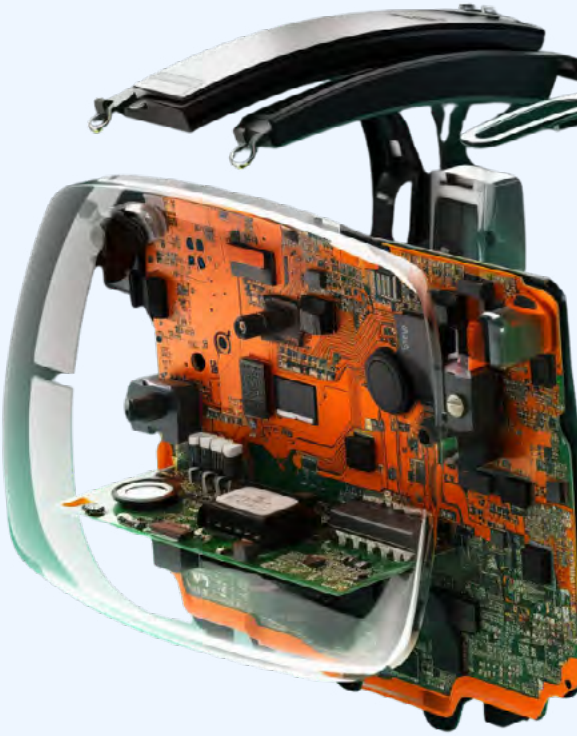
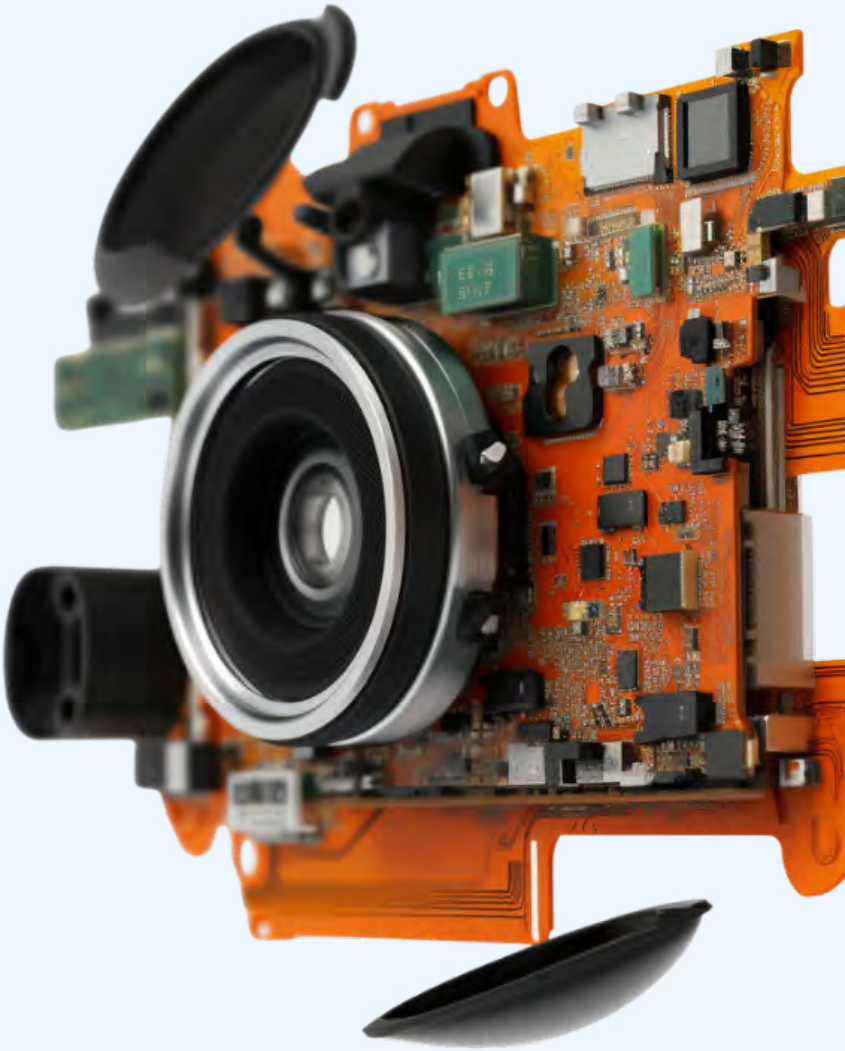
WEVOLVER

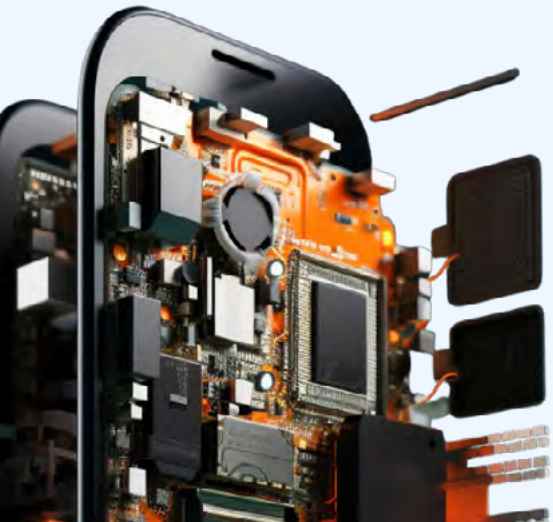
TINY ML

2024 STATE OF EDGE AI REPORT

Exploring the dynamic world of Edge AI applications across industries







About the Contributors

How this report came together

This technology report came to fruition thanks to the collaborative efforts of multiple contributors. Samir Jaber, the editor-in-chief, alongside John Soldatos and Ravi Rao, the supporting authors, have played integral roles in researching, developing, and editing the report content. The insights provided by our sponsors and contributors have significantly enriched the content. Special acknowledgment goes to Jessica Miley, the content director at Wevolver, for her consistent support and leadership throughout the process. Each contributor's dedication and input are deeply valued, highlighting the collective effort that has gone into creating this report.

Samir Jaber, Editor-in-Chief
Leipzig, Germany

Samir Jaber is an editor, writer, and industry expert on technology, science, and engineering topics. He is an online content specialist with an academic background in mechanical engineering, nanotechnology, and scientific research. Samir has comprehensive experience working with major engineering and technology companies as a writer, editor, content manager, and digital marketing consultant. He is a featured author in 30+ industrial magazines with a focus on Artificial Intelligence (AI), the Internet of Things (IoT), 3D printing, Autonomous Vehicles (AV), nanotechnology, materials science, and sustainability. Samir is also an award-winning engineering researcher in the fields of nanofabrication and microfluidics.

John Soldatos, Co-author
Athens, Greece

Honorary Research Fellow at the University of Glasgow

John Soldatos holds a Ph.D. in Electrical & Computer Engineering from the National Technical University of Athens (2000) and is currently an Honorary Research Fellow at the University of Glasgow, UK (2014-present). He was Associate Professor and Head of the Internet of Things (IoT) Group at the Athens Information Technology (AIT), Greece (2006–2019), and Adjunct Professor at the Carnegie Mellon University, Pittsburgh, PA (2007–2010). He has significant experience working closely with large multi-national industries (e.g., IBM, INTRACOM, INTRASOFT International) as an R&D consultant and delivery specialist while being a scientific advisor to various high-tech startup enterprises. Dr. Soldatos is an expert in Internet-of-Things (IoT) and Artificial Intelligence (AI) technologies and applications, including IoT/AI applications in smart cities, finance (Finance 4.0), and industry (Industry 4.0).

Ravi Rao, Co-author
Ahmedabad, India

Ravi is a software engineer at one of the world's largest IT consulting firms. He works on digital transformation projects that help streamline operations for customers. Ravi's education in Power Electronics Engineering and his professional experience have equipped him with expertise in electrical, electronics and computer engineering. Ravi has authored two IEEE research papers: one exploring the integration of AI and IoT in irrigation systems and another assessing the implementation of a curriculum to stimulate innovation among engineering students. Additionally, Ravi has earned a patent for a project aimed at enhancing independence for individuals with mobility challenges. Outside work, Ravi is passionate about writing on technology, gadgets, and gaming. He also loves engaging in conversations about the latest advancements in technology.

Introduction

In the midst of all the technological advancements happening today, one term has taken over almost every discourse about technology: artificial intelligence (AI). Once a thrilling notion for Sci-Fi movies, AI is now an indisputable reality, and it has taken multiple industries by storm. From Large Language Models (LLMs) enabling fantastic chatbots like ChatGPT to Internet-of-Things (IoT) devices bringing about Industry 4.0, AI has found applications and use cases across almost every modern industry. While the majority of AI implementations have taken place centrally on cloud servers, there has been a strong pursuit for intelligence to be deployed locally at the edge.

Artificial intelligence at the edge is referred to as Edge AI, and it is the convergence of AI model deployment with edge computing. By bringing computational power closer to the data source, Edge AI offers real-time data processing on edge devices with reduced latency, higher bandwidth efficiency, enhanced reliability, and more robust security and privacy. This has the potential to propel AI capabilities beyond its current centralized, cloud-based structure. As the adoption of Edge AI continues to grow almost exponentially, we will witness a paradigm shift in how businesses deploy and utilize AI, especially with edge computing decentralizing data processing.

Industries that require real-time data processing capabilities are the primary industries in Edge AI's scope of impact and are expected to experience a radical transformation as a result. Embracing this dramatic shift is essential for businesses that seek to stay ahead of the competition and leverage the benefits of real-time, context-aware decision-making. But knowing that Edge AI impacts every industry differently, how can a business in a specific industry embrace this technology to ensure it keeps pace with its rapid evolution?

Accordingly, Wevolver has partnered with several major companies and thought leaders in this space to produce this holistic report on the state of Edge AI in 2024 and its applications in every relevant industry, from healthcare to industrial IoT and manufacturing, smart cities, retail, energy, agriculture, and automotive. Each industry is covered by a dedicated chapter that provides industry-specific insights, descriptions, and examples, complemented by featured sections of real-world case studies. The report also provides a peek into the exciting generative AI technology and its convergence with edge computing before delving into the challenges still hindering Edge AI today. Finally, it gives a glimpse into the future of Edge AI and how it will develop in the years to come.

This is the second in-depth report on Edge AI by Wevolver after its inaugural report in 2023, which focused primarily on the technological side of Edge AI. This report explores Edge AI's application side and its impact on the various industries mentioned above. Will Edge AI be one of the world's fundamental technologies going forward? Will it continue to reshape our industries and become ubiquitous? If so, how? Let's find the answers in the chapters below.

Samir Jaber
Editor-in-Chief

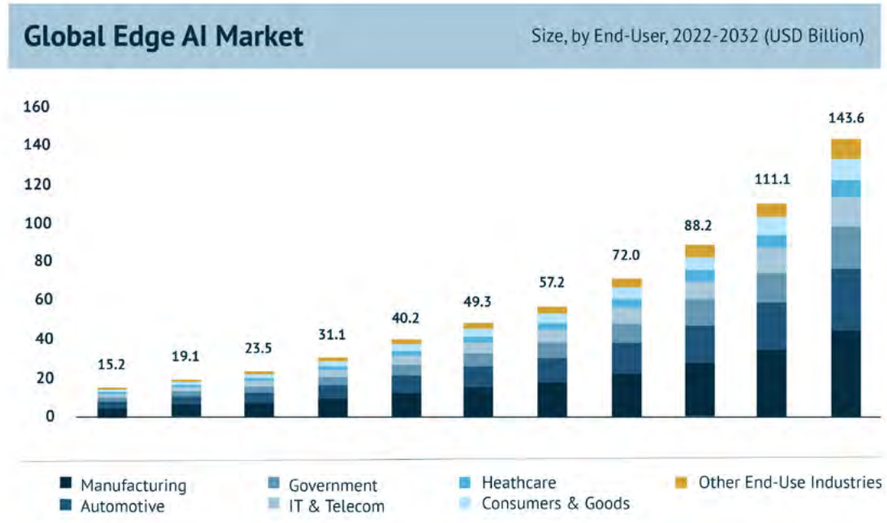
Chapter I: Edge AI Market Analysis and Trends

"Simplicity is key in the tech world: a solution is only as successful as it is widely accepted, adopted, and applied. That's why Arduino's mission is to democratize technologies like Edge AI, making it an accessible option for people with different backgrounds and in all industries to solve problems, create value, and grow."

Fabio Violante, CEO of Arduino

Edge AI's growing momentum over recent years has not been a stroke of luck or an unexpected turn of events. It's been brewing in the research spheres for quite a while now. In fact, the way AI influences technology has been a discussion since the mid-1900s, ever since Alan Turing set the standards for the "thinking machine," and Christopher Strachey wrote the first successful AI computer program, followed by Arthur Samuel of IBM pioneering machine learning. From then on, AI went on a long rollercoaster ride of rises and dips, going through hype cycles and significant interest and funding all the way down to not one but two "AI winters" before the turn of the millennium. Nonetheless, in recent years, especially going into the 2020s, AI has taken yet another leap, but this time, it has taken off.

With major developments in machine learning and the advent of technologies like cloud computing, edge computing, and fog computing, AI has significantly benefited from the data processing capabilities brought about by these technologies. Cloud computing's ability to process massive amounts of data simultaneously and edge computing's ability to process data locally and in real time are both crucial for AI's mass adoption. This helps AI make informed decisions in fractions of the time that it would take humans, thus improving processes and systems across different industries and creating new and optimal ways of working. As a result, AI has witnessed significant growth in adoption rate across various large organizations, reaching 42% in 2023, according to the IBM Global AI Adoption Index 2023, with as many as 40% of organizations actively exploring the use of AI in their business operations. AI has once again captured the spotlight — and this resurgence has brought about a new wave of possibilities and opportunities to explore.



The Edge AI Market is expected to reach USD 143.6 Billion by 2032, an exponential rise from its 2023 value of USD 19.1 Billion (Credit: market.us)

Today, a lot of data processing is being decentralized from large cloud data centers to smaller localized data centers and edge devices. This has enabled the emergence of Edge AI, which processes data at or near the source of data generation. Many organizations are deploying edge functionalities, resulting in energy-efficient, low-latency applications with real-time performance. Edge AI offers significant data protection and security benefits, making it an attractive proposition for organizations across sectors to use edge computing features for various use cases. In the following sections, we take a look at the Edge AI market to see how it's been responding to the technology's potential, and we explore how different industries are adopting Edge AI into their workflows and systems.

Edge AI Market Landscape

For a technology that is moving quickly on the Gartner hype cycle for AI, Edge AI is constantly finding new use cases in various industries, and its adoption is expected to grow further and faster. In fact, Gartner analysts predict that edge computing technologies will gain traction and maturity in 2024, especially with the significant drop in the cost of developing and deploying edge systems thanks to technical innovation in this space. Such improvements in the technology have enabled the Edge AI market to witness remarkable growth over recent years. According to Market.US research, the global Edge AI market is expected to surpass the USD 140 billion mark by 2032, a considerable rise from just over USD 19.1 billion in 2023. That is a compound annual growth rate (CAGR) of almost 26% across nine years.

The growth of the Edge AI market reflects the increasing integration of technology into various aspects of modern life. With the proliferation of IoT devices across industries, from manufacturing to healthcare, vast volumes of data are being generated continuously. This data holds significant potential for insights and optimization, but traditional centralized processing methods often struggle to handle real-time data processing without having to deal with latency issues.

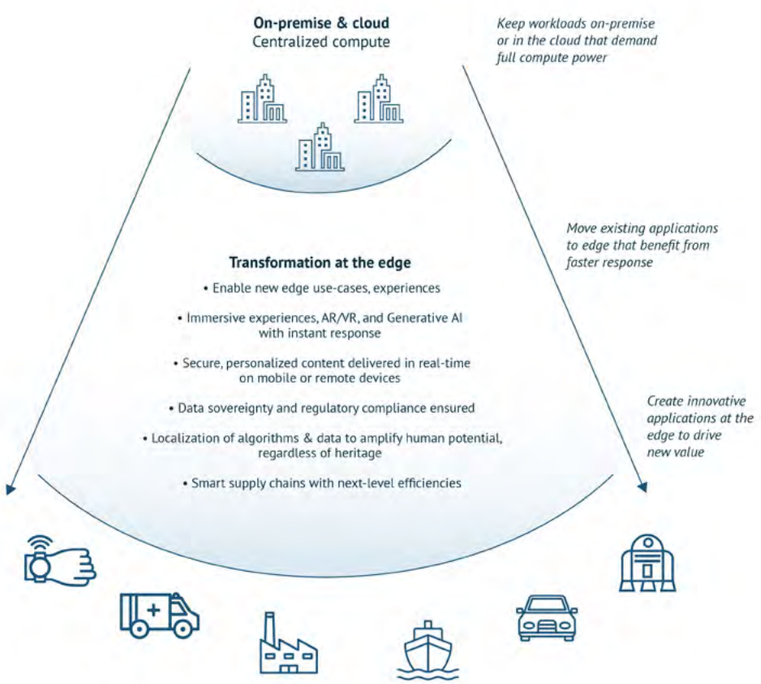
Edge AI addresses this challenge by bringing AI and machine learning algorithms closer to where the data

is generated, at the "edge" of the network. This localized approach allows for real-time processing and analysis, minimizing latency, reducing bandwidth usage, and enabling quicker decision-making. For instance, within the realm of autonomous vehicles, split-second responses to changing road conditions are crucial for safety. Edge AI enables these vehicles to process sensor data onboard, ensuring faster reaction times.

Moreover, advancements in semiconductor technology have played a crucial role in enabling more powerful and energy-efficient edge computing devices. These devices are capable of handling complex AI algorithms while remaining efficient enough to operate in resource-constrained environments, such as remote industrial sites or within wearable devices.

The rollout of 5G technology further amplifies the capabilities of Edge AI solutions. With its significantly enhanced connectivity and data transfer speeds, 5G facilitates seamless communication between edge devices and central systems, enabling faster data transmission and response times. This is particularly beneficial in scenarios such as healthcare monitoring, where timely analysis of patient data can have life-saving implications.

In essence, the growth of the Edge AI market represents a response to the evolving demands of industries and engineers in a data-driven world. It's about leveraging innovation to optimize processes, enhance efficiency, and ultimately improve the way industries operate.

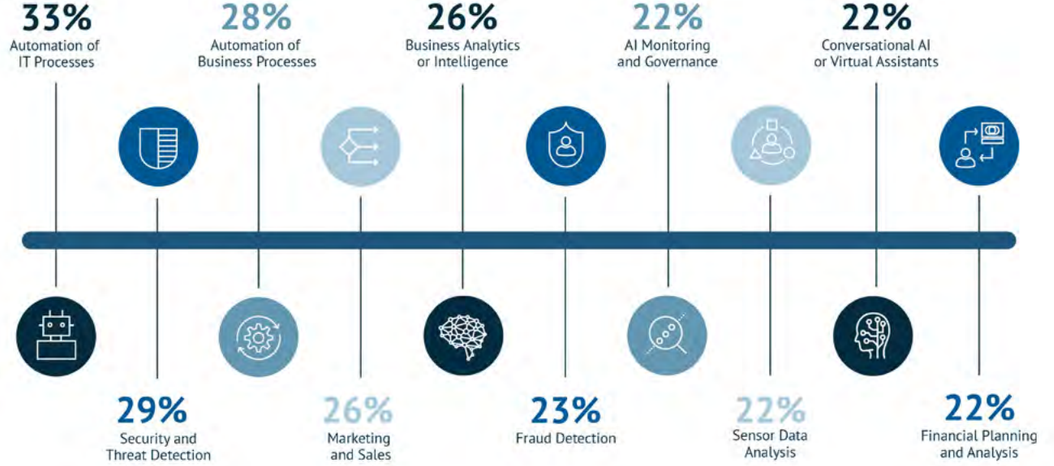


Edge moves computing closer to users and devices at the edge of the network, where it is the closest possible to data sources (Credit: Accenture)

Industry Adoption and Trends of Edge AI

"Under the radar thus far, edge is set to become a ubiquitous lever of scale and reinvention as artificial intelligence (AI)—including generative AI—driven applications become pervasive in enterprise functions and operations." This is what researchers at Accenture articulated in their 2023 study on Leading with Edge Computing. This statement sums up the current state of Edge AI in a nutshell. With edge computing permeating a lot of industries and application areas, it is reaching a level of ubiquity that renders its implementation almost a necessity. In fact, Accenture's survey uncovered that 83% of executives across multiple industries think that in order to stay competitive in the future, edge computing will be essential. And many are fearing missing out on all of Edge AI's benefits if they do not act quickly and incorporate it into their workflows, products, and services.

