Edge Insights Service

Kindly supported by:

Red Hat

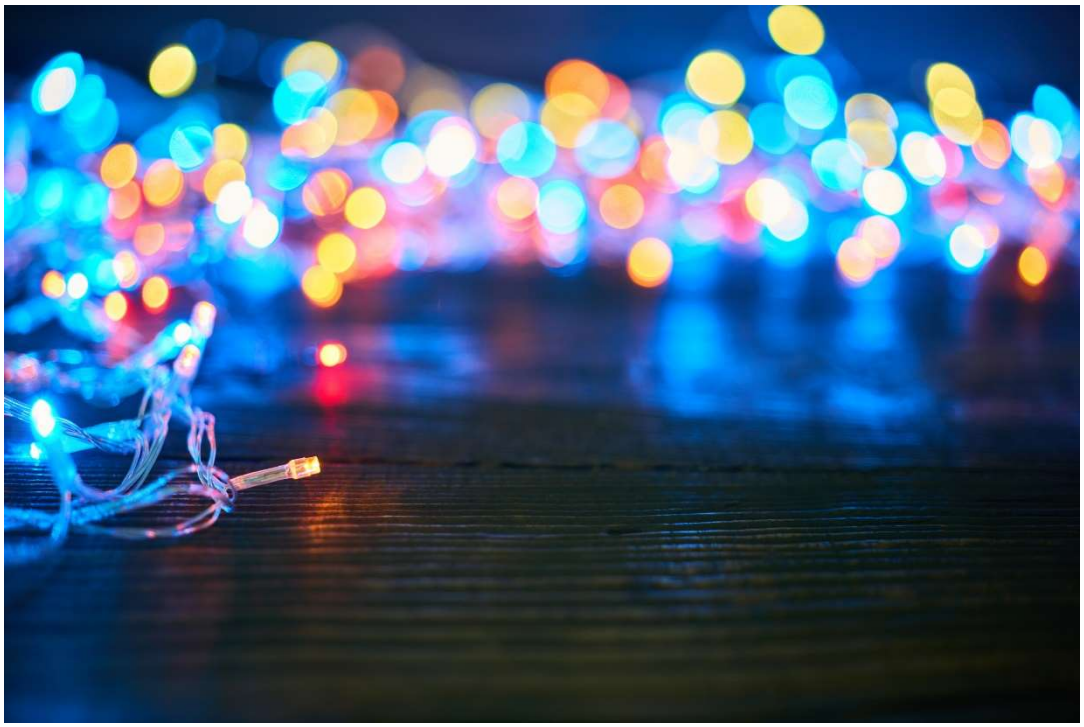
intel.

STL PARTNERS

Executive Briefing

EDGE AI: HOW AI IS SPARKING THE ADOPTION OF EDGE COMPUTING

AI applications will require low-latency, local compute for rapid inferencing and large scale data collection, triage, and engineering. Edge compute will therefore play a key role in AI app delivery. However it's not just about infrastructure - commercial scale for edge AI will depend on effective ecosystem collaboration models.



Tim Otto, Senior Consultant | tim.otto@stlpartners.com | October 2023

Executive Summary

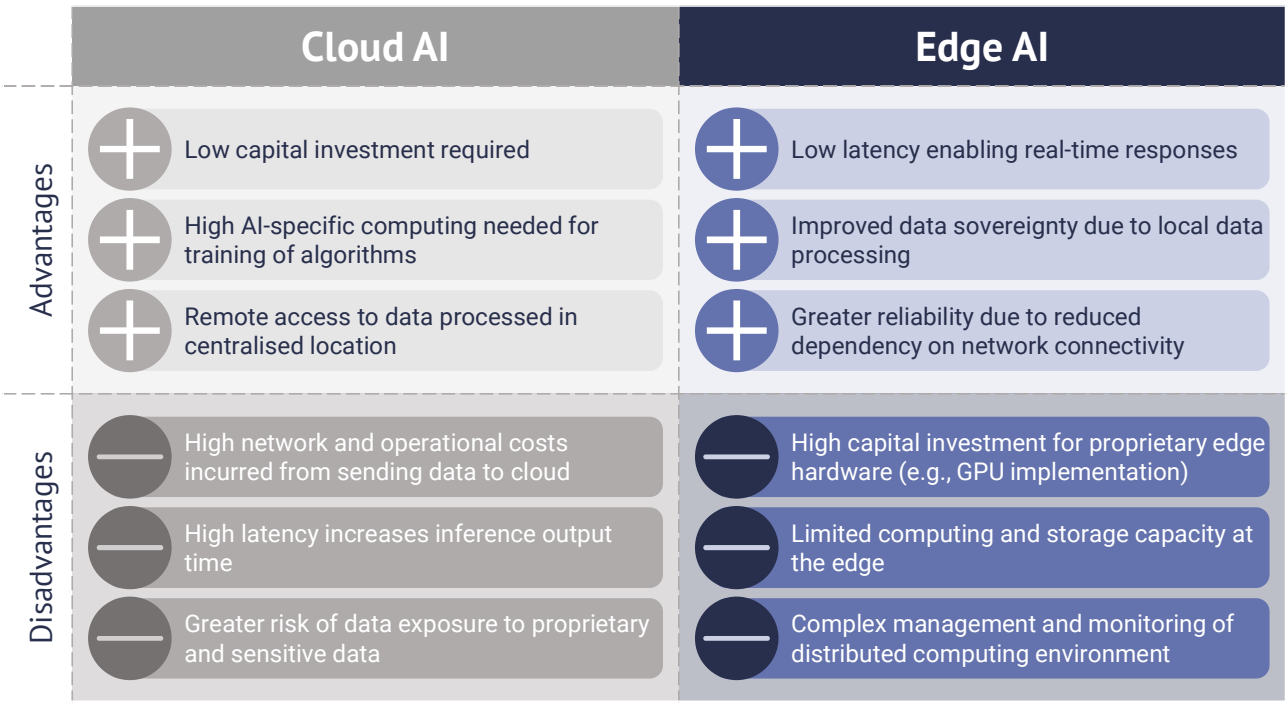
The recent explosion in development and adoption of artificial intelligence (AI) has stemmed from advancements in processing power and the availability of investment capital, both of which have enabled the development of cutting edge, powerful models. As enterprises across the world scramble to assess the value of these AI models to their businesses, developers and infrastructure providers are recognising the need for compute infrastructure that spans from the cloud to the edge.

