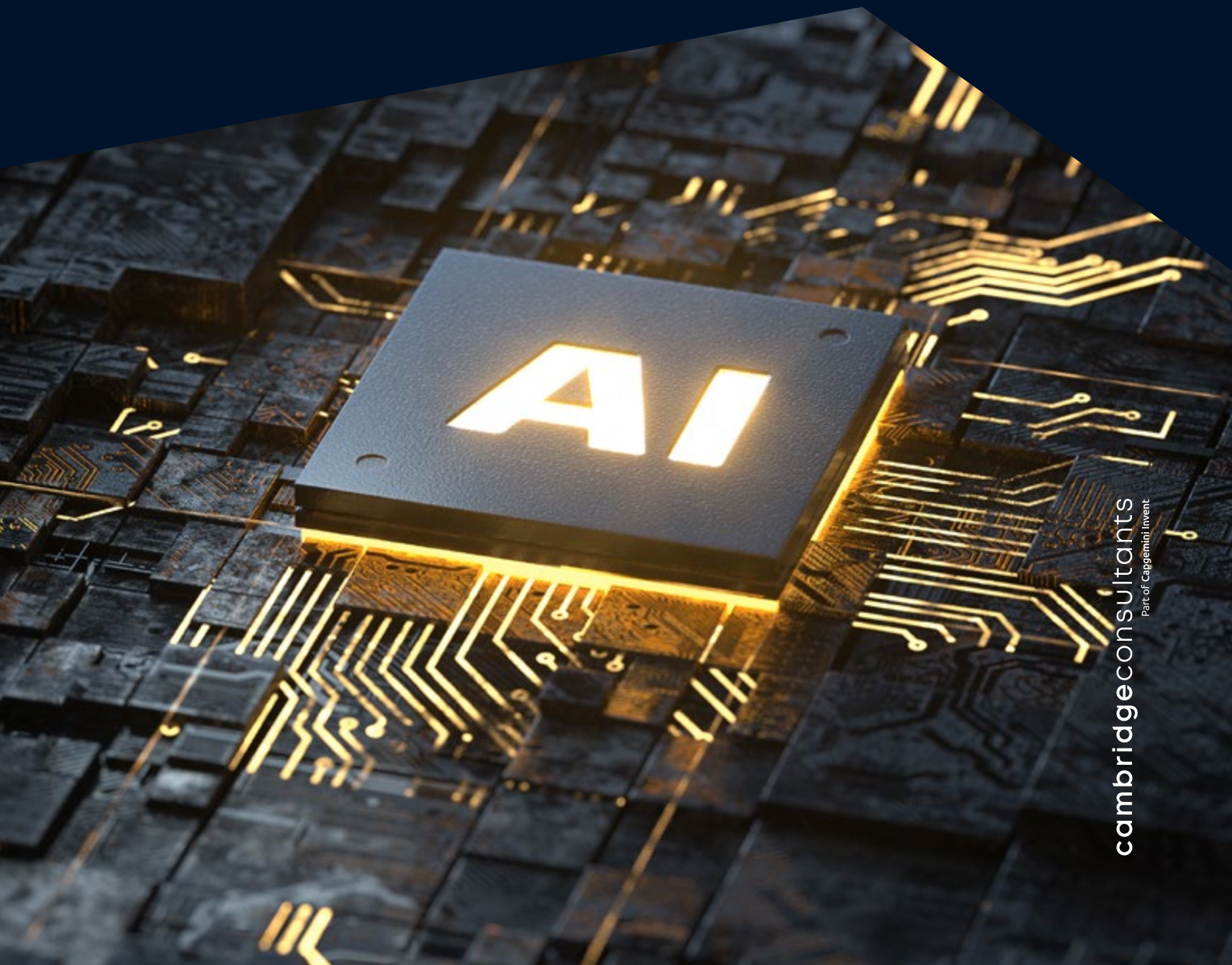
The future of AI is at the edge

How low power and intelligence can go hand in hand



cambridgeconsultants
Part of Capgemini Invent

Contents

Executive summary

Cloud computing is the standard for delivering AI in connected devices. Data is collected by an edge device and uploaded to the cloud for processing. It's flexible and scalable. But now, advances in silicon and machine learning have heralded a transformative opportunity to complement the cloud by putting AI processing into those devices. Through this, a host of new, AI-enhanced products and services is made possible.

The cloud has undoubtedly fueled the proliferation of AI across a variety of market sectors and technology areas, and it will continue to do so. But AI's dependence on cloud computing has resulted in the limits of AI becoming tightly coupled with the limits of the cloud. There are application areas which remain untouched by the advent of AI, not because they do not stand to benefit from increased intelligence, but because they cannot be made interoperable with the cloud.

Edge AI is the principle of performing AI processing closer to the connected device, often on the device itself. It is allowing us to push past the limits of the cloud and it unlocks opportunities to incorporate intelligence into those application areas previously left behind. Three key limitations of cloud AI processing have been identified which can be overcome by embedding intelligence into the edge.

- **Latency** – Relying on the cloud means relying on a constant connection to the cloud. Edge AI enables intelligence in low-latency and safety critical applications
- **Privacy** – Edge AI provides a locally processed alternative to uploading data, enabling intelligence to be applied to data too sensitive to share
- **Power/cost** – Being always-on and always-connected often sets a minimum floor to the cost of cloud-based intelligence. Edge AI enables ultra-low cost, ultra-low power intelligent devices

With all the potential benefits which Edge AI can offer, businesses need to be aware of the management overhead of the edge and how it compares to that of the cloud. This paper outlines the impact of Edge AI on a small selection of industries, highlighting opportunities for value creation by adding low power intelligence to the edge.

1 Edge intelligence is enabling products, services and business models to be reimagined

Cloud computing architectures have become the dominant approach for managing and processing the large volumes of data that underly the analytics and artificial intelligence (AI) algorithms that power today's intelligent products and services. This dominance has made cloud computing adoption a mainstream phenomenon with worldwide public cloud revenue growing by 17% in 2020 to nearly USD 270 billion[1].