However, the adoption of Edge AI is not uniform. While some companies regard edge as a key differentiator to bring AI into their core business, others still struggle with fully leveraging the technology's benefits, mostly due to them considering it a standalone technology and using it in ad-hoc projects. Adopting edge computing strategically and integrating it with existing cloud strategies has shown the best outcomes: Advanced users of edge are four times more likely to achieve accelerated innovation, nine times more likely to increase efficiency, and seven times more likely to reduce costs (Accenture, 2023).



How organizations have been using AI in the past two years (Credit: IBM)

Looking at Edge AI from the industry level, we see a similar distribution, with some industries already utilizing the technology almost ubiquitously and others still exploring its potential. The manufacturing industry seems to take the lion's share of revenue with approximately 31% thanks to the integration of automation and real-time insights, capitalizing on Edge AI's benefits in defect detection, reduced latency, real-time decision-making, cost efficiency, and data security. Following it is the automotive and transportation sector, especially with the recent ACES trends (autonomous driving, connectivity, electric vehicles, and shared mobility) leading the industry, all of which require Edge AI, albeit at varying levels. Furthermore, traffic management is seeing significant developments by integrating Edge AI into its sensors, cameras, and traffic management systems (TMS). While other industries like healthcare, retail, energy, and agriculture acknowledge the potential of Edge AI, they are yet to adopt it at the same level as manufacturing and automotive.

That being said, Edge AI is seemingly showing clear patterns and trends that are influencing the future of data science and machine learning (DSML) across many industries. As Gartner outlined in 2023, one trend that is leading the way is *Edge AI as a promise of responsiveness*. Edge AI promises quicker decision-making by executing AI algorithms locally, bypassing the need for the Cloud or remote data center connections. This reduces latency and enhances system responsiveness. Converging AI and edge computing leads to more efficient and potentially energy-saving solutions. Gartner has forecasted that, by 2025, more than half of data analysis by deep neural networks will occur at the point of capture in an edge system, a significant increase from single-digit percentage points in 2021. This shows the significance of Edge AI in the years to come and how its implementation will continue to grow and penetrate various systems and workflows.

Chapter II: Healthcare and Medical Applications

Edge AI in Healthcare

In recent years, the emergence of Edge AI has played a significant role in the digital transformation of the healthcare sector. With the shift of AI capabilities closer to the source of data generation, Edge AI can enable real-time, data-driven clinical decisions, enhance accuracy, and boost privacy and data protection in various healthcare applications where sensitive data are used. There is a multitude of Edge AI use cases in medical and clinical settings, including use cases that improve existing capabilities and others that enable functionalities that were hardly possible before the advent of AI at the edge. These capabilities fulfill different objectives and serve the needs of various stakeholders, as outlined in the Edge AI use cases presented below.

Real-Time Patient Monitoring

Continuos patient monitoring is crucial to ensuring timely detection of warning signs and providing appropriate interventions. Edge AI plays a vital role in patient monitoring use cases by enabling real-time analysis of data collected from various data sources like wearable devices, sensors, and electronic health records (EHR). Based on data processing at the edge, patient monitoring systems can instantly identify anomalies and trigger appropriate actions, such as alerting healthcare providers or autonomously adjusting medication dosages. Hence, real-time patient monitoring powered by Edge AI can significantly improve patient outcomes and reduce the risk of adverse events.

Edge AI is transforming the medical industry by enabling digital diagnosis and remote patient monitoring (Credit: Jenny Wang, Seeed Studio)

A Fast, Cost-Effective, and Reliable Way to Add Smart Features to Healthcare Devices

The healthcare industry has seemingly infinite vital signs and symptoms to track; finding ways to turn these signals into medically useful information can mean game-changing improvements for the industry and its patients. Edge AI offers an exciting opportunity to propel medical care and healthful living to an impactful new level. It is particularly suitable for the privacy requirements of this industry because data can be kept on the device instead of being sent to the cloud.

Among those testing the way forward is Imagimob, a company intent on making it easier to deploy machine learning on edge devices. With their collection of off-the-shelf models that can be added to healthcare and wearable products with ease and an end-to-end development platform for solving problems with custom Edge AI models, Imagimob's tools can change healthcare for the better.

The Fastest Way to Launch Smart AI Features

A major barrier to creating smarter medical devices is the typically extensive development process. With Imagimob's Ready Models, AI features can easily be deployed onto existing products without the significant time, cost, or machine learning expertise required for custom development.

Ready Models currently available for use in healthcare include audio-based models for coughing and snoring detection. And this is just the beginning — additional Ready Models under development include models based on Radar, IMU, and Capacitive Sensing, which can be used for presence detection, fall identification, and more.

Developing Sharper Models in Real Time

Imagimob's machine learning development platform, Imagimob Studio, covers the entire workflow — from data collection to quick deployment in a healthcare product. Their recently launched Graph UX interface helps produce even higher quality models by letting engineers see them working in real-time.

Take, for instance, a model that identifies coughing in a healthcare setting. In a scenario where coughs are being under-identified, an engineer can pinpoint which data the model is failing to classify and make direct improvements.

"Graph UX makes models more robust as you have greater visibility and can identify problems fast," says Alexander Samuelsson, Imagimob's CTO. "This is a great advantage for the healthcare industry, where the ability to adapt quickly to new scenarios and information can be critical."

The possibilities are limitless. Learn more about how Imagimob's Edge AI solutions are helping to build the intelligent products of the future with applications in healthcare and beyond at www.imagimob.com.

On-Device Medical Imaging and Smart Rehabilitation

Traditional medical imaging techniques require transferring large amounts of data to remote servers for processing, which creates concerns about latency and privacy. With Edge AI, medical imaging can be performed directly on the device, such as ultrasound machines and MRI (Magnetic Resonance Imaging) scanners. This ensures enhanced accuracy and reduced diagnosis time while increasing data protection. In particular, Edge AI algorithms can detect abnormalities in scans and provide immediate feedback to radiologists, enabling more efficient diagnoses. Furthermore, Edge AI addresses privacy concerns by minimizing the transmission of sensitive patient data over the network.

Technology trends like TinyML, embedded machine learning, and on-device neuromorphic computing are likely to increase the number and variety of embedded medical imaging applications in the years to come. On-device inference for medical decision-making is set to become smaller yet faster, smarter, and more privacy-friendly.

Similarly, rehabilitation processes can benefit significantly from real-time feedback in order to optimize therapy routines and improve patient outcomes. Edge AI empowers smart rehabilitation devices with on-device inference, allowing them to analyze sensor data and provide real-time feedback to patients during rehabilitation sessions. For instance, an Edge-AI-powered prosthetic limb could adjust its movements based on the patient's gait analysis toward a more natural and personalized experience. Based on Edge AI, rehabilitation devices can adapt to individual needs in real time, which enhances the effectiveness of therapies and reduces the need for constant supervision.

Clinical Trials with Real-World Data

Real-time medical adherence monitoring is essential in clinical trials to ensure accurate data collection and protocol compliance. Edge AI enables the integration of real-time monitoring devices that track medication intake, adherence to treatment plans, and patient vitals. By leveraging Edge AI algorithms, researchers can assess, in real time, whether participants are following the prescribed protocols.

Such functionalities can provide valuable insights into the effectiveness of treatments while reducing the burden of manual data collection. Most importantly, they are key to ensuring that clinical trials adhere to the prescribed protocols, which increases their credibility. Edge AI on real-world data is, therefore, a technology that will be increasingly adopted and used by Contract Research Organizations (CROs) in a variety of clinical trials worldwide.

Prediction, Detection, and Tracking of Disease Outbreaks

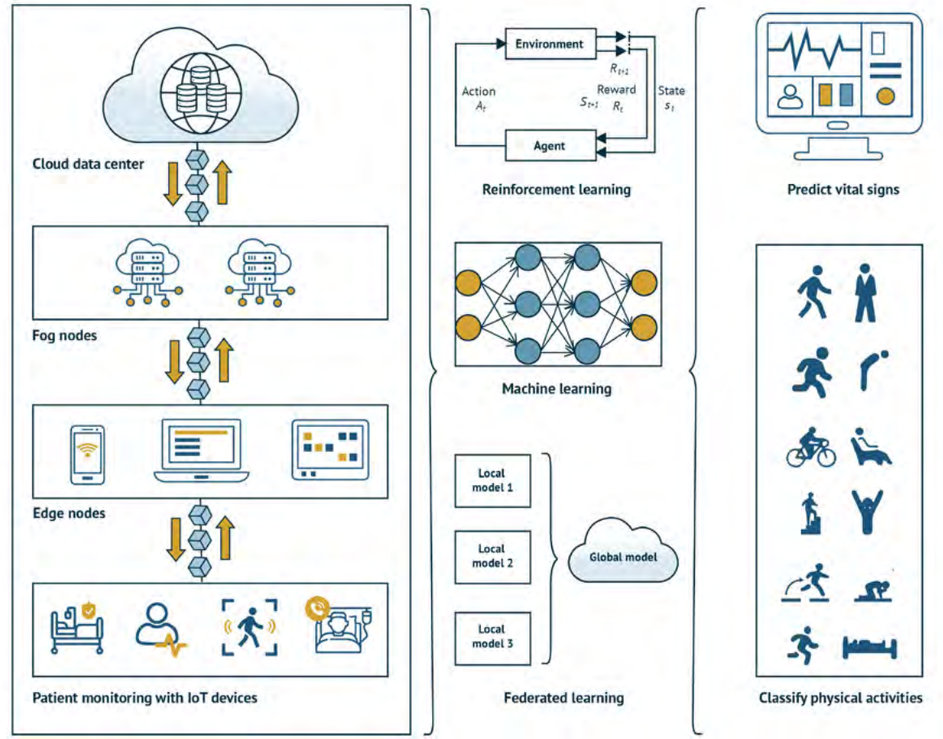
Processing local data at the edge is particularly valuable for predicting and tracking disease outbreaks. Based on the analysis of medical records, wearables data, and environmental factors in real-time at the edge, healthcare organizations can detect early signs of potential outbreaks. This is fundamental to enabling proactive measures to prevent their spread. Edge AI can also ensure privacy and data protection by minimizing the reliance on centralized data repositories, which mitigates the risk of unauthorized access to sensitive healthcare information.

Many healthcare applications can also benefit from decentralized learning models at the edge, such as federated learning. Federated learning is a technique in which global models are developed by combining locally trained models. It is already used in various disease detection and prediction use cases to enhance accuracy. The latter stems from the fact that global models turn out to be more accurate than local models, as they are, in principle, trained on more data.

In the scope of a federated learning approach, edge devices (e.g., smartphones or IoT sensors) locally train AI models using data from individual patients or individual care service providers (e.g., care centers). Encrypted model updates are then sent to a central server, where they are combined to create a global model without compromising the privacy of individual data. This approach has been successfully applied by organizations like the World Health Organization (WHO) during the COVID-19 pandemic, where local models trained on edge devices contribute to a comprehensive global model that aided in disease detection and prediction.

Decentralized learning paradigms will be increasingly adopted as the number of connected medical devices and other healthcare data sources grows rapidly. Managing huge numbers of data sources in a centralized manner increases data protection risks and creates latency and power efficiency concerns. This is why it has to be avoided whenever possible and substituted by decentralized systems powered by Edge AI.

Overall, Edge AI applications in healthcare offer numerous benefits, ranging from real-time patient monitoring and on-device medical imaging to clinical trials with real-world data and smart rehabilitation devices. Based on Edge AI paradigms, healthcare organizations can achieve real-time performance, enhanced accuracy, and improved patient outcomes. Moreover, privacy and data protection are prioritized due to the reduced transmission of sensitive information over networks, minimizing the attack surface of the healthcare application. In the coming years, Edge AI technologies will empower healthcare professionals to work faster and more effectively, delivering tangible benefits to millions of patients worldwide.



Artificial intelligence-enabled remote patient monitoring architectures
(Credit: Shaik, T. et al., WIREs)

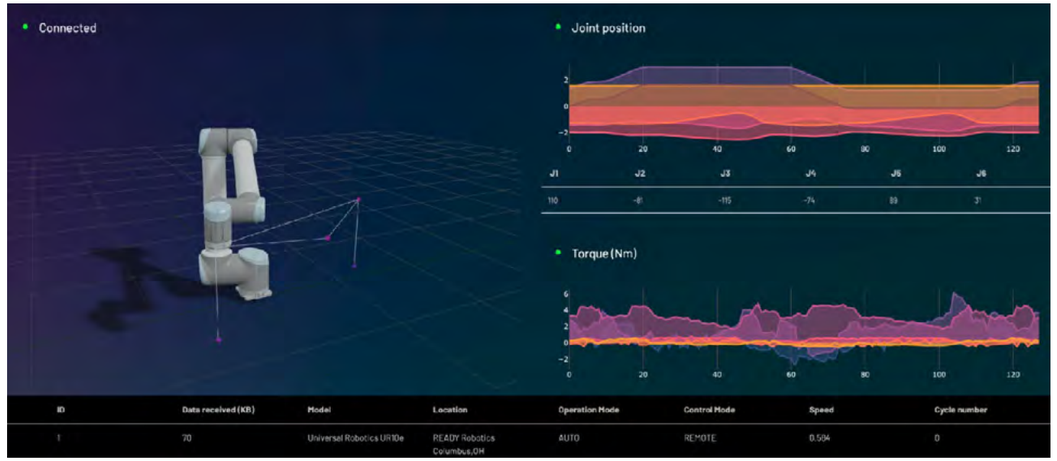
Chapter III: Industrial IoT and Manufacturing

Edge AI Enabling Industry 4.0

The advent of Industry 4.0 is gradually revolutionizing industrial production based on the digitization of physical processes and the introduction of cyber-physical production systems on the manufacturing shop floor. A significant number of Industry 4.0 applications are based on the integration of Artificial Intelligence (AI) and the Internet of Things (IoT). Edge AI is, nowadays, one of the most in-demand implementations of this AI-IoT combination, bringing AI capabilities closer to the network edge and enabling real-time data processing and analysis on edge devices.

Edge AI improves the performance, timeliness, and security of various use cases of the manufacturing sector, including predictive maintenance, real-time quality control, and supply chain optimization. These use cases deliver tangible benefits to key stakeholders such as manufacturers, maintenance teams, quality control departments, supply chain managers, plant operators, and workers.

> “Edge AI isn't just a technology; it's the driving force behind smarter, more efficient, and more responsive industrial ecosystems. By deploying AI capabilities at the network edge, we enable engineers to pioneer the forefront of innovation, leveraging real-time insights and tangible advancements to elevate industrial processes to new heights.”
>
> Richard Curtin, SVP Technology OKdo & RS Group



An example of applying Edge AI for predictive maintenance of a Forge Edge robot by Ready Robotics (Credit: Dominic Pajak).

Enabling Predictive Maintenance with Edge AI

Predictive maintenance is critical to ensuring continuous machine functionality and preventing costly downtime. Today, most predictive maintenance applications deploy machine learning models within cloud infrastructures to predict asset parameters, such as Remaining Useful Life (RUL). Edge AI can be vital in enhancing these use cases to minimize downtime and optimize maintenance. In this direction, it is possible to leverage real-time calculations of Remaining Useful Life (RUL) and End of Life (EOL) to provide real-time insights into the health and performance of machinery and other industrial assets.

For instance, it is possible to use Edge AI to detect anomalies and deviations from expected behaviors within very short timescales, i.e., almost in real time. Based on signals about these deviations, maintenance teams can make timely and informed decisions to proactively address potential failures and optimize the maintenance schedules for key assets. Edge AI also reduces the attack surface of predictive maintenance and intelligent asset management systems; ML models for RUL calculation can be executed within edge clusters or devices instead of within cloud data centers. This is the basis for increasing the security of predictive maintenance and intelligent asset management solutions.

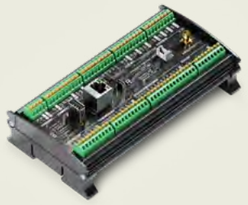
Innovate Predictive Maintenance with Arduino's Open-Source Edge AI Solutions

The convergence of IoT and Edge AI is crucial to address the limitations of traditional cloud-based systems by enabling real-time decision-making closer to the source of data generation – something particularly relevant in manufacturing, where split-second responses can prevent costly machinery failures. Today, **relying on powerful products like Arduino Pro's Opta and Portenta Machine Control represents the most robust and user-friendly catalyst to embrace Edge AI and revolutionize predictive maintenance** in factories big and small.

Opta is gaining huge traction in the industrial world as the innovative micro PLC that is quick and easy to use because it supports PLC standard languages in addition to the Arduino programming experience.

Arduino Opta WiFi

Portenta Machine Control is a fully centralized, low-power industrial control unit that several businesses have successfully chosen to empower new and existing machinery with IIoT capabilities, easily adapting it to a wide range of applications thanks to its modular design.



Arduino Portenta Machine Control

For example, **AROL** – a leading provider of capping machines – has paired the Portenta Machine Control with Arduino Pro's Nicla Sense ME modules to integrate monitoring and predictive maintenance capabilities into the equipment they sell, leveraging efficient data processing and wireless communication to significantly enhance the value they offer customers.

Arduino Nicla Sense ME

Spanish engineering company **Engapplic** chose Arduino's Portenta Machine Control to monitor air compressors' efficiency for a demanding client in the automotive field, allowing for the timely detection and even prediction of any anomalies. The result is a cost-effective and future-proof PoC that reduces downtime and saves energy.

While predictive maintenance is a great reason to add Edge AI capabilities to your machines, it's not the only one. Incorporating the brains of the Portenta Machine Control into the installed base – think professional kitchen appliances or office printers and copier machines – allows manufacturers to not only improve user experience and customer service but also access entirely new business models, such as usage-based rental contracts.

Arduino's commitment to the open-source philosophy brings additional advantages to industrial clients investing in advanced predictive maintenance solutions:

- **No Vendor Lock-In:** The company's open-source approach ensures that you have the flexibility to program, customize, and scale their solutions independently, avoiding vendor lock-in.
- **Shorter Learning Curve:** The user-friendly nature of Arduino's products allows engineers to quickly grasp and implement solutions, even with a limited programming background, facilitating efficient adoption by existing teams.
- **Complete Customization:** You can access, upgrade, and modify Arduino solutions freely, with the company's support, ensuring a seamless process. This flexibility is particularly valuable for companies looking to innovate and adapt to changing requirements.

Integrating IoT and Edge AI for predictive maintenance in manufacturing is a transformative journey – one that Arduino's robust products and open-source approach can help navigate with ease and no limits to innovation.

Real-Time Quality Control

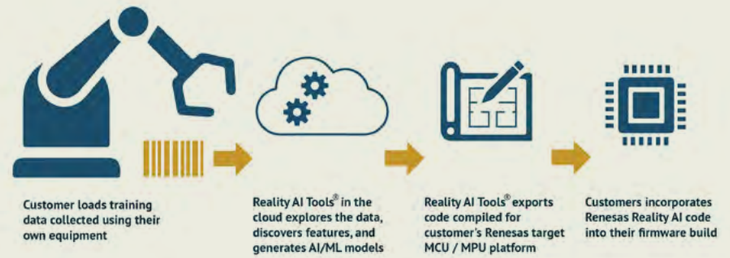
Maintaining high product quality is crucial for manufacturers to meet customer expectations and comply with industry standards. Edge AI enables real-time quality control by utilizing on-device inference for anomaly and defect detection. Based on AI model deployment directly on edge devices, manufacturers can analyze sensor data in real time, identify anomalies or defects, and trigger immediate actions to rectify issues.

This approach significantly reduces latency in quality control processes. Hence, it enables manufacturers to promptly detect and address quality-related issues to reduce waste, foster sustainability, and ensure high-quality products in the market.

Reinforcing Quality Control with Edge AI Tools from Renesas

Edge AI is gaining attention in quality control due to its ability to run AI/ML algorithms on edge devices, enabling scalable, real-time solutions. This approach, compatible with Renesas's diverse MCU and MPU edge devices, is ideal for large-scale manufacturing environments. It offers advantages such as real-time analysis, reduced latency, improved data safety, and cost savings by eliminating the need for extensive data storage and communication infrastructure. Moreover, Edge AI systems can adapt to dynamic manufacturing processes, accommodating variations efficiently.

