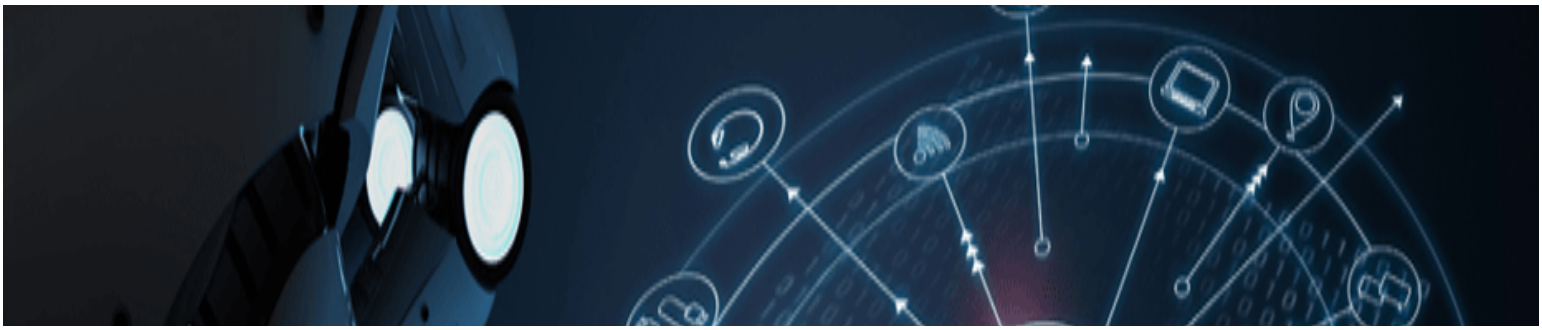
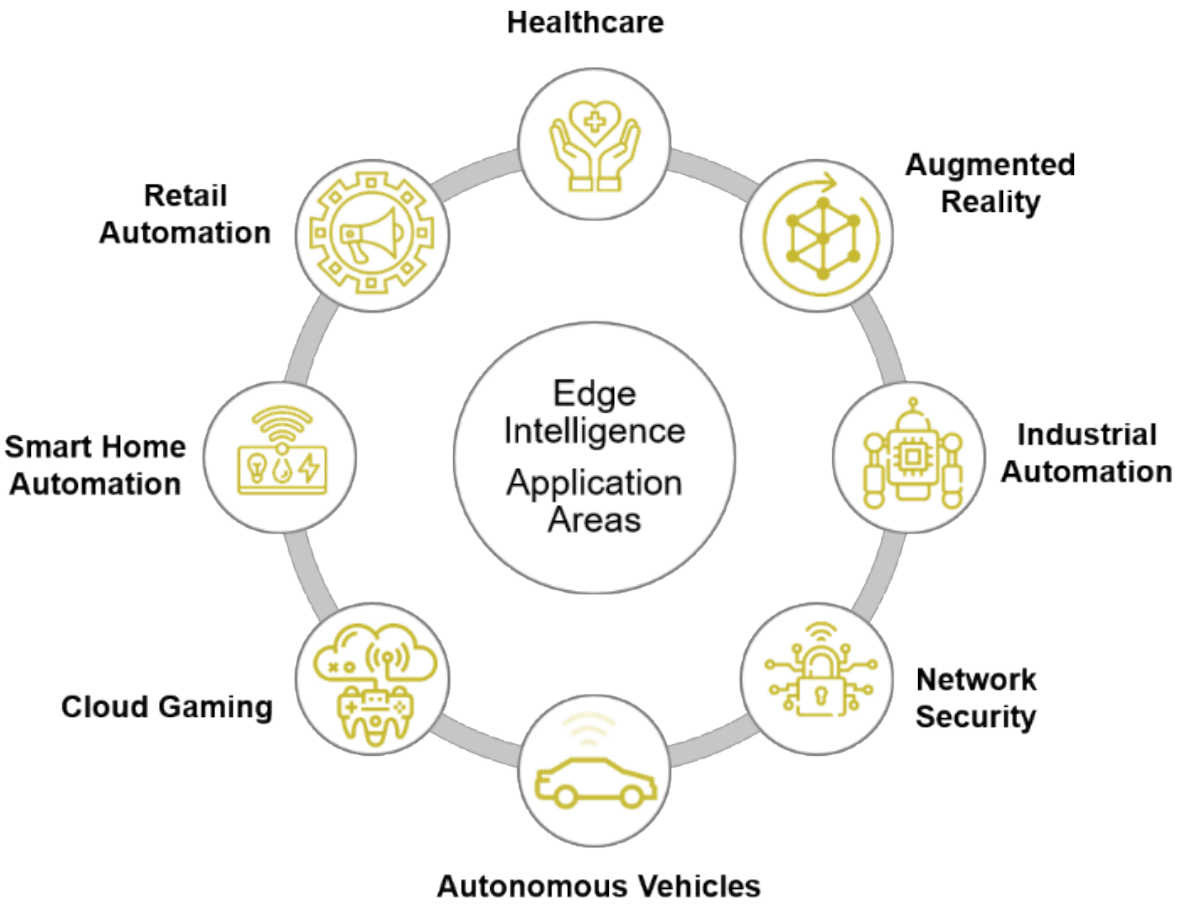