The role of edge AI in driving scale for AI-driven applications

Leveraging edge computing infrastructure for the inference of AI applications promises significant increases in speed and reliability, delivering low-latency for mission-critical applications. Applications such as video analytics for fault detection, production line management, hazard detection, etc. can deliver a much more valuable service when utilising this low-latency infrastructure.

On top of this, the vast amounts of data that must be collected and engineered to maintain and build these AI models incurs a huge cost when operating solely on the cloud. Advancements in chip efficiency and power allows smaller footprint data centres at the edge to do much of this computation, essential for the ROI of these applications. A distributed infrastructure will make use of both cloud and edge data centres. See Figure 1.

Figure 1: Cloud and edge will work in tandem to form AI infrastructure



Source: STL Partners

Key players in the space are aware of these demands and moving quickly. Chip makers like Intel and ARM are designing AI-specific CPUs which can operate at the edge, investors are moving money into

edge data centre companies and orchestration platforms, enterprises are investigating how they can leverage AI applications across their organisations and building the necessary compute infrastructure to run them at scale. As this evolves, close collaboration across the whole ecosystem is necessary to ensure the maximal scale is achieved.

This report provides an overview of **edge AI** – the use of edge computing infrastructure for development and deployment of AI – and how the two technologies must work in tandem to drive scale for the applications they enable.

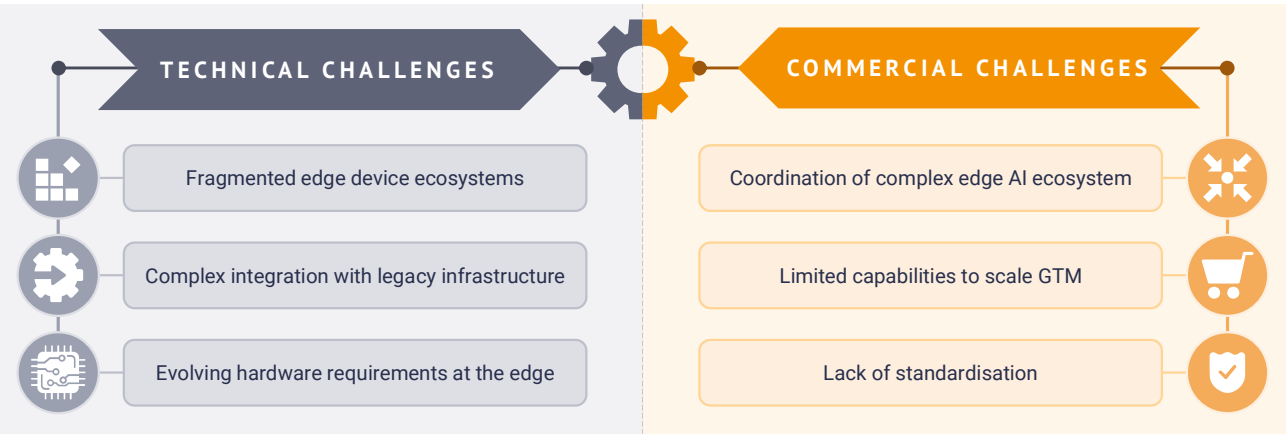
Edge will prove integral to the development of AI applications for the following uses:

1. **Real-time inferencing:** Instantiating outputs from the trained model to deliver recommendations or automated responses across an environment.
2. **AI operations (AIOps)**: Leveraging AI to manage workloads across complex environments, often utilising both edge and cloud compute.
3. **Model fine-tuning:** Adjusting the framework through which the trained model operates for the optimal output response.
4. **Data engineering:** Filtering, structuring and parsing large amounts of data from sensors to create metadata to store or process in the cloud.

Effective ecosystem management will underpin the success of AI at the edge

Building and integrating the right infrastructure is only one side of the coin… Given the complexity and criticality of these AI applications, involving many parties across different technology areas, building a successful ecosystem is integral to driving scale. When the speed of innovation is added to this equation, finding the balance between speed, reliability and security becomes a significant challenge. See Figure 2.

Figure 2: Significant technical and commercial challenges must be addressed before the edge AI ecosystem can achieve meaningful scale



Source: STL Partners

An ecosystem approach addresses some of the technical and commercial challenges facing AI applications.

1. **Technical challenges:**

    1.1. **Device ecosystem:** as the range of edge devices grows, each distinct in architecture and capacity, this complicates the deployment of AI models that can interoperate with the devices in situ.

    1.2. **Existing infrastructure**: the task of integrating with existing infrastructure is significant for businesses and often necessitates resource-intensive system modifications.

    1.3. **Hardware requirements**: new chipsets will be required for AI workloads. More powerful and resource intensive GPUs will be required for the heavier workloads like computer vision driven video analytics, whilst CPUs optimised for AI workloads will be sufficient for lighter-weight applications and models.

2. **Commercial challenges:**

    2.1. **Ecosystem coordination**: The challenge of ecosystem coordination is amplified by the need for harmonising the interests and operations of various stakeholders like hardware producers, software developers and system integrators in the edge AI ecosystem.

    2.2. **Scaling GTM**: scaling GTM models is a hurdle for many start-ups – the central engine of AI innovation – as they are often focused on technology development and customer testing, requiring partnerships or ecosystem leverage to reach a wider client base and achieve commercial scale.

    2.3. **Standards and governance**: the dichotomy of standardisation vs. innovation presents a challenge where the rapid evolution of AI, fostering innovation, clashes with the need for universal standards ensuring interoperability among disparate AI tools and systems. An effective ecosystem can support stakeholders in the definition of deployment standards and by acting as a distribution channel (especially when the orchestrator is a much larger organisation than the suppliers it supports).