While the cloud does offer extensive economies of scale, beneficial for high volume data processing, it does encounter limits arising from latency requirements, power constraints and additional bandwidth costs. Applications that require real-time processing cannot be limited by network latency and therefore may not be well served by the existing cloud infrastructure. A clear solution to this problem is to keep the data close to the user for real-time processing. This can drastically reduce the latency of connections due to the shorter roundtrip that the data must take, and it can also address potential regulatory, privacy and security constraints by avoiding moving the data away from its source at all.

This is where the concept of edge computing comes in. Edge computing can be simply defined as a location that is closer to the user than the cloud, where data storage and processing can take place. The concept dates back more than two decades to the introduction of content delivery networks (CDNs). CDNs are a network of nodes which cache content closer to the end user allowing that content to be served more quickly and internet transit fees to be avoided.

A key use case for this currently is video streaming content. Netflix has built out an extensive CDN for this purpose. The edge concept has since developed from simply storing and serving data to include computational capabilities allowing valuable information to be processed and real-time services to be delivered.[2]

Figure 1 illustrates the trade-offs between compute and latency that arise using different levels of the network. The following definitions describe the different locations ranging from the cloud to a device, reflecting the proximity to the user:

- **Cloud** – Data center grade compute servers which can be in public or private facilities
- **Hybrid cloud** – A mixture of a public cloud accessed through the internet and a private on premise cloud server
- **Edge server** – A traditional edge computing server or device that is at least one hop away from the device and the cloud
- **Edge AI** – A device (usually a sensor) with AI processing capability that is located closer to the user
- **Sensor-only device** – A device with no processing capabilities that simply collects and transmits unprocessed data to either the cloud or an edge server

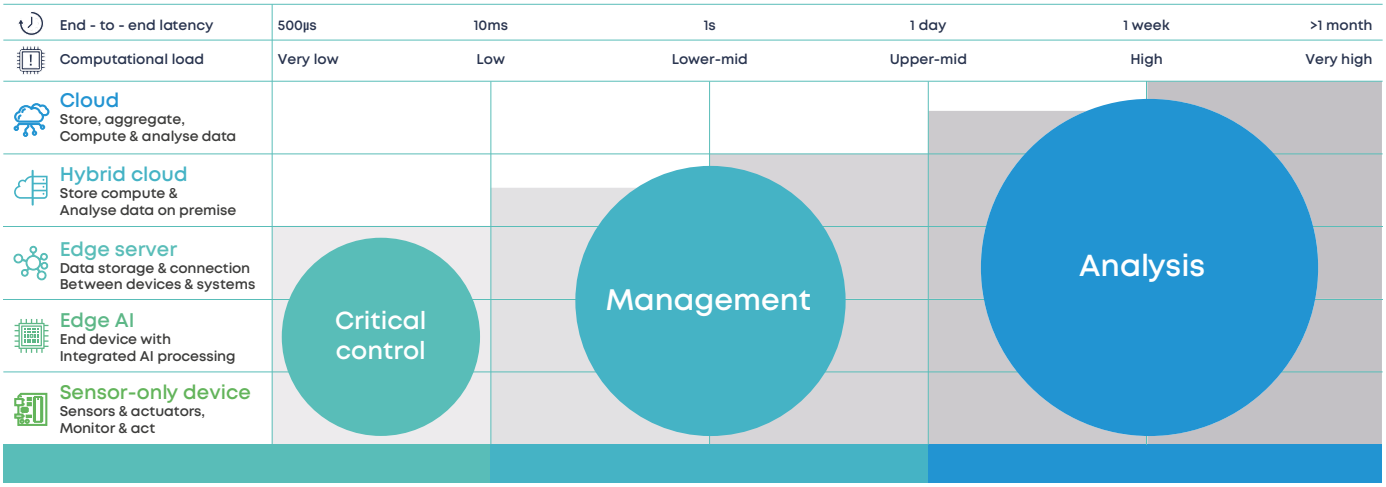


Figure 1: Latency and computation requirements for different network levels (**Source:** Cambridge Consultants)

What has currently been achieved with edge computing is only a glimpse of what companies can achieve when they utilize the power of AI at the edge. At Cambridge Consultants, we envision that AI processing will increasingly be run on low-power edge devices including some reaching less than a dollar in cost. This will create a vast array of use cases that are infeasible in a cloud-only world. These intelligent edge devices will be able to make computations on premise, or on the end device, without relying on a continuous data transfer between device and cloud.

1.1 What exactly do we mean by Edge AI?

'The Edge' is a nebulous term that has different definitions depending on the context in which it is used. It always refers to a location for processing or storage of data that is closer to the user than the cloud, but beyond that it can mean a range of different things. As Figure 1 shows, edge processing could refer to an Internet of Things (IoT) hub that receives data from many devices, but equally it could refer to processing placed on the device itself with further processing taking place at a gateway device (such as an IoT hub).

As the amount of data generated continues to grow, it can become more important to store and process data closer to the user because of the increased data transmission costs. More significantly however, absolute requirements for low latency, privacy and security make edge processing mission critical for certain use cases (e.g. autonomous driving systems, industrial robotics, smart grids). Moreover, adding a level of intelligence to the edge in the form of specific AI algorithms will allow for achieving feature personalization/ customization on premise.

The majority of AI embedded in the edge is one particular type of AI task – machine learning inference. Inference is the point at which all the learning accumulated during training is deployed on real-world data to infer a result. Throughout this paper, when we talk about Edge AI, we are predominantly talking about performing inference at the edge.

Training is a task typically many orders of magnitude more computationally complex than inference, requiring varied data collated from multiple sources. It can be performed as a batch task without strict real-time requirements. As such, it is a task very well suited to cloud computing. Training at the edge is a developing field in its own right (of which Federated Learning and One-Shot Learning are particularly promising examples) but the focus of this paper is on edge inference.

"Absolute requirements for low latency, privacy and security make edge processing mission critical for certain use cases (e.g. autonomous driving systems, industrial robotics, smart grids)."

1.2 What benefits could Edge AI bring to your industry?

While edge devices always process data closer to the user, they take different forms depending on the specific industry use cases. Below we explain the benefits of Edge AI to some of the markets most likely to benefit.