ML Goes Tiny: The Power of Edge Intelligence and Tiny ML

Cindy Waeltermann (C)
Contingent Worker

Summary

The



concept of edge intelligence has been a major buzzword in the tech industry for a few years. Edge intelligence utilizes edge computing. Edge computing decentralizes the processing of data so that it is not reliant on cloud or centralized storage, providing a much faster response time. Edge intelligence is the utilization of Artificial Intelligence (AI) technology on edge devices.

Most of the buzz surrounding edge intelligence involves the potential for IoT devices. Although IoT seems to be drawing all the hype, there are several other applications where edge intelligence is beneficial.

What is edge intelligence?

In simple terms, edge intelligence is the confluence of edge computing and artificial intelligence.

The concept of edge intelligence is based on the idea that data processing and analysis should be performed at the edge of the network, rather than in the central data centers or cloud. This approach brings significant benefits including increased performance, reduced latency, and enhanced privacy and security. Customer demands for personalized, local experiences that exploit data not easily transported to and processed by core infrastructure are also driving demand for edge intelligence capabilities.

Edge computing has been around for years. It's an existing technology that brings data processing closer to the location where data is collected. The traditional approach of sending collected data to a central data center or cloud, waiting for it to process, and sending it back to the source is no longer feasible. Autonomous vehicles, for example, need real-time information to prevent vehicle accidents. Devices such as these require immediate response. Latency is not an option.

The origin of edge computing can be traced back to the 1990's, when Akamai launched its Content Delivery Ncetwork (CDN). The idea back then was to introduce nodes at locations geographically closer to the end user for the

delivery of cached content, such as images and videos.

In 1997, in their work "Agile application-aware adaptation for mobility," Nobel et al. demonstrated how different types of applications (web browsers, video, and speech recognition) running on resource-constrained mobile devices can offload certain tasks to powerful servers (surrogates). The goal was to relieve the load on the computing resources. And, as proposed in a later work, to improve the battery life – of mobile devices. Today, for example, speech-recognition services from Google, Apple, and Amazon work in a similar way.

