Radian Arc

The Edge Evolution:

AI Everywhere

Table of contents

01. Introduction

Edge computing is not a new concept. Ever since the explosion of IoT devices and smart sensors, the potential benefits of processing data closer to where it's gathered have been clear. Edge orchestration specialists like Zededa have long been providing the platforms to manage large, distributed networks of devices and sensors securely and efficiently, extracting and analyzing valuable data across myriad verticals.

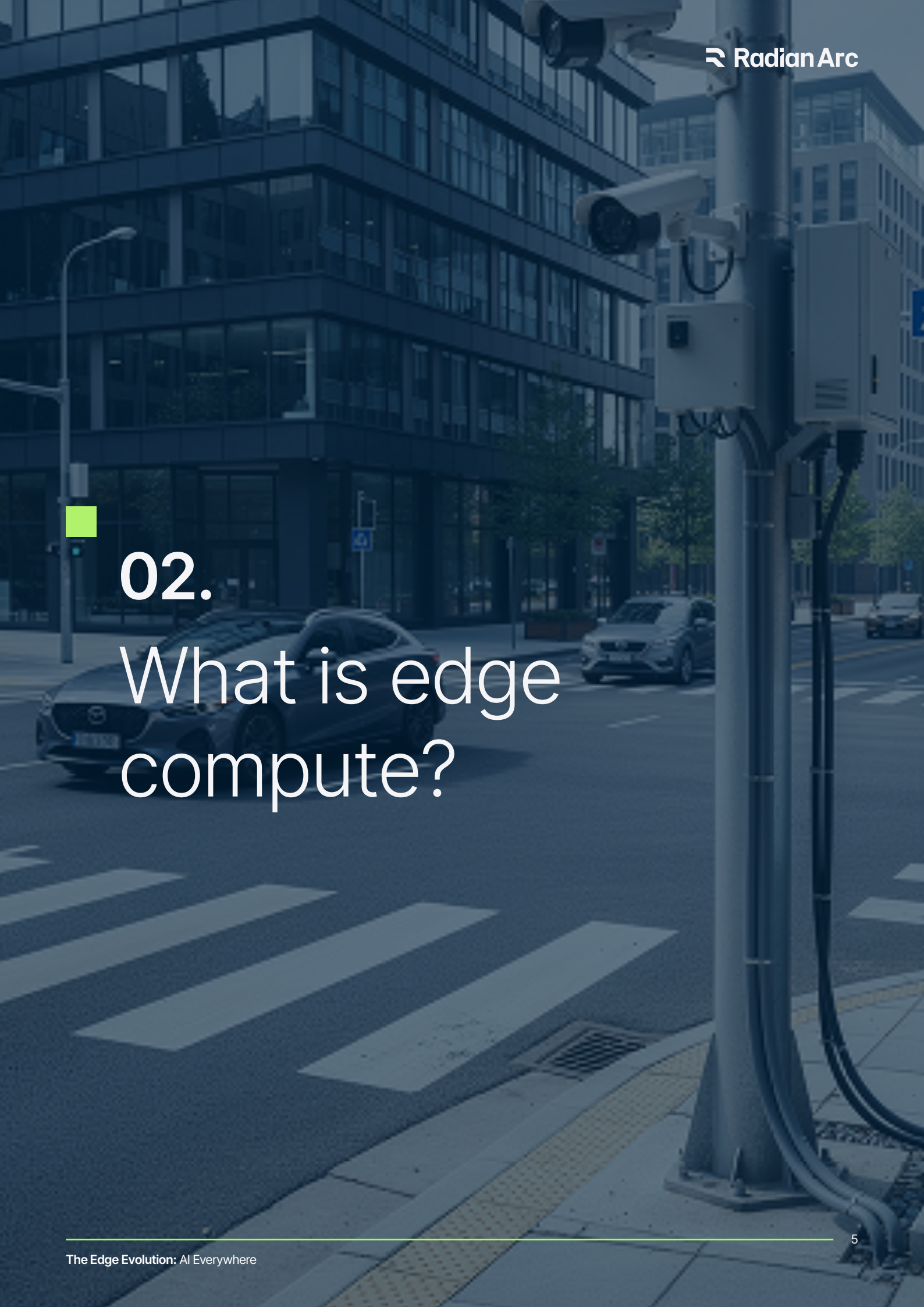
But while the benefits of compute at the edge were evident, they weren't compelling enough to drive cloud computing out of the traditional centralized data centers and into distributed edge nodes.

However, the impact that AI has had on the world is revolutionary and the need to build out ever-more performant AI infrastructure is unrelenting. While huge, hyperscale data centers will always be a core component of AI infrastructure, distributed edge compute is also an important driver for the evolution and growth of AI.