Healthcare

Edge AI will enable healthcare applications to benefit from data gathering which can be turned into real-time actionable insights. One example is continuous health monitoring in hospitals where Edge AI can be in the form of a wearable or ingestible device, smart implant or non-contact sensor (e.g. radar for autonomous monitoring of patient vital signs) which will help reduce the number of medical errors.[3] Such technology could help hospital nurses to do more efficient work and detect potential issues with patients earlier.

Another example relates to medical imaging where large file size images (e.g. radiology Digital Imaging and Communications in Medicine (DICOM) images) can be processed locally by Edge AI rather than sent to the cloud for analysis where bandwidth costs can be significant. NVIDIA has demonstrated this concept using the federated learning support built into Clara, its medical imaging deep learning toolchain.[4]

Mobility

Autonomous vehicles – whether that includes passenger vehicles or heavy vehicles and plant in the agricultural, construction or mining sectors – will rely on high volumes of sensor data to negotiate their complex environments. Overcoming the challenge of processing all this data in real time will be made possible by integrating intelligence at the sensor fusion level, or within the sensor devices themselves (e.g. camera, radar, LiDAR). Furthermore, addressing the safety critical aspects of autonomous driving requires computational and transmission latencies in the order of tens of milliseconds which can only be achieved with intelligent computing located within the vehicle.

Smart grids and distributed generation

The emergence of distributed energy resources (DER) and smart grids is changing the conventional paradigm of a centralized grid management and a unidirectional power flow. In order to manage increasingly distributed energy production, increased monitoring of grid nodes will be required to balance production and storage of electricity. This would be challenging to achieve with current centralized architectures due to the significant roundtrip time delay. Hence, Edge AI devices in the form of local smart nodes can manage DERs more efficiently by forecasting local production peaks and potential grid disconnects which will lead to a smoother grid management operation.

2 Cloud AI processing alone cannot power the expected growth in connected device intelligence

Over the past decade the cloud has been a key driver of growth for the world's most valuable companies. Amazon, Microsoft and Google have all built sizable offerings, generating value by providing cost-effective and innovative scalable compute and storage solutions for a wide range of applications. The public cloud model has been widely successful and has allowed many companies to derive valuable insights from data by using compute on a scalable, consumption basis.

But now, even the public cloud giants are expanding their service offerings for customers who want to process their data closer to the edge.

The cloud is suited for the most computationally intense AI processing tasks and will continue to grow

AI, perhaps more than any other technology area, has been fueled and shaped by the widespread availability of cloud computing and storage. Machine learning models are able to feast on the vast datasets harvested from the billions of connected devices streaming raw data back to central servers. Computationally intensive training and inference operations can be farmed out in batches to high-performance clusters, allowing cost-effective access to the latest advances in parallel processing hardware. Crucially, in a nascent industry with rapidly evolving software ecosystems, AI developers have been able to invest safely in flexible cloud infrastructure where they might otherwise have hesitated for fear of backing the wrong hardware platform in a fragmented market.

The advent of Edge AI processing must not be mistaken for the demise of the cloud. Cloud computing is here to stay, and it will continue to play a vital role in delivering AI functionality for the foreseeable future. Edge acceleration opens up brand new possibilities for AI enabled applications, as well as providing opportunities to improve upon existing ones. But the cloud will continue to provide a backbone of flexible and scalable processing power to the hungriest of AI applications.

There are limitations of cloud computing which can be addressed by edge solutions

There are, however, technical limitations to the level of AI functionality that a cloud-based service can provide – most notably due to network latency and reliability. Use cases with only tens of milliseconds to spare between seeing something and perceiving it cannot afford to wait for a response from a remote server. Reliance on the cloud might harm the usability of an application in the event of poor connectivity, or it may even undermine the efficacy of a safety critical application if a minimum quality of service cannot be guaranteed. Because it can be used to bypass these limitations, edge processing unlocks opportunities to incorporate intelligent features into applications so far untouched by AI.

Even where cloud-based solutions exist, there may be benefits to moving some or all the AI processing towards the edge. One key benefit is a reduction in network traffic, which in turn can contribute to a reduction in cost and an improvement in user privacy. Some of the most exciting Edge AI opportunities arise from exploiting this benefit to its extreme. Where edge intelligence is used to reduce bandwidth requirements so dramatically that a connection is only established when there is something interesting to say, AI functionality can be deployed in ultra-low power connected devices which can run for years on a coin cell.

The volume of data produced will increasingly be done at the edge

According to Gartner, in 2018 only 10% of data generated by enterprises was produced and processed outside the cloud or a centralized data center. This figure is predicted to be around 75% in 2025.[5] This will encourage further edge development where data producers have increased control of their own data.

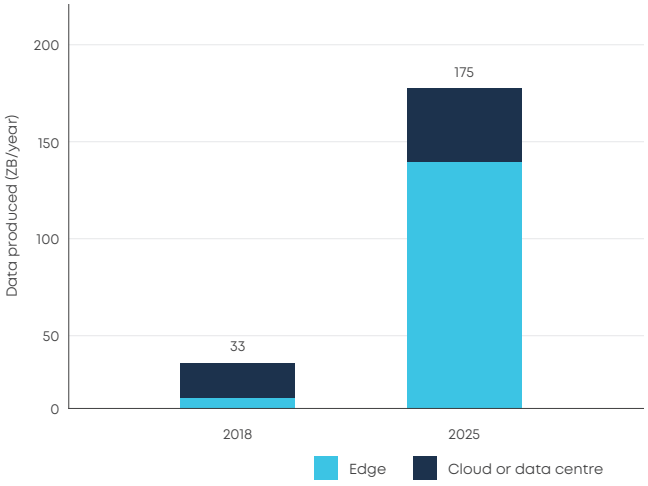


Figure 2: Growth in data at the edge vs in the cloud or a data center (**Source:** Based on IDC, via EC[6])

An increasing proportion of all AI processing will happen at the edge

As intelligence continues to proliferate into all kinds of connected devices across all market sectors, an increasing proportion of that intelligence will be powered by Edge AI processing on the devices themselves. Estimates from IDC, a market intelligence provider, show the Edge AI processor market is expected to grow by a factor of 20 between 2018 and 2023.