In a typical manufacturing line for welding, for example, multiple robots perform various automated tasks. Here, end-of-line testing is crucial for quality control, particularly for manufacturing-dedicated components. The challenge is to detect porosity and burn-through in the welding process using conventional methods. The need, therefore, is to reduce the dependency on human inspections and enable low-cost solutions with real-time anomaly detection and high accuracy. Strict quality control ensures no faulty components are misclassified as good, avoiding costly recalls for automotive OEMs (though some tolerance exists for misclassifying good components as bad, which are discarded before shipment).

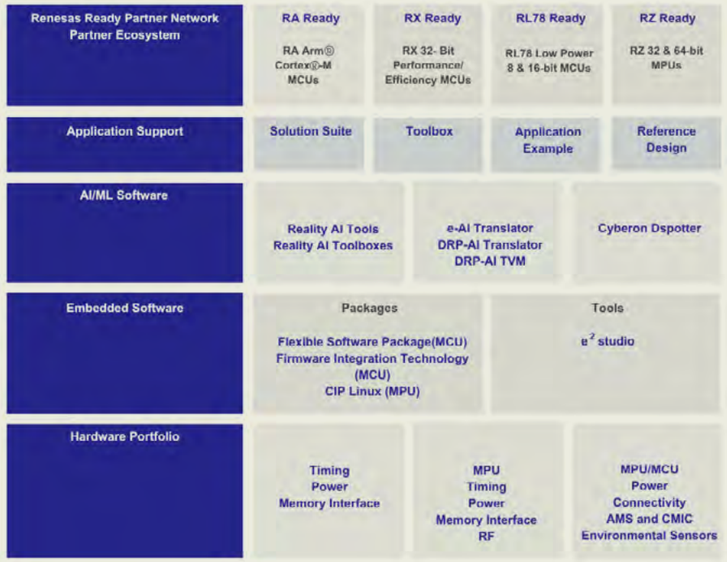


Reality AI Tools® (Credit: Renesas)

Accordingly, the core tasks are identifying, collecting, and analyzing the data set(s), then developing unification of the performance (inference), footprint, and accuracy in the best-fit AI/ML model. To ensure liable and high-quality data set(s), selecting the sensors and their positioning within the system is crucial. Renesas Reality AI Tools® provides the users with an automatic exploration of the sensor data and performs the analytics to find the best sensor (or combination of sensors) and the best placement. With the AI-driven feature Discovery in Reality AI Tools®, users have an advanced automatic exploration of the Sensor Data execution that generates optimized AI/ML Models at the end based on the explored data set(s).

To accelerate the development and deployment of dedicated Edge AI/ML solutions, Reality Al Tools® helps by using a machine-learning-guided process to explore the data and create a set of custom transformations (feature spaces) that define anomalies or maximize separation of classes/correlation to a target variable. The user can inspect the feature spaces and generate a time-frequency heatmap to show the most important structure for model accuracy. Achieving the highest rate of accuracy while meeting 0% False Negatives is a result of increasing the high level of manufacturing efficiency and productivity as a standard.

The Renesas scalable product portfolio comprising 16Bit, 32Bit, and 64Bit MCUs and MPUs, together with the rich ECO system and development infrastructure for embedded systems and Edge AI/ML solutions, is a perfect match for system solution packages. It targets and enables fast prototyping, evaluation, development, and deployment of your next embedded edge AI/ML solution.



The Renesas product portfolio (Credit: Renesas)

Facilitating Supply Chain Optimization

Efficient inventory management and logistics are crucial for maintaining a streamlined manufacturing process. Manufacturers are constantly trying to identify supply chain challenges in order to take remedial actions and implement optimizations. With Edge AI, they have opportunities to identify and confront issues faster than ever before. In particular, Edge AI can revolutionize supply chain optimization by enabling timely detection of real-time events. Specifically, data collection and analysis from various sensors and devices enable Edge AI systems to generate real-time insights into inventory levels, demand fluctuations, and transportation conditions. Manufacturers can use these insights to make data-driven decisions, such as optimizing stock levels, adjusting production schedules, or rerouting shipments. This can greatly improve overall supply chain efficiency, reduce costs, and ensure timely delivery of products.

Transforming Supply Chain Organization: OKdo's Edge AI Solutions in Action

Precision, efficiency, and safety are paramount in supply chain organization, which necessitates integrating cutting-edge technology. OKdo offers a wide range of AI-embedded products that drive faster and better decisions through data collection and analysis at the edge.

Useful Sensors' Tiny Code Reader: A Giant Leap in Inventory Control

Tiny Code Reader from Useful Sensors is a ground-breaking solution in supply chain, equipped with an onboard processor, camera, and AI accelerator. With its compact form factor, Tiny Code Reader – featuring Qwiic connectivity – seamlessly integrates into diverse systems, making it ideal for streamlined stock management. This miniature marvel, with a size akin to a coin, maximises spatial efficiency in warehouses, appealing to engineers who value flexible layouts. Its cost-effective design doesn't compromise technical prowess, offering a comprehensive developer guide with example codes for popular systems. Envision a warehouse with unparalleled accuracy. Tiny Code Reader reduces errors and transforms inventory management for engineers at the forefront of supply chain innovation.

OStream's ROCK 5 AIO: Empowering Predictive Maintenance & Safety Precision

Predictive maintenance is crucial for preventing downtime and ensuring the smooth flow of the supply chain. The ROCK 5 AIO, with its integrated 3 TOPS AI acceleration and pre-integrated 91 open-source AI models, brings a new level of sophistication to this arena.

Real-time equipment analysis predicts failures and prevents downtime whilst mitigating associated costs. Its AI capabilities extend to monitoring conveyor systems, detecting anomalies, vibrations, and wear indications, facilitating predictive maintenance scheduling to ensure uninterrupted goods flow and staff downtime. Furthermore, its robust AI processing enables real-time recognition for worker safety, ensuring compliance with safety equipment rules by monitoring clothing and movement, thus enhancing overall workplace safety.

Radxa ROCK 5A: Real Time Supply Chain Management At Your Fingertips

In today's hyperconnected world, supply chain success relies on real-time data processing. The Radxa ROCK 5A, powered by the advanced Rockchip RK3588S SoC, ensures exceptional processing efficiency with its Octa-core Arm® DynamIQ™ CPU and Arm Mali™ G610MC4 GPU. Enhanced AI capabilities, driven by the onboard 6TOPS NPU, improve digital display interactivity and intelligence, particularly in computer vision and image processing. HDMI outputs support resolutions up to 8Kp60 for seamless real-time inventory updates, while predictive analytics are displayed via dual micro HDMI ports. With 40 pin GPIO interface and versatile USB Type C™ and HDMI connectivity, the ROCK 5A seamlessly integrates with sensors, cameras, and monitors, offering a unified platform for data analysis. Supporting various operating systems including Android 12 and Debian/Ubuntu Linux, it provides flexibility for tailored digital display solutions, advancing supply chain technology.

OKdo Leading the Charge in Supply Chain Innovation

OKdo's Edge AI solutions are actively reshaping conventional practices. The Tiny Code Reader, ROCK 5 AIO, and ROCK 5A signify the practical application of AI, transforming inventory control, predictive maintenance, and ensuring people's safety. Design engineers in the manufacturing sector find a valuable ally in these advanced technologies, offering a gateway to heightened efficiency, unwavering reliability, and enhanced safety protocols within the supply chain.

An engineer using the help of a Kuka robot in a smart manufacturing setting (Credit: Zenoot)

Human-Robot Collaboration and Worker Safety and Training

With the rise of automation in manufacturing, human-robot collaboration (HRC) has become an essential aspect of optimizing operational efficiency. Edge AI is an effective contributor to enabling successful HRC. Deploying AI models on edge devices as part of Edge AI systems provides real-time feedback from the robot to humans and vice versa, facilitating seamless collaboration between human workers and robots.

Real-time feedback improves synchronization, enhances safety measures, and allows efficient task allocation. Overall, this Edge-AI-powered, collaborative approach enables manufacturers to leverage the strengths of both humans and robots, leading to increased productivity and improved operational outcomes.

On the safety front, Edge AI can significantly boost worker safety by leveraging AI models to detect hazardous situations in real time. The continuous monitoring of data from sensors and devices enables Edge AI systems to identify potential risks, such as machine malfunctions, abnormal movements, or hazardous conditions, in a timely manner. Early warnings and notifications can be sent to workers or supervisors, allowing them to take immediate action and prevent accidents. In this direction, Edge AI's ability to analyze data on edge devices ensures minimal latency in detecting safety-related issues, which keeps workers safe in real time. It is also positive that this increased safety comes with improved privacy, as there is no need to transmit workers' personal data outside the factory (e.g., to a cloud data center).

Furthermore, at the core of adapting to technological advancements and improving workforce efficiency is upskilling workers. Edge AI can contribute effectively by enabling real-time feedback during training tasks. With the integration of augmented reality (AR) and virtual reality (VR) applications, Edge AI can provide interactive training experiences, allowing workers to learn and practice in simulated environments. Moreover, real-time feedback based on AI models helps workers understand their performance, identify areas for improvement, and adjust their actions accordingly. This iterative training approach enhances worker skills and knowledge, ultimately leading to increased productivity.

Overall, the emergence and rise of Edge AI in the manufacturing sector offers a wide range of use cases that can significantly improve operational efficiency, product quality, supply chain management, worker safety, and workforce skills. Most of these improvements stem from Edge AI's ability to process and analyze data at the edge, leading to real-time performance, low latency, and enhanced security.

Chapter IV: Smart Cities and Urban Infrastructure

Edge AI Transforming Urban Areas

Urban environments are grappling with complex challenges across transportation, infrastructure, energy, waste management, and public safety, underscoring the pressing need for adaptation in our rapidly evolving cities. For instance, traditional traffic control measures, including infrastructure expansion and static traffic signals, fall short of resolving the complexities of modern urban congestion. The static nature of these systems, which rely on predetermined timing for traffic signals, fails to adapt to real-time traffic conditions, resulting in inefficiencies and unnecessary delays.

Similarly, challenges in energy efficiency and sustainability in smart buildings are made worse by outdated systems, leading to excessive consumption. Meanwhile, waste management systems lag behind in handling urban refuse volumes, and public safety in dense environments calls for more advanced surveillance and responsive emergency services. Recognizing these limitations, there is a growing emphasis on leveraging Edge AI to pioneer intelligent solutions.

> "Edge AI is the architect of change for our cities. As engineers, we hold the power to reshape our cities, infusing intelligence into their very foundations. We can build a future where cities seamlessly adapt, optimize, and protect, where innovation meets necessity, and where every line of code shapes the landscape of progress."
>
> Richard Curtin, SVP Technology OKdo & RS Group

Edge AI can be utilized to collect, visualize, and make decisions based on real-time weather and traffic data. (Credit: Department for Transport (DfT), UK, CC License, no changes made)

Navigating the Future of Traffic Management

Edge AI enhances urban mobility through technologies like AI-powered traffic prediction models that analyze vast amounts of data from cameras and sensors. This data helps optimize public transport routes and schedules, leading to more efficient services. For example, AI-enabled buses equipped with sensors can adjust routes in real time based on traffic conditions, reducing delays and improving passenger experiences. Its integration into public transportation systems has also led to significant advancements in the predictive maintenance of vehicles and effective passenger flow management. By harnessing real-time data, these systems offer commuters updated information, ensuring a more seamless and efficient travel experience. However, the integration of Edge AI extends beyond just route optimization.

A core component of intelligent traffic management is the optimization of traffic signal sequences. Edge AI devices like adaptive traffic lights use real-time data from traffic cameras and sensors to optimize signal timing, reducing congestion. These systems can adjust green light durations based on real-time traffic flow, significantly improving traffic conditions.

Edge AI can also contribute to broader urban planning and daily commutes. The introduction of Edge AI in traffic management has brought about substantial improvements in commute times and overall urban mobility, bringing about reduced congestion, which in turn leads to lower emissions, thus aligning with environmental sustainability goals.

Furthermore, AI-based traffic simulation models can predict the impact of new construction on traffic flow, aiding in more informed urban planning decisions. These advancements aid urban planners in developing more efficient and responsive city layouts, further enhancing the quality of life for urban residents.

Leopard Imaging Pedestrian Detection Solutions for Smart City Powered by Sony AITRIOS™

Traffic safety is a top concern for big cities. Accidents tend to happen both at and between intersections, so a scalable pedestrian detection system is required at an affordable cost. Traffic congestion management is another key issue for the city.

To solve these two challenges, Leopard Imaging, a global leader in embedded vision design and manufacturing, collaborates with Sony AITRIOS™ to develop "Dolphin" — advanced intelligent imaging solutions powered by Sony's IMX500 smart sensors. This state-of-the-art solution has won first place in the 2023 Pedestrian Safety Challenge sponsored by the tinyML Foundation and The City of San José in 2023.

Sony's IMX500 series smart sensor is Sony's latest smart imaging technology to enable smart camera implementation. The smart sensor has built-in image processing and AI acceleration, thus reducing the need for a high-power-consuming processor inside the camera.

Leopard Imaging's Dolphin intelligent solution is created for optimized cost and performance, low maintenance and installation costs, small form factor, and different available system design configurations to meet various applications, price points, and use cases. By taking advantage of Sony's smart sensor's unique capabilities and AITRIOS™ backbone, Leopard Imaging's latest Dolphin intelligent vision solutions can address many challenges facing smart city infrastructure proliferation.

By training the model with 11 gigabytes of data, Leopard Imaging has achieved an average of 90% accuracy in detecting pedestrians, bicyclists, and vehicles. Besides the increase in accuracy, Leopard Imaging has also achieved:

- Tailoring the YOLOV8n model to fit in the IMX500
- Obtaining a large Field of View (FOV)
- Scanning image Region of Interest (ROI) to improve accuracy
- Implementing self-cleaning lensing

Leopard Imaging's vision and detection solution "Dolphin" in-action
(Credit: Leopard Imaging)

The proposed solution also considers making the installation as easy as possible, leading to TCO reduction. The AI camera can be integrated with existing streetlight management systems by connecting a camera to a network lighting controller IoT gateway with a single cable. Thanks to this integration, users can install cameras on the LED lighting pole without having any power and data network installation. This approach helps municipalities not only protect existing investments but also add value to existing infrastructure.

Additionally, this solution is very flexible and can be adapted to various use cases in the city. The same system can be used for pedestrian detection, traffic counting, curbside parking occupancy detection, flood/snow detection for road management, and dynamic lighting control by simply changing the AI model from a remote site.

In conclusion, the partnership between Leopard Imaging and Sony AITRIOS™ represents a transformative stride in the evolution of smart city technology.

Advancing Digital Infrastructure and Smart Buildings

One of the keystones of smart urban infrastructure is energy optimization in buildings, where Edge AI plays a pivotal role. Edge AI enables the real-time management of energy consumption by integrating temperature, motion, and light sensors, along with advanced actuators for heating, ventilation, and air conditioning (HVAC) control. This system fine-tunes energy use based on occupancy, temperature, and lighting conditions, exemplified by adaptive lighting systems that adjust according to time, occupancy, and environmental factors. Such innovations not only reduce energy waste but also contribute to a more sustainable urban ecosystem.

Beyond energy management, Edge AI is instrumental in managing the flow of pedestrians within complex structures like offices, malls, and train stations. By monitoring and analyzing movement patterns, these systems ensure efficient and safe pedestrian traffic. In cases of emergency or security threats, the systems can guide occupants to safety in real time, showcasing Edge AI's critical role in public safety.

The deployment of Edge AI in building management marks a new era of efficiency and operational excellence. With capabilities like predictive maintenance, these intelligent systems preemptively address potential issues, minimizing downtime and extending the lifespan of building infrastructure. Sensors detect early signs of equipment failure, and AI algorithms predict when maintenance is needed, preventing downtime and extending the lifespan of building infrastructure.

Transitioning from building efficiency to broader urban practices, the application of Edge AI in waste management exemplifies another leap toward sustainability. Smart waste collection systems, powered by Edge AI, optimize routes and schedules for waste collection based on real-time data from waste sensor-equipped bins. This approach streamlines the process and supports the broader goal of reducing the carbon footprint and sustainable urban living.

Integrating Edge AI in waste management also has a notable impact on public health and safety. By ensuring more efficient and timely waste collection, these systems contribute to cleaner urban environments, reducing the risk of health hazards associated with accumulated waste. This is particularly important in crowded urban areas, where effective waste management is key to maintaining public health and hygiene standards.

Dustbins can use sensors and Edge AI to check garbage accumulation and predict when they have to be emptied next.

Edge AI Vision Sensor for Buildings with Rapid AI Model Development

Imagine easily setting up battery-powered sensors in your existing commercial building, safely gathering valuable insights into building usage without a disruptive installation process. That was the ambition for this innovative, Edge-AI-powered, vision-based sensor project by Eta Compute, powered by the latest low-power inference semiconductors and cutting-edge software tools to speed the otherwise time-consuming development of optimized ML models for accurate people counting.

Why People-Counting Matters

Understanding how buildings are used optimizes resources and enhances experiences:

- **Optimize Resource Allocation:** Adjust real estate portfolio levels, identify underutilized areas, and streamline layout based on actual measured occupancy, reducing waste and costs.
- **Improved Sustainability:** Optimizing resource allocation, HVAC settings, and energy consumption based on real-time occupancy data contributes to more sustainable building usage.
- **Boosted Safety and Security:** Trigger alerts for overcrowding or unauthorized access to safeguard people and property.

Edge AI: The Power of On-device Intelligence

Our vision sensor leverages low-power Edge AI for on-device people-counting, delivering key advantages:

- **Privacy-First:** Keep sensitive vision data local, ensuring privacy compliance and project acceptance.
- **Reduced Bandwidth:** Minimize network traffic and congestion while lowering power requirements.
- **Real-time Insights:** Make immediate decisions based on near-instantaneous data.

Easy Installation: The Game Changer

The most revolutionary aspect of this solution is its battery-powered, wireless design. Wired sensor installations requiring power and data have been the bane of IoT in buildings, limiting deployments to proof-of-concept trials and blocking widespread rollouts across building portfolios. Our sensor provides up to a 3-year battery life and offers:

- **Minimal Disruption:** Mount anywhere in minutes; no drilling, wires, or professional skills needed.
- **Scalability:** Add or remove sensors as your needs evolve without complex infrastructure adjustments.
- **Better Data:** Place sensors in previously inaccessible areas for richer insights.

Aptos Edge ML: Unleashing AI for Everyone

Developing low-power AI models has been a slow, manual process requiring 'unicorn' expertise in both AI and embedded systems. To overcome the challenges, we developed a radical new Edge AI toolkit tailored for creating such models, and it is called Aptos. Now commercially available, Aptos provides:

- **No-code Tools:** Embedded with the knowledge of capabilities and constraints of target silicon devices and compilers, it automatically builds optimal ML models.
- **Neural Network Algorithmic Search and Optimization:** Automatically explores neural network architectures and hyperparameters for optimal power, latency, and accuracy for your chosen low-power semiconductor.
- **Model Quantization:** Dramatically reduces the model's footprint and computational demand for efficient edge device operation.

By combining Edge AI, low-power technology, and the ease of ML development using Aptos, this battery-powered vision sensor project unlocks the full potential of commercial spaces and mindful use of resources.

Ensuring More Sustainable Cities with Edge AI

Edge AI is crucial for real-time monitoring of urban environmental factors, such as air quality and noise levels. By continuously analyzing data from various sensors spread across the city, these intelligent systems provide valuable insights into the environmental health of urban areas. This information is vital for city administrations to make informed decisions regarding pollution control and urban planning, leading to healthier living conditions for residents.

Edge AI also offers innovative solutions for monitoring and managing water resources. These systems can detect leaks, predict usage patterns, and optimize water distribution, ensuring efficient and sustainable water management. This technology is fundamental in densely populated cities, where water demand is high, and resource management is vital to sustainability.