The impact that AI has had on the world is revolutionary and the need to build out ever-more performant AI infrastructure is unrelenting.

In this report we will explore the edge – from its inception and initial applications, through the use cases that made distributed compute a viable and necessary solution, and ultimately how AI demands have made edge compute a necessity.

In a world where AI is fast becoming a ubiquitous utility, edge infrastructure is vital to delivering the AI services that we rely on today and will need tomorrow.



02. What is edge compute?

Cloud computing changed the way the world worked, transforming IT from on-premises hardware executing locally installed software and applications, to remote hardware processing applications 'in the cloud'. This transition was revolutionary, repositioning business IT from significant upfront CapEx costs to OpEx models[i], via infrastructure-as-a-service (IaaS) and software-as-a-service (SaaS) solutions.

This new cloud era allowed enterprises to move away from building and maintaining their own IT suites and data centers, instead working with cloud providers to deliver always-on infrastructure that workers and customers could access from anywhere. Meanwhile, the cloud allowed smaller businesses to access IT and software solutions that would previously have been out of reach due to prohibitive CapEx costs.

But while the cloud brought undeniable benefits such as convenience and flexibility, it also had certain limitations, the most notable being latency. Executing applications and processing data in cloud facilities that could be hundreds, if not thousands of miles away, resulted in inevitable delays that would impact user experience, and in some cases the overall viability of tasks.

Edge computing aimed to address these issues by processing data as close as possible to where it's gathered, thus significantly reducing latency, improving user experience and potentially making new use cases viable.

One of the core drivers for edge computing was the proliferation of IoT devices[ii] and smart sensors. With these connected sensors often implemented to monitor performance, security or even safety, processing and analyzing the data they gathered as quickly as possible, became vital. The increasing power and complexity of these connected sensors also opened the door to the concept of smart, connected cities, with edge compute analyzing all that localized data in real-time, and arming governments and local authorities with unprecedented insight and understanding.

Smart sensors have had just as big an impact on consumers, too. We take things like satellite navigation with live traffic data for granted[iii], but without distributed edge compute and smart sensors sharing data in real time, it simply wouldn't work – or at least it wouldn't work fast enough to be useful.

If cloud computing and edge computing were analogously compared to the human body, the core cloud data center would be the brain, where complex reasoning is performed, while the edge nodes would be like reflexes, producing near instantaneous reactions, without the need for cerebral processing.

Bringing edge compute as close to the user as possible, drastically reduces latency to deliver those reflexive results. But what constitutes the edge, and how do we harness its potential?

Gaming at the edge

As the world became accustomed to streaming music and video from the cloud, rather than from local, physical media, it wasn't quite so simple with videogames. Whether you're listening to Spotify or watching Netflix, it's essentially one-way traffic – you choose what you want to watch or listen to, and it is streamed to your device. Gaming requires continuous two-way communication, and the speed of that communication is paramount.

A cloud gaming experience lives or dies by latency – controller inputs need to be instantly reflected on screen for both immersion and playability, meaning that the game itself must be processed as close to the player as possible. Edge compute nodes designed specifically to run games for local players can reduce latency to acceptable levels, allowing almost any device with a screen to be a games console.

As broadband speeds increase, along with 5G mobile connectivity, edge compute installations can allow consumers to play their favorite games whenever and wherever they like, on any connected device.

The telco connection

Building out a distributed network of edge nodes or PoPs (Points of Presence) requires both the physical locations and high-speed connectivity. Telecom operators already have both sides of that equation, providing the ideal foundation for edge compute networks. Because edge compute installations are not designed to deliver the extreme levels of performance seen in hyperscale data centers, they can be designed to fit existing telco locations – whether that be within cell tower installations, exchanges, aggregation points, or even network cabinets.

It's not just about fitting existing telco infrastructure, the compute installations at the edge must be resilient, delivering long-term reliability even in hard-to-reach locations, while also being manageable and configurable from anywhere.

Telcos are also the gatekeepers of high-speed connectivity - 5G cellular data, fiber-optic broadband networks, leased lines, etc. - ensuring that any data captured or applications executed at the edge, will deliver results almost instantly. Not only does this provide the best possible user experience for individuals using edge-based services, but it also ensures that services demanding real-time data processing - like smart traffic management or fire detection sensors – operate and trigger action without delay.

The viability of telecom networks as the foundation for edge computing platforms was clearly recognized in 2020 when the Telco Edge Cloud[iv] (TEC) initiative was launched, with 19 global network operators collaborating to design and develop edge computing services. Since its launch, TEC has grown to include leading technology providers alongside network operators, all working towards developing edge services and solutions.

The potential use cases for edge compute have been clear for many years, while the incumbent connectivity and physical locations that telco networks can provide, represent the ideal foundation for deploying edge compute nodes. But the real catalyst for creating extensive, distributed edge compute networks came with the explosion of AI.