Table of Contents

Table of Figures

Introduction

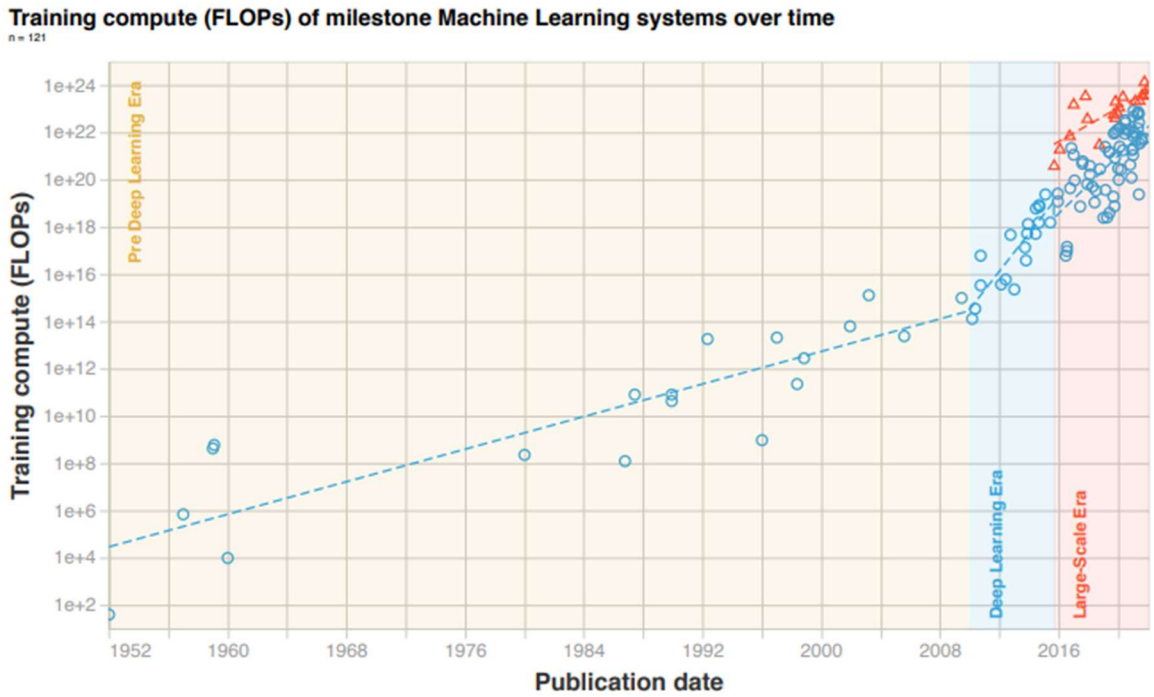
Enterprises across all industries are investigating how they can leverage AI applications

Since the release of GPT-3 from OpenAI, consumer interest in AI has exploded. This change in consumer behaviour follows that of enterprises, who have been investigating AI and its application to their processes for years.

AI has been around for decades. The recent revolution in AI capabilities stems from a combination of innovation in hardware technology and the deep learning era; beginning in 2010, a subset of machine learning used multi-layered neural networks to analyse various forms of data. Since 2012, companies creating these deep learning models have seen an increase in their financing, allowing them to build large data centres with the specific hardware (GPUs) to process larger datasets.

Figure 3 shows this increase in compute power usage by machine learning systems over time.

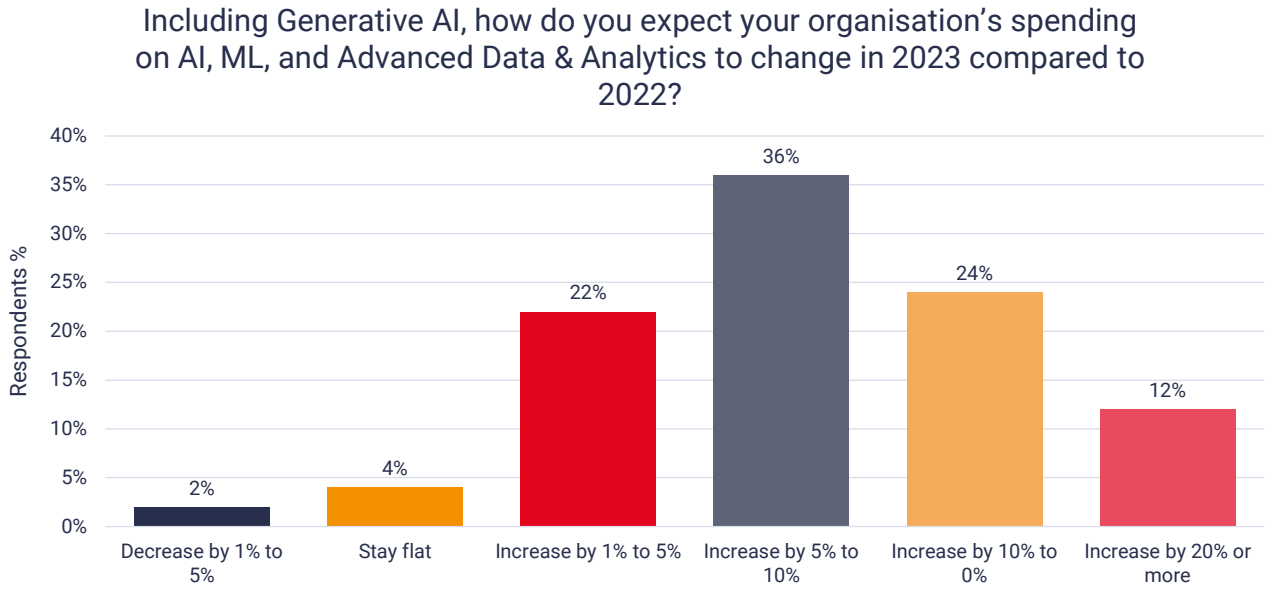
Figure 3: Trends in training of machine learning and AI models



Source: "Compute trends across three eras of machine learning", 2022, J. Sevilla et al. https://arxiv.org/pdf/2202.05924.pdf

Large Language Models (LLMs) like Chat GPT are the most widely known AI models. Training these models on billions of tokens sourced from the public internet (GPT-3: 175 billion parameters, LaMDA [Google]: 137 billion parameters) allows these LLMs to creatively solve a wide variety of problems from equally varied inputs. This ability to ingest diverse inputs and produce outputs which can be utilised in a wide range of environments, is the key differential that has sparked the wave of recent interest.

Figure 4: Horizontal adoption of AI models has exploded in the past 18 months



Source: Wikibon

Enterprises will access centralised AI models, developed and trained in the cloud

AI is not consigned to just language processing. Although LLMs have been the main driver of interest since 2022, AI models such as convolutional neural networks (CNNs) and graph neural networks (GNNs) can ingest different data structures (image/video and graphical, respectively) to provide valuable insights.

Most enterprises are still exploring the possibilities of AI. Initial implementations are centred around simple tasks, such as administration or customer service, but executives are exerting pressure on their organisations to explore how AI can automatically manage mission-critical tasks and augment productivity. To do this effectively and reliably, enterprises will need AI models that are well designed and well trained to carry out the mission-critical tasks that are required of them.

Few enterprises will have the resources to create and train their own AI model for a specific application within a specific industry. This would require a huge amount of compute power, which is both expensive and scarce, as well as a large amount of well documented and parsed data. Bloomberg has created its own LLM, BloombergGPT, pooling its extensive archive of financial data (363 billion tokens) with a public dataset (345 billion tokens) to train the model. Companies like OpenAI and Inflection.AI continue to raise huge sums of capital investment to build their own. However, few enterprises will have the resources to follow suit.