Safeguarding Public Health and Safety

Edge AI devices like smart cameras and drones are used for real-time surveillance, quickly identifying and responding to safety threats. By processing data on the spot, these systems rapidly identify potential safety threats or criminal activities, enabling swift response from law enforcement and emergency services.

Edge AI also plays a pivotal role in disseminating real-time information to the public through smart digital signage. This technology is crucial in a variety of scenarios, from providing daily updates to broadcasting crucial information during disasters or emergencies. These dynamic digital platforms enhance the city's ability to communicate effectively with its residents, fostering a well-informed and prepared community.

Integrating Edge AI into surveillance and information systems not only enhances urban safety but also significantly strengthens community trust and cohesion by fostering a more secure and informed urban environment. The deployment of Edge AI in public safety initiatives has profound implications for community trust and safety. By enhancing surveillance capabilities and improving information dissemination, these technologies make cities safer and bolster the public's trust in urban infrastructure and governance. This, in turn, contributes to a more secure and cohesive urban community.

Throughout this chapter, we've seen Edge AI's multifaceted impact — from revolutionizing traffic management and building operations to advancing waste management and enhancing public safety. As we look to the future, the potential of Edge AI in urban environments continues to expand. With ongoing technological advancements, we can anticipate even more sophisticated applications that will further enhance the quality of urban life.

These advancements promise to streamline city operations and forge deeper connections between the urban landscape and its residents, fostering a more responsive, adaptive, and harmonious living environment. The journey of smart, Edge-AI-powered cities is just beginning, and its full potential is yet to be realized in the quest for smarter, more sustainable urban futures.

Leveraging Edge AI, smart digital signage can adapt content based on real-time data and audience demographics, ensuring tailored and impactful communication for diverse urban populations. (Credit: Kriesten objekt design GmbH, CC License)

Chapter V: Retail and Customer Experience

Edge AI Redefining Retail

The retail sector is undergoing a significant transformation driven by Edge AI, responding to mounting margin pressures and evolving consumer expectations. According to a report by McKinsey & Company, retailers, from grocery to specialty stores, face margin pressures ranging from 100 to 500 basis points, highlighting a critical need for efficiency gains.

Edge AI offers a solution to improve margins by 300 to 500 basis points through technologies like self-checkout systems and digital shelves. By leveraging the capabilities of Edge AI, retailers are not only addressing the immediate challenges of margin pressure but also getting ready for a future where technology-driven solutions redefine the shopping experience.

> "Edge AI transforms retail from transactional to experiential, revolutionizing customer interaction. It's not merely about selling products; it's about crafting personalized journeys that engage shoppers, fostering loyalty and driving growth."
>
> Samir Jaber, Editor-in-Chief

Inventory Management Perfected

Efficient inventory management is a balancing act of critical importance. Retailers are constantly grappling with the twin challenges of stockouts and overstock — each carrying its own set of repercussions.

Stockouts can result in lost sales and erode customer trust and satisfaction, leading to potential long-term implications on brand loyalty. Conversely, overstock situations tie up capital, increase storage costs, and lead to wastage, especially for perishable goods. These inventory misalignments directly impact a retailer's revenue and customer experience, making effective inventory management a foundation of successful retail operations.

In response to these challenges, the integration of Edge AI in inventory management has emerged as a transformative solution. For instance, smart shelves equipped with weight sensors and RFID tags can automatically monitor stock levels. These shelves relay real-time data to Edge AI systems, which then analyze and predict stock requirements, minimizing the occurrences of stockouts and overstock. Another example is the use of AI-powered cameras that track inventory movement, providing instant data on stock levels and customer preferences.

Edge AI also enables retailers to seamlessly integrate online and offline retail experiences. This technological advancement hinges on Edge AI's capability to provide real-time updates on inventory levels, ensuring accurate and current information is available to customers regardless of their shopping channel. Such real-time synchronization is essential in today's omni-channel retail landscape, where a smooth transition between online browsing and in-store purchasing is expected.

For instance, advanced predictive analytics and machine learning algorithms are employed to anticipate stock requirements, while digital price tags updated via Edge AI systems maintain price consistency across platforms. By utilizing these Edge AI-driven solutions, retailers effectively reduce the occurrence of stockouts and overstock, enhancing inventory management efficiency and ensuring a seamless and satisfying shopping journey for the customer.

Shelves with digital displays for price tags (Credit: MediaTile)

SONY

Scaling E-Commerce with Vision-Based AI for Inventory Management and Automation

Retail and e-commerce center around customer satisfaction, reflected through online product reviews and ratings. While scaling retail or e-commerce, hidden pitfalls like inefficient warehouse utilization and inaccurate inventory affect their seller rating by customers, siphon profits, and stifle growth.

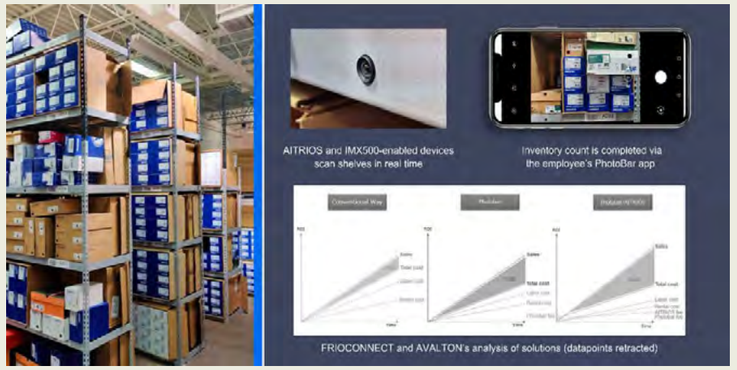
This poses a challenge for online retailers that sell the same goods on multiple websites. The importance of omnichannel for e-commerce means inventory needs to be accurately reflected across all sales channels, instantly. When it's not, order cancellations or wrongly shipped items can decrease seller ratings, resulting in deprioritization of the vendor's product listings, a loss of sales influence of up to 80%, or even vendor suspensions until issues are addressed.

The e-commerce boom brought FrioCONNECT challenges: inventory was inaccurate, counts were tedious, and 2% of listings were oversold or incorrectly shipped.

When 20% of items don't have the potential for restocking, companies like FRIOCONNECT cover the sunk costs associated with the mistake. To manage inventory, the process of counting 100,000+ items and 5,000 SKUs took five employees 10 days to complete. Without the counts, overselling or sending wrong items became more frequent.

Gaining real-time, high-precision inventory visibility, once a luxury, was now a strategic imperative. FrioCONNECT recognized that automation of cycle counts was necessary to ensure correct inventory was represented on sales channels and to support the customer experience.

Partnering with system integrator AVALTON and Sony's AITRIOS™ team, the PHOTOBAR solution was created, leveraging Sony's AITRIOS platform and IMX500 intelligent image sensor. In 2022, FRIOCONNECT dispatched over 75,000 items and realized a net profit of nearly 5-8% per product. They determined, based on their 7,800 square foot Doylestown, PA warehouse, that 348 IMX500-equipped devices would be needed to cover 58 racks and 696 shelves.



Sony's vision-based AI for inventory management and automation (Credit: Sony)

The solution automates cycle counting in real time, using edge devices to capture shelf images, process on-chip, and send only inventory-relevant data to an API for analysis and counting. Now, the 10-day-long process can be completed twice a day or as often as necessary.

Empowered by vision-based inventory management, FrioCONNECT increased order picking by 20%, optimized existing space, and is poised to gain back a projected $31,200 in error-based lost revenue from 2022, breaking even on their deployment of vision AI within a year, now with plans to expand the use of the technology across remaining warehouses.

Customer Behavior Analysis: Personalization at Scale

In today's retail landscape, personalization is increasingly vital, with customers seeking experiences that cater specifically to their preferences and behaviors. Edge AI is instrumental in this shift, enabling the collection and analysis of customer data in real time. Brands leverage AI engines to parse through data gathered from online browsing, past purchases, and in-store interactions, which allows for tailored product recommendations.

This use of data creates personalized shopping experiences and targeted marketing strategies, enabling retailers to meet the growing consumer demand for personalization with greater precision and effectiveness.

Beyond online personalization, Edge AI is transforming the in-store experience. Through the integration of interactive displays and the analysis of data from in-store customer interactions, Edge AI aids retailers in comprehensively understanding and catering to consumer preferences.

Retailers can employ smart cameras and sensors within the store to track and analyze customer movements and their interactions with products in real time. This rich stream of data informs decisions on product placement, store layout, and marketing strategies. By leveraging these insights, shops are empowered to optimize the environment, making it more engaging, tailored, and responsive to the unique behaviors and preferences of their customers.

Implementing Edge AI in customer behavior analysis presents mutual benefits for retailers and consumers. For retailers, the enhanced understanding of customer preferences leads to increased sales and better inventory management. Consumers, on the other hand, enjoy a more personalized shopping experience with recommendations and offers that are relevant to their interests and needs. This synergy is often supported by industry data and case studies, showcasing significant improvements in customer satisfaction and loyalty.

Smart carts can transform the shopping experience by providing customers with product recommendations, navigating them through the store, and facilitating seamless checkout processes (Credit: Caper).

Qualcomm

Qualcomm Technologies: The Power to Transform Retail with On-Device Generative AI

On-device AI fuels a more capable, cost-efficient, reliable, private, secure, and promising path forward for retail. Capable of working in harmony with cloud AI, edge devices deliver a faster, more efficient, and highly optimized AI with computing power you can rely on.

Qualcomm technologies are helping retailers stay competitive in the digital age by transforming retail with on-device, generative AI. Last year, the breakthroughs in generative AI were monumentally exciting. Aware of its vast potential for several years, Qualcomm Technologies, Inc. has made it so that AI is holistically engineered into our processors. This year, in-store and online shopping are expected to remain equally popular. As physical retail regains momentum, we see an opportunity to fulfill consumers' expectations by enabling seamless and integrated in-store experiences.

From on-device Generative AI applications for front-line workers and retail analytics to biometric Point-of-Sales (POS) payments and frictionless self-checkout and loss prevention, Qualcomm processors power many commercially available retail solutions and help deliver real results through:

- Generating insights for retailers' decision-making
- Boosting loss prevention to address the industry's $112B "retail shrink" problem
- Automating manual tasks to augment the workforce and minimize error rates
- Enhancing overall customer experience as consumers increasingly return to in-person shopping

Our ecosystem of top-notch customers and collaborators is making these transformations a reality. Their products address many of the primary challenges retailers face in areas such as frictionless checkout, loss prevention, inventory management, and operational analytics while also enhancing the all-around shopper's journey. Learn more about how Qualcomm technologies are transforming retail in 2024 and explore further in this infographic.

For more information, check out Qualcomm Technologies' retail solutions.

Checkout Automation for Reduced Waiting Times

One of the perennial pain points in the retail experience is the checkout process. Traditional checkout methods are often plagued with long queues and transactional delays, detracting from customer satisfaction and efficiency. These inefficiencies not only frustrate customers but also impact the overall throughput of retail operations — Edge AI steps in as a game-changer, offering innovative solutions to streamline the checkout process.

Examples of such innovations include automated checkout systems and smart carts. Automated checkout systems, which use computer vision and sensor fusion, are capable of identifying items in a cart and processing transactions without the need for manual scanning. This approach significantly speeds up the checkout process, enhancing customer satisfaction by reducing wait times. Smart carts, on the other hand, are equipped with barcode scanners and payment systems that allow customers to scan and pay for items as they shop, further reducing queue times.

Another application of Edge AI in retail is through self-checkout kiosks. These kiosks utilize machine learning algorithms to quickly identify items, offer various payment options, and even provide personalized offers based on the customer's shopping history. This not only enhances the checkout experience but also reduces wait times, thereby improving overall customer satisfaction and efficiency in retail operations.

By minimizing the time spent in queues and making transactions more seamless, retailers can ensure a more positive end-to-end shopping experience for their customers. Moreover, these technologies often lead to better resource allocation within stores, as employees freed from traditional checkout roles can focus on customer service and other value-added activities.

The integration of specific Edge AI devices and technologies in retail is not just a futuristic concept but a present reality. From smart shelves to checkout automation, these innovations are reshaping the retail landscape, offering high levels of efficiency, personalization, and customer satisfaction like never before. As we look ahead, the continued advancement and integration of these technologies promise to revolutionize the retail experience further, making it more adaptive, responsive, and customer-centric.

In addition to security and surveillance, cameras can play a crucial role in gathering valuable insights into customer behavior, enabling personalized shopping experiences.

Chapter VI:
Energy Efficiency and Sustainability

"Edge AI is a crucial technology in this world of finite resources. For example, it allows us to monitor and optimize electricity and water consumption in real time. Manufacturing, agriculture and logistics can thus minimize their impact, leading to huge cost savings and lowering the carbon footprint."

Fabio Violante, CEO of Arduino

Innovating with Edge AI

The global demand for energy is at an all-time high, driven by population growth and economic development. This relentless consumption exerts immense strain on our natural resources and contributes significantly to climate change. In response, improving energy efficiency and sustainability has become an absolute necessity.

Initiatives such as COP28 and the global stocktake that took place in November 2023 highlight this urgency, establishing ambitious benchmarks for international efforts.

The search for solutions must extend beyond traditional methodologies, necessitating innovative approaches.

A notable example of such innovation is highlighted in a McKinsey report, which suggests that adopting AI-driven methods in grids could enhance energy efficiency by 2-5% at the very point of generation, with up to 10% improvement in production and 30% in cost savings. This shift towards Edge AI in areas such as energy consumption monitoring, renewable energy integration, and smart grid management marks the next logical step in harmonizing our energy needs with the planet's health.

Edge-AI-powered smart energy meters revolutionize energy management by providing precise, real-time data on consumption.

Smart Energy Monitoring for Consumer Awareness and Cost Savings

Smart metering, bolstered by Edge AI, signifies a significant advancement in the way energy consumption is monitored. This combination not only facilitates real-time monitoring but also propels consumer engagement to new heights.

By providing detailed insights into energy usage patterns, consumers are empowered to make informed decisions, leading to more sustainable consumption habits. Edge AI plays a crucial role in this ecosystem by processing data on the spot, which helps quickly identify areas of excessive use or inefficiency.

Building on enhanced consumer awareness, Edge AI extends its impact to the operational level, driving efficiency and sustainability in various sectors. Implementing Edge AI in energy management systems has proven to be a game-changer for cost savings. In sectors such as commercial real estate, Edge AI has been instrumental in optimizing (HVAC systems, resulting in significant energy and cost savings.

Manufacturing is another area where Edge AI has made a substantial impact, streamlining energy consumption in line with production needs and thus lowering operational costs. These instances exemplify how Edge AI can foster consumer engagement and effectively reduce energy expenditure, aligning economic benefits with sustainability objectives.

Innovations in Renewable Energy Integration Powered by Edge AI

Renewable energy sources, such as solar and wind, are at the core of our efforts for a sustainable future. Edge AI enhances the integration of these renewable sources by optimizing their operation and efficiency. For solar energy systems, Edge AI algorithms adjust panel angles in real time to capture maximum sunlight. At the same time, such algorithms can optimize the blade rotation of wind turbines to harness wind energy more efficiently. The wind power generation maintenance market has also expanded, reaching approximately 35.7 billion RMB in China in 2022. Utilizing AI visual inspection, manufacturers have developed systems that ensure over 95% accuracy in detecting critical issues like ice accretion and cracks, significantly reducing maintenance costs and enhancing turbine efficiency. By ensuring more reliable and efficient wind power generation, these Edge AI systems support the integration of wind energy into microgrids, which are essential for decentralized energy systems. This not only contributes to the resilience and sustainability of energy supply but also accelerates the transition towards a more diversified and renewable energy mix, aligning with global environmental and energy security goals. This real-time data processing and decision-making capability of Edge AI ensures that renewable energy sources operate at their peak efficiency, significantly increasing their output and reliability.

As Edge AI continues to refine the efficiency and output of renewable sources, it sets the stage for their increased adoption and integration into global energy systems. In countries leading in renewable energy, such as Germany and Denmark, Edge AI has played a pivotal role in managing the variability of renewable sources, ensuring a stable energy supply to the grid.

Similarly, in remote and emerging regions, Edge AI has facilitated the deployment of microgrids powered by renewable sources, providing reliable energy access to communities. The collaboration between Engie Energy Access and Atlas AI in Kenya serves as a compelling example of how AI can facilitate the deployment of microgrids powered by renewable sources, thus providing reliable energy access to remote and emerging regions. By leveraging Atlas AI's predictive analytics and high-resolution geospatial data, Engie was able to identify high-density areas with significant demand potential for off-grid solar solutions. This data-driven approach enabled Engie to target and expand their renewable energy solutions specifically to communities most in need, resulting in a 48% increase in monthly sales in a pilot program. The use of AI in this context not only optimized Engie's operational focus and product offerings but also significantly contributed to enhancing energy access in underserved areas. These examples underline the global applicability of Edge AI in enhancing the scalability and effectiveness of renewable energy, marking a step forward in the global pursuit of sustainability.

Integrating renewable energy poses challenges due to its variability and the need for advanced grid management to ensure reliability and stability.

Proactive Smart Grid Management through Edge AI Solutions

Smart grid management represents a key area where Edge AI is making a significant impact, particularly in balancing supply and demand with greater precision. By leveraging real-time data analysis, Edge AI enables the grid to respond dynamically to changes in energy demand and supply conditions. This not only improves the reliability of the energy supply but also enhances the overall efficiency of the grid. For instance, during times of low demand, Edge AI can facilitate the storage of excess energy or manage its redistribution, optimizing grid operations.

Another area where Edge AI demonstrates its potential is in predictive maintenance, where its capabilities let it identify potential system vulnerabilities before they escalate. This foresight allows for preemptive adjustments to the grid, improving its resilience against disruptions and ensuring a stable energy supply. Predictive maintenance, enabled by Edge AI, can significantly reduce downtime and operational costs, further enhancing grid reliability and efficiency.

The discussion on smart grids just can't be completed without addressing the crucial integration and enhancement of electric vehicle (EV) charging infrastructure. The widespread adoption of EVs hinges heavily on a robust and intelligent charging network. In 2023, the European Commission passed a law

Optimizing EV charging infrastructure with Edge AI is crucial for balancing grid demands and facilitating the transition to electric mobility (Credit: Ather Energy on Unsplash).

mandating the expansion of the EV charging infrastructure to match the growth in EV adoption. For every battery-electric car registered in a member state, public charging stations must offer a power output of 1.3 kW. Furthermore, starting in 2025, there will be a requirement to install fast recharging stations with a minimum power of 150 kW every 60 km along the trans-European transport network (TEN-T), ensuring widespread and efficient charging options for EV users. Taking a look at another initiative, the government of India is focusing on improving the EV charging infrastructure in its capital. The plan involves setting up at least one charging station every 3 square kilometers, supporting the wider goal of electric vehicles making up 25 percent of all new vehicle registrations by the end of this year.

In this regard, Edge AI presents a game-changing opportunity to revolutionize this very infrastructure. Imagine charging stations that predict peak usage, dynamically allocate power based on vehicle needs, and even integrate with renewable energy sources for a truly sustainable experience. Through AI algorithms crunching real-time data on station availability, battery health, and even weather patterns, waiting times can plummet while charging efficiency skyrockets.

Predictive maintenance becomes possible, preventing downtime and ensuring smooth operation. Furthermore, edge AI empowers drivers with features like personalized charging plans and real-time updates on station status, fostering a seamless and stress-free journey. This revolution extends beyond individual stations, optimizing grid stability by intelligently managing power demands across entire networks. By harnessing the power of edge AI, we can unlock a future where EV charging is not just convenient but efficient, sustainable, and personalized, truly paving the way for a cleaner and more electrified tomorrow.

Transforming the Conventional Energy Sector with Edge AI

Edge AI is redefining safety and efficiency in the conventional energy sector, as well. On drilling floors, Edge AI-equipped cameras enhance safety by monitoring hazardous conditions in real time, significantly reducing the risk of accidents. Leak detection, traditionally reliant on extensive sensor networks, is revolutionized through Edge AI, where a single camera can identify leaks and other anomalies, streamlining monitoring processes and reducing equipment needs. Furthermore, in transportation and refinery operations, Edge AI optimizes routes and processes, improving efficiency and reducing environmental impact.