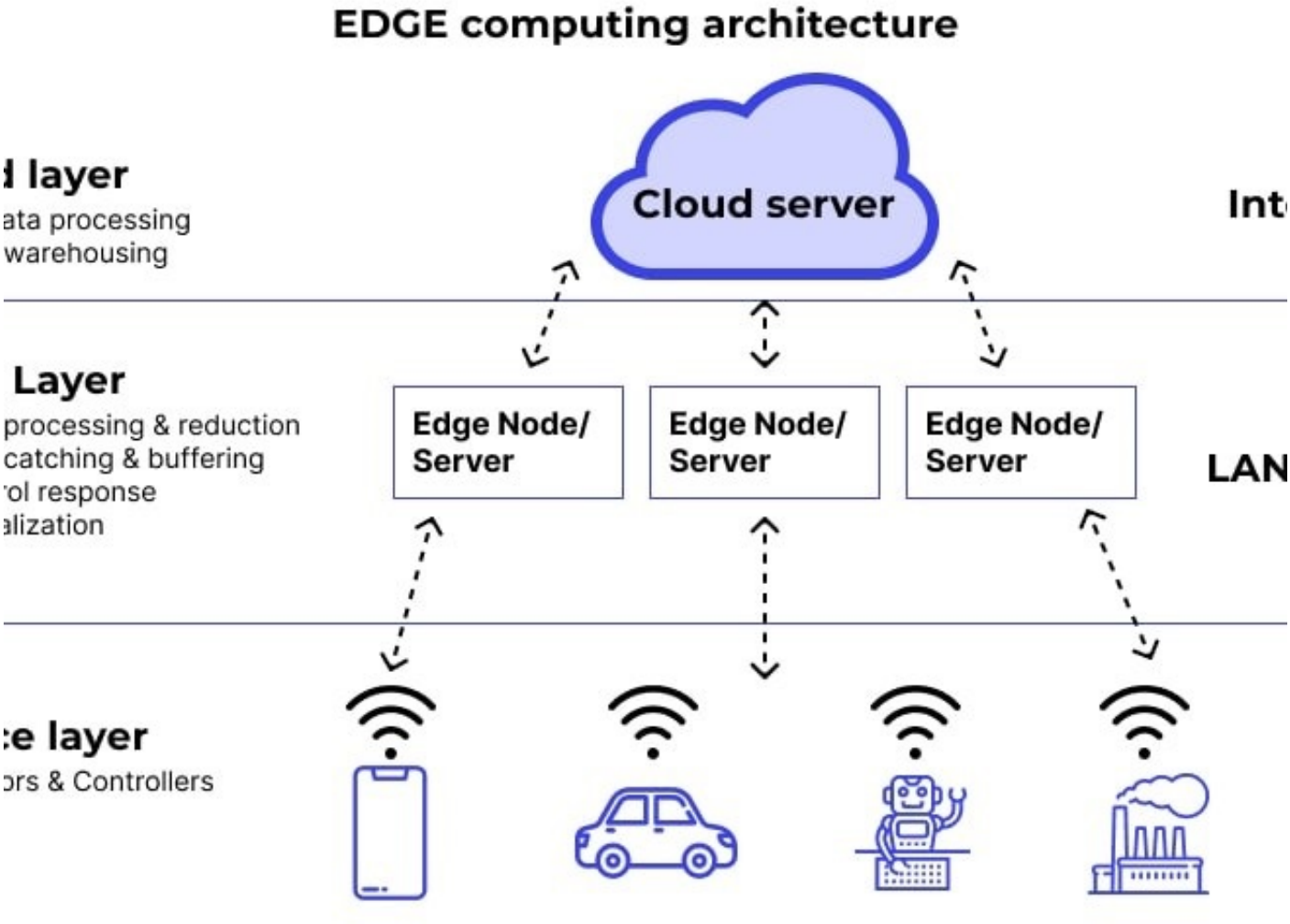
Today's edge intelligence is the deployment of AI applications in devices throughout the physical world. Since the internet has global reach, the edge of the network can connote any location. It can be a retail store, factory, hospital, or devices all around us, like traffic lights, autonomous machines and phones.

Advances in edge AI have opened opportunities for machines and devices, wherever they may be, to operate with the "intelligence" of Machine Learning. AI-enabled smart applications learn to perform similar tasks under different circumstances, much like real life.

According to Tiffany Yeung at NVIDIA, there are three reasons why edge intelligence has recently become popular:

1. Maturation of neural networks: Neural networks and related AI infrastructure have finally developed to the point of allowing for generalized machine learning. Organizations are learning how to successfully train AI models and deploy them in production at the edge.
2. Advances in compute infrastructure: Powerful distributed computational power is required to run AI at the edge. Recent advances in highly parallel GPUs have been adapted to execute neural networks.
3. Adoption of IoT devices: The widespread adoption of the Internet of Things has fueled the explosion of big data. With the sudden ability to collect data in every aspect of a business — from industrial sensors, smart cameras, robots and more — we now have the data and devices necessary to deploy

AI models at the edge. Moreover, 5G is providing IoT a boost with faster, more stable and secure connectivity.



Why is edge intelligence beneficial?

Forrester defines edge intelligence as “capabilities that help capture data, embed inferencing, and connect insight within a real-time network of application, device, and communication ecosystems.” Edge intelligence includes streaming analytics, edge machine learning, and real-time data management on intelligent devices and edge servers.

The following are examples of the benefits of edge intelligence:

- Retailers can integrate manufacturing, supply chain, store, customer, time, and environment data for in-moment personalized customer and employee experience.