> A lot of large, public companies are now talking about AI to appease investors and shareholders. Most are in research mode, trying to understand and evaluate the value of the technology for their business case. AI is not a "one-size-fits-all" scenario – every company needs to do their own evaluation.
>
> *VP of Global Sales, Edge AI platform provider*

Most will leverage models trained on large, centralised datasets (like GPT) that are built in the cloud, fine tuning them through prompts and contextualisation to create outputs which are more tailored for their environment. This approach is considered less secure than building your own, as it involves the sharing of proprietary data with the model, which will then use these inputs to further train its algorithm. However, this approach allows enterprises to benefit from the generalist applications of these models whilst not having to invest time and money in developing and training the model themselves.

> We are seeing a greater mandate from enterprise to invest in edge AI. Clients are becoming more educated about AI's capabilities and are much more comfortable trusting processes and decisions with AI models which they do not fully understand. Although ChatGPT and LLMs don't impact every industry, their adoption from consumers has greatly increased the trust from enterprise.
>
> *CTO, Video analytics solution provider*

Edge computing enables three elements for the application of AI models

Cloud provides the required level of compute and data volumes for deep training of AI models, and subsequently enables enterprises to quickly access centralised AI models and tools. It has drawbacks when it comes to the active implementation and fine tuning of AI applications.

On top of the potential security/privacy risks with cloud models, as mentioned above, the cloud is limited by its ability to provide:

- **Ultra-low latency infrastructure:** Due to its physical location further away from the end user, cloud cannot provide the performance requirements for running low latency, mission-critical AI applications.
- **Guaranteed reliability of connectivity:** Given the reliance on connectivity infrastructure to send data back, the cloud cannot achieve the required five nines (99.999%) availability of mission-critical infrastructure.
- **Cost-effective backhaul options:** As well as possible issues with the reliability and latency of backhauling the necessary data to run and refine AI models in the cloud, enterprises would incur significant costs for transferring (and storing) all the data on the cloud without parsing it.

Edge computing promises to address these challenges by enabling four elements of AI implementation (see Figure 5):

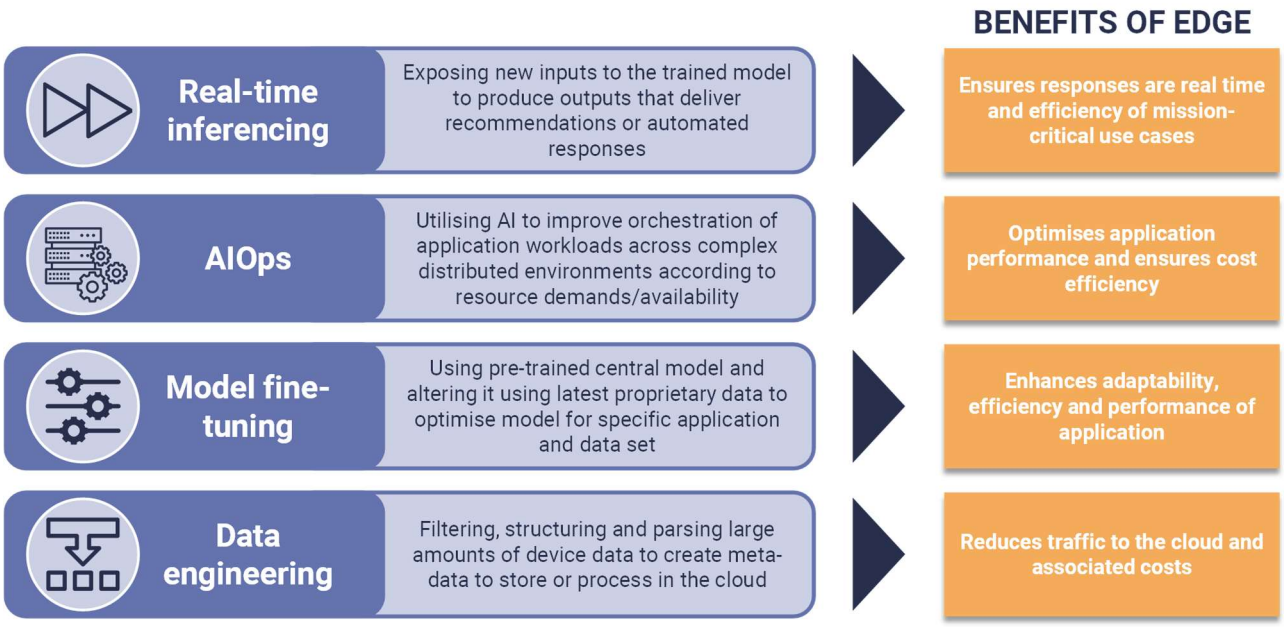
1. **Real-time inferencing:** Instantiating outputs from the trained model to deliver recommendations or automated responses across an environment.

2. **AI operations (AIOps)**: Leveraging AI to manage workloads across complex environments, often utilising both edge and cloud compute.

3. **Model fine-tuning:** Adjusting the framework through which the trained model operates for the optimal output response.

4. **Data engineering:** Filtering, structuring and parsing large amounts of data from sensors to create metadata to store or process in the cloud.

Figure 5: Edge provides low latency, greater reliability and cost efficiency for AI applications



Source: STL Partners

Real-time inferencing

Edge computing becomes important for AI inferencing: instantiating the outputs from the model to trigger real-time actions. Once trained, the model itself is surprisingly small and can be run on minimal compute. Inferencing at the edge will provide the enterprise with the benefits of both worlds, allowing them to run mission-critical applications through the AI effectively.

> AI is driving the adoption of edge infrastructure because of the need for real-time inferencing of LLMs.
>
> *Head of Revenue and GTM Lead, Edge security provider*

Data engineering

Additionally, edge computing can provide the first-pass computation required to filter and compress data *collected* from IoT devices before packets are sent back into the cloud to train the central model further. By filtering the swaths of data collected by end devices, ensuring only valuable metadata is shared back with the model, enterprises save drastically on connectivity costs by lowering the total traffic going to the cloud. This is especially important with the move towards four- and even eight-k video feeds – the strain on the network and resulting cost of backhaul for sending all data back to the cloud without triage at the edge is too large a barrier.