Illustrating the impact of Edge AI, Aramco, a leading hydrocarbon producer, has markedly improved its efficiency by adopting the technology. In its Khurais oil field, thousands of IoT sensors have enhanced oil well monitoring, cut power consumption by 18%, and reduced maintenance costs by 30%. The deployment of drones and wearable tech has also slashed inspection times by up to 90%. These applications of Edge AI not only promote operational safety and efficiency but also support the sector's shift towards more sustainable practices, highlighting the technology's potential to transform conventional energy production and management.

Edge AI's transformative impact on the energy sector is undeniable. Beyond smart metering and cost reduction, it optimizes renewable integration and enhances grid efficiency. While challenges persist, Edge AI represents a pivotal shift towards a sustainable future. Through responsible development and collaborative efforts, we can harness its power to unlock a cleaner, greener, and more resilient energy ecosystem, where sustainability is not a dream but a tangible reality.

Chapter VII: Agriculture and Food Production

Edge AI Enabling Smart Agriculture

As the global population, projected to surge to 8.5 billion by 2030 and 9.9 billion by 2050, demands more food, the agricultural sector faces the dual challenges of increasing production sustainably and ensuring food security. With the dire need to significantly ramp up production to meet this escalating demand, the adoption of Edge AI becomes not just advantageous but essential.

Edge AI is central to transforming food production, enabling advancements in precise management of crop/livestock, efficient resource utilization, enhanced quality assurance and beyond, addressing a spectrum of challenges with innovative solutions.

> "Integrating Edge AI into agriculture is about leveraging technology to optimize resources, maximize yields, and ensure food security for a growing population."
>
> Samir Jaber, Editor-in-Chief

Edge-AI-powered drones are transforming agriculture by enabling precise aerial monitoring and data collection, significantly improving crop health analysis and the efficiency of resource application (Credit: CropWatch).

Improved Crop Monitoring and Analytics for Maximizing Yield

Edge AI enhances agricultural productivity by enabling precise monitoring of crop health, growth patterns, and soil conditions. By processing data from an array of sensors—such as moisture, pH, and nutrient sensors — and high-resolution imaging technologies, Edge AI algorithms can identify early indicators of stress, disease, or nutrient imbalances in crops. This timely intervention allows for tailored management practices, such as targeted application of fertilizers or pesticides, leading to healthier crops and maximized yields.

Edge-AI-based systems can go a step further by harnessing real-time data on soil moisture and nutrient levels to optimize irrigation schedules and nutrient application, ensuring crops receive precisely what they need for optimal growth. This targeted approach not only conserves valuable resources but is also crucial in drought-prone areas, potentially turning the tide between crop failure and a successful harvest.

Furthermore, Edge AI empowers precision farming to become more controlled and accurate. Utilizing data-driven strategies, such as targeted irrigation and fertilization tailored to the unique needs of each plant, Edge AI significantly improves crop performance while reducing environmental footprints. Variable rate technology (VRT), for instance, applies water, fertilizers, and pesticides at the right moment and location, maximizing efficiency and minimizing waste. Together, these Edge AI-driven practices represent a leap forward in sustainable agriculture, combining resource conservation with enhanced crop yields.

Edge AI also plays a pivotal role in promoting sustainable agriculture by enabling practices that conserve resources and reduce chemical usage. By providing detailed insights into crop and soil health, it helps in implementing conservation tillage, cover cropping, and integrated pest management strategies more effectively. This not only supports the health of the ecosystem but also ensures long-term agricultural productivity and food security.

Edge AI's applications extend beyond immediate farm-level benefits, laying

a foundation for broader impacts on climate adaptation, economic sustainability, and informed policy-making. It is instrumental in adapting agricultural practices to changing climatic conditions and managing risks effectively. It also plays a crucial role in biodiversity conservation, monitoring ecosystem health, and promoting practices that sustain biodiversity within agricultural landscapes. The insights garnered from Edge AI applications can empower farmers with improved economic sustainability and influence agricultural policy-making, ensuring practices that are both productive and sustainable.

Streamlining Livestock Management with Edge AI

Edge AI introduces a new era in livestock management, focusing on enhancing animal welfare while optimizing productivity. By employing real-time monitoring technologies, Edge AI systems track the health, behavior, and nutritional status of livestock, enabling early detection and treatment of illnesses, stress, or dietary deficiencies. This proactive approach not only improves the welfare of animals but also contributes to more efficient farming operations.

From facial recognition for individual animal identification to automated systems for tailored nutrition and health management, Edge AI is transforming livestock care with precision and personalization. These technologies enable monitoring and management on a per-animal basis, improving the efficiency of breeding programs, and enhancing meat and milk quality. By leveraging Edge AI, farmers can make informed decisions that boost productivity while adhering to high welfare standards, showcasing a future where technology and traditional farming converge for superior outcomes.

Edge AI and IoT can come together and enhance herd management, enabling real-time tracking of health and behavior. (Credit: Nvidia)

Food Quality Assurance

Edge AI plays a critical role in minimizing food waste and mitigating contamination risks throughout the supply chain. By employing advanced imaging and sensor technologies, Edge AI systems can detect early signs of spoilage or contamination in food products, allowing for immediate corrective actions.

This capability is instrumental in ensuring that food storage conditions are optimized, significantly extending shelf life and reducing waste. Additionally, Edge AI aids in the identification of potential contaminants before products reach consumers, enhancing food safety and quality.

The application of Edge AI in food production can go beyond waste reduction, significantly contributing to higher food safety standards. Through real-time monitoring and analysis, Edge AI can enable the detection of pathogens, toxins, and other harmful substances at various stages of the food supply chain.

This proactive approach to food safety can not only help in preventing health risks but also boost consumer confidence in food products. Edge AI-driven systems can facilitate compliance with stringent food safety regulations, ensuring that products meet all necessary standards before distribution.

Edge AI also streamlines supply chain operations, from demand forecasting to inventory management, ensuring the efficient delivery of fresh products. This operational efficiency is instrumental in reducing food waste and operational expenses.

Furthermore, Edge AI fosters a transparent food system, where consumers gain insights into the food production journey, bolstering confidence and engagement.

Edge AI is at the forefront of today's agricultural revolution, driving advancements in crop monitoring, livestock management, and food quality assurance. Its ability to process and analyze data in real time is transforming traditional farming practices, making agriculture more efficient, sustainable, and productive. As we look to the future, the continuous evolution of Edge AI holds the promise of further innovations, ensuring food security and environmental conservation for generations to come.

Chapter VIII: Automotive and Transportation

An Edge-AI-Powered Automotive Sector

In recent years, Edge AI has proven its value as a game-changer in the automotive industry by enabling real-time performance and various optimizations across different use cases of the transport and mobility ecosystem. Such use cases are found not only within popular AI systems like autonomous vehicles but also in use cases that foster more efficient traffic management.

"Innovating with Edge AI in automotive isn't just about driving smarter cars; it's about redefining the road ahead, enhancing safety, and empowering vehicles to make split-second decisions, making every journey safer and more efficient."

Samir Jaber, Editor-in-Chief

The Holon Mover, an autonomous mini-bus (Credit: Benteler)

Autonomous Vehicles: Enhancing Safety and Efficiency on the Road

Autonomous vehicles are at the forefront of disruptive innovation in the transportation sector. Edge AI is vital in enabling these vehicles to navigate and make critical decisions in real time based on the fast processing of a large volume of sensor data. Edge AI technologies are deployed in all five levels of autonomous vehicles, ranging from Level 0 (no automation) to Level 5 (full automation). Specifically, Edge AI systems improve the real-time functionalities of vehicles of lower automation levels while boosting the autonomy of vehicles that fall into higher levels of automation.

For instance, Edge AI enhances the autonomy of partial-automation vehicles (i.e., Level 3) by boosting their advanced real-time perception, decision-making, and control functionalities. These functionalities help autonomous vehicles drive more safely and improve their environmental performance. Nowadays, relevant Edge AI functionalities can be implemented and deployed on different types of embedded devices of modern vehicles, such as On-Board Units (OBUs), enabling fast and real-time inference.

Furthermore, Edge AI's integration with other technological trends, such as Autonomous, Connected, Electric, and Shared (ACES) mobility, unlocks synergistic benefits for the transport ecosystem. Based on Edge AI, ACES technologies can enable collaborative intelligence functionalities close to the field. Such functionalities include, for example, route optimization, congestion reduction, and enhanced energy efficiency.

Real-Time Traffic Management and Smart Parking

Real-time optimizations are at the heart of efficient traffic management systems, enabling dynamic adaptations to rapidly changing conditions. Here, Edge AI empowers traffic management systems to process data from various sources (e.g., sensors, cameras, transport infrastructure) in real time. Analyzing such data at the network's edge enables the development of intelligent traffic management solutions that optimize signal timings, prioritize emergency vehicles, and dynamically adjust traffic flows. This way, Edge AI solutions help reduce congestion and improve transport efficiency.

The real-time data processing functionalities of Edge AI are essential in use cases that involve busy intersections where split-second decision-making is required. For such cases, Edge AI enables the development of intersection control systems that identify and prioritize vehicles while being able to detect anomalies or potential hazards. Such functionalities are key to ensuring smooth and safe traffic flows.

Furthermore, finding parking spaces in urban environments is one of the most pressing challenges, especially in highly populated megacities. With Edge AI, smart parking systems can provide real-time feedback about parking positions based on on-device AI. These systems leverage edge devices, such as parking sensors and cameras, to monitor parking occupancy and guide drivers to available spaces efficiently and in a timely fashion. Hence, Edge AI's processing power on edge devices enables smart parking systems characterized by low latency and enhanced real-time decision-making capabilities, improving the overall parking experience for drivers.

The Canada Infrastructure Bank (CIB) and Quebec-based charging network operator FLO have announced a plan to install over 2,000 public DC fast charger ports across the country by 2027 (Credit: FLO).

Electric Vehicles: Enhancing Battery Management and Charging Infrastructure Optimization

During the last few years, the world has increasingly embraced electric vehicles (EVs). EVs require battery management optimizations and the availability of EV charging infrastructures. To this end, Edge AI can facilitate real-time monitoring and analysis of EV batteries in ways that enhance their performance, lifespan, and overall reliability.

Moreover, thanks to Edge AI algorithms, EVs can dynamically adjust charging patterns based on factors like battery condition, power availability, and user preferences. At the same time, Edge AI's integration with EV charging stations enables intelligent load management, balances energy demand, and optimizes charging schedules. Also, the proper distribution of the computational load at the edge can improve the grid's stability and efficiency by reducing strain during peak demand.

Fleet Management and Traffic Sign Recognition

Edge AI empowers fleet management systems to collect and process a wealth of vehicle data, including location, fuel consumption, driver behavior, and vehicle health. The processing of these data at the edge provides fleet managers with real-time insights into fleet performance for implementing routing optimizations, reducing fuel consumption, and enhancing the operational efficiency of their fleet. Thus, Edge-AI-driven fleet management solutions can effectively help with cost reduction, improved logistics, and enhanced customer experience.

Similarly, driver safety and compliance rely primarily on traffic sign recognition. Edge AI systems enable vehicles to detect and interpret traffic signs in real time. As such, they can provide drivers with real-time information about speed limits, traffic rules, and potential hazards. In this direction, advanced image processing algorithms can be deployed at the edge based on Edge AI approaches like TinyML for on-device image analysis. Therefore, Edge AI systems can enhance driver awareness, reduce the risk of accidents, and foster compliance with traffic regulations.

The above-listed use cases indicate how Edge AI transforms the transport sector to enable real-time functionalities across various use cases. With its ability to process vast amounts of data at the edge, Edge AI empowers transport stakeholders to enhance safety, optimize operations, and improve the overall efficiency of modern transportation systems. Moreover, it helps reduce the attack surface of transport applications and limit the amount of potentially sensitive data (e.g., driving patterns information) from being shared within a cloud infrastructure. Edge AI can also boost privacy and data protection for consumer-facing applications in the transport sector. In the future, Edge AI systems will continue to play a significant role in shaping the evolution of the automotive and transportation industry.

Chapter IX: Generative AI at the Edge

"Edge AI is changing dramatically; the new generation of hardware is capable of 500 times better performance than before, which has a massive impact on the world. It's difficult to fathom the full potential, but we do know that our ability to create accurate, high-performance Edge AI models is hugely expanded."

Alexander Samuelsson, CTO of Imagimob

Where Gen AI Meets Edge Computing

The convergence of generative AI and edge computing is revolutionizing how AI interacts with its environment. Enterprises are increasingly deploying computing resources at the edge to capitalize on local data collection, filtering, aggregation, analysis, and generation. The trend of deploying generative AI at the edge is driven primarily by two key advancements: the development of relatively small, specialized models tailored for edge deployment and the availability of hardware acceleration for inferencing processes.

These advancements enable the deployment of generative AI in edge environments, leading to applications such as voice-assisted shopper suggestions in retail, translation and emotion analysis of customer feedback, and autonomous decision-making in industrial environments and warehouses. In this chapter, we'll take a deeper look at large language models (LLMs), what hinders their deployment at the edge, and how leading companies are finding ways to overcome these challenges.

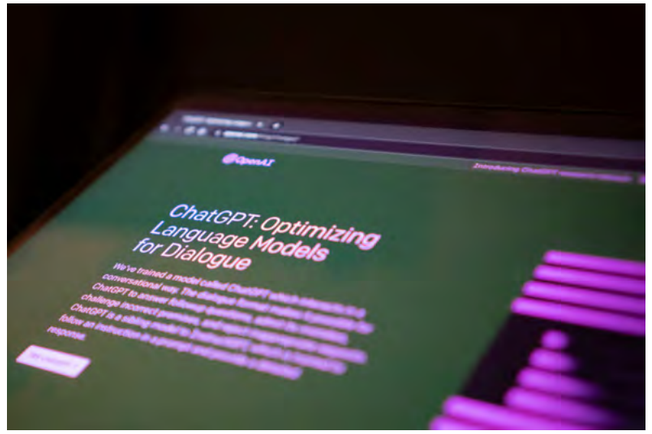
LLMs at the Edge: The Challenges

The emergence of ChatGPT in November 2022 has led to a surge of interest in LLMs and their applications in various sectors. LLMs have revolutionized natural language processing (NLP) and natural language understanding (NLU) by enabling general-purpose conversational interactions while achieving remarkable performance on various language tasks. These models, such as GPT-3, Bard, LLaMa2, GPT-4, and Gemini, leverage neural networks with billions of parameters trained on vast amounts of data. This enables LLMs to generate coherent and contextually relevant text, which drives the development of many innovative applications. These applications are usually deployed within cloud-based infrastructures, as the growing size of LLMs poses challenges that inhibit their deployment on edge devices.

The most prominent challenge to deploying LLMs at the edge is their sheer model size. LLMs with billions of parameters require substantial computing resources, which exceed the computational capabilities of most edge devices like smartphones or IoT devices. For instance, most IoT devices have limited memory and processing power, which makes it difficult to store and execute large models efficiently. Hence, deploying massive models on edge devices requires significant optimization efforts and hardware upgrades.

Another major obstacle is the computational power required to utilize LLMs in real-time interactions. Edge devices often lack the computational capacity to process complex language models quickly. Therefore, applications requiring real-time interactions (e.g., chatbots, voice assistants) may not be feasible on edge devices with limited processing capabilities. The deployment of LLMs on such devices asks for sophisticated strategies to reduce latency and enhance performance.

Despite these challenges, the concept of edge LLMs is very promising, as it comes with significant advantages like reduced latency and enhanced user experience. For instance, voice assistants or predictive text features can run directly on the device, eliminating the need for a constant internet connection and ensuring fast response times. Moreover, Edge LLMs can greatly enhance privacy and data protection for the growing base of LLM users. When processing prompts locally on the device, there is no need to transmit sensitive user information (e.g., information about the users' tasks) to the cloud. This reduces the potential risks associated with data breaches or unauthorized access to user data.



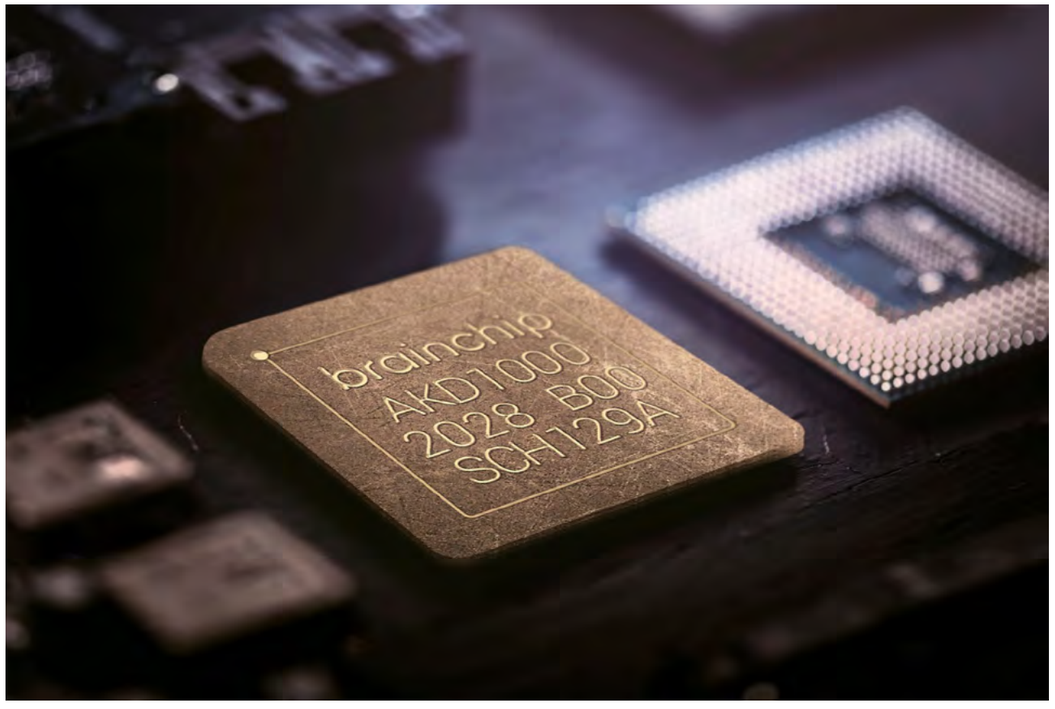
ChatGPT by OpenAI has made a substantial impression on almost every user (Credit: Jonathan Kemper).

Enabling LLMs and Edge Computing Convergence

Nowadays, different solutions are under research and development to address the challenges of LLM deployment at the edge. These include model compression techniques, specialized hardware accelerators, and on-device training methods. They aim to minimize model size, reduce computational requirements, and achieve efficient execution on edge devices.

One of the most prominent approaches to deploying LLMs on edge devices is reducing their model size. This can be achieved through techniques such as pruning, quantization, and distillation to produce smaller models while maintaining high performance. Pruning involves removing redundant parameters from the model, which results in a more compact representation. Quantization reduces the precision of numerical values used for computation, thus requiring less memory and processing power. Distillation compresses large models into smaller ones by using the output of one model as the target of another, leading to significant reductions in model size.

Another approach involves the design and development of domain-specific models. Domain-specific applications enable the deployment of smaller and highly optimized models customized for particular use cases yet offer better efficiency and accuracy. Furthermore, domain-specific models enable edge devices to benefit from LLMs' capabilities without relying on cloud infrastructures, which leads to increased privacy and faster response times.



BrainChip's AKD1000 AIoT chip brings AI and edge together, leveraging its Akida™ advanced neural networking processor (Credit: BrainChip)

Examples of Real-Life Edge LLM Systems

One example of an LLM deployed on edge devices is an MIT-developed model called TinyChat. TinyChat is a neural conversational model that uses a small, domain-specific model to achieve real-time interactions on edge devices. The model is trained on data from a specific domain, such as customer support or technical assistance, which makes it more efficient and accurate in handling domain-specific queries.

NVIDIA has also developed an embedded AI platform called NVIDIA IGX Orin Developer Kit, specifically designed to deploy LLMs on the edge. It offers high-performance computing power and memory capacity, making it ideal for running large models locally on devices like cars or robots. The kit uses NVIDIA's TensorRT high-performance deep-learning inference framework. The latter optimizes models for deployment on edge devices towards faster inference times and lower power consumption.

Qualcomm is also working on optimizing generative AI for edge devices. Their efforts enable large models to perform real-time inferences on low-powered devices, which reduces applications' reliance on cloud infrastructures. Qualcomm uses techniques such as on-device model compression and hardware acceleration to overcome the challenges of deploying LLMs on edge devices.