- Physicians can operate side by side virtually, with systems that monitor patient status, doctor and nurse movement, and past outcomes to guide and coordinate remote surgery.
- Public transportation firms can identify traffic congestion and coordinate emergency response by rerouting traffic using connected vehicle, highway, and environment insight.

Edge intelligence

Mainstream: 2024 to 202

Technology adoption and demand

y-four percent of telecommunications decision-makers say their firm is implementing or expandin
e analytics capabilities in 2022. These firms are looking to edge computing capabilities to: 1) prov
n with real-time customer insights from localized data; 2) empower employees to drive higher and
er values; and 3) provide flexibility to handle present and future AI demands.

ions embracing edge intelligence	How the technology helps
e development, industrial IoT software forms, businesswide software-defined vorks, AI platforms, streaming data, and lytics platforms	Key edge intelligence use cases help firms address a range of key initiatives, including: • Industrial automation • Patient monitoring • Content delivery • Smart buildings and cities • Cybersecurity and monitoring

mple innovations

veva provides supply-chain and logistics platforms for intelligent scheduling and coordination of ipping for the oil and gas industry.

.M uses a distributed flight operations platform to manage airplane fleets and partners for aintenance, flight plans and schedules, products, services, and passenger scheduling globally.

4,049 telecommunications decision-makers

e: Forrester's Networks And Telecom Survey, 2022

: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

By eliminating the need to transfer large amounts of data to a central location, edge intelligence enables devices to operate independently, providing greater speed, reduced latency, increased security, lower bandwidth usage, and

increased responsiveness.

Greater speed and reduced latency

While a centralized location typically struggles to receive a barrage of information threatening to overload it and impact its performance, edge computing decentralizes data processing in order to lower data transport time and increase availability. It also reduces overall network traffic which, in turn, improves performance of other applications and software.

Increased Security

Because the data is decentralized (processed locally) in edge computing, it is distributed among the devices where it is produced. This makes it difficult for hackers to compromise data within a single attack. In the event of a cyberattack, affected areas of the network can be contained so only one device is compromised.

Lower Bandwidth

With edge intelligence, there is a significant reduction in the volume of bandwidth required to process information at the edge. In addition, since edge AI processes data locally, less data is sent to the cloud through the internet, thereby saving bandwidth.

Increased Responsiveness

Although the process of sending data to cloud-based data centers can be done within a few seconds, edge intelligence further reduces the amount of time it takes smart devices to respond to requests by generating and processing the data within the device. With a high response rate, technologies like autonomous vehicles, robots and other intelligent devices can provide instant feedback to automatic and manual requests.

What is TinyML?

Basically, TinyML is machine learning code that is refined and reduced in size to the point where it can run locally on an IoT device (on edge).

Machine learning models take up an inordinate amount of compute space. In order to use machine learning on IoT devices, the code must be compressed. This is accomplished through the use of TinyML.

Tiny machine learning (TinyML), as defined by the TinyML Foundation, is "a fast growing field of machine learning technologies and applications including hardware, algorithms and software capable of performing on-device sensor data analytics at extremely low power typically in the mW range and below, that enables a variety of always-on use-cases and targeting battery operated devices."

Size reduction in TinyML focuses on making models simpler by reducing model parameters, thereby reducing RAM requirements in execution, and storage requirements in memory. These models can also be run using cloud and edge computing, meaning that the personal data used to train such models won't need to be stored in a server.

There are several frameworks available for shrinking a deep learning model into TinyML to fit an embedded device, including:

- AIMET (Qualcomm)
- TensorFlow Lite (Google)
- CoreML (Apple)
- PyTorch (Facebook)

For a brief explanation of how ML code is compressed, refer to Model Compression Techniques for Edge AI.

Why is TinyML beneficial?

So why is TinyML beneficial? The answer is pretty simple – when used in conjunction with edge intelligence, TinyML reduces the computational requirements for machine learning models, making them suitable for deployment on devices with limited resources. It enables offline processing and reduces the dependence on internet connectivity. The smaller size of the models also results in much less power consumption.

Business Impact

Edge intelligence is poised to enable billions of new IoT endpoints and real-time local artificial intelligence/machine learning (AI/ML) for autonomous systems.