Anyone aiming to push the limits of low latency or low power AI must make the most of edge processing in order to overcome the network limitations of the cloud. Anyone with concerns about the bandwidth requirements and privacy implications of continuing to transfer ever-increasing quantities of raw data for cloud processing should be considering Edge AI alternatives. Everyone should be considering whether Edge AI could be providing them with more cost-effective intelligence.

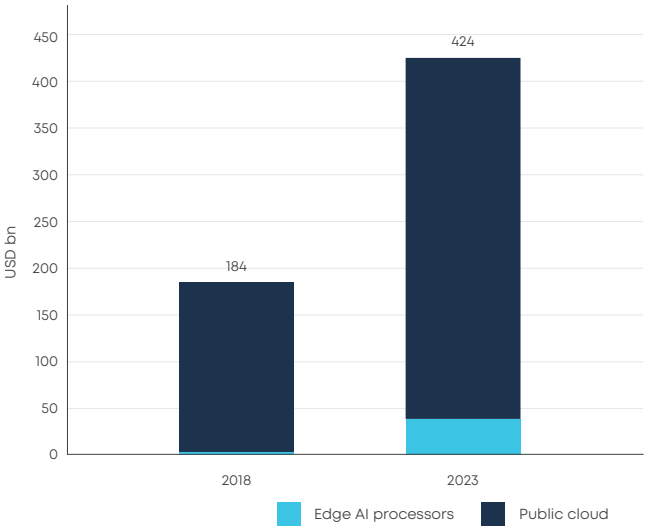


Figure 3: Edge AI processing growth compared to public cloud (**Sources:** IDC[7], Gartner[8])

"The Edge AI processor market is expected to grow by a factor of 20 between 2018 and 2023."

2.1 Cloud architectures are only as responsive and reliable as the network they depend on

Intelligent devices which rely on the cloud for their AI processing are also reliant on their connection to the cloud. Innovation in telecommunications continues to yield impressive improvements in the bandwidth, latency and quality of service that a connected device can expect. But for some applications this reliance still constitutes a fundamental limitation to the level of AI functionality that a device can provide.

Caring about latency means caring about communication delay

The applications most restricted by reliance on the cloud are those with especially low latency requirements. Virtual or augmented reality applications, for example, must operate at higher frame rates than recorded video in order to keep the latency between the camera moving and the image updating below perceptible levels[9]. If an augmented reality headset were to send each frame to a cloud server for processing, then a round trip delay of 35ms, which would be considered good for a typical internet connection[10], would introduce a noticeable lag and possibly induce simulation sickness.

Network latency is especially relevant for safety critical applications with low latency requirements, such as autonomous driving or healthcare robotics. When the consequences of a missed or delayed response are unacceptable, an application must account for the worst-case (rather than typical) roundtrip delay, which eats further into the available window of processing time.

If latency is not an issue, poor network reliability might be

For applications where latency is not a particular concern, the reliability of the network connection may still be a limiting factor. Consider, for example, a wildlife camera in a remote location which includes an AI application for detecting and classifying animals in its view in order to determine when to record footage. Guaranteeing a network connection to that remote location may not be possible.

Even if a network connection is normally reliable, reliance on that connection may still complicate the design of an AI application which must consider what happens in the rare cases when the connection is not available. Imagine a targeted fire suppression system for residential settings, capable of scanning a room for flames and extinguishing sources of fire without blanketing the building with water from a sprinkler system. Ordinarily, a domestic internet connection could be considered to be reliably present, but it would clearly not be fit for purpose if the home Wi-Fi router were engulfed in flames and the flame detection and localization algorithms were running in the cloud.

Reliance on network connectivity can be a vulnerability

Network reliability becomes particularly relevant for applications which have to deal with adversarial human behavior. The threat of a denial of service attack on a fleet of self-driving vehicles (e.g. passenger automobiles on a road network, forklifts in a distribution center or freight vessels at sea) is one reason to ensure that each vehicle has sufficient edge capabilities to at least come to a safe stop independently.

An intruder detection system covering a setting like a museum or warehouse might benefit from using facial or gait recognition to distinguish between an expected staff member and an unexpected stranger. A cloud-based system could probably fail safe in this example by raising a general alarm if it loses connectivity, but a system which could remain operational if the lines were cut would be preferable. Alternatively, consider again the wildlife camera in a remote location and imagine that there are many of them hidden throughout a nature reserve. They could perhaps include an application to detect and record humans in order to prosecute and deter poachers. If they relied on cloud processing to decide whether a human was in the frame or not, then a determined poacher might be incentivized to disrupt the network.

Edge AI processing reduces reliance on network connection

None of the problems posed in this section are insurmountable, and many can be resolved while continuing to take advantage of cloud AI processing, through careful consideration of network latency and reliability. However, as the AI processing capability of edge devices continues to grow – and the proliferation of AI technology permeates into low-latency and/or safety critical applications – the most effective way of dealing with network latency and reliability concerns will be, in an ever-increasing proportion of cases, to reduce reliance on a cloud connection and instead perform some or all of the AI processing on the edge device itself.

2.2 Edge processing can help safeguard an end user's privacy and data security

Cloud-based AI solutions rely on end users uploading raw data for processing, which has wide-ranging implications for their privacy and data security. By keeping processing on the device which acquires the data in the first place, edge-based solutions can reduce or eliminate the need to transfer raw data.

This does not apply in all cases. There are applications where it might be necessary to upload or store the data for some other purpose than AI processing. A security camera system capable of identifying instances of shoplifting would still need to keep hold of the raw video footage as evidence or for further investigation. A single application might be a part of some wider service, and there may be valuable insights from the data to be exploited at a later date, or in combination with other data.

However, for applications where the data is acquired solely for the purpose of feeding a single AI-enabled function, it can be discarded as soon as it has been processed. Once a smartphone has translated an item on a foreign language menu or identified a plant in a flowerbed, the picture is no longer required. If that picture can be processed locally, it need never be uploaded.

To what extent a reduction in raw data transfer translates into a privacy benefit varies between applications, which can be broadly grouped into three categories based on what information is uploaded in lieu of the full raw data set.