Where is the edge?

While the nature of edge computing is clear, the actual location of 'the edge' is somewhat nuanced.

- **Device edge:** Literally the very edge of the equation – where the data or queries originate. The device edge could be your smartphone, your smart doorbell, or the smoke sensors inside an office building. Depending on the device, at least part of the data processing could be executed locally.
- **On-premises edge:** Enterprises or local authorities can deploy data centers on-site at factories or utility plants, ensuring that any data gathered from sensors and applications is processed, analyzed and acted upon instantly.
- **Network edge:** Small compute installations designed for specific workloads – can be deployed within cell tower installations or network hubs.
- **Distributed data centers:** The edge compute data centers embedded within telco networks, delivering low-latency execution of workloads close to the source of data capture.

Each step further away from the origin of the data / query will add latency, but any of the edge locations listed will deliver significantly faster processing compared to traditional, centralized, cloud-based data center scenarios.

03.
AI at the edge

The impact that AI has had on the world is impossible to ignore, and that impact is only going to grow. This is a truly transformative technology that's becoming ubiquitous in its usage across industries and geographies, for both B2B and B2C markets. But that ubiquity requires significant infrastructure, and as the need for AI continues to grow, so must the infrastructure that powers it.

When we think about AI infrastructure the most common image is massive, hyperscale data centers, designed to provide vast levels of high-density compute for LLM training and complex reasoning. There's no denying that the need for hyperscale facilities will continue to grow across all territories – Goldman Sachs predicts that the five highest-spending hyperscalers will invest a combined $400 billion in such facilities in 2026[v].

But while the need for hyperscale data centers will certainly not diminish, there's more to AI infrastructure than these massive, cloud-based compute factories. In a world where the usage models for AI are almost limitless, sometimes there are more important factors than raw compute power.

Targeted inference

When it comes to AI the two most common workloads are training and inference. Training large language models (LLMs) takes an incredible amount of compute power, which is why the need for hyperscale AI data centers is showing no sign of slowing. But inference can be far less compute heavy, making distributed edge compute nodes viable processing platforms. Additionally, when you consider that inference is the production or operational face of AI, it's vital to ensure that workloads run as efficiently as possible, and to make the most of that efficiency, you need low latency, once again, making the edge ideal.

Employing edge compute PoPs creates a distributed AI infrastructure with a tiered execution model – data and workloads that can be processed locally will be, while more complex workloads continue to larger, cloud-based data centers. Crucially, those workloads that are processed locally, will be executed with minimum latency, providing near-real-time response.

The key to ensuring the fastest possible inference performance is predicting the types of data and workloads that will need to be processed. Essentially, an edge compute node can be highly targeted – whether that be to the needs of a local populace, or the specific requirements of an enterprise or local authority.

A manufacturer could, for example, deploy an edge compute PoP that's specifically designed to process data supplied by smart sensors within a factory, allowing for real-time analysis of productivity or even potential mechanical failures.

A local authority could deploy an edge node designed to deliver inference based on a localized small language model (SLM) that can respond instantaneously to the most likely queries from people in the area. If the local area includes tourist attractions, shopping facilities, transport hubs, etc. the SLM would be trained – and learn - accordingly.

National governments could prioritize real-time AI powered translation based on international visitor

metrics and the most common questions and enquiries within a localized area and refined by native language.

But what do these edge AI PoPs look like, and how do they form part of a wider AI infrastructure?

Edge installations

According to the GSMA[vi] 'AI adds a new dimension to the value of edge through distributed inference.' While accepting that edge compute has always been a core aspect of the 5G value proposition, the GSMA believes that 'AI is driving, and will continue to drive, an upward inflection in mobile data traffic growth from consumer and enterprise customers.'

This belief is shared across both the telecommunications and AI infrastructure sectors and is the driving force behind the distributed AI data center strategies that are taking shape today. Put simply, telcos can provide the foundation on which to build out networks of distributed AI factories at the edge.

Utilizing telecom network infrastructure as the basis for distributed AI edge compute requires flexibility and adaptability – designing, building and deploying AI factory installations that fit the physical location and deliver on localized compute needs.

Compute density and cooling are two key factors of any edge AI deployment, ensuring that as much compute power as possible is delivered within the available physical space, and that all the hardware – CPUs, GPUs, memory, storage, interconnects, etc. – is kept cool enough for maximum performance and longevity.

Maintenance schedules are also important at the edge, where simply accessing an installation could be difficult – cell towers are often built in hard-to-reach locations to ensure blanket network coverage, so any edge compute hardware must deliver long-term, maintenance free reliability.