Edge intelligence use-case scenarios

- Autonomous Vehicles – The decision to stop for a pedestrian crossing in front of an autonomous vehicle (AV) must be made immediately. Relying on a remote server to handle this decision is not reasonable. Vehicles that utilize edge intelligence can interact more efficiently because they can communicate with each other first as opposed to sending data on accidents, weather conditions, traffic, or detours to a remote server first.
- Healthcare Devices -- Health monitors and other wearable healthcare devices (such as an insulin pump) can keep an eye on chronic conditions for patients. It can save lives by instantly alerting caregivers when help is required. If these devices rely on transmitting data to the cloud before making decisions, the results could be fatal.
- Security Solutions -- Because it's necessary to respond to threats within seconds, security surveillance systems can identify potential threats and alert users to unusual activity in real-time.
- Retail Advertising -- Targeted ads and information for retail organizations are based on key parameters, such as demographic information, set on field devices. In this use case, edge computing can help protect user

privacy. It can encrypt the data and keep the source rather than sending unprotected information to the cloud.

Limits of edge intelligence

- Edge intelligence can be limited by the processing power and memory of the edge device.
- The limited computing resources on edge devices can also limit the types of algorithms and models that can be run on them.

Drawbacks of edge intelligence

- The deployment and maintenance of edge intelligence systems can be challenging.

Startup activity

Many new startups are focusing on edge intelligence with TinyML, including:

- Useful Sensors – Started by former Google engineer Pete Warden. Useful Sensors has created low-cost, easy-to-integrate hardware modules that bring ML capabilities like gesture recognition, presence detection, and voice interfaces to TVs, laptops, and appliances while preserving users' privacy.
- Edge Impulse – Focuses on building, deploying, and scaling embedded ML applications.
- FogHorn Systems (recently acquired by Johnson Controls)

Top edge computing platforms

- Amazon Web Services (AWS) -- helps in moving components such as storage, data analysis, and data processing closer to endpoints.
- Microsoft Azure -- delivers managed services, taking Azure's competencies, such as computing, intelligence, and storage, to the edge. Azure edge is suitable for machine learning, edge-to-cloud data

transfer, and IoT solutions.

- Dell Technologies – Provides an array of storage, computing, and networking technologies deployed on the edge.
- Cisco Edge Intelligence -- Programmable application to extract, transform, govern. and deliver data from IoT edge devices to applications. Scalable solution simplifying edge to multi-cloud data flows.

Competitor activity

Avalara

Avalara

addressed edge computing as an innovation trend in a blog in February 2022. They also offer their integration capabilities with edge computing, as shown below. This is edge computing, however, not edge intelligence.

Sovos

There is no mention of edge intelligence in Sovos' blogs or on their website.

Thomson Reuters

Thomson Reuters has launched the first edge computing tax engine. This is also not edge intelligence, but edge computing.

- Thomson Reuters Launches First Edge Computing Tax Engine on the Market

Stripe (TaxJar)

There are multiple mentions of edge computing on TaxJar's website and on TaxJar blogs, but nothing about any specific offering.

Conclusion

While the technology offered in edge intelligence is driving massive innovation in IoT devices, edge intelligence can also help technology like control system engineering, where equipment is monitored for health to prevent manufacturing downtime. For critical industries such as energy, in which supply issues can threaten the health and welfare of the general population, intelligent forecasting is key.

The biggest benefit of edge intelligence is for IoT devices, and that is exactly where all the buzz has been. When considering things like insulin pumps that rely on real-time data where someone's health and well-being is in the balance, or in an autonomous vehicle, the interest becomes clear.

While overall this is a fantastic innovation, Vertex could possibly benefit from this technology as its edge computing initiatives and insight applications mature. For example, if an edge tax calculation engine could feed real-time sales transaction to an ML model trained to detect buying trends, a store could offer real-time sales. Like an old-fashioned blue light special with a brain, perhaps a store could combine inventory data with real-time customer preference data to accelerate sales of overstock items.

References

- Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence
- Akamai
- NVIDIA
- TinyML Foundation
- AIMET

- TensorFlow Lite
- CoreML
- PyTorch
- Useful Sensors
- Edge Impulse
- FogHorn Systems (recently acquired by Johnson Controls)
- Model Compression Techniques for Edge AI.

Meet the Innovation Team Behind VX3