> AI can provide sharp contextualising information on the relevance of data captured by IoT sensors. The AI we are seeing today is using another level of visibility, but this requires inferencing at the edge in order to get the most value.
>
> *CEO, Edge IoT platform provider*

Model fine-tuning

Fine-tuning involves taking the pre-trained model and training it on a smaller, much more specific dataset. This allows the model to adapt to a very particular task and offer specific outputs. This is particularly powerful in enterprise scenarios where sites and processes may have unique characteristics and requirements from the model. Fine-tuning allows an organisation to manipulate the larger model for their individual needs.

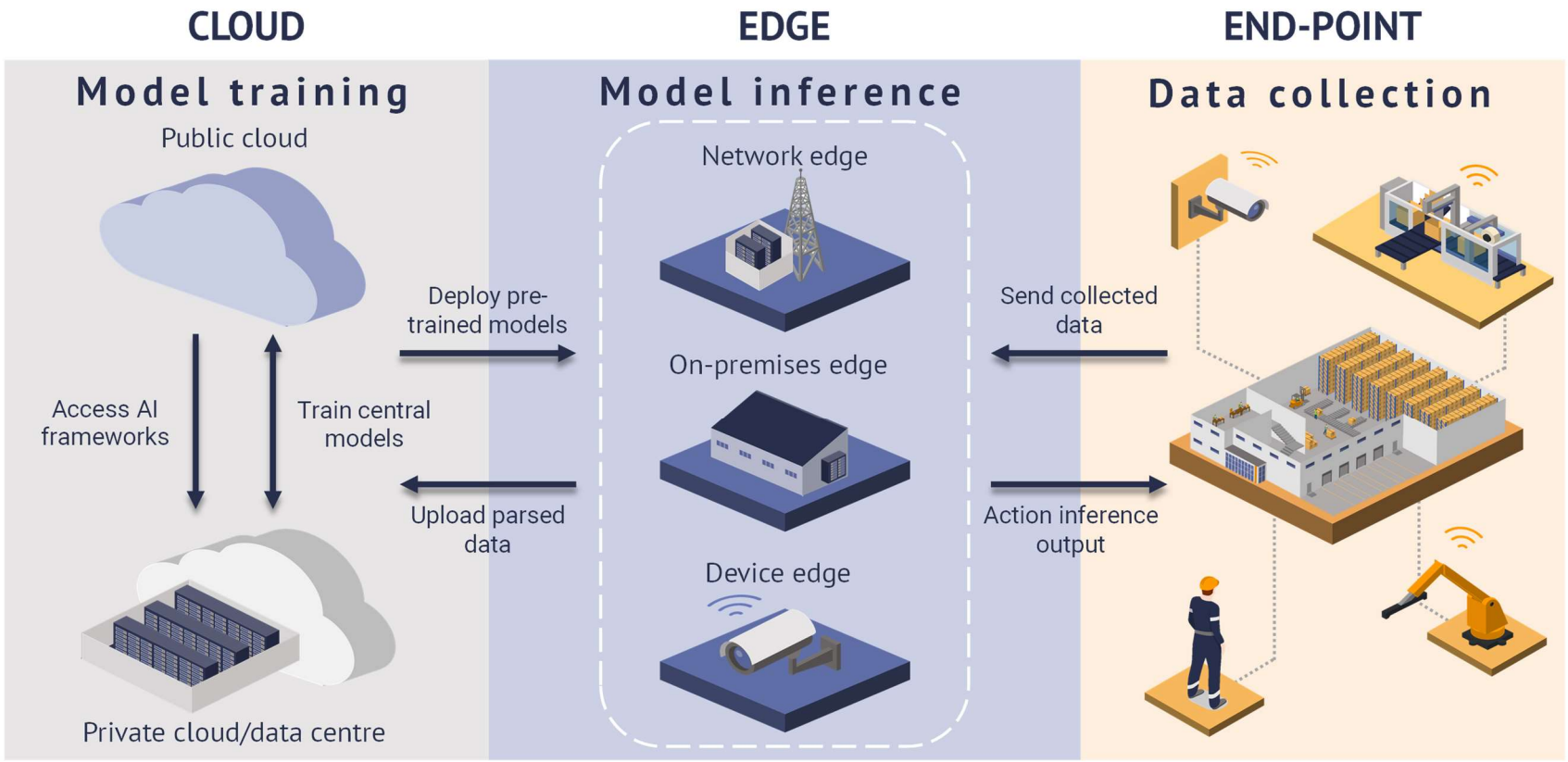
This kind of training could take place at the edge, given the smaller data management and compute requirements. As processing units evolve, AI models can increasingly be hosted and fine-tuned on smaller form factors, including tailored CPUs that are much more cost effective than GPUs at the edge. This advancements ensure edge AI remains a financially viable without losing performance.

Edge and cloud must come hand in hand for the deployment of AI

To address the specific capabilities mentioned above, edge infrastructure is valuable for AI use cases – giving rise to "edge AI". However, edge AI is not a distinct type of AI, but rather one piece of the infrastructure puzzle. It requires an edge-cloud management platform to allow the real-time inferencing to interact with the centralised training which will happen at large, specialised data centres.

Given the higher latency and cost (from a financial and energy perspective) of hosting LLM models in the cloud end-to-end, edge-cloud architecture offers a ready solution which is both cost-effective and fast. However, while enterprises will use the edge to exploit its speed and privacy (which will become increasingly important to enterprise customers), the complex training and hosting of AI models would remain in a central data centre. See Figure 6.

Figure 6: AI will leverage an edge-cloud infrastructure for end-to-end training and inference