Generative AI Meets Axelera's In-Memory Computing at the Edge

At its core, Generative AI's ability to craft new data boosts new applications. Large Language Models (LLMs) like GPT-4 drive conversations, and tools like Dall-E convert natural language input into images. A fast-evolving extension of this development is multi-modal Generative AI, where different input sources of text, audio, and video are analyzed and acted upon to generate new data.

For businesses across the board, AI technology marks a transformative era, enabling dynamic optimization of operations for enhanced efficiency and environmental sustainability. For example, in surveillance and security, along with other sectors that rely heavily on vast volumes of text, audio, and video data, multi-modal Generative AI is set to introduce formidable capabilities for real-time anomaly detection and proactive intervention to prevent incidents from escalating. Furthermore, surveillance footage becomes queryable through simple prompts, allowing for comparing different feeds, generating situational descriptions, and providing simulated outcomes. In parallel, the manufacturing sector leverages machine vision and automation, enhanced by digital twin simulations, while the retail industry is witnessing the emergence of automated checkouts. Moreover, AI's role extends to facilitating personalized shopping experiences and advancing smarter inventory management practices.

How does this impact AI processing at the edge? Today, most of the AI processing is done in centralized data centers. However, compared to typical business computing, AI processing is considerably more compute- and data-demanding and typically resorts to using powerful but power-hungry GPUs (Graphics Processing Units). This creates sustainability challenges, heavy demands on network bandwidth, real-time delays, and privacy concerns.

Enter Generative AI processing at the edge. Localized AI processing, closer to where the data is generated, can enhance privacy by not sending sensitive data to the cloud. Edge AI can also efficiently meet scalability challenges due to its more distributed model. However, implementing Generative AI at the edge presents new challenges, particularly in resource- and cost-constrained environments. As AI demands substantial processing power, a new breed of scalable, purpose-built AI acceleration solutions is emerging. Such novel technologies, like Axelera's Digital In-Memory-Computing, can deliver remarkably high throughput while supporting programming flexibility, yet process AI models at a fraction of the cost and power consumption compared to traditional GPUs.

The evolution of Generative AI at the edge heralds a significant transformation, necessitating a careful equilibrium between the opportunities it presents and the responsibilities it entails. It's imperative to equip the workforce for this transition, ensure AI development is guided by ethical principles, and guarantee that the fruits of this technological evolution are accessible to all.

LLM at the Edge: Transforming Multiple Industries at Once

Edge LLM deployments are likely to have a significant impact on various markets and application areas, bringing advanced conversational AI capabilities to the edge in real time. Here are some specific applications where Edge LLMs are expected to make a significant impact:

- **Chatbots:** Edge LLMs can improve the capabilities of state-of-the-art chatbots by enabling them to generate more intelligent and contextually relevant responses directly on edge devices. This reduces latency and the need for an internet connection, leading to faster and more efficient user interactions.

- **Virtual Assistants for Front Office Applications:** Edge LLMs can enhance virtual assistants and front office applications by allowing them to process user queries locally. This is particularly useful in scenarios requiring real-time conversation, such as with voice assistants.

- **Code Generation and Code Completion:** Edge LLMs can greatly benefit developers by providing code suggestions, code completion, and automatic code generation directly in their development environment. This can streamline the software development process, leading to increased productivity and efficiency in tasks such as Continuous Integration and Continuous Development (CI/CD) and other DevOps support tasks.

- **Text Generation by IoT devices:** When deployed in cloudless environments like rural areas, IoT devices can leverage Edge LLMs to generate text in real time. For example, a smart agriculture device can analyze sensor data locally and provide recommendations on crop management, irrigation, or pest control without relying on an internet connection.

- **Training with Real-time Feedback:** Edge LLMs can enable on-device training with real-time feedback. This is particularly beneficial in applications such as Augmented Reality (AR) and Human-Robot Collaboration (HRC), where instant feedback and adaptation are crucial for providing a pleasant and satisfactory user experience. For example, an AR application can use an Edge LLM to generate realistic virtual objects based on real-time information captured by the device's camera.

Overall, the combination of generative AI and edge computing in hybrid edge-cloud deployments offers a wide range of benefits in terms of reduced latency, improved scalability, enhanced privacy, and cost efficiency. Edge AI vendors have already developed and demonstrated Edge LLM solutions, including their benefits in real-time applications like chatbots, virtual assistance and human-robot collaboration. As more Edge LLM applications emerge across various industries, the integration and use of edge and cloud resources with LLMs will play a crucial role in unlocking the full potential of generative AI.

Scaling Generative or Multi-modal AI

brainchip
Essential AI

The Generative AI phenomenon and the related Multi-modal AI have opened a new way of services and business along with an enormous challenge of exponentially increasing costs for training and inference.

Activities like ASR (Automatic Speech Recognition) through raw audio, Text-to-Speech (and vice versa), and, with the advent of Augmented Reality, more contextual scene generation are among many possibilities driving demand for compute. The challenge is that the level of sensor data, LLMs, and Large Vision Models (LVMs) is too heavy to manage on the Edge device. Therefore, it drives a much larger load at the cloud, not to mention the challenges with real-time response, security, and privacy.

With the slowing down of Moore's Law and especially Dennard's Scaling, power and energy benefits that were taken for granted don't provide the expected gain. Therefore, there is a great deal of architectural innovation needed to empower Edge devices.

BrainChip's DNA is that of brain-inspired, neuromorphic computing, but taking it in a digital and event-based manner rather than traditional analog. In the second generation, BrainChip introduced the Temporal Event-based Neural Networks (TENNs) that are very adept at multi-dimensional streaming data. Initially, TENNs have shown their benefit for a wide variety of streaming data solutions – consuming raw audio signals or health care data without the need for filtering to infer audio or vital signs. Similarly, they show benefits in achieving video object detection in sub-watt power envelopes. Combining the new algorithms with some innovative hardware choices – especially those that make 3D convolutions very efficient – brings a big step in efficiency—demonstrating more than 100-500x improvement in energy efficiency without compromising accuracy.

In terms of disruptive potential, TENNs could revolutionize LLMs and LVMs at the Edge. Tested on prior generation transformer-based models, a TENNs-equivalent has shown that for equivalent perplexity scores (an indicator of correctness), while reducing model size and MACs/token by 3-4 orders of magnitude. More importantly, the training of these models is similar to that of CNN training and yet takes less than 1/10th the time compared to the transformer equivalent.

The result is that Akida™ with TENNs can provide radical alternatives for Edge Devices that can now handle much more complex models in a small footprint solution for vision, surveillance, hearables, automotive, healthcare, and more. Check out our white paper from BrainChip.

Chapter X: Edge AI Challenges and Real-World Mitigations

Edge AI has transformed the landscape of computing by enabling local data processing on devices, offering a plethora of benefits, including reduced data transfer requirements, accelerated response times, and better reliability. However, alongside these advantages, the deployment and integration of Edge AI systems encounter a multitude of challenges that demand thorough consideration and strategic mitigation.

These challenges still hold Edge AI back in terms of data processing and AI modeling. Fortunately, many mitigation strategies and techniques have been suggested and tested, but there's still a lot of work to do. Arguably, one of the biggest challenges facing Edge AI today is resource-constrained environments. Enabling Edge AI to successfully operate in these environments can be transformative, at the very least.

In addition to the technical challenges, several human-related obstacles still stand in the way of the adoption of Edge AI in some industries. In this chapter, we take a closer look at all these challenges, investigate their reasons and how they affect Edge AI's efficiency and performance, and identify real-world mitigations that can help Edge AI navigate its way through to mass adoption.

> "Edge AI is revolutionizing technology – devices can process data locally in real-time, and provide seamless, intelligent experiences. We believe in enabling companies to take part in this revolution by lowering the barriers to using machine learning in products. These lowered barriers create unprecedented opportunities for innovation, efficiency, and productivity."
>
> Anders Hardebring, CEO of Imagimob

Tackling Edge AI's Key Challenges

While the adoption of Edge AI offers significant advantages, it is not without its challenges. Various factors contribute to complexities in Edge AI implementation, introducing hurdles for deployment in practical scenarios. Some of the most prominent challenges include the lack of hardware standards, data management, data privacy and security, and scalability.

1. **The hard problem of hardware:** Integration challenges also loom large, driven by disparities in hardware, software, and communication protocols across edge devices. The lack of standardization, especially in hardware, exacerbates compatibility issues, hindering seamless integration with existing systems. Varying computing capabilities across different edge devices make it difficult for developers striving to create universally compatible Edge AI applications. Establishing industry-wide hardware standards is essential to facilitate seamless integration and scalability of Edge AI solutions across diverse environments and devices. Until then, companies must be careful when choosing the hardware for deploying AI on the edge, especially taking into consideration factors like power consumption, memory requirements, processors, interoperability, and security. Initiatives like the Open Neural Network Exchange (ONNX) offer promising pathways to address these concerns, fostering interoperability and facilitating smoother integration processes. Also, leveraging other technologies, such as Wi-Fi and Bluetooth, can help overcome integration and interoperability challenges.
2. **The persistent challenge of data management:** One of the primary obstacles facing the implementation of Edge AI today is data management. This comes in the form of data movement, data storage, and data governance.

- **Data movement** is influenced by the amount and speed of data transmitted. In turn, issues with data movement can impact efficiency, power, latency, and real-time decision-making. With edge computing, it is necessary to reduce data movement and, in turn, minimize latency to maximize real-time decision-making. This can be done by distributing intelligence. One innovative solution for reducing data movement is called federated learning, which leverages distributed data across multiple edge devices to train AI models and

enhance data quality, privacy, and diversity. Also known as on-device ML, federated learning is capable of training AI models on different data sets without the need to exchange raw data. In other words, the data in its original form always stays on the device and is never gathered in one central location.

- As for **data storage**, edge devices' limited storage capacity necessitates the employment of data compression techniques to optimize memory usage and accommodate larger data volumes. Several compression techniques have been proposed and tested, such as Huffman coding and Lempel-Ziv-Welch (LZW) compression.

- The challenge of **data governance**, while often overlooked, can become more and more problematic as Edge AI applications become more widespread. In order to govern data properly and comply with regulations, especially in complex enterprise environments, dedicated data governance frameworks would be needed. These may have to rely on other state-of-the-art technologies to ensure effective governance, such as blockchain, to ensure edge devices with high efficiency, security, and reliability.

3. **AI's most worrisome challenge - data privacy and security:** Security remains a paramount concern for Edge AI, particularly given the sensitive nature of the data processed by edge devices. Robust security measures, including secure boot mechanisms and hardware root of trust (RoT), are imperative to safeguard the edge device's integrity against potential threats such as data breaches and privacy violations. Likewise, employing secure software development methods such as modeling threats and reviewing codes can mitigate typical vulnerabilities. Interfacing with cloud-based services can pose additional security risks, which can be countered by employing methods such as encrypting data, ensuring secure authentication, and using reliable communication protocols. In terms of AI models, techniques like differential privacy, federated learning, and homomorphic encryption are utilized to train AI models on sensitive data without jeopardizing privacy. In addition, anomaly detection methods can identify and mitigate attacks targeting Edge AI systems, thereby ensuring the security of the entire system.

4. **To scale or not to scale:** Scalability presents another formidable challenge, spanning computational, data, and system scalability limitations. Edge AI systems face scalability challenges in three key areas: computational, data, and system scalability.

   - **Computational scalability** refers to the system's ability to handle increasing data volumes without exceeding device capacities, hindering accuracy and responsiveness.

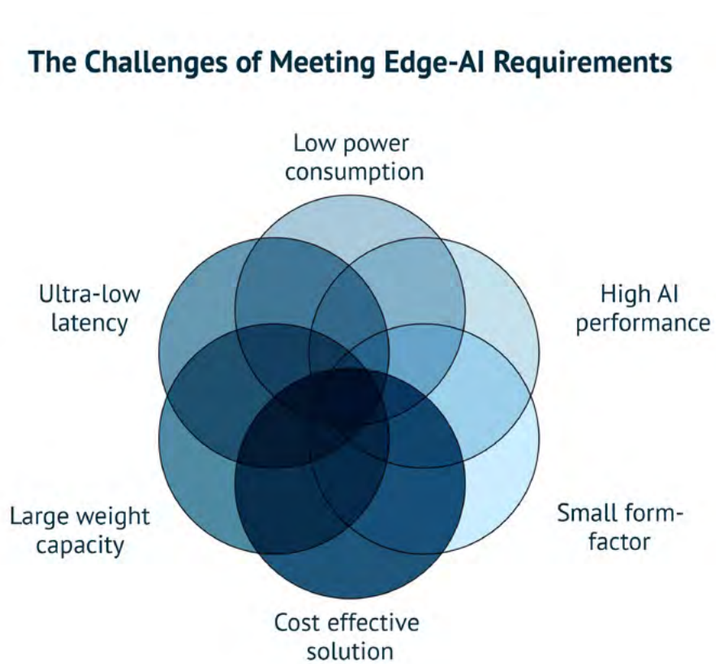
   - **Data scalability** involves managing large data volumes without performance compromise due to limited data transfer capacity and unreliable connectivity.

   - **System scalability** addresses the management of growing device and user numbers, complicated by distributed processing's latency and complexity.

   Solutions like load balancing and distributed computing optimize scalability, while techniques like edge orchestration and network slicing enhance coordination and resource allocation, overcoming Edge AI's scalability barriers.

   Needless to say, other challenges of Edge AI, such as cost considerations, power consumption, form factor, and ultra-low latency, are also worthy of mentioning. These can be addressed with a unique strategy of co-designing hardware and software. This approach, led by companies like Axelera, involves techniques such as data quantization and network pruning to compress data. By doing so, the chip uses less storage and memory to run AI algorithms. This reduction in memory and computational complexity, in turn, improves energy efficiency.

   Most importantly, the successful deployment and integration of Edge AI necessitate a comprehensive understanding of the challenges and opportunities inherent in this transformative technology. By addressing these key challenges, stakeholders can unlock the full potential of Edge AI to drive innovation and efficiency across diverse industries and applications.



Edge AI is still in its early stages, and several challenges remain to be solved (Credit: David Kuo, Industrial Ethernet Book).

> "There is a huge gap between the mainstream AI domain and the constrained environments of embedded systems. This gap, and the lack of 'unicorn' developer resources with expertise in both areas, is blocking the potential of Edge AI from being unleashed. New no-code automation tools are on the horizon to overcome these challenges and enable developers to rapidly bring Edge AI-based products to volume markets."
>
> Evan Petridis, CEO of Eta Compute

SYNTIANT

Overcoming the Challenges of Edge AI

Edge AI is revolutionizing the accessibility of AI by bringing its capabilities closer to data sources on networks. However, beyond the initial hardware challenge of providing ample computational power to constrained environments, three critical obstacles must be addressed for widespread deployment: the scarcity of production models; a shortage of modeling expertise; and the diverse array of applications.

Scarcity of Production Models

Substantial R&D efforts in AI have generated numerous models predominantly designed for cloud-based environments, often exhibiting basic performance. While creating demonstration models can be a swift process, developing production models demands significant time and effort. Adapting these models to operate efficiently at the edge, where computational resources are restricted, presents a formidable challenge. Moreover, the diversity of edge devices, each with its unique computational capabilities, introduces additional complexity. Creating lightweight and efficient models adaptable to various cloud-free environments is crucial for facilitating the widespread adoption of Edge AI.

Shortage of Modeling Knowhow

Crafting models that can effectively operate within the limitations of constrained resources necessitates a specialized skill set. Edge devices, often constrained in processing power, memory, and energy, demand the optimization of algorithms and models. The shortage of experts possessing the knowledge to design and implement AI models tailored for edge deployment poses a substantial barrier. Closing this knowledge gap is imperative to unlock the full potential of Edge AI, broadening its accessibility to a diverse range of applications. An efficient and adaptable Modeling Platform, designed to create production-quality models, represents one solution to empower a broader community of engineers with expert-level performance.

High Diversity of Applications

Edge AI finds applications across diverse fields, such as healthcare, manufacturing, smart cities, and autonomous vehicles. The challenge lies in the varied requirements and constraints of these applications. Designing a one-size-fits-all solution for Edge AI proves impractical due to the unique demands of different use cases. Customization and adaptation of models for specific applications are essential, making it challenging to establish standardized frameworks. The deployment of a versatile Edge AI Modeling Platform, easily tailored to diverse applications, ranging from simple sensors to large language models, is pivotal for realizing the full potential of Edge AI across various industries.

Syntiant's Answer

Syntiant has emerged as an Edge AI solutions provider, surpassing the challenges associated with hardware performance by offering turn-key production models for various sensor types, encompassing audio events and computer vision. For teams seeking to develop their own models, the Syntiant Modeling Platform provides extensive augmentation, rapid iteration, and quality assurance, significantly expediting the deployment of production models across platforms spanning from DSPs to GPUs. Addressing the high diversity of applications, the Syntiant Modeling Platform supports various modalities, including audio, vision, and radar. With these tools, Syntiant is making Edge AI a reality.

Navigating Resource-Constrained Environments with Edge AI

Resource constraints, including limited computing resources and network bandwidth, pose significant challenges to Edge AI implementation. Edge devices often operate in resource-constrained environments with limited processing power, memory, and storage capacity. Moreover, bandwidth constraints in edge networks can restrict the transfer of data between edge devices and centralized servers, affecting the performance and scalability of Edge AI applications.

To address resource constraints, stakeholders must adopt a holistic approach that encompasses hardware optimization, software efficiency improvements, and network optimization strategies. Hardware manufacturers are developing specialized edge computing hardware tailored for AI workloads, featuring low-power processors, accelerators, and memory architectures optimized for edge deployments. Software developers are exploring lightweight AI algorithms and techniques for edge computing, such as edge caching and data compression, to minimize resource utilization and enhance performance. Additionally, network optimization strategies and edge computing architectures can help alleviate bandwidth constraints and improve data transfer efficiency in edge environments.

Ensuring a balance between model complexity and performance is crucial for Edge AI to navigate resource-constrained environments. A 2024 study on resource-constrained deep learning on Edge explains this explicitly: Deploying Convolutional Neural Networks (CNNs) - a type of deep learning algorithms suitable to analyze visual data - on resource-constrained edge devices is "practically infeasible due to their extensive parameter counts and floating-point operations." For this, innovative network pruning methods are necessary to compress the AI models and materialize CNNs on low-power edge devices.

Edge AI on Embedded Low-Power Devices Will Transform IoT

The ability to use artificial intelligence (AI) by performing machine learning (ML) on low-power, embedded devices that can run on batteries is a game changer for IoT. This so-called ‚Edge AI' is where everything is heading behind closed doors on countless product roadmaps that will come to market within the next few years.

It will push everything in IoT towards being much more intelligent, useful, and powerful, and enable new types of products and applications that were previously impossible.

What's transformed this is the development of embedded machine learning or ‚embedded ML.' This enables ML to be performed on constrained-computation devices in a highly simplified way. The industry gold standard for running advanced AI and ML algorithms on resource-constrained devices typical of the IoT is TensorFlow Lite.

In parallel, Nordic Semiconductor has constantly expanded the limits of what ‚resource constrained' means in battery-powered IoT with its multi-protocol Bluetooth LE nRF52, nRF53, and nRF54 Series Systems-on-Chips (SoCs) and nRF91 Series cellular IoT Systems-in-Package (SiPs).

All employ multiple Arm® Cortex® processors and large amounts of Flash and RAM memory, which, when married to embedded ML, suddenly make Edge AI a technological and commercial reality on embedded devices designed to be powered by batteries.

"You can engineer low-power embedded devices to run machine learning in such an optimized way you don't necessarily need to think dedicated ML accelerators," says Kjetil Holstad, EVP Strategy & Product Management at Nordic Semiconductor. "The key is to marry innovative engineering to maximum data processing and minimum power consumption."

"As Edge AI in IoT evolves, so will the demands that will be placed upon it, and so the future may well include the use of dedicated ML accelerators on ultra-low-power embedded devices with dedicated ML cores. But for now, we're demonstrating the power of optimization that can enable Edge AI today. Our efficient SoCs and SiPs showcase that you don't need dedicated ML accelerators to empower embedded ML."