Uploading nothing at all is best

Firstly, in some applications edge processing allows a device to provide AI functionality without the need to upload any user data from the device at all. In these cases, the service provider saves on the costs of handling and protecting that data, and the end user is more likely to trust that they can control its entire chain of custody.

Uploading processed information is preferable to uploading raw data

Secondly, in cases where raw data is processed at the edge and the result is uploaded to the cloud to be actioned or analyzed further, the end user is likely to be equally as concerned about the security of the processed information as they would have been about the raw data – and the service provider must go to the same lengths to protect it. A breach of voice assistant command histories would be just as serious in processed text form as in a raw audio format.

However, edge processing might still reduce an end user's overall exposure in these cases because only the information relevant to application is uploaded. Images, in particular, are information-rich data sources likely to contain potentially sensitive details not pertinent to the application.

Imagine, for example, a cat-whisperer app which can determine a cat's mood from a few seconds of video footage. Raw footage of the cat might be innocuous in the vast majority of cases, but on rare occasions it might be footage of a cat sat on a desk next to a bank statement. A breach of the processed data is less worrying in this case. It probably would not matter if somebody were to gain unauthorized access to a pet's mood swing history.

Uploading some raw data is preferable to uploading all the raw data

Thirdly, in some applications edge processing is used as an initial triage for determining which pieces of raw data are worth uploading for more detailed analysis. In these cases, raw data is still uploaded but in smaller quantities, reducing the user's overall exposure.

Sticking with the cat-whisperer example, perhaps the computationally complex task of determining the cat's mood is performed in the cloud, but a simpler network running on the edge device detects the presence of a cat in the frame and crops it accordingly, so that only the relevant sections of relevant frames are uploaded.

Sensitive data is of paramount concern

The importance of privacy and data security is most immediately apparent in cases where the user data is widely regarded as sensitive, such as medical or financial data. Perhaps a patient using a triaging app on their mobile phone might be more inclined to capture an image of the symptom for identification purposes if they knew that picture would be used only for analysis on their device and not transmitted anywhere else. The principle of moving processing towards data rather than vice-versa is already well established in medical research. Federated approaches to machine learning could allow hospitals to pool results without having to share the original sensitive patient data.[11]

All data is of some concern

Privacy concerns are not, however, limited only to the most sensitive data. Consider, for example, the sheer quantity of rich video and audio data captured by always-watching intelligent cameras and always-listening voice assistants from privileged positions throughout our homes. We each trade our privacy for convenience at different exchange rates. For some, streaming live footage of the living room would be an acceptable price to pay for an intelligent home that anticipates the needs of its inhabitants and reacts to their moods and intentions. To reach a wider market, of more privacy conscious users, an intelligent home service could perform some or all of that processing locally in order to minimize the amount of footage transmitted. This is already the case for voice assistants, which typically perform keyword detection at the edge so that only the speech directed at them is sent back to the cloud for further processing.

2.3 Trade-offs are increasingly falling in favor of the edge

For deployers of AI applications who are neither pushing against the limits of network latency and reliability, nor constrained to operate at the edge for privacy reasons, the decision of whether to perform AI processing at the edge or in the cloud will come down to cost. Cloud processing will continue to be the more cost-effective option for the most computationally intensive tasks, but it is already cheaper to provision devices to perform smaller tasks themselves. When considering the total cost of ownership of an AI-enabled device, edge processing is becoming the more cost-effective solution for an increasing proportion of applications.

The cost of delivering AI functionality at the edge is rapidly decreasing as hardware AI accelerators and their software toolchains mature. Designers of systems-on-chip for mobile phones, such as Qualcomm or Samsung, have now been including dedicated AI acceleration in their flagship chips for at least three generations.[12] The companies ordinarily associated with supplying the computing power behind the cloud, such as Intel and NVIDIA, have also been selling Edge AI solutions into the automotive market. For custom silicon solutions, neural network processors have recently become available as customizable IP blocks from a variety of vendors such as Arm, Imagination and Cadence. The total cost of ownership of an AI application running on a connected device can be simplistically considered to be the sum of three major components:

- **Compute costs** (the upfront cost of on-device AI processing capacity, amortized over the lifetime of the device, and/or the operational costs of hiring cloud computing units as required)
- **Bandwidth costs** (the operational cost of sending raw data back to the cloud for processing, including the power required to establish and maintain a network connection)
- **Power costs** (the cost of powering the AI capabilities of the device, including associated overheads such as manually charging or changing a battery, or providing extra cooling to the device)

Figure 4 is a generalized simplification illustrating the key factors behind the cost trade-offs of edge and cloud solutions. The shape of the curve can be taken to be generally applicable but the numbers on the axes are intended to be indicative only and would vary between different device types and application areas.

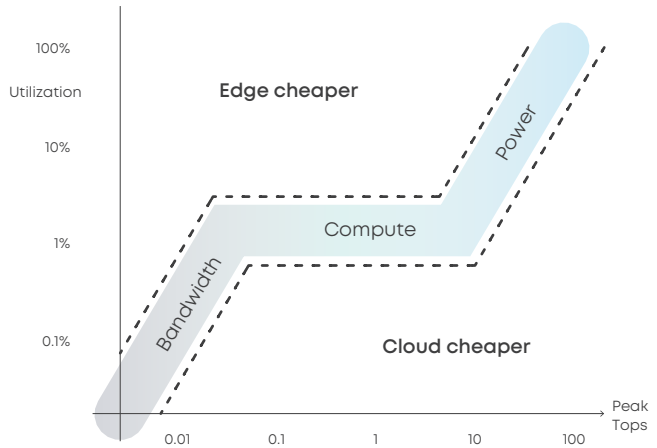


Figure 4: Cost trade-offs of cloud and edge solutions (**Source:** Cambridge Consultants)

Bandwidth costs dominate and the edge is cheaper for very simple tasks

For small AI tasks with low peak performance requirements (e.g. less than 0.01 TOPS), the bandwidth cost of transmitting raw data exceeds the cost of processing it locally, so an edge solution is highly likely to be more cost-effective.

This principle is already well understood in the context of video streaming, where the cost of video codec processing is preferable to the cost of transmitting uncompressed video.