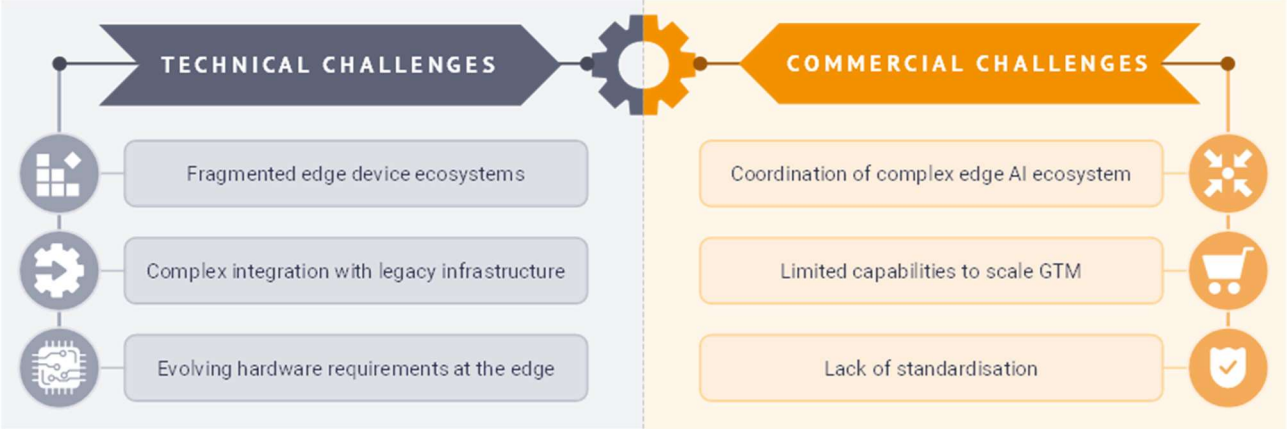


Source: STL Partners

Interest in AI is driving demand for edge but obstacles remain

Edge has the potential to play an important role in the scaled utilisation of AI models. To reach this scale however, there are several challenges, both **technical** and **commercial**, which must be addressed. See Figure 7.

Figure 7: Technical and commercial challenges to scale AI at the edge



Source: STL Partners

Technical challenges

1. **Fragmented device ecosystems:** With industries constantly digitising, the range of edge devices expands – each characterised by distinct architectures and operating capacities. Unlike the relative uniformity of centralised systems, edge environments are diverse by nature, spanning from simple IoT sensors to more sophisticated edge gateways. This heterogeneity is further complicated when introducing AI models, which are not universally adaptable. A model fine-tuned for one device might underperform or even be incompatible with another due to variances in processing power, memory or storage requirements. This specificity in design and deployment demands meticulous planning and adaptable strategies.

2. **Integrating with existing infrastructure:** Transitioning into this environment, businesses face the monumental task of integrating modern AI solutions with existing legacy systems. While some of these systems might have the computational muscle, they may lack the adaptive frameworks or the compatible tools essential for integrating contemporary AI applications. Bridging this gap becomes a resource-intensive venture, often necessitating significant system modifications or complex workarounds.

   Furthermore, edge AI requires the orchestration and distribution of AI/data workloads across an ecosystem of decentralised compute – from the enterprise premises to the cloud. Ensuring this continuum is secure, can operate in real time, and can run on the appropriate cloud architectures is a key challenge for enterprises and developers alike.

3. **New hardware at the edge:** Although inference workloads are relatively light, with the promise of fine-tuning and greater traffic at the edge on the horizon, it seems likely that edge deployments will start to need GPU chipsets. The ability to process data in parallel allows these nodes to operate much more data at greater speed, ensuring the benefits of edge are maintained when workloads become more intense. This change in chipset brings with it changes in power, cooling, design and a whole host of other factors, creating a significant challenge for edge infrastructure providers.

Commercial challenges

1. **Ecosystem coordination:** Further amplifying the challenge is the essential coordination among the myriad stakeholders in the edge AI ecosystem. This intricate dance involves hardware producers, software developers, application vendors and system integrators (SIs). Each of these entities, often operating within their distinct spheres driven by individualised goals and timelines, must find common ground. Synchronising these divergent forces into a singular, cohesive operation is no small feat. Beyond mere alignment of objectives, establishing and maintaining robust communication channels becomes paramount, given that even minor miscommunications can lead to significant inefficiencies or performance issues.

2. **Scaling GTM models:** Given the current state of AI, there are many start-ups innovating and proving the value of their solution. Their resources are focused on technology development and customer testing, making GTM challenging for these organisations. In many cases, they will be unable to reach a wide client base on their own, leveraging partners or ecosystems to encourage scale. Without this kind of approach, much of today's innovation will be unable to reach commercial scale.

3. **Standardisation vs. innovation:** Lastly, the whirlwind pace of AI's evolution inadvertently challenges the very essence of standardisation and interoperability. The dynamism of AI's growth, while fostering innovation, often runs counter to the need for universal standards – standards that ensure disparate AI tools and systems can seamlessly interoperate (and are correctly regulated). The absence of such standards risks creating siloed solutions, where certain AI tools or platforms are incompatible with others, making integrations and expansions labour-intensive endeavours.

Enterprises are concerned about the security of AI and protecting their proprietary data

Even outside of edge AI, there are concerns around the security of AI models which threaten to slow the adoption of these applications. When sharing data for training or fine-tuning AI models, there is a tangible risk that proprietary or sensitive information might be inadvertently integrated into the model's large knowledge repository. Such integration, even if unintentional, could allow the AI to divulge enterprise-specific insights in future interactions, jeopardising confidentiality and competitive advantage.

> "Security of data is a big issue for AI. This is slowly being dealt with, as the ecosystem evolves and standards are created."
>
> *Global Head of Business Development, Cloud-edge platform provider*

Particularly for industrial processes which extend beyond customer service or content creation, this kind of data breach represents a serious risk to enterprise customers who see this data as their differentiating factor.

Additionally, the act of transmitting data to external servers for AI processing amplifies security vulnerabilities. This not only introduces potential interception or breach points but also raises issues of data sovereignty, especially when the data is processed across different jurisdictions with varying protection regulations. This external handling can lead to unintended compliance breaches, carrying both legal and reputational implications for the enterprise.

Despite this, the perceived benefits that AI can bring an organisation are likely to outweigh these concerns, as evidenced by the rapid investment that it has seen across industries.

An ecosystem approach to edge AI

To capture the edge AI opportunity, it is clear that the ecosystem must collaborate to address both the technical and commercial barriers to scale. Working with partners through the development and deployment of applications allows providers to create a solution that is easily flexible and secure, able to provide the real ROI promised by AI, whilst maintaining security and compliance.

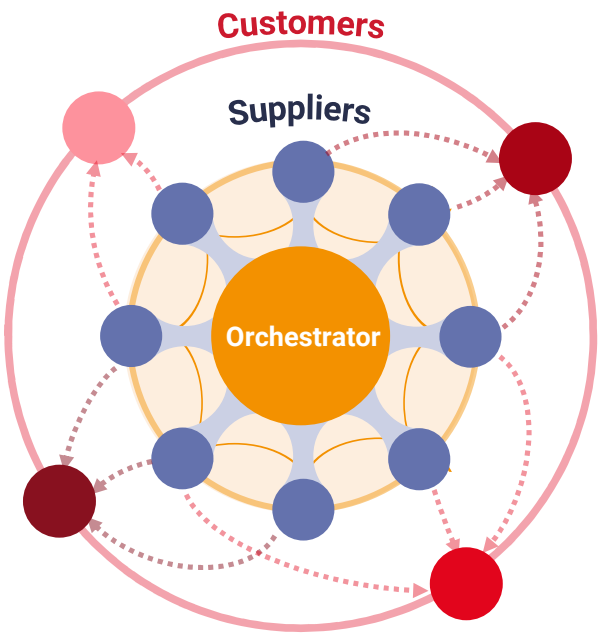
What is an ecosystem?