Immersion liquid cooling is ideal for AI compute at the edge, ensuring optimal hardware performance and long-term reliability, even in the most inhospitable environments. The modular nature of immersion cooling pods also aids flexibility at the edge – from single pod installations at cell towers, to multi-pod deployments in exchanges and telco data centers.

> When it comes to AI the two most common workloads are training and inference.

04.
Data privacy and sovereignty

Edge AI delivers additional advantages beyond the reduction in latency. With traditional cloud computing, where centralized hyperscale data centers could be in another country or territory, or where redundancy and failover protocols could result in data being processed further afield, there are potential risks around data sovereignty and privacy. Certain data – especially personal data – is often subject to local regulation, meaning that it must be stored and processed within certain geographical borders, so any uncertainty of where your data is going once it leaves the source could be problematic.

However, with edge AI compute the data is being processed as close to the user as possible, with the results being returned directly. Consequently, processing at the edge can help ensure data sovereignty, given that the edge compute node is likely to be within a few miles of the user and well within any regulated geographic boundaries.

This isn't just an advantage for end-users, either. Software developers building AI applications will be subject to data privacy and sovereignty regulatory compliance. But if developers ensure their applications run and process all locally collected data at the edge, it will remain within sovereign borders. As far as privacy goes, if data never leaves the edge and is flushed after processing, data privacy compliance should be guaranteed.

The evolution of MEC and split compute

Multi-access Edge Computing (MEC) is essentially the picture of edge that we've been painting thus far. It's the evolution of cloud computing, whereby processing of data is brought closer to the user, reducing both latency and network congestion. Split compute takes the edge concept a step further, bringing 'device edge' into the spotlight.

As already covered, device edge refers to the origin point of the data. If that origin point is a simple sensor that's monitoring and collecting data, all the processing of that data must be done at the next edge layer – whether that be on-premises or network edge. But if the origin point of the data is a device with its own processing power – a smartphone, tablet, laptop, etc. – then the processing load can be split between the device itself and a network edge layer.

In an AI inference environment, a split compute model would divide a workload between the processing power of the device that initiated it and a remote edge compute node. The calculation of that division would depend on the processing power of the device, the energy requirements – think battery life – for local processing, and the latency benefit / impact between the local device and the edge. Additionally, there is also the scenario where personal / sensitive data is kept on the device and either anonymized or simply not shared with the edge compute node.

With mobile devices becoming ever-more powerful, and the latest mobile chipsets including AI specific neural processors, split compute AI inference will undoubtedly gain momentum. And coupled with growing networks of AI edge compute nodes, real-time inference will most likely become the norm.

05.
The future of edge

The promise of edge compute has been there for a very long time, but the revolutionary impact of AI is powering the realization of that promise. The tangible benefits that distributed edge AI data centers and PoPs can deliver are impossible to ignore, driving AI infrastructure specialists and telcos to work together to build a new frontier of edge compute.

"The global edge AI market is on a steep upward trajectory," according to Joshua David, senior director of edge project management at Red Hat[vii]. A belief that's backed up by predictions of growth in the sector – STL Partners estimates edge AI revenue reaching $157 billion by 2030[viii]. And to realize that revenue potential, the infrastructure and applications must be deployed and developed to facilitate it.

> The global edge AI market is on a steep upward trajectory.

The need and appetite for massive, centralized hyperscale data centers will continue to grow – LLMs will keep advancing and training and refining those models will require ever-more powerful compute. But inference, especially refined and targeted inference, will ideally happen at the edge, delivering real-time – or as close to it as possible – processing and analysis of data.

The evolution of AI inference is starting to deliver the smart cities, countries and world we've been promised for so long. From AI powered analysis of meteorological sensor data to predict the weather, to the real-time sensor monitoring required to make autonomous vehicles safe and efficient, all that inference must happen at the edge.

Now is the time to build the edge network for tomorrow so that we can keep pushing the boundaries of what AI is capable of, and how we can best leverage this generation's most disruptive and exciting technology.

Radian Arc

A Submer Group Company

Real-world intelligence

Learn more at our blog

Submer Group

[i] learn.microsoft.com/en-us/training/modules/get-started-with-finops/1-introduction

[ii] ibm.com/think/topics/iot-edge-computing

[iii] tomtom.com/solutions/traffic-and-mobility-intelligence/

[iv] gsma.com/solutions-and-impact/technologies/networks/gsma_resources/tec-value-whitepaper/

[v] goldmansachs.com/insights/articles/how-ai-is-transforming-data-centers-and-ramping-up-power-demand

[vi] gsmaintelligence.com/research/distributed-inference-how-ai-can-turbocharge-the-edge

[vii] infoworld.com/article/4117620/edge-ai-the-future-of-ai-inference-is-smarter-local-compute.html

[viii] stlpartners.com/research/edge-ai-market-forecast-computer-vision-leads-the-way/