"The cost of delivering AI functionality at the edge is rapidly decreasing as hardware AI accelerators and their software toolchains mature."

Compute costs dominate and utilization is key for mid-range tasks

For larger tasks (peak performance requirements in the range 0.1 to 10 TOPS), the dominant cost is that of the computation itself. Whether edge or cloud processing is the more cost-effective solution in this case depends on how often that processing is needed, much like the decision of whether to purchase or hire a piece of equipment. Applications with 100% utilization, such as those operating in real-time on a constant stream of sensor data, are getting good value from their edge processing capabilities. Applications with very low utilization, operating on only occasional batches of data, are better served by consumption-based cloud pricing.

For example, hiring time on an NVIDIA GPU through IBM cloud services costs somewhere between $0.09 and $0.35 per TFLOPS per hour[13]. Adding a high end mobile chip with its own network processing unit (NPU) might increase the materials cost of a product by somewhere between $5 and $50 per TOPS.[14] This corresponds to less than $0.01 per TOPS per hour if run continuously for a year or more, but could cost over $1 per TOPS per hour if used less than 1% of the time for under a year.

Of course, this is not a like-for-like comparison. Floating point operations on a GPU are only comparable to 8-bit operations on an NPU if the application can tolerate quantization. The cost of hiring AI cloud time includes all of the maintenance and management required to provide that computing as a service. Most of a mobile ASIC is not the NPU, and the cost of including it in a product is not just its material cost. But the key point of the comparison is that it can be swung dramatically in either direction by utilization.

Power costs make cloud cheaper for very complex tasks

For very large tasks (100 TOPS or more) the cost of compute is still dominant, but the cost of powering that compute becomes increasingly significant. The scalability of cloud solutions makes them almost certainly the most cost-effective option for computationally intensive tasks. For example, the cost of powering 100s of TOPS of AI compute in a wireless device might be prohibitive, even if that wireless device is as large as a car.

2.4 Edge AI introduces new challenges as well as new opportunities

While edge processing is helping us to circumvent some of the obstacles to delivering intelligent features, it also presents some new challenges of its own – both in terms of product development and service management.

Embedding intelligence in edge devices presents challenges for product designers

Development and maintenance of AI software are made more complex by targeting resource constrained platforms with heterogeneous architectures. Cloud AI processing benefits from being able to run inference on the same hardware as training, which minimizes the development effort required to take a novel neural network and create an application from it. Edge processing, however, requires redeployment and optimization for the edge hardware, probably involving quantization and possibly placing constraints upstream in the workflow on the development of the network itself.

The holy grail of Edge AI software is a turnkey toolchain for converting from the high-level frameworks in which networks are developed and trained (such as TensorFlow or PyTorch) to an optimized deployment for a particular edge device. Each vendor of edge AI solutions is working towards ensuring that this toolchain exists for their platform[15], but this has so far resulted in a fragmented software ecosystem.

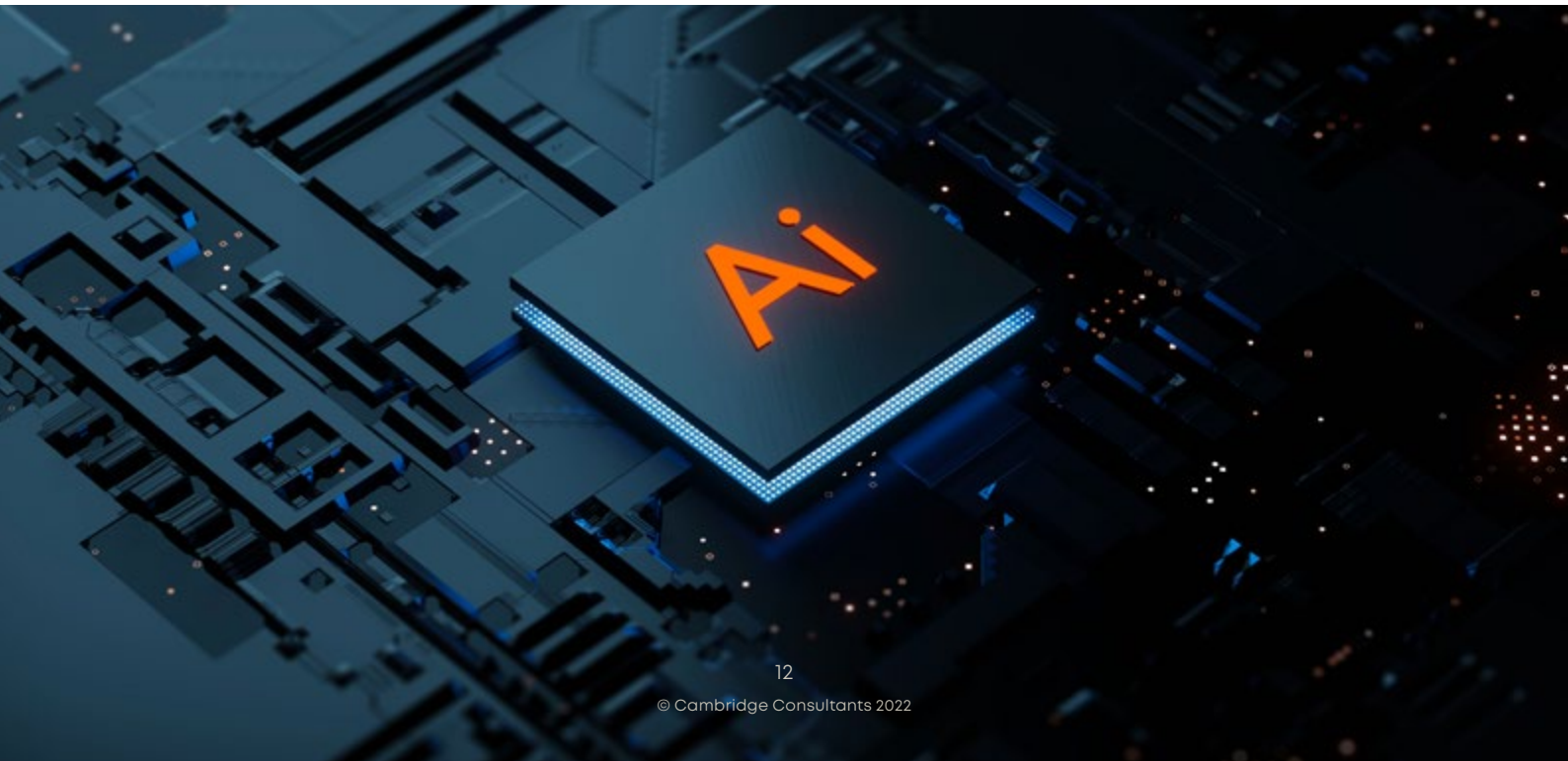
Future-proofing an edge device is harder if that device must contain its own AI processing capabilities. A minimal edge device backed by a cloud compute infrastructure could be both the quick route to market and the long-term roadmap for an AI-enabled product. An edge device, however, must be provisioned with sufficient headroom in its processing capacity to remain relevant throughout its lifetime.