Ecosystem business frameworks present a distinct approach to organising economic activity, contrasting with traditional models like vertical integration or linear value chains. By correctly leveraging these frameworks, companies can achieve enhanced responsiveness, innovation and scalability.

> "Partnerships are absolutely critical to the success of our solution. We provide the software, platform, and technical expertise, but we want to be more aligned with the chip and hardware and network side. This would add significant value to our customers"
>
> *CEO, Edge IoT platform provider*

A successful ecosystem necessitates drawing in external stakeholders into a well-coordinated business setting. In this environment, the orchestrating entity offers robust assets and expertise, aiding participants in delivering value to the end consumers. The design of these ecosystems should prioritise stakeholder engagement and cater to their motivations to avoid supply-side lapses. Concurrently, the onus is on the orchestrator to captivate the end consumers, ensuring their sustained involvement. See Figure 8.

Figure 8: Ecosystem business models unlock value for all stakeholders involved



Orchestrators

- Define pre-integration process
- Support suppliers with technical expertise (e.g. integration and certification) and commercial expertise (e.g. GTM support)
- Bring new customers and suppliers to the ecosystem
- Leverage geographical reach/trusted brand to attract customers

Suppliers

- Bring solution and solution specific expertise to ecosystem
- Collaborate with other suppliers across the value chain
- Produce integrated solution for end-customer

Customers

- End-user of solutions produced from ecosystem
- May be existing customer from orchestrator or from suppliers from ecosystem

Source: STL Partners

For maintaining the engagement of both participants and consumers, the ecosystem needs to prioritise an excellent user experience and consistent innovation. This requires a comprehensive grasp of the end user's behaviours and aspirations. The ecosystem must also exhibit adaptability, allowing shifts in strategy to account for unexpected outcomes or evolving market dynamics.

> To work successfully, we need a developed ecosystem where there is specific funding from a central party, driving vertical expertise, technical support, and integration across different parties
>
> *Business Analyst, Drone solution provider*

What are the benefits of ecosystem approaches?

Ecosystem frameworks are especially beneficial when they enable rapid responsiveness and innovation. For instance, frameworks like Android and the Telecom Infra Project (TIP) emphasise stakeholder engagement, even incorporating this into governance for fostering the required collaboration levels. Ecosystems become especially advantageous when:

- Customer requirements are mutable or evolving.
- A wide range of end-user preferences exists.
- Rapid market entry is essential.
- Revenue generation paths are ambiguous.
- Scalability is essential due to dominant market player dynamics.
- Supporting B2B2X business models becomes necessary.
- There is potential for reutilising resources to offer value to other businesses.

Why are ecosystem approaches necessary for scaling edge AI?

An effective ecosystem allows for a diverse range of use preferences, rapid innovation and protection against dominant single players.

1. **A wide range of end-user preferences:** Given the horizontal nature of the models, large language model AI models offer the ability for the end user to tailor the outputs of the models to their particular needs. This user specificity will be extremely powerful but will require a strong and flexible ecosystem to ensure the underlying infrastructure and deployment cycles can still be efficient enough to achieve commercial scale.

   Our research has found that software developers and infrastructure players felt edge would be an important technology to enable the unique fine-tuning of larger models. Given the greater control of data transmission and the ability to manage the infrastructure (if not on-premises, then in a regional or network edge data centre), edge infrastructure will allow enterprises to create their own unique framework on top of the larger AI models.

   The open-source AI ecosystem is particularly diverse, with a vast array of developers collaborating to iteratively develop new models. This mode of development is increasingly seen as one of the most potent avenues to innovation, allowing developers to work rapidly and ensuring innovation is a top priority. Open-source projects allow enterprises to adopt and adapt these models for their own use cases, providing them with the technology whilst avoiding a large majority of the development process.

2. **Rapid market entry is essential:** The ability to rapidly innovate and go-to-market is essential especially at the early stages of the AI trend. The current pace of development is extraordinary, with new applications and organisations starting operations every week. For these solutions to be able to achieve a global reach, they will need to leverage ecosystems managed by larger players which provide them with greater exposure and the GTM tools which smaller organisations do not have the resource to build themselves.

   Being able to update applications on a universal infrastructure platform is essential for AI applications. Given the pace of new companies and the evolving offerings of incumbent players, platforms are already being built to ensure that the new applications can be added, and existing solutions can be updated, without causing severe disruption to the whole platform.

3. **Scalability is essential due to dominant market player dynamics:** Although the AI market is dominated by a handful of private companies, many believe that the industry will evolve and become much more diverse. Given the demand for AI applications across many industries, each with their own specifications, it seems likely that a “long-tail” of AI models will begin to appear, each with their own specifications and application ecosystems. Ensuring that these models can scale to serve the needs of their specific markets will be integral to their ability to overthrow the current incumbents with their plentiful resources.

Independent software vendor case studies: How ecosystems help address the technical and commercial barriers to scale

Ecosystems of the kind we have described often leverage the scale of a larger organisation. These organisations become the orchestrator of the ecosystem. This company provides the resources and reach to allow start-ups and smaller businesses to serve customers they otherwise would have struggled to reach, providing integration standards and support to ensure that these applications can easily be deployed at sites already working with the larger company. Given the size of the orchestrator customer-base, this significantly decreases the time-to-market for application providers with pre-integrated solutions and strengthens their proposition.

We spoke to several ISVs working within the Red Hat and Intel edge ecosystem who have successfully addressed several barriers to scale by leveraging ecosystem models.

Addressing technical challenges

Due to its reliance on distributed infrastructure, without the power, cooling and operational support of centralised data centres, edge computing requires greater hardware optimisation. Players like Red Hat and Intel have invested in ecosystem co-innovation to try and address this. Through regional vertical-specific lab environments, customers and partners can bring their product into the lab facility and begin testing/integrating seamlessly.

Within these lab environments, established edge players are then able to support the technical integration of new, innovative products from emerging application independent software vendors (ISVs) onto the platform, tuning both sides of the equation to ensure optimal performance through benchmarking of the core performance metrics.

