerous verticals, with each of these domains having unique requirements, constraints, and data types that need specialised solutions. The use cases diverge greatly, with some applications demanding real-time processing with stringent privacy and latency requirements, whereas others can tolerate higher latency affording a different AI deployment strategy. Edge devices also come in varied forms, with no single hardware configuration suiting all scenarios. This influences the type of AI models and algorithms that can be deployed. Other reasons for unique approaches include: *environmental considerations, regulatory and compliance requirements, software ecosystems, and customisation requests.*

7. **What can drive decisions to deploy locally vs. centrally with any workload?**

   The decision to deploy locally over centrally for any workload revolves around edge's core benefits. Applications that demand real-time processing often require local edge deployment with a central data/cloud server taking too long to process the data. In scenarios where sensitive data is involved, local processing can enhance data privacy and security with the risk of breaches during transmission to a central server eliminated. Deploying AI models centrally can strain network bandwidth, particularly when dealing with high-resolution video streams or large volumes of sensor data. Local deployment where multiple edge nodes are available also improves redundancy. If one node fails, then another can step in compared to an architecture which relies on centralised servers. Other benefits of deployment locally include: *improved connectivity functionality, reduced cloud costs, resilience, scalability and compliance.*

   The location of deploying edge AI models should be decided by careful assessment of the specific use case, performance requirements, data privacy considerations, connectivity constraints, and cost factors. A hybrid approach combining edge and cloud processing may be the most effective in balancing these various considerations. For edge computing to be readily consumed by enterprises and developers, it needs to be part of a multi-edge cloud solution with application workloads deployed across centralised cloud servers and the edge as needed.

For more information about our work in edge, visit our hub:

https://stlpartners.com/edge-computing/

...or get in touch to learn more:

- Yesmean Luk, STL Partners yesmean.luk@stlpartners.com
- Jack Hurley, STL Partners jack.hurley@stlpartners.com

STL PARTNERS

Research

Consulting

Events

www.stlpartners.com