What this means is that Nordic Semiconductor has successfully and uniquely broken the traditional tradeoff link between choosing either high computational power and high-power consumption, or low computational power and low power consumption. Now, you can have high processing power and low power consumption: the lowest power on the market for Edge AI and embedded ML.

As such, if your next-generation product or application doesn't incorporate AI and embedded machine learning, it risks being completely outclassed—even invalidated—by your competitors' products that do employ these technologies.

AI Adoption Challenges: Where Are We on the Human Front?

A significant obstacle to Edge AI adoption is the complex interplay of technical and organizational factors that influence decision-making and implementation. Organizational inertia, lack of expertise, and competing priorities within enterprises can impede the adoption and integration of Edge AI solutions. Furthermore, cultural resistance to change and legacy systems entrenched within organizational workflows present additional barriers to adoption.

To overcome such challenges, stakeholders must foster a culture of innovation and collaboration, ensuring alignment between business objectives and technological initiatives. Providing training and education programs to equip employees with the necessary skills and knowledge for Edge AI implementation can also facilitate adoption. Additionally, establishing clear governance structures and accountability mechanisms can help streamline decision-making processes and mitigate organizational barriers.

In recent years, numerous debates (Harvard, 2020, WEF, 2021, Georgia Tech, 2023) have emerged discussing the ethical issues of AI, including violations of privacy, perpetuation of bias, and social impact. Particularly speaking of bias in AI, this takes place when ML algorithms have the potential to reproduce and amplify pre-existing biases, which could lead to inequitable and unethical consequences. This should be addressed by ensuring accountability, transparency, and fairness in AI-based decisions by maintaining a balance between technological advancements and matters of morality and ethics, which encourages ethical innovation.

Chapter XI: The Future of Edge AI

In the forthcoming era, the escalating demand for AI applications characterized by real-time responsiveness, minimal latency, and stringent privacy measures will spur further proliferation of Edge AI implementations. These deployments are poised to become increasingly precise and effective, capitalizing on advancements in communication networks, developments in hardware and software, and breakthroughs in artificial intelligence.

"The maturation of Edge AI is crucial for realizing the promise of the Artificial Intelligence of Things (AIoT). We are approaching the holistic problem of improving the energy efficiency of silicon solutions and software for Edge acceleration, reducing the cost of model training and application development, and simplifying the customization and privacy for the user. This enables the scale necessary for the rapid proliferation of AI for our customers."

Nandan Nayampally CMO, BrainChip

Innovations in Edge AI

Researchers, open-source communities, leading research organizations, and early adopters will play a pivotal role in propelling the evolution of Edge AI, driving innovation, refinement, and further adoption in this domain. However, Edge AI does not stand alone. The trajectory of efficient Edge AI applications hinges on the progression of various supporting technologies tailored for edge devices and streamlined AI functionalities. The continuous evolution of these technologies will shape the future landscape of Edge AI. In turn, Edge AI's growth and maturation will influence various existing industries and maybe even contribute to creating new ones.

In this chapter, we explore the latest innovations in Edge AI and how they are poised to influence its adoption in the future. We also set the stage for new and upcoming developments in this space. Then, we discuss the key considerations companies need to assess Edge AI solutions and how different adopter types will impact the next decade of industry adoption of Edge AI.

Embracing the Latest Innovations in Edge AI

Arguably, one of the most popular AI technologies today is Generative AI. The integration of **Generative AI at the edge**, exemplified by large language models' (LLMs) convergence with edge computing, is unlocking new possibilities for content generation and personalization at the edge. As we discussed in detail in Chapter 9, in order to deploy LLMs on edge devices, their model size needs to be reduced. Techniques like pruning, quantization, and distillation help produce smaller models while maintaining high performance. TinyChat, NVidia IGX Orin Developer Kit, and Qualcomm's Stable Diffusion are all examples of serious efforts being made to bring Generative AI to the edge. By leveraging generative capabilities at the edge, organizations can deliver richer and more immersive user experiences while preserving data privacy and minimizing reliance on centralized cloud infrastructure.

Looking at the latest research and developments in Edge AI, we also find a lot of work being done on coming up with low-power, high-performance computing, such as neuromorphic computing and data-efficient AI. **Neuromorphic computing** mimics the structure and function of the human brain to enable efficient and intelligent processing at the edge. Realistically, neuromorphic chips consist of numerous artificial neurons and synapses, mirroring the behavior of brain spikes. These chips offer significant advantages for scaling Edge AI applications. They consume less power and offer faster processing speeds compared to traditional processors. Crucially, they equip Edge AI systems with human-like reasoning capabilities, which are highly beneficial for various applications like obstacle avoidance and robust acoustic perception.

BrainChip leads the way in neuromorphic computing with its fully digital designs, providing portability and reliability. Their Akida™ neural processor supports on-device learning, allowing for personalization and customization without cloud connectivity. With the addition of Temporal Event-Based Neural Nets (TENNs), BrainChip's technology achieves remarkable speed-ups in complex time-series data applications, enhancing efficiency while maintaining accuracy. Recognized by NASA for in-space autonomy and Mercedes Benz for automotive applications, BrainChip's innovation promises transformative solutions across diverse industries. As neuromorphic computing technology advances, it paves the way for a new era of AI-enabled edge devices capable of real-time learning and adaptation. Similarly, IBM's TrueNorth chip embodies neuromorphic principles, exhibiting remarkable energy efficiency while delivering cognitive capabilities suitable for Edge AI applications.

Additionally, strides in **data-efficient AI** are shaping the future of edge computing, enabling AI algorithms to operate effectively with minimal data requirements. Ongoing research on data-efficient AI explores many techniques, ranging from augmenting and using pre-trained models with domain knowledge (e.g., transfer learning) to paradigms that engage

humans in the data labeling processes (e.g., active learning) as part of human-AI interactions. These methods eliminate the need for extensive data collection and reduce the computational demands on Edge AI systems.

There are also popular data-efficient techniques that reduce the size of the AI model (e.g., model pruning) to enable space-efficient models that can fit on edge devices without compromising performance. Moreover, emerging approaches like one-shot learning and few-shot learning inherently enable models to learn from minimal data samples, further enhancing Edge AI's efficiency and effectiveness. A recent research study has found a way to convert large-scale natural language processing (NLP) models like Google's BERT model into formats that are tailored for deployment on resource-constrained edge devices. As such, they managed to convert a pre-trained and fine-tuned NLP model in Bert into a so-called MobileBERT, with a 160x reduction in footprint and a mere 4.1% reduction in accuracy. All this shows that the integration of such data-efficient AI techniques within Edge AI systems not only optimizes resource utilization but also ensures robust performance in resource-constrained environments.

Beyond advancements in computing capabilities, the **integration of Edge AI with 5G and emerging 6G networks** is poised to revolutionize connectivity at the edge. The ultra-low latency and high bandwidth offered by these networks enable real-time data processing and communication, unlocking new possibilities for applications such as autonomous vehicles and remote healthcare monitoring. For instance, Verizon's 5G Edge with AWS Wavelength facilitates the deployment of ultra-low latency applications by bringing compute and storage closer to the network edge, reducing round-trip time for data transmission.

The Future of Edge AI', capturing a futuristic integration of Edge Artificial Intelligence in a smart city environment (Created by DALL-E 3).

Another technology paradigm that is gaining traction is **distributed AI**, leveraging decentralized learning and collaborative intelligence to enhance edge computing capabilities. Projects like NVIDIA's Federated Learning Toolkit empower edge devices to collaboratively train AI models without sharing raw data, preserving data privacy while improving model accuracy. This distributed approach enables Edge AI systems to adapt dynamically to changing environments and diverse user needs, paving the way for more resilient and responsive edge applications.

Decentralized learning mechanisms like federated learning and swarm learning are revolutionizing Edge AI systems. While federated learning allows edge devices to collaboratively train a shared machine learning model without sharing raw data, swarm learning, inspired by collective behavior in nature, fosters decentralized and self-organizing AI systems. It facilitates information sharing among edge devices in a fully decentralized manner, enhancing Edge AI systems' adaptability and performance in real-time. These decentralized learning approaches promise scalable, efficient, and privacy-preserving solutions for future Edge AI deployments.

In addition to that, the emergence of **Cloud-Edge Hybrid solutions** is bridging the gap between centralized cloud computing and edge devices, offering a balance between scalability and low-latency processing. Companies like Microsoft Azure and Relay2 provide seamless integration between cloud services and edge devices, enabling hybrid AI solutions that leverage the strengths of both environments. This hybrid approach optimizes resource utilization and minimizes data transfer costs, facilitating the deployment of AI applications across distributed edge networks.

We are also seeing exciting advancements in **computer vision at the edge**, enabling real-time analysis of visual data and empowering Edge AI applications with contextual awareness and situational understanding. Platforms like Intel's OpenVINO toolkit enable efficient deployment of computer vision algorithms on edge devices, facilitating applications such as smart surveillance and industrial automation. By processing visual data locally, Edge AI systems can extract actionable insights in real time, enhancing decision-making and response capabilities.

Just as importantly, **natural interfaces**, such as voice commands, gestures, and facial expressions, are becoming increasingly integrated with Edge AI systems. These interfaces allow users to interact with devices intuitively and efficiently without the need for traditional input methods like keyboards or touchscreens. By leveraging Edge AI for on-device processing, natural interfaces can offer low-latency responses and enhanced privacy by minimizing data transmission to the cloud. Moreover, advancements in machine learning algorithms enable Edge AI systems to accurately interpret and respond to various natural inputs, enhancing user experience and expanding the usability of Edge AI across diverse applications and industries.

Future of Edge AI: A New World of Efficiency and Sustainability

The rapid evolution of Edge AI and related applications will continue to be fueled by two powerful driving forces. On the one hand, "bottom-up" technological advancements will gain further momentum in almost all directions:

- Devices
- Advanced hardware
- Sophisticated, compact, and capable models
- Robust, easy-to-use, and easy-to-deploy software and tools

As for "top-down" approaches, there will be more applications, more developers, and use cases, resulting in more adopters of Edge AI in all verticals. This will eventually lead to a "new world with trillions of intelligent devices enabled by tinyML/Edge AI technologies that sense, analyze, and autonomously act together to create a healthier and more sustainable environment for all – a vision of the tinyML Foundation.

On the technology side, in addition to currently used digital accelerators, novel HW architectures will find their way into products. Some examples of such technologies under development nowadays include compute-in-memory, analog compute, and event-driven/neuromorphic architectures. Their advancement will lead to even more energy efficiency (e.g., 10-100x better than the state-of-the-art today) yet with more compute horsepower and on-chip resources.

Natural progression in algorithms and ML models will yield smaller yet more sophisticated and multi-modal models, including transformers, without sacrificing accuracy. These innovations will bring more capabilities to Edge-AI-powered devices without compromising their cost, power, or form factor. Software tools are becoming more robust, universal, and easy to use to the point that no special coding training will be required for developers, engineers, and end users to utilize them. The same will be true for deployment tools to connect and orchestrate many connected Edge AI devices with built-in security.

Devices will become smarter with on-device learning capabilities at the individual and network levels with the help of approaches like federated learning. In particular, there will be continuous progress in the supporting and "peripheral" technologies such as low-power sensors and imagers of all kinds, low-power connectivity technologies, memory technologies, and battery technologies for portable devices, which will assist product developers in designing better and smarter products, enabled by Edge AI.

On the application front, following early adopters of Edge AI in the past several years, Edge AI will find its way into the mainstream, penetrating more verticals, applications, and use cases, including industrial and consumer devices. Some examples can be found in industrial IoT, healthcare/wellness, consumer electronics, and smart everything. Energy efficiency, privacy, low cost, robustness, and ease of deployment will drive such adoption.

Edge AI Adoption Levels Shaping the Future

As organizations integrate Edge AI into their operations, critical considerations emerge. These factors will play a key role in shaping the success and effectiveness of Edge AI deployments in the years to come, ensuring they align with organizational goals and meet evolving technological requirements. Below, we outline six key considerations for technology leaders assessing Edge AI solutions.

Consideration	Description
Open Architecture	An open and vendor-neutral architecture that supports a mix of edge computing technologies, devices, software stacks, and networking solutions
Security and Privacy	Implementation of robust security measures, including encryption, access controls, and adherence to a zero-trust security framework
Scalability and Flexibility	The ability of the chosen Edge AI platform to scale seamlessly and adapt to changing enterprise needs across diverse usage scenarios
Edge Device Capabilities	Evaluation of edge devices' processing power, storage, and connectivity to meet AI application performance requirements and ensure ease of deployment and management at scale
Interoperability	Integration with existing systems and compatibility with diverse devices to facilitate a smooth transition and maximize edge AI benefits
Data Governance and Compliance	Establishment of robust data governance policies and compliance with relevant regulations, addressing data ownership, consent, and adherence to industry standards

Based on these considerations, companies across different industries will be able to assess new technologies and innovations and decide on what and how to adopt. Based on Accenture's research on the adoption of edge computing and AI, they identified four main enterprise approaches to edge:

1. **Type 1: Ad Hoc** – Centralized IT-led edge deployers
2. **Type 2: Tactical** – Specific-need adopters of pre-packaged solutions
3. **Type 3: Integrated** – Integrates with cloud and scales widely
4. **Type 4: Super Integrated** – Ties edge to business in transformative adoption

These approaches correspond to how companies are integrating edge into the digital core of their business. They are predominantly led by factors such as strategic edge implementation, enterprise-level scalability, and technology maturity. Based on their survey results, they found out that super integrated organizations (Type 4) are seeing the most success, in part because "they build edge on their digital core, integrating it with cloud, data, AI, and interoperable applications and platforms. Yet, this was not the most common approach in the survey; it was, in fact, the least common with a mere 6% of edge adopters.

Almost half of the edge adopters surveyed took the integrated approach (Type 3), with 79% aiming for full edge-cloud integration in the next three years. This shows that the intention to incorporate edge into their business is there, but they still need further strategic expansion for enterprise-wide implementation and more investment in talent and partnerships. This leaves 20% of edge adopters taking the tactical approach (Type 2), with only 28% of them integrating their edge strategies into their cloud strategies (partially or fully). The ad-hoc approach adopters constitute the remaining 30%, which shows that there's still a lot of room for Edge AI to find further adoption avenues and better ways to get integrated into business strategies and workflows. For this, Accenture built a three-step framework that can enable adoption across all levels:

1. **Strategize for edge:** As they state in their report, "Approach edge as a foundational capability, not as a bolt-on."
2. **Scale across the enterprise:** Integrate edge on the back of cloud across the whole organization using enterprise data and AI applications.
3. **Strengthen capabilities:** Ensure every employee and process is well-prepared for edge. This requires appropriate classification of job roles, verified qualifications, relevant training, and opportunities for career growth within a learning environment based on trust.

Following this, we expect to witness significant developments and adoption rate increases in many industries, primarily healthcare, retail, and energy. Needless to say, the manufacturing sector and the automotive sector will continue to expand their adoption rate, but it is safe to say that Edge AI has already reached a ubiquitous level in these two industries. By integrating Edge AI with other technologies and strategies, companies and service providers will have the ability to shape the future of their organizations and, eventually, their industries.

Shaping the Future of Industry with WiFi-enabled Edge AI

The future of industry hinges on transformative technologies like Edge AI, promising efficiency and innovation across sectors. By harnessing AI at the network's edge, businesses can unlock productivity and agility, driving real-world impact. This is further enhanced by leveraging other technologies that can boost Edge AI's application and impact.

WiFi-enabled Edge AI from Relay2 exemplifies this potential. By integrating Wi-Fi technologies, Relay2 amplifies Edge AI's capacity for real-time analysis, enabling swift responses to various scenarios while optimizing bandwidth usage for enhanced efficiency. Such a convergence of technologies holds promise for manufacturing, healthcare, transportation, agriculture, education, and retail. It marks a shift in data handling, enabling real-time insights and autonomous decision-making.

Looking at manufacturing, for instance, Relay2's WiFi-enabled Edge AI solution is ushering in a new era of efficiency and safety. At its core, the WiFi Edge AI Service Point application provides real-time video analytics for monitoring and alerting, enabling real-time responses to anomalies and potential hazards within manufacturing facilities.

A standout feature of Relay2's solution is its ability to replace traditional IoT sensors with intelligent cameras equipped with Edge AI capabilities. These cameras play a critical role in measuring machine metrics and detecting operational issues in Surface Mount Technology (SMT) and Assembly lines. By integrating Edge AI into these cameras, manufacturing processes are streamlined, costs are reduced, and operational efficiency is significantly enhanced. Additionally, such an application can minimize irrelevant data from video surveillance or IoT sensors, thereby reducing network bandwidth usage and enhancing operational efficiency and cost-effectiveness for Cloud AI operations.

Relay2's solution also addresses security and privacy concerns by processing and analyzing data locally at the edge, reducing the transmission of sensitive information to the cloud. This safeguarding of confidential data, including video images and IoT sensor data, mitigates the risk of unauthorized access or data breaches. Keeping with the manufacturing example, Relay2's solution boasts adaptability to diverse manufacturing environments. Whether operating in closed environments with limited internet connectivity or facing bandwidth constraints, manufacturers can rely on Relay2's all-in-one infrastructure. It ensures seamless integration and hassle-

free installation through a reliable network (independent of internet condition), AI computing capabilities, and easy-to-access storage.

Harnessing the power of AI-driven analytics at the edge, Relay2 facilitates data-driven decision-making, predictive maintenance, and process optimization, propelling manufacturing through the era of Industry 4.0. Technologies like WiFi-enabled Edge AI represent a pragmatic leap forward in industrial innovation. By seamlessly integrating Wi-Fi technologies with Edge AI capabilities, Relay2 empowers businesses to enhance efficiency, safety, and cost-effectiveness across diverse sectors, just like in manufacturing. As industries embrace this convergence, they unlock new possibilities for real-time monitoring, proactive decision-making, and streamlined operations.

Edge technology stands out as a game-changer, offering the capability for instant decision-making and personalized experiences for both customers and employees. The widespread adoption of decentralized data processing, whether in educational settings, manufacturing facilities, or wearable devices, is expected to occur rapidly and visibly. Many early adopters and those on the brink of implementing Edge AI anticipate its transformative potential, predicting the emergence of new business models or the evolution of existing ones within the next few years. This trend suggests that edge computing will soon become a global phenomenon, with innovation hubs expanding beyond traditional tech centers like Silicon Valley.

Notably, edge technology has already found applications in space missions, highlighting its versatility and potential impact across diverse sectors. Now is the opportunity for businesses across all industries to embrace edge computing to drive innovation, a critical driver of progress. While challenges related to data architecture and talent acquisition may arise, a strategic approach that aligns edge initiatives with broader business goals, integrates seamlessly with existing digital infrastructures, and leverages partnerships and expertise can facilitate rapid and cost-effective innovation, laying the groundwork for a brighter future.

Report Partner

tinyML Foundation

Los Altos, CA
IT Services and IT Consulting

tinyML Foundation is a non-profit professional organization focused on supporting and nurturing the fast-growing branch of ultra-low power machine learning technologies and approaches dealing with machine intelligence at the very edge of the cloud. These integrated “tiny” machine learning applications require “full-stack” (hardware, system, software, and applications) solutions including machine learning architectures, techniques, tools, and approaches capable of performing on-device analytics. A variety of sensing modalities (vision, audio, motion, environmental, human health monitoring, etc.) are used with extreme energy efficiency, typically in the single milliwatt (and below) power range, to enable machine intelligence right at the boundary of the physical and digital worlds. We see a new world with trillions of distributed intelligent devices enabled by energy efficient machine learning technologies that sense, analyze, and autonomously act together to create a healthier and more sustainable environment for all To enable this vision, tinyML Foundation is: + Growing a diverse global community of hardware, software, and system scientists, engineers, designers, product management, and businesspeople. Engaging experts and newcomers alike in developing leading edge ultra-low power machine learning. + Promoting and stimulating open knowledge exchange between researchers and industry to accelerate the field ahead. + Inspiring the capabilities of ultra-low power machine learning and demonstrating the potential of machine intelligence and data analytics at the very edge of the physical and digital world. + Connecting technologies and innovations to enormous product and business opportunities creating value across the whole ecosystem and within industry verticals.

www.tinyml.org

Sponsors

Renesas

Koto-ku, Toyosu, Tokyo
Semiconductor Manufacturing

Renesas scalable product portfolio comprising 16Bit, 32Bit, and 64Bit MCUs and MPUs, together with the rich ECO system and development infrastructure for embedded systems and Edge AI/ML solutions, is a perfect match for system solution packages. It targets and enables fast prototyping, evaluation, development, and deployment of your next embedded edge AI/ML solution.

www.renesas.com

Image credit: Renesas

Synaptics

San Jose, California
Semiconductors Manufacturing

The infusion of AI into IoT edge devices can realize an intelligent, context-aware environment. This is possible only when intelligent IoT islands are knit securely, reliably, and cost-effectively with the right compute solutions and right wireless connectivity at each node. That is why Synaptics, an IoT company dedicated to changing how we interact with technology, has introduced Synaptics Astra™, the AI-native compute platform for the IoT.