Essential to the success of this model is the ability to offer significant, personalised support to ISV partners. For example, offer free commercial licenses to ensure applications can be fully integrated onto the platform without significant cost to the application vendor. This ensures that ecosystem partners can rapidly adapt their applications to operate on the platform, shortening the time-to-market and increasing the overall ROI for the partner.

We spoke with Ipsotek, a video analytics application provider that is working closely with Red Hat on the deployment of its application. The multi-cloud, open-source platform allows Ipsotek to rapidly deploy their solution to a wide range of enterprise customers, also benefiting from Red Hat's increased GTM reach and support.

One of the major technical challenges that Ipsotek faced was the move to a containerised architecture, rebuilding its whole application which previously operated on a virtual machine infrastructure. Through the ecosystem approach, it was able to get access to the expertise, resources and tools of an established and leading player in the space, as well as 24/7 on-call support in dedicated regions. On top of providing Ipsotek with free commercial licenses, this ensured an accelerated integration process over 18 months. Ipsotek is focussed on deploying this containerised platform with real commercial customers.

To mitigate the bifurcation of standards within verticals, Red Hat and Intel have also built "strategic advisory boards" that create a forum for collaboration and engineering. These boards offer blueprints and standards for working that others can utilise, whilst also allowing participants early-access to technical development roadmaps. Whilst not a strict consortium prescribing industry standards, these advisory boards do limit the fragmentation of industries as they are developing their edge environments.

Go-to-market support

In addition to the technical support that ecosystem models can provide ISVs, being able to tap into the resource of established open players can provide support in actually getting solutions to customers. Through ecosystem models, once a partner is certified on the platform, established and credible players lead sales activity with enterprise clients, pushing these applications to a wide range of customers. By activating their sales machine, ecosystem partners benefit and can reach a scale otherwise impossible for smaller organisations without extensive brand leverage or sales resource.

> "Communicating the business case to the customer – proving the value of edge and the AI applications on top – is the most significant obstacle."
>
> *Global Head of Business Development, Cloud-edge platform provider*

We spoke to Guise AI, a small organisation specialising in device edge AI. This company focuses on small compute technologies, specifically for devices on the edge, such as those with just 8 megabytes of capacity and tools like the Raspberry Pi, extracting valuable data in these resource-constrained environments with AI and analytics. Its customers are often operating infrastructure that has inconsistent power and broadband access and are often situated in places where large machinery and infrastructure cannot be accommodated.

Guise AI has been working with Red Hat and Intel since 2021 as part of streamlining its partnership process, limiting the complexity that comes with working in an edge environment. With Red Hat providing the operating system infrastructure on top of Intel hardware, it is moving towards a hybrid cloud approach, focusing on scaling horizontally across devices on the edge instead of vertically in the cloud.

Data generated at the device edge has been available for decades. The proliferation of IoT devices has gradually accelerated over the past 20 years, allowing enterprises to collect data and optimise processes given the right internal expertise. With the application of AI to harness and process this data in real time, without sending it to a central cloud, enterprises are suddenly able to save on bandwidth and associated costs whilst dramatically increasing the efficiency of their sites and processes. This requires edge AI and managed infrastructure, coordinating the varied hardware and software vendors present with an edge environment. To support this, Guise AI has created an "edge ops platform" through the partnership with Red Hat and Intel to manage this complexity.

With support on the silicon side, it is shifting from the need for high-capacity GPUs to smaller compute units, which are more cost efficient. This way, instead of relying on expensive GPUs, companies can use more affordable Intel hardware. Furthermore, working as part of the ecosystem has enabled Guise AI to add security to its platform, which is crucial for industries like oil and gas, where data security is paramount, as well as to facilitate the deployment and management of AI models across various

devices. Initially focused more on AI and inferencing, the partnership now delves into deployment, scalability, and management.

We heard a similar story from Aotu, an edge operating system that provides the consistent platform for computer vision applications. The organisation has a broad collaboration with Intel, with Intel teams providing engineering guidance on selecting chipsets and GPU optimisation. The latter also provide commercial support, leveraging its brand reputation to attract more customers and use cases from larger clients with greater support requirements. It works exclusively with large-scale enterprises, ensuring it only deploys the best hardware and caters to the best customers.

Aotu collaborates not just at a strategic level but also with Intel's medium-level managers to ensure every project's success. This alignment across the organisation facilitates the rapid innovation across the partnership, ensuring joint projects are as efficient and cost effective as possible. The partnership with Intel decreases the risk exposure of Aotu through the stringent assurance of agreements and solid certification cycles. The alignment of strategic decisions with medium-level management is seen as an optimal approach.

Conclusion

AI application developers continue to develop new and exciting applications with a wide range of uses. However, they will put an ever-greater strain on the network infrastructure over which they run. The growing complexity of our digital interactions is already creating an immense amount of data, much of which is increasingly becoming integral to the automation of processes, and no longer limited to information transfer. The footprint of large, centralised data centres will not be able to provide for these applications alone – edge computing is necessary for the proliferation of AI applications.

When it comes to real-time performance, edge infrastructure will enable applications to automatically run environments with new and uncertain data inputs (like a manufacturing site), massively increasing efficiency and driving down operational costs. It is already clear, at this early stage of AI's development, that edge will form an essential pillar of AI infrastructure, enabling real-time app-to-app (A2A) response, limiting data costs to the cloud, and protecting the proprietary data of enterprises. For applications to scale, they will have to navigate a complex array of infrastructure options, as well as other applications running in the same environment. Over time, these applications will negotiate the resources available to them unmonitored, ridding digital environments of the need for human interfaces.

However, for this vision of an AI-managed future, edge infrastructure alone will not be sufficient. Providers will have to play an active role in ecosystems that form around particular verticals and application offerings to overcome the technical and commercial challenges hindering the scale of edge AI solutions.

Operators and other partners in the edge ecosystem must consider how they effectively drive co-innovation, with start ups and with larger scale edge players (e.g. hyperscalers, systems integrators, industry solution providers), to provide both:

- **Technical support:** for the development and delivery of edge AI solutions
- **Go-to-market support:** lending their brand and customer reach to scale smaller ISVs and bring innovative solutions to market.

Should edge players successfully operate within these ecosystem collaboration models that seek to foster co-innovation, they will be able to bring differentiated edge services to market faster and more cost effectively, accelerate the growth of the edge market which has been slower than anticipated to take off, and capture a larger share of the pie by operating across multiple routes to market with a more diverse set of customers.

STL PARTNERS

Research

Consulting

Events