Migrating intelligence away from the cloud presents challenges for service providers

Management and orchestration of AI-enabled services could become more complex as intelligence is distributed out towards the edges of the system. Cloud AI services benefit from being able to collocate their applications, bringing data back to a central point for processing and storage. This shields the end user from the complexities of managing, maintaining and updating those applications and the hardware they run on. If the user is to be similarly shielded in an edge solution, then complex systems will be required for securely deploying and managing applications across multiple edge devices, as well as monitoring, repairing or replacing those devices in the event of a fault.

Processing data at the edge as it is being acquired and then throwing it away requires a service provider to know what insights they hope to extract from that data in advance of acquiring it. This represents a paradigm shift away from big data analytics, where data can be revisited multiple times to extract new insights in light of new information or with new mining techniques.

"A minimal edge device backed by a cloud compute infrastructure could be both the quick route to market and the long-term roadmap or an AI-enabled product."



3 Low power silicon is putting edge intelligence into the hands of more businesses

3.1 Silicon technology improvements will drive more intelligence for a given power budget

Steady advances across a variety of areas within silicon technology have recently converged to begin enabling impactful intelligence in low cost, low power devices. Increasingly sophisticated power optimization techniques in the field of electronic design automation have been able to squeeze maximal advantage out of the new ultra-low leakage processes that have been yielded by advances in semiconductor fabrication. One example of this is the emergence of chiplets, which separate the analogue and digital bits of the board to achieve cost-effective chip integration. Digital silicon designers are tailoring neural network processing architectures to optimally fit AI tasks, often associated with the term 'TinyAI', while machine learning developers are continually innovating new ways to shrink processing loads without compromising on accuracy.

As silicon technologies continue to advance, the type of compute that can be placed on small edge devices, which operate at a low power, will become increasingly sophisticated. Small devices, costing in the single figure dollar range, will be able to execute AI algorithms previously only accessible in the cloud or on larger more expensive devices. This type of processing will unlock a range of opportunities by increasing the level of intelligence that can be applied to devices with extremely restricted power supplies. This will open a plethora of opportunities for device manufacturers to create value for customers.

"Small devices, costing in the single figure dollar range, will be able to execute AI algorithms previously only accessible in the cloud or on larger more expensive devices."

Case Study 1 – Ultra-low power keyword recognition

Cambridge Consultants has demonstrated how the power consumption of an Arm Cortex M-series microcontroller (MCU) based keyword detection (KWD) can be reduced by up to 99% in typical conditions.

By triaging the incoming audio via voice activity detection (VAD), developed within our Sapphyre™ custom core and accelerator framework, we've enabled the design to be compartmentalized into power islands.

This means the more power-hungry KWD, which includes an algorithm trained on the 'Google Speech Commands' dataset, is inactive when it's not needed. Assuming the VAD were to trigger once every two minutes, this would enable the power of the KWD to be reduced by 99%.

The Sapphyre VAD requires as low as 20µW, consuming just a 50th of the power consumed by modern hearing aids. Based on triggering once every two minutes, the system could operate for up to a year on a single coin cell battery. The same system without the VAD would run out of battery a matter of days.

This opens the door to 'always on', powering up AI and more power-hungry applications only when they are absolutely needed. It is likely to bring more intelligent audio, imaging and sensor-based technology for markets as diverse as health, consumer and automotive.

3.2 Low power intelligence is enabling new products, services and business models

Low power and low cost intelligence will enable businesses to adapt more quickly to the increasing computing needs of the world, driven by factors such as the growth of IoT devices and the emergence of 5G networks. This will then allow business to keep their innovative spirit even when digital infrastructures are not yet established, which is the case in many countries.[16] Furthermore, different industry verticals will be able to benefit by adopting low power intelligence in their services and products.

Healthcare devices such as smart implants for continuous monitoring of a medical condition can provide a more personalized treatment for the patient. An important aspect of smart implants, such as neural spinal stimulators, is power, meaning that prior to embedding any intelligence in the device the power costs need to be carefully considered. An ultra-low power edge device with sophisticated AI algorithms will prolong an implant's lifecycle before it is replaced.

Consumer wearables such as sport shoes will benefit from low power Edge AI as it will enable offline data collection and analysis during physical training. For example, Edge AI will be able to coach individuals on how to run properly as well as to adapt to the user's needs. This degree of personalization will be possible without the need to be constantly connected to the cloud and will help reduce the ongoing data transmission costs.[17]

Case Study 2 – Ultra-low power consumption of neural networks

We investigated what could be achieved in terms of power consumption and accuracy levels in common problems such as image classification where neural networks are being used. We used the CIFAR-10 dataset that consists of 32x32 pixel color images where each image is to be identified as one of 10 different image categories. As part of the case study, two neural network architectures – the VGG network and the CIFAR-10 tutorial network were taken and deployed on our Sapphyre platform.

Using cutting-edge AI algorithm techniques allowed us to trade-off network accuracy with the required energy for running the model. By leveraging the XNOR-net algorithm for the neural network layers running in Sapphyre, we achieved a sliding scale of power consumption and accuracy.

As a result, we concluded that it is possible to run 25 million classifications from a single coin cell battery since the energy consumption was just 86uJ. Furthermore, the same calculation for the 7J of energy of the VGG reference network yielded more than 300 classifications. This shows the potential to run ultra-low power intelligent applications, especially at the edge where power is scarce.