Astra is an approach to IoT edge device design based on scalable AI-native hardware, unified software, an adaptive AI framework, a partner-based ecosystem, and seamless and robust wireless connectivity. This enables developers at any stage of their AI journey—from beginner to expert—to deploy AI at the edge at the power, performance, reliability, security, and cost points required to participate successfully in a rapidly unfolding future.

With its multi-generational experience across key IoT functions and applications, Synaptics is accelerating this fundamental shift. Your AI journey begins with the first Astra processors, the SL-Series (SL1680, SL1640, and SL1620) MPUs. Supported by the Astra Machina™ Foundation Series development kit, the SL-Series comprises highly integrated Linux™ and Android™ SoCs optimized for multi-modal consumer, enterprise, and industrial IoT workloads and features hardware accelerators for edge inferencing, security, video, graphics, and audio.

The SL-Series provides high-performance and cost-effective solutions for developers of smart home, conferencing, retail, and industrial applications where secure AI-based processing can deliver an enhanced user experience with low cost of ownership.

The SL-Series will be followed by the SR-Series of high-performance AI-native MCUs with intelligent tiered inferencing to ensure optimized context-aware computing from the home and office to the factory floor.

www.synaptics.com

synaptics

Image credit: Synaptics

Arduino

Turin, Italy
Appliances, Electrical, and Electronics Manufacturing

Arduino is the leading open-source hardware and software company in the world, with a community of over 33 million active users. Born to provide an easy-to-use platform for anyone creating interactive projects, Arduino has grown and adapted to new needs and challenges, branching out into solutions for IoT, wearables, 3D printing, and embedded environments.

Today Arduino offers an end-to-end ecosystem that includes production-ready certified hardware, user-friendly software tools such as the Arduino IDE, a plethora of libraries and sketches ready to use, and Arduino Cloud services. This complete offering, combined with a strong mission to make technology accessible to anyone, makes Arduino the ideal ally for innovators in any field – be they hobbyists, designers, educators, or engineers in any industry.

Indeed, the Arduino Pro range was launched in 2020 to bring professional users the best of two worlds: simplicity of integration and absence of vendor lock-in on the one hand, and scalable, secure, and widely supported solutions on the other. Going beyond the concept and rapid prototyping phase Arduino has always been known for, new products are constantly being added to the ecosystem with the high performance and industrial-grade quality required by mass production and the most demanding fields, from high-speed manufacturing equipment to airport security.

Reducing non-recurring engineering costs, accelerating time to market, and bridging the tech skills gap, Arduino lowers the barriers to innovation by democratizing technology, enabling everyone to solve problems, create value, and grow.

www.arduino.cc

Photo by: Alberto Morici for Arduino

Nordic Semiconductor

Trondheim, Norway
Semiconductor Manufacturing

Nordic Semiconductor stands at the forefront as a complete provider of low power wireless connectivity solutions, offering end-to-end services. Renowned as the leading provider of Bluetooth Low Energy, we are an emerging leader in cellular IoT, DECT NR+, Wi-Fi, Matter, Thread, Zigbee, and power management.

Established in Trondheim in 1983, Nordic has grown to approximately 1500 employees in over 20 countries.

Our innovative low-power wireless solutions connect millions of IoT devices worldwide, enhancing lives by enabling smarter, safer, and more sustainable solutions. Nordic's cutting-edge hardware, software, and development tools rank among the most advanced globally.

Our solutions are trusted by the world's leading brands across various products and applications, from Asset Tracking and Industrial Lighting to Health Care, Audio, and Smart

Agriculture. Nordic is a proud member of the ANT+ Alliance, Bluetooth SIG, Thread Group, Connectivity Standards Alliance, Wi-Fi Alliance, and GSMA.

Recognizing the importance of AI and ML in shaping the future of IoT, Nordic excels in meeting the demand for high processing power and low power consumption. We take pride in offering a comprehensive developer experience, delivering ultra-low power wireless solutions that are easy to implement and backed by world-class support.

www.nordicsemi.com

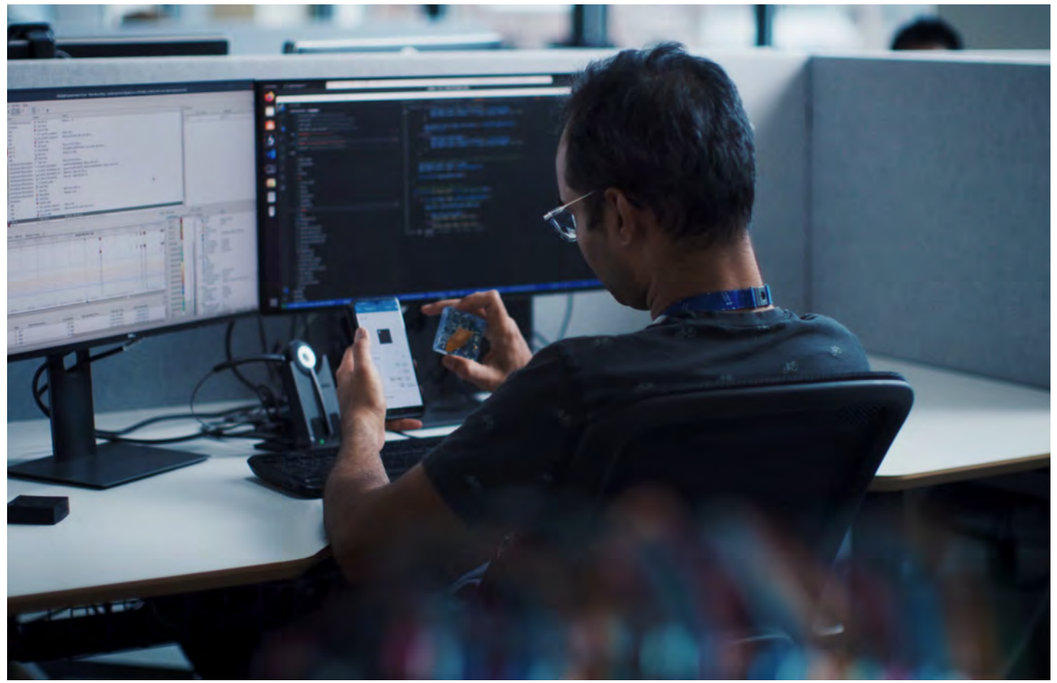


image credit: Nordic Semiconductor

Syntiant

Irvine, California
Semiconductors

Syntiant is making Edge AI a reality with providing the highest performance processors for constrained environments along with hardware agnostic machine learning models that run on most any processor from the smallest DSPs to the largest GPUs. Whether you need the most performance in the smallest energy footprint or already have a processors and need top performing machine learning model, Syntiant has a solution for you.

Syntiant tackled this fundamental challenge by developing custom at-memory deep learning processors that delivers best-in-class performance, while meeting size, power and cost constraints. We streamlined the conversion of raw and synthetic data into quality machine learning models, while providing a training pipeline for optimizing edge applications. We accelerate models ranging from the smallest event detectors all the way up to large language models (LLMs).

Over fifty (50) million products are deployed using Syntiant technology, bringing intelligent edge processing to audio, video, speech and sensors. Our smallest design is in a hearing aid and our largest is in an automobile. We are deployed in consumer, mobility, industrial, security, health and defense applications.

Our technology can equip almost any device with powerful deep learning capabilities, real-time data processing, and decision making with near-zero latency, at orders of magnitude lower power than traditional solutions.

Whether it is an acoustic event detector for security applications, advanced video processing in a teleconferencing device, or real-time monitoring of battery health, we provide developers with proven hardware and software solutions that can take you from concept to product deployment in the shortest time possible.

www.syntiant.com

SYNTIANT

Mouser Electronics

Mansfield, Texas
Semiconductor Manufacturing

Mouser Electronics is a worldwide leading authorized distributor of semiconductors and electronic components from over 1,200 manufacturer brands, with local sales and service centers located around the globe. We specialize in the rapid introduction of new products and technologies for design engineers and buyers. Our extensive product offering includes semiconductors, interconnects, passives, and electromechanical components.

In 2007, Mouser became a part of the Warren Buffett Berkshire Hathaway family of companies. Today, Buffett's holdings include insurance and finance subsidiaries and a host of almost fifty businesses ranging from jewelry and furniture to manufactured homes.

Mouser has a strong commitment to customer service. That's why we've won awards for our legendary worldwide customer service excellence. We understand the value of having a knowledgeable person there to answer your questions quickly. Mouser is redefining customer-focused distribution.

www.mouser.com

Axelera

Eindhoven, Netherlands
Semiconductor Manufacturing

Axelera AI is providing the world's most powerful and advanced solutions for AI at the edge. Our game-changing Metis™ AI platform – a holistic hardware and software solution for AI interference at the edge – enables computer vision and future generative AI applications to become more accessible, powerful and user friendly than ever before.

The core of our platform is our Metis AI Processing Unit (AIPU), which is based on proprietary digital in-memory computing technology (D-IMC) and RISC-V controlled dataflow technology. The AIPU offers industry-leading performance, usability, and efficiency at a fraction of the cost of existing solutions. Our technology is scalable and outperforms any other startup or incumbent. Additionally, our embedded security engine protects data and information through sophisticated encryption, ensuring the security of sensitive biometric data.

Our technology is integrated into AI acceleration cards, boards, and vision-ready systems. This enables small to medium-sized enterprises to speed up adoption and streamline field installation. Our products will be available worldwide – and we're sampling to a select group of customers until general availability – making AI accessible to any developer in the world.

Developed from the ground together with our Metis AIPU, our click-and-run Voyager SDK software stack provides easy-to-use and user-friendly neural networks initially for computer vision applications to software developers aiming to integrate AI into their devices.

The Voyager SDK automatically quantizes and compiles neural networks that have been trained on different frameworks – so you don't need to retrain – generating code that runs on the Metis AI platform with industry-leading accuracy. Optimized networks running on Metis AIPU are indistinguishable from those running on systems with floating-point units.

Headquartered in the AI Innovation Center of the High Tech Campus in Eindhoven, The Netherlands, Axelera AI has R&D offices in Belgium, Switzerland, Italy and the UK, with more than 140 employees in 15 countries. Our team of experts in AI software and hardware hail from top AI firms and Fortune 500 companies.

Axelera AI is providing the world's most powerful and advanced solutions for AI at the edge. Its game-changing Metis™ AI platform – a holistic hardware and software solution for AI interference at the edge – enables computer vision applications to become more accessible, powerful, and user-friendly than ever before.

Headquartered in the AI Innovation Center of the High Tech Campus in Eindhoven, The Netherlands, Axelera AI has R&D offices in Belgium, Switzerland, Italy, and the UK, with more than 125 employees in 15 countries. Its team of experts in AI software and hardware hail from top AI firms and Fortune 500 companies.

www.axelera.ai

OKdo

London, United Kingsom
Appliances, Electrical, and Electronics Manufacturing

OKdo, a global technology company within the RS Group plc, is disrupting the landscape of single board computers (SBCs) and AIoT segments. Offering a distinctive blend of hardware, software, development support, and manufacturing services to empower customers to transform their innovative ideas into industrial and commercial reality. They are on a global mission to deliver the imagination, creativity, and technical expertise forward-thinking engineers crave.

Whether you're a newcomer to the world of SBCs or an industrial designer seeking to push the boundaries of innovation, OKdo is here to propel you forward. Partnering with industry giants like NVIDIA, Arduino, Synaptics, ASUS and BeagleBone, as well as tech powerhouses such as Arm, NXP, Rockchip and Intel, OKdo offers unparalleled access to cutting-edge technology. Our collaboration extends to rising tech startups like OStream, Useful Sensors and LAIIER, ensuring our customers have access to the latest solutions and ideas.

More than a hardware provider, OKdo offers a comprehensive solution hub. Value-added services such as manufacturing support, rapid prototyping, and personalized design configurations are just the start. OKdo's end-to-end solutions simplify your journey from concept to market, unlocking accelerated growth and success.

Whether you're just starting your SBC and IoT journey or looking to scale your operations, OKdo provides everything you need to thrive and succeed.

Visit www.okdo.com to learn more and embark on your journey of innovation.

Brainchip

Laguna Hills, California
Computer Hardware Manufacturing

BrainChip is a leader in edge AI on-chip processing and learning. The company's first-to-market, convolutional, neuromorphic processor, Akida™, mimics the event-based processing method of the human brain in digital technology to classify sensor data at the point of acquisition, processing data with unparalleled energy-efficiency and independent of the CPU or MCU with high precision. On-device learning that is local to the chip without the need to access the cloud dramatically reduces latency while improving privacy and data security. In enabling effective edge computing to be universally deployable across real-world applications, such as connected cars, consumer electronics, and industrial IoT, BrainChip is proving that on-chip AI is the future for customers' products and the planet.

www.brainchip.com

brainchip
Essential AI

Relay2

Milpitas, CA
Computer Networking Products

Relay2 has developed a unique Wi-Fi access and edge computing platform designed to enable Mobile Network Operators, Wi-Fi Carriers and Managed Service Providers to create and monetize horizontal and vertical business applications and services on top of a high-performance, cloud-managed Wi-Fi network. Relay2 provides the only Wi-Fi network solution with the ability to host rich multimedia content like video and presentations as well as custom Applications directly on their Access Points. This approach eliminates latencies and lowers backhaul traffic requirements by delivering content directly from the AP to the connected Wi-Fi device.

Relay2 Inc. was founded in 2011 by technology leaders with backgrounds at Cisco, Juniper Networks, Nokia and Siemens, and is backed by strategic investors public company. The Relay2 headquarter is located in Milpitas, California with branch offices in Japan, Taiwan, China and Europe.

www.relay2.com

Relay2

Imagimob

Stockholm, Sweden
Software Development

Imagimob is a company driving innovation in Edge AI/Machine Learning (ML), and helping customers create the intelligent products of the future. Based in Stockholm, Sweden, Imagimob has served global customers within the automotive, manufacturing, healthcare, and lifestyle industries since 2013. Imagimob Studio was launched in 2020 as a development platform enabling the swift, easy, end-to-end development of Edge AI applications for devices with constrained resources.

Since 2020, Imagimob Studio has seen significant improvements, including launching the Graph UX interface. Graph UX visualizes the ML modeling process in a unique way, making it easier for users to create high-quality models. It also provides model explainability, giving users the ability to zoom in and take a close look inside models they are building throughout the entire development process. This explainability, and resulting transparency, are something Imagimob believes in strongly; the data policy, where customers maintain the right to the data they use in Imagimob Studio, is just one way Imagimob follows through on this belief. Through this and other features, Imagimob Studio guides and empowers users throughout the entire development journey, resulting in game-changing productivity and faster time-to-market.

In late 2023, Imagimob launched a new product line, Ready Models. These production-ready models enable companies to add Edge AI features into their products without spending the time or cost required to develop their own. They even open up the possibility of using Edge AI features for companies that have no internal ML know-how.

Tirelessly dedicated to staying on top of the latest research and finding new ways to deliver the best ML models possible, the experienced Imagimob team is always thinking new, and thinking big. In May 2023, Imagimob became part of Infineon Technologies AG. Together with Infineon's ModusToolbox™, Imagimob truly enables the entire ML journey, from data collection through deployment onto products. And with the recent launch of Infineon's PSoC™ Edge, Imagimob is poised to deliver Edge AI capabilities never before possible.

www.imagimob.com

Additional Contributors

Eta Compute

Sunnyvale, California
Software Development

Eta Compute is at the forefront of empowering edge-AI innovation through groundbreaking solutions that overcome the gap between the fast-moving landscape of AI and the unique attributes of embedded systems. Founded in 2015, Eta Compute is fueled by a team of experts with AI, IoT, and systems design DNA who understand the challenges of deploying machine learning models on resource-constrained edge devices.

Recognizing the immense potential of edge-AI and the challenges hindering its realization, we developed Aptos, a revolutionary SaaS no-code toolchain engineered for embedded inference. Designed to streamline edge-AI model development and deployment, Aptos accelerates the creation of efficient models tailored for low-power edge processors.

What sets Aptos apart is its deep learning into the AI capabilities and performance of edge-AI silicon solutions. By harnessing these insights and abstracting away the hardware details, the Aptos toolset empowers developers to focus on innovation rather than technical intricacies. With Aptos, developers create better models, faster.

www.etacompute.com

Qualcomm

San Diego, CA
Telecommunications

Qualcomm is enabling a world where everyone and everything can be intelligently connected. Our one technology roadmap allows us to efficiently scale the technologies that launched the mobile revolution – including advanced connectivity, high-performance, low-power compute, on-device intelligence and more – to the next generation of connected smart devices across industries. Innovations from Qualcomm and our family of Snapdragon platforms will help enable cloud-edge convergence, transform industries, accelerate the digital economy, and revolutionize how we experience the world, for the greater good.

www.qualcomm.com

Sony

Tokyo, Japan
Technology

Sony stands at the cutting edge of technology, pioneering in the transformative field of Edge AI. This innovative approach, where AI processing is executed at the very edge of the network, right where the data is generated, is revolutionizing how we interact with technology. By reducing latency, minimizing bandwidth use, and prioritizing privacy, Sony's Edge AI technologies ensure smarter, faster, and more efficient operations.

At the heart of Sony's Edge AI advancements are its state-of-the-art smart sensors and AI chips. These powerful innovations are designed to enhance device intelligence, enabling them to make real-time decisions without relying on cloud connectivity. Sony's image sensors, for example, are redefining photography and videography, offering unparalleled autofocus, object recognition, and scene analysis directly on devices. This leap in technology not only enhances user experience but also opens new vistas in camera technology.

Moreover, Sony's commitment to innovation extends to its AI chips, tailored for edge computing. These chips strike the perfect balance between computational power and energy efficiency, essential for powering the next generation of robotics, IoT, and smart devices.

In the realms of entertainment and gaming, Sony leverages Edge AI to create immersive, personalized experiences that respond more intuitively to user interactions. This commitment to enhancing user engagement through technology makes Sony not just a leader in the industry but a visionary, redefining the boundaries of what's possible with Edge AI.

www.sony.com

Leopard Imaging

Fremont, California
Embedded Hardware

Leopard Imaging is a global leader that provides high definition embedded cameras and AI-based camera solutions—focusing on core technologies that improve image processing in autonomous vehicles, drones, IoT, robotics, and healthcare devices. Leopard Imaging works closely with Nvidia, Intel, Xilinx, Qualcomm, Sony, ON Semiconductor, OmniVision, and STMicro sensor companies with Original Equipment Manufacturer and Original Design Manufacturer services. Leopard Imaging serves Microsoft, Google, Amazon, Facebook, Zoox, Daimler, Boston Dynamics, and other well-established companies and organizations.

www.leopardimaging.com

About Wevolver

Wevolver is a global platform and community used by engineers to stay up to date about the latest technologies.

On Wevolver, professional engineers access informative and inspiring content such as articles, videos, podcasts, and reports on robotics, aerospace, semiconductors, advanced manufacturing, and state-of-the-art technologies.

The content on Wevolver is published by tech companies, universities, and individual community members. Next to that, Wevolver collaborates with dozens of technical content creators to develop content for customers and publish that on Wevolver.com.

Every month, millions of engineers leverage Wevolver to stay up to date, find knowledge when they are developing products, and to make meaningful connections with each other and the industry.

Wevolver has won the SXSW Innovation Award, the Accenture Innovation Award, and the Top Most Innovative Web Platforms by Fast Company.

Wevolver is how today's engineers stay cutting edge.

wevolver.com

WEVOLVER