Case Study 3 – Porting AI from the cloud to a microcontroller

Cambridge Consultants is currently porting a medical diagnostic neural network away from a cloud-powered mobile phone application and onto a microcontroller subsystem.

This could allow the diagnostic application to be embedded into a local device, which could include the microscope used to view patient samples, eliminating the need for a smartphone or network connection.

The design takes advantage of two recently available pieces of Arm IP aimed at running neural networks in low power devices – the Cortex-M55 CPU and the Ethos-U55 microNPU (Neural Processing Unit).

This has the potential to broaden the benefits of AI to environments that can't rely on connectivity, addressing practical, high-impact challenges in areas that will improve lives at scale.

4 To stay ahead of the competition, modern businesses must embrace edge AI as part of their strategy

Edge AI is gaining momentum, fueled by the increasing need for lower latency, increased privacy and more tailored personalization demanded by many novel applications. Traditional reliance on the cloud is not enough to address safety critical use cases which require a real-time response. Intelligent edge devices have a role to play in serving the next generation of low power applications with a higher degree of convenience. Edge AI will also allow companies to customize their solution to fit their needs in a way they see fit.

Whilst cloud computing will continue to grow, Edge AI will become increasingly crucial in the new digital world where local processing will provide better efficiency and save data transmission costs. Overcoming the challenges of real-time learning at the edge will enable even more advanced Edge AI applications.[18]

The barriers to embedding intelligence at the edge device are lowering and advances in digital signal processing technology will open the door for companies to create new business models that will disrupt their markets. Understanding the opportunities and the trade-offs is the first step in enabling Edge AI to transform your business.

To discuss how Edge AI will impact your industry, please contact:

Michal Gabrielczyk, Head of Edge AI
michal.gabrielczyk@cambridgeconsultants.com

References

1 https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020

2 https://blog.bosch-si.com/bosch-iot-suite/cloud-and-edge-computing-for-iot-a-short-history/#:~:text=The%20origin%20of%20edge%20computing,such%20as%20images%20and%20videos.

3 https://www.wearable-technologies.com/2019/11/google-and-care-ai-team-up-for-autonomous-monitoring-of-healthcare-facilities-using-ai/

4 https://docs.nvidia.com/clara/tlt-mi/clara-train-sdk-v3.0/nvmidl/additional_features/federated_learning.html

5 https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/

6 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

7 https://www.idc.com/getdoc.jsp?containerId=prUS45124019

8 https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g#:~:text=The%20worldwide%20public%20cloud%20services,%2C%20according%20to%20Gartner%2C%20Inc.&text=Through%202022%2C%20Gartner%20projects%20the,growth%20of%20overall%20IT%20services.%E2%80%9D

9 Typically minimum 60fps (lags under 20ms) to avoid noticeable artefacts in lightweight applications, or 120fps (lags under 10ms) to avoid simulation sickness in highly immersive applications.

10 Try it for yourself! Open a terminal or command prompt and enter 'ping cambridgeconsultants.com' to see the round-trip time.

11 https://www.techradar.com/uk/news/federated-learning-delivers-ai-to-hospitals

12 https://www.androidauthority.com/qualcomm-snapdragon-865-vs-kirin-990-vs-exynos-990-1060885/

13 Based on $1000 for 100 hours on a NVIDIA V100 or 333 hours on a NVIDIA K80, as advertised at https://cloud.ibm.com/catalog/services/machine-learning

14 Based on $81 for a 15 TOPS Snapdragon 865 (https://www.techinsights.com/blog/samsung-galaxy-s20-teardown-analysis) or $70 for a 2 TOPS Exynos 9820 (https://www.techinsights.com/blog/samsung-galaxy-s10-teardown)

15 For example, Qualcomm's Snapdragon Neural Processing Engine SDK (https://developer.qualcomm.com/docs/snpe/overview.html) or Arm's Neural Network SDK (https://www.arm.com/products/silicon-ip-cpu/ethos/arm-nn).

16 https://www.forbes.com/sites/cognitiveworld/2019/07/10/how-artificial-intelligence-is-transforming-business-models/#63db25f26488

17 https://analyticsindiamag.com/sneakers-with-a-pinch-of-ai/

18 https://enterprisersproject.com/article/2020/5/edge-and-ai-7-things-know

Authors

James Peet, Principal Engineer, Digital Design
Ivan Petrov, Consultant, Technology Strategy
Michal Gabrielczyk, Head of Edge AI
Oliver Matthews, Senior Consultant, Technology Strategy

Contributors

Joe Corrigan, Head of Technology, Global Medical Technology
Lucy Archer, Head of Technology, Wireless and Digital Services
Mark Scoones, Senior Consultant, Digital Design
Martin Cookson, Head of Digital Services
Sam Pumphrey, Head of Digital Security
Tim Ensor, Director of AI

About Cambridge Consultants

Cambridge Consultants has an exceptional combination of people, processes, facilities and track record. Brought together, this enables innovative product and services development and insightful technology consulting. We work with companies globally to help them manage the business impact of the changing technology landscape.

We're not content to deliver business strategy based on target specifications, published reports or hype. We pride ourselves on creating real value for clients by combining commercial insight with engineering rigor. We work with some of the world's largest blue-chip companies as well as with innovative start-ups that want to change the status quo fast.

With a team of around 800 staff in Cambridge (UK), Boston, San Francisco and Seattle (USA), Singapore and Tokyo, we have all the in-house skills needed to help you – from creating innovative concepts right the way through to taking your product into manufacturing. Most of our projects deliver prototype hardware or software and trial production batches. Equally, our technology strategy consultants can help you to optimize your product portfolio and technology roadmap, investigate new opportunities or refine your operations.

For more information, or to discuss your requirements, please contact:

Michal Gabrielczyk, Head of Edge AI
michal.gabrielczyk@cambridgeconsultants.com

cambridge consultants
Part of Capgemini Invent

UK — USA — SINGAPORE — JAPAN

www.cambridgeconsultants.com

Lit-WHI-048_v0.4'2022_TheFutureofAiisattheedge_Whitepaper'